Toward Reliable Analysis of Individual Genomes

A Dissertation Presented
by
Jason Anthony O'Rawe
to
The Graduate School
in Partial Fulfillment of the
Requirements
for the degree of
Doctor of Philosophy
in
Genetics

Stony Brook University

August 2016

Stony Brook University
The Graduate School

Jason Anthony O'Rawe

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation

Gholson J. Lyon - Dissertation Advisor
Assistant Professor of Genetics

Joshua Rest - Chairperson of Defense
Associate Professor of Ecology and Evolution

Robert Patro - Committee Member
Assistant Professor of Computer Science

Christopher E. Mason - Outside Member
Associate Professor of Physiology and Biophysics
Weill Cornell Medicine

Scott Ferson - Outside Member
Professor and Chair of Uncertainty in Engineering
University of Liverpool

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Toward Reliable Analysis of Individual Genomes

by
Jason O’Rawe
Doctor of Philosophy
in
Genetics
Stony Brook University
2016

High-throughput DNA sequencing technologies have given us the power to understand genetic disease at extraordinarily detailed resolution. It is now possible to sequence a person’s whole genome and search for the genetic markers that contribute to specific disease, or even markers that contribute to the possibility of developing a new one. However, the task of understanding and sifting through billions of data-points is not a trivial one. There are diverse statistical, algorithmic and practical implementation challenges that must be met so that we can accurately and reliably analyze the vast swaths of data that come from human DNA sequences. Indeed, strategies for detecting human sequence variation in exome and whole genome sequencing data are myriad, but the reliability of these methods, even when applied to the same underlying sequencing data, is unclear. Furthermore, in the context of imperfect agreement among results stemming from these various methods, powerful strategies for assessing and recovering true, but missed, sequence variation have yet to be devised. Most research effort has focused on mitigating false detection. It is in this context that high-throughput sequencing technologies are used for both research and clinical investigations.

In the medical genomics realm, our understanding of the genetic origins of human disease has been empowered by these technologies, but unreliable analyses have led to a number of false positive research findings. The community has since recognized the need for robust and comprehensive sequencing and analysis methods, particularly in cases where only a small number of samples from probands or affected families are available. In the clinical realm, most agree that there exists an enormous amount of potential for these technologies to transform clinical care, but the practicality of their use is currently understudied, particularly for individual patients among complex cohorts, such as those harboring psychiatric afflictions.

In order to move the field of human genetics research forward and to contribute toward the successful implementation of genomics-guided medical care, several key advancements are needed: a characterization of the reliability of current high-throughout analysis methods, methods for recovering missed sequence variants from discordant detection sets, an understanding of current infrastructural deficiencies for implementation, general guidance on how to use diverse sets of analysis results in the context of generating robust relationships between human sequence variation and disease, and new methodological approaches for generating sequence analysis results that accurately characterize uncertainties in the underlying data, so that the reliabilities of their inferences remain robust throughout the lifetime of their use.

Contents

List of figures

List of Tables

Publications

First author publications related to the dissertation

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., & Johnson, W. E. (2013). *Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing*. Genome med, 5(3), 28.

O'Rawe, J. A., Fang, H., Rynearson, S., Robison, R., Kiruluta, E. S., Higgins, G., Eilbeck, K., Reese, M. G., & Lyon, G. J. (2013). *Integrating precision medicine in the study and clinical treatment of a severely mentally ill person*. PeerJ, 1, e177.

O'Rawe, J. A., Ferson, S., & Lyon, G. J. (2015). *Accounting for uncertainty in DNA sequencing data*. Trends in Genetics, 31(2), 61-66.

O'Rawe, J. A., Wu, Y., Dorfel, M. J., Rope, A. F., Au, P. Y., Parboosingh, J. S., Moon, S., Kousi, M., Kosma, K., Smith, C. S., Tzetis, M., Schuette, J. L., Hufnagel, R. B., Prada, C. E., Martinez, F., Orellana, C., Crain, J., Caro-Llopis, A., Oltra, S., Monfort, S., Jimenez-Barron, L. T., Swensen, J., Ellingwood, S., Smith, R., Fang, H., Ospina, S., Stegmann, S., Den Hollander, N., Mittelman, D., Highnam, G., Robison, R., Yang, E., Faivre, L., Roubertie, A., Riviere, J. B., Monaghan, K. G., Wang, K., Davis, E. E., Katsanis, N., Kalscheuer, V. M., Wang, E. H., Metcalfe, K., Kleefstra, T., Innes, A. M., Kitsiou-Tzeli, S., Rosello, M., Keegan, C. E., & Lyon, G. J. (2015). *TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations*. Am J Hum Genet, 97(6), 922-932.

Other publications

Lyon, G. J., & **O'Rawe, J.** (2015). *Human Genetics and Clinical Aspects of Neurodevelopmental Disorders*. The Genetics of Neurodevelopmental Disorders (Vol. 1, pp. 289-317): Wiley-Blackwell.

Narzisi, G., **O'Rawe, J. A.**, Iossifov, I., Fang, H., Lee, Y.-h., Wang, Z., Wu, Y., Lyon, G. J., Wigler, M., & Schatz, M. C. (2014). *Accurate de novo and transmitted indel detection in exome-capture data using microassembly*. Nature methods, 11(10), 1033-1036.

Ferson, S., **O'Rawe, J.**, & Balch, M. (2014). *Computing with confidence: imprecise posteriors and predictive distributions*. Vulnerability, Uncertainty, and Risks: Quantification, Mitigation, and Management, 895-904.

Ferson, S., **O'Rawe, J.**, Antonenko, A., Siegrist, J., Mickley, J., Luhmann, C. C., Sentz, K., & Finkel, A. M. (2015). *Natural language of uncertainty: numeric hedge words*. International Journal of Approximate Reasoning, 57, 19-39.

He, M., Person, T. N., Hebbring, S. J., Heinzen, E., Ye, Z., Schrodi, S. J., McPherson, E. W., Lin, S. M., Peissig, P. L., Brilliant, M. H., **O'Rawe, J.**, Robison, R. J., Lyon, G. J., & Wang, K. (2015). *SeqHBase: a big data toolset for family based sequencing data analysis*. J Med Genet, 52(4), 282-288.

Xiang, G., **O'Rawe, J.**, Kreinovich, V., Hajagos, J., & Ferson, S. (2014). *Protecting patient privacy while preserving medical information for research*.

Jiménez-Barrón, L. T., **O'Rawe, J. A.**, Wu, Y., Yoon, M., Fang, H., Iossifov, I., & Lyon, G. J. (2015). *Genome-wide variant analysis of simplex autism families with an integrative clinical-bioinformatics pipeline*. Molecular Case Studies, 1(1).

Fang, H., Wu, Y., Narzisi, G., **O'Rawe, J.**, Jimenez Barrón, L., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M., & Lyon, G. (2014). *Reducing INDEL calling errors in whole genome and exome sequencing data*. Genome Medicine, 6(10).

Acknowledgments

I would like to thank my thesis mentor Dr. Gholson Lyon, my thesis committee and the world wide web.

The following authors contributed to **Chapter 1**: Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W. Evan Johnson, Zhi Wei, Kai Wang and Gholson J Lyon.

The following authors contributed to **Chapter 2**: Han Fang, Shawn Rynearson, Reid Robison, Edward S. Kiruluta, Gerald Higgins, Karen Eilbeck, Martin G. Reese, and Gholson J. Lyon.

The following authors contributed to **Chapter 3**: Yiyang Wu, Max J. Dörfel, Alan F. Rope, P.Y. Billie Au, Jillian S. Parboosingh, Sungjin Moon, Maria Kousi, Konstantina Kosma, Christopher S. Smith, Maria Tzetis, Jane L. Schuette, Robert B. Hufnagel, Carlos E. Prada, Francisco Martinez, Carmen Orellana, Jonathan Crain, Alfonso Caro-Llopis, Silvestre Oltra, Sandra Monfort, Laura T. Jiménez-Barrón, Jeffrey Swensen, Sara Ellingwood, Rosemarie Smith, Han Fang, Sandra Ospina, Sander Stegmann, Nicolette Den Hollander, David Mittelman, Gareth Highnam, Reid Robison, Edward Yang, Laurence Faivre, Agathe Roubertie, Jean-Baptiste Rivière, Kristin G. Monaghan, Kai Wang, Erica E. Davis, Nicholas Katsanis, Vera M. Kalscheuer, Edith H. Wang, Kay Metcalfe, Tjitske Kleefstra, A. Micheil Innes, Sophia Kitsiou-Tzeli, Monica Rosello, Catherine E. Keegan and Gholson J. Lyon.

The following authors contributed to **Chapter 4**: Scott Ferson and Gholson Lyon

0

Introduction

Genetic disease accounts for 20-30% of infant deaths[18] and 30-50% of post-neonatal deaths[116]. 3-5% of all births result in genetic disease[87] and 12% of all chronic adult aliments have a significant genetic component[93]. Clinical and research investigations into the precise genetic contributors of human disease have historically been difficult, slow, low-yielding and generally imprecise. Since the advent of high throughput and genome-scale DNA sequencing technologies, researchers and clinicians can quickly and cheaply study the mechanisms of genetic disease at nucleotide resolution. The implications for biomedical research and human health care have thus far been invaluable. In 2010, whole genome sequencing (WGS) led to the discovery of the genetic basis of Miller Syndrome[257]. In another instance WGS was used to investigate the genetic basis of Charcot-Marie-Tooth neuropathy[174], and it has since fueled a discussion about returning genome wide test results to study participants[189]. In 2011, WGS led to the diagnosis of a pair of twins with dopa (3,4-dihydroxyphenylalanine) responsive dystonia (DRD; OMIM #128230) and to the discovery that they carried compound heterozygous mutations in the SPR gene encoding sepiapterin reductase. This information led to the clinical supplementation of l-dopa therapy with 5-hydroxytryptophan (a serotonin precursor), which resulted in remarkable clinical improvements in both twins[10]. The application of these technologies has since led to the discovery of the genetic basis of many other human disorders, and it has even shed light on the genetic architecture of complex diseases such as autism[53,126,127], heart disease[223] and cancer[60]. The health care industry has, in parallel to this development, begun to shift its focus from reactive to preventative measures, where the use genomic information, many speculate, will be of immense value (through reactive measures taken in the context of assessing disease risk loci). Some have further speculated that such a shift could improve life span, improve the quality of life people experience and vastly reduce healthcare costs.

Empirical estimates seem to suggest that exome sequencing can identify a putative disease associated variant in only about 10-50% of the cases for which it is applied[182]. However, the diagnostic yield for complex traits is likely to be significantly lower, such as for neuropsychiatric illness where the underlying genetic architecture of these diseases is still largely undefined and controversial[141,197,198,301]. In contrast to the cost of exome sequencing, which targets and sequences only the exonic regions of the genome, the cost of a whole genome is decreasing more rapidly as new technologies emerge. In addition, there is emerging evidence that exon capture and sequencing only achieves high depth of sequencing coverage in about 90% of the exons, whereas whole genome sequencing does not involve a capture step and thus obtains better coverage on >95% of all exons in the genome. Of course, even the definition of the exome is a moving target, as the research community is constantly annotating and finding new exons not previously discovered[312,322]. Indeed, whole genome sequencing is known to yield higher sensitivity for detecting short genomic deletions and insertions even in genomic regions accessible to both technologies[73]. As a consequence, whole genome sequencing is becoming recognized by the community as a much more cost effective and comprehensive way to assess coding and non-coding regions of the genome.

Before high-throughput DNA sequencing can be reliably used in clinical applications, there are still a number of statistical, algorithmic and practical implementation challenges that must be met. In this thesis document, I will discuss four studies that aim to assess the reliability and accuracy of variant calling in exome sequencing data, asses the usefulness and practicality of sequencing a single human genome in the context of clinical care, use high-throughput sequencing technology in a robust way to facilitate the discovery of the genetic basis of a Mendelian disorder and develop methods for comprehensive uncertainty characterization in the context of detecting human sequence variation, so as to increase the reliability of these data.

The first study[225] involves the sequencing and analysis of 15 human exomes and one whole genome from four families. Variant detection reliability is assessed by comparing the detection outputs from a range of sequence analysis tools. Large scale validation data were generated and used to (i) investigate validity of discordant variants and (ii) recover missed sequence variation using machine learning models. Models were also used to characterize the discriminatory power of various predictors of variant calling errors.

The second study[228] aims to assess the usefulness and practical limitations of clinical-grade whole genome sequencing in the context of a person with severe obsessive compulsive disorder (OCD), who has also been treated with deep brain stimulation (DBS) for his OCD. The study involves the sequencing, analysis and interpretation of this persons whole genome, as well as a clinical evaluation of his progress since receiving DBS for his OCD.

The third study[226] aims to contribute to our understanding of the genetic basis of a newly discovered disorder, which is characterized by global developmental delay, intellectual disability (ID), characteristic facial dysmorphology, generalized hypotonia, and variable neurologic features, all in male individuals. Comprehensive genomic analyses involving several affected families, as well as one large family, were undertaken. Functional RNA sequencing studies were undertaken for one of these families in the context of comprehensive whole genome sequencing. In addition, collaborative efforts with another group enabled zebrafish studies of a candidate gene.

The final study[229] reviews the state of uncertainty quantification in DNA sequencing applications, describe sources of error, and proposes methods that can be used for accounting and propagating these errors and their uncertainties through subsequent calculations. The proposed methods are then used in case studies with synthetic data to determine whether these methods are useful for increasing the reliability of some aspects of sequence analysis.

1

Variability in variant detection for individual exomes and genomes.

1.1 Motivation

Recent studies have substantiated the prevalence of rare mutations in the human genome[212,291]. Whole-genome sequencing (WGS) can uncover substantially more genetic variation than traditional single-nucleotide polymorphism (SNP) arrays, thus explaining a larger fraction of human phenotypic diversity[224,14]. This in turn is driving the sequencing of personal genomes aimed at obtaining highly accurate information about each person's genome[13,63].

Given the existence of multiple sequencing platforms and multiple data-analysis pipelines for next-generation sequencing, researchers and clinicians may be under the impression that these methods all work similarly to identify genetic variants from personal genomes. However, one group recently reported that when variants detected in the same sample by the 1000 genomes project (1KGP) and the Complete Genomics (CG) platform were compared, 19% of the single-nucleotide variants (SNVs) derived were unique to one dataset[265]. This is likely due to differences in technology, data collection, read-alignment methods, and variant-calling algorithms. The group further concluded that 'current research resources and informatics methods do not adequately account for the high level of variation that already exists in the human population, and significant efforts are needed to create resources that can accurately assess personal genomes for health, disease, and prediction of treatment outcomes'[265]. As an illustration of the widely differing methods currently being used, one of the above-referenced papers used Illumina sequencing data processed with the Short Oligonucleotide Analysis Package (SOAP) pipeline[212] whereas the other group used Illumina sequencing data processed with the Genome Analysis Toolkit (GATK) pipeline[291]. Neither group published a comparison of the overlap (concordance or discordance) between pipelines. Other researchers have worked on establishing a rigorous filtering pipeline to optimize SNV calling, reporting that the cumulative application of 12 individual filters resulted in a 290-fold reduction in the error rate[249]. Another group has worked to optimize their own pipeline utilizing, among other things, GATK and SAMtools, although it is not clear if this group compared their results with anything from SOAP[155]. This same group published a comparison of data obtained using sequencing from Illumina and CG, which showed an unexpectedly high level of discordance between the two platforms[42], which has been debated in blog postings[248,41].

Despite these previous studies comparing technical platforms, there have not been many published systematic evaluations of a number of currently used bioinformatics pipelines when generating variant calls from the same set of raw sequence data. Additionally, despite the existence of many variant-calling software tools[221], their concordance using near-default settings has not been thoroughly investigated, making it difficult to assess the relative effects on variant calling of differences in sequencing platforms versus differences in implementations of bioinformatics pipelines. Ideally, researchers and clinicians should have little to no uncertainty about the correct pipeline parameterizations for each sequencing experiment, and hence little variability with respect to their pipeline implementations; however this is rarely, if ever, the case. Indeed, knowledge about the perfect and most appropriate parameterization is often not available or easily obtainable when performing in-depth sequence analysis, and, sometimes the ’correct‘ parameters may never be precisely characterized due to the complex nature of the experiment. Researchers, clinicians and policy-makers stand to benefit from a greater understanding of the variability introduced by imperfect and non-standardized implementations of the available bioinformatics pipelines.

There are currently two major misconceptions with respect to variant calling in high throughput sequencing data. The first is that variant calling is a relatively solved problem, and the second is that newer statistical and algorithmic methods will not result in a qualitative difference in the quality of variant calling; new and better *sequencing technology* is needed. These misconceptions are virtually ubiquitous when it comes to detecting differences at the single nucleotide level (SNVs), particularly in the human medical genetics realm. Accurately detecting small insertions or deletions (INDELs), and large duplications or deletions (CNVs) as well as complex structural rearrangements (SVs) is a more difficult problem, however, and it remains an active field of research[209,305]. The implication of these misconceptions is that the development of more robust methods for detecting sequence variation at any scale has slowed, and popular methods are now considered ”gold standard” in terms of their ability to outperform most other methods[240]. Despite this, the recent inception of the US precision medicine initiative has stoked a need to revisit and improve standard methods so that DNA sequencing technologies can be reliably used for routine medical care. A recent report detailing the precision medicine road map for research in sequencing informatics stresses the need for nine major areas of investigation, most notably are areas (iii), (iv) and (v), which call for a better understanding of sequence errors and error modeling, comprehensive comparisons between various high-throughput sequence analysis pipelines, and better evaluations of bench marking strategies, respectively[8].

Substantial work has been dedicated to the study of and mitigation of false positive variant calls[258]. This is, for the most part, due to the fact that putative false positive calls can be validated in a simple and cost-efficient way. One would only need to randomly sample and validate a collection of detected variants. Validating false negative calls and determining the overall false negative rate(s) for a detection algorithm/pipeline is difficult, as false negatives are inherently difficult to detect. Determining how often a particular analysis pipeline misses true sequence variation can be assessed using synthetic data/studies, though this strategy is generally only useful for assessing performance on idealized data sets. Real data tends to be more complex than synthetic data sets. In real use-case scenarios, a truth set is invariably unavailable and the rate in which a detection pipeline misses true variants is generally unknown and poorly characterized. For practical applications of human DNA sequencing to be useful, it is critically important to have good estimates about the rate at which sequence variants are missed, as these could in practice represent dangerous misdiagnoses. Yet, methods for recovering false negative calls from sets of existing variant calls generated by a variety of methods are currently underdeveloped.

To address this issue, we carried out a study of 15 exomes and one whole genome from 15 research participants, analyzing the data with a range of different variant-calling pipelines using near-default parameters. Variant sets derived from the different pipelines will be compared and methods for recovering lost sequence variants will be developed using machine learning techniques. Our results have significant implications for analyzing personal genomes from next-generation sequencing experiments.

1.2 Methods

1.2.1 Ethics approval

The collection and genomic analysis of the DNA were approved by the institutional review board at the University of Utah, and written informed consent was obtained from all study participants. Research was carried out in compliance with the Helsinki Declaration.

1.2.2 Sample collection

The samples used in our study all came from families of human research participants ascertained in clinics at the University of Utah (see Figure 1.1 for pedigree drawings). Blood samples were collected and genomic DNA extracted using alkaline lysis and ethanol precipitation (Gentra Puregene; Qiagen Corp., Valencia, CA USA). DNA was quality-checked on agarose gels and quantified using a microvolume spectrophotometer (NanoDrop 2000; Thermo Fisher Scientific Inc., West Palm Beach, FL, USA).

1.2.3 Whole-genome sequencing and analysis with Complete Genomics

After quality control to ensure lack of genomic degradation, we sent DNA samples (10 ug) to Complete Genomics (CG) (Mountain View, CA, USA) for sequencing. The whole-genome

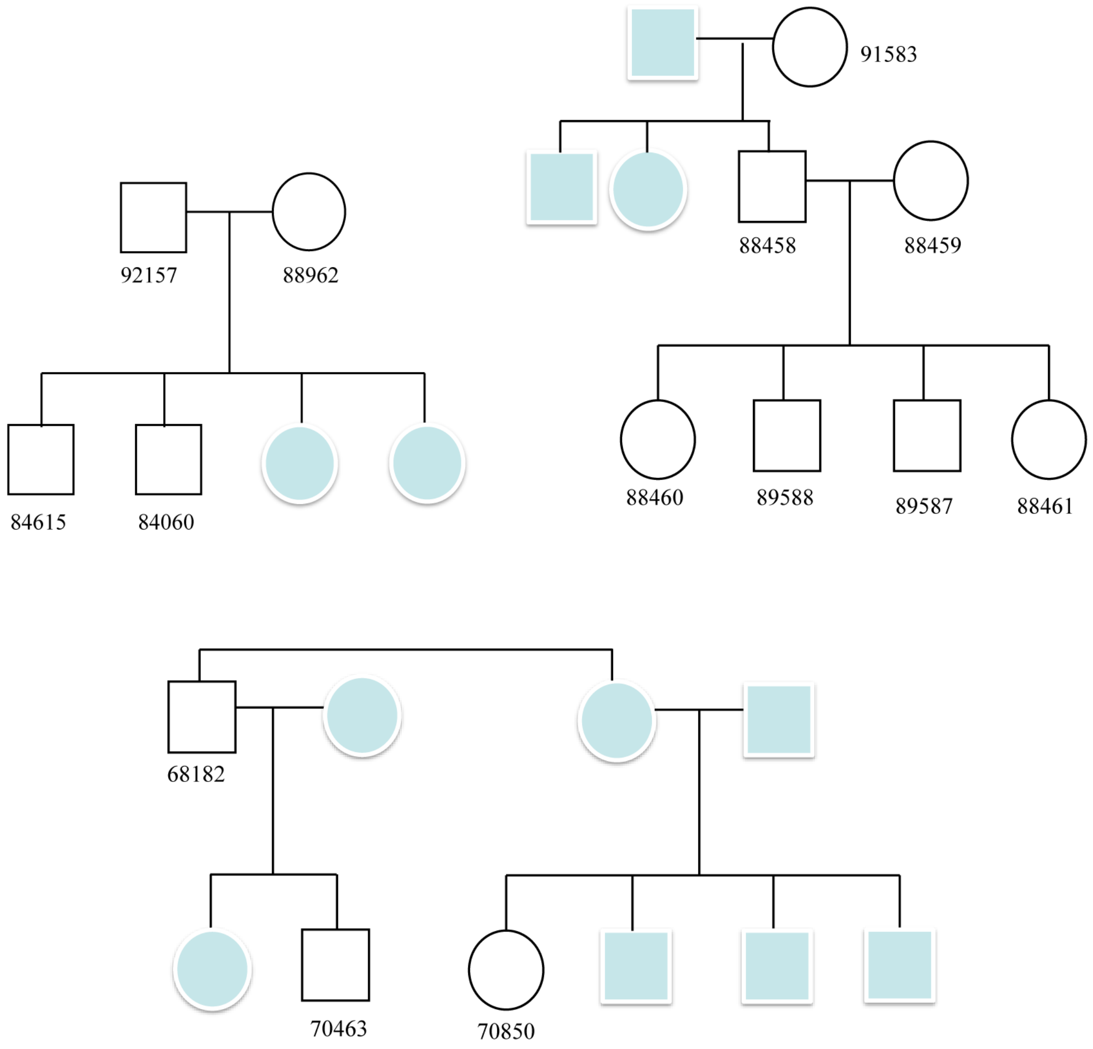


Figure 1.1: Family pedigrees contained within the 15 sequenced exomes. Of the fifteen exomes that were sequenced, 14 were sequenced from families chosen for future disease discovery related work. Each sequenced individual (numbered) is displayed in the context of his or her constituent family pedigree.

DNA was sequenced using nanoarray-based short-read sequencing by ligation technology[64], including an adaptation of the pairwise end-sequencing strategy[256]. Reads were mapped to the Genome Reference Consortium assembly GRCh37. Owing to the proprietary data formats, all the sequencing data quality control, alignment, and variant calling were performed by CG as part of their sequencing service, using their version 2.0 pipeline[34].

1.2.4 Exome capture and sequencing with Illumina HiSeq2000

Exome capture for all 15 samples was carried out using a commercially available in-solution method (SureSelect Human All Exon v2; Agilent Technologies Inc., Wilmington, DE, USA), following the manufacturer's guidelines. This method is designed to target all human exons, regions totaling approximately 44 Mb, covering 98.2% of the Consensus Coding Sequence (CCDS) database. For the capture, a DNA-shearing instrument (focused-ultrasonicator; Covaris Inc., Woburn, MA, USA) was used to randomly fragment the pure and high molecular weight genomic DNA samples (experiments carried out by BGI-Shenzhen, Shenzhen, China), resulting in DNA fragments with a base-pair peak of 150 to 200 bp. Adaptors were then ligated to both ends of the resulting fragments. The adaptor-ligated templates were purified by magnetic beads (Agencourt AMPure SPRI; Beckman Coulter Inc., Brea, CA, USA), and fragments with an insert size of approximately 250 bp were excised. Extracted DNA was amplified by ligation-mediated (LM)-PCR, purified, and hybridized (SureSelect Library; Agilent Technologies) for enrichment. Hybridized fragments bound to the strepavidin beads, whereas the unbound non-hybridized fragments were washed out after 24 hours of hybridization. Captured LM-PCR products were analyzed using a microfluidics-based platform (2100 Bioanalyzer; Agilent Technologies) to estimate the magnitude of the enrichment. Paired-end sequencing was performed using a sequencing platform (HiSeq2000; Illumina Inc., San Diego, CA, USA) with average read lengths of 90 bp. Raw image files were processed (Pipeline version 1.6; Illumina Inc.) for base-calling, using the default parameters. FASTQ files were produced from the pipeline for downstream sequence data analysis. A gender check was compatible with the known genders of the collected human participants.

1.2.5 SNP arrays

DNA samples were genotyped on the SNP arrays (Human610-Quad, version 1; Illumina Inc.) with approximately 610,000 markers (including approximately 20,000 non-polymorphic markers) at the Center for Applied Genomics (Children's Hospital of Philadelphia, Philadelphia, PA USA). Total genomic DNA extracted from whole blood was used in the experiments. Standard data-normalization procedures and canonical genotype-clustering files provided by Illumina were used to process the genotyping signals. Concordance between SNPs from the arrays and SNPs from exome sequencing was determined by calculating the percentage of variants from exome sequencing and comparing this with the same genotype derived from the SNP arrays.

1.2.6 Alignment and variant calling

BWA-GATK variant calling

Burrows-Wheeler aligner (BWA; version 0.5.9[160]) was used to align the sequencing reads, with default parameters, to the human reference genome sequence GRCh37. Alignments were converted from sequence alignment map (SAM) format to sorted, indexed binary alignment map (BAM) files (SAMtools version 0.1.18; sourceforge.net). The Picard tool was used to remove duplicate reads. GATK software tools (version 1.5; www.broadinstitute.org) were used for improvement of alignments and genotype calling and refining with recommended parameters[57]. BAM files were re-aligned with the GATK IndelRealigner, and base quality scores were re-calibrated by the GATK base quality recalibration tool. Genotypes were called by the GATK UnifiedGenotyper, and the GATK VariantRecalibrator tool was used to score variant calls by a machine-learning algorithm and to identify a set of high-quality SNPs using the Variant Quality Score Recalibration (VQSR) procedure. GATK was used to filter high-quality insertions and deletions (indels) by hard criteria, 'QD < 2.0, ReadPosRankSum < -20.0 FS > 200.0'. Finally, we removed SNVs and indels located outside of regions targeted by exome capture. To increase sensitivity, only those indels with depth (DP) of 10 or more, and with more than 4 reads supporting the indel events were included in the final high-confidence indel set. At a later date, one exome was processed with newer versions of the GATK v2.3-9 UnifiedGenotyper and GATK v2.3-9 HaplotypeCaller modules.

BWA-SAMtools genotype calling

Using the above BAM files, we used SAMtools (version 0.1.18) to generate genotype calls[161]. The 'mpileup' command in SAMtools was used to identify SNPs and indels, and we removed variants with DP coverage less than 10, and variants located outside of exome-capture regions.

SOAP pipeline

Adaptor and low-quality sequences were removed before mapping. Sequence reads identified from each individual were then aligned to human reference genome GRCh37 using SOAPaligner (version 2.21[164]) with a maximum of five mismatches. Duplicate reads were removed. Consensus genotypes in target regions were called by SOAPsnp (version 1.03)[162] with recommended parameters. SNV results were filtered (Phred-like SNV quality ≥20, overall depth 8 to 500, copy number estimate <2, and distance between two adjacent SNVs ≥5). For a heterozygous SNV, the quality of the minor allele was required to be at least 20, depth of coverage for the minor allele at least 4, and the ratio of major allele to minor allele less than 5. For indel calling, SOAPindel was used, which adopts local assembly based on an extended de Bruijn graph[165]. For SOAPindel, the aligner BWA was used to align the reads to the human reference sequence with default parameters. Initially, putative indels were assumed to be located near the unmapped reads whose mates mapped to the reference genome. SOAPindel then executed a local assembly (k-mer=25) on the clusters of unmapped reads. Clusters with coverage of less than 5 were not used. The assembly results were aligned to the reference in order to find the potential indels. To distinguish true-positive and false-positive indels, SOAPindel generates Phred quality scores, which take into consideration the depth of coverage, indel size, number of neighboring variants, distance to the edge of the contig, and position of the second different base pair. Only those indels with a quality score of 10 or higher were retained in the final indel call set.

GNUMAP pipeline

Diploid and monoploid SNVs for each individual were called using the GNUMAP pipeline (version 3.1.0[43]). GNUMAP-SNP utilizes a novel probabilistic pair-hidden Markov model, which accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion. Raw reads were initially aligned to the full genome using an alignment score of 260 or greater, which for this dataset allowed for only one SNV per read. A k-mer size of 12 and a jump size of 10 were also used. Only SNVs within exome regions with a $P<0.001$ were reported. The GNUMAP pipeline cannot currently call indels.

BWA-SNVer pipeline

BWA[160] was used to align the sequencing reads to GRCh37 with default parameters. Duplicate reads were removed by Picard, and SNVer (version 0.2.1) was then used for detecting SNVs in each sample[307]. Similar to GATK[57], only the mapped short reads with mapping quality of greater than 20 were considered, and only bases with base quality greater than 17 counted. SNVer estimated the empirical error rate for those selected reads in making variant calls. We set the number of haploids to 2 for analysis of individual samples, and set the variant allele frequency threshold of greater than 0 for detecting both rare and common SNVs. SNVer provides multiplicity control, and we performed Bonferroni correction and controlled the family-wise error rate at the 0.05 level to report identified SNVs. Indels cannot currently be called by the BWA-SNVer pipeline.

1.2.7 Post-variant calling analyses

Post-variant-calling analyses were performed using Golden Helix SVS (version 7.6.10[289], ANNOVAR[306], the R suite of statistical programming tools (www.r-project.org), and custom Perl scripts.

1.2.8 MiSeq sequencing for validation

Validation variants were randomly selected from sets of particularly controversial variants, indels and SNVs unique to GATK, indels and SNVs unique to SOAP, and variants (both SNVs and indels) shared by these two pipelines. PCR primers were designed using the software program Primer 3 (sourceforge.net), to produce amplicons (ranging in size from 100 to 200 bp) containing variants of interest in approximately the center of the amplicon. Primers were obtained in 96-well plate format, 10 μmol/L dilution each (Sigma-Aldrich, St Louis, MO, USA). All primers were first tested for PCR efficiency using a HAPMAP DNA sample (Catalog ID NA12864l Coriell Institute for Medical Research, Camden, NJ, USA) and DNA polymerase (LongAmp® Taq DNA Polymerase; New England Biolabs, Beverly, MA, USA). k8101-49685 genomic DNA was used as template for the validation experiment. After quality-control steps using agarose gel, the product was purified (ExoSAP-IT® reagentsl Affymetrix Inc., Santa Clara, CA, USA) and pooled. Final PCR products were quantified (Qubit® dsDNA BR Assay Kitl Invitrogen Corp., Carlsbad, CA, USA), then library construction for the sequencer platform (MiSeq Personal Sequencer; Illumina Inc.) was performed. Finally, before being loaded onto the MiSeq machine, the quality and quantity of the sample was verified using the Bioanalyzer (Agilent Technologies) and quantitative PCR (Kapa Biosystems Inc., Woburn, MA, USA).

1.2.9 Predictive models from validation data

MiSeq validation data were also use to build predictive models of false negative sequence variation. Note that for this part of the study, MiSeq validation data from two additional publications[73,209] were used. To build predictive models that can distinguish between true negative and false negative sequence variation, training data sets that contain both types are required.

Isolating false negative sequence variants from K8101 exome sequencing data

In order to isolate false negative sequence variants, SNV and INDEL calls were generated using the GATK HaplotypeCaller and the FreeBayes caller from K8101 raw exome sequence data. GATK calls were generated as described above, using the GATK HaplotypeCaller version 3.5. FreeBayes calls were generated by first aligning the raw sequence data to the hg19 human reference sequence using bwa mem version 0.7.10-r789 with default parameters. Duplicate reads were then marked using Picard MarkDuplicates version 2.0.1, and reads were realigned using the GATK IndelRealigner. SNV and INDEL calls were generated using FreeBayes version 1.0.2 with default parameters.

Once SNV and INDEL calls were generated with both the GATK HaplotyperCaller and FreeBayes using the same raw K8101 exome sequencing data, variants that were validated via MiSeq re-sequencing but were unique to each caller were isolated. Note that if a variant was validated but was unique to one caller, this variant represents a false negative calls with respect to the other caller. False negative SNV and INDEL calls were collected for both the GATK and the FreeBayes variant caller.

Isolating true negative sequence variants from K8101 exome sequencing data

Binary classification models work most efficiently if both classes have a similar set of predictors. For the false negative call sets, both GATK and FreeBayes return variant call sets that have a number of informative predictor variables for each variant. However, non-variant sites do not contain the same number or types of characteristic predictor variables. In order to generate a true negative call set with the same predictor variables as the false negative call set, a new procedure was devised. The following procedure to generate true negative call sets with informative predictor variables was performed in the same way for both SNV and INDEL calls, for both GATK and FreeBayes call sets. For each false negative call set, a new reference sequence was generated. For each new reference sequence, genomic locations several bases up or downstream of false negative loci were edited away from the reference base. For each false negative call set and its respective edited reference sequence, raw exome sequencing data for K8101 was then realigned and variants were called as described above. The results should include initial full exome call set, plus additional variants corresponding to locations that were edited in the reference sequence. These ’variant’ sites are in fact true negatives, but now are characterized with the full predictor set, similar to the false negative call set.

Building predictive models using a variety of machine learning techniques

Using false negative and true negative call sets, several predictive models were built using 5 different machine learning techniques. These include, a random forest classifier, a support vector machine classifier, a logistic regression, a k-nearest neighbors classifier and a gradient tree boosting classifier. Models were trained using 10-fold cross validation, and hyperparameters were chosen among the most performant sets among 20,000 models, which were generated using a randomized grid search approach. Models were evaluated by assessing their area under the receiver operator characteristic (AUROC) curve.

A Phylogenetic method of extracting false negative sequence variants

The above approach to building models capable of classifying false negative sequence variation relies on and leverages the existence of a large validation data set for one human sample. In the absence of such a data set, another approach can be taken. Henn et al. 2016 describes a phylogenetically motivated approach for detecting and characterizing false negative sequence variation in human mitochondrial sequence. Briefly, test samples are assigned to a haplogroup on the mtDNA phylogeny[298] using the haplogrep algorithm[144], and the absence of haplogroup-defining sequence variation in the sample is assumed to be the result of a false negative call. This approach was applied to 1072 human samples taken from the 1000 genomes phase 3 project[292], and a large false negative set was generated. True negatives were generated in the same way as describe above. Note that for this set, it was only possible to generate false and true negative sets for SNVs, due to the nature of the human mtDNA phyologeny. In addition, since many false negatives were not present in any form in the original call set, the false negative set here is derived from those sequence variants that were marked as false negative by the phylogenetic method, but were indeed present in pre-filtered sets. Predictive models were built and optimized as previously described for the MiSeq set.

1.2.10 Accessing data

The data used in the analyses performed in this study can be found on the Sequence Read Archive. Accession numbers SRS402291 and SRS402299 correspond to the 15 exomes and the single whole genome analyzed during the course of this study.

1.3 Results

1.3.1 Data production summary

Fifteen DNA samples from four different families (Figure 1.1) were prepared by exon capture (Agilent 44 MB SureSelect protocol; Agilent Technologies), followed by sequencing on (HiSeq2000; Illumina Inc.). On average, we obtained sequence coverage of approximately 120X (range, 100 to 154X) on targeted regions for these 15 samples. For all samples, sequence reads covered more than 80% of the targeted region with a depth of greater than 20 reads per base (Figure 1.2; Table 1.1). Five different pipelines were used for read alignment and variant calling (SNVs and indels when possible) (Table 1.2). In addition, one whole genome was sequenced and analyzed by CG with 95% of the exome region covered by 20 reads or more per base, resulting in greater than 88% of the genome covered with a depth of greater than 20 reads per base. Variant calls were generated by CG with their in-house analysis pipeline (version 2.0).

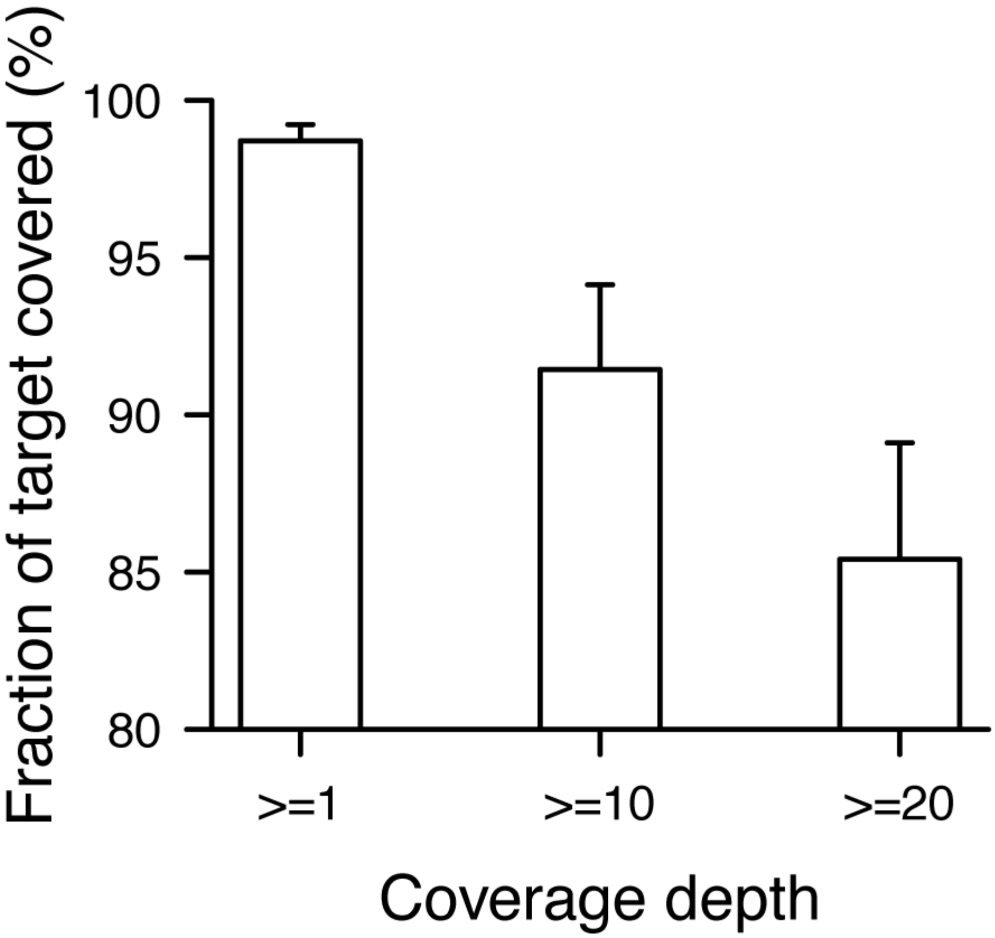


Figure 1.2: Fraction of target capture region covered versus coverage depth for 15 exomes. All exomes have at least 20 reads or more per base pair in > 80% or more of the 44 MB target region.

Exome Capture Statistics	K14349-68182	K14349-70463	K14349-70850-A	K26679-88458	K26679-88459	K26679-88460	K26679-88461	K26679-89588	K8101-49685s	K26679-89587	K26679-91583	K24510-84060	K24510-92157-a	K24510-84615	K24510-88962
Target region (bp)	46,401,121	46,401,121	46,257,379	46,401,121	46,257,379	46,257,379	46,257,379	46,401,121	46,401,121	46,401,121	46,257,379	46,401,121	46,401,121	46,401,121	46,257,379
Raw reads	107,947,768	119,004,986	145,676,240	116,262,804	119,560,838	113,339,390	117,568,124	127,333,280	113,617,288	172,589,226	107,896,114	138,779,950	161,898,170	156,985,870	104,423,704
Raw data yield (Mb)	9,715	10,711	13,112	10,464	10,760	10,200	10,581	11,460	10,225	15,533	9,711	12,490	14,571	14,129	9,398
Reads mapped to genome	90,351,872	100,229,650	98,761,126	95,893,837	95,239,735	94,539,914	95,595,560	102,666,713	93,112,728	134,626,098	79,566,261	110,160,277	135,603,094	135,087,576	83,942,646
Reads mapped to target region	64,588,670	73,432,323	72,237,430	70,257,106	70,448,954	68,513,496	69,533,715	75,684,755	67,891,089	94,864,043	67,033,232	68,042,793	84,379,239	80,347,146	61,207,116
Data mapped to target region (Mb)	4,821.52	5,478.11	5,489.86	5,253.57	5,306.94	5,139.26	5,247.46	5,692.83	5,103.67	7,160.14	5,348.42	5,337.69	6,647.18	6,280.01	4,614.47
Mean depth of target region	103.91	118.06	118.68	113.22	114.73	111.1	113.44	122.69	109.99	154.31	115.62	115.03	143.25	135.34	99.76
Coverage of target region (%)	0.9870	0.9878	0.9814	0.9860	0.9826	0.9819	0.9862	0.9819	0.9816	0.9914	0.9907	0.9948	0.9947	0.9954	0.9828
Average read length (bp)	89.81	89.83	89.71	89.82	89.72	89.78	89.75	89.79	89.80	89.67	89.23	89.91	89.92	89.95	89.75
Rate of nucleotide mismatch (%)	0.31	0.30	0.38	0.36	0.31	0.31	0.40	0.33	0.31	0.39	0.43	0.31	0.24	0.24	0.39
Pct. of target covered >=4X	95.90	96.12	94.03	95.86	94.99	94.83	95.04	94.97	94.73	96.70	96.99	98.17	98.38	98.47	94.25
Pct. of target covered >=10X	91.60	92.11	87.19	91.75	90.41	90.17	88.92	90.47	89.95	90.76	93.28	95.18	95.90	95.97	87.90
Pct. of target covered >=20X	85.87	86.86	80.00	86.35	84.38	83.93	81.97	84.61	83.59	81.88	87.56	90.12	91.62	91.75	80.70
Pct. of target covered >=30X	80.40	81.94	74.35	81.27	78.67	78.10	76.47	79.13	77.61	75.74	81.85	84.98	87.42	87.67	74.69
Capture specificity (%)	71.68	73.48	73.56	73.54	74.12	72.63	73.00	73.94	73.08	70.80	84.52	61.52	62.12	59.25	73.16
Reads mapped to flanking region	12,559,590	13,375,864	11,103,165	12,699,450	10,877,252	12,255,477	11,909,544	12,505,048	11,470,864	17,762,776	2,788,337	4,950,181	5,747,450	6,563,002	11,195,600
Mean depth of flanking region	29.99	33.22	28.55	31.41	28.60	30.15	29.57	31.78	28.86	41.81	12.93	17.69	21.04	22.16	26.91
Coverage of flanking region (%)	95.77	95.79	85.57	95.42	93.81	94.48	91.60	94.19	94.06	92.13	71.16	89.50	89.95	93.71	90.75
Pct. of flanking covered >=4X	84.97	84.92	69.67	84.18	80.38	82.71	75.64	82.13	81.38	77.77	44.95	63.89	66.69	73.05	74.97
Pct. of flanking covered >=10X	66.76	67.24	52.03	65.96	60.96	64.74	57.56	64.19	62.54	60.91	28.17	39.97	43.09	47.84	56.71
Pct. of flanking covered >=20X	46	48	36	46	42	45	40	45	43	46	17	25	28	31	39
Pct. of flanking covered >=30X	33.23	35.77	27.27	33.96	30.10	32.51	30.03	33.19	30.88	36.65	11.83	16.91	19.93	21.67	28.12
Pct. of unique mapped bases	84.57	85.96	83.87	85.88	84.78	84.63	84.53	85.25	84.52	82.99	87.67	65.59	65.98	63.69	85.46
Duplication rate	10.29	9.73	26.79	10.50	12.81	9.22	10.39	12.02	11.05	12.17	10.70	8.90	5.29	3.72	11.99
Mean depth of chrX	74.30	85.75	163.58	81.01	160.60	154.55	158.91	88.94	78.61	107.46	152.32	80.73	100.02	94.71	140.29
Mean depth of chrY	112.83	129.38	-	125.45	-	-	-	142.76	107.81	147.29	-	117.22	151.43	140.53	-
GC rate	45.01	44.85	49.67	45.14	43.42	43.32	48.21	43.29	43.24	50.55	50.00	47.97	48.06	47.18	48.19
Gender test result	M	M	F	M	F	F	F	M	M	M	F	M	M	M	F

Table 1.1: Data production statistics from 15 sequenced exomes

Pipeline name	Alignment method	Variant-calling module	Description of variant-calling algorithm
SOAP	SOAPaligner version 2.21/BWA version 0.5.9	SOAPsnp version 1.03 SOAPindelversion 2.01	SOAP uses a method based on Bayes' theorem to call consensus genotype by carefully considering the data quality, alignment, and recurring experimental errors
GATK version 1.5	BWA version 0.5.9	UnifiedGenotyper version 1.5	GATK employs a general Bayesian framework to distinguish and call variants. Error correction models are guided by expected characteristics of human variation to further refine variant calls
SNVer version 0.2.1	BWA version 0.5.9	SNVer Individual version 0.2.1	SNVer uses a more general frequentist framework, and formulates variant calling as a hypothesis-testing problem
GNUMAP version 3.1.0	GNUMAP version 3.1.0	GNUMAP version 3.1.0	GNUMAP incorporates the base uncertainty of the reads into mapping analysis using a probabilistic Needleman-Wunsch algorithm
SAMtools version 0.1.18	BWA version 0.5.9	mpileup version 0.1.18	SAMtools calls variants by generating a consensus sequence using the MAQ model framework, which uses a general Bayesian framework for picking the base that maximizes the posterior probability with the highest Phred quality score.

Table 1.2: A descriptive summary of the variant calling pipelines included in the comparative analyses.

1.3.2 SNV analysis

Concordance with SNP genotyping arrays

Sensitivity and specificity for detecting common SNPs was calculated for each Illumina variant-calling pipeline for four samples that were genotyped with the Illumina Human610-Quad version 1 SNP array (see Tables 1.3 and 1.4). We caution that this analysis was restricted to a set of common SNPs targeted by the SNP array, and that these tend to be within regions containing little to no repeated sequences and without extreme GC contents. Therefore, although widely used in published literature, concordance with SNP arrays does not adequately measure real-world performance on all variants in personal genomes. With this major caveat in mind, performance for each pipeline was measured by treating the Illumina Human610-Quad version 1 SNP arrays as a true-positive reference, and comparing the exome-capture sequencing results with this reference set. The average specificity for each of the five Illumina pipelines was generally high, ranging from 99.59% to 99.87% (Table 1.4), consistent with the fact that each of these pipelines have been optimized to minimize false negatives for known common SNPs. The average sensitivity ranged among the five pipelines from 86.6% (with GNUMAP) to 95.3% (with GATK1.5). Sensitivity decreased when the variant set was iteratively restricted to the intersection between two or more variant-calling pipelines, whereas specificity naturally shows the opposite trend of increasing values under the same series of intersections (Table 1.4).

Evaluation of performance by inheritance analysis

To explore the validity of SNVs called by each Illumina pipeline, we performed an inheritance analysis for two families contained within the 15 sequenced exomes. Previous calculations have estimated the average expected number of de novo non-synonymous coding mutations

Sample	Software	Compared Sites	Concordance Sites	Concordance rate
Mother-1	SOAPsnp	6088	6074	99.77%
	GATK 1.5	6249	6224	99.60%
	SNVer	5723	5708	99.74%
	GNUMAP	5458	5434	99.56%
	SAMTools	5885	5848	99.37%
Son-1	SOAPsnp	6366	6353	99.80%
	GATK 1.5	6341	6323	99.72%
	SNVer	6255	6239	99.74%
	GNUMAP	5850	5828	99.62%
	SAMTools	6383	6362	99.67%
Son-2	SOAPsnp	6412	6401	99.83%
	GATK 1.5	6426	6413	99.80%
	SNVer	6336	6325	99.83%
	GNUMAP	5906	5889	99.71%
	SAMTools	6477	6450	99.58%
Father-1	SOAPsnp	6247	6238	99.86%
	GATK 1.5	6304	6288	99.75%
	SNVer	6205	6192	99.79%
	GNUMAP	5805	5786	99.67%
	SAMTools	6344	6327	99.73%

Table 1.3: Concordance rates with common SNPs genotyped on Illumina 610K genotyping chips; all pipelines are very good with identifying already known, common SNPs.

	Sensitivity		Specificity	
	Mean*	SD	Mean*	SD
SOAPsnp	94.68	2.26	99.79	0.03
GATK1.5	95.34	1.16	99.72	0.08
SNVer	92.33	4.4	99.78	0.04
GNUMAP	86.6	3.23	99.64	0.06
SAMtools	94.47	4.22	99.59	0.16
Any pipeline	97.67	1.2	99.62	0.11
≥ 2 pipelines*	96.64	2.28	99.69	0.07
≥ 3 pipelines*	95.62	3.13	99.73	0.05
≥ 4 pipelines*	92.6	3.4	99.82	0.04
5 pipelines*	80.58	5.26	99.87	0.01

Table 1.4: Sensitivity and specificity was calculated for each pipeline by comparing Illumina Human610-Quad version 1 SNP arrays with exome-capture sequencing results, based on the four samples whose genotyping data was available. *Intersection of variants contained in the number of pipelines specified.

per individual exome to be approximately 1 to 2[217,215,216,257]. However, we found that the number of putative de novo mutations per child per exome was much higher if only the parents of the child were used to filter out inherited mutations. Adding an additional familial generation to the filtering process, in our case a grandparent, significantly reduced the number of putative de novo variants to a value comparable with that of the previously reported value of expected de novo non-synonymous mutations. In addition, significant variation was seen in the number of putative de novo mutations between the two families (Table 1.5), consistent with previous findings[46].

Variant-calling pipeline concordance

SNV concordance between all 5 Illumina pipelines across all 15 exomes was 57.4% on average, and Ti/Tv ratios showed a generally increasing trend for sets of variants intersected by an increasing number of variant-calling pipelines (Figure 1.3). We found that for novel SNVs (those not found in dbSNP135) the overall concordance (11.4%) was much lower than the overall concordance between known SNVs (59.6%) (Figure 1.3). In a previous paper, we validated with Sanger sequencing or Sequenom genotyping 17 SNVs found in 3 of the current pilot samples[179]. Of these 17 validated SNVs, 16 were detected by all 5 pipelines, and the remaining variant was called by 4 of the 5 pipelines. Additional validation analyses are presented later in this paper.

A more detailed analysis of SNVs of one sample (k8101-49685) revealed that the exome variant calls had moderate to high depth of coverage (Figure 1.4). The range of read depths along with read-depth uniformity of variant calls varied between pipelines (Figure 1.4). Overall concordance between all five pipelines for sample k8101-49685 was 57.5%; however, subsetting variants called by Illumina pipelines using increasingly stringent read-depth thresholds did not increase SNV concordance (Figure 1.5), and overall concordance was lowest when

	Number of putative de novo coding non-synonymous or nonsense SNVs detected	
Family 1	Without using the grandparents as a filter	Using the grandparents as a filter
Child A	241	1
Child B	211	0
Child C	102	6
Child D	242	3
Family 2		
Child A	49	NA[a]
Child B	41	NA[a]

Table 1.5: De novo single-nucleotide variants (SNVs) were detected in two families contained within the 15 study exomes. Family 1 had a grandparent available for filtering purposes, whereas family 2 did not. To minimize false positives in the pool of SNVs associated with each child, only highly concordant SNVs were used (SNVs detected by all five pipelines). To construct a comprehensive set of SNVs for each parent, and hence increase filtering accuracy, false negatives for parent SNVs were reduced by taking the union of all SNV calls from all five pipelines. [a]N/A, no grandparent available.

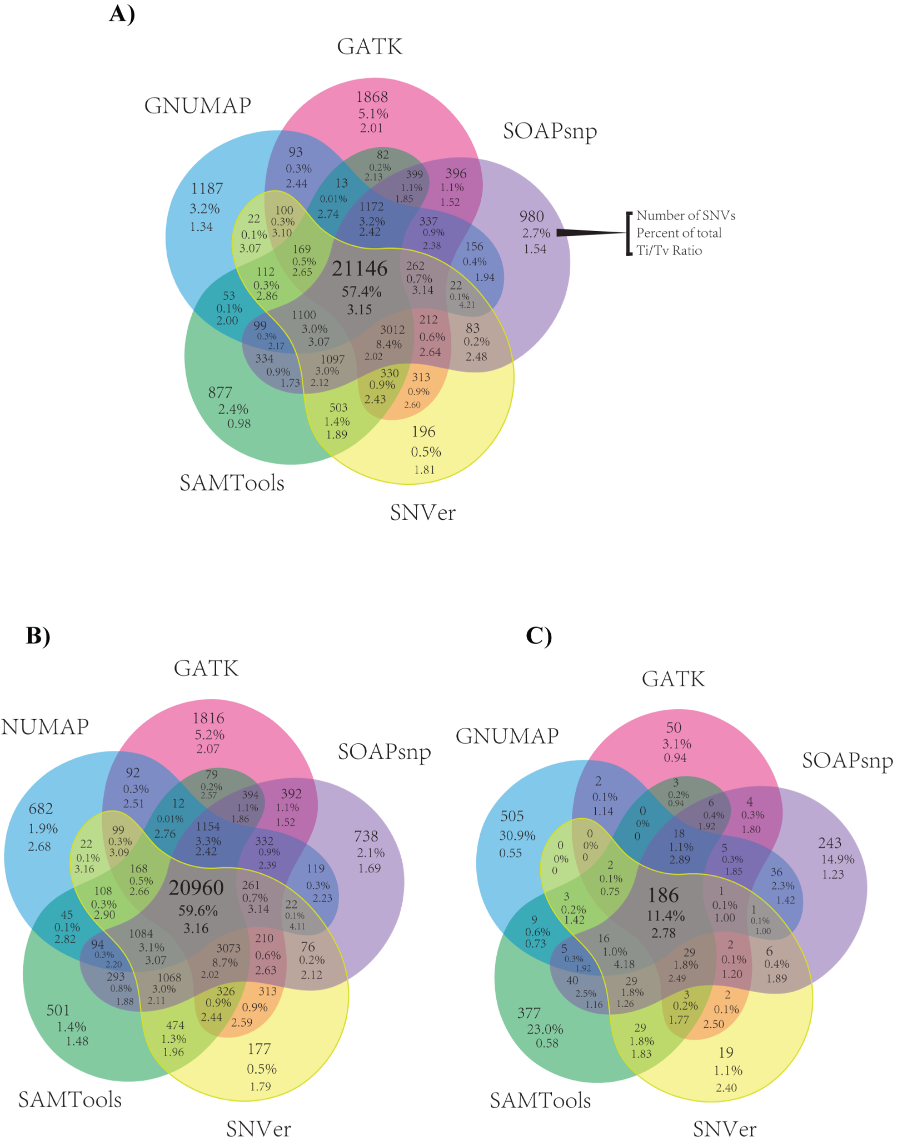


Figure 1.3: Mean single-nucleotide variants (SNV) concordance over 15 exomes between five alignment and variant-calling pipelines. The alignment method used, followed by the SNV variant calling algorithm is annotated here in shorthand: BWA-GATK, SOAP-Align-SOAPsnp, BWA-SNVer, BWA-SAMtools, and GNUMAP-GNUMAP. (**A**) Mean SNV concordance between each pipeline was determined by matching the genomic coordinate as well as the base-pair change and zygosity for each detected SNV. (**B**) The same analysis as in (A) but filtered to include only SNVs already found in dbSNP135. (**C**) The same analysis as in (A), but filtered to include novel SNVs (that is, SNVs not found in dbSNP135).

