Scaling Interpretation of Clinical Sequence Data for Rare Diseases

By Jeanette McCarthy, MPH, PhD, Fabric Genomics

OVER A DECADE has passed since the first rare genetic disease patient was diagnosed using genome sequencing.[1] The landmark story of Nic Volker's diagnostic odyssey, subsequent sequencing, and treatment laid the groundwork for genomics to diagnose rare diseases. Today, clinical sequencing is recommended as standard of care for diagnosing rare diseases in children in both the acute care setting (intensive care) as well as outpatient setting (diagnostic odyssey cases).[2,3] In the United States, the uptake of diagnostic sequencing has been driven by demonstrable clinical validity, utility, and cost-effectiveness, which has led to increased, although not uniform, payer coverage.[4-6]

Many laboratories offer clinical sequencing at a typical price of thousands of dollars per patient with several weeks turn-around time. Commercial laboratories including GeneDx, Ambry, Invitae and Centogene (Europe) perform the majority of these tests. Rady Children's Hospital pioneered (and is the principal provider of) rapid clinical whole genome sequencing to diagnose genetic diseases in the acute setting, such as neonatal and pediatric intensive care units. Children's hospitals that see rare disease patients on an outpatient setting are increasingly bringing clinical sequencing in house to run the

same tests offered by commercial laboratories, but with more control over turn-around time and cost, improved diagnostic yield, and the added benefit of banking the genomic data of their patients for future analysis.

With the cost of sequencing coming down significantly (Illumina just announced the $200 genome), the disparity with the biggest cost-driver, analysis and interpretation, has become even more pronounced. Analysis and interpretation, while still requiring highly skilled individuals, can be streamlined with advanced artificial intelligence that can significantly reduce the interpretation time, and therefore, cost. This reduction in interpretation time and overall cost of clinical sequencing tests will allow laboratories to *scale and expand* the application of genome sequencing beyond the diagnostic setting to broader screening of populations including newborns and ostensibly healthy adults. Another issue that may be addressed with greater access to this technology is the correction of misdiagnosed patients, saving individuals from years of trial-and-error therapies. And while all results may not lead to complete cures, they may indicate treatments that mitigate or allow for palliative intervention.

Early genome interpretation methods

Diagnostic sequencing in rare genetic diseases has been enabled by next generation (short read) sequencing technologies which allow for the analysis of a whole genome or exome in a matter of hours. Following sequencing, the data are aligned to a reference genome and variants are called using bioinformatics tools. Subsequent annotation of variants including their location relative to genes, their predicted effect on the resulting protein, allele frequency, presence in clinical database, available literature and other relevant data provide the context of each variant.

The average genome contains four to five million variants, among which lies a presumed culprit(s) that can be linked to the disease presenting in the patient. The earliest methods used for discovering the disease-causing (pathogenic) variant(s) underlying a rare genetic disease were simple filtering strategies. The idea was to focus on the type of variants that were commonly known to be associated with rare genetic disorders: mono- or bi-allelic rare, protein-altering variants. This approach reduced the variant burden by several orders of magnitude, leaving only a few hundred candidate variants for consideration. From here, a manual review of the genes containing the rare, protein-altering variants would ensue, with variant scientists scouring the literature and public databases like OMIM, looking for clues that the damaged gene could be linked to the disease in the patient. Analysis of a genome could still take several hours per case.

Table 1: Evolution of Fabric Genomics' gene prioritization algorithms

Fabric algorithm	Basis of prioritization	Year	Reference
VAAST	Variant deleteriousness	2011	PMID: 21700766
Phevor	VAAST plus phenotype of patient (HPOs)	2014	PMID: 24702956
GEM	Phevor plus knowledge from OMIM and ClinVar	2021	PMID: 34645491

Algorithmic approaches to prioritize variants

The field has come a long way since then with advances in bioinformatics that now offer alternative approaches to simple filtering (**Table 1**). In 2011, Fabric Genomics introduced the very first variant prioritization tool called VAAST (Variant Annotation, Analysis & Search Tool).[7] VAAST is a gene-ranking algorithm that integrates information about phylogenetic conservation, amino acid substitution effects, allele frequencies, and other factors into a single unified likelihood-framework for ranking genome variants according to deleteriousness. Reviewing the highest VAAST ranked (more deleterious) variants first allowed variant scientists to identify causal variants quicker and more efficiently.

