Next-Generation Roadmap for Patient-Centered Genomics

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable Link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:33789915">http://nrs.harvard.edu/urn-3:HUL.InstRepos:33789915</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:_dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:_dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>

(Article begins on next page)
Next-Generation Roadmap for Patient-Centered Genomics

Niklas Smedemark-Margulies

A Thesis Submitted to the Faculty of

The Harvard Medical School

in Partial Fulfillment of the Requirements

for the Degree of Master of Medical Sciences in Immunology

Harvard University

Boston, Massachusetts.

May, 2016
Next-Generation Roadmap for Patient-Centered Genomics

Abstract

In the era of precision medicine, understanding genetic variation has grown from a topic of research interest into a tangible source of therapeutic benefit for patients. As the list of confirmed links between genetic lesions and disease continues to grow\(^1\)–\(^7\), so does the list of actionable genetic diagnoses\(^8\)–\(^13\).

The workup of a childhood-onset schizophrenia case provides a useful foil for discussion of current methods for genomic diagnostics, both to demonstrate some of the important available analyses, and to highlight areas of ongoing need. In brief, the stages of this case as pertains to the general diagnostic process are: clinical workup, sequencing and technical processing, analysis and interpretation of results, and follow-up research study.

The patient in this case presented with command auditory hallucinations at age 6 and began empirical treatment for schizophrenia; he was subsequently found to have a novel \textit{de novo} heterozygous missense mutation in \textit{ATP1A3} NM\_152296.4\ c.385G\textgreater A, predicted to cause the coding change p.V129M. This gene codes for a neuron-specific isoform of the alpha subunit of the sodium-potassium pump complex that helps establish transmembrane ion gradients necessary for neuronal function. The variant found in this case is now being replicated in a patient-derived iPS-neuron model to seek greater insight into the mechanism of disease and possible therapeutic opportunities.

Generalizing from this case, researchers and clinicians hoping to replicate or improve upon this patient-centric genomics workflow can benefit from reviewing technical and infrastructural best practices. This case may also help illustrate some of the key difficulties in
connecting genomic evidence with appropriate functional validation and other clinical markers to support well-informed decision-making.
Table of Contents

List of Figures .............................................................................................................................. vii
List of Tables ................................................................................................................................. xi
Acknowledgements ....................................................................................................................... xii
Introduction and Overview ........................................................................................................... 1
  Introduction ................................................................................................................................ 1
  Overview ..................................................................................................................................... 2
Framing Example Case: Patient 0515-01, Childhood-onset Schizophrenia and ATP1A3 ............ 3
  Introduction .................................................................................................................................. 3
  Clinical and research domain background ................................................................................. 3
  Clinical presentation and family history ..................................................................................... 5
  Neurological assessment .............................................................................................................. 6
  Treatment history .......................................................................................................................... 7
  Genomic analyses ......................................................................................................................... 8
  Discussion primer .......................................................................................................................... 9
  Discussion: Motor phenotypes of ATP1A3 mutations ................................................................. 11
  Discussion: Psychiatric phenotypes of mutations in gene family ATP1A ...................................... 12
  Discussion: Evidence from tissue, and from immortalized and primary cell lines .................. 13
  Discussion: Downstream affected pathways and their disease relevance .................................. 14
  Ongoing research and functional validation .............................................................................. 15
  Detailed sequencing methods ..................................................................................................... 16
Patient Genomics Roadmap .......................................................................................................... 20
  Roadmap 1. Clinical Phenotyping and Indications for Sequencing ........................................... 22
      Overview and indications for sequencing .............................................................................. 22
      Family history and phenotyping ............................................................................................ 22
  Roadmap 2. Data Generation and Primary Analysis Methods .................................................. 24
      Preparing for sequencing ........................................................................................................ 24
      Chromosomal microarray ....................................................................................................... 25
      Exome and genome sequencing chemistry ........................................................................... 25
      Sanger sequencing .................................................................................................................. 28
  Roadmap 3. Secondary Analysis Methods ................................................................................ 29
Negative findings: Inconclusive variants, research models .......................................................... 60
Routes to biomarkers of disease and high-throughput screening ................................................. 61
Discussion .................................................................................................................................... 62
Future directions .......................................................................................................................... 63
Bibliography ................................................................................................................................. 65
List of Figures

**Figure 1:** Sequence and Protein models of ATP1A3 Variant. A) Models compiling previously published mutations in ATP1A3\textsuperscript{14,15}. AHC - Alternating Hemiplegia of Childhood. RDP - Rapid-onset Dystonia Parkinsonism. CAPOS - Cerebellar ataxia, Areflexia, Pes cavus, Optic atrophy and Sensorineural hearing loss. Mutations modeled using UCSF Chimera package and Phyre2 web portal relative to UniProt P13637\textsuperscript{16,17}. B) V129M mutation. COS - Child Onset Schizophrenia. C) Sequence conservation plot produced using Clustal Omega via EMBL-EBI\textsuperscript{18,19}, (http://www.ebi.ac.uk/Tools/msa/clustalo/). D) Sanger sequencing results plotted using Geneious version 8.1.4\textsuperscript{20}. E-F) Molecular modeling of the p.V129M using the homologous protein ATP1B1 in *Squalus acanthias* (PDB code: 2ZXE) with position numberings relative to ATP1A3. Model visualized using Pymol (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.).

**Figure 2:** Patient Genomics Roadmap. This diagram is intended to highlight some of the key steps for successful genetic workup; these steps are discussed in detail through the following sections. 1) The case workflow begins when patients are identified with indications for possible genetic disease, and are phenotyped before being referred for sequencing. 2) A patient tissue sample is processed on a next-generation sequencing platform to produce raw read data. 3) Raw data must then be analyzed using standard processing pipelines to produce a list of candidate variants and to describe these data with contextual quality information. 4) Candidate variants identified during this analysis must be annotated with evidence from many levels to provide sufficient context to understand possible disease relevance. 5) Once variants have been thoroughly annotated, they must be interpreted in the context of the case by considering these levels of annotation, as well as patient phenotype data or family history. 6) Variants of likely disease relevance should be confirmed, and those variants with known clinical implications may
result in follow-up action such as treatment modification or clinical trial enrollment, while novel or inconclusive variants may result in research-level follow-up such as patient-derived cell modeling.

**Figure 3:** The Broad Institute’s Genome Analysis Toolkit\(^{21}\) (GATK) Best Practices workflow, reprinted from https://www.broadinstitute.org/gatk/guide/best-practices.php. This workflow is considered an accepted industry standard for processing sequence data produced on Illumina sequencing platforms. This workflow begins from raw sequencing reads in FASTQ format, and produces aligned reads in SAM/BAM format and variants with initial annotations in VCF format. The GATK documentation provides thorough instructions for each step of this process.

**Figure 4:** Example of an individual read in a FASTQ file. The FASTQ format contains a sequence identifier describing the machine and location within the machine where a read was produced, a nucleotide sequence, an optional description, and a quality score for each nucleotide described using an ASCII schema (described in detail elsewhere\(^{22}\)).

**Figure 5:** Example read in a SAM/BAM file. See https://samtools.github.io/hts-specs/SAMv1.pdf for a detailed explanation of each field.

**Figure 6:** An example of a pileup of aligned sequence reads, viewed using the WuXi NextCODE Sequence Miner software. Forward and reverse aligned reads are indicated in blue and orange, respectively. Positions highlighted in yellow differ from the reference sequence, shown at bottom.

**Figure 7:** This reprinted diagram shows how several types of variation events will appear in read data; understanding the expected configuration of reads may help an investigator manually evaluate supporting a putative variant. Blue arrows indicate read pairs. For each event shown, the reference sequence is shown at bottom and the test sequence is shown at top\(^{23}\).
Figure 8: Levels of Evidence. A schematic to overview some of the levels of evidence that can give context to candidate variants during case interpretation, shown here for a variant in a hypothetical transmembrane channel. A candidate variant is shown in the underlying raw read data, overlaid on a gene model, and shown on the cDNA for a single splice isoform. Based on the gene model, this variant can be described with a predicted amino acid consequence. This consequence can be examined in the linear amino acid sequence, a model of known or predicted protein functional domains, or in a model of the folded protein. This affected protein and the expected effect of the variant can be described within a relevant pathway, or functional network such as known protein-protein interactions. Finally, past disruptions of this protein, pathway, or variant can be examined in a relevant reference population.

Figure 9: This figure is reprinted from WuXi NextCODE documentation (goo.gl/8Mmm5T), and shows the possible descriptors that the Variant Effect Predictor algorithm\textsuperscript{24} may give when annotating the gene-level consequence of a candidate variant. This list thoroughly covers the effects that are currently assessed for single nucleotide variants and short indels.

Figure 10: Adding charge to a transmembrane domain. This example is meant to convey the importance of considering environmental context when evaluating the protein-level effect of a variant. Whereas adding a charged residue to a lone transmembrane (left) may be strongly destabilizing to the protein, adding charge may be tolerated in the core of a transmembrane ion channel (middle) or at a charged interface between two transmembrane proteins.

Figure 11: An example Genomic Ordered Relations (GOR) query in the WuXi NextCODE Sequence Miner software. This query simultaneously processes multiple exome trio datasets (with data from an affected proband and from both parents). The query selects exonic compound heterozygous SNVs in RefSeq genes with minor allele frequency less than 1%, that passed a
variant quality score filtration, with a phred-scale genotype quality greater than 20 and predicted protein consequence of frameshift, stopgain, stoploss, or splice disruption. Minor allele frequencies are drawn from Broad Institute’s Exome Aggregation Consortium\(^\text{25}\), NHLBI’s Exome Variant Server (http://evs.gs.washington.edu/), Thousand Genomes Project phase 3\(^\text{26}\), Genome of the Netherlands\(^\text{27}\), and Human Genetic Variation Database\(^\text{28}\). Variant phasing is determined by site-specific comparison between subjects and parents.

**Figure 12:** An example of annotated variants produced by a Genomic Ordered Relations (GOR) query and viewed using the WuXi NextCODE Sequence Miner platform. Each row corresponds to one variant, with column headers describing each annotation column. GT: genotype.

**VEP\_Max\_Score:** Variant Effect Predictor pathogenicity score, calculated as the maximum of SIFT, PolyPhen-2, and MutationTaster. **VEP\_Max\_Af:** the maximum Minor Allele Frequency observed across Broad Institute’s Exome Aggregation Consortium\(^\text{25}\), NHLBI’s Exome Variant Server (http://evs.gs.washington.edu/), Thousand Genomes Project Phase 3\(^\text{26}\), Genome of the Netherlands\(^\text{27}\), and Human Genetic Variation Database\(^\text{28}\). **OMIM\_description:** Gene-level descriptions obtained from Online Mendelian Inheritance in Man\(^\text{29}\). **DIAG\_ACMGcat:** variant-level diagnostic category according to American College of Medical Genetics 2013 criteria\(^\text{30}\).
List of Tables

Table 1: Example Case Variant Table. ATP1A3 reference sequence = NM_152296.4, ENST00000302102. HGVS = Human Genome Variation Society. ExAC = Exome Aggregation Consortium (http://exac.broadinstitute.org). RVIS - Residual Variation Intolerance Score. SNV= Single Nucleotide Variant. MAF = Minor Allele Frequency. The ExAC constraint z-score from Broad Institute’s ExAC browser compares the expected frequency of functional variation to the observed frequency, where a large positive z-score indicates a gene significantly depleted for variation31, (http://exac.broadinstitute.org/faq). The Residual Variation Intolerance Score (RVIS) indicates the percentile of variation intolerance, where lower percentiles are more intolerant32, (http://genic-intolerance.org).

Table 2: Example Case Sequencing Coverage Information. Sequencing coverage information for the proband and parents. The target region comprises 44.1MB as defined by the EZ Exome 2.0 capture kit used for sequencing. The coverage at the site of the variant of interest is included to give context for the trio genotype obtained from exome sequencing at that site.
Acknowledgements

For mentorship and support, both personal and technical
Timothy Yu and Yu lab
Peter Park and Park lab
NextCODE team
Claritas team
MMSc Program directors, Shiv Pillai and Michael Carroll
David, Else, Daniel, Maja, Carrie, Charlie
Roommates and Friends
MMSc classmates

Full authors list for ATP1A3 workup
Niklas Smedemark-Margulies
Catherine A. Brownstein, MPH, PhD
Sigella Vargas, MD
Sahil K. Tembulkar
Meghan C. Towne, MS, GC
Jiahai Shi
Elisa Gonzalez-Cuevas
Kevin X. Liu, PhD
Kaya Bilguvar, MD
Robin J. Kleiman, PhD
Min-Joon Han, PhD
Alcy Torres, MD
Gerard T. Berry, MD
Timothy W. Yu, MD, PhD
Alan H. Beggs, PhD
Pankaj B. Agrawal, MD, MMSc
Joseph Gonzalez-Heydrich, MD

Cold Spring Harbor Molecular Case Studies; in press

This work was conducted with support from Students in the Master of Medical Sciences in Immunology program of Harvard Medical School. The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard University and its affiliated academic health care centers.
Introduction and Overview

Introduction

The announcement of the precision medicine initiative serves as a useful rallying point for the era of personalized genomic diagnostics. One of the key goals is to provide targeted treatment through specific, mechanistically-informed diagnosis, such as moving beyond DSM-V to describe mental illness in terms of underlying physiologic mechanisms.

