Comparative Sequencing Strategies
Decreasing sequencing costs and broad dissemination of next-generation (next-gen) equipment and expertise are increasing the availability of massively parallel sequencing of patient DNA samples (short read massively parallel or next-gen sequencing). 1,2 Most rapidly expanding is exome sequencing, where all protein-coding sequences (exons) are selected from total genomic DNA and selectively sequenced. 3 Alternative approaches to next-gen sequencing include targeted sequencing (TS) and whole genome sequencing (fig. 1). Currently, marketed targeted Sanger sequencing panels using traditional individual exon-by-exon sequencing remain expensive and time consuming, and massively parallel next-gen approaches are beginning to supplement Sanger sequencing in the clinic. 4

Here, we compare and contrast the 3 next-gen approaches to molecular diagnostics: TS panels, whole exome sequencing (WES), and whole genome sequencing (WGS). Briefly, TS selects candidate genes that are already known to cause the disorder in question and targets only these for extensive sequence analyses (as high as 1 million bp). Exome sequencing (WES) pulls out most exons (coding sequence) of all genes in the genome for sequencing (as high as 1 million bp). WGS nondiscriminately sequences all 6 billion bp of DNA in a patient, including the large majority of DNA that does not code for proteins and remains problematic to interpret.

Both targeted next-gen sequencing and whole exome next-gen sequencing have recently entered the molecular diagnostics workspace, with multiple private and academic labs offering next-gen sequencing on a clinical, fee-for-service basis. 9,10 The costs of generating targeted and WES data are not very different. If targeted panels covering 50 genes cost about the same as WES covering 20,000 genes, then it would seem that “more is better,” and the exome will become the standard diagnostic test. However, there are some technical differences in how the data are generated and interpreted that make the targeted versus exome choice less clear cut. The key issue is that TS is able to cover 99% of the candidate genes, whereas exomes cover only 90% of the same candidate genes (as well as 90% of all other genes; in-house sequencing data). This 10% difference in sensitivity is significant. For example, as described in more detail below, if a patient with dilated cardiomyopathy (DCM) comes into your clinic and you suspect a titin gene mutation, you may wish to confirm this suspicion by next-gen sequencing. Ordering a TS panel that includes the titin gene will cover 99% of the 363 exons and likely rule in or rule out titin as a cause of the patient’s disorder. However, sending the same patient out for a whole exome analysis will miss 1 of 10 exons; so for the 363 exon titin gene, 36 exons will be missing from the data generated (often called exon drop outs). Thus, a negative result could be because of missing data due to drop out, and thus not rule out a titin mutation. The higher sensitivity of TS panels has been successfully shown in the diagnosis of cardiomyopathies by Meder et al. 11

In addition to the issue of differential drop out, TS and exome sequencing can differ in terms of accuracy of detection of mutations and copy number variations (CNVs). Accuracy for both small mutations (single base changes, small deletions/duplications) and CNVs depends on read depth. Depth refers to the number of independent sequence reads generated.
for each specific region of the genome tested. A typical exome has a median read-depth of 50; each exon shows ≈50 independent reads. In contrast, targeted re-sequencing panels may achieve a median read-depth of ≥500 reads per exon (10x greater than exomes). In general, the greater the depth, the more accurate the detection of mutations, and the easier it is to detect CNVs. It is possible to increase the machine time on an exome sequence to match the depth of TS, but this increases machine time allocated per sample, and thus cost.

Whole complete genome sequencing has not yet entered the molecular diagnostics workspace. There has been a single company marketing WGS, Complete Genomics, but uptake has been limited. The Complete Genomics method that has been used to generate most whole genomes produces large numbers (120 billion) of short (33 bp) reads, shorter than other next-gen sequencing platforms. Interpretation of pathogenic variants in the large majority of noncoding sequence (99% of genome) remains difficult, and this results in relatively little interest in generating these data in the clinical setting. However, as sequencing technologies continue to rapidly advance, and costs decline, this may change in the not so distant future.

