Computational Analysis of Genetic Variation.

Matthew Arnell Field

Submitted December 2015

A thesis submitted for the degree of Doctor of Philosophy of

The Australian National University

Declaration

The thesis is the original work of Matthew Field. The thesis by compilation consists of six publications largely describing software and custom analyses I have done during my candidature. The common theme of my thesis is developing computational methods able to better elucidate the often-complex link between genetic variation and disease.

As of May 2016, all manuscripts have been published. For the six publications presented in detail (chapters 2-7), my specific contribution to each manuscript is detailed in the subsequent pages in the form of a statement signed by all other first and senior authors for each manuscript. Of the six publications, I am either first or senior author on every publication except *"Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models"*. In this instance, its inclusion is justified due to my significant contribution to this body of work, spending one year of software development converting a prototype pipeline into a production ready, high-throughput mouse exome pipeline described in the paper. To date, the high-throughput version of this pipeline has processed over 3000 mouse exomes.

The final section of my thesis (chapter 8) consists of eight additional publications describing significant discoveries about the genetic mechanisms of disease. In all publications, I am included as an author and the software described in the earlier chapters was utilised.

Publication "*Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies*". I designed and performed all the analyses described and also wrote the manuscript.

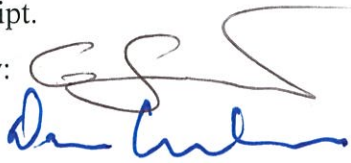
i) Senior author #1 Chris Goodnow:

ii) Senior author #1 Dan Andrews:

Publication "*Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees*". I designed and implemented the software tools described and also wrote the manuscript.

i) Senior author #1 Chris Goodnow:

ii) Senior author #1 Dan Andrews:



Publication "*Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-calling algorithms for human melanoma genomes".* I performed all bioinformatics analyses described, designed and implemented the highthroughput melanoma analysis pipeline descibed, and wrote the corresponding manuscript section. **Also, while the final draft erroneously omitted the joint first author symbol, a correction will be made in the January edition of Pathology**.

i) Senior author Richard Scolyer:

30 November 2015

ii) Joint first author James Wilmott:

Dr James Wilmott

01 December 2015

Publication "*Comparison of predicted and actual consequences of missense mutations*". For this publication, I was involved in the design and implementation of the bioinformatics analyses and contributed significantly to the manuscript."

i) Senior author #1 Chris Goodnow:

ii) Senior author #1 Dan Andrews:

iii) Joint first author Lisa Miosge:

Publication "*DeepSNVMiner: A sequence analysis tool to detect emergent, rare mutations in sub-sets of cell populations*". I designed and implemented all the software tools described and co-author Andrews and I wrote the manuscript"

i) Joint senior author #1 Chris Goodnow:

ii) First author #1 Dan Andrews:

Publication "*Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models*". While I am not first or senior author, I made significant contributions to this body of work, spending one year developing the high-throughput mouse exome pipeline from an earlier prototype pipeline

i) Senior author Chris Goodnow:

ii) First author Dan Andrews:

Abstract

High throughput sequences are generating increasingly detailed catalogues of genetic variation both in human disease and within the larger population. To effectively utilise this rich data set for maximum research benefit, as a discipline we require robust, flexible, and reproducible analysis pipelines capable of accurately detecting and prioritising variants. While data-specific computational algorithms aimed at deriving accurate data from these technologies have reached maturity, two major challenges remain in order to realise the goals of elucidating the underlying genetic causes of disease as a means of developing custom treatment options. The first challenge is the creation of high-throughput variant detection pipelines able to reliably detect sample variation from a variety of sequence data types. Such a system needs to be scalable, flexible, robust, highly automated, and able to support reproducible analyses in order to support both default and custom variant detection workflows. The second challenge is the effective prioritisation of the huge number of variants detected in each sample, a task required to reduce the large search space for causal variants down to variant lists suitable for manual interrogation. This thesis describes six publications describing components of the larger informatics framework I have developed over the last four years to address these challenges, a framework designed from the onset to effectively manage and process large data sets with an end goal of utilising computational analysis of sequence data to further understand the relationship between genetic variation and human disease. The first publication "*Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies*" describes a variant detection strategy designed to minimize false negative variants as is desired when utilising patient variation data in the clinic. The next four publications describe custom workflows developed for detecting variants in sequence data from different sample types, namely paired cancer samples ("*Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-calling algorithms for human melanoma genomes*"), pedigrees ("*Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees*"), mixed cell populations containing ultra-rare mutations ("*DeepSNVMiner: A sequence analysis tool to detect emergent, rare mutations in sub-sets of cell populations*") and mouse exome data containing ENU mutations ("*Massively parallel sequencing of the mouse exome to accurately identify rare,*

induced mutations: an immediate source for thousands of new mouse models") . The last publication, "*Comparison of predicted and actual consequences of missense mutations*" focuses on the validation of computational tools that predict functional impact of missense mutations and further attempts to explain why many missense mutations predicted to be damaging do not result in an observable phenotype as might be expected. Collectively these publications detail efforts to reliably detect and prioritise variants across a wide variety of data types, efforts all based around the significant underlying software framework I have developed to better elucidate the link between genetic variation and disease.

Acknowledgements

I wish to acknowledge the guidance and expertise of my three supervisors, Chris Goodnow, Dan Andrews, and Alistair Rendall. I would also like to acknowledge the help of Anna Cowan, Wendy Riley, and Charani Ranasinghe for guiding me through the PhD by publication process.

For my family, I would first like to thank my parents for their positive support, reassurance, and ongoing belief in me. I also want to thank my children Sedona and Carter for putting up with a little less Dad time and for being amazing about everything.

Finally (and most importantly) I wish to thank my partner Krista. She made this PhD possible with her constant encouragement throughout and her hard work behind the scenes. Without Krista, this PhD never, ever, ever would have happened.

Table of Contents

Chapter 1: General Introduction

1.1 Aims of the Thesis

This thesis describes six publications detailing aspects of the larger informatics framework I developed over the last four years to reliably detect and prioritize variants from next generation sequencing (NGS) data. The framework addresses two main challenges in the field; first how to manage and process increasingly large volumes of data from a variety of sequence data types and second how to prioritize the huge number of variants detected in each sample. Collectively, this framework aims to further elucidate the often-complex relationship between genetic variation and disease.

1.2 Overview

The study of genetic variation and its effect on human disease has undergone a major transformation in the last ten years. This has occurred due to a variety of factors including the arrival of commercially available NGS technologies, the generation of the first human reference genome (1), and the growing databases of known variation in both healthy individuals (2) and individuals with disease (3, 4). With over 250,000 human genomes sequenced to date and a doubling of this number expected in the next year alone (5), health sciences has truly entered the age of big data (6). While these exciting developments open many new research avenues, it is apparent that numerous computational challenges exist in the field that are yet to be resolved. For example, there currently is no dominant software framework able to manage and analyse large volumes of diverse sequence data nor is there a gold standard methodology for identifying disease-causing variants within the large background variation present in each individual.

1.3 Gene-disease link

Historically, the link between genes and disease has long been theorised with Archibald Garrod in 1908 coining the term “inborn errors of metabolism” to describe the increased incidence of alkaptonuria in consanguineous families which he suspected were caused by “transmissible elements within family” (7).

This understanding eventually led to the development of the first genetic linkage maps and linkage analysis methods, giving researchers for the first time a means of mapping the location of these elements responsible for inheritable diseases. During the twentieth century, increasingly complex linkage analysis methods capable of detecting specific genetic regions that co-segregate with affected family members were developed and refined eventually becoming a standard means of detecting the genomic regions linked to inherited disorders. Such approaches proved extremely successful mapping over 1000 monogenic diseases to date (8). Despite these successes, due to a variety of limitations in this approach, the cause of many diseases remained unknown particularly when studying complex/late onset disease as well as diseases with incomplete penetrance.

1.4 DNA sequencing

In addition to the creation of linkage maps, another critical development that greatly increased our understanding of the link between genes and human disease was the arrival of technologies for the sequencing of DNA molecules. The technology to sequence DNA was first developed in the 1970s with the arrival of the Maxam-Gilbert method (9) followed later by the invention of Sanger sequencing in 1977 (10), a technique which became the standard for the next 25 years. Over the ensuing decades, advances in both sequencing technologies and computational analysis of sequence data made DNA sequencing more reliable and cost effective. However sequencing complete mammalian genomes remained difficult and time consuming largely due to technical limitations of the time. As a result, it wasn't until 2003 that the first draft of the human genome was generated using Sanger sequencing by a consortium called the Human Genome Project (HGP), effectively creating the initial human reference genome (1). The success of this project was the culmination of a worldwide collaboration of hundreds of labs internationally ultimately costing 3.4 billion USD and taking 13 years to complete. While costly and time consuming, the creation of a human reference genome by the HGP proved an enormous breakthrough in many ways, offering a new tool for researchers to utilise in the identification of many more disease-causing genes.

By combining linkage data with the annotated reference genome, researchers were, for the first time, able to generate small candidate gene lists and further to identify point mutations in DNA sequence within these genes common to affected family members.

1.5 Draft human genome

While the HGP effort generated the first complete draft of the human genome, another major effort was underway during this time led by Craig Ventner at Celera Genomics (11), a project that aimed to build a draft genome using the less expensive shotgun sequencing technique they pioneered. The eventual completion of this project marked the arrival of the second draft human genome, all of sudden making it possible to generate genome wide catalogues of genetic differences between individuals. Increasing sequencing efforts resulted in more and more sequence variation information being generated (both within and between species) leading to the eventual creation of the first publicly available variation database; the single nucleotide polymorphism database or dbSNP (2), an NCBI resource designed to capture all genetic variation identified in human populations. From its creation in 1998, the number of entries in dbSNP continues to grow each year with the latest human release (v144) containing over 150 million unique variants, over 50 million of which were added in the last year alone (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). Following this effort, interest in cataloguing variants relevant to specific diseases began to emerge, resulting in the creation of repositories such as COSMIC (12) for storing somatic mutations in cancer or ClinVar (4) for aggregating information about relationships among variation and human health.

1.6 Next generation sequencing

While the accumulation of variation information continued after the creation of the initial human reference genome, the pace changed dramatically with the development and widespread adoption of massively parallel sequencing technologies (Next-Generation Sequencing; NGS). NGS is commonly defined as the group of sequencing technologies that parallelize the sequencing of individual sequences in a single experiment, making it possible to sequence

millions or even billions of DNA molecules simultaneously. The first widely available platform was Life Sciences 454 released in 2004, a platform that offered an immediate 6X decrease in per base cost compared to traditional Sanger sequencing (13). Since this time, advances have continued apace with the cost of sequencing a mammalian genome dropping from $100,000,000 USD in 2001 to $10,000,000 USD in 2007 all the way down to $1,000 USD in 2015 via the arrival of Illumina's X-Ten sequencing system (http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html). With the arrival of the $1,000 genome, both population scale and large disease genome sequencing studies have become economically feasible for the first time, although it must be mentioned these figures do not factor in the significant storage and analysis costs required for such large data sets (14).

1.7 Sequence data growth

Decreasing per-base sequencing costs means an ever-growing number of studies are using NGS sequence data to detect causal variants in a variety of diseases (15, 16), and that more and more genomes are being sequenced worldwide every year. In fact, it is estimated that 228,000 human genomes will be sequenced in 2015 which represents more genomes than have been sequenced up to the end of 2014 in total (5). Already, large population studies are being published such as the recent project sequencing 2636 Icelanders where researchers found an over-representation of homozygosity and rare protein-coding variants, many of which proved important for disease (17). Ambitious plans are being formulated around the world to sequence millions more genomes in the coming years; for example Genome England recently announced their intention to sequence 100,000 genomes (http://www.genomicsengland.co.uk/the-100000-genomes-project/), the NIH announced their plans to sequence 1,000,000 genomes (http://news.sciencemag.org/biology/2015/01/white-house-fleshes-out-obama-s-215-million-plan-precision-medicine), and China is discussing similarly ambitious plans (18). The decrease in sequencing cost is fuelling such projects and locally we observe huge increase in use of our infrastructure with almost 1000 whole human genomes analysed as of December 2015, the majority of which have been analysed in the last 18 months following the arrival of the

Illumina X-Ten sequencing cluster at the Garvan Institute in Sydney (https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/clinical-genomics/sequencing-services/hiseqXTen). Collectively, these and other ongoing sequencing projects mark the beginning of the big-data era in the health sciences.

1.8 Computational challenges

While the ability to generate huge amounts of sequence data in a cost effective manner removes a long-standing technical bottleneck, it presents several new computational challenges requiring flexible and robust software solutions. The first major challenge to address with large NGS data sets is how to effectively manage and analyse the data from increasingly large numbers of often-related samples. With each 30X whole genome sample typically generating 0.1 terabyte (TB) of compressed raw sequence data and requiring hundreds of hours of compute time for basic variant detection alone, a great need exists for robust, flexible, and mature high-throughput bioinformatic analysis pipelines. The second major challenge is how to prioritise the huge number of variants detected in an individual genome in order to identify causal variant(s) responsible for the disease in an affected individual. This is a particularly daunting challenge given that each genome has a huge number of mostly benign variants with the average person having roughly 6 million SNVs, 650000 small indels, 14000 structural variants, and 250-300 loss of function variants (19). While numerous solutions exist to address both these challenges, there are currently no gold standard methodologies for either task resulting in an increasing number of in-house custom analysis pipelines and tools that string together a combination of various open source software and custom code.

1.9 Clinical application

One increasingly common application of NGS technology is the sequencing of individuals with rare monogenic diseases, a large worldwide health problem with one in every fifty individuals worldwide affected by one of the estimated 10000 rare monogenic diseases (http://www.who.int/genomics/public/geneticdiseases/en/index2.html). With

recent studies showing how treatment options can be altered based on patient variant information (16), clinicians are increasingly incorporating such information in their practice to identify both disease-causing and risk-factor variants that predispose patients to certain disease (20). Challenges remain however and while algorithms for reliably detecting variants from sequence data are relatively mature (21), the routine use of patient variation data to improve clinical outcome remains elusive with no 'gold-standard' methodology describing a current set of best practices. Initiatives such as the CLARITY challenge (22) are attempting to describe such best practices, while other more general frameworks such as ADAM (23) are tackling genomic data sharing issues, aiming to forge common standards and protocols to make sharing and computing genomic data seamless. Despite these ongoing efforts, global standards remain elusive and in the interim increasingly complex in-house analysis pipelines are being developed to detect clinically significant variants. Standardising such analysis pipeline remains challenging due to numerous factors such as the wide variety of available analysis software (21) and project-specific variant prioritisation strategies. Further, even when software choices are standardised, pipelines may still differ with regard to software version and parameters utilised with such differences often having a significant effect on variant calls. For example, SAMtools v0.1.19 incorporated base alignment quality (BAC) filtering as a default parameter for the first time, a change which produced a huge difference in the total number of variants called relative to earlier software versions. Despite all these challenges, progress toward standardization has begun largely from the widespread use of the GATK system (24), although even within this framework there are two distinct algorithms for detecting variants, namely the unified genotype caller and the haplotype caller, both of which are typically updated during each new GATK release.

1.10 Analysis pipelines

Given the large number of in-house analysis pipelines, numerous options for the overall management of informatics pipelines have been developed which can be divided into easy to use web application like GALAXY (25) and Taverna (26) and language specific frameworks such as GATKs queue

(https://github.com/broadgsa/gatk/) or Bpipe (27). Developing a robust framework that is generalizable for the majority of common use-cases proves challenging however, with most software making assumptions about the inherent nature of the analysis pipeline. For example, frameworks commonly assume that any analysis workflow will be linear and processive, consisting of a series of discrete analysis steps such as sequence data quality control (QC), alignment, variant calling, annotation and variant prioritization. This assumption does not hold for all workflows however, with recent studies demonstrating significant differences in variant call quality between individual software and pipelines (28, 29) meaning any enduring genomics framework must be designed to be highly flexible from the onset.

1.11 Introduction for publication #1:

"*Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies" (30)*

Given these known differences in software/pipeline performance, one approach is to aggregate variant calls from multiple software tools, with such results shown to yield an overall improvement in both sensitivity and specificity compared to any individual tool (31). While running a linear analysis pipeline is suitable in many instances, it has become increasingly clear multiple variant callers are preferable, particularly in circumstances where either minimizing false positives or false negatives variants is a priority. For example, a pipeline designed for clinical use should focus on reducing the number of false negative variants, as failing to detect clinically important variants may represent a missed opportunity to improve clinical outcome. This publication details a comprehensive three-part study first assessing whether the choice of software and variant filtering impacts the overall variant call quality and secondly, using a melanoma cell line, if clinically important melanoma risk factor variants are uniformly detected under all software conditions tested. Lastly, I simulate various sequence contamination levels to determine whether contamination issue impacts variant callers uniformly, an important consideration when sequencing cancer samples which typically suffer from contamination (32). Overall, the results imply careful software selection, variant caller filtering optimization, and

combining variant calls from multiple tools are all required to minimize false negative variants, an important consideration when utilizing variation data in a clinical context.

1.12 Introduction for publication #2:

"*Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees" (33)*

While NGS data have proven extremely successful in studying monogenic diseases, it is increasingly common to sequence large numbers of individuals with polygenic or complex disease such as autism (34) or auto-immune diseases (http://jcsmr.anu.edu.au/research/cpi). While the sequencing of unrelated individuals has resulted in the successful discovery of causal/disease-associated variants for monogenic (15) and complex disease (35), many causal variants remain unidentified and are effectively lost in the noise arising from the millions of benign variants detected in each sample which produce no phenotype or are not related to the phenotype of interest. With 2% of all people carrying a non-damaging missense mutation in any given gene (36), routinely differentiating causal variants from background variation typically requires additional information to reduce the search space for causal variants. One of the simplest ways to reduce the search space is to sequence both affected and unaffected family members of a proband, as we are then able to both prioritize variants common to affected individuals and exclude benign variants shared between affected and unaffected individuals (37). Being able to easily detect such private familial variants is important, particularly with studies identifying such variants as being causal (38). When sequencing a pedigree, in addition to yielding information on variant distribution within a pedigree, we can obtain additional pedigree level annotation information such as disease inheritance, genome phasing, and compound heterozygosity. We are able to utilize this additional information for prioritization that is unavailable unless family members are sequenced. While numerous tools exist for variant prioritization within a single genome (39), the ability to concurrently analyse variants within pedigrees remains a challenge, especially should there be no prior indication of the underlying genetic cause of the disease. While tools are emerging to detect

causal variants in sequenced pedigrees, they tend to focus either on removing variants based on criteria deemed unlikely to be causal (40) or on variants matching specific inheritance models such as auto-dominant (41) or compound heterozygotes (42).

This publication describes the software package I wrote to reliably detect causal variants in sequenced pedigrees for any genetic disease, whether monogenic or complex (https://github.com/mattmattmattmatt/VASP). Designed to aggregate data for genetic variants across the entire pedigree without making any underlying assumptions regarding disease transmission mechanism, VASP enables powerful and customizable variation prioritization, allowing researchers to utilize their knowledge of the disease (e.g. expected inheritance pattern or mapped genomic region) to reduce the search space for causal variants. This tool enables users to greatly reduce the number of candidate causal variants down to a size suitable for manual interrogation, using single variant and pedigree-wide annotation coupled with custom filtering criteria.

1.13 Introduction for publication #3:

"Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-calling algorithms for human melanoma genomes" (43)

While NGS has successfully been utilised to detect causal variants in all types of Mendelian diseases, one of the earliest adopters of NGS sequencing is cancer sequencing and the accurate detection of somatic variants. In this approach, paired tumour and normal whole genome samples are sequenced in order to comprehensively characterise cancer genomes in a variety of tumour types (44). The nature of variation in cancer genomes is fundamentally different from non-cancer genomes (inherited mutations germ line are thought to be important in only 5-10% of cancers (45)) with much of the current understanding of the genetics of cancer based on the understanding that a clone accumulates somatically acquired mutations that ultimately leads to malignant transformation (46). Large-scale cancer genome sequencing is actively being used to identify important somatic variants, some of which has been successfully translated into novel customised therapies. In melanoma for example, activating mutations of the MAPK pathway activator gene BRAF were first identified by routine genetic

screening and later shown to be highly prevalent across a number of cancers including melanoma (47). Following this discovery, BRAF inhibiting drugs were developed and then shown to arrest melanoma growth in a high percentage of BRAF mutation-positive patients and to prolong survival in advanced, metastatic melanoma – representing the first systemic therapy to do so (48). Importantly, this early success story demonstrates how the detection of a single somatic mutation led to the development and widespread adoption of a targeted therapy.

Generally the first step in identifying possible novel therapeutic targets for any cancer type is the characterisation of the cancer genome in question, as is being done for numerous cancers via large-scale cancer sequencing efforts such the International Cancer Genome Consortium (ICGC) (49) and the Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/). Both tumour and normal genomes need to be sequenced and somatic variants identified with tools such as MuSiC (50) and MuTect (51), software designed specifically to detect somatic variants in paired cancer samples. While such tools are suitable for many cancer types, differences with regard to both tumour purity (32) and mutational heterogeneity (52) amongst cancers means custom workflows are often required. In support of Bioplatforms Australia melanoma initiative to characterise the melanoma genome (http://www.bioplatforms.com/melanoma/), I developed a custom melanoma analysis pipeline to analyse 150 melanoma whole genomes, the largest melanoma genome sequencing project at the time. In this project, an initial pilot data set was generated sequencing three cancer genomes (primary tissue, metastatic tissue, and a metastatic cell line) to high coverage in order to empirically determine optimal coverage levels required for tumour and normal samples for each type of sample. Once coverage levels were determined I reviewed both existing and custom solutions for detecting somatic variants in paired cancer genomes, analysing all three samples with a variety of tools before ultimately incorporating the best performing software into the pipeline.

1.14 Introduction for publication #4:

"*Comparison of predicted and actual consequences of missense mutations*". (53)

While fundamentally different sample types such as paired cancer or sequenced pedigrees require custom workflows, a common requirement of all

workflows is the prioritisation of the huge number of variants typically detected in each sample. Most workflows do this by incorporating a post-variant detection annotation step using tools such as AnnoVar (54) or Variant Effect Predictor (VEP) (55) to identify missense, nonsense, and splice-site mutations. With each genome containing thousands of missense variants and very few exhaustive experimental data sets available, practical interpretation of the functional consequence of missense mutations typically relies on a variety of functional inference tools such as Polyphen2 (56), SIFT (57), or CADD (58). Most, if not all of the most accurate tools, rely heavily on multiple protein sequence alignment, with changes to highly conserved amino acids thought to be more damaging to a protein than changes to a non-conserved amino acid. While a useful approximation in many cases, it makes scoring of missense mutations extremely sensitive to the initial transcript selected for each organism included in the alignment (59), an important consideration now that the tools are being used to prioritise variants in the clinic. Unfortunately, clinicians are beginning to discover that in some cases the tool's ability to predict the severity of the disruption to protein function does not always correspond to the observable clinical outcome or disease progression. For example, while prediction of the severity of missense mutations in the tumour suppressor gene TP53 shows some correlation to clinical outcome, the correlation is not significant (60). This observation and others of this nature have led many to conclude that functional inference tools lack the precision required for clinical usage (61). Given these known issues for individual tools, software has been developed to integrate the prediction of various individual tools, tools such as the popular Condel (62) which uses a weighted average of the normalised prediction scores. While such tools have been shown to improve accuracy in functional inference predictions (perhaps due to the minimisation of outlier effects), there is no large reduction in the number of missense mutations predicted to be damaging that do not cause an observable phenotype.

In two separate analyses, we attempted to measure the efficiency of the functional inference tools with the hope of gaining an understanding of the nature of the functional inference tools apparent 'overcalling' problem. The first data set consisted of 33 missense mutations bred to homozygosity in N-ethyl-N-nitrosourea (ENU) treated mice spread across 23 essential immune system genes,

the mice identified as part of our larger system developed to identify protein altering mutations in the exome sequence of progeny from ENU treated G1 mice. In this data set we showed how only ~20% of missense mutations predicted to disrupt protein function produced a discernible phenotype, essentially replicating earlier studies that highlighted this overcalling problem. Several explanations for this overcalling phenomenon have been previously proposed, the most common being variable penetrance. While this experiment sidesteps the issue of variable penetrance (all genes caused a fully penetrant, detectable, and well-characterised phenotype when rendered null), two possible explanations for the overcalling remain. First, the possibility exists that defects in the immune system are being masked *in vivo* by compensation and the environment. The second possible explanation is that the differences between mouse and human proteins are responsible for the overcalling as most tools are calibrated using human protein data, so possibly the tools are inappropriate for scoring missense variants in non-human variants. To address both these concerns, we utilized the exhaustive TP53 data, where transcription enhancing activity measurements are available for all possible missense mutations (63). In this analysis, we compared comprehensive functional assay data with functional inference scores from a variety of tools and conclude that for all tools tested, almost half of all missense mutations predicted to be damaging produce no phenotype. We conclude that there exists a subclass of "nearly neutral" (64) mutations that are subject to purifying selection yet produce little impact on clinical phenotype in any individual, a class of variants likely responsible for the apparent overcalling phenomenon.

1.15 Introduction for publication #5:

"*DeepSNVMiner: A sequence analysis tool to detect emergent, rare mutations in sub-sets of cell populations*" (65)

While NGS sequence data typically is derived from a single homogeneous population, an increasing number of researchers are performing massively parallel sequencing on mixed populations of cells representing both wild type and disease state. Applications for this technology are broad ranging, with current examples including the detection of sequence variation in a variety of cancer subtypes (66) and the identification of the emergence of drug resistant

point mutations in virus (67) to name a few. The central technique that enables the mix of cells to be disambiguated is the attachment of a random unique identifier sequence (UID) to each DNA molecule either prior to or during the amplification step (68). By attaching the UID at the onset, even though subsequent polymerase amplifications may introduce errors, real variants can be differentiated from amplification variants by grouping sequences with common UIDs and detecting variants within each group. While the laboratory techniques for attaching UIDs are relatively mature, there is currently no software capable of routinely detecting variants from mixed cell sequence data with UIDs attached.

To address this, I developed the software DeepSNVMiner to detect variants in these mixed data sets, a resource publicly available on github (https://github.com/mattmattmattmatt/DeepSNVMiner). DeepSNVMiner allows robust and reliable identification of sequence variants present in a subset of sequences within a mixed DNA input sample. This tool runs workflows required to support SafeSeq and similar UID tagged sequence datasets, and enables variant detection from single DNA molecules present within mixed heterogeneous sequence data.

1.16 Introduction for publication #6:

"*Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models*" (69)

While this introduction has thus far focused on sequence data from human samples, another widespread application of NGS data is the sequencing of model organisms such as the mouse, an organism which has traditionally been used for modelling human disease and comparative genome analysis. There are over 450 inbred strains of mice (70) representing a wealth of phenotype and genotype varieties available for genetic study in addition to a large number of mouse models of human disease. Genetic analysis of mouse is typically performed on the C57BL/6 strain or another inbred strain due to the uniform genetic background. One common means of studying disease in the mouse is via treatment with N-ethyl-N-nitrosourea (ENU), a powerful mutagen that generates random single nucleotide germ line mutations allowing observable resultant

changes in traits or disease to be detected by phenotypic screening procedures (71). While ENU screening has proven extremely successful it has traditionally been an extremely expensive and time-consuming process in which affected mice are outcrossed to another inbred mouse strain and causal variants meiotically mapped to a subregion of the genome using a panel of strain specific variants. Following this, genes within each mapped region are identified with all exons Sanger sequenced and the causal variant identified within this reduced genomic region. While this method proved extremely successful, it is unsuitable for high-throughput sample processing primarily due to the investment required to take each mutation forward. Alternatively, it was theorised that if it proved possible to identify the causal variant by sequencing a mutated mouse, the time and cost required would drop dramatically making high-throughput processing possible. Further, as almost all ENU-induced causative mutations have been located either in exons (~75%) or splice donor-acceptor sites (~25%) (72), it is appropriate to utilize targeted exome sequencing to generate sequence data for the causal variants.

In this publication, we describe the high-throughput variant detection pipeline we built to reliably identify rare, ENU-induced de novo mutations in exome data from C57BL/6 mice. While previous studies had suggested that exome sequence alone may not be enough to identify disease-causing induced mutations without extensive SNV validation (73), we have shown that exome sequencing as a sole source of information is sufficient by itself to identify ENU-induced mutations selected by phenotype. The reliable detection of ENU-induced causal mutations from exome sequencing greatly reduces both the time and cost of ENU screening and offers an immediate source of thousands of new mouse models.

Chapter 2: Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies

Field, M. A., V. Cho, T. D. Andrews and C. C. Goodnow. "Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies." PLoS One 2015;10(11):e0143199

PLOS ONE

RESEARCH ARTICLE

Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies

Matthew A. Field[1,3]*, Vicky Cho[1,2], T. Daniel Andrews[1,3☯], Chris C. Goodnow[1,4☯]

1 Department of Immunology, John Curtin School of Medical Research, Australian National University, Canberra, ACT, Australia, **2** Australian Phenomics Facility, Australian National University, Canberra, ACT, Australia, **3** National Computational Infrastructure, Australian National University, Canberra, ACT, Australia, **4** Immunogenomics Group, Immunology Research Program, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia

☯ These authors contributed equally to this work.
* matt.field@anu.edu.au

OPEN ACCESS

Citation: Field MA, Cho V, Andrews TD, Goodnow CC (2015) Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. PLoS ONE 10(11): e0143199. doi:10.1371/journal.pone.0143199

Editor: Yan W. Asmann, Mayo Clinic, UNITED STATES

Received: June 8, 2015

Accepted: November 2, 2015

Published: November 23, 2015

Data Availability Statement: The GIAB data set was downloaded as two FASTQ files 'illumina-100bp-pe-exome-150x' from GCAT (http://www.bioplanet.com/gcat) and variant calls compared against GIAB high-confidence genotypes for NA12878 (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/GIAB_integration/. The melanoma cell line is available from the Australasian Biospecimen Network http://abrn.net/tsl/locations/qimr_cb/

Funding: This work has been in part funded by grants from the National Institutes of Health (Author: CCG; URL: www.nih.gov; Grant ID: U19 AI100627), the National Health and Medical Research Council

Abstract

A diversity of tools is available for identification of variants from genome sequence data. Given the current complexity of incorporating external software into a genome analysis infrastructure, a tendency exists to rely on the results from a single tool alone. The quality of the output variant calls is highly variable however, depending on factors such as sequence library quality as well as the choice of short-read aligner, variant caller, and variant caller filtering strategy. Here we present a two-part study first using the high quality 'genome in a bottle' reference set to demonstrate the significant impact the choice of aligner, variant caller, and variant caller filtering strategy has on overall variant call quality and further how certain variant callers outperform others with increased sample contamination, an important consideration when analyzing sequenced cancer samples. This analysis confirms previous work showing that combining variant calls of multiple tools results in the best quality resultant variant set, for either specificity or sensitivity, depending on whether the intersection or union, of all variant calls is used respectively. Second, we analyze a melanoma cell line derived from a control lymphocyte sample to determine whether software choices affect the detection of clinically important melanoma risk-factor variants finding that only one of the three such variants is unanimously detected under all conditions. Finally, we describe a cogent strategy for implementing a clinical variant detection pipeline; a strategy that requires careful software selection, variant caller filtering optimizing, and combined variant calls in order to effectively minimize false negative variants. While implementing such features represents an increase in complexity and computation the results offer indisputable improvements in data quality.

(Author: CCG; URL: https://www.nhmrc.gov.au/; Grant ID: Australia Fellowship 585490), the National Collaborative Research Infrastructure Strategy (Australia), the Melanoma Institute of Australia (Author: MAF; URL: www.melanoma.org.au/), and Bioplatforms Australia (Author: MAF; URL: www.bioplatforms.com/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Author Matthew Field confirms that he received partial funding from a commercial source, Bioplatforms Australia. This funding does not alter the authors' adherence to PLOS ONE policies on sharing data and materials as detailed online in our guide for authors.

Introduction

Rapid improvements in massively parallel sequencing technologies have dropped the cost of sequencing to the point where it is feasible to use patient sequence data in the clinic in order to identify both disease-causing and risk-factor variants that predispose patients to certain disease [1]. While data-specific computational algorithms aimed at deriving accurate data from these technologies have reached maturity [2], the routine use of genomic data for improving clinical diagnosis and treatment requires formalizing existing research methods into clinical best practices; the goal of recent initiatives like the CLARITY challenge [3]. In the interim, bioinformaticians are increasingly developing local, in-house production variant detection pipelines, the core of which typically consists of a highly customized workflow utilized by a relatively small local user-base, being developed out of necessity as few available software systems support the easy 'pipelining' of large biological datasets through multiple analytical tools to produce the desired end result. As a result, standardization of such pipelines is challenging largely due to the diversity of rapidly developing software tools being utilized, though some progress toward standardization efforts are beginning to emerge from the widespread use of the GATK software package [4].

The last several years has seen the creation of software frameworks for the management of bioinformatics pipelines which generally fall into two categories; easy to use GUI and web based approaches like GALAXY [5] and Taverna [6] compared to language specific frameworks such as GATK's Queue (https://github.com/broadgsa/gatk/), BPipe [7] and Snakemake [8]. While these tools do address many of the key requirements of a high-throughput informatics systems, they typically anticipate that analysis will be linear and processive, an assumption which does not hold true in many instances. In addition, surprisingly few of these tools currently support large numbers of users, with the notable exception of GALAXY, which enjoys active support from a large broad-based community of users and developers (https://biostar.usegalaxy.org/). GALAXY and other web-based tools however, are not always the immediate choice for a high-volume production system due to potential challenges with transferring large volumes of data and the unavailability of highly specialized workflows.

Given the fast moving nature of both sequencing technologies and bioinformatic software development, successful and enduring informatics frameworks must remain flexible with regard to software selection, a central idea in the development of Bioconductor [9]. Such flexibility requires systems to routinely evaluate new software particularly in light of recent publications demonstrating low concordance levels between variant detection pipelines [10, 11]. Such differences arise not only due to software choices but also due to sequencing technology choices with studies demonstrating large differences in variant calls using different exome capture systems [12–14] and whole genome sequencing platforms [15, 16]. While the choice of sequencing technology and software is important, the internal parameters utilized for each algorithm are also important, particularly the filtering options employed by variant callers, a feature known to affect the overall variant call quality [17, 18]. Another important consideration for a framework is the ability to support testing software both in isolation as well as in various combinations (e.g. aligner / variant caller pairs) with previous studies showing the choice of short read aligner significantly impacts downstream variant calling, particularly for indels [19]. Further, with additional studies reporting both platform-specific [15, 20, 21] and software-specific [19] variants, it is clear that ideally a framework will support variant calls from multiple tools and sequencing platforms, particularly when avoiding false negative variants is the highest priority, as is the case when using variation data in a clinical context.

Assessing the quality of variant calls from any variant detection pipeline is greatly facilitated by the recent development of high quality reference data sets such as 'genome in a bottle' or

GIAB [22]. Studies demonstrating low concordance levels amongst variant callers [23, 24] demonstrate the importance of software selection and highlight the need for standardized frameworks such as GCAT [25] which make it possible to assess variant calls relative to a set of validated high quality variants. While such frameworks are useful for accessing the results from a single processive analysis, tools such as BAYSIC [26] have shown that aggregating variants from multiple variant callers yields an overall improvements in both sensitivity and specificity compared to any individual tool, making it increasingly clear multiple variant callers are preferable in circumstances where either minimizing false positives or false negatives are crucial to a project. While many aspects of variant calling algorithms are similar, combining them leads to a substantial improvement in output quality justifying their use in parallel in certain circumstances. The utility of this approach is that through combining the relative sensitivities of different tools, a broader sensitivity is gained through taking the union of multiple calling methods. Similarly, performing variant calling through different algorithms in parallel, the quirks or systematic errors of a single tool may be overcome by taking the intersection of multiple approaches. It is for these reasons we have enabled combined variant calls in our in-house high throuhgput production system, yet, given these advantages, few specific workflows used in practice allow integrating tools in parallel in such a manner. This may be due to the increased computational load or due to structural limitations of workflow management tools.

In this work we first seek to demonstrate the significant impact the choice of aligner, variant caller, and variant caller filtering strategy has on both the number and quality of variant calls using high quality NA12878 genotype calls as a baseline (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/GIAB_integration/). Next, using the same data set, we replace increasing portions of the GIAB sequence data with non-variant reference data to determine the effectiveness of variant callers at increasing levels of simulated sample contamination. This is important given the increased use of sequenced cancer data in the clinic with cancer samples known to differ with regard to tumour purity [27], mutational heterogeneity [28], and subclonality [29]. Finally, using a melanoma cell line control sample, we demonstrate the impact the choice of software and filtering strategy has on the detection of clinically important melanoma risk-factor variants. For all analyses, the three different variant types assessed in this study (SNVs, small insertions, and small deletions) are analyzed independently to assess whether any single algorithm is superior to all others tested for all variant types. We conclude by presenting a cogent strategy for implementing a variant detection pipeline for clinical use; a pipeline focused on minimizing the total number of false negative variants.

Materials and Methods

Samples

GCAT Sample. The two FASTQ files 'illumina-100bp-pe-exome-150x' were downloaded from GCAT (http://www.bioplanet.com/gcat) and variant calls compared against GIAB high-confidence genotypes for NA12878 (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/GIAB_integration/). The sequence data consists of 45 million 100bp read pairs for estimated exome coverage of 150X.

Melanoma cell line control sample. Biospecimens were provided by members of the ABN-Oncology group, which is funded by the National Health and Medical Research Council. Sample C001 is a control lymphocyte derived sample from a female patient available for download at https://ccgapps.com.au/bpa-metadata/melanoma/sample/102.100.100.7688/ with the corresponding tumour sample (not used in this analysis) also available at https://ccgapps.com.au/bpa-metadata/melanoma/sample/102.100.100.7687/. For sequencing, the library construction was performed using TruSeq DNA Sample Preparation kits as per Illumina instructions.

1 μg of sample DNA was fragmented into 300–400 bp average insert size with 3' or 5' overhangs. End repair mix was then used to convert the fragmented DNA into blunt ends by removing the 3′ overhangs and the polymerase activity fills the 5' overhang. The 3' ends were then acetylated to add a single "A" nucleotide to the 3' to reduce chimera formation. Ligate adapters were then used to attach adapters to the DNA fragments so they could be loaded into a flow cell and purified to remove unligated adapters to generate a final product with an insert size of 300–400 bp. PCR was then used to selectively enrich DNA fragments with adapter molecules at both ends for sequencing. Post amplification quality controls were performed using DNA High Sensitivity Labchips (Agilent Bioanalyzer). The libraries were then pooled and clustered using the iBOT and ready for sequencing. The 100 bp pair-end library was sequenced on a Hiseq2000 using a Truseq SBS V3-HS kit. The sequence data for C001 consists of 925 million 100bp read pairs for an estimated genome-wide coverage of 60X.

System and Implementation

All analyses described was performed using an in-house production genomics analysis framework; a framework which bundles analytical processes as externally developed, compiled binary objects, driven by a custom Perl module layer that wraps individual tools. Each step in the workflow is driven at a scripting level and the command to these driver scripts and the associated parameters are defined within an XML configuration file—which allows customization, quick tool substitution, and straightforward change and extension of the workflow. Furthermore, archived workflows may be easily recreated with this static XML file from a particular analysis, along with the code repository revision number, allowing quick reproduction of archived analyses. Initially designed to detect SNVs in the exome sequence of progeny of C57BL/6 laboratory mice exposed to the spermatogonial point-mutagen *N*-ethyl-*N*-nitrosourea (ENU) [30], the expanded framework includes workflows for SNV/indel detection in human exomes and genomes, as well as custom multi-sample workflows to identify causal variation across sequenced human pedigrees [31] and paired tumour-normal analyses [32]. The total versioned code-base currently utilizes an ever-changing catalogue of open-source components combined with bespoke analysis tools–currently all linked via an underlying MySQL (http://www.mysql.com) tracking database (S1 Fig) designed for persistence, archiving and querying of summary results. A more detailed description of the design and implementation of the system can be found in supplementary data (S1 File).

Expanded Software Assessment

The default variant calling workflow (S2 Fig) was expanded to pair three short read aligners (BWA [33], Bowtie 2 [34], and Isaac-aligner [35]) with three variant callers (GATK [4], Isaac-variant-caller [35], and SAMtools [36]), each of which was run with both with and without additional filtering (Fig 1). For all 18 possible aligner/variant caller pairs, software was run using either default options or as per suggestions in associated documentation (S1 Table). For each aligner, reads were aligned to the human reference genome (assembly GRCh37) and a sorted, indexed BAM file generated using SAMtools. Each BAM file was provided as input to each variant caller to generate a VCF file of unfiltered variant calls. To obtain filtered variant calls, GATK was run through variant quality score recalibration steps (VQSR) as documented at https://www.broadinstitute.org/gatk/guide/article?id=2805, SAMtools was run with and without full BAQ filtering and Isaac variant had all LowGQX annotated variants removed (defined as variants with a GQX score less than 30 or not present with GQX being the minimum of genotype quality score assuming variant and non-variant locus). Next, all variants with quality scores of less than 40 were removed and variants regularized using custom code

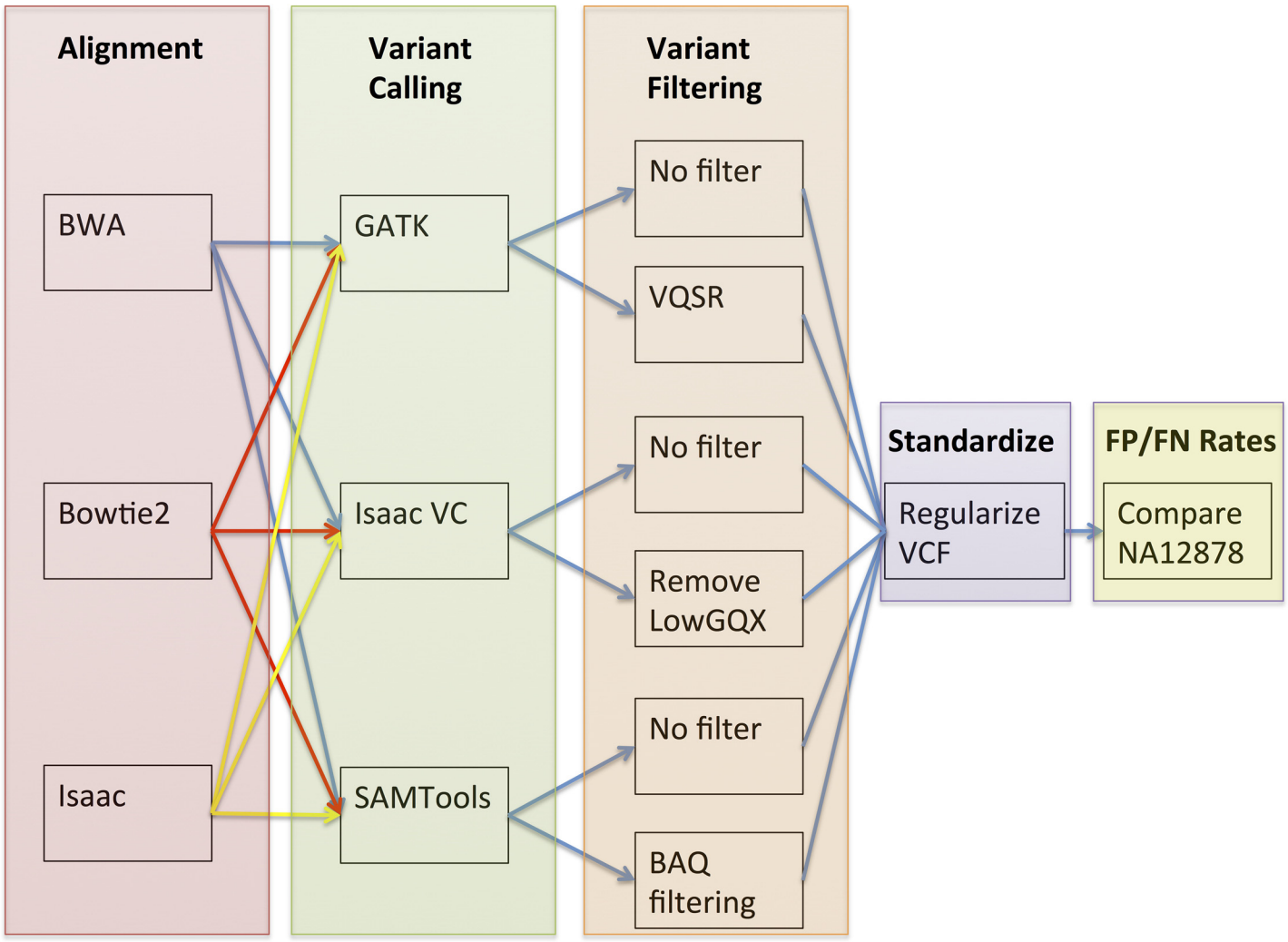


Fig 1. Analysis Workflow. BAM files from BWA, Isaac aligner, and Bowtie2 were paired with each of GATK, Isaac variant caller, and SAMTools run both with and without additional filtering (VQSR, BAQ, and LowGQX respectively). Output vcf files were regularized using custom code and variants from GIAB high quality regions taken forward to generate false positive and false negative rates.

doi:10.1371/journal.pone.0143199.g001

and further divided into SNVs, insertions, and deletions lists. These lists were reduced to only variant contained in the high quality genotype GIAB regions and finally compared to the corresponding GIAB variant type to obtain false positive rates. False negative rates were estimated by taking all variants in the high confidence variant calling regions found to also overlap ENSEMBL v75 canonical transcripts. This reduction in variant search space was required to account for the fact that exome sequence data was utilized in this analysis.

Variants were classified as matching when the following criteria were met:

1. both variants are of the same types; either SNV, insertion, or deletion
2. both variants share the same start coordinate
3. both variants share the same end coordinate

To ensure consistent reporting of genomic coordinates for indels (an issue particularly problematic with indels falling in repetitive genomic sequences), vcf files are regularized with

indels assigned the lowest genomic coordinate; for example a missing GC in a string of repeating GC's is recorded as a deletion of the first GC encountered in the 5' to 3' direction within the larger repeat sequence range. Results are summarized for SNVs, deletions, and insertions (S2–S4 Tables respectively).

Individual Software Assessment

To assess both the number and quality of variant calls for each aligner and variant caller in isolation, for each tool the union of all variants from all pairing was calculated and overlapped with GIAB variants and false positive and false negative rates calculated. For example, to assess GATK SNV calling efficiency, the union of SNVs called by GATK (paired with BWA, Bowtie2, or Isaac aligner input) was calculated and the merged list subsequently compared with high quality GIAB SNVs. Variant calls unique to a single software tool were also identified (referred to as 'tool-specific variants') and independently compared to GIAB variants to obtain false positive rates.

Overall Software Concordance

To measure the concordance of variant calls between the three aligners, the union of variants detected by each aligner was calculated and both Venn diagrams (Fig 2) and ROC curves plotted using genotype quality score (Fig 3) with the R packages VennDiagram [37] and ROCR [38] respectively. Similarly, variant caller concordance compared the output of GATK, Isaac variant caller, and SAMtools for both unfiltered and filtered variant calls. To assess how overall variant quality changed relative to the frequency of detection the number of times a variant was detected by all aligner / unfiltered variant caller pairs was calculated and divided into four categories; variants detected by at least 1 pair (union), variants detected by all 9 pairs (intersection), variants detected by 2–8 pairs, and variants detected by only 1 pair.

Heterogeneous Sample Simulation

To determine the impact of sample heterogeneity on variant detection, increasing portions of GIAB sequence data was replaced with mutation-free reference genome reads generated using the SAMtools utility wgsim. For each simulated contamination level, the total sequence coverage level was kept constant at 150X with contamination levels of 0%, 25%, 50%, 75%, 90%, 95%, 98%, and 99% generated. At each level, reads were aligned to the reference genome with BWA and SNVs called with GATK, Isaac variant caller, and SAMtools with filtering applied. False negative and false positive levels were obtained by comparing variant calls to the GIAB high quality variant regions.

Clinically Important Variants

To identify clinically important variants, all melanoma cell line control variants were overlapped with ClinVar (downloaded on October 24, 2014; [39]) and ClinVar entries classified as 'pathogenic' or 'risk factor' were identified with any melanoma risk factors variants prioritized.

Results

Individual Software Assessment

For each software tool, the union of all variants and tool-specific variants was calculated with false positive and false negative rates determined for SNVs (Table 1), deletions (Table 2), and insertions (Table 3). For each variant type, the false positive and false negative rates differed substantially with even greater differences observed for the tool-specific variants. For the

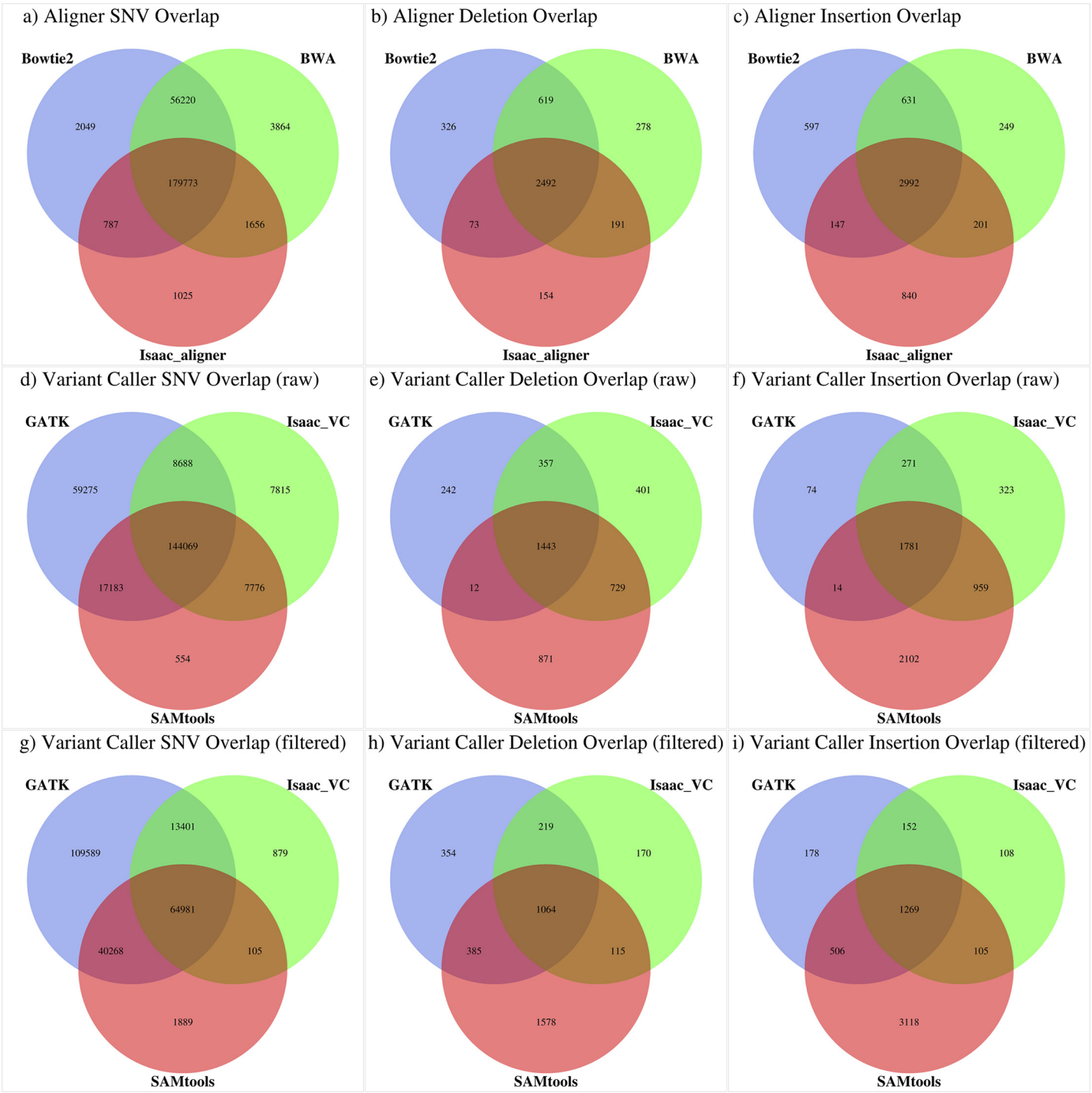


Fig 2. Software concordance Venn diagrams. Merged variant calls for each tool were overlapped with other tools of the same variety for each variant type. Aligners are compared in row 1, variant callers without filtering in row 2 and variant callers with filtering in row 3.

doi:10.1371/journal.pone.0143199.g002

aligners, BWA called the greatest number of total SNVs and deletions while Bowtie2 generated the most insertion calls. While Isaac aligner had fewer variant calls in general it had the lowest false positive rates for SNVs and deletions, with BWA having the lowest false positive rate for

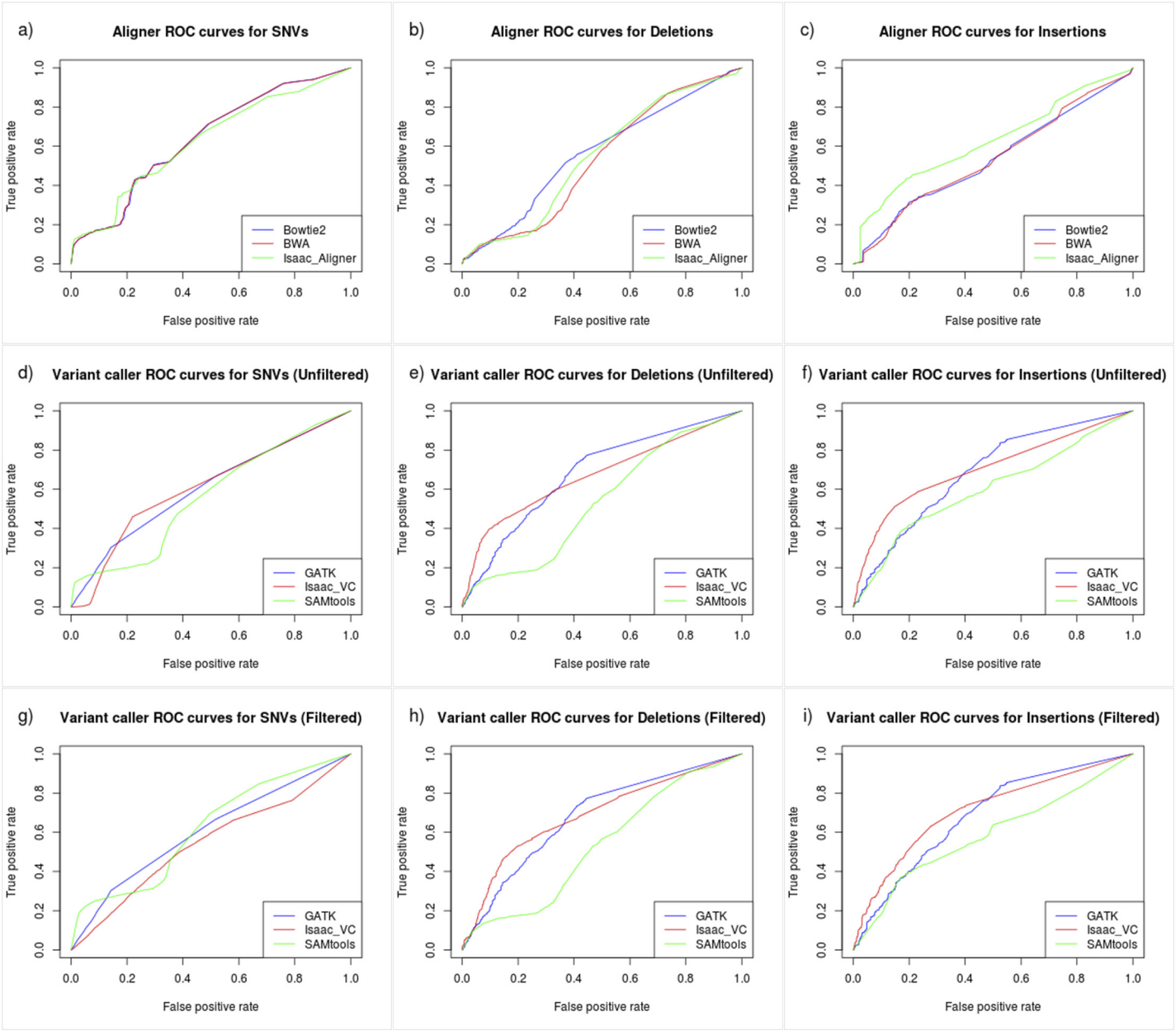


Fig 3. Software concordance ROC curves. Merged variant calls for each tool were calculated and ROC curves generated using the genome quality score. Aligners are compared in row 1, variant callers without filtering in row 2, and variant callers with filtering in row 3.

doi:10.1371/journal.pone.0143199.g003

insertions. Considering false negative rates, BWA had the lowest rate for SNVs and deletions while Bowtie2 had the lowest rate for insertions. For the tool-specific variants, the results differed relative to variant type with BWA calling the most SNVs with the second lowest false positive rate, Bowtie2 calling the most deletions with the lowest false positive rate, and Isaac aligner calling the most insertions with the lowest false positive rate. For the unfiltered variant calls from the three variant callers, GATK called ~25% (59,275) more SNVs than either SAMtools or Isaac variant caller while SAMtools called the greatest number of insertions and deletions. SAMtools had the lowest false positive rate for SNVs while Isaac variant caller had the lowest false positive rate for insertions and deletions. For the false negative rate, SAMtools

Table 1. Total SNV calls and tool-specific SNV calls for each aligner and variant caller run with and without filtering.