This movement seeks to advance the current idiom, where the entry point for treatment and diagnosis are often downstream molecular markers that give an incomplete view of the full state of biological networks in a patient. Instead, precision medicine aspires to leverage network-scale measurements of multiple omics datasets to provide a systems-level description of disease states and compensatory mechanisms.

Given the rapidly increasing diagnostic yield of DNA and RNA sequencing, practitioners of precision medicine today have begun to take advantage of next-generation sequencing.

Everyone knows there is no free lunch; making best use of genomic information requires substantial infrastructure to: 1) process huge datasets consistently and efficiently at the whole-exome and whole-genome scale; 2) build and maintain reference databases; 3) design and validate pathogenicity-prediction algorithms to annotate processed data for interpretation; and 4) design data visualization tools that provide real-time diagnostic power so that domain-experts can make use of these interpreted datasets.

Largely similar workflows can support the analysis of genomic, transcriptional or even epigenetic datasets. These datasets can be further supported by evidence from other contexts,
such as experimental systems for validating putative mechanisms or disease associated cellular signatures, and computational models of affected cellular components such as simulations of protein folding, protein-protein interaction, and chromatin geometry over time.

As an ultimate goal, these sets of evidence may help identify specific readouts or biomarkers of disease\textsuperscript{38}. Such markers can then be translated into a high-throughput drug-screening platform\textsuperscript{39}. These platforms may help identify candidate treatments\textsuperscript{40} or help match candidate treatments to specific patients\textsuperscript{41}.

**Overview**

To provide a grounded example of some of the topics mentioned above, this discussion will first cover the specific example of a patient with a striking presentation of childhood-onset schizophrenia. This patient was subjected to genetic workup, which uncovered a likely pathogenic mutation in a neuron-specific protein. This mutation is now under functional study in a patient-derived tissue model. The intent of discussing this patient’s case is both to highlight the interesting features of this specific patient’s presentation, and also to concretely demonstrate some portions of this general workflow.

Following an explanation of this example case, the discussion will make another pass through this workflow for a generalized case. This second pass will focus on a refresher of the methods and keys to success at each stage, as well as emphasizing some areas where this specific case workup was lacking or where work for this individual patient is ongoing.

Finally, the discussion will broaden to outline some areas with tremendous potential for future development and to emphasize some ways that current research may naturally feed into these future applications.
Framing Example Case: Patient 0515-01, Childhood-onset Schizophrenia and ATP1A3

Introduction

The general process of a genetic workup is best understood in the context of a real example. The following example case comes from the clinical cohort of Dr. Joseph Gonzalez-Heydrich. His cohort consists of 49 patients with severe early-onset psychosis in the absence of trauma or other likely environmental causes. One of the overall goals of this cohort is enroll patients whose disease is likely to be of genetic origin, and search for recurrent genetic features across cases. The proband in this case is referred to by his research ID number, 0515-01. This case illustrates some of the spectrum of issues and opportunities for genomic characterization of a disorder that is currently described in syndromic rather than mechanistic terms. The full list of contributors to this case workup is provided in the acknowledgements section.

Clinical and research domain background

Patients with childhood-onset schizophrenia (COS) meet the same DSM criteria as typical late adolescent-adult onset schizophrenia (SZ) patients but with onset of psychosis before 13 years. COS is rare, having a prevalence of around 1 in 40,000\textsuperscript{42}. The disease presents with a premorbid phase characterized by impairment in motor, social, and cognitive functioning\textsuperscript{43,44}. It progresses to include a characteristic combination of symptoms that can include delusions, hallucinations, disorganized speech, grossly disorganized or catatonic behavior, negative symptoms (i.e., diminished emotional expression or avolition) and diminished functioning\textsuperscript{45}. While antipsychotics are the mainstay of treatment for COS, response rates are only moderate at
best and most patients still suffer from marked functional impairments and lack effective treatment\(^\text{46}\).

Efforts to understand the genetic architecture of SZ have included genome wide association studies of large patient cohorts and controls seeking to identify candidate loci in linkage with common variants\(^\text{47,48}\). Rare genetic variants also play a role in SZ risk; chromosomal microarray and exome sequencing studies have identified excesses of \textit{de novo} protein-impactful variants in COS cases\(^\text{44,47}\). When these mutations occur in highly conserved or mutation-intolerant genes, the probability that they are disease-relevant mutations may be greatly increased\(^\text{49,32,50}\). Moving forward from these high-level observations about the burden of mutation in this disease will require identifying candidate genotypes for targeted functional studies to uncover the physiologic role of disease-associated variants and examine how they give rise to the disease phenotypes. Studies of COS cases may reveal a powerful set of disease-associated mutations, given the severity and early onset of that phenotype, and may thereafter provide a window into the etiology of schizophrenia more broadly.

The following workup describes a case of childhood-onset schizophrenia with a novel heterozygous \textit{de novo} mutation in the \textit{ATP1A3} gene. This gene codes for the catalytic alpha subunit of a neuron-specific ATP-dependent transmembrane sodium-potassium pump. Previously, mutations in this gene have been associated with both sporadic and inherited forms of Alternating Hemiplegia of Childhood (AHC) and Rapid-onset Dystonia Parkinsonism (RDP). This case is noteworthy for the early and severe onset of psychotic symptoms associated with \textit{ATP1A3} mutation, without previous or simultaneous onset of motor phenotypes previously associated with that gene.
Clinical presentation and family history

The proband is a 9-year-old Caucasian boy with a history of selective mutism and severe aggression who presented with command auditory hallucinations and behavioral worsening meeting full DSM-5 criteria for COS at 6 years of age. He was born full-term via emergency cesarean section after maternal pre-eclampsia and fetal tachycardia. At birth he weighed 8-pounds 11 ounces, was noted to have difficulty breathing, and was admitted to the Neonatal Intensive Care Unit for 24 hours. At 2 months of age, he was diagnosed with decreased muscle tone for which he began receiving Early Intervention Services including physical therapy. Around 2 years of age, he exhibited severe head-banging.

At the age of 3 he was diagnosed with selective mutism, pervasive developmental disorder (NOS), and depression. He was described as having mood swings, lack of emotional control, and severe separation anxiety. He also had severe self-injurious behaviors – for example, he tried to pull his teeth out and to cut his gingiva out with scissors. For the management of anxiety, he was started on clonazepam, which was then discontinued due to increased aggressive behavior.

At age 6 years and 2 months, he reported auditory hallucinations. He was found hitting himself in the head and said he was trying to silence the voices of two small boys that he described as having a “bed in his head.” These voices often said “bad things,” told him to hurt himself and others, and he felt he needed to obey them. He had delusional conviction that the boys in his head were real. At this time he also began experiencing diurnal enuresis, although he had toilet trained at age 4. The proband’s history of aggression towards his sister and dog worsened at this time and became highly unpredictable, to the extent that he could not be left alone with them or any other children. His mother noted that he had episodes of stiffness while
sleeping. His physical exam at this time was unremarkable. He met diagnostic criteria for DSM 5 Schizophrenia with hallucinations and delusions (Criterion A) and decreased level of functioning (Criterion B) persisting for 9 months (Criterion C) with no major mood episode present during the majority of this time (Criterion D) and no discernable pharmacologic or medical cause explaining his symptoms (Criterion E)\textsuperscript{51}

There is no family history of birth defects, recurrent miscarriages, stillbirths, infant deaths, or consanguinity. The proband has one full sister who is one year younger and healthy.

The proband's father is in his 30's and healthy. The proband has a paternal cousin with bipolar disorder and ADHD. The proband’s father has a maternal half-sister who is healthy. This half-sister has five children: one healthy teenage son, one teenage son with impulsive behavior disorder, one teenage son with bipolar disorder and ADHD but no known psychotic symptoms, one healthy prepubertal son, and one healthy infant daughter.

The proband's mother is in her late 20’s and has depression. She has a twin brother with ADHD. The proband has a maternal aunt with epilepsy and developmental delay, a maternal uncle with anxiety, and maternal grandfather with a history of addiction and maternal grandmother with depression.

**Neurological assessment**

Following recognition of psychosis, the proband was referred for neurological assessment from a specialist at BCH. His neurologic examination showed intact extraocular movements and pupils that were equally round and reactive. His facial strength was symmetric and normal. His jaw strength was normal. His palate raised symmetrically. His tongue was midline and had normal strength bilaterally. His red reflex was noted bilaterally. He had normal muscle bulk and tone and full strength in the upper and lower extremities bilaterally. His sensory exam was intact
to light touch. His deep tendon reflexes were normal and symmetric, as were his finger to nose movements and gait. He was able to walk on his heels, toes, and tandem gait, as well as hop on either foot, and run 20 feet without difficulty.

Given the literature documenting motor phenotypes in association with mutations in the \textit{ATP1A3} gene (discussed below), the proband was specifically screened for any motor or autonomic symptoms, including episodic symptoms and has had none other than the episodes of stiffness in sleep noted in his clinical presentation.

He had a clinical brain MRI that was read as within normal limits and an EEG that was read as abnormal due to arrhythmic diffuse slowing. This latter finding may be explained by his medication regimen, but is also consistent with excess theta and delta activity in the EEG of patients with SZ$^{52}$.

\textbf{Treatment history}

The proband’s auditory hallucinations initially responded to risperidone (with benztropine added to prevent extrapyramidal symptoms), but he was switched to olanzapine to control his aggression. His selective mutism and enuresis resolved completely. A few months later he developed depressive symptoms that resolved with the addition of fluoxetine. After being stable for 7 months, he began hearing voices again, thus Haloperidol was added to olanzapine and fluoxetine. Guanfacine and atomoxetine were added to manage his ADHD symptoms, but resulted in increased aggression. Lithium was prescribed to address his mood fluctuations with some initial benefit. At age 9, he began to exhibit echolalia. After 18 months without auditory hallucinations, and while continuing to take olanzapine, haloperidol, fluoxetine, benztropine, and lithium, he started hearing voices again. This was accompanied by increased aggression and frequent diurnal enuresis and encopresis, similar to when his psychotic symptoms
first became evident. Increasing his antipsychotic medication has seemed to reverse this relapse for now.

**Genomic analyses**

The proband was first assessed using chromosomal microarray, which did not identify any copy number variants. Next, the proband and both parents underwent whole exome sequencing (WES) from peripheral blood lymphocytes, which identified a high-confidence de novo missense change in *ATP1A3* corresponding to NM_152296.4:c.385G>A and p.V129M. To validate this initial finding, Sanger sequencing from peripheral blood was performed on the proband, both parents, and the proband’s healthy sibling; this confirmed the presence of the variant in the proband, and its absence in his parents and sister (Figure 1D). This variant has not been previously described, but other *ATP1A3* variants have been documented in association with a range of autosomal dominant neurological and psychomotor phenotypes discussed below.

This gene was examined using the Exome Aggregation Consortium’s browser, and the Genic Intolerance database, which each provide a quantification of the probability of gene-level intolerance to functional variation. Both analyses show that this gene is highly intolerant to variation, increasing the likelihood that mutation in this gene is associated with disease (Table 1). Analysis of the p.V129M variant using the SIFT and Polyphen-2 algorithms, led to predictions that this change is deleterious by both methods (SIFT score 0, PolyPhen-2 score 0.999). Additional details of this variant are summarized in Table 1.

Molecular modeling of ATP1A3 p.V129M was performed on several models to examine the potential effect of this mutation on the function of the protein. First, the variant of interest was modeled alongside previously described variants using a predicted structure produced from the known amino acid sequence (Figure 1A, B). Next, the affected residue was examined for
conservation across species (Figure 1C). Finally, since a crystal structure of the human protein is not available, the proband’s amino acid change was modeled relative to the crystal structure of the homologous sodium-potassium pump ATP1B1 in *Squalus Acanthias*\(^{55}\), which has 92% amino acid sequence homology with human ATP1A3 (see Figure 1 E-F). Homologous positions given here are relative to ATP1A3 numbering, but the relationships described between residues were examined in the ATP1B1 crystal structure. This modeling by homology shows that the mutation of interest may affect the potassium-binding residues in this channel; residue V129 is located in a trans-membrane domain, and forms hydrophobic interactions with residues L798 and L802 near the potassium-binding residue D801. The additional side-chain bulk from the V129M mutation may thereby push the transmembrane helix containing L798 and L802 towards the sodium-potassium channel and alter channel function.

