UNIVERSITY OF CALIFORNIA, SAN DIEGO

Limiting Bias in Biological Data Analysis by Pooling Information

A Dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Kunal Bhutani

Committee in charge:

Professor Nicholas J. Schork, Chair
Professor Vineet Bafna, Co-Chair
Professor Vikas Bansal
Professor Lawrence Goldstein
Professor Olivier Harismendy
Professor Diane Nugent

2017

The Dissertation of Kunal Bhutani is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____
Co-Chair

____
Chair

University of California, San Diego

2017

DEDICATION

To my parents.

EPIGRAPH

In the depth of winter,

I finally learned that within me there lay an invincible summer.

— Albert Camus

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks to all the friends, family, collaborators, committee members, students, teachers, and strangers that made this journey worthwhile.

Chapter 2, in full, is a reprint of the material as it appears in *Nature Communications* 2016. Kunal Bhutani, Kristopher L Nazor, Roy Williams, Ha Tran, Heng Dai, Zeljko Dzakula, Edward H Cho, Andy WC Pang, Mahendra Rao, Han Cao, Nicholas J Schork, Jeanne F Loring. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part is currently being prepared for submission for publication of the material. Kunal Bhutani, Abhishek Sarkar, Yongjin Park, Manolis Kellis, Nicholas J Schork. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part is currently being prepared for submission for publication of the material. Kunal Bhutani, Victoria Magnuson, Alexandra Buckley, Danjuma Quarless, Laura Goetz, Nicholas J Schork. The dissertation author was the primary investigator and author of this paper.

VITA

2011	B. S. in Biomedical Engineering, University of Texas at Austin
2017	Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

Kim S*, Jeong K*, **Bhutani K**, Lee JH, Patel A, Scott E, Nam H, Lee H, Gleeson JG, Bafna V. "Virmid: accurate detection of somatic mutations with sample impurity inference". *Genome biology.* 14(8). 2013

Bhutani K, Nazor KL, Williams R, Tran H, Dai H, Dzakula Z, Cho EH, Pang AW, Rao M, Cao H, Schork NJ, Loring JF. "Whole-genome mutational burden analysis of three pluripotency induction methods". *Nature Communications.* 7. 2016.

Krishnaswami SR*, Grindberg RV*, Novotny M, Venepally P, Lacar B, **Bhutani K**, Linker S, Pham S, Erwin JA, Miller JA, Hodge R, McCarthy JK, Kelder M, McCorrison J, Aevermann BD, Fuertes FDR, Scheuermann RH, Lee J, Lein ES, Schork NJ, McConnell MJ, Gage FH, Lasken RS. "Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons". *Nature protocols.* 11(3). 2016.

Buckley AR, Standish KA, **Bhutani K**, Ideker T, Harismendy O, Carter H, Schork NJ. "Pan-Cancer Analysis Reveals Technical Artifacts in The Cancer Genome Atlas (TCGA) Germline Variant Calls". *bioRxiv.* 2016.

Bhutani K*, Sarkar A*, Park Y, Kellis M, Schork NJ. "Modeling prediction error improves power of transcriptome-wide association studies". *bioRxiv.* 2017.

Park Y*, Sarkar A*, **Bhutani K**, Kellis M "Multi-tissue polygenic models for transcriptome-wide association studies". *bioRxiv.* 2017.

Bhutani K*, Magnuson V*, Buckley A, Quarless D, Goetz L, Schork NJ. "Longitudinal metabolome, microbiome, and transcriptome profiling of a germline TP53 mutation carrier". *In Preparation.* 2017.

ABSTRACT OF THE DISSERTATION

Limiting Bias in Biological Data Analysis by Pooling Information

by

Kunal Bhutani

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Nicholas J. Schork, Chair
Professor Vineet Bafna, Co-Chair

Innovations in the design and implementation of high-throughput technologies has shifted biological research from hypothesis-driven inquiries to large data-driven studies. Scientists can now jointly interrogate the genome, transcriptome, metabolome, microbiome, and dozens of other molecular systems to develop more complete, interconnected pictures of biological states. However, accurate interpretation of each state requires a thorough understanding of the sources of variation associated with the underlying assays and experimental approaches used. Here, I pool information from related sources based on known generative processes to model variation and limit biases in the analysis of three different biological phenomenon. First, I discuss jointly identifying genomic variants in induced pluripotent stem cells derived from the same fibroblast population to assess the mutational burden of three different reprogramming methods: retroviral transfection, Sendai virus, and non-integrating mRNA. The research suggests that each method induces new mutations, but there are no obvious systematic differences in the types of mutations nor in the genomic regions harboring them. Shifting to transcriptomics, I next model uncertainty and variation in imputed expression in transcriptome-wide association studies. I show through simulations that a novel Bayesian method that pools multiple models of transcription regulation outperforms current methodologies in identifying associations between imputed gene expression and a phenotype. In an application to seven diseases from the Wellcome Trust Case Control Consortium, the method finds 42 associations, 17 of which have not yet been previously identified by GWAS or differential gene expression analyses in case-control cohorts. Finally, I describe results from a study exploring longitudinal profiles of the metabolome, microbiome, and transcriptome of a young female germline TP53 mutation carrier. The motivation for this study was to determine if any health status changes might occur in this carrier that could be indicative of tumor formation given her extremely high cancer susceptibility. I utilize a Bayesian model to separate metabolite variation from instrumentation variation by calculating latent metabolite levels across multiple instrumentation runs. Fortunately, I do not find obvious and statistically deviations from baseline for any biomarker indicate of cancer, but I highlight power limitations in such study designs. Together, these three works demonstrate the importance and utility of pooling information to limit biases in contemporary high-throughput, data intensive biological analyses.

Chapter 1

Introduction

Innovations in the design and implementation of high-throughput technologies have revolutionized approaches to the biomedical sciences. For example, with an estimated 10^{21} bases sequenced a year by next-generation sequencing instruments [SLF+15], genomically-guided biological research has shifted from hypothesis-driven inquiries to large data-driven studies. In addition, scientists can now jointly interrogate the genome, transcriptome, metabolome, microbiome, and dozens of other molecular systems to develop more complete, interconnected pictures of biological states, such as pathologies and precursors to disease. However, accurate interpretation of a biological state requires a thorough understanding of the sources of variation associated with each assay and experimental approach used in order to avoid false claims about that state. In this thesis, I discuss pooling information from different, yet relevant, sources based on known generative processes in order to model and account for variation in, e.g., genomics, transcriptomics, and metabolomics data in specific applications. In the sections below, I briefly describe settings in which I was able to leverage different sources of information, model their influence on the phenomenon under study using computational and data analysis techniques and address important needs in the biomedical sciences. These settings include the identification somatically-acquired cancer mutations against a background of normal cells contaminating a cancer cell sample, the characterization of the mutational landscape of induced pluripotent stem cells, the identification of intermediate phenotypes impacting a specific clinically-relevant phenotype using combined GWAS and imputation methods, and the analysis of the longitudinal metabolic profile of a patient who is highly susceptible to cancer.

1.1 Genomics and the Identification of Somatically Acquired Variations

With the advent of next-generation sequencing, it is possible to characterize the genomes of a population of cells from a single individual or across a set of individuals. Studying somatic cells from an individual could lead to insights about not only the natural diversity of such cells, but also about the events, probably occurring during cell replication, that could lead to disease states. This has far-reaching implications for studying the origins and pathogenesis of many diseases, such as cancer, where there is known to be a large somatic mutation-induced cell dysfunction component. In addition, characterization of cell populations can also lead to insights into the genomic stability of cells being considered in therapeutic cell replacement or transplant strategies. However, the analysis of the genomic and mutational landscape of such cells is limited by technological constraints associated with contemporary DNA sequencing-based assays. For example, the depth of coverage achieved in a sequencing study, as well as base call and sequencing read-mapping errors, must be considered when applying next-generation sequencing (NGS) to the study of cell populations. In the presence of the many potential factors that may confound a study, Bayesian methods have been developed and employed to accurately accommodate them [KZL$^+$12, MHB$^+$10].

1.1.1 Somatic Mutations in Cancer

In the specific context of cancer, emphasis is placed on identifying somatic variants that are unique to a population of diseased cells, but do not exist in normal cells. For such studies, several tools have been developed that compare the sequencing reads matching the reference genome from the normal and disease (i.e., cancerous) tissues and highlight variants unique to diseased population with a level of confidence[CLC$^+$13, KZL$^+$12, RDM$^+$12, SWS$^+$12]. However, most of these tools fail when the samples cellular composition is heterogeneous, particularly when the relative fraction of diseased and normal cells is unknown.

In order to overcome limitations with many of the available tools for identifying DNA sequence variants present in cancerous cells but not present in normal cells, I worked with a team to create the Virtual Microdissection for Variant Calling (Virmid) analysis tool, which leverages a Bayesian framework for detecting somatically-acquired single nucleotide variants (SNVs) in paired normal/disease samples[KJB$^+$13]. As shown in Figure 1.1, the program explicitly models the fraction of normal cells contaminating the diseased sample, α, by pooling information across genomic loci.

This contamination, as a generative process, affects the relative fraction of diseased cells carrying a mutant (i.e., somatically acquired variant) allele in the sample. Virmid models this contamination in a Bayesian probabilistic setting for evaluating the likelihood of a site being mutated in the the faction of the sampled cells that are cancer cells. This formulation is a viable option for cancerous diseases such as focal cortical dysplasia and hemimegalencephaly, in which its impossible to acquire normal brain cells in addition to diseased ones. Additionally, gastric and breast cancer tissues are normally contaminated by stromal cells and hence require some way of accommodating the non-cancerous cells in an analysis. In such scenarios, if the cancer is dominated by one subclone, the proposed method should be able to accurately identify mutated sites.

In addition to the Bayesian framework, Virmid also applies a set of filters that limits biases arising from poorly aligned reads and/or platform-specific errors. Briefly, these focus on mapping quality, biases in the location of the variant allele in a read, proximity to indels and base quality statistics[CLC$^+$13]. Virmid has been shown to be able to effectively identify low frequency variants in a tumor sample in many settings, for example achieving 98.61% sensitivity to identify somatically acquired mutations in breast cancer samples that were missed by other variant callers, such as Mutect[CLC$^+$13] (97.16%), JointSNVMix[RDM$^+$12] (84.58%) and Strelka[SWS$^+$12] (97.82%), because of high contamination of normal cells.

1.1.2 Induced Pluripotent Stem Cells

Multi-sample DNA sequence variant calling techniques have their roots in computational methods for use in population genetic studies seeking to pool DNA sequencing data across a number of individuals and jointly identify variants in all those individuals simultaneously. These strategies underlying assumption is that the population of individuals are part of the same lineage and share many variants. If a genomic position has several reads matching a variant allele in an individual, but

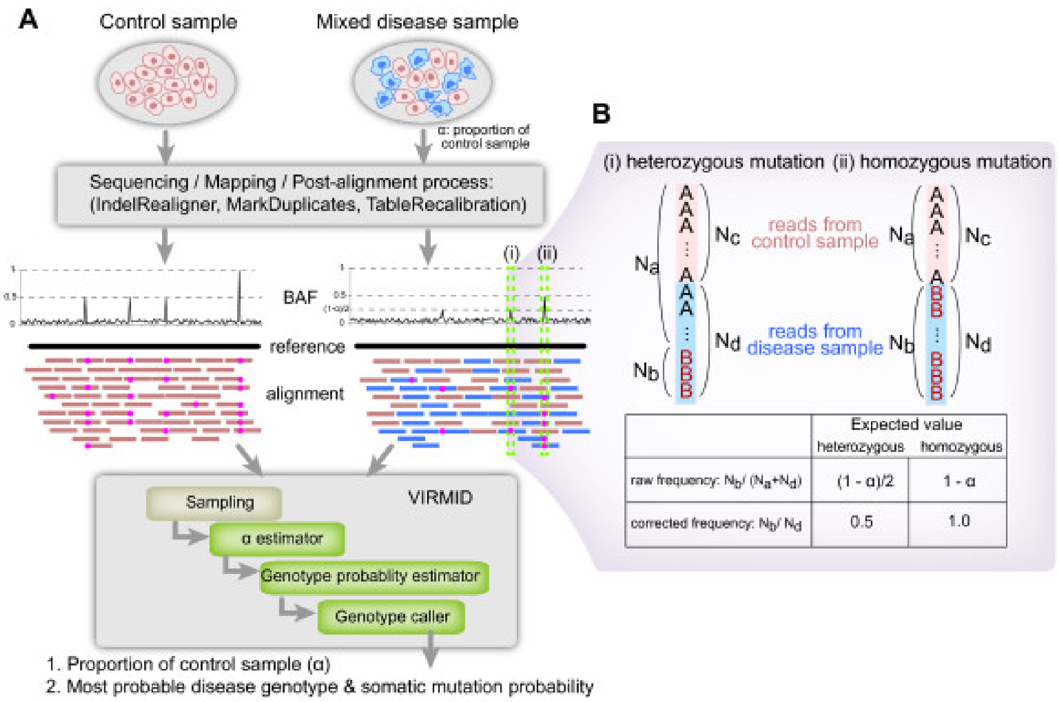

A
Control sample
Mixed disease sample
α: proportion of control sample
Sequencing / Mapping / Post-alignment process:
(IndelRealigner, MarkDuplicates, TableRecallibration)
1
0.5
0
BAF
1
0.5
(1-α)/2
0
(i)
(ii)
reference
alignment
Sampling
VIRMID
α estimator
Genotype probablity estimator
Genotype caller
1. Proportion of control sample (α)
2. Most probable disease genotype & somatic mutation probability
B
(i) heterozygous mutation (ii) homozygous mutation
Na
Nc
reads from control sample
Na
Nc
Nb
Nd
reads from disease sample
Nb
Nd
A A A ⋮ A A A ⋮ B B B
A A A ⋮ A B B ⋮ B B B
Expected value
heterozygous
homozygous
raw frequency: Nb/ (Na+Nd)
(1 - α)/2
1 - α
corrected frequency: Nb/ Nd
0.5
1.0

Figure 1.1: Overall Virmid workflow. (a) Disease/control paired data are used (top) to generate an alignment (BAM) file. The mixed disease sample produces short reads of mixed types (blue and orange rectangles). Somatic mutations, where the control has the reference genotype (AA) and the disease has the non-reference (AB or BB, red dots in the alignment), are hard to detect if there is high contamination due to the significant drop in B allele frequency (BAF). Virmid takes the disease/control paired data and analyzes: (1) the proportion of control cells in the disease sample (α) and (2) the most probable disease genotype for each position that can be used to call somatic mutations. (b) An example BAF drop. Without contamination, the expected BAF is 0.5 and 1.0 for heterozygous and homozygous mutations sites, respectively. When there is control sample contamination, α, mutation alleles are observed only in (1 - α) of the whole reads. So the expected BAF drops to (1 - α)/2 and (1 - α). With an accurate estimate of α, Virmid can detect more true somatic mutations, which would be missed by conventional tools due to insufficient observation of B alleles. BAF, B allele frequency.

only a few in another individual, most models consider the variant as present in both individuals with high probability. However, for a variant to be assigned only in a single individual, the models would require several reads supporting the existence of the variant allele.

In Chapter 2, I discuss an application of multi-sample variant calling to study the genomic stability of cells derived from induced pluripotent stem cell (iPSC) technology that might be considered in therapeutic cell replacement strategies [BNW$^+$16]. The study compared the mutational burden induced by three reprogramming methods for transforming the same population of fibroblasts into iPSCs: a technique based on the use of retroviruses, Sendai virus, and non-integrating mRNAs. Previous studies have assessed the genomic integrity of iPSCs by using paired normal and disease cell methodologies and subtracting mutations in fibroblasts from mutations found in each iPSC colony separately. In contrast, I identified mutations associated with the reprogramming techniques by considering all samples simultaneously.

It is important to note that the predominant theory behind techniques for inducing pluripotency is that the overexpression of the Yamanaka transcription factors (Oct3/4, Sox2, Klf4, c-Myc) leads to a cell-state switch for a single fibroblast (or other cell) into a progenitor cell, which divides several times into a colony of iPSCs [TY06]. Under the aforementioned theory, if a mutation occurs as a result of reprogramming, it will be present in all cells of a colony and, under the infinite sites assumption [McV02], no two colonies will share a mutation resulting from reprogramming. By treating each iPSC colony and the progenitor cell fibroblasts as 'individuals', I show that pooling information using multi-sample variant calling is better suited for identifying mutations resulting from reprogramming than other methods.

1.2 Transcriptomics and Transcriptome-wide Association Studies

Genome-wide association studies (GWASs) test for linear associations between variant alleles and phenotypes using large populations of individuals[LS94]. Variant alleles occur throughout the genome and number in the millions. Most risk loci identified to-date in large GWAS initiatives lie outside of protein-coding regions and provide no clear mechanistic or intervention insights to treat diseases[WMM^{+}14]. Transcriptome-wide association studies (TWASs) extend the GWAS strategy by finding linear associations between imputed gene expression values and phenotypes[GWS^{+}15]. They use reference expression data to build models of transcription regulation that relate variant alleles to gene expression values and then impute or assign gene expression values to individuals with genotype data in GWAS cohorts where expression has not been measured directly. Ultimately, a TWAS tests the association between imputed or assigned gene expression values and phenotype in GWAS data sets to, e.g., identify possible therapeutic targets.

Current methods for TWASs use a single model of transcriptional regulation, and do not take into account the uncertainty in imputed expression. In Chapter 3, I pool information across multiple models of transcription regulation to quantify uncertainty in imputed expression values and propagate it through the association testing with the phenotype of interest. I show that this strategy of combining or pooling the different transcriptional models when testing associations results in increased power and led to the identification of several possible therapeutic targets for seven different complex diseases that are missed by current analytical methods used in TWAS.

1.3 Metabolomics and Li-Fraumeni Syndrome

Single subject or N-of-1 studies and trials monitor health information in single individuals over time and study deviations from baseline measurements that might be indicative of a health status change worth further scrutiny or provide evidence that a therapeutic intervention is working. Such approaches are the hallmark of new initiatives to promote and test precision (i.e., or personalized or individualized) medicine and care, where interventions are tailored to an individuals profile instead of relying on the belief that interventions will work on everyone in the general population[Sch15]. In Chapter 4, I discuss a suite of analytical techniques for an n-of-1 molecular surveillance study of an individual a germline p53 mutation, which causes Li-Fraumeni syndrome[LFJ69]. Li-Fraumeni patients have a high incidence of cancer during their lifetime and often undergo regular clinical surveillance, including several imaging and biochemical screens for cancer. The analytical techniques were designed to accommodate a wide variety of assays and sources of variation, particularly those associated with longitudinal sampling.

To showcase the analytical techniques, we longitudinally profiled cancer biomarkers in the metabolome, microbiome, and transcriptome of the patient over 16 months. I noticed incongruence across instrumentation runs when identifying potentially clinically meaningful outliers and linear trends in the metabolome data. To combat this, I created a Bayesian method that pools information across multiple metabolome instrumentation runs to separate biologically-meaningful metabolite variation from instrumentation or technical variation. Fortunately, I did not find any significant deviations from baseline which might be indicative of tumor formation, but I highlight power limitations in such study designs.

Chapter 2

Whole Genome Mutational Burden Analysis of Three Pluripotency Induction Methods

2.1 Abstract

There is concern that the stresses of inducing pluripotency may lead to deleterious DNA mutations in induced pluripotent stem cell (iPSC) lines, which would compromise their use for cell therapies. Here we report comparative genomic analysis of nine isogenic iPSC lines generated using three reprogramming methods: integrating retroviral vectors, non-integrating Sendai virus, and synthetic mRNAs. We used whole genome sequencing and de novo genome mapping to identify single nucleotide variants, insertions and deletions, and structural variants. Our results show a moderate number of variants in the iPSCs that were not evident in the parental fibroblasts, which may result from reprogramming. There were only small differences in the total numbers

and types of variants among different reprogramming methods. Most importantly, a thorough genomic analysis showed that the variants were generally benign. We conclude that the process of reprogramming is unlikely to introduce variants that would make the cells inappropriate for therapy.

2.2 Introduction

Cell replacement therapies using cells derived from human pluripotent stem cells (embryonic stem cells (hESCs) and induced pluripotent stem cells (iPSCs) have been approved for clinical trials for macular degeneration, spinal cord injury, and Type 1 diabetes. The limited data available to date indicate that there are no adverse events in patients receiving these cells [HBV+11]. However, there continues to be discussion about the theoretical chance that these transplanted cells may develop into tumors or cause other pathologies. The discussion has come to focus on induced pluripotent stem cells, largely due to concerns that the massive epigenetic remodeling that occurs during reprogramming might cause genomic mutations that could make the cells tumorigenic. These concerns have led multiple groups to study the genomic integrity of iPSCs using methods that include SNP genotyping [HBV+11, LUS+11], CGH[MBDL+10], karyotyping [TNN11], and exome sequencing[GLF+11] (reviewed in[PL14, SDB+14]). In each case the focus has been exclusively on a single type of genomic alteration, rather than considering the combined effects of single nucleotide variants (SNVs), structural variants (SVs), and copy number variations (CNVs). Further, detailed comparative genomic analyses of iPSC lines that have been generated via distinct reprogramming methodologies have yet to be reported. In this study we assessed genome-wide mutation rates from replicate isogenic cell lines generated by three distinct methods. We used integrating viral (retrovirus), non-integrating viral (Sendai virus), and non-integrating non-viral (mRNA) reprogramming strategies to introduce exogenous expression of POU5F1, SOX2, KLF4 and MYC in separate fractions of a single fibroblast population (Figure 2.1a). Three clonal lines were established from iPSCs generated by each method and determined to be pluripotent by standard measures. In order to detect SNVs, SVs, and CNVs within each line and the parental fibroblast population, we generated whole genome sequencing (WGS) data for each iPSC line and the parental fibroblasts at an an average read depth of 39 fold, with 93.7% of the autosomal genome covered by at least 10 reads. In addition, to assess chromosomal rearrangements and large structural variants with high resolution, we performed whole genome mapping using the recently developed Irys technology (BioNano Genomics (BNG), San Diego, CA). We detected subtle differences in the numbers of variants depending on the method, but rarely found mutations in genes that had any association with increased cancer risk. We conclude that mutations that have been reported in iPSC cultures are unlikely to be caused by their reprogramming, but instead are probably due to the well-known selective pressures that occur when hPSCs are expanded in culture.

2.3 Results

2.3.1 Identification of single nucleotide variants

To characterize the mutational burden in iPSCs, we identified SNVs that were unique to each iPSC cell line by integrating results from HaplotypeCaller[MHB^{+}10] and MuTect[CLC^{+}13], as described in the Methods. For variant calling with HaplotypeCaller, we treated all ten samples (the parent fibroblast population and three biological replicates for each reprogramming method) as part of a single population using the multi-sample option. This pipeline was tuned for the identification of recurrent variations in population studies, and therefore enabled us to have higher specificity in classifying reprogramming-induced mutations by accounting for mosaicism in the parental fibroblast population. To gain a more sensitive assessment of the SNV landscape across the iPSC samples, we also called variants using MuTect, wherein each iPSC cell line was compared to the parental fibroblast population in an analogous manner to which tumor samples are compared to normal tissue in oncogenomic studies Figure (2.1c; Supplementary Figure S2.1). Taken together, the results from these two distinct variant calling pipelines gave us higher confidence in our ability to identify true variants through the moderation of type I and type II errors, respectively.

The identified set of putative unique variants was split into three groups according to our confidence in the variant calls: Variant Set 1 was called unique by both MuTect and HaplotypeCaller; Variant Set 2 had coverage between 20-60x and allele frequency distribution between 0.4 and 0.6 but was called only by MuTect, and Variant Set 3 comprised those with allele frequencies between 0.2-0.4 and 30-50x coverage. Variant validation by qPCR from each of these three groups indicated that the mutations from Variant Set 1 had the highest likelihood of being true somatic variants (Supplementary Data 1; Methods). Therefore, subsequent analyses were restricted to these variants, which likely occurred during the initial doublings of the founder populations of the iPSCs compared to variants that are in lower allelic fractions that could arise as a result of proliferation of the cells. Our filtering strategy focuses on these high confidence variants and also removes variants that are found in low complexity regions and dbSNP (Figure 2.1c).

2.3.2 Identification of Insertions and Deletions

Due to the specificity of multi-sample HaplotypeCaller in identifying high confidence SNVs, HaplotypeCaller was also used for identifying Insertions and Deletions (indels).

2.3.3 Identification of Structural Variants

Structural variations (SVs), 10 kbp to 1 Mbp in size, are common in the human genome, but challenging to assess by sequencing alone. Genome mapping in nanochannel arrays provides a single-molecule platform complementary to DNA sequencing for accurate genome assembly and SV analysis (Irys System, BioNano Genomics[DAA$^+$10, HDS$^+$13, LHL$^+$12, XPH$^+$07]. Unamplified genomic DNA, fluorescently labeled at a seven-base sequence motif and by the intercalating dye YOYO-1, was linearized by electrophoresis into 50 nm channels. An array of approximately 12,000 nanochannels was imaged for each sample, and the process was repeated multiple times, producing data at a throughput of about 1.5 Gbp/hour. Only molecules 150 kbp and longer were used to create a de novo assembly of the complete genome. We analyzed the fibroblast control and one iPSC line chosen at random from each of the three reprogramming methods.

We collected 160 Gbp (50x) of high molecular weight DNA for each sample, approximately 500,000 single molecule maps with the minimal length of 150 kbp. Data from each sample were assembled using an in-house developed de novo assembler based on Overlap-Layout-Consensus paradigm[Ana01, Ngu10, VSZW06]. The N50 length is the length for which the collection of genome maps of that length or greater cover more than 50% of the total genome length. The genome maps in this study had a N50 of ¿ 0.9 Mbp, overlapping close to 90% of the Genome Reference Consortium Human genome (GRch37) (Methods; Supplementary Figure S2.2; Supplementary Table S2.1).

2.3.4 Assessing pathogenicity of unique variants

To assess the functional consequences of the variants, the sets of high confidence unique SNVs and indels were annotated using the SGAdviser[PSE$^+$15] and Oncotator[RLG$^+$15] program suites. They were further characterized based on overlaps with ENCODE annotated genomic regions. As expected, most mutations fell within intergenic or intronic regions, with the rate of coding mutations in the range of two to ten mutations per cell line (Figure 2.2d). The potential for variants to be oncogenic was assessed by measuring their overlap with cancer genes, transcription factor binding sites (TFBS), MutSig genes, and Familial Syndrome Cancer Genes (Methods). We then compared the frequency of the annotated somatic mutations across the three reprogramming methods using an exhaustive permutation testing procedure and one vs. all contrasts in ANOVA . All of the iPSC lines had hundreds of variants compared to the parental fibroblasts, with the mRNA-derived lines having fewer high confidence mutations on average than the other methods (Figure 2.2a; Table 2.1, 2.2; Supplementary Table S2.2). We did not find evidence that any one of the methods was more likely than the others to cause oncogenic or deleterious mutations, but we identified trends based on one versus all ANOVA contrasts that linked reprogramming strategies to certain mutations. The cell lines reprogrammed using mRNA had several variants that overlapped binding sites for a transcription factor (EZH2; enhancer of zeste 2 polycomb repressive complex 2 subunit), although it is unknown if these would disrupt transcription factor binding. The retroviral induced cell lines harbored more potentially damaging mutations than the other methods (Table 2.3). In addition, the three retrovirally reprogrammed cell lines contained integrated vectors in 7, 8, and 12 sites in the genome, and while the majority of integration sites were intergenic, some integrations mapped to coding regions (Supplementary Figure S2.3). In Sendai virus reprogrammed cell lines there was a trend toward fewer coding mutations. However, we want to emphasize that based on the variance and means of the aggregated statistics, this study was not sufficiently powered to assess the differences among the different reprogramming methods for some variant classifications (Table 2.1).

We also characterized the context of the mutations as well as the overall transition/transversion rates for the different samples (Figure 2.2b). Mutational context analysis revealed no realizable difference among the different reprogramming methods and identified no links to known cancer-related mutational signatures[KMV+13]. Finally, we looked at the Combined Annotation Dependent Depletion (CADD)[KWJ+14] score distribution of the variants, and although we saw statistically significant differences among the three methods, the scores for all methods fell mostly within the non-deleterious range of less than 15 CADD Score (Supplementary Figure S2.4).

Using the variant-calling algorithms for the data from de novo whole genome mapping of four of the samples, we called 259 insertions and deletions in the parental fibroblast sample, 239 in the retroviral sample, 248 in the Sendai sample and 268 in the mRNA sample. The size of the variants ranged from 2.8 kbp to 4.9 Mbp. Using a 50% reciprocal size overlap cutoff, we identified variants that were shared among the samples. After manual inspection to eliminate false positives and false negatives, we found no cell line-specific variants in the retroviral and Sendai samples, but one deletion in the mRNA-reprogrammed sample. The deletion in this line was a heterozygous 228.8 kbp deletion at Xp22.11, which removed one copy of the PHEX gene (Phosphate-regulating neutral endopeptidase) and one copy of Mir_548. This deletion is illustrated in Figure 2.3, which shows the assembly of part of the X chromosome for each cell line, as well as the single-molecule data supporting two haplotypes for this region in the mRNA sample. We subsequently looked for this deletion in the sequencing data and identified it in the same sample; it was not present in any of the other samples. This suggests that this iPSC line was derived from a rare fibroblast containing the deletion or that the deletion was acquired very early in the reprogramming process.

2.4 Discussion

Our assessment of the mutation profiles associated with three widely used reprogramming methods for generating iPSCs indicates that all of the reprogramming approaches add to the mutational load of cells, but there were subtle differences among the methods. Although we found that the non-integrating mRNA reprogramming technique resulted in fewer total mutations than either retrovirus or Sendai virus-based reprogramming methods, mRNA-iPSCs had a greater number of mutations in binding sites for a transcription factor (EZH2). In addition, the only large structural variation (a 228.8 kbp deletion) we detected was in an mRNA-reprogrammed iPSC line. In contrast, the Sendai virus samples had fewer coding mutations than the other methods. Our results using retroviral vectors showed that this method caused a similar number of mutations as the other methods, but was slightly more likely to introduce mutations that are classified as deleterious. But the main concern with retroviral vectors is the fact that they insert into the genome. Genomic insertions have a low but finite chance of disruption of active genes or regulatory regions, and can cause cancers if they activate endogenous oncogenes[BvKS$^+$04]. Retroviral vectors were used in the first methods developed for reprogramming[TY06] but were reported to be capable of reactivation, which resulted in tumors in mice[OIY07]. Use of retroviral vectors for reprogramming has become less popular as delivery methods for transiently expressing the reprogramming factors without being inserted, such as the Sendai viral vectors and mRNA used here, have become more efficient.