Fabric followed this up in 2014 with another variant prioritization algorithm called PHEVOR (the Phenotype Driven Variant Ontological Re-ranking tool).[8] PHEVOR builds upon the output of VAAST, integrating it with knowledge resident in multiple diverse biomedical ontologies. Ontology annotations, such as those found in Human Phenotype Ontology (HPO) and Gene Ontology (GO), are readily available for many human and model organism genes. Starting with a list of phenotypes or gene function terms, PHEVOR leverages a network of these ontologies to automatically derive a candidate gene list from these terms, including genes that are not explicitly annotated to a given phenotype, but for which the ontology graph structure implies potential latent linkages. It then re-ranks the variants from VAAST output accordingly, reprioritizing them considering gene function, disease, and phenotype knowledge.

Gene prioritization algorithms have resulted in a reduction in the number of candidate genes for a given patient to <100, reducing the review time to hours instead of days. While this represents a significant improvement, these methods still miss a fraction of disease genes, and therefore are commonly used in parallel with filtering approaches instead of completely replacing them.

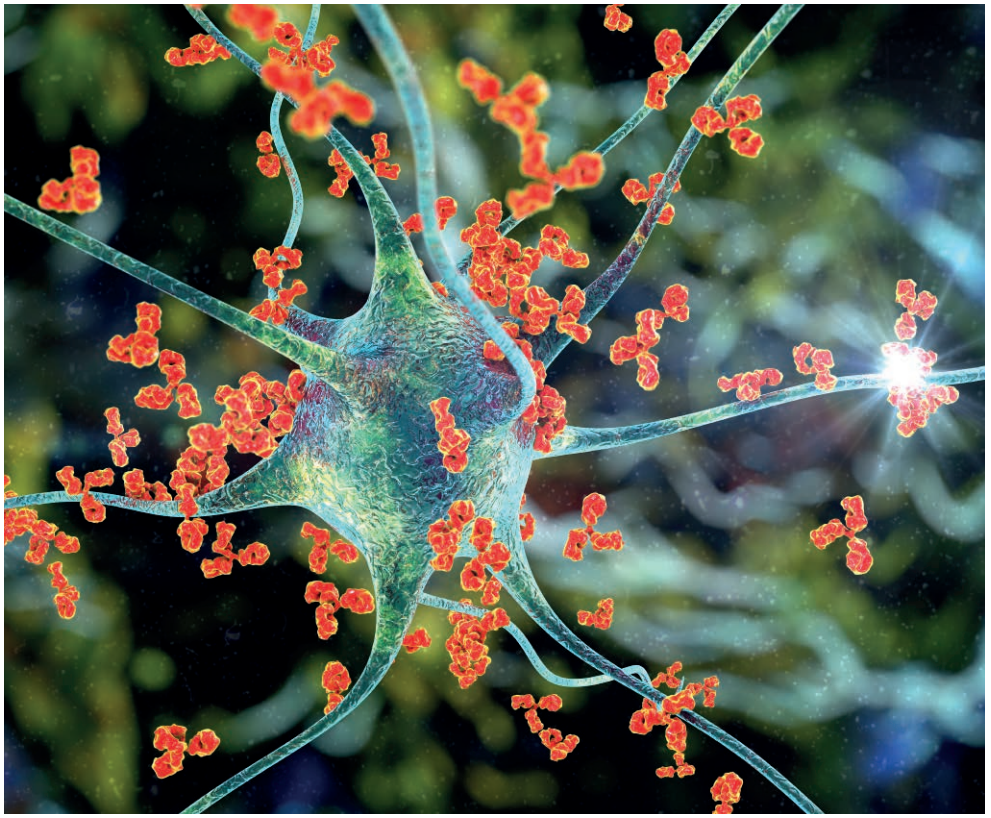
A truly scalable interpretation platform