It should be noted that the Platypus variant calling algorithm used performs local assembly to identify candidate variation, and the authors of this software report that this algorithm is sensitive to deletions up to one kilobase and insertions up to several hundred bases\(^{56}\). Thus, for the capture region of the exome performed and within the resolution limits of the CMA and exome variant calling performed, the patient was deemed to have a normal copy number.

**Discussion primer**

The proband in this case presented with childhood-onset schizophrenia with suspected genetic abnormalities. Trio whole-exome sequencing identified a single high-confidence *de novo* variant of strong predicted impact in the *ATP1A3* gene, which codes for isoform 3 of the alpha subunit of the Na+/K+-ATPase (NKA) complex. Figure 1 shows the location of this patient’s mutation in relation to previously observed disease-associated mutations in this protein, as well
as known ion-binding sites. Using the ACMG guidelines for variant interpretation, this variant meets the criteria for “likely pathogenic,” based on the presence of de novo data, multiple lines of computational evidence, population evidence on the prevalence of disease-associated missense variation in this gene, and the proximity of the variant to a well-established functional domain for the protein. The variant has never been seen in healthy controls, and the gene is highly depleted for functional variation compared to a model of expected background variation (ExAC constraint z-score 7.38), and relative to common variation in comparable genes (RVIS percentile score 3.37%), and missense variation in the gene is previously well-documented to be associated with disease. Furthermore, the variant was confirmed to be de novo, and no one else in this family is known to have a psychotic phenotype. Additionally, the variant is predicted to be deleterious by both SIFT (SIFT score 0) and PolyPhen-2 algorithms (PolyPhen-2 score 0.999). As stated, the mutation observed in this patient meets the ACMG criteria for “likely pathogenic,” and given the observed association with the patient’s phenotype as well as supporting literature discussed below, we investigate the plausibility of a connection between the mutation and the phenotype. However, it should be noted that the evidence presented here is not sufficient to make a strong causal link between this patient’s mutation and his phenotype.

The ATP1A3 protein complex plays a key role in establishing the resting membrane voltage in neurons and other electrically active cells by primary active transport of 3 sodium ions out of the cell and 2 potassium ions into the cell. There are four genes encoding this alpha subunit (ATP1A1, ATP1A2, ATP1A3, ATP1A4), which is the major subunit responsible for ATP hydrolysis and for ion binding. The alpha-3 gene is neuron-specific and primarily expressed in central nervous system (CNS) neurons, notably in GABAergic projection neurons of the basal ganglia. Mutations in this subunit of the protein complex have been documented in
association with a spectrum of psychological, psychomotor, and neuromuscular abnormalities\textsuperscript{61–64}. Therefore, it is valuable to review the documented breadth of phenotypic variation associated with mutations found in this protein.

**Discussion: Motor phenotypes of $ATP1A3$ mutations**

Alternating Hemiplegia of Childhood (AHC)\textsuperscript{15,65} is characterized by paroxysmal eye movements with onset in the first few months of life, and paroxysmal focal dystonia and flaccid alternating hemiplegia with onset before 18 months\textsuperscript{66}. Symptoms of AHC are often episodic, and may be triggered by physical stressors such as extreme change in body temperature, fatigue and infection, as well as psychological or emotional stressors\textsuperscript{61}. There is some suggestion that episodes may be alleviated by sleep\textsuperscript{67}. Panagiotakaki et al. have recently reported that a cumulative total of 132 out of 155 observed patients with AHC exhibit likely causal \textit{de novo} $ATP1A3$ mutations\textsuperscript{67}.

Rapid-onset Dystonia Parkinsonism (RDP)\textsuperscript{61,68} is diagnosed by bulbar weakness with dysphagia and dysarthria, dystonia, and Parkinsonism. Typical onset is during adolescence or adulthood in response to a physiological stressor\textsuperscript{63,69}. A variety of triggers have been reported to produce symptom onset, including acute alcohol consumption\textsuperscript{70}, extreme temperature change, head trauma, fever, strenuous exercise, and psychological stressors\textsuperscript{71}. Brashear et al. reported that 36 out of a group of 49 individuals referred with symptoms of RDP had likely causal mutations in $ATP1A3$\textsuperscript{61}.

A third clinical disorder recently associated with mutations in $ATP1A3$ is CAPOS syndrome. Diagnostic criteria for CAPOS include the constellation described within the name - cerebellar ataxia, areflexia, pes cavus, optic atrophy and sensorineural hearing loss - as well as abnormal eye movements. Patients were described broadly to exhibit episodic ataxic
encephalopathy, commonly triggered by fever\textsuperscript{72}. CAPOS syndrome has been observed in only 10 patients from 3 unrelated families to date\textsuperscript{73}.

Importantly, these past descriptions all show a causal segregation of mutations in \textit{ATP1A3} with syndromic motor phenotypes. Remarkably, to date, the proband in this study has not experienced motor disturbances resembling any of these previously described phenotypes. While the proband is older than the typical age of onset for AHC, he is still younger than the typical age of onset for RDP, and will continue to be monitored for the possible appearance of motor symptoms. The proband’s young age and psychosis precludes his ability to provide accurate history needed to determine whether his psychotic symptoms have episodic characteristics as is seen with motor phenotypes in AHC.

\textbf{Discussion: Psychiatric phenotypes of mutations in gene family \textit{ATP1A}}

Mutations in \textit{ATP1A3} and in the larger \textit{ATP1A} gene family have also been linked with psychiatric symptoms. Brashear et al. compared psychiatric histories for a cohort of 26 patients manifesting motor symptoms due to \textit{ATP1A3} mutations and 27 non-carrier control family members. They found that the affected group had approximately two times greater incidence of bipolar disorder and depression than the control group (13 of 26 affected patients also had mood disorder compared to 6 of 27 controls), and a substantial incidence of psychosis where the control group had none (5 of 26 affected patients compared to 0 of 27 controls)\textsuperscript{72}. This represents a dramatically increased incidence of psychiatric symptoms in association with mutations in \textit{ATP1A3}. Furthermore, Goldstein et al. found that when comparing 126 subjects with bipolar disorder to their unaffected family members, the presence of single nucleotide polymorphisms (SNPs) in \textit{ATP1A1}, \textit{ATP1A2}, and \textit{ATP1A3} correlated significantly with disease state\textsuperscript{71}.
Mutations in the functionally homologous alpha-2 isoform (ATP1A2) have are associated with type 2 Familial Hemiplegic Migraine (FHM2) and Benign Familial Infantile Epilepsy (BFIE)\textsuperscript{74–77}. These disorders have motor phenotypes similar to AHC and RDP\textsuperscript{78}, and FHM2 may additionally be associated with psychotic symptoms in the form of psychotic migraine auras, which may last for days and consist of complex delusions\textsuperscript{79,80}. Thus it is plausible to consider that variation in ATP1A3 may also present with similar psychiatric symptoms.

**Discussion: Evidence from tissue, and from immortalized and primary cell lines**

There is evidence for changes in the level of ATP1A3 protein in post-mortem samples of auditory cortex from schizophrenia patients compared to controls. Liquid-chromatography/mass-spectrometry analysis showed a significant decrease in ATP1A3 protein level, specifically in subjects with a history of auditory hallucinations. This alteration is likely not due to antipsychotic treatment, since chronic treatment with antipsychotics in rhesus monkeys produced an increase in post-mortem protein level\textsuperscript{81}. ATP1A3 mRNA levels were also reduced in a microarray study of post-mortem samples of prefrontal cortex from schizophrenia patients who committed suicide\textsuperscript{82}.

Consistent with findings of decreased mRNA and protein in patient-derived samples, introducing RDP-associated mutant ATP1A3 into HEK293T cells results in less protein expression compared to wild type. Furthermore, the effect of the potent Na+/K+ ATPase inhibitor ouabain can be reversed in HEK cells by transfecting in mutant ATP1A3 protein that is resistant to ouabain. However, if the ouabain-resistant ATP1A3 protein also contains an RDP-associated mutation, it fails to rescue HEK cells from ouabain treatment, indicating that these mutations reduce function\textsuperscript{83}. In COS-7 cells, AHC- and RDP-associated mutations both caused a reduction in ATP1A3 enzyme function as measured by an ATPase assay, while only RDP-
associated mutations also caused a decrease in protein levels measured on western blot\textsuperscript{15}, suggesting that mutations in different regions of the protein may impair function via different mechanisms. Finally, slice cultures from heterozygous \textit{Atp1a3} knockout mice show abnormal synaptic behavior, including an increased frequency of mini-iPSC release and a decreased threshold for firing from electrical stimulation, suggesting that reductions in this protein impact neuronal physiology\textsuperscript{84}. Similar functional studies of human ATP1A3 p.V129M will be necessary to help further establish the pathogenicity of this variant.

\textbf{Discussion: Downstream affected pathways and their disease relevance}

To understand the potential impact of ATP1A3 p.V129M on this patient and to identify potential drug targets, it is useful to consider the potential downstream effects of an altered sodium gradient in affected neurons. At the cellular level, there may be impaired recovery of resting membrane potential after action potential firing, such that prolonged stimulation without recovery could result in a loss of neuronal excitability\textsuperscript{85}. In addition, downstream secondary active transporters such as the sodium calcium exchanger (NCX), are dependent upon the sodium gradient; depletion of this gradient can lead to reversal of NCX\textsuperscript{85,86} as is seen with ouabain-induced toxicity in cardiac cells\textsuperscript{87}. Ouabain-impairment also causes membrane depolarization, disruption of normal synaptic activity\textsuperscript{88}, and increased calcium influx in glutamatergic and cholinergic neurons in response to applied neurotransmitter\textsuperscript{89}. Transporters for many neurotransmitters are downstream of ATP1A3 function, including glutamate, GABA, glycine, serotonin, and dopamine\textsuperscript{90,91}. Therefore, impaired sodium gradients may result in impaired neurotransmitter clearance from the synapse\textsuperscript{92,93}. Additionally, a number of proteins have been observed to directly interact with ATP1A3, including Src kinase\textsuperscript{94}, a glycine transporter (GlyT)\textsuperscript{95}, and Agrin\textsuperscript{96}. Mutations of \textit{ATP1A3} may
produce exacerbated functional changes if these interactions are disrupted, and this may indicate some candidate loci associated with a predisposition toward this phenotype.

At the neural circuit level, expression of ATP1A3 gene product is strongest in GABAergic neurons of basal ganglia and, to some extent, dopaminergic neurons of ventral tegmental area\textsuperscript{60}, suggesting these neurons may be particularly vulnerable to functional disturbances due to mutations in the ATP1A3 gene. Consistent with this distribution, each of the described phenotypes of ATP1A3 mutations share some features with disturbances of the basal ganglia that also present with a combination of motor and psychiatric abnormalities, including Parkinson’s, Huntington’s, and Progressive Supranuclear Palsy\textsuperscript{97–99}. Furthermore, dysfunction in these circuits fit with existing understanding of the etiology of motor gating abnormalities, schizophrenia and other sensory gating abnormalities involving GABAergic neurons of basal ganglia and dopaminergic neurons of ventral tegmentum\textsuperscript{100–102}. Given evidence for high expression localized to these regions, and their past implication in relevant disease phenotypes, functional defects in this circuit may explain a portion of the phenotype observed in the proband of this study. Further research is needed to demonstrate whether cell type-specific functional impairment due to Na+/K+ ATPase mutation may explain the observed phenotypes in patients.

**Ongoing research and functional validation**

Following the identification of the variant of interest in this case, another key objective was to confirm the functional impact of this variant. Since direct study of this patient’s neural tissue is not possible, a collaborating researcher at the BCH Translational Neuroscience Center (http://www.childrenshospital.org/research-and-innovation/researchcenters/translational-neuroscience-center) was contacted to produce iPS-neurons as a surrogate tissue model for the purpose of functional study of the variant of interest. The proband and his family were consented
for a supplementary blood draw, from which peripheral blood mononuclear cells were extracted. These cells were dedifferentiated *in vitro* to create induced pluripotent stem cells using a standard lentiviral method (e.g.\(^{103}\)), and the resulting iPS cells will subsequently be differentiated into neurons which may be assessed using transcriptomic, electrophysiological, optophysiological or other techniques to look for changes in baseline characteristics, such as resting membrane potential, ability to sustain spike trains, frequency or magnitude of miniature synaptic vesicle release, or postsynaptic potential characteristics\(^{39,104}\). Since this work is ongoing, the results of this functional study are not yet available.