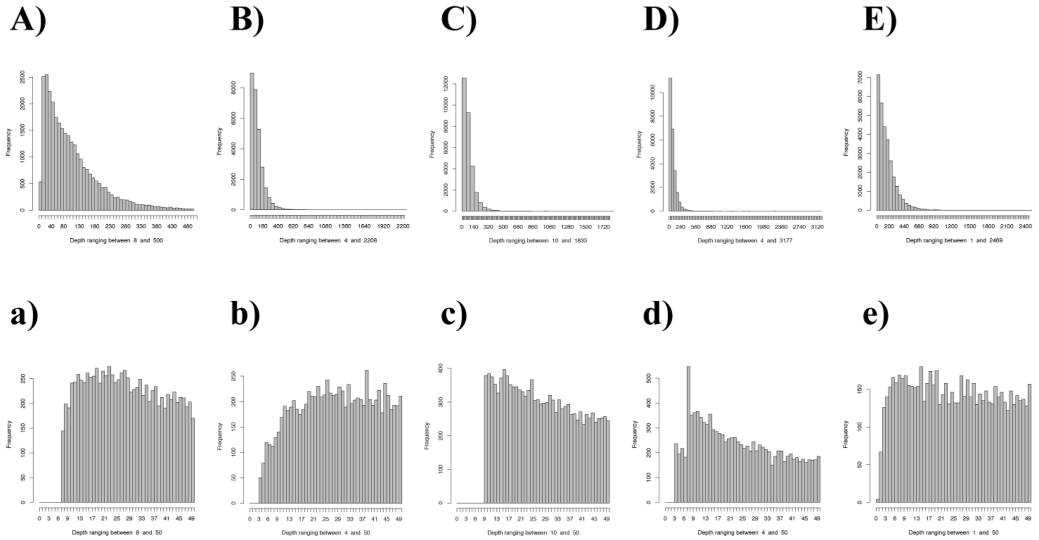


Figure 1.4: Histograms of Illumina read depth at SNV coordinates. Read depth taken from each pipeline's independently aligned BAM file at genomic coordinates of SNVs called by each of the 5 alignment and variant calling pipelines. A) SOAPsnp, B) SNVer, C) SAMtools, D) GNUMAP and E) GATK, respectively. Frequency of read depths for all SNVs (A, B, C, D, and E) as well as for SNVs having depths between 0 and 50 (a, b, c, d, and e) were plotted.

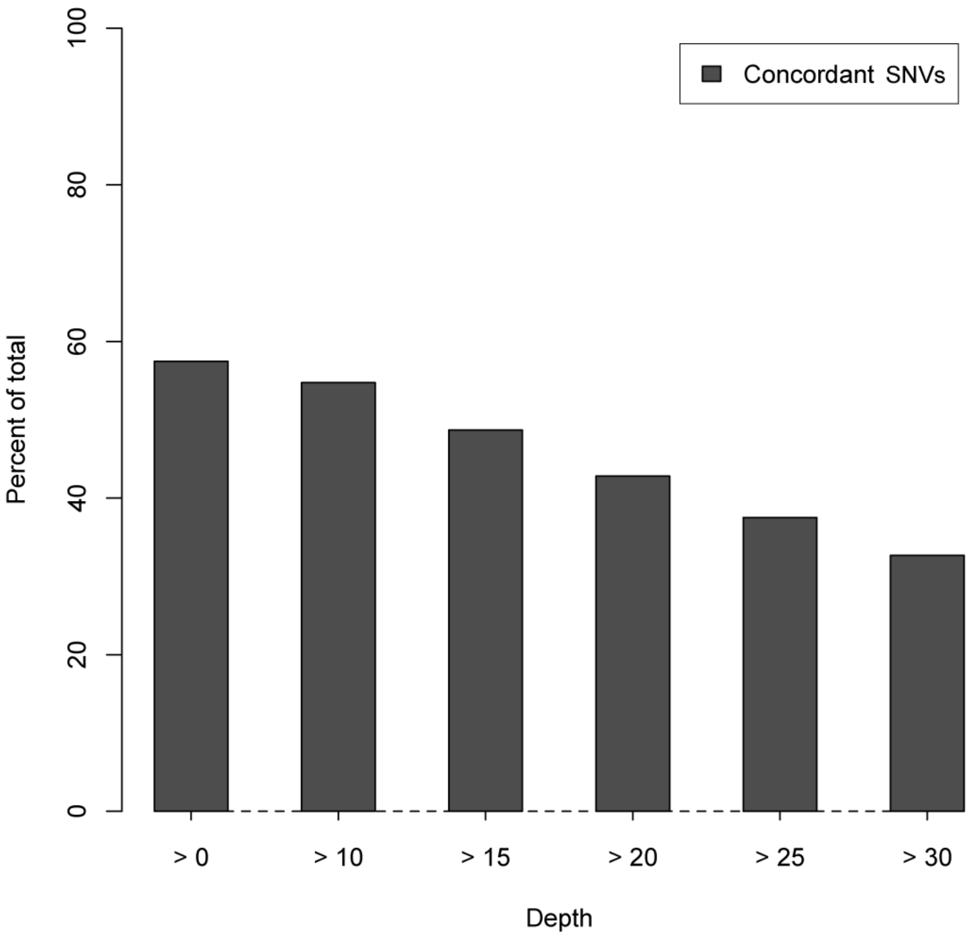


Figure 1.5: SNV concordance measured at varying Illumina read depth threshold values. SNV concordance for a single exome, ”k8101-49685”, between five alignment and variant detection pipelines: GATK, SOAPsnp, SNVer, SAMtools, and GNUMAP. Concordance between each pipeline was determined by matching the genomic coordinate as well as the base pair change and zygosity for each detected SNV. Concordance was measured at varying Illumina read depth threshold values in each independently aligned BAM file, ranging from > 0 (no threshold) to > 30 reads.

read-depth threshold values were at their highest (32.7% concordance when depth was required to be greater than 30 supporting reads).

Sequencing platform concordance

For sample k8101-49685, we selected variants generated by the CG pipeline that fell within the exon-capture regions of the Agilent SureSelect version 2 capture kit. We found that of the 21,050 SNVs identified by CG and located within the UCSC refGene regions, 19,407 (92%) were also within regions targeted for capture by the Agilent SureSelect version 2 kit. Of these, 2,085 (11%) were not called by any of the Illumina-based exome-analysis pipelines, despite computed high mappability scores for these variants (Figure 1.6)[159]. Of these 2,085 SNVs uniquely called by CG, an average of 558 had no sequence coverage as mapped by any of the Illumina-based exome-analysis pipelines. The Illumina exome read-depth for the remaining 1,527 CG-unique SNVs was calculated, and the majority of these SNVs were found to be in regions of very low Illumina sequence coverage (<20 reads) in the exome data sets (Figure 1.7).

We found that 89.3% of CG SNVs (17,322 of 19,407) were contained within the union of all five Illumina pipelines (35,653 putative SNVs), whereas 18,331 of these 35,653 putative Illumina SNVs were not called by CG, suggesting a high false positive rate in the union of the Illumina calls and/or conversely a high false-negative rate in the CG calls (Figure 1.6). Overall concordance displayed marginal increases when VQLOW SNVs (low-quality CG variants) were removed from the pool of CG SNVs (Figure 1.6). Overall concordance remained stable as the depth of coverage threshold value associated with Illumina data calls increased (Figure 1.8A).

When only highly concordant Illumina SNVs (SNVs called by all five Illumina pipelines) were compared with the CG SNVs, only 64.4% (12507) of CG SNVs were contained within

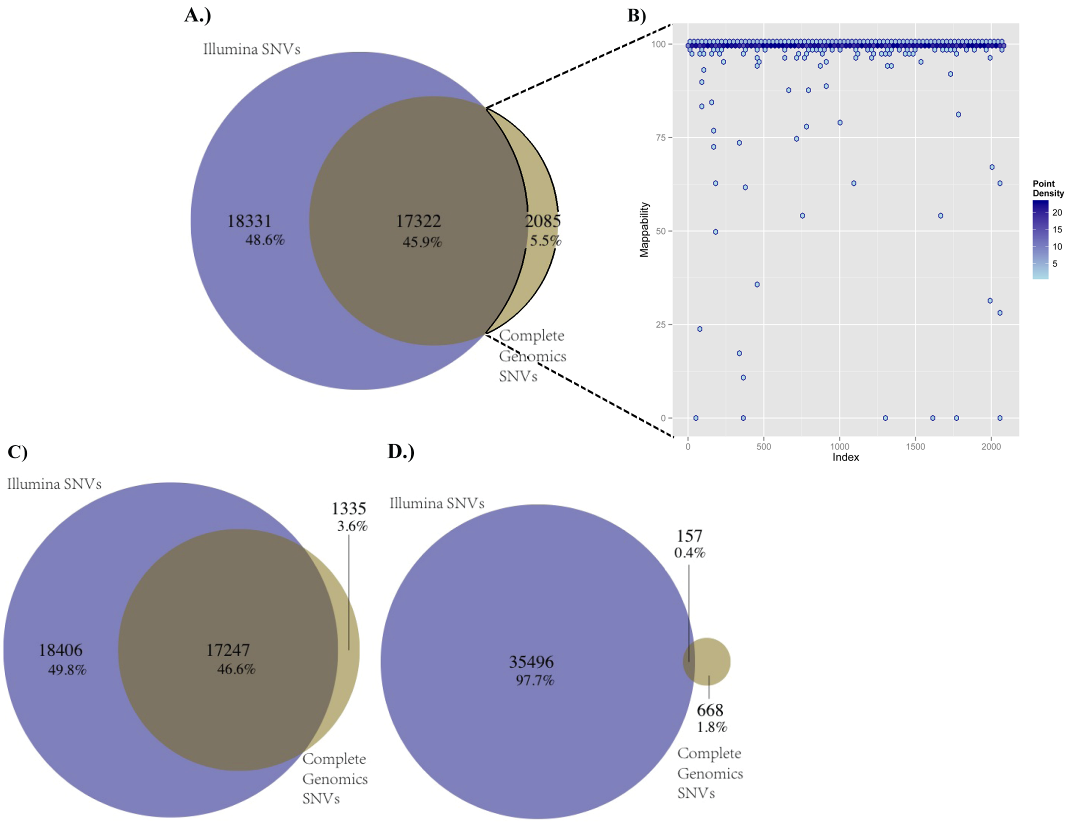


Figure 1.6: Single-nucleotide variant (SNV) concordance, between two sequencing pipelines (Illumina and Complete Genomics (CG)) for a single exome, k8101-49685. For the Illumina sequencing, exons were captured using the Agilent SureSelect version 2 panel of capture probes. CG SNVs consisted of a subset of all SNVs called by CG that fell within the Agilent SureSelect version 2 exons. Concordance was determined by matching the genomic coordinates, base-pair composition, and zygosity status for each detected SNVs. Illumina SNVs consisted of all SNVs (the union) called by the five variant-calling pipelines GATK, SAMtools, SOAPsnp, SNVer, and GNUMAP, which increased the false positives but decreased the false negatives. Concordance was measured between Illumina SNVs and (**A**) all CG SNVs, (**C**) only high-quality (VQHIGH) CG SNVs, and (**D**) only low quality (VQLOW) CG SNVs. (**B**) Genome mappability analyses were performed on 2,085 discordant SNVs, which were found by the CG pipeline and not found by any of the five Illumina data pipelines.

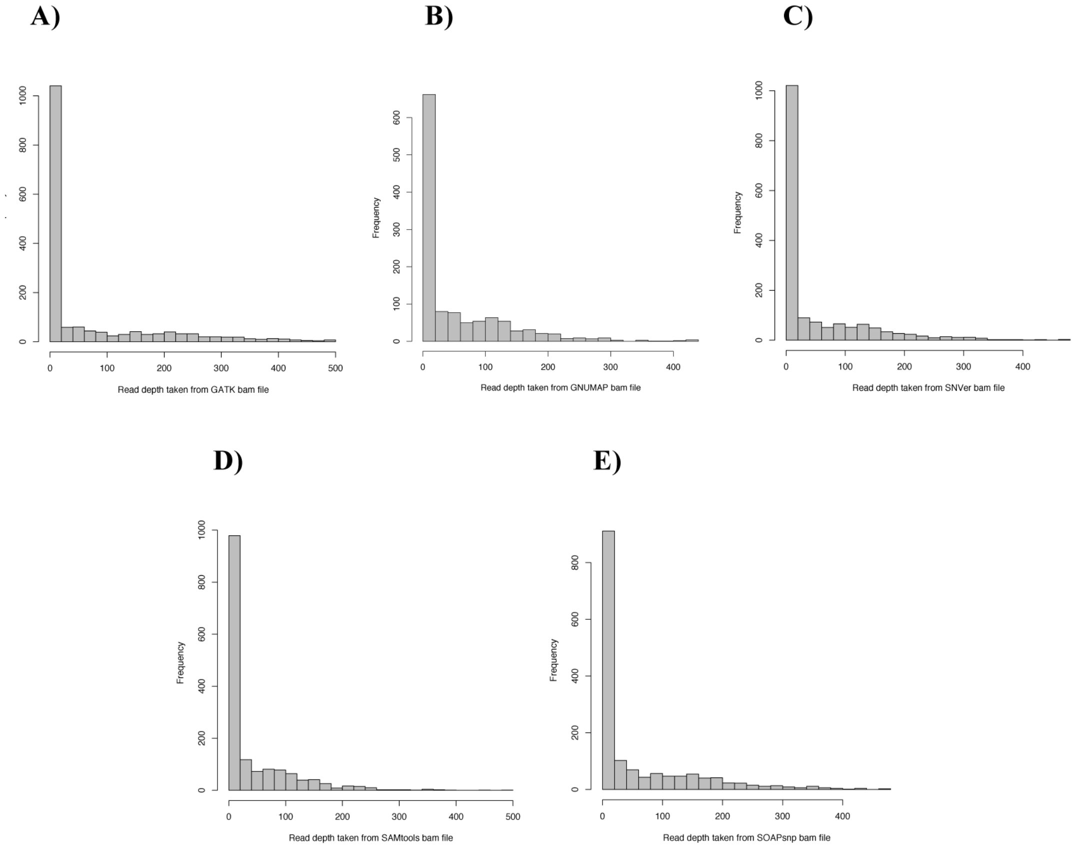


Figure 1.7: Histograms of illumina read depth at genomic coordinates of the unique to Complete Genomics SNV calls. Histograms of read depth taken from each of the five Illumina pipeline's independently aligned BAM file at genomic coordinates of SNVs that were found by Complete Genomics but not by any of the 5 Illumina pipelines: GATK, GNUMAP, SNVer, SAMtools and SOAPsnp, A, B, C, D and E respectively. All coordinates fell within the range of the Agilent SureSelect v.2 exons.

the concordant Illumina set, suggesting a high false-negative rate in this highly concordant Illumina set. Overall agreement decreased as the depth of coverage threshold value for Illumina calls increased, consistent with an increasing false-negative rate (Figure 1.8B).

Cross-platform comparison of unique-to-pipeline SNVs

SNVs from sample k8101-49685 that were uniquely detected by only one of the five Illumina variant-calling pipelines were compared with SNVs called by CG (Figure 1.9). Of the SNVs uniquely called by GATK, 809 of 1671 (48%) were concordant with CG data. The concordance was much lower for the other four pipelines, 49 of 1,102 SNVs (4%) for GNUMAP, 45 of 886 (5%) for SAMtools, 29 of 226 (12%) for SNVer, and 24 of 908 (3%) for SOAPsnp. Concordance improved for SNVs that were called by more than a single Illumina data pipeline, and the concordance was the highest for variants found by all five Illumina pipelines (Figure 1.9).

For variants that were novel as well as unique to a single Illumina pipeline, concordance with CG data was low (see Additional file 1, Figure S7). For GATK, 25% (13 of 51) of novel and unique-to-pipeline SNVs were concordant with CG data; for GNUMAP and SOAPsnp, no novel and unique-to-pipeline SNVs were concordant (0 of 470 and 0 of 229 respectively); for SAMtools, 0.2% (1 of 418) of novel and unique-to-pipeline SNVs were concordant; and for SNVer, 6% (1 of 18) of novel and unique-to-pipeline SNVs were concordant. Concordance rates of novel and unique-to-pipeline SNVs increased for variants called by an increasing number of pipelines (Figure 1.9).

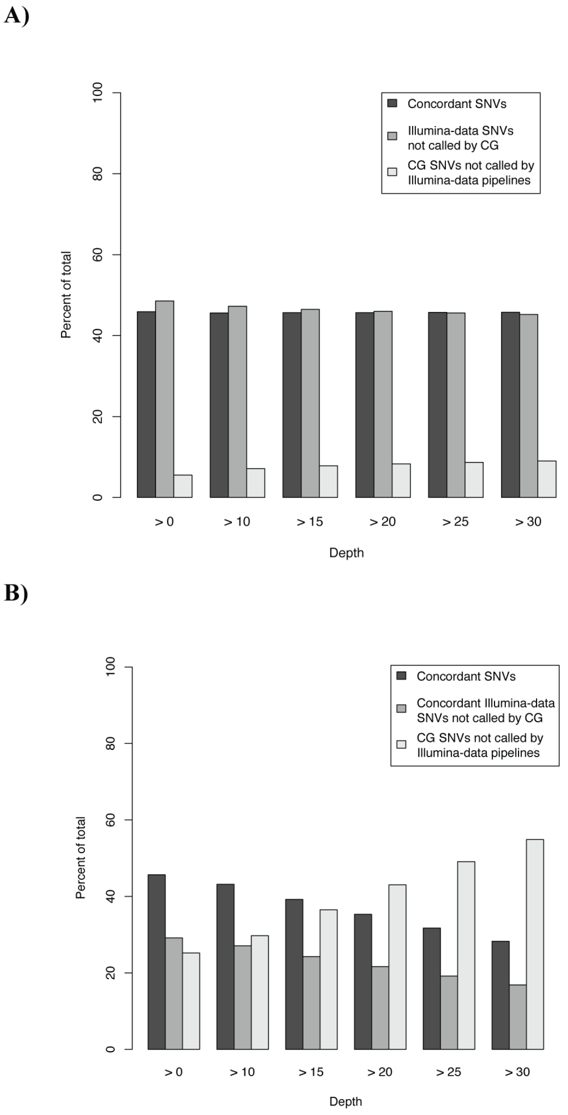


Figure 1.8: SNV concordance for a single exome, ”k8101-49685”, between two sequencing pipelines: Illumina and Complete Genomics For the Illumina sequencing, exons were captured using the Agilent SureSelect v.2 panel of capture probes. Complete Genomics SNVs consist of a subset of all SNVs called by CG that fell within the Agilent SureSelect v.2 exons. Concordance was determined by matching the genomic coordinates, base pair composition, and zygosity status for each detected SNV. Concordance was measured between CG SNVs and A) the union of all SNVs called by 5 variant calling pipelines (”Illumina-data calls”) and B) only SNVs that all 5 Illumina pipelines collectively called (”concordant Illumina-data calls”).

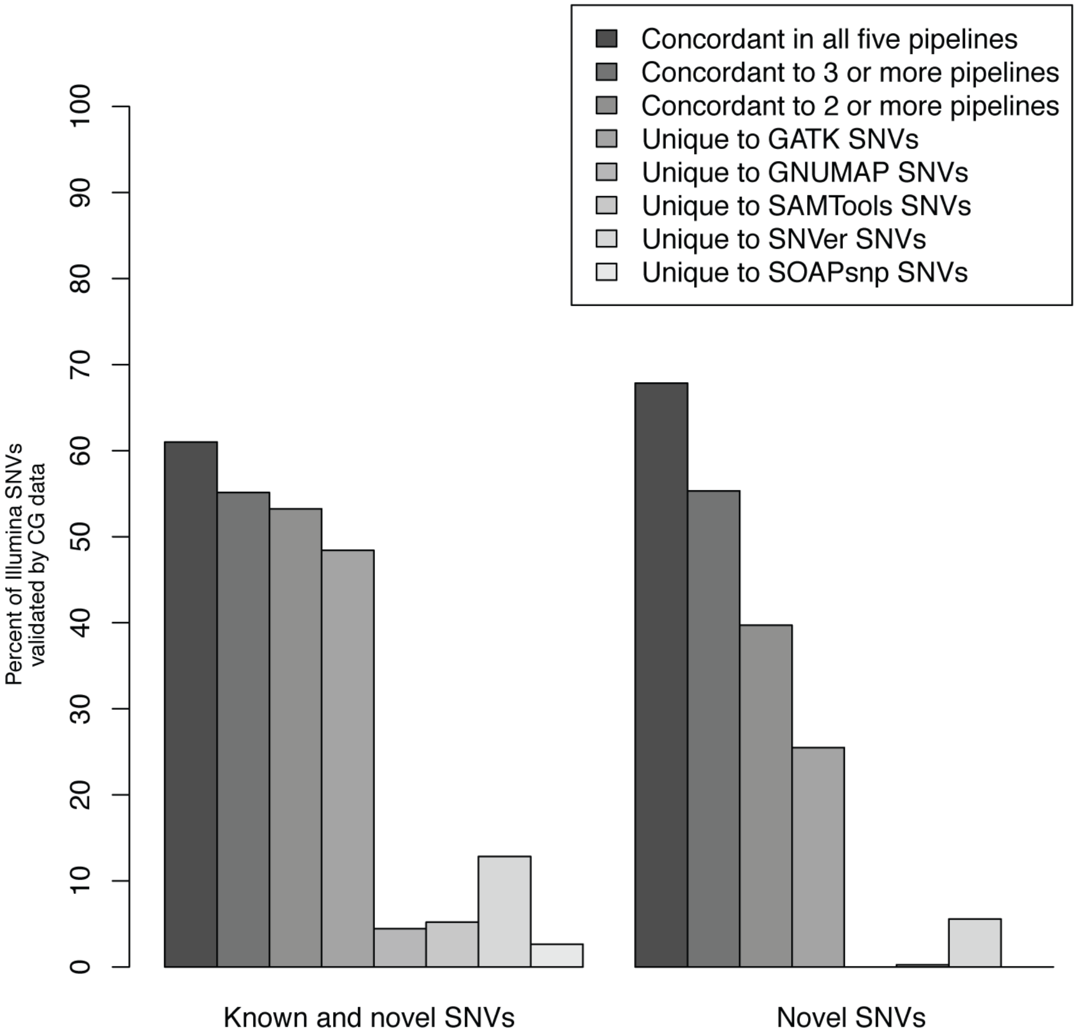


Figure 1.9: Cross-validation of illumina SNV calls using Complete Genomics SNV calls. SNVs called by each Illumina-data pipeline were cross-validated using SNVs called by Complete Genomics, an orthogonal sequencing technology, in sample ”k8101-49685”. The percentage of Illumina SNVs that were validated by CG sequencing was measured for variants having varying degrees of Illumina-data pipeline concordance. The same analysis was performed for variants that were considered novel (absent in dbSNP135).

1.3.3 Indel analysis

Variant-calling pipeline concordance

For indel calls, initial agreement between SOAPindel, SAMtools and GATK was very low at 3.0% (Figure 1.10). Indel coordinates were subsequently left-normalized and intervalized using a total range of 20 genomic coordinates (10 bp in each direction of their genomic coordinates). We found that increasing the intervalized indel range to as much as 60 genomic coordinates only marginally differed from having 20, so we chose to use 20 as a reasonable and conservative range for intervalizing indels. This method increased the overall concordance to 26.8% between the three indel-calling pipelines (Figure 1.11). For novel indels, the concordance (4.7%) was much lower than the overall concordance among known indels (43.3%). In an earlier paper, we previously validated with Sanger sequencing three indels found in three of the current pilot samples[179]. These three validated indels were detected by all three indel-calling pipelines.

Sequencing platform concordance

Indels falling within the range of the Agilent SureSelect version 2 exons were excised from the whole genome of sample k8101-49685 sequenced and analyzed by CG. Indels were again left-normalized and intervalized using a total range of 20 genomic coordinates, 10 in either direction. The normalized and intervalized CG indel calls were compared with all normalized and intervalized indels detected by the three Illumina data pipelines, and 32% were in agreement (Figure 1.12).

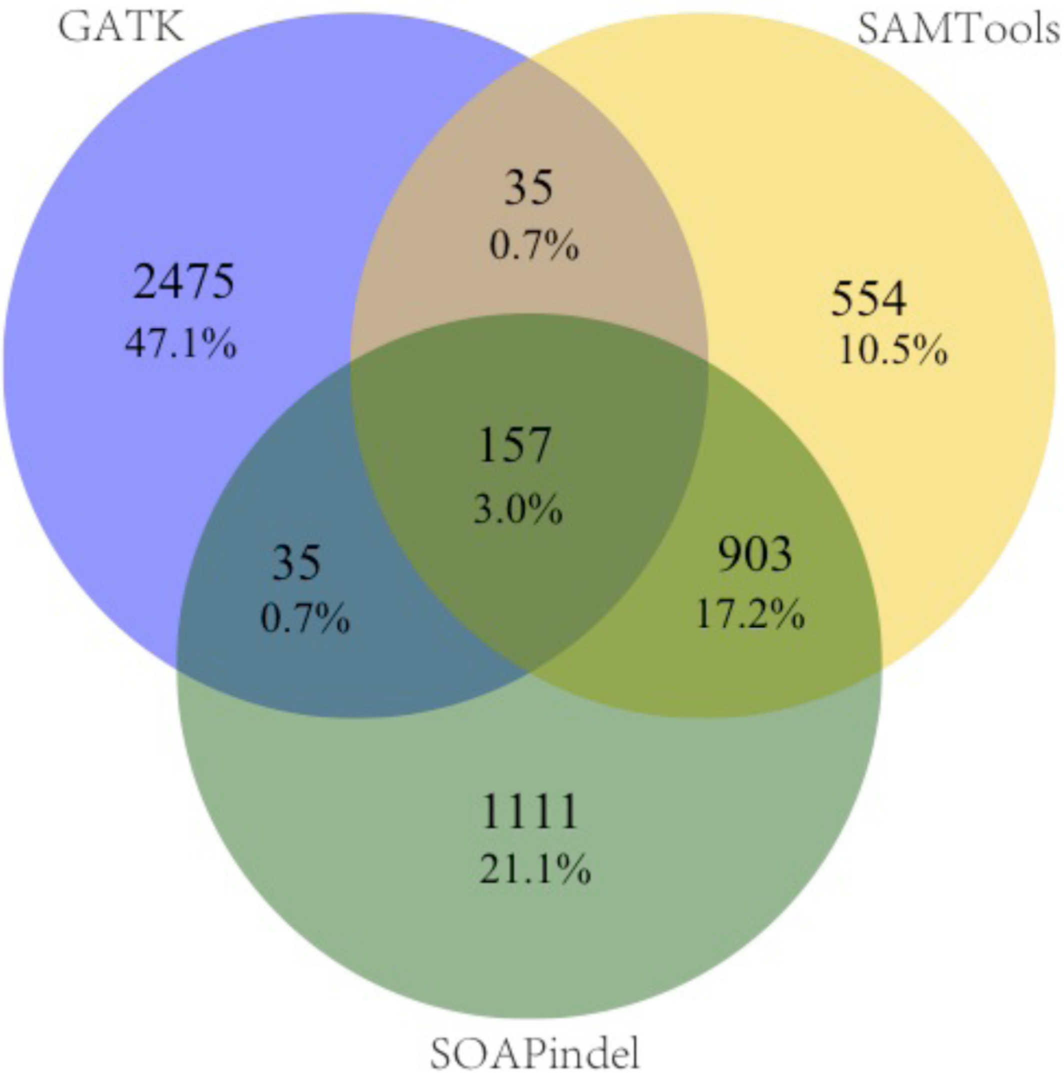


Figure 1.10: Mean indel concordance among 15 exomes between three indel detecting pipelines: GATK, SAMtools and SOAPindel. Concordance was measured between raw, pre-standardized, indel calls. Indels were considered in agreement if the genomic coordinates, length and composition of indels matched between pipelines

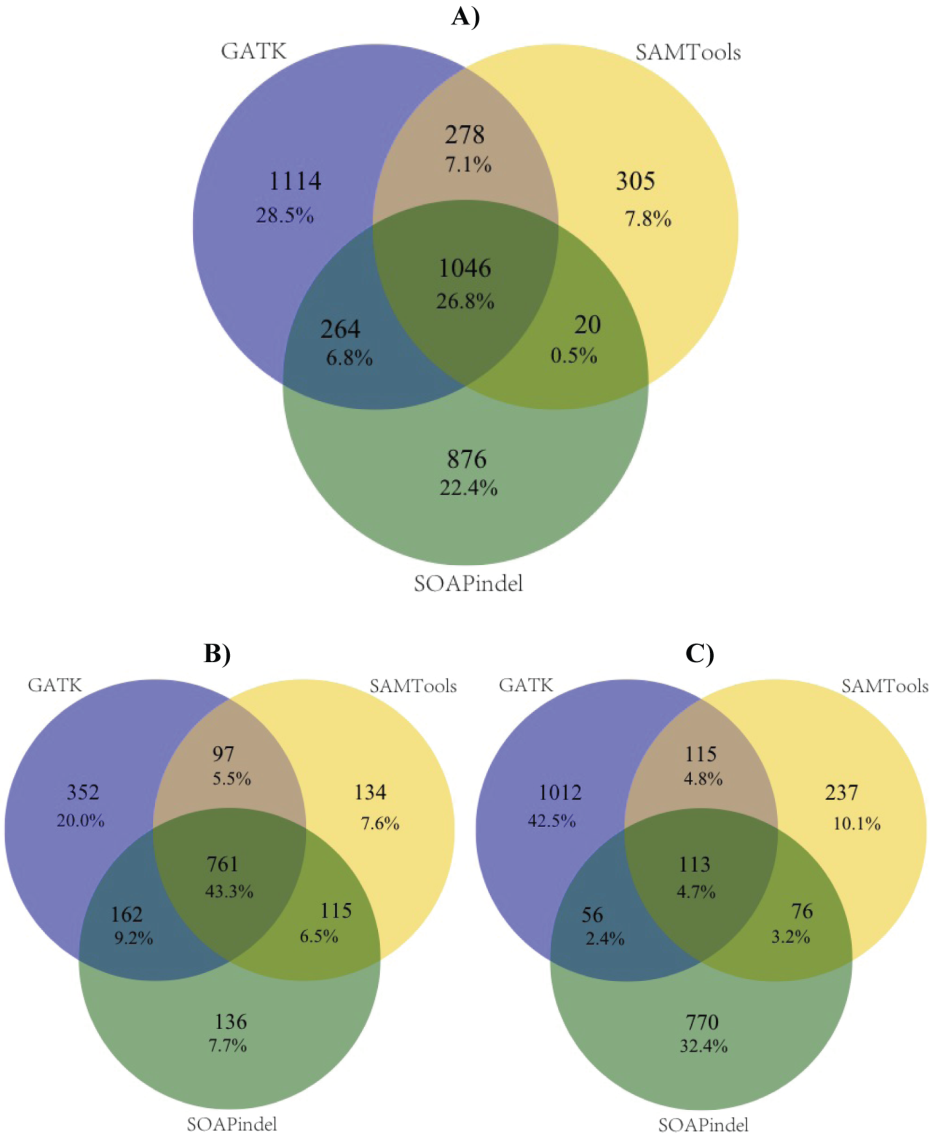


Figure 1.11: Mean indel concordance over 15 exomes between 3 indel-calling pipelines: GATK, SOAPindel, and SAMtools. Mean concordance was measured between (**A**) all indels, (**B**) known indels (indels found in dbSNP135), and (**C**) unknown indels (indels not found in dbSNP135). Indels were left normalized and intervalized to a range of 20 genomic coordinates (10 coordinates on each side of the normalized position) to allow for a reasonably standardized indel metric for comparison. To determine whether or not indels were matching, the genomic coordinates as well as the base length and composition of each indel were considered.

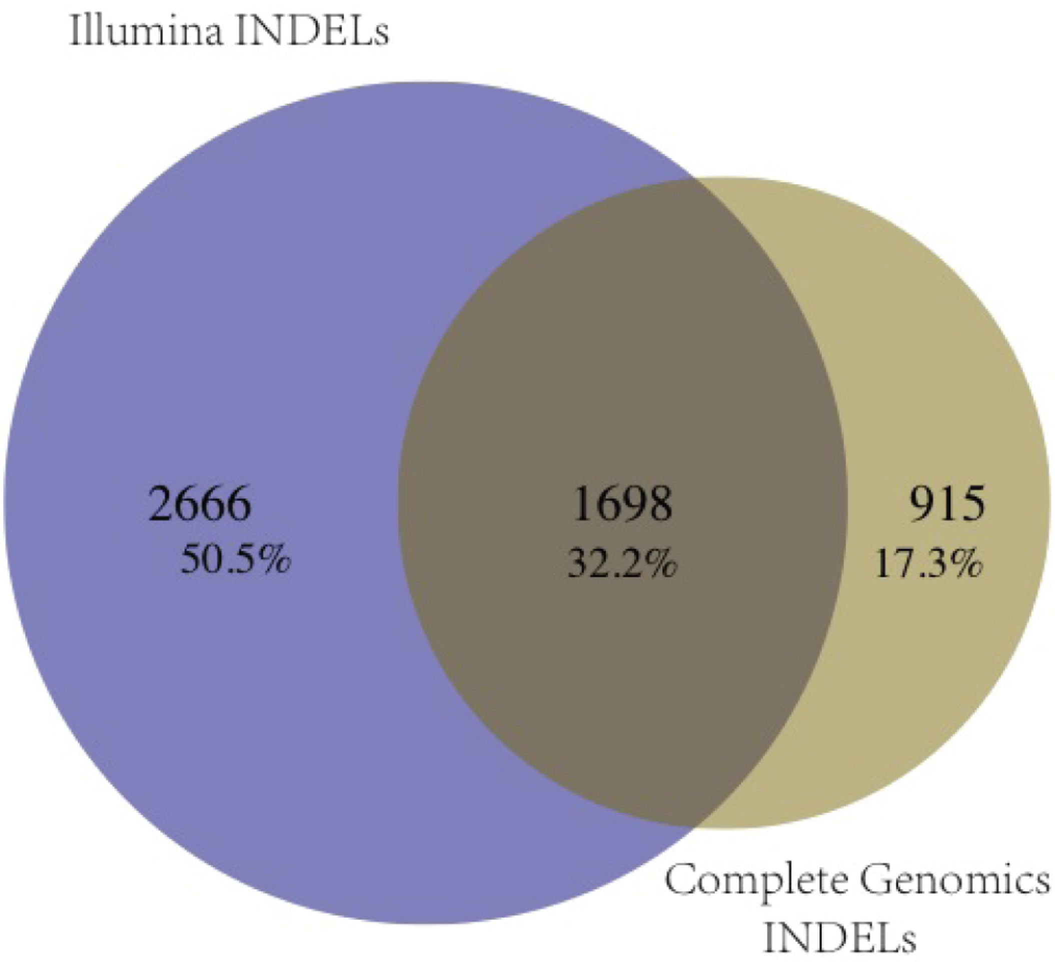


Figure 1.12: Indel concordance for a single exome, k8101-49685, between two sequencing pipelines: Illumina and Complete Genomics (CG). Illumina indels consist of a union of all indels called by each of the three indel-calling pipelines GATK, SOAPindel, and SAMtools. CG indels consisted of a subset of indels called by CG that fell within the Agilent SureSelect version 2 exons. Both Illumina and CG indels were left normalized and intervalized to a range of 20 genomic coordinates (10 coordinates on each side of the normalized position). To determine whether or not indels were matching, the genomic coordinates as well as the base length and composition of each indel were considered.

Cross-platform comparison of unique-to-pipeline indels

For the three Illumina data indel-calling pipelines, unique-to-pipeline indels were compared with indels discovered by CG (Figure 1.13). Concordance with CG indels was relatively low for all three pipelines with unique-to-GATK indels, showing a concordance of 24% with CG indels (324 of 1366), unique-to-SAMtools showing 29% concordance with CG indels (142 of 498) and unique-to-SOAPindel having 12% of called indels being concordant with CG indels (147 of 1246). Concordance rates improved for variants that were called by two Illumina data pipelines and further improved for variants called by all three of the indel-calling pipelines, 63% (1241 of 1986) and 90% (963 of 1069), respectively.

Novel (to dbSNP135) and unique-to-pipeline indels were also compared to CG indel calls, and the concordance rates for each pipeline were similar to those of the unique-to-pipeline only variants (Figure 1.13). Novel and unique-to-GATK indels had 24% of its indels concordant with CG indels (299 of 1236), novel and unique-to-SAMtools had 28% of its indels concordant with CG indels (96 of 343) and novel and unique-to-SOAPindel had 5% of its indels concordant with CG indels (53 of 1056). Novel indels that were called by two Illumina data pipelines displayed an increased concordance rate, 54% (229 of 423). Variants called by all three of the indel-calling pipelines showed the highest concordance rate for novel, unique-to-pipeline indels, 84% (103 of 122).

1.3.4 MiSeq validation of pooled PCR amplicons

To validate variants called by the two more widely used pipelines (SOAP and GATK), we used the orthogonal approach of PCR amplification of genomic DNA regions containing selected SNVs and indels, followed by pooled MiSeq sequencing. The PCR amplification (instead of exon capture), longer read lengths on the MiSeq platform, and the much higher depth

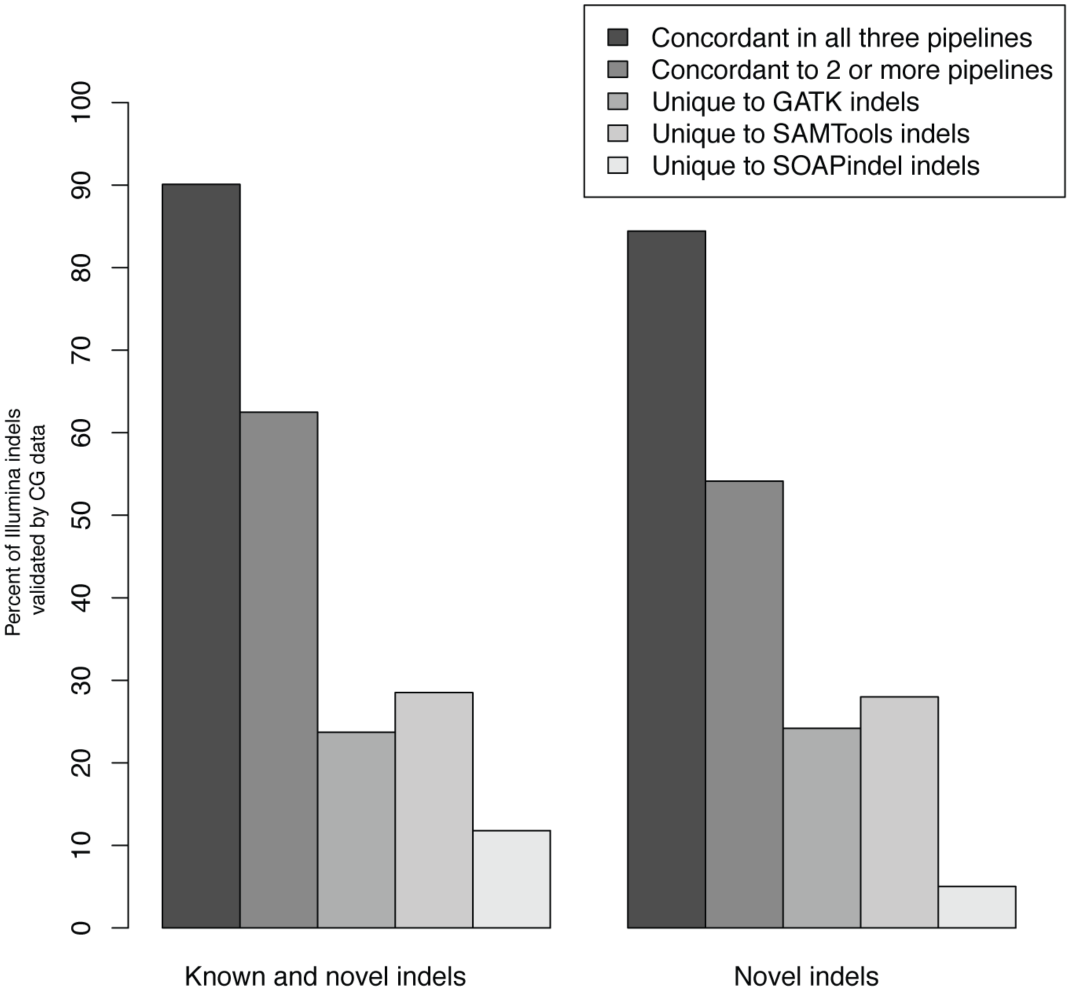


Figure 1.13: Cross-validation of illumina indel calls using Complete Genomics indel calls. Indels called by each Illumina-data pipeline were cross-validated using indels called by Complete Genomics for sample ”k8101-49685”. The percentage of Illumina indels that were validated by CG sequencing was measured across varying degrees of Illumina pipeline concordance. The same analysis was done for novel indels (indels not found in dbSNP 135).

of coverage provided a strong method of validation for SNVs and indels. A total of 1,140 SNVs found in sample k8101-49685 were selected for MiSeq validation; 760 of these SNVs were randomly selected from the set of SNVs that were unique to the GATK version 1.5 and SOAPsnp version 1.03 pipelines, 380 SNVs from each pipeline respectively. An additional 380 SNVs were randomly selected from the set of variants that were in agreement between GATK and SOAPsnp. After some analysis of the quality of the MiSeq data using FASTX, the MiSeq paired-end read data (version 2 sequencing kit, 250 × 250 bp reads) was trimmed to 150 bp and then aligned with BWA version 0.6.2 to the human reference genome sequence GRCh37, and variants were called with GATK UnifiedGenotyper version 2.3-9.

Of the 1,140 SNVs targeted for MiSeq validation, 919 (81.0%) were successfully amplified and sequenced, with an average read depth of 5,392. Validation rates for unique-to-GATK SNVs were high, with 306 of 315 (97.1%,) being successfully validated. For unique-to-SOAPsnp, 174 of 289 SNVs (60.2%) were validated. SNVs that were called by both GATK and SOAP were validated in 312 of 315 instances (99.1%) (Figure 1.14).

For indels found in sample k8101-49685, 960 were randomly selected for validation. Of these, 386 were randomly selected from the unique-to-GATK indel set, 387 were randomly selected from the unique-to-SOAPindel set, and 187 were randomly selected from set of indels overlapping between the two (SOAPindel and GATK). Of the 960 indels that were targeted for sequencing, 841 (83.5%) were successfully amplified and sequenced, with an average coverage of 4,866.

Unique-to-GATK indels had a validation rate of 180 of 336 (54.0%), being validated. The validation rate for unique-to-SOAPindel was found to be 44.6% ,with 148 of 332 validating. For indels that were called by both SOAPindel and GATK, 132 of 169 (78.1%) were successfully validated (Figure 1.14).

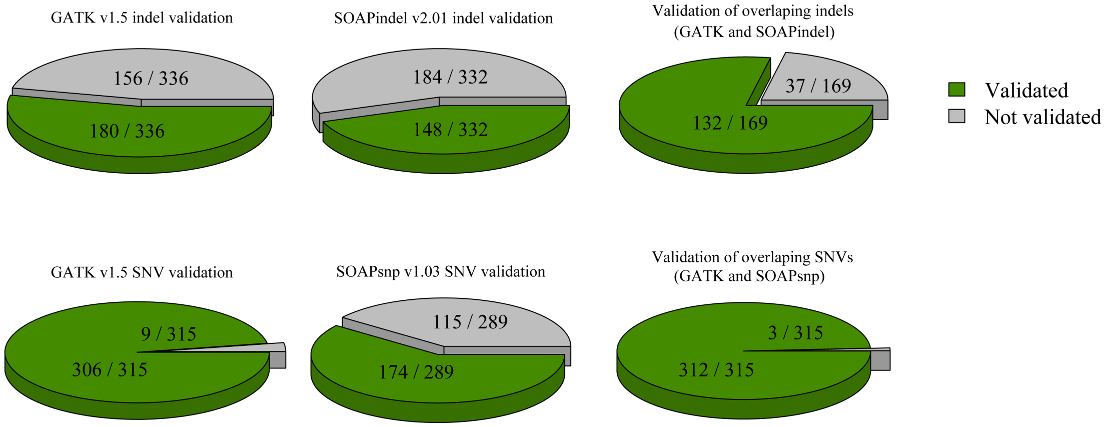


Figure 1.14: MiSeq validation experiment on a subset of Illumina-data calls. A total of 1,140 SNVs from sample k8101-49685 were randomly sampled for MiSeq validation, with 380 sampled from the set of unique-to-GATK SNVs, 380 sampled from the set of unique-to-SOAPsnp SNVs, and 380 sampled from the set that were overlapping between these two pipelines. There were 919 (81.0%) of these variants that were successfully amplified and sequenced. BWA version 0.6.2 and GATK version 2.3-9 were used to process the sequencing data for variant-calling. Additionally, 960 indels from sample k8101-49685 were randomly selected for validation. Of these, 386 were randomly selected from the unique-to-GATK indel set, 387 were randomly selected from the unique-to-SOAPindel set, and 187 were randomly selected from the set of indels overlapping between the two (SOAPindel and GATK). There were 841 (83.5%)of these indels that were successfully amplified and sequenced. BWA version 0.6.2 and GATK version 2.3-9 were used to determine the number of variants that were also successfully validated across these sets.

1.3.5 GATK v2.3-9 and the new HaplotyperCaller

Newer implementations of SNV and indel-calling pipelines continually advance the field of variant discovery and analysis by increasing the accuracy by which variants can be reliably called. Here, we show an example of the differences between previous versions of GATK with respects to SNV calls and indel calls on the same sample, k8101-49685. The vast majority of SNV calls made by both the GATK UnifiedGenotyper version 2.3-9 and the GATK HaplotypeCaller version 2.3-9 modules overlapped with the SNV calls made by the GATK UnifiedGenotyper version 1.5, showing an overall concordance of 91.0% (27,150 of 29,912) and 87.0% (26,751 of 30779) respectively. However, for indel calls, the picture was quite different, with the GATK UnifiedGenotyper version 2.3-9 and GATK HaplotypeCaller version 2.3-9 modules showing an overall concordance with the GATK UnifiedGenotyper version 1.5 calls of 54.7% (1,688 of 3,085) and 54.6% (1,858 of 3,404) respectively (Figure 1.15).

1.3.6 Predictive models

For the GATK HaplotypeCaller, 12 false negative INDEL calls were isolated and no true negative calls were recovered. For SNVs, 88 false negative calls were isolated, and 75 true negative calls were recovered. For the FreeBayes caller, 315 false negative INDEL calls were isolated and 266 true negative calls were recovered. For SNVs, 291 false negative calls were isolated, and 277 true negative calls were recovered. For the phylogenetic method, 2886 false negative and 2788 true negative SNVs were isolated and recovered, respectively. Because the GATK HaplotypeCaller set is small, we have excluded it from further analysis. The subsequent analyses focus instead on the FreeBayes and phylogenetic sets.

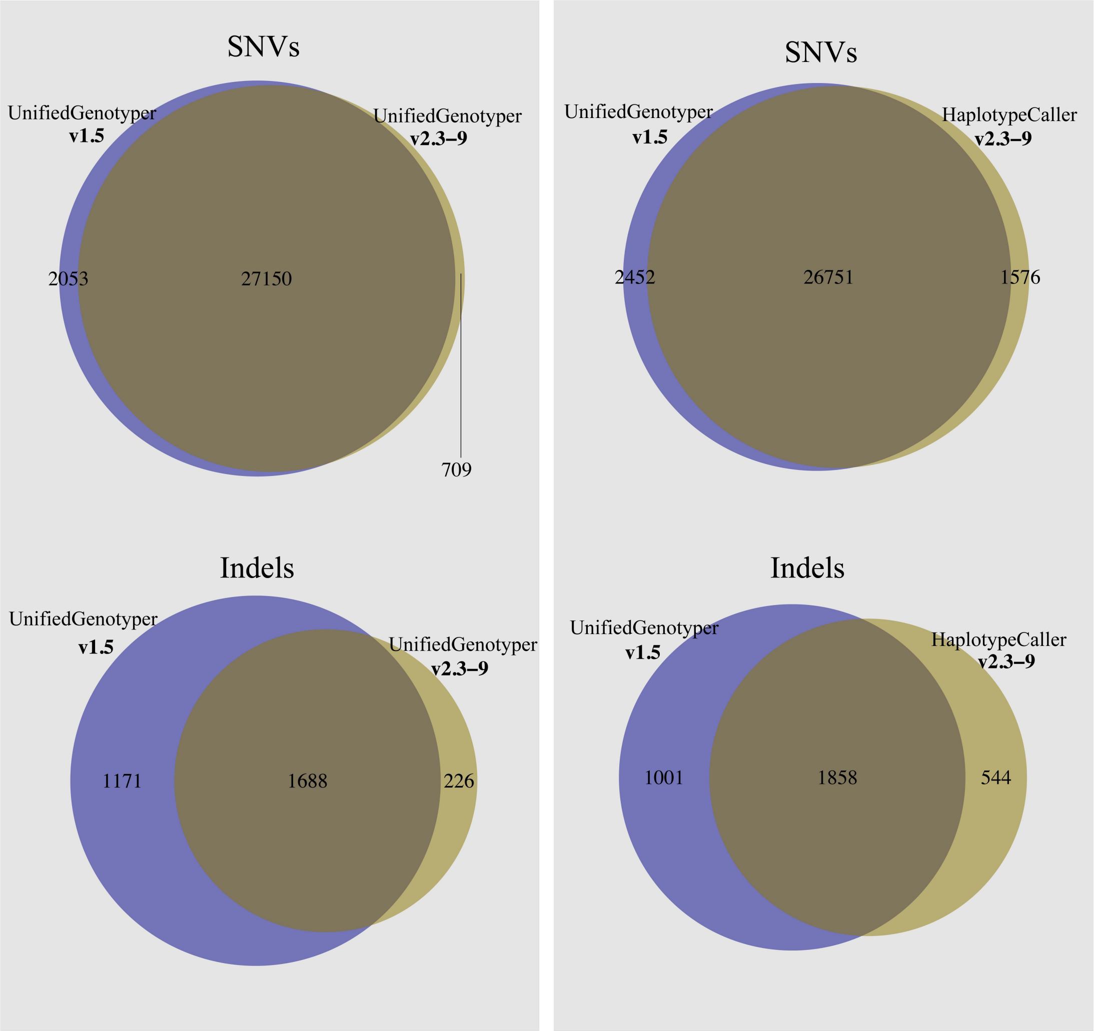


Figure 1.15: A comparison between recent versions of various GATK variant calling modules. The similarity between SNV and indel calls made between two versions of GATK, v1.5 and v2.3-9, was measured. SNV and indel calls were made using both the UnifiedGenotyper and HaplotypeCaller modules on the same k8101-49685 participant sample. Pairwise comparisons were made between the GATK UnifiedGenotyper v1.5 and each of the GATK v2.3-9 modules (the UnifiedGenotyper and HaplotypeCaller).

Model performance

For the models built using FreeBayes false and true negative SNVs, the random forest and gradient tree boosting classifiers performed similarly with a mean AUROC of 0.92 (sd=0.04), k-nearest neighbors performed well with a mean AUROC of 0.91 (sd=0.04), followed by the support vector machine classifier which had an AUROC of 0.90 (sd=0.05), and lastly the logistic regression had an AUROC of 0.87 (sd=0.08). For the models built using FreeBayes false and true negative INDELs, the random forest classifier performed best with an AUROC of 0.76 (sd=0.06), followed by the gradient tree boosting classifier with an AUROC of 0.74 (sd=0.04) and the k-nearest neighbor classifier with an AUROC of 0.71 (sd=0.06). Both the logistic regression and support vector machine classifiers performed poorly with AUROC scores of 0.50 (sd=0.05) and 0.48 (sd=0.13), respectively (Figure 1.16).

For the models built using the set of false and true negatives constructed using the phylogenetically motivated approach, the gradient tree boosting classifier performed the best with a mean AUROC of 0.58 (sd=0.03). The k-nearest neighbor classifier achieved an AUROC of 0.57 (sd=0.03), and the random forest and support vector machine classifiers had AUROC scores of 0.57 (sd=0.03) and 0.56 (sd=0.03), respectively (Figure 1.16). Lastly, the logistic regression classifier achieved an AUROC of 0.54 (sd=0.02).