Over the years, a vast amount of disease-related information has accumulated in databases like OMIM and Clinvar, as well as in published scientific literature. The OMIM database has been growing at a rate of ~250 new disease gene discoveries per year. ClinVar has been growing at a rate of close to 10,000 new variant disease associations per year. The Mastermind database contains information from millions of scientific articles. These and other resources have contributed to the enhanced efficiency and accuracy of interpretation and improved performance of diagnostic algorithms (see **Table 2**).

In 2020, Fabric introduced its latest gene prioritization algorithm, GEM. GEM builds on prior algorithms VAAST and PHEVOR, but also considers clinical knowledge resident in the

Table 2: Resources contributing to the enhanced efficiency and accuracy of interpretation of rare disease cases

Resource	Description	Website
Human Phenotype Ontology (HPO)	HPO provides a standardized vocabulary of phenotypic abnormalities encountered in human disease	hpo.jax.org
Online Mendelian Inheritance in Man (OMIM)	OMIM is a compendium of human genes and genetic phenotypes	www.omim.org
ClinVar	ClinVar archives and aggregates information about relationships among variation and human health	www.ncbi.nlm.nih.gov>clinvar
Mastermind	Mastermind is a search engine to identify gene, variant, disease, phenotype and therapy evidence from scientific literature	www.genomenon.com/mastermind/



ClinVar and OMIM databases. In addition, GEM utilizes an algorithm that measures consanguinity, along with molecularly-defined ancestry of the patient and a model that controls for artifacts that are common for short-read technology. All these inputs are integrated into a Bayesian Network that determines the likelihood of a variant in a gene to be disease causing. The algorithm takes patient HPO terms as input and evaluates the whole genome (or exome), typically returning a very short list of viable candidates, on average fewer than ten per case. These candidate variants are reviewed to find the best match with the disorder. This cuts the review time for a case from hours to sometimes minutes. GEM has been shown to have excellent sensitivity, finding 98% of causal single nucleotide variants (SNVs) in a large benchmarking set of whole genome cases from the NICU.[9]

GEM is also a powerful tool for ranking causal copy number variants (CNVs) and other structural variants. Variant calling algorithms like Illumina's Dragen can produce over 5000 CNVs per whole genome, many of which are false positives. GEM can take unfiltered CNVs as input, along with SNVs and produce a single ranked list of candidates, on average ranking only one to two CNVs per case. Benchmarking GEM's performance in over 75 cases with causal CNVs demonstrated over 97% sensitivity for CNV prioritization (unpublished data), including large multigenic CNVs, gene-level CNVs and compound heterozygotes (CNV/SNV).

Knowing when to stop

The diagnostic yield of exome and genome sequencing is estimated to be about 35-40%.[5] The inclusion of CNVs can improve this yield modestly. If a causal variant is present in the genome, it will likely be found relatively quickly, especially when using an AI algorithm like Fabric's GEM. In many laboratories, most of the interpretation time is spent on the ~60% of negative cases, where an obvious candidate is elusive.

Table 3: Websites for evaluating treatable genetic disorders at childbirth

Project name	Website
BeginNGS	https://radygenomics.org/begin-ngs-newborn-sequencing/
Guardian	https://guardian-study.org/
Early Check	https://earlycheck.org/
Screen4Care	https://screen4care.eu/

After reviewing the top candidates in a case and coming up empty handed, it's tempting to continue to evaluate additional variants that may be less compelling in the hopes of finding a diagnostic variant, however, this can add significant reviewer time and reduces the throughput of a laboratory.

Using a smart, streamlined workflow that incorporates AI algorithms like Fabric GEM provides variant scientists with a stopping point, beyond which there is a diminishing return of definitive diagnoses. If no good candidates are found in a case after reviewing the GEM candidates, the likelihood that reviewing additional variants will yield a diagnosis is unlikely. A simple workflow that incorporates GEM has proven to be an efficient workflow that reduces the overall time and cost of variant interpretation while maximizing the diagnostic yield.