**Detailed sequencing methods**

Upon enrollment in the Manton Center for Orphan Disease Research, a standard assessment was performed, which documented physical features and recorded medical, developmental, psychiatric, and family history, supplemented by medical records.

The proband underwent chromosomal microarray analysis (CMA) of peripheral blood lymphocytes (Claritas Genomics, Cambridge MA). Briefly, the patient was assessed using an Agilent custom 4x180k CGH+SNP chip with a standard Agilent protocol. CNV analysis was performed using Agilent CytoGenomics (v3.0.0.27). The thresholds used for sample quality required derivative log\(_2\) ratio spread less than 0.2 and SNP call rate greater than 0.85, and at least 5 consecutive oligonucleotide probes were required to support a putative copy number alteration (Claritas Genomics, Cambridge, MA). It should be noted that CMA is a coarse-grained assay and not intended to identify copy number alterations at the single nucleotide or single exon scale.

WES was provided by Yale University Centers for Mendelian Genomics on an Illumina HiSeq 2000 instrument with blood samples pooled 6 per lane, using 74 base pair paired-end sequencing. Libraries (TruSeq DNA v2 Sample Preparation kit; Illumina, San Diego, CA) and
whole exome capture (EZ Exome 2.0, Roche) were performed according to manufacturer protocols. FASTQs were filtered, aligned, and variants were called, filtered and annotated by Codified Genomics (Houston, TX). Briefly, reads were aligned to UCSC’s hg19 reference genome using BWA-MEM (v0.7.5-a). BAMs were duplicate marked, realigned, and recalibrated with GATK v.2.5.2 and Picard Tools v.1.38 (http://picard.sourceforge.net). Variants were identified in the proband and parents using pooled variant calling with Platypus using default parameters; the resulting variants were processed to obtain de novo variants as described previously. Analysis showed that sequencing in the proband achieved an average of 61.9-fold coverage; for the 44.1 MB target region in the exome capture kit used, 96.3% of sites were covered at least 10-fold in the proband, and 89.8% of sites were covered at least 20-fold in the proband. Coverage information for the trio is provided in Table 2. Sanger confirmation of the candidate variant was performed at Boston Children’s Hospital Manton Center Gene Discovery Core. Sanger sequencing data was viewed using Geneious (v8.1.4). Protein models were constructed using the Phyre2 web portal, and viewed using the UCSF Chimera package. The variant identified in the patient was also modeled using Pymol (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.). Multiple sequence alignment was performed using Clustal Omega via EMBL-EBI.

WES data for this case has been submitted to dbGaP (accession number phs000744.v4) and ClinVar (SCV000267646).
Table 1. Example Case Variant table.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>HGVS DNA</th>
<th>HGVS protein</th>
<th>Variant Type</th>
<th>Variant Allele Fraction</th>
<th>SIFT Score</th>
<th>PolyPhen-2 score</th>
<th>Genotype</th>
<th>ExAC MAF</th>
<th>ExAC constraint z-score</th>
<th>RVIS Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP1A3</td>
<td>19</td>
<td>c.385G&gt;A</td>
<td>p.V129M</td>
<td>SNV</td>
<td>42% of 40 reads</td>
<td>0</td>
<td>0.999</td>
<td>Het</td>
<td>0%</td>
<td>7.38</td>
<td>3.37</td>
</tr>
</tbody>
</table>

Table 1: ATP1A3 reference sequence = NM_152296.4, ENST00000302102. HGVS = Human Genome Variation Society. ExAC = Exome Aggregation Consortium (http://exac.broadinstitute.org). RVIS - Residual Variation Intolerance Score. SNV= Single Nucleotide Variant. MAF = Minor Allele Frequency. The ExAC constraint z-score from Broad Institute’s ExAC browser compares the expected frequency of functional variation to the observed frequency, where a large positive z-score indicates a gene significantly depleted for variation\(^{31}\), (http://exac.broadinstitute.org/faq). The Residual Variation Intolerance Score (RVIS) indicates the percentile of variation intolerance, where lower percentiles are more intolerant\(^{32}\), (http://genic-intolerance.org).

Table 2. Example Case Sequencing Coverage Information

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number of Reads (Millions)</th>
<th>Mean Coverage</th>
<th>% Unmapped Reads</th>
<th>% target region &gt;10x</th>
<th>% target region &gt;20x</th>
<th>Coverage at ATP1A3 c.385</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proband</td>
<td>68.4</td>
<td>61.9</td>
<td>0.16%</td>
<td>96.30%</td>
<td>89.80%</td>
<td>40 reads</td>
</tr>
<tr>
<td>Mother</td>
<td>57.7</td>
<td>51.5</td>
<td>0.90%</td>
<td>94.90%</td>
<td>84.90%</td>
<td>21 reads</td>
</tr>
<tr>
<td>Father</td>
<td>65.1</td>
<td>59.3</td>
<td>0.15%</td>
<td>95.90%</td>
<td>88.60%</td>
<td>49 reads</td>
</tr>
</tbody>
</table>

Table 2: Sequencing coverage information for the proband and parents. The target region comprises 44.1MB as defined by the EZ Exome 2.0 capture kit used for sequencing. The coverage at the site of the variant of interest is included to give context for the trio genotype obtained from exome sequencing at that site.
Figure 2: Patient Genomics Roadmap. This diagram is intended to highlight some of the key steps for successful genetic workup; these steps are discussed in detail through the following sections. 1) The case workflow begins when patients are identified with indications for possible genetic disease, and are phenotyped before being referred for sequencing. 2) A patient tissue sample is processed on a next-generation sequencing platform to produce raw read data. 3) Raw data must then be analyzed using standard processing pipelines to produce a list of candidate variants and to describe these data with contextual quality information. 4) Candidate variants identified during this analysis must be annotated with evidence from many levels to provide sufficient context to understand possible disease relevance. 5) Once variants have been thoroughly annotated, they must be interpreted in the context of the case by considering these
levels of annotation, as well as patient phenotype data or family history. 6) Variants of likely disease relevance should be confirmed, and those variants with known clinical implications may result in follow-up action such as treatment modification or clinical trial enrollment, while novel or inconclusive variants may result in research-level follow-up such as patient-derived cell modeling.

The workup of patient 0515-01 in the case report above is an instance of a more general procedure, and followed only one of the potential paths through this diagnostic and patient-care roadmap. Figure 2 shows a high-level overview of the steps involved in a generalized workup, starting from indications for referring a patient for sequencing and ending with some of the follow-up action that may help translate findings into benefits for the patient. The next portion of this discussion will broaden from this specific patient to a walkthrough of a general case with the goal of clarifying the objectives and pitfalls for each portion of the workflow. Furthermore, it is useful to emphasize some of the components that were either missing from that case, or are currently under development.

The stages of this roadmap are as follows. First, patient must be phenotyped in the clinic. The purpose of this stage is to determine whether or not sequencing may benefit a given patient and to describe that patient’s phenotype for the purpose of interpreting downstream results of sequencing. Second, a tissue sample is collected and raw sequencing data is generated. Third, this raw sequencing data is processed to calculate the patient genotype and identify sites of variation. Fourth, these variant sites are annotated with descriptions such as their predicted effect on known gene models and the frequency of that variant genotype in a reference population. Fifth, these annotated variants are considered in the context of the case history and patient
phenotype to identify variants that are likely to be disease-associated. Lastly, high-quality variants of interest are confirmed, and may either result in clinical action, or may necessitate further exploratory research.

**Roadmap 1. Clinical Phenotyping and Indications for Sequencing**

**Overview and indications for sequencing**

For patient 0515-01, there were several aspects of his initial presentation that supported his referral for sequencing study. His phenotype did not correlate with any known environmental factors and plausible environment contributors such as trauma had been ruled out. His presentation was both severe and early relative to other patients with comparable phenotypes, indicating the plausibility of recessive or *de novo* genetic disease.

In general, good candidates for genetic study include with perinatal patients with early-onset syndromic illness without known environmental causes, patients with a family history of inherited disease, or patients otherwise on a ‘diagnostic odyssey’ without other options. Pre-conception or prenatal genetic study and counseling may also be warranted for individuals with known or suspected genetic risk factors, such as past environmental mutagen exposure, advanced age, past fertility issues, documented family histories of genetic illness, or ethnic risk factors for genetic illness\(^{108,109}\).

**Family history and phenotyping**

Obtaining phenotypic characterizations of the patient and available relatives upstream of genomic analysis is crucial for successfully identifying potential signals of disease amidst a large
and noisy genomics data set. Relevant family histories of disease can help inform the suspected mode of inheritance of any candidate disease genotypes. In addition, given a positive family history of relevant illness, it may be possible to measure (by genotyping information across the family) or estimate the penetrance of the disease genotype. A negative family history of relevant illness may indicate that a genetic lesion arose de novo. Phenotypic information collected from the individual under study should provide thorough detail about clinically apparent features of the disease and the details of relevant family disease history. ‘Relevant’ is the crucial word here; since the extent of pathways disrupted in a patient may not be clear, this set of pathways cannot necessarily be used to crystallize phenotype data from the family history.

Ideally, phenotypic information should be collected at all available levels, from behavioral observation, to biopsy study, lab testing, and structural imaging. Collating the resulting trove of EMR data in a useful format poses another challenge; one part of the solution is to provide phenotype information in a standardized format that can be parsed by machine, such as Human Phenotype Ontology (HPO). The translation from natural language phenotype descriptors to standardized phenotype terminology may be performed as a software-assisted manual process, or via natural language techniques to parse hand-written or free-text notes about clinical observations and case histories. Other software platforms enable selection of disease-specific phenotypes and diagnostic terms from a rich case description, or parsing and statistical interrogation of EMR data in real time. As an additional proactive step to support future patient-focused disease modeling, clinicians may seek to derive representative cell lines.
Roadmap 2. Data Generation and Primary Analysis

Methods

Preparing for sequencing

After identifying a patient as a good candidate for sequencing and collecting phenotype information from the individual and available family members, the next phase of genetic workup is generating raw DNA sequence data. In current practice, as a first pass method before using genome-scale sequencing methods, it may be worthwhile in some cases to begin with a more narrowly focused test for a region of interest or for a particular type of expected variation. There are several factors that would favor the use of targeted testing. First, there must be a defined region of interest, such as a distinct phenotype whose heritability has been comprehensively mapped to a small set of candidate genes. Given availability of a test and assurance that it will definitively cover all loci of potential genetic relevance, the available resources for targeted testing must provide data of equal or greater sensitivity and specificity compared to genome-scale sequencing. Lastly, the issue of cost must be considered, which is a function both of the cost and price of sequencing, and the status of insurance coverage for different types of genetic test. As a peripheral consideration, comprehensive genotyping may provide future utility to the patient, family of the patient, and other future patients in the disease cohort. It is worth noting that the rapid ongoing developments in high-throughput sequencing technologies will affect each point in this consideration.
Chromosomal microarray

Chromosomal microarray ("CMA"), also known as array comparative genomic hybridization ("aCGH") can assay for copy number variation with a resolution in the kilobase to megabase range\(^\text{119}\). This technique uses a microarray with clusters of DNA probes, and compares the amount of hybridization of two differently labeled samples – one control sample and one test sample. If, at a certain locus, the test sample’s copy number is greater than that of the control sample, this will be apparent by an abundance of test sample binding to the corresponding probe(s), and vice-versa if the test sample has lower copy number at that locus. Depending on the density and spread of probes across the target region, chromosomal microarray can detect copy number changes in the kilobase to megabase range. This technique is reviewed in greater detail elsewhere\(^\text{120,121}\).

Exome and genome sequencing chemistry

The basic strategy in genome-scale sequencing is to perform massively-parallel sequencing-by-synthesis of a set of fragments spanning the subject’s DNA, reconstruct these sequenced fragments into a representation of the subject’s whole genotype, identify candidate sites of variation relative to a control sequence, and then search among these candidates for potentially disease-associated variants. This process is briefly reviewed here, since the details of sequencing provide a crucial context for interpreting candidate variants. The process is well-reviewed in greater depth elsewhere\(^\text{122–126}\).

The first stage of the process, up to and including the measurement of raw sequencing reads, is referred to as “primary analysis” or “primary pipeline processing.” The process from raw tissue to fragments ready for measurement on a sequencing machine is referred to as “library preparation.” DNA is first extracted from peripheral blood or saliva, though one may consider a
more disease-relevant tissue source if there is a reason to suspect there may be somatic mosaicism relevant to the disease under study, such as in cancer sequencing. Once a tissue source has been selected, DNA is extracted and purified, and may then be amplified if further samples are unavailable and there insufficient DNA for sequencing – with modern techniques, even a single cell may provide sufficient starting material to obtain a reasonably accurate genotype measurement\textsuperscript{127}. Next, sequencing primers are ligated onto fragments to facilitate DNA polymerase binding during sequencing. For exome sequencing, the pool of sequence fragments is restricted to the coding portions of the genome by either hybridization probe capture or selective amplification\textsuperscript{128}. Using unbiased methods for library preparation, such as for sample amplification, fragmentation, and exome capture, can help produce data with even coverage, whereas targeted fragmentation, uneven amplification, or biased capture methods may result in uneven coverage or gaps in coverage\textsuperscript{129}. For pooled sequencing, library preparation also involves uniquely barcoding individual samples such that data from multiple subjects may be sequenced together and then processed separately.