Predictor importance

To obtain some insight as to which predictors were important in determining class, a measure of predictor importance was used (Figure 1.17). Predictor importance was obtained from the random forest classifier by computing the total increase in node purity weighted by the probability of reaching said node, averaged across all trees. The resulting value can be interpreted as the discriminatory power of the predictor, where higher values indicate a higher

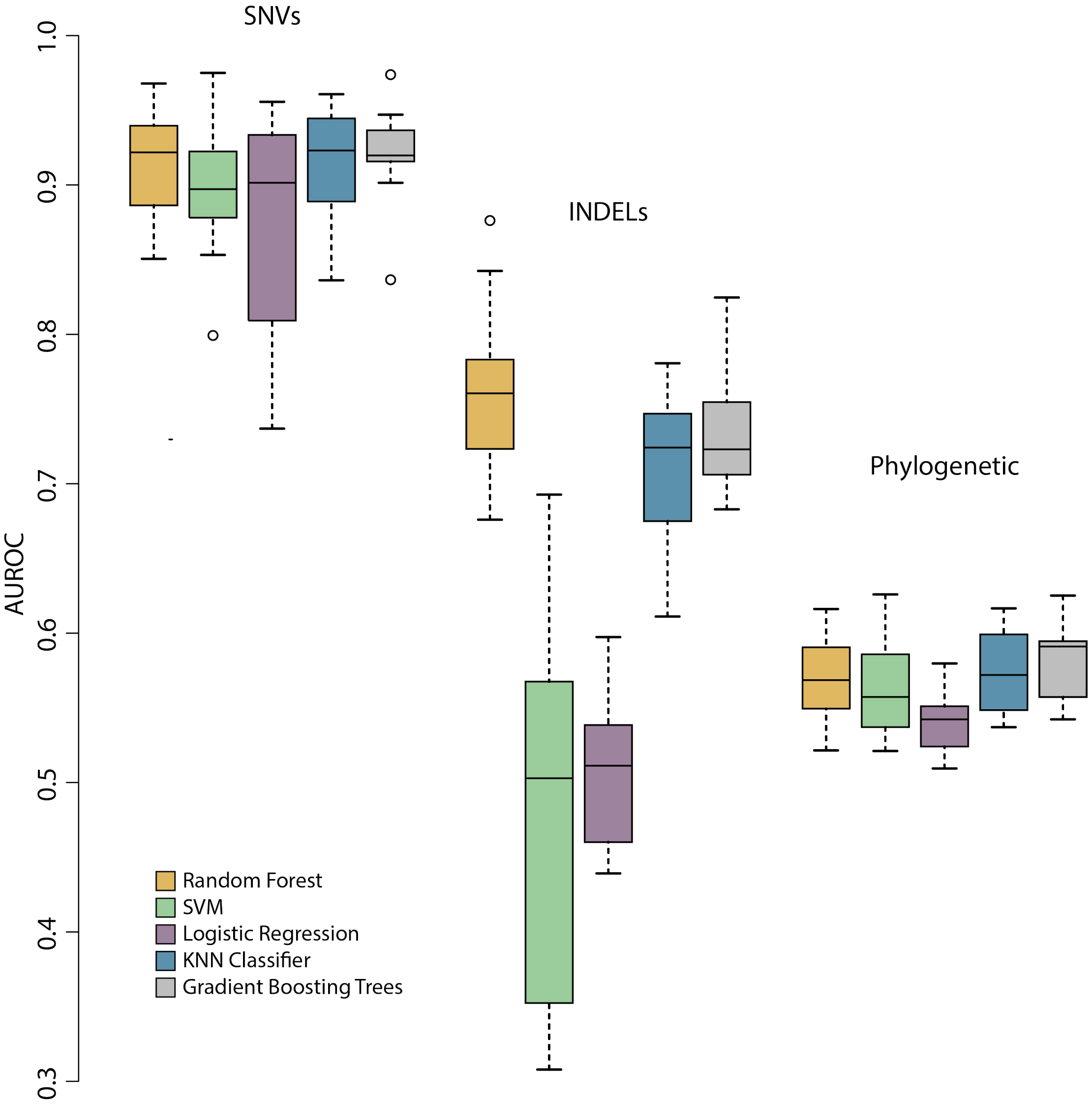


Figure 1.16: Machine learning model performance across all true and false negative call sets. Model performance (AUROC) for is plotted for each model built across the three distinct call sets. Box plots are generated using the performance from each fold in a 10-fold cross-validation experiment. All models perform quite well for the FreeBayes SNV set, but perform less well for the INDEL set. All models perform only moderately better than chance alone for the sets of false and true negative SNVs that were isolated using the phylogenetic methodology.

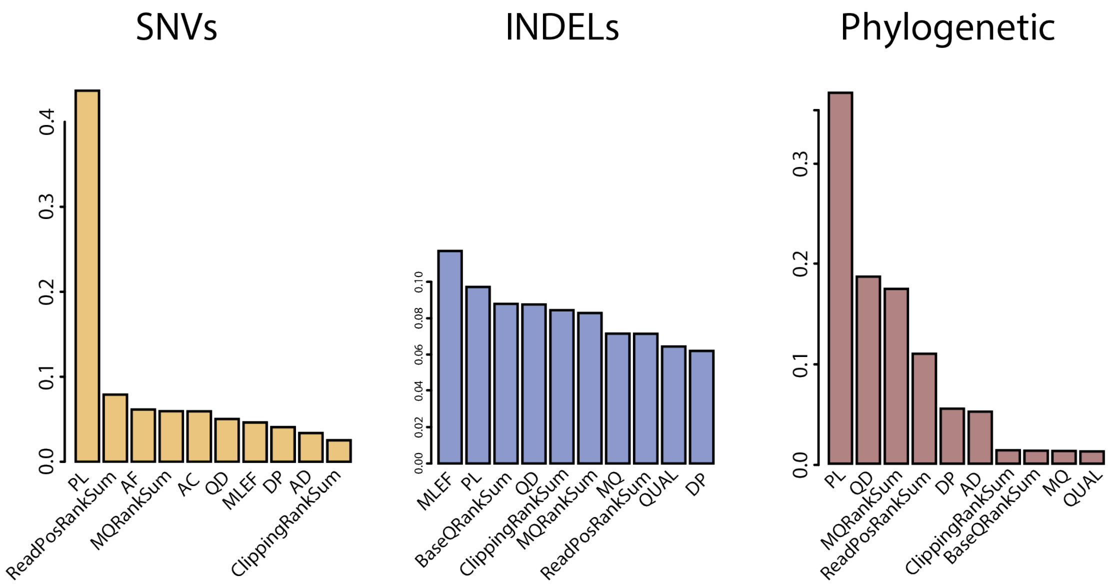


Figure 1.17: Predictor importance across the three different modeling tasks. Predictor importance was measured using the random forest classifier. The total increase in node purity weighted by the probability of reaching that node, averaged across all trees, is plotted for the top 10 most important variables for models built using FreeBayes false and true negative SNVs INDELS and false and true negative SNVs that were isolated using the phylogenetically motivated approach.

degree of discriminatory power. Interestingly, for models built using FreeBayes false and true negative SNVs, there seems to be a dominant predictor, followed by a number of weaker predictors. This is in contrast to models built using FreeBayes false and true negative INDELs, where there are seems to be a large number of weakly informative predictors. For the models built using the set of false and true negatives constructed using the phylogenetically motivated approach, there seems to be one dominant predictor and several intermediately informative predictors. See Table 1.6 for the predictors used for model building.

1.4 Discussion

Significant advances have been made in detecting genomic variation in ‘next-generation’ sequencing data, despite considerable sources of uncertainty[221]. However, we have found that

Predictor	Description
AC	Allele count in genotypes
AD	Allelic depths for the ref and alt alleles
AF	Allele Frequency, for each ALT allele
AN	Total number of alleles in called genotypes
BaseQRankSum	Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities
ClippingRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases
DP	Approximate read depth
ExcessHet	Phred-scaled p-value for exact test of excess heterozygosity
FS	Phred-scaled p-value using Fisher's exact test to detect strand bias
GQ	Genotype Quality
MLEAC	Maximum likelihood expectation (MLE) for the allele counts
MLEAF	Maximum likelihood expectation (MLE) for the allele frequency
MQ	Mapping Quality
MQRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
PL	Normalized, Phred-scaled likelihoods for genotypes
QD	Variant Confidence/Quality by Depth
QUAL	The Phred-scaled probability that a REF/ALT polymorphism exists at this site given sequencing data
ReadPosRankSum	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
SOR	Symmetric Odds Ratio of 2x2 contingency table to detect strand bias
AA_len	Alternative allele length

Table 1.6: A list and description of predictors that were consistently present among all false and true negative sets used in model building

there still exists significant discrepancy between the overall variant sets called by five different variant-calling pipelines applied to the same raw sequencing data using near-default parameters, along with differences noted between two next-generation sequencing (NGS) platforms. There are, of course, relatively large regions of overlap between all pipelines even when highly specialized parameterizations are not used. This suggests that there exists a 'region' of variants that can be called robustly by many different pipelines regardless of meticulous parameterization. The field has naturally focused on this robust set of calls, but we wish to highlight here the considerable degree of discordance as well as the high false-negative rates.

1.4.1 A discussion about variant quality and the case for multiple methods for sequence analysis in personal genomes

For the five variant-calling pipelines included in our study, a large number of calls (both SNV and indel) are shared among them in each exome, 21,146 on average for SNVs. Although all five pipelines converge on a relatively large number of SNVs, this still represents less than 60% of the total SNV call set determined by all five pipelines, and hence there still exists a considerable degree of variation between pipelines used, with near-default parameterization on the same raw sequencing data. This disagreement is likely to be the result of many factors including alignment methods, post-alignment data processing, parameterization efficacy of alignment and variant-calling algorithms, and the underlying models utilized by the variant-calling algorithm(s).

SAMtools, SOAPsnp, and GATK use similar Bayesian methods to compute the posterior probability of the true genotype[161,190,163,164,167], but they differ in the prior probabilities used. For example, SAMtools (MAQ) sets the prior probability for a heterozygous SNV at 0.001 for novel variants, and 0.2 for known SNVs. SOAPsnp uses a more complex method of assigning prior probabilities by distinguishing the homozygous genotype for the reference allele from the homozygous genotype for the alternative, and distinguishing transition (A↔G, C↔T) mutations from transversion (A/G↔C/T) mutations. GATK is similar to SAMtools but utilizes more advanced pre-processing and post-processing steps, such as local re-alignment around possible indel loci, and quality recalibration of both base quality and variant quality to improve overall variant-call performance. By modeling allele frequency, SNVer uses a frequentist framework for calling variants[307]. SNVer formulates variant calling as a hypothesis-testing problem so that a prior probability is not required, and SNVer could act as a complementary method to Bayesian methods. GNUMAP, which employs a probabilistic Needleman-Wunsch algorithm, might also be considered an orthogonal method, as even its computational framework for sequence alignment is novel[43].

Each variant-calling pipeline detects variants that others do not, and the accuracy of these discordant variants is expected to be low, but not zero. Indeed, for our comparison of SNVs called by SOAP, GATK, or both, MiSeq validation of unique-to-GATK and unique-to-SOAP variants demonstrated relatively high rates of validation, with 306 of 315 of the randomly selected SNVs from unique-to-GATK and 174 of 289 SNVs from unique-to-SOAPsnp being validated. Indels had lower, but still non-zero, validation rates, with 180 of 336 unique-to-GATK indels and 148 of 332 unique-to-SOAPindel indels being validated. Given that 'unique-to-pipeline' variants exist even in regions of relatively high sequence coverage, it is necessary to develop other approaches for including or excluding these variants from downstream analyses.

In the realm of biomedical research, every variant call is a hypothesis to be tested in light of the overall research design. Missing even a single variant can mean the difference between discovering a disease-contributing mutation or not[182]. For this reason, our data suggest that using a single bioinformatics pipeline for discovering disease-related variation is not always sufficient. A more comprehensive approach can be taken; all variants discovered by multiple variant-calling pipelines, when coupled with appropriate no-calling and quality filtering, could be included in downstream analyses, so as to not miss potentially disease-contributory variants. This is something that we intuitively implemented in a prior project[263], but for which these data now provide empirical support.

One can minimize false positives by increasing stringency filters, but this automatically and correspondingly increases the false-negative rate. The intersection between multiple variant-calling pipelines will minimize the false-positive rate, but we have also shown that each pipeline uniquely identifies some true variants. Hence, clinicians and policy-makers should be aware that propagating forward the intersection of variant calls will result in a high false-negative rate with exome sequencing and WGS, and we discuss below the advantage brought to bear on this issue by studying families. Therefore, processing single sample datasets from one sequencing platform with multiple variant-calling pipelines should not be the long-term solution for generating variant calls with high sensitivity and specificity, and we discuss alternative approaches below.

1.4.2 The case for indel standardization

Although the focus of most variant-calling software has been on detecting SNVs, it is the case that large-scale structural copy number variants and small indels are known to also be a biologically relevant and prevalent form of genetic variation[206,196]. Indeed, initial indel mapping efforts revealed upwards of 800,000 indels in a diverse population that map to known human genes, some of which can be associated with genetic disease[194,195], while recent estimates from the 1000 Genomes Project[47] suggest a 10:1 ratio of SNVs to indels in individual human genomes. Reliably detecting indels is therefore a crucial component of constructing a comprehensive set of clinically relevant genetic variants.

In contrast to SNVs, few indel-calling tools have been developed, so current knowledge of the existing variation due to indels, as well as the clinical implications of indels, has lagged. In spite of the fact that indel detection is becoming an important aspect of structural-variant analysis[206], indel calling is relatively imprecise and inaccurate. For example, the position of an indel with respect to its reference is, in many cases, ambiguous. An indel can often be represented at any of multiple locations. Krawitz et al.[150] designed an indel coordinate comparison metric, the equivalent indel region (eir), for comparing indel calls between pipelines, and GATK provides a tool which attempts to normalize indel position by left-justifying the indel within its multiple possible coordinate representations. Indeed, commonly used databases such as dbSNP have not yet entirely addressed the imprecision of indel-calling pipelines[150] and report only a single position for an indel, which could lead to disparate clinical diagnoses/outcomes between similarly affected individuals. We suggest that a more comprehensive approach should be taken, with all potential positions for each indel expressed and accounted for, so that downstream analysis can take advantage of the known existing ambiguity.

Our data demonstrates large discrepancies between indel-calling pipelines and suggests potentially high numbers of false positives and/or false negatives. Although putative false positives can be tested via modest resequencing efforts, false negatives and ‘no-calls‘ require large-scale, often impractical, resequencing projects to discover them. Because of this limitation, few data exist on false-negative rates across pipelines, and inferring these rates from unique calls between pipelines is likely to be inadequate. Indel frequency, indel size, read length, and read depth are all known to affect the accuracy of indel-calling pipelines, and the performance of different pipelines also depends, in part, on experimental conditions[214]. We show that a relatively simple method can increase comparison accuracy for indels between pipelines and between individuals; left-normalized and intervalized indel calls allow rapid and reasonable comparison of called indels between different indel-calling pipelines, as well as between individuals who have had their genome or exome sequenced. The issues highlighted in our indel comparisons demonstrate the difficulties associated with attaining accurate and standardized indel calls, and our data illustrate the need for robust and ubiquitous indel standardization metrics/methods to allow for objective comparisons across pipelines and across sequencing projects.

1.4.3 Recovering lost variants from diverse data sets

We found that machine learning models perform relatively well when trained to extract missed variants from existing data sets (Figure 1.16). Our modeling results further support the notion that using several variant calling algorithms on the same underlying sequencing data is ideal. Indeed, in conjunction with validation data, predictive models are relatively straight forward to implement, and can lead to improved overall sensitivity without risking the addition of a substantial amount of false positive sequence variants. Ideally, our models could be effectively trained without the need for large-scale validation data sets. However, we found that building models using phylogenetic relationships between mtDNA sequences were not as effective as the validation-based approach (Figure 1.16).

In our analysis of FreeBayes false negatives, we found that QUAL scores (Table 1.6) were not a strong predictor of false negative calls (Figure 1.17). We also found that measures and proxies of sequencing read depth did not strongly inform our models. Both QUAL and measures of sequence read depth are, counter intuitively, among the most widely used criteria for filtering unreliable calls. The data used to train the models were of a high average read depth (with a mean of over 100 sequence reads at the variation position for both training class sets), so it is not surprising that depth is not a substantial predictor in this case, as the data are generally sufficient in quantity. However, the QUAL score is a model based statistic which corresponds to a general characterization of the uncertainty about each call. Our data suggests that current formulations of QUAL scores are not good at representing the reliability of calls with respects to the likelihood of a false negative detection. This is important from a practical perspective because it suggests that QUAL scores should not necessarily be used as a criteria for variant filtration.

It is important to note that even with a small amount of feature engineering, our validation-based models performed quite well for SNV data sets and well, although comparatively worse, for INDEL sets (Figure 1.16). There are several other likely important features in high-throughput sequencing data that would be informative for model building. As one example for alignment based sequence analysis pipelines, local sequence characteristics are likely to be of paramount importance in building predictive models for detecting both false negative and false positive sequence variants. Indeed, others have shown that that sequence complexity limits sensitivity for alignment based callers «citation». As a consequence, features characterizing sequence similarity within a pre-defined window might be strong predictors of calling errors.

Assembly based variant calling methods represent an important algorithmic improvement over existing alignment based methods because they perform better in repetitive genomic regions. Consequently, feature design for predictive models should be tailored to the underlying algorithmic and statistical methodology employed by the different callers. It is likely that callers using distinct methodologies will be sensitive to different data characteristics, and so models may perform better if features were designed in a caller-specific manner.

One could, of course, also argue that some portion of the feature space should be variant type specific. Indeed, our results show that model performance was not only worse for recovering false negative INDEL calls, but that the relationship between the features and the response was in stark contrast to that of the SNV models. INDEL predictors were universally weakly informative while SNV predictors were dominated by a single strong predictor and other moderate to weak predictors (Figure 1.17). This suggests to us that there are important differences between SNVs and INDELs in terms of (i) how the predictors relate to calling errors and (ii) the set of predictors that are important for detecting these errors. These differences do not seem to be perfectly represented by a uniform set of predictors. INDELs, unlike SNVs, have a number of additional characteristic dimensions. They are defined by a length that can be greater than 1, and those INDELs that consist of multiple bases can be characterized by their base content, their sequence similarity to the region of interest, as well as other practical informatics based representations (such as the size of the equivalence region[150]). To obtain a better understanding of false negatives, and variant calling errors in general, more work is needed toward developing a comprehensive set of predictors for both SNVs and INDELs.

1.4.4 The case for studying large families for discovering disease-related genetic variation

Our analysis of two families, one containing only two generations (parents and children) and the other containing three generations (one grandparent, both parents, and children) demonstrates that the ability to accurately distinguish de novo variants from familial inherited variants may be more strongly limited by high false-negative rates in the parents than by high false-positive rates in the children. This can be significantly improved by having sequence data from one or more grandparents or other relatives. This finding is particularly salient for single-generation de novo studies, which attempt to characterize novel variants that are associated with genetic disease observed in the children of healthy parents[227,127,210,273,313]. While such 'no-call' or false-negative errors in parents can be ameliorated somewhat with higher sequencing depth and/or more comprehensive variant-calling strategies, larger and more comprehensive pedigrees provide a powerful, complementary source for discovering and studying human genetic variation. Although most studies utilizing NGS data to date have focused on ‘quads’ or ’trios’[227,127,210,273,313], studying large and/or consanguineous families can maximize the utility of filtering strategies and statistical approaches for identifying disease-contributory variants in genetic disease[36].

1.4.5 The case for different platforms for a more comprehensive “exome”

The relative merits of WGS are expanding as both the cost of the technique decreases and as more scientists and clinicians use the technique in an increasing number of studies/analyses[13,257,237,58]. We have shown that WGS with the version 2.0 CG pipeline delivers SNVs and indels not discovered by the Agilent exon capture and Illumina sequencing, and that these variants have a very low number of reads (<20) at those positions in the Agilent exomes, arguing that the capture of those regions was not efficient. Conversely, there were a significant number of variants in our data that were unanimously called by the five Illumina variant-calling pipelines but not by the CG WGS. Exon capture and sequencing at high depth with one platform can yield a much higher depth of coverage in most exonic regions, whereas WGS offers a more uniform and comprehensive coverage that appears to cover regions missed by exon capture and sequencing. To attain a truly comprehensive set of exonic variants, WGS on one platform could be combined with exon capture and sequencing on a different platform. This combination of the depth of exonic sequencing provided by exon capture with the breadth of coverage of WGS on a different sequencing platform, alongside the use of multiple variant-calling pipelines, provides a powerful means to maximize sensitivity and specificity for any one personal genome. However, as costs for sequencing are reduced, we anticipate that sequencing whole genomes on two or more platforms may become a feasible option for similarly maximizing accuracy.

1.4.6 The current state of variant discovery

Many of the most recent advancements that have been made in variant discovery and sequencing analysis are those related to indel discovery and analysis. Indeed, newer versions of GATK have improved upon the false-negative and false-positive rates of their indel calls in both UnifiedGenotyper and with the newer HaplotypeCaller. By leveraging local de novo assembly, similarly to the SOAPindel pipeline used in this study, the new HaplotypeCaller from GATK potentially greatly improves upon its indel-calling accuracy. The more distinct differences observed between GATK indel calls by different versions reflects the fact that indel discovery is in the earlier stages of development, when large differences are often observed between and within pipelines, with accuracy potentially remaining relatively low. For example, in each pipeline, SNV calling relies on set algorithms, which are not dramatically changed in updated versions of the software. Therefore, we do not see great leaps in accuracy for SNVs with newer versions of GATK, despite the fact that we found in the current study that at least one other pipeline (SOAP) did uniquely discover some validated SNVs that were not discovered by any other pipeline or any version of BWA-GATK tested here, and vice versa. We also note that GATK discovered comparatively more unique SNVs not discovered by any other pipeline. One caveat is that we processed the MiSeq data with the newest BWA-GATK pipeline, so this might favor the exome variants previously called by GATK, but it is nonetheless the case that SOAP identified variants that GATK missed in the same exome data using near-default parameters.

1.4.7 Some limitations of the study

It is important to note that our study did not examine somatic mutations in tumor samples, so our conclusions and our testing pipelines focus only on germline mutations from diploid genomes. We recognize that variant calling in cancer genomes represents similar but somewhat distinct challenges, and that software tools (such as SNVMix[105]) are typically developed specifically for somatic mutations in cancer genomes. Our efforts were designed to evaluate whether or not rare variants within personal genomes can be reliably and comprehensively generated from sequencing data, with or without sequencing data from other family members. Hence, we did not evaluate pipelines that specifically employ imputation or multi-unrelated-sample variant-calling algorithms, such as Thunder[168], IMPUTE2[309,119,120], or BEAGLE[27,28], or specific procedures within the software tools used that allow for calling of multiple unrelated sample (such as those available by GATK)[221]. Additionally, software tools that are only commercially available (such as CLCBio) are not evaluated in our comparative study. Predictive models built in this work are in the context of a single human sample of mtDNA, and as a consequence the results are not automatically generalizable to all sequenced human samples. Additional classification validation experiments across a large number of human samples is needed to build more informative and general models. It is also important to note that the predictive power of both sets of models built for FreeBayes INDELs and the phylogenetic set may be substantially improved by using a more sophisticated set of predictors. This study was conducted to see whether a naive model building approach would result in powerful and accurate models.

1.5 Concluding remarks

We have shown that there remains significant discrepancy in SNV and indel calling between many of the currently available variant-calling pipelines when applied to the same set of Illumina sequence data under near-default software parameterizations, thus demonstrating fundamental methodological variation between these commonly used bioinformatics pipelines. In spite of this methodological variation between pipelines, there exists a set of robust calls that are shared between all pipelines even under lax parameterization. We have further shown that the relatively recent CG version 2.0 WGS pipeline detects a set of exon variants that are not detected by several variant-calling pipelines with an Illumina-based exon-capture sequencing strategy, even in regions of high mappability[159]. Therefore, each single existing exon-capture, NGS platform, or variant-calling pipeline is likely to miss some true functional rare variants. Some authors have suggested using two separate sequencing platforms on the same samples[154], while others have suggested that technical replicates for exon capture may help to further improve accuracy of heterozygote variant calls[111]. With current technologies and cost considerations, exon capture and deep sequencing combined with WGS is still too expensive for most laboratories, and is therefore not likely to be a practical solution in the short term, despite providing the best combination of depth and breadth of coverage for genetic analyses of selected exonic regions. As an alternative, considering current prices, we suggest that utilizing the total list of variants derived from multiple (and orthogonal) variant-calling pipelines is a more feasible first option for reducing false negatives in a discovery setting. However, we fully acknowledge that in this scenario, the rates of false positives and false negatives are inversely and directly dependent on one another; that is decreasing the false-positive rate with filters will increase the false-negative rate, and vice versa[97]. We have demonstrated that studying larger multi-generational families can increase the accuracy for de novo variants. We also note that the standardization of indel discovery and reporting in a way that allows more accurate comparison of indels between sequencing platforms, variant-calling pipelines, and most importantly between individuals in a population is a critical step that needs to be addressed before this functionally important class of variants can be comprehensively assessed in either a research or a clinical setting.

2

Integrating a personal genome into clinical treatment.

2.1 Motivation

Deep brain stimulation (DBS) has emerged as a relatively safe and reversible neurosurgical technique that can be used in the clinical treatment of traditionally treatment resistant psychiatric disorders. DBS enables the adjustable and stable electrical stimulation of targeted brain structures. A recent paper by Höflich et al[117] notes variability in treatment outcomes for DBS patients, which is likely due to variable responses to differences in targeted stimulation regions and in post-operative stimulation parameters. Both sources of variation, the authors note, will effect the stimulation of different brain tissue fibers having different anatomical and functional connections. Furthermore, the authors suggest that not every target will be suitable for every person, as there exists a large degree of inter-individual variability of brain region activation during a reward task in healthy volunteers, and suggest that future work could (and should) focus on developing surgical plans based on individual-specific activations, functional connectivity and/or tractography. This work exempli-fies the large degree of clinically relevant biological variability that exists in terms of individual clinical characteristics.

Ongoing clinical trials testing the "Effectiveness of Deep Brain Stimulation for Treating People With Treatment Resistant Obsessive-Compulsive Disorder" detail the below exclusion criteria:

- current or past psychotic disorder,
- a clinical history of bipolar mood disorder, and/or
- an inability to control suicide attempts, imminent risk of suicide in the investigator's judgment, or a history of serious suicidal behavior, which is defined using the Columbia-Suicide Severity Rating Scale (C-SSRS) as either: one or more actual suicide attempts in the 3 years before study entry with the lethality rated at 3 or higher, or one or more interrupted suicide attempts with a potential lethality judged to result in serious injury or death.

These study criteria exclude the most severe cases of OCD, as many people with severe OCD also have severe depression, usually with passive (and sometimes active) suicidal ideation[294,7,12].

Obsessions and compulsions can be quite severe, with very poor insight, sometimes to a delusional or psychotic degree, and there can also be co-occurring psychoses in any individual. Each person is to some degree unique in his or her psychiatric presentation, and a tailored evaluation schema could prove more effective in clinical treatment. Due in part to these above hurdles, there are few detailed descriptions of the efficacy of DBS for OCD, with the number of published case studies on the efficacy of DBS for OCD covering upwards of 100 people[261,101,22,30,193,109,102,56,146,133,55,29,276,80,170,171].

An explosive growth in exome and whole genome sequencing (WGS)[182] has occurred in parallel to the emergence of DBS for OCD, led in part by dramatic cost reductions. This in turn has given medical practitioners an efficient and comprehensive means to medically assess coding and non-coding regions of the genome, leading to much promise in terms of assessing and treating human disease. In our own efforts to push forward the field of precision medicine, we report here one effort to integrate the areas of clinical neuropsychiatry, brain machine interfaces and personal genomics in the individualized care of one person. We evaluate and treat an individual with DBS for treatment refractory OCD, gauge the feasibility and usefulness of the medical integration of genetic data stemming from whole genome sequencing, and search for rare variants that might alter the course of medical care for this person. As mentioned above, there have been relatively few reports on studies detailing the effective application of DBS for OCD; we report here one such study.

2.2 Methods

2.2.1 Ethics compliance

Research was carried out in compliance with the Helsinki Declaration. Dr. Gholson J. Lyon (GJL) conducted all clinical evaluations and he is an adult psychiatry and child/adolescent psychiatry diplomate of the American Board of Psychiatry and Neurology. GJL obtained IRB approval number 00038522 at the University of Utah in 2009-2010 to evaluate candidates for surgical implantation of the Medtronic Reclaim® DBS Therapy for OCD, approved under a Humanitarian Device Exemption (HDE) for people with chronic, severe, treatment-resistant OCD[?] . The interdisciplinary treatment team consisted of one psychiatrist (GJL), one neurologist and one neurosurgeon. Implantation ultimately occurred on a clinical basis at another site. Written consent was obtained for phenotyping and whole genome sequencing through Protocol number 100 at the Utah Foundation for Biomedical Research, approved by the Independent Investigational Review Board, Inc. Informed and written consent was also obtained using the Illumina Clinical Genome Sequencing test consent form, which is a clinical test ordered by the treating physician, G.J.L.

2.2.2 Evaluation and recruitment for DBS for treatment-refractory OCD

GJL received training regarding DBS for OCD at a meeting hosted by Medtronic in Minneapolis, Minnesota, in September 2009. The same author attended a Tourette Syndrome Association meeting on DBS for Tourette Syndrome, Miami, Florida, in December 2009. Approximately ten candidates were evaluated over a one-year period in 2010. The individual discussed herein received deep brain stimulation surgery at another site, and then returned for follow-up with GJL. Another psychiatrist, Dr. Reid Robison, provided ongoing consultation throughout the course of this study. Although other candidates have since returned for follow-up (with GJL), no others have been surgically treated.

2.2.3 CLIA WGS and the Management of Results from sequencing data

CLIA WGS using the Illumina Individual Genome Sequencing test

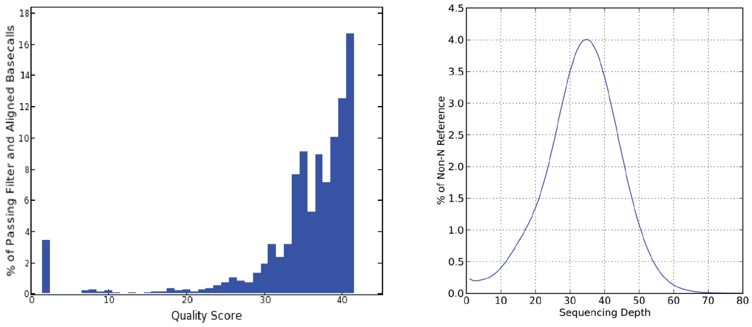
Whole genome sequencing was ordered on this individual as part of our ongoing effort to implement precision medicine in the diagnosis, treatment, and preventive care for individuals. His genome was sequenced in the Illumina Clinical Services Laboratory (CLIA-certified, CAP-accredited) as part of the TruSight Individual Genome Sequencing (IGS) test, a whole-genome sequencing service using Illumina's short- read sequencing technology. Although clinical genome sequencing was ordered by GJL on a clinical basis (thus not requiring IRB approval), the clinical phenotyping and collection of blood and saliva for other research purposes was approved by the Institutional Review Board (iIRB) (Plantation, Florida) as part of a study protocol at the Utah Foundation for Biomedical Research (UFBR). Consistent with laboratory-developed tests, WGS has not been cleared or approved by the U.S. Food and Drug Administration[181].

The entire procedure included barcoded sample tracking of the blood collected by GJL from this person, followed by DNA isolation and sequencing in the Illumina CLIA lab. The sequence was generated from extracted DNA, sequenced using the Illumina HiSeq 2000 sequencer. Briefly, DNA was fragmented, and the fragments were attached to the surface of a glass microscope slide. The fragments were then sequenced using fluorescently labeled nucleotides, which were excited by a laser and imaged using digital equipment. These fragments were then assessed for quality using a variety of metrics to ensure that only robust sequences were analyzed. Fragments were aligned to the NCBI reference sequence. Fragments that aligned to more than one region of the reference genome were excluded from the report. Additionally, fragments were excluded from the analysis on the basis of quality and alignment scores. Each nucleotide site reported was sequenced an average of 30 times, so there was

Data Volume and Quality

	Yield (Gigabases)	% Bases ≥ Q30	% Bases Aligned
Passing Filter	113.10	87.10%	87.80%

	% Callable	% ≥ 5x depth	% ≥ 10x depth	% ≥ 20x depth	Mean depth(x)
Non-N Reference	93.28%	97.57%	96.22%	88.54%	33.35



SNP Assessment

Total	Het/Hom	% in dbSNP	% in Genes	% in Coding
3,308,246	1.61	98.13%	45.47%	0.63%

Variant Statistics

	SNVs
Total Number	3,308,246
Number in Genes	1,504,121
Number in Coding Regions	20,879
Number in UTRs	24,946
Splice Site Region	2,917
Stop Gained	72
Stop Lost	16
Non-synonymous	9,884
Synonymous	10,907
Mature miRNA	36

Figure 2.1: Data statistics and SNP characteristics for the Illumina CLIA WGS pipeline WGS was performed using the Illumina CLIA WGS pipeline. We report the volume of data, the quality of the data as well as whole genome SNP characteristics and more general characteristics of SNVs reported by the Illumina CLIA WGS pipeline, including: the total number of SNVs, the total number of SNVs that are within genes, coding regions, UTRs, splice site regions as well as the number of SNVs that were stop gained, stop lost, non-synonymous, synonymous and mature mRNA.

on average 30-fold redundancy for each base pair reported. Additionally, no positions were called when the genotype quality score was less than 30 or depth was less than 6. Only single nucleotide variants are called and validated in the Illumina Clinical Services Laboratory. Data statistics are summarized in Figure 2.1.

WGS data analysis and variant prioritization

For the bioinformatics analyses, Illumina utilized the internal assembler and variant caller CASAVA (short for Consensus Assessment of Sequence And VAriation). Reads were mapped to the Genome Reference Consortium assembly GRCh37. Data for sequenced and assembled genomes was provided on one hard drive, formatted with the NTFS file system and encrypted using the open source cross platform TrueCrypt software (www.truecrypt.org) and the Advanced Encryption Standard (AES) algorithm (Federal Information Processing Standards Publication 197).

Genotyping array data was generated in parallel of the CLIA whole genome sequencing, using the Illumina HumanOmni2.5-8 bead chip. The encrypted hard drive contains several files, including a “genotyping folder” within which there is a genotyping report in a text-based tab-delimited format.

Insertions, deletions and structural alterations are not validated variant types in the Illumina Clinical Services Laboratory. Insertions and deletions provided in the gVCF file are for investigative or research purposes only. A medical report and the raw genomic data were provided back to the ordering physician (GJL) on an encrypted hard drive as part of the Illumina Understand your Genome Symposium, held in October 2012, which included the clinical evaluation of 344 genes (Table 2.1).

To perform more comprehensive downstream analyses using a greater portion of the genomic data, all of the variants that were detected by the Illumina CLIA WGS pipeline were

Genes																	
AARS	APOB	CAPN3	COL6A1	DTNA	FANCG	GJB3	HFE	KIF1B	MEFV	MYL3	NR2E3	PKHD1	PSEN2	SCN4B	SLC26A4	TMC1	USH2A
ABCA4	APP	CAV3	COL6A2	DYSF	FANCI	GJB6	HFE2	KLHL7	MERTK	MYO15A	NRL	PKP2	RAB7A	SCN5A	SLC26A5	TMEM43	VAPB
ABCC8	ARG1	CDH1	COL6A3	EGR2	FANCM	GLA	HGSNAT	LAMP2	MFN2	MYO1A	OPA3	PLN	RAD51C	SEMA4A	SLC3A1	TMIE	VCL
ABCC9	ARSA	CDH23	CPS1	ELOVL4	FGD4	GLB1	HNF1A	LDB3	MFSD8	MYO3A	OTC	PLOD1	RAG1	SERPINA1	SLC40A1	TMPRSS3	WFS1
ACADL	ARSB	CDK4	CPT2	ENG	FIG4	GM2A	HPD	LDLR	MLH1	MYO6	OTOF	PMP22	RAG2	SERPINC1	SLC7A9	TNNC1	
ACADM	ASAH1	CERKL	CRB1	EPB42	FKTN	GNPTAB	HSPB1	LHFPL5	MLH3	MYO7A	PAH	PMS1	RBM20	SERPINE1	SLX4	TNNI3	
ACADS	ASPA	CFTR	CRX	ESRRB	FUCA1	GNPTG	HSPB8	LIPA	MMAA	MYOT	PALB2	POMGNT1	RDH12	SETX	SMAD4	TNNT2	
ACADVL	ASS1	CHEK2	CTF1	EYA4	G6PD	GNS	IDUA	LITAF	MMAB	MYOZ2	PARK2	POMT1	RDX	SGCA	SMPD1	TP53	
ACTC1	ATP7B	CHMP2B	CTNS	EYS	GAA	GPD1L	IGHMBP2	LOXHD1	MMADHC	NAGA	PAX8	POU3F4	RET	SGCB	SNCA	TPM1	
ACTN2	BCKDHA	CLDN14	CTSA	F11	GALC	GPSM2	IKBKAP	LRAT	MPI	NAGLU	PCCA	POU4F3	RGR	SGCD	SNTA1	TPO	
ACVRL1	BCKDHB	CLN3	CYP27A1	F2	GALK1	GRN	IVD	LRRK2	MPZ	NDRG1	PCCB	PPT1	RHO	SGCG	SOD1	TPP1	
ADA	BEST1	CLN6	DBT	F5	GALNS	GRXCR1	IYD	LRTOMT	MSH2	NEFL	PCDH15	PRKAG2	ROM1	SGSH	SPTA1	TRIM32	
ADAMTS2	BRCA1	CNGA1	DES	F7	GALT	GUCA1B	JAG1	MANBA	MSH6	NEXN	PCSK9	PROC	RP1	SH3TC2	SPTB	TRPV4	
AGA	BRCA2	CNGB1	DFNB31	FAH	GAMT	GUSB	JUP	MAPT	MTMR2	NF1	PDE6A	PROS1	RP2	SIX1	STK11	TSHR	
AKAP9	BRIP1	COCH	DFNB59	FANCA	GARS	HADHA	KCNE1	MARVELD2	MUT	NF2	PDE6B	PRPF3	RPE65	SLC16A1	SUMF1	TTN	
ALS2	BSND	COL11A2	DHCR7	FANCB	GATM	HADHB	KCNE2	MAT1A	MYBPC3	NHLRC1	PEX7	PRPF31	RYR2	SLC17A5	TAT	TTR	
ANK1	BTD	COL1A2	DMD	FANCC	GCDH	HAMP	KCNE3	MCCC1	MYH14	NLRP12	PHOX2B	PRPF8	SAG	SLC17A8	TAZ	TULP1	
ANK2	CA4	COL3A1	DSC2	FANCD2	GCK	HBB	KCNJ11	MCCC2	MYH7	NLRP3	PHYH	PRPH2	SBF2	SLC22A5	TCAP	UBA1	
ANO5	CACNA1C	COL5A1	DSG2	FANCE	GDAP1	HEXA	KCNJ2	MCEE	MYH9	NPC1	PINK1	PRPS1	SCN1B	SLC25A13	TECTA	UGT1A1	
APC	CACNB2	COL5A2	DSP	FANCF	GJB2	HEXB	KCNQ1	MED25	MYL2	NPC2	PKD2	PSEN1	SCN3B	SLC25A20	TJP2	USH1C	

Table 2.1: A list of 344 genes analyzed by Illumina as part of the Understand your Genome Symposium in 2012

imported and analyzed within the Omicia Opal web-based clinical genome interpretation platform, version 1.5.0. The Omicia system annotates variants and allows for the identification and prioritization of potentially deleterious alleles. Omicia Scores, which are computationally derived estimates of deleteriousness, were calculated by using a decision-tree based algorithm, which takes as input the Polyphen, SIFT, MutationTaster and PhyloP score(s), and derives an integrative score between 0 and 1. Receiver operating characteristic (ROC) curves are plotted for that score based on annotations from HGMD. For further details on the method and the program see www.omicia.com.

We also applied a recently published method, ERDS (Estimation by Read Depth with SNVs) version 1.06.04[321], in combination with genotyping array data, to generate a set of CNV calls (Figure 2.2). ERDS starts from read depth information inferred from BAM files, but also integrates other information including paired end mapping and soft-clip signature, to call CNVs sensitively and accurately. We collected deletions and duplications that were >200 kb in length, with confidence scores of >300. CNVs that were detected by the ERDS method were visually inspected by importing and visualizing the read alignment data in the Golden Helix Genome Browser, version 1.1.1. CNVs were also independently called from Illumina HumanOmni2.5-8v1 genotyping array data. Array intensities were imported and analyzed within the Illumina GenomeStudio software suite, version 1.9.4. LogR values were exported from GenomeStudio and imported into Golden Helix SVS, version 7.7.5. A Copy Number Analysis Method (CNAM) optimal segmentation algorithm was used to generate a list of putative CNVs, which was then restricted to include only CNVs that were >200kb in length with average segment LogR values of > 0.15 and < -0.15 for duplications and deletions, respectively. LogR and covariate values were plotted and visually inspected at all genomic locations where the CNAM method detected a CNV. CNVs that were simultaneously detected by both methods (ERDS and CNAM) were considered to be highly confident CNVs. Highly confident CNVs were, again, visually inspected within Golden Helix Genome Browser to further eliminate any artefactual CNV calls.

Managing sequencing results

There are multiple steps involved in the management of clinical test results, beginning with bar-coded tracking of orders and the return of results to the clinician's office from the outside CLIA-certified testing facility. The results are transferred to the clinician, who reviews, signs, and interprets the results and incorporates them into the medical health record. The patient is notified, and needed follow-up is arranged.

In an ongoing effort to develop ways to incorporate genomic data into clinical EHR, we also collaborated with the Sequence Ontology Group to convert the data into the GVFclin format. The Genome variant format (GVF), which uses Sequence Ontology to describe genome variation[247], has been extended for use in clinical applications. This extended file format, called GVFClin[69], adds the necessary attributes to support Health Level 7 compatible data for clinical variants. The GVF format represents genome annotations for clinical applications using existing EHR standards as defined by the international standards consortium: Health Level 7. Thus, GVFclin can describe the information that defines genetic tests, allowing seamless incorporation of genomic data into pre-existing EHR systems.

2.3 Results

2.3.1 Pertinent clinical symptoms and treatment

A 37-year old man and U.S. veteran (here named with pseudonymous initials M.A.) was evaluated by GJL in 2010 for severe, treatment-refractory obsessive compulsive disorder (OCD), which is an illness that can be quite debilitating[207]. M.A. had a lifelong history of severe ob-

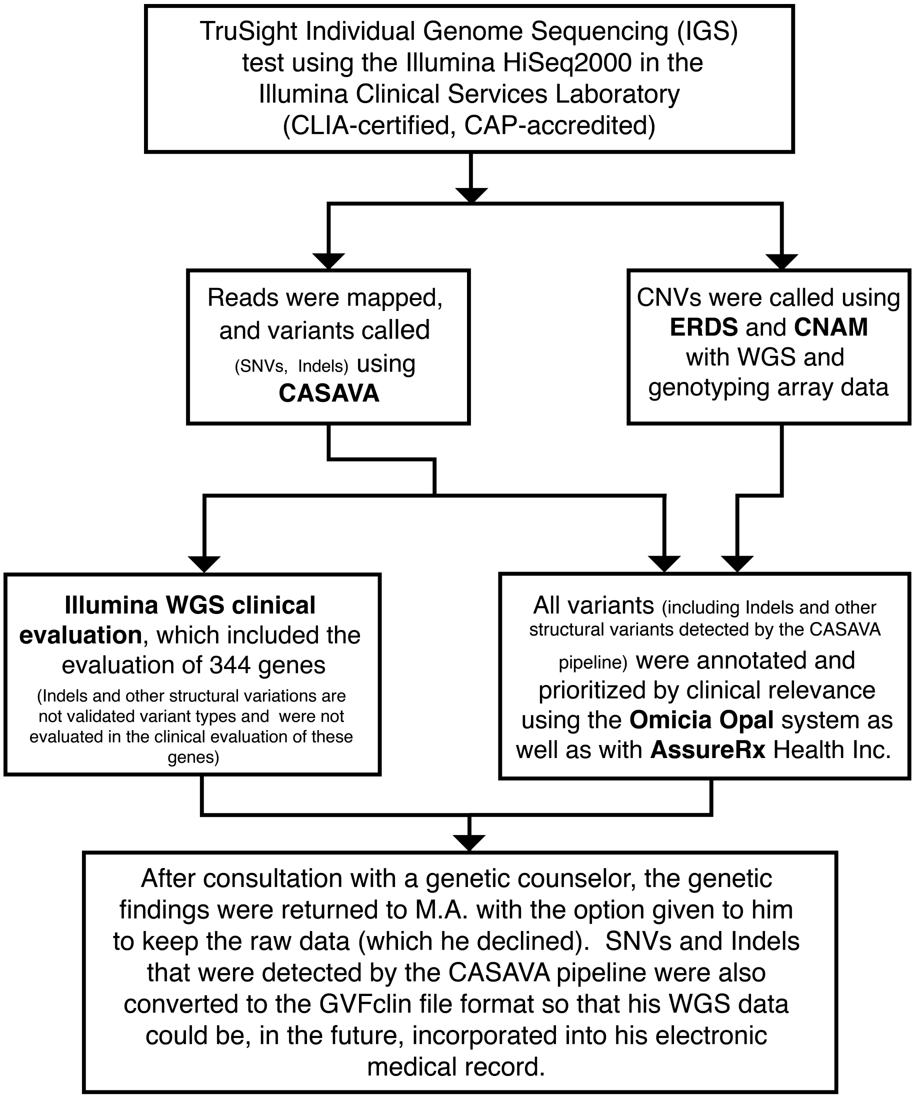

TruSight Individual Genome Sequencing (IGS) test using the Illumina HiSeq2000 in the Illumina Clinical Services Laboratory (CLIA-certified, CAP-accredited)
Reads were mapped, and variants called (SNVs, Indels) using CASAVA
CNVs were called using ERDS and CNAM with WGS and genotyping array data
Illumina WGS clinical evaluation, which included the evaluation of 344 genes (Indels and other structural variations are not validated variant types and were not evaluated in the clinical evaluation of these genes)
All variants (including Indels and other structural variants detected by the CASAVA pipeline) were annotated and prioritized by clinical relevance using the Omicia Opal system as well as with AssureRx Health Inc.
After consultation with a genetic counselor, the genetic findings were returned to M.A. with the option given to him to keep the raw data (which he declined). SNVs and Indels that were detected by the CASAVA pipeline were also converted to the GVFclin file format so that his WGS data could be, in the future, incorporated into his electronic medical record.

Figure 2.2: Implementation of the analytic-interpretive split model for the clinical incorporation of a whole genome. We have implemented the analytic-interpretive split model here with MA, with WGS being performed in a CLIA certified and CAP accredited lab at Illumina as part of the Individual Genome Sequencing test developed by them. The WGS acts as a discrete deliverable clinical unit from which multiple downstream interpretive analyses were performed. We used the ERDS CNV caller, the Golden Helix SVS CNAM for CNV calling, and the Omicial Opal and the AssureRx Health Inc. pipelines for variant annotation and clinical interpretation of genomic variants. By archiving and offering to him the encrypted hard drive containing his "raw" sequencing data, any number of people, including the individual and/or his/her health care providers can analyze his genome for years to come. Abbreviations: CLIA, Clinical Laboratory Improvement Amendments; CAP, College of American Pathologists; CASAVA, Consensus Assessment of Sequence and Variation; ERDS, Estimation by Read Depth with SNVs; CNAM, Copy Number Analysis Method; WGS, Whole Genome Sequencing.

sessions and compulsions, including contamination fears, scrupulosity, and the fear of harming others, with much milder symptoms in childhood that got much worse in his early 20's. His Yale-Brown Obsessive Compulsive Scale (YBOCS)[103,104] ranged from 32-40, indicating extremely severe OCD. Perhaps the worst period of OCD included a 5-day, near continuous, period of tapping on his computer keyboard as a compulsion to prevent harm from occurring to his family members. M.A. had suffered throughout his life from significant periods of depression with suicidal ideation, and he had attempted suicide at least three times. His prior psychiatric history also includes episodes of paranoia relating to anxieties from his OCD, and he continues to be treated with biweekly injections of risperidone.

His treatment history included over 15 years of multiple medication trials, including clomipramine and multiple SSRIs at high doses, including fluoxetine at 80 mg by mouth daily, along with several attempts with outpatient exposure and ritual prevention (ERP) therapy[100]. M.A. inquired and was evaluated by GJL at the University of Utah and then at two other centers independently offering deep brain stimulation for OCD. One of these centers required (as a condition for eligibility for an ongoing clinical trial) a two-week inpatient hospitalization with intensive ERP, which subsequently occurred and was documented as improving his YBOCS score to 24 at discharge. He maintains that he actually experienced no improvement during that hospitalization, but rather told the therapists what they wanted to hear, as they were "trying so hard".

The teams at the University of Utah and two other centers declined to perform surgery due to his prior history of severe depression, suicide attempts and possible psychoses with paranoia. Through substantial persistence of M.A. and his family members, a psychiatrist and neurosurgeon at a fourth center decided that he was an appropriate candidate for surgical implantation of the Medtronic Reclaim® DBS Therapy device for OCD, approved under a Humanitarian Device Exemption (HDE) for people with chronic, severe, treatment-resistant

OCD[?] , and he was implanted in January of 2011 (Figure 2.3). The device targets the nucleus accumbens / anterior limb of the internal capsule (ALIC).

2.3.2 Clinical results for DBS for treatment-refractory OCD

After healing for one month, the implanted device (equipped with the Kinetra Model 7428 Neurostimulator) was activated on February 14, 2011, with extensive programming by an outpatient psychiatrist, with bilateral stimulation of the ALIC. Final settings were case positive, contact 1 negative on the left side at 2.0 V, frequency 130 Hz, and pulse width 210 usec, and case positive, contact 5 negative on the right side with identical settings.

Over the next few months, his voltage was increased monthly in increments of 0.2-0.5 V by an outpatient psychiatrist. He returned to one of the author's (GJL) for psychiatric treatment in July 2011, at which time his voltage was set at 4.5 V bilaterally. His depression had immediately improved after the surgery, along with many of his most irrational obsessions, but his YBOCS score still remained in the 35-38 range. From July 2011-December 2011, his voltage was increased bilaterally on a monthly basis in increments of 0.2 V, with steady improvement with his OCD until his battery started to lose charge by December 2011. This caused him considerable anxiety, prompting him to turn off his battery in order to "save battery life", which unfortunately led to a complete relapse to his baseline state in a 24 hour period, which was reversed when he turned the battery back on. The battery was surgically replaced with a rechargeable Activa RC neurostimulator battery in January 2012, and the voltage has been increased monthly in 0.1-0.2 V increments until the present time (May 2013).

At every visit, M.A. has reported improvements, with reductions of his obsessions and compulsions, marked by a steady decline in his YBOCS score (Figure 2.4). M.A. has started to participate in many activities that he had never previously been able to engage in. This includes: exercising (losing 50 pounds in two years) and volunteering at the church and other

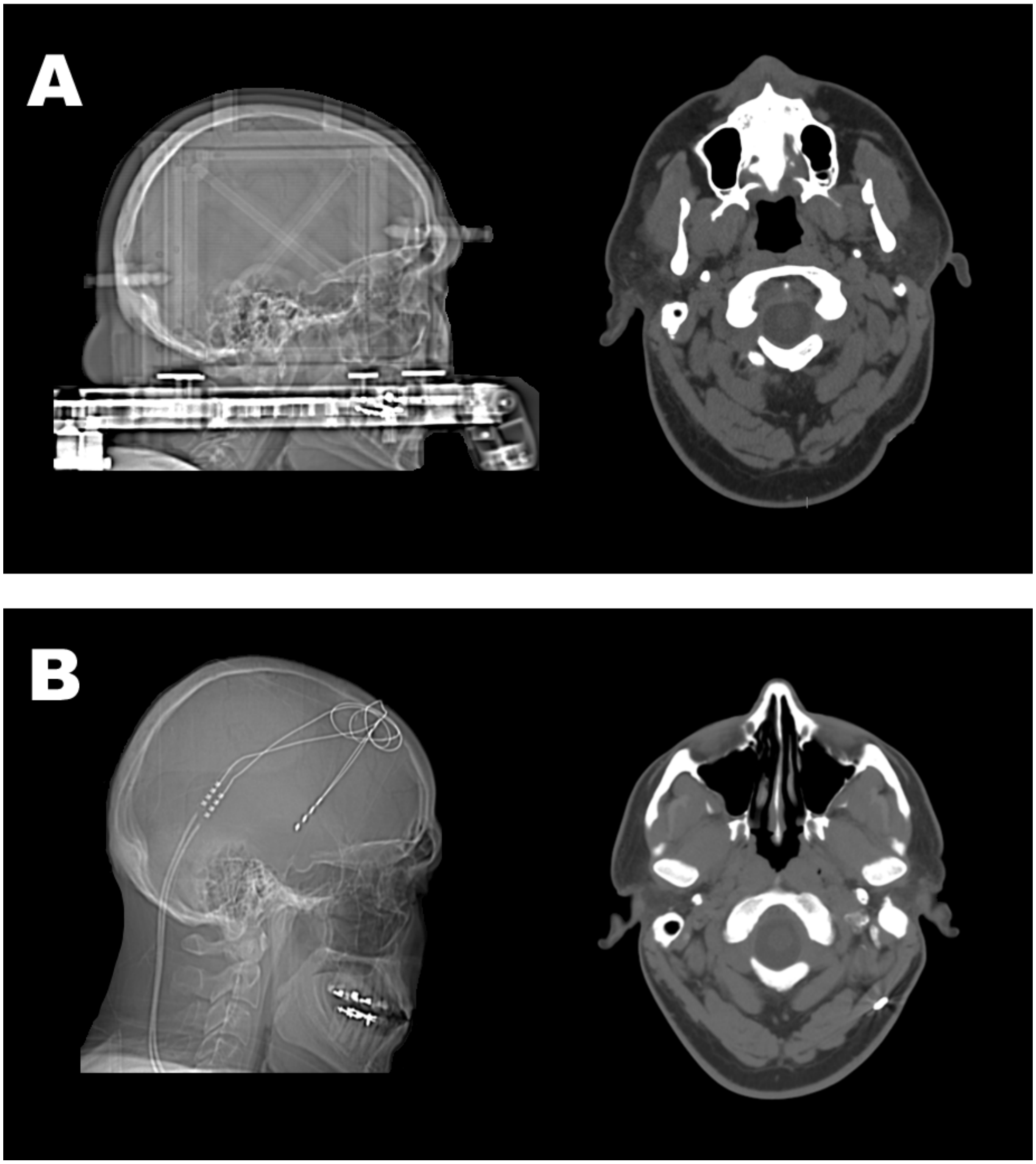


Figure 2.3: Sagittal and transverse computed tomography (CT) images of the brain and skull of MA. We show here sagittal and transverse sections taken from CT scans. Imaging was performed before (**A**) and after (**B**) MA received deep brain stimulation surgery for his treatment refractory OCD. Two deep brain stimulator probes can be seen to be in place from a bifrontal approach (B), with tips of the probes located in the region of the hypothalamus. Leads traverse through the left scalp soft tissues. Streak artifact from the leads somewhat obscures visualization of the adjacent bifrontal and left parietal parenchyma. We did not observe any intracranial hemorrhage, mass effect or midline shift or extra-axial fluid collection. Brain parenchyma was normal in volume and contour.