Given the potential practical significance of our findings to clinical applications for stem cells, it is important to appreciate some of the biological context surrounding our experiments. We focused on the high confidence variants that differed from the parental fibroblast population. These variants likely arose during the initial doublings of the founder population of iPSCs, but we cannot rule out their origin in a minority population in the heterogeneous parental fibroblasts that was undetectable by sequencing. We note that there were sequencing reads that support several additional lower variant allele frequency mutations in the cell lines as well. It should be noted that the cells we analyzed were cultured for a relatively short time, and that variants in an early-arising small subpopulation could become more dominant over time if they give the cells a selective advantage in culture.

In order to move forward toward applying stem cell-based therapies for human disease, it is important to focus our efforts on improving the likelihood that there will be no adverse effects of these therapies. Our study was more extensive than previous analyses, but larger studies of this depth are still needed. While results of our study do not rule out the possibility that reprogramming cells could introduce oncogenic mutations that compromise the safety of iPSCs, they should alleviate some of the concern about how likely it is that reprogramming itself would cause dangerous genomic changes that could lead to harm to transplant recipients. It is important to note that genomic aberrations, some of which could be oncogenic, are known to occur during the considerable expansion of cells that is required for clinical applications. While development of new methods for reprogramming will make the process simpler and less expensive, it is critical at this stage that we concentrate on monitoring the appearance and potential consequences of mutations that arise during cell division and differentiation in culture and are selected for by the culture conditions.

2.5 Methods

The design of our study was to evaluate mutation profiles associated with the three different reprogramming strategies (see below for details on each method) we considered, as well as the characterization of different forms of variation and the analysis of the variation within and across the different reprogramming strategies

2.5.1 Retroviral Reprogramming

PLAT-A packaging cells (Cell Biolabs, Inc.) were plated onto six well plates coated with Poly-D-Lysine at a density of 1.5x106 cells per well without antibiotics and incubated overnight. Cells were transfected with 4 μg of Moloney murine leukemia based retroviral vectors (pMXs) containing the human cDNA of POU5F1, SOX2, KLF4 or MYC (Addgene catalog number 17217, 17218, 17219, and 17220 respectively) by Lipofectamine 2000 (Life Technologies) according to the manufacturers instructions. Viral supernatants were collected at 48 and 72 hours post-transfection, filtered through a 0.45 μm pore-size filter. 200,000 Human dermal fibroblasts (Science cell Catalog #2300) were seeded onto each well of a 6 well plate overnight prior transfection. Equal volumes of fresh 48 hour and 72 hour viral supernatants containing each of four retroviruses supplemented with 6 μg/ml of Polybrene (Sigma) were added onto the cells on day one and day two, respectively. On day five, the transduced cells were split onto MEFs at a density of 104 cells per well of a six well plate in hESC medium supplemented with 0.5mM Valproic Acid (Stemgent). Cells were fed every other day with valproic acid (VPA) supplemented hES medium for 14 days before VPA was withdrawn. Individual iPSC colonies were manually picked and clonally expanded three weeks post transduction and transferred onto MEF plates.

2.5.2 Sendai Virus Reprogramming

Human dermal fibroblasts (Science cell Catalog #2300) were reprogramed according to the manufacturers instructions (CytoTune-iPS 2.0 Sendai Reprogramming Kit, Life technology catalog number A1378001). Cells were transduced with Sendi viruses containing the Yamanaka factors, and individual iPSC colonies were identified by morphology,. The colonies were manually picked, expanded for three weeks post transduction, and transferred onto a feeder layer of irradiated mouse embryonic fibroblasts (MEFs).

2.5.3 mRNA Reprogramming

Human dermal fibroblasts (Science cell Catalog #2300) were reprogramed according to the manufacturers instructions (The Stemgent mRNA Reprogramming Kit catalog number 00-0071). Individual iPSC colonies were manually picked and clonally expanded three weeks post transfection and transferred onto MEF feeder layers.

2.5.4 Cell Culture

Plat-A Packaging cells (Cell Biolabs, Inc.) were maintained according to the manufacturers instructions. Human dermal fibroblasts (Science cell Catalog #2300) were cultured in Dulbeccos modified eagle medium (DMEM), 2mM GlutaMax, 10% fetal bovine serum and 0.1 mM non-essential amino acids (Life Technologies). iPSCs were generated and maintained in standard hESC medium containing DMEM/F12 supplemented with 20% Knockout Serum Replacement (Life Technologies), 2mM GlutaMAX, 0.1 mM nonessential amino acids, 0.1mM 2-Mercaptoethanol, and 12 ng/ml of Human Recombinant Fibroblast Growth Factor-basic (bFGF , Stemgent). HDFiPS cells were cultured on MEF feeder layers in hESC medium and mechanically passaged once a week. The hESC medium was changed daily. All cultures were tested and were negative for mycoplasma.

2.5.5 DNA Extraction and Sequencing

The Qiagen DNeasy Blood and Tissue Kit (cat. no. 69504) was used to prepare genomic DNA from 2 million cells of each cell line, as recommended by the manufacturer. Template DNA fragments (3ug) were hybridized to the surface of flow cells HiSeq Paired-End cluster Kits (v2.5 or v3) and amplified to form clusters using the Illumina cBot. Paired-end libraries were sequenced for 2 x 101 cycles of incorporation and imaging using TruSeq SBS kits. Sequencing was performed at Illumina, Inc. (San Diego).

2.5.6 Realignment of Illumina Reads and Recalibration

Reads were extracted from Illumina Casava aligned BAM files using the HTSLib by first shuffling the reads and then extracting interleaved reads. These reads were then processed through the GATK Best Practices workflow for Variant Calling v.2.6 , which included first aligning the reads using BWA 0.7 with the mem option, marking duplicates, pursuing local realignment, considering base quality recalibration and finally using the reduce reads options. The BAM files generated from this process were used with the different variant calling methods.

2.5.7 SNVs and Indel Variant Calling

HaplotypeCaller, as bundled with GATK v2.7, was used to call all ten samples together. The variant calls were recalibrated using files in the GATK bundle which included data from HapMap, Omni, dbSNP, Mills Indels, 1000Genomes Indels databases. Variants falling in tranche level 0-90, and 90-99.00 were used for downstream analysis. Unique variants were identified by looking at positions where only one sample had a non-reference allele.

It should be noted that the HaplotypeCaller multisample calling pipeline described sacrifices sensitivity for specificity, as it is meant for population scale studies. The pipeline may favor the identification of variants common across the samples rather than variants or mutations unique to each. To gain a more sensitive assessment of the somatic SNVs landscape across the iPSC samples, we also ran MuTect by treating the parent fibroblast population as the normal and each iPSC cell line as a different derived cell (in an analogous manner to which tumors are treated with respect to germline samples in oncogenomics studies). MuTect was run with default settings and the calls were filtered using the judgment KEEP option. Unique variants from the analysis were determined by intersecting the calls based on chromosomal coordinates and the variant calls.

2.5.8 SNVs and Indel Validation and Filtering

To validate the identified variants, we split them into three groups of confidence: high confidence variants had coverage between 40-60x with the variant allele frequency in the range of 0.4-0.6; low confidence variants were classified as having coverage 20-40x with variant allele frequency between 0.4-0.6; and subclonal variants were those with variant allele frequencies between 0.2-0.4 and 40-60x. We validated these results using qPCR. The ABL files associated with the variants were read in using the abifpy package (https://github.com/bow/abifpy), and looking at the calls made by the SOLID software. For any variants that appeared heterozygous, the relative amount of noise in the ten upstream and downstream bases was evaluated based on the assumption that they should be homozygous. Any amplitude values that were two standard deviations higher than the mean background noise were called as variant calls. Most of the calls made in this manner were validated through manual inspection (Supplementary Table1).

2.5.9 SNVs and Indel Variant Annotation

The variants were run through the SGAdviser[PSE$^+$15] pipeline for annotation. A SNV was considered damaging if it had a harmful designation by Condel, Polyphen or SIFT. To assess the likely oncogenic potential of any variant, we assessed overlaps with the MKCC Cancer Genes, Atlas Oncology, and Sanger Cancer Genes. Variants that overlapped Transcription Factor Binding Sites (TFBS) were assessed as being damaging by looking at changes calculated by the MOODS algorithm[KMP$^+$09]. High confidence TFBS altering mutations were those that changed the binding affinity by more than 7. We also ran the variants through the Onconator annotation web service (http://www.broadinstitute.org/oncotator/) and found overlaps with MutSig genes and Familial Syndrome Cancer Genes. Finally, the variants were fed through CADD annotation service through their online web service.

2.5.10 ENCODE Transcription Factor Binding Sites Annotation

Genomic coordinates (hg19) for Transcription Factor (TF) bound regions of DNA, as curated by ENCODE (wgEncodeRegTfbsClusteredV3.interval), were downloaded from the UCSC genome browser Main Page (via Galaxy) in *.bed format. Additionally, the genomic coordinates (hg19) for all iPSC variants (SNVs/indels) were concatenated into a single bed file and intersected with the TF genomic intervals (bedtools) in order to determine whether specific TFbinding profiles were disproportionally enriched with/depleted of mutations in iPSCs generated via one reprogramming method in comparison to the others.

2.5.11 Identification of Integration Sites

Integration sites were identified by looking at the paired end reads in which one end mapped to the reference genome hg19 and one end maps to the viral sequences. The one end mapped reads were extracted from the BAM files using the command samtools view[LD09], and then aligned to the viral genomes using BWA v. 0.7 aligner[LHW$^+$09]. The readnames identifiers were used to find the pair in the hg19 aligned reads. These were further filtered down to require at least 5 reads matching on either end of the integration site and were annotated by intersecting the knownGenes track on UCSC.

2.5.12 Permutation Testing

We leveraged permutation testing in an analysis of variance (ANOVA) setting to find the probability of observing the differences in mutation and variant rates across the three reprogramming methods. We exhaustively relabeled all the reprogramming types, recomputing the ANOVA statistic each time we relabeled the cell lines, thus determining exactly how likely it would be to observe our mutation profile differences across the reprogramming cell types given the 3 replicates for each of the 3 reprogramming methods (a total of 280 permutations). We tested each genomic feature separately using this permutation strategy. To calculate the relative rates of different types of mutations (e.g., coding, non-coding, etc.), we divided by the total number of mutations.

2.5.13 ANOVA Contrasts

We also tested if one method was different from the other two methods using contrasts. Three different tests were employed for mRNA versus all, retroviral versus all, and Sendai virus vs all using the contrasts (-2, 1, 1), (1, -2, 1), and (1, 1, -2), respectively. We employed an ANOVA fit to estimate the impact of the reprogramming strategy on the aggregated counts of the different variant classifications. We employed this strategy to look at the difference in the nominal counts for the different variant classifications, but we also subdivided by the type of variants (SNV, Insertion, or Deletion) as well as the relative rates of each variant type. We further filtered reported results to classifications that had at least 10 variants.

2.5.14 Comparison of Combined Annotation Dependent Depletion Score Distributions

To compare the distribution of Combined Annotation Dependent Depletion (CADD) Scores across the different samples, we employed a Kruskal-Wallis test. First, we looked at the differences between the replicates of the same reprogramming method to ensure that the variance within a group was not high. Next, we looked at pairwise comparisons between the different reprogramming methods by pooling all the variants for the replicates into one distribution for the reprogramming method. This analysis revealed that synthetic mRNA had a distribution skewed towards lower scores compared to the retroviral vectors and non-integrating Sendai virus. There was no realizable difference between the retroviral vectors and non-integrating Sendai virus. However, further inspection of the distributions revealed that most of the variants still fell in the non-damaging designation of CADD scores, indicating that the results were not biologically significant.

2.5.15 BioNano High Molecule Weight DNA Extraction

HDF51iPS11, HDF51iPS509, HDF51iPS1003, and HDF51 cell lines termed R3, S3, M3 and F cell lines, respectively were done on-site at BioNano Genomics where they were washed with 1x PBS, placed in resuspension buffer, and embedded into agarose gel plugs (BioRad, Hercules, CA). Embedded cells were incubated with lysis buffer (BioNano Genomics, San Diego, CA) and proteinase K for four hours at 50C. Agarose was solubilized with GELase (Epicentre, Madison, WI) and extracted DNA was drop dialyzed for four hours. DNA concentrations were measured using the Quant-iT dsDNA Assay Kit (Life Technologies, Carlsbad, CA).

2.5.16 BioNano DNA Labeling

DNA was labeled following the IrysPrep Reagent Kit protocol (BioNano Genomics, San Diego, CA). Briefly, 900 ng of DNA was digested with 10 U of Nt.BspQI nicking endonuclease (New England BioLabs, Ipswich, MA) for two hours at 37 C. Nick digested DNA was then incubated for one hour at 72 C with fluorescently labeled dUTP and Taq Polymerase (New England BioLabs, Ipswich, MA). Taq ligase (New England BioLabs, Ipswich, MA) was used in the presence of dNTPs for ligation of nicks. DNA was counterstained with YOYO-1 (Life Technologies, Carlsbad, CA).

2.5.17 BioNano Data collection

Labeled and counterstained DNA samples were loaded into IrysChips (BioNano Genomics, San Diego, CA) and run on the Irys (BioNano Genomics, San Diego, CA) imaging instrument. Data was collected for each sample until 50-fold coverage of long molecules (¿150 kbp) was achieved. The IrysView (BioNano Genomics, San Diego, CA) software package was used to detect individual linearized DNA molecules using the YOYO-1 counterstain and determine the localization of labeled nick sites along each DNA molecule. Sets of single-molecule maps, equivalent to 50x haploid coverage, for each sample were then used to build a full genome assembly.

2.5.18 BioNano De Novo Assembly

De novo assembly of single molecules is accomplished using BioNanos custom assembler software program based on an Overlap-Layout-Consensus paradigm. First, we started with pair-wise comparison of all molecules longer than 150kbp and ≥ 5 labels to find all overlaps with a p-value $\leq 5 \times 10^{-10}$, then we could construct a draft consensus genome map based on these overlaps. The draft map could be further refined by mapping single molecules to it and recalculating the label positions. Next, the maps were extended by aligning overhanging molecules to the maps and calculating a consensus in the extended regions. Finally, the genome maps were compared and merged where patterns match with a p-value $\leq 10^{-15}$. The process of extension and merge was repeated five times before a final refinement was applied to finish all genome maps. The result of this assembly is a genome map set entirely independent of any known reference or external data (Figure 3). Statistics about N50 and percentage coverage of Genome Reference Consortium Human Build 37 (GRCh37) are described in Supplementary Table S2.1.

2.5.19 BioNano Structural Variation Calls

Structural variation was detected by examining the alignment profiles between the de novo assembled genome maps against the GRCh37 human reference assembly. Significant discrepancies in: a) the distance or b) the number of unaligned labels between adjacent aligned labels would indicate the presence of insertion and deletion events. We used two algorithms to call SV, and they differ in the way discrepant regions in the alignment (termed outliers) were handled. In the first algorithm, the reference and maps were split at outliers, and split maps were iteratively re-aligned.

The alignments of the newly split maps would then pinpoint the locations of the insertion and deletion variants. We used an alignment p-value of 10^{-12} and an outlier cutoff of 10^{-4} to call variants in all four cell line samples. In the second algorithm, the reference and genome maps were not split; instead, the global alignment profiles were kept with insertions and deletion events being intra-alignment gaps. For all four cell line samples, the alignment p-value was 10^{-12} and the outlier cutoff was 10^{-4}. Alignment p-values are calculated using the algorithm described in Anantharaman et al[Ana01].

2.5.20 BioNano Identification of Cell Line Specific Calls

A series of steps were used to identify cell line specific variants. First, we conservatively selected only insertions and deletions detected by both calling algorithms. Using a 50% reciprocal size overlap cutoff, we cross-compared variants detected among all cell lines to identify those that were putatively cell line-specific. Finally, we manually curated the candidate variants to ensure that a) there were molecule supporting the variant allele in the cell line of interest, and b) there was no molecule supporting the variant allele in all other cell line samples.

2.6 Figures and Tables

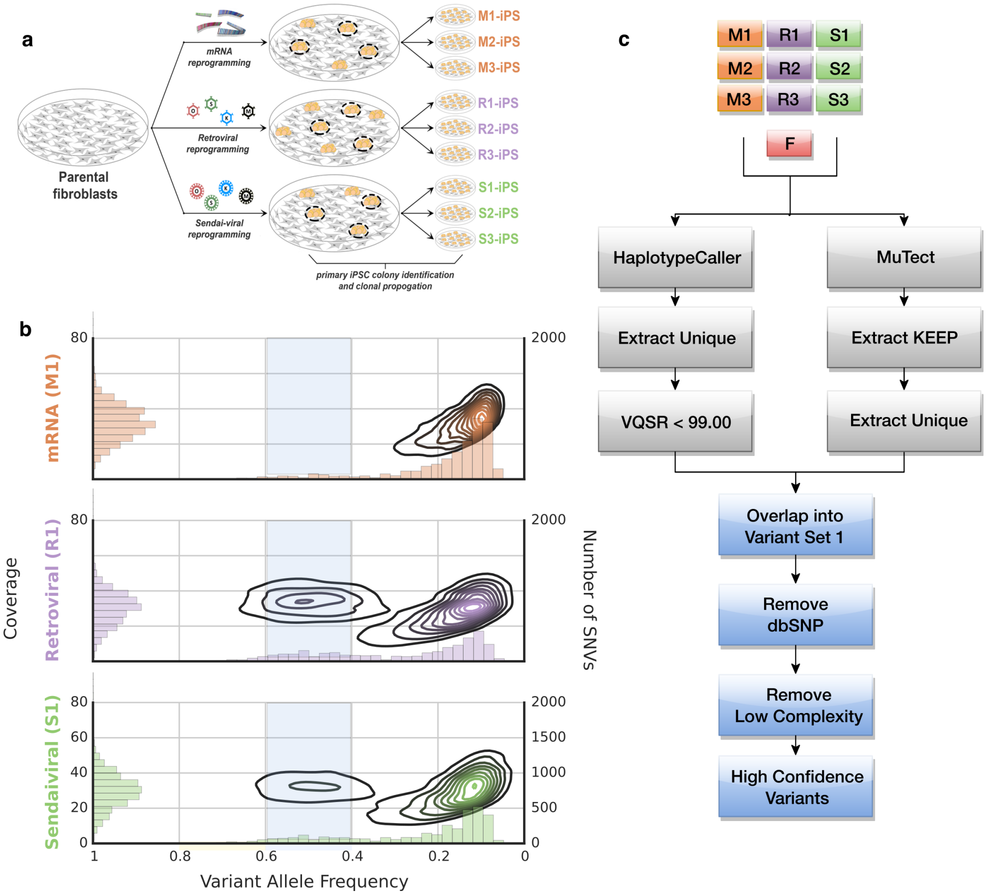


Figure 2.1: Experimental and computational design for identifying variants caused by reprogramming a) Diagram describing the derivation of three biological replicates of each three reprogramming methods: retrovirus, Sendai virus, and non-integrating mRNA b) Kernel density estimation for VAF and coverage for a constituent sample from each reprogramming method: M1 (mRNA), R1 (retro-virus), S1 (sendai virus). For R1 and S1, there are denser clusters near 40x coverage and 40% - 60% VAF than the M1 sample, which indicates they had a higher mutational load during initial doublings. However, it should be noted that all these samples also contained several subclonal variants that are not considered in further analyses. The histograms are intended to aid the readers in interpreting the results of the kernel density estimations. c) Flow diagram detailing the filtering strategy employed to arrive at high confidence set of SNVs unique to each reprogrammed cell line using MuTect and HaplotypeCaller.

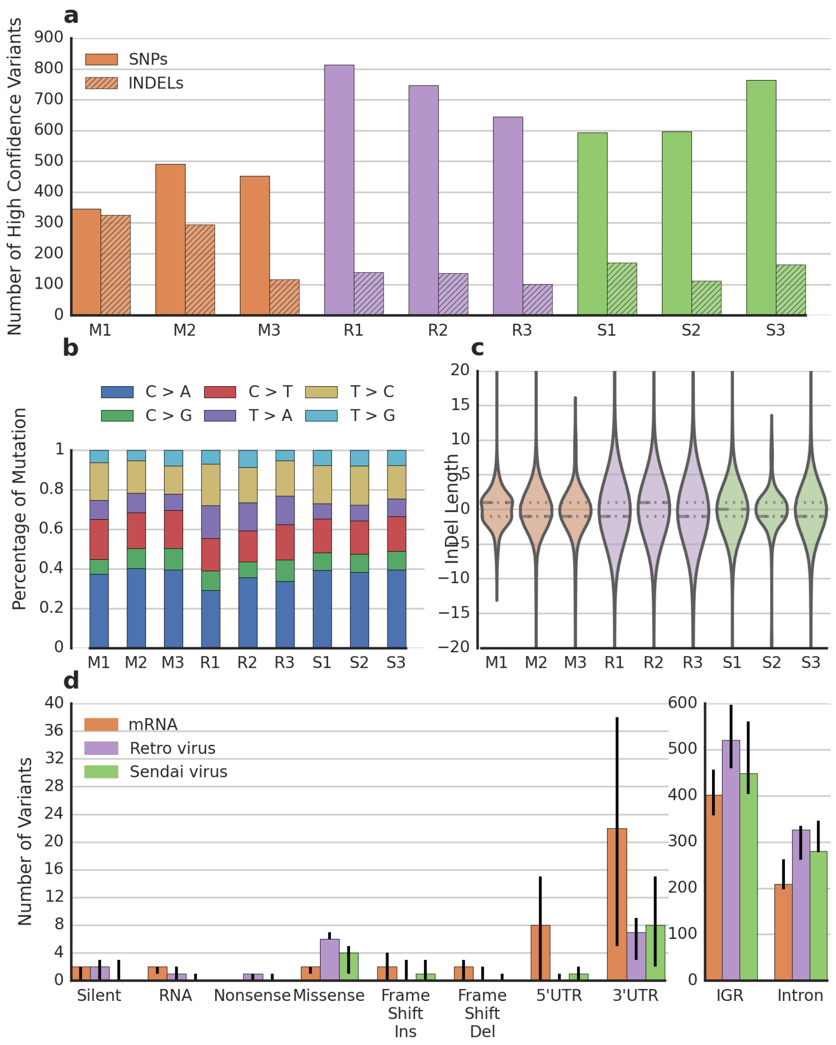


Figure 2.2: Characterization of variants caused by reprogramming method a) Overall counts for the number of high confidence SNVs and indels per sample. b) The relative percentage of mutational subtypes for the SNVs in each sample. c) A violin plot and box plot for the indel size distributions in the sample, a positive length indicates an insertion, whereas a negative one is a deletion. d) Variant classifications based on their relative locations in the genome. The error bars indicate the low, median, and high replicate for each reprogramming method. Introns and IGR variants are plotted on a different scale.

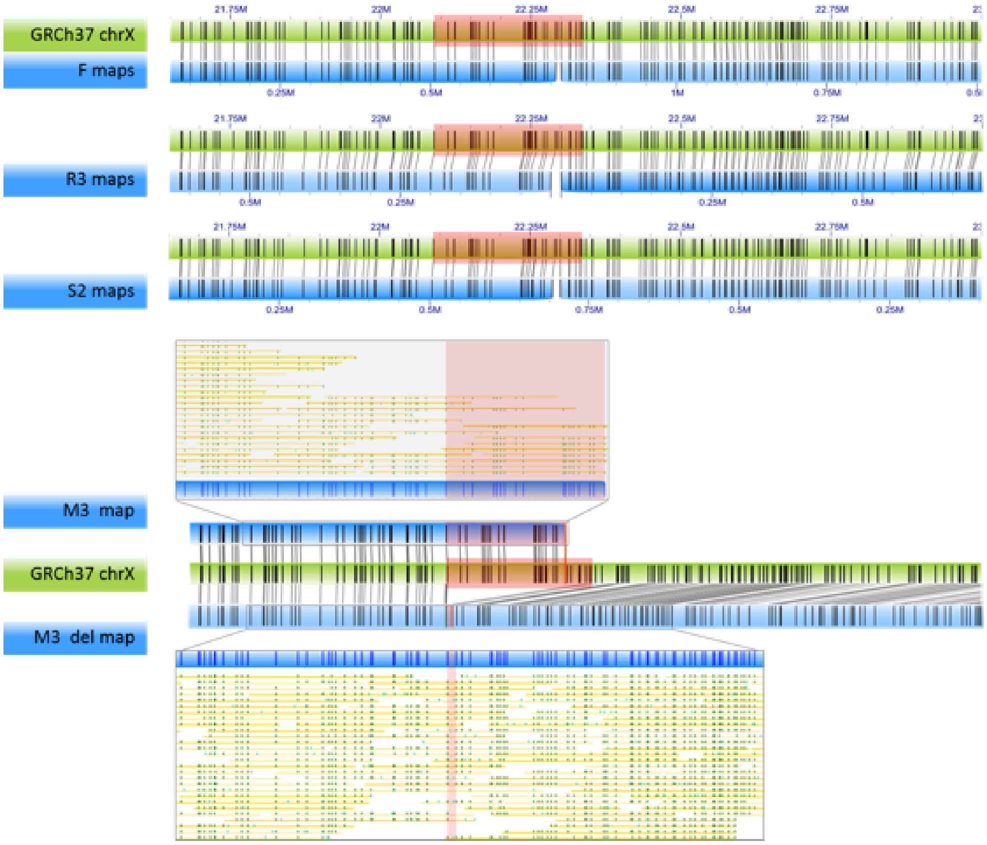


Figure 2.3: A 228.8 kb deletion at Xp22.11 in sample M3 detected by BioNano genome mapping. Each assembly is compared to the GRCh37 reference genome. Black vertical marks show the position of the fluorescently-labeled 7-base motif. For the M3 sample, observed individual DNA molecules and their labels are represented, showing the support for two haplotypes, one with the deletion at Xp22.11.

Table 2.1: P-values from the permutation-based ANOVA test for variant type differences across the three reprogramming methods. Rates were determined by dividing the number of SNPs by the total number of variants. ND=Not determined either because it is not consistent with the calculations or there were too few variants to analyze. "All" is the sum of SNVs and indels. The last column lists the sample size estimates necessary based on 80% power for an ANOVA statistic given the current mean and variance for the grouping by reprogramming methods. This is based on the combined counts.

	SNVs	Rates	Insertions	Deletions	Indels	All	Samples for 80% Power
CADD Phred >15	0.12	0.36	ND	ND	ND	0.12	4
Coding	0.03	0.18	0.34	0.23	0.24	0.16	4
Damaging	0.09	0.3	ND	ND	ND	0.09	4
Near Cancer Gene	0.1	0.94	0.11	0.2	0.14	0.39	10
Total	0.01	ND	0.18	0.16	0.16	0.2	6

Table 2.2: High confidence variants in coding regions. The number of high confidence synonymous and non-synonymous coding mutations identified with high confidence SNVs in each sample. Non-synonymous variants in protein coding regions are listed. M1-3: mRNA vector; R1-3: Retrovirus; S1-3: Sendai virus.

Sample	Number of high confidence variants		Coding Regions
	Synonymous	Nonsense & nonsynonymous	
M1	0	2	Nonsynonymous: BC068088, C14orf159
M2	2	2	Nonsynonymous: BPIFB1, MACROD1
M3	2	1	Nonsynonymous: RPAP2
R1	2	7	Nonsense: SALL1; Nonsynonymous: C2orf91, CCDC150, SSC5D, SYT4, UTRN, WDR72
R2	0	7	Nonsense: PRR12; Nonsynonymous: ADAM18, CATSPERG*, IKBIP, NCR3LG1, PRR12, SPTA1
R3	3	7	Nonsynonymous: ALPI, HIST1H2BD, ITGB8, OR5AP2, PKP2, RNF10, TMPRSS5
S1	0	5	Nonsynonymous: KCNC3, OVOS2, SDR16C5, XPR1, ZNF660
S2	3	4	Nonsynonymous: CLEC5A, FAM208B, MMEL1, PCDHB16
S3	0	2	Nonsense: FSCN3; Nonsynonymous: FSCN3, PRDM4

Table 2.3: Functional impacts as calculated by ANOVA contrasts using a One versus All approach. ANOVA contrasts were set up to compare one reprogramming methods against the other two. For each reprogramming method, the most significant difference is presented. The EZH2 binding site overlaps were determined by ENCODE annotations (Methods).