Software	Filter	Total SNV Calls	False Positive Rate	False Negative Rate[a]	Tool-specific SNVs	Tool-specific FP Rate
Bowtie2	N/A	238829	8.44%	3.15%	2050	53.41%
bwa	N/A	241513	8.55%	2.96%	3860	39.27%
Isaac_align	N/A	183241	4.02%	3.31%	1030	36.12%
GATK	None	229215	7.43%	2.91%	59275	15.19%
GATK	VQSR	228239	7.06%	2.99%	N/A	N/A
Isaac_VC	None	168348	6.35%	4.27%	7815	32.50%
Isaac_VC	LowGQX	79366	4.56%	7.47%	N/A	N/A
Samtools	None	169582	6.12%	3.68%	554	86.64%
Samtools	BAQ	107243	3.84%	6.52%	N/A	N/A

The union of variant calls for each tool was calculated and false positive and false negative rates determined relative to high quality GIAB variants. Tool-specific calls were also calculated, defined as SNVs specific to a single tool.

[a]From bases overlap ENSEMBL v75 canonical transcripts.

doi:10.1371/journal.pone.0143199.t001

had the lowest rate for deletions, Isaac variant caller the lowest rate for insertions, and GATK the lowest rate for SNVs. Considering the tool-specific variants, SAMtools called the most insertions and deletions and GATK the most SNVs with GATK-specific variants having consistently low false positive rates, much lower than either SAMtools or Isaac variant caller specific variants. Finally, for the filtered variant calls, filtering uniformly increased the false negative rate and decreased the false positive rate for SNVs while the effect on indel calls was mixed. For indels only GATK filtering resulting in a decrease in the false positive rate with SAMtools and Isaac variant caller filtering yielding higher false positive rates.

Overall Software Concordance

To assess the concordance of variant calls generated by the alignments from BWA, Bowtie2, and Isaac aligner, the merged variant lists for each aligner were overlapped to generate Venn diagrams with a similar approach taken to measure concordance levels between both filtered and unfiltered variant calls generated by GATK, Isaac variant caller, and SAMTools (Fig 2). For the three aligners, 73.26% of SNV calls were unanimously detected compared to only

Table 2. Total deletion calls and tool-specific deletion calls for each aligner and variant caller run with and without filtering.

Software	Filter	Total Deletion Calls	False Positive Rate	False Negative Rate[a]	Tool-specific Dels	Tool-specific FP Rate
Bowtie2	N/A	3617	33.26%	23.73%	379	29.29%
bwa	N/A	3676	30.88%	23.73%	252	46.03%
Isaac_align	N/A	2976	25.5%	27.12%	157	43.95%
GATK	None	2084	28.69%	27.12%	260	17.31%
GATK	VQSR	2049	27.57%	28.18%	N/A	N/A
Isaac_VC	None	2964	26.05%	23.73%	414	45.17%
Isaac_VC	LowGQX	1602	31.34%	35.59%	N/A	N/A
Samtools	None	3160	29.4%	27.12%	958	50.21%
Samtools	BAQ	3247	30.49%	27.12%	N/A	N/A

The union of variant calls for each tool was calculated and false positive and false negative rates determined relative to high quality GIAB variants. Tool-specific calls were also calculated, defined as deletions specific to a single tool.

[a]From bases overlap ENSEMBL v75 canonical transcripts.

doi:10.1371/journal.pone.0143199.t002

Table 3. Total insertion calls and tool-specific insertion calls for each aligner and variant caller run with and without filtering.

Software	Filter	Total Insertion Calls	False Positive Rate	False Negative Rate[a]	Tool-specific Ins	Tool-specific FP Rate
Bowtie2	N/A	4367	27.11%	22.39%	565	35.04%
bwa	N/A	4073	20.94%	31.34%	215	61.40%
Isaac_align	N/A	4180	31.27%	31.34%	843	14.95%
GATK	None	2140	20.56%	35.82%	74	14.86%
GATK	VQSR	2105	19.62%	37.31%	N/A	N/A
Isaac_VC	None	3334	19.26%	25.37%	323	56.97%
Isaac_VC	LowGQX	1634	23.32%	34.33%	N/A	N/A
Samtools	None	4856	34.1%	22.39%	2102	41.29%
Samtools	BAQ	4998	34.87%	20.90%	N/A	N/A

The union of variant calls for each tool was calculated and false positive and false negative rates determined relative to high quality GIAB variants. Tool-specific calls were also calculated, defined as insertions specific to a single tool.

[a]From bases overlap ENSEMBL v75 canonical transcripts.

doi:10.1371/journal.pone.0143199.t003

58.71% of unfiltered SNVs from the three variant callers. For deletions, the three aligners shared 55.89% of all deletion calls compared to only 35.58% of unfiltered deletions from the three variant callers. Lastly for insertions, the three aligners shared 52.89% of all insertion calls compared to only 32.24% of unfiltered insertions from the three variant callers. In addition to Venn diagrams, ROC plots were generated for each variant type using genotype quality as input (Fig 3).

Variant Type Comparison

Variants were segregated based on the frequency of detection for each variant types (Table 4). Overall, SNV calls exhibiting the greatest concordance levels with 43.45% of the total 245,360 SNVs detected by all pairs, 54.20% detected by 2–8 pairs, and 2.35% detected by a single pair. Deletions had the next highest concordance levels with 27.61% of the total 4194 deletions calls

Table 4. Variant calls grouped by frequency of detection.

Variant Type	Number of times variant detected (out of 9 total software pairs)	Total Variant Calls	False Positive Rate[a]	False Negative Rate
SNV	9 (Intersection)	106594	3.29%	6.78%
	> = 1 (Union)	245360	9.09%	2.90%
	2–8	132989	12.18%	N/A
	1	5777	46.43%	N/A
Deletion	9 (Intersection)	1158	14.68%	35.59%
	> = 1 (Union)	4194	35.13%	23.72%
	2–8	2326	36.54%	N/A
	1	710	63.80%	N/A
Insertion	9 (Intersection)	1415	12.16%	46.26%
	> = 1 (Union)	5524	35.36%	19.40%
	2–8	2612	26.03%	N/A
	1	1497	73.55%	N/A

Unfiltered variants were grouped based on the frequency of detection within nine possible aligner/variant caller pairs and segregated into four bins; variants in all 9 pairs, variants in at least 1 pair, variants in 2–8 pairs, and variants unique to 1 pair.

[a]From bases overlap ENSEMBL v75 canonical transcripts.

doi:10.1371/journal.pone.0143199.t004

detected by all pairs, 55.46% detected by 2–8 pairs, and 16.93% unique to a single pair. Lastly, insertions exhibited the lowest concordance levels with only 25.62% of the total 5524 insertions detected by all pairs, 47.28% detected by 2–8 pairs, and 27.10% unique to a single pair.

Considering the union of all variant calls results in 245,360 unique SNV calls, 5524 unique insertion calls, and 4194 unique deletion calls with an average of 205,115 SNVs, 3787 insertions, and 3055 deletion calls per aligner/variant caller combination. Within these variant lists, SNVs had the lowest false positive and false negative rates (9.09% and 2.90% respectively), followed by deletions (35.13% FP and 23.72% FN), and insertions (35.36% FP and 19.40% FN). Considering the intersection of all variants (i.e. only unanimously called variants) generated 106,594 unique SNV calls, 1415 unique insertion calls, and 1158 unique deletion calls, all of which has lower false positive rates and higher false negative rates as expected. The false positive rate was 3.29% for SNVs, 14.68% for deletions, and 12.16% for insertions and the false negative rate was 6.78% for SNVs, 35.59% for deletions, and 42.26% for insertions. Finally, the variants unique to a single pair contained large numbers of false positives for SNVs (46.43% FP), deletions (63.80% FP), and insertions (73.55% FP).

Heterogeneous Sample Simulation

Using BWA alignments filtered SNV lists for GATK, Isaac variant caller, and SAMtools were generated at simulated contamination levels of 0%, 25%, 50%, 75%, 90%, 95%, 98%, and 99% (Table 5). Each variant caller detected substantially fewer variants as contamination levels increased resulting in increasing false negative rates. The increase in the false negative rate was not consistent across all variant callers however, with GATK still able to detect 60% of all SNVs using only 10% of the original GIAB data compared to only 45% for Isaac variant caller and 35% for SAMtools. In general, GATK coped better with higher contamination levels than either SAMtools or Isaac variant caller with similar patterns observed for both deletions and insertions.

Clinically Important Variants

All variants calls from melanoma cell line control C001 were overlapped to ClinVar yielded 2266 matching SNVs and 9 matching deletions, all but four of which corresponded to existing dbSNP entries. Of the 2266 SNV calls, 1894 were unanimously detected (83.6%), with the remaining 372 (16.4%) missed by at least one software pair. From the ClinVar annotations, three variants were annotated as known melanoma risk factors with only one of such variants unanimously detected by all software combinations (Table 6).

Discussion

Ongoing efforts to characterize genetic variants for clinical action [40] and the ever-increasing number of previously characterized variants routinely being used for decisions in the clinic [41, 42] illustrate the potential importance of a single variant. From the analysis of the melanoma cell line control, we identified three important melanoma risk factor variants, two of which were not detected unanimously under all software conditions (Table 6). One missed risk factor variant, rs861539, was not detected by Isaac aligner when paired with any of the three variant callers (SAMtools assigned a failing variant score of 11, Isaac variant caller assigned a failing variant score of 8, and GATK did not flag the base as variant) highlighting the danger of relying on a single tool for any analysis step. The other missed risk factor variant, rs1126809, was missed by a single software combination (Bowtie2 and Isaac variant caller with filtering applied) as a result of the variant being annotated as low quality, illustrating the danger of applying aggressive filtering when considering variation data in a clinical context. While these

Table 5. SNV calls for GIAB data at simulated contamination levels.

Variant Caller	Simulated Contamination Level	Variant Calls	False Positive Rate[a]	False Negative Rate
GATK	0%	228239	6.62%	3.23%
	25%	163041	6.22%	3.76%
	50%	113263	5.50%	5.54%
	75%	69416	4.77%	13.45%
	90%	34018	4.13%	40.74%
	95%	17321	4.08%	67.55%
	98%	5662	4.06%	88.87%
	99%	1696	4.25%	96.46%
Isaac VC	0%	79366	4.51%	7.90%
	25%	73259	4.26%	12.84%
	50%	61458	4.01%	18.56%
	75%	43091	3.90%	31.59%
	90%	23146	3.67%	54.76%
	95%	10888	3.77%	76.29%
	98%	3246	4.13%	92.60%
	99%	887	5.41%	97.96%
SAMtools	0%	106332	3.78%	6.39%
	25%	83337	3.66%	9.17%
	50%	63382	3.46%	14.76%
	75%	39853	3.29%	30.92%
	90%	16660	3.16%	64.83%
	95%	6365	3.03%	85.64%
	98%	1534	3.46%	96.45%
	99%	371	6.47%	99.18%

BWA alignments were used to generate filtered SNV lists for GATK, Isaac variant caller, and SAMtools at simulated contamination levels of 0%, 25%, 50%, 75%, 90%, 95%, 98%, and 99%. Variant lists were overlapped to GIAB high quality variants to determine false positive and false negative rates.
[a]From bases overlap ENSEMBL v75 canonical transcripts.

doi:10.1371/journal.pone.0143199.t005

two variants are significant in this instance, broader analysis of the GIAB data set showed that all aligners and variant callers assessed in this study fail to detect true variants that are detected by other algorithms, implying clinically important variants may be missed even when the top performing software is utilized. Given the importance of detecting clinically important variants accurately, the utility of first optimizing filtering strategies for individual algorithms and then

Table 6. Melanoma cell line control variant calls overlapping annotated ClinVar melanoma risk factors.

GRCh37 Coordinate (dbSNP)	dbSNP id	Missing Aligner / Variant Caller Pair (F = filtered, U = unfiltered)	ClinVar Annotation
5:33951693	rs16891982	None	Malignant melanoma of skin
11:89017961	rs1126809	Bowtie2/Isaac_vc (F)	Increased risk of cutaneous melanoma
14:104165753	rs861539	Isaac_aligner/GATK (U, F) Isaac_aligner/Isaac_vc (U, F) Isaac_aligner/ SAMtools (U, F)	Increased risk of cutaneous melanoma

Variant calls from melanoma cell line C001 were overlapped to ClinVar and all annotated melanoma risk factors examined. Software pairs failing to detect these variants are reported in column 3 with variants listed as unfiltered (U) or filtered (F) to reflect whether variant caller filtering was applied.

doi:10.1371/journal.pone.0143199.t006

combining variant calls from multiple variant callers is apparent as the increased output quality quickly justifies any increased computation.

Using the GIAB reference data, we compared the false positive and false negative rates of three short-read aligners paired with three variant callers run both with and without variant caller filtering for SNVs, deletions, and insertions (S2–S4 Tables). Here we confirmed previous work that the choice of aligner, variant caller, and variant caller filtering affects both the number of variants detected and their subsequent quality [11, 21, 23]. To better assess the performance of each individual algorithm in isolation the union of all variants generated by each tool was calculated for SNVs (Table 1), deletions (Table 2), and insertions (Table 3). While each algorithm tested caused an observable effect on variant call quantity and quality, this effect differed with regard to the type of variant, the overall false positive and false negative rates, and the quality of the tool-specific variants. For total variant calls, the variant caller choice had a greater impact than the aligner choice with 36% more SNVs called by GATK than Isaac variant caller, 51% more deletions called by SAMtools than GATK, and 127% more insertions called by SAMtools than GATK. While the choice of aligner had less impact than the choice of variant caller on total variant number, the effect was still significant with 32% more SNVs called by BWA than Isaac aligner, 24% more deletions called by BWA than Isaac aligner, and 7% more insertions called by Bowtie2 than BWA. The false positive rate also differed significantly ranging for SNVs from 3.84% for SAMtools to 8.55% for BWA, for deletions from 25.5% for Isaac aligner to 33.26% for Bowtie2, and for insertions from 19.26% for Isaac variant caller to 34.87% for SAMtools. Similarly, the false negative rate varied significantly ranging for SNVs from 2.91% for GATK to 7.47% for Isaac variant caller, for deletions from 23.73% for BWA and Bowtie2 to 35.59% for Isaac variant caller, and for insertions from 20.90% for SAMtools to 37.31% for GATK. When considering all variants called the range in quality is large, however the range is even greater when considering tool-specific variants. Tool-specific variants represent an important variant subset as they serve to highlight differences amongst the individual algorithm, with such variants being important to consider particularly when minimizing false calls is a priority. For example, there are 59,275 GATK-specific SNVs of which almost 85% are true variants meaning these will be missed unless GATK is utilized for SNV calling. While GATK seems an excellent choice for SNV detection, it calls substantially less tool-specific indels than SAMtools meaning no single algorithm is optimal for minimizing false calls across all variant types. Such results demonstrate the utility of tool-specific variants for making informed decisions about software selection; for example the consistently low false positive rate of ~15% for all types of GATK-specific variants led us to incorporate it into our production system. Collectively these results illustrate the significant and often-unpredictable effects the choice of aligner and variant caller has on variant call quality and highlights the importance of ongoing software appraisal and optimization.

Another important factor known to affect variant call quality is the filtering strategy applied by each of the variant caller software suites [17, 18]. Each variant caller employs a distinctive strategy for filtering; GATK uses variant quality score recalibration (VQSR), SAMtools uses BAQ filtering, and Isaac aligner annotates questionable variants having low scoring genotype quality as 'lowGQX'. While these filtering strategies differ significantly algorithmically, they share the common goal of trying to remove false positive variants from the original variant lists. In our analysis, applying these filters mostly resulted in a reduced number of total calls as expected, with the exception of BAQ filtering which resulted in a slight increase in the number of indel calls as might be expected given it's focus on removing false positive SNV calls around candidate indels. Interestingly, only GATK VQSR filtering reduced the false positive rate and increased the false negative rate across all variant types with both SAMtools and Isaac variant caller filtering yielding increased false positive rates for indels indicating potential issues with

their respective indel filtering strategies. While the reduction in false positive rate is important, from a clinical context it is important to note that all filtering strategies removes true variants; ranging from as few as 73 true positive SNVs with GATK to the extreme case of the Isaac variant caller filtering removing over 100,000 true positive SNVs. Overall, these results highlight the large effect software choices make in both the precision and recall of variants.

In our analysis, no aligner / variant caller pair outperformed all other pairs across all variant types, illustrating the importance of selecting software specific to each variant type. In general, SNV calls yielded significantly lower false positive and false negative rates compared to indel calls with unanimously detected SNVs having FP/FN rates of 3.29% and 6.78% compared to 14.68% and 35.59% for deletions and 12.16% and 46.26% for insertions (Table 4). This difference is even more apparent when considering the union of all SNV calls with FP/FN rates of 9.09% and 2.90% compared to 35.13% and 23.72% for deletions and 35.36% and 19.40% for insertions. The higher false positive rate of indel calls might be explained if the GIAB reference set was under-reporting indels however this seems unlikely with GIAB reporting a SNV to indel ratio of 6:1 (2.89 million SNVs and ~465,000 million indels), a lower ratio than the 10:1 estimated by the 1000 genomes project [43]. A more likely explanation is that indels are scarcer than SNVs and more difficult to detect due to challenges of aligning short reads around indels [44] meaning the distinct patterns in alignments of short-read data requires distinct workflows for SNV and indel detection [35]. A final possibility is that the variant callers assessed in this analysis do not employ the most up to date methodologies for indel detection with new tools such as Scalpel [45] and ABRA [46] utilizing microassembly in the detection of indels. Such tools are reporting lower false positive rates than existing software and likely do offer improvements in the overall quality of indel calls. Regardless, with variant detection software rapidly improving, the need to select software specific to each variant type is clear.

Another increasingly common application utilizing variation data in the clinic is the use of cancer samples as a way of advancing personalized treatment of cancer [47]. Working with cancer samples has additional complications however, with factors such as sample contamination known to be a problem [27]. To measure the effect of contamination on variant detection, an additional analysis was undertaken where GIAB sequence data was replaced with increasing levels of non-variant reference data to simulate increasing levels of contamination (Table 5).

For simplicity, only the filtered SNV calls from GATK, Isaac variant caller, and SAMtools were assessed and as expected, we observe large increases in false negative rates as the contamination level increases. The ability to cope with contamination differed for the three variant callers however with GATK outperforming the others significantly; for example at 50% contamination levels GATK had a false negative rate of only 5.54% compared to 14.76% for SAMtools and 18.56% for Isaac variant caller. These results suggest additional considerations are required when analyzing samples likely to suffer from contamination, as is often the case for cancer samples. While this study focused on the detection of germ line variation, the nature of variation in cancer is fundamentally different from non-cancer with much of the current thinking based on the understanding that a clone accumulates somatically acquired mutations that ultimately leads to malignant transformation [48], with inherited germ line mutations thought to be important in only 5–10% of cancers [49]. Somatic mutation detection is not addressed in this study, however the design of this study could be easily applied to assess software such as MuSiC [50] or MuTect [51], software specifically designed to detect somatic mutations in paired samples.

While we have demonstrated the impact the choice of software, filtering strategy, and combined variant calls have on resultant variant calls, differences in sample and library preparation as well as coverage levels are also known to significantly affect variant calling [52]. For sample preparation, differences in input DNA amount, sample age, and sample preservation method

are known to be important as are library issues such as PCR amplification errors, primer biases, chimeric reads, and barcode/adaptor errors. Depth of coverage has also been shown to have a large impact on variant detection [53] and is of particular importance when considering contaminated samples such as tumors, with additional coverage typically required to compensate for contamination. While in this study it is not feasible to exhaustively examine the impact of all such factors on variant detection, the sequencing of larger and larger number of samples will allow us to begin to understand the effect of such factors on variant detection.

Finally, we present a cogent strategy for creating a variant detection pipeline for clinical use that focuses on minimizing the total number of false negative variants from the onset. For all workflows, a single aligner is sufficient with BWA generating the greatest number of both SNV and deletion calls and only 7% less insertion calls than Bowtie2. In this study, we show the single most effective measure for minimizing false negative variants is combining the results from multiple variant callers. Implementing this approach requires running multiple variant callers in parallel and combining the results to take forward for analysis, an approach already implemented by tools such as BAYSIC [26]. While combined variant calls is the ideal strategy, it must be acknowledged that computational resources are often limited in which case it is preferable to run BWA paired with GATK due to the low false negative rates, particularly for SNVs and deletions. Finally, when sample contamination is likely to be an issue (as in cancer samples) GATK should be utilized as it outperforms Isaac variant caller and SAMtools at increasing contamination levels. While the strategy presented represents the most effective use of the tools assessed in this study, it is important to routinely reassess both new software and new versions of existing software to remain current in this fast moving field. Doing this routinely requires that any framework designed to support clinical usage of variation information must be able to easily run and benchmark a wide variety of software. Further, such a framework must be able to run multiple variant callers and combine their output, a feature missing in almost all modern pipeline frameworks.

Conclusion

Analysis of the GIAB reference dataset shows that the choice of aligner, variant caller, and variant caller filtering strategy significantly affects the quality of variant calls and that true variants can be missed by individual software or removed during variant caller filtering. Analysis of the melanoma cell line control shows only one of three melanoma risk factor variants is detected unanimously with one variant missed completely by a single software tool and the other variant removed during variant caller filtering. These results demonstrate the importance of developing a strategy based on reducing false negative variants when utilizing variation data in a clinical context; a strategy that requires careful software selection, variant caller filtering optimization, and combined variant calls from multiple variant callers.

Supporting Information

S1 Fig. Tracking database schema. Tracking database schema generated using MySQL Workbench. The database records all sample metadata, sequence data information, and the analysis steps performed. Any previous analysis can be completely reproduced solely from the information contained in the database.
(TIFF)

S2 Fig. Default pipeline workflow. Default workflow for the production in-house pipeline. When new data is received the metadata is parsed to determine whether the sample is new and if so, what type of sample it is with available options for single human, single mouse, human

pedigree, or human cancer. If new data is added to an existing sample it is linked to the original data and a new analysis run commences.
(TIFF)

S1 File. Detailed Variant Detection Workflow. Detailed description of design and implementation of high-throughput in-house variant detection pipeline.
(DOCX)

S1 Table. Short read aligner and variant caller versions and commands. All short read aligners and variant callers commands utilized in our example. The commands listed match the exact commands run with the exception of the shortening of file names. The commands were chosen by either following documentation suggestions, or else by using default options.
(DOCX)

S2 Table. SNV call overlaps with GIAB. SNV stats from all eighteen possible software combinations derived from the pairing of the three aligners with each of the three variant callers run both with and without filtering. SNVs were overlapped to GIAB SNVs and false positive and false negative rates calculated.
(DOCX)

S3 Table. Deletion call overlaps with GIAB. Deletion stats from all eighteen possible software combinations derived from the pairing of the three aligners with each of the three variant callers run both with and without filtering. Deletions were overlapped to GIAB deletions and false positive and false negative rates calculated.
(DOCX)

S4 Table. Insertion call overlaps with GIAB. Insertion stats from all eighteen possible software combinations derived from the pairing of the three aligners with each of the three variant callers run both with and without filtering. Insertions were overlapped to GIAB insertions and false positive and false negative rates calculated.
(DOCX)

Acknowledgments

We thank the National Computational Infrastructure (Australia) for continued access to significant computation resources and technical expertise. We also would like to thank Queensland Institute of Medical Research for access to the melanoma cell line generated as part of the larger Australian Melanoma Genome Project.

Author Contributions

Conceived and designed the experiments: MAF VC TDA CCG. Performed the experiments: MAF. Analyzed the data: MAF. Contributed reagents/materials/analysis tools: MAF. Wrote the paper: MAF TDA CCG. Obtained permission for use of cell line: MAF.

References

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends in genetics: TIG. 2014; 30(9):418–26. doi: 10.1016/j.tig.2014.07.001 PMID: 25108476.
2. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in bioinformatics. 2014; 15(2):256–78. doi: 10.1093/bib/bbs086 PMID: 23341494; PubMed Central PMCID: PMC3956068.
3. Brownstein CA, Beggs AH, Homer N, Merriman B, Yu TW, Flannery KC, et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical

genome sequencing results in the CLARITY Challenge. Genome biology. 2014; 15(3):R53. doi: 10.1186/gb-2014-15-3-r53 PMID: 24667040; PubMed Central PMCID: PMC4073084.

4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010; 20(9):1297–303. doi: 10.1101/gr.107524.110 PMID: 20644199; PubMed Central PMCID: PMC2928508.
5. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology. 2010; 11(8):R86. doi: 10.1186/gb-2010-11-8-r86 PMID: 20738864; PubMed Central PMCID: PMC2945788.
6. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004; 20(17):3045–54. doi: 10.1093/bioinformatics/bth361 PMID: 15201187.
7. Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics. 2012; 28(11):1525–6. doi: 10.1093/bioinformatics/bts167 PMID: 22500002.
8. Koster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012; 28(19):2520–2. doi: 10.1093/bioinformatics/bts480 PMID: 22908215.
9. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004; 5(10):R80. doi: 10.1186/gb-2004-5-10-r80 PMID: 15461798; PubMed Central PMCID: PMC545600.
10. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome medicine. 2013; 5(3):28. doi: 10.1186/gm432 PMID: 23537139; PubMed Central PMCID: PMC3706896.
11. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Human genomics. 2014; 8:14. doi: 10.1186/1479-7364-8-14 PMID: 25078893; PubMed Central PMCID: PMC4129436.
12. Chilamakuri CS, Lorenz S, Madoui MA, Vodak D, Sun J, Hovig E, et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics. 2014; 15:449. doi: 10.1186/1471-2164-15-449 PMID: 24912484; PubMed Central PMCID: PMC4092227.
13. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nature biotechnology. 2011; 29(10):908–14. doi: 10.1038/nbt.1975 PMID: 21947028; PubMed Central PMCID: PMC4127531.
14. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. A comparative analysis of exome capture. Genome biology. 2011; 12(9):R97. doi: 10.1186/gb-2011-12-9-r97 PMID: 21958622; PubMed Central PMCID: PMC3308060.
15. Lam HY, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, et al. Performance comparison of whole-genome sequencing platforms. Nature biotechnology. 2012; 30(1):78–82. doi: 10.1038/nbt.2065 PMID: 22178993; PubMed Central PMCID: PMC4076012.
16. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13:341. doi: 10.1186/1471-2164-13-341 PMID: 22827831; PubMed Central PMCID: PMC3431227.
17. Jia P, Li F, Xia J, Chen H, Ji H, Pao W, et al. Consensus rules in variant detection from next-generation sequencing data. PloS one. 2012; 7(6):e38470. doi: 10.1371/journal.pone.0038470 PMID: 22715385; PubMed Central PMCID: PMC3371040.
18. Reumers J, De Rijk P, Zhao H, Liekens A, Smeets D, Cleary J, et al. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nature biotechnology. 2012; 30(1):61–8. doi: 10.1038/nbt.2053 PMID: 22178994.
19. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. BioMed research international. 2014; 2014:309650. doi: 10.1155/2014/309650 PMID: 24779008; PubMed Central PMCID: PMC3980841.
20. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome biology. 2013; 14(5):R51. doi: 10.1186/gb-2013-14-5-r51 PMID: 23718773; PubMed Central PMCID: PMC4053816.
21. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, Schuster SC. Comparison of sequencing platforms for single nucleotide variant calls in a human sample. PloS one. 2013; 8(2):e55089. doi: 10.1371/journal.pone.0055089 PMID: 23405114; PubMed Central PMCID: PMC3566181.

22. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature biotechnology. 2014; 32(3):246–51. doi: 10.1038/nbt.2835 PMID: 24531798.

23. Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM. Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. Nucleic acids research. 2014; 42(12):e101. doi: 10.1093/nar/gku392 PMID: 24831545; PubMed Central PMCID: PMC4081058.

24. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant callers for next-generation sequencing data: a comparison study. PloS one. 2013; 8(9):e75619. doi: 10.1371/journal.pone.0075619 PMID: 24086590; PubMed Central PMCID: PMC3785481.

25. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, et al. An analytical framework for optimizing variant discovery from personal genomes. Nat Commun. 2015; 6:6275. doi: 10.1038/ncomms7275 PMID: 25711446; PubMed Central PMCID: PMC4351570.

26. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC bioinformatics. 2014; 15:104. doi: 10.1186/1471-2105-15-104 PMID: 24725768; PubMed Central PMCID: PMC3999887.

27. Liotta L, Petricoin E. Molecular profiling of human cancer. Nat Rev Genet. 2000; 1(1):48–56. doi: 10.1038/35049567 PMID: 11262874.

28. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013; 501(7467):338–45. doi: 10.1038/nature12625 PMID: 24048066.

29. Yates LR, Campbell PJ. Evolution of the cancer genome. Nat Rev Genet. 2012; 13(11):795–806. doi: 10.1038/nrg3317 PMID: 23044827; PubMed Central PMCID: PMCPMC3666082.

30. Andrews TD, Whittle B, Field MA, Balakishnan B, Zhang Y, Shao Y, et al. Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. Open biology. 2012; 2(5):120061. doi: 10.1098/rsob.120061 PMID: 22724066; PubMed Central PMCID: PMC3376740.

31. Field MA, Cho V, Cook MC, Enders A, Vinuesa C, Whittle B, et al. Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees. Bioinformatics. 2015. doi: 10.1093/bioinformatics/btv135 PMID: 25755272.

32. Wilmott JS, Field MA, Johansson PA, Kakavand H, Shang P, De Paoli-Iseppi R, et al. Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes. Pathology. 2015. doi: 10.1097/PAT.0000000000000324 PMID: 26517638.

33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. doi: 10.1093/bioinformatics/btp324 PMID: 19451168; PubMed Central PMCID: PMC2705234.

34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9(4):357–9. doi: 10.1038/nmeth.1923 PMID: 22388286; PubMed Central PMCID: PMC3322381.

35. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013; 29(16):2041–3. doi: 10.1093/bioinformatics/btt314 PMID: 23736529.

36. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27(21):2987–93. doi: 10.1093/bioinformatics/btr509 PMID: 21903627; PubMed Central PMCID: PMC3198575.

37. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC bioinformatics. 2011; 12:35. doi: 10.1186/1471-2105-12-35 PMID: 21269502; PubMed Central PMCID: PMC3041657.

38. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21(20):3940–1. doi: 10.1093/bioinformatics/bti623 PMID: 16096348.

39. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014; 42(Database issue):D980–5. doi: 10.1093/nar/gkt1113 PMID: 24234437; PubMed Central PMCID: PMC3965032.

40. Ramos EM, Din-Lovinescu C, Berg JS, Brooks LD, Duncanson A, Dunn M, et al. Characterizing genetic variants for clinical action. Am J Med Genet C Semin Med Genet. 2014; 166C(1):93–104. doi: 10.1002/ajmg.c.31386 PMID: 24634402; PubMed Central PMCID: PMC4158437.

41. Jeck WR, Parker J, Carson CC, Shields JM, Sambade MJ, Peters EC, et al. Targeted next generation sequencing identifies clinically actionable mutations in patients with melanoma. Pigment Cell

Melanoma Res. 2014; 27(4):653–63. doi: 10.1111/pcmr.12238 PMID: 24628946; PubMed Central PMCID: PMC4121659.

42. Thomas A, Rajan A, Lopez-Chavez A, Wang Y, Giaccone G. From targets to targeted therapies and molecular profiling in non-small cell lung carcinoma. Ann Oncol. 2013; 24(3):577–85. doi: 10.1093/annonc/mds478 PMID: 23131389; PubMed Central PMCID: PMC3574546.

43. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. doi: 10.1038/nature09534 PMID: 20981092; PubMed Central PMCID: PMC3042601.

44. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome research. 2011; 21(6):961–73. doi: 10.1101/gr.112326.110 PMID: 20980555; PubMed Central PMCID: PMC3106329.

45. Fang H, Wu Y, Narzisi G, O'Rawe JA, Barron LT, Rosenbaum J, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome medicine. 2014; 6(10):89. doi: 10.1186/s13073-014-0089-z PMID: 25426171; PubMed Central PMCID: PMC4240813.

46. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. ABRA: improved coding indel detection via assembly-based realignment. Bioinformatics. 2014; 30(19):2813–5. doi: 10.1093/bioinformatics/btu376 PMID: 24907369; PubMed Central PMCID: PMC4173014.

47. Guan YF, Li GR, Wang RJ, Yi YT, Yang L, Jiang D, et al. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. Chin J Cancer. 2012; 31(10):463–70. doi: 10.5732/cjc.012.10216 PMID: 22980418; PubMed Central PMCID: PMCPMC3777453.

48. Burnet M. Somatic Mutation and Chronic Disease. Br Med J. 1965; 1(5431):338–42. PMID: 14237898; PubMed Central PMCID: PMC2165357.

49. Nagy R, Sweet K, Eng C. Highly penetrant hereditary cancer syndromes. Oncogene. 2004; 23(38):6445–70. doi: 10.1038/sj.onc.1207714 PMID: 15322516.

50. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome research. 2012; 22(8):1589–98. doi: 10.1101/gr.134635.111 PMID: 22759861; PubMed Central PMCID: PMC3409272.

51. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013; 31(3):213–9. doi: 10.1038/nbt.2514 PMID: 23396013; PubMed Central PMCID: PMC3833702.

52. Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet. 2014; 15(1):56–62. doi: 10.1038/nrg3655 PMID: 24322726; PubMed Central PMCID: PMCPMC4103745.

53. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15(2):121–32. doi: 10.1038/nrg3642 PMID: 24434847.

S1 Text. Detailed Variant Detection Workflow

Sequence data for each sample is initially copied to a staging area on raijin (http://nci.org.au/nci-systems/national-facility/peak-system/raijin/) at the National Computational Infrastructure, the 38th largest supercomputer in the world (http://www.top500.org/). Metadata for each sample is loaded from an external sample tracking database and read directories created automatically with all sequence data copied from the staging area to the read directory using regular expressions to determine the destination for each file. Next, sample metadata is loaded into a master configuration file containing information such the sample project, type of sample (cancer, pedigree, single, mouse, etc), read directory, and other relevant metadata specific to each sample. A cronjob running on the headnode of raijin checks for changes to the master config file and when new samples are identified the metadata is loaded into an internal tracking database (**Fig S1**) with all necessary HPC job files created simultaneously (similar to a tool like COSMOS (1)) and the initial BWA alignment jobs submitted (2). In order to reduce the overall run time multiple samples are analyzed in parallel, making the sole-limiting factor the amount of available compute resources. If the sample contains more than one lane of sequence data each lane is analyzed independently which offers two main advantages; lane-specific QC (which frequently serves in preventing corrupt lanes being merged with otherwise uncorrupted BAM files) as well as allowing new lanes to be added to existing samples at a later date if higher coverage levels are desired. To generate a BAM files, repetitively aligned reads are filtered out (identified by the XT:A:R tag) and a sorted BAM file generated using SAMtools (3). Alignment statistics are generated for each BAM file using the samtools flagstat command and only lanes with 90% or more of the reads aligning pass lane QC. When all individual lanes for a sample have passed QC, BAM files are merged into a single BAM file and run through samtools (3) rmdup to remove candidate PCR duplicates. Variants are next called using either samtools/bcftools (4) or GATK (5) best practices. Jobs are split up into chromosome-by-chromosome jobs (using a map-reduce like approach (6)), which run independently while a cronjob running on the headnode detecting when all jobs are complete and proceeds to merge the chromosome-by-chromosome results and submit the next job to run. Variants are annotated using ENSEMBL variant effect predictor (VEP) (7) and overlapped with ENSEMBL canonical transcripts and splice site variants, defined as the 10bp either side of a coding exon. Additional annotations not

available from standard annotation tools are next added, annotations such as ExAC (http://exac.broadinstitute.org/about) or mouse phenotype information from MGI (http://www.informatics.jax.org/). Lastly, variant reports for SNVs and indels are generated and prioritized by several measures such as novel or low frequency variants, nonsense and missense mutations, high polyphen2 (8) and low SIFT (9) scores amongst others. Numerous custom filtering options to reduce the variant search space with filters based on genomic region, gene name, population allele frequencies, and polyphen/SIFT score are also available.

For running individual system commands, a generic framework was developed to reflect the fact that any large workflow can be broken down into smaller analysis components and ultimately individual system commands. To run a single analysis step, the wrapper first reads the sample's configuration file to retrieve relevant step information such as job pre-requisites, command arguments, and compute requirements. After checking prerequisite conditions have been met, job(s) are submitted and monitored until completion when step-specific QC is performed with the exact command recorded in both the tracking database (**Fig S1**) and the detailed log file. The final task performed by the wrapper is the submission of the next analysis job, a process that continues until the final analysis step is completed.

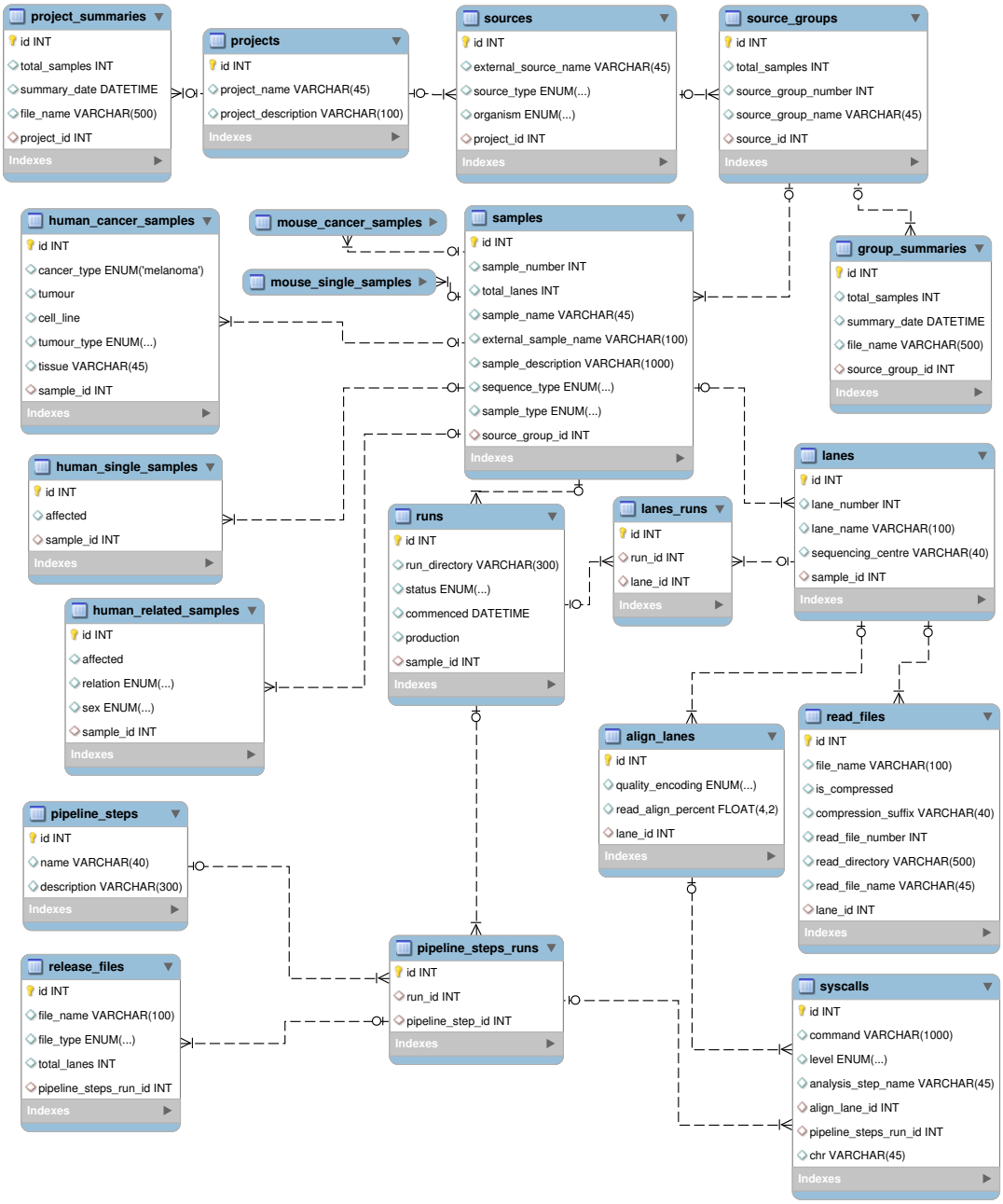
The detection and recovery from both processing and hardware errors is also implemented in our system, using an approach similar to eHive (10). The system works to detect both catastrophic errors (errors that cause the entire analysis to halt) and non-catastrophic errors (errors that do not cause analysis to halt but generate incorrect output). Non-catastrophic errors can be particularly problematic both in wasting ensuing CPU cycles and in making it difficult to pinpoint the exact point of failure in long running analysis.

Importantly, the system output is entirely reproducible, a feature which cannot be assumed in biological sciences as highlighted by a study discovering less than half of the microarray studies published in Nature Genetics were reproducible (11). To achieve reproducibility (a main goal of the GALAXY framework (12)), our system employs multiple tracking methods, an underlying mysql database and a detailed log file, with each method capable of independently reproducing previous results. While this design decision was initially implemented to ensure output consistency, it has proven useful in diagnosing pipeline crashes by offering two distinct reference points for pinpointing the exact point of failure. Full reproducibility requires more than

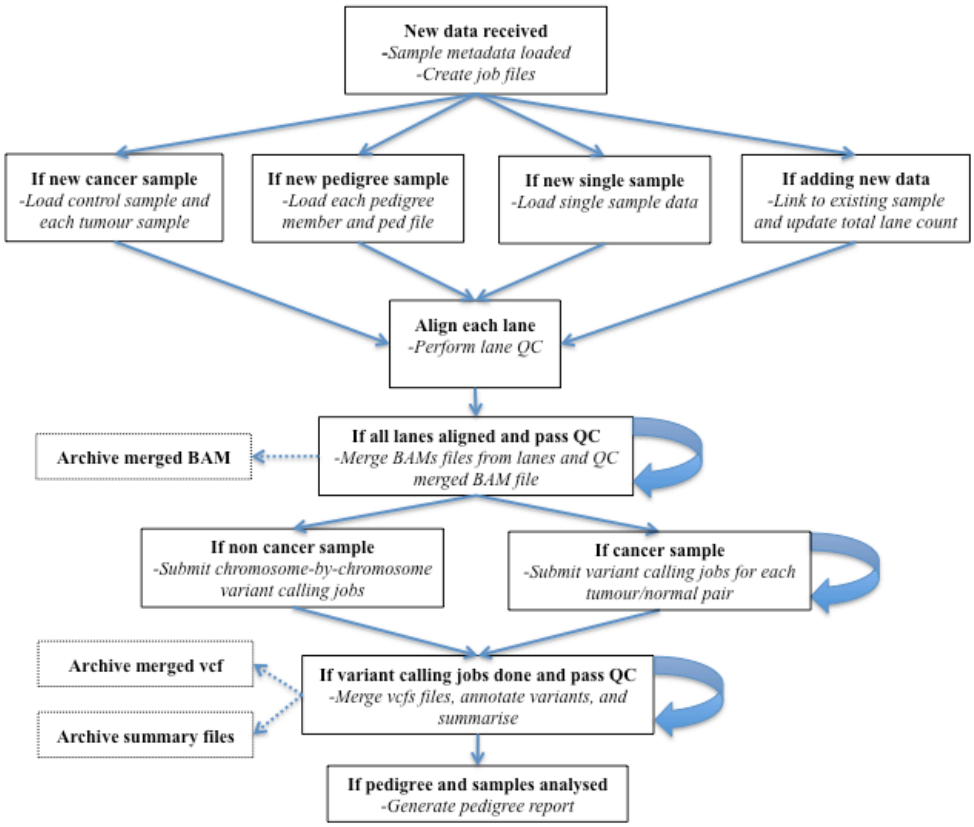
detailing how commands were run, it requires the versioning of all files and binaries with all relevant version numbers recorded. This applies to the version of external binaries, the version of the larger code base, and the version of external annotation data sets. External annotation data in particular represents a challenge for reproducibility with rapid updates and changes to format typical of these external data sets. To manage such annotations consistently, all annotations are read from the original data source and stored locally in a standardized file format with standard file names. Annotation files from a single source are stored in time-stamped directories thus allowing the most up to date information to be automatically detected and utilized during analysis. This consistent handling allows for seamless updates as new annotations become available while additionally simplifying batch re-analysis when important annotation sets like dbSNP (13) are updated.

While here we describe the default workflow (**Fig S2**), the system is flexible with regard to software selection thus providing a ready-made testing environment, a feature currently available with tools like Nestly (14). This testing environment can be utilized both in determining optimal parameters for individual algorithms but also in determining effective software combinations by testing aligners and variant callers together for example. Currently a selection of some of the most utilized aligners and variant callers are available to facilitate such analyses (bowtie2 (15), BWA (2), isaac (16), GATK (5), isaac_variant_caller (16), and SAMtools (3)).

To address the often-low concordance rates between variant callers (17) and variant calling pipelines (18) our system allows the use of multiple variant callers with options for taking forward either the union or the intersection of raw variant calls depending on project-specific circumstances, similar to the approach taken by BAYSIC (19). In general taking the union of variants gives higher false positive rates and lower false negative rates while taking the intersection of variants gives lower false positive rates and higher false negative rates.



S1 Figure. Tracking database schema. Tracking database schema generated using MySQL Workbench. The database records all sample metadata, sequence data information, and the analysis steps performed. Any previous analysis can be completely reproduced solely from the information contained in the database.



S2 Figure. Default pipeline workflow. Default workflow for the production in-house pipeline. When new data is received the metadata is parsed to determine whether the sample is new and if so, what type of sample it is with available options for single human, single mouse, human pedigree, or human cancer. If new data is added to an existing sample it is linked to the original data and a new analysis run commences.

S1 Table. Short read aligner and variant caller versions and commands.

Software	Version	Commands
bowtie2	2.2.3	1) bowtie2 –p 16 -1 fq1 -2 fq2 –x hg19 | samtools view –bt hg19.fa.fai - | samtools sort - > bt2.bam
bwa	0.7.10	1) bwa aln –t 16 hg19 fq1 > r1.sai 2) bwa aln –t 16 hg19 fq2 > r2.sai 3) bwa sampe hg19 r1.sai r2.sai fq1 fq2 | samtools view –bt hg19.fa.fai - | samtools sort - > bwa.bam
isaac aligner	01.14.08.28	1) isaac-align -r sorted_reference.xml --base-calls-format fastq-gz –j 16 –m 40 –keep-aligned back –realign-gaps yes -o isaac_align –b readdir
GATK (raw)	3.2.2	1) java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -I bwa.bam -o bwa_gatk.vcf -R hg19.fa –glm BOTH -metrics bwa_gatk.metrics -rf MappingQuality -mmq 10
GATK	3.2.2	1) java -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R hg19.fa -input bwa_gatk.vcf -recalFile bwa_gatk.snv.recal -tranchesFile

(VQSR)		bwa_gatk.snv.tranches - resource:hapmap,known=false,training=true,truth=true,prior=15.0 hapmap_3.3.hg19.sites.vcf - resource:omni,known=false,training=true,truth=true,prior=12.0 1000G_omni2.5.hg19.sites.vcf - resource:1000G,known=false,training=true,truth=false,prior=10.0 1000G_phase1.snps.high_confidence.hg19.sites.vcf - resource:dbsnp,known=true,training=false,truth=false,prior=2.0 dbsnp_138.hg19.vcf -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an DP -mode SNP 2) java -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R hg19.fa - input bwa_gatk.def.vcf -tranchesFile bwa_gatk.snv.tranches -recalFile bwa_gatk.snv.recal -o bwa_gatk.vqsr.snv.vcf --ts_filter_level 99.5 - mode SNP 3) java -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R hg19.fa - input bwa_gatk.vcf -recalFile bwa_gatk.indel.recal -tranchesFile bwa_gatk.indel.tranches - resource:mills,known=false,training=true,truth=true,prior=12.0 Mills_and_1000G_gold_standard.indels.hg19.sites.vcf - resource:dbsnp,known=true,training=false,truth=false,prior=2.0 dbsnp_138.hg19.vcf --maxGaussians 4 -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an DP -mode INDEL 4) java -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R hg19.fa - input bwa_gatk.def.vcf -tranchesFile bwa_gatk.indel.tranches - recalFile bwa_gatk.indel.recal -o bwa_gatk.vqsr.indel.vcf -- ts_filter_level 99.0 -mode INDEL
isaac variant caller (raw)	1.0.6	1) configureWorkflow.pl –bam bwa.bam –ref hg19.fa --config config.ini 2) make –j 16
Isaac variant caller (No LowGQX variants)	1.0.6	1) grep -v LowGQX bwa_isaac.vcf > bwa_isaac.noLowGQX.vcf
Samtools (no BAQ filtering)	0.1.18	1) samtools mpileup -C50 -uDBf hg19.fa bwa.bam | bcftools view -vcg - > bwa_samools.noBAQ.vcf
Samtools (BAQ filtering)	0.1.18	1) samtools mpileup -C50 -uDf hg19.fa bwa.bam | bcftools view -vcg - > bwa_samools.vcf

All short read aligners and variant callers commands utilized in our example. The commands listed match the exact commands run with the exception of the shortening of file names. The commands were chosen by either following documentation suggestions, or else by using default options.

S2 Table. SNV call overlaps with GIAB.

Aligner	Variant Caller	Total SNVs	SNVs Match GIAB	% SNVs Match GIAB
Bowtie2	GATK (raw)	224436	208759	93.01%
Bowtie2	GATK (filtered)	223085	208232	93.34%
Bowtie2	isaac (raw)	159982	150434	94.03%
Bowtie2	Isaac (filtered)	74316	71354	96.01%
Bowtie2	Samtools (filtered)	105900	101920	96.24%
Bowtie2	Samtools (raw)	166722	156807	94.05%
BWA	GATK (raw)	225546	210043	93.13%
BWA	GATK (filtered)	224671	209788	93.38%
BWA	isaac (raw)	166528	156026	93.69%
BWA	isaac (filtered)	78002	74484	95.49%
BWA	Samtools (filtered)	106322	102301	96.22%
BWA	Samtools (raw)	167295	157330	94.04%
isaac_	GATK (raw)	182928	175666	96.03%
isaac	GATK (filtered)	182703	175546	96.08%
isaac	isaac (raw)	129105	124657	96.55%
isaac	isaac (filtered)	70499	67964	96.40%
isaac	Samtools (filtered)	82164	79728	97.04%
isaac	Samtools (raw)	120000	115689	96.41%

Snv stats from all eighteen possible software combinations derived from the pairing of the three aligners with each of the three variant callers run both with and without filtering. SNVs were overlapped to GIAB SNVs and the percent matching calculated.

S3 Table. Deletion call overlaps with GIAB.

Aligner	Variant Caller	Total Deletions	Deletions Match GIAB	% Deletions Match GIAB
Bowtie2	GATK (raw)	1997	1455	72.86%
Bowtie2	GATK (filtered)	1934	1441	74.51%
Bowtie2	isaac (raw)	2829	2146	75.86%
Bowtie2	Isaac (filtered)	1471	1041	70.77%
Bowtie2	Samtools (filtered)	2594	1898	73.17%
Bowtie2	Samtools (raw)	2565	1886	73.53%
BWA	GATK (raw)	1918	1435	74.82%
BWA	GATK (filtered)	1862	1415	75.99%
BWA	isaac (raw)	2680	2074	77.39%
BWA	Isaac (filtered)	1376	1013	73.62%
BWA	Samtools (filtered)	2753	2018	73.30%
BWA	Samtools (raw)	2657	1976	74.37%
isaac	GATK (raw)	1615	1295	80.19%
isaac	GATK (filtered)	1603	1288	80.35%

isaac	isaac (raw)	2223	1763	79.31%
isaac	Isaac (filtered)	1156	886	76.64%
isaac	Samtools (filtered)	2390	1822	76.23%
isaac	Samtools (raw)	2356	1808	76.74%

Deletion stats from all eighteen possible software combinations derived from the pairing of the three aligners with each of the three variant callers run both with and without filtering. Deletions were overlapped to GIAB deletions and the percent matching calculated.

S4 Table. Insertion call overlaps with GIAB.

Aligner	**Variant Caller**	**Total Insertion**	**Insertions Match GIAB**	**% Insertions Match GIAB**
Bowtie2	GATK (raw)	2114	1689	79.90%
Bowtie2	GATK (filtered)	2052	1666	81.19%
Bowtie2	isaac (raw)	3245	2647	81.57%
Bowtie2	Isaac (filtered)	1577	1218	77.24%
Bowtie2	Samtools (filtered)	3793	2842	74.93%
Bowtie2	Samtools (raw)	3678	2808	76.35%
BWA	GATK (raw)	1936	1601	82.70%
BWA	GATK (filtered)	1901	1591	83.69%
BWA	isaac (raw)	2985	2488	83.35%
BWA	Isaac (filtered)	1447	1174	81.13%
BWA	Samtools (filtered)	3553	2865	80.64%
BWA	Samtools (raw)	3451	2803	81.22%
isaac	GATK (raw)	1718	1460	84.98%
isaac	GATK (filtered)	1702	1448	85.08%
isaac	isaac (raw)	2542	2160	84.97%
isaac	Isaac (filtered)	1284	1082	84.27%
isaac	Samtools (filtered)	3826	2604	68.06%
isaac	Samtools (raw)	3796	2592	68.28%

Insertion stats from all eighteen possible software combinations derived from the pairing of the three aligners with each of the three variant callers run both with and without filtering. Insertions were overlapped to GIAB insertions and the percent matching calculated.

References:

1. Gafni E, Luquette LJ, Lancaster AK, Hawkins JB, Jung JY, Souilmi Y, et al. COSMOS: Python library for massively parallel workflows. Bioinformatics. 2014.
2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
4. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.
6. Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC bioinformatics. 2010;11 Suppl 12:S1.
7. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248-9.
9. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073-81.
10. Severin J, Beal K, Vilella AJ, Fitzgerald S, Schuster M, Gordon L, et al. eHive: an artificial intelligence workflow system for genomic analysis. BMC bioinformatics. 2010;11:240.
11. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nature genetics. 2009;41(2):149-55.
12. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology. 2010;11(8):R86.
13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308-11.
14. McCoy CO, Gallagher A, Hoffman NG, Matsen FA. Nestly--a framework for running software with nested parameter choices and aggregating results. Bioinformatics. 2013;29(3):387-8.
15. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9.
16. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29(16):2041-3.
17. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant callers for next-generation sequencing data: a comparison study. PloS one. 2013;8(9):e75619.

18. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome medicine. 2013;5(3):28.
19. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC bioinformatics. 2014;15:104.

2.2 Further discussion

In this two-part study described in this publication, I aimed to answer the following questions.

1) Does the choice of aligner, variant caller, and variant caller filtering significantly affect the quantity and quality of resultant variant calls?
2) In general, does software perform consistently for SNVs, insertions, and deletions?
3) Do we get higher quality variant calls by combining the output of multiple variant callers?
4) Can we demonstrate using a real patient data set, how software choices affect the detection of clinically important variants?
5) What is the ideal strategy for variant detection pipelines being used for clinical application?

This publication describes two separate analyses to answer the five main questions. First, using Illumina's validated platinum HapMap (74) genome NA12878 (http://www.illumina.com/platinumgenomes/) I generated variant calls for SNVs, deletions, and insertions resulting from pairing three aligners (bowtie2 (75), BWA (76), and isaac_aligner (77)) with each of three variant callers (GATK (24), isaac_variant_caller (77), and SAMtools (78)) run both with and without additional filtering steps. A merged list of variants was generated for each software tool assessed and compared to NA12878 high-confidence variant calls to obtain false positive and false negative rates and all variant lists overlapped and displayed using Venn diagrams. To determine whether combining the output of multiple tools generated higher quality variants, four subsets of the total variants were analysed independently: the union of all variant calls, the intersection of all variant calls (i.e. unanimous variant calls), variants detected by a single software pair, and variants detected by at least two software combinations but not unanimously. In the second part of the study, to demonstrate the impact software choices have on the detection of clinically important variants, a melanoma cell line was analysed using a similar methodology with ClinVar (4) 'melanoma risk-factor' variants identified and examined across all software combinations. In the concluding section, I propose

a cogent strategy for implementing a variant detection pipeline for clinical use, a strategy that focuses on minimizing false negative variants from the onset.