Having an efficient and cost-effective means of clinically interpreting genomes opens the possibility to scale the analysis of rare disease cases, provide routine re-analysis of negative cases and launch into new areas like newborn sequencing.

"A simple workflow that incorporates GEM has proven to be an efficient workflow that reduces the overall time and cost of variant interpretation while maximizing the diagnostic yield"

The rise of expanded newborn genome sequencing

The application of genome sequencing to rare disease cases has recently expanded beyond diagnostic testing to the screening setting. State-mandated newborn screening to prevent childhood onset of a panel of severe but treatable, genetic conditions is hailed as one of the most successful public health programs in the United States. Newborn screening currently utilizes biochemical (mass-spectrometry) methods to detect analytes from dried blood spots. Several studies have assessed the feasibility of using next-generation sequencing as a complementary test, either to confirm biochemical test results or as a means to expand the number of genetic disorders that could be detected at birth.[10-13]

Currently, hundreds of treatable genetic disorders are being considered across multiple pilot projects in the United States, the United Kingdom, Europe, Australia and elsewhere. Studies including BeginNGS at Radys Children's Hospital, Guardian in New York City, Early Check in North Carolina, and the Screen4Care consortium in Europe are just

Table 4: Studies evaluating sensitivity of next-gen sequencing for expanded newborn screening

Study	Genes (backbone)	N affected newborns	Sensitivity	Source of known P/LP variants	Filter to find novel P/LP variants
Bodian (2016)	163 (WGS)	34 NBS+	91%	Clinvar	• Rare • LOF • Damaging
Roman (2020)	466 (WES)	17 IEM+	88%	Clinvar	• Rare • LOF • Damaging
Adhikari (2020)	78 (WES)	674 IEM+	85%	Clinvar	• Rare • LOF • Damaging
Kingsmore (2022)	388 (WGS)	119 NICU+	88%	Clinvar, Mastermind	None (Fabric GEM)

WGS – whole genome sequencing; WES – whole exome sequencing; NBS - newborn screening; IEM - inborn errors of metabolism; NICU – neonatal intensive care unit; LOF – loss of function; P/LP – pathogenic or likely pathogenic variants

a few of the efforts that were highlighted in the first International Conference on Newborn Sequencing (IcoNS) held in Boston last October (see **Table 3** for websites).

Need to scale interpretation for newborn sequencing

Several aspects of newborn sequencing present challenges for interpretation. The first consideration is the need to return results relatively quickly, on the order of days. Second, if used as a population-screening, a too must allow for processing large numbers of cases (over three million live births per year in the United States alone). The third aspect to be weighed for interpreting large scale newborn sequencing is the need for high specificity, to reduce false positives that lead to unnecessary and expensive follow up and treatment. A scalable interpretation platform should address all three.

Using a traditional panel-based approach, newborn sequencing of hundreds of genes performed off a whole genome or exome backbone will yield several hundred single nucleotide variants and several dozen copy number variants requiring interpretation. Strategies that have been used by different groups to identify reportable pathogenic (P) and likely pathogenic (LP) variants include a combination of look-up tables for known P/LP variants, followed by filtering to identify novel rare loss of function or other predicted deleterious variants.

Table 4 summarizes the major studies evaluating next generation sequencing for expanded newborn screening. There is remarkable consistency between the studies, showing a sensitivity near 90% for sequence-based approaches to identifying reportable variants. All these studies included Clinvar as a source of known P/LP variants, while the study by Kingsmore, et. al. also included curated variants from Genomenon's Mastermind database. Three of the studies supplemented this with analysis of novel P/LP variants identified by filtering on a combination of allele frequency, loss of function and predicted deleteriousness. The Kingsmore study took a different approach to identifying novel P/LP variants by employing Fabric's GEM algorithm, allowing for the prioritization of suspicious missense variants along with loss of function variants.