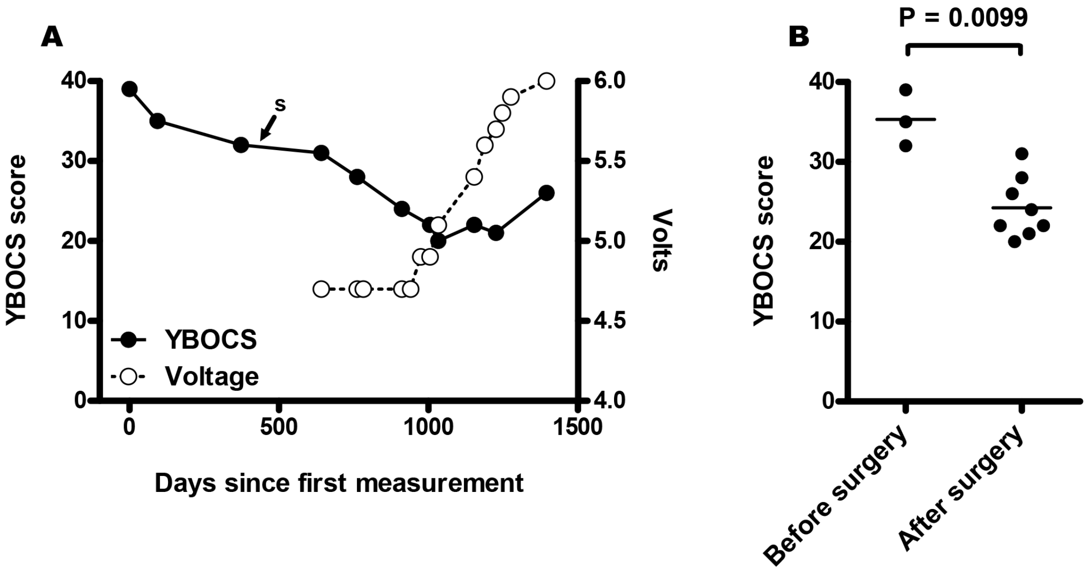


Figure 2.4: Yale Brown Obsessive Compulsive Scale (YBOCS) scores were measured for MA over a three year and seven months period of time. A time series plot (**A**) shows a steady decline in YBOCS scores over the period of time spanning his DBS surgery (s) and treatment. Incremental adjustments to neurostimulator voltage are plotted over a period of time following DBS surgery. Mean YBOCS scores are plotted for sets of measurements taken before and after Deep Brain Stimulation (DBS) surgery (**B**) (p = 0.0099, one tailed unpaired t-test with Welch's correction).

organizations. In fact, M.A. started dating and recently became engaged to be married, highlighting his improvement in daily functioning. New issues that M.A. reports are consistent tenesmus, occasional diarrhea (which he can now tolerate despite prior contamination obsessions) and improved vision (going from 20/135 to 20/40 vision, as documented by his optometrist), with him no longer needing to wear glasses. It is unknown whether the DBS implant has contributed to any of these issues. Attempts to add fluoxetine at 80 mg by mouth daily for two months to augment any efficacy from the DBS and ERP were unsuccessful, mainly due to no discernible benefit and prominent sexual side effects. M.A. still receives an injection of 37.5 mg risperidone every two weeks for his past history of psychoses; otherwise, he no longer takes any other medications. There has not been any exacerbation of psychoses in this individual during the two years of treatment with DBS.

2.3.3 CLIA certified WGS results

Illumina WGS clinical evaluations

The Illumina WGS clinical evaluation included manual annotation of 344 genes, which led to the following conclusion:

- “No pathogenic or likely pathogenic variants were found in the 344 genes evaluated that are expected to be clinically significant for the patient. The coverage for these 344 genes is at least 99%. Therefore, significant variants could exist that are not detected with this test.”

The clinical evaluation did, however, identify M.A. as a carrier for a variant (c.734G>A ,p.Arg245Gln) in PHYH, which has been associated in the autosomal recessive or compound heterozygote states with Refsum disease, which is an inherited condition that can lead to vision loss, anosmia, and a variety of other signs and symptoms. In silico prediction programs suggest little impact; however, the variant is rare with a 1000 Genomes frequency of 0.18%. In this regard, it is worth noting that M.A. has always had poor night vision and enlarged pupils, and, as a result of this genetic finding, we met with M.A.'s treatment team at his Veteran's Affair's (V.A.) medical center and learned that he had recently been diagnosed with bilateral cataracts, enlarged pupils, and vision loss. We also learned that M.A.'s mother and maternal grandfather have a history of enlarged pupils with poor vision, and we are currently following up whether this might be related in any way to this particular variant and Refsum disease.

Disease variant discovery

Further downstream analyses identified and prioritized several other potentially clinically relevant variants (Figure 2.5).

Variants that were imported into the Omicia Opal system were filtered to include those that had a high likelihood of being damaging (as defined by an Omicia score > 0.7) and those that have supporting Online Mendelian Inheritance in Man (OMIM; an online database of human genetics and genetic disorders) evidence. We chose to filter based on an Omicia Score of > 0.7 as this value derives a slightly more inclusive list of variants which still cannot be dismissed, but for which we have additional corroborating evidence (i.e., Illumina Genome Network (IGN) validation and annotation). A screen shot of the Omicia web-app showing the prioritized variants is shown in Figure 2.6. See table 2.2 for a list of those variants that met the filtering criteria. We highlight here some of the findings.

M.A. was found to be a heterozygote for a p.Val66Met change in BDNF, which encodes a protein that is a member of the nerve growth factor (NGF) family. The protein is induced by cortical neurons, and is deemed necessary for the survival of striatal neurons in the brain. In drug naïve patients, BDNF serum levels were found to be significantly decreased in OCD pa-

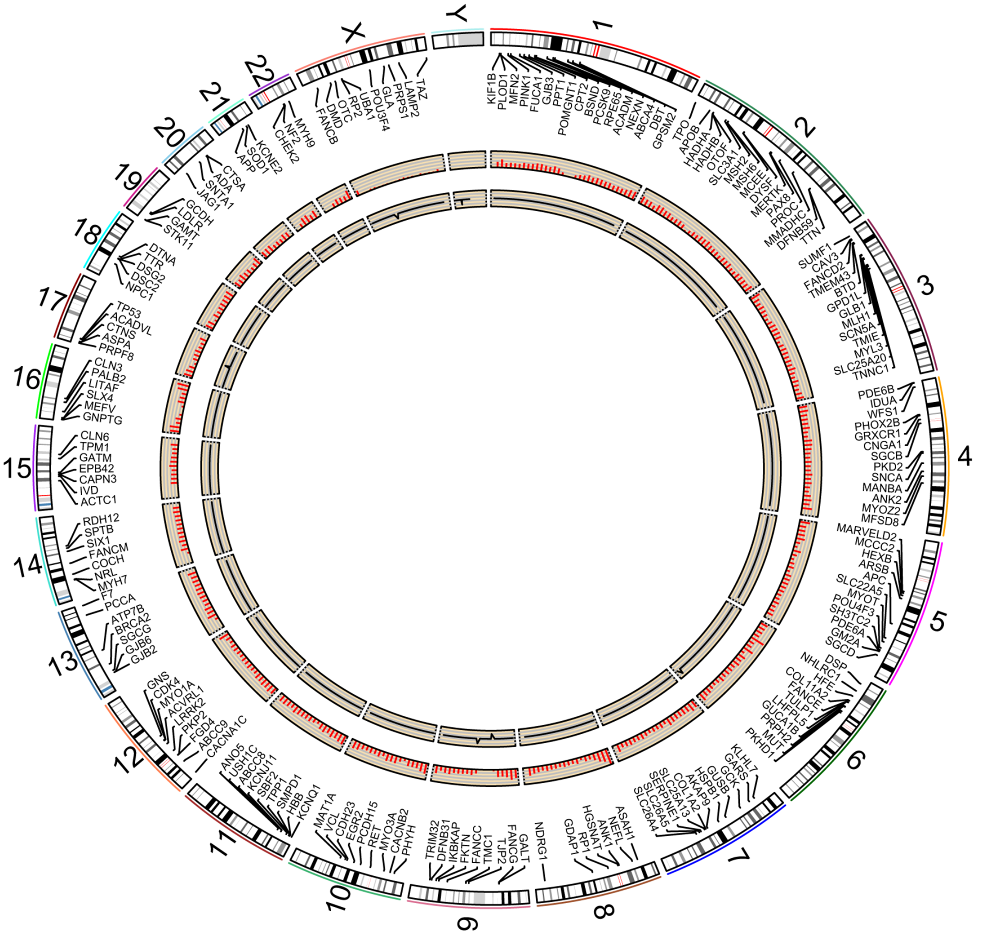
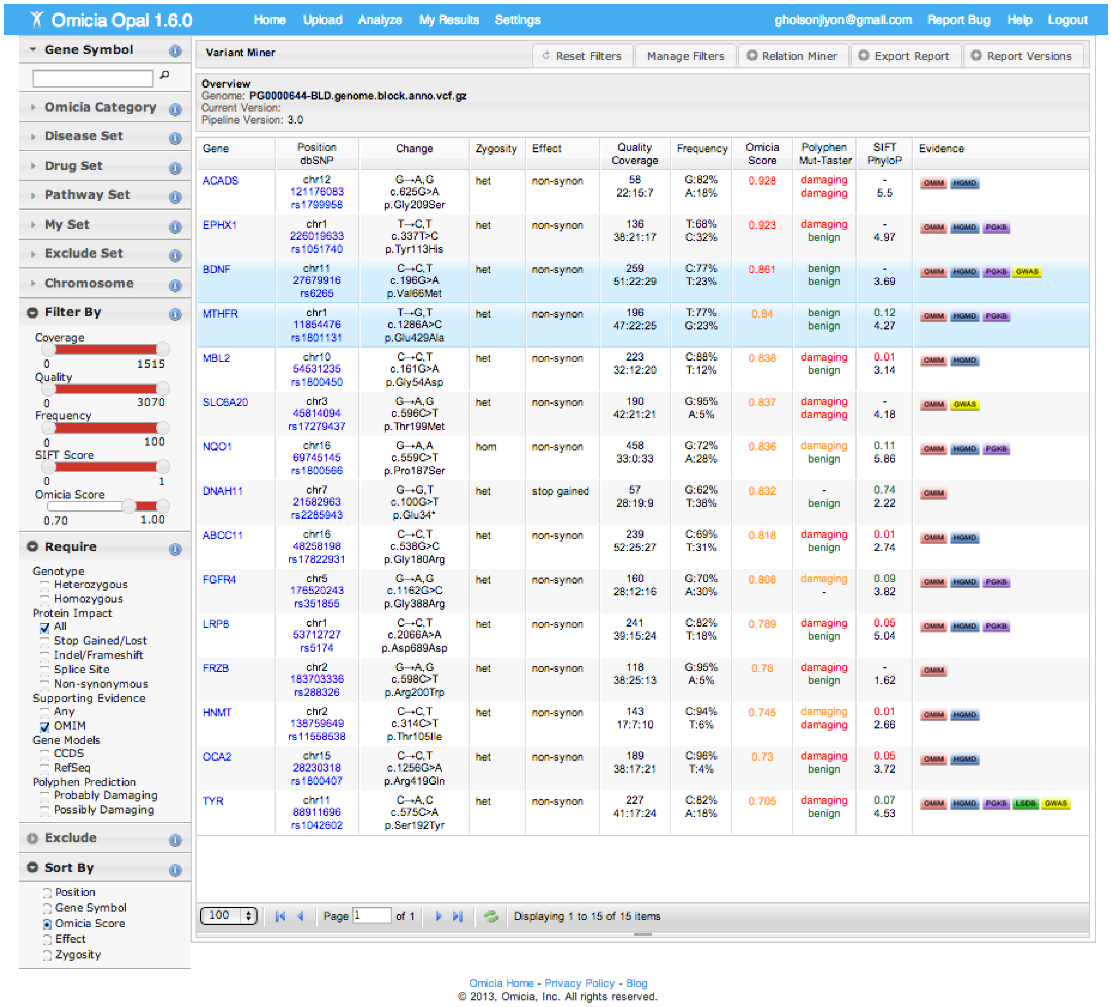


Figure 2.5: Illumina CLIA Whole genome sequencing data summarized in the form of a Circos plot. We show here a summary of the genomic coordinates corresponding to the 344 genes that were clinically evaluated by the Illumina CLIA WGS pipeline, the frequency of IGN validated SNVs across the genome (plotted in red) and a summary of highly confident copy number variants (CNVs) that were simultaneously detected by the Estimation by Read Depth with SNVs (ERDS) and Copy Number Analysis Method (CNAM) detection methods (plotted in black). Duplications and deletions are depicted as elevations and declinations, respectively.



Coverage: 0 – 1515
Quality: 0 – 3070
Frequency: 0 – 100
SIFT Score: 0 – 1
Omicia Score: 0.70 – 1.00

Genotype
□ Heterozygous
□ Homozygous

Protein Impact
☑ All
□ Stop Gained/Lost
□ Indel/Frameshift
□ Splice Site
□ Non-synonymous

Supporting Evidence
□ Any
☑ OMIM

Gene Models
□ CCDS
□ RefSeq

Polyphen Prediction
□ Probably Damaging
□ Possibly Damaging

Sort By
○ Position
○ Gene Symbol
● Omicia Score
○ Effect
○ Zygosity

Overview
Genome: **PG0000644-BLD.genome.block.anno.vcf.gz**
Current Version:
Pipeline Version: 3.0

Gene	Position dbSNP	Change	Zygosity	Effect	Quality Coverage	Frequency	Omicia Score	Polyphen Mut-Taster	SIFT PhyloP	Evidence
ACADS	chr12 121176083 rs1799958	G→A,G c.625G>A p.Gly209Ser	het	non-synon	58 22:15:7	G:82% A:18%	0.928	damaging damaging	- 5.5	OMIM HGMD
EPHX1	chr1 226019633 rs1051740	T→C,T c.337T>C p.Tyr113His	het	non-synon	136 38:21:17	T:68% C:32%	0.923	damaging benign	- 4.97	OMIM HGMD PGKB
BDNF	chr11 27679916 rs6265	C→C,T c.196G>A p.Val66Met	het	non-synon	259 51:22:29	C:77% T:23%	0.861	benign benign	- 3.69	OMIM HGMD PGKB GWAS
MTHFR	chr1 11854476 rs1801131	T→G,T c.1286A>C p.Glu429Ala	het	non-synon	196 47:22:25	T:77% G:23%	0.84	benign benign	0.12 4.27	OMIM HGMD PGKB
MBL2	chr10 54531235 rs1800450	C→C,T c.161G>A p.Gly54Asp	het	non-synon	223 32:12:20	C:88% T:12%	0.838	damaging benign	0.01 3.14	OMIM HGMD
SLC6A20	chr3 45814094 rs17279437	G→A,G c.596C>T p.Thr199Met	het	non-synon	190 42:21:21	G:95% A:5%	0.837	damaging damaging	- 4.18	OMIM GWAS
NQO1	chr16 69745145 rs1800566	G→A,A c.559C>T p.Pro187Ser	hom	non-synon	458 33:0:33	G:72% A:28%	0.836	damaging benign	0.11 5.86	OMIM HGMD PGKB
DNAH11	chr7 21582963 rs2285943	G→G,T c.100G>T p.Glu34*	het	stop gained	57 28:19:9	G:62% T:38%	0.832	- benign	0.74 2.22	OMIM
ABCC11	chr16 48258198 rs17822931	C→C,T c.538G>C p.Gly180Arg	het	non-synon	239 52:25:27	C:69% T:31%	0.818	damaging benign	0.01 2.74	OMIM HGMD
FGFR4	chr5 176520243 rs351855	G→A,G c.1162G>C p.Gly388Arg	het	non-synon	160 28:12:16	G:70% A:30%	0.808	damaging -	0.09 3.82	OMIM HGMD PGKB
LRP8	chr1 53712727 rs5174	C→C,T c.2066A>A p.Asp689Asp	het	non-synon	241 39:15:24	C:82% T:18%	0.789	damaging benign	0.05 5.04	OMIM HGMD PGKB
FRZB	chr2 183703336 rs288326	G→A,G c.598C>T p.Arg200Trp	het	non-synon	118 38:25:13	G:95% A:5%	0.76	damaging benign	- 1.62	OMIM
HNMT	chr2 138759649 rs11558538	C→C,T c.314C>T p.Thr105Ile	het	non-synon	143 17:7:10	C:94% T:6%	0.745	damaging damaging	0.01 2.66	OMIM HGMD
OCA2	chr15 28230318 rs1800407	C→C,T c.1256G>A p.Arg419Gln	het	non-synon	189 38:17:21	C:96% T:4%	0.73	damaging benign	0.05 3.72	OMIM HGMD
TYR	chr11 88911696 rs1042602	C→A,C c.575C>A p.Ser192Tyr	het	non-synon	227 41:17:24	C:82% A:18%	0.705	damaging benign	0.07 4.53	OMIM HGMD PGKB LSDB GWAS

100 Page 1 of 1 Displaying 1 to 15 of 15 items

Figure 2.6: Screen shot of the Omicia Opal system showing the list of prioritized variants.

Gene	Chrom	Position	Ref SNP	Nuc Change	AA Change	Zygosity	Effect	Validated	Quality	Coverage	Frequency	Transcript HGVS	Protein HGVS	Omicia Score	Polyphen	MutationTaster	PhyloP	SIFT
MTHFR	chr1	11854476	rs1801131	T>G,T	Glu>Ala	heterozygous	non-synon	IGN validated	196	(47:22:25)	T:77% G:23%	NM_005957.4:c.1286A>C	NP_005948:p.Glu429Ala	0.84	benign (B)	polymorphism (P)	4.27	0.12
LRP8	chr1	53712727	rs5174	C>C,T	Arg>Gln	heterozygous	non-synon	IGN validated	241	(39:15:24)	C:82% T:18%	NM_017522.3:c.2066A>A	NP_059992:p.Asp689Asp	0.789	probably damaging (D)	polymorphism (P)	5.04	0.05
EPHX1	chr1	226019633	rs1051740	T>C,T	Tyr>His	heterozygous	non-synon	IGN validated	136	(38:21:17)	T:68% C:32%	NM_000120.3:c.337T>C	NP_000111:p.Tyr113His	0.923	probably damaging (D)	polymorphism (P)	4.97	
HNMT	chr2	138759649	rs11558538	C>C,T	Thr>Ile	heterozygous	non-synon	IGN validated	143	(17:7:10)	C:94% T:6%	NM_006895.2:c.314C>T	NP_008826:p.Thr105Ile	0.745	possibly damaging (P)	disease causing (D)	2.66	0.01
FRZB	chr2	183703336	rs288326	G>A,G	Arg>Trp	heterozygous	non-synon	IGN validated	118	(38:25:13)	G:95% A:5%	NM_001463.3:c.598C>T	NP_001454:p.Arg200Trp	0.76	probably damaging (D)	polymorphism (P)	1.62	
SLC6A20	chr3	45814094	rs17279437	G>A,G	Thr>Met	heterozygous	non-synon	IGN validated	190	(42:21:21)	G:95% A:5%	NM_020208.3:c.596C>T	NP_064593:p.Thr199Met	0.837	probably damaging (D)	disease causing (D)	4.18	
FGFR4	chr5	176520243	rs351855	G>A,G	Gly>Arg	heterozygous	non-synon	IGN validated	160	(28:12:16)	G:70% A:30%	NM_002011.3:c.1162G>C	NP_002002:p.Gly388Arg	0.808	possibly damaging (P)		3.82	0.09
DNAH11	chr7	21582963	rs2285943	G>G,T	Glu>Ter	heterozygous	stop gained	IGN validated	57	(28:19:9)	G:62% T:38%			0.832		polymorphism (P)	2.22	0.74
MBL2	chr10	54531235	rs1800450	C>C,T	Gly>Asp	heterozygous	non-synon	IGN validated	223	(32:12:20)	C:88% T:12%	NM_000242.2:c.161G>A	NP_000233:p.Gly54Asp	0.838	probably damaging (D)	polymorphism (P)	3.14	0.01
BDNF	chr11	27679916	rs6265	C>C,T	Val>Met	heterozygous	non-synon	IGN validated	259	(51:22:29)	C:77% T:23%	NM_170735.5:c.196G>A	NP_733931:p.Val66Met	0.861	benign (B)	polymorphism (P)	3.69	
TYR	chr11	88911696	rs1042602	C>A,C	Ser>Tyr	heterozygous	non-synon	IGN validated	227	(41:17:24)	C:82% A:18%	NM_000372.4:c.575C>A	NP_000363:p.Ser192Tyr	0.705	probably damaging (D)	polymorphism (P)	4.53	0.07
ACADS	chr12	121176083	rs1799958	G>A,G	Gly>Ser	heterozygous	non-synon	IGN validated	58	(22:15:7)	G:82% A:18%	NM_000017.2:c.625G>A	NP_000008:p.Gly209Ser	0.928	probably damaging (D)	disease causing (D)	5.5	
OCA2	chr15	28230318	rs1800407	C>C,T	Arg>Gln	heterozygous	non-synon	IGN validated	189	(38:17:21)	C:96% T:4%	NM_000275.2:c.1256G>A	NP_000266:p.Arg419Gln	0.73	probably damaging (D)	polymorphism (P)	3.72	0.05
ABCC11	chr16	48258198	rs17822931	C>C,T	Gly>Arg	heterozygous	non-synon	IGN validated	239	(52:25:27)	C:69% T:31%	NM_033151.3:c.538G>C	NP_149163:p.Gly180Arg	0.818	probably damaging (D)	polymorphism (P)	2.74	0.01
NQO1	chr16	69745145	rs1800566	G>A,A	Pro>Ser	homozygous	non-synon		458	(33:0:33)	G:72% A:28%	NM_000903:c.559C>T	NP_000894:p.Pro187Ser	0.836	possibly damaging (P)	polymorphism (P)	5.86	0.11

Table 2.2: Variant prioritization was performed on all variants discovered by the Illumina CLIA WGS pipeline using the Omicia Opal version 1.5.0 platform. Variants were imported into the Omicia Opal cloud based clinical annotation and variant prioritization platform, and subsequently prioritized by requiring each variant to have prior evidence in OMIM and by additionally requiring each variant to be scored as having an Omicia Score of greater than 0.7.

tients when compared to controls (36.90 ± 6.42 ng/ml versus 41.59 ± 7.82 ng/ml; p = 0.043)[184], suggesting a link between this protein and OCD. Moreover, a study including 164 proband-parent trios with obsessive-compulsive disorder[107] uncovered significant evidence of an association between OCD and all of the BDNF markers that were tested, including the exact variant found here in this person, p.Val66Met. This particular variant has been further studied in a sample of 94 nuclear families[264], which included 94 probands with schizophrenia-spectrum disorders and 282 family members. The results of this study suggest that the p.Val66Met polymorphism may play a role in the phenotype of psychosis. Similar anxiety-related behavioral phenotypes have also been observed among mice and humans having the p.Val66Met variant in BDNF[286]. In humans, the amygdala mediates conditioned fear[52], normally inhibited by 'executive centers' in medial prefrontal cortex[205]. Deep brain stimulation of the pathways between medial prefrontal cortex and the amygdala increased the extinction of conditioned fear in a rat model of OCD[259]. Studies using functional magnetic resonance imaging (fMRI) demonstrate that humans with the p.Val66Met variant exhibit exaggerated activation of the amygdala in response to an emotional stimulus in comparison to controls lacking the variant[200,156]. It is thought that this variant may influence anxiety disorders by interfering with the learning of cues that signal safety rather than threat and may also lessen efficacy of treatments that rely on extinction mechanisms, such as exposure therapy[286]. In this regard, it is interesting to note that this person did indeed obtain very little benefit from exposure therapy prior to surgery.

M.A heterozygously carries the p.Glu429Ala allele in MTHFR, encoding a protein that catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co-substrate for homocysteine remethylation to methionine, and which has been shown to confer an elevated susceptibility to psychoses. Variants in MTHFR influence susceptibility to occlusive vascular disease, neural tube defects, colon cancer and acute leukemia. Variants in this gene are associated with methylenetetra-hydrofolate reductase deficiency. In addition, a meta-analysis comparing 1,211 cases of schizophrenia with 1,729 controls found that the MTHFR p.Glu429Ala allele was associated with susceptibility to schizophrenia[5] (odds ratio, 1.19; 95% CI, 1.07- 1.34; p = 0.002). According to the Venice guidelines for the assessment of cumulative evidence in genetic association studies, the MTHFR association exhibited a strong degree of epidemiologic credibility[89]. Pharmacogenetic studies have found a consistent association between the MTHFR p.Glu429Ala allele and metabolic disorder in adult, adolescent and children taking atypical antipsychotic drugs[50,299].

M.A. is also heterozygous for the p.Val108Met variant in catechol-O-methyltransferase (COMT), which catalyzes the transfer of a methyl group from S-adenosylmethionine to catecholamines, including the neurotransmitters dopamine, epinephrine, and norepinephrine. The minor allele A of this 472G>A variant produces a valine to methionine substitution, resulting in a less thermostable COMT enzyme that exhibits a 3-fold reduction in activity. A substantial body of literature implicates this variant as possibly elevating the risk for various neuropsychiatric disorders in some Caucasian populations but not necessarily in other genetic backgrounds[44,65,153,246,172,281,128]. There is some evidence that MTHFR x COMT genotype interactions might also be occurring in M.A. to influence his neuropsychiatric status[260], and the same is true for BDNF x COMT interactions[6].

Pharmacogenetic variants

Pharmacogenetic analyses were performed using the Omicia Opal platform. Pharmacogenetic variants were identified and prioritized by activating the “Drugs and Pharamcology” track within the Opal system and by requiring these variants to have prior evidence within any one of several supporting databases (i.e., OMIM, HGMD, PharmGKB, LSDB and GWAS). Prioritized variants are shown in Table 2.3. Below, we highlight pharmacogenetic variants found to be informative in terms of future medication choices for M.A.

M.A. is heterozygous for a c.19G>A p.Asp7Asn allele in ChAT, encoding choline O-acetyltransferase, which synthesizes the neuro-transmitter acetylcholine. This particular variant (rs1880676) is significantly associated with both risk for schizophrenia in Caucasians (P = 0.002), olanzapine response (P = 0.04) and for other psychopathology (P = 0.03)[186]. Allele A is associated with increased response to olanzapine in people with schizophrenia as compared to allele G. This association was significant (p= 0.04) in the Spanish cohort[186].

M.A. is homozygous for a p.Ile359Leu change in CYP2C9, and this variant has been linked to a reduction in the enzymatic activity of CYP2C9[173]. CYP2C9 encodes a member of the cytochrome P450 superfamily of enzymes. Cytochrome P450 proteins are mono-oxygenases, which catalyze many reactions associated with drug metabolism as well as reactions associated with the synthesis of cholesterol, steroids and other lipids[280]. CYP2C9 localizes to the endoplasmic reticulum and its expression is induced by rifampin. CYP2C9 is known to metabolize xenobiotics, including phenytoin, tolbutamide, ibuprofen as well as S-warfarin. Studies identifying individuals who are poor metabolizers of phenytoin and tolbutamide suggest associations between metabolism and polymorphisms found within this gene. CYP2C9 is located within a cluster of cytochrome P450 genes on chromosome 10[211]. Fluoxetine is commonly used in the treatment of OCD; it has been shown to be as effective as clomipramine and causes less side effects[239,238]. CYP2C9 acts to convert fluoxetine to R-norfluoxetine[254], and so M.A. may not be able to adequately biotransform fluoxetine[320]. However, CYP2C9 does not play a rate-limiting role for other SSRIs or clomipramine. In his own treatment experience, M.A. had no response to an 80 mg daily dose of fluoxetine, although he did experience sexual side effects at that dosage.

The protein encoded by DPYD is a pyrimidine catabolic enzyme and it acts as the initial and rate-limiting factor in uracil and thymidine catabolism pathways. M.A. was found to be a

Gene	Chrom	Position	Ref SNP	Nuc Change	AA Change	Zygosity	Effect	Validated	Quality	Coverage	Frequency	Transcript HGVS	Protein HGVS	Omicia Score	Polyphen	MutationTaster	PhyloP	SIFT
DPYD	chr1	97981395	rs1801159	T>C,T	Ile>Val	heterozygous	non-synon	IGN validated	153	(24:11:13)	T:80% C:20%	NM_000110.3:c.1627A>G	NP_000101:p.Ile543Val	0.295	benign (B)	polymorphism (P)	0.86	1
DPYD	chr1	98348885	rs1801265	G>A,A	Arg>Cys	homozygous	non-synon	IGN validated	317	(20:0:20)	G:23% A:77%	-	-	0.708	-	-	2.55	0.18
OGG1	chr3	9798773	rs1052133	C>C,G	Ile>Met	heterozygous	non-synon	IGN validated	146	(30:16:14)	C:70% G:30%	-	-	0.258	-	-	-0.25	0.01
OGG1	chr3	9798773	rs1052133	C>C,G	Pro>Ala	heterozygous	non-synon	IGN validated	146	(30:16:14)	C:70% G:30%	NM_016820.3:c.994C>G	NP_058213:p.Pro332Ala	0.258	-	-	-0.25	0.01
OGG1	chr3	9798773	rs1052133	C>C,G	Ser>Cys	heterozygous	non-synon	IGN validated	146	(30:16:14)	C:70% G:30%	NM_002542.5:c.977C>G	NP_002533:p.Ser326Cys	0.258	-	-	-0.25	0.01
CYP3A7	chr7	99306685	rs2257401	C>G,G	Arg>Thr	homozygous	non-synon	IGN validated	331	(22:0:22)	C:27% G:73%	NM_000765:c.1226G>C	NP_000756:p.Arg409Thr	0.163	benign (B)	polymorphism (P)	0.35	0.16
NAT2	chr8	18257854	rs1801280	T>C,T	Ile>Thr	heterozygous	non-synon	IGN validated	191	(39:20:19)	T:70% C:30%	NM_000015.2:c.341T>C	NP_000006:p.Ile114Thr	0.467	benign (B)	polymorphism (P)	0.74	0.08
NAT2	chr8	18258103	rs1799930	G>A,G	Arg>Gln	heterozygous	non-synon	IGN validated	220	(37:16:21)	G:76% A:24%	NM_000015.2:c.590G>A	NP_000006:p.Arg197Gln	0.653	probably damaging (D)	polymorphism (P)	3.11	0.08
NAT2	chr8	18258316	rs1208	G>A,G	Arg>Lys	heterozygous	non-synon	IGN validated	165	(26:12:14)	G:32% A:68%	NM_000015.2:c.803G>A	NP_000006:p.Arg268Lys	0.087	benign (B)	polymorphism (P)	-0.12	1
NBN	chr8	90990479	rs1805794	C>C,G	Glu>Gln	heterozygous	non-synon	IGN validated	193	(30:12:18)	C:67% G:33%	NM_002485.4:c.553G>C	NP_002476:p.Glu185Gln	0.172	benign (B)	polymorphism (P)	0.5	1
CYP11B2	chr8	143996539	rs4539	T>C,T	Lys>Arg	heterozygous	non-synon	IGN validated	78	(16:8:8)	T:64% C:36%	NM_000498.3:c.518A>G	NP_000489:p.Lys173Arg	0.081	benign (B)	polymorphism (N)	-1.15	0.63
ABCA1	chr9	107562804	rs2230808	T>C,C	Lys>Arg	homozygous	non-synon	IGN validated	536	(38:0:38)	T:41% C:59%	NM_005502:c.4760A>G	NP_005493:p.Lys1587Arg	0.7	benign (B)	polymorphism (P)	4.87	1
ABCA1	chr9	107589255	rs2066718	C>C,T	Val>Met	heterozygous	non-synon	IGN validated	195	(40:19:21)	C:94% T:6%	NM_005502.3:c.2311G>A	NP_005493:p.Val771Met	0.562	benign (B)	disease causing (D)	1.4	1
ABCA1	chr9	107620867	rs2230806	C>C,T	Arg>Lys	heterozygous	non-synon	IGN validated	131	(30:18:12)	C:58% T:42%	NM_005502.3:c.656G>A	NP_005493:p.Arg219Lys	0.187	benign (B)	polymorphism (P)	0.16	0.32
CYP2C19	chr10	96602623	rs3758581	G>A,A	Val>Ile	homozygous	non-synon	IGN validated	634	(48:0:48)	G:96% A:4%	NM_000769:c.992G>A	NP_000760:p.Val331Ile	0.082	benign (B)	polymorphism (P)	-1.53	1
CYP2C9	chr10	96741053	rs1057910	A>C,C	Ile>Leu	homozygous	non-synon	IGN validated	496	(36:0:36)	A:96% C:4%	NM_000771:c.1076A>C	NP_000762:p.Ile359Leu	0.189	benign (B)	disease causing (D)	-	0.11
CETP	chr16	57016092	rs5882	G>A,G	Val>Ile	heterozygous	non-synon	IGN validated	203	(36:15:21)	G:45% A:55%	NM_000078.2:c.1264G>A	NP_000069:p.Val422Ile	0.088	benign (B)	polymorphism (P)	-1.43	1
NQO1	chr16	69745145	rs1800566	G>A,A	Pro>Ser	homozygous	non-synon	IGN validated	458	(33:0:33)	G:72% A:28%	NM_000903:c.559C>T	NP_000894:p.Pro187Ser	0.836	possibly damaging (P)	polymorphism (P)	5.86	0.11
CYP4F12	chr19	15789140	rs609290	A>G,G	Ile>Val	homozygous	non-synon	IGN validated	578	(44:0:44)	A:6% G:94%	NM_023944:c.269A>G	NP_076433:p.Ile90Val	0.126	-	polymorphism (P)	-0.6	0.7
CYP4F12	chr19	15789140	rs609290	A>G,G		homozygous	splice site	IGN validated	578	(44:0:44)	A:6% G:94%	-	-	0.172	-	-	-0.6	
CYP4F12	chr19	15793235	rs79526533	T>C,T	Cys>Arg	heterozygous	non-synon	IGN validated	155	(30:16:14)	T:43% C:57%	NM_023944.2:c.562T>C	NP_076433:p.Cys188Arg	0.083	benign (B)	polymorphism (N)	-0.36	1
CYP4F2	chr19	15990431	rs2108622	C>C,T	Val>Met	heterozygous	non-synon	IGN validated	183	(30:12:18)	C:78% T:22%	NM_001082.3:c.1297G>A	NP_001073:p.Val433Met	0.473	probably damaging (D)	polymorphism (P)	2.31	0.01
CYP2B6	chr19	41515263	rs28399497	A>A,G	Lys>Arg	heterozygous	non-synon	IGN validated	54	(17:8:9)	-	NM_000767.4:c.785A>G	NP_000758:p.Lys262Arg	0.178	benign (B)	polymorphism (N)	0.84	1
CYP2B6	chr19	41522715	rs28399500	C>C,T	Arg>Cys	heterozygous	non-synon	IGN validated	190	(26:10:16)	C:95% T:5%	NM_000767.4:c.1459C>T	NP_000758:p.Arg487Cys	0.088	benign (B)	polymorphism (N)	-0.88	1

Table 2.3: Variants discovered by the Illumina CLIA WGS pipeline to have pharmacogenomic significance were evaluated and prioritized using the Omicia Opal version 1.5.0 platform. Pharmacogenomic variant prioritization was performed by importing all variants called by the Illumina CLIA WGS and bioinformatics pipeline into the Omicia Opal cloud based variant prioritization platform. Variants were filtered by activating the "Drugs and Pharmacology" track in Opal, and further filtered to those that also had prior evidence in a variety of supporting databases, including: OMIM, HGMD, PharmGKB, LSDB and GWAS.

carrier of two variants in this gene, p.Ile543Val and p.Arg29Cys, for which he is a heterozygote and homozygote, respectively. Variants within DPYD result in dihydropyrimidine dehydrogenase deficiency, an error in pyrimidine metabolism associated with thymine-uraciluria and an increased risk of toxicity in cancer patients receiving 5-fluorouracil chemotherapy. Two transcript variants encoding different isoforms have been described for DPYD[134,278].

M.A. is heterozygous both for a c.590G>A p.Arg197Gln allele (rs1799930) and a c.803G>A p.Arg268Lys allele (rs1208) in NAT2, encoding an enzyme that functions to both activate and deactivate arylamine and hydrazine drugs and carcinogens[267,25]. Genotype AG for rs1799930 is associated with increased risk of toxic liver disease in people with tuberculosis when treated with ethambutol, isoniazid, pyrazinamide and rifampin as compared to genotype GG. Allele G for rs1208 is not associated with risk of hypersensitivity when treated with sulfamethoxazole and trimethoprim in people with infection.

Copy number variants

ERDS identified 60 putative CNVs, all of which were visually inspected within the Golden Helix Genome Browser. Many of the CNVs detected by the ERDS method were found to be located within chromosomal boundary regions and were determined to be false positives due to highly variable read depth in these regions. The CNAM method detected 35 putative CNVs, which were visually inspected by plotting the LogR and covariate values in Golden Helix SVS. Only six CNVs were simultaneously detected by both the ERDS and CNAM methods, and were visually inspected as further confirmation to be included among the set of highly confident CNVs (Table 2.4). To our knowledge, these CNVs have not been previously associated in any way with M.A.'s disease phenotype, but we are archiving these results for future analysis as knowledge of CNVs and disease associations expands.

Region	Gene	Type	Length
chr6:161601836-161816660	AGPAT4,PARK2	DUP	214824
chr9:69513401-69788000	ncRNA_exonic: LOC100133920	DUP	274599
chr9:43425001-43705991	CNTNAP3B,SPATA31A6	DEL	280990
chr17:41409959-41635571	ARL4D,DHX8,ETV4	DUP	225612
chrX:52103601-52395800	XAGE1A,XAGE1B,XAGE1C,XAGE1D,XAGE1E,XAGE2,XAGE2B	DEL	292199
chrY:13239001-13547000	GYG2P1(dist=970915)	DEL	307999

Table 2.4: A list of 6 high confidence copy number variants (CNVs) that were called by both the ERDS and CNAM CNV detection methods. ERDS (version 1.06.04) derived CNVs were required to be >200 kb in length, with confidence scores of >300. CNAM (Golden Helix SVS version 7.7.5) CNVs were also required to be >200 kb in length with average segment LogR values of > 0.15 and < -0.15 for duplications and deletions, respectively. CNVs detected by both methods were visually inspected to eliminate obvious false positive calls. The 6 CNVs shown here were detected by each method, visually confirmed, and are thus considered high confidence.

Return of results

A board-certified genetic counselor was consulted by GJL prior to returning results, and all therapy and counseling was provided by GJL. Although we believe in archiving and managing all genetic results and not just a small subset of genes, we did analyze the 57 genes that are currently recommended for "return of results" by the American College of Medical Genetics. These results are shown in Table 2.5, and one of us (GJL) met with M.A. to go over the results with him, along with adding some of the findings into his paper-based medical record. Lastly, we did contact the physicians and other officials at the U.S. Veterans Affairs office to offer to incorporate these data into the electronic medical record for M.A. at the VA, but we were informed that the VistA health information system (HIS)[45,242,151,26] does not currently have the capability to incorporate any genomic variant data.

2.4 Discussion

2.4.1 DBS for treatment-refractory OCD

Deep brain stimulation for M.A.'s treatment refractory OCD has provided a quantifiable and significant improvement in the management of his symptoms. M.A. has regained a quality of life that he had previously not experienced in over 15 years, which is highlighted by him participating in regular exercise, working as a volunteer in his local church, dating, and eventually getting married, all of which act to illustrate a dramatic improvement in his daily functioning since receiving DBS treatment for his OCD.

One significant aspect of this study is the rechargeable, and hence depleteable, nature of the Activa RC neurostimulator battery, which serves to illustrate the efficacy of DBS for OCD for this individual. On one such illustrative occasion, M.A. forgot to take the recharging de-

Gene	Chrom	Position	Ref SNP	Nuc Change	AA Change	Zygosity	Effect	Quality	Coverage	Frequency	Transcript HGVS	Protein HGVS	Omicia Score	Polyphen	MutationTaster	PhyloP	SIFT
MUTYH	chr1	45797505	rs3219489	C>G,G	Gln>His	homozygous	non-synon	347	(24:0:24)	C:68% G:32%	NM_001048172:c.933G>C	NP_001041637:p.Gln311His	0.096	benign (B)	polymorphism (P)	-0.08	0.14
PCSK9	chr1	55505732	rs2495482	A>G,G	-	homozygous	splice site	471	(36:0:36)	A:9% G:91%	NM_174936:c.207+15A>G	-	0.144	-	-	0.02	-
PCSK9	chr1	55524197	rs540796	A>G,G	Tyr>Cys	homozygous	non-synon	396	(27:0:27)	A:14% G:86%	-	-	0.341	-	-	1.21	0.16
PCSK9	chr1	55524237	rs562556	G>A,A	Pro>Pro	homozygous	non-synon	356	(24:0:24)	G:15% A:85%	NM_174936:c.1421G>A	NP_777596:p.Pro474Pro	0.129	-	-	-1.6	-
PCSK9	chr1	55529187	rs505151	G>A,A	Gly>Glu	homozygous	non-synon	505	(37:0:37)	G:10% A:90%	NM_174936:c.2010G>A	NP_777596:p.Gly670Glu	0.093	-	polymorphism (P)	0.07	1
CACNA1S	chr1	201009182	rs12139527	A>A,G	Leu>Ser	heterozygous	non-synon	208	(38:16:22)	A:79% G:21%	NM_000069.2:c.5399T>C	NP_000060:p.Leu1800Ser	0.61	probably damaging (D)	polymorphism (P)	2.47	-
CACNA1S	chr1	201020105	rs6702590	A>G,G	-	homozygous	splice site	443	(31:0:31)	A:16% G:84%	NM_000069:c.4114+5T>C	-	0.131	-	-	-1.82	-
CACNA1S	chr1	201020165	-	T>A,T	Thr>Ser	heterozygous	non-synon	225	(52:26:26)	-	NM_000069:c.4060A>T	NP_000060:p.Thr1354Ser	0.551	benign (B)	polymorphism (N)	2.68	1
CACNA1S	chr1	201052310	rs12742169	A>A,T	Leu>His	heterozygous	non-synon	142	(22:8:14)	A:83% T:17%	NM_000069.2:c.1373T>A	NP_000060:p.Leu458His	0.508	benign (B)	polymorphism (N)	2.16	0.2
RYR2	chr1	237620049	rs2045955	T>C,C	-	homozygous	splice site	416	(29:0:29)	T:46% C:54%	-	-	0.134	-	-	-0.53	-
RYR2	chr1	237841390	rs34967813	A>G,G	Gln>Arg	homozygous	non-synon	522	(36:0:36)	A:86% G:14%	-	-	0.59	-	polymorphism (N)	3.25	0.42
RYR2	chr1	237893674	rs2797445	C>T,T	-	homozygous	splice site	354	(25:0:25)	C:1% T:99%	-	-	0.145	-	-	0.17	-
RYR2	chr1	237957161	rs790901	A>G,G	-	homozygous	splice site	540	(41:0:41)	A:25% G:75%	-	-	0.145	-	-	0.04	-
RYR2	chr1	237957309	rs790900	A>C,C	-	homozygous	splice site	487	(35:0:35)	A:26% C:74%	-	-	0.244	-	-	-0.38	-
APOB	chr2	21225281	rs1042034	C>C,T	Ser>Asn	heterozygous	non-synon	98	(31:22:9)	C:34% T:66%	NM_000384.2:c.13013G>A	NP_000375:p.Ser4338Asn	0.087	benign (B)	polymorphism (P)	0.23	1
APOB	chr2	21231524	rs676210	G>A,G	Pro>Leu	heterozygous	non-synon	194	(52:29:23)	G:66% A:34%	NM_000384.2:c.8216C>T	NP_000375:p.Pro2739Leu	0.774	possibly damaging (P)	polymorphism (P)	5.95	-
APOB	chr2	21232803	rs584542	T>C,C	Ile>Val	homozygous	non-synon	556	(41:0:41)	T:3% C:97%	NM_000384:c.6937A>G	NP_000375:p.Ile2313Val	0.081	benign (B)	polymorphism (N)	-1.24	0.57
APOB	chr2	21235475	rs568413	T>C,C	Tyr>Cys	homozygous	non-synon	496	(35:0:35)	-	NM_000384:c.4265A>G	NP_000375:p.Tyr1422Cys	0.085	benign (B)	polymorphism (N)	0.21	0.46
APOB	chr2	21250914	rs679899	G>A,G	Ala>Val	heterozygous	non-synon	154	(35:19:16)	G:51% A:49%	NM_000384.2:c.1853C>T	NP_000375:p.Ala618Val	0.763	benign (B)	polymorphism (P)	4.94	-
MSH6	chr2	48026308	rs2020908	C>C,G	Leu>Val	heterozygous	non-synon	153	(37:21:16)	-	NM_000179.2:c.1186C>G	NP_000170:p.Leu396Val	0.59	benign (B)	disease causing (D)	-0.04	0.11
COL3A1	chr2	189875421	rs1516446	T>G,G	His>Gln	homozygous	non-synon	630	(47:0:47)	-	NM_000090:c.4060T>G	NP_000081:p.His1354Gln	0.182	-	polymorphism (N)	1.02	1
TGFBR2	chr3	30686414	rs1155705	A>G,G	-	homozygous	splice site	594	(43:0:43)	A:57% G:43%	NM_003242:c.338+7A>G	-	0.129	-	-	-1.33	-
MLH1	chr3	37053568	rs1799977	A>A,G	Ile>Leu	heterozygous	non-synon	183	(31:14:17)	A:83% G:17%	NM_000249.3:c.655A>C	NP_000240:p.Ile219Leu	0.591	benign (B)	polymorphism (P)	2.59	0.14
MLH1	chr3	37053568	rs1799977	A>A,G	Ile>Val	heterozygous	non-synon	183	(31:14:17)	A:83% G:17%	NM_000249.3:c.655A>G	NP_000240:p.Ile219Val	0.549	benign (B)	polymorphism (P)	2.59	0.27
MYLK	chr3	123440967	rs820355	G>A,A	-	homozygous	splice site	501	(36:0:36)	G:7% A:93%	NM_053026:c.1805+6C>T	-	0.446	-	-	0.86	-
MYLK	chr3	123451773	rs9833275	G>C,C	Leu>Val	homozygous	non-synon	534	(39:0:39)	-	-	-	0.346	benign (B)	polymorphism (N)	2.47	0.46
MYLK	chr3	123451932	rs35156360	G>A,G	Pro>Ser	heterozygous	non-synon	188	(38:17:21)	G:99% A:1%	NM_053025.3:c.1327C>T	NP_444253:p.Pro443Ser	0.841	probably damaging (D)	disease causing (D)	2.77	0.07
MYLK	chr3	123457893	rs9840993	G>A,A	Pro>Ser	homozygous	non-synon	538	(39:0:39)	G:12% A:88%	NM_053026:c.439C>T	NP_444254:p.Pro147Ser	0.191	benign (B)	polymorphism (P)	0.84	0.35
MYLK	chr3	123512627	rs28497577	G>G,T	Pro>His	heterozygous	non-synon	142	(37:22:15)	G:84% T:16%	NM_053026.3:c.62C>A	NP_444254:p.Pro21His	0.187	benign (B)	polymorphism (P)	0.02	0.3
APC	chr5	112176756	rs459552	T>A,A	Val>Asp	homozygous	non-synon	360	(24:0:24)	T:14% A:86%	NM_000038:c.5466T>A	NP_000029:p.Val1822Asp	0.182	benign (B)	polymorphism (P)	1.16	-
PMS2	chr7	6026775	rs41534544	T>C,C	Lys>Glu	homozygous	non-synon	358	(24:0:24)	T:12% C:88%	NM_000535:c.1621A>G	NP_000526:p.Lys541Glu	0.303		polymorphism (P)	2.04	1
PMS2	chr7	6026988	rs1805321	G>A,G	Pro>Ser	heterozygous	non-synon	158	(22:9:13)	G:61% A:39%	NM_000535.5:c.1408C>T	NP_000526:p.Pro470Ser	0.109	benign (B)	polymorphism (N)	-0.01	0.46
RET	chr10	43610119	rs1799939	G>A,G	Gly>Ser	heterozygous	non-synon	217	(41:17:24)	G:84% A:16%	NM_020975.4:c.2071G>A	NP_066124:p.Gly691Ser	0.297	possibly damaging (P)	polymorphism (P)	0.84	0.62
MEN1	chr11	64572018	rs2959656	T>C,C	Thr>Ala	homozygous	non-synon	490	(35:0:35)	T:15% C:85%	-	-	0.234	-	polymorphism (P)	1.38	0.93
SDHD	chr11	111963860	rs592626	A>G,G	Gln>Arg	homozygous	non-synon	368	(26:0:26)	A:4% G:96%	-	-	0.129	-	-	-2.52	0.17
BRCA2	chr13	32906480	rs766173	A>A,C	Asn>Asp	heterozygous	non-synon	132	(32:19:13)	A:94% C:6%	NM_000059.3:c.865A>G	NP_000050:p.Asn289Asp	0.201	probably damaging (D)	polymorphism (N)	0.18	0.04
BRCA2	chr13	32906480	rs766173	A>A,C	Asn>His	heterozygous	non-synon	132	(32:19:13)	A:94% C:6%	NM_000059.3:c.865A>C	NP_000050:p.Asn289His	0.204	probably damaging (D)	polymorphism (N)	0.18	0.01
BRCA2	chr13	32906480	rs766173	A>A,C	Asn>Tyr	heterozygous	non-synon	132	(32:19:13)	A:94% C:6%	NM_000059.3:c.865A>T	NP_000050:p.Asn289Tyr	0.208	probably damaging (D)	polymorphism (N)	0.18	-
BRCA2	chr13	32911463	rs1799944	A>A,G	Asn>Asp	heterozygous	non-synon	193	(31:13:18)	A:94% G:6%	NM_000059.3:c.2971A>G	NP_000050:p.Asn991Asp	0.095	benign (B)	polymorphism (P)	0.18	1
BRCA2	chr13	32929387	rs169547	T>C,C	Val>Ala	homozygous	non-synon	540	(38:0:38)	T:2% C:98%	-	-	0.346	-	polymorphism (N)	2.22	1
FBN1	chr15	48720526	rs363832	G>C,C	-	homozygous	splice site	400	(28:0:28)	G:34% C:66%	NM_000138:c.6998+15C>G	-	0.332	-	-	1.11	-
FBN1	chr15	48807637	rs4775765	C>T,T	Cys>Tyr	homozygous	non-synon	465	(33:0:33)	-	NM_000138:c.1415G>A	NP_000129:p.Cys472Tyr	0.35	-	polymorphism (N)	3.17	1
MYH11	chr16	15820863	rs16967494	C>C,T	Ala>Thr	heterozygous	non-synon	209	(43:21:22)	C:76% T:24%	NM_001040114.1:c.3721G>A	NP_001035203:p.Ala1241Thr	0.462	benign (B)	polymorphism (P)	1.05	0.39
TP53	chr17	7579472	rs1042522	G>C,C	Pro>Arg	homozygous	non-synon	356	(26:0:26)	G:40% C:60%	-	-	0.489	benign (B)	polymorphism (P)	3.31	0.17
BRCA1	chr17	41223094	rs1799966	T>C,T	Ser>Gly	heterozygous	non-synon	155	(35:20:15)	T:67% C:33%	NM_007297:c.4696A>G	NP_009228:p.Ser1566Gly	0.201	possibly damaging (P)	polymorphism (P)	0.26	0.05
BRCA1	chr17	41234470	rs1060915	A>A,G	Leu>Pro	heterozygous	non-synon	161	(37:19:18)	A:70% G:30%	-	-	0.257	-	-	-0.05	-
BRCA1	chr17	41244000	rs16942	T>C,T	Lys>Arg	heterozygous	non-synon	124	(36:23:13)	T:68% C:32%	NM_007300.3:c.3548A>G	NP_009231:p.Lys1183Arg	0.081	benign (B)	polymorphism (P)	-0.68	1
BRCA1	chr17	41244435	rs16941	T>C,T	Glu>Gly	heterozygous	non-synon	149	(34:20:14)	T:70% C:30%	NM_007300.3:c.3113A>G	NP_009231:p.Glu1038Gly	0.191	possibly damaging (P)	polymorphism (P)	-0.3	0.04
BRCA1	chr17	41244936	rs799917	G>A,G	Pro>Leu	heterozygous	non-synon	172	(50:29:21)	G:52% A:48%	NM_007297:c.2471C>T	NP_009228:p.Pro824Leu	0.322	benign (B)	polymorphism (P)	1.73	1
BRCA1	chr17	41245471	rs4986850	C>C,T	Asp>Asn	heterozygous	non-synon	208	(34:12:22)	C:96% T:4%	NM_007300.3:c.2077G>A	NP_009231:p.Asp693Asn	0.205	benign (B)	polymorphism (P)	0.27	0.01
LDLR	chr19	11221454	rs2738442	T>C,C	-	homozygous	splice site	491	(38:0:38)	-	NM_000527:c.1060+7T>C	-	0.142	-	-	-0.15	-
LDLR	chr19	11221457	rs12710260	G>C,C	-	homozygous	splice site	465	(36:0:36)	G:72% C:28%	NM_000527:c.1060+10G>C	-	0.245	-	-	-1.19	-
TNNI3	chr19	55665584	rs7252610	A>C,C	-	homozygous	splice site	480	(35:0:35)	-	NM_000363:c.373-12T>G	-	0.132	-	-	-0.45	-

Table 2.5: 57 genes recommended by the ACMG as candidates for returning results were analyzed and annotated by the Omicia Opal system. Only two variants, one in CACNA1S and one in MYLK, were interpreted as being of putative interest but not rising to the level of "pathogenicity".

vice on a four-day weekend trip. Once his battery was depleted, all of his symptoms gradually returned to their full level over a 24 hour period, including severe OCD, depression and suicidality. Since that episode, M.A. always takes his recharging device with him on extended trips, but there have been other such instances in which his battery has become depleted for several hours, with the noticeable and intense return of his OCD symptoms and the cessation of his tenesmus. The electrical stimulation is having a demonstrable effect on his OCD, and these data are complementary to other data-sets involving turning DBS devices off for one week at a time[80].

There are many ethical and regulatory issues relating to deep brain stimulation that have been discussed elsewhere[83,290,86,84,85,82,70], and we report here our one positive experience, made possible when the US Food and Drug Administration granted a Humanitarian Device Exemption (HDE) to allow clinicians to use this intervention. The rechargeable nature of the new battery has been reassuring to M.A., as he is able to exert self-control over his battery life, whereas he previously had no control with the original “single-use” battery that must be replaced when the battery depletes (usually at least once annually). We assume that other persons treated with DBS for OCD will likely also start receiving rechargeable batteries. In this regard, it is worth noting that the recent development of an injectable class of cellular-scale optoelectronics paves the way for implanted wireless devices[140], and we fully expect that there will be more brain-machine neural interfaces used in humans in the future[3,4,231,293,218].

2.4.2 Clinical WGS

There are still many challenges in showing how any one mutation can contribute toward a clear phenotype, particularly in the context of genetic background and possible environmental influences[204]. Bioinformatics confounders, such as poor data quality[221], sequence inaccuracy, and variation introduced by different methodological approaches[225] can further complicate biological and genetic inferences. Although the variants discussed in the results section of our study have been previously associated with mental disease, we caution that the data presented are not sufficient to implicate any particular mutation as being necessary or sufficient to lead to the described phenotype, particularly given that mental illness results from a complex interaction of any human with their surrounding environment and social support structures. The genetic architecture of most neuropsychiatric illness is still largely undefined and controversial[141,198,197,301]. We provide our study as a cautionary one: WGS cannot act as a diagnostic and prognostic panacea for neuropsychiatric disorders, but instead could act to elucidate risk factors for psychiatric disease and pharmacogenetic variants that can inform future medication treatments.

During our study, we found that M.A. carries at least three alleles that have been associated with neuropsychiatric phenotypes, including variants in BDNF, MTHFR, and ChAT (Table 2.6). And, although we have discovered informative phamacogenetic variants in this person, these discoveries have not led to the immediate alteration of this person's medication schema. We have archived these discoveries and expect that these variants will be useful over the course of his life-long medical care. We feel that this information is inherently valuable, as one can never predict with certainty what the future might hold, and a more complete medical profile on individual patients will facilitate more informed medical choices.