The first analysis replicates previous studies demonstrating the significant impact software selection has on overall variant quality (29, 79) and how this impact is inconsistent across different variant types (80). In this analysis, no single aligner / variant caller pair was found to outperform all other combinations across the three types of variants assessed, namely SNVs, deletions, and insertions. Thus, in order for a pipeline to utilise the most effective tool for detecting each specific variant type, it must minimally support multiple variant detection tools for different variant types. This analysis also confirms previous work (31) showing that combined variant calls from multiple tools yield higher quality resultant variant sets, for either specificity or sensitivity, depending on whether the intersection or union, of all variant calls is used respectively. Building on these findings, the second part of the study demonstrates using a real melanoma data set how two of the three total ClinVar annotated melanoma risk factor variants are not detected unanimously by all software assessed. Importantly, the two variants not detected unanimously are missed for entirely different reasons; a single aligner fails to detect one variant completely no matter what variant caller is used, while the other variant is filtered out as low quality by a single variant caller. While in this instance these two variants are particularly important, the work on the NA12878 reference set demonstrated how each tool detected high-confidence variants unique to that software, meaning no single tool is immune to false negative variants that are detectable with other tools. These results demonstrate the importance of initial software selection, optimized filtering strategies, and combined variant calls in minimizing false negative variants. Further, the increased output quality from such a strategy quickly justifies any increased computation costs given the importance of accurately detecting clinically actionable variants when incorporating variation information in the clinic. Collectively, these results indicate that any framework designed to support clinical usage of variation information needs to be capable of benchmarking and supporting a wide variety of software. Further, such a framework is ideally able to run multiple variant callers and combine their output, a feature missing in almost all modern pipeline frameworks.

The results of this and similar earlier analyses played a significant role in both the initial design and ongoing development of the in-house production genomics analysis framework I have developed and managed over the last four years. Internally, this framework works by bundling modular analytical processes as externally developed, compiled binary objects, with everything driven by a custom Perl module layer designed to wrap the individual tools. These custom modules manage the entire workflow in a highly automated fashion: automating steps such as data import, setup, execution, result parsing, logging, archiving, and even error handling. The system is extremely flexible with each step run in the workflow driven at a scripting level by a pre-defined list of system commands contained in an XML configuration file - a design that allows customisation, quick tool substitution, and straightforward change and extension of the workflow. Importantly, all workflows are entirely reproducible using a combination of XML configuration files, versioning of code base and annotations, and recording the version of all external binaries utilised. Designed to originally detect ENU SNVs in mouse exomes (69), flexible design has enabled numerous custom workflows to be incorporated including sequenced human pedigrees (33) and paired tumour-normal cancer samples (43). The total code-base currently consists of 30,000+ lines of predominantly Perl code and utilizes an ever-changing catalogue of open-source components combined with bespoke analysis tools – currently all linked via an underlying MySQL (http://www.mysql.com) tracking database designed to manage sample metadata, sample progress within the pipeline, and summary of results. Overall the flexibility, high level of automation, and reproducibility of results has allowed the framework to endure in a high-throughput production environment for five years. To date, over 3000 exomes and 1000 genomes have been analysed resulting in eight publications to date detailing discoveries made by this framework (see Chapter 8).

Chapter 3: Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees

Field, M. A., V. Cho, M. C. Cook, A. Enders, C. Vinuesa, B. Whittle, T. D. Andrews and C. C. Goodnow. "Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees." Bioinformatics. 2015; 15;31(14):2377-9

Bioinformatics, 2015, 1–3
doi: 10.1093/bioinformatics/btv135
Advance Access Publication Date: 8 March 2015
Applications Note

OXFORD

Genome analysis

Reducing the search space for causal genetic variants with VASP

Matthew A. Field[1,*], Vicky Cho[2], Matthew C. Cook[1,3], Anselm Enders[1,4], Carola G. Vinuesa[1], Belinda Whittle[2], T. Daniel Andrews[1,†] and Chris C. Goodnow[1,5,†]

[1]Department of Immunology, John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia, [2]Australian Phenomics Facility, Australian National University, Canberra, ACT 2601, Australia, [3]Department of Immunology, The Canberra Hospital, Canberra, ACT 2605, Australia, [4]Rammaciotti Immunisation Genomics Laboratory, John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia and [5]Immunogenomics Group, Immunology Research Program, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.
Associate Editor: Alfonso Valencia

Received on November 2, 2014; revised on February 10, 2015; accepted on March 2, 2015

Abstract

Motivation: Increasingly, cost-effective high-throughput DNA sequencing technologies are being utilized to sequence human pedigrees to elucidate the genetic cause of a wide variety of human diseases. While numerous tools exist for variant prioritization within a single genome, the ability to concurrently analyze variants within pedigrees remains a challenge, especially should there be no prior indication of the underlying genetic cause of the disease. Here, we present a tool, variant analysis of sequenced pedigrees (VASP), a flexible data integration environment capable of producing a summary of pedigree variation, providing relevant information such as compound heterozygosity, genome phasing and disease inheritance patterns. Designed to aggregate data across a sequenced pedigree, VASP allows both powerful filtering and custom prioritization of both single nucleotide variants (SNVs) and small indels. Hence, clinical and research users with prior knowledge of a disease are able to dramatically reduce the variant search space based on a wide variety of custom prioritization criteria.

Availability and implementation: Source code available for academic non-commercial research purposes at https://github.com/mattmattmattmatt/VASP.

Contact: matt.field@anu.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

While exome sequencing has been successfully utilized in the discovery of causal variation using small numbers of unrelated individuals (Ng *et al.*, 2010) it is becoming clear that this approach is insufficient to reliably identify the genetic causes in many cases (Yang *et al.*, 2013). Increasingly, genetic variation information from related individuals is being employed to reduce the search space for causal variants, by both the prioritization of variants common to affected individuals and the exclusion of benign variants shared between affected and unaffected individuals. The effective analysis of sequenced pedigrees requires new tools capable of combining variant specific and pedigree wide annotations with powerful filtering options to dramatically reduce the causal variation search space. Current tools focus on either progressively removing variants based

on criteria deemed unlikely to be causal (Li *et al.*, 2012) or by focusing on variants matching specific inheritance models [compound heterozygotes (Kamphans *et al.*, 2013); autosomal dominant (Koboldt *et al.*, 2014)]. Here, we present variant analysis of sequenced pedigrees (VASP), a tool that integrates and summarizes information across a sequenced pedigree without making any assumptions regarding disease inheritance further providing the full integrated pedigree variation information for subsequent prioritization. VASP is standalone software written in Perl derived from our original variant detection pipeline (Andrews *et al.*, 2012).

2 Methods

VASP integrates variant information from each pedigree member and aggregates this information on a per variant basis, describing, among other things, the likely inheritance pattern while simultaneously incorporating external annotation. Furthermore, VASP uses parental allele segregation patterns to determine phasing of variants and identifies genomic blocks common to affected pedigree individuals. VASP and other tools that aggregate information across a pedigree are critical for successful causal variant detection as they automate a complex task too labor intensive to perform manually.

2.1 Input

2.1.1 Input files

Two input files are required; a pedigree (PED) file representing pedigree structure and a variant call format (VCF) file containing variant information. The third required argument is either a variant effect predictor (VEP) annotation file (McLaren *et al.*, 2010) or the path to a local VEP installation when no annotation file is readily available. Optional input files include a text file with the path to individual binary sequence alignment/map (BAM) files used for determining zygosity and inheritance patterns. VASP is able to support any genome build providing all input files refer to the same reference genome.

2.1.2 Input filters

A key feature of VASP is flexible options with prioritization and ordering of variants based on stipulated criteria being specified at the outset. Filtering options include specific inheritance patterns, genomic region, gene name, population allele frequency, phasing information, number of affected/unaffected variant individuals and certain polyphen2 and sorting intolerant from tolerant (SIFT) categories. Combining parameters results in a dramatic reduction in the length of candidate variant lists—and the remaining variants will better correspond to hypotheses about the genetic basis of a particular disease.

2.1.3 Input variant set

Variant callers typically employ a quality score as a cutoff, above which lie a set of presumed high-quality variant calls. In reality, variants above this cutoff often include false positives, whereas variants below this cutoff include true positives (Weisenfeld *et al.*, 2014). To minimize and potentially avoid this problem, the union of all variants called within the pedigree is used as the initial input variant set. This approach allows variant calls at a particular genomic position to be reconciled across the pedigree, potentially correcting calls lying near the cutoff for a single individual.

2.2 External variant annotation

Individual variants are annotated with VEP data including population frequency information from dbSNP (Sherry *et al.*, 2001) and the 1000 genomes project (Genomes Project *et al.*, 2010) as well as SIFT (Kumar *et al.*, 2009) and PolyPhen2 (Adzhubei *et al.*, 2010) scores for estimating the functional effect of missense mutations.

2.3 Pedigree-wide annotation

2.3.1 Genome phasing and de novo mutations

Whenever possible the parental allele inherited by each child is determined, and this information is clustered into genomic blocks of presumed shared inheritance to obtain genome-phasing information. Variants demonstrating segregation differences between affected and unaffected children are further prioritized. In addition, any variant exhibiting Mendelian inconsistencies is annotated, with special attention paid to putative *de novo* mutations (i.e. non-mutant, unaffected parents and a heterozygous offspring).

2.3.2 Disease inheritance patterns and compound heterozygosity

For each variant in the pedigree the inheritance is determined and annotated accordingly. The zygosity of particular variants is preferentially determined from raw sequence data obtained from BAM files using SAMtools (Li et al., 2009) but failing this the required genotype field (GT) and optional allele depth (AD) tags from the VCF file are utilized. Compound heterozygote genes are also annotated, defined as genes containing at least one heterozygous SNV or indel inherited from each parent with unaffected and affected siblings not sharing identical heterozygous variants. These variants must further be heterozygous in all affected individuals and not be homozygous in any unaffected individuals. These compound heterozygous genes are further prioritized in cases where each parent contributes rare or novel alleles.

2.3.3 Gene variability statistics

For each gene three measures of variability are reported; total number of variants, total number of unique variant coordinates and percentage of total transcript bases found variant. Increased gene variability may be relevant to particular diseases but also may be indicative of read alignment issues (often due the presence of gene duplicates) or may indicate the gene is functionally redundant and thus not functionally constrained.

2.4 VASP output and ordering

VASP reports contain all variants detected in at least one pedigree member and categorises variants as either novel, rare (0–2% population frequency), no frequency (known variant but no frequency data available) or common (>2% frequency). For each variant VASP reports both pedigree-wide information (such as inheritance pattern or phasing data) as well as variant-specific information (such as population frequency or polyphen score). By default VASP reports are sorted progressively on four measures: variant category (novel, rare, no frequency and common), the number of variant affected samples (in descending order), the number of unaffected variant samples and lastly the variant population frequency.

3 Results

VASP makes no assumptions regarding the underlying disease transmission mechanism, an apparent strength when compared with similar software (Supplementary Table S1). Instead, VASP provides powerful filters with the aim of allowing researchers to harness their additional knowledge of the disease to generate reduced variant lists

suitable for manual interrogation. One current limitation of VASP is that it can only be run on the command line.

Five pedigrees (Supplementary Table S2) were analyzed to calculate variant segregation statistics with pedigree G1 (Supplementary Figure S1) variant lists (Supplementary Table S3) taken forward to illustrate the effect of various filtering strategies (Supplementary Table S4). To date VASP has been used to analyze 45 pedigrees and found strong candidate causal variants in 15 of these (33.3%). These 15 pedigrees exhibit a wide array of disease transmission mechanisms including autosomal dominant and recessive inheritance, *de novo* mutations, compound heterozygosity and more complex multi-gene cases. This variety in transmission mechanisms within this relatively small group sharing similar diseases illustrates the importance of flexible pedigree analysis software.

We present VASP, a flexible tool for identifying putative causal variants from pedigree sequence data. Through aggregation of data for genetic variants across pedigree members, VASP allows powerful, custom variation prioritization, taking advantage of external datasets and prior knowledge of disease incidence and inheritance patterns. With this tool users have the opportunity to greatly narrow the number of candidate causal variants, using custom criteria, to a size suitable for manual interrogation.

Acknowledgement

We thank the National Computational Infrastructure (Australia) for access to significant computation resources and technical expertise.

Funding

This work was supported by National Health and Medical Research Council Australia Fellowship 585490, National Institutes of Health [grant number U19 AI100627] and Bioplatforms Australia.

Conflict of Interest: none declared.

References

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Andrews,T.D. *et al.* (2012) Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. *Open Biol.*, **2**, 120061.

Genomes Project,C. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Kamphans,T. *et al.* (2013) Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. *PloS One*, **8**, e70151.

Koboldt,D.C. *et al.* (2014) Exome-based mapping and variant prioritization for inherited Mendelian disorders. *Am. J. Human Genet.*, **94**, 373–384.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,M.X. *et al.* (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.*, **40**, e53.

McLaren,W. *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.

Ng,S.B. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Weisenfeld,N.I. *et al.* (2014) Comprehensive variation discovery in single human genomes. *Nat. Genet.*, **46**, 1350–1355.

Yang,Y. *et al.* (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, **369**, 1502–1511.

Supplemental Material

Table 1. Pedigree Analysis Software Comparison

Software	CLI/GUI*	Filter Output	Inher-itance	Com-het	Phasing/ Linkage	Expression Data
Gemini	CLI	Y	Y	Y	Y	N
KGGSeq	CLI	Y	Y	Y	Y	N
Mendel Scan	CLI	Y	Y	N	Y	Y
plink	CLI	Y	Y	N	Y	N
pVAAST	CLI	Y	Y	N	N	N
VariantStudio/ TruSight One	GUI	Y	Y	N	Y	N
VAR-MD	CLI	N	Y	Y	Y	N
VASP	CLI	Y	Y	Y	Y	N

*Command line interface (CLI) or graphical user interface (GUI)

Table 2. Sequenced Pedigree Composition

Family ID	# Affected (Parents)	#Unaffected (Parents)	Total Samples
A8	4 (2)	0 (0)	4
A10	3 (1)	1 (0)	4
G1	2 (0)	3 (2)	5
R2	2 (1)	2 (1)	2
S1	2 (0)	3 (2)	5

Variant segregation statistics

On average, a union of 34,049 variants were in identified in each affected pedigree (32,761 snvs and 1288 small indels) overlapping either a coding base or a splice site (defined as within 10bp of a coding base) with reference to the canonical transcript from ENSEMBL v73.

An average of 32,500 variants (95.5%) were consistent with Mendelian inheritance, with an additional 768 variants on average (2.2%) being of ambiguous provenance due to insufficient read coverage in at least one pedigree member. The remaining 798 variants (2.3%) on average displayed some form of Mendelian inconsistency in at least one parent-child pair of which an average of 159 variants (0.5%) were classified as *de novo.*

Disease inheritance was calculated in these five pedigrees. An average of 31,769 variants (93.3%) did not match any known disease inheritance pattern with the remaining 2280 (6.7%) variants classified as either autosomal dominant, autosomal recessive, X-linked dominant, or X-linked recessive. Of these 2280 variants an average of 991 were autosomal dominant (43.5%), 1254 were autosomal recessive (55%), 13 were X-linked dominant (0.6%), with 22 X-linked recessive (0.9%) variants on average.

Compound heterozygosity classification yielded an average of 987 variants (2.8%) distributed across 170 genes or 1.8% of all total variant genes.

Effects of Filtering Strategies

Pedigree G1 was further analysed to illustrate the effect of applying some of the different filters available in VASP on the raw variant lists. Pedigree G1 contains two unaffected parents, one proband, one affected brother, and one unaffected brother. For each filter tested the following fields are reported; the number of variants passing each filter, the percentage of variants removed by the filter, and the number of distinct genes overlapping the variants that pass each filter.

Table 3. Pedigree G1 Variants

Family Member	**Total SNVs**	**Total Indels**	**Total Variants**
Mother	22660	646	23306
Father	21935	679	22614
Proband	22477	661	23138
Brother (affected)	22132	640	22772
Brother (unaffected)	23932	663	24595

In total pedigree G1 contains 32418 distinct variants distributed across 11577 genes.

Figure 1: Pedigree G1

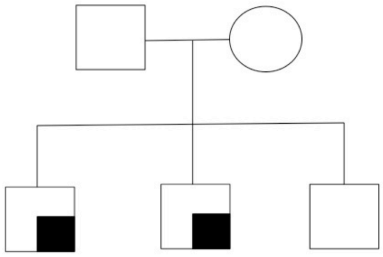


Table 4. Filtering Effect on Raw Variant

Filter	Number Variants	Per cent Filtered Out	Total Genes
De novo	60	99.81%	55
Autosomal dominant	4	99.99%	4
Autosomal recessive	145	99.55%	101
Compound heterozygous	469	98.55%	115
Mother allele to affected+	643	98.02%	411
Father allele to affected+	1145	96.47%	810
Parental allele blocks*	337	98.96%	156
Damaging polyphen	1268	^90.54%	1104
Deleterious SIFT	2472	^81.57%	1948
Polyphen AND SIFT	918	^93.15%	830
Nonsense Mutation	101	99.69%	96

+Heterozygous parent exclusively passes variant allele to affected children and reference allele to unaffected children

*Blocks of 3 or more variants matching the above definition

^Percentages based on 13412 missense mutations

3.2 Further discussion

This publication describes a new variant data prioritisation tool, Variant Analysis of Sequenced Pedigrees (VASP), which is software written to detect causal variants within sequenced families or pedigrees. With an increasing number of projects sequencing families in order to detect causal variants, no simple automated mechanism existed to examine variants in the context of the entire pedigree, meaning researchers were resorting to manually manipulating individual sample variant lists to obtain this information for variants of interest. While such an approach is appropriate for a select number of variants, it proved unsuitable for the examination of the huge number of variants contained across the entire pedigree. In addition to the difficulty of automating this approach, another shortcoming resulted from the fact that pedigree members were analysed in isolation, meaning no pedigree wide information (e.g. disease inheritance, genome phasing, and compound heterozygosity) was available despite the fact that calculating such information is highly suitable to automation. Further, while pedigree wide annotation is informative for specific variants of interest, it also represents a powerful mechanism for prioritising variants, particularly when researchers have existing knowledge of the disease transmission mechanism and large lists of variants unsuitable for manual interrogation.

Prior to the implementation of VASP an exhaustive search of existing tools was performed to determine whether suitable tools existed for routinely identifying causal variants in pedigrees. While a handful of tools exist, none proved flexible enough for our needs particularly with regard to disease transmission mechanism, pedigree structure, and filtering options. Existing tools tended to either focus on a single disease transmission mechanism (e.g. compound heterozygosity (42) or autosomal dominant (41)) or else employ a winnowing strategy by successively removing variant deemed unlikely to be causal (40). While effective in many instances, such strategies inevitably make certain assumptions about the disease that will not hold in many cases. For example, tools that focus on a single disease transmission mechanism will not be effective when the mechanism differs from the expected model while tools that progressively filter out variants often encounter issues due to incomplete penetrance, pedigree sample mix-ups, and the addition of rare and disease-

causing variants to dbSNP (2). Given these known limitations, VASP was created to be free from assumptions regarding the underlying disease mechanism and to feature a huge number of available filtering options.

VASP has a number of unique features compared to existing software. First, as mentioned, it makes no assumptions regarding the transmission of the disease making it suitable for the examination of either monogenic or polygenic disease, an important distinction from other existing tools. Another important feature is that VASP takes the union of all variants called across the entire pedigree and obtains raw sequence information from BAM files for every member of the pedigree regardless of whether a variant was detected in that particular pedigree member. This is an important feature as variant callers typically employ a quality score as a cut-off, above which lie a set of presumed high-quality variant calls. In reality identifying a perfect cut-off value is impossible however, as there are inevitably false positive variants above the cut-off and true positive variants below the cut-off (81). By taking the union of all pedigree variants as input, calls at a particular genomic position can be reconciled across the pedigree, potentially correcting calls lying near the cut-off. Importantly, raw sequence data is used to determine zygosity for each pedigree member with this information used for calculating disease inheritance, thus giving the most accurate representation of the inheritance based on the actual sequence data, not whether a variant fell above or below an imperfect cut-off value. Another novel feature of VASP is the use of genome phasing information to identify genomic regions consistent with certain disease inheritance patterns. To do this, VASP obtains genome phasing information by using parental sequence information to determine the allele inherited by each child, with the phasing information clustered into haplotype blocks. Variants demonstrating segregation differences between affected and unaffected children are further prioritized and identified as candidate autosomal dominant variant hotspots. To ensure flexibility, users are able to precisely define the nature of a hotspot with regard to the number of consecutive variants segregating in this manner and the minimum genomic length of these blocks.

One of VASP's greatest strengths is the variety of custom filters available, filters allowing researchers to dramatically reduce the search space for causal variants based on their inherent knowledge of the disease. Filters are

broadly divided into two categories, pedigree wide annotation filters and variant-specific annotation filters. Pedigree wide filters include disease inheritance mechanism (autosomal or X-linked dominant / recessive, compound heterozygosity, and *de novo* mutations as well as genomic block hotspots as described above. Variant specific filters include variant score cut-off, SIFT and polyphen function inference predictions, population allele frequency, specific gene(s), or genomic regions. Collectively, these filters are run either individually or in combination resulting in a dramatic reduction in the length of candidate variant lists with the remaining variants better corresponding to the hypotheses about the genetic basis of the particular disease.

To date VASP has been used to analyse 45 pedigrees and identified strong candidate causal variants in 15 cases for a success rate of 33.3%, similar to the 25% success rate reported by existing tools (36). The program has been used to study pedigrees from a wide variety of projects with resultant publications coming from studies of lupus (16), Wiedermann-Steiner syndrome (82), and serrated polyposis (38) to date. Almost 75% of all pedigrees studied to date are part of a wide scale study to identify the genetic cause of complex auto-immune diseases and interestingly, candidate causal variants exhibit numerous disease transmission mechanisms such as autosomal dominant, autosomal recessive, *de novo* mutations, compound heterozygosity and even complex multi-gene interactions. This variation in transmission mechanisms within this relatively small group sharing similar diseases illustrates the importance of flexible pedigree analysis software.

Chapter 4: Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-calling algorithms for human melanoma genomes

Wilmott, J. S., **M. A. Field**, P. A. Johansson, H. Kakavand, P. Shang, R. De Paoli-Iseppi, R. E. Vilain, G. M. Pupo, V. Tembe, V. Jakrot, C. A. Shang, J. Cebon, M. Shackleton, A. Fitzgerald, J. F. Thompson, N. K. Hayward, G. J. Mann and R. A. Scolyer. "Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes." Pathology. 2015; 47(7), pp. 683–693

Pathology (Month 2014) **00(xx)**, pp. 1–11

MOLECULAR PATHOLOGY

Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes

James S. Wilmott[1,2], Matthew A. Field[3], Peter A. Johansson[4], Hojabr Kakavand[1,2], Ping Shang[1], Ricardo De Paoli-Iseppi[1], Ricardo E. Vilain[1,2], Gulietta M. Pupo[1,5], Varsha Tembe[1,5], Valerie Jakrot[1], Catherine A. Shang[6], Jonathan Cebon[7], Mark Shackleton[8], Anna Fitzgerald[6], John F. Thompson[1,2,9], Nicholas K. Hayward[4], Graham J. Mann[1,2,5] and Richard A. Scolyer[1,2,10]

[1]Melanoma Institute Australia, North Sydney, NSW, [2]Sydney Medical School, The University of Sydney, Camperdown, NSW, [3]Immunogenomics Laboratory, Australian National University, Canberra, ACT, [4]Oncogenomics Laboratory, QIMR Berghofer Medical Research Institute, Herston, Brisbane, Qld, [5]Centre for Cancer Research, The University of Sydney at Westmead Millennium Institute, Westmead, NSW, [6]Bioplatforms Australia, Macquarie University, North Ryde, NSW, [7]Ludwig Institute for Cancer Research, Olivia Newton-John Cancer and Wellness Centre, Austin Health, Heidelberg, Vic, [8]The Cancer Development and Treatment Laboratory, Peter MacCallum Cancer Centre and Sir Peter MacCallum Department of Oncology, The University of Melbourne, Vic, [9]Departments of Melanoma and Surgical Oncology, and [10]Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Camperdown, NSW, Australia; these authors contributed equally

Summary

Whole genome sequencing (WGS) of cancer patients' tumours offers the most comprehensive method of identifying both novel and known clinically-actionable genomic targets. However, the practicalities of performing WGS on clinical samples are poorly defined. This study was designed to test sample preparation, sequencing specifications and bioinformatic algorithms for their effect on accuracy and cost-efficiency in a large WGS analysis of human melanoma samples. WGS was performed on melanoma cell lines ($n=15$) and melanoma fresh frozen tumours ($n=222$). The appropriate level of coverage and the optimal mutation detection algorithm for the project pipeline were determined. An incremental increase in sequencing coverage from 36X to 132X in melanoma tissue samples and 30X to 103X for cell lines only resulted in a small increase (1–2%) in the number of mutations detected, and the quality scores of the additional mutations indicated a low probability that the mutations were real. The results suggest that 60X coverage for melanoma tissue and 40X for melanoma cell lines empower the detection of 98–99% of informative single nucleotide variants (SNVs), a sensitivity level at which clinical decision making or landscape research projects can be carried out with a high degree of confidence in the results. Likewise the bioinformatic mutation analysis methodology strongly influenced the number and quality of SNVs detected. Detecting mutations in the blood genomes separate to the tumour genomes generated 41% more SNVs than if the blood and melanoma tissue genomes were analysed simultaneously. Therefore, simultaneous analysis should be employed on matched melanoma tissue and blood genomes to reduce errors in mutation detection. This study provided valuable insights into the accuracy of SNV with WGS at various coverage levels in human clinical cancer specimens. Additionally, we investigated the accuracy of the publicly available mutation detection algorithms to detect cancer specific SNVs which will aid researchers and clinicians in study design and implementation of WGS for the identification of somatic mutations in other cancers.

Key words: Cancer, coverage, genomics, melanoma, methods, mutation, pathology, sequencing, treatment, whole genome.

Received 18 March, revised 22 June, accepted 28 June 2015

INTRODUCTION

Whole-genome sequencing (WGS) of fresh-frozen tumours has enabled the characterisation of the entire genomic profile of a patient's cancer, theoretically allowing the identification of virtually all possible genomic drivers, disease modifiers and risk factors, thereby facilitating the selection of the most appropriate treatment options for an individual patient's disease.[1,2] Unlike whole exome sequencing (WES), which only analyses the portion of DNA that is transcribed into proteins (~1.2% of the genome), WGS covers the entire genome. Until very recently, the use of WES or WGS in the clinical setting has been restricted by the prohibitive costs of sequencing and the additional necessary computing resources needed to perform downstream data analysis. However, with the advent of new technologies, the cost of WGS of a patient's genome has recently dropped below US$1000.[3] Therefore, we can expect its clinical use to rapidly expand due to greater accessibility to clinicians and affordability for patients. In fact many small scale trials have already implemented WGS in the clinic, resulting in altered clinical care based on the increase in knowledge of the cancer genome and the development of novel therapies that target the protein products of specifically mutated genes.[4–6]

The successful implementation of the technology in the clinic or in a research setting relies upon sound methodological

Print ISSN 0031-3025/Online ISSN 1465-3931
DOI: 10.1097/PAT.0000000000000324

approaches and despite technological advances, some of the practicalities of performing WGS are not well outlined for clinical samples.[7] Methodological factors that need to be considered in performing WGS on clinical samples include: appropriate sample selection, storage, preservation and macrodissection for tumour enrichment, DNA preparation, sequencing specifications and accuracy of the variant detection algorithms. The importance of sample selection and DNA extraction can be overlooked, leading to unusable or inaccurate data.[8] Patient biopsies often contain an assortment of tumour and non-tumour cells such as immune, stromal and endothelial cells. The presence of the latter in substantial numbers reduces the ability to detect somatic mutations within tumours by genome sequencing and can lead to false negative mutation calls.[9–11] This can have serious consequences for patient care, as patients with false negative results could be incorrectly designated as ineligible for a personalised treatment that could have been effective.

Another major factor to consider when performing WGS is the number of replicate sequences (or reads) of the genome that is required to accurately identify all mutations present. Sequencing coverage (X) expresses the average number of times an average nucleotide base will have been read in a given sample.[12] However, coverage varies across the genome due to variability in sampling and ease of sequencing; it is desirable to maximise coverage to obtain data on a greater proportion of the genome. Selecting the appropriate coverage for a project is often a balancing act between sensitivity and costs, with increased levels of coverage producing more reliable and sensitive variant detection capabilities but at an increased cost.[7] For sequencing projects, this results in a choice between increased coverage per sample but lower overall sample numbers or higher sample numbers and reduced sensitivity to detect rare or low frequency events. Deciding upon the appropriate coverage levels for a WGS assay is a vital aspect of clinical and research WGS. Coverage levels of 10–30X to detect putative germline mutations (usually derived from blood DNA) have been suggested, while higher coverage levels (>40X) are considered necessary for mutation detection with tumour samples due to contamination by normal cells, tumour heterogeneity and amplification bias.[7,13]

The algorithms and processing tools that are used to convert raw sequencing data into annotated lists of mutations can often be the most expensive and infrastructure-intensive aspect of the WGS process. The sequencing platforms produce large text files of nucleotide sequences that need to be aligned to the reference genome using algorithms such as the Burrows–Wheeler aligner (BWA).[14] Variant detection algorithms such as SAMtools can then detect alterations between the genomes based on the probability of a variant occurring in that genomic region,[15] whilst taking into account sequencing error and predicted polymorphism rates. There are over 205 tools for WGS data analysis that differ in their statistical approach, number of variants identified and type of mutations detected by each algorithm.[16,17] However, for matched blood and melanoma tissue samples, these algorithms are founded on two basic principles of detecting somatic mutations in cancer: (1) subtraction method, whereby the mutations are compared to a synthetic reference genome separately for the matched blood and the tumour samples, then mutations present in the blood are removed from the tumour calls by subtraction or filtering; (2) simultaneous detection of the matched tumour and blood samples uses probability based statistics to filter mutations that are unique to the tumour sample.[18] The choice of the bioinformatics variant detection algorithm and methodology has an effect on the data produced from the raw WGS.[19]

In the present study, part of the Australian Melanoma Genome Project (AMGP) which will complete WGS of 400 human melanomas by the end of 2015, we performed a rigorous optimisation process for sample preparation, quality control, sequencing coverage level analysis and the mutation detection methodology via the WGS of 222 patients' matched melanoma and normal white blood cell genomes. The results of this study provided some valuable insights into the accuracy to detect single nucleotide variants (SNVs) at various coverage levels in different sample types. Additionally, we investigated the accuracy and specificity of the simultaneous versus the subtraction methodology using publicly available mutation detection algorithms.

METHODS AND MATERIALS

Study overview

WGS was performed on genomic DNA extracted from matching blood and melanoma tissue samples on the Illumina HiSeq2000 platform (Illumina, USA). The study was designed to optimise the sample preparation, sequencing specifications and bioinformatic algorithms to decide upon the most accurate and cost-efficient methodology to be used in the AMGP. We began with a pilot study that performed WGS on a cell line and melanoma tissue samples to determine the appropriate level of coverage and optimise the mutation detection algorithms. Samples were sequenced in individual flow cell lanes and the data later combined to simulate incremental levels of coverage, 20–60X for blood genomes and 40–100X for tumour genomes. Variant detection was then performed at increasing coverage levels to determine the appropriate coverage for each DNA source. We then optimised the mutation detection algorithms to determine the methodology that produced the most reliable and accurate variant detection.

Specimen collection

The tissue and blood samples analysed in the current study were obtained from Australian melanoma biospecimen banks, which include the Melanoma Institute Australia (MIA) ($n = 198$), Queensland Institute of Medical Research (QIMR) ($n = 15$), Ludwig Institute for Cancer Research ($n = 5$), Peter MacCallum Cancer Centre/Victorian ($n = 4$) biospecimen banks. All tissues and bloods form part of a prospective collection of fresh-frozen samples accrued with written informed patient consent and institutional review board approval [MIA is covered by the Sydney South West Area Health Service institutional ethics review committee (Royal Prince Alfred Hospital zone), the Ludwig Institute for Cancer Research is covered by the Austin Hospital committee and the Peter MacCallum Cancer Center]. Clinical and follow-up details were collected on all patients as approved by the aforementioned research ethics committees.

Study population

Patients were selected for WGS based on the availability of fresh frozen melanoma tissue and blood that fulfilled the following clinical criteria:

1. Primary melanoma with a patient matched melanoma metastasis (any metastatic site).
2. Primary melanoma with greater than 3 years clinical follow-up and prior mRNA array data or known sentinel lymph node biopsy (SLNB) status.
3. Regional lymph node metastasis with greater than 3 years of clinical follow-up data and prior mRNA array data.[20]
4. Distant metastasis to the small intestine (limited to 8 samples) or brain (unlimited number of samples).
5. Melanomas of any stage of disease that arose from mucosal epithelium. Any sample with a primary melanoma that bordered cutaneous epithelium was excluded.
6. Melanomas of any stage of disease that arose from acral skin. Any sample with a primary melanoma that bordered non-acral skin was excluded.
7. Human melanoma cell lines with prior drug screen data were also included (limited to 15 samples).

Patients were excluded if they had received prior chemotherapy or radiotherapy to the biopsy site or if they had existing exome or WGS data available.

Tissue DNA extraction and quality assessment

Fresh surgical specimens were macro-dissected and tumour tissue was procured (with as little contaminating normal tissue as possible) and snap frozen in liquid nitrogen within 1 hour of surgery. For primary tumours, only macroscopically visible tumour was banked. The fresh-frozen tumour samples were sectioned on a cryostat (CM1520; Leica, Germany) and the slides were stained with haemotoxylin and eosin (H&E). Areas with high tumour content (>80% if possible) were marked and reviewed by pathologists (RV and RAS). The following features were evaluated for the selected area: percentage of tumour nuclei, percentage of non-tumour nuclei, percentage area displaying necrosis, degree of pigmentation (Fig. 1A–D; absent, mild, moderate, severe), predominant cell shape (Fig. 1E–G; epithelioid, spindle and mixed epithelioid and spindle), tumour cell size of the most cellular portion of the tumour, and the density of tumour-infiltrating lymphocytes (TILs) as previously described and depicted in Fig. 1H–K.[20–23] Cell size was measured as the longest dimension of the nuclei in ten representative cells using the measure tool in Leica's LAS software on photomicrographs taken using a 20× objective (Fig. 1L; DM2000; Leica).

The minimum tissue criterion needed for inclusion in the study was a macro or microdissectible tumour area containing greater than 80% tumour content and less than 30% necrosis as marked. Samples that needed tumour enrichment underwent macrodissection using a marked H&E slide as a reference to remove non-tumour or necrotic tissue under sterile conditions (Fig. 1M,N). Tumour DNA was then extracted using DNeasy Blood and Tissue Kits (69506; Qiagen, Germany), according to the manufacturer's instructions. Blood DNA was extracted from whole blood using the Flexigene DNA Kit (51206; Qiagen). All individual samples were quantified using the NanoDrop (ND1000; Thermo-Scientific, USA) and Qubit dsDNA HS Assay (Q32851; LifeTechnologies, USA) and DNA size and quality were tested using electrophoresis gels. Samples with a concentration of less than 50 ng/μL or absence of a high molecular weight band in electrophoresis gels were excluded from further analyses.

Cetyltrimethyl ammonium bromide (CTAB) DNA clean-up

Excessive melanin pigment was removed from tissue-derived DNA using a modified CTAB clean-up process.[24] Briefly, fresh CTAB was made with 50 mM Tris-HCL, 1% CTAB (#52365; Sigma, USA), 4 M urea (#U5378; Sigma), 1 mM EDTA and RNAse-free water. Per 100 μL of DNA elution buffer, 39 μL of 5 M NaCl was added, then 500 μL of CTAB-Urea solution added, mixed and left overnight at 4°C on a rotator. Samples were then spun at 15,000 g for 15 min at 4°C. Supernatant was discarded and the DNA pellet resuspended in 200 μL of buffer ATL, to which 200 μL of buffer AL and 200 μL of absolute ethanol were added and the subsequent solution pipetted onto a DNA extraction column. Columns were spun and washed in buffer AW1, AW2 and the DNA was eluted in 100 μL of buffer AE.

Whole genome sequencing

WGS was undertaken at three Australian sequencing facilities (Australian Genomic Research Facility, Ramaciotti Centre for Genomics, John Curtin School of Medical Research) managed by the Australian government infrastructure enabling body, Bioplatforms Australia, and also at Macrogen (South Korea). All facilities carried out the following protocols. Library construction was performed using TruSeq DNA Sample Preparation kits as per Illumina instructions. Sample DNA (1 μg) was fragmented into 300–400 base pair (bp) average insert size with 3′ or 5′ overhangs. End repair mix was then used to convert the fragmented DNA into blunt ends by removing the 3′ overhangs and the polymerase activity fills the 5′ overhang. The 3′ ends were then acetylated to add a single 'A' nucleotide to the 3′ to reduce chimera formation. Ligate adapters were then used to attach adapters to the DNA fragments so they could be loaded into a flow cell and purified to remove unligated adapters to generate a final product with an insert size of 300–400 bp. PCR was then used to selectively enrich DNA fragments with adapter molecules at both ends for sequencing. Post-amplification quality controls were performed using DNA High Sensitivity Labchips (Agilent Bioanalyzer; Agilent, USA). The libraries were then pooled and clustered using the iBOT and ready for sequencing. The 100 bp pair-end library was sequenced on a Hiseq2000 using a Truseq SBS V3-HS kit (Illumina). Each sample was analysed in the minimum number of lanes to generate the total data output (i.e., full lanes dedicated to one sample rather than multiplexing unless a half lane is needed to achieve the final desired output). FASTQ files were then sent to the bioinformatics pipeline for analysis.

Detection of somatic nucleotide variation

To detect somatic variants in each sample, a variant detection pipeline was implemented that was derived from an existing pipeline originally designed for mouse exome analysis.[25] The new pipeline consists of a MySQL tracking database Supplementary Fig. 1, http://links.lww.com/PAT/A38 in addition to a versioned code base containing custom Perl code, external annotation files, and external binaries. For each new patient, an automated script enters sample and sequencing metadata into the tracking database, information such as tumour and tissue type as well as the number of lanes for the each tumour and control sample. The script further creates all the necessary files for submission of jobs onto the 57,000-processor compute cluster raijin (http://nci.org.au/nci-systems/national-facility/peak-system/raijin/) hosted at the National Computational Infrastructure (NCI) at the Australian National University. The utilisation of the NCI was required as analysis of each pair of patient samples requires 1000 CPU hours and 1 terabyte of storage. Once entered into the system, each sample followed a pre-defined workflow (Fig. 2) ensuring reproducible results and further allowing users to make consistent comparisons across samples analysed at different stages of the project. This workflow consisted of three main phases: alignment, variant detection and annotation.

i) Alignment

Each individual paired lane was aligned to the human reference genome GRCh37 using the short read aligner BWA[14] with default parameters unless a non-standard base quality encoding was utilised. Repetitively aligned reads were filtered out from the SAM file and a sorted BAM file generated. Alignment statistics were generated from the BAM file using the SAMtools[15] flagstat command and only lanes with 90% or more of the reads aligning to the reference genome were further processed. When all the sequencing lanes for a patient had been aligned and had passed quality control steps, single merged BAM files were created for the control sample and each tumour sample using SAMtools merge. Duplicate reads were subsequently removed from each merged BAM file using the SAMtools rmdup command, a step that aims to avoid the detection of false positive variants arising due to PCR duplication. Coverage levels for each sample were computed and checked against minimum requirements of the project (40X for control / cell line tumour samples and 60X for tissue derived tumours).

ii) Variant detection

Variant detection utilised SAMtools and BCFtools to generate a VCF file for each chromosome. Variants were filtered first on their variant quality score (>40, higher score indicating reduced risk of false base call) and then their CLR score (>60), a score that measures the probability that the tumour and control had different genotypes. Variants from each chromosome were next merged and classified according to variant type (SNVs or indels) and further segregated into germline, LOH, or somatic categories with only somatic variants processed further.

iii) Annotation

All variants were overlapped with ENSEMBL's canonical transcripts and run through the variant effect predictor (VEP).[26] to determine whether the SNVs represented non-synonymous changes and to obtain additional annotations. Polyphen and SIFT scores were extracted in order to prioritise variants by predicting the functional impact of the missense mutations on the protein. Variants that overlapped with exon splice sites (defined as 10 bases either side of exon) were considered the most likely to be pathogenic. Finally, variants were overlapped with dbSNP[27] to determine whether the variant had been previously reported in the general population and to record allele frequencies of variants matching dbSNP.

To compare the somatic variants with previous cancer studies, we annotated genes in Vogelstein's influential paper[28] and provided a link to the COSMIC repository.[29] Variants found to have an exact coordinate match to existing COSMIC entries were further annotated with special attention paid to variants previously detected in melanoma studies.

The final step was the generation of separate variant reports for SNVs and small indels that integrated all the information calculated during analysis which was accessed from the tracking database. The reports contained sequence information for the variants in the tumour and control samples as well as extensive gene annotations and links to external resources.

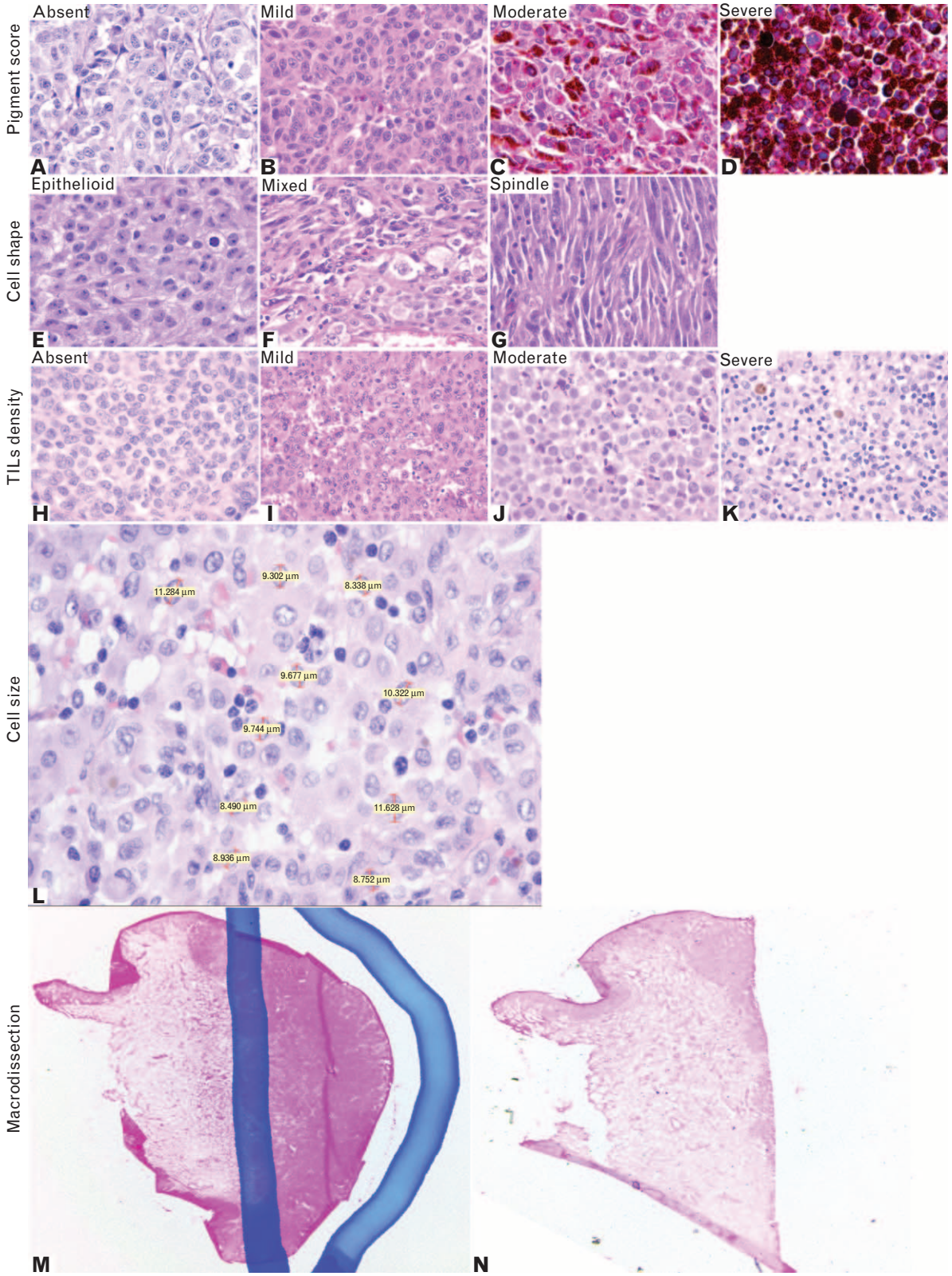


Fig. 1 Guide for pathological assessment of frozen tissue sections for WGS (H&E). (A–D) Increasing degrees of melanin pigmentation; (E–G) epithelioid, spindle, and mixed cell shape; (H–K) increasing density of tumour infiltrating lymphocytes; (L) an example of the measurements of cell size; (M) pre-macrodissection marked slide; and (N) post-macrodissection quality control slide.

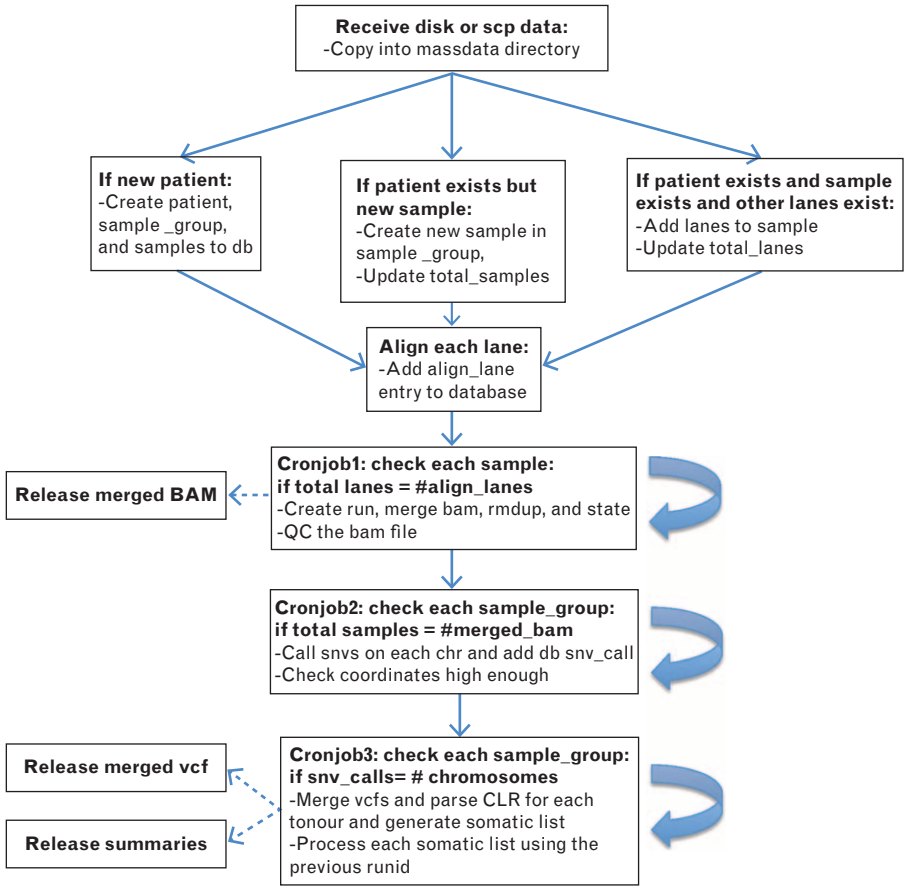


Fig. 2 Workflow of patient data through the bioinformatics process.

RESULTS

Patients and melanoma samples

At the time of writing this report, 222 matched bloods and tissue samples from 204 patients had been contributed from the biospecimen banks of MIA ($n = 198$), QIMR ($n = 15$), Peter MacCallum Cancer Centre ($n = 5$) and the Ludwig Institute for Cancer Research ($n = 4$), all of which had undergone WGS. The melanoma samples comprised 18 primary melanomas with patient matched metastasis, 30 primary melanomas with long-term (>3 years) clinical follow-up data, 18 primary melanomas with known SLNB status, 62 regional lymph node metastases with long-term (>3 years) clinical follow-up data, 19 cerebral metastases, eight small intestine metastases, 15 melanoma cell lines with prior drug screen data, 25 (8 primary and 17 metastatic) acral melanomas and nine (6 primary and 3 metastatic) mucosal melanomas. Figure 3 depicts the selection and exclusion process of the samples that underwent WGS. In total, over 445 patients were identified as having fresh frozen melanoma tissue in the MIA biospecimen bank that fitted the criteria for the study. The major causes of exclusion from the study were lack of a blood or other germline sample ($n = 93$) and prior radiotherapy or chemotherapy ($n = 17$). Tissue and DNA quality controls excluded 83 samples due to low tumour content (<80%), high necrosis (>30%) and consequent poor quality DNA. A proportion of patients had undergone prior genome sequencing in other studies and therefore were excluded from this analysis ($n = 30$).

Seventy-nine primary melanomas were suitable for WGS, the subtypes of which included 26 nodular (33%), 14 superficial spreading (SS) (18%), 11 acral lentiginous (14%), nine desmoplastic (11%), five SS with a dominant dermal nodule (6%) and one lentigo maligna melanoma (1%). The subtype of 13 cases (17%) was unknown at time of analysis (Fig. 4A, B). Additionally, 142 metastatic melanoma samples were included in the WGS study, comprising 21 in-transit metastases, 72 regional lymph node metastases, 34 distant metastases and 15 cell lines (Fig. 4A, C). The subtype of the likely causal antecedent ('culprit') primary was determined using a published algorithm,[30] indicating that the metastatic samples were derived from 19 acral lentiginous, four desmoplastic, 40 SS, 40 NM and eight SS with a dominant dermal nodule. In eight cases the primary melanoma was occult. The primary melanoma subtype for 23 samples is still to be determined.

Genomic DNA quantification and quality controls

During the DNA extraction process it became apparent that the eluted DNA remained pigmented when extracted from pigmented melanoma tissue. This was a concern as melanin has been shown to inhibit PCR reactions and can interfere with downstream library preparation.[13] Additionally, the DNA from pigmented samples demonstrated a significant over-estimation in DNA concentration when measured with the spectrometry-based NanoDrop compared with the fluorometric based Qubit assay ($p = 0.008$; Supplementary Fig. 2A, http://links.lww.com/PAT/A38): pigmented samples often measured many fold higher using the NanoDrop compared to the Qubit assay due to the absorbance attributable to melanin. The pigmented DNA

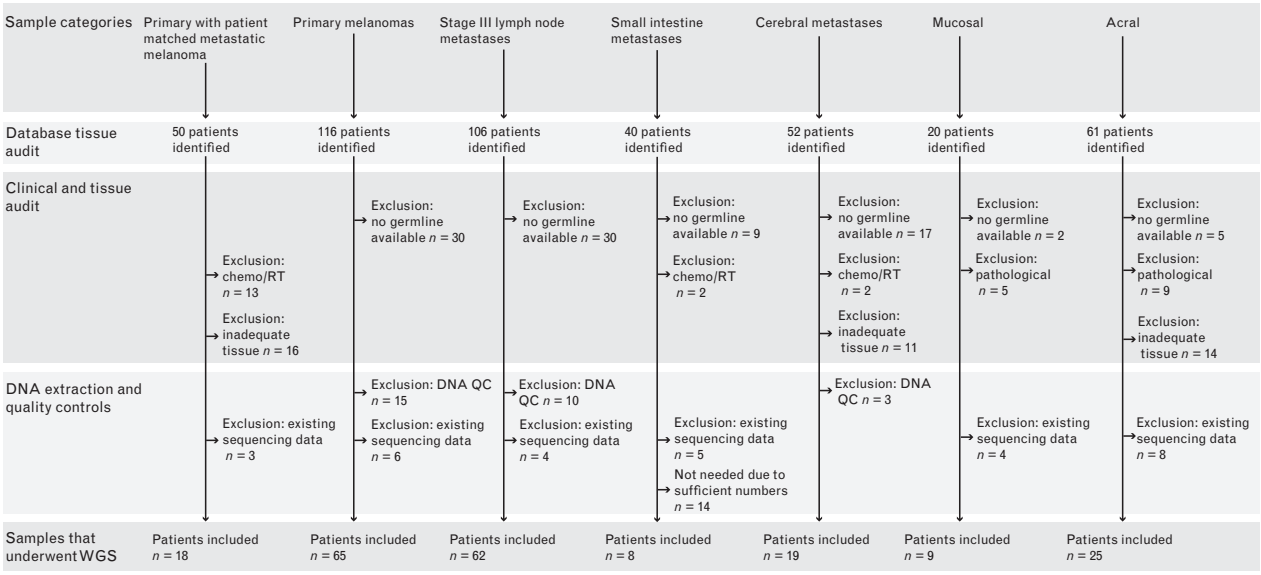


Fig. 3 Patient inclusion and exclusion process for the patient cohorts. A database audit of the biospecimen banks yielded candidate samples. Patient samples were excluded on the presence of appropriate germline samples, prior radio or chemotherapy, poor quality tumour (<80% tumour content or >30% necrosis), poor quality or quantity of genomic DNA (lack of high molecular weight band in electrophoresis gel, concentration <50 ng/μL and <1 μg total) or existence of prior exome or whole genome sequencing data.

also appeared heavily smeared on electrophoresis gels (Supplementary Fig. 2B, http://links.lww.com/PAT/A38).

Therefore any pigmented tissue samples that produced pigmented DNA, smearing on electrophoresis gels or had abnormally high NanoDrop readings underwent a cleanup process using a modified CTAB method.[24] We measured the DNA concentration before and after the clean-up process on both assays. The concentrations of the CTAB clean-up DNA was reduced in the NanoDrop readings but the Qubit readings were often stable or increased due to differences in elution volumes, suggesting minimal loss of double stranded DNA and removal of melanin ($p = 0.476$). Subsequent electrophoresis gels of the cleaned-up DNA show less DNA fragmentation following CTAB treatment (Supplementary Fig. 2B, http://links.lww.com/PAT/A38) and the copy number plots from the WGS data shows an intact genome (Supplementary Fig. 2C, http://links.lww.com/PAT/A38).

Optimal coverage level determination

In order to determine optimal coverage levels, two melanoma tissue samples and one melanoma cell line were sequenced to depths of at least 100X. Sequence data were then subsampled to test tumour coverage levels in one sequencing lane increments, ranging from 30–132X against a control cover level of three lanes of sequencing. For each coverage level, somatic variants were identified, average quality scores calculated and variants compared to dbSNP v135.[31]

For the melanoma tissue samples (SCC09 and SMU11), Table 1 shows <1% additional variants were identified at 124–132X compared to 61–64X coverage levels. Furthermore, somatic variants identified at 61–64X have mean SNV qualities of 212 and 216 while the variants added at 124–132X are of lower quality with means of 90 and 65 (Fig. 5A–D). However, 0.7–2.4% of SNVs were added at coverage levels less than 60X (Table 1). Consequently, for melanoma tissue samples we concluded that coverage above 60X only results in a minor increase in the total number of additional variants, the majority of which have quality scores that are indicative of false positive variants. The coverage analysis was then performed on the genome derived from melanoma cell line A15. This showed that increasing coverage from 30X to 40X resulted in a 1% increase in the total number of SNVs that were detected at 103X (Fig. 5E), with the gain in the number of SNVs levelling off below 1% beyond 43X. Less than 1% more variants were identified at 103X compared to 43X and these variants had a mean quality of 70 compared to 209 mean quality for variants at 43X (Fig. 5F). Therefore, we selected a minimum coverage of 40X for all cell line derived samples. Likewise, coverage below 40X for blood derived DNA resulted in a reduction of 1.5–5% of total SNVs compared to coverage greater than 40X (Table 1 and Fig. 5G–I). Therefore coverage of at least 40X was recommended for blood derived genomes Fig. 4D.

Variant detection optimisation: independent versus paired variant detection

Two options for somatic variant detection were initially considered. The first option was to call variants independently on the tumour and normal samples using SAMtools/BCFtools followed by extraction of tumour unique somatic variants by detecting variants present in the tumour sample and absent in the normal sample. The second option was to call variants simultaneously using SAMtool's paired mode followed by an additional filtering step on the CLR score. Several samples were analysed to determine which method generated more reliable tumour specific variant calls.