M1	M2	M3	R1	R2	R3	S1	S2	S3	P-value
Non-integrating mRNA vs All others Transcription Factor EZH2 Binding Sites									
13	11	6	1	7	2	3	0	3	0.008
Retrovirus vs All others Damaging mutations assessed by Condel, Polyphen or SIFT									
5	3	1	5	9	13	5	3	2	0.014
Sendai virus vs All others Mutations in coding regions									
27	16	9	16	20	24	12	9	5	0.044

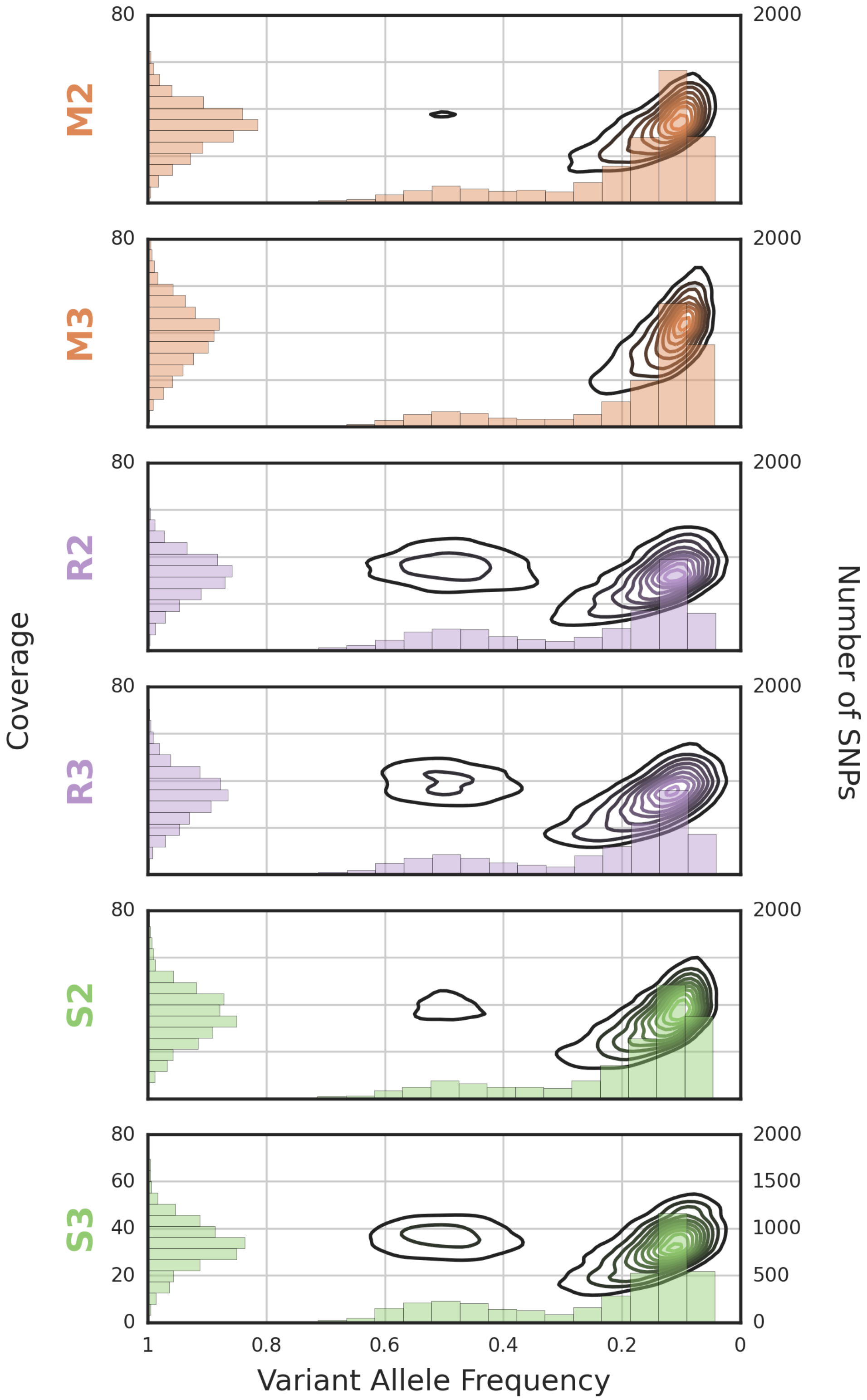


Figure S2.1: Kernel density estimation based on variant allele frequency (VAF) and coverage for samples not presented in main text Figure 1.

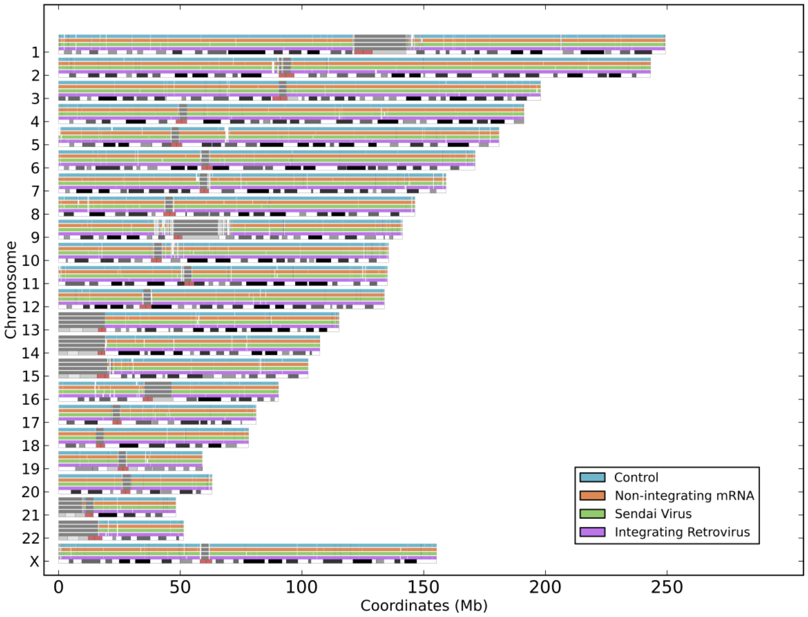


Figure S2.2: De novo assemblies of the fibroblast control and one of each iPSC type from nanochannel mapping data. Genome maps of the four cell lines were aligned to the GRch37 reference map. Ideogram and Giemsa banding is plotted at the bottom of each chromosome in grey scale, with centromeres highlighted in light red. In each de novo assembly, white spaces separate contigs, and N base gaps in the reference are shaded with grey.

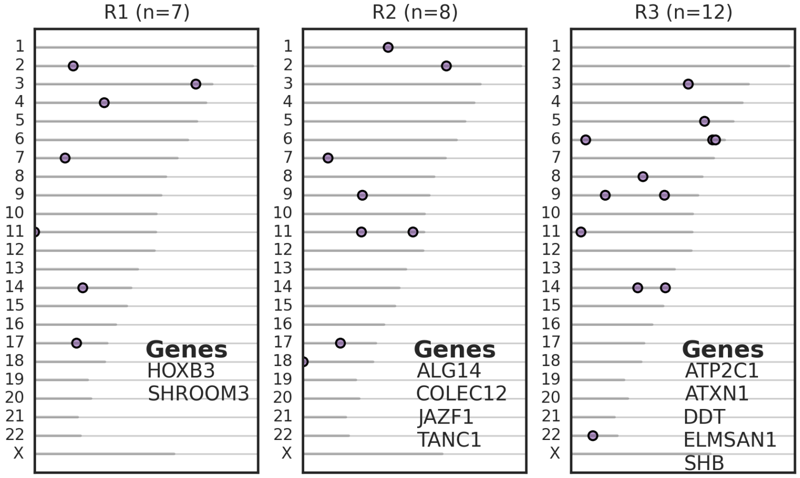


Figure S2.3: Integration sites of Retrovirus in the three replicates. Integration sites in the three retroviral induced cell lines based on the methods outlined in the Methods. A list of genes is also provided for integration sites that were not intergenic.

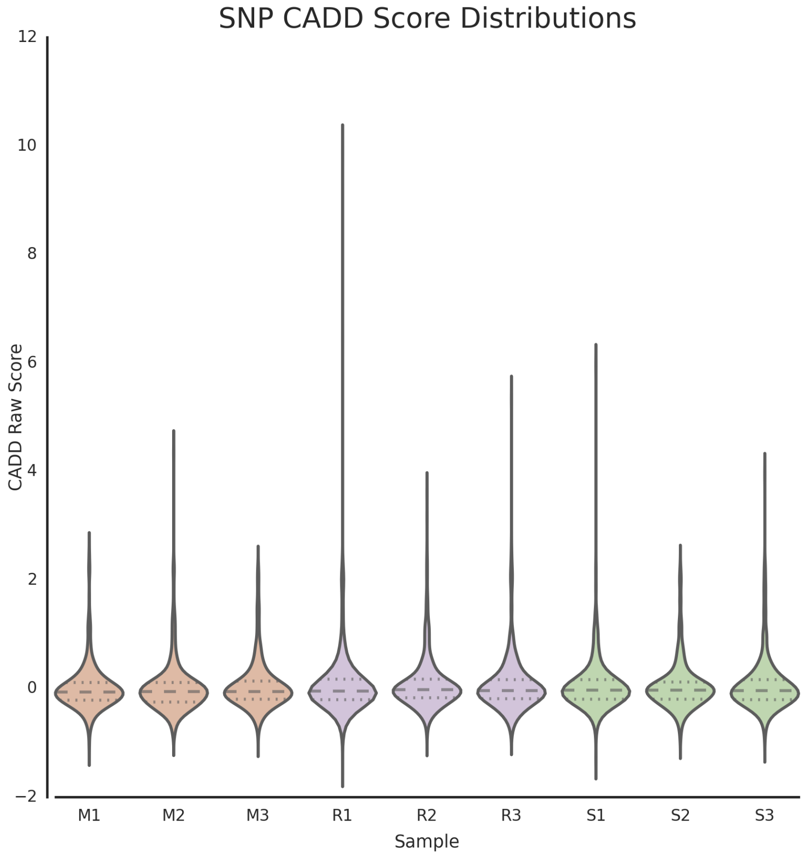


Figure S2.4: To investigate the potential pathogenicity of the variants, we compared the CADD scores of SNVs across the three different reprogramming methods. Although there is a statistical difference between the reprogramming methods (Kruskal-Wallis p-value 0.02) the results are biologically insignificant based on the criteria that most known damaging SNPs fall above a CADD Score of 15.

Table S2.1: Statistics of molecules and de novo assembled genome

	Control (F)	Retrovirus (R3)	Sendai virus (S2)	mRNA (M3)
Molecules >150 kb (Gb)	156.98	160.69	160.69	160.66
Molecule N50 (kb)	284.15	261.35	259.91	255.55
Molecule Coverage (X)	50.71	51.91	51.91	51.9
Assembly Genome Maps (#)	3886	3832	3930	3799
Total Genome Map Len (Mb)	2855.3	2820.79	2829.19	2847.77
Genome Map N50 (Mb)	0.93	0.95	0.9	0.97
Total Genome Map Len / Ref Len	0.92	0.91	0.91	0.92
Total Unique Len / Ref Len	0.89	0.89	0.88	0.89

Table S2.2: Coding mutations identified by MuTect. The number of synonymous and non-synonymous coding mutations identified in each sample using MuTect.

Sample	Number of variants identified by MuTect	
	Synonymous	Nonsense and nonsynonymous
M1	92	32
M2	133	42
M3	24	20
R1	8	13
R2	25	20
R3	7	21
S1	44	17
S2	31	20
S3	7	9

2.7 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in *Nature Communications* 2016. Kunal Bhutani, Kristopher L Nazor, Roy Williams, Ha Tran, Heng Dai, Zeljko Dzakula, Edward H Cho, Andy WC Pang, Mahendra Rao, Han Cao, Nicholas J Schork, Jeanne F Loring. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Modeling prediction error improves power of transcriptome-wide association studies

3.1 Abstract

Transcriptome-wide association studies (TWAS) test for associations between imputed gene expression levels and phenotypes in GWAS cohorts using models of transcriptional regulation learned from reference transcriptomes. However, current methods for TWAS only use point estimates of imputed expression and ignore uncertainty in the prediction. We develop a novel two-stage Bayesian regression method which incorporates uncertainty in imputed gene expression and achieves higher power to detect TWAS genes than existing TWAS methods as well as standard methods based on missing value and measurement error theory. We apply our method to GTEx whole blood transcriptomes and GWAS cohorts for seven diseases from the

Wellcome Trust Case Control Consortium and find 45 TWAS genes, of which 18 do not overlap previously reported GWAS associations. Surprisingly, we replicate only 2 of 40 previously reported TWAS genes after accounting for uncertainty in the prediction. Software implementing our methods and fitted model parameters are available at https://github.com/Schork-Lab/mediator-was.

3.2 Introduction

Thousands of loci associated with hundreds of complex diseases have been reported in the NHGRI catalog of genome-wide association studies [KAT$^+$11, WMM$^+$14] (GWASs). However, most genome-wide significant loci are devoid of protein-coding alterations [HSJ$^+$09] and likely instead affect gene regulation by mechanisms such as transcriptional, post-transcriptional, or epigenetic regulation. Several studies have directly investigated the role of transcriptional regulation on complex diseases by jointly considering genotypes, expression, and phenotypes using Mendelian randomization [SFP$^+$14, ACF$^+$16]. However, such studies require genetic, transcriptomic, and phenotypic data to be measured in all samples, which is still prohibitive at the scale of GWAS.

Recent large-scale efforts such as the Gene-Tissue Expression Project (GTEx) have generated reference transcriptomes across multiple human tissues[ADS$^+$15]. These data have enabled transcriptome-wide association studies (TWAS), which use the reference expression data to build models of transcription regulation, impute gene expression into GWAS cohorts where expression is not measured, and directly test for association between predicted gene expression and phenotype[GWS$^+$15, GKS$^+$16]. However, current methods are limited to using only point estimates of imputed expression, while ignoring the uncertainty in the predicted expression.

The impact of not incorporating uncertainty of imputed predictors on genetic association analysis has been previously studied in the context of imputed genotypes in GWAS. Although taking the best-guess genotype (posterior mode) is standard practice for GWAS, using posterior mean dosages increases power to detect associations [ZLAS11, AS13]. Standard methods from missing data theory such as multiple imputation have been applied in this setting yielding reductions in bias [SSK+07, PPB+16]. More sophisticated (and computationally expensive) methods such as SNPTEST [MHM+07] can analytically integrate over the full posterior distribution of the imputed genotypes, further improving power.

Here, we develop a novel Bayesian method for modeling uncertainty in imputed expression and propagating this uncertainty through TWAS. We compare our method to existing methods for TWAS and standard methods from missing data and measurement error theory and show that our method increases power to detect genes associated with phenotype. We apply our methods to GWAS for seven diseases from the Wellcome Trust Case Control Consortium [BCC+07] and find 42 TWAS genes, replicating only 2 of 40 previously reported TWAS genes. We find 17 of the 42 genes have not yet been identified by GWAS or differential gene expression in case-control cohorts. We provide an implementation of all of the methods in a Python package (https://github.com/Schork-Lab/mediator-was).

3.3 Results

3.3.1 Uncertainty in TWAS

The key insight enabling TWAS is that one can train models to predict gene expression from genotype in reference cohorts, and use these models to impute unobserved gene expression values in GWAS samples using the genotype information. Direct tests for association between gene expression and phenotype can be pursued, more directly identifying putative causal genes for the phenotype of interest. However, current methods only use a point estimate of the predicted gene expression in TWAS and ignore the uncertainty in the prediction. The uncertainty arises from two sources: not learning the correct model for transcriptional regulation (e.g., omitting trans-regulatory effects), which we do not consider here, and not correctly estimating the model parameters due to sample size, linkage disequilibrium, or biological and technical confounders.

Our main contribution is a novel two-stage Bayesian regression model which incorporates uncertainty for each SNP effect (BAY-TS). The key idea of BAY-TS is to use the posterior distribution of SNP effects from the first-stage regression of expression against genotype as the prior distribution on effects in the second stage regression of phenotype against expression (Methods). After performing the second stage regression, we compute a Bayes factor comparing the fitted model against a null model where gene expression has no effect on phenotype.

We compared our method against existing methods, which merely calculate association statistics using ordinary least squares. We compared two strategies: using the full elastic net model trained on the entire GTEx dataset (OLS-E), and using the mean value of imputed expression across 50 bootstrapped models (OLS-M). We note that existing methods for TWAS implement OLS-E only.

We then compared our method to multiple imputation (MI), a standard method for handling completely missing (i.e., unobserved) data such as gene expression in TWAS [LD02]. For each gene, we imputed 50 expression levels for each individual using the bootstrapped models described above and estimated 50 effect sizes. We then combined these effect sizes into a single association statistic for evaluation incorporating both the mean and the variance of the 50 estimates (Methods).

Finally, we compared our method to regression calibration (RC), a standard method from measurement error theory. Measurement error theory explicitly models the error in observations (here, imputed expression) and predicts the impact of not including the errors on statistical inference [Ful87]. Briefly, not explicitly including error in the model leads to a violation of the model assumptions and therefore leads to bias in the estimated regression coefficients. Applying this theory to TWAS, we modeled each imputed expression value as the true expression value plus additive error. We estimated the distribution of the error as the variance in predicted expression across the 50 bootstrapped models. We then performed RC, estimating the true expression based on the estimated measurement (imputation) errors and regressing phenotype on the true expression.

3.3.2 Simulation study

We used real genotype data to jointly simulate gene expression at both causal and non-causal genes in simulated reference and GWAS cohorts and continuous phenotypes in the GWAS cohorts, as done in prior work [GKS$^+$16]. To calculate the recall and the effective false discovery rate (FDR), we tested single-gene associations against a phenotype generated using all of the simulated genes (Methods). For each simulated data set, we computed the the area under the precision-recall curve (AUPRC) of each method. We compared the AUPRC rather than the area under the receiver operating characteristic (AUROC) curve because the AUROC is not appropriate when the proportion of positive and negative examples is not 0.5 [DM07]. We ranked the genes according to the association statistic computed by each method, then computed the cumulative precision and recall (based on the simulated ground truth) for each position in the ranked list (Figure S3.2).

We simulated a reference cohort of 300 individuals and a GWAS cohort of 5,000 individuals for which 40 genes were causal and 1,000 were non-causal, and found that Bay-TS outperformed all other methods across the entire range of cis-regulatory architectures (Figure 3.1a). Interestingly, neither MI nor RC improved performance over OLS-E, which is likely explained by the fact that in this setting a central assumption in measurement error theory is violated. Specifically, when modeling the error in imputation, the second-stage predictors (means of imputed expression) and their associated errors (variances of imputed expression) are correlated because they both depend on genotype.

We then investigated the recall of each method at FDR 10%. To control the FDR of BAY-TS, we calibrated a threshold for the Bayes factor (BF > 24) which controlled the average FDR over the entire range of simulation parameters at the desired level (Methods). For the other methods, we used the Benjamini-Hochberg procedure to control the FDR. We found that BAY-TS again consistently outperformed all the methods, with average recall equal to 15% (Figure 3.1b). Surprisingly, MI had the worst performance, likely due to deflated association statistics (Figure S3.3).

3.3.3 Application to seven diseases

Before applying our method to real data, we sought to evaluate the impact of expression normalization on the trained first-stage models in TWAS. We first asked which genes had gene expression reliably predictable from cis-genotypes. We used whole blood expression in 338 individuals from the GTEx project [ADS+15] and compared models trained on the published normalized expression values (GTEx-Norm) to models trained on regularized log transformed expression (RLog) (Methods). We found 1,666 and 1,655 genes with $R^2 \geq 0$ for GTEx-Norm and RLog, respectively, and found 1,043 genes common between the two sets (Figure 3.2a). However, only 987 of the 1,043 genes had enough non-missing data in the GWAS cohorts to successfully impute expression into GWAS.

We next considered the ratio between the variance of within-individual imputed expression estimates to across-individual estimates, which we define as the variance ratio (VR) (Methods). Measurement error theory predicts VR is correlated with power to detect associations, which we confirmed in simulation (Figure S3.4). Intuitively, after estimating the error variance of imputed expression, we can denoise the imputed expression by subtracting variation due to error in the imputed values. For the set of 987 genes, RLog based models had higher VR than GTEx-Norm based models (Figure 3.2b, S3.5), suggesting RLog normalization increases power to detect TWAS genes.

We then asked whether the same cis-SNPs were reliably selected in the first stage model fitting for the 987 genes. We note that in this setting first-stage regression predictors are correlated due to linkage disequilibrium, and therefore regularized regression techniques such as elastic net will in general not select the same non-zero regression coefficients for replicate data sets. We calculated the discordance fraction between the predictors selected in our first-stage bootstrap training against the predictors included in a single model trained on the entire dataset (Methods). We found not only that GTEx-Norm models have more predictors per gene on average, but also have higher discordance in the selected predictors. Since RLog models find a more consistent set of cis-SNPs and have larger VR (Figure S3.5), we perform TWAS based on RLog models.

We performed TWAS on seven disease cohorts from the Wellcome Trust Case Control Consortium [BCC+07]: bipolar disorder (BD), Crohn's disease (CD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), Type 1 Diabetes (T1D), and Type 2 Diabetes (T1D). We first sought to replicate PrediXcan by using the published cis-regulatory model weights on our imputed genotypes [GWS+15]. We replicated only 18 of 40 reported TWAS genes (Bonferroni correction, $p < 5.76 \times 10^{-6}$), likely due to differences in imputation pipelines between the two studies. 13 of the discordant genes are in the Major Histocompatability Complex (MHC) region, and 15 have estimated effect sizes with the same sign as previously reported (Table S3.1).

We then performed TWAS using BAY-TS in each of the seven diseases and found 45 associated genes (FDR 10%, Table 3.1). Surprisingly, we only replicated 2 of 40 reported TWAS genes (Table S3.1). Moreover, we found that only 9 of the 40 reported genes in our set of high confidence genes with reliably predictable expression. The discrepancy in TWAS genes is due partly to differences in genotype imputation, expression normalization, and first stage model fitting as described above, and highlight the importance of data processing choices when performing TWAS. We compared BAY-TS to the other methods proposed above (including OLS-E, equivalent to Predixcan) and found that 23 of the 45 genes are not found by any of the other methods (Table S3.2, S3.3, S3.4). The other methods collectively also found TWAS genes for CD, HT, RA, T1D, and T2D , but did not find associations for the other diseases likely due to lack of power.

We discovered three genes significantly associated with BD: SARDH, ZNF79, and WDR25. A higher burden of CNVs in the SARDH gene has been previously linked to BD and Schizophrenia[Lac08], and its loss is the principal cause for sarcosinemia, which often leads to mental impairment[SCTS70]. However, ZNF79 and WDR25 are both poorly characterized and have not been studied in the context of BD previously.

We discovered four genes associated with CAD: DAGLB, CYTH3, PARL, and TMEM158. DAGLB is active in the triglyceride lipase activity pathway, and has been linked previously to high-density lipoprotein levels levels[WSS+13]. CYTH3 is a regulator of PI-3 kinase signaling, which mediates many pathways in the cardiovascular system [STW+15]. Variants in PARL have been associated with increased levels of plasma insulin and predisposition to CAD [PWA+08]. In Chinese populations, the gene was also linked to higher levels of triglyceride and total cholestrol in both T2D and control populations [LHL+14]. Lower TMEM158 expression is associated with increased risk in both of the CAD and T2D populations in our study. It has previously transcriptionally associated with T1D, T2D, and gestational diabetes based on a case-control differential expression study performed using peripheral lymphomononuclear cells[CEX+13].

We discovered seven genes associated with CD (Figure 3.3): PIGC (1q24..33), FAAH (1p33), HLA-DQA1 (6pq21.32), HLA-B (6p21.33), PRKAB1 (12q24.23), LRRC37A2 (17q21.31), and HSBP1L1 (18q23). PIGC has been associated with inflammatory bowel disease as part of a larger Immunochip meta-analysis[JRW$^+$12]. It is also significantly associated with T2D in our study. There has been no direct genetic evidence supporting the role of FAAH in CD, but recent experiments have shown that drugs targeting FAAH are effective against mouse models of colitis [SMZ$^+$14]. The HLA region has a known role in autoimmune disorders such as CD [FTA$^+$08, AAB$^+$02]. PRKAB1 encodes the noncatalytic beta subunit of the AMP-activated protein kinase (AMPK), which has been previously experimentally validated to play an important role in IBD and is a therapeutic target for drugs used to treat CD and T2D [LLY$^+$15]. Neither LRRC37A2 nor HSBP1L1 have been characterized in CD.

We found two genes associated with HT: SHMT1 and CIAO1. SHMT1 has previously been associated with hypertension, is a general prognostic marker for hypertension [OAM$^+$15], is a marker for intra-cranial hypertension during space flight [SGZ$^+$16], and mediates the response to the angiogenesis inhibitor Bevacizumab[BKA$^+$12]. CIAO1 encodes a protein in the iron-sulfur protein assembly complex and modulates activity of WT1, an oncogone associated with nephroblastomas [JWTV98]. It has not been studied in relation to HT.

We found multiple associations for RA in the MHC region: HLA-B, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HLA-G, IER3, and MICA. Similarly, we found multiple MHC associations with T1D: BAK1, BTN3A2, HISTH1H2AG, HLA-B, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HMGN4, and IER3. However, we discovered T1D associations with C1QTNF6, RBMS2, and RSP26. Larger meta-analysis of T1D has found association in the C1QTNF6 and RSP26 loci[CSS$^+$08, BCC$^+$09]. However, RBSM2 has not been previously linked to T1D.

Finally, we found four genes are associated with T2D: KHK, PIGC, POLR2J3, and TMEM158. Ketohexokinase (KHK) plays a role in fructose metabolism and has been studied extensively as a possible cause for T2D [CGM$^+$09, KK13]. A sub-threshold association in its locus has been identified in a recent study of T2D in a Japanese population [ITY$^+$16]. The PIGC locus has been previously linked to BMI[SVSV15], but not to T2D. POLR2J3 also has not been studied in T2D previously.

We further characterized our TWAS associations by looking at overlaps with GWAS loci (Methods). For each significant gene, we looked for a GWAS hit within 1 MB of the gene body ($p < 5 \times 10^{-8}$), and found overlaps for 23 of 42 genes. We then asked how many of the 42 genes were later discovered by larger meta-analyses in the seven diseases using the NHGRI GWAS Catalog [HSJ$^+$09, WMM$^+$14] and found 10 TWAS genes overlap loci reported in the catalog. However, most of the common genes found by both TWAS and GWAS lie in the MHC region. We finally sought to validate our TWAS associations using orthogonal case-control expression data sets[IKB$^+$04, BFT$^+$12, WDC$^+$07, BCM$^+$04, TOMM$^+$09, KAT$^+$11].

Surprisingly, we found only 3 TWAS genes are differentially expressed between cases and controls (limma modified t-test[Smy05], Benjamini-Hochberg FDR < 0.1). We investigated the ranking of the top 250 genes by TWAS and differential expression and found no significant overlap between the ranked lists (Methods). There are several possible explanations for this discrepancy, including the tissue in which expression was measured, sample size, and technical confounders.

3.4 Discussion

Transcriptome-wide association studies (TWAS) have proven to be a powerful approach for identifying new genes associated with a phenotype by cleverly combining reference expression data and available GWAS data. TWAS directly associate genes to disease, revealing new biological insights. Indeed, these ideas have been extended to additional levels of mediation, incorporating histone modification data to study genetic effects on epigenomic control of transcription [GMF+16]. However, the field has not yet fully appreciated the importance of uncertainty in these multi-stage regression models. We showed that the state-of-the-art methods for TWAS adequately control type I error, but lose power due to uncertainty. We proposed a novel two-stage Bayesian method, BAY-TS, which outperforms not only existing methods but also standard methods from both missing data and measurement error theory. In applications to seven diseases from WTCCC, our method identified new genes not identified by previous methods, which do not incorporate uncertainty.

Our results reveal that uncertainty arises from many sources, not only differences in the trained models of regulation. Using the same GWAS genotypes, different imputation pipelines did not yield the same gene associations. We showed that expression normalization has an impact on trained models used for TWAS, and demonstrated that models trained on regularized-log transformed data were better than those trained on published GTEx expression data. Other studies have shown inconsistency between different population of reference transcriptomes from the same tissue [GKS+16]. These differences can be attributed to technical variation, environmental noise, sequencing technology, processing pipelines, and population differences. There is a pressing need to develop methods which adequately account for all these sources of uncertainty.

3.5 Methods

3.5.1 Uncertainty in TWAS

We assume a continuous phenotype y_i with zero mean collected on n individuals, and regress phenotype on predicted expression w_i for each gene whose expression levels can be predicted from genotype information. To handle binary phenotypes, we estimate latent liabilities using LEAP[WLGH15] and regress predicted expression against individual liabilities. For ease of exposition, we describe a model with no additional covariates; these can be included as additional terms in the model with no modification to the algorithms.

$$\begin{aligned} y_i &= \alpha w_i + \epsilon_i \\ y_i &= \text{phenotype of individual } i \\ w_i &= \text{predicted expression of individual } i \\ \epsilon_i &= \text{error in equation} \end{aligned}$$

Current models use only one prediction for w_i, based on a model of transcriptional regulation learned for the gene.

$$\hat{w}_i = G_i\hat{\beta}$$

$$G_i = \text{cis-regulatory SNP dosages}$$

$$\hat{\beta} = \text{eQTL effect size}$$

Here, we investigate models which account for the distribution of w_i and β. Assuming access to only one training cohort, we estimate these distributions by fitting k bootstrapped models regressing observed gene expression E on genotype G. Here, the regressions are performed using elastic net with regularization penalty tuned using cross-validation.

$$\hat{\beta}^{(k)} = \underset{\beta^{(k)}}{\arg\min}\, ||E - G\beta^{(k)}||^2 + \lambda_1||\beta^{(k)}||_1 + \lambda_2||\beta^{(k)}||_2$$

We estimate the first two moments of the distributions of w_i and β using the following equations:

$$
\begin{aligned}
\hat{w}_i^{(k)} &= G\hat{\beta}^{(k)} \\
\bar{w}_i &= \frac{1}{K}\sum_k w_i^{(k)} \\
\hat{\sigma}_{ui}^2 &= \frac{1}{K-1}\sum_j (w_i^{(k)} - \bar{w}_i)^2
\end{aligned}
$$

3.5.2 Bayesian two-stage regression model

To fit the Bayesian two-stage regression model for a gene (BAY-TS, Figure S3.1), we use the distributions of β_j learned using the k bootstrapped models for that gene as the prior in the second-stage regression.

$$
\begin{aligned}
P(\beta_j) &= \mathcal{N}(\mathbb{E}[\beta_j], \mathbb{V}[\beta_j]) \\
P(\alpha) &= \mathcal{N}(0, 1) \\
P(\epsilon_i) &= \mathcal{N}(0, \sigma_e^2) \\
P(\sigma_e^2) &= C^+(0, 10)
\end{aligned}
$$

Here, $\mathcal{N}(\cdot,\cdot)$ denotes the Gaussian density and $C^+(\cdot,\cdot)$ denotes the half-Cauchy density. Our inference goal is to estimate a Bayes factor comparing the model described above to a null model where $\alpha = 0$. Rather than estimating the intractable model evidences and taking a ratio, we added a model indicator variable z and combined the null model $p(x, y) = N(0, \sigma_e^2)$ and the alternate model (described above). Then, the

Bayes factor is given by the ratio $p(z = 1)/p(z = 0)$. We implemented the model using PyMC3 [SWF16] and used the Metropolis-Hastings algorithm to perform inference. We ran the MCMC chain for 100,000 steps and used the last 10,000 samples to compute the Bayes factor.