Integrating WGS data into the Electronic Medical Health Record

In the context of the incomplete, and sometimes proprietary, nature of human gene mutation databases, it is likely that analyses and medical guidance gleaned from these WGS data will differ from institution to institution. It is therefor important that people be given the opportunity, like with many other traditional medical tests, to obtain "second opinions". For this to be possible, one must accurately describe the contents of short-read sequencing data in terms

Gene name	Genomic coordinates	Amino acid change	Zygosity	Variation type	Population frequency	Clinical significance
MTHFR	chr1: 11854476	Glu > Ala	heterozygous	non-synon	T:77% G:23%	Susceptibility to psychoses, schizophrenia occlusive vascular disease, neural tube defects, colon cancer, acute leukemia, and methylenetetrahydrofolate reductase deficiency
BDNF	chr11: 27679916	Val > Met	heterozygous	non-synon	C:77% T:23%	Susceptibility to OCD, psychosis, and diminished response to exposure therapy
CHAT	chr10: 50824117	Asp > Asn	heterozygous	non-synon	G:85% A:15%	Susceptibility to schizophrenia and other psychopathological disorders.

Table 2.6: A summary of three clinically relevant alleles found in the sequencing results of MA. Variations in MTHFR, BDNF, and ChAT were found to be of potential clinical relevance for this person as they are all implicated in contributing to the susceptibility and development of many neuropsychiatric disorders that resemble those present within MA. A brief summary of the characteristics of each variation is shown, including the gene name, genomic coordinates, amino acid change, zygosity, variation type, estimated population frequency and putative clinical significance.

of the existing electronic medical health standards, so that these data can be incorporated into an electronic medical health record. Accurately describing the contents of next generation sequencing (NGS) results is particularly critical for clinical analysis of genomic data. However, genomics and medicine use different, often incompatible terminologies and standards to describe sequence variants and their functional effects. In our efforts to treat this one person with severe mental illness, we have implemented the GVFclin format for the variants that were discovered during the sequencing of his whole genome. We hope to eventually incorporate his genetic data into his electronic health record, if and when the VistA health information system (HIS)[45,242,151,26] is upgraded to allow entry of such data. We did already counsel M.A. regarding several genetic variants that may be clinically relevant to predisposing him to his psychiatric disorder[21].

Returning genetic results

There is considerable controversy in the field of medical genetics concerning the extent of return of genetic results to people, particularly in the context of “secondary”, “unrelated”, “unanticipated” or “incidental” findings stemming from new high-throughput sequencing techniques. Some people have concerns regarding the clinical utility of much of the data, and in response have advocated for selectively restricting the returnable medical content. One such set of recommendations has been provided by the American College of Medical Genetics which recently released guidelines in which they recommended the “return of secondary findings” for 57 genes, without detailed guidance for the rest of the genome. These types of recommendations take a more paternalistic approach in returning test results to people, and generally involve a deciding body of people that can range in size from a single medical practitioner to a committee of experts. We believe that anyone should be able to access and manage their own genome data[316], just like how anyone should be able to own and manage their medical and radiology test results, particularly if the testing is performed with suitably appropriate clinical standards in place, i.e. CLIA in America[178,177]. In this regard, we found by means of WGS that M.A carries a variant in PHYH, this revelation ended up improving his care despite not being related in any known or direct way to his psychiatric disorder, which is the main focus of this study. As stated in our Results section, M.A. has been diagnosed with bilateral cataracts and has been counseled in ways to reduce further damage to and deterioration of his vision.

2.5 Concluding remarks

One can learn a substantial amount from detailed study of particular individuals[268,269,175,176,297,245,68,311], and we believe that we are entering an era of precision medicine in which we can learn from and collect substantial data on informative individual cases. Incorporating insights from a range of scientific and clinical disciplines into the study and treatment of any one person is therefore beginning to emerge as a tractable, and more holistic, approach, and we document here what we believe to be the first integration of deep brain stimulation and whole genome sequencing for precision medicine in the evaluation, treatment and preventive care for one severely mentally ill individual, M.A. We have shown that DBS has been successful in aiding in the care and beneficial clinical outcome of his treatment refractory OCD, and we have also demonstrated that it is indeed feasible, given current technologies, to incorporate health information from WGS into the clinical care of one person with severe mental illness, including with the return of these health information to him directly. On a comparative level, deep brain stimulation has thus far been a more direct and effective intervention for his mental illness than anything discovered from his whole genome sequencing. Despite this, health information stemming from these WGS data was nevertheless immediately useful in the care of this person, as a variant associated with his ophthalmologic phenotype did indeed inform and enrich his care, and we expect that these data will continue to inform his care as our understandings of human biology and the genetic architecture of disease improves. Of course, the genomic data would have been more helpful if obtained much earlier in his medical course, as it could have provided guidance on which medications to avoid or to provide in increased doses.

3

Variants in *TAF1* are associated with a new syndrome.

3.1 Motivation

Transcription factor II D (TFIID) consists of the TATA binding protein (TBP) and 12-14 TBP associated factors (TAFs). TFIID promotes transcriptional initiation by recognizing promoter DNA and facilitating the nucleation of other general transcription factors for assembly into a functional preinitiation complex[40,136,149,232,20,19], and it also functions as a coactivator by interacting with transcriptional activators[232]. Subunits of TFIID have been suggested to play a possible role in neurodegenerative diseases and developmental delay when disrupted[262,185,208,271,131,112,114]. Indeed, variants in TAF2 [MIM 604912] and TBP [MIM 600075] have been implicated in intellectual disability (ID) and developmental delay, with or without corpus callosum hypoplasia[262,1,112]. Recent work toward understanding the molecular basis of Cornelia de Lange syndrome (CdLS [MIM 122470, 300590, 300882, 614701, and 610759]) has also implicated mutations in TAF6 [MIM 602955], a component of TFIID, as playing an important role in the pathogenesis of this syndrome[317]. CdLS is a phenotypically and genetically heterogeneous syndrome characterized by distinct facial features, hirsutism, developmental delay, intellectual disability and limb abnormalities[54], with mutations in several different genes implicated in contributing to this heterogeneous clinical presentation[187,142,98,129]. Two mutations in TAF6 were implicated in the pathogenesis of some cases with phenotypic features of CdLS and were shown, through biochemical investigations, to reduce binding of TAF6 to other core components of TFIID[317].

A recent paper nominated TAF1 [MIM 313650] as a candidate gene for intellectual disability (ID), based on segregation of missense variants in two different pedigrees; however, no clinical information other than intellectual disability was provided[121]. TAF1 is the largest subunit of the TFIID complex, and has been ranked 53rd among the top 1,003 constrained human genes in a recent population-scale study[272], suggesting a critical role for this protein in normal cellular function. Previous work in Drosophila cells has shown that TAF1 depletion increases the magnitude of the initial transcription burst and causes delay in the shutoff of transcription upon removal of the stimulus[233]. The authors showed that the magnitude of the transcriptional response to the same signaling event, even at the same promoter, can vary greatly depending on the composition of the TFIID complex in the cell. In addition and consistent with the notion that TAF1 is important in controlling the binding patterns of TFIID to specific promoter regions, this study showed that the set of genes conferring increased expression were enriched for TATA-containing promoters, suggesting an association between the depletion of TAF1 and increased expression of genes with the TATA-motif. The genomic region containing TAF1 has also been suggested to play an important role in X-linked dystonia-parkinsonism (XDP [MIM 314250]), although the exact mechanism remains undetermined[185,114,62,67]. XDP is an X-linked recessive movement disorder characterized by adult onset dystonia and parkinsonism, which leads to eventual death due to oropharyngeal dystonia or secondary infections[222]. Studies investigating the molecular basis of XDP demonstrated aberrant neuron-specific TAF1 isoform expression levels in neuronal tissue containing TAF1 variants. Herzfeld et. al (2013) corroborated previous reports which suggested that a reduction in TAF1 expression is associated with large-scale expression differences across hundreds of genes[114], and studies in rat and mice brain also corroborate the importance and relevance of TAF1 expression patterns specific to neuronal tissues[271,131].

In this study, we describe a recognizable syndrome attributed to mutations in TAF1. This work represents a collaborative research effort between independent groups engaged in studying the molecular basis of human disease. A "genotype-first" approach[287] was taken to find families with variants in TAF1. This approach included phenotypic evaluations and the screening of families with individuals harboring mutations in TAF1. This process was facilitated by utilizing databases such as DECIPHER and reporting initial findings on the BioRxiv preprint server[226]. These efforts culminated in a study cohort of 14 affected individuals from 11 unrelated families, 12 of which (from 9 unrelated families) contain single nucleotide changes in TAF1 and 2 of which contain large duplications involving TAF1 (from 2 unrelated families).

3.2 Methods

Several strategies were used to identify candidate disease-related sequence variation. These included whole genome sequencing, exome sequencing, targeted gene panel sequencing and microarray-based strategies. Sanger sequencing was used to validate sequence variations. Many of the families studied here underwent genotyping for a small number of genomic regions using clinical microarrays or gene-specific sequencing.

To evaluate structural CNS defects, our colleagues performed in vivo functional modeling of TAF1 using transient knockdown in developing zebrafish embryos. The optic tectum is a neuroanatomical structure that occupies the majority of the space within the midbrain and its size is a proxy for head size (microcephaly)[24], which is one of the most robustly observed phenotype.

A splice-blocking morpholino (MO) targeting the donor site of exon 9 of the sole zebrafish taf1 ortholog (ENSDART00000051196) was designed, stainings were performed for acetylated tubulin to evaluate the structure of the optic tectum. In total, 134 control embryos, 133 taf1 MO-injected, 109 taf1 MO+WT TAF1 RNA-injected and 78 WT TAF1 mRNA-injected embryos were analyzed by measuring the area of the optic tectum at three days post fertilization.

3.2.1 Cluster analysis

Nonsynonymous TAF1 hemi or homozygote variants from European and Latino populations were taken from the ExAC database. Unique genomic positions were collected. Following Cucala 2008[51], these locations were set such that

$$0 = X_0 <= X_1 <= \cdots X_n <= X_n + 1 = 1$$

we then compute all $j - i$ ordered spacing's such that

$$D_i,j = X_j - X_i \sum_{k=i+1}^{j} D_k, \quad 1 <= i < j <= n$$

we also let

$$U_i = B_{inc}(D_{i,j}, j - i, n + 1, j + i), \quad 1 <= i < j <= n$$

and compute the hypothesis-free scan statistic, which is

$$\wedge_{HF} = \sup_{1<=i<j<=n} \frac{1}{U_{i,j}}$$

p-values are computed via a Monte Carlo procedure. To identify more than one cluster, a multiple procedure is introduced. If there exists an interval that represents a significant cluster in the initial search, which Cacala 2008 and we note here as $[X_{i^*}, X_{j^*}]$, then let $T^* = 1 - X_{j^*} + X_{i^*}$. We will then transform the data such that

$$X_k^2 = \begin{cases} \frac{X_k}{T^*} & \text{if} \quad 1 <= k <= i^* \\ \frac{X_{k+j^*-j^*} - X_{j^*} + X_{i^*}}{T^*} & \text{if} \quad i^* + 1 <= k <= n - j^* + i^* \end{cases}$$

and test for clusters as described above

3.2.2 Family 1 specific methods

Below, we described sequencing and analysis methods that are specific to Family 1, one of the larger of the study families.

Complete Genomics whole genome sequencing and variant detection for Family 1

After quality control to ensure lack of genomic degradation, we sent 10 g DNA of each sample to Complete Genomics (CG) at Mountain View, California for sequencing. The whole-genome DNA was sequenced with a nanoarray-based short-read sequencing-by-ligation technology, including an adaptation of the pairwise end-sequencing strategy. Reads were mapped to the Genome Reference Consortium assembly GRCh37. Due to the proprietary data formats, all the sequencing data QC, alignment and variant calling were performed by CG as part of their sequencing service, using their version 2.0 pipeline. Complete Genomics WGS was optimized to cover 90% of the exome with 20 or more reads and 85% of the genome with 20 or more reads.

Illumina HiSeq 2000 whole genome sequencing and variant detection for Family 1

After the samples were quantified using Qubit® dsDNA BR Assay Kit (Invitrogen), 1 μg of each sample was sent out for whole genome sequencing using the Illumina® Hiseq 2000 platform. Sequencing libraries were generated from 100 ng of genomic DNA using the Illumina TruSeq Nano LT kit, according to manufacturer recommendations. The quality of each library was evaluated with the Agilent bioanalyzer high sensitivity assay (less than 5% primer dimers), and quantified by qPCR (Kappa Biosystem, CT). The pooled library was sequenced in three lanes of a HiSeq2000 paired end 100 bp flow cell. The number of clusters passing initial filtering was above 80%, and the number of bases at or above Q30 was above 85%. Illumina reads were mapped to the hg19 reference genome using BWA v0.6.2-r126, and variant detection was performed using the GATK v. 2.8-1-g932cd3a. Illumina WGS resulted in an average mapped read depth coverage of 37.8X (SD=1.3X). >90% of the genome was covered by 30 reads or more and >80% of the bases had a quality score of >30. A second analytical pipeline was used to map the Illumina reads and detect variants using novoalign v3.00.04 and the FreeBayes caller v9.9.2-43-ga97dbf8. Additional variant discovery procedures included Scalpel v0.1.1 for insertion or deletion (INDEL) detection, RepeatSeq v0.8.2 for variant detection in short tandem repeat regions, and the ERDS (estimation by read depth) method v1.06.04 and PennCNV (2011Jun16 version) for detecting larger copy number variants (CNVs).

Sanger Sequencing

PCR primers were designed using Primer 3 (http://primer3.sourceforge.net) to produce amplicons of around 700 bp in size, with variants of interest located approximately in the center of each amplicon. Primers were obtained from Sigma-Aldrich®. Upon arrival, all primers were tested for PCR efficiency using a HAPMAP DNA sample (Catalog ID NA12864, Coriell Institute for Medical Research, USA) and LongAmp® Taq DNA Polymerase (New England Biolabs, USA). PCR products were visually inspected for amplification efficiency using agarose gel electrophoresis. PCR products were further purified using QIAquick PCR Purification Kit (QIAGEN Inc., USA), quantified by Qubit® dsDNA BR Assay Kit (Invitrogen Corp., USA), and diluted to 5 - 10 ng/µl in water for Sanger sequencing using the ABI 3700 sequencer. The resulting *.ab1 files were loaded into the CodonCode Aligner V4.0.4 for analysis. All sequence traces were manually reviewed to ensure the reliability of the genotype calls (Figure 3.1).

DNA Microarrays

DNA samples were genotyped on Illumina Omni 2.5 DNA microarrays (which contain approximately 2.5 million markers). Total genomic DNA extracted from whole blood was used in the experiments. Standard data-normalization procedures and canonical genotype-clustering

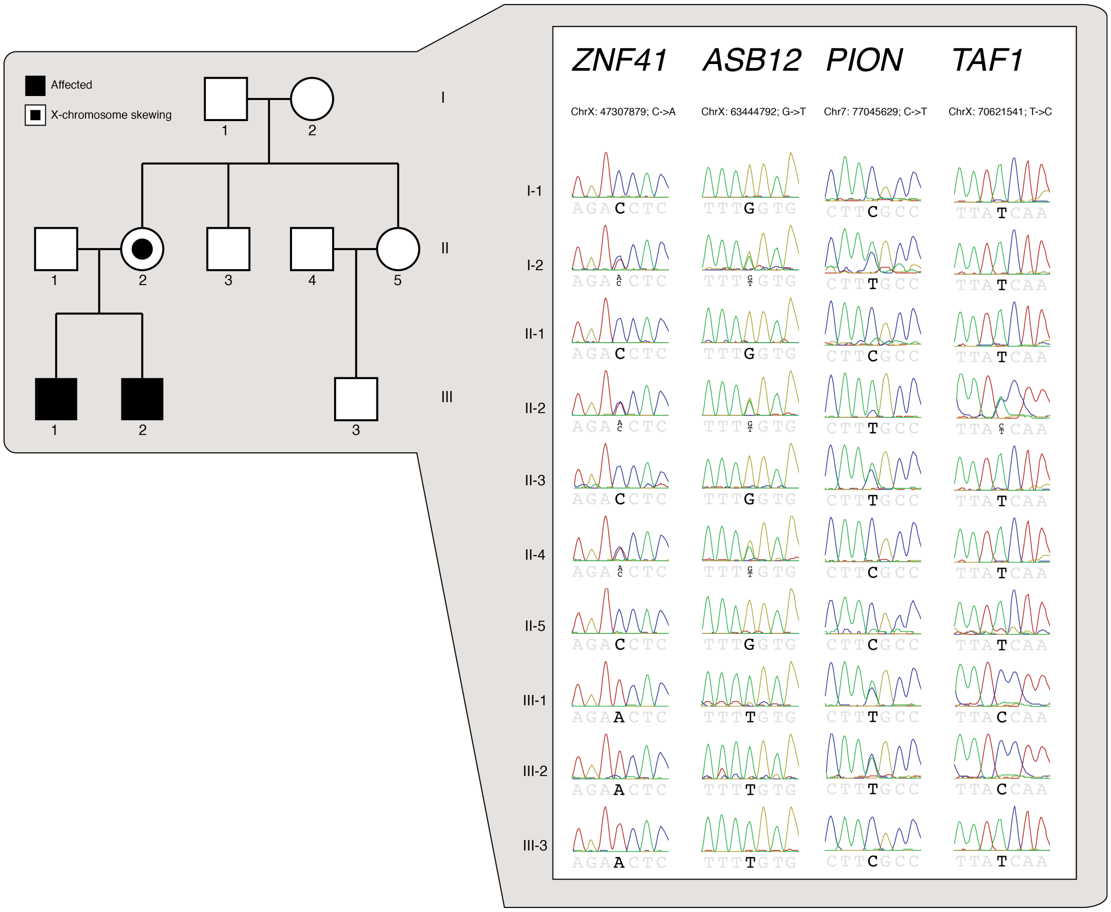

Affected
X-chromosome skewing
I
II
III
1
2
1
2
3
4
5
1
2
3
ZNF41
ASB12
PION
TAF1
ChrX: 47307879; C->A
ChrX: 63444792; G->T
Chr7: 77045629; C->T
ChrX: 70621541; T->C
I-1
AGACCTC TTTGGTG CTTCGCC TTATCAA
I-2
AGA A C CTC TTT G T GTG CTTTGCC TTATCAA
II-1
AGACCTC TTTGGTG CTTCGCC TTATCAA
II-2
AGA A C CTC TTT G T GTG CTTTGCC TTA C T CAA
II-3
AGACCTC TTTGGTG CTTTGCC TTATCAA
II-4
AGA A C CTC TTT G T GTG CTTCGCC TTATCAA
II-5
AGACCTC TTTGGTG CTTCGCC TTATCAA
III-1
AGAACTC TTTTGTG CTTTGCC TTACCAA
III-2
AGAACTC TTTTGTG CTTTGCC TTACCAA
III-3
AGAACTC TTTTGTG CTTCGCC TTATCAA

Figure 3.1: Sanger sequencing traces for all 10 family members from family 1 for variants found in ZNF41, ASB12, PION and TAF1.

files provided by Illumina were used to process the genotyping signals.

Variant detection

Human sequence variation ranges in manifestation from differences that can be detected at the single nucleotide level, to whole chromosome differences. In our study, we used a number of bioinformatics software packages to extract signals for differences seen at the levels of single nucleotide variants (SNV), small insertions/deletions (INDELs), variants in short tandem repeat structure (STRs), and variants in copy number (CNVs) (see Figure 3.3 and 3.3 for a general map of the analyses performed). When possible, we used more than a single bioinformatics software package to detect different classes of genetic variants, so as to arrive at a comprehensive and high-quality set of variants for each person sequenced. Standard data quality filtering approaches were used for all genetic variants detected by the various different methods. This includes, when appropriate, requiring sequencing to be at a depth of 10 or more reads at the location of a sequence variant, and a variant phred quality score of 30 or above. Specific variant detection parameters, which themselves detail internal or pipeline specific variant detection thresholds, are described below or in the documentation of the software which has been described in detail elsewhere. We expect variants where all pipelines agree to be more reliable in terms of their validation rate, whereas those variants that were unique to single pipelines will likely have lower yet potentially vastly different validation rates. This expectation has been shown to be true in our previous studies that have used high-throughput MiSeq validation methods[225]. This information was carried through to the various stages of the variant prioritization and functional annotation stages of the study. If a variant was annotated as being highly deleterious by functional annotation or by frequency inference, sequence error was more easily identified by first checking how many detection pipelines found it. In contrast, if a variant was detected by one sequencing platform and not the other, sequence depth and quality variation between platforms contributed to these instances and were not as easily dismissed as errors.

bwa-GATK

Illumina reads were mapped to the hg19 reference genome using BWA v. 0.7.5a using default 'mem' parameters. BWA was directed to mark shorter split hits as secondary, so as to make the output compatible with Picard and the Genome Analysis Took Kit (GATK). BWA sequence alignments were converted into binary format using SAMtools v0.1.19-44428cd, and duplicate reads were marked using Picard tools v1.84. GATK 2.8-1-g932cd3a was used to realign the reads around putative INDELs, and base quality scores were then recalibrated. Variants were detected using the GATK HaplotypeCaller, and variant quality scores were then recalibrated using the GATK variant quality score recalibration (VQSR) protocol. The GATK HaplotypeCaller works by generating a reference graph assembly, which starts out as a directed DeBruijn graph. The GATK HaplotypeCaller then tries to match each sequence read to a path in the reference graph, this is called the 'read-threading' graph. The graph is then pruned by removing sections of the graph that are supported by fewer than 2 reads, which are considered to be the result of stochastic errors. Haplotype sequences are then constructed using likelihood scores for each path in the graph. A Smith-Waterman alignment of each haplotype to the original reference sequence is used to generate potential variant calls, which are then modeled using a genotype likelihood framework.

novoalign-FreeBayes

SNP and INDELs were also detected with a novoalign-FreeBayes pipeline, using novoalign v3.00.04 to map reads to the hg19 reference genome, and the FreeBayes caller v9.9.2-43-

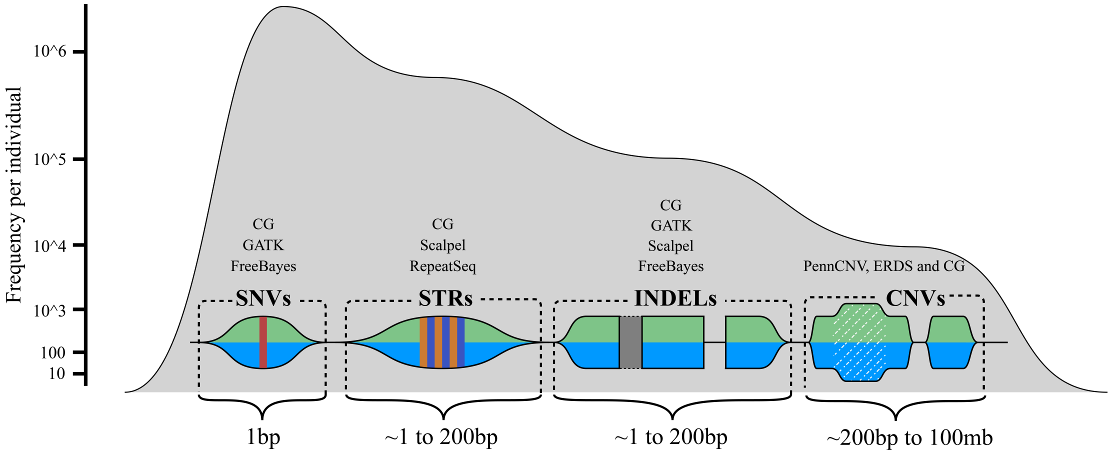


Figure 3.2: A conceptual map of human sequence variation. Here, we show approximate sizes, as well as the associated signature, of the various different types of human sequence variation that can be currently detected with the WGS and informatics technologies employed in this work. The frequency axis shows the approximate frequency of the various genetic variation types that currently detectable via germline WGS. Above the visual signatures of the different types of human sequence variation, the general names of the different informatics software tools for detecting the variation are noted which include, the Genome Analysis Took Kit (GATK), Scalpel, RepeatSeq PennCNV, the estimation by read depth with single-nucleotide variants (ERDS) CNV caller and the FreeBayes caller. We do not differentiate here by raw sequencing data generated by different sequencing technologies, but its important to note that Complete Genomics (CG) is listed here as a software tool but in actuality what we are referring to is the CG sequencing technology as well as its own proprietary sequence analysis.

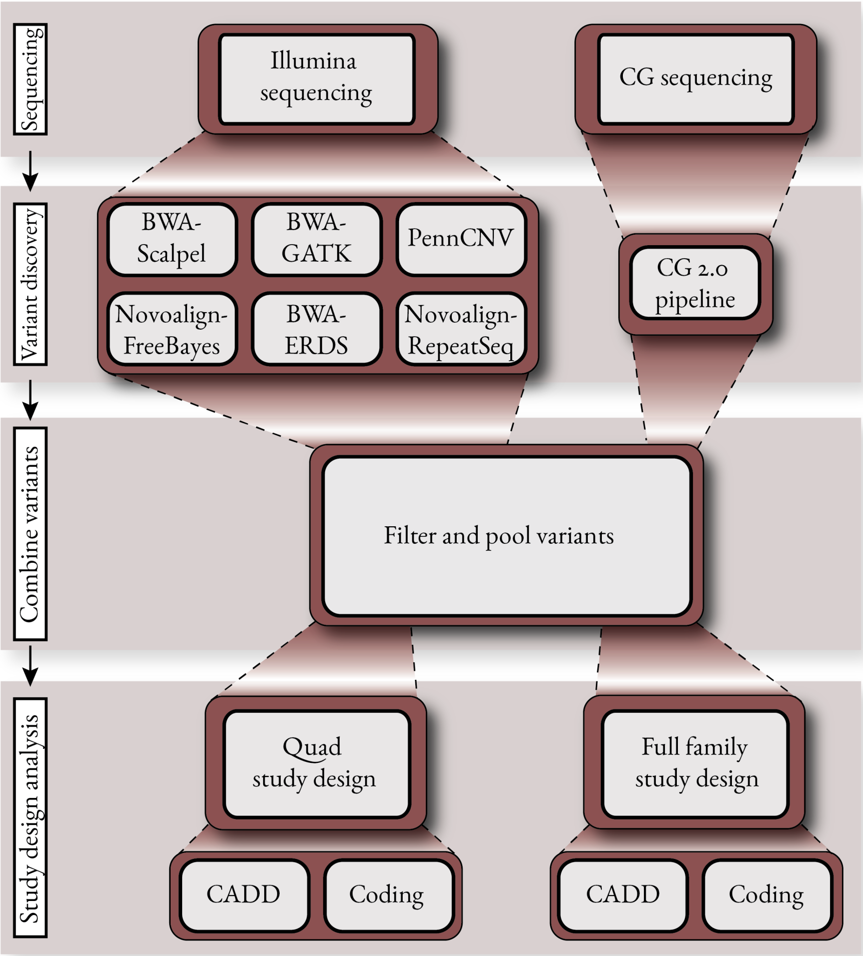


Figure 3.3: A generalized map of the flow of work performed during the course of the study for Family 1. Briefly, the family was sequenced using two different sequencing technologies, the Illumina (which includes WGS on the HiSeq 2000 and genotyping array data from Illumina Omni 2.5 microarrays) and the Complete Genomics (CG) sequencing platforms. Raw sequencing data resulting from the CG sequencing was processed using the internal CG informatics pipeline v. 2.0. Six variant discovery pipelines analyzed raw sequencing data resulting from the Illumina-based sequencing. Variants resulting from all of the post-sequencing data analysis and variant discovery pipelines were filtered using standard filtering methods/thresholds (see Methods section) and pooled for further post-variant discovery analyses. From the pooled variant data set, two family-based study designs were performed, a quad based study design and a study design that incorporates data from all of the sequenced members of the family. For these two studies, two variant prioritization strategies were employed, 'CADD' and 'Coding'. Both prioritization schemes required variants to be in low population frequencies (minor allele frequency of less than 1%), but the CADD strategy further required variants to have a CADD score of greater than 20 whereas the coding scheme required variants to be coding and non-synonymous.

ga97dbf8 to detect variants. Novoalign was used to map the first 50,000 reads to the reference sequence in order to determine the empirical insert size for the Illumina paired end reads. Once the insert size was determined, novoalign was then used to map all of the Illumina reads to the reference sequence using default parameters. Sequence alignment output from novoalign was used to generate variant calls using FreeBayes with default parameters. FreeBayes uses a Bayesian genotype likelihood approach, but generalizes its use to perform over haplotype sequences, which is in contrast to precise alignment based implementations.

bwa-Scalpel

Sequence alignments obtained from the above ‘bwa-GATK’ pipeline were used in conjunction with Scalpel v0.1.1[209] to extract INDELS from the WGS data. Scalpel was run with near-default parameterizations in ‘single’ mode. The minimum coverage threshold and minimum coverage ratio for emitting a variant was set to 3 and 0.1 respectively, and the threshold at which low coverage nodes were removed was set to 1. Scalpel uses both sequence mapping and assembly to detect INDELs. First, Scalpel extracts aligned sequence reads to construct a de Bruijn graph. Low coverage nodes and sequencing errors are then removed and a repeat analysis of the each region is performed to tune the k-mer size. Assembled sequences are then aligned back to the reference genome where a standard Smith-Waterman-Gotoh alignment algorithm with affine gap penalties is used to detect candidate variants.

RepeatSeq and Scalpel

We used RepeatSeq[115] v0.8.2 to extract variants in short tandem repeats across the genome using default settings. RepeatSeq uses a Bayesian model selection approach to assign the most probable genotype using information about the full length of the sequence repeat, the repetitive unit size and the average base quality of mapped reads. Scalpel has been shown to perform well in terms of detecting variants in short tandem repeat regions. For this reason, we also consider Scalpel to be a good informatics pipeline for use in profiling STR regions. Scalpel was used to detect sequence variants in STRs using the same methods described in the section detailing pipelines on SNVs and INDELs (above).

CNVs

Bedtools v2.17.0[244] was used to compare CNVs. CNVs were required to overlap reciprocally by 90%. Hypervariable and invariant CG CNVs were excluded from the analysis, and CG CNVs were required to have a 'CNVTypeScore' of greater than 30.

PennCNV

The PennCNV[305] software package (2011Jun16 version) was used to perform Copy Number Variant CNV calling using the Illumina Omni 2.5 microarray data for all the samples. For kilobase-resolution detection of CNVs, PennCNV uses an algorithm that implements a hidden Markov model, which integrates multiple signal patterns across the genome and uses the distance between neighboring SNPs and the allele frequency of SNPs. The two signal patterns that it uses are the Log R Ratio (LRR), which is a normalized measure of the total signal intensity for two alleles of the SNP and the B Allele Frequency (BAF), a normalized measure of the allelic intensity ratio of two alleles. The combination of both signal patterns is then used to infer copy number changes in the genome.

Microarrays often show variation in hybridization intensity (genomic waves), that is related to the genomic position of the clones, and that correlates to GC content among the genomic features considered. For adjustment of such genomic waves in signal intensities, the cal_gc_snp.pl PennCNV program was used to generate a GC model that considered the GC content surrounding each Illumina Omni2.5 marker within 500kb on each side (1Mb total). The detect_cnv.pl program in the mode – test for individual CNV calling was used, the Hidden Markov Model used is contained in the hhall.hmm file provided by the latest PennCNV package, and custom Population Frequency of B allele (PFB) file for all the SNPs in the Illumina Omni2.5 array was generated from 600 controls (which consists of 600 unaffected parents from the Simons Simplex Collection), the GC model described above was also used during CNV calling. Chromosome X CNVs were called separately using the – test mode with the –chrx option. We excluded CNVs with an inter-marker distance of >50kb and required each CNV to be supported by at least 10 markers.

ERDS

To detect CNVs from the Illumina WGS data, the Estimation by Read Depth with SNVs (ERDS) v1.06.04[321] method was employed, using default pipeline parameterization. ERDS uses WGS read depth information contained within sequence alignment files, along with soft-clip signatures, to detect CNVs. CNVs that were detected by the ERDS method were filtered to include CNVs that were greater than 200 kilobases in scale and CNVs called with a confident score of greater than 300.

Disease variant prioritization, post variant discovery analyses

We performed analyses to prioritize sequence variants conforming to three disease model pathways: de-novo, autosomal recessive and x-linked models of transmission. X-chromosome skewing in the mother of the two affected boys suggests that genetic components of the disease phenotype are most likely segregating and following an X-linked mode of inheritance.

Recent work illustrates the existence of a substantial amount of complexity in elucidating genetic factors of human disease, with many syndromes likely being the result of an array of different genetic aberrations in conjunction with environmental effects and modification of gene/variant function by ancestral background[95,180,275]. There are also many still uncharacterized noncoding regions of the genome[35], along with continuous re-annotation of protein coding portions[72]. In light of this complexity, we sought to identify variants following de novo, autosomal recessive and x-linked models of transmission that may be contributing, together or alone, to the disease phenotype. It is possible that a disease-contributory variant in the germline of a somatically mosaic parent could pass on to both children, appearing as ’de novo’ when compared to DNA from the blood of the parents[32,192,277,284]. Similarly, variants benign in the heterozygous state might prove deleterious if present in the homozygous state in the two children, so we sought to identify these autosomal recessive variants as well.

In general, to identify de-novo variants, we isolated genetic variants shared by both affected boys. Variants in common with all other healthy people in the family were then filtered out. For X-linked variants, all X-chromosome variants shared by the two affected boys as well as their mother were identified. Then, variants in common with all other healthy males were removed. Autosomal recessive variants were identified by first selecting all heterozygous variants shared by the mother and father of the two affected children. Then, homozygous variants shared by the two affected boys where both parents were heterozygous were selected. Finally, homozygous variants that were also present in any other healthy family member were removed.

Variant prioritization for Family 1

We used several methods to prioritize and identify possible disease-contributory germ-line variants, including VAAST[122,137,263,315], Golden Helix SVS v8.1.4[289], ANNOVAR (2013Aug23 version)[306], and GEMINI v0.9.1[230]. VAAST employs a likelihood-based statistical framework for identifying the most likely disease-contributory variants given genomic makeup and population specific genomic information. SVS, ANNOVAR and GEMINI employ more traditional annotation and filtering-based techniques that leverage data stored in public genomic databases (i.e., dbSNP 137, 1000 Genomes phase 1 data, NHLBI 6500 exomes, etc.).

More detailed methodology for the analysis of Family 1 is available below. The Illumina data were also re-analyzed in the recent development of SeqHBase, a big data-based tool set for analyzing family based sequencing data to detect de novo, inherited homozygous, or compound heterozygous mutations that may contribute to disease manifestations[110].

SVS, ANNOVAR and GEMINI

Recent work has identified differences between results generated by available annotation software packages, which can, in part, be the result of differences in choice of transcript by the user[188]. To capture and analyze this variability, we used three annotation and filtering software packages with similar databases to filter and prioritize disease variants. For variants conforming to each disease model, ANNOVAR, SVS and GEMINI were used to filter and annotate them.

To be consistent among the different software tools, we used the same filtering strategy for each of the three software packages. Depending on the analysis, filtering criteria required each variant to be characterized by a population frequency of less than 1 percent in the available variant frequency databases (this includes genotype frequencies derived from the 1000 genomes project as well as genotype frequencies reported in dbSNP 138 and the Exome Sequence Project, which includes genotype frequencies, among other information, on 6500 individuals of various recently derived ancestral lineages), a CADD (Combined Annotation Dependent Depletion) score of greater than 20 (top 1% of all possible human genomic variants in terms of deleteriousness) or be either a non-synonymous or a splice site variant. Variants that passed these filters were then annotated by gene and variant type using UCSC's Known Genes table for annotation.

VAAST

VAAST v2.0 was used to identify variants that are likely contributing to disease[219,315]. SNPs and INDELs were converted into the GVF file format using the vast_converter tool, annotated using the the VAAST annotation tool (VST) and then converted into a condenser file (*.cdr). VAAST was run in CLRT mode without grouping variants when they are located within the same feature. Amino acid substation frequencies were included in the likelihood ratio test when scoring variants, and the maximum expected frequency of the 'causal' allele in the background population was set to 0.01. 10,000 permutations were performed. The VAAST background file that was used contains 1057 "1000 genome project" genomes, 54 Complete Genomics genomes, 184 genomes from Danish exomes, and 9 genomes from 10 Gen data[247]. Variants from dbSNP and NHLBI ESP that have a sample size >= 100 were randomly spiked into the dataset based on their allele frequencies. Coding variants only within CDS regions of the RefSeq gene set with 10 nts around each exon splice regions". The background file used in this study is public and freely available for download on the VAAST website (http://www.yandell-lab.org/software/VAAST).

RNA sequencing and analysis

We conducted RNA sequencing with RNA isolated from blood from Family 1. Blood was collected from the two probands, their parents and the maternal grandparents. Except for the single blood draws for the grandparents, two blood draws were taken on separate days for each subject. The blood was collected in PAXgene Blood RNA tubes and the RNA was isolated with the PAXgene Blood RNA kit (QIAGEN) according to the manufacturer's recommendations. The final pooled library was measured by qPCR using the KAPA SYBR® Fast Universal qPCR kit (Kapa Biosystem, Wilmington, MA) and sequenced on a HiSeq 2000 across three lanes (paired-end 100bp). A mean of 49,792,652 (sd = 11,666,119) properly paired reads were generated and a mean spliced mapping percentage of 85.41 (sd = 7.1) per sample was observed. HISAT[138] was used for spliced alignment to the UCSC human reference sequence hg19 using 10 alignment threads and the –rna-strandness RF flag for stranded libraries. Stringtie[236] was used to quantify transcripts using the UCSC hg19 transcript annotations, with estimates of abundances being restricted to these annotated transcripts. Cuffdiff[295] was used to perform differential expression analysis using 32 threads with the –library-type fr-firststrand flag for our stranded libraries. The R package CummeRbund[152] was used to analyze, filter and visualize the results from the Cuffdiff differential expression analysis. WebGestalt[304] was used to perform gene set enrichment analyses using Molecular Signatures Database (MSigDB) for Transcription Factor Targets[288], KEGG[135], GO[48], and HPO[145] databases for gene annotations and set inclusion information. We used various analysis tools in the CummeRbund package[152] to evaluate the quality of our RNA sequencing data/analysis for Family 1.

3.3 Results

Shared phenotypic features representing the cardinal features of this syndrome (see Figures 3.4 and 3.6, and Tables 3.1 and 3.2) include global developmental delay, intellectual disability, characteristic facial dysmorphologies, and generalized hypotonia. Shared facial dysmorphologies include prominent supraorbital ridges, down slanted palpebral fissures, sagging cheeks, long philtrum, low-set and protruding ears, long face, high palate, pointed chin and anteverted nares. The probands also share a characteristic gluteal crease, with a sacral caudal remnant, although spine MRI imaging on two probands did not show any major underlying defect. The affected individuals have generalized hypotonia, as well as joint hypermobility. Other widely shared features include hearing impairment, microcephaly and hypoplasia of the corpus callosum. Interestingly, probands 8A (in Figure 3.5 as II-1 in family 8) and 11A (in Figure 3.6 as II-1 in family 11) are also affected by abnormal thoracic cage development. Some additional neurological features include spastic diplegia, dystonic movements, and tremors. Individuals 8A, 10A (in Figure 3.6 as IV-3 in family 10), and 11A had progressive symptoms, and one individual (11A) died of severe cardiopulmonary insufficiency attributed to an infection. Importantly, probands bearing duplications of TAF1 (10A and 11A) not only demonstrated severe and progressive neurodegeneration, they also do not share some of the more common features of the rest of the probands (see Table 3.1,and Figures 3.4 and 3.6 for comparison).

All 14 affected individuals were found to contain sequence variants in TAF1, the majority of which are missense variants (11 out of 14). One proband (5A in Figure 3.5 as II-1 in family 5) was found to have a variant which influences *TAF1* splicing, and two probands (10A and 11A) have a duplication that involves *TAF1*. As stated above, the two CNV duplication probands exhibit less overlap with those that harbor single nucleotide changes (Table 3.1) and exhibit severe progressive neurologic impairment. All of the mutations reported here, including the duplications, are de novo or co-segregate with the phenotype in other affected male individuals (see Figures 3.5 and 3.6).

Families 1 and 10 could be tested for X-chromosome inactivation, which showed that female carriers of TAF1 mutations and duplications demonstrated highly skewed inactivation (99:1). The female carriers in families 5 and 11 were not informative for the polymorphic

Features	Proband													
(Human Phenotype Ontology Nos.)	1A	1B	2A	3A	4A	5A	6A	7A	8A	8B	8C	9A	10A[a]	11A[a]
Sex	M	M	M	M	M	M	M	M	M	M	M	M	M	M
Age (years)	15	13	5	6	9	3	22	11	9	4	1	3	16	8
Postnatal growth retardation (HP: 0008897)	+	+	+	+	+	+	−	−	+	+	+	+	UK	+
Delayed gross motor development (HP: 0002194)	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Delayed speech and language development (HP: 0000750)	+	+	+	+	+	+	+	+	+	+	+	UK	+	+
Oral-pharyngeal dysphagia (HP: 0200136)	+	+	+	+	UK	+	UK	−	+	+	+	UK	+	UK
Prominent supraorbital ridges (HP: 0000336)	+	+	−	+	UK	−	+	−	+	+	+	+	+	+
Downslanted palpebral fissures (HP: 0000494)	+	+	+	−	+	+	+	−	+	+	+	−	+	UK
Sagging cheeks	+	+	−	−	−	+	−	−	+	+	+	+	+	+
Long philtrum (HP: 0000343)	+	+	+	+	+	+	−	+	+	+	+	+	−	−
Low-set ears (HP: 0000369)	+	+	+	+	+	+	+	−	+	+	+	+	−	+
Protruding ears (HP: 0000411)	+	+	+	+	+	+	+	−	+	+	+	−	−	+
Long face (HP: 0000276)	+	+	−	−	UK	+	+	−	+	+	+	+	+	+
High palate (HP: 0000218)	UK	UK	+	+	−	+	+	−	+	+	+	+	+	+
Pointed chin (HP: 0000307)	+	+	−	−	+	+	+	−	+	+	+	−	+	+
Anteverted nares (HP: 0000463)	−	−	+	+	+	+	−	+	+	+	+	+	−	+
Hearing impairment (HP: 0000365)	+	+	+	+	UK	+	−	−	+	+	+	−	UK	−
Chromic otitis media (HP: 0000389)	+	+	+	−	+	+	−	+	+	+	+	−	UK	−
Strabismus (HP: 0000486)	+	+	+	+	UK	+	+	−	+	−	−	+	+	−
Microcephaly (HP: 0000252)	+	+	+	+	+	−	−	+	+	+	+	+	−	−
Hypoplasia of the corpus callosum (HP: 0002079)	+	+	+	UK	+	+	+	UK	+	+	+	+	UK	−
Generalized hypotonia (HP: 0001290)	+	+	+	+	+	+	+	−	+	+	+	+	−	+
Unusual gluteal crease with sacral caudal remnant and sacral dimple (abnormal sacral segmentation [HP:0008468] and prominent protruding coccyx [HP: 0008472])	+	+	+	+	+	+	+	+	+	+	+	+	UK	−
Joint hypermobility (HP: 0001382)	+	+	−	+	UK	+	−	−	+	+	+	+	−	UK
Autistic behaviors (HP: 0000729)	+	+	+	−	UK	UK	+	+	+	+	+	−	+	+
Intellectual disability (HP: 0001249)	+	+	+	+	+	UK	+	+	+	+	+	+	+	+

Table 3.1: This table demonstrates clinical features shared by eight or more probands across all affected individuals in the families. Abbreviations are as follows: M, male; and UK, unknown. [a] Probands containing duplications; they are generally less similar to the probands containing SNVs and share fewer common clinical features.

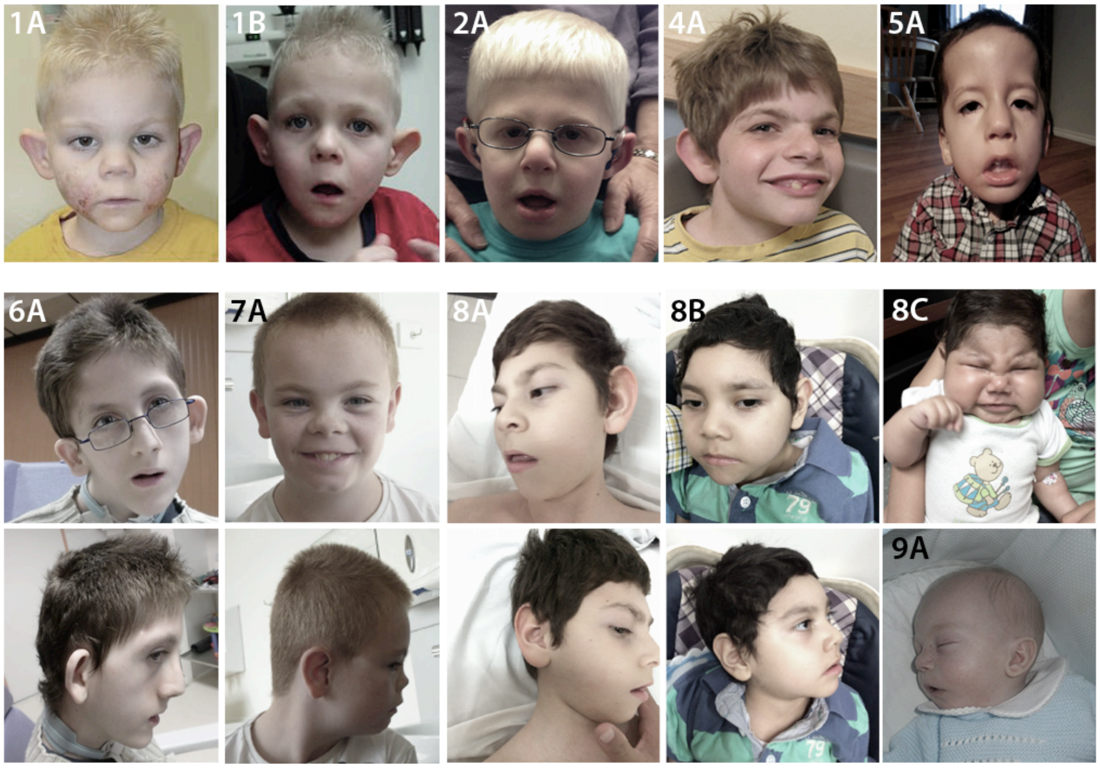
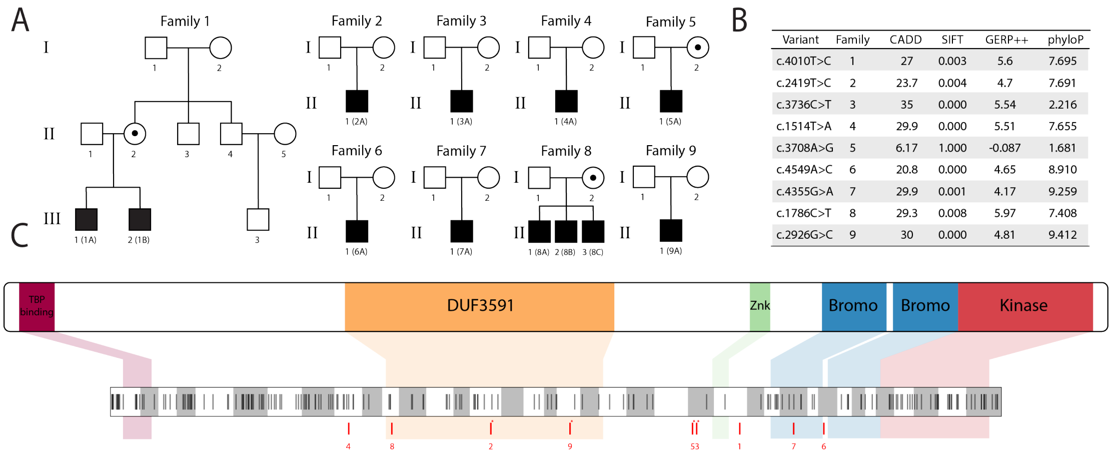


Figure 3.4: Images of the Facial Phenotype from Families 1, 2, and 4–9. Cardinal facial dysmorphologies include prominent supraorbital ridges (seen in 1A, 1B, 2A, 6A, 8A–8C, and 9A), down-slanted palpebral fissures (1A, 1B, 2A, 4A, 5A, 6A, 8A–8C, and 9A), sagging cheeks (1A, 1B, 5A, 8A–8C, and 9A), a long philtrum (1A, 1B, 2A, 4A, 5A, 8A–8C, and 9A), low-set and protruding ears (1A, 1B, 2A, 4A, 5A, 6A, 8A–8C, and 9A), a long face (1A, 1B, 2A, 5A, 6A, 8A–8C, and 9A), a high palate (5A, 6A, 8A–8C, and 9A), a pointed chin (1A, 1B, 2A, 4A, 5A, 6A, and 8A–8C), and anteverted nares (2A, 4A, 5A, 7A, 8A–8C, and 9A).

Proband	Inheritance	Genetic Background	TAF1 Mutation (hg19)
1A	maternal	European decent	chrX: g.70621541T>C (c.4010T>C; p.Ile1337Thr)
1B	maternal	European decent	chrX: g.70621541T>C (c.4010T>rC; p.Ile1337Thr)
2A	de novo	European decent	chrX: g.70607243T>C (c.2419T>C; p.Cys807Arg)
3A	de novo	European decent	chrX: g.70618477C>T (c.3736C>T; p.Arg1246Trp)
4A	de novo	European decent	chrX: g.70601686T>A (c.1514T>A; p.Ile505Asn)
5A	maternal	Ecuadorian	chrX: g.70618449A>G (c.3708A>G; r.[3708a>g; 3681_3708del28]; p.Arg1228Ilefs□16)
6A	de novo	European decent	chrX: g.70643003A>C (c.4549A>C; p.Asn1517His)
7A	de novo	British	chrX: g.70627912G>A (c.4355G>A; p.Arg1431His)
8A	maternal	Colombian	chrX: g.70602671C>T (c.1786C>T; p.Pro596Ser)
8B	maternal	Colombian	chrX: g.70602671C>T (c.1786C>T; p.Pro596Ser)
8C	maternal	Colombian	chrX: g.70602671C>T (c.1786C>T; p.Pro596Ser)
9A	de novo	Spanish	chrX: g.70612503G>C (c.2926G>C; p.Asp976His)
10A	maternal	Albanian	0.423 Mb duplication including TAF1 and other genes at Xq13.1(70,370,794–70,794,385); deletion containing KANSL1 and other genes at 17q21.31 (0.63 Mb)
11A	de novo	Greek	0.42 Mb duplication including TAF1 and other genes: arr Xq13.1(70,287,519–70,711,110)×2

Table 3.2: Summary of TAF1 variants across all affected individuals in this study



Variant	Family	CADD	SIFT	GERP++	phyloP
c.4010T>C	1	27	0.003	5.6	7.695
c.2419T>C	2	23.7	0.004	4.7	7.691
c.3736C>T	3	35	0.000	5.54	2.216
c.1514T>A	4	29.9	0.000	5.51	7.655
c.3708A>G	5	6.17	1.000	-0.087	1.681
c.4549A>C	6	20.8	0.000	4.65	8.910
c.4355G>A	7	29.9	0.001	4.17	9.259
c.1786C>T	8	29.3	0.008	5.97	7.408
c.2926G>C	9	30	0.000	4.81	9.412

Figure 3.5: TAF1 Domains, Variant Scores, and ExAC Sequence Variation Plot. (A) Pedigree drawings of the nine families who were found to harbor TAF1 SNVs (NCBI Gene ID: 6872 according to the GRCh37.p13 assembly). Black dots indicate maternal carriers. (B) All nine SNVs are listed and annotated with CADD, SIFT, GERP++, and phyloP scores (which indicate conservation across 99 vertebrate genomes and humans). All of the SNVs are considered to be potentially deleterious by all of the listed annotations, except for c.3708A>G, which is a splice-site variant and as a consequence is not necessarily expected to be categorized as deleterious by any of the listed scores, because it does not affect amino acid composition of the predicted protein. (C) Known TAF1 domains are shown with respect to their corresponding genic positions. All but non-synonymous variants reported in the ExAC Browser for TAF1 are plotted below as lines; white and gray indicate exon boundaries. Red lines indicate the relative positions of the eight missense variants described in this paper (see Table 2). Numerals link the sequence variants shown on the ExAC plot to their familial origin, and those noted with a star fall within TAF1 regions that are significantly underrepresented by non-synonymous sequence variation in the ExAC Browser in European and Latin populations (p values of 0.032 and 0.037 for the first [c.2419T>C, c.2926G>C, and c.3736C>T] and second [c.3708A>G] clusters, according to Cucala's hypothesis-free multiple scan statistic with a variable window 27).

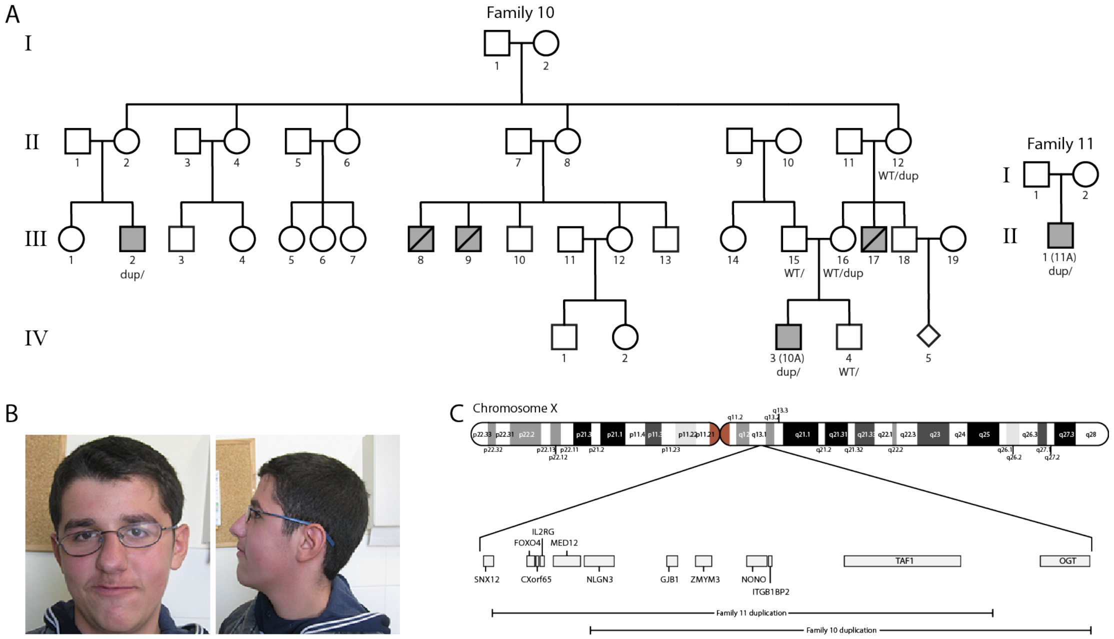


Figure 3.6: Duplications Involving TAF1 from Families 10 and 11. (A) Pedigree drawings of families 10 and 11. (B) The facial phenotype of proband 10A is notable for prominent supraorbital ridges, down-slanted palpebral fissures, sagging cheeks, a long face, a high palate, and a pointed chin. (C) Chromosome X cytobands are plotted above a more focused view of the region containing duplications that involve TAF1 in families 10 and 11. UCSC refGenes (from the UCSC Genome Browser tables) whose canonical transcript start or stop sites overlap either of the two duplications are plotted.