Detection of variants simultaneously yielded an average of 40% less somatic variants with virtually all variant calls being a subset of variants detected using the independent detection method (Fig. 6). Further examination of the 40% of variants detected solely with the independent method revealed that they generally fell into the following categories: (1) the tumour sample contained a variant while the normal sample had insufficient coverage to make a variant call either way, or (2) the tumour sample variant was slightly above the variant cut-off score while the normal sample was slightly below the variant cut-off (Supplementary Fig. 3, http://links.lww.com/

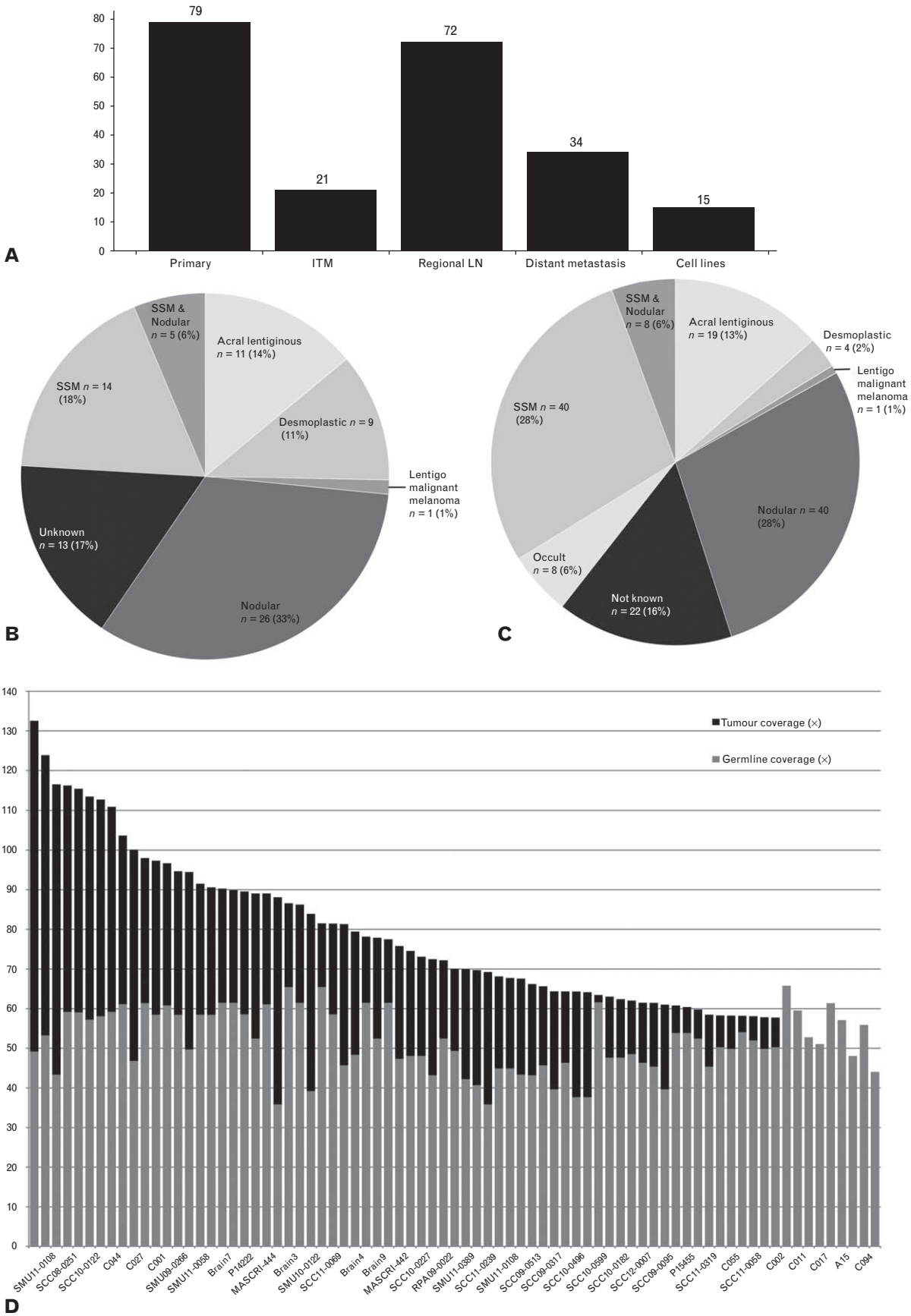


Fig. 4 Summary of melanoma sample type, melanoma subtype of the primary melanoma and coverage levels of the samples within the study. (A) Histogram of the samples included in the WGS study and their various tissue types; (B) melanoma subtype of the primary melanoma samples; (C) melanoma subtype of the culprit primary melanoma of all the metastatic melanoma samples; and (D) histogram depicting the coverage levels of the blood (mean 51X) and tumour (mean 79X) genomes of the samples process at time of publication.

Table 1 Depth of coverage and single nucleotide variant (SNV) detection

Sample	Coverage (Lanes)	Total somatic SNVs	Number SNVs added at this coverage level	Mean SNV quality of added SNVs	Percentage of total SNVs detected
A15 (tumour)	30X (3)	98297	N/A	N/A	98.4
A15 (tumour)	43X (4)	99267	970	76	99.4
A15 (tumour)	56X (5)	99543	276	67	99.7
A15 (tumour)	70X (6)	99738	158	65	99.9
A15 (tumour)	80X (7)	99740	2	62	99.9
A15 (tumour)	93X (8)	99868	128	66	100.0
A15 (tumour)	103X (9)	99884	16	64	
SCC09 (tumour)	38X (3)	157338	N/A	N/A	97.6
SCC09 (tumour)	51X (4)	159352	2014	85	98.8
SCC09 (tumour)	64X (5)	160255	903	87	99.4
SCC09 (tumour)	79X (6)	160699	444	90	99.7
SCC09 (tumour)	92X (7)	160955	256	90	99.8
SCC09 (tumour)	105X (8)	161098	143	97	99.9
SCC09 (tumour)	119X (9)	161157	59	97	100.0
SCC09 (tumour)	132X (10)	161214	57	97	
SMU11 (tumour)	36X (3)	322406	N/A	N/A	98.3
SMU11 (tumour)	47X (4)	325846	3440	77	99.3
SMU11 (tumour)	61X (5)	327149	1303	69	99.7
SMU11 (tumour)	74X (6)	327660	511	66	99.9
SMU11 (tumour)	86X (7)	327891	231	63	99.9
SMU11 (tumour)	99X (8)	328051	160	61	100.0
SMU11 (tumour)	111X (9)	328089	38	61	
SMU11 (tumour)	124X (10)	328142	53	60	
A15 (Lymphoblast cell line)	33X (3)	93664	N/A	N/A	93.9
A15 (Lymphoblast cell line)	46X (4)	98212	4548	178	98.5
A15 (Lymphoblast cell line)	55X (5)	99738	1526	136	
SCC09 (blood)	36X (3)	152682	N/A	N/A	95.0
SCC09 (blood)	49X (4)	160699	8017	183	
SMU11 (blood)	39X (3)	322659	N/A	N/A	98.5
SMU11 (blood)	53X (4)	327660	5001	159	

PAT/A38). As variants in these categories are likely to either be false positives or inconclusive we concluded that the simultaneous detection method yielded more reliable somatic variant calls and was utilised in our analysis.

DISCUSSION

The accuracy and sensitivity of WGS data depends on appropriate sample preparation, sequencing specifications and the selection of an appropriate mutation detection algorithm. The current study determined the optional WGS coverage for human melanoma tissue and blood genomes and optimised freely available mutation detection algorithms to accurately call mutations unique to the tumour genome, whilst limiting false positive calls.

Our study found that an increase in sequencing coverage from 60X to 120X in melanoma samples resulted in a minimal increase in the detection of SNVs, the quality of which indicated they could be false positive calls or in a minor subpopulation of tumour cells. Likewise, melanoma cell line genomes reach saturation for SNV discovery at around 40X, most likely due to their greater purity and increased homogeneity. The study suggests that these coverage levels empower the detection of 99% of informative SNVs, which would represent a sensitivity level at which clinical decision making or landscape research projects can be carried out with a high degree of confidence. However, lower coverage WGS (5–15X) has been used to detect high frequency variants in genome wide association studies and by The Cancer Genome Atlas,[32] but the sensitivity to detect rare or low frequency events is diminished. Nevertheless, variant detection algorithms and sequencing technologies are constantly evolving, with improvements being made to the accuracy, speed and ability to accommodate lower levels of coverage.[33,34] Therefore, the appropriate coverage levels for a specific project may vary depending on the tissue type, sequencing technologies, detection algorithms and the type of questions the data needs to address.

Likewise the detection methodology, subtraction versus the simultaneous variant detection, had significant effect on the number and quality of SNVs detected, finding that the subtraction methodology with SAMtools resulted in a 41% increase in the number SNVs detected over the simultaneous detection method. These results are similar to those of other studies that found the additional SNV calls are the result of library preparation bias, sequencing errors and mutation detection artifact due to read depth imbalances that represent false positive calls.[25] Therefore, the simultaneous detection approach should be employed on matched tumour and control genomes to reduce false positive variant calls.

The current study highlights some basic steps that are required to achieve high quality input DNA and accurate sequencing data from clinical melanoma specimens. The key to achieving this is the accurate evaluation and selection of areas of tumour for macrodissection by a skilled pathologist. With careful macrodissection of the pathologist-selected tumour region and subsequent quality controls, factors that contribute to erroneous results such as low tumour content, excessive necrosis or misclassified samples can be better avoided. If the tissue analysed is suboptimal, this can have long-term implications for further research as the genome sequencing results are made publicly available for use by the wider cancer research community. Therefore the process of sample selection and preparation is critical in WGS as subsequent results depend upon the accuracy of these steps.

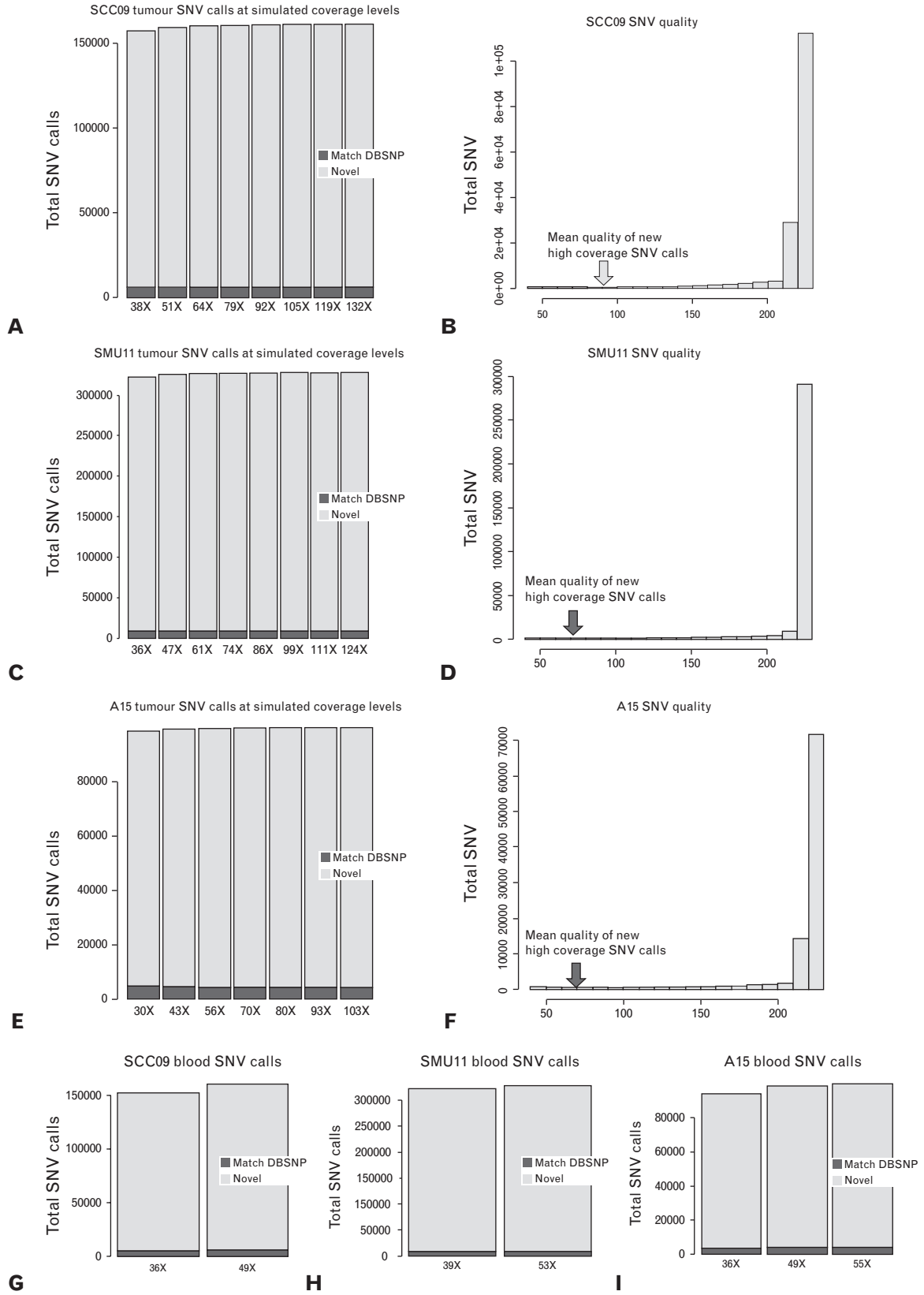


Fig. 5 Number and quality of SNV detection at simulated coverage levels. (A,B) SCC09 tumour genome: increasing coverage from 38 to 132X only generated 2.5% additional SNVs, with a total of 161,214 SNVs identified at 132X at an average quality of 212 (arrow, mean quality of SNVs detected greater than 32X). (C,D) SMU11 tumour genome: increasing coverage from 36 to 124X only generated 1.8% additional SNVs, with a total of 328,142 SNVs identified at 124X at an average quality of 216 (arrow, mean quality of SNVs detected greater than 36X). (E,F) A15 tumour genome: increasing coverage from 30 to 103X only generated 1.6% additional SNVs, with a total of 99,543 SNVs at an average quality of 209 (arrow, mean quality of SNVs detected greater than 30X). (G) SCC09 blood genome: increasing the coverage from 36 to 49X added 8017 (5%) additional SNVs with a mean quality of 212 for all SNVs identified. (H) SMU11 blood genome: increasing the coverage from 39 to 53X added 5001 (1.5%) additional SNVs with a mean quality of 216 for all SNVs identified. (I) A15 lymphoblast cell line: increasing the coverage from 33 to 46X added 4548 (5%) and from 46 to 55X added another 1526 (1.5%) SNVs with a mean quality of 209 for all SNVs.

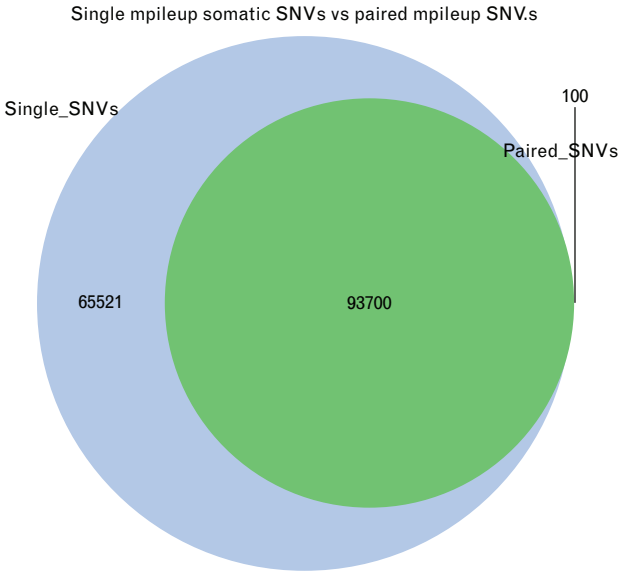


Fig. 6 Venn diagram depicting the overlap between the subtraction and simultaneous detection methods using SAMtools mpileup. 70% more candidate SNVs were identified with the subtraction method (SNV single) and virtually no new variants added with paired method (paired SNVs).

Melanoma tissue preparation for WGS is especially problematic for sequencing due to naturally occurring melanin which interferes with spectrometer-based DNA quantification and PCR reactions.[24] The presence of melanin in the library preparation can inhibit library preparation and may even induce sequencing artifact. We sought to remove pigmentation prior to library preparation using the CTAB clean-up method. Results show this process removes pigment and improves the quality of the genomic material on electrophoresis gels, which is then suitable for WGS. The results show that cleaning the DNA with CTAB results in minimal loss of genomic material as measured via Qubit and the copy number plots of the WGS of these samples appears to be consistent with non-pigmented tumour genomes. Therefore, a fluorometric based method should be used to accurately quantify double-stranded DNA concentrations in melanoma-derived DNA and a modified CTAB clean-up protocol can be used to produce suitable genomic material for WGS from heavily pigmented samples.

The Australian Melanoma Genome Project has performed WGS on 222 human melanoma samples and gained valuable insight into WGS and analysis of clinical melanoma samples. Data generated by the project will be made publicly available in order to make a substantial impact on future melanoma diagnosis and treatment. The current study outlines practical guidelines for sample preparation, quality control, sequencing methodology and mutation detection algorithms that will aid researchers and clinicians in sequencing patient-derived melanoma samples on a large scale.

Conflicts of interest and sources of funding: This work was supported by Melanoma Institute Australia, the Australian Government through Bioplatforms Australia, the NSW Ministry of Health, and a Cancer Council NSW project grant (RG12/13), as well as by Program grants (402761, 633004) of NHMRC (to JT, GM, RS, NH), Translational Research Program Grants (05TPG1-01, 10TPG1-02) of Cancer Institute NSW (to GM, JT, RS), Australian Research Council discovery grant (DP1301004 to GM), a NHMRC Research Fellowship (to NH) and a Cancer Institute NSW Clinical Research Fellowship and NHMRC Practitioner Fellowships (to RS).

Address for correspondence: Prof Richard A. Scolyer, Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Missenden Road, Camperdown, NSW 2050, Australia. E-mail: richard.scolyer@sswahs.nsw.gov.au

References

1. Scolyer RA, Thompson JF. Biospecimen banking: The pathway to personalized medicine for patients with cancer. *J Surg Oncol* 2013; 107: 681–2.
2. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time. *Genet Med* 2011; 13: 499–504.
3. Hayden EC. Is the $1,000 genome for real? *Nature News* 15 Jan 2014. http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530.
4. Van Allen EM, Wagle N, Stojanov P, *et al*. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014; 20: 682–8.
5. Roychowdhury S, Iyer MK, Robinson DR, *et al*. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* 2011; 3: 111–21.
6. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *N Engl J Med* 2014; 370: 2418–25.
7. Rehm HL, Bale SJ, Bayrak-Toydemir P, *et al*. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013; 15: 733–47.
8. Hirsch FR, Wynes MW, Gandara DR, Bunn PA. The tissue is the issue: personalized medicine for non-small cell lung cancer. *Clin Cancer Res* 2010; 16: 4909–11.
9. Long GV, Wilmott JS, Capper D, *et al*. Immunohistochemistry is highly sensitive and specific for the detection of V600E BRAF mutation in melanoma. *Am J Surg Pathol* 2013; 37: 61–5.
10. Song S, Nones K, Miller D, *et al*. Qpure: a tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS One* 2012; 7: e45835.
11. Van Loo P, Nordgard SH, Lingjaerde OC, *et al*. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* 2010; 107: 16910–5.
12. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 1988; 2: 231–9.
13. Bentley DR, Balasubramanian S, Swerdlow HP, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; 456: 53–9.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; 25: 1754–60.
15. Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; 25: 2078–9.
16. Ruffalo M, LaFramboise T, Koyuturk M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 2011; 27: 2790–6.
17. Pabinger S, Dander A, Fischer M, *et al*. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 2014; 15: 256–78.
18. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* 2014; 15: 244.
19. O'Rawe J, Jiang T, Sun G, *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013; 3: 28.
20. Mann GJ, Pupo GM, Campain AE, *et al*. BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. *J Invest Dermatol* 2013; 133: 509–17.
21. Viros A, Fridlyand J, Bauer J, *et al*. Improving melanoma classification by integrating genetic and morphologic features. *PLoS Med* 2008; 5: e120.
22. Wilmott JS, Long GV, Howle JR, *et al*. Selective BRAF inhibitors induce marked T-cell infiltration into human metastatic melanoma. *Clin Cancer Res* 2012; 18: 1386–94.
23. Long GV, Wilmott JS, Haydu LE, *et al*. Effects of BRAF inhibitors on human melanoma tissue before treatment, early during treatment, and on progression. *Pigment Cell Melanoma Res* 2013; 26: 499–508.
24. Stefania Lagonigro M, Cecco LD, Carninci P, *et al*. CTAB–Urea method purifies RNA from melanin for cDNA microarray analysis. *Pigment Cell Res* 2004; 17: 312–5.
25. Andrews TD, Whittle B, Field MA, *et al*. Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. *Open Biol* 2012; 2: 120061.

26. McLaren W, Pritchard B, Rios D, *et al*. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010; 26: 2069–70.
27. Sherry ST, Ward MH, Kholodov M, *et al*. DbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; 29: 308–11.
28. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med* 2004; 10: 789–99.
29. Forbes SA, Bhamra G, Bamford S, *et al*. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* 2008; Chapter 10: Unit 10.11.
30. Murali R, Brown PT, Kefford RF, *et al*. Number of primary melanomas is an independent predictor of survival in patients with metastatic melanoma. *Cancer* 2012; 118: 4519–29.
31. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (Oct 13, 2011 edn, Vol. dbSNP 135: dbSNPv135). http://www.ncbi.nlm.nih.gov/SNP/.
32. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* 2015; 161: 1681–96.
33. Bizon C, Spiegel M, Chasse SA, *et al*. Variant calling in low-coverage whole genome sequencing of a Native American population sample. *BMC Genomics* 2014; 15: 85.
34. Li Y, Sidore C, Kang HM, Boehnke MA, Becasis GR. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res* 2011; 21: 940–51.

4.2 Further discussion

This publication describes work to build a new variant detection pipeline for 150 whole melanoma genome paired samples, which at the time, represented the largest melanoma genome sequencing project anywhere. To do this the following analyses were required:

1) Identify the best software for reliably detecting somatic variants within paired normal/tumour samples
2) For all library types (cell line, primary tissue, and metastatic tissue) determine optimal coverage levels for both tumour and normal samples
3) Create a complete variant detection pipeline for detecting somatic mutations that is robust, high-throughput, automated, and reproducible

First, in order to determine an appropriate method for somatic variant detection in paired tumour/normal samples, a review of existing tools was required. There exists a plethora of genomic analysis tools with a recent publication surveying 205 such tools (21), many of which prove suitable for detecting somatic variants. The challenge is that the underlying statistical approach of these tools differs, resulting in output variant lists highly variable with regard to both quality and quantity of variant calls. This means there is not, as of yet, any gold standard methodology for detecting somatic variants in cancer; in fact even within large cancer genome projects such ICGC (49), multiple workflows using a variety of tools are supported. In general, detecting somatic mutations in paired tumour/normal samples follows one of two fundamentally different approaches. The first approach involves calling variants independently in both the tumour and the normal and obtaining somatic mutations by filtering out variants from the tumour sample that are shared with the germ line sample. This method is sometimes called the 'subtraction method' and has been used in several high profile cancer studies (83) and has the advantage of not requiring any modifications to existing variant detection workflows, only requiring an additional filtering step. The second approach is to simultaneously analyse matched tumour and normal samples using a joint probability-based statistical approach, the basis of methods such as MuTect (51), or SAMtools somatic mode (78). This method is called the 'pairwise method' and has the advantage of considering the nature of the sequence content from both the tumour and the

normal sample before classifying a variant as somatic. To determine which method was more suitable for melanoma samples, I ran all three-sample types (cell line, primary and metastatic tumour samples) through SAMtools using both the default mode and the paired somatic mode for comparison. Overlapping the variant lists from the two method showed the subtraction method generated substantially more somatic variant calls than the pairwise method, typically in the order of 40% more variants, and further that all variants detected by the pairwise method were also detected using the subtraction method. Closer examination of the additional variants detected by the subtraction method revealed the variants fell into two categories: 1) the tumour sample contained sufficient depth and non-reference bases to score above the variant cut-off while the normal sample had insufficient coverage to make a definitive variant call either way or 2) both tumour and normal contained variant bases with the tumour variant score slightly above the variant cut-off and the normal variant score slightly below the variant cut-off. As variants in these categories are likely to be either inconclusive (as in case 1) or false positives (as in case 2) the pairwise method was ultimately chosen due to it yielding a more reliable set of somatic variant calls containing a smaller number of false positives. These conclusions support previous studies demonstrating the superior performance of pairwise variant detection compared to subtraction variant detection (84).

Once the variant detection method was selected, the optimal coverage levels needed to be determined for all library types utilised in this project; namely tumour and normal samples from cell lines, primary tissues, and metastatic tissue. These levels were important to establish specifically for melanoma, as there existed no large-scale melanoma whole genome sequencing repositories at the time. Melanoma-specific levels needed to be determined empirically as tumour purity (32) and mutational heterogeneity (52) are different for each cancer type, with both factors known to affect the required coverage levels. To determine the values, a pilot project was undertaken where one of each sample type was sequenced to extremely high depth for both tumour and normal libraries. Sub-sampling was utilised to simulate lower coverage levels and variants called at each level, with each level designed to be roughly 10X apart. First, to establish baselines, variants detected both at the lowest and highest simulated coverage level (30X and 120X respectively) were identified. Then for

each 10X increment in coverage, variants not detected at previous coverage levels were identified and analysed independently to access the quality of the variants unique to each coverage level. The final summary contained the per cent of the total variants detected at each coverage level as well as the number and quality of the new variants detected at that level, thus allowing an informed decision to be made. As sequencing depth decisions typically boil down to a cost-benefit analysis between the sensitivity/specificity of variant calls and the number of samples able to be sequenced given a fixed budget, it was important to make an informed decision based on real empirical data from melanoma samples. These pilot analyses were subsequently used as the sole information on which to base the required sequence depth to be obtained for tumour and normal tissue samples. This analysis yielded several important findings that ultimately resulted in more effective use of the limited resources available for the project. First, optimal coverage levels differ for cell line libraries and libraries from tissue samples, likely explained by issues with cross contamination of both tumour and normal samples. Second, there is very little gain in the number of variant calls beyond 60X for tumour tissue samples and 40X for cell line tumour samples with at least 98% of the total variants are detected by these coverage levels, results that differ from Illumina's current suggestion of 80X minimum tumour coverage (http://www.illumina.com/technology/next-generation-sequencing/deep-sequencing.html). Further, variants added at the higher coverage levels were uniformly of lower quality than variants detected at lower coverage levels and more likely to be false positives.

Lastly, an analysis pipeline that both manages and analyses the large number of melanoma samples was needed, a pipeline that was flexible, high-automated, and supported reproducible bioinformatic workflows. At the time, we had a mature variant detection pipeline for mouse exomes (see Chapter 7) (69) that met many of these requirements as it was running in an high performance computing (HPC) environment on the supercomputer raijin, housed within the National Computational Infrastructure, the largest cluster in the southern hemisphere (https://nci.org.au/systems-services/national-facility/peak-system/raijin/). In the end, borrowing many of the same design principles, a more sophisticated melanoma variant detection pipeline was created to handle the increasingly large and complex sequence data sets. Numerous additional

features were incorporated including the ability to manage multiple sequencing lanes per sample, the ability to call pairwise variant calls from tumour and normal samples, the addition of cancer annotation information, and the parallelisation of the variant detection step previously requiring 10 days to run in serial. The final version of the pipeline effortlessly processed the melanoma samples and remains in production today following the addition of several configurable components that allow it to be used with a variety of human and mouse cancer types.

Chapter 5: Comparison of predicted and actual consequences of missense mutations

Miosge, L. A., **M. A. Field**, Y. Sontani, V. Cho, S. Johnson, A. Palkova, B. Balakishnan, R. Liang, Y. Zhang, S. Lyon, B. Beutler, B. Whittle, E. M. Bertram, A. Enders, C. C. Goodnow and T. D. Andrews. "Comparison of predicted and actual consequences of missense mutations." Proceedings of the National Academy of Sciences of the United States of America. 2015;15;112(37):E5189-98

Comparison of predicted and actual consequences of missense mutations

Lisa A. Miosge[a,1], Matthew A. Field[a,1], Yovina Sontani[a], Vicky Cho[a,b], Simon Johnson[a,b], Anna Palkova[a,b], Bhavani Balakishnan[b], Rong Liang[b], Yafei Zhang[b], Stephen Lyon[c], Bruce Beutler[c], Belinda Whittle[b], Edward M. Bertram[b], Anselm Enders[d], Christopher C. Goodnow[a,e,2,3], and T. Daniel Andrews[a,2,3]

[a]Immunogenomics Laboratory, John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia; [b]Australian Phenomics Facility, John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia; [c]Center for Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX 75390; [d]Ramaciotti Immunisation Genomics Laboratory, John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia; and [e]Immunogenomics Laboratory, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia

Contributed by Christopher C. Goodnow, July 9, 2015 (sent for review February 24, 2015; reviewed by Jean-Laurent Casanova and Alain Fischer)

Each person's genome sequence has thousands of missense variants. Practical interpretation of their functional significance must rely on computational inferences in the absence of exhaustive experimental measurements. Here we analyzed the efficacy of these inferences in 33 de novo missense mutations revealed by sequencing in first-generation progeny of *N*-ethyl-*N*-nitrosourea–treated mice, involving 23 essential immune system genes. PolyPhen2, SIFT, MutationAssessor, Panther, CADD, and Condel were used to predict each mutation's functional importance, whereas the actual effect was measured by breeding and testing homozygotes for the expected in vivo loss-of-function phenotype. Only 20% of mutations predicted to be deleterious by PolyPhen2 (and 15% by CADD) showed a discernible phenotype in individual homozygotes. Half of all possible missense mutations in the same 23 immune genes were predicted to be deleterious, and most of these appear to become subject to purifying selection because few persist between separate mouse substrains, rodents, or primates. Because defects in immune genes could be phenotypically masked in vivo by compensation and environment, we compared inferences by the same tools with the in vitro phenotype of all 2,314 possible missense variants in *TP53*; 42% of mutations predicted by PolyPhen2 to be deleterious (and 45% by CADD) had little measurable consequence for *TP53*-promoted transcription. We conclude that for de novo or low-frequency missense mutations found by genome sequencing, half those inferred as deleterious correspond to nearly neutral mutations that have little impact on the clinical phenotype of individual cases but will nevertheless become subject to purifying selection.

de novo mutation | immunodeficiency | evolution | nearly neutral | cancer

The genome sequence of any particular person contains extensive protein-altering genetic variation and de novo point mutations (1), of which only a minority are unambiguously deleterious and introduce premature stop codons or disrupt normal mRNA splicing. When considering a person with a suspected genetic illness, the first clinical question is whether any of the mutations identified in their genome sequence involve essential genes whose disruption is known to cause a phenotype resembling that of the patient in question. The most numerous class of protein-altering mutations is missense mutations, where a single codon is altered to encode a different amino acid. On average, 2% of people carry a missense mutation in any given gene (2). Hence, by chance, missense mutations will often be found in genes that are seemingly relevant to a person's disease phenotype, and the next key clinical question is whether or not these substitutions alter the function of the corresponding protein. Short of mutational studies of all possible amino acid substitutions coupled with comprehensive functional assays, the sheer number and diversity of missense mutations present in each person's genome means that their functional importance must presently be addressed primarily by computational inference.

Many tools now exist that use diverse information to make inferences of the functional importance of single amino acid substitutions (3, 4). The majority of better-performing tools use a protein multiple sequence alignment and judge the importance of residues depending on their conservation across available homologous sequences. Notably, these tools are sensitive to the sequence choice in the input alignment (5). Some examples of commonly used tools that use a sequence conservation approach, with diverse methods to calculate conservation, are SIFT (6), MutationAssessor (7), MAPP (8), AlignGVGD (9), PANTHER (10), and GERP (11). Protein structural information is also included in other tools to judge whether an amino acid substitution may importantly alter protein stability or catalysis, but few commonly used tools rely just on these data alone (3).

The widely used PolyPhen2 and CADD tools (12, 13) integrate a number of different information sources, including sequence- and structure-based features (and in the case of CADD, the results of other tools), and use a machine learning approach to categorize variants as benign or deleterious. It is worth noting that

Significance

Computational tools applied to any human genome sequence identify hundreds of genetic variants predicted to disrupt the function of individual proteins as the result of a single codon change. These tools have been trained on disease mutations and common polymorphisms but have yet to be tested against an unbiased spectrum of random mutations arising de novo. Here we perform such a test comparing the predicted and actual effects of de novo mutations in 23 genes with essential functions for normal immunity and all possible mutations in the *TP53* tumor suppressor gene. These results highlight an important gap in our ability to relate genotype to phenotype in clinical genome sequencing: the inability to differentiate immediately clinically relevant mutations from nearly neutral mutations.

Author contributions: L.A.M., M.A.F., B. Beutler, B.W., E.M.B., A.E., C.C.G., and T.D.A. designed research; L.A.M., M.A.F., Y.S., V.C., S.J., A.P., B. Balakishnan, R.L., Y.Z., S.L., B.W., and T.D.A. performed research; S.L., B. Beutler, and T.D.A. contributed new reagents/analytic tools; L.A.M., M.A.F., Y.S., V.C., S.J., A.P., B. Balakishnan, R.L., Y.Z., S.L., B. Beutler, B.W., E.M.B., A.E., C.C.G., and T.D.A. analyzed data; and M.A.F., C.C.G. and T.D.A. wrote the paper.

Reviewers: J.-L.C., The Rockefeller University; and A.F., Imagine Institute, Hôpital Necker Enfants Malades.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

See Commentary on page 11426.

[1]L.A.M. and M.A.F. contributed equally to this work.

[2]C.C.G. and T.D.A. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: c.goodnow@garvan.org.au or dan.andrews@anu.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1511585112/-/DCSupplemental.

inference tools that integrate diverse information, such as PolyPhen2 and SNAP (14), do not necessarily have efficacy better than simpler tools (3). Interestingly, there is further utility in integrating the predictions of individual methods (15), even those that are very similar or that already integrate wide ranges of information, potentially due to minimization of outlier effects.

Functional inferences of the severity of protein disruption only weakly predict disease incidence, disease severity, or clinical outcome. For example, functional inferences of missense variants in the cystic fibrosis gene, *CFTR*, are not well correlated with disease incidence or severity (16). Similarly, the functional inference of mutation severity in the tumor suppressor gene, *TP53*, does not correlate significantly with patient clinical outcome (17). An emerging consensus has formed that functional inference scores lack the sensitivity and specificity for their clinical use (18–20). The human genome of any particular individual likely contains many apparently unambiguous disease-associated variants (1), and these very often occur in people without symptoms (21). Variable penetrance is most frequently cited to explain the presence of deleterious variants in asymptomatic individuals, although a prevalent viewpoint is that functional inference tools overcall pathogenic variants (22). Most validation of the various predictive algorithms have used test sets of selected mutations that are known or likely to be disease causing (such as recurrent mutations inferred to be driver mutations in cancer) and likely to be benign (such as germ-line polymorphisms present at high frequency in the human population). What remains to be tested, in essential genes that produce a clear mutant phenotype, is how often do predicted deleterious variants produce the expected mutant phenotype? This question sidesteps the thicket of variable penetrance issues and allows direct appraisal of whether functional inferences accurately predict disruption of the gene/protein in question.

Here we addressed this question by breeding to homozygosity and phenotyping mice with 1 of 33 potentially disruptive de novo point mutations residing in 23 essential immune system genes. Should a mutation disrupt the function of the gene in which it occurs, we expected to observe a previously characterized immune system phenotype specific to that gene. We found that only 4 of 20 missense mutations predicted to be "probably damaging" or "possibly damaging" by PolyPhen2 produced the expected mutant phenotype. This apparent overcalling of deleterious variants by computational tools was not simply due to masking of immune defects in vivo, because when the algorithms were compared with systematic in vitro phenotyping of all possible missense mutations in *TP53*, only half of the mutations predicted to be deleterious caused a clear reduction of transcription-enhancing activity. Interestingly, the large set of apparent false-positive *TP53* inferences has a mean distribution of measured transcriptional activities that is 14.2% lower than the set of *TP53* mutations predicted to be benign, consistent with population genetics models that predict a large class of nearly neutral mutations. We present evidence that the apparent overcalling of deleterious mutations by functional inference tools is not a failure of the inference tools per se, but elucidates a critical gap between these inferences and our understanding of how mutations yield phenotypes measured in individuals as opposed to small differences in fecundity compounded over hundreds of generations in large populations of competing individuals.

Results

Identification of Induced De Novo Mutations in Essential Immune Genes. To identify large numbers of de novo point mutations that have yet to be subject to phenotypic selection, we previously developed a system to exome sequence and accurately identify single nucleotide, protein-altering variants in the progeny of C57BL/6 laboratory mice exposed to the spermatogonial point-mutagen *N*-ethyl-*N*-nitrosourea (ENU) (23). As part of a larger project aimed at producing an ENU-induced mouse mutant for each gene in the mouse genome, we developed a resource of mice with identified mutations and corresponding exome sequence data (databases.apf.edu.au/mutations and mutagenetix.org/incidental/incidental_list.cfm). First generation (G1) offspring from male mice treated with ENU have a mean of 45 nonsynonymous de novo mutations (23), although the functional effect of each mutation is largely unknown. Mutations induced by exposure to ENU are likely to have diverse functional consequences, from benign to severely deleterious. This diversity replicates that of spontaneous nonsynonymous de novo mutations that arise in humans at an average rate of 0.75 mutations per child (24). It is also likely to replicate the functional spectrum of inherited nonsynonymous variants in recessive genes that occur at frequencies below 1% in the human population.

To gain an understanding of the functional effects of these phenotypically unselected point mutations, we propagated 33 mutations within 23 genes and bred mouse pedigrees to bring these to homozygosity. Mutations were propagated if they produced a nonsynonymous alteration in genes that, when rendered null, were already known to cause a fully penetrant, well-characterized phenotype in the mouse immune system that was readily detectable by flow cytometry of peripheral blood lymphocytes (details of these genes and the expected phenotypes are given in Table S1). Loss-of-function mutations in 11 of the 23 studied genes (*BTK, DCLRE1C, DOCK8, IL2RA, IL7R, JAK3, LIG4, PRKDC, PTPRC, RAG1,* and *RAG2*) are also already known to cause human immune deficiency with Mendelian inheritance.

Mouse-Specific Calculation of Mutation Functional Impact. The majority of tools that infer the functional consequences of missense sequence variants are understandably built to analyze human sequence data. By default, the inferences made with these tools for model organisms, such as the mouse, using human-specific data sources are inherently of lower and mixed accuracy. We substituted the internal data sources within PolyPhen2 to produce mouse-specific values (see *Materials and* Methods for details). Table S2 shows PolyPhen2 values for the 33 de novo mutations in immune genes, calculated using both human- and mouse-specific implementations, and human inferences for the mouse mutations made with CADD. Although the mouse- and human-specific PolyPhen2 values are generally well correlated, several values cannot be calculated in humans due to lack of conservation between mouse and human amino acids at the position in question (which is also a limitation when calculating CADD scores for mouse mutant orthologs from human sequence information). Of the values that can be calculated, several PolyPhen2 values change between deleterious and benign categories (*Il7r, Lig4, Rasgrp1,* and *Tnfaip3*) depending on the species of database used. Human-specific PolyPhen2 and CADD scores are also highly similar.

We also calculated functional impact inferences using the sequence homology-based methods SIFT (6) GERP (11), MutationAssessor (7), and PANTHER (10), using mouse input sequences (Table 1 and Table S3) and generated weighted average scores from these diverse measures using Condel (15). Table 1 shows that the functional inference scores for each of the de novo mutations are generally in agreement. There are small differences between functional inference tools regarding the deleterious/benign cutoff used by each different tool, but these are not a predominant feature of the data (Table S3). Occasionally, a single tool produces a discordant inference to the other tools, such as the benign call by PolyPhen2 of the *Ptpn6* variant at chr6:124682374. Potentially one tool is making a better call than each of the other tools, although as Condel validation shows, removal of these outlier functional inferences improves the accuracy of the inferences overall (15).

Table 1. Predicted and observed effect of unselected de novo ENU mutations in 23 essential immune genes

Gene symbol	Num hom tested	Mut phen obs?	UniProt identifier	AA change	Polyphen category	Polyphen score	SIFT cat.	GERP score*	Mut Assessor cat.	PANTH subPSEC[†]	Condel cat.	CADD Phred-like score[‡]
Dclre1c	3	Yes	Q8K4J0	L95Stop								
Dock8	1	Yes	Q8C147	R1630Stop								
Dock8	3	Yes	Q8C147	F1885Stop								
Ets1	4	Yes	P27577	P334L	Prob. Dam.	1	Del.	3.97	Medium	−2.3993	Del.	19.09
Hnrnpl	5	No	Q8R081	A118D	Prob. Dam.	1	Del.	3.74	Medium	−4.06474	Del.	22
Tnfaip3	3	No	Q60769	F394S	Prob. Dam.	1	Del.	3.27	Medium	[§]	Del.	16.78
Satb1	2	No	Q60611	V99A	Prob. Dam.	0.999	Toler.		Low	−4.11365	Neutral	20.7
Btk	3	Yes	P35991	Y152H	Prob. Dam.	0.998	Del.	3.91	Low	−2.39271	Del.	27
Prkdc	8	No	P97313	D382G	Prob. Dam.	0.997	Del.	3.91	Medium	−3.61416	Del.	23.6
Il7r	3	Yes	P16872	T56P	Prob. Dam.	0.985	Del.	2.92	Low	[§]	Del.	12.47
Lig4	4	No	Q8BTF7	Y335C	Prob. Dam.	0.985	Del.	2.32		[§]		25.7
Itch	1	No	Q8C863	H563L	Prob. Dam.	0.972	Del.	3.61	Medium	−3.98029	Del.	21.5
Tbx21	7	No	Q9JKD8	E481A	Prob. Dam.	0.971	Toler.	3.22	Low	−0.97263	Neutral	15.79
Prkdc	4	No	P97313	I1010T	Poss. Dam.	0.935	Del.	3.17	Medium	−1.55087	Del.	25.9
Bcl2	3	No	Q4VBF6	T175A	Poss. Dam.	0.924	Toler.	3.04	Neutral	−2.00435	Neutral	14.68
Dock8	5	No	Q8C147	N1567Y	Poss. Dam.	0.904	Del.	3.66	Medium	−4.15438	Del.	26
Dock8	5	No	Q8C147	K26E	Poss. Dam.	0.894	Del.	3.92	Low	[§]	Del.	[¶]
Rag1	1	Yes	P15919	E803G	Poss. Dam.	0.879	Del.	3.83	High	−2.75501	Del.	29.2
Il2ra	2	No	P01590	V230A	Poss. Dam.	0.763	Del.	1.89	Low	−2.25488	Del.	[¶]
Rasgrp1	4	No	Q9Z1S3	K659R	Poss. Dam.	0.712	Toler.	3.49	Low	−1.80793	Neutral	22.2
Jak3	2	No	Q62137	I663V	Poss. Dam.	0.705	Del.		Low	−3.23261	Del.	[‖]
Cd74	1	No	P04441	I203F	Poss. Dam.	0.434	Toler.	−2.02	Medium	−2.04615	Neutral	17.91
Prkdc	5	No	P97313	V3389L	Benign	0.346	Toler.	3.73	Medium	[§]	Neutral	21.5
Ptprc	3	No	P06800	K921R	Benign	0.322	Toler.	2.64	Low	−0.99775	Neutral	21.9
Rag2	2	No	P21784	D424G	Benign	0.151	Toler.	3.51	Medium	−2.32166	Neutral	[¶]
Prkdc	2	No	P97313	Y2044F	Benign	0.097	Toler.	3.35	Medium	−2.36465	Neutral	13.5
Tnfaip3	1	No	Q60769	K41E	Benign	0.049	Toler.		Neutral	−0.78466	Neutral	22.3
Ptpn6	4	No	P29351	D90E	Benign	0.015	Del.	−5.96	Medium	[§]	Neutral	12.48
Lig4	1	No	Q8BTF7	N158K	Benign	0.013	Toler.	−0.667		[§]		15.95
Prkdc	3	No	P97313	V3589A	Benign	0.007	Toler.	−1.75	Low	−1.69134	Neutral	[‖]
Dock8	4	No	Q8C147	T1748A	Benign	0.001	Toler.	2.51	Neutral	−1.77255	Neutral	14.89
Cd22	1	No	Q3UP36	M157V	Benign	0	Toler.	2.65	Neutral	−0.93903	Neutral	[¶]
Itpkb	2	No	B2RXC2	L228P	Benign	0	Toler.	−1.04	Neutral	[§]	Neutral	[¶]

AA, amino acid; cat., category; Del., deleterious; Mut Assessor cat., MutationAssessor category; Mut phen obs?, mutant phenotype observed?; Num hom, number of homozygotes; Poss. Dam., possibly damaging; Prob. Dam., probably damaging; Toler., tolerated.

*Larger GERP scores denote variants more likely to be deleterious.

[†]Smaller subPSEC scores denote increasingly deleterious variants.

[‡]Phred-like scores calculated using coordinates of mouse mutation lifted over to the orthologous human protein.

[§]Missing values were not contained in the alignment created by the HMM.

[¶]Mouse and human orthologs have a different amino acid at the variant site.

[‖]Absent in GRCh37.

In Vivo Phenotypic Consequences of Nonsense and Missense Mutants. For each mutation, heterozygous G1 animals were bred with unrelated mice and heterozygous mutant offspring identified by allele-specific genotyping. These mutant offspring were intercrossed to yield third-generation (G3) offspring, of which 25% were expected to be homozygous for the mutation. Peripheral blood, or the spleen when necessary for particular mutants (e.g., *Dock8*), was collected from homozygous mice, and leukocyte subsets were analyzed by flow cytometry. A panel of antibodies was used for flow cytometry capable of detecting abnormalities in lymphocyte subsets that characterize mice with well-defined, homozygous loss-of-function mutations in each of the 23 genes (Table S1). Particular attention was also paid to the detection of subtle hypo- and hypermorphic alterations within the lymphocyte subsets as detectable by flow cytometry of blood samples, such as the compensating shift toward the $CD44^{hi}$ subset of T cells that occurs when thymic T-cell output or peripheral T-cell survival is subtly decreased. We nevertheless recognize that these in vivo phenotyping tests may fail to detect some hypomorphic alleles due to compensating processes in the immune system and lack of exposure to pathogenic microbes in the environment where these mice were raised.

Three of the 33 de novo mutations taken to homozygosity introduced premature stop codons, two in *Dock8* and one in *Dclre1c* (also called *Artemis*), and all three resulted in the expected immune system in vivo blood cell phenotype (Table 1). Introduction of a premature stop codon is rarely ambiguous because these eliminate a portion of the protein and often cause the transcript to be degraded by nonsense-mediated decay unless the premature stop codon occurs within 55 nucleotides 5′ to the terminal intron (25). All three of the nonsense mutations studied here bear this out.

By contrast with the nonsense mutations, only 4 of 30 missense variants (13%, in *Btk*, *Ets1*, *Il7r*, and *Rag1*) screened for an expected gene-specific loss-of-function blood cell phenotype had a detectable change in the expected lymphocyte subpopulations relative to WT mice analyzed in parallel (Table 1). This result is consistent with previous evidence comparing the frequency of overt null mutations to missense mutations in sets of unphenotyped incidental ENU mutations to sets of ENU mutations

known to cause a mouse phenotype, where it was estimated that only 21% of missense mutations cause a measurable individual phenotype (26). Of the four missense mutations here that produced a phenotype, three (*Btk, Ets1,* and *Il7r*) were inferred by PolyPhen2 to be probably damaging and one (*Rag1*) as possibly damaging. One missense variant (*Il7r*), which produced the expected mutant phenotype in mice, was predicted to be a benign change with human-specific PolyPhen2, whereas it was predicted probably damaging by the mouse-specific calculation (Table S2). Conversely, none of the 12 missense mutations predicted to be benign had a measurable blood cell immune phenotype by flow cytometry. Hence, from this set of tested mutations in essential immune genes, none of the inference tools appear to generate a high rate of false-negative calls. In contrast, 10 of the 30 de novo missense mutations were called as probably damaging by Polyphen2, yet only 3 of these (30%) resulted in the expected phenotype. A further nine missense mutations were inferred as possibly damaging, yet only one of these (11%) produced the expected mutant phenotype. Fifteen mutations received a Polyphen2 score of 0.85 or greater, yet only 4 of these (27%) were sufficient to cause a discernable immune cell abnormality in individual homozygous mice. PolyPhen2 probably damaging predictions appeared the most reliable indicator of a mutant phenotype, although SIFT predictions (20%) and Condel aggregated predictions (23%) were of similar utility. Table 1 and Table S2 include Phred-like scores calculated with the CADD method (13) for each orthologous human mutation corresponding to each mouse mutation. The CADD scores correctly prioritized only two of the four mutations with a measurable in vivo phenotype (*Btk* and *Rag1*), assigning these the highest two scores, whereas the other two received relatively low scores. By contrast, Polyphen2 correctly predicted all four as probably damaging. CADD assigned a score of greater than 20 to 13 missense mutations, yet only 2 of these (15%) had a measurable phenotype. Thus, use of CADD did not improve the prediction of in vivo immune cell phenotypes.

Functional Inferences for All Potential Mutations in 23 Immune Genes. The disparity between the predicted and observed effects of the missense mutations above led us to ask what the range of functional inferences was like for all possible missense mutations in the same set of essential immune genes. We determined the exhaustive set of all 89,887 possible single missense mutations in the 23 essential immune system genes studied above and calculated their PolyPhen2 scores (labeled All possible in Fig. 1). These potential mutations produce a characteristic "hourglass" shape previously observed (23) where the range of Polyphen2 scores is concentrated toward being either deleterious or benign and less likely to be of intermediate effect. Half of all potential missense variants in these genes receive a PolyPhen2 score greater than 0.85.

As opposed to all possible mutations, we also analyzed all missense mutations generated by ENU in the same set of 23 essential immune genes but independent of any phenotypic testing. These ENU-induced mutations were identified by exome sequence analysis of 2,081 G1 offspring of ENU-treated C57BL/6 mice and this allowed us to consider the impact of the characteristic T/A → A/T and A/T → C/G substitutions produced by ENU. Of a total of 136,970 ENU-induced coding variants across the mouse genome, 388 nonsynonymous variants were identified in the 23 essential immune genes (ENU observed in Fig. 1). Only 33 of these were bred to homozygosity and tested for in vivo immune phenotypes as described above, and the remaining 355 represent de novo mutations of unknown phenotype. Comparison of kernel density plots in Fig. 1 between these variant sets indicates that the ENU Observed set tended toward higher PolyPhen2 scores than the All Possible set. Restricting the All Possible set to just T/A → A/T and A/T → C/G substitutions did not replicate this effect; hence, this appears not to be due to the

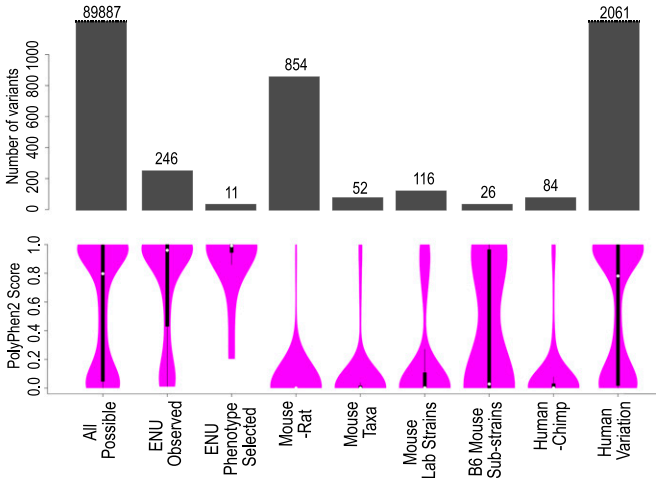


Fig. 1. Spectrum of functional consequences predicted by PolyPhen2 for different sources of missense variants in 23 essential immune system genes. PolyPhen2 scores were calculated for the following sets of missense variants: All possible, the complete set of 89,887 possible amino acid substitutions caused by single missense changes in the 23 mouse immune system genes listed in Table 1; ENU observed, 388 mostly unphenotyped, de novo mutations found in the 23 genes by exome sequencing of 2,081 G1 mice; ENU phenotype selected, mutations in the 23 genes discovered by flow cytometric screening of thousands of G3 offspring of ENU-exposed mice; Mouse-rat, missense variants between the C57BL/6J mouse and Brown Norway rat genome sequences; Mouse taxa, missense variants between the four wild-derived, inbred strains representing *Mus spretus, Mus musculus musculus, Mus musculus domesticus,* and *Mus musculus castaneus*; Mouse Lab strains, missense variants between the genomes of 14 inbred laboratory mouse strains; B6 mouse substrains, missense variants between the genomes of inbred strains C57BL/6J and C57BL/6N; Human-Chimp, missense variants in the orthologous 23 immune genes between the reference human and chimpanzee genome; Human variation, missense variants in the same immune genes detected by population-scale human exome sequencing (36). The barplots in gray indicate the number of variants present in each set. Purple violin plots are kernel density plots representing the distribution of PolyPhen2 scores for each variant set. The white dot indicates the median PolyPhen2 value for each set. The black bars are box-and-whisker plots: thick black bar extends from the first to the third quartile; thin black lines extend to the lowest and highest data points within 1.5 times the interquartile range.

characteristic pattern of ENU-induced mutations. The disparity in sample sizes between these variant sets is large (89,887 All Possible vs. 246 ENU Observed). Iterative random sampling (with replacement) of the All Possible variant set to produce a subset of the same size as the ENU Observed set indicated that the median PolyPhen2 value of the ENU Observed set is not significantly higher than the All Possible set at $P = 0.02$.

For comparison, we also computed Polyphen2 scores for an independent set of ENU-induced mutations in the same 23 genes, but representing alleles discovered by phenotype-driven blood cell screening followed by meiotic mapping and DNA sequencing (ENU phenotype selected in Fig. 1 and Table S4). These alleles were identified using the same flow cytometric panel by screening peripheral blood samples from thousands of G3 progeny from hundreds of G1 mice, where the G1 parents had not been exome sequenced to detect their ENU-induced mutations. When G3 animals were identified with discernible abnormalities in their lymphocyte subsets, and these were proven heritable by identifying the same immune defect in first- or second-degree relatives from later generations of the same pedigree, exon and exome sequencing followed by allele-specific genotyping was used to identify the causative mutation (23, 26). The phenotypically selected set comprised 19 mutations in 12 of the 23 genes studied here: *Bcl2, Btk, Cd22, Il7r, Jak3, Lig4, Prkdc, Ptpn6, Ptprc, Rasgrp1, Satb1,* and *Tnfaip3* (Table S4). Eleven (58%) were missense mutations,

4 (21%) were stop-gain or nonsense mutations (compared with 9% of unphenotyped mutations in the same genes), and 3 (16%) were mutations that disrupted exon splice sites. Compared with the distribution of all possible missense mutations, the PolyPhen2 scores for phenotypically selected missense mutations were strongly biased toward higher values: 82% received scores greater than 0.85.

Functional Inference of Rodent Interspecies and Interstrain Variants. Given the distribution of scores for all possible missense mutations, we next asked what range of PolyPhen2 scores would be calculated for inherited nonsynonymous variants between mouse and rat in the same 23 essential immune genes. Alignments of these mouse genes with orthologous rat genes (27) were obtained from Ensembl Compara (28), and PolyPhen2 scores were calculated for all missense variants between the two rodent species, which have been separate for 12–24 My (29, 30). Individual neutrally evolving nucleotide positions within the mouse and rat genomes have been mutated on average between 0.15 and 0.2 times since species divergence (27), so it would be expected that between 9,427 and 12,569 missense variants are likely to have arisen in the 62.8 Kb of coding sequence within the 23 essential immune genes during this time. Because the rat sequence comes from highly inbred individuals of the Brown Norway strain (27), the set of mouse-rat missense immune gene variants is expected to have been subject to many intensified generations of purifying selection to remove deleterious variants. Fig. 1 shows that functional inference scores calculated with PolyPhen2 for mouse-rat variants were almost exclusively benign. In the small number of cases where a deleterious prediction was made, investigation of the orthologous mouse and rat sequences demonstrated that the deleterious variants originated from areas of low homology in the pairwise protein sequence alignment. In the investigated subset of these low-homology protein regions, it was apparent that the UniProt reference sequences chosen for the Compara mouse/rat alignment were not the identical splice form, and the deleterious substitutions lay within short regions due to alignment of nonorthologous exons. Even without excluding these seemingly nonhomologous nucleotides and assuming that all variants are nonrecent and have been subject to extensive purifying selection, the deleterious overcall rate of PolyPhen2 is still just 4.68% (deleterious variants/total variants = 40/854 = 0.0468). Similarly, among other functional inference tools, the rates of deleterious overcalling were determined: MutationAssessor, 3.30% (25/758 = 0.0330); SIFT, 2.05% (13/634 = 0.0205; excluding low confidence predictions); PANTHER, 4.44% (15/338 = 0.0444; $P_{del} > 0.5$). Overall, these tools appear remarkably accurate by this measure, and these values are likely an overestimate as some of the variants will be of recent origin or spurious due to local misalignment.