3.5.3 Multiple imputation

To perform multiple imputation (MI), we fit the k bootstrapped models for w against the phenotype using linear regression, calculate an aggregated test statistic θ_{mi}, and compute a Wald test statistic.

$$\begin{aligned}
\hat{\alpha}_{MI} &= \frac{1}{K}\sum_k \alpha_k \\
\mathrm{Var}(\hat{\alpha}_{MI}) &= \mathrm{Var}(\alpha) + \left(1 + \frac{1}{K}\right)\frac{1}{K}\sum_k \mathrm{Var}(\alpha_k) \\
\theta_{MI} &= \frac{\hat{\alpha}_{MI}}{\mathrm{Var}(\hat{\alpha}_{MI})} \\
H_0 &: \theta_{MI} = 0 \\
H_1 &: \theta_{MI} > 0
\end{aligned}$$

3.5.4 Regression calibration

We assume additive measurement error on the predicted expression value:

$$w_i = x_i + u_i$$

$$x_i = \text{true (latent) expression of individual } i$$

$$u_i = \text{error in predicted expression of individual } i$$

We assume measurement errors have zero mean and finite variance:

$$E[u_i] = 0$$

$$V[u_i] = \sigma_{ui}^2$$

To perform regression calibration (RC), we impute the true expression value and regress phenotype against this estimated true expression. Given $\hat{\sigma}_{ui}^2$, we regress y_i on $\hat{x}_i = \bar{w} - \hat{\kappa}(w_i - \bar{w})$, yielding estimate $\hat{\alpha}^*$. To estimate the association p-value, we perform a Wald test. We estimate Σ_α, the covariance of $\hat{\alpha}^*$, using a robust estimator [Ful87]:

$$\begin{aligned}
\Sigma_\alpha &= M_{XX}^{-1} H M_{XX}^{-1} \\
M_{XX} &= \frac{W'W}{n} - \hat{\Sigma}_u \\
W &= n \times p \text{ design matrix (including intercept)} \\
\hat{\Sigma}_u &= \text{covariance of measurement errors} \\
H &= \frac{1}{n(n-p)} \sum_i \Delta_i \Delta_i' \\
r_i &= y_i - W_i \hat{\alpha} \\
\Delta_i &= W_i' r_i + \Sigma_{ui} \hat{\alpha} \\
\theta_{RC} &= \left(\frac{\hat{\alpha}^*}{SE(\hat{\alpha}^*)} \right)^2 \sim \chi^2(\theta; 1) \\
H_0 &: \theta_{RC} = 0 \\
H_1 &: \theta_{RC} > 0
\end{aligned}$$

3.5.5 Simulation study

We used imputed dosages for 4,884 samples from the Hypertension, 58C, and NBS cohorts as described below. We selected 193 genes with cis-heritable gene expression (likelihood ratio test, GREML) in all of three studies as previously reported[GKS$^+$16]: Metabolic Syndrome in Men (METSIM), Netherlands Twin Registry (NTR), and Young Finns Study (YFS). We held out 350 individuals as the training cohort and used the rest as the test cohort.

For each gene, we sampled the causal fraction of eQTLs from (Single, 1%, 5%, 10%)

of SNPs from the cis-regulatory window. We computed the genetic value of each individual $X = G\beta$ and added i.i.d. Gaussian noise to achieve proportion of variance explained (PVE) equal to 0.17 in expectation by sampling from $\mathcal{N}(0, \mathbb{V}[G\beta] * (1/.17 - 1))$, where $\mathbb{V}[G\beta]$ is the sample variance of the genetic values.

For each simulation, we sampled 40 causal genes and added i.i.d Gaussian noise to achieve PVE = 0.2 using the procedure described above. We computed the genetic value of each individual as $y = X\alpha$ and add Gaussian noise as described above. We evaluated the method using sample sizes of 5000 individuals. We also tested the performance varying the number of non-causal genes from 400 to 4000.

3.5.6 GWAS processing

We downloaded Affymetrix genotypes for the Wellcome Trust Case Control Consortium seven diseases study in OXSTATS format called using the Chiamo algorithm from the European Genome Archive. We downloaded probe identifiers, hg19 positions, and strand information (http://www.well.ox.ac.uk/~wrayner/strand/) to convert positions to hg19 and used GTOOL version 0.7.5 to align all genotypes. We used PLINK version 1.09b to produce hard genotype calls with genotype probability threshold 0.99 and remove all SNPs and samples excluded from the original study.

We used SHAPEIT2 v2.r644 (ref. [HFS$^+$12]) to exclude unalignable SNPs and phase the case and control cohorts independently for each autosome. We used default values for all model parameters. We used IMPUTE2 version 2.3.0 (ref. [HDM$^+$09]) to impute into all SNPs and indels with MAF in European samples > 0.01. We divided the autosomes into 5 MB windows and threw out windows with fewer than 100 array probes.

For each cohort, we hard-called imputed dosages with genotype probability threshold 0.9. For each disease cohort, we produced a case-control set of hard-called genotypes for both the array genotypes and imputed genotypes by merging all chromosomes with the shared controls (1958 Birth Cohort and National Blood Services). We used GCTA 1.24 (ref. [YLGV11]) to estimate a genetic relatedness matrix on the case-control array genotypes and prune pairs of individuals with relatedness > 0.05. We used plink to remove these individuals from the imputed genotypes, and further remove indels and SNPs with missingness > 0.01, differential missingness ($p < 0.05$) or HWE $p < 10^{-5}$.

We used LEAP version 0.1.8.9 (ref. [WLGH15]) to estimate latent liabilities for each chromosome of each case-control dataset separately (holding out that chromosome) using the array genotypes. We used FastLMM version 0.2.26 (ref. [LLL^{+}11]) to compute association p-values for the imputed genotypes using kinship matrices estimated from the array genotypes (described above). We made extensive use of GNU parallel[Tan11] to facilitate the analysis.

3.5.7 Reference expression processing

We downloaded genotypes and RNA-Seq read counts in the v6 release of GTEx from dbGaP. We restricted our analysis to only those genes which had RNASeqC gene-level read count >= 10 in at least 10 individuals, resulting in 12,049 genes. We transformed the counts to regularized log-transform values (RLog) using DESEq2, adjusting for sequencing depth using the median-of-ratios method [LHA14]. We trained our models on these normalized values to models trained on the published expression values (GTEX-Norm).

We extracted genotypes for SNPs within 500kb upstream and downstream of the transcription start and end sites for each gene using plink. We filtered sites with missingness > 0.01 or HWE $p < 10^{-5}$. For models fit on the published expression values (GTEx-Norm), we included 3 genotype principal components (PCs), 35 expression PEER factors, gender, and sequencing platform as covariates. For models fit on RLog, we used the same covariates but used 10 expression PCs instead of the PEER factors.

To find the optimal elastic net penalty parameter and l1/l2 regularization ratio, we used ElasticNetCV from the scikit-learn package [PVG^{+}11] with possible l1 ratio values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.95, and 0.99 and 100 uniformly distributed penalty parameters from 0.1 to 1. Using the best fitted penalty parameters parameters, we fit 50 bootstrapped models by sampling 300 individuals from the 338 samples.

We imputed expression using PrediXcan software as well as an independent implementation. In our independent implementation, we filtered all GWAS sites that had higher than 10% missing genotypes. Additionally, we assigned the average value to all missing genotypes (which is possible for best-guess imputed genotypes after thresholding on the posterior probability of any genotype call). We note that the implementation of PrediXcan assumes there is no missingness in the data.

To calculate cross-validation prediction accuracy, we predicted hold-out gene expression using only genotype (omitting covariates). We used 5-fold cross-validation and calculated the average R^2 across the folds.

We define the variance ratio (VR) as the ratio of variance of mean imputed expression to the mean of the variance of imputed expression. Intuitively, the VR compares the within-individual variation in imputed expression to between-individual variation. We estimate VR by estimating the necessary means and variances over the

50 bootstrapped models described above.

3.5.8 GEO data

We downloaded case-control expression datasets for the seven diseases from GEO and processed them using the GEO2R web service: GSE12654 (BD) [IKB+04], GSE20681 (CAD) [BFT+12], GSE6731 (CD) [WDC+07], GSE703 (HT) [BCM+04], GSE15573 (RA) [TOMM+09], GSE55100 (T1D) [YYW+15], GSE21321 (T2D). For each data set, we found differentially expressed genes between cases and controls using limma and the GEO2R web service. We assessed significance in the ranking between TWAS lists and GEO lists using the R package OrderedList [YXC], restricting to the top 250 overlapping genes for each disease.

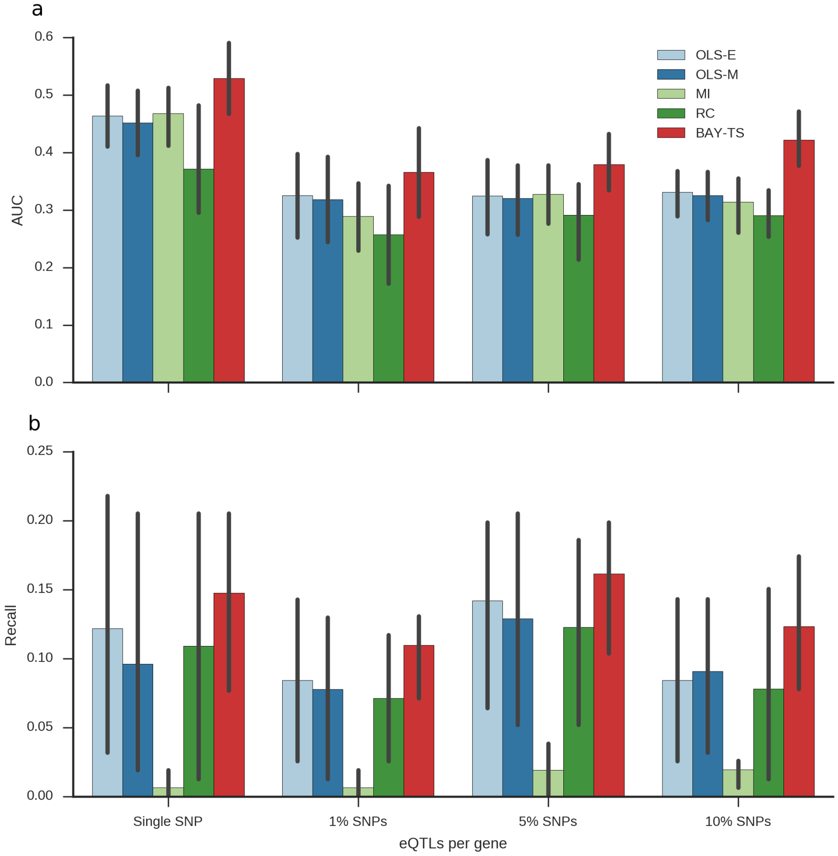


Figure 3.1: Simulation results. a) Area under precision-recall curve (AUPRC) for each method. Error bars represent standard error of AUPRC estimated over 4 replicates. b) Recall of each method controlling FDR at 10%. Error bars represent standard error of recall over 4 replicates.

Table 3.1: TWAS results for Bayesian two-stage regression model. BF: Bayes factor of BAY-TS. We only report genes with BF > 24, corresponding for FDR 10%. Effect size: Estimate of gene effect on disease liability. R^2: cross-validation prediction accuracy. Ratio: variance ratio, the ratio of within-individual variance in imputed expression to across-individual variance. GWAS: best p-value in the WTCCC cohort within 1MB of the gene body. Catalog: best p-value reported in the NHGRI GWAS catalog within 1MB of the gene body. Diff. expr.: p-value for differential expression in independent case-control expression cohort

Gene	Disease	Chr	BAY-TS	Effect-Size	R2	Ratio	GWAS	Catalog	GEO
SARDH	BD	9	76.52	-0.43	0.03	156.52	7.71e-04	-	0.98
ZNF79	BD	9	25.18	-1.01	0.04	23.03	1.40e-03	-	0.38
WDR25	BD	14	28.07	-0.69	0.15	41.10	6.76e-04	-	-
PARL	CAD	3	46.39	0.38	0.03	136.61	1.02e-04	-	0.26
TMEM158	CAD	3	48.50	-1.03	0.01	1.82	2.09e-03	-	0.092
CYTH3	CAD	7	28.15	-1.47	0.02	0.71	1.42e-04	-	0.47
DAGLB	CAD	7	77.12	2.13	0.09	4.13	1.42e-04	-	0.39
FAAH	CD	1	44.66	0.62	0.01	8.39	1.86e-04	-	0.38
PIGC	CD	1	311.50	-1.73	0.12	34.47	5.93e-08	3e-22	0.0075
HLA-B	CD	6	9999.00	-0.76	0.02	33.40	9.22e-09	7e-32	0.34
HLA-DQA1	CD	6	10000.00	0.59	0.39	27.64	6.12e-08	9e-59	0.98
PRKAB1	CD	12	26.17	-0.87	0.03	74.76	1.33e-03	-	0.34
LRRC37A2	CD	17	32.00	-0.47	0.34	47.37	3.55e-05	-	0.086
HSBP1L1	CD	18	29.86	-0.88	0.25	32.46	3.05e-05	-	-
CIAO1	HT	2	49.00	1.07	0.03	48.08	6.04e-05	-	-
SHMT1	HT	17	9999.00	0.47	0.15	45.07	1.48e-04	-	-
HLA-B	RA	6	311.50	-1.65	0.02	16.04	9.54e-33	3e-10	0.45
HLA-DQA1	RA	6	10000.00	3.90	0.39	8.21	2.25e-86	1e-299	-
HLA-DQA2	RA	6	10000.00	1.00	0.37	16.46	2.25e-86	1e-299	0.29
HLA-DQB1	RA	6	10000.00	4.97	0.61	6.58	2.25e-86	1e-299	0.85
HLA-DQB2	RA	6	10000.00	3.16	0.37	12.05	2.25e-86	1e-299	-
HLA-DRB1	RA	6	10000.00	-1.20	0.26	15.60	2.25e-86	1e-299	0.33
HLA-DRB5	RA	6	10000.00	-3.14	0.62	8.28	2.25e-86	1e-299	0.26
HLA-G	RA	6	4999.00	-0.30	0.34	24.16	4.51e-16	-	0.75
IER3	RA	6	10000.00	1.55	0.05	16.03	9.13e-12	-	0.0089
MICA	RA	6	10000.00	0.94	0.29	62.73	9.54e-33	3e-10	0.27
BAK1	T1D	6	10000.00	-1.10	0.14	32.53	2.35e-31	-	0.21
BTN3A2	T1D	6	10000.00	-0.47	0.19	148.71	1.94e-13	-	0.44
HIST1H2AG	T1D	6	10000.00	-3.33	0.02	3.26	2.26e-15	-	-
HLA-B	T1D	6	10000.00	-4.84	0.02	2.24	1.44e-91	-	0.37
HLA-DQA1	T1D	6	10000.00	5.82	0.39	2.23	0.00e+00	5e-134	0.045
HLA-DQA2	T1D	6	10000.00	9.79	0.37	5.93	0.00e+00	5e-134	-
HLA-DQB1	T1D	6	10000.00	8.49	0.61	1.34	0.00e+00	5e-134	0.00018
HLA-DQB2	T1D	6	10000.00	7.40	0.37	26.58	0.00e+00	5e-134	0.11
HLA-DRB1	T1D	6	10000.00	-3.97	0.26	8.23	0.00e+00	5e-134	0.24
HLA-DRB5	T1D	6	554.56	-2.08	0.62	0.46	0.00e+00	5e-134	-
HMGN4	T1D	6	10000.00	-1.44	0.09	9.59	1.94e-13	-	0.22
IER3	T1D	6	10000.00	5.08	0.05	15.09	1.16e-48	-	0.13
RBMS2	T1D	12	10000.00	-3.31	0.23	1.23	8.93e-12	-	0.026
RPS26	T1D	12	10000.00	-3.10	0.02	12.37	8.93e-12	2e-25	-
C1QTNF6	T1D	22	27.82	0.83	0.06	45.02	1.17e-05	2e-08	0.73
PIGC	T2D	1	33.01	-0.98	0.12	30.33	1.73e-05	-	-
KHK	T2D	2	28.15	-0.60	0.04	36.63	8.13e-04	2e-06	0.087
TMEM158	T2D	3	103.17	-1.06	0.01	1.77	9.53e-04	-	0.2
POLR2J3	T2D	7	624.00	-0.40	0.08	4.27	1.13e-03	-	-

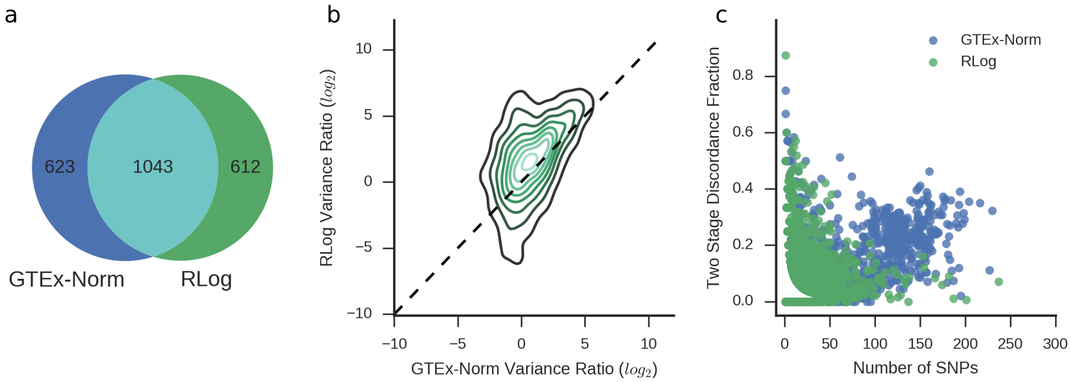


Figure 3.2: Impact of expression normalization on trained models of transcriptional regulation. a) The number of genes with cross-validation $R^2 \geq 0$. b) Density of the distribution of variance ratio (VR) for each of the 1043 common genes with cross-validation $R^2 \geq 0$. Contours denote surfaces of equal density. c) The total number of selected predictors per gene and the fraction of discordant predictors selected per gene.

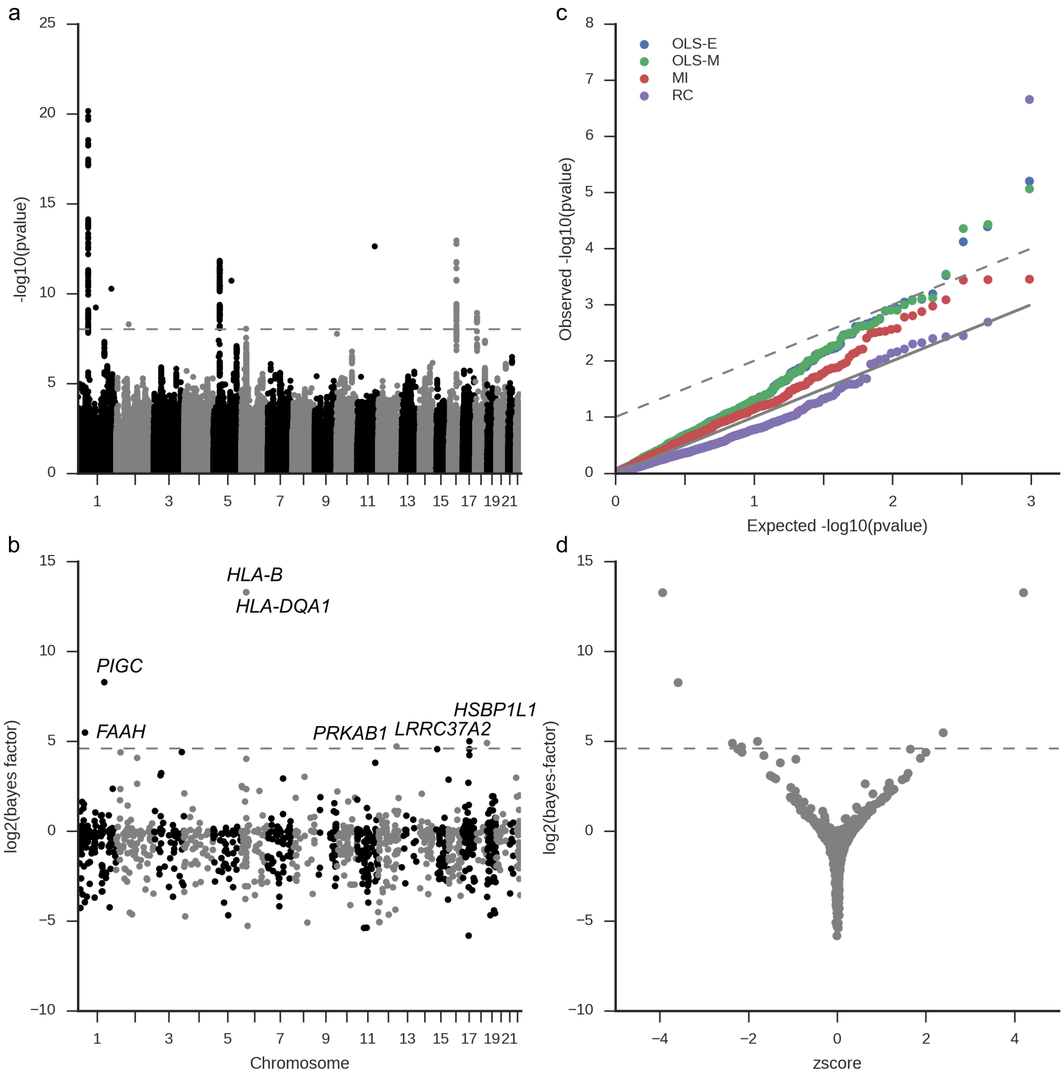


Figure 3.3: TWAS results for Crohn's Disease. a) Manhattan plot of GWAS summary statistics estimated using FastLMM and LEAP on the WTCCC GWAS cohort. b) BAY-TS TWAS on the same samples reveals 7 genes that are significantly associated with Crohn's Disease. c) Quantile-quantile plot of TWAS test statistics for frequentist methods. d) Bayes factor and z-score of BAY-TS associations. A Bayes factor of at least 24 corresponds to FDR 10%.

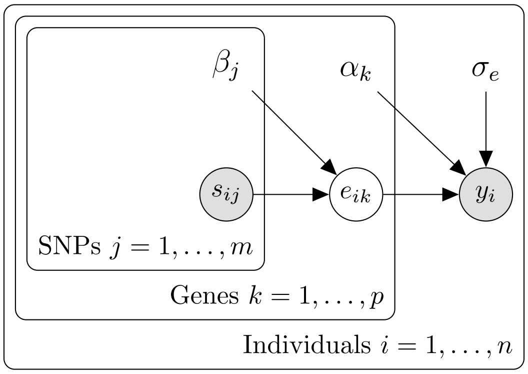


Figure S3.1: Graphical model for BAY-TS.

Table S3.1: Replication of PrediXcan TWAS genes by our imputed genotypes and BAY-TS. Logit Z: TWAS z-statistic using logistic regression with published cis-regulatory models and our imputed genotypes. Logit p-value: TWAS logistic regression p-value. R^2: Computed R^2 in the Rlog trained models. OLS-E: TWAS p-value computed using OLS-E with liability estimates using LEAP and RLog trained models. BAY-TS: TWAS Bayes factor computed using BAY-TS. A Bayes factor > 24 corresponds to FDR 10%.

Disease	Gene	Chr	P-Xcan z	P-Xcan p-value	Logit Z	Logit p-value	R^2	OLS-E	BAY-TS
RA	DCLRE1B	1	-6.68	2.46e-11	-6.42	1.35e-10	-	-	-
RA	PTPN22	1	5.67	1.44e-8	5.78	7.61e-09	0.0097	0.0012	1.8
BD	PTPRE	10	4.94	7.71e-7	4.27	1.95e-05	-	-	-
CD	IL23R	1	5.23	1.74e-7	5.98	2.22e-09	-	-	-
CD	APEH	3	5.14	2.77e-7	5.13	2.92e-07	-	-	-
CD	ZNF300	5	-4.98	6.29e-7	-4.40	1.07e-05	0.027	0.49	0.94
CD	NKD1	16	-4.91	8.91e-7	-4.97	6.78e-07	-	-	-
CD	BSN	3	-4.68	2.89e-6	-3.37	7.39e-04	-	-	-
CD	GPX1	3	-4.62	3.87e-6	-3.26	1.13e-03	-	-	-
CD	SLC22A5	5	-4.54	5.75e-6	-3.62	2.98e-04	0.19	0.00088	0.87
HT	KCNN4	19	-4.70	2.62e-6	-2.99	2.83e-03	0.043	8.7e-05	19
T1D	DCLRE1B	1	-7.84	4.34e-15	-6.53	6.75e-11	-	-	-
T1D	ZNF165	6	7.30	2.92e-13	7.00	2.58e-12	-	-	-
T1D	ERBB3	12	-6.81	1.01e-11	-6.83	8.67e-12	0.016	0.0066	1
T1D	EGFL8	6	6.33	2.52e-10	-11.36	6.39e-30	-	-	-
T1D	C6orf136	6	-6.33	2.52e-10	-3.60	3.22e-04	-	-	-
T1D	HCG27	6	-6.33	2.52e-10	3.36	7.93e-04	-	-	-
T1D	GTF2H4	6	6.33	2.52e-10	-1.89	5.85e-02	-	-	-
T1D	DDR1	6	6.33	2.52e-10	0.22	8.27e-01	-	-	-
T1D	AGER	6	-6.33	2.52e-10	5.91	3.41e-09	-	-	-
T1D	POU5F1	6	6.33	2.52e-10	0.25	8.00e-01	-	-	-
T1D	ATP6V1G2	6	6.33	2.52e-10	1.28	2.01e-01	-	-	-
T1D	TUBB	6	6.33	2.52e-10	4.92	8.51e-07	-	-	-
T1D	AIF1	6	6.33	2.52e-10	0.02	9.81e-01	-	-	-
T1D	CYP21A2	6	-6.33	2.52e-10	10.31	6.51e-25	-	-	-
T1D	LSM2	6	6.33	2.52e-10	7.85	4.24e-15	-	-	-
T1D	VARS2	6	6.33	2.52e-10	-9.70	3.01e-22	-	-	-
T1D	APOM	6	-6.33	2.52e-10	4.03	5.57e-05	-	-	-
T1D	DDAH2	6	-6.33	2.52e-10	-6.81	9.63e-12	-	-	-
T1D	NCR3	6	-6.33	2.52e-10	-0.05	9.62e-01	-	-	-
T1D	ZSCAN16	6	6.16	7.37e-10	5.94	2.93e-09	-	-	-
T1D	ZKSCAN4	6	6.15	7.73e-10	5.90	3.57e-09	-	-	-
T1D	PTPN22	1	5.83	5.41e-9	6.28	3.28e-10	0.0097	0.0003	2
T1D	RPS26	12	5.82	6.00e-9	5.76	8.34e-09	0.022	3.6e-10	1e+04
T1D	GDF11	12	-5.75	9.11e-9	-4.01	6.09e-05	-	-	-
T1D	SUOX	12	-5.47	4.49e-8	-5.31	1.09e-07	-	-	-
T1D	BTN3A2	6	-5.11	3.30e-7	-4.94	7.75e-07	0.19	8.3e-08	1e+04
T1D	PRSS16	6	4.83	1.34e-6	3.15	1.62e-03	-	-	-
T1D	FAM109A	12	-4.76	1.94e-6	-4.43	9.59e-06	0.033	-	1
T1D	SH2B3	12	4.67	3.05e-6	5.50	3.81e-08	-	-	-

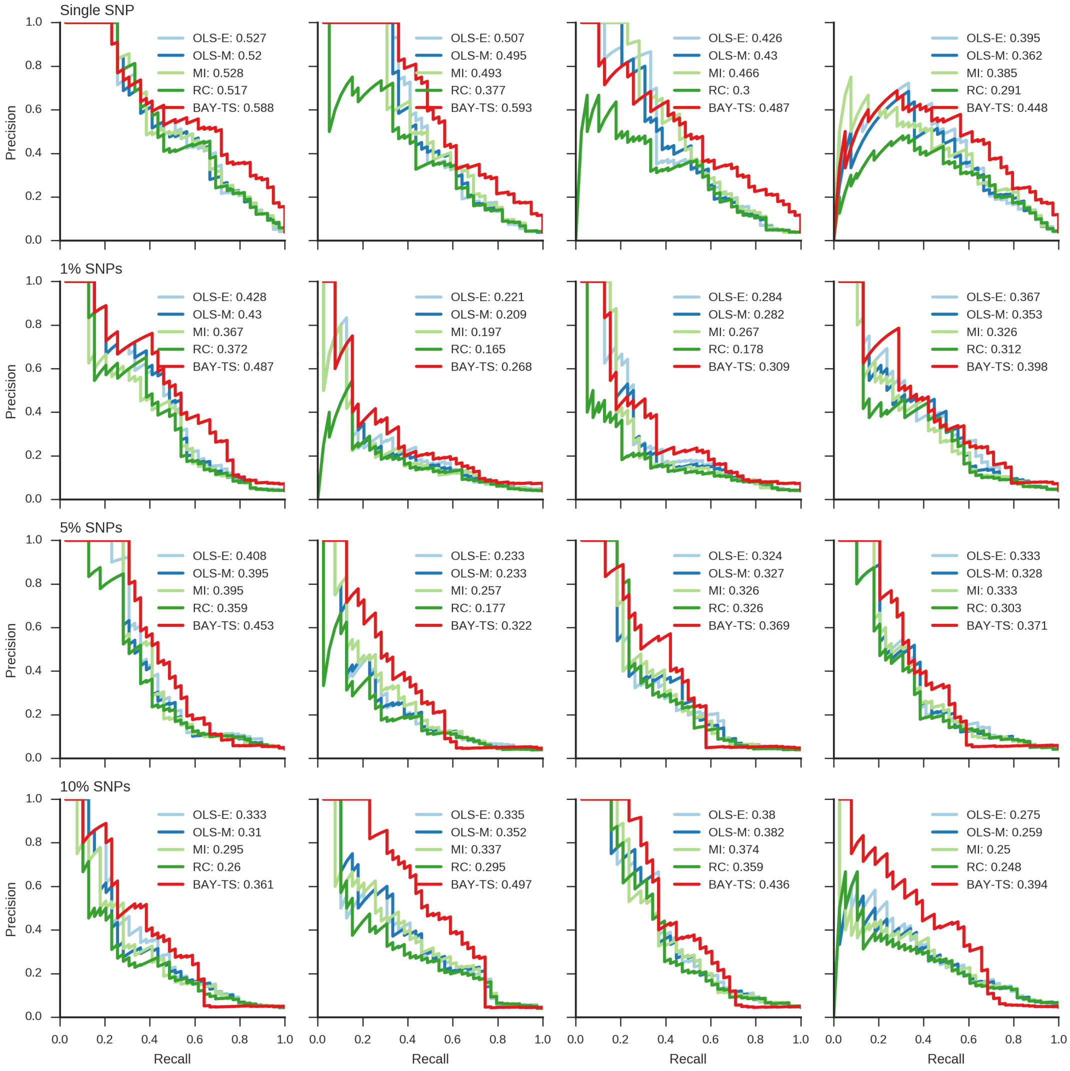


Figure S3.2: Precision-recall curves for the proposed methods over four simulated cis-regulatory architectures and four trials.