**Variant Calling: Benign, Pathogenic, Variant of Unknown Significance, or Incidental**

Once next-gen sequencing data are obtained with adequate depth, the bioinformatics process of interpretation of the data begins to define potential mutations, dubbed variant calling (Figure 2). Variants are typically grouped into 4 categories: benign, pathogenic, variant of unknown significance (VOUS/VUS), or incidental findings (IFs). Pathogenic variants are mutations—variants in a gene that are likely or known to cause the patient’s symptoms (Table 2). Benign polymorphisms are variants that are common in healthy populations and are not considered candidates for disease-causing variants. Variants of unknown significance are sequence changes that are not seen at high frequency in general populations, but where there is little or no supportive evidence of the variant being pathogenic. An IF is defined as a sequence variant that...
has potential health or reproductive significance to the patient being tested but is secondary to the patient’s primary clinical complaint (eg, a variant known to be associated with drug sensitivity but unrelated to the complaint of cardiac symptoms; Table 2).13

A growing number of software tools and database resources greatly aid in variant calling and variant filtering, and these are increasingly adept at predicting which variant has the highest probability of causing disease. Predictive software, such as Mutation Taster,14 Polyphen-2,15 likelihood ratio test16 (of codon constraint), and SIFT, [Sorting Intolerant From Tolerant]17 takes into account evolutionary conservation and amino acid substitution to predict the probability of the mutation being disease causing or benign. Different functional prediction tools use different methods to assign likelihood of functional significance with each tool having its own strengths and weaknesses. The database of non-synonymous functional prediction18 includes functional prediction scores from 4 predictive software (MutationTaster, Polyphen-2, SIFT, [Sorting Intolerant From Tolerant] and likelihood ratio test) and a conservation score (PhyloP [phylogenetic P-values]). Therefore, it gives likelihood scores for functional prediction of non-synonymous single-nucleotide polymorphisms from multiple databases with a single query.

Different mutations in a single gene can cause varying phenotypes (phenotypic and allelic heterogeneity), and the bioinformatics pipeline to classify variants as likely pathogenic, possibly pathogenic or variant of unknown significance, benign polymorphism, or clinically actionable (or nonactionable) incidental is not black and white. Typically, the larger a gene becomes, the more variants it contains, and categorizing variants and defining what is pathogenic is more difficult for larger genes like titin. We define and discuss the classification of variants into these categories, using titin as an example in Table 2. The clinical phenotype of a patient can also be considered a highly relevant filter for classifying variants. There can be a top candidate gene in the list of those containing potential pathogenic variants (eg, disease-causative variant) given concordance of the patient’s clinical symptoms to previous reports. Confirming variants as pathogenic is easier if the variants are previously reported and known to be causative of the disease being studied. With novel variants in genes related to the disease where the variant is predicted to be pathogenic (evolutionarily conserved region or frame-shifting variant), further analysis may be required. Testing can include segregation analysis in families, functional testing for reduced protein levels in tissue/biopsy, checking databases for known polymorphisms, or in vitro protein remodeling studies.19

<table>
<thead>
<tr>
<th>Table 1. Comparison of Existing Next-Generation Sequencing Strategies for Molecular Diagnosis of a Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approach</strong></td>
</tr>
<tr>
<td><strong>Approx size</strong></td>
</tr>
<tr>
<td><strong>Capture technique</strong></td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
</tr>
<tr>
<td><strong>Success</strong></td>
</tr>
</tbody>
</table>