As the likely evolutionary age of a mutation decreases, the time for purifying selection to occur on deleterious variants is diminished and the number of deleterious variants as a proportion of the total variants is expected to be greater. Much genomic data exist for mouse strains, especially the recently diverged laboratory strains of *Mus musculus* (31, 32). Missense variants between mouse strains in the 23 immune genes above were collected from the Sanger Institute mouse genomes resource (31). The divergences between the four sequenced wild-derived mouse strains, representing four divergent mouse taxa (*Mus spretus, Mus musculus musculus, Mus musculus domesticus,* and *Mus musculus castaneus*), are much less than the mouse-rat divergence and are estimated to be not greater than 1.6 My (for *musculus*/*spretus*) (33). Each of the four sequenced, wild-derived mouse strains has nevertheless undergone many generations of laboratory-based inbreeding before genome sequencing, potentially exerting strong purifying selection against deleterious missense mutations that might have arisen since strain/subspecies divergence. Fig. 1 shows the range of PolyPhen2 scores for 52 missense variants in the 23 essential immune genes between the mouse subspecies (Mouse Taxa): only 5.7% (3/52) of those observed between mouse taxa received a PolyPhen2 score greater than 0.85 compared with 50% of all possible missense variants in the same immune genes. This result implies that most amino acid substitutions in the 23 immune genes that receive a PolyPhen2 score of greater than 0.85 are indeed sufficiently deleterious to be removed over many generations by purifying selection, either during the divergence of wild-mouse strains or during fixation to homozygosity by laboratory inbreeding.

We extended this analysis to missense variations of very recent origin that exist between inbred strains of laboratory mice. Between the 14 *Mus musculus* laboratory strains for which a full genome sequence has been obtained (31) 116 missense variants were identified within the set of 23 essential immune genes. The divergences between laboratory *Mus* strains are complicated, but a large number of the variants identified between strains will be of much more recent origin than those that exist between *Mus* taxa—of the order of a hundred years. The range of PolyPhen2 scores for these variants (Fig. 1, Mouse Lab Strains) includes 12.9% (15/116) with a PolyPhen2 score of 0.85 or greater, representing an apparently increased fraction of possibly damaging or probably damaging variants than the interspecies *Mus* variants, but still much lower than the random set of all possible variants. The range also includes a great many benign variants that may predate the strain divergences or are evidence for the effect of purifying selection even over a few hundred generations, especially during the inbreeding conducted over the last century to produce these strains. The inbred C57BL/6J and C57BL/6N substrains have only been genetically separate for ~220 generations, since 1951 (32), and their genomes have only 32 missense variants in the 23 essential immune genes studied. Although the number of variants is small and the median score is benign, these variants will be mostly of very recent origin, and a higher proportion is computationally inferred to be deleterious (Fig. 1, B6 Mouse Substrains).

Functional Inference of Human Missense Variants in 23 Immune Genes. Because missense mutations in many of the immune genes studied above are already known to cause devastating human immune deficiency or autoimmune disorders, we performed a parallel set of analyses of human missense variants in the same genes. Within these 23 genes, PolyPhen2 scores were calculated for the 84 missense variants identified between the human reference genome and the chimpanzee genome sequence (34) [*Pan troglodytes*; human-chimp divergence 5–7 Mya (35)]. The range of scores was similar to those for missense variants between mouse taxa: the median score being benign (0.001), and only 5.6% (5/84) receiving a score greater than 0.85 (Fig. 1, Human-Chimp). However, a very different distribution was observed when we calculated PolyPhen2 scores for the set of all missense variants within these same genes detected in 6,503 human genomes by exome sequencing (36) (data release ESP6500SI-V2). The distribution of scores for all observed human variants was similar to the distribution observed for all potential missense variants in these immune genes, with approximately half (987/2,061 = 47.9%) receiving a score of 0.85 or greater (Fig. 1, Human Variation). Most of these human population variants are likely to be of relatively recent origin, because those with a minor allele frequency of 5% or greater account for 1.7% (35/2,026) of the total variants. Of these more prevalent and presumably older missense mutations, only 20% (7/35) have a PolyPhen2 score >0.85.

Systematic Comparison of Prediction and in Vitro Phenotype for Mutations in TP53. As noted above, a limitation of the in vivo immune cell phenotyping tests is the potential for clinically significant partial loss-of-function mutations to be masked by compensating processes, such as lymphocyte homeostatic expansion to counter diminished lymphocyte differentiation, and by shielding of the mice in the laboratory from normal exposure to a wide range of

pathogens and dietary fluctuations. We therefore sought an independent approach to compare computational inferences with experimentally measured effect in a random set of unselected missense mutations, using experimental testing in vitro by quantitative assays with the least possibility for compensation or environmental masking. A comprehensive mutation dataset meeting these criteria exists for the human tumor suppressor gene and protein, TP53 (also called p53) (37). Loss-of-function mutations in *TP53* are the most frequent somatic mutation in human cancer and cause highly penetrant immune system cancers when experimentally bred to homozygosity in the germ line of mice. WT TP53 suppresses tumor development by blocking cell division and inducing apoptosis, primarily by binding to defined DNA sequence motifs in numerous target genes and enhancing their transcription into mRNA (38). The transcription-enhancing activity of WT TP53, and each possible single amino acid substitution arising from single nucleotide mutations, has been systematically quantified by expressing each mutant in yeast in vitro and measuring transcription-enhancing activity against different TP53-binding enhancer sequences from the human TP53 target genes *p21WAF1*, *MDM2, BAX, 14-3-3σ*, *p53AIP1, GADD45,* NOXA, and *p53R2.* Each *TP53* target sequence has been inserted into an enhancerless reporter gene encoding a fluorescent protein (37). When WT or mutant TP53 was expressed, transcription and mean accumulation of the fluorescent protein in a population of yeast cells was measured with a spectrophotometer (37).

We calculated the PolyPhen2 score for all 2,314 potential TP53 missense mutations arising from single nucleotide substitutions and plotted the score for each mutation against its measured $p21^{WAF1}$ transcriptional enhancer activity, the latter normalized to the activity of WT TP53 (Fig. 2*A*). The transactivation (TA) assay results are similar among the differing TP53 target binding sequences (Table S5), as are the plots of PolyPhen2 score against TA activity (Fig. S1). Fig. 2*A* shows that the set of all possible mutations in the resulting plots cluster into three clear density regions. In a first cluster, 640 *TP53* mutations (31.6%) had TA activity diminished to less than 50% of WT and received PolyPhen2 scores of 0.8 or greater. The mutations in this first cluster appear to represent true-positive (TP) predictions. A second cluster includes 831 mutations (41.0%) and had good transcriptional activity, greater than 50% of WT, and received a PolyPhen2 score of 0.2 or less. This cluster appears to represent true-negative (TN) predictions. However, the third clear cluster of 462 mutations (22.8%) also had WT or near WT TA activity yet received a PolyPhen2 score of 0.8 or higher. This cluster represents apparent false-positive (FP) predictions of the effect of missense mutations, and its range of PolyPhen2 scores was comparable to the TP predictions (Fig. 2*A*). Of the 1,102 mutations predicted to be deleterious with score of 0.8 or greater, 42% had good TA activity measured in yeast with a reporter carrying the TP53-binding sequence from $p21^{WAF1}$. When these 1,102 predicted deleterious mutations were tested for activity against TP53-binding sequences from other target genes, the fraction that were FP predictions ranged from 34% for *MDM2* sequences to 61% for *P53R2* sequences (Table S5 and Fig. S1). By contrast, false-negative (FN) predictions, where mutants have less than 50% of WT TA activity yet are predicted not to be damaging (PolyPhen2 ≤ 0.2), accounted for only 93 of the 2,026 mutations (4.6%).

We performed parallel analyses of all possible 2,314 *TP53* missense mutations using MutationAssessor and CADD (Fig. 2 *B* and *C* and Tables S6 and S7), yielding a similar clustering of all possible mutations into three main groups of TN, FP, and TN predictions. However, unlike PolyPhen2 that produces a simple bimodal distribution of scores, CADD and MutationAssessor gave a broader spread of scores, and a trimodal distribution in the case of CADD. Both of these measures appeared to spread out scores at the upper end of the deleterious range so that the distribution of scores for FPs tended to be lower than the range of scores for TPs. This was especially pronounced for CADD scores: 45% of mutations with a score of 20 or greater were FP predictions

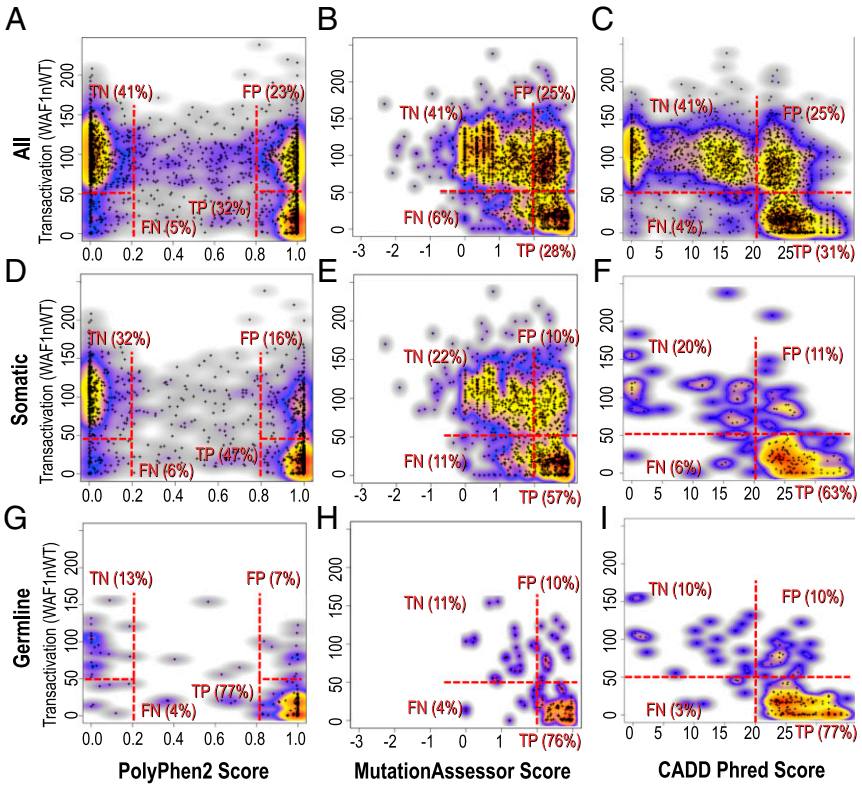


Fig. 2. Comparison of computationally predicted damage with experimentally measured activity of TP53. (*A–C*) Analysis of all 2,314 amino acid substitutions in human TP53 that can arise from a single nucleotide substitution. The transcription-enhancing activity of each mutant, acting on the $p21^{WAF1}$ target sequence in yeast (37), is shown on the *y* axis normalized to the activity of WT TP53 and multiplied by 100. Predicted damage for each possible mutation is plotted on the *x* axis, calculated by (*A*) PolyPhen2; (*B*) MutationAssessor; and (*C*) CADD. Regions marked with dashed lines denote true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). (*D–F*) The same analyses performed on 1,191 somatically acquired *TP53* missense mutations from human cancers in the TP53 database. (*G–I*) The same analyses of 172 germ-line *TP53* missense mutations found by clinical sequencing, primarily performed when the clinical phenotype was consistent with a germ-line *TP53* mutation.

(retaining at least 50% of WT activity), but only 19.5% of mutations with a score of 30 or greater were FP by the same measure.

Evidence for Nearly Neutral TP53 Mutations. As there are a large number of unique *TP53* mutations in both the FP (CADD score > 20; TA activity > 50%) and the group of TN mutations with CADD score < 5 but TA activity > 50%, it was possible to compare the distribution of transcriptional activity values between these groups in aggregate (Fig. 3). This comparison showed that the mean activity of the FP set of mutations was 86% of the mean activity in the TN set, with the difference being statistically significant. A smaller difference, although still statistically significant, was observed in the groups of TN and FP mutants resolved by Polyphen2 and MutationAssessor scores (Fig. S2). Hence a large fraction of mutations predicted to be deleterious by the various algorithms may not be FP functional inferences with no actual effect, but instead appear to be nearly neutral mutations of small effect (39). Fig. 2*C* also shows that many intermediate points lie between the CADD TN and FP categories (Phred-like score > 5 and < 20). When these intermediate values are included in the TN category, this category remains significantly distinct from the FP category (Fig. S2).

Some apparent FP mutations might have lost some other function critical for TP53 in human cells not measured in the yeast assay. To explore this possibility, we prepared comparable plots for a set of 1,191 distinct amino acid substitutions identified in human cancers and obtained from the curated *TP53* mutation database (IARC TP53 Database, R17) (40). Compared with the 42% FP rate found for all mutations with PolyPhen2 score > 0.8, in the cancer-selected somatic mutation set, this was reduced to 25% (Fig. 2*C* and Table S5). The decrease in FP predictions is likely to be due to the cancer mutation set being enriched for mutations that diminish TP53 transcriptional activity to less than 50% of WT (667/1,192, 53%) compared with the full set of mutations (733/2,026, 36%), consistent with previous analysis using a cutoff of 20% of WT activity (40); 82.8% of the TP subset of *TP53* mutations was found in two or more independent cases of cancer (reported in the *TP53* mutation database), whereas only 27.7% of the FP nearly neutral subset was found recurrently (Fig. S3*A*). The small FN subset also had a higher fraction found recurrently in cancer (60.0%), emphasizing the importance of experimental testing to identify this small but important subset of mutations. TN mutations (26.1%) were no more likely to be found recurrently in cancer than FP mutations, suggesting that these subsets represent passenger mutations that confer no neoplastic advantage.

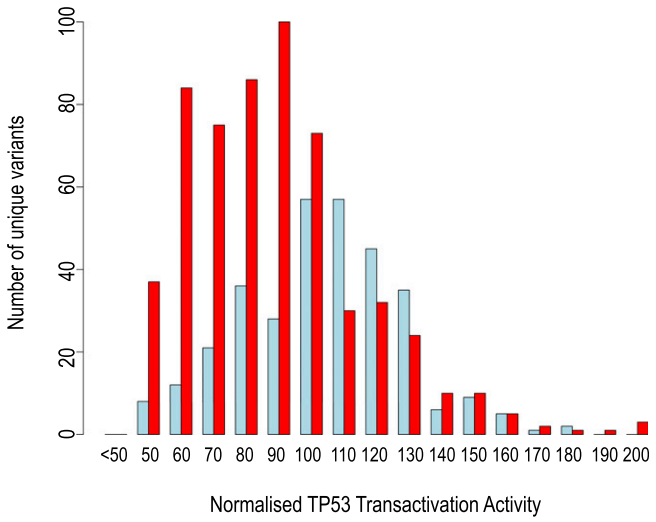


Fig. 3. Distribution of measured TP53 transactivation activity from true-negative and false-positive TP53 mutations. From all possible *TP53* missense mutations, the distribution of measured transcriptional activity values is shown for those inferred as true negative (TN, CADD score < 5 and transactivation activity >50; blue, $n = 322$) or as false positive (FP, CADD score > 20 and transactivation activity >50; red, $n = 573$). The mean normalized activity was 108.8 for the TN set and 93.4 for the FP set. The distribution of values for each set is unlikely to be the same (Kolmogorov–Smirnov $D = 0.3427$ $P < 2.2e^{-16}$; Wilcox $W = 57{,}261.5$, $P < 2.2e^{-16}$).

The same analysis was performed on a phenotypically selected set of 172 *TP53* missense mutations found by sequencing germline DNA in patients with types of tumors and age of onset suggesting a loss-of-function *TP53* germ-line mutation (IARC TP53 Database, R17) (40). Only 8% of those that were predicted to be deleterious by PolyPhen2 retained >50% activity in the yeast assay, representing a 5.25-fold lower fraction of FP predictions than for all possible mutations. The particularly low rate of FP predictions in the germ-line set is likely to reflect the high proportion of these mutations that lack transcriptional activity, especially in families with highly penetrant cancer meeting the criteria of familial Li-Fraumeni syndrome, where 92% of mutations have less than 20% of WT transcriptional activity (40). In other words, when the clinical phenotype is sufficiently clear and specific for profound loss-of-function mutations in a particular gene, simply finding a rare missense mutation in that gene is already highly predictive, so that computational predictions pose a lower risk of FPs.

The higher concordance between prediction and actual effect among cancer-selected somatic mutations and in phenotypically selected germ-line mutations in *TP53* implies that the yeast transcriptional assay measures the relevant *TP53* function for tumor suppression. However, it does not rule out the possibility that some FP predictions might disrupt other evolutionarily conserved *TP53* functions. Interestingly, FP mutations were concentrated outside the core DNA-binding domain residues 102–292 (41) (Fig. S3*B*), in contrast to the 95% of oncogenic mutations recorded in TP53 that lie within the DNA-binding domain (42). An example of a functionally important mutation that lies outside the DNA-binding domain and does not have a direct effect on transactivation is the E388A substitution that disrupts conjugation of SUMO1 to K396 in TP53 (43). SUMO1 conjugation at this site has a regulatory effect on TP53 activity (44), but the transactivation assays in yeast (37) indicate this has little effect on the core transactivation mechanism. The PolyPhen2 score for this change (0.893) predicts it is probably damaging, but as the change has little effect of transactivation ($p21^{WAF1}$ assay value = 105.5), this data point is classified as a FP inference in Fig. 2*A*.

Discussion

The discordance between the predicted and actual effect of missense mutations revealed here creates the potential for many FP conclusions in clinical whole genome sequencing. Due to the large number of candidate missense variants revealed by whole exome or genome sequencing of any individual, computational prediction of the functional importance of each mutation is widely used to prioritize candidates for further analysis. However, current functional inference tools rely heavily on sequence conservation information. As we have found, these functional predictions do not effectively differentiate between mutations that are immediately clinically relevant, because they ablate or markedly reduce function of an essential protein, and those that are nearly neutral because they only decrease the function of the corresponding protein by 10% (39). FN predictions were also a small but significant subset of mutations in the systematic analysis of *TP53*. Hence, for interpretation of a clinical genome sequence at present, it is essential to measure experimentally the consequence of any missense mutation thought to be causal.

Two lines of experimental evidence from our work indicate that this lack of differentiation has the potential for overcalling causative mutations. First, 73% of predicted deleterious missense mutations arising from de novo single nucleotide substitutions in 23 essential immune system genes did not sufficiently disrupt protein function to produce an observable defect in the expected subset of blood lymphocytes in individual homozygous mice. Absence of a phenotype in vivo may occur by compensation: for example, a lymphocyte production defect can be compensated by lymphocyte homeostatic expansion. A discernable in vivo phenotype may also require an environmental cofactor such as malnutrition or exposure to a particular infectious agent. For that reason, we turned to the *TP53* gene for a complementary analysis of predicted versus actual effects of all possible single nucleotide, missense mutations, taking advantage of a dataset where the experimentally determined phenotype of each mutation has been measured in a simple in vitro assay. Even under these simplified conditions, 42% of the missense TP53 mutations that were predicted to be deleterious caused little or no measurable decrease in the protein's transcription-enhancing activity.

The plots in Fig. 2, especially those of CADD Phred-like scores, show that mutations that are computationally predicted to be deleterious have a bimodal distribution. Half drastically decreases protein activity to a mean of 17% of WT, and it is this group that is of interest in clinical cases with onset in childhood or adolescence as suspected monogenic or oligogenic disease. The other half, despite most being scored just as likely to be deleterious, actually only slightly decrease protein activity, to a mean of 86% of WT. Discerning the subtle shift in activity caused by these nearly neutral mutations required us to consider the FP mutants as a group, comprising hundreds of independent measurements. Because they have effects on protein activity that are barely measurable in the laboratory, these nearly neutral mutations are unlikely to cause monogenic diseases but may be relevant to diseases that have a more complex basis involving interaction of many weakly damaged genes and environmental factors.

Presumed nearly neutral FP mutations were frequent among unselected sets of missense mutations (de novo ENU mutations, all possible *TP53* mutations) but were rare in sets of mutations that had been subject to phenotypic selection. When applied to missense variants in the 23 essential immune system genes between inbred mouse and rat species, or between inbred mouse taxa, Polyphen2 only predicted 4.7% and 5.7%, respectively, to be deleterious. Because 50% of random mutations in the same 23 genes are predicted to be deleterious, the much lower frequencies among mutations that have become fixed in these species/strains implies that most of the FP predictions in the random mutation sets are indeed deleterious over evolutionary scales. This feature was not restricted to mouse mutations, because a similarly low rate of 5.6% was predicted to be deleterious in the set of variants in the 23 essential immune genes between the consensus human and chimpanzee genome sequences compared with 48% of the predominantly rare variants in the same 23 genes present in a population of 6,503 people for whom exome sequences had been obtained.

At first sight, it appears paradoxical that nearly neutral mutations in essential immune genes will be efficiently removed from the species' gene pool despite having no easily discernible impact on the immune system of individual homozygotes. One possible explanation is that these subtle variants are individually sufficient to cause a serious problem in conjunction with particular environmental stressors such as malnutrition and pathogenic microbes. Another explanation comes from decades of research into the evolution of protein molecules and the nearly neutral theory of molecular evolution (39, 45). Mathematical models predict that slightly disadvantageous nearly neutral alleles will be lost over many generations through random drift in large populations, even when these alleles reduce fecundity by as little as 1% (39). For essential immune genes, that small difference in fecundity could result from one extra bout of influenza per lifetime, which would not be perceived as clinically significant. In addition to genetic drift in large populations, truncating selection has the potential to remove nearly neutral mutations from the gene pool more rapidly (46). This model considers that, as the number of slightly damaging alleles increases in the population, by Poisson distribution a subset of individuals will inherit a larger burden of nearly neutral alleles affecting a critical function (46), for example, in *RAG1, RAG2, LIG4, DCLRE1C*, and *PRKDC* and other genes required for VDJ recombination. Although none of these mutations would be of clinical consequence individually, the chance inheritance of three or more subtle defects in the same pathway may be sufficient to cause recurrent infections or Omenn's syndrome-like autoimmune manifestations. Previously, we demonstrated experimentally how three heterozygous loss-of-function mutations in sequential steps in a biochemical pathway—affecting Lyn kinase, its substrate CD22, and the CD22-binding tyrosine phosphatase SHP1—only precipitate B-cell deficiency in individuals that inherit all three but not any pair or single mutation (47). Some of the 23 immune genes studied here have essential roles outside the immune system, and those other functions may be more sensitive to disruption by apparently FP mutations. For example *LIG4, DCLRE1C*, and *PRKDC* are critical for DNA damage repair in all cells, and it is conceivable that small decreases in their activity could subject these variants to purifying selection over evolutionary timescales because of more rapid aging of stem cells (48).

The same question about subtle loss-of-function arises for the apparent FP predictions among TP53 mutations. These apparent FP predictions were frequent in the set of random *TP53* mutations (42%) but much lower in sets of *TP53* mutations that had been positively selected for *TP53* loss-of-function either as a result of being found by selective resequencing in cancer cells (25%) or in the germ line of young cancer patients with suspected Li-Fraumeni syndrome (8%). Most of the TP somatic mutations were found recurrently in different cancers, consistent with these being driver mutations that confer a growth advantage for neoplastic cells. By contrast, most of the apparent FP somatic mutations that retained greater than 50% of transactivation activity were singleton observations, as would be expected if these were random passenger mutations. These singleton FPs might nevertheless represent driver mutations that provide a more subtle growth advantage. A 10% decrease in TP53 transcriptional activity is difficult to distinguish from WT in individual laboratory tests but may nevertheless confer a selective advantage when this effect is compounded over hundreds of cell divisions or when combined with other partial loss-of-function mutations in the TP53 tumor suppression pathway that may have arisen earlier in the evolution of the neoplastic cell clone.

A similar explanation may apply to the 10% of apparent FP predictions for germ-line *TP53* mutations. Indeed one of this set came from a clinical case that did not display the age of onset and pattern of sarcomas typical of Li-Fraumeni syndrome. Instead this individual displayed a syndrome of familial adenomatous polyposis (FAP) that was at the severe end of the spectrum and carried an additional germ-line mutation in the *APC* tumor suppressor gene that is typically inactivated in FAP (40).

In clinical genome sequencing, the risk of FP calling of missense mutations is likely to be highest when the clinical phenotype does not match a distinct Mendelian syndrome but could be explained by defects in hundreds of genes, for example, in sporadic cases of autoimmune disease or in common variable immune deficiency. Here it will be particularly critical to validate computational predictions experimentally by biochemical tests and recreating the mutations in animals (49). By contrast, the FP problem demonstrated here is minimized when the clinical phenotype can be refined sufficiently that only one or two genes could explain it—for example, when a person develops multiple cancers at an early age

including uncommon sarcomas typical of Li-Fraumeni syndrome. Our findings underscore the importance of acquiring two additional types of information to interpret missense variation identified by clinical genome sequencing: (*i*) direct experimental measurement of the consequences of a candidate mutation, using as specific and sensitive an assay as possible; and (*ii*) much more specific human phenotyping, capable of narrowing the set of candidate genes to a handful related to a particular biochemical pathway or syndrome. Experimental analysis of the connection between nearly neutral mutations and in vivo immune or cancer phenotypes poses a major challenge and will likely require very large numbers of replicate or iterative measurements.

Materials and Methods

Generation of Random Mouse Mutants, Sequencing of Mouse Exomes, and Variant Calling. Generation of pedigrees from mice treated with ENU, sequencing of exomes of these mice, and computational identification of ENU-induced point mutants were performed as previously described (23). This research was approved by the animal experimentation and ethics committee of the Australian National University (protocol number A2014/61).

Detailed Description of Mouse-Specific PolyPhen2 Changes. PolyPhen2 (12) scores were calculated from a local mouse-specific installation of PolyPhen version 2.1.0. Local installation required using the mouse specific UniProt (50) and Pfam (51) databases during setup followed by the subsequent mapping of all mouse UniProt protein sequences to mouse assembly (mm9) coordinates. This mapping allows mouse variant genomic coordinates to be converted directly to UniProt amino acid locations giving PolyPhen2 access to mouse-specific protein annotations (often different from the human homolog annotations), thus often resulting in more accurate PolyPhen2 scores. The mapping was accomplished by searching the CCDS (consensus coding sequence; ref. 52) gene sets for exact UniProt protein sequence matches and was successful for 93% of all mouse UniProt entries. The remaining 7% of nonmapping UniProt entries were provided as input to PolyPhen2 as a UniProt protein FASTA-formatted file and relative protein coordinates (which is a less specific input option available with PolyPhen2). To calculate many thousands of PolyPhen2 scores in a timely manner, UniProt-specific PolyPhen2 calculations were cached, hence avoiding duplicate calculations for variants occurring in UniProt entries previously encountered in other samples.

Calculation of Functional Inferences. Mouse PolyPhen2 scores were calculated as described above. SIFT (6) scores were obtained from the Variant Effect Predictor (53) and from the SIFT Web server (www.siftdna.org). MutationAssessor (mutationassessor.org) (7) and PANTHER (www.pantherdb.org) (10) scores were calculated for mouse-specific UniProt entries using public webservers provided by the method authors. CADD scores (13), which are presently only calculated from human data, were determined by mapping the coordinate of each mouse mutant to the orthologous base in the human genome with liftOver [via the University of California, Santa Cruz (UCSC) genome browser] (54), providing that the orthologous amino acid in the human reference was identical to that in the mouse reference . Scores for these humanized mutations were obtained from the CADD Web server (cadd.gs.washington.edu). GERP (11) scores were derived from the UCSC Genome Browser (54). Condel weighted average scores were calculated using PolyPhen2, SIFT, and MutationAssessor scores using the Perl code provided by the Condel authors (15).

Allele-Specific Genotyping. Competitive allele-specific genotyping was performed using the KASP system (Kbioscience). A pair of WT and mutant allele-specific oligonucleotide primers were designed to anneal to sequence flanking the variant site. These primers were conjugated with a tail sequence that contained a FRET cassette labeled with allele-specific dyes. With these primers, sample DNA was amplified with a thermal cycler and dye values read with a FluoStar Optima fluorescent microplate reader (BMG Labtechnologies). Homozygous and heterozygous genotypes were distinguished by whether one or two fluorescent signals, respectively, were detected.

Mouse Phenotyping. Mouse phenotypes were appraised for most mutations by eight-color flow cytometry on peripheral blood. Two hundred microliters of blood was collected by retro-orbital bleeding into tubes containing 20 μL heparin (Sigma: 1,000 U/mL in PBS). Red blood cells were lysed, and samples were stained in 96-well plates alongside WT C57BL/6 mouse controls as previously described (55). Granulocytes were enumerated by forward and side scatter of laser light. The antibodies used to detect B, T, and NK populations were as follows: anti-B220 (RA3-6B2; BD Pharmingen), anti-IgM (R6-60.2; BD Pharmingen), anti-IgD (11-26c.2a; BioLegend), anti-CD3 (17A2; eBioscience), anti-CD4 (RM4-5; BioLegend), anti-CD44 (IM7; BioLegend), anti-KLRG1 (2F1; eBioscience), and anti-NK1.1 (PK136; BD Pharmingen). To detect splenic marginal zone B (MZB) cells in *Dock8* mutant mice, spleens were harvested, and single cell suspensions were stained with the following antibodies to distinguish MZB cells (B220+, IgM high, CD21 high, CD23 negative) from follicular B cells (B220+, IgM intermediate, CD21 intermediate, CD23 positive): anti-B220 (RA3-6B2; BD Pharmingen), anti-IgM (II/41; eBioscience), anti-CD21 (7E9; BioLegend), and anti-CD23 (B3B4; Biolegend). Samples were acquired using a LSR II flow cytometer (BD Bioscience) and analyzed using FlowJo software (FlowJo LLC).

ACKNOWLEDGMENTS. We thank the National Computational Infrastructure (Australia) for continued access to significant computation resources and technical expertise; the staff of the Australian Phenomics Facility for animal husbandry, DNA preparation, exome sequencing, and genotyping; and the staff of the Microscopy and Cytometry Resource Facility for help with flow cytometry. This work has been funded by National Institutes of Health Grant AI100627, Australian National Health and Medical Research Council Grant 585490, Wellcome Trust Grant 082030/B/07/Z, and the National Collaborative Research Infrastructure Strategy (Australia).

1. MacArthur DG, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335(6070):823–828.
2. Andrews TD, Sjollema G, Goodnow CC (2013) Understanding the immunological impact of the human mutation explosion. *Trends Immunol* 34(3):99–106.
3. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z (2013) Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* 14(Suppl 3):S7.
4. Thusberg J, Olatubosun A, Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat* 32(4):358–368.
5. Hicks S, Wheeler DA, Plon SE, Kimmel M (2011) Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum Mutat* 32(6):661–668.
6. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4(7):1073–1081.
7. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res* 39(17):e118.
8. Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 15(7):978–986.
9. Mathe E, et al. (2006) Computational approaches for predicting the biological effect of p53 missense mutations: A comparison of three sequence analysis based methods. *Nucleic Acids Res* 34(5):1317–1325.
10. Thomas PD, Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci USA* 101(43):15398–15403.
11. Cooper GM, et al.; NISC Comparative Sequencing Program (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15(7):901–913.
12. Adzhubei IA, et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7(4):248–249.
13. Kircher M, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46(3):310–315.
14. Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. *Bioinformatics* 24(20):2397–2398.
15. González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 88(4):440–449.
16. Dorfman R, et al. (2010) Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene? *Clin Genet* 77(5):464–473.
17. Masica DL, et al. (2015) Predicting survival in head and neck squamous cell carcinoma from TP53 mutation. *Hum Genet* 134(5):497–507.
18. Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL (2014) Organizing knowledge to enable personalization of medicine in cancer. *Genome Biol* 15(8):438.
19. Manolio TA, et al. (2013) Implementing genomic medicine in the clinic: The future is here. *Genet Med* 15(4):258–267.
20. Rehm HL (2013) Disease-targeted sequencing: A cornerstone in the clinic. *Nat Rev Genet* 14(4):295–300.
21. Cassa CA, Tong MY, Jordan DM (2013) Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum Mutat* 34(9):1216–1220.
22. Stanley CM, Sunyaev SR, Greenblatt MS, Oetting WS (2014) Clinically relevant variants - identifying, collecting, interpreting, and disseminating: the 2013 annual scientific meeting of the Human Genome Variation Society. *Hum Mutat* 35(4):505–510.
23. Andrews TD, et al. (2012) Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: An immediate source for thousands of new mouse models. *Open Biol* 2(5):120061.

24. Kong A, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412):471–475.
25. Khajavi M, Inoue K, Lupski JR (2006) Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur J Hum Genet* 14(10):1074–1081.
26. Bergmann H, et al. (2013) B cell survival, surface BCR and BAFFR expression, CD74 metabolism, and CD8- dendritic cells require the intramembrane endopeptidase SPPL2A. *J Exp Med* 210(1):31–40.
27. Gibbs RA, et al.; Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428(6982):493–521.
28. Hubbard T, et al. (2005) Ensembl 2005. *Nucleic Acids Res* 33(Database issue):D447–D453.
29. Adkins RM, Gelke EL, Rowe D, Honeycutt RL (2001) Molecular phylogeny and divergence time estimates for major rodent groups: Evidence from multiple genes. *Mol Biol Evol* 18(5):777–791.
30. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci USA* 100(3):1056–1061.
31. Keane TM, et al. (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477(7364):289–294.
32. Simon MM, et al. (2013) A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biol* 14(7):R82.
33. Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K (2004) Temporal, spatial, and ecological modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. *Mol Phylogenet Evol* 33(3):626–646.
34. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055): 69–87.
35. Kumar S, Filipski A, Swarna V, Walker A, Hedges SB (2005) Placing confidence limits on the molecular age of the human-chimpanzee divergence. *Proc Natl Acad Sci USA* 102(52):18842–18847.
36. Tennessen JA, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.
37. Kato S, et al. (2003) Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci USA* 100(14):8424–8429.
38. Bieging KT, Mello SS, Attardi LD (2014) Unravelling mechanisms of p53-mediated tumour suppression. *Nat Rev Cancer* 14(5):359–370.
39. Ohta T (1992) The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst* 23:263–286.
40. Petitjean A, et al. (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database. *Hum Mutat* 28(6):622–629.
41. Cho Y, Gorina S, Jeffrey PD, Pavletich NP (1994) Crystal structure of a p53 tumor suppressor-DNA complex: Understanding tumorigenic mutations. *Science* 265(5170): 346–355.
42. Bullock AN, Fersht AR (2001) Rescuing the function of mutant p53. *Nat Rev Cancer* 1(1):68–76.
43. Rodriguez MS, Dargemont C, Hay RT (2001) SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *J Biol Chem* 276(16): 12654–12659.
44. Rodriguez MS, et al. (1999) SUMO-1 modification activates the transcriptional response of p53. *EMBO J* 18(22):6455–6461.
45. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* 8(8):610–618.
46. Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* 1(1):40–47.
47. Cornall RJ, et al. (1998) Polygenic autoimmune traits: Lyn, CD22, and SHP-1 are limiting elements of a biochemical pathway regulating BCR signaling and selection. *Immunity* 8(4):497–508.
48. Nijnik A, et al. (2007) DNA repair is limiting for haematopoietic stem cells during ageing. *Nature* 447(7145):686–690.
49. Casanova J-L, Conley ME, Seligman SJ, Abel L, Notarangelo LD (2014) Guidelines for genetic studies in single patients: Lessons from primary immunodeficiencies. *J Exp Med* 211(11):2137–2149.
50. Bairoch A, et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res* 33(Database issue):D154–D159.
51. Finn RD, et al. (2014) Pfam: The protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230.
52. Pruitt KD, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 19(7):1316–1323.
53. McLaren W, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26(16):2069–2070.
54. Karolchik D, et al. (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 42(Database issue):D764–D770.
55. Yabas M, et al. (2011) ATP11C is critical for the internalization of phosphatidylserine and differentiation of B lymphocytes. *Nat Immunol* 12(5):441–449.
56. Nakayama K, et al. (1993) Disappearance of the lymphoid system in Bcl-2 homozygous mutant chimeric mice. *Science* 261(5128):1584–1588.
57. Veis DJ, Sentman CL, Bach EA, Korsmeyer SJ (1993) Expression of the Bcl-2 protein in murine and human thymocytes and in peripheral T lymphocytes. *J Immunol* 151(5): 2546–2554.
58. Hendriks RW, et al. (1996) Inactivation of Btk by insertion of lacZ reveals defects in B cell development only past the pre-B cell stage. *EMBO J* 15(18):4862–4872.
59. Kerner JD, et al. (1995) Impaired expansion of mouse B cell progenitors lacking Btk. *Immunity* 3(3):301–312.
60. Khan WN, et al. (1995) Defective B cell development and function in Btk-deficient mice. *Immunity* 3(3):283–299.
61. O'Keefe TL, Williams GT, Davies SL, Neuberger MS (1996) Hyperresponsive B cells in CD22-deficient mice. *Science* 274(5288):798–801.
62. Otipoby KL, et al. (1996) CD22 regulates thymus-independent responses and the lifespan of B cells. *Nature* 384(6610):634–637.
63. Sato S, et al. (1996) CD22 is both a positive and negative regulator of B lymphocyte antigen receptor signal transduction: Altered signaling in CD22-deficient mice. *Immunity* 5(6):551–562.
64. Bikoff EK, et al. (1993) Defective major histocompatibility complex class II assembly, transport, peptide acquisition, and CD4+ T cell selection in mice lacking invariant chain expression. *J Exp Med* 177(6):1699–1712.
65. Viville S, et al. (1993) Mice lacking the MHC class II-associated invariant chain. *Cell* 72(4):635–648.
66. Rooney S, et al. (2002) Leaky Scid phenotype associated with defective V(D)J coding end processing in Artemis-deficient mice. *Mol Cell* 10(6):1379–1390.
67. Randall KL, et al. (2009) Dock8 mutations cripple B cell immunological synapses, germinal centers and long-lived antibody production. *Nat Immunol* 10(12):1283–1291.
68. Bories JC, et al. (1995) Increased T-cell apoptosis and terminal B-cell differentiation induced by inactivation of the Ets-1 proto-oncogene. *Nature* 377(6550):635–638.
69. Muthusamy N, Barton K, Leiden JM (1995) Defective activation and survival of T cells lacking the Ets-1 transcription factor. *Nature* 377(6550):639–642.
70. Gaudreau M-C, Heyd F, Bastien R, Wilhelm B, Möröy T (2012) Alternative splicing controlled by heterogeneous nuclear ribonucleoprotein L regulates development, proliferation, and migration of thymic pre-T cells. *J Immunol* 188(11):5377–5388.
71. Willerford DM, et al. (1995) Interleukin-2 receptor α chain regulates the size and content of the peripheral lymphoid compartment. *Immunity* 3(4):521–530.
72. Peschon JJ, et al. (1994) Early lymphocyte expansion is severely impaired in interleukin 7 receptor-deficient mice. *J Exp Med* 180(5):1955–1960.
73. Fang D, et al. (2002) Dysregulation of T lymphocyte function in itchy mice: A role for Itch in TH2 differentiation. *Nat Immunol* 3(3):281–287.
74. Pouillon V, et al. (2003) Inositol 1,3,4,5-tetrakisphosphate is essential for T lymphocyte development. *Nat Immunol* 4(11):1136–1143.
75. Park SY, et al. (1995) Developmental defects of lymphoid cells in Jak3 kinase-deficient mice. *Immunity* 3(6):771–782.
76. Thomis DC, Gurniak CB, Tivol E, Sharpe AH, Berg LJ (1995) Defects in B lymphocyte maturation and T lymphocyte activation in mice lacking Jak3. *Science* 270(5237): 794–797.
77. Gao Y, et al. (1998) A targeted DNA-PKcs-null mutation reveals DNA-PK-independent functions for KU in V(D)J recombination. *Immunity* 9(3):367–376.
78. Kurimasa A, et al. (1999) Requirement for the kinase activity of human DNA-dependent protein kinase catalytic subunit in DNA strand break rejoining. *Mol Cell Biol* 19(5):3877–3884.
79. Taccioli GE, et al. (1998) Targeted disruption of the catalytic subunit of the DNA-PK gene in mice confers severe combined immunodeficiency and radiosensitivity. *Immunity* 9(3):355–366.
80. Shultz LD, et al. (1993) Mutations at the murine motheaten locus are within the hematopoietic cell protein-tyrosine phosphatase (Hcph) gene. *Cell* 73(7):1445–1454.
81. Tsui HW, Siminovitch KA, de Souza L, Tsui FW (1993) Motheaten and viable motheaten mice have mutations in the haematopoietic cell phosphatase gene. *Nat Genet* 4(2):124–129.
82. Byth KF, et al. (1996) CD45-null transgenic mice reveal a positive regulatory role for CD45 in early thymocyte development, in the selection of CD4+CD8+ thymocytes, and B cell maturation. *J Exp Med* 183(4):1707–1718.
83. Mombaerts P, et al. (1992) RAG-1-deficient mice have no mature B and T lymphocytes. *Cell* 68(5):869–877.
84. Shinkai Y, et al. (1992) RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *Cell* 68(5):855–867.
85. Dower NA, et al. (2000) RasGRP is essential for mouse thymocyte differentiation and TCR signaling. *Nat Immunol* 1(4):317–321.
86. Alvarez JD, et al. (2000) The MAR-binding protein SATB1 orchestrates temporal and spatial expression of multiple genes during T-cell development. *Genes Dev* 14(5): 521–535.
87. Townsend MJ, et al. (2004) T-bet regulates the terminal maturation and homeostasis of NK and Valpha14i NKT cells. *Immunity* 20(4):477–494.
88. Szabo SJ, et al. (2002) Distinct effects of T-bet in TH1 lineage commitment and IFN-gamma production in CD4 and CD8 T cells. *Science* 295(5553):338–342.

Supporting Information

Miosge et al. 10.1073/pnas.1511585112

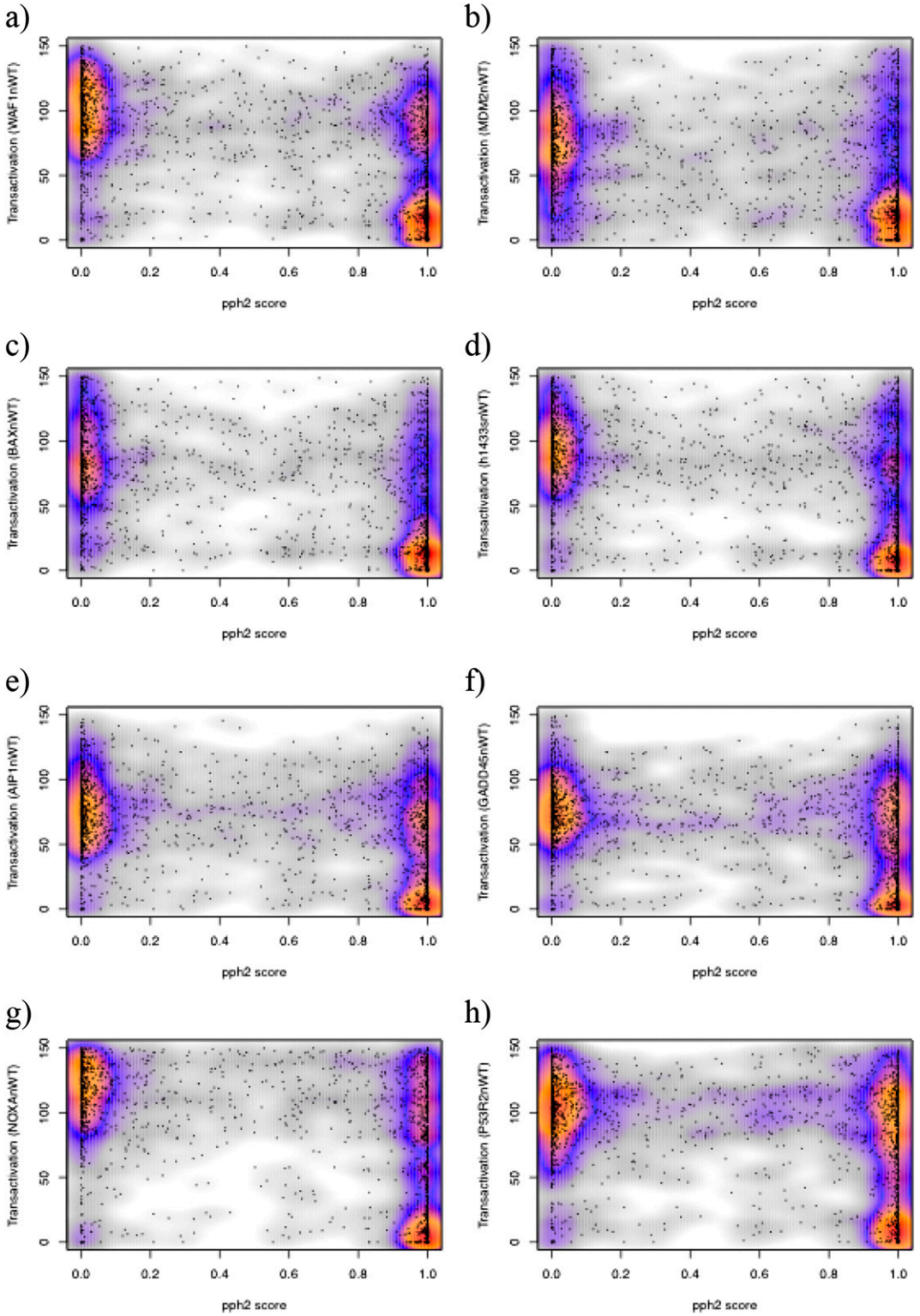


Fig. S1. Analysis of PolyPhen2 predicted damage scores compared with the measured transactivation activity of all possible TP53 missense mutations using binding sequences from eight different TP53 target genes. The target binding sequences used in each transactivation assay are (*A*) $p21^{WAF1}$, (*B*) MDM2, (*C*) BAX, (*D*) *14–3-3σ*, (*E*) *p53AIP1*, (*F*) *GADD45*, (*G*) *Noxa*, and (*H*) *p53R2*.

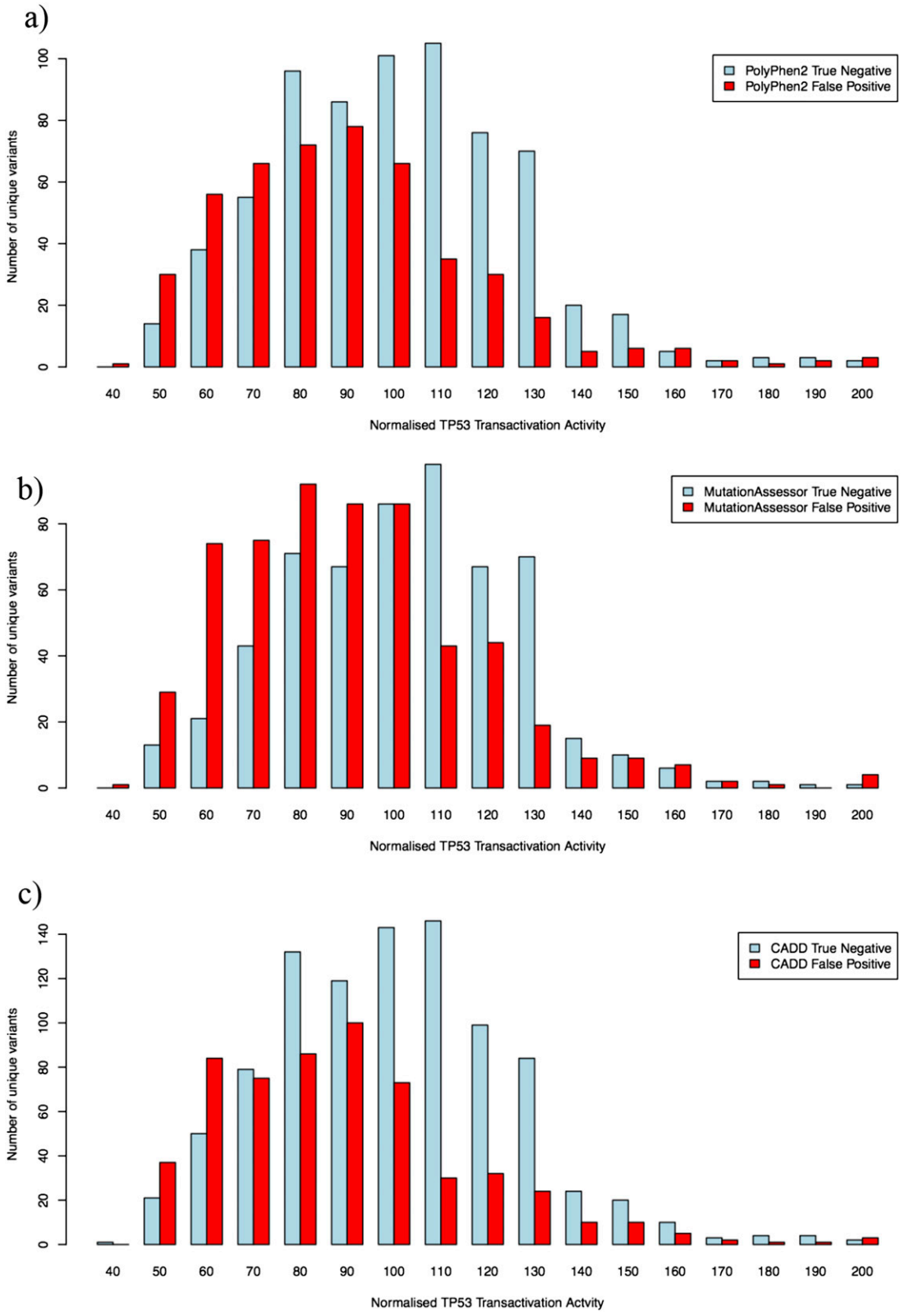


Fig. S2. Distributions of residual TP53 transactivation activities of PolyPhen2-, MutationAssessor-, and CADD-assigned true-negative and false-positive mutations. Measured TP53 transactivation activity for true-negative (TN; blue) and false-positive (FP; red) categories of *TP53* missense mutations identified in Fig. 2 *A*–*C* using inference scores calculated with (*A*) PolyPhen2, (*B*) MutationAssessor, and (*C*) CADD. For PolyPhen2, the mean normalized activity was 106.1 for the TN set and 94.6 for the FP set (Kolmogorov–Smirnov $D = 0.2385$, $P = 2.454e^{-14}$; Wilcox $W = 210{,}343.5$, $P = 3.541e^{-16}$). For MutationAssessor, the mean normalized activity was 104.1 for the TN set and 95.4 for the FP set (Kolmogorov–Smirnov $D = 0.1792$, $P = 2.204e^{-10}$; Wilcox $W = 327{,}022.5$, $P = 3.364e^{-12}$). For CADD, when intermediate values (score >5 and <20) are included with the TN category, the mean normalized activity was 105.3 for the TN set and 93.4 for the FP set (Kolmogorov–Smirnov $D = 0.2401$, $P < 2.2e^{-16}$; Wilcox $W = 190{,}543.5$, $P < 2.2e^{-16}$).

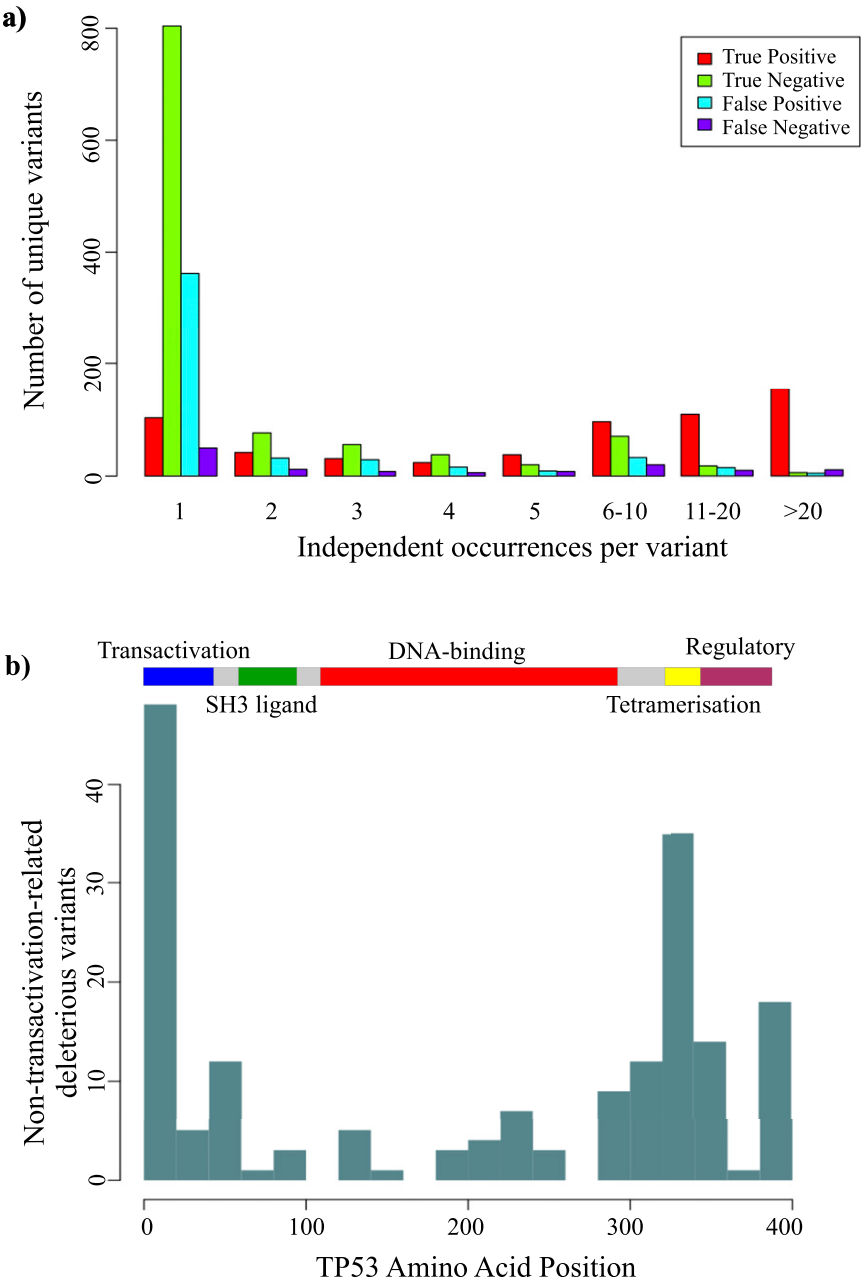


Fig. S3. Incidence and location of apparent false-positive TP53 somatic mutations. (*A*) Unique *TP53* somatic mutations found in human cancers, separated into categories as shown in Fig. 2*D* and ranked by the number of independent occurrences in separate cancer cases. Red, true positive; green, true negative; blue, false positive; purple, false negative. (*B*) The ideogram shows the major functional domains of the TP53 protein. Plotted against this are the number of false-positive mutations in 20 amino acid bins: from the set of all possible TP53 mutations, those that were predicted to be deleterious with Polyphen2 score >0.8 yet actually retained >40% of WT transactivation activity. Unlike most oncogenic TP53 mutations that occur in the DNA-binding domain (red), false-positive mutations appear concentrated in the extremities of the transactivation and tetramerization domains.