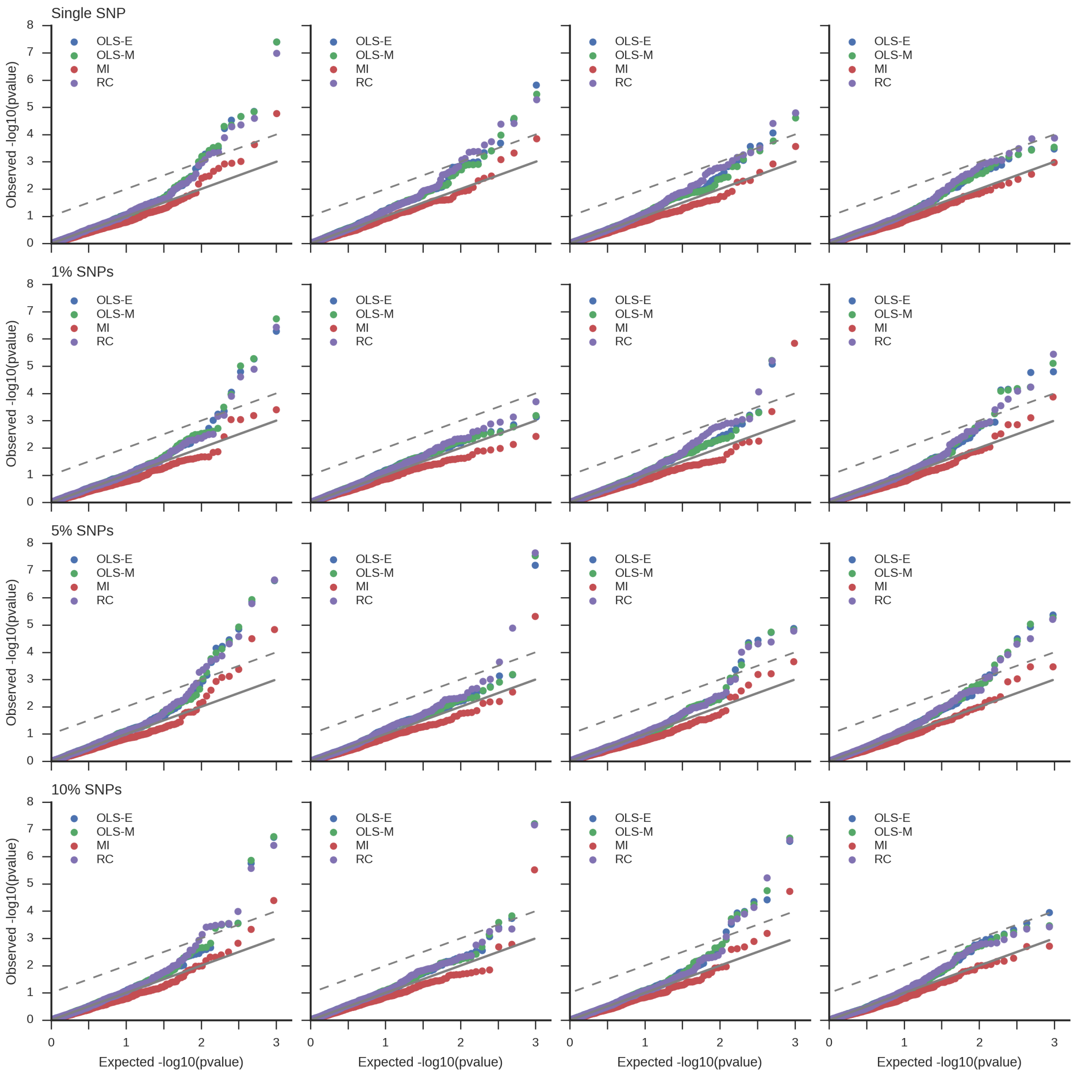


Figure S3.3: Quantile-quantile plots for the frequentist methods across four simulated cis-regulatory architectures and four trials.

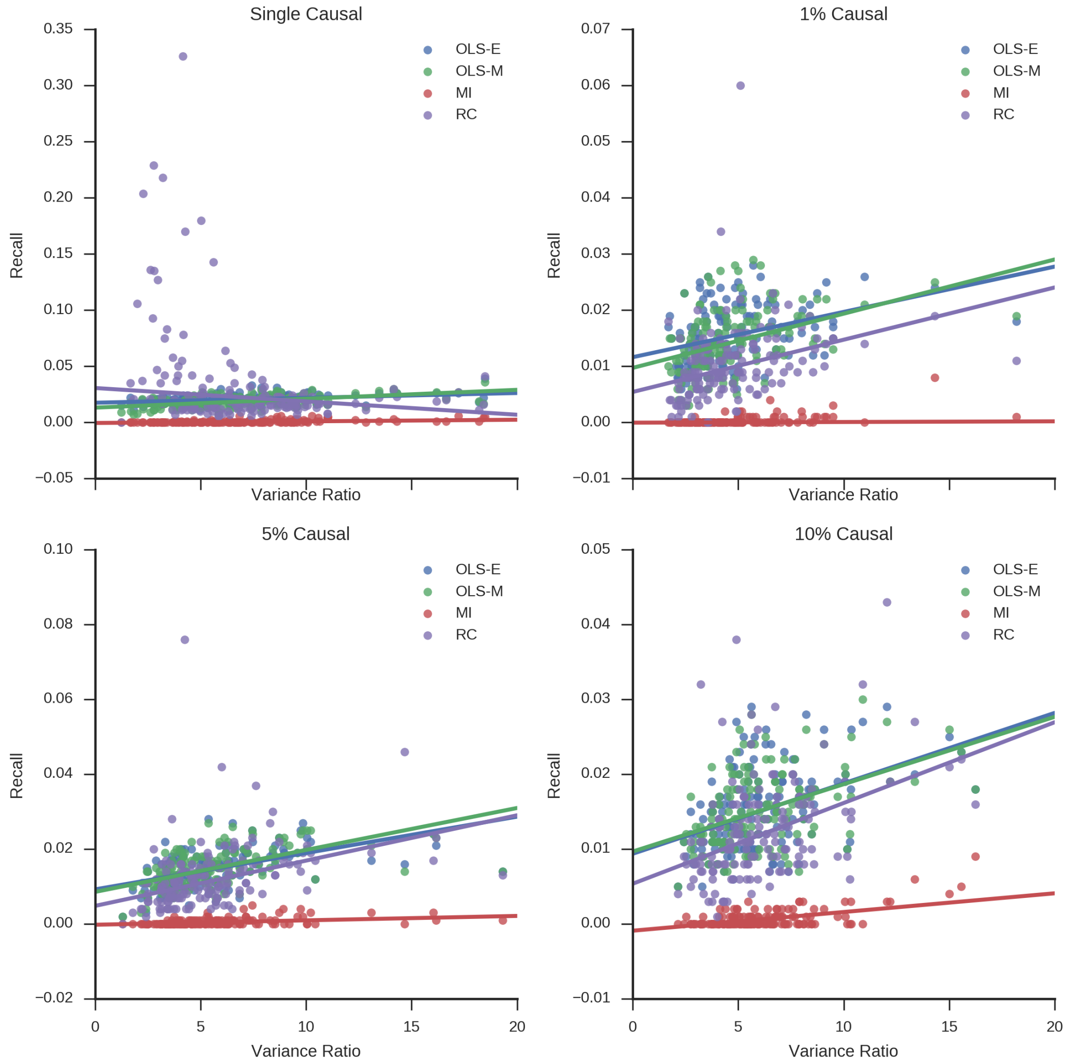


Figure S3.4: Power vs. variance ratio across four simulated cis-regulatory architectures. Each of 150 consistently heritable genes is simulated 1,000 times, and power is estimated as the number of trials for which the gene was significantly associated ($p < 0.05$) with the simulated phenotype after Bonferroni correction for 1,000 tests.

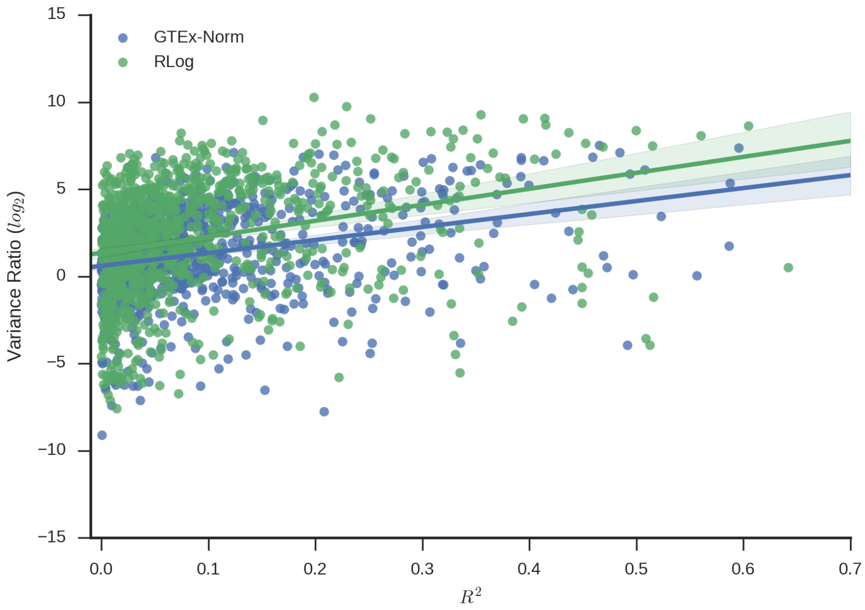


Figure S3.5: Mean variance ratio across the diseases. For genes with $R^2 \geq 0$, the mean variance ratio across the diseases. As expected, there is a high correlation between variance ratio and R^2. However, variance ratio is higher for RLog compared to GTEx-Norm transcription models.

Table S3.2: OLS-E results using RLog learned models of transcription.

Gene	Disease	Chr	OLS-E	Effect-Size	R2	Ratio	GWAS	Catalog	GEO
MCM6	CD	2	4.03e-05	1.62	0.02	5.70	1.64e-05	-	0.67
IER3	CD	6	7.54e-05	-0.54	0.05	18.59	1.59e-06	7e-07	1.8e-05
ZC3HAV1	CD	7	6.30e-06	-4.69	0.04	8.30	5.57e-05	-	0.00018
LRRC37A2	CD	17	2.98e-04	-0.48	0.34	47.37	3.55e-05	-	0.086
MIF4GD	CD	17	2.60e-05	-160.54	0.03	0.06	5.26e-06	-	-
SHMT1	HT	17	3.40e-05	0.31	0.15	45.07	1.48e-04	-	-
KCNN4	HT	19	8.70e-05	-0.58	0.04	5.99	1.63e-04	-	-
AP4B1	RA	1	7.05e-04	2.70	0.04	5.80	3.82e-24	-	1.7e-05
HLA-B	RA	6	7.62e-04	-0.38	0.02	16.04	9.54e-33	3e-10	0.45
HLA-C	RA	6	1.38e-05	-0.18	0.40	150.82	1.26e-32	3e-10	0.75
HLA-DQA1	RA	6	1.34e-26	1.16	0.39	8.21	2.25e-86	1e-299	-
HLA-DQA2	RA	6	3.37e-43	0.73	0.37	16.46	2.25e-86	1e-299	0.29
HLA-DQB2	RA	6	3.62e-05	0.42	0.37	12.05	2.25e-86	1e-299	-
HLA-DRB1	RA	6	8.42e-33	-1.08	0.26	15.60	2.25e-86	1e-299	0.33
HLA-DRB5	RA	6	3.66e-06	0.98	0.62	8.28	2.25e-86	1e-299	0.26
IER3	RA	6	7.92e-05	0.55	0.05	16.03	9.13e-12	-	0.0089
MICA	RA	6	6.52e-08	0.33	0.29	62.73	9.54e-33	3e-10	0.27
ZNF783	RA	7	4.89e-04	-52.68	0.02	0.08	2.02e-04	-	-
CLEC12A	RA	12	2.73e-04	-0.09	0.35	583.64	1.75e-06	-	0.84
CLEC12B	RA	12	3.00e-04	-0.08	0.39	545.61	1.75e-06	-	0.23
NYNRIN	RA	14	8.64e-04	2.02	0.07	1.57	1.78e-04	-	-
AP4B1	T1D	1	9.92e-05	3.08	0.04	5.93	3.18e-25	-	0.0027
BMP8A	T1D	1	1.15e-03	2.60	0.02	2.52	2.79e-03	-	0.5
CDC7	T1D	1	8.38e-04	0.67	0.05	13.84	1.15e-03	-	0.025
LBX2	T1D	2	5.14e-05	1.25	0.05	37.19	5.40e-05	-	-
BTN3A2	T1D	6	8.31e-08	-0.22	0.19	148.71	1.94e-13	-	0.44
HIST1H2AG	T1D	6	2.92e-10	-7.90	0.02	3.26	2.26e-15	-	-
HLA-B	T1D	6	4.09e-06	-0.79	0.02	2.24	1.44e-91	-	0.37
HLA-DQA1	T1D	6	4.70e-61	2.75	0.39	2.23	0.00e+00	5e-134	0.045
HLA-DQA2	T1D	6	1.42e-59	1.63	0.37	5.93	0.00e+00	5e-134	-
HLA-DQB1	T1D	6	3.99e-28	-1.60	0.61	1.34	0.00e+00	5e-134	0.00018
HLA-DQB2	T1D	6	5.60e-22	0.83	0.37	26.58	0.00e+00	5e-134	0.11
HLA-DRB1	T1D	6	2.98e-61	-2.17	0.26	8.23	0.00e+00	5e-134	0.24
HMGN4	T1D	6	6.68e-05	-0.85	0.09	9.59	1.94e-13	-	0.22
IER3	T1D	6	9.03e-28	1.52	0.05	15.09	1.16e-48	-	0.13
MICA	T1D	6	2.08e-04	0.25	0.29	65.78	1.44e-91	-	0.83
RBMS2	T1D	12	2.48e-11	-10.08	0.23	1.23	8.93e-12	-	0.026
RPS26	T1D	12	3.58e-10	3.39	0.02	12.37	8.93e-12	2e-25	-
XRCC6BP1	T1D	12	1.10e-03	0.26	0.21	76.21	4.16e-04	-	-
HEATR6	T1D	17	1.32e-03	-0.56	0.13	29.73	4.82e-04	-	0.18

Table S3.3: MI results using RLog learned models of transcription.

Gene	Disease	Chr	MI	Effect-Size	R2	Ratio	GWAS	Catalog	GEO
SHMT1	HT	17	5.17e-05	0.31	0.15	45.07	1.48e-04	-	-
HLA-C	RA	6	1.24e-04	-0.18	0.40	150.82	1.26e-32	3e-10	0.75
HLA-DQA1	RA	6	1.24e-07	1.28	0.39	8.21	2.25e-86	1e-299	-
HLA-DQA2	RA	6	5.07e-28	0.71	0.37	16.46	2.25e-86	1e-299	0.29
HLA-DRB1	RA	6	2.40e-15	-1.07	0.26	15.60	2.25e-86	1e-299	0.33
MICA	RA	6	2.63e-06	0.32	0.29	62.73	9.54e-33	3e-10	0.27
CLEC12A	RA	12	3.04e-04	-0.09	0.35	583.64	1.75e-06	-	0.84
CLEC12B	RA	12	3.59e-04	-0.09	0.39	545.61	1.75e-06	-	0.23
LBX2	T1D	2	2.16e-04	1.28	0.05	37.19	5.40e-05	-	-
BTN3A2	T1D	6	3.08e-07	-0.22	0.19	148.71	1.94e-13	-	0.44
HLA-DQA1	T1D	6	4.97e-12	2.87	0.39	2.23	0.00e+00	5e-134	0.045
HLA-DQA2	T1D	6	3.81e-09	1.58	0.37	5.93	0.00e+00	5e-134	-
HLA-DQB2	T1D	6	1.21e-08	0.82	0.37	26.58	0.00e+00	5e-134	0.11
HLA-DRB1	T1D	6	1.33e-30	-2.20	0.26	8.23	0.00e+00	5e-134	0.24
IER3	T1D	6	9.30e-16	1.60	0.05	15.09	1.16e-48	-	0.13
RPS26	T1D	12	3.45e-06	3.46	0.02	12.37	8.93e-12	2e-25	-

Table S3.4: RC results using RLog learned models of transcription. For RC, we additionally filter on Ratio $>$ 1 due to known biases for Ratio $\leq$ 1.

Gene	Disease	Chr	RC	Effect-Size	R2	Ratio	GWAS	Catalog	GEO
CARKD	BD	13	4.81e-03	0.63	0.07	7.25e+01	0.00024	-	-
LGALS9B	BD	17	9.23e-04	-0.21	0.04	3.54e+01	9.1e-05	-	-
NT5M	BD	17	4.71e-03	0.45	0.04	1.95e+01	0.001	-	-
GSTT1	BD	22	1.50e-03	9.87	0.30	2.22e+00	0.00097	-	0.26
LBX2	CD	2	2.04e-03	1.77	0.05	1.02e+01	3.1e-05	-	-
ZC3HAV1	CD	7	2.18e-07	-6.53	0.04	8.30e+00	5.6e-05	-	0.00018
KCNN4	HT	19	2.89e-03	-0.76	0.04	5.99e+00	0.00016	-	-
AP4B1	RA	1	4.98e-04	3.49	0.04	5.80e+00	3.8e-24	-	1.7e-05
HLA-DQA1	RA	6	1.27e-05	1.52	0.39	8.21e+00	2.3e-86	1e-299	-
HLA-DQA2	RA	6	3.19e-05	0.74	0.37	1.65e+01	2.3e-86	1e-299	0.29
HLA-DRB1	RA	6	6.04e-05	-1.12	0.26	1.56e+01	2.3e-86	1e-299	0.33
HLA-DRB5	RA	6	8.09e-05	1.30	0.62	8.28e+00	2.3e-86	1e-299	0.26
IER3	RA	6	5.27e-03	0.68	0.05	1.60e+01	9.1e-12	-	0.0089
MICA	RA	6	1.50e-07	0.33	0.29	6.27e+01	9.5e-33	3e-10	0.27
ASNS	RA	7	2.28e-03	2.89	0.07	2.43e+00	0.00031	-	0.02
MRPS16	RA	10	3.60e-03	-5.55	0.02	6.18e+00	0.027	-	0.44
CLEC12A	RA	12	4.64e-03	-0.09	0.35	5.84e+02	1.8e-06	-	0.84
AP4B1	T1D	1	1.17e-05	4.41	0.04	5.93e+00	3.2e-25	-	0.0027
DRAM2	T1D	1	5.03e-03	-6.31	0.03	3.04e+00	0.0039	-	0.51
LBX2	T1D	2	1.06e-04	1.39	0.05	3.72e+01	5.4e-05	-	-
BTN3A2	T1D	6	3.22e-06	-0.22	0.19	1.49e+02	1.9e-13	-	0.44
HIST1H2AG	T1D	6	1.70e-09	-11.63	0.02	3.26e+00	2.3e-15	-	-
HLA-DQA1	T1D	6	8.10e-06	5.00	0.39	2.23e+00	0	5e-134	0.045
HLA-DQA2	T1D	6	1.02e-04	1.94	0.37	5.93e+00	0	5e-134	-
HLA-DQB2	T1D	6	2.76e-04	0.87	0.37	2.66e+01	0	5e-134	0.11
HLA-DRB1	T1D	6	2.56e-06	-2.54	0.26	8.23e+00	0	5e-134	0.24
IER3	T1D	6	4.08e-12	1.76	0.05	1.51e+01	1.2e-48	-	0.13
MICA	T1D	6	8.71e-04	0.23	0.29	6.58e+01	1.4e-91	-	0.83
RPS26	T1D	12	3.71e-06	3.79	0.02	1.24e+01	8.9e-12	2e-25	-
HEATR6	T1D	17	1.56e-03	-0.59	0.13	2.97e+01	0.00048	-	0.18
C1QTNF6	T1D	22	2.55e-03	0.41	0.06	4.50e+01	1.2e-05	2e-08	0.73
GSTT1	T2D	22	3.07e-04	8.29	0.30	2.16e+00	0.003	-	0.97

3.6 Acknowledgements

Chapter 3, in part is currently being prepared for submission for publication of the material. Kunal Bhutani, Abhishek Sarkar, Yongjin Park, Manolis Kellis, Nicholas J Schork. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Longitudinal metabolome, microbiome, and transcriptome profiling of a germline TP53 mutation carrier

4.1 Abstract

Patients suffering from Li-Fraumeni syndrome have a high incidence rate of cancer during their lifetime as a result of a germline *TP53* mutation. Clinical standard of care includes regular surveillance, including several imaging and biochemical screens for cancer. Here, we test the utility of molecular surveillance for a 16 year old female with a T167C *TP53* mutation through longitudinal profiling of her metabolome, microbiome, and transcriptome. We devised analytical techniques to accommodate several sources of variation, particularly those associated with longitudinal sampling.

We did not find any significant deviations from healthy baseline measurements indicate of cancer risk, but we highlight power limitations and propose several strategies for increasing sensitivity in N-of-1 study designs.

4.2 Introduction

Li-Fraumeni syndrome (LFS) is a rare autosomal dominant disorder that is caused by a germline mutation in the tumor protein 53 gene (*TP53*), which encodes tumor protein 53 (p53) [LFJ69]. As a transcription factor that responds to cellular stress, p53 plays a role in several important cancer-related pathways such as DNA repair, cell-cycle arrest, senescence, and apoptosis[Geo11]. Carriers of many nonsynonymous or nonsense germline p53 mutations have a high incidence of cancer with 68% of males and 93% of females developing multiple cancers in their lifetime. Compared to the general population, the average onset of cancer in LFS is much younger, whereas 2% of the cancer cases in the general population occur before age 30, and 11% before age 50, those figures rise to 56% and 100%, respectively in LFS families [HKHH02]. There is high predisposition towards certain cancer subtypes including soft tissue sarcoma, osteosarcoma, breast cancer, brain tumors, adrenocortical carcinoma, and leukemias [SZNG93]. Patients are typically instructed to avoid known carcinogens including tobacco, excessive alcohol, long sun explore, and other known occupational exposures.

Early detection of cancer is an integral part of the clinical strategy to combat LFS. Regular surveillance for mutation carriers includes several imaging and biochemical screens: ultrasound, brain MRI, total body MRI, mammography and breast MRI, and blood tests for several cancer biomarkers. A prospective study on the efficacy of added cancer surveillance showed a significant increase in 3-year overall survival rate for the patients in the surveillance group compared to the non-surveillance group [VTS+11]. To our knowledge, there has been no previous study that has additionally surveyed the metabolome, microbiome, and transcriptome of a LFS patient and used the resulting information to assess health status changes that might be indicative of early signs of cancer.

Here, we show results from a 16 month molecular surveillance study of a 16 year old female with a germline T167C p53 mutation. We assessed her mutational burden by analyzing whole genomes of the entire nuclear family and, via longitudinal monitoring, also sought outliers and linear trends in her metabolome, microbiome, and transcriptome profiling. We did not find any significant deviations that might be indicative of the presence of a cancer in our study, but we highlight challenges in achieving high sensitivity in small n-of-1 study designs due to inherent biological, assay and instrumentation noise.

4.3 Results

The father is a healthy carrier of the heterozygous T167C variant, and transmitted it to the daughter and son in the nuclear family (Figure 4.1a). The mother is unaffected. The son developed neuroglioblastoma at age 12 and is deceased. The daughter, the proband, is a healthy, 16 year old with no past indications of cancer.

4.3.1 Whole Genome Sequencing

We sequenced the father, mother, son, and daughter at an average read depth of 39x, with 97% of the autosomal genome covered by at least 15 reads. We used

GATK best practices (Methods) for group calling to align the reads and call single nucleotide variants (SNVs) and insertions or deletions (INDELs) [MHB$^+$10].

We hypothesized that the p53 variant will lead to malfunction of DNA repair mechanism, and characterized the de-novo mutations in the offspring to study the impact of the p53 variant. De-novo mutations, or mutations present in the offspring germline that are not present in the parents' germline, occur either in the parents' sex cells or in progenitor cells of the offspring blood cells. We are interested in the latter case, or somatic de-novo variants that occurred in the presence of the p53 variant in the offspring. There are 3071 and 886 de-novo SNVs and 3402 and 2962 de-novo INDELs in the daughter and son, respectively using variant recalibration steps that favor sensitivity (Methods, Figure 4.2a). We studied the mutational signatures of the SNVs by categorizing the mutations based on the upstream and downstream bases. This results in 96 total mutational subtypes: 4 upstream bases, 6 mutation types, and 4 downstream bases. The highest fraction of mutations for both offspring occur in homopolymer contexts such 3':T, C→T, 5':T (Figure 4.2b, c).

Previous studies have shown it is possible to deconvolve mutational signatures using matrix factorization into constitutive DNA repair malfunctions in cancer [AS14]. We used deconstructSigs to compare the mutational signatures of the offspring to 30 published COSMIC cancer signatures [FBG$^+$15b, RMH$^+$16]. Signature 3 and 12 are highly abundant in both offspring, whereas the daughter's signature further factorizes into Signature 2, 8 and 11. Signature 3 is linked to DNA double-stand break repair by homologous recombination and to mutations in the DNA repair genes BRCA1 and BRCA2 in cancer patients. p53 is a known mediator of this DNA repair pathway, but has not been linked to the mutational signature [MP14]. The aetiology of Signature 12 is unknown. After subtracting the aggregated effect of COSMIC signatures, the residual mutational signature is dominated by the homopolymer mutations (Supplementary Figure S4.1).

We additionally surveyed the genome for cancer predisposition variants to guide molecular and phenotype surveillance. We identified 265 (5 de-novo) and 267 (2 de-novo) loss-of-function variants using LOFTEE [MBF$^+$12] in the genome of the daughter and son, respectively using a variant recalibration steps targeted towards high specificity (Methods). Two shared variants impact known oncogenes: HNF1A and SETBP1[FBG$^+$15b]. Mutations in HNF1A are linked to hepatic adenoma and hepatocellular carcinoma [ILZR10], whereas mutational burden in SETBP1 is associated with Schnizel-Giedion syndrome[HvBG$^+$10]. Patients suffering from the condition have a higher prevalence for neuroepithelial cancers than the general population.

4.3.2 Longitudinal Profiling

We conducted longitudinal profiling of the microbiome, metabolome, and transcriptome of the daughter over a 16-month period from April 2014 to August 2015 (Figure 4.1b). As part of real-time monitoring, when possible, we profiled each new sample in the context of all previous samples to identify any outliers or linear trends for known cancer biomarkers. For brevity and completeness, we present results here from the final time-point. However, we highlight deviations that would be impactful during surveillance.

Metabolomics

We performed real-time monitoring of the metabolome over 14 months across 13 samples using nine different instrumentation runs (Figure 4.3a). At each run, we re-profiled all available samples due to known issues with comparing values across instrumentation runs (Methods, Supplementary Figure S4.2). On average, we measured 601 metabolites per instrumentation run, but here we only consider the 279 metabolites that were measured confidently in each instrumentation run.

In addition to analyzing each instrumentation run separately, we devised a Bayesian strategy to pool information across runs to separate metabolite level variation from instrumentation variation. The model produces a latent abundance of each metabolite, and finds a scaling factor for each run to allow direct comparisons across runs (Methods). We highlight scaling and latent levels for Gycocholenate Sulfate as an example in Figure 4.3b). The latent levels have a significant robust linear regression association between the metabolite abundance and days (p-value: 4.1e-5; Bonferroni-corrected p-value cutoff: 1.8e-4), but the added variance for Runs 8 and 9 lead to non-significant associations (p-value: 3.97e-3 and 2.99e-3, respectively). In general, the Bayesian strategy yields more power for discovering associations between metabolite levels and time by removing instrumentation variation from metabolite levels.

Overall, we find 15, 9, and 13 significant associations for latent, Run 8, and Run 9 metabolite levels, respectively after a Bonferroni correction for 279 metabolites. There is high overlap across the different measurements, but Run 8 produced no unique associations (Figure 4.3c). On average, Run 8 added 33% more variance to each metabolite compared to 13% for Run 9 (Figure 4.3d). In the absence of any variation from instrumentation, the current study design of 13 samples has 80% to detect correlation of 0.9 between metabolite levels and days. If we add instrumentation noise, only 65 metabolites require less than 20 samples to achieve 80% power for the same latent correlation (Figure 4.3e, Table 4.1, Methods). We did not find any significant deviations for metabolites that have been previously associated with abnormal abundance in cancer samples compared to normal samples in population studies: N1-methyladenosine and hydrocinnamate (3-phenylpropionate) [WJG^{+}13]. Hydrocinnamate had very low instrumentation noise with only 0.04% relative variance added, whereas N1-methyladenosine had 167% relative noise added for Run 9 (Supplementary Figure S4.2). Given the high amount of instrumentation noise for N1-methyadenosine, there would be a huge reduction in power to detect true deviations from baseline in an n-of-1 study design with a single instrumentation run.