Next, there are a variety of technologies available to perform the sequencing measurements. These methods differ in their performance characteristics, and also differ in cost per base pair sequenced, time required per sample, and maximum throughput or capacity for processing simultaneous samples\textsuperscript{130}. This discussion will focus on just two technologies that are currently in widespread use, namely Illumina chain-terminating sequencing (http://www.illumina.com) and Life Technologies ion semiconductor sequencing\textsuperscript{131} (https://www.thermofisher.com/us/en/home/brands/ion-torrent.html).

Illumina sequencing uses fluorescently labeled chain-terminating nucleotides to allow single base incorporation. First, diluted, primer-ligated, barcoded sequence fragments are washed
onto the sequencing surface. After annealing to the surface, they are amplified into clonal fragment clusters via bridge amplification. Sequencing of these fragment clusters then proceeds via cyclic application of a single type of nucleotide, fluorescence measurement to detect which clusters underwent nucleotide incorporation, and then removal of the reversible terminator. The sequence of incorporations for each location on the sequencing chip is converted to raw output data that describes the location of the read cluster, the measured base, and a quality score associated with the confidence of the machine’s assignment of a particular label\(^\text{132}\).

The Life Technologies Ion Torrent sequencing platform uses emulsion PCR to amplify sequence fragments into a library of beads, each containing many copies of one fragment. On the sequencing surface, each well is filled with one clonal bead and a reaction mixture containing DNA polymerase. These wells are connected to an ion-sensitive field-effect transistor, whose transmitted voltage varies with the pH of a measured solution. Since incorporation of a nucleotide during DNA strand elongation involves the concomitant release of a proton, the pH in the small volume of reaction mixture will change measurably, and the detector translates this change in pH to a change in voltage that can be translated into output data\(^\text{132}\).

The above description for both technologies covers the generation of a single directional measurement read per original DNA fragment, which is known as single-end sequencing. The accuracy and sensitivity of downstream processing can be greatly increased by using paired-end sequencing, in which reads are measured from both ends of original sequence fragments. Since the accuracy of measurement varies at different positions in a read, this parallax measurement gives additional confidence in the sequence. Furthermore, in protocols where the size of the fragments and the relative position of pairs of reads can be controlled, paired-end sequencing also provides additional power to resolve structural variation in the genome of the subject\(^\text{133}\).
Both the Illumina and Proton methods are constrained in the length of read that can be accurately sequenced. Sequencing an individual DNA fragment would produce a signal too small to be detected by the high-throughput detection mechanisms, and thus it is necessary to measure sequence from clonal read clusters so that the signal will increase additively. However, this also means that the sequencing process requires each cluster to proceed in sync. Even with high fidelity DNA polymerase enzyme and well-controlled reaction conditions, there will always be a probability of errors such as side reactions, mismatch nucleotide incorporation, failed incorporation, or failure to unblock a nucleotide in the case of Illumina. Therefore, as sequencing progresses, read clusters will become out-of-sync, and the accuracy of base calling will decrease as a function of position in the read.

Since Ion Torrent sequencing does not use blocked nucleotides, regions with homopolymer runs will undergo incorporation of multiple simultaneous nucleotides. However, since the raw signal does not scale linearly with the number of nucleotides incorporated at a single position (due in part to the loss of synchronous measurement mentioned above), measurement of these regions is more prone to error\textsuperscript{134}.

**Sanger sequencing**

Based on the range of uncertainties and errors that may arise at the level of raw data generation, candidate variants must be confirmed by independent re-measurement with another platform, often using Sanger sequencing\textsuperscript{57}. Sanger sequencing chemistry is also the basis of the Illumina sequencing platform, though these confirmatory experiments are performed on a much smaller scale by designing PCR primers to capture a locus of interest and sequence only that narrow region. The design of Sanger sequencing experiments is reviewed in detail elsewhere\textsuperscript{135}.
Roadmap 3. Secondary Analysis Methods

Processing sequence data:

There are a few overall stages of processing required to convert raw sequencing reads into a list of high-quality candidate variants along with an alignment map describing the underlying sequencing reads. Overall, this procedure is known as “secondary analysis” or “secondary pipeline processing.” The specific steps in this secondary analysis process depend on the underlying data source, and the workflow described here is specific to Illumina. However, other short read sequencing platforms, including Life Technologies ion semiconductor sequencing, share the general data processing steps.

Since library preparation upstream of sequencing loses information about the original location of reads in the genome, the first key step is to assign reads to a genomic location by comparing their sequence to a reference genome. Once located, the total set of read evidence at a candidate locus can be analyzed to identify candidate variants. These candidate variants can subsequently be annotated with any information that may be relevant to the variant or affected gene, and filtered by relevance criteria to identify a list of likely candidate variants. The choice of annotations to provide sufficiently powerful context for variant interpretation is a point of ongoing research, and is further discussed below in the section on tertiary analysis methods.

The Broad Institute curates a ‘best practices’ workflow that is the industry standard for high-quality data processing, and they provide operating instructions and troubleshooting help for each step of this process. The bulk of this workflow is performed using their Genome Analysis Toolkit (GATK) software package. As with the technical details of producing raw sequence data, making well-informed interpretations of candidate variants and genotypes,
requires an understanding of the processing steps involved in producing these candidate results. Therefore, this portion of the discussion will walk through an abridged version of the processing steps outlined in the GATK best practices (shown in full in figure 3).

Figure 3: The Broad Institute’s Genome Analysis Toolkit\textsuperscript{21} (GATK) Best Practices workflow, reprinted from https://www.broadinstitute.org/gatk/guide/best-practices.php. This workflow is considered an accepted industry standard for processing sequence data produced on Illumina sequencing platforms. This workflow begins from raw sequencing reads in FASTQ format, and produces aligned reads in SAM/BAM format and variants with initial annotations in VCF format. The GATK documentation provides thorough instructions for each step of this process.

**Raw sequencing data**

The process of converting raw physical sequence measurements into calculated bases described with a statistical confidence is described as “base-calling.” This process produces raw
sequencing reads that are the input for secondary analysis. Each sequencing run on an Illumina machine produces reads pooled from multiple samples in BCL format, which can be demultiplexed into a FASTQ-format file\(^2\) for each individual sample before further processing.

Within reads in a FASTQ file, each position contains a nucleotide, and a phred-scaled quality score, which indicates the probability of an erroneous measurement at that position. This quality score informs downstream statistical operations on the data, but is agnostic to the underlying sequencing method. Reads may also be labeled by the location within the sequencing machine where they were produced for the purpose of visualizing errors due to technical defects in one region of a sequencing machine. For paired-end sequencing, read pairs are typically separated and stored in adjacent files such that they can be directly collated to bring read pairs together. Figure 4 shows an example of a single read from a FASTQ.

Figure 4: Example of an individual read in a FASTQ file. The FASTQ format contains a sequence identifier describing the machine and location within the machine where a read was produced, a nucleotide sequence, an optional description, and a quality score for each nucleotide described using an ASCII schema (described in detail elsewhere\(^2\)).

**Upstream quality control**

Early stage quality control can be performed on samples at this stage to identify errors at the level of an entire data set. Software packages like FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) can give such high-level statistics, both for the purpose of evaluating raw data quality before expending resources on downstream processing, and to better inform later interpretation of the results. For example, physical
problems affecting the reaction conditions during sequencing may impair the fidelity of sequence elongation or reaction cycling, resulting in uneven read length distributions. For Illumina sequencing, especially, the length of reads should be precisely controlled by the number of reaction cycles, and any deviations indicate reduced precision. Another effect of reduced reaction fidelity is an increased chance for a given cluster of sequence fragments to become out-of-sync; this is visible in the distribution of base quality scores as a function of position in a read. This effect will be visible in the overall distribution of base call qualities and the distribution of base call quality as a function of position in reads. Technical problems relating to the physical measurement apparatus may be apparent by identifying patterns in data quality as a function of spatial location on the sequencing surface. Errors in sample preparation, such as contamination with viral or microbial nucleic acids, barcode oligonucleotides, or biases in amplification, may be uncovered by tallying the number of times a given sequence is represented in the data set and looking for overrepresented sequences.

Mapping raw reads

The purpose of next-generation sequencing is to achieve high throughput and speed, though this is achieved at the expense of producing spatially scrambled data; output sequencing reads are measured without storing information about their original position in the linear sequence of genome. Therefore, the next step is to align these reads against a reference genome map. This task could be compared to reconstructing a well-shredded encyclopedia in the hopes of finding typos.

Raw reads or read pairs are typically aligned to a genomic location using a software package such as Burrows Wheeler Aligner (BWA)\textsuperscript{136} and a curated reference genome, such as one released by NCBI’s Genome Reference Consortium. The alignment results are typically
produced in Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) format, which contains information about each sequence read from the input FASTQ, plus the newly computed information about alignment characteristics, and characteristics of that read's alignment relative to the reference genome. Figure 5 shows an example of a single read from a SAM/BAM file.

Therefore, it is useful to collect another set of quality metrics describing the raw BAM using a software package such as Picard (http://picard.sourceforge.net). This is useful because quality metrics previously obtained by examining the raw read data at the FASTQ stage could not have revealed any information about the distribution of read coverage across the sequenced region. Ideally, the coverage obtained will be deep and even across the target region; shallow coverage may prevent accurate variant calling. To give an example of raw read data underlying a variant, figure 6 shows an example of aligned reads viewed as a pileup, where a candidate variant can be seen in the center of the frame.
Figure 6: An example of a pileup of aligned sequence reads, viewed using the WuXi NextCODE Sequence Miner software. Forward and reverse aligned reads are indicated in blue and orange, respectively. Positions highlighted in yellow differ from the reference sequence, shown at bottom, and a candidate variant (C>T) is clearly visible in the center of the frame; this variant looks real because there are many supporting reads that otherwise align perfectly and span the region evenly, and there is approximately even support on both strands.

At this point, the calculated alignment may reveal discrepancies between the test data and the reference genome. Read pairs may be split up and mapped to distant locations in the reference genome, potentially indicating that a deletion occurred in the test data relative to the reference. Reads may map in an incorrect configuration, potentially indicating structural rearrangements such as an inversion or tandem duplication in the test data relative to the reference. One mate of a read pair may not map well, potentially indicating an insertion in the test data relative to the reference. Reads may also be ‘clipped,’ such that only a portion of the
read is considered, in order to achieve a good alignment in cases where part of the read is of lower quality or came from a location where the test sequence differs from the reference sequence due to variation or insertion. Figure 7 shows several examples to give a sense of how structural variation events will appear in sequencing reads.

**Figure 7**: This reprinted diagram\textsuperscript{23} shows how several types of variation events will appear in read data; understanding the expected configuration of reads may help an investigator manually evaluate supporting a putative variant. Each row describes one type of variation event, and each column shows how that type of event will appear in raw data. Blue arrows indicate read pairs. For each event shown, the reference sequence is shown at bottom and the test sequence is shown at top.

**Bam refinement**

After doing the bulk of the alignment work, the resulting raw BAM can be refined to correct for several sources of error. First, any identical reads are likely to originate from identical fragments of DNA in the original library when it was prepared for sequencing, due to the huge and sparsely covered sequence space for long oligonucleotides in the human genome\textsuperscript{138}. When
aligned, these reads will appear to increase the coverage depth (and therefore the amount of potential genotype information) at a certain location, but in fact do not add information; therefore, they are identified and removed by sorting the BAM lexically and selecting only unique reads.

Additionally, while the preliminary alignment accurately place each read in the global space of the genome, it may suffer from inaccuracy at a fine-grained detail in regions of repetitive sequence, or in the region of an indel variant. Thus, realignment is done to fine-tune alignments in the region around known indels. Once the alignment of reads has been finalized, it may be possible to identify other biases in the base quality scores by finding correlations between base quality scores and parameters such as position in a read or sequence context. The phred-scaled base quality score is meant to indicate the probability that an error was made in measuring the base, and this probability will be used when calculating genotypes to identify variation, and when reporting the probability of error in putative variants called at that position – thus it is beneficial to ensure that this reported base quality score is as accurate as possible. If the reported base quality scores show a correlation with such parameters, this may indicate biases that can be reduced through score adjustment.