Table S1. Expected cellular phenotypes of null mutants

Gene symbol	Expected phenotype	Gene name
Bcl2	Low B and T cells (56, 57)	B cell leukemia/lymphoma 2
Btk	Reduced mature B cells (58–60)	Bruton agammaglobulinemia tyrosine kinase
Cd22	Low IgM expression on mature B cells (61–63)	CD22 antigen
Cd74	Low CD4 T cells (64, 65)	CD74 antigen
Dclre1c	Absence of B and T cells (66)	DNA cross-link repair 1C
Dock8	Absence of MZB cells in spleen in ENU mutant mice (67)	dedicator of cytokinesis 8
Ets1	Reduced T cells (68, 69)	E26 avian leukemia oncogene 1
Hnrnpl	Reduced T cells in conditional knockout (70)	heterogeneous nuclear ribonucleoprotein L
Il2ra	Activated (increased CD44hi) T cells and enlargement of spleen and LN (71)	interleukin 2 receptor, alpha chain
Il7r	Low lymphocytes (72)	interleukin 7 receptor
Itch	Activated (increased CD44hi) T cells (73)	itchy, E3 ubiquitin protein ligase
Itpkb	Absence of T cells (74)	inositol 1,4,5-trisphosphate 3-kinase B
Jak3	Low lymphocytes (75, 76)	Janus kinase 3
Lig4	Low B and T cells in ENU mutant (48)	ligase IV, DNA, ATP-dependent
Prkdc	Absence of B and T cells (77–79)	protein kinase, DNA activated, catalytic polypeptide
Ptpn6	Early death, immunodeficiency, autoimmunity in spontaneous mutants (80, 81); reduced B cells, down-regulation of IgM and IgD on mature B cells in ENU mutants	protein tyrosine phosphatase, nonreceptor type 6
Ptprc	Low B and T cells (82)	protein tyrosine phosphatase, receptor type, C
Rag1	Absence of B and T cells (83)	recombination activating gene 1
Rag2	Absence of B and T cells (84)	recombination activating gene 2
Rasgrp1	Low mature T cells (85)	RAS guanyl releasing protein 1
Satb1	Small mice, early death, low and abnormal T cells (86)	special AT-rich sequence binding protein 1
Tbx21	Low NK and NK T cells (87) and Th1 response (88)	T-box 21
Tnfaip3	Increased granulocytes and activated (CD44hi) T cells in ENU mutant	tumor necrosis factor, alpha-induced protein 3

Table S2. Comparison of scores calculated with mouse- and human-specific PolyPhen2 and with human-specific CADD

		Mouse (mm9 assembly)								Human (GRCh37 assembly)											
							Polyphen2											Polyphen2*		CADD†	
Gene symbol	Mutant phenotype mouse	Chr	Coordinate	Base substitution	UniProt identifier	Amino acid substitution	Category	Score	Chr	Coordinate	Strand	Ref nt	Ref codon	Var nt	Ref AA	Var AA		Category	Score	C score	Phred-like score
Bcl2		1	108608935	T/C	Q4VBF6	T175A	Poss. Dam.	0.924	18	60985368	−1	A	ACT	G	T	A		Poss. Dam.	0.898	1.254779	14.68
Btk	Yes	X	131093803	A/G	P35991	Y152H	Prob. Dam.	0.998	X	100617615	−1	T	TAT	C	Y	H		Prob. Dam.	0.985	5.052784	27
Cd22		7	31661166	T/C	Q3UP36	M157V	Benign	0	19	35826965	1	C	CTC		L**			‡			‡
Cd74		18	60970980	A/T	P04441	I203F	Poss. Dam.	0.434	5	149782843	−1	A	ATC	T	I	F		Benign	0.006	1.847686	17.91
Dock8		19	25125999	A/G	Q8C147	K26E	Poss. Dam.	0.894	9	271649	1	A	AAA		F**			‡			‡
Dock8		19	25248733	A/T	Q8C147	N1567Y	Poss. Dam.	0.904	9	432235	1	A	AAT	T	N	Y		Poss. Dam.	0.578	4.777098	26
Dock8		19	25257466	A/G	Q8C147	T1748A	Benign	0.001	9	441301	1	A	ACA	G	T	A		Benign	0.05	1.291052	14.89
Ets1	Yes	9	32560393	C/T	P27577	P334L	Prob. Dam.	1	11	128333513	−1	C	CCA	T	P	L		Prob. Dam.	0.983	2.002118	19.09
Hnrnpl		7	29598392	C/A	Q8R081	A118D	Prob. Dam.	1	19	39337980	−1	C	GCC	T	A	V		Benign	0.079	2.390845	22
Il2ra		2	11604726	T/C	P01590	V230A	Poss. Dam.	0.763	10	6061417	−1	A	ATA		I**			‡			‡
Il7r	Yes	15	9454864	T/G	P16872	T56P	Prob. Dam.	0.985	5	35861037	1	A	ACC	C	T	P		Prob. Dam.	0.917	0.914559	12.47
Itch		2	155034818	A/T	Q8C863	H563L	Prob. Dam.	0.972	20	33067458	1	A	CAT	T	H	L		Poss. Dam.	0.692	2.276928	21.5
Itpkb		1	182263124	T/C	B2RXC2	L228P	Benign	0	1	226924492	−1	A	CCC		P**			‡			‡
Jak3		8	76731835	A/G	Q62137	I663V	Poss. Dam.	0.705	Absent in GRCh37§												
Lig4		8	9972775	T/C	Q8BTF7	Y335C	Prob. Dam.	0.985	13	108862613	−1	A	ATG		M**			Prob. Dam.	0.923	4.648035	25.7
Lig4		8	9973305	A/T	Q8BTF7	N158K	Benign	0.013	13	108863143	−1	T	AAT	A	N	L		Benign	0.016	1.494994	15.95
Prkdc		16	15664332	A/G	P97313	D382G	Prob. Dam.	0.997	8	48849046	−1	A	GAC	G	D	G		Benign	0.059	3.592267	23.6
Prkdc		16	15697319	T/C	P97313	I1010T	Poss. Dam.	0.935	8	48817433	−1	T	ATA	C	I	T		Poss. Dam.	0.894	4.722739	25.9
Prkdc		16	15809737	G/T	P97313	V3389L	Benign	0.097	8	48733488	−1	A	TAC	T	Y	F		Benign	0.142	2.284577	21.5
Prkdc		16	15790559	A/T	P97313	Y2044F	Benign	0.346	8	48711903	−1	G	GTG	C	V	L		Benign	0.06	1.065498	13.5
Prkdc		16	15816851	T/C	P97313	V3589A	Benign	0.007	Absent in GRCh37§												
Ptpn6		6	124682374	A/T	P29351	D90E	Benign	0.015	12	7061284	1	C	GAC	A	D	E		Poss. Dam.	0.511	0.916457	12.48
Ptprc		1	139975069	T/C	P06800	K921R	Benign	0.322	1	198713286	1	A	AAA	G	K	R		Benign	0.274	2.360073	21.9
Rag1	Yes	2	101482545	T/C	P15919	E803G	Poss. Dam.	0.879	11	36597271	1	A	GAG	G	E	G		Prob. Dam.	0.988	5.433662	29.2
Rag2		2	101470774	A/G	P21784	D424G	Benign	0.151	11	36614448	−1	G	CCT		P**			‡			‡
Rasgrp1		2	117110866	T/C	Q9Z1S3	K659R	Poss. Dam.	0.712	15	38786866	−1	A	AAG	G	K	R		Benign	0.389	2.452775	22.2
Satb1		17	51944614	A/G	Q60611	V99A	Prob. Dam.	0.999	3	18458486	−1	T	GTG	C	V	A		Prob. Dam.	0.997	2.179183	20.7
Tbx21		11	96960237	T/G	Q9JKD8	E481A	Prob. Dam.	0.971	17	45822581	1	A	GAA	C	E	A		Poss. Dam.	0.551	1.464169	15.79
Tnfaip3		10	18724943	A/G	Q60769	F394S	Prob. Dam.	1	6	138199763	1	T	TTC	C	F	S		Prob. Dam.	0.999	1.65794	16.78
Tnfaip3		10	18731461	T/C	Q60769	K41E	Benign	0.049	6	138192485	1	A	AAA	G	K	E		Benign	0.01	2.538327	22.3

Chr, chromosome; Poss. Dam., possibly damaging; Prob. Dam., probably damaging; Ref AA, reference amino acid; Ref nt, reference nucleotide; Var AA, variant amino acid; Var nt, variant nucleotide.

*Human PolyPhen2 values calculated using orthologous human protein from UniProt and were only calculated where the reference amino acid was the same between species. Human calculations performed using public PolyPhen2 web server at genetics.bwh.harvard.edu/pph2/bgi.shtml.

†CADD scores obtained from public web server at cadd.gs.washington.edu.

‡Absent values are a consequence of the mouse and human orthologs having a different amino acid at the variant site.

§Nucleotide position unable to mapped from mm9 to hg19 with LiftOver (genome.ucsc.edu/cgi-bin/hgLiftOver).

Table S3. Predicted and observed effect of unselected de novo ENU mutations in 23 essential immune genes

Gene symbol	Number homozygotes tested	Mutant phenotype observed?	Coordinate (mm9)*	Base substitution	UniProt identifier	Amino acid position	Gene symbol		Number homozygotes tested		GERP	MutationAssessor		PANTHER*		Condel	
							Category	Score	Category	Score		Category	Score	subPSEC	Pdel.	Category	Score
Dclre1c	3	Yes	Chr2:3351056	T/A	Q8K4J0	L95Stop		NA		NA	NA		NA	NA	NA		NA
Dock8	1	Yes	Chr19:25249949	A/T	Q8C147	R1630Stop		NA		NA	NA		NA	NA	NA		NA
Dock8	3	Yes	Chr19:25262899	G/T	Q8C147	F1885Stop		NA		NA	NA		NA	NA	NA		NA
Ets1	4	Yes	Chr9:32560393	C/T	P27577	P334L	Prob. Dam.	1	Del.	0	3.97	Medium	2.685	−2.3993	0.35418	Del.	0.638
Hnrnpl	5	No	Chr7:29598392	C/A	Q8R081	A118D	Prob. Dam.	1	Del.	0	3.74	Medium	2.19	−4.06474	0.7436	Del.	0.612
Tnfaip3	3	No	Chr10:18724943	A/G	Q60769	F394S	Prob. Dam.	1	Del.	0	3.27	Medium	1.955	−3.25635	0.56374	Del.	0.6
Satb1	2	No	Chr17:51944614	A/G	Q60611	V99A	Prob. Dam.	0.999	Toler.	0.6		Low	1.285	−4.11365	0.75281	Neutral	0.446
Btk	3	Yes	ChrX:131093803	A/G	P35991	Y152H	Prob. Dam.	0.998	Del.	0	3.91	Low	1.32	−2.39271	0.35268	Del.	0.567
Prkdc	8	No	Chr16:15664332	A/G	P97313	D382G	Prob. Dam.	0.997	Del.	0.04	3.91	Medium	2.35	−3.61416	0.64889	Del.	0.612
Il7r	3	Yes	Chr15:9454864	T/G	P16872	T56P	Prob. Dam.	0.985	Del.	0.01	2.92	Low	1.445			Del.	0.569
Lig4	4	No	Chr8:9972775	T/C	Q8BTF7	Y335C	Prob. Dam.	0.985	Del.	0	2.32						
Itch	1	No	Chr2:155034818	A/T	Q8C863	H563L	Prob. Dam.	0.972	Del.	0	3.61	Medium	3.105	−3.98029	0.72716	Del.	0.654
Tbx21	7	No	Chr11:96960237	T/G	Q9JKD8	E481A	Prob. Dam.	0.971	Toler.	0.53	3.22	Low	1.04	−0.97263	0.11636	Neutral	0.442
Prkdc	4	No	Chr16:15697319	T/C	P97313	I1010T	Poss. Dam.	0.935	Del.	0	3.17	Medium	2.6	−1.55087	0.19013	Del.	0.62
Bcl2	3	No	Chr1:108608935	T/C	Q4VBF6	T175A	Poss. Dam.	0.924	Toler.	0.33	3.04	Neutral	0.395	−2.00435	0.2698	Neutral	0.439
Dock8	5	No	Chr19:25248733	A/T	Q8C147	N1567Y	Poss. Dam.	0.904	Del.	0	3.66	Medium	2.67	−4.15438	0.76031	Del.	0.618
Dock8	5	No	Chr19:25125999	A/G	Q8C147	K26E	Poss. Dam.	0.894	Del.	0.01	3.92	Low	1.585			Del.	0.558
Rag1	1	Yes	Chr2:101482545	T/C	P15919	E803G	Poss. Dam.	0.879	Del.	0	3.83	High	3.71	−2.75501	0.43906	Del.	0.666
Il2ra	2	No	Chr2:11604726	T/C	P01590	V230A	Poss. Dam.	0.763	Del.	0.01	1.89	Low	1.04	−2.25488	0.32189	Del.	0.504
Rasgrp1	4	No	Chr2:117110866	T/C	Q9Z1S3	K659R	Poss. Dam.	0.712	Toler.	0.24	3.49	Low	0.805	−1.80793	0.23289	Neutral	0.436
Jak3	2	No	Chr8:76731835	A/G	Q62137	663V	Poss. Dam.	0.705	Del.	0.04		Low	0.96	−3.23261	0.55789	Del.	0.482
Cd74	1	No	Chr18:60970980	A/T	P04441	I203F	Poss. Dam.	0.434	Toler.	0.13	−2.02	Medium	1.965	−2.04615	0.27811	Neutral	0.462
Prkdc	5	No	Chr16:15809737	G/T	P97313	V3389L	Benign	0.346	Toler.	0.31	3.73	Medium	2.28			Neutral	0.424
Ptprc	3	No	Chr1:139975069	T/C	P06800	K921R	Benign	0.322	Toler.	0.18	2.64	Low	0.805	−0.99775	0.11897	Neutral	0.37
Rag2	2	No	Chr2:101470774	A/G	P21784	D424G	Benign	0.151	Toler.	0.06	3.51	Medium	1.995	−2.32166	0.33663	Neutral	0.421
Prkdc	2	No	Chr16:15790559	A/T	P97313	Y2044F	Benign	0.097	Toler.	0.16	3.35	Medium	2.175	−2.36465	0.3463	Neutral	0.399
Tnfaip3	1	No	Chr10:18731461	T/C	Q60769	K41E	Benign	0.049	Toler.	0.18		Neutral	0.205	−0.78466	0.09838	Neutral	0.284
Ptpn6	4	No	Chr6:124682374	A/T	P29351	D90E	Benign	0.015	Del.	0.02	−5.96	Medium	2.45			Neutral	0.425
Lig4	1	No	Chr8:9973305	A/T	Q8BTF7	N158K	Benign	0.013	Toler.	0.21	−0.667						
Prkdc	3	No	Chr16:15816851	T/C	P97313	V3589A	Benign	0.007	Toler.	0.49	−1.75	Low	0.895	−1.69134	0.21271	Neutral	0.249
Dock8	4	No	Chr19:25257466	A/G	Q8C147	T1748A	Benign	0.001	Toler.	1	2.51	Neutral	−0.555	−1.77255	0.22663	Neutral	0.072
Cd22	1	No	Chr7:31661166	T/C	Q3UP36	M157V	Benign	0	Toler.	0.18	2.65	Neutral	0.345	−0.93903	0.11295	Neutral	0.282
Itpkb	2	No	Chr1:182263124	T/C	B2RXC2	L228P	Benign	0	Toler.	0.16	−1.04	Neutral	−0.55			Neutral	0.24

Del., deleterious; NA, not applicable; Pdel., $P_{deleterious}$; Prob. Dam., probably damaging; Poss. Dam., possibly damaging; Toler., tolerated.
*Missing values were not contained in the alignment created by the HMM.

Table S4. Predicted effect of phenotypically selected ENU mutations in 23 essential immune genes

Gene symbol	Coordinate (mm10)	Base substitution	UniProt identifier	Amino acid position	Polyphen2		SIFT		GERP	PANTHER		Condel		PANTHER	
					Category	Score	Category	Score		Category	Score	subPSEC	Pdel.	Category	Score
Btk	ChrX:134558685	C to T	P35991	Q196Stop	NA	NA	NA	NA	3.91	NA	NA	NA	NA	NA	NA
Btk	ChrX:134570797	G to A	P35991	Intronic	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA
Cd22	Chr7:30870310	T to A	P35329	C511Stop	NA	NA	NA	NA	−0.765	NA	NA	NA	NA	NA	NA
Prkdc	Chr16:15714375	T to C	P17313	Intronic	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA
Prkdc	Chr16:15777109	T to A	P17313	Y3442Stop	NA	NA	NA	NA	2.21	NA	NA	NA	NA	NA	NA
Ptprc	Chr1:139980367	T to A	P06800	Intronic	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA
Ptprc	Chr1:140007727	T to C	P06800	Intronic	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA
Rag1	Chr2:101845708	C to T	P15919	R328Stop	NA	NA	NA	NA	−0.906	NA	NA	NA	NA	NA	NA
Jak3	Chr8:71680943	T to C	Q62137	L398P	Prob. Dam.	1	Not Tol.	0.00	−3.3	Medium	2.545	−5.78447	0.94183	Deleterious	0.635
Jak3	Chr8:71685302	G to T	Q62137	G825V	Prob. Dam.	1	Not Tol.	0.00	3.08	High	4.145	−6.22242	0.96167	Deleterious	0.717
Ptpn6	Chr6:124722095	C to T	P29351	T462M	Prob. Dam.	1	Not Tol.	0.03	−4.72	High	3.61	−4.82869	0.86161	Deleterious	0.689
Ptprc	Chr1:138099437	T to C	P06800	F526S	Prob. Dam.	1	Not Tol.	0.01	0.292	Medium	2.095	*	*	Deleterious	0.607
Tnfaip3	Chr10:19005539	T to A	Q60769	I325N	Prob. Dam.	1	Not Tol.	0.00	−0.294	Low	1.04	−3.71158	0.67075	Deleterious	0.556
Rasgrp1	Chr2:117287984	A to G	Q9Z1S3	R519G	Prob. Dam.	0.991	Not Tol.	0.01	1.23	Low	1.935	*	*	Deleterious	0.597
Satb1	Chr17:51782585	T to A	Q60611	F393L	Prob. Dam.	0.984	Not Tol.	0.00	3.63	Low	1.875	−2.081	0.28516	Deleterious	0.594
Lig4	Chr7:13311208	A to G	Q8BTF7	Y288C	Prob. Dam.	0.975	Not Tol.	0.04	3.83	Low	1.805	−5.10454	0.89134	Deleterious	0.592
Ptpn6	Chr6:124732942	C to T	P29351	V2M	Poss. Dam.	0.913	Not Tol.	0.00	2.91	Neutral	0.68	*	*	Deleterious	0.517
Ptprc	Chr1:138101016	T to C	P06800	C463R	Benign	0.56	Tolerated	0.11	1.59	Low	1.095	*	*	Deleterious	0.469
Bcl2	Chr1:106711797	T to C	P10417	S24P	Benign	0.203	Not Tol.	0.00	−0.17	Low	1.355	−2.05643	0.28018	Neutral	0.411

NA, not applicable; Not Tol., not tolerated; Pdel., $P_{deleterious}$; Poss. Dam., possibly damaging; Prob. Dam., probably damaging.
*Position not aligned in the HMM.

Table S5. Apparent accuracy of PolyPhen2 predictions for different sets of TP53 mutations

Variant set	True positive (pph2 ≥ 0.8 and TA ≤ 50)	True negative (pph2 ≤ 0.2 and TA > 50)	False Positive (pph2 ≥ 0.8 and TA > 50)	False negative (pph2 ≤ 0.2 and TA ≤ 50)	% false-positive predictions among mutations with pph2 ≥ 0.8
All possible variants, target gene *TP53* binding sequence					
WAF1	640 (31.6%)	831 (41.0%)	462 (22.8%)	93 (4.6%)	42%
MDM2	728 (35.9%)	709 (35.0%)	373 (18.4%)	217 (10.7%)	34%
BAX	660 (32.6%)	778 (38.4%)	441 (21.8%)	147 (7.3%)	40%
14–3-3σ	619 (30.6%)	817 (40.3%)	482 (23.8%)	107 (5.3%)	44%
p53AIP1	621 (30.6%)	753 (37.1%)	482 (23.8%)	173 (8.5%)	44%
GADD45	566 (27.9%)	809 (39.9%)	536 (26.5%)	115 (5.7%)	49%
NOXA	470 (23.2%)	869 (42.9%)	631 (31.2%)	55 (2.7%)	57%
p53R2	433 (21.4%)	864 (42.7%)	668 (33.0%)	60 (3.0%)	61%
Clinically observed variants, *WAF1* binding sequence					
Somatic	560 (47%)	377 (31.6%)	188 (15.8%)	67 (5.6%)	25%
Germ line	132 (76.7%)	22 (12.8%)	12 (7.0%)	6 (3.5%)	8%

Table S6. Apparent accuracy of MutationAssessor predictions for different sets of TP53 mutations

Variant set	True positive (MA ≥ 2 and TA ≤ 50)	True negative (MA < 2 and TA ≥ 50)	False positive (MA ≥ 2 and TA ≥ 50)	False negative (MA < 2 and TA ≤ 50)	% false-positive predictions among mutations with MA ≥ 2
All possible variants, target gene *TP53* binding sequence					
WAF1	651 (28.2%)	935 (40.5%)	581 (25.2%)	142 (6.1%)	20%
MDM2	765 (33.1%)	796 (34.5%)	466 (20.2%)	283 (12.3%)	38%
BAX	678 (29.4%)	877 (38.0%)	554 (24.0%)	200 (8.7%)	27%
14-3-3σ	632 (27.4%)	929 (40.3%)	599 (26.0%)	148 (6.4%)	20%
p53AIP1	626 (27.1%)	848 (36.7%)	605 (26.2%)	233 (10.1%)	28%
GADD45	579 (25.1%)	923 (40.0%)	653 (28.3%)	155 (6.7%)	19%
NOXA	484 (21.0%)	999 (43.3%)	747 (32.4%)	78 (3.4%)	9%
p53R2	449 (19.4%)	998 (43.2%)	783 (33.9%)	79 (3.4%)	9%
Clinically observed variants, *WAF1* binding sequence					
Somatic	93 (57.1%)	35 (21.5%)	17 (10.4%)	18 (11%)	51%
Germ line	138 (75.4%)	20 (10.9%)	18 (9.8%)	7 (3.8%)	28%

Table S7. Apparent accuracy of CADD predictions for different sets of TP53 mutations

Variant set	True positive (Phred ≥ 20 and TA ≤ 50)	True negative (Phred < 20 and TA ≥ 50)	False positive (Phred ≥ 20 and TA ≥ 50)	False negative (Phred < 20 and TA ≤ 50)	% false-positive predictions among mutations with Phred ≥ 20
All possible variants, target gene *TP53* binding sequence					
WAF1	708 (30.7%)	941 (40.8%)	573 (24.8%)	85 (3.7%)	45%
MDM2	832 (36.0%)	812 (35.2%)	449 (19.5%)	215 (9.3%)	35%
BAX	731 (31.7%)	879 (38.1%)	550 (23.8%)	147 (6.4%)	43%
14-3-3σ	671 (29.1%)	916 (39.7%)	610 (26.5%)	109 (4.7%)	48%
p53AIP1	702 (30.4%)	870 (37.7%)	582 (25.2%)	156 (6.8%)	45%
GADD45	617 (26.7%)	909 (39.4%)	666 (28.9%)	116 (5.0%)	52%
NOXA	508 (22.0%)	971 (42.1%)	773 (33.5%)	54 (2.3%)	60%
p53R2	470 (20.4%)	968 (42.0%)	812 (35.2%)	57 (2.5%)	63%
Clinically observed variants, *WAF1* binding sequence					
Somatic	102 (62.6%)	34 (20.9%)	18 (11.0%)	9 (5.5%)	15%
Germ line	140 (76.5%)	19 (10.4%)	19 (10.4%)	5 (2.7%)	12%

5.2 Further discussion

In this two-part study we aimed to answer the following questions.

1) How effective are the tools designed to predict the functional significance of missense mutations?
2) Do the tools work as effectively on non-human samples as human samples?
3) Why do most tools seem to call certain mutations damaging despite there being no discernable phenotype?
4) Can we offer an explanation of this "overcalling" phenomenon using real data sets?

To attempt to answer these questions two separate data sets measuring real protein function were utilized and compared to the predictions from a variety of functional inference computational tools such as Polyphen2 (56) and CADD (58). The first data set was generated in-house as part of the larger ENU mouse project (69), consisting of 33 mice breed to homozygosity, each containing a unique missense mutation spread across 23 essential immune system genes. Mutations were selected that produced a missense mutation in genes that, when rendered null, caused a fully penetrant and well-characterized phenotype in the mouse immune system easily detectable by flow cytometry of peripheral blood lymphocytes. The second data set consisted of a comprehensive mutation data set for the tumour suppressor gene TP53. This data set was generated by systematically quantifying the transcription enhancing activity of WT TP53 by expressing each mutant in yeast *in vitro* and measuring transcription-enhancing activity against different TP53-binding enhancer sequences for all possible missense mutations (63).

With the mouse data we were able to directly compare the prediction of the functional inference tools with real functional data and importantly, by focusing on genes that when rendered null caused a fully penetrant and well-characterized detectable phenotype, determine whether variable penetrance is responsible for the often poor performance of these tools as has been suggested (85). While an extremely valuable data set for ruling out variable penetrance as the cause of overcalling, several other possible explanations remain: *i)* With known redundancy in the immune system (86), uncharacterized compensatory

processes may be masking the mutation effects *ii)* Sterile laboratory conditions were preventing the mice from exposure to pathogens required for the immune response *iii)* The 'human-centric' focus of the computational tools means subtle differences either in mouse/human homologs or the transcripts selected for each organism may affect the scoring. To address the human-centric nature of the tools we installed the complex polyphen analysis pipeline locally, replacing all human protein data with the appropriate mouse protein data. We applied both the default human and the mouse specific versions of polyphen to 30 of the 33 mouse mutations, with the other 3 mutations creating stop codons that produced a phenotype as expected. For the 30 mutations tested, 20 were predicted to be damaging by the default human polyphen yet only 4 produced a phenotype (20% success) with the remaining 10 correctly predicted to be not damaging, confirming that false positives are a bigger problem than false negatives. The mouse specific version of polyphen made some small improvements (e.g. Rag1 mutant which produced a phenotype changed from 'possibly damaging' to 'probably damaging' using mouse polyphen) but overall the scores were well correlated between versions of polyphen meaning this difference is unlikely responsible for the overcalling we observe.

Lastly, to investigate whether environmental masking or compensation were responsible for overcalling, an exhaustive TP53 functional data set was analysed, a valuable data set as loss of function TP53 mutations are the most common somatic mutation in human cancer and known to cause highly penetrant immune system cancers when bred to homozygosity in the germ line of mice. In this analysis the entire set of all missense mutations was compared to functional assay scores for all possible 2314 missense mutations. The results showed that while the overcalling problem is less pronounced than in the mouse data set, there are still ~25% of all mutations predicted to be damaging that have no impact on transactivation function compared to the relatively small ~5% of non-damaging predictions that did impact transactivation function. After discounting the last remaining commonly held explanation for the observed overcalling we considered why this issue persisted regardless of the tool selected and the functional data set utilized for comparison. In the end, we proposed that the persistent overcalls result from the reliance of the tools on conservation across species, and that the overcalls corresponds to nearly neutral mutations of small

effect (62); mutations that are subject to purifying selection yet do not necessarily produce a phenotype in a single individual. To test this hypothesis, additional data sets were analysed for the 23 genes studied in the mouse data set, focusing on data sets subject to differing degrees of purifying selection. At one extreme we considered all possible missense mutations across the 23 genes while at the other extreme we considered all missense variants between the mouse and rat, species which have been separate for 12–24 million years (87). These results appear to support this hypothesis with ~50% of all possible mutations in the 23 genes predicted to be damaging contrasting with the mouse-rat mutations being classified as almost exclusively benign. A similar comparison for humans generated similar results with ~50% of all possible missense variants in the 23 genes had damaging predictions while almost all human-chimp missense variants had benign predictions. Both these results support the hypothesis that the overcalls correspond to nearly neutral mutations as data subject to long periods of purifying selection (e.g. mouse-rat and human-chimp missense mutations) are largely predicted to be benign while data sets subject to shorter periods of purifying selection (e.g. all possible mutations) are often predicted to be damaging.

These results highlight a critical gap in our ability to relate genotype to phenotype in a clinical setting and make clear that current functional inference tools are unable to differentiate immediately actionable relevant mutations from nearly neutral mutations. The results further highlight the problem with overreliance on such tools in any variant prioritization strategy, particularly when being used in the clinical setting. By proposing a novel explanation for the observed overcalling, we are hopeful this will lead to an understanding that using such tools in isolation may not be appropriate for use in the clinical. Instead these tools need to be part of a broader overall strategy. Finally, using this analysis and others, it may be possible to implement a next generation of functional inference software able to differentiate between truly damaging and nearly neutral mutations based on currently unknown differences in the signature of their mutations.

Chapter 6: DeepSNVMiner: A sequence analysis tool to detect emergent, rare mutations in sub-sets of cell populations

TD Andrews, Y Jeelall, D Talaulikar, CC Goodnow, MA Field. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. PeerJ. 2016; 24;4:e2074

DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations

T. Daniel Andrews[1,2], Yogesh Jeelall[1,3], Dipti Talaulikar[1,4,5], Christopher C. Goodnow[1,6,7] and Matthew A. Field[1]

[1] Department of Immunology, John Curtin School of Medical Research, Australian National University, Canberra ACT, Australia
[2] National Computational Infrastructure, Canberra ACT, Australia
[3] School of Medicine and Pharmacology, University of Western Australia, Harry Perkins Institute, Perth, Australia
[4] Haematology Translational Research Unit, Haematology Unit, ACT Pathology, Canberra ACT, Australia
[5] ANU Medical School, Australian National University, Canberra ACT, Australia
[6] Immunology Division, Garvan Institute of Medical Research, Sydney NSW, Australia
[7] St Vincent's Clinical School, University of New South Wales, Darlinghurst NSW, Australia

ABSTRACT

Background. Massively parallel sequencing technology is being used to sequence highly diverse populations of DNA such as that derived from heterogeneous cell mixtures containing both wild-type and disease-related states. At the core of such molecule tagging techniques is the tagging and identification of sequence reads derived from individual input DNA molecules, which must be first computationally disambiguated to generate read groups sharing common sequence tags, with each read group representing a single input DNA molecule. This disambiguation typically generates huge numbers of reads groups, each of which requires additional variant detection analysis steps to be run specific to each read group, thus representing a significant computational challenge. While sequencing technologies for producing these data are approaching maturity, the lack of available computational tools for analysing such heterogeneous sequence data represents an obstacle to the widespread adoption of this technology.

Results. Using synthetic data we successfully detect unique variants at dilution levels of 1 in a 1,000,000 molecules, and find DeeepSNVMiner obtains significantly lower false positive and false negative rates compared to popular variant callers GATK, SAMTools, FreeBayes and LoFreq, particularly as the variant concentration levels decrease. In a dilution series with genomic DNA from two cells lines, we find DeepSNVMiner identifies a known somatic variant when present at concentrations of only 1 in 1,000 molecules in the input material, the lowest concentration amongst all variant callers tested.

Conclusions. Here we present DeepSNVMiner; a tool to disambiguate tagged sequence groups and robustly identify sequence variants specific to subsets of starting DNA molecules that may indicate the presence of a disease. DeepSNVMiner is an automated workflow of custom sequence analysis utilities and open source tools able to differentiate somatic DNA variants from artefactual sequence variants that likely arose during DNA amplification. The workflow remains flexible such that it may be customised to variants of the data production protocol used, and supports reproducible analysis through detailed logging and reporting of results. DeepSNVMiner

Submitted 12 January 2016
Accepted 3 May 2016
Published 24 May 2016

Corresponding author
Matthew A. Field,
matt.field@anu.edu.au

Academic editor
Elena Papaleo

Additional Information and Declarations can be found on page 10

DOI **10.7717/peerj.2074**

is available for academic non-commercial research purposes at https://github.com/mattmattmattmatt/DeepSNVMiner.

Subjects Bioinformatics, Computational Biology, Genomics
Keywords NGS, Deep sequencing, Rare mutations, Variant detection

INTRODUCTION

Deep sequencing of a restricted set of gene targets in a large population of cells has rapidly become a key application of second-generation sequencing, allowing a census of variation to be conducted on an *in vivo* biological system (*Fu et al., 2011*; *Hiatt et al., 2013*; *Jabara et al., 2011*; *Kinde et al., 2011*; *Kivioja et al., 2012*; *Schmitt et al., 2012*). Applications of this technology allow polling of sequence variation in cancer subtypes (*Forshew et al., 2012*), ascertainment of minimal residual disease (*Bidard, Weigelt & Reis-Filho, 2013*), ascertainment of malignancies or antibody specificity in the immune system (*Georgiou et al., 2014*) and observation of the emergence of drug resistant virus point-mutants (*Al-Mawsawi et al., 2014*).

The central technique in molecule tagging that allows disambiguation of these deep sequence datasets is the attachment of a random unique sequence identifier (UID) to the end(s) of input DNA, either prior to or simultaneously with amplification of target sequences (Fig. 1). Hence, even though subsequent polymerase amplification of target sequences may introduce errors, mapping these sequences to their UID sequence allows easy differentiation of sequence variation that was originally present in the input DNA from variation that has been introduced during subsequent amplification steps. Recently developed methods for molecule tagging rely on digital PCR, a process where individual DNA molecules are assessed individually (*Vogelstein & Kinzler, 1999*). Several variants of this technique have now been described (*Dressman et al., 2003*; *Ottesen et al., 2006*) with the common thread being the binding of oligonucleotide to each individual input DNA molecule prior to or during amplification. This technique is not to be confused with sample barcoding or multiplexing, a process where individual samples are tagged with small oligonucleotides and pooled in a single lane for sequencing.

In comparison to traditional massively parallel sequencing, molecule tagging has an additional step where a small unique oligonucleotide is attached to each DNA molecule prior to polymerase chain reaction (PCR) amplification. While both techniques generate huge numbers of sequenced DNA molecules in parallel a potential issue with traditional sequencing is that the introduction of erroneous base calls into a single DNA molecule can result in inaccurate sequence information being amplified in subsequent PCR steps. Such issues are not necessary prohibitive for reliable variant detection when samples are relatively homogeneous however, mainly due to the relatively low base error and PCR bias rates (*Ross et al., 2013*; *Schirmer et al., 2015*), and the ability to remove candidate PCR duplicates reads using tools such as SAMTools (*Li et al., 2009*) or SAMBLASTER (*Faust & Hall, 2014*). When the samples being sequenced are heterogeneous however, traditional

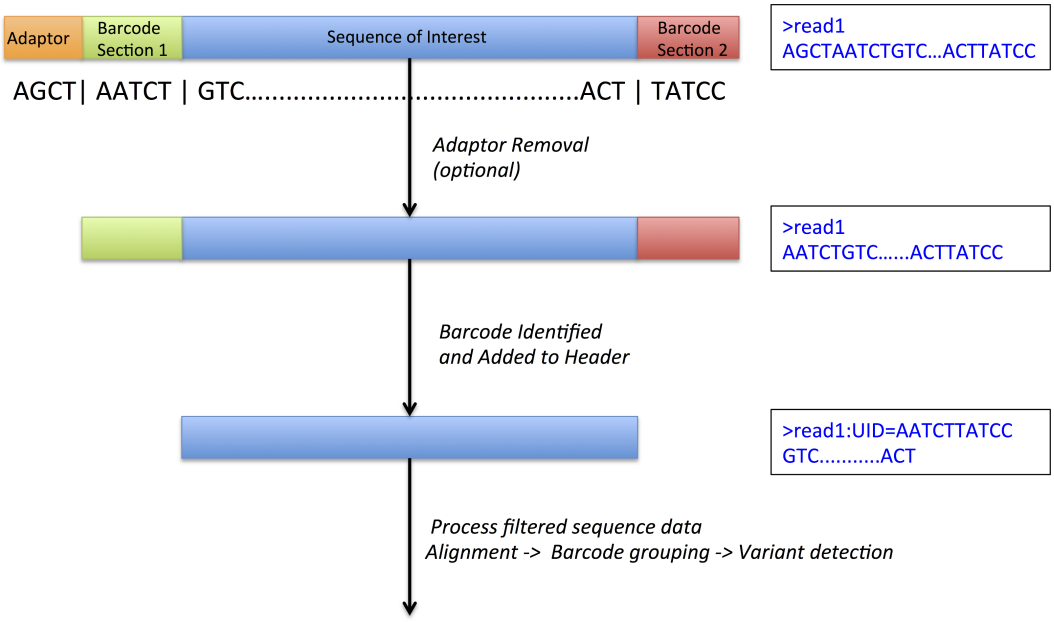


Figure 1 DeepSNVMiner barcode and adaptor processing. A sample sequence read which has undergone molecule tagging with a ten base pair UID consisting of five bases each attached to the 5′ and 3′ end. DeepSNVMiner first removes any adapters followed by the removal of the unique sequence identifier (UID) from the raw sequence data. The UID sequence information is preserved in the FASTQ header allowing for variant detection on read groups sharing a common UID.

variant detection methods often fail to reliably detect rare variants due to the small fraction of the original material containing the variant of interest and differences in the variant detection algorithms (*Field et al., 2015*). With molecule tagging techniques, we are able to overcome PCR issues and detect rare variants within heterogeneous samples due to the attachment of UIDs, effectively allowing the differentiation of amplification error from variation present in the original DNA molecules (*Kinde et al., 2011*).

While the utility of sequencing tagged samples is clear, the analysis of sequence data generated with UID tags is non-trivial and, as yet, software or a computational workflow does not exist in the public domain to allow easy calling and tallying of this mutation information. The fundamental technical challenge of working with such data is largely due to the wide variety of methods for attaching UIDs, methods that generate vastly different UIDs with regard to total sequence length and their position on the molecules relative to the sequence of interest and/or adaptors. The ability to work with such data requires software where users can first define the nature of the specific UID in their experiment, followed by an analysis workflow where UIDs are temporarily removed from raw sequence data for the alignment step and later restored as a means of grouping the individual reads by common UID. Finally, variants must be called within each group of input molecules sharing a common UID, a computationally intensive task given the huge numbers of groups often generated in a single experiment. To address this need we present DeepSNVMiner, a tool able to detect rare single nucleotide variants and small indels specific to a single amplified DNA molecule identified by a unique tagged sequence identifier. The DeepSNVMiner

workflow consists of grouping reads by UID tag sequences and calling variant bases in UID groups, thus identifying mutations that existed in single molecules from the original heterogeneous input DNA. DeepSNVMiner is a standalone-automated workflow that runs in a Linux or Macintosh environment and has been successfully used even on modest desktop hardware.

MATERIALS & METHODS

Cell lines used

Two cell lines were utilised in the dilution series experiment, HEK293 and OCI-LY10. HEK293 is available from ATCC (accession CRL-1573; http://www.atcc.org/products/all/CRL-1573.aspx) and OCI-LY10 from Ontario Cancer Institute (accession CVCL_8795; https://www.abmgood.com/OCI-Ly10-Cell-Lysate-Data-Sheet-L134.html).

Software input

Running DeepSNVMiner requires three input files; paired-end FASTQ read files and a BED file containing the specific locations of targeted genomic region(s). An initial configuration step is also required to determine the location of three required external resources; Burrows-Wheeler Aligner (BWA) (*Li & Durbin, 2009*), SAMtools (*Li et al., 2009*), and a reference genome FASTA file with BWA index files.

Workflow design

The workflow to disambiguate sequence variants from their unique sequence ID tags groups involves multiple steps involving both purpose built tools and calls to external binaries (Fig. 2).

First, the sequence read dataset is subjected to preliminary quality control, to remove low quality reads or those containing predominantly N calls (and hence avoid assigning UID groups of consecutive N's). The data is next interrogated for the presence of obvious adaptor sequence, which may contaminate UID tags if left untrimmed. Each UID tag is then identified based on the user defined input and removed from the FASTQ sequence line and appended to the existing FASTQ read header. These filtered reads and headers are written to new FASTQ files with the UID header information later used to detect variants specific to common UID groups. DeepSNVMiner is flexible with regard to the structure of the UID tag as both the expected UID length and strand location of the UID typically vary depending on the tagging protocol and/or sequencing technology used. For example, frequently the UID is appended solely at the 5′ end of the amplified region, but in other protocols the sequence from both the 5′ and the 3′ ends needs to be concatenated to derive the UID. Next, the modified reads are aligned to a reference genome sequence with BWA (*Li & Durbin, 2009*) using a set of alignment parameters that are permissive of mismatches but which penalise opening a gap within the alignment, especially at the ends of sequence reads. Variant bases are then identified base-by-base using the SAMTools `calmd` command (*Li et al., 2009*) within a predefined set of user-specified target genomic locations input from a BED-format file. SAMtools `calmd` is used instead of the more standard SAMtools/BCFtools workflow, as running the millions of common UID groups we typically observe through the

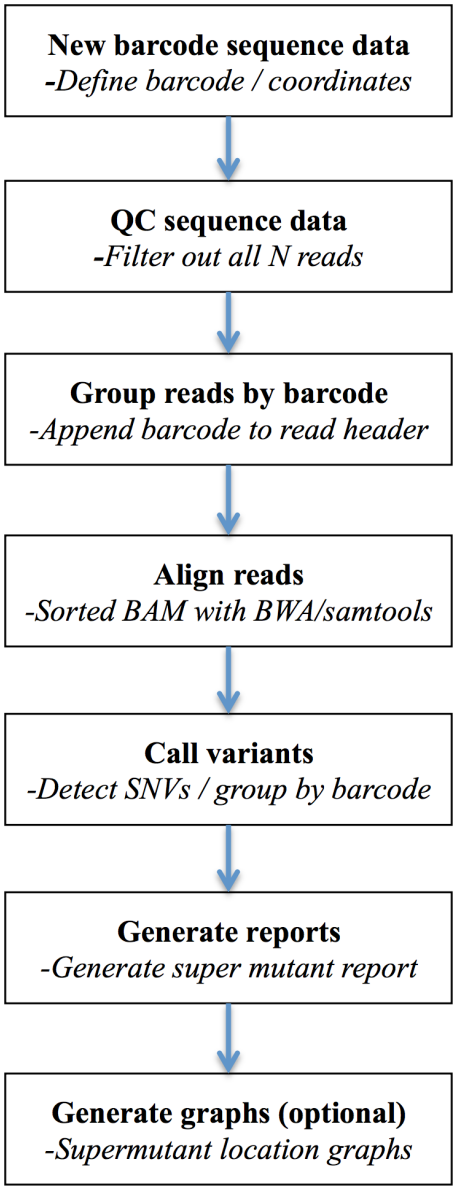


Figure 2 DeepSNVMiner workflow. DeepSNVMiner Workflow consists of seven major steps: UID definition, sequence data QC, UID processing and grouping, alignment, variant detection, report generation, and optionally graphing.

entire SAMtools/BCFtools workflow is computationally prohibitive. Output from `calmd` is next parsed and variant positions and the reads in which they occur are tallied and grouped according to read UIDs. By default, common UID groups of 5 or more reads of which at least 40% detect the same variant are classified as a variant 'super mutant' and variants reported in two or more super mutants further classified as a 'super group.' This default value of 40% was chosen to allow successful super mutant detection even in the rare cases where a super mutant was missed due to an identical UID tag being added to two distinct DNA molecules. Examination of pilot data determined lowering this threshold did not add

any false positive super mutants and further, users of the software are able to override the defaults and determine the appropriate value for each cut-off based on the nature of their dataset and the method by which it was generated. Finally, optional summary graphics of the variants in the context of sequence read depth and chromosomal position are created using R.

Workflow implementation

The overall workflow is comprised of discrete command-line interface calls to both custom- and open source-tools as well as UNIX utilities. All commands are stored within a configuration module and executed by a Perl wrapper script to allow chaining together of each command and automation of the multifaceted workflow. This also allows for easy, frequent customisation of workflow commands (should this be desired) and the capturing of specific commands and run-level information into a log file that contributes to analysis reproducibility. The workflow has the facility to allow it to be resumed or re-run from any point midway through the analysis.

The workflow commands, specifically, are various calls to several purpose-built tools (implemented in Perl), external open-source bioinformatics software tools and UNIX utilities. Custom Perl scripts are used to perform workflow steps to identify, remove and store UID tags from each read, to aggregate and summarise variant calls within UID groups and to generate final reports and graphs. Alignment of sequence reads is accomplished using BWA, variant calling is done with SAMtools `calmd`, and graphing performed with R. Identification and cleaning of reads containing runs of Ns is performed using sed and awk commands piped to other UNIX utilities such as cut, sort, uniq, and cat which are required to manipulate the output of these tools.

Output

The final report contains a listing of all variants detected, based on either the user-configured expected variant frequency or default parameters (e.g., a common UID group must contain at least five reads with at least 40% sharing the same variant). For each called variant, the output super mutant summary reports the chromosome and genomic coordinate(s), the variant base, the UID, the number of total variant reads in the groups, the number of reads in the group and the fraction of variant reads. The super group report contains information on recurrent super mutants (grouped by common genomic coordination and variant base) and additionally reports their frequency.

RESULTS AND DISCUSSION

We developed DeepSNVMiner to disambiguate tagged sequence groups within mixed cell populations and detect sequence variants specific to individual amplified DNA molecules. To assess the performance of DeepSNVMiner we first compare it to the well known variant callers FreeBayes, (http://arxiv.org/abs/1207.3907), Genome analysis toolkit (GATK) (*McKenna et al., 2010*), SAMTools/BCFtools (*Li, 2011*), and LoFreq (*Wilm et al., 2012*) using simulated tagged sequence data at increasing variant dilution levels. Next, we test DeepSNVMiner by running a dilution series with genomic DNA from two cells lines: one of which is known to contain a known heterozygous somatic variant.

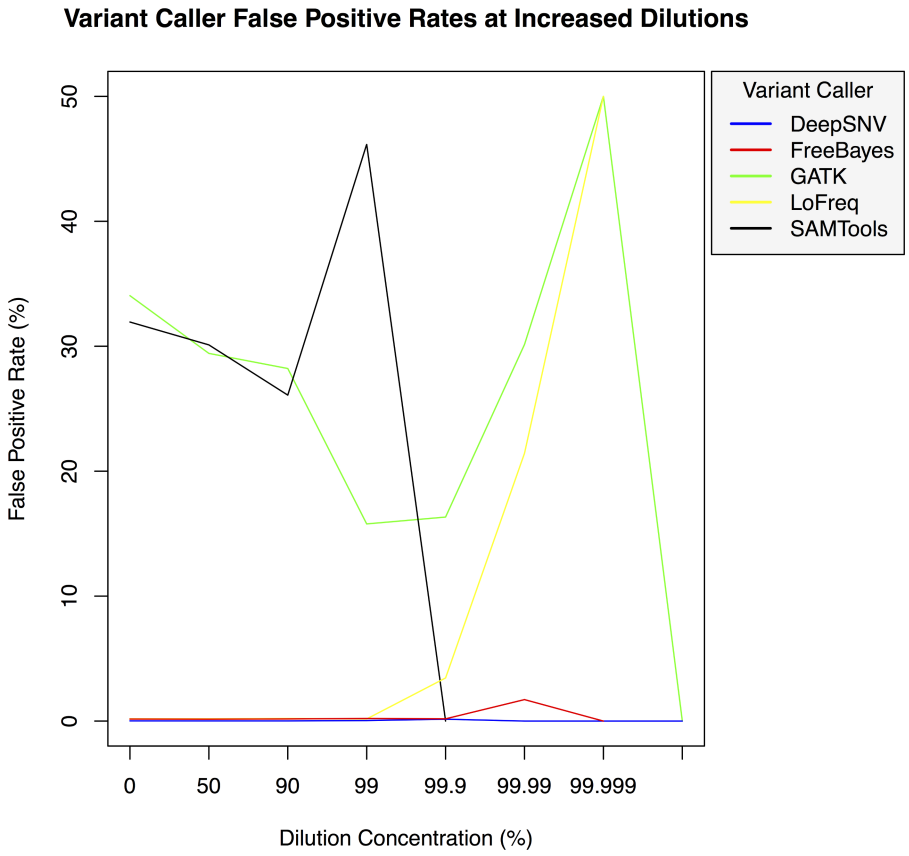


Figure 3 Variant caller false positive rate at increased dilution levels. False positive rates for DeepSNVMiner compared to FreeBayes, GATK, LoFreq, and SAMTools at increasing variant dilution levels using synthetic data.

Comparison with existing software

To test the effectiveness of DeepSNVMiner at increasing dilution levels two datasets containing 100-bp paired-end reads from chromosome 22 of the human reference genome (GRCh37) were created, with each read-pair having a randomly generated 10 bp barcodes attached at the 5′ end to simulate the attachment of a UID sequence tag to the original DNA molecule. The first input data set contained no mutations while the second input data set contained randomly generated single nucleotide variants (SNVs) with each mutated read duplicated randomly between 1 to 50 times within the FASTQ files to simulate the polymerase chain reaction (PCR) replication process of initial DNA fragments. Mixing the two data sets in appropriate concentrations simulated dilution levels of 0%, 50%, 90%, 99%, 99.9%, 99.99%, 99.999%, and 99.9999% with 4,000,000 million total paired end reads ultimately added to each FASTQ file. For each dilution level the FASTQ files were first aligned to chromosome 22 and variants called using DeepSNVMiner, FreeBayes, GATK, SAMTools, and LoFreq run with default parameters or as suggested in documentation (Table S1). False positive (Fig. 3) and false negative rates (Fig. 4) were then calculated (Table S2).

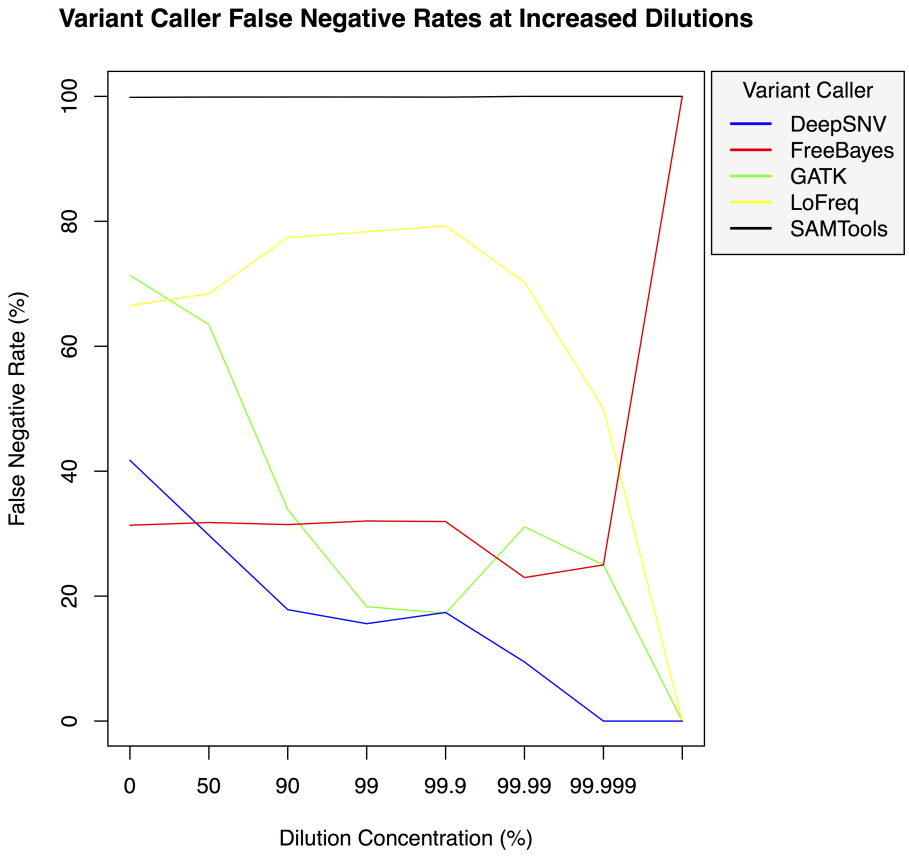


Figure 4 Variant caller false negative rate at increased dilution levels. False negative rates for DeepSNVMiner compared to FreeBayes, GATK, LoFreq, and SAMTools at increasing variant dilution levels using synthetic data.

Table 1 Variant caller ability to detect known heterozygous mutation at different dilution levels. A dilution series was performed with genomic DNA from two cells lines: HEK293 containing wild-type MYD88 and OCI-LY10 containing known heterozygous L265P MYD88 mutation. The ability to detect the known heterozygous mutation was determined for DeepSNVMiner, FreeBayes, GATK, LoFreq, and SAMTools at increasing dilution levels.

Dilution Percent	Deep- SNVMiner	FreeBayes	GATK	LoFreq	SAMTools
0	Detected	Detected	Detected	Detected	Detected
90	Detected	Not detected	Detected	Detected	Not detected
99	Detected	Not detected	Not detected	Detected	Not detected
99.9	Detected	Not detected	Not detected	Not detected	Not detected
99.99	Not detected	Not detected	Not detected	Not detected	Not detected
99.999	Not detected	Not detected	Not detected	Not detected	Not detected
99.9999	Not detected	Not detected	Not detected	Not detected	Not detected

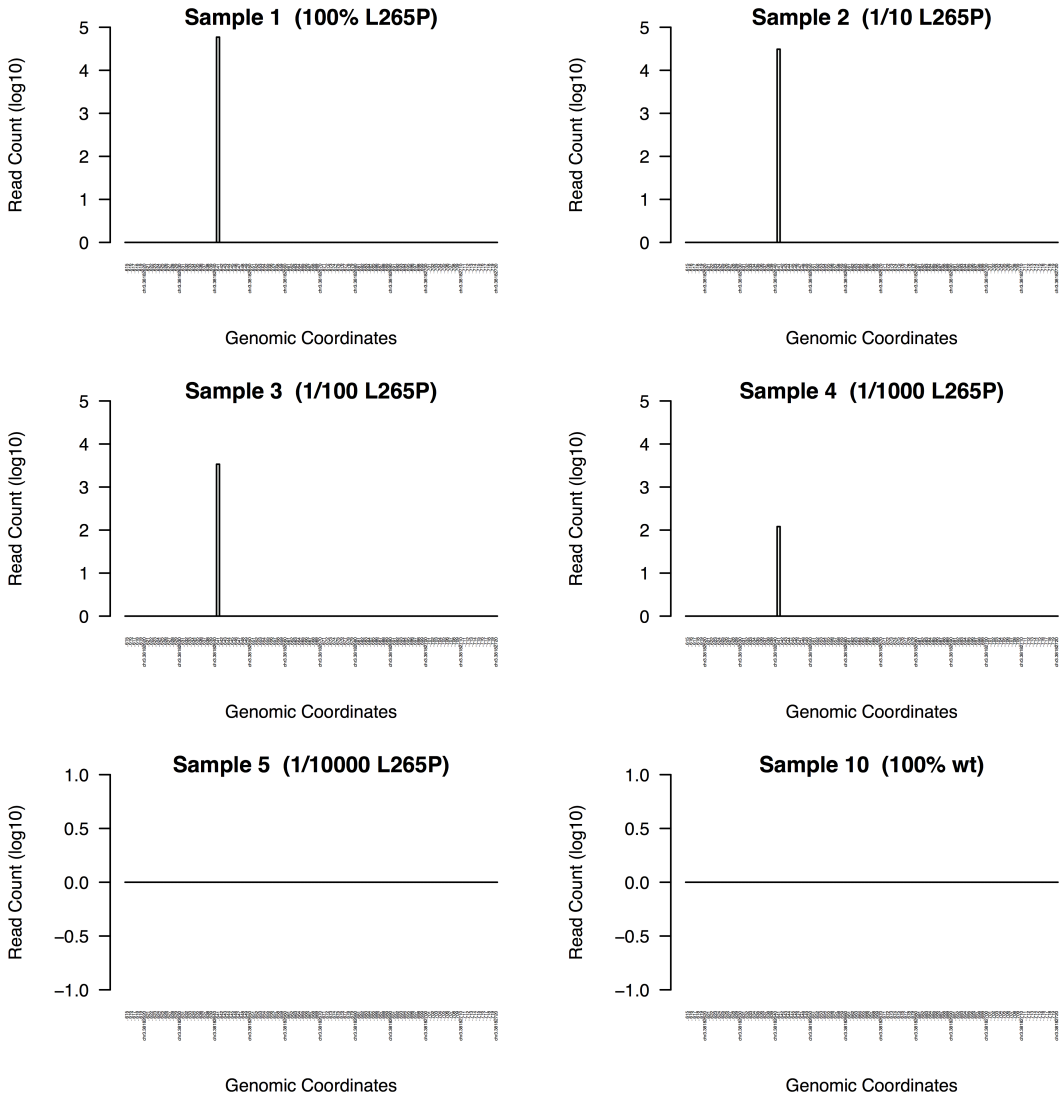


Figure 5 DeepSNVMiner dilution series results. To measure the sensitivity of DeepSNVMiner a dilution series containing known heterozygous L265P MYD88 mutation was performed. DeepSNVMiner was able to detect the mutation in a sample where only 1 in 1,000 samples contained the mutation.

Real data evaluation

To evaluate the performance of DeepSNVMiner on real data versus other variant callers, we performed a dilution series using a mixture of genomic DNA from two cell lines HEK293 and OCI-LY10 (Table S3). Cell line OCI-LY10 carries a heterozygous point mutation within the *MYD88* gene at L265P or chr3:38172641 (GRCh37), a somatic mutation occurring frequently in non-Hodgkin lymphoma (*Ngo et al., 2011*). It would be clinically useful to have a method to detect and enumerate rare cells carrying this mutation in samples of blood or bone marrow. For each cell mixture in the dilution series, a 116 bp genomic region surrounding chr3:38172641 was amplified using primers with UID tags and sample ID tags and a per-sample average of 183 thousand paired-end reads were sequenced on an Illumina MiSeq. The resulting sequence reads were analysed with DeepSNVMiner,

FreeBayes, GATK, LoFreq, and SAMTools and the ability to detect the heterozygous mutation was measured.

DeepSNVMiner was successfully able to detect the mutation in dilution levels down to 1/1000 compared to 1/100 for LoFreq, 1/10 for GATK, and only in the non-diluted sample for FreeBayes and SAMTools (Table 1). In the non-diluted sample, DeepSNVMiner was able to detect the mutation in 4,055 separate super mutants consisting of 59,038 total DNA sequences and at the lower range of detection (1/1000), DeepSNVMiner detected the variant in 6 separate super mutants consisting of 120 total DNA sequences (Fig. 5)

The mutation was reliably detected at concentrations of 1/1000 by DeepSNVMiner but not in concentrations of 1/10000 indicating the lower detection limit lies somewhere in this range. However, it should be noted this limit is imposed by current laboratory methodology however, as DeepSNVMiner remains capable of achieving the theoretical limits of the technology imposed by the chosen length of UID sequences.