Microbiome

We profiled 11 fecal microbiome samples over five months from July 2014 to December 2014 (Methods). On average, we generated 12,103,824 shotgun sequencing reads per sample. The relative abundance of the 6 phylum exhibited high variations with no significant time-variant properties (Figure 4.4a). We analyzed 6 genera that have been previously associated with colorectal cancer: Porphyromonas, Peptostreptococcus, Parvimonas, Fusobacterium, Collinsella, and Anaerococcus [BRRS16]. Collinsella had a high deviance on September 16, but quickly reverted to baseline 8 days later (Figure 4.4b). Similar deviations occurred at later time points for the other genera, but in all cases, there was a rebound back into baseline range. We note previous studies have shown there is known high variation in microbiome samples due to diet [DMC+13], and larger sample sizes are necessary to quantify baseline variation. We also analyzed outliers and linear trends by mapping sequences to the TigrFam database [HSW03] of sequence functions. We did not observe any significant deviations for any biological processes.

Transcriptome

We analyzed 11 samples of the patient's transcriptome over 12 months 4.1b. Additionally, we profiled the mother's transcriptome three times. We sequenced the RNA collected from whole blood in three different batches, and principal component analysis of the transcriptome revealed a large variation between batches that contained a globin-clearing library preparation step and those that did not (Figure 4.5a, Methods). All samples were retrospectively sequenced using the same library preparation and sequencer.

We looked for deviance in the expression levels of 222 curated oncogenes [FBG+15a] for each of the time points using a simulated surveillance study design, where we analyzed each time point after February 2015 using only time points up to the given date. May 11, 2015 had a large number of outlying oncogenes with high Cook's distances compared to all other samples using both the simulated surveillance (Supplementary Figure S4.3) and retrospective analysis of all samples together (Figure 4.5b). However, the genes reverted to baseline expression levels at later time points (4.5c). It is unclear if these changes are caused by underlying biological changes or by other cofactors such as diet and exercise or by cyclical deviations resulting from circadian rhythm or seasonal patterns of expression. We looked for enrichment in KEGG pathways[KG00] outliers across all genes separately for each sample (Methods).

We did not find any significant associations for cancer-related pathways.

The most significant associations for linear trends were similarly found for May 11 and June 10, 2015 time-points using a simulated surveillance strategy. However, these genes similarly reverted to baseline expression on later dates. Enrichment analysis did not reveal any significant associations across top 100 associations from each simulated surveillance for each sample.

4.4 Discussion

During a 16-month molecular surveillance study of a 16 year-old germline TP53 gene mutation carrier, we did not find any significant trends in her biomarker profile that might be indicative of the presence of a tumor. We additionally characterized the genome of the carrier using blood cells and found a high rate of de-novo mutations. Although she had less than one-third the number of de-novo mutations as her brother, both offspring had mutational signatures that deconvolved into malfunctions to homologoous DNA repair. However, we note that the bulk of the de-novo mutations occurred in homopolymer contexts, which have been previously associated with sequencing errors [RRC+13].

We employed two strategies for longitudinal profiling: a simulated real-time monitoring strategy and a retrospective analysis strategy. The goal of these analyses was to see if any changes in her profile occurred that deviated from her previous profile in ways that might be indicative of a significant health status change. Although we identified several instances where outlier detection using real-time monitoring yielded a signal, we found that this signal quickly receded back to baseline levels at the next time point, suggesting that those signals were likely false positives. Such deviations can occur as a result of both natural biological variation or measurement noise inherent to the biological system being interrogated. Any recommendations for clinical care should properly account for these sources of variation. The larger number of samples in retrospective analysis yielded refined estimates of baseline range and we found fewer number of deviations from baseline. For future study designs, we recommend a strategy of first defining a baseline range through collection of several samples and then real-time monitoring to limit false positives. The number of samples necessary for defining a baseline range depends on the biological variation and measurement error associated with the particular biomarker. Due to uneven sampling and limited sample sizes in our study, we did not consider autocorrelation in the data, but biological processes should exhibit such properties and analytical methods should properly account for them.

As we highlight in our analysis of metabolomics, it is possible to model instrumentation noise using multiple measures of the same quantity. In metabolomics, the instrumentation noise results from the detection of the m/z spectra, the mapping of the m/z spectra to a specific metabolite, and then the quantification of the area under the curve for the metabolite. In cases where there are multiple metabolites that have overlapping m/z spectra or metabolites with low signal, these steps yield high variation in abundance values. By intelligently averaging signal over multiple measurements, it is possible to separate instrumentation noise from true biological variation. We identified 65 metabolites with low instrumentation noise that could be sensitive enough for detecting health status changes via linear regression analysis for less than 20 samples. Similar multiple measurement strategies could be pursued for microbiome and transcriptome profiling.

Although previous studies have shown that retrospective n-of-1 studies can yield insights for linking molecular and macroscopic phenotypes [CMLPT+12], real-time molecular surveillance currently provides limited utility due to several technical and analytical challenges. Researchers interested in pursuing such study designs should carefully consider issues such as homogeneity of experimental procedures, biological and experimental variation, the number of samples and data time point collections necessary for proper inference, and the ease with which such procedures might be integrated into clinical care.

4.5 Methods

4.5.1 Whole Genome Sequencing

Whole blood samples from each family member were sent to Illumina for sequencing. Reads were extracted from Illumina Casava aligned BAM files using the HTSLib [Li11] by first shuffling the reads and then extracting interleaved reads. These reads were then processed through the GATK Best Practices workflow for Variant Calling v.2.6, which included first aligning the reads using BWA 0.7[LD10] with the mem option, marking duplicates, pursuing local realignment, considering base quality recalibration and finally using the reduce reads options using Genome Analysis ToolKit v. 3.1[MHB+10]. The multi-sample option in HaplotypeCaller was used to call variants in the family with variant recalibration applied at tranch levels 0 - 90.00 and 90.00 - 99.00 to filter to a high confidence set of variants.

De-novo variants were identified using gemini v. 16 [PCKQ13] and its mendelian errors option. We selected only plausible de novo variants from the output, further filtering down to posterior probability of mutation greater than 0.99. We did not include cases where the parents were homozygous alternate and the children had a different alternate mutation or a mutation to heterozygous reference.

Putative LOF variants, defined here as stop-gained, nonsense, frameshift, and splice site disrupting, were identified using the LOFTEE plugin for VEP and Ensembl release 85. LOFTEE assigns confidence to loss of function annotations based on position of variant in the transcript, proximity to canonical splice sites, and conservation of the putative LOF allele across primates. For our analysis we used default LOFTEE filter setting and only included high confidence predicted LOF variants. A variant was called LOF if it received a high confidence LOF prediction in any Ensembl transcript. We further filtered variants down to tranch levels 0 - 90.00 only.

4.5.2 Metabolome

We sent whole blood samples to Metabolon, Inc. for metabolomics profiling. The laboratory divided each sample into 12 homogeneous aliquots to allow new sample would be re-profiled with all previous samples. Each instrument run reports the raw abundance of each metabolite as produced by their internal processing pipeline of the m/z spectra. To pool information across the runs, we devised a Bayesian latent model as illustrated in Supplementary Figure **??**. Briefly, there is a hidden, latent metabolite level z_t for each time point, but we only observe y_{tr} for each time point and run. y_{tr} adds additional noise specific to the instrument σ_r that is sampled from a larger hierarchical space of all instrument noise. It additionally scales this value by a scaling factor β_r.

$$\begin{aligned}
z_t &\sim \mathcal{N}(\text{Constant}, \sigma_m^2) \\
y_{tn} &\sim \mathcal{N}(\beta z_t, \beta^2 \sigma_r^2) \\
P(\beta) &= \mathcal{N}^+(0, 1e7) \\
P(\sigma_m^2) &= C^+(0, 10) \\
P(\sigma_n^2) &= C^+(0, \zeta) \\
P(\zeta) &= U(0, 100)
\end{aligned}$$

Here, $\mathcal{N}(\cdot, \cdot)$ denotes the Gaussian density, $\mathcal{N}^+(\cdot, \cdot)$ the half-Gaussian density, and $C^+(\cdot, \cdot)$ the half-Cauchy density. We implemented the model using PyMC3 [SWF16] and used the Metropolis-Hastings algorithm to perform inference. We ran the MCMC chain for 100,000 steps and used the last 10,000 samples to compute the posterior means for z_t, σ_m, σ_r, and ζ for downstream analysis.

We fit Robust Linear Regression using the statsmodels python package with a Huber's T M-estimator [SP10]. We used the R package 'pwr' [Cha16] to perform power calculations. We estimated the number of samples necessary for 80% power with Ordinary Least Squares given the instrumentation noise by calculating a projected R^2

$$R^2_{\text{projected}} = \frac{R^2_{\text{original}} \sigma_m^2}{\sigma_m^2 + \sigma_r^2}$$

and using a significance level of 0.05/200 to account for multiple hypothesis correction.

4.5.3 Microbiome

We used Human Longevity Inc's Microbiome service to profile the fecal microbiome samples. Their proprietary software estimates abundance of 3785 species, and aggregates information to get relative abundances of genera, orders, families, phyla, and kingdoms. They further map the sequences to orthogonal sequences in the TigrFam database to quantify relative function of different molecular processes across the bacterial species.

4.5.4 Transcriptome

We sent whole blood samples to The Scripps Research Institute's Sequencing Center for RNA Sequencing. We used bcbio-nextgen v. 0.9 to align and generate counts for paired end reads. The packaged pipeline uses cutadapt for adapter trimming [Mar11], aligns the reads using TopHat2[KPT+13], and generates gene-level counts with HTSeq [APH15].

We normalized the samples using DESeq2 [LHA14] using all samples included not globin-cleared for illustrative purposes in 4.5a, but all subsequent analysis was done using normalization performed on all samples that were globin-cleared only. We identified outliers and linear trends in the data using DESeq2 provided Cook's distance and its likelihood-ratio test comparing a generalized linear model using a negative binomial link with days as a predictor to a reduced model of only variance.

4.6 Figures and Tables

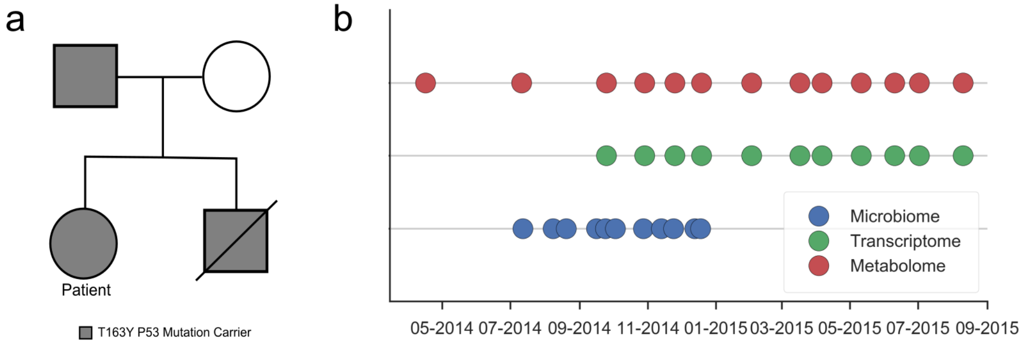


Figure 4.1: Pedigree of nuclear family and overview of sample collection for patient a) The father is an affected, but healthy heterozygous germline T163Y p53 mutation carrier and the mother is not affected. The offspring are carriers of the mutation. The son developed a glioblastoma at 12 years of age and is deceased. The daughter is a healthy 16 year old with no signs of cancer. b) We profiled the metabolome, microbiome, and transcriptome of the daughter on 13, 11, and 11 occasions, respectively over the course of 16 months.

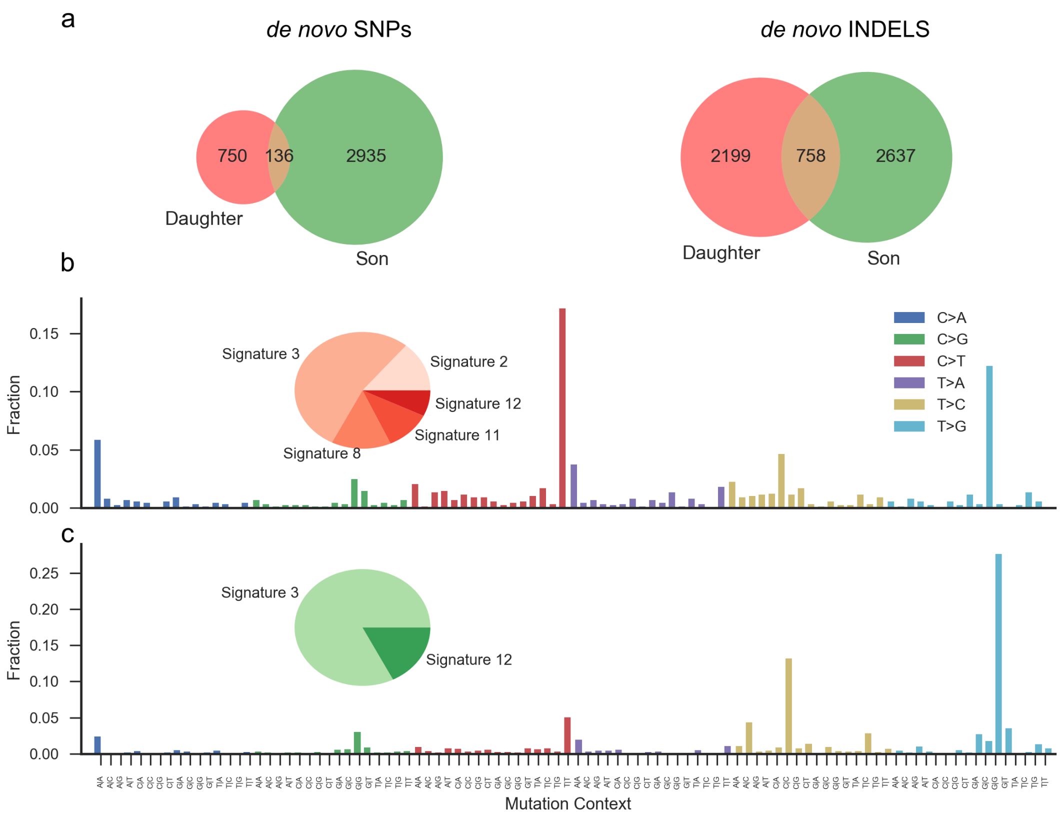


Figure 4.2: De-novo mutation context analysis in the offspring a) We identified 886 and 3071 de novo SNVs and 2957 and 3395 de novo INDELs in the daughter and son, respectively. Of these only 136 SNVs and 758 INDELs were shared between the offspring. b, c) Each bar represents a different mutational context. The contexts are lexical sorted, i.e. A, C→A, A $\cdots$ A, C→A, T $\cdots$ T, C→, A $\cdots$ T, C→, T. We characterized these mutations by deconvolving their mutational context signatures using deconstructSigs and comparing them to COSMIC curated mutational signatures. Signature 3 and 12 explain a high fraction of the variation of the mutational signatures in both offspring. Signature 3 has been previously associated with homologous DNA repair, whereas Signature 12 has no known etiology.

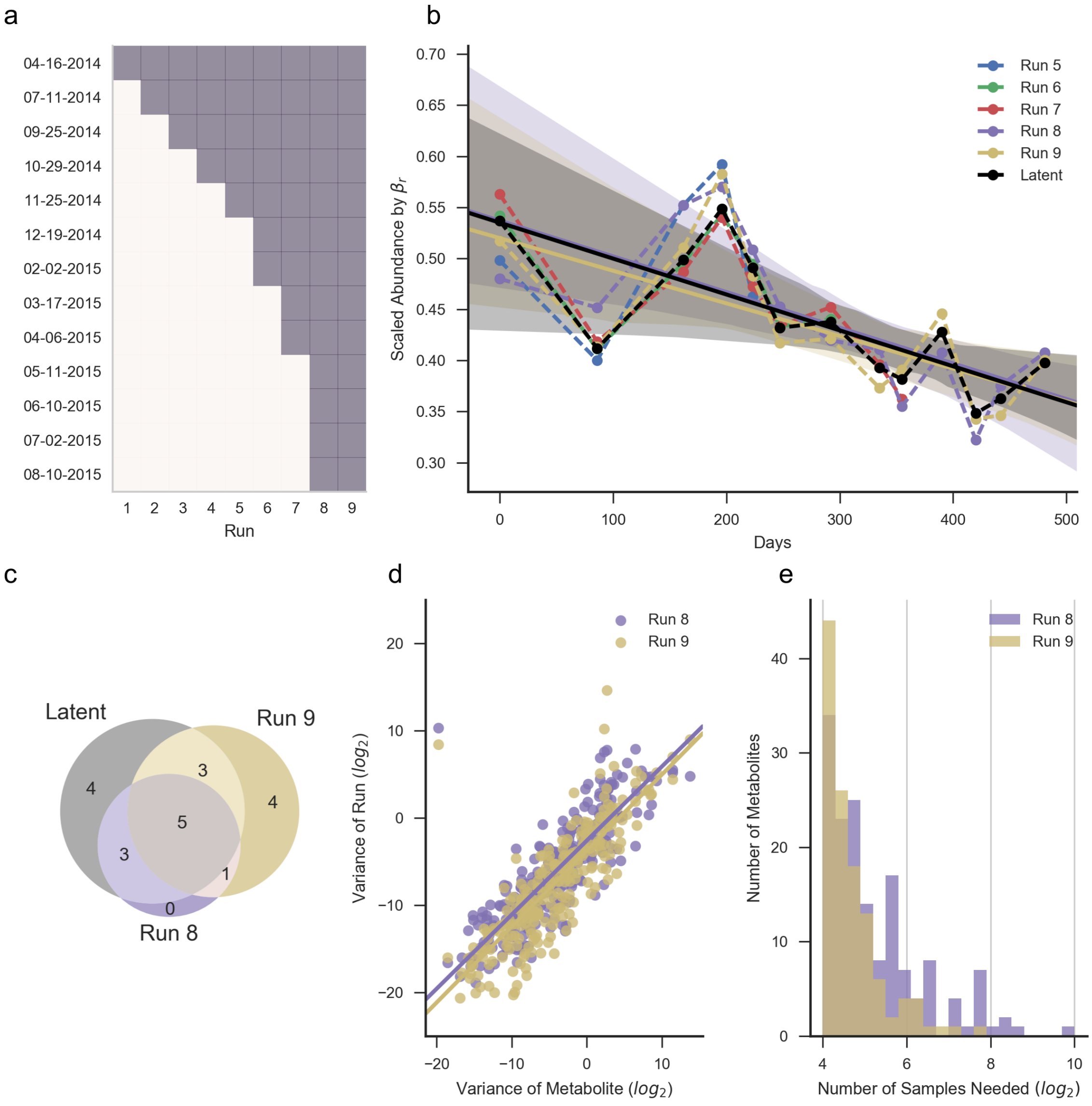


Figure 4.3: Metabolome profiling reveals metabolite-dependent instrumentation noise a) Due to known scaling issues across instrumentation runs, a profiling run for a new sample also reprofiled all previous samples. Runs 8 and 9 were performed using all samples. b) The latent and scaled abundance values of Glycocholenate Sulfate. Run 8 and 9 have higher variation than the latent model, which leads to non-significant linear associations. c) Over all metabolites, the latent model recovers the most associations using robust linear regression. d) Run 8, on average, added more variability to metabolite levels than Run 9. e) Number of samples necessary to achieve 80% power for observing a significant association with a latent correlation of 0.9 due to added instrumentation noise.

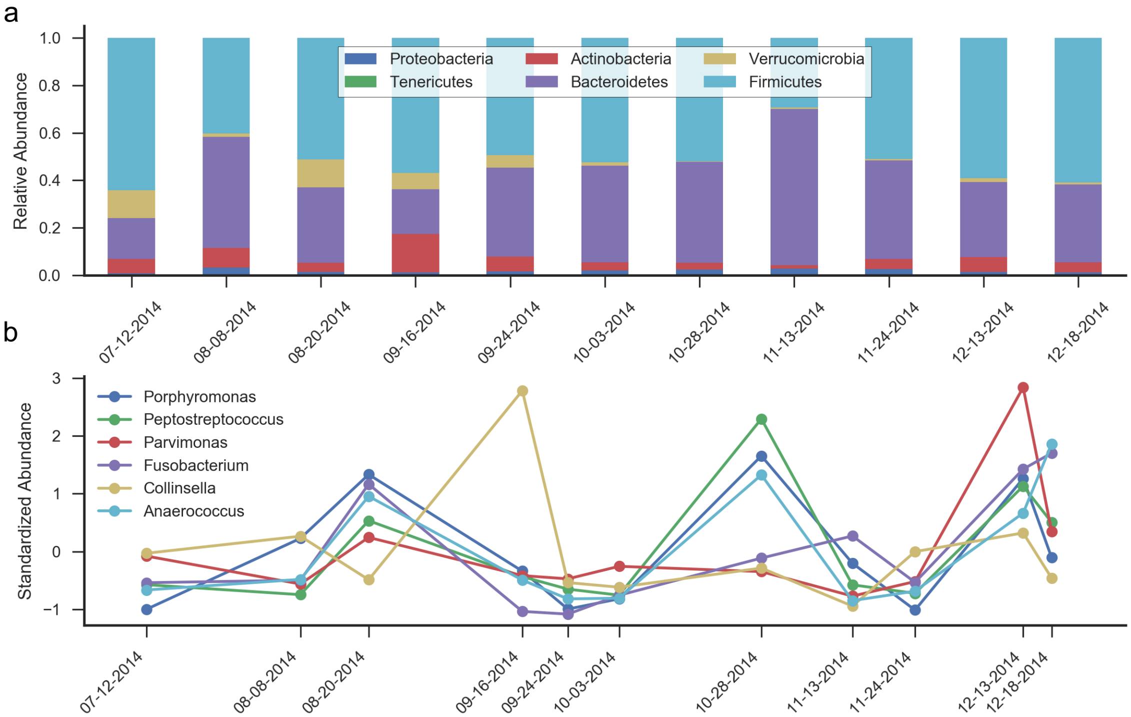


Figure 4.4: High variation in Microbiome profiling a) There was high variation in relative abundances of the 6 phyla across the samples, but no obvious trends. b) Although we did observe deviations from baseline in six colorectral cancer associated genera, there is high variability overall.

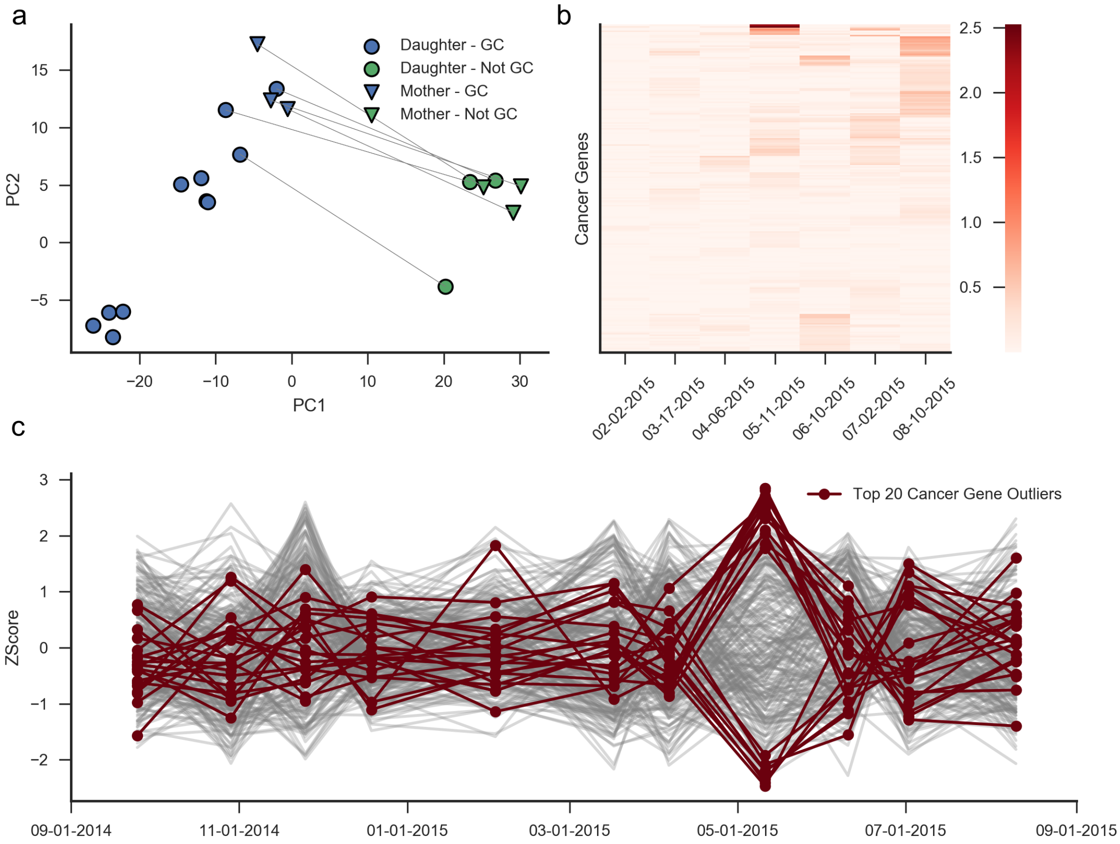
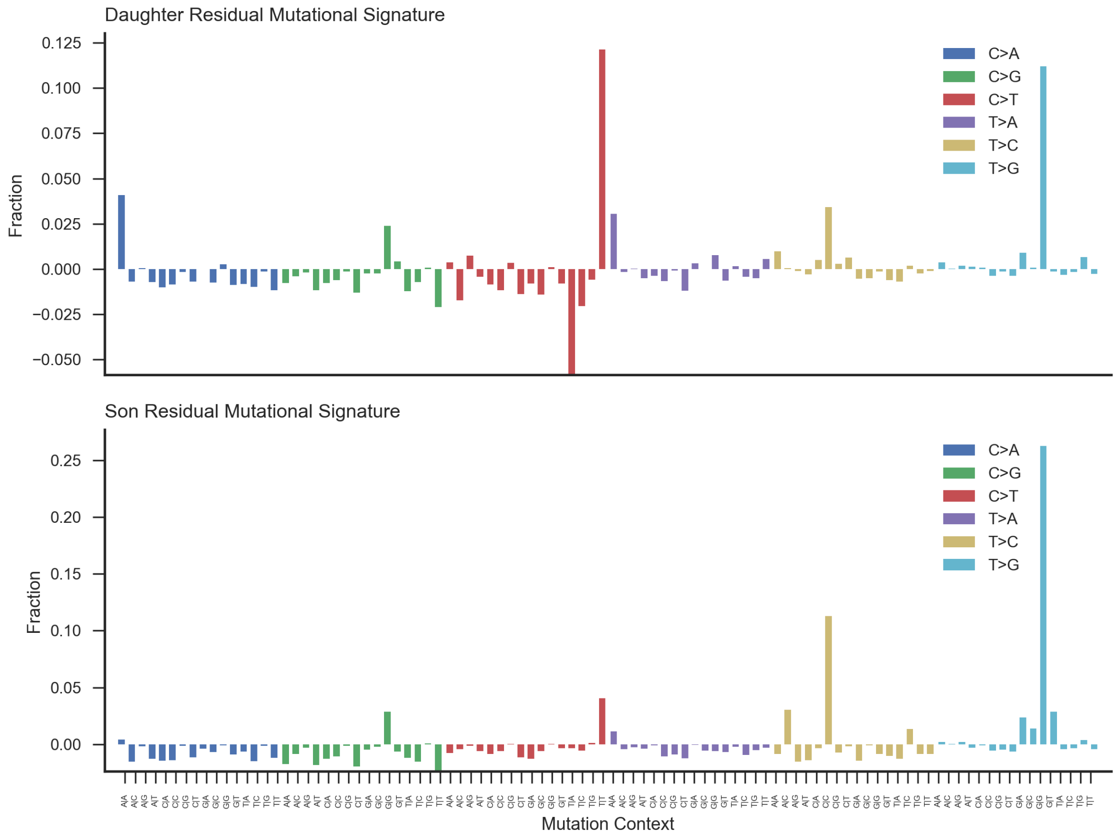


Figure 4.5: Outliers detected by transcriptome profiling a) Library preparation using a globin-clearing (GC) step was a source of large variation in initial sequencing of RNA. All samples were retrospectively profiled using the same RNA-Seq protocol. Grey lines connect technical replicates. b) Cook's distance for a curated list of cancer genes for the last seven samples. 05-11-2015 had several large outliers. c) Top 20 gene outliers from 05-11-2015 all returned to baseline expression at later time points.

Table 4.1: Low instrumentation variance metabolites.65 metabolites that require less than 20 samples to detect a correlation of 0.9 between days and metabolite levels with 80% power.