**Variant calling**

The next step in the process is to assign a genotype to the individual being measured, and identify candidate sites of variation; this process is described as “variant calling.” For each position in the sequenced region, a variant calling algorithm considers the evidence available from reads spanning that position.

Recall that these reads contain information about the quality of their constituent bases, as well as information about their alignment in that region. Variant calling algorithms may also
consider cues such as the proportion of reads on each strand at the candidate region, the position of the putative variant in each read, or many other characteristics to contextualize the available evidence. Additionally, algorithms may consider the local sequence context as a factor in variant calling; local sequence repetition or high homology to other locations may impair read alignment and result in errors during variant calling.\textsuperscript{139}

The output of this process are usually stored using the Variant Call Format (VCF)\textsuperscript{140}; this format is quite flexible and facilitates description of many attributes about the candidate variants to give context downstream. Additionally, at this stage, investigators may choose to describe the calculated genotype at all measured positions in a patient, to describe genotype at a set of positions relevant to the cohort they are processed with, or to describe only the positions where a variant was identified in the patient or group of patients currently being sequenced.

**Variant callset refinement**

This initial list of candidate variant sites can then be refined further. One important goal of this refinement is to provide users the flexibility to define their own balance of sensitivity and specificity. The GATK software achieves this flexibility by classifying variants into tranches with a range of sensitivity/specificity thresholds. Grouping variants into analysis tranches relies on a training set of known common variants that are likely to be found in the test data set. First, the algorithm identifies which of these common (and therefore, likely trustworthy) variants were observed in the individual. It then re-attempts variant calling while modulating the evidence parameters being considered. As it modulates the evidence parameters, the fraction of this trustworthy set that are successfully called shows how sensitive the variant calling performance is to these “true” variants, and therefore how sensitive the variant calling is likely to be for any other novel variants. At this point, the algorithm can report variants in “tranches”; for example, if
a variant appears in a 0.90 tranche, the calling parameters required to identify that variant are also sensitive enough to capture 90% of all “trustworthy” training variants. That variant from the 0.90 tranche is more likely to be real than a variant in a more inclusive tranche such as 0.99; at the more inclusive 0.99 level, the variant calling parameters have been tuned to be more sensitive, and false-positive variants will included more easily in that tranche.

By analogy, consider how a musician may perform audio equalization in a recording studio. Suppose the musician seeks to play a new piece of sheet music and scrutinize the recording for notes that stand out, just as the geneticist seeks to read a target sequence and search for variants that stand out. For the musician, the specific acoustics of the recording room may be unfamiliar and unique to that particular day and particular studio, just as the specific characteristics of a genetic data set are somewhat unique to the sample, sequencing machine, and data processing steps. Therefore, in order to get the best possible recording of a new song, the musician may first play a very familiar song in the room and equalize the audio until that familiar “training” song is recreated appropriately. At that point, knowing how the equipment processes that familiar song, the musician can form an expectation of how the equipment will process a new song.

**Downstream QC**

Secondary analysis is completed at this point, and it may be useful to collect high-level summary statistics describing the resulting dataset to give context for future interpretation of the data. This analysis may be done as validation for the setup of a particular processing pipeline, where the results of a standard sample such as the NIST Genome-In-A-Bottle sample NA12878\(^1\) may be compared to the known results derived from other sources.
This quality control may be performed for an individual sample by examining high-level statistics such as the distribution of coverage depth across the whole genome or a region of interest. Additionally, the extent of variation in a given individual can be compared to the expected rate of variation by examining the number of SNV and indel variants observed per base pair sequenced. The accuracy of genotyping may also be measured by comparing the observed and expected percentage of guanine-cytosine content, or Ti:Tv – the ratio of transition variation (SNVs within nucleotide classes, i.e. from a purine to a purine, or from a pyrimidine to a pyrimidine) to transversion variation (SNVs across nucleotide classes). No single one of these 1-dimensional statistics is sufficient to broadly describe the quality of such a complex dataset, but they may be helpful for identifying large deviations in quality.

**Scalable infrastructure**

The previous sections overviewed the complexity of the processing required to make use of shotgun sequencing data. Another key dimension of this work is implementing these processing steps in an efficient, scalable way that is also flexible enough to incorporate growing numbers of reference data sets, variant and sequence annotation algorithms, and high-dimensional phenotypic data sets. The choice of hardware architecture affects the cost, processing time, and accessibility of clinical genomics data, and sets the scale of possible investigation in research genomics applications.

These issues are extremely complex – they will not be solved here and will likely continue to evolve in years to come alongside changing cloud computing paradigms. There are a large number of infrastructure vendors working in this domain, with offerings varying from pure hardware support\textsuperscript{142}, to workflow management (https://www.dnanexus.com), novel data formats\textsuperscript{143}, and even full-stack solutions (https://cloud.google.com/genomics/).
Roadmap 4. Tertiary Analysis Methods

Annotating sequence data to support case interpretation

After generating a list of candidate variants, the next step is to annotate and filter the list of results in order to identify variants that are likely to be pathogenic. This procedure is broadly referred to as “tertiary analysis.” Whereas up to this point the variant call set has been refined based on sampleset-intrinsic measures, the majority of information to be added for tertiary analysis comes from a wide range of external sources, such as population reference databases, gene-centric statistics about the likelihood of disease association, and past research on animal research models. After applying these annotations, variants can be pruned by setting thresholds for variants of interest. ACMG defines some thresholds for classifying the pathogenicity of variants according to the strength of available supporting evidence\(^{57}\). In order to complement their framework for variant interpretation, it is useful to discuss the categories of evidence that may be available to annotate a variant.
Figure 8: Levels of Evidence. A schematic to overview some of the levels of evidence that can give context to candidate variants during case interpretation, shown here for a variant in a hypothetical transmembrane channel. A candidate variant is shown in the underlying raw read data, overlaid on a gene model, and shown on the cDNA for a single splice isoform. Based on the gene model, this variant can be described with a predicted amino acid consequence. This consequence can be examined in the linear amino acid sequence, a model of known or predicted protein functional domains, or in a model of the folded protein. This affected protein and the expected effect of the variant can be described within a relevant pathway, or functional network.
such as known protein-protein interactions. Finally, past disruptions of this protein, pathway, or variant can be examined in a relevant reference population.

One approach to itemizing the variety of evidence available for interpreting candidate variants is to categorize them into coarse levels of perspective, from the narrow technical features of a genotype at single-nucleotide resolution, up to broad statistical observations such as the degree of gene-level sensitivity to functional variation in a control population. Figure 8 gives an overview of these levels of evidence to emphasize that these levels of perspective cover a wide range of scales, and therefore a large set of evidence is necessary to give proper context about the expected impact of a candidate variant. For each of these levels, one can consider a dichotomy between positive evidence supporting a previous association with disease, and negative evidence supporting the absence of the candidate variation in healthy conditions.

**Technical level evidence**

Starting with fine-grained details at single-nucleotide resolution provides an opportunity to review the technical aspects of the sequencing workflow that identified a given candidate variant. Tersely, a variant that is likely to be real is one that was called with high stringency, from high quality, well aligned reads, in a proportion appropriate for the designated mode of inheritance (e.g. ~50% for heterozygous variants). The originating sample should have been well prepared, and sequenced using an industry-accepted sequencing technology. Even with the sophisticated statistical approaches taken to align and realign sequencing reads, it is useful to examine the underlying read data by eye to confirm that the evidence underlying a given variant call looks reasonable. Furthermore, this can give an opportunity to examine the local sequence context to understand whether this region may have been problematic for data processing.
Recall that there are a number of high-level heuristics to estimate the noise in a reported variant or the sample set it came from, such as the sensitivity tranche in which a variant was called, the global frequency of SNVs and indels, the global distribution and evenness of coverage, measured GC content, and Ti:Tv ratio.

**Variant level: Empirical evidence**

Still at single-nucleotide resolution, it is also important to consider the variant itself. Conservation across evolutionary time among homologous genes or functional domains can provide an approximate surrogate measure of functional importance; a nucleotide which is crucial to function is likely to be more well conserved, and variation in that position may be more likely deleterious to the fitness of the organism overall.

Similarly, the conservation of the affected nucleotide and the severity of possible deleterious effects due to mutation at the position can be estimated by measuring the minor allele frequency (MAF) of the candidate variant nucleotide(s) in a healthy control population; a variant that is strongly linked to rare disease should also be rare in healthy populations.

Another empirical source of evidence for the disease-relevance of variants are hand-curated databases of known disease variants, such as ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), ClinGen, OMIM, and HGMD. It is important to note that while these resources provide gold-standard information on known relationships between variants and disease, these sources may not provide information relevant for novel disease genotypes, and may be limited in their coverage based on the relatively high cost and low throughput of manual curation efforts.

In some cases, there may be an expected mechanism of mutation, in which case there may be a corresponding mutational signature that is expected, and candidate variants can be
assessed for how well they match this expected signature – for example, this has been demonstrated in cancer sequencing or in cases of known exposure to chemical mutagens\textsuperscript{147,148}.

**Variant level: Predictive evidence**

In addition to empirical evidence describing the relationship between variant genotypes and disease, there are also a number of algorithms that have been created for annotating variants with predictions of pathogenicity. While the specific set of evidence and statistical method used may differ substantially between different algorithms, it is useful to understand the categories of evidence used by these algorithms. Broadly, these algorithms may consider conservation of nucleotide or amino acid sequence. These algorithms may provide a score by comparing the native and mutant nucleotide or amino acid sequence; therefore, variants may be scored for the extent of chemical dissimilarity to the original residues. Variant severity may be scored according to local sequence context, or according to nearby gene-specific functional domains (e.g. ligand binding domains, membrane-spanning domains, or enzyme active sites). The regulatory motifs and functional domains used in these considerations may be drawn from computational predictions or from hand-curated reference sources.

Given the complexity of methods and data sources, users may seek an appropriately weighted integration of multiple well-validated algorithms. For example, dbNSFP curates a large set of algorithms for predicting pathogenicity and for describing multi-species conservation. The SNV portion of the database is available for download and covers all possible mutations across the approximately 83 million coding positions currently documented in the genome\textsuperscript{149}. Other sources discuss the interpretation of variant pathogenicity algorithms in greater detail\textsuperscript{150}.
Gene level evidence

Considering the level of the affected gene is one of the most informative contexts for evaluating protein-coding variation. Any previously observed genotype-phenotype correlations in that gene might be relevant.

Figure 9: This figure is reprinted from WuXi NextCODE documentation (goo.gl/8Mmm5T), and shows the possible descriptors that the Variant Effect Predictor algorithm\(^{24}\) may give when annotating the gene-level consequence of a candidate variant. This list thoroughly covers the effects that are currently assessed for single nucleotide variants and short indels.

The CCDS project\(^{151}\) aggregates predicted gene models from multiple sources so that an investigator can overlay a gene model on a candidate variant and describe the expected impact on that gene according to the local reading frame, exon boundaries, known splicing sites, and or regulatory binding sites. Typically, a variant is described by its expected effect at the codon level as ‘silent’ or ‘synonymous’ when the predicted amino acid sequence is not changed, and ‘non-synonymous’ otherwise. Within non-synonymous mutations, ‘missense’ indicates that a variant changes the predicted amino acid sequence, and ‘nonsense’ indicates that a variant introduces an
early stop codon. Variants may also disrupt the border of a canonical splice site, potentially causing intron retention in the expressed transcript. Out-of-frame insertions or deletions can cause a shift of reading frame, causing likely disruption of all downstream portions of the gene. The descriptor “loss of function” is often used to refer to the combination of nonsense, frameshift, and splice-disrupting variants. Figure 9 shows the variant effect categories described by the Variant Effect Predictor algorithm, which thoroughly covers the types of effects currently considered for single nucleotide variants and short indels.

It is important to note that there may be multiple splice isoforms for a given gene, and the exon containing a candidate variant may not be present in all isoforms. Furthermore, different transcripts may have different expression distributions, and the affected transcript may or may not be functional in a disease-relevant tissue. Therefore, when describing variants according to nucleotide change at the cDNA level, it is important to also describe the variant relative to a transcript of interest. The longest transcript is sometimes used by default\(^\text{152}\), though deeper investigation on transcript-specific expression patterns can do a better job to select the transcript with the most functionally important expression level and distribution in the tissue of interest. Resources like GTEx can help easily identify the most highly expressed transcript in a tissue of interest\(^\text{153}\).