CONCLUSIONS

We present DeepSNVMiner; an integrated tool set and automated workflow to allow robust and reliable identification of sequence variants present in a subset of sequences within a tagged input DNA sample. This tool makes available the analysis procedure required to support SafeSeqs and similar UID tagged sequence datasets. DeepSNVMiner has been built to allow easy automation and reproducibility and makes this technique available to a wide range of applications.

ACKNOWLEDGEMENTS

We thank the National Computational Infrastructure (Australia) for continued access to significant computation resources and technical expertise.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

National Institutes of Health Grant U19 AI100627, NHMRC Australian Fellowship 585490, and Bioplatoforms Australia supported this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Institutes of Health Grant: U19 AI100627.
NHMRC Australian Fellowship: 585490.
Bioplatoforms Australia.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- T. Daniel Andrews conceived and designed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper.
- Yogesh Jeelall conceived and designed the experiments, performed the experiments, reviewed drafts of the paper.
- Dipti Talaulikar conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Christopher C. Goodnow conceived and designed the experiments, reviewed drafts of the paper.
- Matthew A. Field conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

Data is available at GitHub: https://github.com/mattmattmattmatt/DeepSNVMiner

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2074#supplemental-information.

REFERENCES

Al-Mawsawi LQ, Wu NC, Olson C, Shi V, Qi H, Zheng X, Wu T, Sun R. 2014. High-throughput profiling of point mutations across the HIV-1 genome. *Retrovirology* **11**:124 DOI 10.1186/s12977-014-0124-6.

Bidard FC, Weigelt B, Reis-Filho JS. 2013. Going with the flow: from circulating tumor cells to DNA. *Science Translational Medicine* **5**: 207ps214 DOI 10.1126/scitranslmed.3006305.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**:8817–8822 DOI 10.1073/pnas.1133470100.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**:2503–2505 DOI 10.1093/bioinformatics/btu314.

Field MA, Cho V, Andrews TD, Goodnow CC. 2015. Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies. *PLoS ONE* **10**:e0143199 DOI 10.1371/journal.pone.0143199.

Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, Dawson SJ, Piskorz AM, Jimenez-Linan M, Bentley D, Hadfield J, May AP, Caldas C, Brenton JD, Rosenfeld N. 2012. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Science Translational Medicine* **4**: 136ra168 DOI 10.1126/scitranslmed.3003726.

Fu GK, Hu J, Wang PH, Fodor SP. 2011. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences of the United States of America* **108**:9026–9031 DOI 10.1073/pnas.1017621108.

Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* **32**:158–168 DOI 10.1038/nbt.2782.

Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. 2013. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Research* **23**:843–854 DOI 10.1101/gr.147686.112.

Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America* **108**:20166–20171 DOI 10.1073/pnas.1110064108.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **108**:9530–9535 DOI 10.1073/pnas.1105422108.

Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. 2012. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**:72–74 DOI 10.1038/nmeth.1778.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**:2987–2993 DOI 10.1093/bioinformatics/btr509.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760 DOI 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079 DOI 10.1093/bioinformatics/btp352.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303 DOI 10.1101/gr.107524.110.

Ngo VN, Young RM, Schmitz R, Jhavar S, Xiao W, Lim KH, Kohlhammer H, Xu W, Yang Y, Zhao H, Shaffer AL, Romesser P, Wright G, Powell J, Rosenwald A, Muller-Hermelink HK, Ott G, Gascoyne RD, Connors JM, Rimsza LM, Campo E, Jaffe ES, Delabie J, Smeland EB, Fisher RI, Braziel RM, Tubbs RR, Cook JR, Weisenburger DD, Chan WC, Staudt LM. 2011. Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**:115–119 DOI 10.1038/nature09671.

Ottesen EA, Hong JW, Quake SR, Leadbetter JR. 2006. Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* **314**:1464–1467 DOI 10.1126/science.1131370.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biology* **14**:R51 DOI 10.1186/gb-2013-14-5-r51.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* **43**:e37 DOI 10.1093/nar/gku1341.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109**:14508–14513 DOI 10.1073/pnas.1208715109.

Vogelstein B, Kinzler KW. 1999. Digital PCR. *Proceedings of the National Academy of Sciences of the United States of America* **96**:9236–9241 DOI 10.1073/pnas.96.16.9236.

Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research* **40**:11189–11201 DOI 10.1093/nar/gks918.

Table S1: Variant caller commands

Variant caller	Version	Command
DeepSNVMiner	1.0	run_deepsnv.pl -read1_fastq R1.fq -read2_fastq R2.fq -coord_bed chr22.bed -filename_stub test
FreeBayes	1.0.2-6	freebayes -f chr22.fa bam_file > freebayes.vcf
GATK	3.2.2	1) java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R chr22.fa -L chr22.intervals -I bam_file -o gatk.gvcf -variant_index_type LINEAR -variant_index_parameter 128000 --minPruning 3 -ERC GVCF -contamination 0.0 --maxNumHaplotypesInPopulation 200 --max_alternate_alleles 3 2) java -jar -T GenotypeGVCFs -R chr22.fa -L chr.intervals -V gatk.gvcf-o gatk.vcf
LoFreq	2.1.2	lofreq call -f chr22.fa -o lofreq.vcf bam_file
SAMTools	0.1.19	samtools mpileup -C50 -uDEf chr22.fa bam_file | bcftools view -vcg - > sam.vcf

Variant caller commands utilized in our example. The commands listed match the exact commands run with the exception of the shortening of file names. The commands were chosen by either following documentation suggestions, or else by using default options.

Table S2: False positive rates for variant callers at increasing dilution levels

Dilution Percent	Total Variants	Deep-SNVMiner	FreeBayes	GATK	LoFreq	SAMTools
0	799962	0.014	0.16	34.05	0.18	31.94
50	408518	0.012	0.15	29.43	0.18	30.11
90	81708	0.012	0.17	28.22	0.21	26.09
99	8211	0.043	0.21	15.78	0.17	46.15
99.9	811	0.149	0.18	16.33	3.45	0
99.99	74	0	1.72	30.14	21.43	N/A
99.999	8	0	0	50.00	50.00	N/A
99.9999	2	0	N/A	0	N/A	N/A

False positive rates for DeepSNVMiner compared to FreeBayes, GATK, LoFreq, and SAMTools at increasing variant dilutions.

Table S3: False negative rates for variant callers at increasing dilution levels

Dilution Percent	Total Variants	Deep-SNVMiner	FreeBayes	GATK	LoFreq	SAMTools
0	799962	41.74	31.35	71.35	66.52	99.85
50	408518	29.76	31.79	63.48	68.40	99.89
90	81708	17.83	31.45	33.86	77.41	99.90
99	8211	15.59	32.03	18.32	78.32	99.91
99.9	811	17.39	31.94	17.26	79.28	99.88
99.99	74	9.46	22.97	31.08	70.27	100
99.999	8	0	25.00	25.00	50.00	100
99.9999	2	0	100	0	0.00	100

False negative rates for DeepSNVMiner compared to FreeBayes, GATK, LoFreq, and SAMTools at increasing variant dilutions.

Table S4: Dilution series for cell lines HEK293 and OCI-LY10

Sample	**HEK293 wt MYD88**	**OCI-LY10 L265P MYD88**
Sample1	0%	100%
Sample2	90%	10%
Sample3	99%	1%
Sample4	99.9%	0.1%
Sample5	99.99%	0.01%
Sample6	99.999%	0.001%
Sample7	99.9999%	0.0001%
Sample8	99.99999%	0.00001%
Sample9	99.999999%	0.000001%
Sample10	100%	0%

To measure the sensitivity of DeepSNVMiner a dilution series was performed with genomic DNA from two cells lines: (i) HEK293 (Human Embryonic Kidney): wild-type MYD88 (ii) OCI-LY10 (Ontario Cancer Institute, lymphoma cell line 10): heterozygous L265P MYD88 mutation.

6.2 Further discussion

This publication describes DeepSNVMiner, software to detect rare variants contained in sequence data from a heterogeneous population of cells. While the laboratory techniques for attaching unique identifiers to each DNA molecule prior to the amplification step have reached maturity, there is currently no available software to reliably detect rare variants that are present in sub-populations of cells. Existing standard variant detection tools are unable to work with mixed sequence data as the underlying statistical models typically expect homogeneous sequence data from a single population, an assumption that fails with a mixed population of cells. With most sequencing platforms yielding error rates in the order of ~1% (88), such tools are incapable of differentiating the real signal from the rare variants from the general background noise inherent in the data. Cancer variant detection software is more suitable for the task (given that its model expects sequence data from both tumour and normal cells) yet is still unable to detect variants in very highly heterogeneous sequence data due to the thousands of cells contained in the mixed sequence data. Given the lack of readily available software solutions for this increasingly important sequencing application, DeepSNVMiner was initially written to support an internal project trying to detect rare variants within mixed cell populations of Sjogren syndrome patients, with the software later generalised and released as an open-source tool.

Internally, DeepSNVMiner is a complete analysis pipeline for detecting variants in mixed sequence data sets and implemented to output completely reproducible bioinformatic workflows with detailed logging occurring from the onset. The workflow consists of five major components; FASTQ checking, barcode grouping, alignment to the reference genome, variant detection, and reporting. DeepSNVMiner works by accepting raw FASTQ sequence files with UIDs, extracting the unique identifier from the sequence column, and generating a new filtered FASTQ file with the identifier removed from the sequence row and added to the header row. The filtered FASTQ files are next aligned to the reference genome using BWA (76), reads are grouped by common UID, and variants called using SAMtools (89) fillmd on each group of reads in user specified genomic regions only. Variant detection output is next parsed, variants in read groups are identified, summary reports are generated, and plots generated

using the software package R (90).

The software is designed to be extremely flexible with regard to the nature of the input sequence data, an important feature given the increasing diversity of laboratory techniques available for generating data sets of this nature. Users are able to precisely define the unique identifying barcode (UID), whether this barcode consists of sequence from the 5′ end, the 3′ end, or some combination of sequence from both ends. The software is also flexible with regard to defining exactly what a variant looks like within each group of reads sharing a UID. Users are able to define the minimum number of reads required to constitute a read group and further define what fraction of reads must contain the same variant base to be considered a true rare variant. The flexibility in defining both what a UID looks like and what constitutes a variant in each UID group allows users to customise input parameters in such a way that the software is able to effectively detect rare variants in their unique data sets.

While grouping reads by common UID and searching for variants within each read group ultimately allow the detection of rare variants, the data sets presented several technical challenges needing custom solutions. First, given the huge number of common UID groups within each data set (typically 10000s of unique read groups), traditional variant detection for each group became computationally prohibitive and was not feasible without the use of significant HPC resources. To remedy this, DeepSNVMiner incorporated the very fast SAMtools view/fillmd command, a command capable of identifying bases not matching the reference genome from a single read. While not designed to detect variants per se, by parsing the command's output, grouping reads by common UID, and searching for variants within each group, I was able to reduce the run time by three orders of magnitude, essentially making the software able to run in a standard desktop environment. The other major technical challenge to solve was how to effectively search only the genomic regions of interest; as such data sets target a restricted set of gene targets. For this, users are able to pass in a bed file containing the genomic regions of interest meaning the run time scales linearly with the size of the genomic regions to search. By limiting the search space to the regions of interest, run time is further minimized compared to exhaustively searching the entire genome.

In real control cell line data sets, DeepSNVMiner is able to detect

variants as scarce as 1 in 1000 molecules present in input material, something no existing publicly available software is able to do. Using simulated data sets, DeepSNVMiner does even better and detects variants present in 1 in 1000000 molecules, a feat other standard variant callers are unable to do. The lack of available computational tools for analysing such heterogeneous sequence data has represented an obstacle to the widespread adoption of this technology, an obstacle which DeepSNVMiner aims to remove. With the widespread adoption of this tool, researchers will hopefully be able to utilise this technology to gain a greater understanding of disease progression within mixed populations of cells.

Chapter 7: Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models

Andrews, T. D., B. Whittle, **M. A. Field**, B. Balakishnan, Y. Zhang, Y. Shao, V. Cho, M. Kirk, M. Singh, Y. Xia, J. Hager, S. Winslade, G. Sjollema, B. Beutler, A. Enders and C. C. Goodnow. "Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models." Open Biology. 2012; 2(5): 120061.

open Biology

rsob.royalsocietypublishing.org

Research

Cite this article: Andrews TD, Whittle B, Field MA, Balakishnan B, Zhang Y, Shao Y, Cho V, Kirk M, Singh M, Xia Y, Hager J, Winslade S, Sjollema G, Beutler B, Enders A, Goodnow CC. 2012 Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models.
Open Biol 2: 120061.
http://dx.doi.org/10.1098/rsob.120061

Received: 7 March 2012
Accepted: 16 April 2012

Subject Area:
genomics/bioinformatics/immunology/biotechnology

Keywords:
exome sequencing, DNA capture, *N*-ethyl-*N*-nitrosourea mutagenesis, variation detection, mutation detection, mouse

Author for correspondence:
T. D. Andrews
e-mail: dan.andrews@anu.edu.au

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsob.120061.

Royal Society Publishing
Informing the science of the future

Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models

T. D. Andrews[1,3], B. Whittle[3], M. A. Field[1,3], B. Balakishnan[3], Y. Zhang[3], Y. Shao[1,3], V. Cho[1,3], M. Kirk[1,3], M. Singh[2], Y. Xia[4,5], J. Hager[6], S. Winslade[3], G. Sjollema[3], B. Beutler[4,5], A. Enders[2] and C. C. Goodnow[1]

[1]Immunogenomics Laboratory, and [2]Ramaciotti Immunisation Genomics Laboratory, John Curtin School of Medical Research, Australian National University, GPO Box 334, Canberra City, Australian Capital Territory, 2601, Australia
[3]Australian Phenomics Facility, Australian National University, Hugh Ennor Building, Building 117, Garran Road, Canberra City, Australian Capital Territory, 0200, Australia
[4]Center for the Genetics of Host Defense, University of Texas Southwestern, 6000 Harry Hines Boulevard, Dallas, TX 75930-8505, USA
[5]Department of Genetics, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA
[6]Centre National de Génotypage, 2 rue Gaston Crémieux, CP5721, 91057, Evry Cedex, France

1. Summary

Accurate identification of sparse heterozygous single-nucleotide variants (SNVs) is a critical challenge for identifying the causative mutations in mouse genetic screens, human genetic diseases and cancer. When seeking to identify causal DNA variants that occur at such low rates, they are overwhelmed by false-positive calls that arise from a range of technical and biological sources. We describe a strategy using whole-exome capture, massively parallel DNA sequencing and computational analysis, which identifies with a low false-positive rate the majority of heterozygous and homozygous SNVs arising de novo with a frequency of one nucleotide substitution per megabase in progeny of *N*-ethyl-*N*-nitrosourea (ENU)-mutated C57BL/6j mice. We found that by applying a strategy of filtering raw SNV calls against known and platform-specific variants we could call true SNVs with a false-positive rate of 19.4 per cent and an estimated false-negative rate of 21.3 per cent. These error rates are small enough to enable calling a causative mutation from both homozygous and heterozygous candidate mutation lists with little or no further experimental validation. The efficacy of this approach is demonstrated by identifying the causative mutation in the *Ptprc* gene in a lymphocyte-deficient strain and in 11 other strains with immune disorders or obesity, without the need for meiotic mapping. Exome sequencing of first-generation mutant mice revealed hundreds of unphenotyped

protein-changing mutations, 52 per cent of which are predicted to be deleterious, which now become available for breeding and experimental analysis. We show that exome sequencing data alone are sufficient to identify induced mutations. This approach transforms genetic screens in mice, establishes a general strategy for analysing rare DNA variants and opens up a large new source for experimental models of human disease.

2. Introduction

Genetic traits in mammals have long posed a great challenge in connecting them to their causal DNA variant. This is especially true when that variant is a single-nucleotide substitution and is present on only one of the two copies of a chromosome. Finding such a single-nucleotide substitution in a genome as large as humans or mice without huge numbers of false positives and without reducing the search to a sub-chromosomal region by meiotic mapping has been an unattainable goal. Single-nucleotide variants (SNVs) represent a major source of de novo and inherited genomic variation in humans, mice and other mammals, and, as such, new strategies are needed to identify and analyse these variants accurately on a genome-wide scale.

Genetic analyses of mammalian traits are often performed in inbred C57BL/6 laboratory mice. These mice have a known homogeneous reference genome sequence and have a uniform genetic background that allows experimental reproducibility and transplantation experiments. In these mice, treatment with the chemical mutagen *N*-ethyl-*N*-nitrosourea (ENU) efficiently generates random single-base mutations in the germline DNA (reviewed in [1]). Diseases and traits resulting from these ENU-induced mutations can be detected by phenotypic screening procedures relevant to an area of biological investigation.

The bottleneck of the ENU mutagenesis approach has long been in identifying a single disease-causing mutation in an entire genome of possibilities. Until recently, the approach employed has been arduous: to out-cross affected mice to another inbred strain and then use a panel of common strain-specific variants to meiotically map the causal mutation to a sub-region of an individual chromosome of less than 20 megabases (Mb). Once limited to a relatively short list of positional candidate genes, PCR amplification of all exons in the mapped interval followed by Sanger sequencing could then be performed and variants identified by a combination of automated and manual review of the sequence traces. This has proven to be an effective strategy, although it can take several years and is labour-intensive, expensive and often confounded by modifier genes introduced during the cross to another inbred strain.

To date, all but the smallest minority of causative ENU-induced mutations have been shown to reside in the exonic portion of the genome. Approximately 75 per cent are caused by SNVs in protein-coding exons that result in missense or nonsense mutations and the remaining approximately 25 per cent are SNVs in splice donor–acceptor sites that disrupt correct mRNA splicing to cause protein truncations, deletions or nonsense-mediated decay [2]. Hence, sequencing of the exome rather than the whole genome should identify almost all interesting ENU-induced variants. Array- and solution-based DNA capture technologies [3,4] can now reliably enrich a DNA sample for coding regions, enabling massively parallel sequencing to be undertaken on a greatly reduced proportion of the genome. Exome capture followed by sequencing has already become an established technique in human genetics and an early vanguard of reports has identified the genetic cause of a number of monogenic diseases (reviewed in [5]). In most of these studies, prior information regarding a general chromosomal location of the genetic lesion was known, heritability information was available or a candidate gene approach was used. One feature of all of these studies was the difficulty in discerning causative, deleterious mutations from normal genetic variation and sequencing errors.

In the mouse, early studies [6–8] using slightly different approaches have identified ENU-induced mutations using massively parallel sequencing information. Zhang *et al.* [8] identified a previously known ENU-induced mutant by sequencing cloned bacterial artificial chromosomes from a 2.2 Mb genomic region that had first been defined by meiotic mapping. Arnold *et al.* [6] applied shallow sequencing of the entire mouse genome to detect putative mutations and, following this, they performed extensive validation by Sanger sequencing and meiotic mapping. Yabas *et al.* [7] mapped a novel ENU mutation to a region of the X-chromosome, and identified the mutation by oligonucleotide bait-mediated capture and deep sequencing of exonic DNA fragments within this region. Fairfield *et al.* [9] provided an extensive demonstration of the utility of exome capture technology for identifying both homozygous and heterozygous ENU-induced and spontaneous mutations in nine mouse strains. However, in all cases these studies relied on at least coarse meiotic mapping information or considerable validation of SNV calls to identify the causative mutation. Fairfield *et al.* [9] suggest that an exome sequence as a sole source of information may not be enough to identify disease-causing induced mutations without extensive SNV validation.

In this study, we have investigated whether exome capture followed by sequencing provides sufficient information alone to reliably identify the rare, ENU-induced, de novo mutations in C57BL/6j mice. We generated exome datasets for 12 mutant mouse strains, including a matched technical and biological replicate dataset for one strain. We present methodology developed to identify both homozygous and heterozygous ENU-induced mutations and use this to identify 12 primary causative mutations and two disease-causing incidental mutations. We also reveal hundreds of potentially deleterious ENU mutations in first-generation (G1) mice that are immediately available for phenotypic and experimental analysis in their progeny. Our results demonstrate that exome sequencing provides highly reliable information which by itself is sufficient to identify ENU-induced mutations selected either by phenotype or by the nature of the gene that is mutated. These results provide an immediate source for thousands of new experimental models for understanding human diseases and establish a strategy that can be extended for identifying rare SNVs in outbred mice, humans and other species.

3. Results

3.1. Generation and detection of induced, de novo single-nucleotide variants

Many parallel mouse pedigrees, each segregating a different set of random, de novo mutations induced in the C57BL/6j

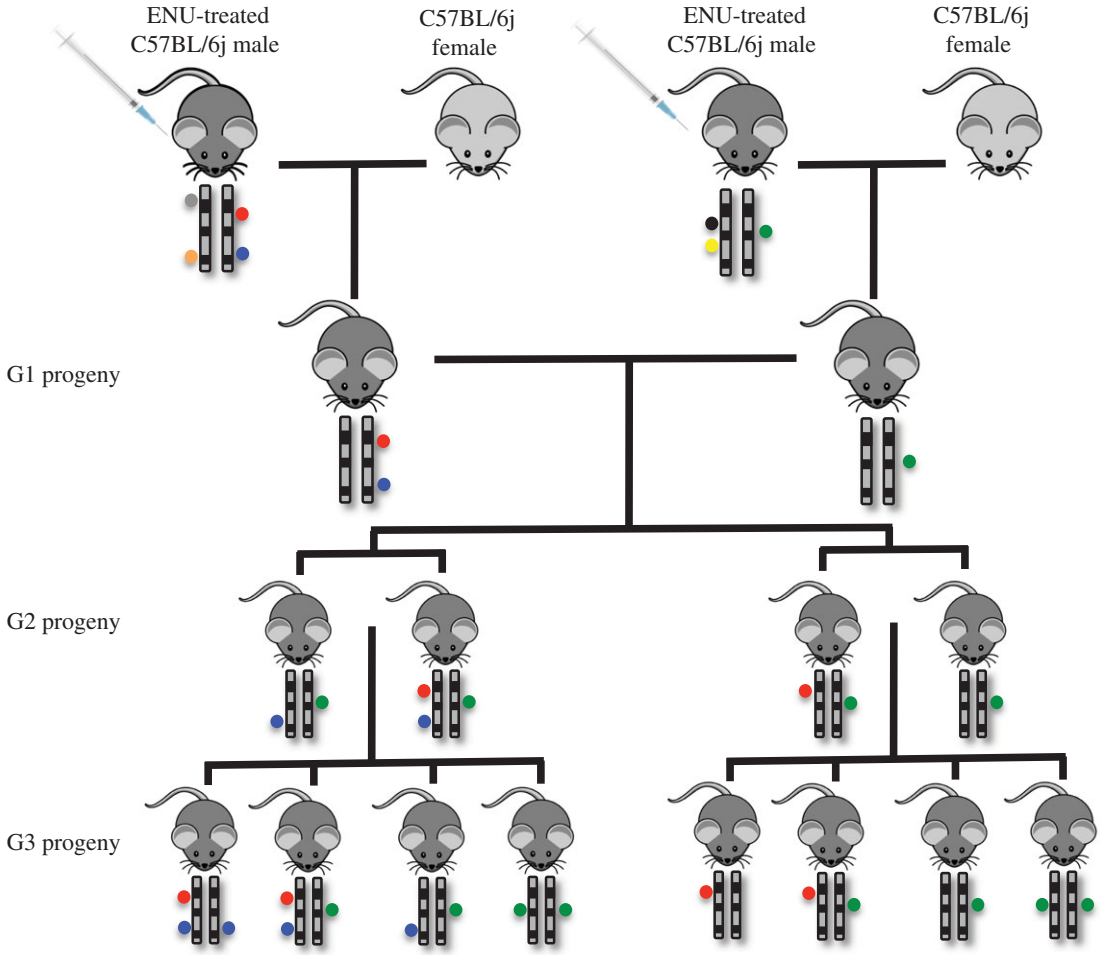


Figure 1. Summary of the structure of ENU-mutated mouse pedigrees. Each pedigree is initiated by two unrelated G1 founders. Each of these founders inherits a random set of de novo point mutations (coloured circles) on the paternal chromosomes, induced by ENU treatment of their male parent. These G1 founders will carry on average one to two DNA variants per Mb and 90 exonic ENU-induced mutations. Second-generation (G2) progeny of these mice inherit a theoretical 45 ENU-induced exonic mutations, all of which are carried in the heterozygous state. Two productive sibling–sibling matings of the G2 mice result in third-generation (G3) progeny that carry approximately 94% of the founding ENU-induced, protein-coding mutations, of which on average five are homozygous in any given mouse.

genome by ENU were established using the breeding strategy shown in figure 1. Each pedigree was founded by two unrelated G1 mice conceived from male C57BL/6j mice that had been treated with three doses of ENU administered at 90 mg kg^{-1} to induce random point mutations in spermatogonial stem cells [2,10,11]. Based on published mutation rates [12–14], we estimated that each of these G1 animals would carry approximately one de novo SNV per Mb of the paternal genome, of which around 45 would result in a non-synonymous exonic mutation. Intercrossing of the G1 animals transmitted half of these mutations in heterozygous state to each of their second-generation (G2) offspring. Intercrossing the G2 animals subsequently transmitted approximately 94 per cent of the mutations to offspring, a subset of which was inherited in homozygous state in third-generation (G3) animals (figure 1).

We developed a workflow (figure 2*a*) to use massively parallel sequencing reads as a sole data source to identify exonic ENU-induced mutations in 15 DNA samples taken from mutated mice (see electronic supplementary material, table S1). These samples were prepared and enriched for exonic sequences using either Agilent or Nimblegen solution-based capture technologies. Each exome sample was then sequenced as paired-end reads in a full lane of an Illumina GAIIx sequencer or as a multiplexed, bar-coded sample in an Illumina HiSeq sequencer, and the resultant reads aligned to the C57BL/6 mouse reference genome using the BWA aligner [16]. Table S1 in the electronic supplementary material shows the numbers of reads sequenced and the number of reads aligned to exonic target regions per sample. The exome capture efficiency was uniformly high with approximately 40 to 55 per cent of all DNA sequenced being exonic. Based on a mouse genome size of 2493 Mb [15] and 37 Mb of exonic sequence, using consensus coding sequence (CCDS) exons [18], this represents on average a 30.6-fold ($\sigma = 3.3$) sequence enrichment. Across the coding portion of the genome sequence, coverage was generally better than 85 per cent at 5 times depth and better than 70 per cent at 20 times depth, although coverage was distinctly less for the sex chromosomes (see electronic supplementary material, figure S1).

Raw SNVs relative to the C57BL/6 reference sequence were called using SAMTools [17]. In the inbred C57BL/6j mice we analysed, we would expect the number of true variant calls to be low (approx. 50 exonic SNVs) and almost entirely due to ENU treatment of the G0 male mouse that founded their line. However, in each animal, of the order of 10 000 raw SNVs were called across the entire genome, of which 500–750 SNV calls were located in exons and/or near exon splice sites (see electronic supplementary material, table S2). Multiple sources can be attributed to these variant calls, potentially being due to genetic drift of the C57BL/6j

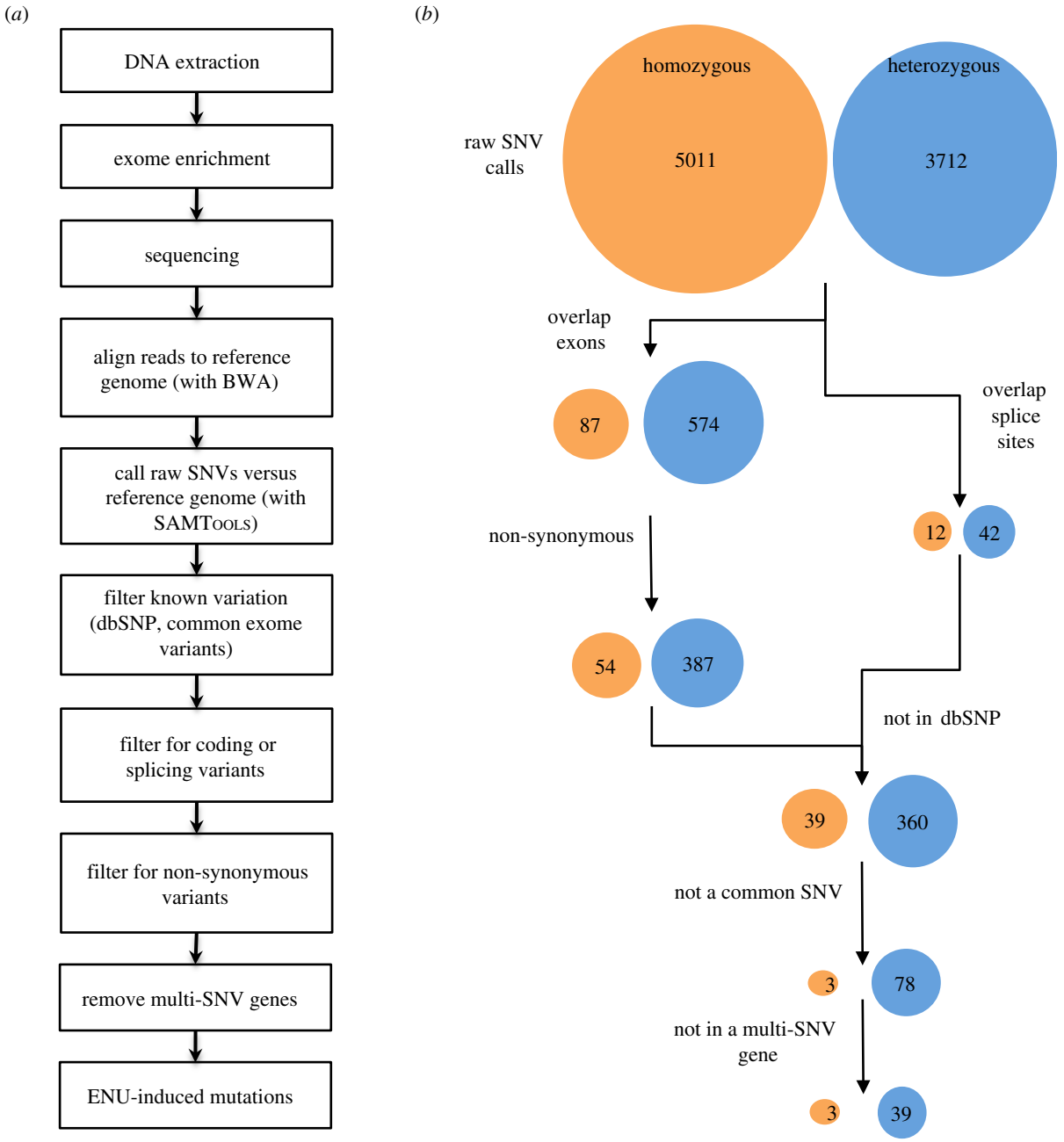


Figure 2. Workflow and filtering strategy used to identify de novo protein-changing mutations. (*a*) Following DNA extraction, exome enrichment and sequencing, reads were aligned to the mouse reference genome [15] using BWA [16] and variation between the two genomes identified using SAMTOOLS [17]. The set of raw SNVs was subsequently filtered to annotate known variation and other apparent SNVs known not to be ENU-induced. SNVs were further filtered to annotate those that fell within coding regions (or adjacent splice donor/acceptor sites) and were non-synonymous changes. Finally, as ENU treatment is known to introduce a uniform genomic distribution of mutations, genes that contained multiple SNVs were filtered from the final set of variants. (*b*) Using this cumulative filtering strategy against a single replicate exome sequence of the *nimbus* mouse, the initial 8723 variant calls reduced to a final set of three homozygous and 39 heterozygous putative mutations. Circles representing homozygous and heterozygous SNV numbers are coloured orange and blue, respectively.

mouse strain versus the reference genome and the frequency of sequencing errors in massively parallel sequencing. However, many of the variants appear to be called because of technical issues associated with aligning large numbers of short reads to a large genome containing repeated or highly similar sequence regions. To reduce these raw variant calls to a smaller number highly enriched for ENU-induced mutations, we applied a series of filters to remove known variants (present in dbSNP) and/or recurrent false-positive variants (figure 2*a*). We assert that between multiple, unrelated mouse exome sequences, de novo ENU-induced nucleotide changes should be unique to individual pedigrees, whereas other sources of variants should recur. Based on this reasoning, we collated a list of SNVs that recurred in more than one unrelated mouse and found this list to be a very effective filter for false-positive and potentially sequencer- and enrichment-specific variants. A further filter was applied to remove variants where they originate from a gene with multiple SNV calls, assuming that in any single ENU-mutated mouse it is highly unlikely that the same gene will have multiple mutations and that the calls are due to incorrect alignment of sequence reads between members of gene families. Figure 2*b* shows the efficacy of each individual filtering step applied and the outcome of the filters applied in a

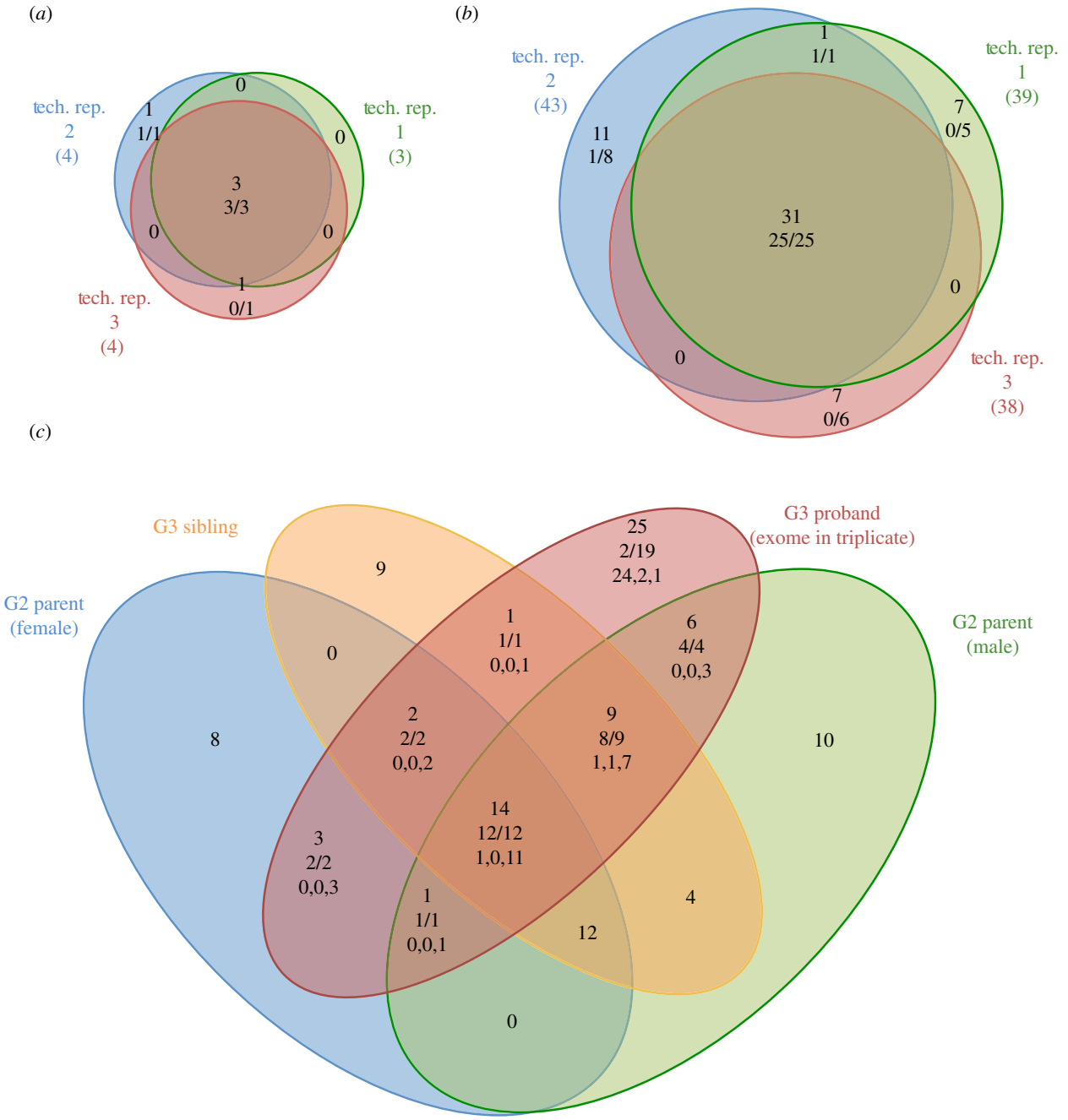


Figure 3. Sensitivity and specificity of mutation detection in the *nimbus* mutant mouse pedigree assessed through technical and biological replicate datasets. Venn diagrams of overlap of filtered variant calls between three technical replicate exome sequence datasets, showing putative (*a*) homozygous and (*b*) heterozygous ENU-induced mutations. The red, green and blue circles each indicate separate technical replicates, and the coloured numbers associated with each denote the total number of variants called in each dataset. Upper numbers within each sector show the number of filter-passing SNVs called in one, two or all three technical replicates. The numbers below show the fraction of these SNVs that were validated as true mutations by independent, custom, SNV-specific PCR assays. The denominator in each case is the number of SNVs where an SNV-specific PCR assay was established successfully. (*c*) Overlap of filtered variant calls from a set of four biological replicates, representing two parental G2 *nimbus* mice and two of their G3 offspring. One of the G3 offspring (labelled G3 proband) is the same mouse as that sequenced in the technical replicates shown in (*a*) and (*b*). The variant numbers shown for this mouse are pooled values from the three technical replicates. Both G2 *nimbus* mice and the sibling of the G3 proband (labelled G3 sibling) are unaffected by the lymphopaenia phenotype. Upper numbers within each sector of the four-way Venn diagram show the total number of filter-passing heterozygous and homozygous SNVs called in one or more of the replicates from this pedigree. The numbers immediately below show the fractions of biologically replicated SNVs that were validated as true mutations by independent, custom, SNV-specific PCR assays. In the case of technically replicated data from the proband (the red circle), the third line of data in each region of overlap shows the number of times a variant was seen in one, two or three replicates (formatted as: single count, double count and triple count).

cumulative manner. Overall, from a set of several thousands of raw variant calls, the cumulative filtering reduced this number mostly to less than 10 homozygous and 50 heterozygous exonic variants per mouse (see electronic supplementary material, table S2), closely approximating the expected rate (figure 1).

3.2. Sensitivity and specificity of single-nucleotide variant detection

To assess the reliability of SNV calls made from a single exome dataset, we performed a technical and biological replication experiment on G2 and G3 animals from a pedigree

Table 1. Validation results of all SNVs detected in proband replicate exome sequences. chr, chromosome; coord, coordinate; het, heterozygous; hom, homozygous; wt, wild type.

chr	coord	hom/het	wt genotype	genotype in G3 proband	genotype in G2 parent (female)	genotype in G2 parent (male)	genotype in G3 sibling	validated	number of replicates SNV detected	gene	amino acid change	PolyPhen2 score
1	6264360	het	C/C	C/T	C/C	C/T	C/C	yes	3	Rb1cc1	splice	—
1	59115542	het	A/A	A/T	A/T	A/T	T/T	yes	3	Als2cr11	S→T	0.126
1	139986182	hom	C/C	T/T	C/T	C/T	C/T	yes	3	Ptprc	splice	—
1	153543280	hom	T/T	A/A	T/A	T/A	T/A	yes	3	Fam129a	V→D	0.998
1	155590911	het	G/G	G/A	G/G	G/A	G/G	yes	3	Rgs16	D→N	0.812
2	13387686	hom/het	C/C	C/T	C/C	C/C	C/C	yes	2	Cubn	splice	—
2	14210637	het	A/A	A/T	A/A	A/T	A/T	yes	3	Mrc1	M→L	0.725
2	26263680	het	T/T	T/C	T/C	T/C	T/C	yes	3	Inpp5e	E→G	0.008
2	49590962	het	G/G	G/T	G/T	G/T	G/T	yes	3	Kif5c	V→L	0
2	70417236	het	A/A	A/G	A/G	A/G	A/G	yes	3	Gad1	T→A	0.512
2	89592814	het	A/A	A/T	A/T	A/A	A/A	yes	3	Olfr1253	M→K	0.988
3	19910826	het	C/C	C/A	C/C	C/A	C/A	yes	3	Hps3	V→L	0
3	88347367	het	A/A	A/G	A/A	A/G	A/G	yes	3	Rab25	V→A	0.098
3	108230315	het	A/A	A/T	A/A	A/A	A/T	yes	3	Sars	splice	—
4	15925782	het	A/A	A/T	A/T	A/A	A/T	yes	3	Osgin2	M→K	0.045
4	140271822	het	A/A	A/G	A/G	A/A	A/A	yes	3	Rcc2	I→V	0.178
6	34974302	het	A/A	A/G	A/G	A/G	G/G	yes	3	Cnot4	W→R	0.991
6	56952536	het	A/A	A/T	A/T	A/T	A/T	yes	1	Vmn1r6	D→V	0.028
6	116597609	het	A/A	A/T	A/T	T/T	A/A	yes	3	Rassf4	V→E	0.069
6	124882799	het	G/G	G/T	G/G	G/G	G/G	yes	3	Mlf2	G→V	0.999
9	108817748	het	T/T	T/C	T/T	T/C	T/T	yes	3	Tmem89	V→A	0.860
11	70159403	het	T/T	T/A	T/T	T/A	T/T	yes	3	Alox15	D→V	0.871
11	100051095	hom	T/T	C/C	T/C	T/C	T/C	yes	3	Krt9	K→E	0.144
11	120576524	het	T/T	T/A	T/A	T/T	T/A	yes	3	Lrrc45	S→T	0
13	34010014	het	T/T	T/A	T/T	T/A	T/A	yes	3	Serpinb6a	M→L	0
13	41141476	het	T/T	T/C	T/T	T/C	T/C	yes	3	Mak	N→S	0.001
14	99585815	het	A/A	A/G	A/G	A/G	G/G	yes	3	Pibf1	K→R	0.953

(*Continued.*)

Table 1. (*Continued.*)

chr	coord	hom/het	wt genotype	genotype in G3 proband	genotype in G2 parent (female)	genotype in G2 parent (male)	genotype in G3 sibling	validated	number of replicates SNV detected	gene	amino acid change	PolyPhen2 score
15	88956462	het	T/T	T/A	T/A	T/A	T/A	yes	3	Hdac10	Q⟶L	0.396
15	101398409	het	T/T	T/C	T/T	T/C	T/C	yes	3	Krt75	T⟶A	0.998
17	37362339	het	G/G	G/A	G/G	G/A	G/A	yes	3	Olfr96	E⟶K	0.001
17	37436474	het	T/T	T/C	T/T	T/C	T/C	yes	2	Olfr101	S⟶G	0
1	3661021	het	G/G	G/T	G/G	G/G	G/G	no	1	Xkr4		
1	26744177	het	A/A	A/G	A/A	A/G	A/G	no	1	4931408C20Rik		
4	43429551	het	A/A	A/C	A/A	A/A	A/A	no	1	Rusc2		
5	14934071	hom	G/G	C/C	G/G	G/G	G/G	no	1	RP23-239L21.1		
7	13629965	het	G/G	G/A	G/G	G/G	G/G	no	1	Mzf1		
7	66046516	het	C/C	C/A	C/C	C/C	C/C	no	1	Atp10a		
7	136751431	het	A/A	A/C	A/A	A/A	A/A	no	1	Wdr11		
9	40703661	het	C/C	C/A	C/C	C/C	C/C	no	1	4931429I11Rik		
10	36717792	het	C/C	C/A	C/C	C/C	C/C	no	1	Hdac2		
10	57861777	het	C/C	C/A	C/C	C/C	C/C	no	1	Lims1		
11	61265622	het	G/G	G/C	G/G	G/G	G/G	no	1	Rnf112		
11	69717525	het	C/C	C/A	C/C	C/C	C/C	no	1	Neurl4		
11	102527771	het	C/C	C/A	C/C	C/C	C/C	no	1	Gm1564		
12	21316065	het	C/C	C/G	C/C	C/C	C/C	no	1	Cpsf3		
14	65377451	het	G/G	G/T	G/G	G/G	G/G	no	1	Kif13b		
15	96846616	het	C/C	C/A	C/C	C/C	C/C	no	1	Slc38a4		
18	21288527	het	C/C	C/A	C/C	C/C	C/C	no	1	Fam59a		
X	121242115	het	A/A	A/G	A/A	A/A	A/A	no	1	Vmn2r121		
X	121246252	het	T/T	T/G	T/T	T/T	T/T	no	1	Vmn2r121		
1	56954912	het	A/A	A/G	A/G	A/G	G/G	no data	3	Satb2	C⟶R	0.537
1	74442070	het	A/A	A/C	A/A	A/A	A/A	no data	1	Ctdsp1	E⟶A	0.974
1	145475019	het	G/G	G/A	G/G	G/G	G/G	no data	1	Cdc73	splice	—
2	50148343	het	G/G	G/A	G/A	G/A	G/A	no data	3	Mmadhc	A⟶V	0
2	91831062	het	T/T	T/C	T/T	T/T	T/T	no data	3	Creb3l1	splice	—

(*Continued.*)

Table 1. (*Continued.*)

chr	coord	hom/het	wt genotype	genotype in G3 proband	genotype in G2 parent (female)	genotype in G2 parent (male)	genotype in G3 sibling	validated	number of replicates SNV detected	gene	amino acid change	PolyPhen2 score
2	161523850	het	C/C	C/G	C/C	C/C	C/C	no data	1	Ptprt	splice	—
4	40933721	het	T/T	T/C	T/T	T/T	T/T	no data	1	Nfx1	splice	—
4	137106002	het	A/A	A/T	A/T	A/A	A/A	no data	3	Hspg2	T→S	0.024
6	113233356	het	G/G	G/A	G/A	G/A	G/A	no data	3	Cpne9	E→K	0.987
9	61783268	het	A/A	A/G	A/A	A/G	A/A	no data	3	Kif23	V→A	0.999
10	78045769	het	G/G	G/T	G/G	G/G	G/G	no data	1	Ilvbl	G→V	0.407
19	39808875	het	T/T	T/C	T/T	T/C	T/T	no data	3	Cyp2c68	I→V	0

(*nimbus*) that had shown mild lymphopaenia in the blood of some G3 offspring. These *nimbus* mutant animals displayed a fourfold reduction in the percentage of $CD3^+$ T cells and represented 8 of a total of 30 phenotyped individuals, suggesting that *nimbus* was a recessive trait. We sequenced the exome of one proband G3 affected *nimbus* mouse in triplicate (technical replicates) and also sequenced the exome of both G2 parents and an unaffected G3 sibling (loosely termed biological replicates). Figure 3*a,b* shows that the SNVs called in each of the technical replicates of the proband's exome were highly replicable. The total number of coding changes called in each replicate was 47, 42 and 42, of which 34 were called in all three replicates, representing 72, 81 and 81 per cent of the SNVs called in each individual exome analysis. The triplicated SNV calls comprised three homozygous and 31 heterozygous mutations. We successfully established custom, SNV-specific PCR assays (Amplifluor assays; see §5.4) for 50 of the SNVs called in one or more of these replicates. From 50 successful assays, 100 per cent (28 of 28) of the triplicated SNV calls were validated as true mutations in this pedigree, whereas of the SNV calls that were present in only one or two of the replicate analyses only 14 per cent (3 of 22) were validated and the remainder were established to be false positives (figure 3*a,b* and table 1). From these technical replicate data the false-positive call rate among our filtered variants can be estimated as 19.4 per cent, calculated from an average of six false-positive calls per replicate exome as a proportion of the 31 true-positive SNVs.

In mouse spermatogonial stem cells and the mice conceived from the resulting sperm, ENU has been found to induce a biased set of nucleotide substitutions. Several previous studies have shown an abundance of TA–CG transitions and TA–AT transversions (ranging between 36–43% and 22–44% of changes, respectively [2,12,14,19]) and GC–CG transversions very rarely or never occur [14]. Of the validated 31 true-positive SNVs shown in table 1, 35.5, 38.7 and 0 per cent were TA–CG, TA–AT and GC–CG changes, respectively. Of the remaining 19 non-replicated, false-positive SNV calls, 26.3, 0 and 15.8 per cent were TA–CG, TA–AT and GC–CG changes, respectively.

Exome analysis of the G2 parents of our G3 *nimbus* proband mouse would be expected to reveal all the true ENU variants present in the proband mouse. Likewise, approximately half of the true variants should have also been inherited by the G3 sibling of the proband. Figure 3*c* shows a Venn diagram detailing the overlap between the SNVs called in the exome sequence of the two parents and sibling compared with those in the pooled technical replicate exome sequence of the proband. As expected, all of the seven homozygous mutations called in the proband or its sibling (three in proband + four in G3 sibling) were also called in heterozygous state in both parents. Of the total of 31 validated mutations present in the G3 proband, 28 were called in one or both parents (table 1). Inspection of the sequence data for the two parent G2 exomes revealed that the false-negative mutations were present, but the number of variant reads fell below the required coverage and/or read ratio thresholds used for SNV calling. That three of the 31 true mutations were not identified in one or more of the replicate analyses indicates a technical false-negative rate of 9.7 per cent per exome analysis. However, this estimate does not accommodate

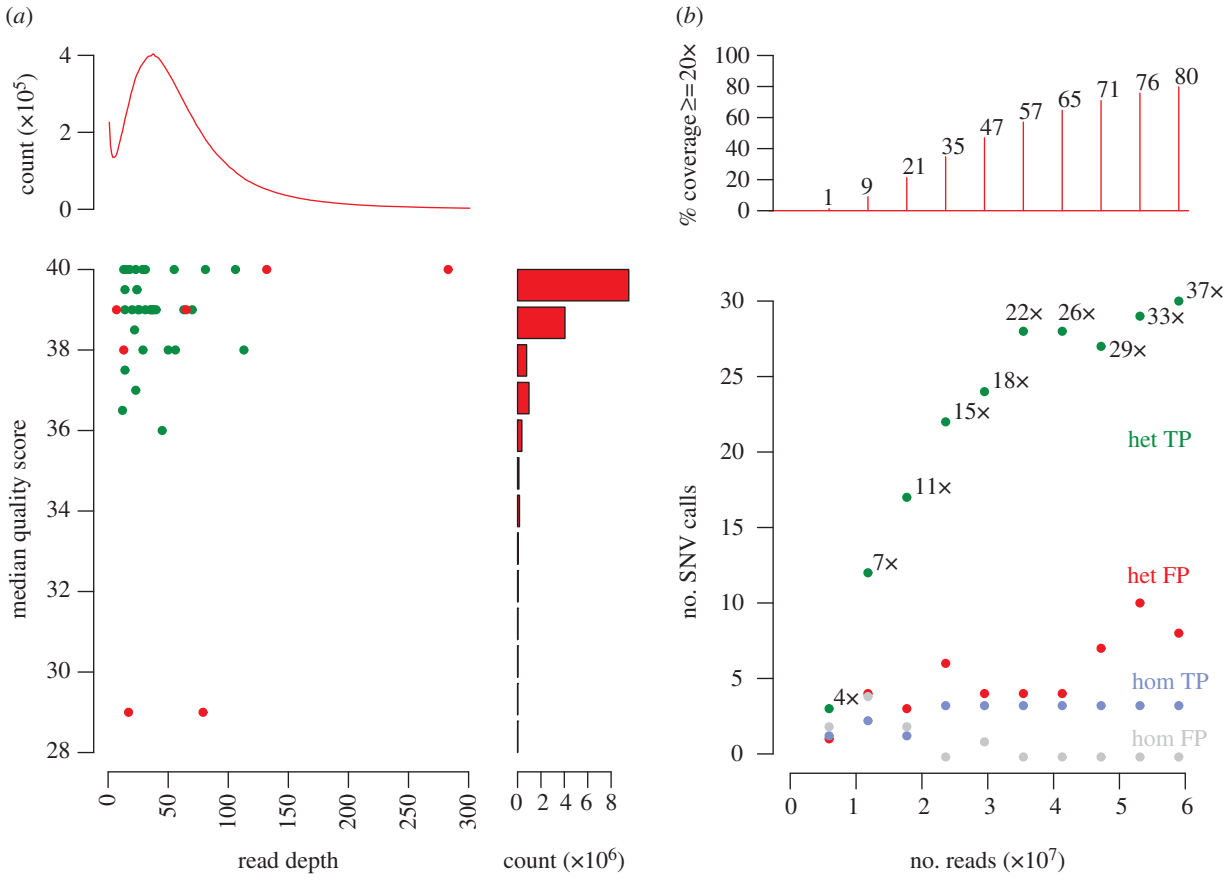


Figure 4. The influence of sequence quality scores and read depth on the identification of true-positive and false-positive SNVs. (*a*) False-positive calls with respect to read depth and quality score, shown for a single exome dataset generated from the G3 *nimbus* mouse (technical replicate 1 from figure 3). Variant calls on this dataset were compared with the PCR-validated true-positive and false-positive SNVs called in the technical replicate exome datasets of the G3 *nimbus* proband. Green and red points are true- and false-positive SNV calls, respectively. The distribution of read depth frequencies over all exonic bases is indicated by the red line in the top graph. The red bars in the right-hand graph indicate the distribution of quality scores also ascertained for all exonic bases. (*b*) Results of simulation experiment performed to generate random subsets of a single exome dataset, being one of the triplicate exome runs for the *nimbus* proband (technical replicate 1). The panel shows tallies of true-positive heterozygous (green), false-positive heterozygous (red), true-positive homozygous (blue) and false-positive homozygous (grey) SNV calls plotted against the number of input reads, which are incremental proportions of an Illumina GAIIx lane. Numbers alongside the green dots indicate the median read depth determined for each true-positive data point. Plotted above are the proportions of the exome covered at 20× depth or better for each proportion of the input read set.

the percentage of true mutations that might be missed consistently because they lie in exons that are inefficiently captured and exhibit low sequence coverage.

The false-negative rate of SNV detection can also be estimated from the distribution of sequence read depths generated at random across a whole genome. The depth of reads obtained from random short-read sequencing approximates a Poisson distribution and the probability of observing both alleles at a single site in a diploid genome is a binomial function of read depth [20,21]. A combination of these two distributions can be used to estimate the false-negative SNV call rate based simply on the mean read depth [20–22]. While the distribution of read depths obtained from exome capture appears to be mostly Poisson distributed, this approximation does not hold for sites that are poorly enriched by hybridization to exomic baits (see electronic supplementary material, figure S2), which are also the sites where low coverage is likely to result in the greatest incidence of false-negative calls. In order to estimate an accurate false-negative SNV call rate we used the observed distribution of sequence read depths rather than that derived from a Poisson function. In this manner, we calculated the false-negative SNV call rate in the *nimbus* G3 proband mouse as 21.3 per cent. The average read depth from this dataset was 39.5, but 14.6 per cent of CCDS exomic bases were not covered at all, this being the major source of missing SNV calls. Increasing the amount of sequence data does reduce the false-negative rate slightly, but still a large number of genomic sites will remain poorly covered, either owing to it being difficult to design capture baits to these regions or owing to extreme GC content reducing the efficiency of hybridization of some areas of the genome (data not shown).

Taking the SNV calls from a single replicate exome from the *nimbus* proband G3 mouse, we investigated whether or not validated true- and false-positive SNVs differed in sequence coverage or quality. Figure 4*a* shows that false-positive SNVs had unusually high or low read depth, or had lower quality scores, relative to the depth and quality of reads across all exonic nucleotides. However, in these data the read depths and quality scores of false-positive variants overlap with those of true-positive calls. While we have chosen to minimize the false-negative rate as much as possible, if it were desirable to reduce the false-positive call rate at the expense of the false-negative rate, this could potentially be achieved with more stringent filtering against read depth and quality score.

To evaluate how deeply an exome should be sequenced, we simulated an exome sequencing experiment where incremental proportions of one lane of exome sequence reads were randomly sampled from a full lane of G3 *nimbus* exome data (figure 4*b*). While reliable homozygous variant calls (blue dots in figure 4*b*) were made at even shallow read depths, a substantially greater depth was required for reliable heterozygous variant calls. True-positive heterozygous variant calls (green dots in figure 4*b*) increased significantly with increasing depth up to a total of 30 million reads. Ninety-three per cent of true-positive mutations were detected with 35–40 million reads (22–26 times median depth). With increasing the read depth beyond this value, relatively few additional true positives were called but the number of false-positive heterozygous SNV calls doubled.

3.3. Functional validation of causative mutation in *nimbus* strain

To identify the mutation causing the recessive lymphopaenia phenotype in the *nimbus* strain, we performed Amplifluor assays on each of the three homozygous mutations identified in the proband exome sequence to trace their inheritance in the pedigree. Homozygosity for a C-to-T mutation identified at Chr1 : 139 986 182 bp was found to co-segregate with the lymphopaenia phenotype (table 2). This change lies 1 bp upstream of exon 18 of the *Ptprc* gene and disrupts the intronic-1 G nucleotide of the consensus splice acceptor sequence [34], which is otherwise absolutely conserved across vertebrates. PCR amplification of the mutant $Ptprc^{nim}$ mRNA showed the first 14 bp of exon 18 were deleted compared with the spliced wild-type mRNA and putatively the AG nucleotides at +13 to 14 of exon 18 from an alternative splice acceptor site. This altered splicing leads to a frameshift in the mutant transcript from the truncated start of exon 18 onwards. *Ptprc* encodes the CD45 protein, which is a tyrosine phosphatase receptor type C. CD45 is an abundant protein in the plasma membrane of leukocytes and plays critical roles in lymphocyte development in mice and humans (reviewed in [35]). Mice homozygous for the $Ptprc^{nim}$ mutation indeed had almost no CD45 protein on the surface of their B-lymphocytes (2% of wild-type controls) as measured by flow cytometric staining with antibodies to CD45 (figure 5*b*), while heterozygous mice showed an approximately 50 per cent reduction in the expression of CD45. The lymphopaenia in *nimbus* homozygotes matches that in mice and humans with other null or severe loss-of-function mutations in *Ptprc* [29,36,37].