Although Fabric's GEM algorithm was originally developed for indication-based testing and requires phenotypes (HPO terms) of affected individuals as input, a subsequent iteration of the GEM algorithm, GEM-NBS, obviates the need for HPO terms, making it a viable solution for newborn sequencing. GEM-NBS accurately ranks both known and novel P/LP variants in these cases and in early benchmarking studies shows a 93% sensitivity. The number of candidates ranked by GEM-NBS is considerably lower as well, with an average of <1 candidate per case (compared to <10 for GEM used in indication-based testing). GEM-NBS addresses the need for speed, efficiency and accuracy in interpretation of in silico newborn sequencing panels.

Summary

In the time since Nic Volker's odyssey, we have witnessed the demonstration, deployment and routine adoption of clinical sequencing for rare diseases, resulting in thousands of newborns and children receiving diagnoses and, in some cases, life-altering treatments. As clinical sequencing is reduced to practice, AI-driven analysis workflows offer an efficient means of interpreting cases, opening the door for more hospital laboratories to offer these tests at the point of care. Reduction in the cost of sequencing along with improved scalability of interpretation will enable broader uptake and expanded use of clinical sequencing for early detection of genetic disease in newborns. JoPM

Jeanette McCarthy, MPH, PhD

VP, Precision Medicine, Fabric Genomics

Founder, Precision Medicine Advisors

Adjunct Associate Professor, Community and Family Medicine, Duke University

Jeanette McCarthy joined Fabric Genomics in 2020 and serves as Vice President of Precision Medicine, leading Fabric's Clinical Services as well as Product. She is a genetic epidemiologist by training, receiving her MPH and PhD at UC Berkeley under the mentorship of geneticist Dr. Mary-Claire King. Prior to joining Fabric, she worked as a consultant, helping numerous organizations educate and train their workforce in the use of genetic testing in clinical practice. She is an internationally-recognized leading educator in precision medicine, reaching thousands of learners through her online Precision Medicine Academy. Prior to that, she served on the faculty at Duke University in the Institute for Genome Sciences and Policy where she ran research programs on the genetic underpinnings of complex diseases, both infectious and chronic. She began her career running clinical genomic studies at Millennium Pharmaceuticals, a pioneer in genomic-based drug discovery, and it's subsidiary, Millennium Predictive Medicine.

References

1. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 2011;13(3):255-262.
2. Manickam K, McClain MR, Demmer LA, et al. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2021;23(11):2029-2037.
3. Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med. 2018;20(4):435-443.
4. Dimmock D, Caylor S, Waldman B, et al. Project Baby Bear: Rapid precision care incorporating rWGS in 5 California children's hospitals demonstrates improved clinical outcomes and reduced costs of care. Am J Hum Genet. 2021;108(7):1231-1238.
5. Clark MM, Stark Z, Farnaes L, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med. 2018;3:16.
6. Douglas MP, Parker SL, Trosman JR, Slavotinek AM, Phillips KA. Private payer coverage policies for exome sequencing (ES) in pediatric patients: trends over time and analysis of evidence cited. Genet Med. 2019;21(1):152-160.
7. Yandell M, Huff C, Hu H, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. 2011;21(9):1529-1542.
8. Singleton MV, Guthery SL, Voelkerding KV, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. 2014;94(4):599-610.
9. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med. 2021;13(1):153.
10. Bodian DL, Klein E, Iyer RK, et al. Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. Genet Med. 2016;18(3):221-230.
11. Roman TS, Crowley SB, Roche MI, et al. Genomic Sequencing for Newborn Screening: Results of the NC NEXUS Project. Am J Hum Genet. 2020;107(4):596-611.
12. Adhikari AN, Gallagher RC, Wang Y, et al. The role of exome sequencing in newborn screening for inborn errors of metabolism. Nat Med. 2020;26(9):1392-1397.
13. Kingsmore SF, Smith LD, Kunard CM, et al. A genome sequencing system for universal newborn screening, diagnosis, and precision medicine for severe genetic diseases. Am J Hum Genet. 2022;109(9):1605-1619.