CNV indicates copy number variation; PCR, polymerase chain reaction; TS, targeted sequencing; WES, whole exome sequencing; and WGS, whole genome sequencing.
algorithms at once and giving a probability score ranging from 0 to 1 (0 being benign and 1 being disease causing). For detection of pathogenic variants, further variant filtration is required. This is achieved by looking at only the coding sequences (CDS); splice junctions (±5 bases from exons), removing synonymous SNPs, and removing polymorphisms already reported in dbSNP and 1000 genomes. We then create a filter to first look at the 51 genes already reported to cause DCM as there is a greater probability of finding a mutation in 1 of these genes given the patient’s phenotype. If variants are detected in the known DCM-causing genes, the variants are confirmed by Sanger sequencing and cross-referenced with the patient’s physician. Because exome sequencing has a high rate of false positives, confirming that the variant is not a polymorphism on the exome variant server is the next step. If no variants in known genes are detected in the patient, we then proceed to analyzing variants in the whole exome.

**CNVs From Next-Gen Data**

Detection of deletions or duplications of exons (CNVs) is necessary for high sensitivity of detection for all types of mutations. With the aid of newer bioinformatics tools, it is possible to detect CNVs in exome sequencing datasets using CNV calling algorithms; however, these tools are still being evaluated before moving to the clinical realm. Next-gen sequencing does not detect CNVs directly via sequence data, but the number of sequence reads per exon can be used to derive the presence of CNVs. A log2 ratio of reads per kilobase of exon model per million mapped reads for 1 patient’s next-gen reads can be compared with the average of the other patients for each exon, with significant deviations from baseline indicating the presence of a deletion or duplication. The accuracy and sensitivity of all CNV detection methods depend on the depth of the sequence data, where targeted next-gen data typically give greater depth, and hence greater sensitivity, for CNVs compared with exome or whole genome data.

Interpretation of CNVs and small mutation pathogenic data is often simplified by the sequencing of parents and siblings, as well as the proband. Having the sequence data of the parents permits the testing of different inheritance models. For example, de novo mutations are increasingly seen as a cause of genetic disease. However, detection of de novo

---

**Table 2. Classification of Variants Obtained From NGS Data Using TTN as an Example**

<table>
<thead>
<tr>
<th>Variant</th>
<th>Definition</th>
<th>Example</th>
<th>Significance</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathogenic</td>
<td>Disease-causing mutation</td>
<td>Frame-shifting variant in TTN A-band</td>
<td>Diagnosis</td>
<td>Yes</td>
</tr>
<tr>
<td>Benign polymorphism</td>
<td>Natural variation in DNA with no known adverse effect</td>
<td>Variant in TTN present in dbSNP with high frequency</td>
<td>Could be a genetic modifier in future studies</td>
<td>No</td>
</tr>
<tr>
<td>VOUS/VUS</td>
<td>Variant of uncertain or unknown significance</td>
<td>Missense variant in a less evolutionarily conserved region of TTN</td>
<td>May change status to pathogenic with additional data</td>
<td>Yes</td>
</tr>
<tr>
<td>Incidental findings</td>
<td>Variants that have been associated with some risk of a disease state but are not relevant to the patient’s diagnostic questions</td>
<td>Carrier state for mutations in CFTR gene</td>
<td>Not diagnostic. Assessment of clinically actionable defined by ACMG guidelines</td>
<td>Yes, if clinically actionable</td>
</tr>
<tr>
<td>Genetic modifier</td>
<td>Variant that modifies a disease state because of a second gene</td>
<td>Hypothetical variant in a gene that increases or decreases TTN expression</td>
<td>Genetic modifiers of monogenic disorders remain few</td>
<td>No</td>
</tr>
</tbody>
</table>

ACMG indicates American College of Medical Genetics and Genomics; CFTR, cystic fibrosis; dbSNP, single-nucleotide polymorphism database; NGS, next-generation sequencing; TTN, transmembrane conductance regulator; titin; and VOUS, variant of unknown significance.
mutational events is all but impossible without the sequence of both parents in hand.

The current ambiguities concerning sensitivity and specificity of detecting and reporting pathogenic mutations and CNVs, as well as incidental data, require consent forms that speak to these ambiguities. Before offering WGS tests, medical diagnostic laboratories have to be equipped with detailed consent forms and resources for data storage, analysis, and interpretation.