Type	Metabolites
Amino Acids	3-(4-hydroxyphenyl)lactate, 3-methylhistidine, 3-phenylpropionate (hydrocinnamate), creatine, ethylmalonate, indolepropionate, isobutyrylcarnitine, N-methyl proline, p-cresol sulfate, serotonin (5HT), S-methylcysteine, taurine, tryptophan betaine
Carbohydrates	glucuronate, lactate
Cofactors and Vitamins	gamma-CEHC, pyridoxate
Lipids	10-heptadecenoate (17:1n7), 10-nonadecenoate (19:1n9), 10-undecenoate (11:1n1), 1-linoleoylglycerophosphoethanolamine*, 3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF), 3-hydroxyoctanoate, 4-androsten-3beta,17beta-diol monosulfate (1), 5alpha-pregnan-3beta,20alpha-diol disulfate, 5alpha-pregnan-3beta,20alpha-diol monsulfate (2), 5-dodecenoate (12:1n7), androsterone sulfate, cholate, choline phosphate, cis-4-decenoyl carnitine, decanoylcarnitine, dihomo-linoleate (20:2n6), eicosanodioate, eicosenoate (20:1n9 or 11), epiandrosterone sulfate, glycochenodeoxycholate, glycocholate, glycolithocholate sulfate*, glycoursodeoxycholate, laurate (12:0), laurylcarnitine, margarate (17:0), myristate (14:0), myristoleate (14:1n5), myristoleoylcarnitine*, octanoylcarnitine, palmitate (16:0), pregn steroid monosulfate*, taurochenodeoxycholate, taurocholenate sulfate, taurodeoxycholate, taurolithocholate 3-sulfate
Xenobiotics	2-aminophenol sulfate, 2-hydroxyhippurate (salicylurate), 4-methylcatechol sulfate, 4-vinylphenol sulfate, catechol sulfate, cinnamoylglycine, hippurate, methyl glucopyranoside (alpha + beta), O-methylcatechol sulfate, stachydrine, theobromine, thymol sulfate


Daughter Residual Mutational Signature

Son Residual Mutational Signature

Figure S4.1: Residual mutational signatures of de-novo variants in offspring. Residuals mutational signatures after subtracting the effects of Signature 2, 3, 8, 11, and 12 and Signatures 3 and 11 for the daughter and son, respectively. Each bar represents a different mutational context. The contexts are lexical sorted, i.e. A, C→A, A ⋯ A, C→A, T ⋯ T, C→, A ⋯ T, C→, T

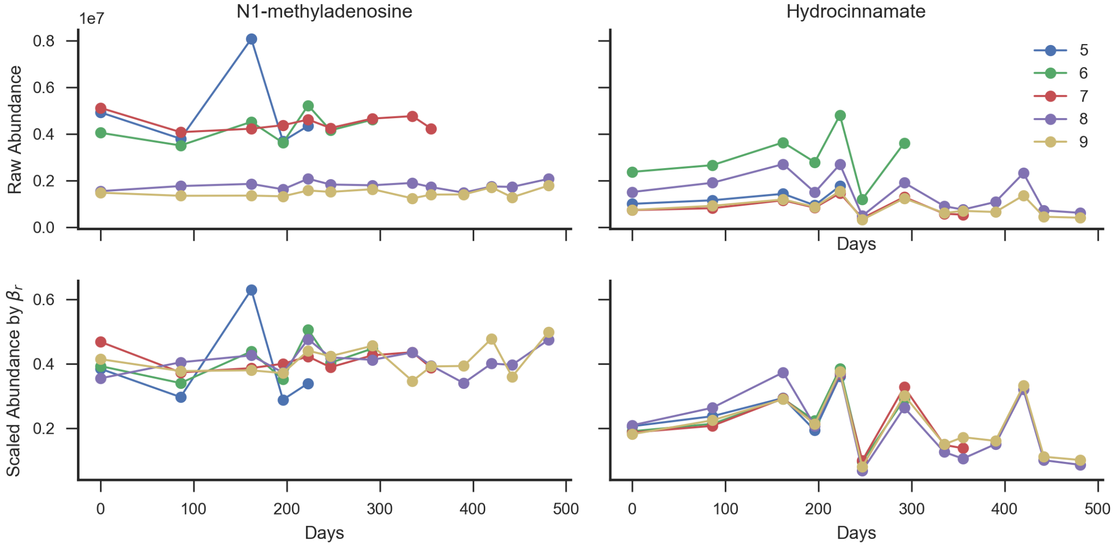


Figure S4.2: Metabolomic profiling of two known cancer-associated metabolites: N1-methyladenosine and Hydrocinnamate [WJG+13]. The top panel is the reported raw abundance values for the metabolites for each instrumentation run. The bottom panels are the scaled metabolite levels based on the Bayesian latent model fit.

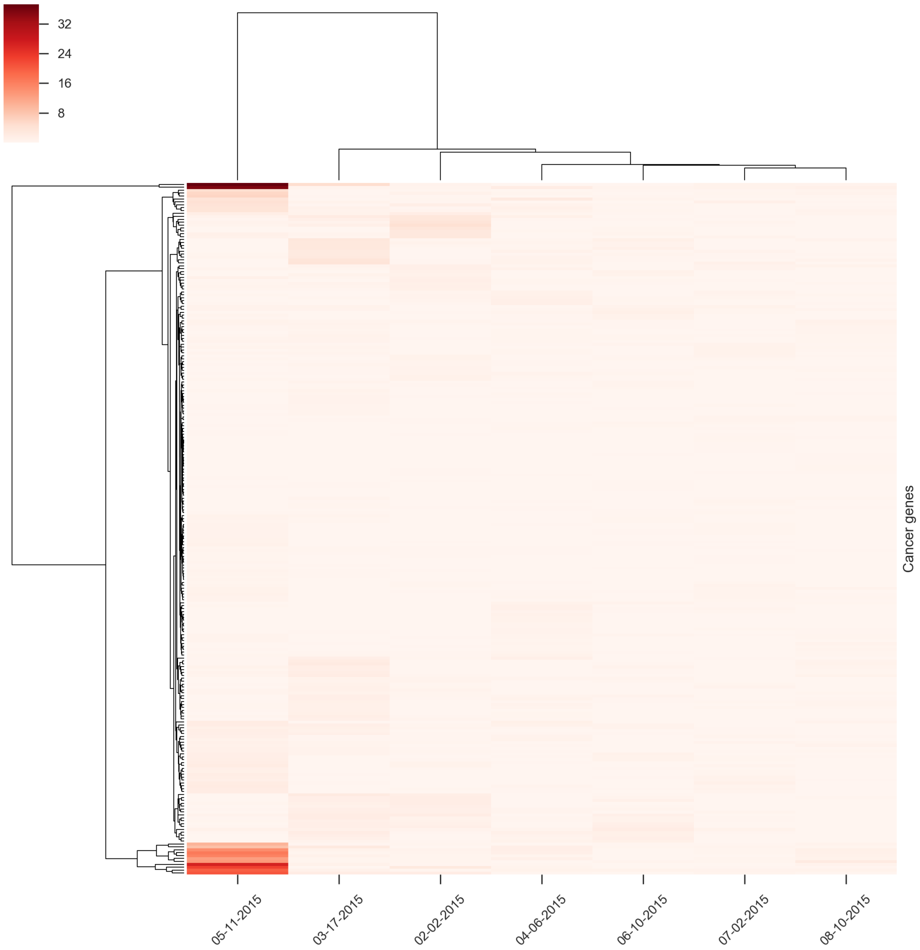


Figure S4.3: Outliers in simulated real-time monitoring of oncogene transcriptome. Cook's distance and hierarchical clustering of samples based on expression levels in 222 oncogenes

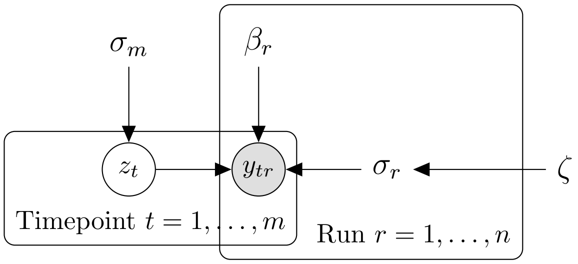


Figure S4.4: Graphical model for latent metabolite levels

4.7 Acknowledgements

Chapter 4, in part is currently being prepared for submission for publication of the material. Kunal Bhutani, Victoria Magnuson, Alexandra Buckley, Danjuma Quarless, Laura Goetz, Nicholas J Schork. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Closing Remarks

As the biomedical sciences continue to generate data describing biological phenomena at astronomical scales, robust analysis and interpretation of data will require an intimate understanding of biases inherent to biological experiments and assays. In this dissertation, I have illustrated one widely-applicable strategy for separating biological signal from other sources of variation: pooling information. Through general analysis pipelines and several hierarchical Bayesian models, I have shown that it is possible to pool information to explicitly model variation that is inherent to the experimental setup, and differentiate it from the biological signal of interest.

In Chapter 2, I discuss pooling information across iPSC colonies to better characterize the single nucleotide variants in the initial fibroblast populations. This analysis pipeline allowed us to find variants that were specifically caused by the reprogramming step of iPSC creation. In another application, in Chapter 3, I focus on pooling information across bootstrapped models of transcriptional regulation to estimate uncertainty and then propagate this uncertainty in a second stage hierarchical

Bayesian regression model for associating imputed gene expression to a continuous phenotype. Finally, in Chapter 4, I discuss pooling information across multiple metab olomic instrumentation runs to separate instrumentation variation from biological signal in a latent Bayesian model. I also show that given the inherent noise in the assay, it is not possible to find highly correlated relationships between metabolite levels and time for several metabolites using regular blood draws.

Although I discuss only post-data generation application of pooling information, the technique could be incorporated as part of the initial study design before any experimental work is performed. Currently, many experiments designs utilize replicates to confirm biological results and prevent false positives. The underlying assumption in this design is that noise in the experiment is random and more replicates will yield better estimates of the true biological signal. Scientists could similarly design studies to utilize pooling information across replicates or other shared characteristics to increase interpretability and understanding of functional forms of the variation inherent to the experimental setup or assay. I mainly highlight Bayesian formulations of pooling information in application-dependent ways, and I foresee similar efforts of creating mathematical models for pooling that are application specific and unique to the system of interest. I hope that this dissertation has served as an introduction to the power of pooling information to limit biases in biological data analysis, and believe the importance and application of pooling information will continue to grow alongside the data revolution in the biomedical sciences.

Bibliography

[AAB+02] Tariq Ahmad, Alessandro Armuzzi, Mike Bunce, Kim MulcahyHawes, Sara E. Marshall, Timothy R. Orchard, Jonathan Crawshaw, Oliver Large, Arjuna De Silva, Julia T. Cook, Martin Barnardo, Sue Cullen, Ken I. Welsh, and Derek P. Jewell. The molecular classification of the clinical manifestations of Crohn's disease. *Gastroenterology*, 122(4):854–866, 2002.

[ACF+16] Mariet Allen, Minerva M. Carrasquillo, Cory Funk, Benjamin D. Heavner, Fanggeng Zou, Curtis S. Younkin, Jeremy D. Burgess, High-Seng Chai, Julia Crook, James A. Eddy, Hongdong Li, Ben Logsdon, Mette A. Peters, Kristen K. Dang, Xue Wang, Daniel Serie, Chen Wang, Thuy Nguyen, Sarah Lincoln, Kimberly Malphrus, Gina Bisceglio, Ma Li, Todd E. Golde, Lara M. Mangravite, Yan Asmann, Nathan D. Price, Ronald C. Petersen, Neill R. Graff-Radford, Dennis W. Dickson, Steven G. Younkin, Nilfer Ertekin-Taner, M. M. Carrasquillo, D. Harold, J. C. Lambert, S. Seshadri, A. C. Naj, P. Hollingworth, J. C. Lambert, G. U. Hoglinger, J. Simon-Sanchez, F. Zou, A. L. Dixon, V. Emilsson, A. J. Saykin, F. Zou, M. Allen, M. Allen, M. Allen, M Allen, A. J. Myers, J. A. Webster, J. Chapuis, L. N. Hazrati, A. Ramasamy, S. B. Montgomery, J. M. Derry, M. Allen, G McKhann, LA Farrer, H. Braak, E. Braak, J. J. Hauw, S. S. Mirra, J. Wang, D. W. Dickson, J. Q. Trojanowski, V. M. Lee, P. Du, W. A. Kibbe, S. M. Lin, S. Purcell, A. T. Magis, C. C. Funk, N. D. Price, and C Funk. Human whole genome genotype and transcriptome data for Alzheimers and other neurodegenerative diseases. *Scientific Data*, 3:160089, 10 2016.

[ADS+15] K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M.

Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 5 2015.

[Ana01] T Anantharaman. Probabilistic Analysis of False Positives in Optical Map Alignment and Validation. *WABI2001*, 2001.

[APH15] S. Anders, P. T. Pyl, and W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 1 2015.

[AS13] Elif F. Acar and Lei Sun. A Generalized Kruskal-Wallis Test Incorporating Group Uncertainty with Application to Genetic Association Studies. *Biometrics*, 69(2):427–435, 6 2013.

[AS14] Ludmil B Alexandrov and Michael R Stratton. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development*, 24:52–60, 2 2014.

[BCC$^+$07] Paul R. Burton, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, Nilesh J. Samani, John A. Todd, Peter Donnelly, Jeffrey C. Barrett, Paul R. Burton, Dan Davison, Peter Donnelly, Doug Easton, David Evans, Hin-Tak Leung, Jonathan L. Marchini, Andrew P. Morris, Chris C. A. Spencer, Martin D. Tobin, Lon R. Cardon, David G. Clayton, Antony P. Attwood, James P. Boorman, Barbara Cant, Ursula Everson, Judith M. Hussey, Jennifer D. Jolley, Alexandra S.

Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V. Prowse, Helen E. Stevens, Niall C. Taylor, Graham R. Walters, Neil M. Walker, Nicholas A. Watkins, Thilo Winzer, John A. Todd, Willem H. Ouwehand, Richard W. Jones, Wendy L. McArdle, Susan M. Ring, David P. Strachan, Marcus Pembrey, Gerome Breen, David St Clair, Sian Caesar, Katherine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K. Green, Detelina Grozeva, Marian L. Hamshere, Peter A. Holmans, Ian R. Jones, George Kirov, Valentina Moskvina, Ivan Nikolov, Michael C. O'Donovan, Michael J. Owen, Nick Craddock, David A. Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H. Young, I. Nicol Ferrier, Stephen G. Ball, Anthony J. Balmforth, Jennifer H. Barrett, D. Timothy Bishop, Mark M. Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S. Hall, Peter S. Braund, Paul R. Burton, Richard J. Dixon, Massimo Mangino, Suzanne Stevens, Martin D. Tobin, John R. Thompson, Nilesh J. Samani, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W. Lees, Elaine R. Nimmo, Jack Satsangi, Sheila A. Fisher, Alastair Forbes, Cathryn M. Lewis, Clive M. Onnie, Natalie J. Prescott, Jeremy Sanderson, Christopher G. Mathew, Jamie Barbour, M. Khalid Mohiuddin, Catherine E. Todhunter, John C. Mansfield, Tariq Ahmad, Fraser R. Cummings, Derek P. Jewell, John Webster, Morris J. Brown, David G. Clayton, G. Mark Lathrop, John Connell, Anna Dominiczak, Nilesh J. Samani, Carolina A. Braga Marcano, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L. Lee, Patricia B. Munroe, Stephen J. Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, , The Biologics in RA Genetics Genomics (BRAGGS), Ian N. Bruce, Hannah Donovan, Steve Eyre, Paul D. Gilbert, Samantha L. Hider, Anne M. Hinks, Sally L. John, Catherine Potter, Alan J. Silman, Deborah P. M. Symmons, Wendy Thomson, Jane Worthington, David G. Clayton, David B. Dunger, Sarah Nutland, Helen E. Stevens, Neil M. Walker, Barry Widmer, John A. Todd, Timothy M. Frayling, Rachel M. Freathy, Hana Lango, John R. B. Perry, Beverley M. Shields, Michael N. Weedon, Andrew T. Hattersley, Graham A. Hitman, Mark Walker, Kate S. Elliott, Christopher J. Groves, Cecilia M. Lindgren, Nigel W. Rayner, Nicholas J. Timpson, Eleftheria Zeggini, Mark I. McCarthy, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V. S. Hill, Linda A. Bradbury, Claire Farrar, Jennifer J. Pointon, Paul Wordsworth, Matthew A. Brown, Jayne A. Franklyn, Joanne M. Heward, Matthew J. Simmonds, Stephen C. L. Gough, Sheila Seal, Breast Cancer Susceptibility Collaboration (UK), Michael R. Stratton, Nazneen Rahman, Maria Ban, An Goris, Stephen J. Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Melanie Newport, Giorgio Sirugo, Kirk A. Rockett, Dominic P. Kwiatkowski, Suzannah J. Bumpstead, Amy Chaney, Kate Downes, Mohammed J. R. Ghori, Rhian Gwilliam, Sarah E. Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker, Claire Widden, David Withers, Panos Deloukas, Hin-Tak Leung, Sarah Nutland, Helen E. Stevens, Neil M. Walker, John A. Todd, Doug Easton, David G. Clayton, Paul R. Burton, Martin D. Tobin, Jeffrey C. Barrett, David Evans, Andrew P. Morris, Lon R. Cardon, Niall J. Cardin, Dan Davison, Teresa Ferreira, Joanne Pereira-Gale, Ingileif B. Hallgrimsdóttir, Bryan N. Howie, Jonathan L. Marchini, Chris C. A. Spencer, Zhan Su, Yik Ying Teo, Damjan Vukcevic, Peter Donnelly, David Bentley, Matthew A. Brown, Lon R. Cardon, Mark Caulfield, David G. Clayton, Alistair Compston, Nick Craddock, Panos Deloukas, Peter Donnelly, Martin Farrall, Stephen C. L. Gough, Alistair S. Hall, Andrew T. Hattersley, Adrian V. S. Hill, Dominic P. Kwiatkowski, Christopher G. Mathew, Mark I. McCarthy, Willem H. Ouwehand, Miles Parkes, Marcus Pembrey, Nazneen Rahman, Nilesh J. Samani, Michael R. Stratton, John A. Todd, and Jane Worthington. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 6 2007.

[BCC$^+$09] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Ccile Julier, Grant Morahan, Jrn Nerup, Concepcion Nierras, Vincent Plagnol, Flemming Pociot, Helen Schuilenburg, Deborah J Smyth, Helen Stevens, John A Todd, Neil M Walker, Stephen S Rich, and Type 1 Diabetes Genetics Consortium. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, 41(6):703–707, 6 2009.

[BCM$^+$04] Todd M Bull, Christopher D Coldren, Mark Moore, Sylk M Sotto-Santiago, David V Pham, S Patrick Nana-Sinkam, Norbert F Voelkel, and Mark W Geraci. Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. *American journal of respiratory and critical care medicine*, 170(8):911–9, 10 2004.

[BFT$^+$12] Philip Beineke, Karen Fitch, Heng Tao, Michael R Elashoff, Steven Rosenberg, William E Kraus, James A Wingrove, and PREDICT Investigators. A whole blood gene expression-based signature for smoking status. *BMC medical genomics*, 5:58, 12 2012.

[BKA$^+$12] Barna Budai, Viktor Komlósi, Vilmos Adleff, va Pap, Andrea Réti, Tnde Nagy, Judit Kralovánszky, Istvn Láng, and Erika Hitre. Impact of SHMT1 polymorphism on the clinical outcome of patients with metastatic colorectal cancer treated with first-line

FOLFIRI+bevacizumab. *Pharmacogenetics and Genomics*, 22(1):69–72, 1 2012.

[BNW$^+$16] Kunal Bhutani, Kristopher L. Nazor, Roy Williams, Ha Tran, Heng Dai, eljko Džakula, Edward H. Cho, Andy W. C. Pang, Mahendra Rao, Han Cao, Nicholas J. Schork, Jeanne F. Loring, S. M. Hussein, L. C. Laurent, Y. Mayshar, S. M. Taapken, A. Gore, S. E. Peterson, J. F. Loring, T. M. Schlaeger, A. McKenna, K. Cibulskis, S. K. Das, A. R. Hastie, E. T. Lam, M. Xiao, A. Valouev, D. C. Schwartz, S. Zhou, M. S. Waterman, G. A. Erikson, N. Deshpande, B. G. Kesavan, A. Torkamani, A. H. Ramos, C. Kandoth, M. Kircher, C. Baum, K. Takahashi, S. Yamanaka, K. Okita, T. Ichisaka, S. Yamanaka, J. Korhonen, P. Martinmaki, C. Pizzi, P. Rastas, E. Ukkonen, H. Li, H. Li, and R. Durbin. Whole-genome mutational burden analysis of three pluripotency induction methods. *Nature Communications*, 7:10536, 2 2016.

[BRRS16] Nielson T. Baxter, Mack T. Ruffin, Mary A. M. Rogers, and Patrick D. Schloss. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, 8(1):37, 12 2016.

[BvKS$^+$04] Christopher Baum, Christof von Kalle, Frank J T Staal, Zhixiong Li, Boris Fehse, Manfred Schmidt, Floor Weerkamp, Stefan Karlsson, Gerard Wagemaker, and David A Williams. Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. *Molecular therapy : the journal of the American Society of Gene Therapy*, 9(1):5–13, 1 2004.

[CEX$^+$13] C. V. A. Collares, A. F. Evangelista, D. J. Xavier, P. Takahashi, R. Almeida, C. Macedo, F. Manoel-Caetano, M. C. Foss, M. C. Foss-Freitas, D. M. Rassi, E. T. Sakamoto-Hojo, G. A. Passos, and E. A. Donadi. Transcriptome meta-analysis of peripheral lymphomononuclear cells indicates that gestational diabetes is closer to type 1 diabetes than to type 2 diabetes mellitus. *Molecular Biology Reports*, 40(9):5351–5358, 9 2013.

[CGM$^+$09] Pietro Cirillo, Michael S Gersch, Wei Mu, Philip M Scherer, Kyung Mee Kim, Loreto Gesualdo, George N Henderson, Richard J Johnson, and Yuri Y Sautin. Ketohexokinase-dependent metabolism of fructose induces proinflammatory mediators in proximal tubular cells. *Journal of the American Society of Nephrology : JASN*, 20(3):545–53, 3 2009.

[Cha16] Stephane Champely. pwr: Basic Functions for Power Analysis, 2016.

[CLC$^+$13] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew

Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–9, 3 2013.

[CMLPT$^+$12] Rui Chen, George I Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo Y K Lam, Rong Chen, Elana Miriami, Konrad J Karczewski, Manoj Hariharan, Frederick E Dewey, Yong Cheng, Michael J Clark, Hogune Im, Lukas Habegger, Suganthi Balasubramanian, Maeve O'Huallachain, Joel T Dudley, Sara Hillenmeyer, Rajini Haraksingh, Donald Sharon, Ghia Euskirchen, Phil Lacroute, Keith Bettinger, Alan P Boyle, Maya Kasowski, Fabian Grubert, Scott Seki, Marco Garcia, Michelle Whirl-Carrillo, Mercedes Gallardo, Maria A Blasco, Peter L Greenberg, Phyllis Snyder, Teri E Klein, Russ B Altman, Atul J Butte, Euan A Ashley, Mark Gerstein, Kari C Nadeau, Hua Tang, and Michael Snyder. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–307, 3 2012.

[CSS$^+$08] Jason D Cooper, Deborah J Smyth, Adam M Smiles, Vincent Plagnol, Neil M Walker, James E Allen, Kate Downes, Jeffrey C Barrett, Barry C Healy, Josyf C Mychaleckyj, James H Warram, and John A Todd. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics*, 40(12):1399–1401, 12 2008.

[DAA$^+$10] Somes K. Das, Michael D. Austin, Matthew C. Akana, Paru Deshpande, Han Cao, and Ming Xiao. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Research*, 38(18), 2010.

[DM07] S. Davis and P. S. Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14):1846–1847, 7 2007.

[DMC$^+$13] Lawrence A. David, Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, A. Sloan Devlin, Yug Varma, Michael A. Fischbach, Sudha B. Biddinger, Rachel J. Dutton, and Peter J. Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, 12 2013.

[FBG$^+$15a] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811, 1 2015.

[FBG$^+$15b] Simon A Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W Teague, Michael R Stratton, Ultan McDermott, and Peter J Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(Database issue):805–11, 1 2015.

[FTA$^+$08] Sheila A Fisher, Mark Tremelling, Carl A Anderson, Rhian Gwilliam, Suzannah Bumpstead, Natalie J Prescott, Elaine R Nimmo, Dunecan Massey, Carlo Berzuini, Christopher Johnson, Jeffrey C Barrett, Fraser R Cummings, Hazel Drummond, Charlie W Lees, Clive M Onnie, Catherine E Hanson, Katarzyna Blaszczyk, Mike Inouye, Philip Ewels, Radhi Ravindrarajah, Andrew Keniry, Sarah Hunt, Martyn Carter, Nick Watkins, Willem Ouwehand, Cathryn M Lewis, Lon Cardon, Alan Lobo, Alastair Forbes, Jeremy Sanderson, Derek P Jewell, John C Mansfield, Panos Deloukas, Christopher G Mathew, Miles Parkes, and Jack Satsangi. Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. *Nature Genetics*, 40(6):710–712, 6 2008.

[Ful87] Wayne A. Fuller. *Measurement Error Models*. John Wiley & Sons, Inc., 1987.

[Geo11] Philomena George. P53 HOW CRUCIAL IS ITS ROLE IN CANCER ? 3(2), 2011.

[GKS$^+$16] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Pivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2 2016.

[GLF$^+$11] Athurva Gore, Zhe Li, Ho-Lim Fung, Jessica E Young, Suneet Agarwal, Jessica Antosiewicz-Bourget, Isabel Canto, Alessandra Giorgetti, Mason a Israel, Evangelos Kiskinis, Je-Hyuk Lee, Yuin-Han Loh, Philip D Manos, Nuria Montserrat, Athanasia D Panopoulos, Sergio Ruiz, Melissa L Wilbert, Junying Yu, Ewen F Kirkness, Juan Carlos Izpisua Belmonte, Derrick J Rossi, James a Thomson, Kevin Eggan, George Q Daley, Lawrence S B Goldstein, and Kun Zhang. Somatic coding mutations in human induced pluripotent stem cells. *Nature*, 471(7336):63–7, 3 2011.

[GMF+16] Alexander Gusev, Nick Mancuso, Hilary K Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Edwin Oh, Steven McCaroll, Benjamin Neale, Roel Ophoff, Michael C O'Donovan, Nicholas Katsanis, Gregory E Crawford, Patrick F Sullivan, Bogdan Pasaniuc, and Alkes L Price. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *bioRxiv*, 2016.

[GWS+15] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 8 2015.

[HBV+11] Samer M Hussein, Nizar N Batada, Sanna Vuoristo, Reagan W Ching, Reija Autio, Elisa Närvä, Siemon Ng, Michel Sourour, Riikka Hämäläinen, Cia Olsson, Karolina Lundin, Milla Mikkola, Ras Trokovic, Michael Peitz, Oliver Brüstle, David P Bazett-Jones, Kari Alitalo, Riitta Lahesmaa, Andras Nagy, and Timo Otonkoski. Copy number variation and selection during reprogramming to pluripotency. *Nature*, 471(7336):58–62, 2011.

[HDM+09] Bryan N. Howie, Peter Donnelly, Jonathan Marchini, JD Rioux, RJ Xavier, KD Taylor, MS Silverberg, P Goyette, J Gudmundsson, P Sulem, A Manolescu, LT Amundadottir, D Gudbjartsson, KA Frazer, DG Ballinger, DR Cox, DA Hinds, LL Stuve, B Servin, M Stephens, J Marchini, B Howie, S Myers, G McVean, P Donnelly, DY Lin, Y Hu, BE Huang, DL Nicolae, E Zeggini, LJ Scott, R Saxena, BF Voight, JL Marchini, JC Barrett, S Hansoul, DL Nicolae, JH Cho, RH Duerr, P Scheet, M Stephens, SR Browning, BL Browning, BL Browning, SR Browning, JC Barrett, DG Clayton, P Concannon, B Akolkar, JD Cooper, S Purcell, B Neale, K Todd-Brown, L Thomas, MA Ferreira, N Li, M Stephens, S Myers, L Bottolo, C Freeman, G McVean, P Donnelly, J Wakeley, A Kong, G Masson, ML Frigge, A Gylfason, P Zusmanovich, Y Guan, M Stephens, J Marchini, D Cutler, N Patterson, M Stephens, E Eskin, W Chen, Y Li, G Abecasis, Z Zhao, N Timofeev, SW Hartley, DH Chui, S Fucharoen, L Huang, Y Li, AB Singleton, JA Hardy, and G Abecasis. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6):e1000529, 6 2009.

[HDS+13] Alex R. Hastie, Lingli Dong, Alexis Smith, Jeff Finklestein, Ernest T. Lam, Naxin Huo, Han Cao, Pui Yan Kwok, Karin R. Deal, Jan Dvorak, Ming Cheng Luo, Yong Gu, and Ming Xiao. Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo

Sequence Assembly of the Complex Aegilops tauschii Genome. *PLoS ONE*, 8(2), 2013.

[HFS+12] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955–959, 7 2012.

[HKHH02] M Hollstein, MA Khan, C.C Harris, and P Hainaut. The IARC TP53 Database: new online mutation analysis and recommendations to users. *Human Mutation*, 19(6):607–614, 2002.

[HSJ+09] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 6 2009.

[HSW03] Daniel H Haft, Jeremy D Selengut, and Owen White. The TIGRFAMs database of protein families. *Nucleic acids research*, 31(1):371–3, 1 2003.

[HvBG+10] Alexander Hoischen, Bregje W M van Bon, Christian Gilissen, Peer Arts, Bart van Lier, Marloes Steehouwer, Petra de Vries, Rick de Reuver, Nienke Wieskamp, Geert Mortier, Koen Devriendt, Marta Z Amorim, Nicole Revencu, Alexa Kidd, Mafalda Barbosa, Anne Turner, Janine Smith, Christina Oley, Alex Henderson, Ian M Hayes, Elizabeth M Thompson, Han G Brunner, Bert B A de Vries, and Joris A Veltman. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature Genetics*, 42(6):483–485, 6 2010.

[IKB+04] K Iwamoto, C Kakiuchi, M Bundo, K Ikeda, and T Kato. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Molecular psychiatry*, 9(4):406–16, 4 2004.

[ILZR10] Sandrine Imbeaud, Yannick Ladeiro, and Jessica Zucman-Rossi. Identification of Novel Oncogenes and Tumor Suppressors in Hepatocellular Carcinoma. *Seminars in Liver Disease*, 30(01):075–086, 2 2010.

[ITY+16] Minako Imamura, Atsushi Takahashi, Toshimasa Yamauchi, Kazuo Hara, Kazuki Yasuda, Niels Grarup, Wei Zhao, Xu Wang, Alicia Huerta-Chagoya, Cheng Hu, Sanghoon Moon, Jirong Long, Soo Heon Kwak, Asif Rasheed, Richa Saxena, Ronald C. W. Ma, Yukinori Okada, Minoru Iwata, Jun Hosoe, Nobuhiro Shojima, Minaka Iwasaki, Hayato Fujita, Ken Suzuki, John Danesh, Torben Jørgensen, Marit E.

Jørgensen, Daniel R. Witte, Ivan Brandslund, Cramer Christensen, Torben Hansen, Josep M. Mercader, Jason Flannick, Hortensia Moreno-Macías, Nol P. Burtt, Rong Zhang, Young Jin Kim, Wei Zheng, Jai Rup Singh, Claudia H. T. Tam, Hiroshi Hirose, Hiroshi Maegawa, Chikako Ito, Kohei Kaku, Hirotaka Watada, Yasushi Tanaka, Kazuyuki Tobe, Ryuzo Kawamori, Michiaki Kubo, Yoon Shin Cho, Juliana C. N. Chan, Dharambir Sanghera, Philippe Frossard, Kyong Soo Park, Xiao-Ou Shu, Bong-Jo Kim, Jose C. Florez, Teresa Tusié-Luna, Weiping Jia, E Shyong Tai, Oluf Pedersen, Danish Saleheen, Shiro Maeda, and Takashi Kadowaki. Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nature Communications*, 7:10531, 1 2016.