CAG repeat in the human androgen-receptor gene.

All missense variants were found to affect evolutionarily conserved residues (Figure 3.5B), and were not present in any frequency in public databases such as dbSNP 137, 1000 Genomes phase 1 data, NHLBI 6500 exomes, or ExAC version 0.2, which contains allelic information derived from 60k exome sequences. The TAF1 missense variants were also predicted to be deleterious by a range of prediction scores (CADD, SIFT, GERP++ and phyloP) (Figure 3.5B). The splice site variant discovered in Family 5 was not predicted to be deleterious by the prediction scores listed; however, this variant does not change the amino acid content of the predicted protein, but instead affects TAF1 splicing in both the mother and proband. In addition, the SNVs described here fall within regions of the TAF1 gene that are relatively sparsely covered by non-synonymous sequence variations reported in the ExAC database (Figure 3.5C). Indeed, four of these missense variants (c.2419T>C, c.2926G>C, c.3736C>T, and c.3708A>G) were found to be within the two regions of TAF1 that are significantly underrepresented by non-synonymous sequence variation reported in the ExAC database when restricted to European and Latin populations (Figure 3.5C). Additional variants found in Families 6 and 7 (c.4549A>C, c.4355G>A) also fall within a region that is underrepresented by non-synonymous sequence variation, although this last cluster is not statistically significant (p value of 0.29).

For the zebrafish studies,the relative area of the optic tectum was reduced by approximately 10% in embryos injected with a MO targeting the donor site of exon 9 of the *D. rerio* taf1 (p<0.0001) (Figure 3.7A-B). The effect is specific to the MO knockdown, as the phenotype could be reliably re-stored by co-injection of MO and wild-type human TAF1 mRNA (p<0.0001). Notably, overexpression of WT human TAF1 mRNA did not result in a phenotype that was significantly different from controls (p=0.79).

To confirm these findings, CRISPR/Cas9 was used to disrupt taf1. For the CRISPR experiments, taf1 guide RNA was generated as described[132]. In agreement with observations of a relative reduction in the area of the optic tectum in taf1 MO embryos with respects to control embryos, F0 taf1 CRISPR mutant embryos showed a relative reduction in the area of this structure with respect to un-injected controls (p<0.001) (Figure 3.7C). Unfortunately, the relatively small effects seen with MO knockdown and CRISPR-mediated mosaicism in F0 embryos precluded observations about whether differences exist between WT and mutated TAF1 constructs in terms of rescuing the neuroanatomical defect.

3.3.1 Family 1 specific results

Whole genome sequencing

Complete Genomics WGS was optimized to cover 90% of the exome with 20 or more reads and 85% of the genome with 20 or more reads (Table 3.3). Illumina WGS resulted in an average mapped read depth coverage of 37.8X (SD=1.3X). >90% of the genome was covered by 30 reads or more and >80% of the bases had a quality score of >30.

Concordance among variant detection pipelines

SNP and INDEL concordance across SNP and INDEL detecting pipelines applied to Illumina raw data was computed. In agreement with various other studies that have focused on computing SNP and INDEL concordance across pipelines, the mean concordance for SNPs across the two SNP detecting pipelines (GATK and FreeBayes) among the 10 sequenced individuals was 81.8%, whereas the mean concordance for INDELs between GATK and FreeBayes was 62.2% (with a mean of 80.3% of Scalpel calls being detected by the other two pipelines). Agreement between CNV detecting pipelines was low; with 6.3% percent of PennCNV found by ERDS and 0.9% percent of ERDS CNVs found by PennCNV. No known

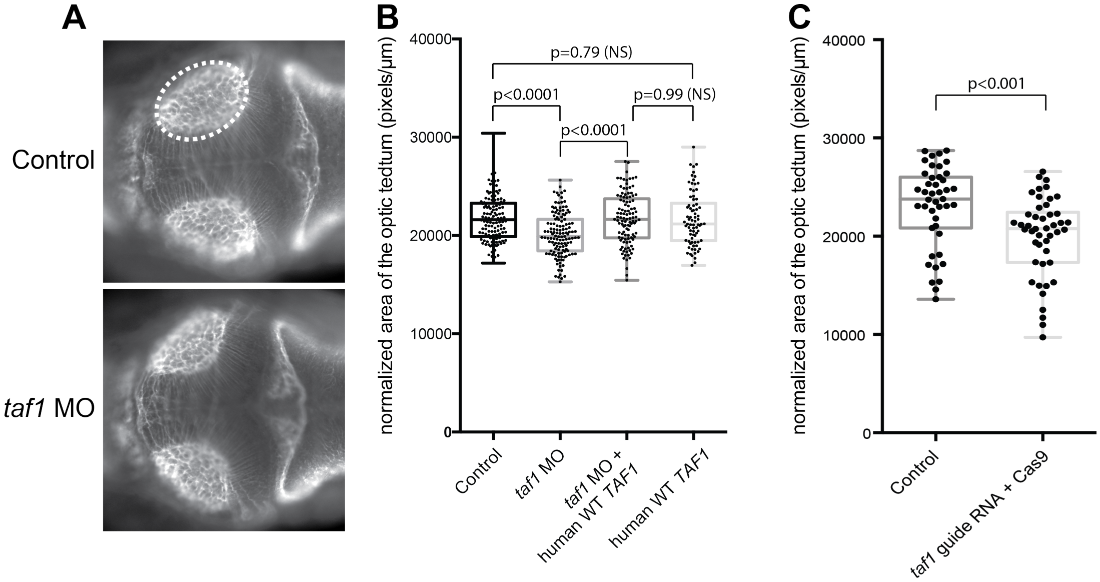


Figure 3.7: Suppression or Genetic Mutation of Endogenous taf1 Induces Decreased Size of the Optic Tectum In Vivo. (A) Dorsal view of a control embryo (top) and an embryo injected with a morpholino (MO) targeting the donor site of exon 9 of D. rerio taf1 3 days after fertilization. An antibody against α-acetylated tubulin was used for visualizing the axon tracts in the brain of evaluated embryos. The assay consisted of measuring the area of the optic tectum (highlighted with the dashed ellipse), a neuroanatomical structure that occupies the majority of the space within the midbrain. (B) A boxplot shows quantitative differences in the size of the optic tectum for each condition tested across three biological replicates. Suppression of taf1 consistently induced a decrease of □10% in the relative area of the optic tectum ($p < 0.0001$). The MO phenotype could be restored by co-injection of MO and wild-type (WT) human TAF1 mRNA ($p < 0.0001$), denoting the specificity of the phenotype due to taf1 suppression. Overexpression of WT human TAF1 mRNA alone did not induce a phenotype that was significantly different from that of controls ($p = 0.79$). The numbers of embryos evaluated per condition were as follows: control, 134; taf1 MO, 133; taf1 MO + WT TAF1 RNA, 109; and WT TAF1 RNA, 78. (C) A boxplot shows quantitative differences in the size of the optic tectum between uninjected controls and F0 embryos with CRISPR-disrupted taf1. The phenotype observed for both MO-injected embryos and embryos with CRISPR-disrupted taf1 was concordant and reproducible across different experiments and across the two different methodologies. The p values were calculated with a Student's t test.

Sample ID	Est library size	% PCR duplicates	Genome coverage of mapped bases	% bases Passing filter >Q30	% reads Passing filter aligning to genome	% of non-N ref bases covered	Ins size	Tight insertsize distribution
II-3	20703336864	2.05	40	0.883566667	0.9435	0.949	261	81
II-5	21316370174	1.67	38.15	0.903266667	0.9447	0.962	268	86
III-1	21317347309	1.75	37.76	0.8971	0.9459	0.941	262	84
III-3	21209346469	1.71	37.2	0.895866667	0.9439	0.938	253	80
II-4	22830743528	1.73	39.4	0.82	0.94	0.952	261	79
I-2	21587072423	1.72	37.1	0.85	0.94	0.963	269	82
II-1	20937437047	1.74	35.4	0.87	0.94	0.936	265	86
III-2	22626476447	1.8	38.2	0.84	0.94	94.6	262	85
I-1	19959631060	1.9	38.14	-	93.99	94.2	257	86
II-2	20095126759	2	36.68	-	94.04	95.8	271	93

Table 3.3: A table summarizing the Illumina HiSeq 2000 sequencing WGS sequencing data for Family 1. We report basic sequencing statistics generated by initial sequence data analysis performed by the sequencing facility, including estimated library size, percent PCR duplicates, genome coverage of mapped bases, percent bases with a base quality of above 30, percent reads which were passing filter aligning to genome, initial estimates of the number of SNPs and the number of novel SNPs along with initial transition to transversion ratios. We also report initial estimates of the percentage of heterozygous sites, the heterozygous to homozygous ratio, the percent of non-N ref bases covered, insert size and the tight insert size distribution.

disease-contributory CNVs were discovered, but we archive in our study 8 de-novo CNVs that are not currently associated with any biological phenotype (see Table 3.4 for the list of CNVs).

Between Illumina and CG sequencing platforms, the SNP concordance was 77.1% whereas the INDEL concordance was 44.8%. CNV concordance between the two sequencing platforms was 5.7%. To make these cross platform comparisons, variants generated from the different informatics pipelines applied to the Illumina raw data were combined into a larger set that only included unique calls from each caller.

Study design comparisons

We explored differences between study-design scenarios in their output in terms of variants conforming to the disease models (de-novo, autosomal recessive and X-linked, see Figure 3.8). Results were compared between two distinct study designs: a quad study design and a full family study, which integrates data from all of the sequenced family members as well as all of the variant detecting pipelines previously described. We found that there was a mean fold difference of 2.4 to 14.0 in the number of variants that were segregating in terms of the three different disease models (a mean fold difference of 2.4 was observed for the autosomal recessive disease model, 3.1 for the de-novo disease model and 14.0 for the X-linked disease model). For each disease model, simple python set operations, SVS operations and GEMINI operations were used to divide variants segregating according to each model. For the de-novo disease model, python set operations, SVS and GEMINI identified 52,360, 40,440 and 42,625 variants respectively for a quad-based study design and 17,908, 12,298 and 14,249 variants for a study design incorporating all of the family members recruited into the study. Similarly for the autosomal recessive disease model, 59,111, 57,614 and 64,366 variants were found using a quad study design and 37,678, 20,579 and 22,302 variants were found when using all of

Disease model	Location	Ploidy	CNV-type	Software	Function
De-novo	chr2:177266000-177272000	0	DEL	CG	intergenic
De-novo	chr6:256000-292000	3	DUP	CG	intergenic
De-novo	chr6:62200000-62206000	0	DEL	CG	intergenic
De-novo	chr8:11895601-12091800	0	DEL	ERDS	DEFB130,FAM86B1, LOC100133267,USP17L2, USP17L7,ZNF705D
De-novo	chr11:50326000-50440000	3	DUP	CG	LOC646813:ncRNA
De-novo	chr16:33846000-33848000	3	DUP	CG	intergenic
De-novo	chr16:55796000-55822000	1	DEL	CG	CES1P1:ncRNA
De-novo	chr22:24274601-24398600	0	DEL	ERDS	DDT,DDTL,GSTT1, GSTT2,GSTT2B,LOC391322

Table 3.4: CNVs were detected with Illumina HiSeq 2000 WGS data using the Estimation by Read Depth with SNVs (ERDS) pipeline, Illumina Omni 2.5 microarray data using the PennCNV software package and Complete Genomics sequencing. We report 8 CNVs following in the de-novo disease model. No CNVs were found to be segregating in an autosomal recessive or X-linked fashion.

the family members. Lastly, for the x-linked model, 26,228, 27,121 and 39,316 variants were found under a quad study design whereas 2,322, 2,538 and 1,958 variants were found when all family members were incorporated into the analysis.

We also explored differences in disease variation discovery due to varying prioritization schemes. We looked at these differences in combination with applying a quad or full family study design. In general, 40 variants were identified using a quad based study design and using the CADD scheme, whereas 14 variants were found using the same study design but instead using the Coding prioritization scheme, with only two variants being identified by both; a non-frameshift substitution in NLGN4X and a nonsynonymous variant in TAF1 conferring a p.Ile1337Thr change (Table 3.5). 8 variants were identified using the full-family based study design in combination with the CADD prioritization scheme whereas 7 variants were found using this same study design and the coding scheme, only one of which was found using both schemes; a nonsynonymous variant in TAF1 conferring a p.Ile1337Thr change.

Multi-generational pedigrees reduce erroneous findings

More variants are reliably eliminated when a greater portion of the family is incorporated into the analysis. This is likely due to varying false positive and false negative rates across sequencing and informatics platforms due in part to variation in data quality across the sequenced portion of each genome in each individual. Trio and quad-based study designs are prevalent in the literature[99,127,210,227], and many human genetics studies using high-throughput sequencing technologies only employ the use of a single, or a limited number, of variant detection pipelines. Our findings highlight the need for more comprehensive family-based study designs, and we demonstrate benefits in focusing high-throughput sequencing efforts on studying large related cohorts, where intra-familial relationships allow for more rigorous variant filtering and identification of true positive alleles that might be contributing to a disease

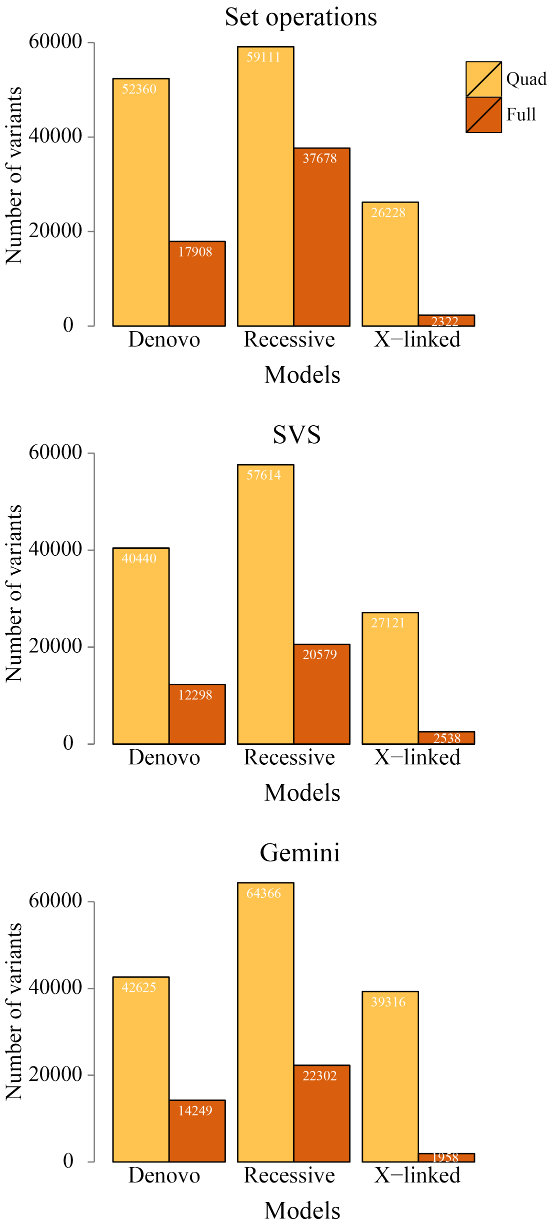


Figure 3.8: Bar plots showing differences in the number of variants conforming to de-novo, autosomal recessive and X-linked disease models using python set operations, SVS and Gemini software between quad and full family-based study designs in the Family 1 study.

Model	Location	Ref	Alt	VariantCaller	Annotation	Function	Scheme
Recessive	chr1:210851705	TT	T	CG, GATK, FreeBayes, RepeatSeq	ANNOVAR, GEMINI, SVS	KCNH1:UTR3	CADD, score:27.5
Recessive	chr1:224772440	AATAATTTG	TA	CG, GATK, FreeBayes	GEMINI	intergenic	CADD, score:22.1
Recessive	chr2:60537356	TTTTATTT	ATTATTA	CG, FreeBayes, GATK, RepeatSeq	GEMINI	intergenic	CADD, score:22.3
Recessive	chr8:109098066	AT	A	CG, FreeBayes, GATK, RepeatSeq	GEMINI	intergenic	CADD, score:24.6
Recessive	chr15:66786022	ACAAA	A	FreeBayes, GATK	GEMINI	SNAPC5:intronic	CADD, score:23.6
Recessive	chr16:49061346	TA	T	CG, GATK, FreeBayes	ANNOVAR, GEMINI	intergenic	CADD, score:25.3
Recessive	chr16:49612367	GAC	G	CG, GATK, FreeBayes	GEMINI, SVS	ZNF423:intronic	CADD, score:20.5
Recessive	chr10:135438929	T	G	CG, GATK, FreeBayes	ANNOVAR, GEMINI, SVS	I171L	Coding, gene:FRG2B
Recessive	chr10:135438951	GGCCC	AGCCT	FreeBayes, Scalpel	GEMINI, SVS	sub	Coding, gene:FRG2B
Recessive	chr10:135438967	C	T	FreeBayes, GATK	GEMINI, SVS	R158Q	Coding, gene:FRG2B
Recessive	chr15:85438314	C	CTTG	CG, FreeBayes, GATK, Scalpel	GEMINI	K141delinsIE	Coding, gene:SLC28A1
De-novo	chr1:53925373	G	GCCGCCC	CG, FreeBayes, Scalpel	GEMINI, SVS	A83delinsAAP	Coding, gene:DMRTB1
X-linked	chrX:34961492	T	C	CG, GATK, FreeBayes	GEMINI	Y182H	Coding, gene:FAM47B
X-linked	chrX:70621541	T	C	CG, GATK, FreeBayes	ANNOVAR, GEMINI, SVS	I1337T	Coding, gene: TAF1; CADD, score: 22.9

Table 3.5: A table of prioritized genetic variations in TAF1 intellectual disability syndrome. Variants conforming to the three disease models, de-novo, autosomal recessive and X-linked were identified. We show a list resulting from the CADD prioritization scheme as well as from the coding prioritization scheme. Both schemes required each variant to have a low population frequency (minor allele frequencey of less than 1%). The coding scheme required all variants to also be within a coding region of the genome and to be a non-synonymous change. The CADD scheme requires all variants to have a CADD score of greater than 20, along with the aforementioned population frequency. A variation in TAF1 was the only variation to be reliably detected using both prioritization schemes.

phenotype.

We were able to minimize false negative variant detections by using many orthogonal informatics pipelines, as each alone miss some true and possibly functional sequence variants but together capture a greater portion of the true call set. The multi-generational pedigree structure allowed us to minimize false positive findings by using expanded disease model operations that included three generations, effectively reducing false positive findings by corroborating genotypic evidence across the familial generations. In general, reductions in false negative and false positive calls should increase the efficacy of prioritization strategies, and reduce the number of candidate variants to a manageable and robust number in terms of performing validation and functional follow-up studies. In our study, we found this reduction to vary across disease models, with autosomal recessive, de novo and X-linked models having candidate variant reductions of 2.4, 3.1 and 14.0 fold respectively. Further, the number of final prioritized variants was reduced by a factor of 3.8 across both of the prioritization schemes that were employed (53 unique variants were identified through prioritization using a quad study design and 14 were identified using the full-family study design).

Before WGS was performed, a SNV of unknown significance was detected by clinical gene-panel sequencing: ZNF41; p.Asp397Glu. This variant was determined to be a variant of unknown significance due to the clinical ambiguity of the variant as well as the limited scope of the gene panel. There is some previous work implicating other variants in this gene as contributing to X-linked mental retardation[279], although during the course of this study, the significance of this finding was challenged[241]. When the study was expanded to include WGS data generated by CG on the two affected children and their parents, this variant was still identified as a putative disease-contributory variant. Only when a larger portion of the family was recruited for genotyping and additional Illumina-based WGS performed were we able to show that this variant was observed in other, unaffected, family members, including a male cousin.

We found this to be the case for other variants detected under a quad-based study design. For example, functional prediction algorithms (Polyphen and SIFT) indicated that another variant, located in ASB12, was deleterious and thus suspected as a potential disease-contributory variant. This inference was found to be invalid due to its presence in other unaffected members of their family (see Figure 3.1 for Sanger sequence traces which show ZNF41 and ASB12 variants to be present in other members of the family, despite being identified as important in disease using a quad-based study design). In another instance, a variant in PION was thought to be de novo in the children, but was found to be the result of poor sequencing coverage at that position, as this variant was indeed present in the mother, hence not de novo in the children (Figure 3.1). We have observed that some studies use trio or quad based designs and assert genetic "causality" when there is very little evidence supporting their case. This runs the risk of polluting the literature further with many false positive findings[125].

An extensive literature review was conducted in pursuit of genotype-phenotype correlations with the above variants. FRG2B and FAM47B are not known to be involved in the pathogenesis of any human disease, although the detailed molecular function of these genes has been largely unexplored. FRG2B is homologous to FRG2, which locates on chromosome 4 and has been implicated in playing a role in the pathogenesis of facioscapulohumeral muscular dystrophy (FSHD) in patients with substantial reductions in a 11-150 unit 4q35 microsatellite repeat[92,250,252]. However, reductions in the homologous 10qt26 microsatellite repeat, proximal to FRG2B, have not been associated with FSHD. ZNF423 acts as a transcriptional regulator, and variants in ZNF423 coding regions have been implicated in the pathogenesis of Joubert syndrome[37,106]. The variant that we have identified in ZNF423 is located within an intron, and its molecular function is unknown. SLC28A1 is thought to mediate sodium-depedent fluxes of uridine, adenosine and azidodeoxythymidine[255], whereas SNAPC5, also known as SNAP19, plays a scaffolding role in the forming the complete SNAP complex, which is required for the transcription of snRNA genes[113]. The molecular functions of KCHN1 and DMRTB1 are not well understood or studied.

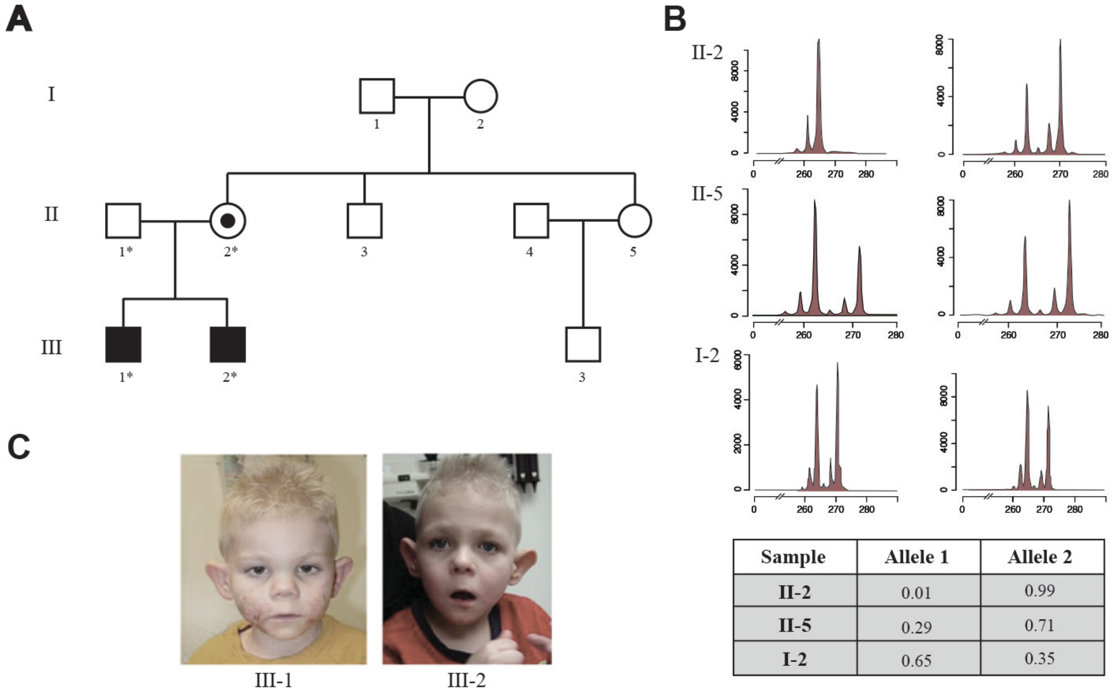
X-chromosome Skewing in Family 1

The X-chromosome skewing assay revealed that the mother of the two affected boys has skewed, 99:1, X-chromosome inactivation (Figure 3.9). The grandmother, as well as the aunt of the affected boys, does not show any appreciable X-chromosome skewing, suggesting the possibility of a newly arising deleterious X-chromosome variant.

RNA sequencing for family 1

RNA libraries were generated and the final pooled library was sequenced on a HiSeq 2000 across three lanes (paired-end 100bp). A mean of 49,792,652 (sd = 11,666,119) properly paired reads were generated and a mean spliced mapping percentage of 85.41 (sd = 7.1) per sample was observed (Table 3.6). HISAT[138] was used for spliced alignment to the UCSC human reference sequence hg19 and Stringtie[236] was used to quantify known transcripts. Cuffdiff[295] was used to perform differential expression analysis and CummeRbund[152] was used to analyze, filter and visualize the results from the Cuffdiff differential expression analysis. WebGestalt[304] was then used to perform gene set enrichment analyses using the Molecular Signatures Database (MSigDB) for transcription factor targets[288], KEGG[135], GO[48], and HPO[145] databases for gene annotations and set inclusion information. We used various analysis tools in the CummeRbund package[152] to evaluate the quality of our RNA sequencing data/analysis for Family 1 (Figure 3.10).

213 genes were found to be differentially expressed in the two affected male probands from Family 1 in comparison to their unaffected parents and grandparents. 179 out of the 213 genes



Sample	Allele 1	Allele 2
II-2	0.01	0.99
II-5	0.29	0.71
I-2	0.65	0.35

Figure 3.9: (A) Pedigree structure of all individuals in Family 1 that were sequenced during the course of this study and images of the two affected siblings, who display strikingly similar facial dysmorphology. Affected brothers, III-1 III-2, are sons to mother II-2, who tested positive for extreme X-chromosome skewing (B). Individuals with a star next to their number indicates that their whole genomes were sequenced with both the Complete Genomics sequencing and analysis pipeline as well as with Illumina sequencing technology and the various downstream analysis pipelines. All other numbered individuals had their whole genomes sequenced only with the Illumina WGS technology, followed by the downstream analysis pipelines described in the methods section. (C) The affected have distinctive shared facial features, including broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, prominent periorbital ridges, deep-set eyes, relative hypertelorism, a high-arched palate, and prominent ears.

Sample	I-1	I-2	II-2a	II-2b	II-1a	II-1b	U1a	U1b	U2a	U2b
LID	295782	295783	295784	295785	295786	295787	295788	295789	295790	295791
Illumina index	AR001	AR002	AR003	AR004	AR005	AR006	AR007	AR008	AR009	AR010
Index sequence	ATCACG	CGATGT	TTAGGC	TGACCA	ACAGTG	GCCAAT	CAGATC	ACTTGA	GATCAG	TAGCTT
Mapping percent	66.44	88.43	87.88	91.73	89.62	84.76	83.92	84.92	88.01	88.37
Properly paired reads	34561120	50224027	50812379	66555506	60522016	35767249	45865914	60357217	35409852	57851237

Table 3.6: RNA sequencing library numbers, adapter sequences and general data quality statistics.

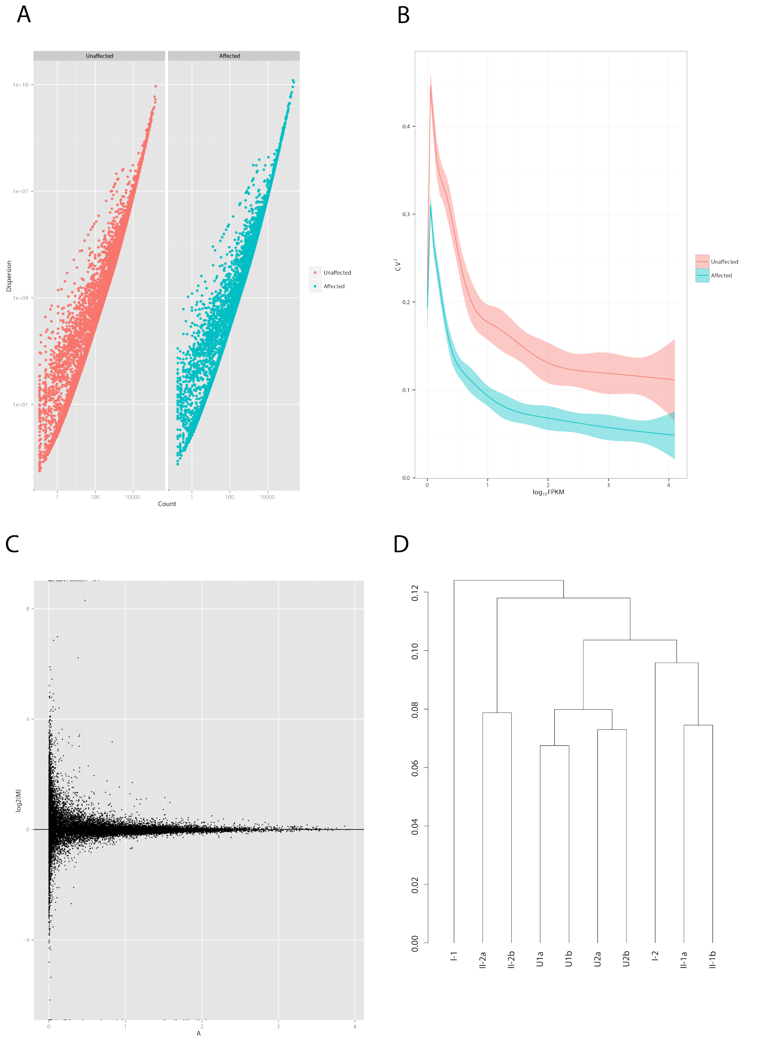
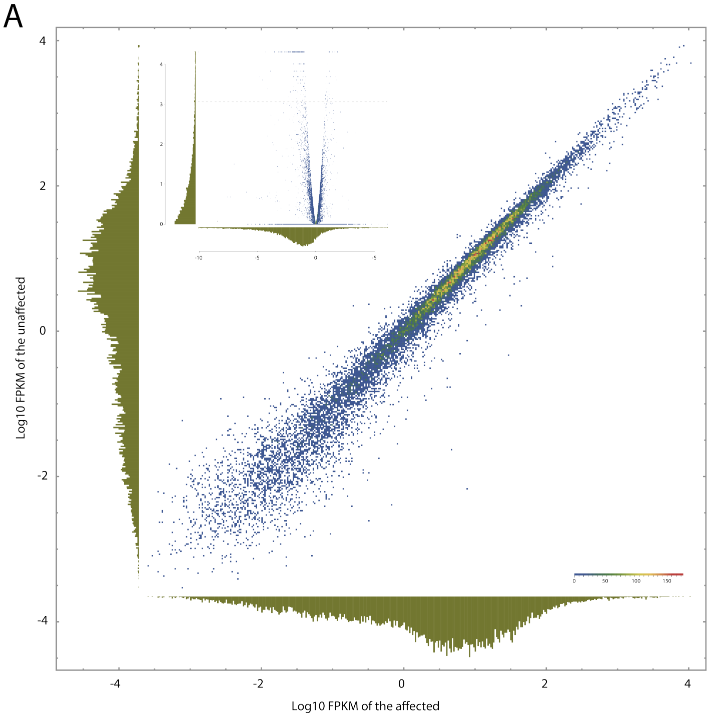


Figure 3.10: RNA sequencing data quality evaluations for Family 1. Various analysis tools in the CummeRbund R package47 were used to evaluate the quality of our RNA sequencing-based differential expression analysis for Family 1. (A) The degree of read count dispersion between the affected and unaffected groups was plotted; both groups appear to be quite similar in this regard. (B) The squared coefficient of variation in log base 10 FPKM values, used here as a measure of cross-replicate variability, was plotted. In general, we saw a higher degree of variability in the unaffected group than the affected group. (C) MA plot shows no obvious evidence of systematic bias between conditions. (D) The Jensen-Shannon distance (shown on the y-axis of D) was used to constructed Dendrograms between replicates (a and b) and groups (affected vs unaffected). The replicates are generally closer to each other than they are to other samples and their replicates. The replicates among the affected group (U1a-b and U2a-b) appear closer to each other than they are with any of the other replicate samples in the unaffected group (I-1, II-2a, II-2b, I-2, II-1a, II-1b).

were expressed less (down-regulated) in the affected, whereas 34 out of the 213 were expressed more (up-regulated) in the affected. Among those genes that were up-regulated in the affected, 24 out of the 34 were expressed with a log base 2 fold change value of greater than 1, and no genes were found to be up-regulated with a log base 2 fold change value of greater than 2. Among those genes that were down-regulated in the affected, 161 out of 179 were expressed with a log base 2 fold change value of less than -1, and 41 out of 179 were expressed with a log base 2 fold change value of less than -2.

Among the genes that were up-regulated in the affected boys, we found no obvious interpretable biological signal. We present the results from the gene set enrichment performed on the set of genes that were down-regulated in the affected with a log base 2 fold change of less than -1, as this set represent the genes that are significantly down-regulated and to a potentially biologically relevant level (Figure 3.11). Transcription factor target enrichment analysis revealed a significant enrichment for genes regulated by E-box proteins (CANNTG promoter motifs, BH corrected p-value of 0.0052). KEGG pathway enrichment analysis revealed an enrichment for genes involved in Parkinson’s, Alzheimer’s, Huntington’s disease, and cardiac muscle contraction (BH corrected p-values of 2.95e-07, 1.54e-06, 2.13e-06, and 2.13e-06, respectively). Enrichments in genes associated with HPO annotations include type 1 muscle fiber predominance, laxity of ankles, fingers and wrists (all four phenotypes were reported with a BH corrected p-value of 0.0481).

It is important to note that the RNA sequencing results are preliminary and potentially confounded by age or sex-specific expression differences between the affected and unaffected groups, and the results here were derived from a single family, as these were the samples that we were able to collect to date. Complete blood counts were not performed on the blood samples used for RNA sequencing; thus, our result could be also confounded by secondary differences in mRNA abundances. The RNA and whole genome sequencing data have been



B

Transcription factor target: CANNTG (E-box), p = 0.0052

Gene	Gene name	log2 fold change
TRPM4	Transient receptor potential cation channel, subfamily M, member 4	-2.17
CACNA2D3	Calcium channel, voltage-dependent, alpha 2/delta subunit 3	-1.18
KCNJ2	Potassium inwardly-rectifying channel, subfamily J, member 2	-1.06
GNB3	Guanine nucleotide binding protein, beta polypeptide 3	-1.92
SLC1A7	Solute carrier family 1 (glutamate transporter), member 7	-2.05
PTPRS	protein tyrosine phosphatase, receptor type, S	-1.31
IGFBP3	Insulin-like growth factor binding protein 3	-2.50
CLEC4D	C-type lectin domain family 4, member D	-1.52
XCL2	Chemokine (C motif) ligand 2	-1.04
PTGDS	Prostaglandin D2 synthase	-1.49
GPR162	G protein-coupled receptor 162	-1.25
LAMB2	Laminin, beta 2 (laminin S)	-2.14
COL5A3	Collagen, type V, alpha 3	-1.78
PVALB	Parvalbumin	-1.94
LEPREL2	Leprecan-like 2	-1.72
TNNT1	Troponin T type 1	-3.23
GZMB	Granzyme B	-1.19

Parkinson's, Alzheimer's, Huntington's disease; p = 2.95e-07, 1.54e-06, and 2.13e-06

Gene	Gene name	log2 fold change
NDUFB3	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 3	-1.28
COX7C	Cytochrome c oxidase subunit VIIc	-1.04
UQCRB	Ubiquinol-cytochrome c reductase binding protein	-1.91
COX7B	Cytochrome c oxidase subunit VIIb	-1.49
NDUFS5	NADH dehydrogenase (ubiquinone) Fe-S protein 5	-1.05
NDUFB1	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 1	-1.55
COX6C	Cytochrome c oxidase subunit VIc	-1.33
NDUFS4	NADH dehydrogenase (ubiquinone) Fe-S protein 4	-1.22

Type 1 muscle fiber predominance (HP:0003803); p = 0.0481

Gene	Gene name	log2 fold change
TNNT1	Troponin T type 1	-3.23
COL6A1	Collagen, type VI, alpha 1	-2.98
COL6A2	Collagen, type VI, alpha 2	-1.32

Laxity of ankles, fingers, wrists (HP:0006460,0006149,0005072); p = 0.0481

Gene	Gene name	log2 fold change
COL6A1	Collagen, type VI, alpha 1	-2.98
COL6A2	Collagen, type VI, alpha 2	-1.32

Figure 3.11: RNA sequencing results from Family 1. Blood extracted RNA from Family 1 was sequenced with stranded sequencing libraries and 100bp paired end reads on the HiSeq2000 platform. (A, large panel). Transcript quantifications (FPKM) were similar between the affected (U1 and U2) and unaffected individuals, suggesting no strong signal for large scale transcription differences involving thousands of genes between them. (A, small panel) 215 genes were found to be differentially expressed. The small panel in A represents a "volcano" plot, where the x-axis is the log base 2 fold change and the y-axis is the –log base 10 p-vale. The dashed line represents the threshold for significance. There were many more genes found to be down-regulated (180) than up-regulated (35). (B) Gene set enrichment analyses performed with WebGestalt(J. Wang et al., 2013) for down-regulated genes with a log base 2 fold change value of less than -1 revealed enrichments in transcription factor binding targets, disease pathways, and phenotypes related to TAF1 syndrome.

deposited to Sequence Read Archive (SRA) under BioProject ID PRJNA301337.

3.4 Discussion

Recent structural work in yeast points to an epigenetic role of the TAF1-TAF7 complex in general TFIID function and/or pre-initiation complex (PIC) assembly[20,19]. Human TAF1 is a 1893 amino acid long multifunctional protein that has been reported to possess DNA promoter binding, histone acetylation and protein phosphorylation activities[59,199,38,157]. The histone acetyltransferase (HAT) activity of TAF1 can be blocked by TAF7 binding[39,96]. Studies suggest that phosphorylation of TAF7 at Ser264 causes release from TFIID, alleviating its inhibitor effect on TAF1[143]. Therefore, mutations that disrupt this inhibitory protein-protein interaction could have devastating effects on gene expression profiles during human development. Intriguingly, four of the eight missense variants in TAF1 change residues that are conserved in higher eukaryotes and map to domains important for TAF7 binding. Variants p.Cys807Arg, p.Pro596Ser, and p.Asp976His from families 2, 8 and 9 fall within an evolutionarily conserved central domain (DUF3591) that spans residues 586-1049. DUF3591 encompasses the TAF1 HAT domain and numerous points of contact with TAF7[303]. The recently reported human TAF1-TAF7 crystal structure reveals that Cys807 is buried in the center of a hetero-dimeric triple barrel formed by segments of TAF1 and TAF7[303]. Replacement of the cysteine, a polar amino acid capable of disulfide bond formation, with the large basic amino acid arginine (p.Cys807Arg) is predicted to destabilize the triple barrel fold and potentially interfere with the interaction of TAF1 and TAF7. The p.Asp976His variant also has the potential to disrupt TAF1-TAF7 binding. The acidic Asp976 maps to a separate TAF1-TAF7 protein interface within DUF3591 and undergoes intermolecular hydrogen binding in the highly conserved glycine-rich loop of TAF1. This loop interacts extensively with a highly conserved Arg-rich motif in TAF7. The acidic to basic amino acid change (p.Asp976His) has the potential to disrupt the architecture of the glycine-rich motif and its ability to effectively bind to TAF7. The p.Arg1246Trp mutation described in this work and the published p.Arg1190Cys TAF1 variant[121] reside in the RAP74 interacting domain (RAPiD) of TAF1 (residues 1120-1279), which also has been shown to be important for TAF7 binding[266,303]. The number of TAF1 de novo missense mutations co-segregating with intellectual disability syndromes and predicted to affect TAF7 binding is quite striking, further strengthening the importance of TAF7 in the regulation of TAF1 function. TAF1 is a difficult protein to work with in isolation and it is likely to be unstable in the absence of a binding partner, such as TAF7[96,19,303]. Thus we speculate that variants that are not within any known protein domain (i.e p.Ile1337Thr) may affect domain packing of TAF1, which may interfere with the TAF1-TAF7 interacting surface[19,303] or mark the protein for proteolytic degradation.

Bromodomains are a common feature of transcription factors that possess HAT activity. TAF1 contains two bromodomains (Bromo1: 1397-1467 and Bromo2: 1520-1590), each of which consist of a bundle of four alpha helices that form a hydrophobic pocket that can recognize acetylated lysines found on the N-terminal tails of histones[130,81]. A close examination of the published TAF1 bromodomain structure reveals that the p.Arg1431His mutation in Family 7 is a surface residue on one face of the alpha helix. It forms hydrogen bonds with two residues on a nearby helix, most likely playing a supporting role in maintaining the bromodomain fold. Mutation of p.Arg1431His to a histidine could affect the stability of the acetyl-lysine ligand binding site in TAF1 and alter promoter recognition.

3.4.1 Speculation about the phenotypic spectrum of *TAF1* syndrome

Phenotypic variance among TAF1 syndrome probands is not unexpected, as it is not uncommon for syndromes whose pathogenesis is linked to mutations in large genes or in genes with many interacting partners spanning many functional domains to vary widely in their phenotypic presentation[180,270,319]. Indeed, proteins affecting fundamental and more global cellular processes, if disrupted, are expected to exert an abnormal effect on a wide range of cellular processes. Given that a protein can be mutated in many different ways, mutations affecting proteins with important global functions may have varied effects, particularly for large proteins with many functional domains[270,319]. If mutations in such a protein lead to disease, the implication is that the disease phenotype, depending on the particular mutation and environmental influencers, could be quite variable, with a range of disease phenotypes manifesting across affected individuals[180] (Figure 3.12; G1). In contrast, some proteins have more constrained cellular functions that are responsible for regulating or affecting a more limited number of cellular processes. If mutations in these proteins lead to human disease, one would expect that, in most or the majority of cases, the disease phenotype will be less variable to reflect the constrained function of the protein (Figure 3.12; G0). TAF1 is a large, multi-domain protein that is involved in general transcription initiation, a ubiquitous cellular process. Consequently, mutations spanning different TAF1 domains are expected to result in potentially discordant, but related, clinical presentations for TAF1 syndrome. Although the phenotypic overlap among all probands in this study is remarkable, each family has their own idiosyncrasies that can be explained by differences brought about by mutations occurring in different TAF1 functional domains or regions, alongside other genetic and environmental influences. The genetic link between the disease phenotype and the larger duplications that were found in two probands included in our study suggests a larger role for the surrounding genomic region in various human disease phenotypes.

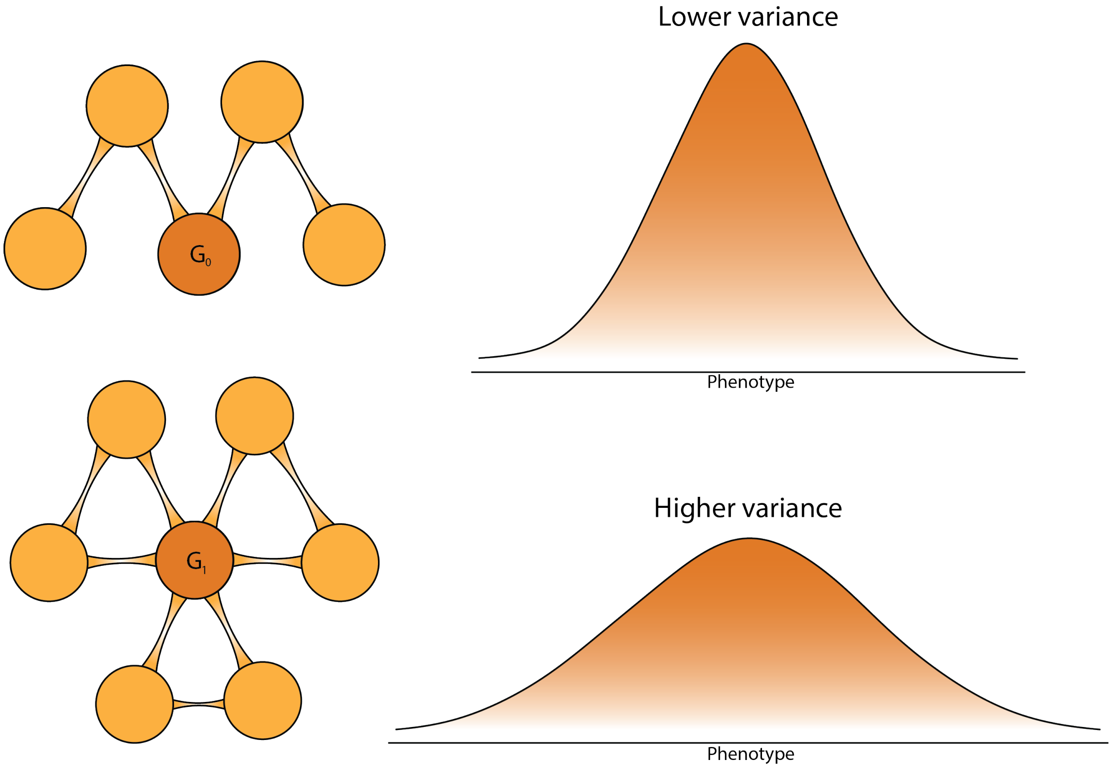


Figure 3.12: Conceptual figure linking phenotypic variability to characteristics of prototypic genes and proteins associated with human disease. Proteins that affect fundamental or global cellular processes, if functionally mutated, are expected to result in a range of cellular consequences. Mutations leading to human disease that are due, in part, to changing such proteins may have more varied molecular and as a consequence phenotypic effects. For example, gene G1 is shown to have a higher node degree than G0, resulting in a phenotypic spectrum (depicted here as a continuous distribution) with a higher variance than G1, which, in contrast, has a more constrained cellular function, leading to a disease phenotype that is less variable, reflecting the constrained function of the gene and its protein product.

3.5 Concluding remarks

In conclusion, we have presented evidence showing that mutations involving TAF1 contribute to the phenotype described here. Differences in genetic background and the environment can certainly account for the phenotypic differences between the various males in these families[308,118,123,23,33,183,180]. Other studies also suggest a functional link between developmental delay and the TFIID multi-protein complex[262,208,112], although the phenotypic variability and expression of other variants in TAF1 and in other TAFs remains to be determined. We have also provided the results of initial studies that suggests a possible regulatory role for at least one variant presented here, and have shown that zebrafish knockdown and mutant studies for this gene have a quantifiable, albeit small, effect on a neuronal phenotype. In our current study we do not see a clear mechanistic link between the missense and splice-site variants versus the duplications in two of the probands, and the duplication probands share less phenotypic features with the rest of the disease cohort. There are many possible explanations for why a duplication of a gene might mimic the effect of a disruptive missense or splice site variant, and our current study lacks the necessary experimental evidence to point to any particular scenario with certainty. Further work is needed to tease apart the contributions of these duplications and their constituent genes to this more complex phenotype. It should also be noted that exome and/or whole genome sequencing have not yet been performed in Family 10 or 11, so there could be other mutations contributing to the phenotypes of the probands.

4

Better accounting for uncertainty in DNA sequencing data

4.1 Motivation

Personalized and genomics-guided medical care has the potential to enhance the way we treat and prevent human disease by relying more heavily on highly accurate and rich characterizations of individuals, rather than on population scale phenomenon. Thus, it is becoming increasingly important and relevant for analysis methods to guarantee rigorous accounts of individuals, whose medical treatment will be shaped by these results. At the forefront of personalized medicine is DNA sequencing, and relatively standardized methods for analyzing these data have been developed[9,295,49,221,314,108]. As sequencing becomes more routine in the clinic, it is important to consider the accuracy of these data and the validity of the conclusions based on them.

DNA sequencing data contain two major types of quantitative uncertainty, which we refer to here as aleatory and epistemic. Aleatory uncertainty refers to the variability that is inherent to most biological systems, such as stochastic fluctuations in a quantity through time or variation across space. This is considered to be a form of uncertainty because the value of the quantity can change each time a measurement is taken, and we cannot predict precisely what the next value will be[77]. In the context of DNA sequencing studies, variability can arise from sequencing DNA that has been extracted from tissues with differing genotypes, sequencing large populations for the purpose of determining population allele frequencies, and from varying RNA expression levels across space or time, just to name a few sources. Epistemic uncertainty, on the other hand, refers to incomplete knowledge about a quantity, which can arise from imperfect measurement, limited sampling effort, or ignorance about the underlying processes that influence a quantity[77]. Again, in the context of DNA sequencing, this type of uncertainty can result from poor base detection, sparse sequence data, or from a fundamental lack of understanding about how sources of error arise in these data and how they are related. These two forms of uncertainty have important practical differences. For example, epistemic uncertainty can in principle be reduced by empirical effort, and although aleatory uncertainty (variability) can sometimes be better characterized through repeated experimentation, it cannot generally be reduced by empirical effort. The differences between these two forms of uncertainty are often significant in practical settings.

Substantial progress has been made toward quantifying and propagating uncertainty through calculations and inferences made on human sequencing data, and this progress has already yielded more accurate characterizations of population scale phenomenon[282,300,283,91,148,139]. However, these methods are limited in that they are not easily extended for use in the many different, and often times piecemeal, analyses that are currently necessary for implementing personalized genomics. Furthermore, the treatment of uncertainty is currently limited in scope, due to the difficulties inherent in modeling different sources and types of uncertainty that often arise in personalized genomics-based analyses. A general framework for propagating uncertainty through calculations is needed, so that computations can be made with as much rigor as is possible given the available technologies.

We describe sources of error that arise in high-throughput sequencing data, describe components of these errors that are the consequence of manifestations of different types of uncertainties, advocate for the development of new DNA sequencing analysis methods for computing with uncertain data in the context of personalized genomics applications, and describe prospective methods that allow for incorporation of uncertainties into their computational frameworks so that genomic inferences can more accurately represent the true state of knowledge.

4.1.1 Sources of uncertainty in DNA sequencing

DNA is composed of categorically defined units, nucleotide bases. Nucleotides, or sequences of nucleotides, consist primarily of adenine (A), guanine (G), thymine (T), or cytosine (C) bases. In practice, the reliability of DNA sequence detection varies from base to base and is usually influenced by the specific sequencing technology being employed. High-throughput and short-read sequencing technologies generally quantify detection reliability using probability values that characterize the chance that a base was correctly detected[71,251]. This value depends in part on the chemistry used in the sequencing, the particular equipment being used to detect DNA, sequence composition, among other things. Relatively short sequence reads (150 consecutive bases) are generated through iterative and consecutive base detection, which are then aligned to a reference sequence. The alignment procedure also generally produces probability values that characterize the chance of having correctly aligned a sequence to its respective genomic location. These two values taken together give the analyst information about the chance that a base is correct.

Errors in high-throughput sequencing data (base calling errors, variant calling errors, etc.) result from complicated and sometimes unforeseen technical and data processing factors. Several empirical studies have identified a number of different important and quantifiable sources of error. In a recent perspective piece[258], the authors review these different sources, which include upstream steps during sample preparation and sequence library preparation, as well as from the sequencing, imaging, data processing and bioinformatics steps[225]. More specifically, errors originating from sample preparation are sometimes due to a combination of human errors in sample handling (which can include sample swaps or DNA/RNA degradation), sample contamination and low quantities of input DNA. During the preparation of sequence libraries, human errors can result in cross-contamination of DNA samples across different library preparations, and errors can occur when PCR amplification incorporates an incorrect base during early synthesis cycles[258]. Primer-mediated sequence amplification biases, the synthesis of chimeric reads, barcode or adapter errors and machine failures are among the other sources of error that originate during sequence library preparation. During base imaging and sequencing, user errors combined with the incorporation of additional bases during single sequence cycles, DNA damage, overlapping signals, strand biases, sequence complexity[243] and machine failures can contribute to sequence error. Moreover, errors in bioinformatics steps resulting from poor sequence alignment in regions where mapping is difficult[159] also contribute to sequence error.

Analysts use a variety of algorithmic and statistical approaches for mitigating errors and for quantifying uncertainties about DNA sequence related estimates. Contemporary tools leverage efficient sequence alignment-based frameworks for detecting similarities and differences between sample and reference sequences[161,57]. Initial analysis steps entail excluding putative sequence error or low-quality sequences using data quality thresholds (i.e., if a sequence is of x or less quality, then exclude it from the analysis). Variations of the Smith-Waterman[285] algorithm are then used to match and align similar sequences. Once the sample sequence has been aligned to the reference sequence, various statistical approaches are used to identify the most likely genotype, including Bayesian inference[57,94], frequentist hypothesis testing[307], and others[318]. More recent algorithmic enhancements use local sequence assembly to mitigate errors caused by aberrant alignments and to detect complicated sequence differences between the sample and the reference sequence[253,209,166]. 'Error correction' and hybrid-sequencing approaches generate high quality sequence data by correcting error-prone long read technologies with high-fidelity short read sequencing technologies[158,147]. These error correction techniques enable reference-free assemblies of larger genomes, reduce sequence alignment artifacts and allow for the sequencing of genomes with no known reference sequence.

There are two major deficiencies in current analysis approaches with respect to the appreciation of uncertainty and errors in DNA sequencing data. The first deficiency is that epistemic uncertainties are currently not well quantified. The second is that uncertainty, even if quantified, is often not properly incorporated into subsequent analyses and calculations. These limitations are compounded by the fact that there are no software implementations that allow for uncertainty quantifications to be carried through and used in routine downstream calculations and analyses. Moreover, software libraries for computing with epistemically uncertain data are almost non-existent and are not widely available to most practicing bioinformaticians.

We view the accurate representation and incorporation of uncertainties into DNA sequencing analyses as a necessary piece of a more general computational solution that generates full, honest and robust determinations of the reliabilities of inferences stemming from these data. It is important to recognize that heuristic filtering approaches essentially discard imprecise or poorly collected/understood data, but these data should and can be included in the analysis.

4.1.2 Quantifying uncertainty

Among the various sources of error and uncertainty discussed above, those that are known to arise through random processes or are a result of system-level variability can be modeled using established statistical approaches. They are often times aleatory in nature because more sequencing data will not reduce the resulting allelic variability; it would simply result in a more precise characterization of it.

When errors are not known to be random and could instead be systematic in nature, then established statistical approaches may not always conveniently apply. Systematic errors are, in some cases, due to inaccuracies in instrument calibration or due to a bias in the way that the instrument takes particular measurements[191]. As just one example in the context of DNA sequencing in personal genomics applications, if either the reverse or forward DNA strand is sequenced more than what one would expect through a random sampling of each, it is considered evidence of potential sequence bias. This type of bias leads to uncertainty about the underlying sequence composition that is epistemic in nature; the analyst simply does not have representative samples of both DNA strands. In these situations, the analyst has less or sometimes even no information about what the other strand looks like in terms of its allelic composition. This creates epistemic uncertainty about the true genetic sequence, and therefore one cannot reasonably make any distributional assignments in lieu of the missing data.