**IFs: What Is Considered Clinically Actionable?**

Incidental findings (IFs) are variants that have defined functions associated with them, such as the sequence change in the Factor V clotting protein that is associated with changes in thrombosis in general populations (Factor V Leiden; ≈5% of individuals of European descent). Another example of an IF is identification of the carrier state for the common deltaF508 cystic fibrosis mutation in a patient where cystic fibrosis is not in the differential diagnosis. A key issue with incidental data is whether they are actionable and how the definition of actionable changes as a function of the patient’s age. For example, to a child, neither Factor V Leiden nor deltaF508 CF carrier status is clinically actionable (nothing would be done differently in the clinical management of the child given knowledge or lack of knowledge of these genotypes). Yet, both become clinically actionable in the context of a young adult. A pregnant woman who is a carrier for Factor V Leiden has a 5- to 10-fold increase in the risk of venous thrombosis events, whereas homozygotes have nearly a 100-fold increased risk. Similarly, the carrier status for deltaF508 becomes clinically actionable when considering pregnancy and having children, and genetic counseling of the carrier is warranted. The definition of clinically actionable versus nonactionable incidental data from next-gen sequencing studies is under active discussion by national regulatory and academic groups and is hotly debated. Recently, the American College of Medical Genetics and Genomics has issued a set of guidelines for reporting of IFs after exome and WGS testing a clinical setting. The authors provide a list of 57 genes where known pathogenic or expected pathogenic mutations should be reported back to the physician, although these genes are not related to the initial clinical complaint or differential diagnosis of the patient (eg, incidental data). Some examples include BRCA1 (breast cancer 1, early onset risk for breast cancer), MYH7 (cardiomyopathy risk), and RYR1 (malignant hyperthermia susceptibility). There are concerns that this will increase costs of these tests in the future because of increased burden on data analyses and interpretation. For example, the RYR1 gene has 106 exons, and variants are frequently found in this gene by exome sequencing. Assigning pathogenic significance to a variant in RYR1 in terms of risk for malignant hyperthermia is extraordinarily challenging. One must also question the cost/benefit of reporting back RYR1 expected pathogenic incidental data, as the risk for malignant hyperthermia is generally only after halothane anesthetics, and only a small subset of the population is ever exposed to this.

A recent publication outlines a proposed bioinformatics tool with minimal researcher effort that looks for IFs (defined as variants known to cause Mendelian diseases) from WGS data as a possible solution to reduce the burden on laboratory personnel. The interpretation, reporting, and retention in medical records are currently quite variable, and most remains in the context of research studies rather than routine molecular diagnostics. It will take a while for laboratories and institutional review boards to incorporate the new American College of Medical Genetics and Genomics guidelines.

**Data Storage**

WES on an individual generates ≈6 to 12 gigabases (10⁹ bp) of data with >30000 variants. WES data analysis strategies focus on filtering out variants at different levels to only look at novel mutations in coding regions for prioritizing possible disease-causing variants. With increasing amounts of data from next-generation sequencing efforts shared across research sites, the database grows and data interpretation of new samples becomes more robust. One can easily imagine that integration of next-gen data from hundreds of thousands of individuals worldwide creates a powerful world genetic knowledge base, where defining the nature of any particular variant becomes increasingly accurate.

A database specific for WES data, the Exome Variant Server by NHLBI GO Exome Sequencing Project, has become an important resource for filtering variants as it contains data from WES (non-pathogenic variants) of >6000 individuals from black and European American populations. Although database resources are quickly improving, there remain technical differences between different next-gen sequencing equipment and bioinformatics methods that make the bioinformatics data resources a work in progress. For example, there was a recent report of a study where the same DNA samples were sent to a number of the top sequencing laboratories in the world, with comparison of the variant classification between sites. There was surprisingly little agreement in the variant analysis of the same individual (www.rd-neuromics.eu; January 24, 2013 workshop).