If the gene is part of a gene family, the other members of this gene family may also be sources of relevant evidence, though one must consider how these genes overlap in structure, function, or spatiotemporal expression distribution. For example, from the example case of patient 0515-01, the affected gene, \textit{ATP1A3}, was part of a gene family \textit{ATP1A}. This gene family comprises four known ATP-dependent sodium potassium pumps, which are functionally very similar with a high degree of sequence overlap. Despite these similarities, the phenotypes of
mutations in different $ATP1A$ family members may differ substantially, since they are expressed in different tissues and at different times throughout the body; disruption of neuron-specific $ATP1A3$ may present very differently than disruption of the highly muscle-expressed $ATP1A1$ (see OMIM entry for $ATP1A1$, http://www.omim.org/entry/182310).

**Gene level: Functional variation intolerance**

Another source of descriptive evidence at the gene level comes from algorithms that statistically evaluate the observed frequency of different types of mutations in a given gene (usually synonymous, missense, or loss-of-function). Broadly, these methods seek to calculate the expected frequency of functional mutation in candidate genes, and then compare this expected frequency to the measured frequency in a reference population. Finding an abundance of functional mutations may indicate that a given gene is tolerant to mutation, and that mutations in this gene are less likely to be deleterious for an individual patient. Conversely, finding a depletion of mutations may indicate that the gene is highly sensitive to functional variation and mutations in an individual may be more likely disease-relevant. Algorithms for detecting functional variation intolerance differ in the set of background evidence and method for calculating the predicted or expected number of mutations in a given functional category of interest.

**Gene level: dN/dS**

A long-standing metric to identify regions that are sensitive to functional variation is the dN/dS ratio (also called Ka/Ks), which is the ratio of non-synonymous mutations per non-synonymous site to the ratio of synonymous mutations per synonymous site$^{154}$. This metric considers only a codon table and the predicted reading frame in the region of interest, and forms an expectation of local mutation rate based on silent mutations.
For a population of interest, the metric is computed at each region of interest (such as a gene, exon, or functional domain) by counting the total number of non-synonymous mutations observed per possible non-synonymous site and dividing by the total number of synonymous mutations observed per possible synonymous site. Thus the rate of synonymous mutation provides a baseline to evaluate the rate of functionally impactful mutation. Finding that functional mutation is relatively abundant in a region may indicate that region is under positive selection, i.e. that mutations in this region may confer increased fitness. Conversely, finding a relative depletion of functional mutation in a region may indicate the region is under purifying selection, and that the fitness of the individual is adversely affected by variation in that region\textsuperscript{155}.

**Gene level: RVIS**

Whereas dN/dS ratio uses the rate of silent mutation as a basis of comparison for the rate of functional mutation, the Residual Variation Intolerance Score created by Petrovski et al. considers only functional mutation; this method compares the rates of “common” functional variation to “uncommon” functional variation.

This method involves first calculating a regression for the number of common functional variants as a function of the total number of variant sites in a given gene, considering all available genes. Then they calculate the studentized residual of the number of common functional variants in a given gene, and designate this as the gene’s score. In this metric, an individual gene’s y-value is compared to the expected y-value based on all other genes at that x-value, providing a more reasonable and sensitive comparison than comparing to the global model. With this score, finding that a gene has an abundance of common functional variation indicates that it is likely tolerant to functional change and candidate variants in that gene may be less likely to be disease-relevant.
**Gene level: Constraint**

Like the dN/dS metric, the “constraint” metric created by Samocha et al. considers only codon context and the local reading frame when describing the expected rate of various mutation types at a given position. However, the expected rate of mutation is derived from a measurement of mutation rates over evolutionary time (from primates to humans), rather than the local rate of mutation.

To calculate this metric, they first use intergenic regions of homology between a human and primate reference sequence, and tabulate a transition table that describes, for a given nucleotide of interest, and considering the two neighboring nucleotides, how frequently that nucleotide is expected to mutate to any of the other possible nucleotides. Next, they use this transition table to form an expectation of mutation rates in different categories for genes of interest in humans. Finally, for each region of interest, they can compare the expected number of mutation events to the number of events truly measured in a large healthy reference population. Finding a significant difference between the expected rate and the rate observed in a reference population helps reveal the influence of other functional constraints on changes in that gene.

**Protein level evidence**

Another important perspective for evaluating a candidate variant is the structure and function of the affected protein. One simple proxy measure that may cover the entire connection between amino acid sequence and health or evolutionary fitness at the organism level is to examine the evolutionary conservation of the amino acid sequence around a candidate variant. As discussed previously, this property may be included in black-box algorithms for predicting variant pathogenicity. Considering the disease context, a protein-altering variant that is disease
relevant should be expressed in an appropriate tissue at a suitable developmental time-point, and the candidate variant should substantially disrupt the function of that protein.

**Protein level: Amino acid link to protein structure and function**

Though it is difficult to conclusively determine whether a variant is likely to have a functional impact without empirical testing, there are both predicted and empirical sets evidence that may describe the relationship amongst the nucleotide sequence of the affected gene, its linear amino acid sequence, the 3-dimensional structure of the resulting protein, and that protein’s interactions with the local environment, substrates, or protein interacting partners.

A very rough impression can be formed by simply comparing the side chain groups of the original and variant amino acids for differences in size and charge. This comparison can then be informed by understanding how this residue fits into its local environment and secondary structures in the protein\textsuperscript{157}.

Using nucleotide sequence information, for example, one could identify potential membrane-spanning domains\textsuperscript{158}; a candidate variant in a lone membrane-spanning domain could destabilize this domain by adding a charged residue to a region that should be hydrophobic\textsuperscript{159}. However, the naïve requirement for a generic membrane-spanning domain to be hydrophobic should be further scrutinized in consideration of the entire protein fold and its environment. In an ion channel, for example, while the transmembrane residues facing the lipid bilayer should indeed be hydrophobic, the interior of the channel pore may be natively charged. Thus, for mutations on an inward-facing residue, changes in volume may be more important to the protein function than changes in charge. Similarly, even an outward-facing residue may, in fact, lie at the interaction surface between two membrane-spanning proteins, and if this protein-protein interface is stabilized by charge-charge interactions, then the addition of a charged residue may
again be tolerated. To emphasize the importance of protein context for variant interpretation, Figure 10 shows an example of a variant that adds a charged residue in several scenarios. Note that a holistic picture of protein-level effect requires a rich set of information about the residues and structures required for the protein’s function, its native environment, and its interaction partners.

![Diagram showing different scenarios of charge addition to transmembrane domains](image)

**Figure 10:** Adding charge to a transmembrane domain. This example is meant to convey the importance of considering environmental context when evaluating the protein-level effect of a variant. Whereas adding a charged residue to a lone transmembrane (left) may be strongly destabilizing to the protein, adding charge may be tolerated in the core of a transmembrane ion channel (middle) or at a charged interface between two transmembrane proteins.

To examine the structure of the protein, especially where a crystal structure or other model may not be available, one can use structural models from homologous proteins where a crystal structure is available or where functional domains have been thoroughly defined through experimentation. For example, for the ATP1A3 protein in the example case workup described above, there is not an available crystal structure for the human protein, though there are structures available for several highly homologous proteins, such as the SERCA pump and the ATP1B1 protein from *Squalus Acanthias*. Furthermore, the main functional domains of the ATP1A3 protein and its homologs have been well-described, including some of the key residues,
such that a variant can be described in terms of its expected effect on the protein structure, and a qualitative description of functionally relevant residues can also be given.

In addition to structural models of a protein of interest, it can be useful to consider resources describing evolutionary conservation of amino acid sequence, as well as resources that identify known functional domains\textsuperscript{161}.

**Protein level: Functional evidence from cell and animal models**

In addition to qualitative or predictive evidence supporting the relationship between a certain amino acid residue and the function of a protein overall, there may also be conclusive past experimentation showing this relationship. Experiments in cell lines can demonstrate the impact of mutation, knockout, or overexpression of the gene of interest, revealing a cellular phenotype to support or refute the putative connection between the affected protein and the organism-level phenotype of interest. Previous animal models may also be available to show whole-organism phenotypes of variation in the gene of interest. Past studies in pre-existing cell or animal models may reveal a relevant mechanistic biomarker, or even suggest potential methods for successful rescue of function (see FlyBase\textsuperscript{162}; Wormbase, http://www.wormbase.org; Jackson Laboratory Mouse search, https://www.jax.org/mouse-search).

In addition to examining evidence from previous animal or cell line experiments, it may be necessary or useful to create new functional models to explore or validate the disease relevance of a candidate variant. This may involve patient-derived cells, existing immortalized cell lines, or new animal models. Given that these methods requires greater investment of time, money, and interpretation, this avenue is not typically pursued until existing evidence has been thoroughly reviewed, and is discussed in the later section on research follow-up.
Protein level: Primary evidence *in situ*

In addition to circumstantial evidence and past experimental evidence for these links between amino acid sequence, protein function, and the function of disease-relevant tissue function, one can also seek primary confirmation of changes in protein level or function by assaying available patient tissue. Feasible assays include purified enzyme function tests or immunofluorescence measurements of protein or mRNA levels or distribution. Protein levels are a function of multiple independent processes, including the rate of synthesis, post-translational modification, stability, and the rate of degradation, all of which may change due to DNA sequence variation, and RNA expression offers an easy and often cheaper proxy to estimate protein levels. There are several repositories compiling expression measurement of many genes across diverse tissues, including NCBI’s GEO, and the GTEx project. For the example of the ATP1A3 protein, a number of past experiments had shown that the protein is restricted primarily to the central nervous system, and expression measurements among the ATP1A gene family showed that ATP1A3 mRNA made up a substantial portion of this functional pool.

In addition to consulting existing evidence for the expression distribution of a candidate disease gene, it may be useful or necessary to confirm changes in protein function or expression by making measurements in model systems; this is discussed further in the later section on research follow-up.

Pathway and network level evidence

Even with good evidence connecting a candidate variant to the function of one protein, it may still be difficult to connect that protein to the affected tissue and overall disease phenotype of a patient. Examining whether the function of an affected protein is important to the function of a disease-relevant pathway can help understand this connection, and also help link the affected
gene or protein to known disease mechanisms or treatments. Without identifying past disease contexts, known mechanisms, or treatments, using pathway labeling can still help group novel cases together for study. Furthermore, grouping diverse cases can help identify core essential biomarkers or pathological mechanisms.

There are a variety of sources to help connect a protein to the functional pathways. Some sources like Kyoto Encyclopedia of Genes and Genomes\textsuperscript{165} and Reactome\textsuperscript{166} provide hand-curated information about functional pathways. Other sources such as IMP\textsuperscript{167}, GIANT\textsuperscript{168}, and STRING\textsuperscript{169}, provide predicted protein pathway and tissue-specific protein functional network models. These resources may identify signals of functional interaction from high-throughput datasets, such as transcriptomic measurement and yeast two-hybrid assays or other protein-protein interaction screens\textsuperscript{170}, and may integrate disparate data sources into inclusive network models through techniques like PANDA\textsuperscript{171}.

**Cohort and population level**

Given the many sources of noise that influence analysis of a candidate disease-associated variant, it is important and very beneficial to confirm findings in multiple cases where possible. Specifically, forming a cohort of relevant cases may help reinforce what elements of a disease presentation are central, while pruning confounding factors and comorbidities. Whereas it is straightforward to collect repeat technical measurements, identifying cases that overlap sufficiently for this type of informative combination may be more difficult.

Consider, for example, the domain of Autism Spectrum Disorders (ASD). In this domain, there are a number of branches of research into underlying mechanism, and both the diversity of findings in these research models and attempts at endo-phenotyping have provided evidence to indicate a heterogeneous set of underlying mechanisms for an umbrella set of behavioral and
cognitive phenotypes. Thus, assembling an informative cohort of ASD patients may require very careful selection to avoid adding mechanistic diversity.

**Roadmap 5. Case Interpretation**

**Difficulties in variant interpretation**

As discussed above, there are many levels of evidence to consider when evaluating a candidate disease-associated variant, and each level is a complex research topic in its own right, with heterogeneous methods that require careful interpretation. Therefore, unassisted investigators may have difficulty making effective use of the full range of available evidence.

The search for diagnostically relevant variants is further complicated by the difficulty of forming an expectation about the results of a given genomic workup. Investigators may use expectations of the results of sequencing workup to define what variants are plausible enough to warrant follow-up investigation or to report to a patient, or when to seek additional measurement.

**Decision support in variant interpretation**

Consider a hypothetical disease-causing variant, tracked from the point of a patient’s referral for sequencing until the discovery of that variant, and note the difficulty of predicting success. DNA must have been extracted from an appropriate tissue source and using appropriate library preparation methods to ensure that fragments containing the variant of interest were present, and present at sufficient fraction to be detected. The sequencing platform must have had sufficient accuracy and depth at the location of the variant to produce high quality reads containing a measurable signal from the variant. Data processing and variant calling needed to be sensitive enough to detect the variant, while still being specific enough that the variant would not
be buried among irrelevant results. Furthermore, there needed to be sufficient supporting evidence for the variant and enough phenotypic information about the patient to conclusively link the variant to the patient’s disease.