In situations where distributional assignments are not well justified, epistemic uncertainties can often be conveniently modeled using intervals[78]. Intervals are a reflection of our inability to assign any distributional information to the various possible states of some variable of interest. The state of the art in DNA sequence analysis simply discards or ignores data that show evidence of systematic error. But this is unnecessarily strict and throws away potentially useful information. In the case of sequence strand bias, one conservative strategy might be to construct an interval (in this case, a set) that includes all possible bases. Although such an interval is a quantitative admission of ignorance, it can accurately represent the epistemic uncertainty that results from one strand being sequenced less often than expected. This uncertainty concerns one of the two DNA strands, the other of which may be well characterized. Instead of discarding unreliable data, charactering its uncertainty and carrying these uncertain data through analyses, although not always resulting in something easily interpreted, can in many cases be useful. It remains an open question as to how one should model systematic uncertainties in sequence data. Schemes are needed for understanding and better modeling systematic uncertainties so that they can be accurately quantified and propagated through calculations.

In the sub-sections below we suggest two, of many possible, aspects of DNA sequencing analysis where improvements in quantifying and propagating uncertainties can be achieved using existing computational tools. We provide two hypothetical situations, in Box 1 and 2, which exemplify how analysts might perform these computations in situations of pervasive uncertainty. These methods and examples should not be taken as comprehensive solutions to the problem at hand, but rather propositions. The challenge of accurately quantifying both epistemic and aleatory uncertainties in DNA sequence related data sets and then subsequently propagating them through analyses is not a solved problem. We hope to stimulate some discussion so that comprehensive solutions might, in the future, be found.

4.1.3 Uncertainty about dependencies

Uncertainties about dependency relationships are an important consideration when combining sources of error for making inferences based off of high-throughput DNA sequencing data, as errors can result from varied or shared processes. As an example, multi-sample variant callers use information spanning many different samples to generate calls. This means that the variant detection between individual samples is not a completely independent process, and so performing logical operations on these variants using information about their probability of being correct cannot be done assuming their independence (Figure 4.1). When knowledge about the dependency relationship between two variables is unavailable, their conjunction (e.g., the probability of the joint event A & B) can be computed using the formula , where a represents the probability of some event, A, to occur and and b represents the probability of another event, B, to occur[78,90]. Similarly, the logical disjunction can be computed using the formula . These represent the tightest possible bounds on calculations when we do not know the dependence; they quantify our epistemic uncertainty about the underlying relationship between A and B. It is in these cases that we are left to compute with intervals or even imprecise probability distributions (if the input operands are distributions), rather than simple point values and precise distributions. Interval analysis[213,2,202,201,66] is the simplest method for performing arithmetic and logical operations on interval data, and the results can be made as precise as possible given the input values. Probability bounds analysis[77,88,310,74,76] allows quantities with epistemic uncertainty represented by intervals to be combined with random distributions representing aleatory uncertainty in mathematical expressions. This method is part of the theory of imprecise probabilities[302]; it allows calculations when only bounds on the input distributions are known.

4.1.4 Probability of shared alleles under conditions of uncertainty

Genomic variants can act as predictors of disease, disease progression, and outcome. In cases of inherited genetic disease, even seemingly trivial uncertainty calculations may make an important difference.

For instance, imagine a case where a pathogenic variant has been detected in a mother with a 0.9 probability of being correct, but was detected in the son with a 0.5 probability of being correct (Figure 4.1). Typically, analysts filter out low-quality variant calls, which, in this case, would result in an analysis that says the mother and son do not share this variant. Instead of filtering the lower-quality variant, one could instead calculate the chance that this variant is present in both the child and the mother. If we assume that the variant detection's for the mother and son are independent, then the probability that they both have the variant is $0.9 \times 0.5 = 0.45$.

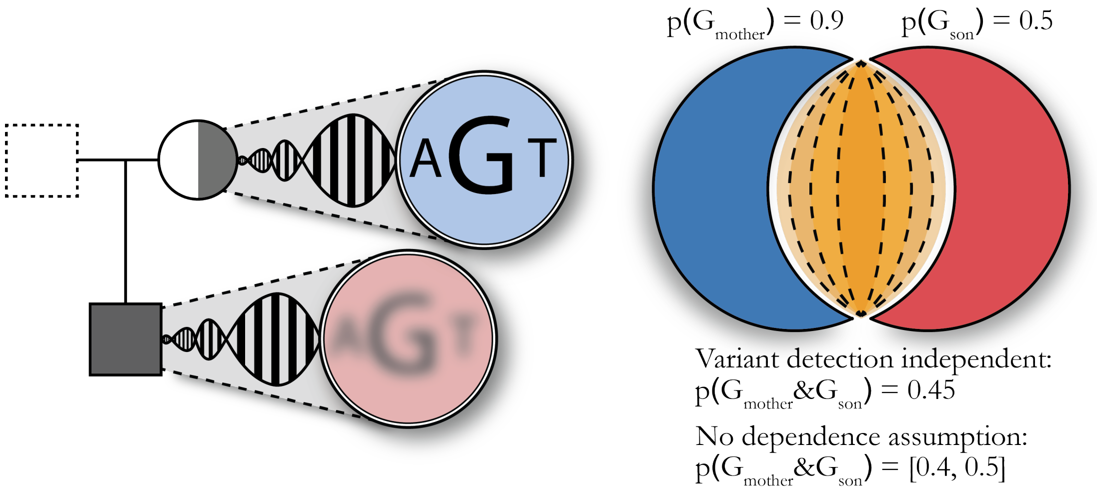

A G T
A G T
p(G_mother) = 0.9
p(G_son) = 0.5
Variant detection independent:
p(G_mother&G_son) = 0.45
No dependence assumption:
p(G_mother&G_son) = [0.4, 0.5]

Figure 4.1: Here, we demonstrate how uncertainty about dependencies results in requiring the necessary framework for computing with interval values, which are quantitative representations of epistemic uncertainty. Uncertainty about variant detection depicted as blurring (left) and encoded as probabilities that are logically conjoined (right) using calculations that assume independence or calculations that make no assumptions about their relation.

In many practical sequencing applications, variants are detected using information spanning multiple related or unrelated cohorts, so genotype inferences are no longer made in an entirely independent manner. Performing the same logical operations, but instead assuming nothing about the dependence relation between the two estimates, results in the simple conjunction 0.5 & 0.9 degenerating into an interval answer, in this case [0.4, 0.5].

It is important to note that for pedagogical purposes, we assumed in this example that the variant quality scores (which are translated into probability values) were computed with absolute precision, and that they accurately represent uncertainty about the call. In practice, this is rarely if ever the case, despite the fact that they are almost always reported as such. Indeed, Phred quality scores that are less than 30 are considered unreliable, which translates into a 0.999 lower bound on variant calling accuracies. These quality scores have been shown to underestimate the probability of variant calling errors made by various different variant-calling algorithms[225,209]. This is perhaps a more pervasive and important issue needing attention from the field, although some statistical and algorithmic approaches have been developed for generating more accurate quality scores[57,220,296].

4.1.5 Model uncertainty

Statistical models inform most DNA analysis algorithms, but this task is often made difficult by sparse data sets, a lack of prior knowledge for use in inferences, and the presence of imprecise or otherwise uncertain input data (due to, for example, noisy or poor raw sequence data). Methods that can make inferences in the presence of these complicating difficulties are needed.

Robust Bayesian[16,17,124,234,235,203] inference allows for the consideration and analysis of imprecise sample data or ignorance about the appropriate prior assumptions (both of which are the manifestations of epistemic uncertainty and can be modeled using intervals, bounding approaches and robust Bayesian approaches). In a robust Bayesian analysis, results are considered robust if neither imprecision in the input data nor differences in the prior probability distribution have large effect on the output.

Like robust Bayesian approaches, confidence structures[79,11,75] characterize inferential uncertainty about statistical estimates from sparse or imprecise data, but the confidence structure approach does not require the use of prior knowledge. Confidence structures are similar to Bayesian posterior distributions because they estimate distribution parameters from sample data. They give us confidence intervals for all levels of confidence (see Figure 4.2), and, with them, analysts can guarantee statistical performance through their repeated use. Confidence structures are useful in practical applications because they can be used in arithmetic or logical calculations, and the results will still guarantee statistical performance, i.e., they will still yield true confidence intervals.

4.1.6 Statistical inference under conditions of pervasive uncertainty

A common task in sequence analysis is to determine sequence composition in single- or multisample data sets. This task is often made difficult by sparse data sets, a lack of prior knowledge for use in statistical inferences, and the presence of imprecise or otherwise uncertain input data (due to, for example, noisy or poor raw sequence data).

Imagine a data set comprising sparse DNA sequences generated from an individual of a never-before sequenced population. The analyst is initially faced with two distinct problems inherent in the particularities of the experiment: sparse data and no prior information about expected allele frequencies. Suppose the available data come from eight sequence reads, and we observe {T, T, A, T, T, T, T, T} at a particular locus (Figure 4.2), with each base characterized as being detected with some degree of uncertainty, due either to known systematic error or from combining estimates of error from multiple sources. The analyst is now faced with a third, and perhaps more difficult, analysis challenge: figuring out how to incorporate epistemic uncertainty about base detection in the quantification of allele frequencies.

For the sake of simplifying this example, let us say that the bases are recorded with accuracies represented by interval probabilities: [0.9, 0.99], [0.8, 0.9], [0.4, 0.6], [0.8, 0.9], [0.98, 0.99], [0.7, 0.9], [0.4, 0.9], [0.8, 0.85], respectively. An analyst could then use a simple model that computes allele frequencies as probabilities from a multinomial distribution. With each base itself modeled as a Bernoulli process with interval probabilities, a confidence structure on allele frequencies can be obtained using the formula $I_x([k_A, k_A + 1], [nk_A + 1, nk_A])$, where k_A and n are the number of observed alleles of interest out of the total, respectively, and I_x is the regularized incomplete beta function [66]. The resulting confidence structure does not assume or require prior knowledge about the expected allele frequencies and we can, from it, compute confidence intervals about the estimate. In particular, we can say with 95% confidence that the allele frequency of adenine at this locus is between 0.002 and 0.526. This broad range reflects the fact that there were only eight data samples, no prior allele frequency information, and uncertainty about each base call.

4.2 Simulation studies

4.2.1 Simulation study design

Although conceptually appealing, the confidence structure approach to quantifying uncertainty in high-throughput sequencing data has not yet been applied even to simulated data. To determine whether this methodology (i) accurately propagates uncertainty in sequence data, and (ii) does so in a way that improves estimates based off of these data in comparison to standard methods, we performed a simulation study. Illumina sequencing data was generated for three people, each with a data set that mimics 10X, 20X and 30X average sequencing coverage.

1,000 variant sites were randomly chosen across the genome to be the focus of the simulation study. To frame the simulation in terms of an analysis that is routinely performed in human genomics labs and in clinical study, two of the three individuals will be considered ‘parent’ samples, while one will be considered the ‘child’ sample. Both parent samples will be given 500 unique SNVs, with 250 being in the heterozygote form and 250 being in the homozygote form. The child sample will receive all variants in the same form as they are represented in each parent. Variants will be detected using the GATK HaplotypeCaller and de novo variants will then subsequently isolated by returning the set of variants that are present in the child, but not in either parent. Note that since all variants were ‘inherited’ (that is, present in one or the other parent), there are no true de novo variants. As a consequence, any de novo variants detected in the child represent false negative calls in the parents. Comparisons will be made between the false de novo calls made by GATK, and the confidence structure analog.

It is important to note that the simulations will be evaluating two fundamental tasks in variant calling. The first is task is accurately determining an individual genotype given limited sequence data, and the second is computing the probability of a shared genotype across related individuals, where the genotype is determined with a degree of uncertainty for each member. The de novo analysis is designed to efficiently compare both aspects of variant calling between GATK and the confidence structure method for sites where GATK failed to provide a correct answer. De novo calls will be accurate if and only if genotype calls in the individual members of the family are accurate. Exploring characteristics of the joint calls (the de novo calls) will serve to evaluate the methods ability to propagate uncertainty in the underlying genotypes to the joint inference.

Although limited in terms of its ability to fully evaluate the performance of the confidence structure methodology, this pilot study will reveal key features and differences between the confidence structure method and the conventional analysis performed with GATK. Further studies will be needed in order to evaluate the sensitivity and specificity of genome-scale variant calls with the confidence structure method. Of course, the confidence structure method is also flexible in terms of being able to model a wide variety of error processes (both aleatory and epistemic in nature), and so a careful consideration and evaluation of these errors will be needed before the confidence structure methodology can be applied in more practical settings.

Illumina sequence data were simulated using ART, version 03.19.15, with 150 bp paired end reads. For both the confidence structure method and GATK, raw sequence reads were aligned to the hg19 human reference sequence, and duplicate reads were marked using Picard MarkDuplicates version 2.0.1. GATK HaplotypeCaller was used to call variants according to the developers recommendations, and de novo calls were isolated by returning the set of variants that are present in the child, but not in either parent.

Confidence structure calls were made by considering, for each sample, all possible genotypes. For each sample, supporting reads were extracted from the same alignments used for the GATK pipeline, using the samtools pileup function version 0.1.19-44428cd. Homozygous genotype probabilities were formulated in the following way: $p(G_g) = p(A_a|d)^2$, where d represents reads that support the A_a allele. g, in this case, represents the homozygous subset of all possible genotypes and a represents their corresponding alleles. $p(A_j|d)$ is modeled with a multinomial confidence structure, which is constructed using a special formulation of the regularized incomplete beta function; $I_x([k_A, k_A + 1], [nk_A + 1, nk_A])$ (see above for more details). Heterozygote genotype probabilities were formulated in a similar way: $p(G_g) = 2 \cdot [p(A_{ai}|d) \cdot p(A_{aj}|d)]$, where i and j are matrix coordinates of the non-homozygous components of a Cartesian product between two $\{$A, T, G, C$\}$ sets. De novo calls were generated by performing pairwise logical conjunctions using the resultant individual genotypes that were computed for both the child and parent samples. If the most likely genotype was anything other than the true underlying genotype, then it was considered a false de novo call.

For the purposes of this small pilot study, all probability bounds operations were performed assuming variable independence.

4.2.2 Results

Out of the 1,000 loci assessed, GATK calls resulted in 21, 49 and 113 false de novo calls in the 10X, 20X and 30X data sets, respectively. In comparison, the confidence structure method resulted in 2, 5 and 26 false de novo calls in the 10X, 20X and 30X data sets, respectively. In other words, using the confidence structure methodology recovered 94%, 90% and 77% of the variants missed by GATK.

The vast majority of variant sites that were miss-called as reference by the GATK have a 0 value for non-reference genotype probabilities. Of course, this invariance is not very informative. In contrast, the confidence structure method accurately characterizes the underlying uncertainty, and can be expressed in simple terms by looking at the width of the resulting confidence interval (Figure 4.3). Indeed, if we plot the distribution of confident interval widths from the true positive sites and false negative sites in the parents (with respect to GATK calls), the widths of the false negative sites are significantly larger than those of the true negative sites (see Figure 4.5, for all panes $p<0.05$, Kolmogorov–Smirnov test).

Importantly, if genotypes that are used as input for the de novo analysis are uncertain, the result of the de novo analysis is also uncertain. In general, this operation using GATK calls does not result in an accurate representation of the uncertainty in such an analysis, because computing with input genotyps from GATK calls in such a way that accurately propagates the uncertainty is not a straight forward process. For confidence structure genotype inferences, that uncertainty propagation is straightforward, and our initial analyses suggest that these structures accurately represent the underlying uncertainty (Figure 4.4).

4.3 Concluding remarks

4.3.1 Uncertainty in DNA sequencing data

As DNA sequencing technologies migrate from science to commerce, incentives to publish full accounts of error rates diminish. Even if commercial entities have access to large-scale in-house sequence validation data, proprietary and business interests may prevent open publication of these critical data. Commercial entities now routinely provide fee-for-analysis services, but cannot, or otherwise choose not to, release specific information about data processing procedures used in the analyses. Although practically useful, and perhaps even reliable for the problems at hand, this practice fundamentally violates basic conventions required to call the endeavor science. Furthermore, many of the most accurate and most widely used software tools for sequence alignment and variant detection are, in fact, closed-source and their algorithms are described only in general terms without explicit, published definitions (e.g., novoalign, GATK HaplotypeCaller). As such, they are not available even to scientists working on applied problems. In principle, software tools can be partially validated against synthetic data whose true nature is known and specifically described, but validation experiments performed on real, and often more complex, data are generally necessary for making post-hoc determinations of sequencing error rates. Indeed, methods that perform similarly with synthetic data may perform decidedly differently on real data[209]. Thus, costly empirical work is generally needed for better characterizing specific or more global sequencing error rates arising from the various sources of error.

Our view is that error estimates and related uncertainties should be incorporated into analyses as they arise, from uncertainties in initial measurements to uncertainties and errors from combinations and mathematical manipulations of data and inferences. This strategy allows for dynamic control over error rates and false findings, and it leverages the available data for responsive and real-time estimates of putative errors. We view uncertainty propagation and post-hoc error quantifications as complementary approaches, and many studies have found practical use in quantifying errors and accuracies of inferences in a post-hoc manner[209,225,73]. Here we have focused on describing uncertainties arising in detecting DNA sequences and in differences between sample and reference sequences, but other techniques uncover information on higher-order biological phenomenon. One such example is Hi-C[169], which is a method that allows researchers to better understand the three-dimensional organization of DNA in the cell[169,61], microbial community composition[15,31], and others[274]. Higher-order inferences should be made with algorithmic and statistical strategies that allow for the full appreciation of uncertainties in underlying measurements and determinations, as validating these inferences can sometimes be difficult due to the complexity of these systems and the rarity and precious nature of the underlying biological samples.

There is a need for rigorous uncertainty accounting across DNA detection and downstream sequencing-related analyses. As we progress through a scientific era in which nucleotide resolution studies are becoming the normal means of genetic dissection, science has found that even single nucleotide mutations can result in serious human disease[182,263]. It is therefore dangerous for an analysis to be wrong about any given variant call, particularly for biological samples that cannot be used for secondary studies or for orthogonal validations. Robust and full uncertainty accounting allows the analyst to better understand and predict when data and inferences are reliable and when they are not. This in turn informs data collection efforts, improves the reliability of published biological inferences based on error-prone DNA sequencing technologies and improves the quantitative rigor associated with all analyses based on DNA sequencing.

4.3.2 Confidence structure models

We have presented confidence structures as one way to improve uncertainty modeling in DNA sequencing data. Unlike current strategies, confidence structures provide a rich suit of methods that enable the simple use of and propagation of comprehensive characterizations of uncertainty through a number of complex downstream calculations. Indeed, the methods described here are general in that they can be used to estimate genetic disease risk, compute population parameters and facilitate the robust reporting of genetic results to patients.

Our simulation studies provide an initial demonstration of one simple application of confidence structure methods. Confidence structure based genotypes were computed for individual samples and across related samples for identifying de novo sequence variation. The results are promising in that the confidence structure methods seem to accurately characterize and propagates uncertainty about genotypes through simple calculations. Current methods do not provide obvious methods by which uncertainty can be propagated through downstream calculations and, as a consequence, these methods represent an important improvement over current approaches.

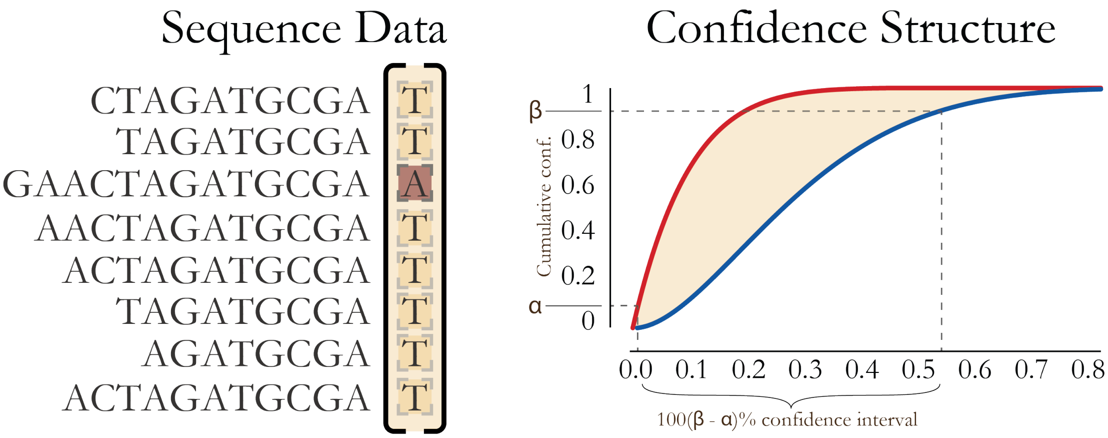


Figure 4.2: Here, we show how variability in sequence composition can be characterized using a statistical inference routine that can compute over epistemically uncertain input data, which in this case refers to interval probability values characterizing base calling errors. Sequence data (**A**) are used to estimate the frequency of an allele (**B**), in this case for adenine, at a particular locus using a multinomial model and confidence structures for statistical inference.

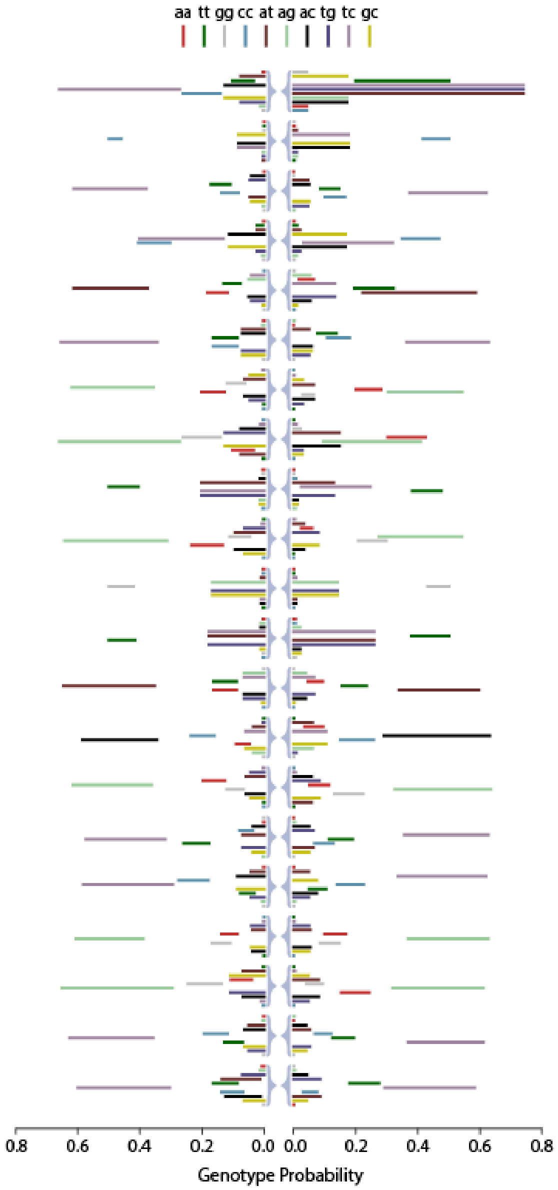


Figure 4.3: Confidence structure genotype inference for loci associated with 21 GATK false de novo calls in the 30X sequence data set. 50% confidence intervals are plotted for genotype probabilities for all possible genotype combinations for each the 21 loci for the 'father' (right) and 'child' (left). The vast majority of the loci contain confidence structure genotype calls that are easily distinguishable from random sequence error, although loci with large degrees of uncertainty are represented by correspondingly large intervals.

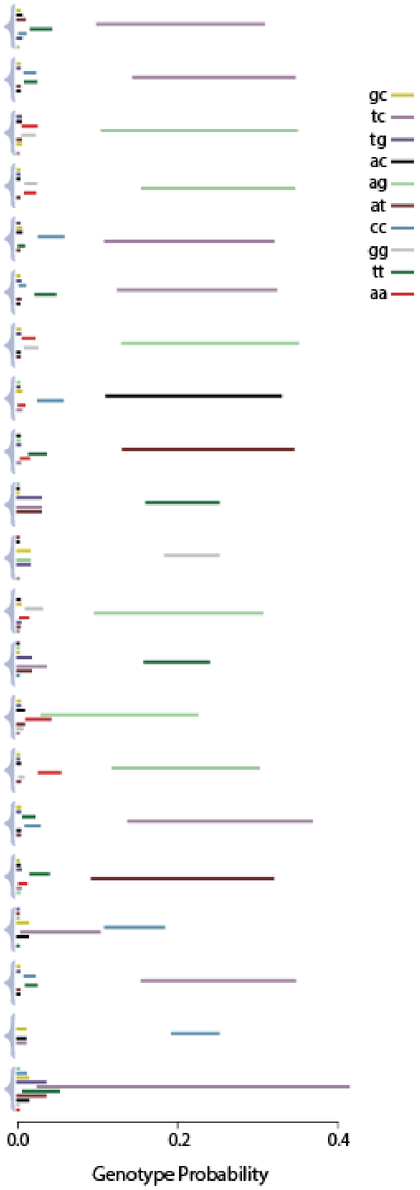


Figure 4.4: Joint genotype probabilities between the ’father’ and ’child’ using confidence structure genotype inference are plotted for 21 loci spanning false de novo calls made by GATK on 30X coverage sequence data. Uncertainty in the joint genotypes is encoded in the width of the 50% confidence intervals for each loci. Note that all but 2 loci result in clearly distinguishable genotype calls. The locus on the bottom has large uncertainty due to the fact that input genotypes from both the ’child’ and ’father’ samples contained a relatively large degree of uncertainty. As a consequence, the joint probability is also represented as being uncertain.

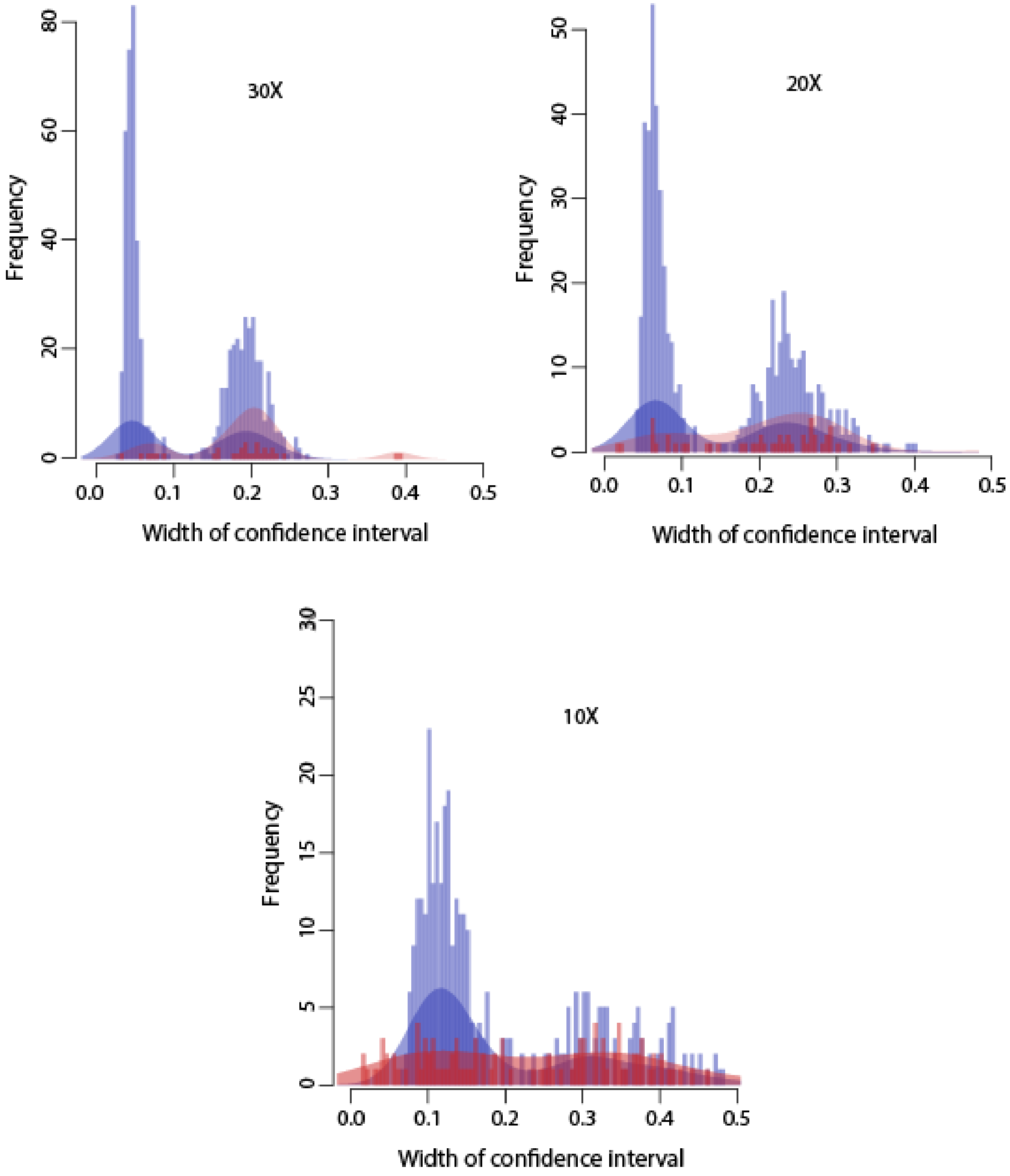


Figure 4.5: Histograms and density plots of the widths of confidence intervals for true negative de novo genotypes and false de novo genotypes with respect to GATK calls. For each coverage level, the widths of 50% confidence intervals for confidence structure estimates of true negative de novo genotypes are plotted in blue and false de novo genotypes are plotted in red. In general, the widths of intervals for the false de novo calls are larger, and as a consequence represent loci whose underlying sequence data are more uncertain than true negative calls ($p<0.05$ for all coverage levels, Kolmogorov–Smirnov test).

5

Conclusions

In this doctoral thesis, we pursued four distinct but intimately related studies aimed at assessing and improving the reliability of analyzing individual genomes, and assessing the feasibility of using individual genomes for genomic-guided medical care. The first study focused on assessing the reliability and accuracy of variant calling in exome and whole genome sequencing data. Predictive models were also designed for recovering missed sequence variation (false-negative detection). The second study focused on assessing the usefulness and practicality of sequencing a single human genome in the context of clinical care. The third study focused on using high-throughput sequencing technology in a robust way to facilitate the discovery of the genetic basis of a Mendelian disorder. The fourth and final study focused on developing methods for comprehensive uncertainty characterization in the context of detecting human sequence variation, so as to increase the reliability of these data.

We have found that there are still several challenges in using high-throughput DNA sequencing technologies for applied tasks. For example, SNV detection is thought to be a relatively simple analysis task, but applying the most popular analysis methods to the same data set results in sets of SNVs that are widely discordant. Agreement between different INDEL detection methods applied to the same data set is even worse. Furthermore, if different sequencing technologies are used to sequence the same sample, the SNVs and INDELs that are detected in the resulting sequence data are also discordant. Variant calling is of course ***far*** from a solved problem, and new methods are needed to generate robust results that contain full characterizations of the underlying uncertainty. Predictive methods in the context of results from many callers can offset the likelihood of missing important sequence variation, but the generality of models developed with validation data is thus far unclear, and more work is needed. In chapter 1, guidance on how to generate reliable analysis results given current technologies and cost considerations have been crafted, and important future directions for the field are discussed.

In chapter 2, we found that incorporating the use of an individual genome into the study and clinical treatment of a person is indeed feasible, given current technologies. However, storing genome-scale sequencing data in their full and most informative form is currently not possible, and the informativeness of a single genome is to some degree dependent on and limited by the condition being studied, as well as the time in which the assay was performed. Early intervention, we believe, will be where genome-scale sequencing will exert most of its value. In the study case, genomic information was useful for his ophthalmological care, but weakly so for his neuropsychiatric disorder. We found pharmacologically related sequence variation to be quite informative and useful for current and future care, and his DBS for his OCD demonstrated a positive influence over his OCD symptoms.

In chapter 3 we show evidence that mutations involving *TAF1* contribute to a new syndrome characterized by global developmental delay, intellectual disability, characteristic facial dysmorphologies, generalized hypotonia and other recognizable features. During the variant prioritization steps of one large pedigree containing affected individuals, we found that more variants are reliably eliminated when a greater portion of the family is incorporated into the analysis. Based partly on findings in chapter 1, we suspect that this observation is likely due to varying false positive and false negative rates across sequencing and informatics platforms. In using a large pedigree for variant prioritization in the context of applying multiple analysis pipelines to each sample, we were able to simultaneously eliminate likely false-positives while maximizing the detection of true positives. This study highlights the need for more comprehensive family-based study designs, and it demonstrates the benefits of focusing high-throughput sequencing efforts on studying large related cohorts, where intra-familial relationships allow for more rigorous variant filtering and identification of true positive alleles that might be contributing to a disease phenotype.

In the last chapter and final study, new statistical algorithms for characterizing high-throughout sequence data were developed, and simulation studies were undertaken. Methods for propagating holistic uncertainty quantifications of sequence variants are needed so that downstream analyses can take advantage of the known existing uncertainty. Confidence-structure based models were formulated to account for the most basic forms of sequence uncertainty, and these methods were used to detect de novo sequence variants in a simulated trio cohort. Comparisons with GATK, a popular analysis tool, show that the new methods are more sensitive and at the same time propagate uncertainty in a way that makes the reliability of the underlying data apparent. This preliminary work is promising, as the application of these methods represent an entirely new approach to sequence analysis which enables simple and obvious procedures for propagating uncertainty through all downstream calculations. The application of these methods can be of use in a variety of different fields and analyses. For example, these methods can be used to assess the genetic risk of disease, estimate population characteristics and in general enable reliable analysis for use in applied settings where clinical tasks require robust results that are honest about the underlying uncertainties.

References

[1] Abu-Amero, K. K., Hellani, A., Salih, M. A., Al Hussain, A., al Obailan, M., Zidan, G., Alorainy, I. A., & Bosley, T. M. (2010). Ophthalmologic abnormalities in a de novo terminal 6q deletion. *Ophthalmic Genet*, 31(1), 1–11.

[2] Alefeld, G. & Herzberger, J. (1984). *Introduction to interval computation*. Academic Press.

[3] Alivisatos, A. P., Andrews, A. M., Boyden, E. S., Chun, M., Church, G. M., Deisseroth, K., Donoghue, J. P., Fraser, S. E., Lippincott-Schwartz, J., Looger, L. L., Masmanidis, S., McEuen, P. L., Nurmikko, A. V., Park, H., Peterka, D. S., Reid, C., Roukes, M. L., Scherer, A., Schnitzer, M., Sejnowski, T. J., Shepard, K. L., Tsao, D., Turrigiano, G., Weiss, P. S., Xu, C., Yuste, R., & Zhuang, X. (2013a). Nanotools for neuroscience and brain activity mapping. *ACS Nano*, 7(3), 1850–1866.

[4] Alivisatos, A. P., Chun, M., Church, G. M., Deisseroth, K., Donoghue, J. P., Greenspan, R. J., McEuen, P. L., Roukes, M. L., Sejnowski, T. J., Weiss, P. S., & Yuste, R. (2013b). Neuroscience. The brain activity map. *Science*, 339(6125), 1284–1285.

[5] Allen, N. C., Bagade, S., McQueen, M. B., Ioannidis, J. P. A., Kavvoura, F. K., Khoury, M. J., Tanzi, R. E., & Bertram, L. (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature Genetics*, 40(7), 827–834.

[6] Alonso, P., Lopez-Sola, C., Gratacos, M., Fullana, M. A., Segalas, C., Real, E., Cardoner, N., Soriano-Mas, C., Harrison, B. J., Estivill, X., & Menchon, J. M. (2013). The interaction between Comt and Bdnf variants influences obsessive-compulsive-related dysfunctional beliefs. *Journal of Anxiety Disorders*, 27(3), 321–327.

[7] Alonso, P., Segalas, C., Real, E., Pertusa, A., Labad, J., Jimenez-Murcia, S., Jaurrieta, N., Bueno, B., Vallejo, J., & Menchon, J. M. (2010). Suicide in patients treated for obsessive-compulsive disorder: a prospective follow-up study. *Journal of Affective Disorders*, 124(3), 300–308.

[8] Altman, R. B., Prabhu, S., Sidow, A., Zook, J. M., Goldfeder, R., Litwack, D., Ashley, E., Asimenos, G., Bustamante, C. D., Donigan, K., Giacomini, K. M., Johansen, E., Khuri, N., Lee, E., Liang, X. S., Salit, M., Serang, O., Tezak, Z., Wall, D. P., Mansfield, E., & Kass-Hout, T. (2016). A research roadmap for next-generation sequencing informatics. *Science Translational Medicine*, 8(335), 10–335.

[9] Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., & Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protocols*, 8(9), 1765–1786.

[10] Bainbridge, M. N., Wiszniewski, W., Murdock, D. R., Friedman, J., Gonzaga-Jauregui, C., Newsham, I., Reid, J. G., Fink, J. K., Morgan, M. B., Gingras, M. C., Muzny, D. M., Hoang, L. D., Yousaf, S., Lupski, J. R., & Gibbs, R. A. (2011). Whole-genome sequencing for optimized patient management. *Science Translational Medicine*, 3(87), 87re3.

[11] Balch, M. S. (2012). Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning*, 53(7), 1003–1019.

[12] Balci, V. & Sevincok, L. (2010). Suicidal ideation in patients with obsessive-compulsive disorder. *Psychiatry Research*, 175(1-2), 104–108.

[13] Ball, M. P., Thakuria, J. V., Zaranek, A. W., Clegg, T., Rosenbaum, A. M., Wu, X., Angrist, M., Bhak, J., Bobe, J., Callow, M. J., Cano, C., Chou, M. F., Chung, W. K., Douglas, S. M., Estep, P. W., Gore, A., Hulick, P., Labarga, A., Lee, J. H., Lunshof, J. E., Kim, B. C., Kim, J. I., Li, Z., Murray, M. F., Nilsen, G. B., Peters, B. A., Raman, A. M., Rienhoff, H. Y., Robasky, K., Wheeler, M. T., Vandewege, W., Vorhaus, D. B., Yang, J. L., Yang, L., Aach, J., Ashley, E. A., Drmanac, R., Kim, S. J., Li, J. B., Peshkin, L., Seidman, C. E., Seo, J. S., Zhang, K., Rehm, H. L., & Church, G. M. (2012). A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(30), 11920–11927.

[14] Bearn, A. G. (1993). *Archibald Garrod and the individuality of Man*. Oxford, New York: Clarendon Press; Oxford University Press.

[15] Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., & Darling, A. E. (2014). Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*, 2, e415.

[16] Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer.

[17] Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., & Betrò, B. (1994). An overview of robust Bayesian analysis. *Test*, 3(1), 5–124.

[18] Berry, R. J., Buehler, J. W., Strauss, L. T., Hogue, C. J., & Smith, J. C. (1987). Birth weight-specific infant mortality due to congenital anomalies, 1960 and 1980. *Public Health Reports*, 102(2), 171–181.

[19] Bhattacharya, S., Lou, X., Hwang, P., Rajashankar, K. R., Wang, X., Gustafsson, J. A., Fletterick, R. J., Jacobson, R. H., & Webb, P. (2014). Structural and functional insight into TAF1-TAF7, a subcomplex of transcription factor II D. *Proceedings of the National Academy of Sciences*, 111(25), 9103–9108.

[20] Bieniossek, C., Papai, G., Schaffitzel, C., Garzoni, F., Chaillet, M., Scheer, E., Papadopoulos, P., Tora, L., Schultz, P., & Berger, I. (2013). The architecture of human general transcription factor TFIID core complex. *Nature*, 493(7434), 699–702.

[21] Biesecker, B. B. & Peay, H. L. (2013). Genomic sequencing for psychiatric disorders: promise and challenge. *The International Journal of Neuropsychopharmacology*, (pp. 1–6).

[22] Blomstedt, P., Sjoberg, R. L., Hansson, M., Bodlund, O., & Hariz, M. I. (2012). Deep Brain Stimulation in the Treatment of Obsessive-Compulsive Disorder. *World Neurosurgery*.

[23] Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L., & Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436), 234–237.

[24] Borck, G., Hög, F., Dentici, M. L., Tan, P. L., Sowada, N., Medeira, A., Gueneau, L., Thiele, H., Kousi, M., Lepri, F., Wenzeck, L., Blumenthal, I., Radicioni, A., Schwarzenberg, T. L., Mandriani, B., Fischetto, R., Morris-Rosendahl, D. J., Altmüller, J., Reymond, A., Nürnberg, P., Merla, G., Dallapiccola, B., Katsanis, N., Cramer, P., & Kubisch, C. (2015). BRF1 mutations alter RNA polymerase III–dependent transcription and cause neurodevelopmental anomalies. *Genome Research*, 25(2), 155–166.

[25] Bozok Cetintas, V., Erer, O. F., Kosova, B., Ozdemir, I., Topcuoglu, N., Aktogu, S., & Eroglu, Z. (2008). Determining the relation between N-acetyltransferase-2 acetylator phenotype and antituberculosis drug induced hepatitis by molecular biologic tests. *Tuberkuloz ve toraks*, 56(1), 81–86.

[26] Brown, S. H., Lincoln, M. J., Groen, P. J., & Kolodner, R. M. (2003). VistA–U.S. Department of Veterans Affairs national-scale HIS. *International Journal of Medical Informatics*, 69(2-3), 135–156.

[27] Browning, B. L. & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, 88(2), 173–182.

[28] Browning, S. R. & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5), 1084–1097.

[29] Burdick, A., Goodman, W. K., & Foote, K. D. (2009). Deep brain stimulation for refractory obsessive-compulsive disorder. *Frontiers in Bioscience*, 14, 1880–1890.

[30] Burdick, A. P. & Foote, K. D. (2011). Advancing deep brain stimulation for obsessive-compulsive disorder. *Expert Review of Neurotherapeutics*, 11(3), 341–344.

[31] Burton, J. N., Liachko, I., Dunham, M. J., & Shendure, J. (2014). Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3: Genes|Genomes|Genetics*, 4(7), 1339–1346.

[32] Campbell, I. M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M. E., Nagamani, S. C., Erez, A., Bartnik, M., Wisniowiecka-Kowalnik, B., Plunkett, K. S., Pursley, A. N., Kang, S. H., Bi, W., Lalani, S. R., Bacino, C. A., Vast, M., Marks, K., Patton, M., Olofsson, P., Patel, A., Veltman, J. A., Cheung, S. W., Shaw, C. A., Vissers, L. E., Vermeesch, J. R., Lupski, J. R., & Stankiewicz, P. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *American Journal of Human Genetics*, 95(2), 173–182.

[33] Carayol, J., Schellenberg, G. D., Dombroski, B., Amiet, C., Genin, B., Fontaine, K., Rousseau, F., Vazart, C., Cohen, D., Frazier, T. W., Hardan, A. Y., Dawson, G., & Rio Frio, T. (2014). A scoring strategy combining statistics and functional genomics supports a possible role for common polygenic variation in autism. *Frontiers in Genetics*, 5, 33.

[34] Carnevali, P., Baccash, J., Halpern, A. L., Nazarenko, I., Nilsen, G. B., Pant, K. P., Ebert, J. C., Brownley, A., Morenzoni, M., Karpinchyk, V., Martin, B., Ballinger, D. G., & Drmanac, R. (2012). Computational techniques for human Genome Researchequencing using mated gapped reads. *Journal of Computational Biology*, 19(3), 279–292.

[35] Cech, T. R. & Steitz, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 157(1), 77–94.

[36] Chahrour, M. H., Yu, T. W., Lim, E. T., Ataman, B., Coulter, M. E., Hill, R. S., Stevens, C. R., Schubert, C. R., Greenberg, M. E., Gabriel, S. B., & Walsh, C. A. (2012). Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genetics*, 8(4), e1002635.

[37] Chaki, M., Airik, R., Ghosh, A. K., Giles, R. H., Chen, R., Slaats, G. G., Wang, H., Hurd, T. W., Zhou, W., Cluckey, A., Gee, H. Y., Ramaswami, G., Hong, C. J., Hamilton, B. A., Cervenka, I., Ganji, R. S., Bryja, V., Arts, H. H., van Reeuwijk, J., Oud,

M. M., Letteboer, S. J., Roepman, R., Husson, H., Ibraghimov-Beskrovnaya, O., Yasunaga, T., Walz, G., Eley, L., Sayer, J. A., Schermer, B., Liebau, M. C., Benzing, T., Le Corre, S., Drummond, I., Janssen, S., Allen, S. J., Natarajan, S., O'Toole, J. F., Attanasio, M., Saunier, S., Antignac, C., Koenekoop, R. K., Ren, H., Lopez, I., Nayir, A., Stoetzel, C., Dollfus, H., Massoudi, R., Gleeson, J. G., Andreoli, S. P., Doherty, D. G., Lindstrad, A., Golzio, C., Katsanis, N., Pape, L., Abboud, E. B., Al-Rajhi, A. A., Lewis, R. A., Omran, H., Lee, E. Y., Wang, S., Sekiguchi, J. M., Saunders, R., Johnson, C. A., Garner, E., Vanselow, K., Andersen, J. S., Shlomai, J., Nurnberg, G., Nurnberg, P., Levy, S., Smogorzewska, A., Otto, E. A., & Hildebrandt, F. (2012). Exome capture reveals ZNF423 and CEP164 mutations, linking renal ciliopathies to DNA damage response signaling. *Cell*, 150(3), 533–548.

[38] Chalkley, G. E. & Verrijzer, C. (1999). *DNA binding site selection by RNA polymerase II TAFs: a TAFII250–TAFII150 complex recognizes the Initiator*, volume 18.

[39] Chiang, C. M. & Roeder, R. G. (1995). Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. *Science*, 267(5197), 531–536.

[40] Ciancarelli, I., Tozzi Ciancarelli, M. G., & Carolei, A. (2013). Effectiveness of intensive neurorehabilitation in patients with Huntington's disease. *European Journal of Physical and Rehabilitation Medicine*, 49(2), 189–195.

[41] Clark, M. J. (2012). Cliff Reid on CG vs Illumina .

[42] Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Euskirchen, G., Butte, A. J., & Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, 29(10), 908–914.

[43] Clement, N. L., Snell, Q., Clement, M. J., Hollenhorst, P. C., Purwar, J., Graves, B. J., Cairns, B. R., & Johnson, W. E. (2009). The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*, 26(1), 38–45.

[44] Collip, D., van Winkel, R., Peerbooms, O., Lataster, T., Thewissen, V., Lardinois, M., Drukker, M., Rutten, B. P., Van Os, J., & Myin-Germeys, I. (2011). COMT Val158Met-stress interaction in psychosis: role of background psychosis risk. *CNS Neuroscience and Therapeutics*, 17(6), 612–619.

[45] Conn, J. (2011). VA to update VistA EHR. *Modern Healthcare*, 41(22), 17.

[46] Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., & Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, 43(7), 712–714.

[47] Consortium, T. . G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56–65.

[48] Consortium, T. G. O. (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1), D1049–D1056.

[49] Cooper, G. M. & Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9), 628–640.

[50] Correll, C. U., Manu, P., Olshanskiy, V., Napolitano, B., Kane, J. M., & Malhotra, A. K. (2009). Cardiometabolic risk of second-generation antipsychotic medications during first-time use in children and adolescents. *Journal of the American Medical Association*, 302(16), 1765–1773.

[51] Cucala, L. (2008). A Hypothesis-Free Multiple Scan Statistic with Variable Window. *Biometrical Journal*, 50(2), 299–310.

[52] Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience*, 15, 353–375.

[53] De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Fu, S.-C., Aleksic, B., Biscaldi, M., Bolton, P. F., Brownfeld, J. M., Cai, J., Campbell, N. G., Carracedo, A., Chahrour, M. H., Chiocchetti, A. G., Coon, H., Crawford, E. L., Crooks, L., Curran, S. R., Dawson, G., Duketis, E., Fernandez, B. A., Gallagher, L., Geller, E., Guter, S. J., Sean Hill, R., Ionita-Laza, I., Jimenez Gonzalez, P., Kilpinen, H., Klauck, S. M., Kolevzon, A., Lee, I., Lei, J., Lehtimaki, T., Lin, C.-F., Ma/'ayan, A., Marshall, C. R., McInnes, A. L., Neale, B., Owen, M. J., Ozaki, N., Parellada, M., Parr, J. R., Purcell, S., Puura, K., Rajagopalan, D., Rehnstrom, K., Reichenberg, A., Sabo, A., Sachse, M., Sanders, S. J., Schafer, C., Schulte-Ruther, M., Skuse, D., Stevens, C., Szatmari, P., Tammimies, K., Valladares, O., Voran, A., Wang, L.-S., Weiss, L. A., Jeremy Willsey, A., Yu, T. W., Yuen, R. K. C., The, D. D. D. S., Homozygosity Mapping Collaborative for, A., Consortium, U. K., The Autism Sequencing, C., Cook, E. H., Freitag, C. M., Gill, M., Hultman, C. M., Lehner, T., Palotie, A., Schellenberg, G. D., Sklar, P., State, M. W., Sutcliffe, J. S., Walsh, C. A., Scherer, S. W., Zwick, M. E., Barrett, J. C., Cutler, D. J., Roeder, K., Devlin, B., Daly, M. J., & Buxbaum, J. D. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526), 209–215.

[54] Deardorff Matthew, A., Krantz Ian, D., & Press, O. U. (2008). NIPBL and SMC1L1 (now SCM1A) and the Cornelia de Lange Syndrome. In J. Epstein Charles, P. Erickson Robert, & A. Wynshao-Boris (Eds.), *Inborn Errors of Development*. Oxford University Press.

[55] Denys, D. & Mantione, M. (2009). Deep brain stimulation in obsessive-compulsive disorder. *Progress in Brain Research*, 175, 419–427.

[56] Denys, D., Mantione, M., Figee, M., van den Munckhof, P., Koerselman, F., Westenberg, H., Bosch, A., & Schuurman, R. (2010). Deep brain stimulation of the nucleus accumbens for treatment-refractory obsessive-compulsive disorder. *Archives of General Psychiatry*, 67(10), 1061–1068.

[57] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–498.

[58] Dewey, F. E., Chen, R., Cordero, S. P., Ormond, K. E., Caleshu, C., Karczewski, K. J., Whirl-Carrillo, M., Wheeler, M. T., Dudley, J. T., Byrnes, J. K., Cornejo, O. E., Knowles, J. W., Woon, M., Sangkuhl, K., Gong, L., Thorn, C. F., Hebert, J. M., Capriotti, E., David, S. P., Pavlovic, A., West, A., Thakuria, J. V., Ball, M. P., Zaranek, A. W., Rehm, H. L., Church, G. M., West, J. S., Bustamante, C. D., Snyder, M., Altman, R. B., Klein, T. E., Butte, A. J., & Ashley, E. A. (2011). Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genetics*, 7(9), e1002280.

[59] Dikstein, R., Ruppert, S., & Tjian, R. (1996). TAFII250 Is a Bipartite Protein Kinase That Phosphorylates the Basal Transcription Factor RAP74. *Cell*, 84(5), 781–790.

[60] Ding, L., Wendl, M. C., McMichael, J. F., & Raphael, B. J. (2014). Expanding the computational toolbox for mining cancer genomes. *Nature Reviews Genetics*, 15(8), 556–570.

[61] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376–380.

[62] Domingo, A., Westenberger, A., Lee, L. V., Braenne, I., Liu, T., Vater, I., Rosales, R., Jamora, R. D., Pasco, P. M., Cutiongco-Dela Paz, E. M., Freimann, K., Schmidt, T. G., Dressler, D., Kaiser, F. J., Bertram, L., Erdmann, J., Lohmann, K., & Klein, C. (2015). New insights into the genetics of X-linked dystonia-parkinsonism (XDP, DYT3). *European Journal of Human Genetics*.

[63] Drmanac, R. (2011). The advent of personal genome sequencing. *Genetics in Medicine*, 13(3), 188–190.

[64] Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M., & Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327(5961), 78–81.

[65] Dumontheil, I., Roggeman, C., Ziermans, T., Peyrard-Janvid, M., Matsson, H., Kere, J., & Klingberg, T. (2011). Influence of the COMT genotype on working memory and brain activity changes during development. *Biological Psychiatry*, 70(3), 222–229.

[66] Dwyer, P. S. (1951). Linear computations.

[67] Dy, M. E., Talkowski, M. E., Multhaupt-Buell, T. J., Paul, L. R., Bragg, D. C., & Sharma, N. (2015). Genotype□phenotype correlation in X□linked dystonia□Parkinsonism (XDP/DYT3) .

[68] Eichenbaum, H. (2013). What H.M. taught us. *Journal of Cognitive Neuroscience*, 25(1), 14–21.

[69] Eilbeck, K. (2013). GVFclin.

[70] Erickson-Davis, C. (2012). Ethical concerns regarding commercialization of deep brain stimulation for obsessive compulsive disorder. *Bioethics*, 26(8), 440–446.

[71] Ewing, B. & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, 8(3), 186–194.

[72] Ezkurdia, I., Vazquez, J., Valencia, A., & Tress, M. (2014). Analyzing the First Drafts of the Human Proteome. *Journal of Proteome Research*.

[73] Fang, H., Wu, Y., Narzisi, G., O'Rawe, J. A., Jimenez Barrón, L. T., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M. C., & Lyon, G. J. (2014). *Reducing INDEL calling errors in whole-genome and exome sequencing data*.

[74] Ferson, S. (1995). Quality assurance for Monte Carlo risk assessment. In *Uncertainty Modeling and Analysis, 1995, and Annual Conference of the North American Fuzzy Information Processing Society. Proceedings of ISUMA-NAFIPS'95., Third International Symposium on* (pp. 14–19).: IEEE.

[75] Ferson, S., Balch, M., Sentz, K., & Siegrist, J. Computing with Confidence.

[76] Ferson, S. & Hajagos, J. G. (2004). Arithmetic with uncertain numbers: rigorous and (often) best possible answers. *Reliability Engineering & System Safety*, 85(1), 135–152.

[77] Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D. S., & Sentz, K. (2002). *Constructing probability boxes and Dempster-Shafer structures*, volume 835. Sandia National Laboratories.

[78] Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., & Ginzburg, L. (2007). *Experimental uncertainty estimation and statistics for data having interval uncertainty*. Sandia National Laboratories.

[79] Ferson, S., O'Rawe, J., & Balch, M. Computing with Confidence: Imprecise Posteriors and Predictive Distributions. In *Vulnerability, Uncertainty, and Risk@ sQuantification, Mitigation, and Management* (pp. 895–904).: ASCE.

[80] Figee, M., Luigjes, J., Smolders, R., Valencia-Alfonso, C. E., van Wingen, G., de Kwaasteniet, B., Mantione, M., Ooms, P., de Koning, P., Vulink, N., Levar, N., Droge, L., van den Munckhof, P., Schuurman, P. R., Nederveen, A., van den Brink, W., Mazaheri, A., Vink, M., & Denys, D. (2013). Deep brain stimulation restores frontostriatal network activity in obsessive-compulsive disorder. *Nature Neuroscience*, 16(4), 386–387.

[81] Filippakopoulos, P., Picaud, S., Mangos, M., Keates, T., Lambert, J.-P., Barsyte-Lovejoy, D., Felletar, I., Volkmer, R., Müller, S., Pawson, T., Gingras, A.-C., Arrowsmith, C. H., & Knapp, S. (2012). Histone Recognition and Large-Scale Structural Analysis of the Human Bromodomain Family. *Cell*, 149(1), 214–231.

[82] Fins, J. J. (2010). Deep Brain Stimulation, Free Markets and the Scientific Commons: Is It time to Revisit the Bayh-Dole Act of 1980? *Neuromodulation*, 13(3), 153–159.