[JRW$^+$12] Luke Jostins, Stephan Ripke, Rinse K. Weersma, Richard H. Duerr, Dermot P. McGovern, Ken Y. Hui, James C. Lee, L. Philip Schumm, Yashoda Sharma, Carl A. Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L. Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N. Ananthakrishnan, Vibeke Andersen, Jane M. Andrews, Leonard Baidoo, Tobias Balschun, Peter A. Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro DAmato, Dirk De Jong, Kathy L. Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R. Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jrgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H. Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C. Lawrance, Charlie W. Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R. Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y. Ponsioen, Uros Potocnik, Natalie J. Prescott, Miguel Regueiro, Jerome I. Rotter, Richard K. Russell, Jeremy D. Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A. Simms, Jurgita Sventoraityte, Stephan R. Targan, Kent D. Taylor, Mark Tremelling, Hein W. Verspaget, Martine De Vos, Cisca Wijmenga, David C. Wilson, Juliane Winkelmann, Ramnik J. Xavier, Sebastian Zeissig, Bin Zhang, Clarence K. Zhang, Hongyu Zhao, Mark S. Silverberg, Vito Annese, Hakon Hakonarson, Steven R. Brant, Graham Radford-Smith, Christopher G. Mathew, John D. Rioux, Eric E. Schadt, Mark J. Daly, Andre Franke, Miles Parkes, Severine Vermeire, Jeffrey C. Barrett, Judy H Cho, and Judy H Cho. Hostmicrobe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 10 2012.

[JWTV98] RW Johnstone, J Wang, N Tommerup, and H Vissing. Ciao 1 is a novel

WD40 protein that interacts with the tumor suppressor protein WT1. *Journal of Biological*, 1998.

[KAT+11] Dwi Setyowati Karolina, Arunmozhiarasi Armugam, Subramaniam Tavintharan, Michael T K Wong, Su Chi Lim, Chee Fang Sum, and Kandiah Jeyaseelan. MicroRNA 144 impairs insulin signaling by inhibiting the expression of insulin receptor substrate 1 in type 2 diabetes mellitus. *PloS one*, 6(8):e22839, 2011.

[KG00] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 1 2000.

[KJB+13] Sangwoo Kim, Kyowon Jeong, Kunal Bhutani, Jeong Ho Lee, Anand Patel, Eric Scott, Hojung Nam, Hayan Lee, Joseph G Gleeson, and Vineet Bafna. Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome biology*, 14(8):R90, 8 2013.

[KK13] Zeid Khitan and Dong Hyun Kim. Fructose: a key factor in the development of metabolic syndrome and hypertension. *Journal of nutrition and metabolism*, 2013:682673, 2013.

[KMP+09] Janne Korhonen, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics (Oxford, England)*, 25(23):3181–2, 12 2009.

[KMV+13] Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F. McMichael, Matthew A. Wyczalkowski, Mark D. M. Leiserson, Christopher A. Miller, John S. Welch, Matthew J. Walter, Michael C. Wendl, Timothy J. Ley, Richard K. Wilson, Benjamin J. Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 10 2013.

[KPT+13] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013.

[KWJ+14] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–5, 3 2014.

[KZL+12] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher a Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, pages 568–576, 2 2012.

[Lac08] H M Lachman. Copy variations in schizophrenia and bipolar disorder. *Cytogenetic and genome research*, 123(1-4):27–35, 2008.

[LD02] Roderick J. A. Little and Rubin B. Donald. *Statistical Analysis with Missing Data.* Wiley Online Library, 2002.

[LD09] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, 7 2009.

[LD10] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–95, 3 2010.

[LFJ69] Frederick Li and Joseph Fraumeni Jr. Soft-Tissue Sarcomas, Breast Cancer, and Other NeoplasmsA Familial Syndrome? *Annals of Internal Medicine*, 71(4):747–752, 1969.

[LHA14] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 12 2014.

[LHL$^+$12] Ernest T Lam, Alex Hastie, Chin Lin, Dean Ehrlich, Somes K Das, Michael D Austin, Paru Deshpande, Han Cao, Niranjan Nagarajan, Ming Xiao, and Pui-Yan Kwok. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly, 2012.

[LHL$^+$14] Jing Liu, Xiao-feng Huang, Ju-xiang Liu, Jin-xing Quan, Li-min Tian, Xiao-juan Huang, Jia Liu, Yan-jia Xu, Qi Zhang, Shu-lan Zhang, Xiao-hui Chen, and Rui-lan Niu. Association of PARL Gene Rs3732581, Rs73887537 Polymorphisms with Type 2 Diabetes Mellitus, Insulin Resistance and Blood Lipid Levels in Chinese Population. *Journal of Metabolic Syndrome*, 03(01):1–7, 2014.

[LHW$^+$09] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, 8 2009.

[Li11] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[LLL^{+}11] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 9 2011.

[LLY^{+}15] Seon-Yeong Lee, Seung Hoon Lee, Eun-Ji Yang, Eun-Kyung Kim, Jae-Kyung Kim, Dong-Yun Shin, Mi-La Cho, DC Baumgart, WJ Sandborn, A Dignass, R Eliakim, F Magro, C Maaser, Y Chowers, K Geboes, A Raza, W Yousaf, R Giannella, MT Shata, J Galvez, CJ Wruck, A Fragoulis, A Gurzynski, LO Brandenburg, YW Kan, K Chan, FM Gu, QL Li, Q Gao, JH Jiang, K Zhu, XY Huang, XO Yang, AD Panopoulos, R Nurieva, SH Chang, D Wang, SS Watowich, EJ Yang, J Lee, SY Lee, EK Kim, YM Moon, YO Jung, J Jhun, J Lee, JK Byun, EK Kim, JW Woo, JH Lee, H Capiralla, V Vingtdeux, J Venkatesh, U Dreses-Werringloer, H Zhao, P Davies, C Kitabayashi, T Fukada, M Kanamoto, W Ohashi, S Hojyo, T Atsumi, L Zhou, JE Lopes, MM Chong, II Ivanov, R Min, GD Victora, MS Maddur, P Miossec, SV Kaveri, J Bayry, C Selmi, CS Park, BR Bang, HS Kwon, KA Moon, TB Kim, KY Lee, HA Hirsch, D Iliopoulos, K Struhl, K Isoda, JL Young, A Zirlik, LA MacFarlane, N Tsuboi, N Gerdes, A Hahn-Windgassen, V Nogueira, CC Chen, JE Skeen, N Sonenberg, N Hay, DB Shackelford, RJ Shaw, HJ Son, J Lee, SY Lee, EK Kim, MJ Park, KW Kim, DA van Heel, IA Udalova, AP De Silva, DP McGovern, Y Kinouchi, J Hull, I Marafini, F Zorzi, S Codazza, F Pallone, G Monteleone, L Deng, JF Zhou, RS Sellers, JF Li, AV Nguyen, Y Wang, ME Himmel, Y Yao, PC Orban, TS Steiner, MK Levings, N Eastaff-Leung, N Mabarrack, A Barbour, A Cummins, S Barry, H Yu, D Pardoll, R Jove, R Izutani, EY Loh, HC Reinecker, Y Ohno, RD Fusunyan, GR Lichtenstein, HC Reinecker, M Steffen, T Witthoeft, I Pflueger, S Schreiber, RP MacDermott, F Scaldaferri, S Vetrano, M Sans, V Arena, G Straface, E Stigliano, J Zhou, J Wulfkuhle, H Zhang, P Gu, Y Yang, J Deng, D Micic, G Cvijovic, V Trajkovic, LH Duntas, S Polovina, Y Zheng, AY Rudensky, P Miossec, JK Kolls, WA Goodman, AB Young, TS McCormick, KD Cooper, A Levine, L Passerini, SE Allan, M Battaglia, S Di Nunzio, AN Alstad, MK Levings, SJ Koh, JM Kim, IK Kim, SH Ko, and JS Kim. Metformin Ameliorates Inflammatory Bowel Disease by Suppression of the STAT3 Signaling Pathway and Regulation of the between Th17/Treg Balance. *PLOS ONE*, 10(9):e0135858, 9 2015.

[LS94] Es Eric S Lander and Nicholas J. Nj Schork. Genetic dissection of complex traits. *Science*, 265:2037–2048, 1994.

[LUS^{+}11] Louise C. Laurent, Igor Ulitsky, Ileana Slavin, Ha Tran, Andrew Schork, Robert Morey, Candace Lynch, Julie V. Harness, Sunray Lee, Maria J. Barrero, Sherman Ku, Marina Martynova, Ruslan Semechkin, Vasiliy

Galat, Joel Gottesfeld, Juan Carlos Izpisua Belmonte, Chuck Murry, Hans S. Keirstead, Hyun-Sook Park, Uli Schmidt, Andrew L. Laslett, Franz-Josef Muller, Caroline M. Nievergelt, Ron Shamir, and Jeanne F. Loring. Dynamic Changes in the Copy Number of Pluripotency and Cell Proliferation Genes in Human ESCs and iPSCs during Reprogramming and Time in Culture. *Cell Stem Cell*, 8(1):106–118, 1 2011.

[Mar11] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, 2011.

[MBDL+10] Y Mayshar, U Ben-David, N Lavon, J C Biancotti, B Yakir, A T Clark, K Plath, W E Lowry, and N Benvenisty. Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell*, 7(4):521–531, 2010.

[MBF+12] Daniel G MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, Stephen B Montgomery, Cornelis A Albers, Zhengdong D Zhang, Donald F Conrad, Gerton Lunter, Hancheng Zheng, Qasim Ayub, Mark A DePristo, Eric Banks, Min Hu, Robert E Handsaker, Jeffrey A Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H A Barnes, Clara Amid, Denise R Carvalho-Silva, Alexandra H Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David N Cooper, Yali Xue, Irene Gallego Romero, 1000 Genomes Project 1000 Genomes Project Consortium, Jun Wang, Yingrui Li, Richard A Gibbs, Steven A McCarroll, Emmanouil T Dermitzakis, Jonathan K Pritchard, Jeffrey C Barrett, Jennifer Harrow, Matthew E Hurles, Mark B Gerstein, and Chris Tyler-Smith. A systematic survey of loss-of-function variants in human protein-coding genes. *Science (New York, N.Y.)*, 335(6070):823–8, 2 2012.

[McV02] Gilean A T McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–91, 10 2002.

[MHB+10] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–303, 9 2010.

[MHM+07] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913, 7 2007.

[MP14] Vijay Menon and Lawrence Povirk. Involvement of p53 in the repair of DNA double strand breaks: multifaceted Roles of p53 in homologous recombination repair (HRR) and non-homologous end joining (NHEJ). *Sub-cellular biochemistry*, 85:321–36, 2014.

[Ngu10] J.V. Nguyen. *Genomic Mapping: A statistical and algorithmic analysis of the optical mapping system.* PhD thesis, University of Southern California, 2010.

[OAM$^+$15] Tobore Onojighofia, Natasha Anand, Brian Meshkin, Sanford Silverman, Derrick Holman, John Hubbard, May Hafez, and Svetlana Kantorovich. Abstract P156: A Novel Gene-based Tool to Predict the Risk of Essential Hypertension and Initial Validation. *Hypertension*, 66(Suppl 1), 2015.

[OIY07] Keisuke Okita, Tomoko Ichisaka, and Shinya Yamanaka. Generation of germline-competent induced pluripotent stem cells. *Nature*, 448(7151):313–317, 7 2007.

[PCKQ13] Umadevi Paila, Brad A Chapman, Rory Kirchner, and Aaron R Quinlan. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS computational biology*, 9(7):e1003153, 7 2013.

[PL14] S. E. Peterson and J. F. Loring. Genomic Instability in Pluripotent Stem Cells: Implications for Clinical Applications. *Journal of Biological Chemistry*, 289(8):4578–4584, 2 2014.

[PPB$^+$16] Cameron Palmer, Itsik Peer, SR Browning, J Marchini, B Howie, S Myers, G McVean, P Donnelly, AL Price, A Tandon, N Patterson, KC Barnes, N Rafaels, I Ruczinski, HM Kang, JH Sul, SK Service, NA Zaitlen, Sy Kong, NB Freimer, S Prabhu, I Peer, AL Price, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, D Reich, GR Abecasis, SS Cherny, WO Cookson, LR Cardon, BM Neale, MA Rivas, BF Voight, D Altshuler, B Devlin, M Orho-Melander, RJA Little, DB Rubin, YF Pei, J Li, L Zhang, CJ Papasian, HW Deng, BL Browning, SR Browning, B Howie, C Fuchsberger, M Stephens, J Marchini, GR Abecasis, M Nothnagel, D Ellinghaus, S Schreiber, M Krawczak, A Franke, K Nho, L Shen, S Kim, S Swaminathan, SL Risacher, AJ Saykin, AE Locke, B Kahali, SI Berndt, AE Justice, TH Pers, FR Day, AR Wood, T Esko, J Yang, S Vedantam, TH Pers, S Gustafsson, EF Acar, L Sun, Y Li, CJ Willer, J Ding, P Scheet, GR Abecasis, BN Howie, P Donnelly, J Marchini, C Fuchsberger, GR Abecasis, DA Hinds, CJ Willer, Y Li, GR Abecasis, O Harel, JL Schafer, JW Graham, AE Olchowski, TD Gilreath, D Lee, TB Bigdeli, BP Riley, AH Fanous, SA Bacanu, B Pasaniuc, N Zaitlen, H Shi, G Bhatia, A Gusev, J Pickrell, D Lee, TB Bigdeli, VS Williamson, VI Vladimirov, BP Riley, AH Fanous, MD Mailman, M Feolo, Y Jin, M Kimura, K Tryka, R Bagoutdinov, S Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, D Bender, O Delaneau, J Marchini, and JF Zagury. Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation. *PLOS Genetics*, 12(6):e1006091, 6 2016.

[PSE+15] Phillip H. Pham, William J. Shipman, Galina A. Erikson, Nicholas J. Schork, and Ali Torkamani. Scripps Genome ADVISER: Annotation and Distributed Variant Interpretation SERver. *PLOS ONE*, 10(2):e0116815, 2 2015.

[PVG+11] Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[PWA+08] Brenda L. Powell, Steven Wiltshire, Gillian Arscott, Pamela A. McCaskie, Joseph Hung, Brendan M. McQuillan, Peter L. Thompson, Kim W. Carter, Lyle J. Palmer, and John P. Beilby. Association of PARL rs3732581 genetic variant with insulin levels, metabolic syndrome and coronary artery disease. *Human Genetics*, 124(3):263–270, 10 2008.

[RDM+12] Andrew Roth, Jiarui Ding, Ryan Morin, Anamaria Crisan, Gavin Ha, Ryan Giuliany, Ali Bashashati, Martin Hirst, Gulisa Turashvili, Arusha Oloumi, Marco a Marra, Samuel Aparicio, and Sohrab P Shah. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(7):907–13, 4 2012.

[RLG+15] Alex H Ramos, Lee Lichtenstein, Manaswi Gupta, Michael S Lawrence, Trevor J Pugh, Gordon Saksena, Matthew Meyerson, and Gad Getz. Oncotator: cancer variant annotation tool. *Human mutation*, 36(4):2423–9, 4 2015.

[RMH+16] Rachel Rosenthal, Nicholas McGranahan, Javier Herrero, Barry S. Taylor, and Charles Swanton. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, 12 2016.

[RRC+13] Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51, 2013.

[Sch15] Nicholas J. Schork. Personalized medicine: Time for one-person trials. *Nature*, 520(7549):609–611, 4 2015.

[SCTS70] C R Scott, S H Clark, C C Teng, and K R Swedberg. Clinical and cellular studies of sarcosinemia. *The Journal of pediatrics*, 77(5):805–11, 11 1970.

[SDB$^+$14] Thorsten M Schlaeger, Laurence Daheron, Thomas R Brickler, Samuel Entwisle, Karrie Chan, Amelia Cianci, Alexander DeVine, Andrew Ettenger, Kelly Fitzgerald, Michelle Godfrey, Dipti Gupta, Jade McPherson, Prerana Malwadkar, Manav Gupta, Blair Bell, Akiko Doi, Namyoung Jung, Xin Li, Maureen S Lynes, Emily Brookes, Anne B C Cherry, Didem Demirbas, Alexander M Tsankov, Leonard I Zon, Lee L Rubin, Andrew P Feinberg, Alexander Meissner, Chad A Cowan, and George Q Daley. A comparison of non-integrating reprogramming methods. *Nature Biotechnology*, 33(1):58–63, 12 2014.

[SFP$^+$14] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, Craig L Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, Melanie Waldenberger, J Brent Richards, Robert P Mohney, Michael V Milburn, Sally L John, Jeff Trimmer, Fabian J Theis, John P Overington, Karsten Suhre, M Julia Brosnan, Christian Gieger, Gabi Kastenmüller, Tim D Spector, Nicole Soranzo, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, 5 2014.

[SGZ$^+$16] S. M. Smith, J. F. Gregory, G. H. Zeisel, C. R. Gibson, T. H. Mader, J. Kinchen, P. Ueland, R. Ploutz-Snyder, M. Heer, and S. R. Zwart. Risk of Visual Impairment and Intracranial Hypertension After Space Flight: Evaluation of the Role of Polymorphism of Enzymes Involved in One-Carbon Metabolism. 2016.

[SLF$^+$15] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson, G.E. Robinson, G.R. Abecasis, L. Chin, J.N. Andersen, P.A. Futreal, J. Giles, C. Mora, J.Y. Li, J. Wang, R.S. Zeigler, J. Zhu, J.A. Eisen, D. Gevers, P. Sulem, W.W. Soon, M. Hariharan, M.P. Snyder, R. Chen, M. Hsi-Yang Fritz, P.-R. Loh, M. Baym, B. Berger, S. Koren, A.M. Phillippy, S. Christley, R. Patro, S.M. Mount, C. Kingsford, A. Golden, S. Djorgovski, J. Greally, C.H. Djoerd Hiemstra, A. Sboner, M. Baker, Y. Erlich, A. Narayanan, B. Langmead, S. Kurtz, O. Trelles, M.C. Schatz, B. Langmead, and S.L. Salzberg. Big Data: Astronomical or Genomical? *PLOS Biology*, 13(7):e1002195, 7 2015.

[Smy05] G. K. Smyth. limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer-Verlag, New York, 2005.

[SMZ+14] M. Sałaga, A. Mokrowiecka, P.K. Zakrzewski, A. Cygankiewicz, E. Leishman, M. Sobczak, H. Zatorski, E. Małecka-Panas, R. Kordek, M. Storr, W.M. Krajewska, H.B. Bradshaw, and J. Fichna. Experimental colitis in mice is attenuated by changes in the levels of endocannabinoid metabolites induced by selective inhibition of fatty acid amide hydrolase (FAAH). *Journal of Crohn's and Colitis*, 8(9), 2014.

[SP10] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python, 2010.

[SSK+07] Bertrand Servin, Matthew Stephens, P Kraft, P Pharoah, SJ Chanock, D Albanes, LN Kolonel, N Li, M Stephens, P Scheet, M Stephens, L Almasy, J Blangero, EI George, RE McCulloch, JD Nott, PJ Green, X Yang, TR Belin, WJ Boscardin, M Stephens, N Smith, P Donnelly, M Stephens, P Scheet, M Lynch, B Walsh, H Jeffreys, AE Raftery, DJ Lunn, JC Whittaker, N Best, RE Kass, AE Raftery, I Good, G Hellenthal, M Stephens, CS Carlson, MA Eberle, MJ Rieder, Q Yi, L Kruglyak, JM Chapman, JD Cooper, JA Todd, DG Clayton, ME Weale, C Depondt, SJ Macdonald, A Smith, PS Lai, SK Tate, C Depondt, SM Sisodiya, GL Cavalleri, S Schorge, J Marchini, D Cutler, N Patterson, M Stephens, E Eskin, EL Heinzen, W Yoon, SK Tate, A Sen, NW Wood, S Zöllner, JK Pritchard, N Soranzo, GL Cavalleri, ME Weale, NW Wood, C Depondt, J Marchini, S Myers, G McVean, P Donnelly, I Pe'er, PIW De Bakker, J Maller, K Jones, MD Altshuler, B Servin, M Stephens, DL Nicolae, JY Dai, I Ruczinski, M LeBlanc, C Kooperberg, MJ Sillanpaa, M Bhattacharjee, AP Morris, J Cohen, A Pertsemlidis, IK Kotowski, R Graham, and CK Garcia. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genetics*, 3(7):e114, 2007.

[STW+15] Shutong Shen, Lichan Tao, Xiuzhi Wang, Xiangqing Kong, and Xinli Li. Common Variants for Heart Failure. *Current genomics*, 16(2):82–7, 4 2015.

[SVSV15] Charles Savona-Ventura and Stephanie Savona-Ventura. The inheritance of obesity. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 29(3):300–308, 2015.

[SWF16] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. 2016.

[SWS+12] Christopher T Saunders, Wendy Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)*, pages 1–7, 5 2012.

[SZNG93] Katherine Schneider, Kristin Zelley, Kim E Nichols, and Judy Garber. *Li-Fraumeni Syndrome*. University of Washington, Seattle, 1993.

[Tan11] O Tange. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine*, 36(1):42–47, 2 2011.

[TNN11] S M Taapken, B S Nisler, and M A Newton. Karyotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. *Nature Biotechnology*, 29:313–314, 2011.

[TOMM+09] Vitor Hugo Teixeira, Robert Olaso, Marie-Laure Martin-Magniette, Sandra Lasbleiz, Laurent Jacq, Catarina Resende Oliveira, Pascal Hilliquin, Ivo Gut, Franois Cornelis, and Elisabeth Petit-Teixeira. Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. *PloS one*, 4(8):e6803, 8 2009.

[TY06] Kazutoshi Takahashi and Shinya Yamanaka. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676, 8 2006.

[VSZW06] Anton Valouev, David C Schwartz, Shiguo Zhou, and Michael S Waterman. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15770–15775, 2006.

[VTS+11] Anita Villani, Uri Tabori, Joshua Schiffman, Adam Shlien, Joseph Beyene, Harriet Druker, Ana Novokmet, Jonathan Finlay, and David Malkin. Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: a prospective observational study. *The Lancet Oncology*, 12(6):559–567, 6 2011.

[WDC+07] Feng Wu, Themistocles Dassopoulos, Leslie Cope, Anirban Maitra, Steven R. Brant, Mary L. Harris, Theodore M. Bayless, Giovanni Parmigiani, and Shukti Chakravarti. Genome-wide gene expression differences in Crohns disease and ulcerative colitis from endoscopic pinch biopsies: Insights into distinctive pathogenesis. *Inflammatory Bowel Diseases*, 13(7):807–821, 7 2007.

[WJG+13] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov,

D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807, 1 2013.

[WLGH15] Omer Weissbrod, Christoph Lippert, Dan Geiger, and David Heckerman. Accurate liability estimation improves power in ascertained case-control studies. *Nature Methods*, 12(4):332–334, 2 2015.

[WMM+14] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 1 2014.

[WSS+13] Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, Martin L Buchkovich, Samia Mora, Jacques S Beckmann, Jennifer L Bragg-Gresham, Hsing-Yi Chang, Aye Demirkan, Heleen M Den Hertog, Ron Do, Louise A Donnelly, Georg B Ehret, Tnu Esko, Mary F Feitosa, Teresa Ferreira, Krista Fischer, Pierre Fontanillas, Ross M Fraser, Daniel F Freitag, Deepti Gurdasani, Kauko Heikkilä, Elina Hyppönen, Aaron Isaacs, Anne U Jackson, sa Johansson, Toby Johnson, Marika Kaakinen, Johannes Kettunen, Marcus E Kleber, Xiaohui Li, Jian'an Luan, Leo-Pekka Lyytikäinen, Patrik K E Magnusson, Massimo Mangino, Evelin Mihailov, May E Montasser, Martina Müller-Nurasyid, Ilja M Nolte, Jeffrey R O'Connell, Cameron D Palmer, Markus Perola, Ann-Kristin Petersen, Serena Sanna, Richa Saxena, Susan K Service, Sonia Shah, Dmitry Shungin, Carlo Sidore, Ci Song, Rona J Strawbridge, Ida Surakka, Toshiko Tanaka, Tanya M Teslovich, Gudmar Thorleifsson, Evita G Van den Herik, Benjamin F Voight, Kelly A Volcik, Lindsay L Waite, Andrew Wong, Ying Wu, Weihua Zhang, Devin Absher, Gershim Asiki, Ins Barroso, Latonya F Been, Jennifer L Bolton, Lori L Bonnycastle, Paolo Brambilla, Mary S Burnett, Giancarlo Cesana, Maria Dimitriou, Alex S F Doney, Angela Döring, Paul Elliott, Stephen E Epstein, Gudmundur Ingi Eyjolfsson, Bruna Gigante, Mark O Goodarzi, Harald Grallert, Martha L Gravito, Christopher J Groves, Gran Hallmans, Anna-Liisa Hartikainen, Caroline Hayward, Dena Hernandez, Andrew A Hicks, Hilma Holm, Yi-Jen Hung, Thomas Illig, Michelle R Jones, Pontiano Kaleebu, John J P Kastelein, Kay-Tee Khaw, Eric Kim, Norman Klopp, Pirjo Komulainen, Meena Kumari, Claudia Langenberg, Terho Lehtimäki, Shih-Yi Lin, Jaana Lindström, Ruth J F Loos, Franois Mach, Wendy L McArdle, Christa Meisinger, Braxton D Mitchell, Gabrielle Müller, Ramaiah Nagaraja, Narisu Narisu, Tuomo V M Nieminen, Rebecca N Nsubuga, Isleifur Olafsson, Ken K Ong, Aarno Palotie, Theodore Papamarkou, Cristina Pomilla, Anneli Pouta, Daniel J Rader, Muredach P Reilly, Paul M Ridker, Fernando Rivadeneira, Igor Rudan, Aimo Ruokonen, Nilesh Samani, Hubert Scharnagl, Janet Seeley, Kaisa Silander, Alena Stancáková, Kathleen Stirrups, Amy J Swift, Laurence Tiret, Andre G Uitterlinden, L Joost van Pelt, Sailaja Vedantam, Nicholas Wainwright, Cisca Wijmenga, Sarah H Wild, Gonneke Willemsen, Tom Wilsgaard, James F Wilson, Elizabeth H Young, Jing Hua Zhao, Linda S Adair, Dominique Arveiler, Themistocles L Assimes, Stefania Bandinelli, Franklyn Bennett, Murielle Bochud, Bernhard O Boehm, Dorret I Boomsma, Ingrid B Borecki, Stefan R Bornstein, Pascal Bovet, Michel Burnier, Harry Campbell, Aravinda Chakravarti, John C Chambers, Yii-Der Ida Chen, Francis S Collins, Richard S Cooper, John Danesh, George Dedoussis, Ulf de Faire, Alan B Feranil, Jean Ferrières, Luigi Ferrucci, Nelson B Freimer, Christian Gieger, Leif C Groop, Vilmundur Gudnason, Ulf Gyllensten, Anders Hamsten, Tamara B Harris, Aroon Hingorani, Joel N Hirschhorn, Albert Hofman, G Kees Hovingh, Chao Agnes Hsiung, Steve E Humphries, Steven C Hunt, Kristian Hveem, Carlos Iribarren, Marjo-Riitta Järvelin, Antti Jula, Mika Kähönen, Jaakko Kaprio, Antero Kesäniemi, Mika Kivimaki, Jaspal S Kooner, Peter J Koudstaal, Ronald M Krauss, Diana Kuh, Johanna Kuusisto, Kirsten O Kyvik, Markku Laakso, Timo A Lakka, Lars Lind, Cecilia M Lindgren, Nicholas G Martin, Winfried März, Mark I McCarthy, Colin A McKenzie, Pierre Meneton, Andres Metspalu, Leena Moilanen, Andrew D Morris, Patricia B Munroe, Inger Njølstad, Nancy L Pedersen, Chris Power, Peter P Pramstaller, Jackie F Price, Bruce M Psaty, Thomas Quertermous, Rainer Rauramaa, Danish Saleheen, Veikko Salomaa, Dharambir K Sanghera, Jouko Saramies, Peter E H Schwarz, Wayne H-H Sheu, Alan R Shuldiner, Agneta Siegbahn, Tim D Spector, Kari Stefansson, David P Strachan, Bamidele O Tayo, Elena Tremoli, Jaakko Tuomilehto, Matti Uusitupa, Cornelia M van Duijn, Peter Vollenweider, Lars Wallentin, Nicholas J Wareham, John B Whitfield, Bruce H R Wolffenbuttel, Jose M Ordovas, Eric Boerwinkle, Colin N A Palmer, Unnur Thorsteinsdottir, Daniel I Chasman, Jerome I Rotter, Paul W Franks, Samuli Ripatti, L Adrienne Cupples, Manjinder S Sandhu, Stephen S Rich, Michael Boehnke, Panos Deloukas, Sekar Kathiresan, Karen L Mohlke, Erik Ingelsson, and Gonalo R Abecasis. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, 10 2013.

[XPH$^+$07] Ming Xiao, Angie Phong, Connie Ha, Ting Fung Chan, Dongmei Cai, Lucinda Leung, Eunice Wan, Amy L. Kistler, Joseph L. DeRisi, Paul R. Selvin, and Pui Yan Kwok. Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Research*, 35(3), 2007.

[YLGV11] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, 1 2011.

[YXC] Scheid S Yang X and Lottaz C. OrderedList: Similarities of Ordered Gene Lists.

[YYW$^+$15] Minglan Yang, Lei Ye, Bokai Wang, Jie Gao, Ruixin Liu, Jie Hong, Weiqing Wang, Weiqiong Gu, and Guang Ning. Decreased miR-146 expression in peripheral blood mononuclear cells is correlated with ongoing islet autoimmunity in type 1 diabetes patients 1miR-146. *Journal of Diabetes*, 7(2):158–165, 3 2015.

[ZLAS11] Jin Zheng, Yun Li, Gonalo R. Abecasis, and Paul Scheet. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic Epidemiology*, 35(2):102–110, 2 2011.