3.4. Identification of causal mutations in 11 additional strains

The successful use of exome analysis to identify causative mutations without meiotic mapping was repeated for 11 other ENU pedigrees with immune disorders or obesity, applying the same analysis to individual exome sequences from proband G3, G4 or G5 mice (table 2). In each of these pedigrees, the causative mutation was revealed solely using exome sequence data followed by SNV-specific Amplifluor PCR typing to correlate the SNV genotypes with the phenotype in the pedigree, without the need for meiotic mapping. The mutations found in each of these strains variously included premature stop codons, disrupted splice donor or acceptor sites and missense changes. The correlation between genotype and phenotype, together with the similar phenotype of independent mutant alleles of the same genes, provided strong corroboration that the mutations identified by exome sequencing were indeed responsible for the phenotypes observed in these mice.

A mean of 6 homozygous and 36 heterozygous mutations were called in the exome sequence of each of the proband individuals from the strains analysed in table 2. These numbers are small enough to contemplate exhaustive validation of each SNV and typing of siblings by Amplifluor PCR assays to test phenotype–genotype concordance, although in many cases a knowledge of the function of the mutant genes allowed candidate mutations to be prioritized. Of the nine strains for which a recessive mutant was sought, the causative variant needed to be selected from on average only 6.4 ($\sigma = 3.8$) candidate mutations. Two of the strains required the causative variant to be identified in a heterozygous form. In these two strains the heterozygous candidate mutation lists were tractably just 40 and 13 variants long.

The incidental mutations revealed by exome sequencing of proband mice in each pedigree represent a remarkable resource for gene-driven testing for other phenotypes. On average, 35.5 ($\sigma = 13.7$) heterozygous exonic mutations were identified in the G3, G4 and G5 mice presented in table 2. Applying the false-positive rate of 19.4 per cent deduced above, on average each G3, G4 or G5 mutant mouse will carry around 29 incidental heterozygous mutations. This gene-driven strategy was successfully reduced to practice in the strain ENU16NI3b, where the original phenotype of low KLRG1 protein on the surface of NK cells occasionally co-occurred with ashen coat colour or stunted growth, neither of which could be explained by a mutation in the KLRG1 gene. With reference to the mutation list obtained from exome sequencing of the G3 proband mouse in this strain, two additional incidental mutations were found by Amplifluor PCR to segregate with each incidental phenotype. A homozygous missense mutation in *Rab27a* co-segregated with ashen coat colour in this pedigree, and an independent *Rab27a* mutation has previously been shown to cause the same trait through a defect in melanosome transport [31]. A homozygous nonsense mutation in the thyroglobulin gene, Tg^{R1471X}, was found to co-segregate in the same pedigree with stunted growth, and this complements an independent study that showed that a spontaneous missense mutation in the *Tg* gene caused stunted growth, hypothyroidism and goiter in an AKR mouse substrain [32]. The new Tg^{R1471X} strain provides a C57BL/6j mouse model for human thyroid dyshormonogenesis 3 syndrome (OMIM: 274700), which was first shown to result from a similar R1510X nonsense mutation in thyroglobulin [38].

3.5. Mutant first-generation mouse resource

The sensitivity and specificity of detecting heterozygous de novo mutations established above opened up a broader strategy to develop mouse experimental models based on tracking specific mutations in gene-driven phenotypic screens, as had been done for the *Tg* and *Rab27a* mutations. To make it possible to do this in a systematic way, we extended the exome sequencing approach to identify novel protein-changing mutations arising in the G1 founders of ENU mutagenized pedigrees, prior to any phenotypic

Table 2. Mutations identified using exome sequence data.

sample identifier	capture	hom calls	het calls	causal or incidental mutation	gene	detected zygosity	observed phenotype	published allele	published allele phenotype	ref	chr	coord	ref allele	var allele	AA change	PolyPhen score	observed genotype–phenotype correlation
ENU16CH51a	Agilent	3	46	causal	Prkdc	hom	few T and B cells in blood	$Prkdc^{SCID}$	few B or T cells	[5]	16	15810811	T	A	Y3442→X	—	6 hom affected, 19 het unaffected, 3 wt unaffected
ENU14CH36b	Agilent	14	21	causal	CD22	hom	fewer mature and more immature B cells in blood	$Cd22^{tm1Eac}$	fewer mature B cells	[23]	7	31655399	A	T	C512→X	—	1 hom affected, 2 het unaffected, 1 wt unaffected
ENU16NI19a	Agilent	6	32	causal	Dock2	hom	decreased naive T cells and B cells in blood	$Dock2^{tm1Tsas}$	decreased naive T and B cells in blood	[24]	11	34414481	C	A	E775→X	—	14 hom affected, 16 het unaffected, 5 wt unaffected
ENU16CH85a	NG	2	20	causal	Reln	hom	ataxia and small body size	$Reln^{rl-tg}$	tremors, dystonia and ataxia	[25]	5	21408594	A	G	splice	—	4 hom affected, 2 het unaffected, 6 wt unaffected
ENU16CH17a	Agilent	3	45	causal	Lyn	hom	decreased blood B cells, increased percentage immature	Lyn^{Mld4}	decreased B cells	[26]	4	3710143	A	G	T410→A	0.990	10 hom affected, 6 het unaffected, 1 wt unaffected
ENU14CH48	NG	7	62	causal	Prkdc	hom	few T and B cells in blood	$Prkdc^{SCID}$	few B or T cells	[27]	16	15714375	T	C	splice	—	4 hom affected, 12 het unaffected, 3 wt unaffected
ENU16NI24a	NG	9	37	causal	Lepr	hom	obese	$Lepr^{db}$	obese	[28]	4	101452668	T	A	N429→K	1.000	3 hom affected, 17 het unaffected, 24 wt unaffected
nimbus	Agilent	3	40	causal	Ptprc	hom	decreased naive T cells and B cells in blood	$Ptprc^{tm1Holm}$	decreased naïve T cells and B cells	[29]	1	139986182	C	T	splice	—	12 hom affected, 18 het unaffected, 3 wt unaffected
ENU16CH71a	NG	10	30	causal	Pax5	hom	few blood B cells	$Pax5^{tm1Mbu}$	arrest of B cell development	[30]	4	44704884	G	A	I78→N	0.266	24 hom affected, 23 het unaffected
ENU18CH65a	NG	4	40	causal	Fcer2a	het	decreased Fcer2a (CD23) on B cells				8	3690110	G	T	C18→X	—	11 hom affected, 29 het intermediate, 20 wt unaffected

(*Continued.*)

Table 2. (*Continued.*)

sample identifier	capture	hom calls	het calls	causal or incidental mutation	gene	detected zygosity	observed phenotype	published allele	published allele phenotype	ref	chr	coord	ref allele	var allele	AA change	PolyPhen score	observed genotype–phenotype correlation
ENU16N3b	NG	7	41	causal	KLRG1	hom	low KLRG1 on NK cells				6	122232913	G	A	5′ UTR	—	20 hom affected, 1 het affected, 9 het unaffected
"	"	"	"	incidental	Rab27a	het	coat colour (Ashen)	$Rab27a^{ash}$	grey coat colour	[31]	9	72930272	T	A	W73→R	1.000	3 hom affected,13 het unaffected, 6 wt unaffected
"	"	"	"	incidental	Tg	hom	small body size	Tg^{cog}	hypothyroidism, goiter, impaired growth	[32]	15	66536802	C	T	R1471→X	—	4 hom affected, 10 het unaffected, 2 wt unaffected
ENU15CH72a	NG	0	13	causal	Ptpn6	het	decreased IgM on mature B cells	$Ptpn6^{mev}$	decreased IgM on mature B cells in hets	[33]	6	124672073	G	A	T464→I	1.000	4 hom affected, 29 het affected, 21 wt unaffected

screening or selection of their G2 and G3 progeny, and when all the mutations are heterozygous (figure 1). We sequenced the enriched exomes of eight different G1 mice as a barcoded, pooled sample on an Illumina HiSeq sequencing run. This provided a greater number of reads per exome than the datasets generated on the GAIIx sequencers, and yielded better than 20 times sequence depth over 80.7 per cent ($\sigma = 1.8\%$) CCDS exons. As expected, very few homozygous variants were identified in the filtered variant lists, presumably being rare variants previously unobserved in the parental C57BL/6j stock. The numbers of heterozygous variants in the G1 mice ($\mu = 59.6$, $\sigma = 13.1$) were higher than those found in G3, G4 or G5 mice ($\mu = 36.5$, $\sigma = 13.7$; table 2), which was as expected since a fraction of ENU-induced alleles will be lost in each subsequent generation owing to random drift and purifying selection. Hence, given the information presented in figure 4*b*, we would expect that the majority of true ENU-induced mutations have been detected from these datasets.

Of the 454 unique mutations detected across these eight G1 mice, 18 (4%) created a premature stop codon, 65 (14%) putatively disrupted an mRNA splice donor/acceptor site and 370 (81%) caused an amino acid substitution (see electronic supplementary material, table S4). We altered PolyPhen2 [39] to use mouse sequence databases (rather than the default human inputs) and calculated scores for missense G1 mutations. Figure 6 shows a comparison of these scores with those calculated for a set of previously characterized ENU-induced mutations known to cause immunological traits. For the causal missense mutations, PolyPhen2 correctly assigned a very high score (greater than 0.95) of 'probably damaging' to 75 per cent and an intermediate to high score (0.44–0.95) of 'possibly damaging' to a further 15 per cent. This result validates the predictive accuracy of PolyPhen2 when applied to novel mouse mutations. Of the 370 de novo missense mutations identified in G1 mice, 134 (36%) were assigned a 'probably damaging' score of greater than 0.95 and 59 (16%) were classified as 'possibly damaging' with a score of 0.505–0.897. The genes affected by these 272 potentially damaging mutations include those known to cause human disease through to entirely unexplored genes with intriguing expression patterns and protein domains (see electronic supplementary material, table S3). By identifying de novo ENU mutations in G1 founders in this way and then breeding, genotyping and phenotyping their G2 and G3 offspring, this approach provides an immediate source for new experimental models for understanding human diseases and traits.

4. Discussion

The pursuit of gene function that starts with the identification of medically important phenotypes displayed by individual mammals (the so-called forward-genetics) has until now been constrained by the time-consuming and expensive bottleneck of mapping these traits to their underlying genetic cause. Conversely, reverse genetics approaches based on knocking out individual genes in embryonic stem cells remain constrained by a comparably time-consuming and expensive bottleneck of converting the embryonic stem cells into a pedigree of mice that can be phenotypically evaluated. Here we have shown that exome capture followed by

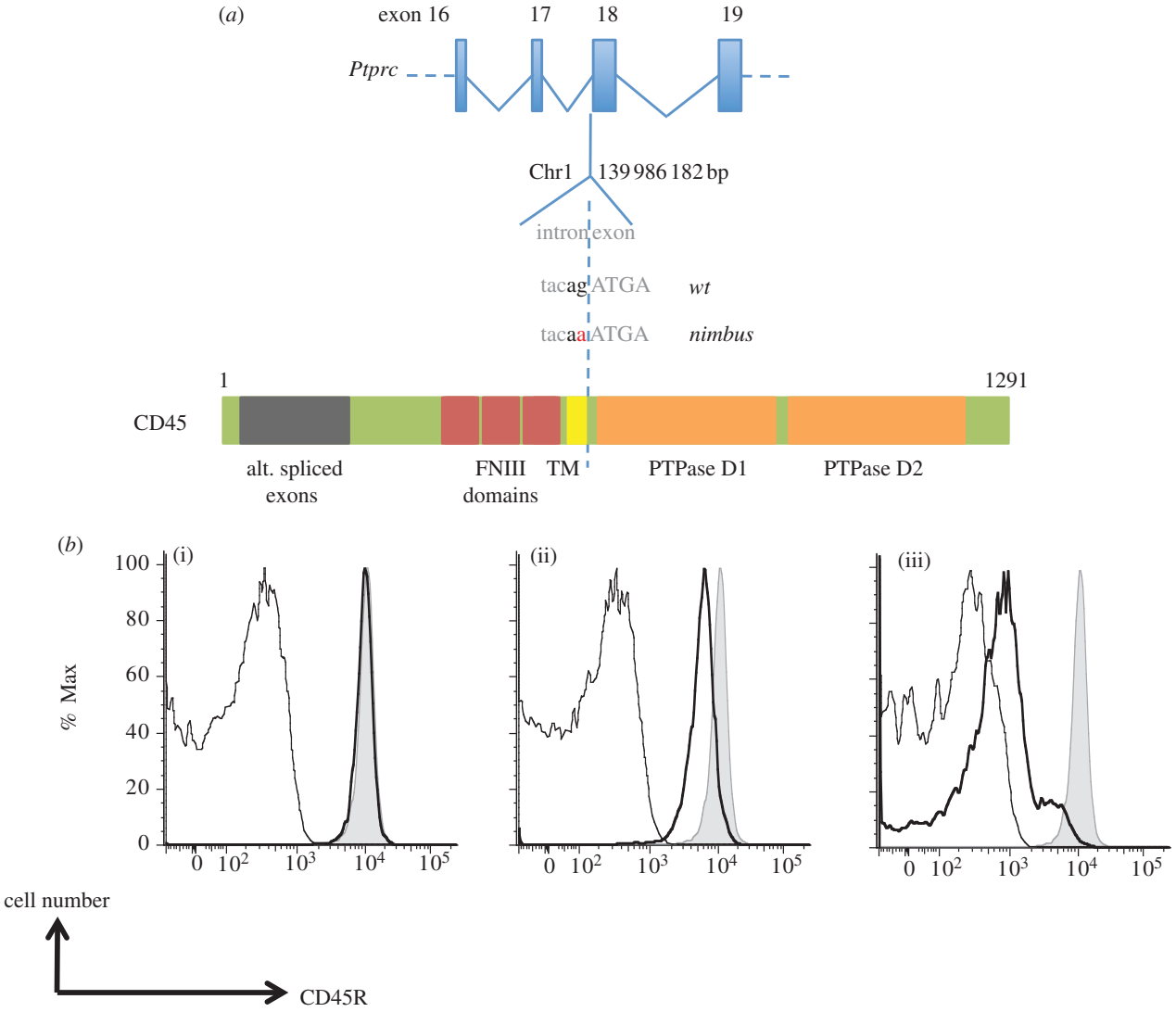


Figure 5. *Nimbus* results from a loss of function mutation in the *Ptprc* gene. (*a*) Schematic diagram showing the location of single nucleotide mutation at Chr1:139986183 at the +1 intronic position of the exon 17 splice donor sequence and the location of the corresponding region in the encoded CD45 protein (TM, transmembrane domain; FNIII, fibronectin III-like domain; PTP, protein tyrosine phosphatase). (*b*) Loss of CD45 protein expression. Bold black lines show flow cytometric staining with antibody to the B-cell-specific CD45R isoform on IgM^+, IgD^+ B lymphocytes in blood from (i) $Ptprc^{+/+}$ wild-type (wt), (ii) $Ptprc^{nimbus/+}$ heterozygous or (iii) $Ptprc^{nimbus/nimbus}$ homozygous mouse, compared with negative control staining on CD3+ T cells in the same mouse (thin black line) and compared with positive control staining with the same antibody on B cells in a wt mouse (grey shaded area).

massively parallel DNA sequence analysis reliably identifies the majority of homozygous and heterozygous ENU-induced mutations. Not only does this eliminate the bottleneck to forward genetics by identifying causal mutations without the need for meiotic mapping, but also it bypasses a key restriction for reverse genetics by revealing thousands of possibly damaging mutations in live-breeding C57BL/6j mouse pedigrees that are immediately available for experimental analysis of gene function.

By technical and biological replication of exome analyses and confirmation of individual SNV calls by PCR, we have shown that both homozygous and heterozygous protein-changing mutations induced by ENU de novo in live-breeding pedigrees of C57BL/6j mice can be called reliably with an estimated sensitivity of 78.7 per cent and a specificity of 80.6 per cent. In 11 separate C57BL/6j mutant strains from forward genetics screens for immune system disorders or obesity, we were able to bypass the need for meiotic mapping and identify short lists of protein-changing ENU-induced mutations that were heterozygous or homozygous in proband individuals from these pedigrees, among which we were able to identify a causative mutation that explained the immunological or obesity phenotype. In identifying ENU-induced mutations, we found massively parallel sequencing data to be highly reliable and sources of error were predictable, such that by filtering commonly called variants (along with previously observed genetic variation) we were able to restrict the false-positive call rate to less than 20 per cent while not incurring a disproportionate false-negative call rate. In terms of the read depth required to reliably identify heterozygous mutations, we found that around 35 million paired-end sequence reads are sufficient to identify more than 90 per cent of these changes.

Fairfield *et al.* [9] have also produced an extensive demonstration of exome capture and sequencing in mice to identify causative mutations. In their study, exome sequence data were used in combination with meiotic mapping information to identify causal mutations without a large validation burden. Our results both confirm and extend this study. Laudably, the Fairfield *et al.* [9] study describes three

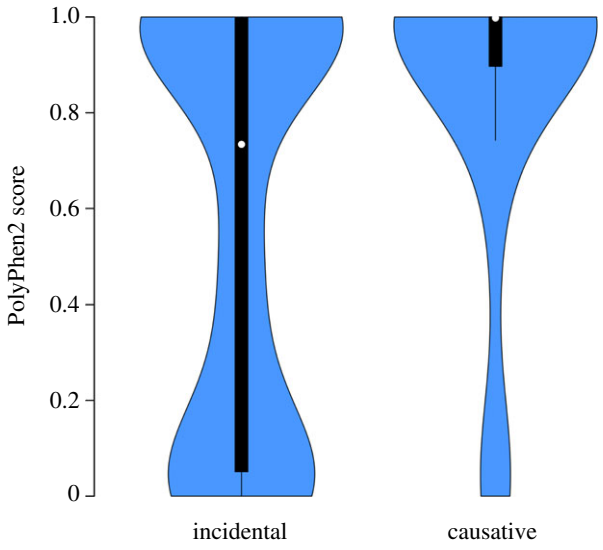


Figure 6. Violin plot comparing PolyPhen2 scores for incidental and causative mutations. The black bars represent a boxplot where 50% of values lie within the main bar. The white dot indicates the median polyphen value for each set of scores. The blue region is a kernel density plot representing the distribution of PolyPhen2 scores. The numbers of mutations included in the plot were: incidental mutations, $n = 325$ and causative mutations, $n = 40$. A Mann–Whitney test for the equality of the mean PolyPhen2 score of the incidental and causative mutations indicated a significant difference in score ($W = 4168$, $p = 0.0000862$).

mutant strains where they did not identify a causal mutation, even with the aid of meiotic mapping information. They speculated that, in those strains where the causal mutation could not be identified, it probably lay outside the chromosomal regions enriched by exome capture. Our analysis provides further insight into this problem and shows that, in approximately one of five mouse strains, we can expect a causal mutation to remain undetected owing to it not being efficiently captured prior to sequencing and/or subsequently detected. We found that solution capture methods commercialized by Agilent and Nimblegen are both effective at specifically concentrating the coding part of the mouse genome, but that a consistent approximately 15 per cent portion of exonic regions is absent from subsequently sequenced reads, regardless of how deeply the captured DNA is sequenced. This may be a fundamental limitation of exome enrichment technologies, perhaps indicating that some genomic regions may be resistant to efficient hybridization with capture baits and/or the PCR amplification steps in the capture and library preparation protocols. From analysis of exome datasets from related mice, in a small number of cases known heterozygous variants were only poorly detected owing to a very few reads supporting the mutant genotype. This effect may indicate that in some local sequence contexts the mutant genotype is out-competed by the reference genotype during sequence capture.

Mutated C57BL/6j inbred mice provide an ideal system for tackling the challenges of identifying rare, de novo mutations from a background of normal genetic variation. While the laboratory mouse is an inbred organism with very little genetic variation, we found that it was necessary to control for even this small amount of variation through a series of data filtering strategies employing catalogues of known strain variants and other sources or recurring false positives in order to identify true mutations with high specificity. Given sufficient data for a specific mouse strain (10–20 individual exome sequences), this strategy of cataloguing recurrent variants has also proven effective in identifying ENU-induced mutations in mice out-crossed to strains other than C57BL/6j (data not shown). We found that detection of ENU-induced mutations can be further enhanced by technical replication of exome analysis and by biological replication taking advantage of heritability information in closely related individuals. Taken together, this information makes pathogenic mutation detection in outbreeding mammals (such as humans) a more tractable possibility.

We have shown that it is feasible to also perform these exome analyses in multiplexed, bar-coded samples from many separate G1 founder mice. This makes it straightforward to analyse the exomes of hundreds of G1 founder mice per year and propagate the mutations they carry in live-breeding pedigree structures such as the ones employed here (figure 1). Given the number of protein-changing mutations we identified in each G1 mouse (table 3), a live-breeding resource of 350 pedigrees bred for two generations from 700 G1 mice each year would reveal 42 000 new protein-changing mutations per year, of which around half are expected to be deleterious. Hence, reliable identification of induced mutations has the potential to transform genetic screens of genes of unknown function and produce mouse models of hundreds of human diseases.

5. Material and methods

5.1. Mutant mouse generation

The *nimbus* mouse strain was generated by treating pure C57BL/6j male mice with the mutagen ENU at the Australian Phenomics Facility of the Australian National University as previously described [10]. Briefly, adult male animals received 90 mg of ENU per kilogram of body weight by three weekly intraperitoneal injections. Once fertility was regained after a further eight weeks, the animals were mated with C57BL/6j females to generate G1 offspring carrying a unique cohort of heterozygous SNVs. A subset of SNVs was brought to homozygosity through unrelated G1 crosses followed by intercrossing to G3 (as shown in figure 1). A peripheral blood screen for lymphocyte subsets identified the *nimbus* strain at G3 as displaying a mild lymphopaenia. All other mutated mouse strains sequenced were generated via this protocol.

5.2. Exome enrichment and sequencing

DNA was extracted from ear tissue of affected mice and 3.5 µg prepared as paired-end genomic libraries (PE-102-1001: Illumina, San Diego, CA). Technical replicates were produced from the same DNA sample. Exome enrichment was performed using either the SureSelect Mouse Exome kit (G7550A-001: Agilent, CA) or the SeqCap Mouse Exome kit (early access: Nimblegen, Madison, WI) following the manufacturer protocols. Four amplification cycles were used in the library pre-capture PCR using Herculase II fusion polymerase (600677, Stratagene) and eight cycles in the post-enrichment amplification for both capture technologies. Enriched libraries were diluted to 10 nM concentrations before further dilution to 7–8 pM for cluster generation and sequencing-by-synthesis on either the Illumina Genome

Table 3. Sequencing statistics and variant calls for G1 mice.

sample identifier	total reads sequenced	CCDS on-target efficiency	median read depth over CCDS exons	CCDS bases covered 5 times or better depth (%)	CCDS bases covered 20 times or better depth (%)	raw variant calls	filtered homozygous variant calls	filtered heterozygous variant calls
MMP-1	94013861	0.539	87	85.9	82.3	10 463	3	55
MMP-2	96872136	0.532	89	86.0	82.6	11 834	0	59
MMP-3	97528301	0.538	89	85.9	82.3	11 396	2	79
MMP-4	71404847	0.536	66	85.3	79.7	9349	0	54
MMP-5	72606249	0.525	64	84.9	78.2	13 404	1	74
MMP-6	92123751	0.531	84	85.9	82.1	10 328	0	42
MMP-7	71220780	0.531	65	85.3	79.5	9050	3	68
MMP-8	67858625	0.537	62	85.2	79.0	8929	0	46

Analyser IIx as 75 bp PE reads or the Illumina HiSeq as 100 bp reads. Each library sequenced on an Illumina GAIIx was sequenced on a single lane of an eight-lane flow-cell, whereas libraries sequenced on the Illumina HiSeq were multiplexed in a pool of 10 samples and sequenced together, and disambiguated using sample bar-coding.

5.3. Single-nucleotide variant detection workflow

A custom workflow was developed to process sequence reads to detect ENU-induced mutations. This workflow holds together a number of open-source analysis tools and employs a Perl code-base to perform custom filtering, reporting and job process control (figure 2*a*). BWA (v. 0.5.9-rc16; [16]) with default settings was chosen to align paired-end reads to the reference mouse genome (mm9/NCBIM37). Reads aligning to multiple genomic locations were removed and raw SNV calls were made using SAMTools (v. 0.1.15; [17]) with parameters set to allow a less conservative calling rate than the default settings, which significantly involved disabling the base alignment quality filtering function. Raw SNV calls were classified as homozygous or heterozygous on the basis of the ratio of alleles (hom $>$ 0.8 variant allele; het two alleles $>$ 0.3) and then annotated as to whether they were also present in dbSNP (v. 128; http://www.ncbi.nlm.nih.gov/snp/), whether they commonly occurred in our exome data and, where appropriate, whether they were strain-specific variants identified from the Sanger Institute mouse genomes sequencing project (http://www.sanger.ac.uk/resources/mouse/genomes/). Commonly occurring variants were collated from all exome data collected by our laboratory. Further annotation of variants was performed to determine overlap with CCDS exons [18] and denote non-synonymous changes (using Annovar [40]). Changes that lay in potential splice donor–acceptor sites immediately adjacent to exon boundaries (out to 10 intronic bases) were also annotated. Using these annotations, we filtered the variant list to only include non-synonymous or splice donor–acceptor site changes that were novel to a particular sample. From this filtered list of variants, for each exome a list of genes containing more than one variant was compiled for each sample and then used to further filter variants across all samples that were found in these multi-SNV genes.

5.4. Variant validation

SNVs were validated using Amplifluor assays (Chemicon, Temecula, CA). Primers were designed using the Assay architect online tool (http://apps.serologicals.com/AAA/mainmenu.aspx). Fluorescent intensities were detected using a Fluostar optima (BMG). The individual affected mice used in the study and a C57BL/6j control were analysed for each SNV assay.

6. Acknowledgements

The authors acknowledge funding from the National Health and Medical Research Council (Australia), the Wellcome Trust, the National Institutes of Health (USA) and the Australian Government.

References

1. Acevedo-Arozena A, Wells S, Potter P, Kelly M, Cox RD, Brown SDM. 2008 ENU mutagenesis, a way forward to understand gene function. *Annu. Rev. Genomics Hum. Genet.* **9**, 49–69. (doi:10.1146/annurev.genom.9.081307.164224)
2. Justice MJ, Noveroske JK, Weber JS, Zheng B, Bradley A. 1999 Mouse ENU mutagenesis. *Hum. Mol. Genet.* **8**, 1955–1963. (doi:10.1093/hmg/8.10.1955)
3. Albert TJ *et al*. 2007 Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905. (doi:10.1038/nmeth1111)
4. Gnirke A *et al*. 2009 Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189. (doi:10.1038/nbt.1523)
5. Ng SB, Nickerson DA, Bamshad MJ, Shendure J. 2010 Massively parallel sequencing and rare disease. *Hum. Mol. Genet.* **19**, R119–R124. (doi:10.1093/hmg/ddq390)
6. Arnold CN, Xia Y, Lin P, Ross C, Schwander M, Smart NG, Müller U, Beutler B. 2011 Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis. *Genetics* **187**, 633–641. (doi:10.1534/genetics.110.124586)
7. Yabas M *et al*. 2011 ATP11C is critical for the internalization of phosphatidylserine and differentiation of B lymphocytes. *Nat. Immunol.* **12**, 441–449. (doi:10.1038/ni.2011)
8. Zhang Z *et al*. 2009 Massively parallel sequencing identifies the gene Megf8 with ENU-induced mutation causing heterotaxy. *Proc. Natl Acad. Sci. USA* **106**, 3219–3224. (doi:10.1073/pnas.0813400106)
9. Fairfield H *et al*. 2011 Mutation discovery in mice by whole exome sequencing. *Genome Biol.* **12**, R86. (doi:10.1186/gb-2011-12-9-r86)
10. Nelms KA, Goodnow CC. 2001 Genome-wide ENU mutagenesis to reveal immune regulators. *Immunity* **15**, 409–418. (doi:10.1016/S1074-7613(01)00199-6)
11. Probst FJ, Justice MJ. 2010 Mouse mutagenesis with the chemical supermutagen ENU. *Methods Enzymol.* **477**, 297–312. (doi:10.1016/S0076-6879(10)77015-4)
12. Boles MK *et al*. 2009 Discovery of candidate disease genes in ENU-induced mouse mutants by large-scale sequencing, including a splice-site mutation in nucleoredoxin. *PLoS Genet.* **5**, e1000759. (doi:10.1371/journal.pgen.1000759)
13. Quwailid MM *et al*. 2004 A gene-driven ENU-based approach to generating an allelic series in any gene. *Mamm. Genome* **15**, 585–591. (doi:10.1007/s00335-004-2379-z)
14. Takahasi KR, Sakuraba Y, Gondo Y. 2007 Mutational pattern and frequency of induced nucleotide changes in mouse ENU mutagenesis. *BMC Mol. Biol.* **8**, 52. (doi:10.1186/1471-2199-8-52)
15. Waterston RH *et al*. 2002 Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562. (doi:10.1038/nature01262)
16. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760. (doi:10.1093/bioinformatics/btp324)
17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
18. Pruitt KD *et al*. 2009 The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323. (doi:10.1101/gr.080531.108)
19. Guénet J-L. 2004 Chemical mutagenesis of the mouse genome: an overview. *Genetica* **122**, 9–24. (doi:10.1007/s10709-004-1442-8)
20. Levy S *et al*. 2007 The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254. (doi:10.1371/journal.pbio.0050254)
21. Wheeler DA *et al*. 2008 The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876. (doi:10.1038/nature06884)
22. Wendl MC, Wilson RK. 2008 Aspects of coverage in medical DNA sequencing. *BMC Bioinform.* **9**, 239. (doi:10.1186/1471-2105-9-239)
23. Otipoby KL, Draves KE, Clark EA. 2001 CD22 regulates B cell receptor-mediated signals via two domains that independently recruit Grb2 and SHP-1. *J. Biol. Chem.* **276**, 44 315–44 322.
24. Fukui Y, Hashimoto O, Sanui T, Oono T, Koga H, Abe M, Inayoshi A, Noda M, Oike M, Shirai T, Sasazuki T. 2001 Haematopoietic cell-specific CDM family protein DOCK2 is essential for lymphocyte migration. *Nature* **412**, 826–831. (doi:10.1038/35090591)
25. D'Arcangelo G, Miao GG, Chen SC, Soares HD, Morgan JI, Curran T. 1995 A protein related to extracellular matrix proteins deleted in the mouse mutant reeler. *Nature* **374**, 719–723. (doi:10.1038/374719a0)
26. Verhagen AM *et al*. 2009 A kinase-dead allele of Lyn attenuates autoimmune disease normally associated with Lyn deficiency. *J. Immunol.* **182**, 2020–2029. (doi:10.4049/jimmunol.0803127)
27. Bosma GC, Custer RP, Bosma MJ. 1983 A severe combined immunodeficiency mutation in the mouse. *Nature* **301**, 527–530. (doi:10.1038/301527a0)
28. Chen H *et al*. 1996 Evidence that the diabetes gene encodes the leptin receptor: identification of a mutation in the leptin receptor gene in db/db mice. *Cell* **84**, 491–495. (doi:10.1016/S0092-8674(00)81294-5)
29. McNeill L, Salmond RJ, Cooper JC, Carret CK, Cassady-Cain RL, Roche-Molina M, Tandon P, Holmes N, Alexander DR. 2007 The differential regulation of Lck kinase phosphorylation sites by CD45 is critical for T cell receptor signaling responses. *Immunity* **27**, 425–437. (doi:10.1016/j.immuni.2007.07.015)
30. Urbánek P, Wang ZQ, Fetka I, Wagner EF, Busslinger M. 1994 Complete block of early B cell differentiation and altered patterning of the posterior midbrain in mice lacking Pax5/BSAP. *Cell* **79**, 901–912. (doi:10.1016/0092-8674(94)90079-5)
31. Wilson SM *et al*. 2000 A mutation in Rab27a causes the vesicle transport defects observed in ashen mice. *Proc. Natl Acad. Sci. USA* **97**, 7933–7938. (doi:10.1073/pnas.140212797)
32. Kim PS, Hossain SA, Park YN, Lee I, Yoo SE, Arvan P. 1998 A single amino acid change in the acetylcholinesterase-like domain of thyroglobulin causes congenital goiter with hypothyroidism in the *cog/cog* mouse: a model of human endoplasmic reticulum storage diseases. *Proc. Natl Acad. Sci. USA* **95**, 9909–9913. (doi:10.1073/pnas.95.17.9909)
33. Cornall RJ, Cyster JG, Hibbs ML, Dunn AR, Otipoby KL, Clark EA, Goodnow CC. 1998 Polygenic autoimmune traits: Lyn, CD22, and SHP-1 are limiting elements of a biochemical pathway regulating BCR signaling and selection. *Immunity* **8**, 497–508. (doi:10.1016/S1074-7613(00)80554-3)
34. Roca X, Olson AJ, Rao AR, Enerly E, Kristensen VN, Børresen-Dale A-L, Andresen BS, Krainer AR, Sachidanandam R. 2008 Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics. *Genome Res.* **18**, 77–87. (doi:10.1101/gr.6859308)
35. Hermiston ML, Xu Z, Weiss A. 2003 CD45: a critical regulator of signaling thresholds in immune cells. *Annu. Rev. Immunol.* **21**, 107–137. (doi:10.1146/annurev.immunol.21.120601.140946)
36. Kung C *et al*. 2000 Mutations in the tyrosine phosphatase CD45 gene in a child with severe combined immunodeficiency disease. *Nat. Med.* **6**, 343–345. (doi:10.1038/73208)
37. Zikherman J, Jenne C, Watson S, Doan K, Raschke W, Goodnow CC, Weiss A. 2010 CD45-Csk phosphatase-kinase titration uncouples basal and inducible T cell receptor signaling during thymic development. *Immunity* **32**, 342–354. (doi:10.1016/j.immuni.2010.03.006)
38. Targovnik HM, Medeiros-Neto G, Varela V, Cochaux P, Wajchenberg BL, Vassart G. 1993 A nonsense mutation causes human hereditary congenital goiter with preferential production of a 171-nucleotide-deleted thyroglobulin ribonucleic acid messenger. *J. Clin. Endocrinol. Metab.* **77**, 210–215. (doi:10.1210/jc.77.1.210)
39. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010 A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249. (doi:10.1038/nmeth0410-248)
40. Wang K, Li M, Hakonarson H. 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164. (doi:10.1093/nar/gkq603)

7.2 Further discussion

In this publication we describe our production pipeline to detect rare, ENU induced mutations in mice using targeted exome sequencing. While historically forward genetic approaches have required huge amounts of time and expense to map the traits of interest to their underlying genetic cause, here we demonstrate how we are able to detect most homozygous and heterozygous ENU-induced mutations using targeted exome capture and sequencing without any pre-existing meiotic mapping information. Further, we are able to perform this task in a highly automated, high-throughput manner with low estimated false positive and false negative rates (~20%), all run in a completely reproducible in-house framework.

To determine whether detecting ENU variants solely from exome sequence data in mice was feasible, a bespoke prototype pipeline was constructed in early 2010 to demonstrate proof of concept. This pipeline consisted of a MySQL tracking database and detailed configuration files containing sample metadata as well as the exact commands to run for the alignment, variant detection, annotation, and summary steps. The first iteration of the pipeline detected a huge numbers of variants; yielding variant lists too large for manual interrogation essentially making causal variant identification impossible without additional mapping information. We expected to observe only ~35 true ENU mutations per exome given the homogeneous genetic background of C57BL/6 mice and the ENU mutation rate of 1 mutation per mega base. Yet many more mutations were being detected per exome indicating the likely presence of both technical and biological replicate variants. To identify and remove variants resulting from technical replicates a strategy was devised where mutations common to unrelated mice would be catalogued and de-prioritised from subsequent variant lists. While this strategy effectively removed most technical replicates, special care had to be taken to correctly track mouse relationships in order to prevent the erroneous removal of important ENU mutations common to related mice. To identify and remove variants arising from biological replicates, a similar strategy was devised where each new group of mice from a previously un-encountered strain was used to generate a list of common strain specific variants that could be used to filter out such variants for

all mice of this strain encountered in the future. Collectively these custom strategies reduce the number of candidate causal mutations to less than 50 variants per exome, which, when combined with extensive annotation allow researchers to routinely identify the important causal variant.

Next, a production pipeline based on the prototype was required, with the new version able to process huge number of samples in a highly automated and reproducible manner. The implementation of the production system proved a significant undertaking requiring reengineering and redesign of the prototype, a task that ultimately took one year to complete. The first major challenge was making the pipeline highly automated and capable of processing large number of samples simultaneously, a task we accomplished by deeply integrating the system with high-performance computing resources (HPC) and minimising the requirements for manual intervention. The production system was designed to run primarily on HPC clusters at the National Computational Infrastructure, and to run on their large raijin system (http://nci.org.au/nci-systems/national-facility/peak-system/raijin/), the 38th largest supercomputer in the world (http://www.top500.org/). In order to cope with the wide variety of analysis steps required a generic wrapper for running a single step was implemented to reflect the fact that any large workflows can be broken down into smaller analysis components and ultimately individual system commands. The process for running a step consists of reading the pre-generated sample-specific configuration file; a file that contains all information required to run each step such as job pre-requisites, command arguments, external modules to load, and compute requirements. After ensuring prerequisite conditions have been met, job(s) are submitted and monitored until completion when step-specific quality control is performed with the exact command run recorded in both the tracking database and log file. The final task performed by the wrapper is the submission of the next analysis job, a process that continues until the final analysis step is complete. Total workflow automation is crucial not only during routine sample analysis but also in both detecting and recovering from errors that inevitably occur. Our system works to detect and recover from both catastrophic errors (errors that cause the entire analysis to halt) as well as non-fatal errors that generate incorrect output with non-fatal errors being particularly problematic both in wasting ensuing CPU cycles and in making it difficult to identify the

exact point of failure. Overall, the focus on automation makes possible both the seamless addition of new analysis steps as well as the ability to resume analysis from any failed step facilitating easy recovery from unforeseen events such as hardware failure.

The ability to unequivocally reproduce past analysis results is of the utmost importance in informatics pipelines particularly in light of the fast moving nature of both sequencing technologies and bioinformatics software development. Unfortunately reproducibility is far from the norm in biological sciences with a study discovering less than half of the microarray studies published in Nature Genetics were reproducible (91). To ensure complete reproducibility our system employs multiple tracking methods using an underlying MySQL database and a detailed log file, with each method capable of reproducing any previous workflow independently. While this design decision was initially implemented to ensure output consistency, it has proven useful in diagnosing pipeline crashes by offering two distinct reference points for identifying the exact point of failure. In addition to recording the system commands run during analysis steps, full reproducibly requires all files and external binaries utilised to be versioned with this information recorded in both log files and the database. This includes the version of external software used, the version of the in-house code base, and external annotation data set versions. External annotation data in particular represents a challenge for truly reproducible results with both rapid updates and changes in format common the norm for annotation data sets. In our system, to manage external annotations effectively and consistently, annotations are parsed from the original data source and converted into a standard file format utilising a standardised naming scheme. Groups of common annotation files are stored in time-stamped directories thus allowing the most up to date annotation information to be automatically detected and incorporated into each new analysis. This consistent handling of annotations allows for seamless updates when new annotation versions become available and also simplifies the reanalysis of large numbers of samples when important annotation sets like dbSNP (2) are updated. Collectively, these features allow the system to reproduce any previous result and thus meet this important system requirement.

The production version of the pipeline has been extremely successful since its inception, being run in high-production mode for four years while undergoing very little change. To date it has analysed 666 ENU mice and 2075 G1 mice and detected causal variants in over three quarters of all ENU mice sequenced with numerous resultant publications, some of which are detailed in Chapter 8. The overall flexibility of the design of this pipeline allowed it to serve as the template for the creation of a human exome analysis pipeline, a pipeline that has analysed 525 human exomes to date and generated four additional publications detailed in Chapter 8. Overall the implementation of this approach has transformed genetic screens in mice, and establishes a general strategy for analysing rare DNA variants while opening up a large new source for experimental models of human disease. This success of this system and the resultant discoveries will potentially lead to a much greater understanding of the underlying mechanism of human disease.

Chapter 8: Other publications

I am included as an author on the additional eight publications summarized below. These publications detail discoveries made by the software developed during my PhD candidature.

1) Taupin, D, W. Lam, D. Rangiah, L. McCallum, B. Whittle, Y. Zhang, D. Andrews, **M. Field**, C. C. Goodnow and M. C. Cook (2015). "A deleterious RNF43 germ line mutation in a severely affected serrated polyposis kindred." Human Genome Variation 2. 2015; 16;2:15013

Summary: "We report a germ line nonsense mutation within the extracellular domain of the RING finger ubiquitin ligase RNF43, segregating with a severe form of serrated polyposis within a kindred. The finding provides evidence that inherited RNF43 mutations define a familial cancer syndrome."
Contribution: This work made extensive use of our variant identification system and the causal variant was identified using a prototype of the VASP tool.

2) Johar, A. S, C. Mastronardi, A. Rojas-Villarraga, H. R. Patel, A. Chuah, K. Peng, A. Higgins, P. Milburn, S. Palmer, M. F. Silva-Lara, J. I. Velez, D. Andrews, **M. Field**, G. Huttley, C. Goodnow, J. M. Anaya and M. Arcos-Burgos. (2015). "Novel and rare functional genomic variants in multiple autoimmune syndrome and Sjogren's syndrome." Journal of Translational Medicine 2015; 13: 173.

Summary: "Novel and rare exonic mutations that may account for autoimmunity were identified. Among those, the *LRP1/STAT6* novel mutation has the strongest case for being categorised as potentially causative of MAS given the presence of intriguing patterns of functional interaction with other major genes shaping autoimmunity."
Contribution: This work used our variant detection system to identify the variants deemed likely causative.

3) Dunkerton, S., **M. Field**, V. Cho, E. Bertram, B. Whittle, A. Groves and H. Goel (2015). "A de novo mutation in KMT2A (MLL) in monozygotic twins with Wiedemann-Steiner syndrome." American Journal of Medical Genetics Part A. 2015; 167A(9):2182-7.

Summary: "In this study, we have identified a *de novo* mutation in KMT2A associated with psychomotor developmental delay, facial dysmorphism, short

stature, hypertrichosis cubiti, and small kidneys. This finding in monozygotic twins gives specificity to the WSS. The description of more cases of WSS is needed for further delineation of this condition. Small kidneys with normal function have not been described in this condition in the medical literature before."

Contribution: This work made extensive use of our variant identification system and the causal variant was identified using a prototype of the VASP tool.

4) Lee, C. E, D. A. Fulcher, B. Whittle, R. Chand, N. Fewings, **M. Field**, D. Andrews, C. C. Goodnow and M. C. Cook (2014). "Autosomal dominant B cell deficiency with alopecia due to a mutation in NFKB2 that results in non-processible p100." Blood. 2014; 124(19):2964-72.

Summary: "A novel NFKB2 mutation confers a severe B cell deficiency but antibody production is partially preserved. Unprocessed p100 results in an IκB-like action on the canonical NF-κB pathway."

Contribution: This work used our variant detection system to identify the causative variant.

5) Enders, A, A. Short, L. A. Miosge, H. Bergmann, Y. Sontani, E. M. Bertram, B. Whittle, B. Balakishnan, K. Yoshida, G. Sjollema, **M. A. Field**, T. D. Andrews, H. Hagiwara and C. C. Goodnow. (2014). "Zinc-finger protein ZFP318 is essential for expression of IgD, the alternatively spliced Igh product made by mature B lymphocytes." Proceedings of the National Academy of Science of the United States of America. 2014; 111(12): 4513-4518.

Summary: "Mammalian B lymphocytes make antibodies of five different heavy chain isotypes, IgM, IgD, IgG, IgE, and IgA. The different isotypes are produced at discrete stages in B-cell development from a single immunoglobulin heavy chain (Igh) gene, either by irreversible rearrangement of the gene to make IgG, IgE, or IgA or by alternative splicing of the RNA transcribed from the Igh gene to coexpress IgM and IgD. Developmentally regulated trans-acting factors have been hypothesized to control IgM and IgD expression from large Igh RNAs, but these factors have remained elusive for several decades. Here, using a genome wide mutation screen in mice, we identify an obscure gene, Zfp318, as encoding a specific and essential factor promoting IgD expression in mature B cells."

Contribution: The mouse exome pipeline was utilised to discover the candidate variant and due the unknown nature of the gene prior to publication, extensive

custom work was performed in order to confirm no other candidate variants existed in the mapped region.

6) Ellyard, J. I., R. Jerjen, J. L. Martin, A. Lee, **M. A. Field**, S. H. Jiang, J. Cappello, S. K. Naumann, T. D. Andrews, H. S. Scott, M. G. Casarotto, C. C. Goodnow, J. Chaitow, V. Pascual, P. Hertzog, S. I. Alexander, M. C. Cook and C. G. Vinuesa (2014). "Whole exome sequencing in early-onset cerebral SLE identifies a pathogenic variant in TREX1." Arthritis & Rheumatology. 2014; 66(12):3382-6

Summary: "Our study is the first to demonstrate that whole exome sequencing can be used to identify rare or novel deleterious variants as genetic causes of SLE and, through a personalized approach, improve therapeutic options."

Contribution: This work used our variant detection system to identify the causal variant. Custom work to identify rare variants (<0.02 MAF) was required as the causal variant was reported in dbSNP at low frequencies in the general population.

7) Daley, S. R., K. M. Coakley, D. Y. Hu, K. L. Randall, C. N. Jenne, A. Limnander, D. R. Myers, N. K. Polakos, A. Enders, C. Roots, B. Balakishnan, L. A. Miosge, G. Sjollema, E. M. Bertram, **M. A. Field**, Y. Shao, T. D. Andrews, B. Whittle, S. W. Barnes, J. R. Walker, J. G. Cyster, C. C. Goodnow and J. P. Roose. (2013). "Rasgrp1 mutation increases naive T-cell CD44 expression and drives mTOR-dependent accumulation of Helios+ T cells and autoantibodies." Elife 2013; 2: e01020.

Summary: "Here we analyze a new mouse missense variant, *Rasgrp1* with an ENU-mutated EF hand in the Rasgrp1 Ras guanine nucleotide exchange factor. *Rasgrp1* mice exhibit anti-nuclear autoantibodies and gradually accumulate a CD4 Helios$^+$ PD-1$^+$ CD4$^+$ T cell population that is dependent on B cells. Despite reduced Rasgrp1-Ras-ERK activation in vitro, thymocyte selection in *Rasgrp1* is mostly normal in vivo, although CD44 is overexpressed on naïve thymocytes and T cells in a T-cell-autonomous manner."

Contribution: The mouse exome pipeline was utilised to discover the causal variant

8) Bergmann, H., M. Yabas, A. Short, L. Miosge, N. Barthel, C. E. Teh, C. M. Roots, Bertram, F. Mackay, A. J. Rimmer, R. J. Cornall**, M. A. Field**, T. D. Andrews, C. C. Goodnow and A. Enders (2013). "B cell survival, surface BCR and BAFFR expression, CD74 metabolism, and CD8- dendritic cells require the

intramembrane endopeptidase SPPL2A." Journal of Experimental Medicine 2013; 210(1): 31-40.

Summary: "In this study, we show that mice with an inactivating mutation in the intramembrane protease signal peptide peptidase–like 2A (SPPL2A) unexpectedly exhibit profound humoral immunodeficiency and lack mature B cell subsets, mirroring deficiency of the cytokine B cell–activating factor (BAFF). The findings illuminate an important role for the final step in the CD74–MHC II pathway and a new target for protease inhibitor treatment of B cell diseases."

Contribution: The mouse exome pipeline was utilised to discover the causal variant

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
2. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001;29(1):308-11.
3. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Human mutation. 2003;21(6):577-81.
4. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. 2014;42(Database issue):D980-5.
5. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biology. 2015;13(7):e1002195.
6. Greene CS, Tan J, Ung M, Moore JH, Cheng C. Big data bioinformatics. Journal of Cellular Physiology. 2014;229(12):1896-900.
7. Scriver CR. Garrod's Croonian Lectures (1908) and the charter 'Inborn Errors of Metabolism': albinism, alkaptonuria, cystinuria, and pentosuria at age 100 in 2008. Journal of Inherited Metabolic Disease. 2008;31(5):580-98.
8. Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. Science. 2001;291(5507):1224-9.
9. Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(2):560-4.
10. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
11. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biology. 2007;5(10):e254.
12. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British Journal of Cancer. 2004;91(2):355-8.
13. Barba M, Czosnek H, Hadidi A. Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses. 2014;6(1):106-36.
14. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Medicine. 2010;2(11):84.
15. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics. 2010;42(1):30-5.
16. Ellyard JI, Jerjen R, Martin JL, Lee A, Field MA, Jiang SH, et al. Whole exome sequencing in early-onset cerebral SLE identifies a pathogenic variant in TREX1. Arthritis & Rheumatology. 2014. Dec;66(12):3382-6
17. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nature Genetics. 2015;47(5):435-44.

18. Zhu J. A year of great leaps in genome research. Genome Medicine. 2012;4(1):4.
19. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
20. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends in Genetics : TIG. 2014;30(9):418-26.
21. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics. 2014;15(2):256-78.
22. Brownstein CA, Beggs AH, Homer N, Merriman B, Yu TW, Flannery KC, et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biology. 2014;15(3):R53.
23. Paten B, Diekhans M, Druker BJ, Friend S, Guinney J, Gassner N, et al. The NIH BD2K center for big data in translational genomics. Journal of American Medical Informatics Assosiation 2015;22(6):1143-7.
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297-303.
25. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology. 2010;11(8):R86.
26. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004;20(17):3045-54.
27. Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics. 2012;28(11):1525-6.
28. Lam HY, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, et al. Performance comparison of whole-genome sequencing platforms. Nature Biotechnology. 2012;30(1):78-82.
29. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, Schuster SC. Comparison of sequencing platforms for single nucleotide variant calls in a human sample. PloS One. 2013;8(2):e55089.
30. Field MA, Cho V, Andrews TD, Goodnow CC. Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. PloS One. 2015;10(11):e0143199.
31. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics. 2014; 12;15:104
32. Liotta L, Petricoin E. Molecular profiling of human cancer. Nature Review Genetics. 2000;1(1):48-56.
33. Field MA, Cho V, Cook MC, Enders A, Vinuesa C, Whittle B, et al. Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees. Bioinformatics. 2015;15;31(14):2377-9.
34. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, State MW, et al. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron. 2012;76(6):1052-6.
35. Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, et al. Association between variants of PRDM1 and NDP52 and Crohn's disease, based

on exome sequencing and functional studies. Gastroenterology. 2013;145(2):339-47.
36. Andrews TD, Sjollema G, Goodnow CC. Understanding the immunological impact of the human mutation explosion. Trends in Immunology. 2013;34(3):99-106.
37. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. The New England Journal of Medicine. 2013;369(16):1502-11.
38. Taupin D, Lam W, Rangiah D, McCallum L, Whittle B, Zhang Y, et al. A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. Human Genome Variation 2. 2015; 16;2:15013
39. Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, et al. eXtasy: variant prioritization by genomic data fusion. Nature Methods. 2013;10(11):1083-4.
40. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Research. 2012;40(7):e53.
41. Koboldt DC, Larson DE, Sullivan LS, Bowne SJ, Steinberg KM, Churchill JD, et al. Exome-based mapping and variant prioritization for inherited Mendelian disorders. American Journal of Human Genetics. 2014;94(3):373-84.
42. Kamphans T, Sabri P, Zhu N, Heinrich V, Mundlos S, Robinson PN, et al. Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PloS One. 2013;8(8):e70151.
43. Wilmott JS, Field MA, Johansson PA, Kakavand H, Shang P, De Paoli-Iseppi R, et al. Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes. Pathology. 2015;47(7), pp. 683–693
44. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010;465(7297):473-7.
45. Nagy R, Sweet K, Eng C. Highly penetrant hereditary cancer syndromes. Oncogene. 2004;23(38):6445-70.
46. Burnet M. Somatic Mutation and Chronic Disease. British Medical Journal. 1965;1(5431):338-42.
47. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417(6892):949-54.
48. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. The New England Journal of Medicine. 2011;364(26):2507-16.
49. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464(7291):993-8.
50. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Research. 2012;22(8):1589-98.
51. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013;31(3):213-9.

52. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467):338-45.
53. Miosge, L. A., M. A. Field, Y. Sontani, V. Cho, S. Johnson, A. Palkova, B. Balakishnan, R. Liang, Y. Zhang, S. Lyon, B. Beutler, B. Whittle, E. M. Bertram, A. Enders, C. C. Goodnow and T. D. Andrews. Comparison of predicted and actual consequences of missense mutations. Proceedings of the National Academy of Sciences of the United States of America 2015;15;112(37):E5189-98
54. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010;38(16):e164.
55. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70.
56. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010;7(4):248-9.
57. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009;4(7):1073-81.
58. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014;46(3):310-5.
59. Hicks S, Wheeler DA, Plon SE, Kimmel M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Human Mutation. 2011;32(6):661-8.
60. Masica DL, Li S, Douville C, Manola J, Ferris RL, Burtness B, et al. Predicting survival in head and neck squamous cell carcinoma from TP53 mutation. Human Genetics. 2015;134(5):497-507.
61. Manolio TA, Chisholm RL, Ozenberger B, Roden DM, Williams MS, Wilson R, et al. Implementing genomic medicine in the clinic: the future is here. Genetics in Medicine. 2013;15(4):258-67.
62. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. American Journal of Human Genetics. 2011;88(4):440-9.
63. Kato S, Han SY, Liu W, Otsuka K, Shibata H, Kanamaru R, et al. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(14):8424-9.
64. Ohta T. Evolution by nearly-neutral mutations. Genetica. 1998;102-103(1-6):83-90.
65. TD Andrews, Y Jeelall, D Talaulikar, CC Goodnow, MA Field. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. PeerJ. 2016 24;4:e2074
66. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Science Translational Medicine. 2012;4(136):136ra68.

67. Al-Mawsawi LQ, Wu NC, Olson C, Shi V, Qi H, Zheng X, et al. High-throughput profiling of point mutations across the HIV-1 genome. Retrovirology. 2014;11(1):124.
68. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(23):9530-5.
69. Andrews TD, Whittle B, Field MA, Balakishnan B, Zhang Y, Shao Y, et al. Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. Open Biology. 2012;2(5):120061.
70. Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, et al. Genealogies of mouse inbred strains. Nature Genetics. 2000;24(1):23-5.
71. Acevedo-Arozena A, Wells S, Potter P, Kelly M, Cox RD, Brown SD. ENU mutagenesis, a way forward to understand gene function. Annual Review of Genomics and Human Genetics. 2008;9:49-69.
72. Justice MJ, Noveroske JK, Weber JS, Zheng B, Bradley A. Mouse ENU mutagenesis. Human Molecular Genetics 1999;8(10):1955-63.
73. Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, et al. Mutation discovery in mice by whole exome sequencing. Genome Biology. 2011;12(9):R86.
74. International HapMap C. Integrating ethics and science in the International HapMap Project. Nature Review Genetics. 2004;5(6):467-75.
75. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357-9.
76. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
77. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29(16):2041-3.
78. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.
79. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Human Genomics. 2014;8:14.
80. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. BioMed Research International. 2014;2014:309650.
81. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, et al. Comprehensive variation discovery in single human genomes. Nature Genetics. 2014;46(12):1350-5.
82. Dunkerton S, Field M, Cho V, Bertram E, Whittle B, Groves A, et al. A de novo mutation in KMT2A (MLL) in monozygotic twins with Wiedemann-Steiner syndrome. American Journal of Medical Genetics Part A. 2015;167A(9):2182-7
83. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463(7278):191-6.

84. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15:244.
85. Stanley CM, Sunyaev SR, Greenblatt MS, Oetting WS. Clinically relevant variants - identifying, collecting, interpreting, and disseminating: the 2013 annual scientific meeting of the Human Genome Variation Society. Human Mutation. 2014;35(4):505-10.
86. Casadevall A, Pirofski LA. Exploiting the redundancy in the immune system: vaccines can mediate protection by eliciting 'unnatural' immunity. The Journal of Experimental Medicine. 2003;197(11):1401-4.
87. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(3):1056-61.
88. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(49):19872-7.
89. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
90. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996;5(3):299-314.
91. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nature Genetics. 2009;41(2):149-55.