[83] Fins, J. J., Dorfman, G. S., & Pancrazio, J. J. (2012). Challenges to deep brain stimulation: a pragmatic response to ethical, fiscal, and regulatory concerns. *Annals of the New York Academy of Sciences*, 1265, 80–90.

[84] Fins, J. J., Mayberg, H. S., Nuttin, B., Kubu, C. S., Galert, T., Sturm, V., Stoppenbrink, K., Merkel, R., & Schlaepfer, T. E. (2011a). Misuse of the FDA's humanitarian device exemption in deep brain stimulation for obsessive-compulsive disorder. *Health Affairs*, 30(2), 302–311.

[85] Fins, J. J. & Schiff, N. D. (2010). Conflicts of interest in deep brain stimulation research and the ethics of transparency. *The Journal of Clinical Ethics*, 21(2), 125–132.

[86] Fins, J. J., Schlaepfer, T. E., Nuttin, B., Kubu, C. S., Galert, T., Sturm, V., Merkel, R., & Mayberg, H. S. (2011b). Ethical guidance for the management of conflicts of interest for researchers, engineers and clinicians engaged in the development of therapeutic deep Brain Stimulation. *Journal of Neural Engineering*, 8(3), 33001.

[87] Fitzpatrick, D. (1994). Clinical Genetics Handbook, 2nd edn. By Arthur Robinson and Mary G. Lindon. Blackwell Scientific Publications. 1993. 614 pages. Paperback. Price £32.50. ISBN 08 654 219 43. *Genetics Research*, 64(03), 219.

[88] Frank, M. J., Nelsen, R. B., & Schweizer, B. (1987). Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probability Theory and Related Fields*, 74(2), 199–211.

[89] Frayling, T. M. (2008). Commentary: Genetic association studies see light at the end of the tunnel. *International Journal of Epidemiology*, 37(1), 133–135.

[90] Fréchet, M. (1935). *Généralisations du théorème des probabilités totales*, volume 25. Fundamenta Mathematica.

[91] Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., & Nielsen, R. (2013). Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. *Genetics*, 195(3), 979–992.

[92] Gabellini, D., Green, M. R., & Tupler, R. (2002). Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell*, 110(3), 339–348.

[93] Gardner-Medwin, D. (1991). Principles and practice of medical genetics, 2nd edn.: Edited by Alan E. H. Emery and David L. Rimoin. Published 1990 by Churchill Livingstone, Edinburgh. ISBN 0 443 03583 0, 2 volumes, 2035 pp. Price £195. *Neuromuscular Disorders*, 1(5), 383–384.

[94] Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

[95] Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., Ripke, S., Sandin, S., Sklar, P., Svantesson, O., Reichenberg, A., Hultman, C. M., Devlin, B., Roeder, K., & Buxbaum, J. D. (2014). Most genetic risk for autism resides with common variation. *Nature Genetics*, 46(8), 881–885.

[96] Gegonne, A., Weissman, J. D., & Singer, D. S. (2001). TAFII55 binding to TAFII250 inhibits its acetyltransferase activity. *Proceedings of the National Academy of Sciences*, 98(22), 12432–12437.

[97] Gigerenzer, G. (2002). *Calculated risks : how to know when numbers deceive you*. New York: Simon & Schuster.

[98] Gil-Rodriguez, M. C., Deardorff, M. A., Ansari, M., Tan, C. A., Parenti, I., Baquero-Montoya, C., Ousager, L. B., Puisac, B., Hernandez-Marcos, M., Teresa-Rodrigo, M. E., Marcos-Alcalde, I., Wesselink, J. J., Lusa-Bernal, S., Bijlsma, E. K., Braunholz, D., Bueno-Martinez, I., Clark, D., Cooper, N. S., Curry, C. J., Fisher, R., Fryer, A., Ganesh, J., Gervasini, C., Gillessen-Kaesbach, G., Guo, Y., Hakonarson, H., Hopkin, R. J., Kaur, M., Keating, B. J., Kibaek, M., Kinning, E., Kleefstra, T., Kline, A. D., Kuchinskaya, E., Larizza, L., Li, Y. R., Liu, X., Mariani, M., Picker, J. D., Pie, A., Pozojevic, J., Queralt, E., Richer, J., Roeder, E., Sinha, A., Scott, R. H., So, J., Wusik, K. A., Wilson, L., Zhang, J., Gomez-Puertas, P., Casale, C. H., Strom, L., Selicorni, A., Ramos, F. J., Jackson, L. G., Krantz, I. D., Das, S., Hennekam, R. C., Kaiser, F. J., FitzPatrick, D. R., & Pie, J. (2015). De novo heterozygous mutations in SMC3 cause a range of Cornelia de Lange syndrome-overlapping phenotypes. *Human Mutation*, 36(4), 454–462.

[99] Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., de Vries, B. B., Kleefstra, T., Brunner, H. G., Vissers, L. E., & Veltman, J. A. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509), 344–347.

[100] Gillihan, S. J., Williams, M. T., Malcoun, E., Yadin, E., & Foa, E. B. (2012). Common Pitfalls in Exposure and Response Prevention (EX/RP) for OCD. *Journal of Obsessive-Compulsive and Related Disorders*, 1(4), 251–257.

[101] Goodman, W. K. & Alterman, R. L. (2012). Deep brain stimulation for intractable psychiatric disorders. *Annual Review of Medicine*, 63, 511–524.

[102] Goodman, W. K., Foote, K. D., Greenberg, B. D., Ricciuti, N., Bauer, R., Ward, H., Shapira, N. A., Wu, S. S., Hill, C. L., Rasmussen, S. A., & Okun, M. S. (2010). Deep brain stimulation for intractable obsessive compulsive disorder: pilot study using a blinded, staggered-onset design. *Biological Psychiatry*, 67(6), 535–542.

[103] Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Delgado, P., Heninger, G. R., & Charney, D. S. (1989a). The Yale-Brown Obsessive Compulsive Scale. II. Validity. *Archives of General Psychiatry*, 46(11), 1012–1016.

[104] Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., Heninger, G. R., & Charney, D. S. (1989b). The Yale-Brown Obsessive Compulsive Scale. I. Development, use, and reliability. *Archives of General Psychiatry*, 46(11), 1006–1011.

[105] Goya, R., Sun, M. G., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., Senz, J., Crisan, A., Marra, M. A., Hirst, M., Huntsman, D., Murphy, K. P., Aparicio, S., & Shah, S. P. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6), 730–736.

[106] Gupta, R. K., Arany, Z., Seale, P., Mepani, R. J., Ye, L., Conroe, H. M., Roby, Y. A., Kulaga, H., Reed, R. R., & Spiegelman, B. M. (2010). Transcriptional control of preadipocyte determination by Zfp423. *Nature*, 464(7288), 619–623.

[107] Hall, D., Dhilla, A., Charalambous, A., Gogos, J. A., & Karayiorgou, M. (2003). Sequence Variants of the Brain-Derived Neurotrophic Factor (BDNF) Gene Are Strongly Associated with Obsessive-Compulsive Disorder. *American Journal of Human Genetics*, 73(2), 370–376.

[108] Hawkins, R. D., Hon, G. C., & Ren, B. (2010). Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7), 476–486.

[109] Haynes, W. I. & Mallet, L. (2010). High-frequency stimulation of deep brain structures in obsessive-compulsive disorder: the search for a valid circuit. *The European Journal of Neuroscience*, 32(7), 1118–1127.

[110] He, M., Person, T. N., Hebbring, S. J., Heinzen, E., Ye, Z., Schrodi, S. J., McPherson, E. W., Lin, S. M., Peissig, P. L., Brilliant, M. H., O'Rawe, J., Robison, R. J., Lyon, G. J., & Wang, K. (2015). SeqHBase: a big data toolset for family based sequencing data analysis. *Journal of Medical Genetics*.

[111] Heinrich, V., Stange, J., Dickhaus, T., Imkeller, P., Kruger, U., Bauer, S., Mundlos, S., Robinson, P. N., Hecht, J., & Krawitz, P. M. (2012). The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Research*, 40(6), 2426–2431.

[112] Hellman-Aharony, S., Smirin-Yosef, P., Halevy, A., Pasmanik-Chor, M., Yeheskel, A., Har-Zahav, A., Maya, I., Straussberg, R., Dahary, D., Haviv, A., Shohat, M., & Basel-Vanagaite, L. (2013). Microcephaly thin corpus callosum intellectual disability syndrome caused by mutated TAF2. *Pediatric Neurology*, 49(6), 411–416.

[113] Henry, R. W., Mittal, V., Ma, B., Kobayashi, R., & Hernandez, N. (1998). SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by RNA polymerases II and III. *Genes & Development*, 12(17), 2664–2672.

[114] Herzfeld, T., Nolte, D., Grznarova, M., Hofmann, A., Schultze, J. L., & Muller, U. (2013). X-linked dystonia parkinsonism syndrome (XDP, lubag): disease-specific sequence change DSC3 in TAF1/DYT3 affects genes in vesicular transport and dopamine metabolism. *Human Molecular Genetics*, 22(5), 941–951.

[115] Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., & Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research*, 41(1), e32.

[116] Hoekelman, R. A. & Pless, I. B. (1988). Decline in mortality among young Americans during the 20th century: prospects for reaching national mortality reduction goals for 1990. *Pediatrics*, 82(4), 582–595.

[117] Hoflich, A., Savli, M., Comasco, E., Moser, U., Novak, K., Kasper, S., & Lanzenberger, R. (2013). Neuropsychiatric deep brain stimulation for translational neuroimaging. *NeuroImage*, 79(0), 30–41.

[118] Hogben, L. T. (1933). *Nature and nurture*. New York,: W.W. Norton Company.

[119] Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through prephasing. *Nature Genetics*, 44(8), 955–959.

[120] Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3: Genes | Genomes | Genetics*, 1(6), 457–470.

[121] Hu, H., Haas, S. A., Chelly, J., Van Esch, H., Raynaud, M., de Brouwer, A. P., Weinert, S., Froyen, G., Frints, S. G., Laumonnier, F., Zemojtel, T., Love, M. I., Richard, H., Emde, A. K., Bienek, M., Jensen, C., Hambrock, M., Fischer, U., Langnick, C., Feldkamp, M., Wissink-Lindhout, W., Lebrun, N., Castelnau, L., Rucci, J., Montjean, R., Dorseuil, O., Billuart, P., Stuhlmann, T., Shaw, M., Corbett, M. A., Gardner, A., Willis-Owen, S., Tan, C., Friend, K. L., Belet, S., van Roozendaal, K. E., Jimenez-Pocquet, M., Moizard, M. P., Ronce, N., Sun, R., O'Keeffe, S., Chenna, R., van Bommel, A., Goke, J., Hackett, A., Field, M., Christie, L., Boyle, J., Haan, E., Nelson, J., Turner, G., Baynam, G., Gillessen-Kaesbach, G., Muller, U., Steinberger, D., Budny, B., Badura-Stronka, M., Latos-Bielenska, A., Ousager, L. B., Wieacker, P., Rodriguez Criado, G., Bondeson, M. L., Anneren, G., Dufke, A., Cohen, M., Van Maldergem, L., Vincent-Delorme, C., Echenne, B., Simon-Bouy, B., Kleefstra, T., Willemsen, M., Fryns, J. P., Devriendt, K., Ullmann, R., Vingron, M., Wrogemann, K., Wienker, T. F., Tzschach, A., van Bokhoven, H., Gecz, J., Jentsch, T. J., Chen, W., Ropers, H. H., & Kalscheuer, V. M. (2015). X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Molecular Psychiatry*.

[122] Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., & Yandell, M. (2013). VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*, 37(6), 622–634.

[123] Huang, W., Richards, S., Carbone, M. A., Zhu, D., Anholt, R. R., Ayroles, J. F., Duncan, L., Jordan, K. W., Lawrence, F., Magwire, M. M., Warner, C. B., Blankenburg, K., Han, Y., Javaid, M., Jayaseelan, J., Jhangiani, S. N., Muzny, D., Ongeri, F., Perales, L., Wu, Y. Q., Zhang, Y., Zou, X., Stone, E. A., Gibbs, R. A., & Mackay, T. F. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39), 15553–15559.

[124] Insua, D. R. & Ruggeri, F. (2000). *Robust Bayesian Analysis*. Springer.

[125] Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., & Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*, 383(9912), 166–175.

[126] Iossifov, I., O/'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., Sullivan, C. A., Walker, M. F., Waqar, Z., Wei, L., Willsey, A. J., Yamrom, B., Lee, Y.-h., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M. C., Ye, K., McCombie, W. R., Shendure, J., Eichler, E. E., State, M. W., & Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526), 216–221.

[127] Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-h., Narzisi, G., Leotta, A., Kendall, J., Grabowska, E., Ma, B., Marks, S., Rodgers, L., Stepansky, A., Troge, J., Andrews, P., Bekritsky, M., Pradhan, K., Ghiban, E., Kramer, M., Parla, J., Demeter, R., Fulton, L. L., Fulton, R. S., Magrini, V. J., Ye, K., Darnell, J. C., Darnell, R. B., Mardis, E. R., Wilson, R. K., Schatz, M. C., McCombie, W. R., & Wigler, M. (2012). De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron*, 74(2), 285–299.

[128] Ira, E., Zanoni, M., Ruggeri, M., Dazzan, P., & Tosato, S. (2013). COMT, neuropsychological function and brain structure in schizophrenia: a systematic review and neurobiological interpretation. *Journal of Psychiatry and Neuroscience*, 38(3), 120178.

[129] Izumi, K., Nakato, R., Zhang, Z., Edmondson, A. C., Noon, S., Dulik, M. C., Rajagopalan, R., Venditti, C. P., Gripp, K., Samanich, J., Zackai, E. H., Deardorff, M. A., Clark, D., Allen, J. L., Dorsett, D., Misulovin, Z., Komata, M., Bando, M., Kaur, M., Katou, Y., Shirahige, K., & Krantz, I. D. (2015). Germline gain-of-function mutations in AFF4 cause a developmental syndrome functionally linking the super elongation complex and cohesin. *Nature Genetics*, 47(4), 338–344.

[130] Jacobson, R. H., Ladurner, A. G., King, D. S., & Tjian, R. (2000). Structure and Function of a Human TAFII250 Double Bromodomain Module. *Science*, 288(5470), 1422–1425.

[131] Jambaldorj, J., Makino, S., Munkhbat, B., & Tamiya, G. (2012). Sustained expression of a neuron-specific isoform of the Taf1 gene in development stages and aging in mice. *Biochemical Biophysics Ressearch Communications*, 425(2), 273–277.

[132] Jao, L.-E., Wente, S. R., & Chen, W. (2013). Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proceedings of the National Academy of Sciences*, 110(34), 13904–13909.

[133] Jimenez-Ponce, F., Velasco-Campos, F., Castro-Farfan, G., Nicolini, H., Velasco, A. L., Salin-Pascual, R., Trejo, D., & Criales, J. L. (2009). Preliminary study in patients with obsessive-compulsive disorder treated with electrical stimulation in the inferior thalamic peduncle. *Neurosurgery*, 65(6 Suppl), 203–9.

[134] Johnson, M. R., Wang, K., Tillmanns, S., Albin, N., & Diasio, R. B. (1997). Structural Organization of the Human Dihydropyrimidine Dehydrogenase Gene. *Cancer Research*, 57(9), 1660–1663.

[135] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1), D199–D205.

[136] Katsanis, S. H. & Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature Reviews Genetics*, 14(6), 415–426.

[137] Kennedy, B., Kronenberg, Z., Hu, H., Moore, B., Flygare, S., Reese, M. G., Jorde, L. B., Yandell, M., & Huff, C. (2014). Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Current Protocols in Human Genetics*, 81, 1–6.

[138] Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4), 357–360.

[139] Kim, S., Lohmueller, K., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J., & Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, 12(1), 231.

[140] Kim, T. I., McCall, J. G., Jung, Y. H., Huang, X., Siuda, E. R., Li, Y., Song, J., Song, Y. M., Pao, H. A., Kim, R. H., Lu, C., Lee, S. D., Song, I. S., Shin, G., Al-Hasani, R., Kim, S., Tan, M. P., Huang, Y., Omenetto, F. G., Rogers, J. A., & Bruchas, M. R.

(2013). Injectable, cellular-scale optoelectronics with applications for wireless optogenetics. *Science*, 340(6129), 211–216.

[141] Klei, L., Sanders, S. J., Murtha, M. T., Hus, V., Lowe, J. K., Willsey, A. J., Moreno-De-Luca, D., Yu, T. W., Fombonne, E., Geschwind, D., Grice, D. E., Ledbetter, D. H., Lord, C., Mane, S. M., Lese Martin, C., Martin, D. M., Morrow, E. M., Walsh, C. A., Melhem, N. M., Chaste, P., Sutcliffe, J. S., State, M. W., Cook Jr., E. H., Roeder, K., & Devlin, B. (2012). Common genetic variants, acting additively, are a major source of risk for autism. *Molecular Autism*, 3(1), 9.

[142] Kline, A. D., Calof, A. L., Schaaf, C. A., Krantz, I. D., Jyonouchi, S., Yokomori, K., Gauze, M., Carrico, C. S., Woodman, J., Gerton, J. L., Vega, H., Levin, A. V., Shirahige, K., Champion, M., Goodban, M. T., O'Connor, J. T., Pipan, M., Horsfield, J., Deardorff, M. A., Ishman, S. L., & Dorsett, D. (2014). Cornelia de Lange syndrome: further delineation of phenotype, cohesin biology and educational focus, 5th Biennial Scientific and Educational Symposium abstracts. *American Journal of Medical Genetics Part A*, 164A(6), 1384–1393.

[143] Kloet, S. L., Whiting, J. L., Gafken, P., Ranish, J., & Wang, E. H. (2012). Phosphorylation-Dependent Regulation of Cyclin D1 and Cyclin A Gene Transcription by TFIID Subunits TAF1 and TAF7. *Molecular and Cellular Biology*, 32(16), 3358–3369.

[144] Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., & Kronenberg, F. (2011). HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutation*, 32(1), 25–32.

[145] Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C. M., Brown, D. L., Brudno, M., Campbell, J., FitzPatrick, D. R., Eppig, J. T., Jackson, A. P., Freson, K., Girdea, M., Helbig, I., Hurst, J. A., Jähn, J., Jackson, L. G., Kelly, A. M., Ledbetter, D. H., Mansour, S., Martin, C. L., Moss, C., Mumford, A., Ouwehand, W. H., Park, S.-M., Riggs, E. R., Scott, R. H., Sisodiya, S., Vooren, S. V., Wapner, R. J., Wilkie, A. O. M., Wright, C. F., Vulto-van Silfhout, A. T., Leeuw, N. d., de Vries, B. B. A., Washingthon, N. L., Smith, C. L., Westerfield, M., Schofield, P., Ruef, B. J., Gkoutos, G. V., Haendel, M., Smedley, D., Lewis, S. E., & Robinson, P. N. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1), D966–D974.

[146] Komotar, R. J., Hanft, S. J., & Connolly Jr., E. S. (2009). Deep brain stimulation for obsessive compulsive disorder. *Neurosurgery*, 64(4), N13.

[147] Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., & Phillippy, A. M. (2012).

Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7), 693–700.

[148] Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 14(1), 289.

[149] Kotani, T., Miyake, T., Tsukihashi, Y., Hinnebusch, A. G., Nakatani, Y., Kawaichi, M., & Kokubo, T. (1998). Identification of Highly Conserved Amino-terminal Segments of dTAFII230 and yTAFII145 That Are Functionally Interchangeable for Inhibiting TBP-DNA Interactions in Vitro and in Promoting Yeast Cell Growth in Vivo. *Journal of Biological Chemistry*, 273(48), 32254–32264.

[150] Krawitz, P., Rodelsperger, C., Jager, M., Jostins, L., Bauer, S., & Robinson, P. N. (2010). Microindel detection in short-read sequence data. *Bioinformatics*, 26(6), 722–729.

[151] Kuzmak, P. M. & Dayhoff, R. E. (1998). The Department of Veterans Affairs integration of imaging into the healthcare enterprise using the VistA Hospital Information System and Digital Imaging and Communications in Medicine. *Journal of Digital Imaging*, 11(2), 53–64.

[152] L. Goff D. Kelley, C. T. (2013). cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. *R package version 2.10.0.*

[153] Lajin, B., Alachkar, A., Hamzeh, A. R., Michati, R., & Alhaj, H. (2011). No association between Val158Met of the COMT gene and susceptibility to schizophrenia in the Syrian population. *North American Journal of Medical Sciences*, 3(4), 176–178.

[154] Lam, H. Y., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., Gerstein, M. B., Butte, A. J., Ji, H. P., & Snyder, M. (2012a). Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology*, 30(1), 78–82.

[155] Lam, H. Y., Pan, C., Clark, M. J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M. B., Kidd, J. M., Bustamante, C. D., & Snyder, M. (2012b). Detecting and annotating genetic variations using the HugeSeq pipeline. *Nature Biotechnology*, 30(3), 226–229.

[156] Lau, J. Y., Goldman, D., Buzas, B., Hodgkinson, C., Leibenluft, E., Nelson, E., Sankin, L., Pine, D. S., & Ernst, M. (2010). BDNF gene polymorphism (Val66Met) predicts amygdala and anterior hippocampus responses to emotional faces in anxious and depressed adolescents. *NeuroImage*, 53(3), 952–961.

[157] Lee, D.-H., Gershenzon, N., Gupta, M., Ioshikhes, I. P., Reinberg, D., & Lewis, B. A. (2005). Functional Characterization of Core Promoter Elements: the Downstream Core Element Is Recognized by TAF1. *Molecular and Cellular Biology*, 25(21), 9674–9686.

[158] Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W. R., & Schatz, M. (2014). *Error correction and assembly complexity of single molecule sequencing reads*.

[159] Lee, H. & Schatz, M. C. (2012). Genomic Dark Matter: The reliability of short read mapping illustrated by the Genome Mappability Score. *Bioinformatics*.

[160] Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.

[161] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, G. P. D. P. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.

[162] Li, R., Li, Y., Fang, X., Yang, H., Wang, J., & Kristiansen, K. (2009b). SNP detection for massively parallel whole-Genome Researchequencing. *Genome Research*, 19(6), 1124–1132.

[163] Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5), 713–714.

[164] Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., & Wang, J. (2009c). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15), 1966–1967.

[165] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Yang, H., & Wang, J. (2010a). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2), 265–272.

[166] Li, S., Li, R., Li, H., Lu, J., Li, Y., Bolund, L., Schierup, M. H., & Wang, J. (2013). SOAPindel: Efficient identification of indels from short paired reads. *Genome Research*, 23(1), 195–200.

[167] Li, Y., Hu, Y., Bolund, L., & Wang, J. (2010b). State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human Genomics*, 4(4), 271–277.

[168] Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research*, 21(6), 940–951.

[169] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950), 289–293.

[170] Lipsman, N., Gerretsen, P., Torres, C., Lozano, A. M., & Giacobbe, P. (2012). A psychiatric primer for the functional neurosurgeon. *Journal of Neurosurgical Sciences*, 56(3), 209–220.

[171] Lipsman, N., Neimat, J. S., & Lozano, A. M. (2007). Deep brain stimulation for treatment-refractory obsessive-compulsive disorder: the search for a valid target. *Neurosurgery*, 61(1), 1–3.

[172] Lopez-Garcia, P., Young Espinoza, L., Molero Santos, P., Marin, J., & Ortuno Sanchez-Pedreno, F. (2012). Impact of COMT genotype on cognition in schizophrenia spectrum patients and their relatives. *Psychiatry Research*.

[173] Lundblad, M. S., Stark, K., Eliasson, E., Oliw, E., & Rane, A. (2005). Biosynthesis of epoxyeicosatrienoic acids varies between polymorphic CYP2C enzymes. *Biochemical and Biophysical Research Communications*, 327(4), 1052–1057.

[174] Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A., McGuire, A. L., Zhang, F., Stankiewicz, P., Halperin, J. J., Yang, C., Gehman, C., Guo, D., Irikat, R. K., Tom, W., Fantin, N. J., Muzny, D. M., & Gibbs, R. A. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England Journal of Medicine*, 362(13), 1181–1191.

[175] Luria, A. R. (1972). *The man with a shattered world; the history of a brain wound*. New York,: Basic Books.

[176] Luria, A. R. (1976). *The mind of a mnemonist : a little book about a vast memory*. Chicago: H. Regnery.

[177] Lyon, G. J. (2012a). Guest post: Time to bring human genome sequencing into the clinic.

[178] Lyon, G. J. (2012b). Personalized medicine: Bring clinical standards to human-Genetics Research. *Nature*, 482(7385), 300–301.

[179] Lyon, G. J., Jiang, T., Van Wijk, R., Wang, W., Bodily, P. M., Xing, J., Tian, L., Robison, R. J., Clement, M., Lin, Y., Zhang, P., Liu, Y., Moore, B., Glessner, J. T., Elia,

J., Reimherr, F., van Solinge, W. W., Yandell, M., Hakonarson, H., Wang, J., Johnson, W. E., Wei, Z., & Wang, K. (2011). Exome sequencing and unrelated findings in the context of complex disease research: ethical and clinical implications. *Discovery Medicine*, 12(62), 41–55.

[180] Lyon, G. J., O'Rawe, J., & Wiley (2015). Human genetics and clinical aspects of neurodevelopmental disorders. In K. Mitchell (Ed.), *The Genetics of Neurodevelopmental Disorders.* Wiley).

[181] Lyon, G. J. & Segal, J. P. (2013). Practical, ethical and regulatory considerations for the evolving medical and research genomics landscape. *Applied & Translational Genomics*.

[182] Lyon, G. J. & Wang, K. (2012). Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genomeic Medicine*, 4(7), 58.

[183] Mackay, T. F. & Moore, J. H. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, 6(6), 42.

[184] Maina, G., Rosso, G., Zanardini, R., Bogetto, F., Gennarelli, M., & Bocchio-Chiavetto, L. (2010). Serum levels of brain-derived neurotrophic factor in drug-naive obsessive-compulsive patients: a case-control study. *Journal of Affective Disorders*, 122(1-2), 174–178.

[185] Makino, S., Kaji, R., Ando, S., Tomizawa, M., Yasuno, K., Goto, S., Matsumoto, S., Tabuena, M. D., Maranon, E., Dantes, M., Lee, L. V., Ogasawara, K., Tooyama, I., Akatsu, H., Nishimura, M., & Tamiya, G. (2007). Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. *American Journal of Human Genetics*, 80(3), 393–406.

[186] Mancama, D., Mata, I., Kerwin, R. W., & Arranz, M. J. (2007). Choline acetyltransferase variants and their influence in schizophrenia and olanzapine response. *American Journal of Medical Genetics Part B*, 144B(7), 849–853.

[187] Mannini, L., Cucco, F., Quarantotti, V., Krantz, I. D., & Musio, A. (2013). Mutation spectrum and genotype-phenotype correlation in Cornelia de Lange syndrome. *Human Mutation*, 34(12), 1589–1596.

[188] McCarthy, D., Humburg, P., Kanapin, A., Rivas, M., Gaulton, K., Consortium, T. W., Cazier, J.-B., & Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3), 26.

[189] McGuire, A. L. & Lupski, J. R. (2010). Personal Genome Research : what should the participant be told? *Trends in Genetics*, 26(5), 199–201.

[190] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303.

[191] Meacham, F., Boffelli, D., Dhahbi, J., Martin, D., Singer, M., & Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12(1), 451.

[192] Meyer, K. J., Axelsen, M. S., Sheffield, V. C., Patil, S. R., & Wassink, T. H. (2012). Germline mosaic transmission of a novel duplication of PXDN and MYT1L to two male half-siblings with autism. *Psychiatric Genetics*, 22(3), 137–140.

[193] Mian, M. K., Campos, M., Sheth, S. A., & Eskandar, E. N. (2010). Deep brain stimulation for obsessive-compulsive disorder: past, present, and future. *Neurosurgical Focus*, 29(2), E10.

[194] Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, 16(9), 1182–1190.

[195] Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., & Devine, S. E. (2011a). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Research*, 21(6), 830–839.

[196] Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y., Leng, J., Li, R., Li, Y., Lin, C. Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stutz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., & Korbel, J. O. (2011b). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332), 59–65.

[197] Mitchell, K. J. (2012). What is complex about complex disorders? *Genome Biology*, 13(1), 237.

[198] Mitchell, K. J. & Porteous, D. J. (2011). Rethinking the genetic architecture of schizophrenia. *Psychological Medicine*, 41(1), 19–32.

[199] Mizzen, C. A., Yang, X.-J., Kokubo, T., Brownell, J. E., Bannister, A. J., Owen-Hughes, T., Workman, J., Wang, L., Berger, S. L., Kouzarides, T., Nakatani, Y., & Allis, C. (1996). The TAFII250 Subunit of TFIID Has Histone Acetyltransferase Activity. *Cell*, 87(7), 1261–1270.

[200] Montag, C., Reuter, M., Newport, B., Elger, C., & Weber, B. (2008). The BDNF Val66Met polymorphism affects amygdala activity in response to emotional stimuli: evidence from a genetic imaging study. *NeuroImage*, 42(4), 1554–1559.

[201] Moore, R. E. (1966). *Interval analysis*, volume 4. Prentice-Hall Englewood Cliffs.

[202] Moore, R. E. & Moore, R. E. (1979). *Methods and applications of interval analysis*, volume 2. SIAM.

[203] Moreno, E. & Pericchi, L. R. (1993). Bayesian robustness for hierarchical ε-contamination models. *Journal of Statistical Planning and Inference*, 37(2), 159–167.

[204] Moreno-De-Luca, A., Myers, S. M., Challman, T. D., Moreno-De-Luca, D., Evans, D. W., & Ledbetter, D. H. (2013). Developmental brain dysfunction: revival and expansion of old concepts based on new genetic evidence. *The Lancet Neurology*, 12(4), 406–414.

[205] Moscarello, J. M. & LeDoux, J. E. (2013). Active avoidance learning requires prefrontal suppression of amygdala-mediated defensive reactions. *The Journal of Neuroscience*, 33(9), 3815–3823.

[206] Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, 19(R2), R131–R136.

[207] Murphy, T. W., Zine, E. E., & Jenike, M. A. (2009). *Life in rewind : the story of a young courageous man who persevered over OCD and the Harvard doctor who broke all the rules to help him*. New York: William Morrow, 1st edition.

[208] Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S. S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P., Zecha, A., Mohseni, M., Puttmann, L., Vahid, L. N., Jensen, C., Moheb, L. A., Bienek, M., Larti, F., Mueller, I., Weissmann, R., Darvish, H., Wrogemann, K., Hadavi, V., Lipkowitz, B., Esmaeeli-Nieh, S., Wieczorek, D., Kariminejad, R., Firouzabadi, S. G., Cohen, M., Fattahi, Z., Rost, I., Mojahedi, F., Hertzberg, C., Dehghan, A., Rajab, A., Banavandi, M. J., Hoffer, J., Falah, M., Musante, L., Kalscheuer, V., Ullmann, R., Kuss, A. W., Tzschach, A., Kahrizi, K., & Ropers, H. H. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature*, 478(7367), 57–63.

[209] Narzisi, G., O'Rawe, J. A., Iossifov, I., Fang, H., Lee, Y.-h., Wang, Z., Wu, Y., Lyon, G. J., Wigler, M., & Schatz, M. C. (2014). Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nature Methods*, 11(10), 1033–1036.

[210] Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., Lin, C. F., Stevens, C., Wang, L. S., Makarov, V., Polak, P., Yoon, S., Maguire, J., Crawford, E. L., Campbell, N. G., Geller, E. T., Valladares, O., Schafer, C., Liu, H., Zhao, T., Cai, G., Lihm, J., Dannenfelser, R., Jabado, O., Peralta, Z., Nagaswamy, U., Muzny, D., Reid, J. G., Newsham, I., Wu, Y., Lewis, L., Han, Y., Voight, B. F., Lim, E., Rossin, E., Kirby, A., Flannick, J., Fromer, M., Shakir, K., Fennell, T., Garimella, K., Banks, E., Poplin, R., Gabriel, S., DePristo, M., Wimbish, J. R., Boone, B. E., Levy, S. E., Betancur, C., Sunyaev, S., Boerwinkle, E., Buxbaum, J. D., Cook Jr., E. H., Devlin, B., Gibbs, R. A., Roeder, K., Schellenberg, G. D., Sutcliffe, J. S., & Daly, M. J. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397), 242–245.

[211] Nelson, D. R., Zeldin, D. C., Hoffman, S. M. G., Maltais, L. J., Wain, H. M., & Nebert, D. W. (2004). Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics and Genomics*, 14(1), 1–18.

[212] Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S. A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M. D., Nangle, K., Wang, J., Abecasis, G., Cardon, L. R., Zollner, S., Whittaker, J. C., Chissoe, S. L., Novembre, J., & Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090), 100–104.

[213] Neumaier, A. (1990). *Interval methods for systems of equations*, volume 37. Cambridge university press.

[214] Neuman, J. A., Isakov, O., & Shomron, N. (2012). Analysis of insertion,deletion from deep-sequencing data: software evaluation for optimal detection. *Briefings in Bioinformatics*.

[215] Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., Beck, A. E., Tabor, H. K., Cooper, G. M., Mefford, H. C., Lee, C., Turner, E. H., Smith, J. D., Rieder, M. J., Yoshiura, K., Matsumoto, N., Ohta, T., Niikawa, N., Nickerson, D. A., Bamshad, M. J., & Shendure, J. (2010a). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics*, 42(9), 790–793.

[216] Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., & Bamshad, M. J. (2010b). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1), 30–35.

[217] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., & Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), 272–276.

[218] Nicolelis, M. A. (2012). Mind in motion. *Scientific American*, 307(3), 58–63.

[219] Niederriter, A. R., Davis, E. E., Golzio, C., Oh, E. C., Tsai, I. C., & Katsanis, N. (2013). In Vivo Modeling of the Morbid Human Genome using Danio rerio. (78), e50338.

[220] Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PloS One*, 7(7), e37558.

[221] Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443–451.

[222] Nolte, D., Niemann, S., & Müller, U. (2003). Specific sequence changes in multiple transcript system DYT3 are associated with X-linked dystonia parkinsonism. *Proceedings of the National Academy of Sciences*, 100(18), 10347–10352.

[223] O'Donnell, C. J. & Nabel, E. G. (2011). Genomics of Cardiovascular Disease. *New England Journal of Medicine*, 365(22), 2098–2109.

[224] Olson, M. V. (2012). Human genetic individuality. *Annual Review of Genomics and Human Genetics*, 13, 1–27.

[225] O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., & Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, 5(3), 28.

[226] O'Rawe, J., Wu, Y., Rope, A., Barrón, L. T. J., Swensen, J., Fang, H., Mittelman, D., Highnam, G., Robison, R., Yang, E., Wang, K., & Lyon, G. (2015). A variant in TAF1 is associated with a new syndrome with severe intellectual disability and characteristic dysmorphic features. *bioRxiv*.

[227] O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., Smith, J. D., Turner, E. H., Stanaway, I. B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J. M., Borenstein, E., Rieder, M. J., Nickerson, D. A., Bernier, R., Shendure, J., & Eichler, E. E. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397), 246–250.

[228] O'Rawe, J. A., Fang, H., Rynearson, S., Robison, R., Kiruluta, E. S., Higgins, G., Eilbeck, K., Reese, M. G., & Lyon, G. J. (2013). Integrating precision medicine in the study and clinical treatment of a severely mentally ill person. *PeerJ*, 1, e177.

[229] O'Rawe, J. A., Ferson, S., & Lyon, G. J. (2015). Accounting for uncertainty in DNA sequencing data. *Trends in Genetics*.

[230] Paila, U., Chapman, B. A., Kirchner, R., & Quinlan, A. R. (2013). GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLOS Computational Biologyogy*, 9(7), e1003153.

[231] Pais-Vieira, M., Lebedev, M., Kunicki, C., Wang, J., & Nicolelis, M. A. (2013). A brain-to-brain interface for real-time sharing of sensorimotor information. *Scientific Reports*, 3, 1319.

[232] Papai, G., Weil, P. A., & Schultz, P. (2011). New insights into the function of transcription factor TFIID from recent structural studies. *Current Opinion in Genetics and Development*, 21(2), 219–224.

[233] Pennington, K. L., Marr, S. K., Chirn, G. W., & Marr 2nd, M. T. (2013). Holo-TFIID controls the magnitude of a transcription burst and fine-tuning of transcription. *Proc Natl Acad Sci U S A*, 110(19), 7678–7683.

[234] Pericchi, L. R. (1998). Sets of Prior Probabilities and Bayesian Robustness ". *Imprecise Probability Project (http://ippserv. rug. ac. be/home/ipp. html), http://ippserv. rug. ac. be/documentation/robust/robust. html*.

[235] Pericchi, L. R. & Pérez, M. E. (1994). Posterior robustness with more than one sampling model. *Journal of Statistical Planning and Inference*, 40(2), 279–294.

[236] Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3), 290–295.

[237] Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., Robasky, K., Zaranek, A. W., Lee, J. H., Ball, M. P., Peterson, J. E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M. I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K. P., Ebert, J. C., Nilsen, G. B.,

Baccash, J., Halpern, A. L., Church, G. M., & Drmanac, R. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487(7406), 190–195.

[238] Pigott, T. A., Pato, M. T., Bernstein, S. E., Grover, G. N., Hill, J. L., Tolliver, T. J., & Murphy, D. L. (1990). Controlled comparisons of clomipramine and fluoxetine in the treatment of obsessive-compulsive disorder. Behavioral and biological results. *Archives of General Psychiatry*, 47(10), 926–932.

[239] Pigott, T. A. & Seay, S. M. (1999). A review of the efficacy of selective serotonin reuptake inhibitors in obsessive-compulsive disorder. *The Journal of Clinical Psychiatry*, 60(2), 101–106.

[240] Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, 8(1), 1–10.

[241] Piton, A., Redin, C., & Mandel, J. L. (2013). XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *American Journal of Human Genetics*, 93(2), 368–383.

[242] Protti, D. & Groen, P. (2008). Implementation of the Veterans Health Administration VistA clinical information system around the world. *Healthcare Quarterly*, 11(4), 83–89.

[243] Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S. E., Kahler, A., Duncan, L., Stahl, E., Genovese, G., Fernandez, E., Collins, M. O., Komiyama, N. H., Choudhary, J. S., Magnusson, P. K., Banks, E., Shakir, K., Garimella, K., Fennell, T., DePristo, M., Grant, S. G., Haggarty, S. J., Gabriel, S., Scolnick, E. M., Lander, E. S., Hultman, C. M., Sullivan, P. F., McCarroll, S. A., & Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487), 185–190.

[244] Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842.

[245] Ratiu, P., Talos, I. F., Haker, S., Lieberman, D., & Everett, P. (2004). The tale of Phineas Gage, digitally remastered. *Journal of Neurotrauma*, 21(5), 637–643.

[246] Raznahan, A., Greenstein, D., Lee, Y., Long, R., Clasen, L., Gochman, P., Addington, A., Giedd, J. N., Rapoport, J. L., & Gogtay, N. (2011). Catechol-o-methyl transferase (COMT) val158met polymorphism and adolescent cortical development in patients with childhood-onset schizophrenia, their non-psychotic siblings, and healthy controls. *NeuroImage*, 57(4), 1517–1523.

[247] Reese, M. G., Moore, B., Batchelor, C., Salas, F., Cunningham, F., Marth, G. T., Stein, L., Flicek, P., Yandell, M., & Eilbeck, K. (2010). A standard variation file format for human genome sequences. *Genome Biology*, 11(8), R88.

[248] Reid, C. (2012). Comparing Performance Data – Taking a Different Perspective .

[249] Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., Sabbe, B., Despierre, E., Vergote, I., Hilbush, B., Lambrechts, D., & Del-Favero, J. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature Biotechnology*, 30(1), 61–68.

[250] Richards, M., Coppee, F., Thomas, N., Belayew, A., & Upadhyaya, M. (2012). Facioscapulohumeral muscular dystrophy (FSHD): an enigma unravelled? *Human Genetics*, 131(3), 325–340.

[251] Richterich, P. (1998). Estimation of errors in ”raw” DNA sequences: a validation study. *Genome Research*, 8(3), 251–259.

[252] Rijkers, T., Deidda, G., van Koningsbruggen, S., van Geel, M., Lemmers, R. J., van Deutekom, J. C., Figlewicz, D., Hewitt, J. E., Padberg, G. W., Frants, R. R., & van der Maarel, S. M. (2004). FRG2, an FSHD candidate gene, is transcriptionally upregulated in differentiating primary myoblast cultures of FSHD patients. *Journal of Medical Genetics*, 41(11), 826–836.

[253] Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Consortium, W. G. S., Wilkie, A. O. M., McVean, G., & Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8), 912–918.

[254] Ring, B. J., Eckstein, J. A., Gillespie, J. S., Binkley, S. N., VandenBranden, M., & Wrighton, S. A. (2001). Identification of the human cytochromes p450 responsible for in vitro formation of R- and S-norfluoxetine. *The Journal of Pharmacology and Experimental Therapeutics*, 297(3), 1044–1050.

[255] Ritzel, M. W., Yao, S. Y., Huang, M. Y., Elliott, J. F., Cass, C. E., & Young, J. D. (1997). Molecular cloning and functional expression of cDNAs encoding a human Na+-nucleoside cotransporter (hCNT1). *American Journal of Physiology*, 272(2 Pt 1), 707–14.

[256] Roach, J. C., Boysen, C., Wang, K., & Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, 26(2), 345–353.

[257] Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., & Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978), 636–639.

[258] Robasky, K., Lewis, N. E., & Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1), 56–62.

[259] Rodriguez-Romaguera, J., Do Monte, F. H., & Quirk, G. J. (2012). Deep brain stimulation of the ventral striatum enhances extinction of conditioned fear. *Proceedings of the National Academy of Sciences of the United States of America*, 109(22), 8764–8769.

[260] Roffman, J. L., Gollub, R. L., Calhoun, V. D., Wassink, T. H., Weiss, A. P., Ho, B. C., White, T., Clark, V. P., Fries, J., Andreasen, N. C., Goff, D. C., & Manoach, D. S. (2008). MTHFR 677C –> T genotype disrupts prefrontal function in schizophrenia through an interaction with COMT 158Val –> Met. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45), 17573–17578.

[261] Roh, D., Chang, W. S., Chang, J. W., & Kim, C. H. (2012). Long-term follow-up of deep brain stimulation for refractory obsessive-compulsive disorder. *Psychiatry Research*, 200(2-3), 1067–1070.

[262] Rooms, L., Reyniers, E., Scheers, S., van Luijk, R., Wauters, J., Van Aerschot, L., Callaerts-Vegh, Z., D'Hooge, R., Mengus, G., Davidson, I., Courtens, W., & Kooy, R. F. (2006). TBP as a candidate gene for mental retardation in patients with subtelomeric 6q deletions. *European Journal of Human Genetics*, 14(10), 1090–1096.

[263] Rope, A. F., Wang, K., Evjenth, R., Xing, J., Johnston, J. J., Swensen, J. J., Johnson, W. E., Moore, B., Huff, C. D., Bird, L. M., Carey, J. C., Opitz, J. M., Stevens, C. A., Jiang, T., Schank, C., Fain, H. D., Robison, R., Dalley, B., Chin, S., South, S. T., Pysher, T. J., Jorde, L. B., Hakonarson, H., Lillehaug, J. R., Biesecker, L. G., Yandell, M., Arnesen, T., & Lyon, G. J. (2011). Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *American Journal of Human Genetics*, 89(1), 28–43.

[264] Rosa, A., Cuesta, M. J., Fatjó-Vilas, M., Peralta, V., Zarzuela, A., & Fañanás, L. (2006). The Val66Met polymorphism of the brain-derived neurotrophic factor gene is associated with risk for psychosis: Evidence from a family-based association study. *American Journal of Medical Genetics Part B*, 141B(2), 135–138.

[265] Rosenfeld, J. A., Mason, C. E., & Smith, T. M. (2012). Limitations of the human reference genome for personalized genomics. *PloS One*, 7(7), e40294.

[266] Ruppert, S. & Tjian, R. (1995). Human TAFII250 interacts with RAP74: implications for RNA polymerase II initiation. *Genes and Development*, 9(22), 2747–2755.

[267] Sacco, J. C., Abouraya, M., Motsinger-Reif, A., Yale, S. H., McCarty, C. A., & Trepanier, L. A. (2012). Evaluation of polymorphisms in the sulfonamide detoxification genes NAT2, CYB5A, and CYB5R3 in patients with sulfonamide hypersensitivity. *Pharmacogenetics and Genomics*, 22(10), 733–740.

[268] Sacks, O. W. (1995). *An anthropologist on Mars : seven paradoxical tales*. New York: Alfrd A. Knopf, 1st edition.

[269] Sacks, O. W. (1998). *The man who mistook his wife for a hat and other clinical tales*. New York, NY: Simon & Schuster, 1st touchs edition.

[270] Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovacs, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y. Y., Yachie, N., Zhong, Q., Shen, Y., Palagi, A., San-Miguel, A., Fan, C., Balcha, D., Dricot, A., Jordan, D. M., Walsh, J. M., Shah, A. A., Yang, X., Stoyanova, A. K., Leighton, A., Calderwood, M. A., Jacob, Y., Cusick, M. E., Salehi-Ashtiani, K., Whitesell, L. J., Sunyaev, S., Berger, B., Barabasi, A. L., Charloteaux, B., Hill, D. E., Hao, T., Roth, F. P., Xia, Y., Walhout, A. J., Lindquist, S., & Vidal, M. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3), 647–660.

[271] Sako, W., Morigaki, R., Kaji, R., Tooyama, I., Okita, S., Kitazato, K., Nagahiro, S., Graybiel, A. M., & Goto, S. (2011). Identification and localization of a neuron-specific isoform of TAF1 in rat brain: implications for neuropathology of DYT3 dystonia. *Neuroscience*, 189, 100–107.

[272] Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnstrom, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook Jr, E. H., Gibbs, R. A., Schellenberg, G. D., Sutcliffe, J. S., Devlin, B., Roeder, K., Neale, B. M., & Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9), 944–950.

[273] Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., Carriero, N. J., Meyer, K. A., Bilguvar, K., Mane, S. M., Sestan, N., Lifton, R. P., Gunel, M., Roeder, K., Geschwind, D. H., Devlin, B., & State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397), 237–241.

[274] Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414), 109–113.

[275] Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510), 421–427.

[276] Shah, D. B., Pesiridou, A., Baltuch, G. H., Malone, D. A., & O'Reardon, J. P. (2008). Functional neurosurgery in the treatment of severe obsessive compulsive disorder and major depression: overview of disease circuits and therapeutic targeting for the clinician. *Psychiatry*, 5(9), 24–33.

[277] Shanske, A. L., Goodrich, J. T., Ala-Kokko, L., Baker, S., Frederick, B., & Levy, B. (2012). Germline mosacism in Shprintzen-Goldberg syndrome. *American Journal of Medical Genetics Part A*, 158A(7), 1574–1578.

[278] Shestopal, S. A., Johnson, M. R., & Diasio, R. B. (2000). Molecular cloning and characterization of the human dihydropyrimidine dehydrogenase promoter. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1494(1,Äì2), 162–169.

[279] Shoichet, S. A., Hoffmann, K., Menzel, C., Trautmann, U., Moser, B., Hoeltzenbein, M., Echenne, B., Partington, M., van Bokhoven, H., Moraine, C., Fryns, J.-P., Chelly, J., Rott, H.-D., Ropers, H.-H., & Kalscheuer, V. M. (2003). Mutations in the ZNF41 Gene Are Associated with Cognitive Deficits: Identification of a New Candidate for X-Linked Mental Retardation. *The American Journal of Human Genetics*, 73(6), 1341–1354.

[280] Sim, S. C. & Ingelman-Sundberg, M. (2013). Update on allele nomenclature for human cytochromes P450 and the Human Cytochrome P450 Allele (CYP-allele) Nomenclature Database. *Methods in Molecular Biology*, 987, 251–259.

[281] Singh, J. P., Volavka, J., Czobor, P., & Van Dorn, R. A. (2012). A meta-analysis of the Val158Met COMT polymorphism and violent behavior in schizophrenia. *PloS One*, 7(8), e43423.

[282] Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2012). Association Testing for Next-Generation Sequencing Data Using Score Statistics. *Genetic Epidemiology*, 36(5), 430–437.

[283] Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics*, 195(3), 693–702.

[284] Slavin, T. P., Lazebnik, N., Clark, D. M., Vengoechea, J., Cohen, L., Kaur, M., Konczal, L., Crowe, C. A., Corteville, J. E., Nowaczyk, M. J., Byrne, J. L., Jackson, L. G., & Krantz, I. D. (2012). Germline mosaicism in Cornelia de Lange syndrome. *American Journal of Medical Genetics Part A*, 158A(6), 1481–1485.

[285] Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1), 195–197.

[286] Soliman, F., Glatt, C. E., Bath, K. G., Levita, L., Jones, R. M., Pattwell, S. S., Jing, D., Tottenham, N., Amso, D., Somerville, L. H., Voss, H. U., Glover, G., Ballon, D. J., Liston, C., Teslovich, T., Van Kempen, T., Lee, F. S., & Casey, B. J. (2010). A Genetic Variant BDNF Polymorphism Alters Extinction Learning in Both Mouse and Human. *Science*, 327(5967), 863–866.

[287] Stessman, H. A., Bernier, R., & Eichler, E. E. (2014). A genotype-first approach to defining the subtypes of a complex disease. *Cell*, 156(5), 872–877.

[288] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550.

[289] SVS (2012). SNP & Variation Suite (Version 7.6.10) [Software].

[290] Synofzik, M., Fins, J. J., & Schlaepfer, T. E. (2012). A neuromodulation experience registry for deep brain stimulation studies in psychiatric research: rationale and recommendations for implementation. *Brain Stimulation*, 5(4), 653–655.

[291] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., & Akey, J. M. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*.

[292] The Genomes Project, C. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74.

[293] Thomson, E. E., Carra, R., & Nicolelis, M. A. (2013). Perceiving invisible light through a somatosensory cortical prosthesis. *Nature Communications*, 4, 1482.

[294] Torres, A. R., Ramos-Cerqueira, A. T., Ferrao, Y. A., Fontenelle, L. F., do Rosario, M. C., & Miguel, E. C. (2011). Suicidality in obsessive-compulsive disorder: prevalence and relation to symptom dimensions and comorbid conditions. *The Journal of Clinical Psychiatry*, 72(1), 17–20.

[295] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript ex-

pression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3), 562–578.

[296] Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

[297] Van Horn, J. D., Irimia, A., Torgerson, C. M., Chambers, M. C., Kikinis, R., & Toga, A. W. (2012). Mapping connectivity damage in the case of Phineas Gage. *PloS One*, 7(5), e37454.

[298] van Oven, M. & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30(2), E386–E394.

[299] van Winkel, R., Moons, T., Peerbooms, O., Rutten, B., Peuskens, J., Claes, S., van Os, J., & De Hert, M. (2010). MTHFR genotype and differential evolution of metabolic parameters after initiation of a second generation antipsychotic: an observational study. *International Clinical psychopharmacology*, 25(5), 270–276.

[300] Vieira, F. G., Fumagalli, M., Albrechtsen, A., & Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Research*, 23(11), 1852–1861.

[301] Visscher, P. M., Goddard, M. E., Derks, E. M., & Wray, N. R. (2011). Evidence-based Psychiatric Genetics, AKA the false dichotomy between common and rare variant hypotheses. *Molecular Psychiatry*.

[302] Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall London.

[303] Wang, H., Curran, E. C., Hinds, T. R., Wang, E. H., & Zheng, N. (2014). Crystal structure of a TAF1-TAF7 complex in human transcription factor IID reveals a promoter binding module. *Cell Research*, 24(12), 1433–1444.

[304] Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research*, 41(W1), W77–W83.

[305] Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., & Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11), 1665–1674.

[306] Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16), e164.

[307] Wei, Z., Wang, W., Hu, P., Lyon, G. J., & Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*, 39(19), e132.

[308] Weldon, W. F. R. (1902). Mendel's laws of alternative inheritance in peas. *Biometrika*, 1, 228–254.

[309] Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H., & Reich, D. (2012). Phasing of many thousands of genotyped samples. *American Journal of Human Genetics*, 91(2), 238–251.

[310] Williamson, R. C. & Downs, T. (1990). Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2), 89–158.

[311] Worthey, E. A., Mayer, A. N., Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., Serpe, J. M., Dasu, T., Tschannen, M. R., Veith, R. L., Basehore, M. J., Broeckel, U., Tomita-Mitchell, A., Arca, M. J., Casper, J. T., Margolis, D. A., Bick, D. P., Hessner, M. J., Routes, J. M., Verbsky, J. W., Jacob, H. J., & Dimmock, D. P. (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine*, 13(3), 255–262.

[312] Wu, P. Y., Phan, J. H., & Wang, M. D. (2013). Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics*, 14 Suppl 1, S8.

[313] Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J. A., & Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics*.

[314] Yandell, M. & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5), 329–342.

[315] Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L. B., & Reese, M. G. (2011). A probabilistic disease-gene finder for personal genomes. *Genome Research*, 21(9), 1529–1542.

[316] Yu, J. H., Jamal, S. M., Tabor, H. K., & Bamshad, M. J. (2013). Self-guided management of exome and whole-genome sequencing results: changing the results return model. *Genetics in Medicine*.

[317] Yuan, B., Pehlivan, D., Karaca, E., Patel, N., Charng, W.-L., Gambin, T., Gonzaga-Jauregui, C., Sutton, V. R., Yesil, G., Bozdogan, S. T., Tos, T., Koparir, A., Koparir, E., Beck, C. R., Gu, S., Aslan, H., Yuregir, O. O., Al Rubeaan, K., Alnaqeb, D., Alshammari, M. J., Bayram, Y., Atik, M. M., Aydin, H., Geckinli, B. B., Seven, M., Ulucan, H., Fenercioglu, E., Ozen, M., Jhangiani, S., Muzny, D. M., Boerwinkle, E., Tuysuz, B., Alkuraya, F. S., Gibbs, R. A., & Lupski, J. R. (2015). Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes. *The Journal of Clinical Investigation*, 125(2), 636–651.

[318] Zhao, Z., Wang, W., & Wei, Z. (2013). An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *The Annals of Applied Statistics*, 7(4), 2229–2248.

[319] Zhong, Q., Simonis, N., Li, Q., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D., Swearingen, V., Yildirim, M. A., Yan, H., Dricot, A., Szeto, D., Lin, C., Hao, T., Fan, C., Milstein, S., Dupuy, D., Brasseur, R., Hill, D. E., Cusick, M. E., & Vidal, M. (2009). *Edgetic perturbation models of human inherited disorders*, volume 5.

[320] Zhou, S. F., Liu, J. P., & Chowbay, B. (2009). Polymorphism of human cytochrome P450 enzymes and its clinical impact. *Drug Metabolism Reviews*, 41(2), 89–295.

[321] Zhu, M., Need, A. C., Han, Y., Ge, D., Maia, J. M., Zhu, Q., Heinzen, E. L., Cirulli, E. T., Pelak, K., He, M., Ruzzo, E. K., Gumbs, C., Singh, A., Feng, S., Shianna, K. V., & Goldstein, D. B. (2012). Using ERDS to infer copy-number variants in high-coverage genomes. *American Journal of Human Genetics*, 91(3), 408–421.

[322] Zumbo, P. & Mason, C. E. (2014). Molecular methods for profiling the RNA world. . In *Genome Analysis: Current Procedures and Applications.* Horizon Press.