The difficulty of conclusively interpreting variants underscores the need for decision-support platforms that enable visualization and facile data exploration on many levels of information, from processed sequencing results, to clinical phenotypes, disease-specific literature, population references, and multiple omics datasets\textsuperscript{173,174}.

During the interpretation of candidate variants, investigators may follow the guidelines set by the American College of Medical Genetics\textsuperscript{6}. These guidelines divide available evidence into broad categories (population data, computational and predictive data, functional data, segregation data, \textit{de novo} data, allelic data, and all other data types), and define thresholds to apply variant labels ranging from ‘likely benign’ to ‘likely pathogenic’. Currently, high-dimensional and unstructured datasets (such as spatial expression distribution in a tissue of interest, evidence from homologous genotypes in animal models, or high-dimensional phenotype data) must be searched and accessed through disparate data portals, and investigators must manually relate these types of data to candidate variants, since there are not reliable methods for using variant information to automatically identify and present relevant evidence from these domains.

**Variant interpretation software**

The bulk of information related to candidate variants – such as technical quality information, predictions of pathogenicity, and reference allele frequencies – can be processed, filtered, and queried within integrated systems such as the WuXi NextCODE software platform (https://www.nextcode.com). This platform enables programmatic annotation and filtration of
variants. Files of variants and annotations are stored as normalized data tables and streamed for joins and filtration. The query system they implement is otherwise agnostic to the content of the underlying data and flexible enough to support any evidence set that can be formatted compatibly, leaving room to incorporate new annotations for predicted disease-relevance of variants, genes, or pathways. Furthermore, their position-indexed data format results in performance that can easily scale with additional computing power.

Figure 11 shows the structure of an example query to apply relevant annotations to variants, apply an expected inheritance model, and then filter based on criteria for disease relevance. Figure 12 shows an example set of variants from a similar query; each row of output corresponds to one variant, and each column describes one annotation. The set of annotations applied here is completely customizable, and this example was created to demonstrate some of the characteristics that an investigator might examine when selecting candidate variants, such as sequencing depth and quality at that location, parental genotype, allele frequency in a reference population, and predicted pathogenicity score according to the Variant Effect Predictor algorithm.
Figure 11: An example Genomic Ordered Relations (GOR) query in the WuXi NextCODE Sequence Miner software. This query simultaneously processes multiple exome trio datasets (with data from an affected proband and from both parents). The query selects exonic compound heterozygous SNVs in RefSeq genes with minor allele frequency less than 1%, that passed a variant quality score filtration, with a phred-scale genotype quality greater than 20 and predicted protein consequence of frameshift, stopgain, stoploss, or splice disruption. Minor allele frequencies are drawn from Broad Institute’s Exome Aggregation Consortium, NHLBI’s Exome Variant Server (http://evs.gs.washington.edu/), Thousand Genomes Project phase, Genome of the Netherlands, and Human Genetic Variation Database. Variant phasing is determined by site-specific comparison between subjects and parents.

| CHR | POS | Reference Allele | GOR_symbol | GT_Depth | GT_CallScore | VEP_Max_Score | VEP_Max_Af | MOTHER_Call | FATHER_Call | OMIM_Description | DIA гос | DIACOS
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>T</td>
<td>H</td>
<td>0</td>
<td>0</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>C</td>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>G</td>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 12: An example of annotated variants produced by a Genomic Ordered Relations (GOR) query and viewed using the WuXi NextCODE Sequence Miner platform. Each row corresponds to one variant, with column headers describing each annotation column. GT: genotype.

- VEP_Max_Score: Variant Effect Predictor pathogenicity score, calculated as the maximum of SIFT, PolyPhen-2, and MutationTaster.
- VEP_Max_Af: the maximum Minor Allele Frequency
observed across Broad Institute’s Exome Aggregation Consortium\textsuperscript{25}, NHLBI’s Exome Variant Server (http://evs.gs.washington.edu/), Thousand Genomes Project Phase 3\textsuperscript{26}, Genome of the Netherlands\textsuperscript{27}, and Human Genetic Variation Database\textsuperscript{28}. OMIM\textsubscript{description}: Gene-level descriptions obtained from Online Mendelian Inheritance in Man\textsuperscript{29}. DIAG\textsubscript{ACMGcat}: variant-level diagnostic category according to American College of Medical Genetics 2013 criteria\textsuperscript{30}

**Roadmap 6. From Findings to Action**

**Confirmation**

If an investigator identifies a candidate variant of high technical quality with sufficient evidence to be deemed relevant to the patient being studied, this variant should then be confirmed by repeat measurement. Variant confirmation is often performed using Sanger sequencing (discussed in an earlier section). Confirmation may also be performed by prospective orthogonal sequencing\textsuperscript{175}; this technique is appealing because further development along these lines may make costly and manually-intensive Sanger confirmation unnecessary, and may also shorten the turnaround time for validated sequencing results.

In addition to confirming the existence of a candidate variant, it may also be worthwhile to confirm the functional relationship between the variant and the patient phenotype before taking follow-up actions. This is especially true in cases where the apparent diagnostic implication is severe, such as in cases of primary immunodeficiency. Confirming that the identified variant is truly a primary causative factor in the disease will provide key information before considering radical intervention such as bone marrow transplant.
Positive report: Clinical follow-up for actionable findings

Primarily, the skill of translating a conclusive finding into clinical-grade actions is the domain of a trained physician, and therefore this portion of the roadmap will only highlight some of the key steps currently in practice. First, the investigator can engage with resources to assist in finding available clinical-grade treatments based on the variants identified, such as OMIM\textsuperscript{29}, ClinGen\textsuperscript{145}, and UpToDate (www.uptodate.com). At the same time, the physician may refer the patient to a genetic counselor to help understand what to expect from genetic testing, and how to proceed after results have been received.

After pursuing available firsthand resources, the primary care provider may consult with a medical geneticist to help interpret the identified variants in the context of the patient’s record and any other available functional testing. As mentioned above, in cases where the apparent findings have significant implications for diagnosis or follow-up treatment, the primary care provider may also refer the patient to a clinical or research specialist for further testing. Additionally, the physician may seek to enroll the patient in a relevant clinical trial if any relevant trials exist; the NIH provides a searchable index of current clinical trials (https://clinicaltrials.gov).

Negative findings: Inconclusive variants, research models

In cases with confirmed variants of unknown significance, there may not be clinical-grade treatment options available. Instead, it may be beneficial to seek research-grade options by engaging with existing research models, or by creating new customized models. Broadly, the long-term benchmarks in a customized research model are to recapitulate key elements from the patient phenotype, attempt to identify biomarkers that are relevant to the disease mechanism, identify candidate therapeutics to modify these biomarkers, and translate these therapies back to
the patient. Recreating the patient phenotype and identifying useful biomarkers requires a system with substantial similarity to the affected tissue in the patient.

**Routes to biomarkers of disease and high-throughput screening**

In some circumstances it may be possible to use primary tissue samples to identify altered pathways\textsuperscript{176}. It may also be possible derive a cell line directly from biopsied patient cells, or to make measurements of distal tissues as a proxy for the tissue of interest\textsuperscript{117}. When primary tissue is not accessible for study and distal tissues are not a viable proxy, it may also be possible to derive iPS cells and use these cells to generate model tissues\textsuperscript{103}.

Patient-derived iPS cells offer a source of model tissue that matches the patient’s full genotype, and these cells may be differentiated into a target cell type for study. The generation of iPS-derived tissue models is an incredibly active area of research\textsuperscript{104,177–179}. This approach may be especially useful if investigators also have techniques for high-throughput measurement of relevant disease phenotypes in model cells. For example, recent work on all-optical electrophysiology shows promise as an approach for high-throughput measurement of patient-derived iPS-neuron models\textsuperscript{180}. With methods for reliably and efficiently creating patient-derived iPS cells and downstream tissue models, as well as methods for high-throughput phenotypic measurement on these models, investigators may be able to search for drugs with disease-modifying activity in a model that is highly likely to translate to the patient.

In cases where cell-line models are insufficient to recreate the phenotype of interest, it may instead be worthwhile to pursue animal models of disease. There is a huge and growing set of available research models; investigators may start by exploring existing models through species-specific catalogs, such as WormBase (https://www.wormbase.org/), FlyBase (http://www.flybase.org), and the Mouse Genome Informatics database of Cre driver lines.
Animal models can be adapted to a specific patient by replicating a variant of interest using techniques such as TALENs, CRISPR-Cas9, or ZFN\textsuperscript{181}.

For many model systems, there are also resources to start browsing existing tissue specific datasets, such as the Allen Brain Atlas\textsuperscript{182} for mouse and human brain expression data, or ViBE-Z\textsuperscript{183} and the Zebrafish Brain Browser\textsuperscript{184} for Zebrafish brain expression data.

Each of the exploratory stages discussed here requires significant expertise; some institutions offer relevant resources such as core facilities for deriving iPS cell lines from patient cells (e.g. http://ipscore.hsci.harvard.edu), and facilities for creating iPS-derived tissue models (e.g. http://www.childrenshospital.org/research-and-innovation/research/centers/translational-neuroscience-center). Beyond the available resources at an investigator’s home institution, it may be possible to find domain expertise through contract research organizations (e.g. http://www.wuxiapptec.com, http://www.criver.com).

**Discussion**

Understanding how genomics data is generated, processed, and filtered can help investigators accurately identify disease-associated genetic variation in patients who are otherwise lacking diagnosis. The subsequent search for treatments is a multi-pronged effort, especially for novel variation. Finding treatments depends in part on individualized functional characterization, both in patient tissue samples and in research models, and also depends on drawing connections between patient-specific biomarkers and existing research models or existing libraries of FDA-approved perturbagens. These connections may be drawn empirically through high-throughput screening approaches, or through knowledge-guided intersections of pathways and mechanisms.
This personalized medicine approach to genetics is already producing tangible results, from pharmacogenomics applications, to single-patient drug screening platforms and animal models. This movement draws substantial support from the large-scale collaborative efforts at many levels, including data generation and processing platforms, annotation frameworks and software support for case interpretation, and flexible research models of rare disease. Investigators can tap into this community of resources through institutional facilities, as well as through public-facing web portals and open-source software packages. Collaborative leaders have piloted projects to build shared resources across multiple large institutions and assemble patient cohorts, and the potential for benefit to patients continues to grow alongside the growth of these projects\textsuperscript{25,185,186}.

Responsible open data sharing will be a crucial element of long-term success of statistical learning on genotype-phenotype associations, as will the use of compatible platforms for data generation and processing. Furthermore, describing cases today with rich phenotypic data for the benefit of future analysis methods will help enable create lasting value from current datasets.

**Future directions**

There is potential for advancement in many of the technical aspects of this work. For instance, given the metaphorical ‘bandwidth’ on current techniques for data generation, processing, and storage, investigators must only apply sequencing judiciously. Thus, investigators can only rely on sparse, static measurements to attempt to unravel the dynamics of potentially subtle disturbances in a complicated biological system.

Two key areas under development may provide investigators more power to resolve signatures of disease from omics datasets. The first area are infrastructural methods, such as the development of cheaper and faster sequencing platforms capable of higher resolution, and
algorithms for processing data more efficiently and storing datasets more densely. With these developments, it may become feasible to obtain time-course measurements at the single-cell scale, even across multiple tissues\textsuperscript{187}. This will provide a data substrate for the other, second area of development in identifying disease mechanisms, prognostic markers, and therapeutic perturbations from high-dimensional multi-omics datasets. In an oft-cited analogy, the dramatic success of Google Maps to provide answers to travelers on a ‘geographic odyssey’ provides a powerful model for precision medicine. Following this model, the development of infrastructural methods will populate the multiple data layers necessary to describe the state of the biological world, and the latter set of methods will provide the routing algorithms to identify key network roadblocks and alterations and make informed decisions based on this multiplicity of data.
Bibliography


19. Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic...*


39. Prilutsky, D. *et al.* IPSC-derived neurons as a higher-throughput readout for autism:


60. Bøttger, P. et al. Distribution of Na/K-ATPase alpha 3 isoform, a sodium-potassium P-


81. MacDonald, M. L. et al. Altered Glutamate Protein Co-Expression Network Topology


91. Shi, L., Quick, M., Zhao, Y., Weinstein, H. & Javitch, J. A. The Mechanism of a


121. Pinkel, D. & Albertson, D. G. Comparative genomic hybridization. *Annu. Rev. Genomics*


132. Liu, L. et al. Comparison of next-generation sequencing systems. *Journal of Biomedicine*


143. Massie, M. et al. *ADAM: Genomics Formats and Processing Patterns for Cloud Scale*


163. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from