**Exemplar in Cardiogenomics: Genetics of DCM**

Dilated cardiomyopathy (also CMD) is a condition in which the heart’s ability to pump blood is lessened because of its weakened and enlarged state, eventually leading to heart failure. DCM is characterized by enlargement of the left ventricle and systolic dysfunction. It is the third most common cause of heart failure with an estimated frequency of 1:2500 in the general population. Around 20% to 48% of idiopathic DCM cases were found to have an underlying genetic cause (familial DCM/familial dilated cardiomyopathy). FDC was estimated to be found in 20% to 35% of first-degree relatives of patients diagnosed with idiopathic DCM and may be inherited in autosomal dominant, autosomal recessive, X-linked, or mitochondrial inheritance patterns. To date, mutations in 51 genes have been reported to cause DCM. Diverse inheritance patterns and a large number of causative genes make molecular diagnostics of DCM a challenge.

Until recently, patients and physicians only had the option of iteratively sequencing genes 1 at a time with high costs and long turn-around times. This involved send-outs to multiple laboratories and companies, often internationally, with each
negative result leaving a long list of genes yet to be tested. This was especially challenging for cardiomyopathies because of the inclusion of some of the largest and most complex genes in the genome (DMD, dystrophin; TTN, titin).

Over the last few years, several DCM gene testing panels have been introduced commercially in the United States—Harvard Medical School and Partners Healthcare: DCM panel, 28 genes; GeneDx: DCM panel, 38 genes; AmbryGenetics DCM panel, 37 genes; and Transgenomic DCM panel, 13 genes (GeneTests; DCM). These panels offer TS for coding regions (exons) of genes related to DCM. The results are available in a timely manner and are relatively cost-effective ($4000) as multiple genes are tested at once.

An example of where next-gen sequencing is making a major impact on the diagnosis of DCM is with regard to mutations in the titin gene and protein. The titin protein, as its name belies, is the largest known protein in humans, with 34000 amino acids, and a molecular weight of 3810 kDa (100x larger than average). For comparison, the average protein in the human body is 300 amino acids and 30 kDa in molecular weight.53 The gene encoding titin is also the most complex in the genome, with 363 exons used in many different alternatively spliced transcripts (Figure 3).54 The titin protein is a key component of muscle tissue (both skeletal muscle and cardiac muscle), where it serves as a structural and functional linker between adjacent sarcomeres (Figure 3), and giving structural integrity to the characteristic striated appearance of the actin/myosin network.55–57 Mutations in titin can be a cause of both skeletal muscle and cardiac muscle disease. Titin mutations can be autosomal dominant or recessive with phenotypes ranging from dilated/hypertrophic cardiomyopathy to skeletal myopathy/dystrophy, or both (Online Mendelian Inheritance in Man #188840). Titin’s enormous size has made it a molecular diagnostic nightmare using more classic Sanger sequencing approaches where each of the 363 exons was analyzed individually. As such, it was a prime target for next-gen sequencing approaches where the role of titin in cardiomyopathies and skeletal myopathies could be better defined.

A recent landmark study by Herman et al64 used both next-gen and Sanger sequencing methods to sequence all coding exons and splice sites of the titin gene in a large cohort of patients with DCM, hypertrophic cardiomyopathy, and normal controls. They found that a large proportion of DCM cases (27%) showed likely autosomal-dominant mutations in titin that result in shortening of full-length protein (truncations). Titin-truncating variants (nonsense, frameshift, splicing variants) were seen in >25% of familial cases (inherited mutations) and >18% of sporadic cases (de novo mutations). In the familial DCM cases, co-segregation analysis was highly supportive of a causative role of the identified titin mutations, with strong statistical support (combined logarithm of odds score of 11.1), and high penetrance of the mutations (>95% after 40 years of age). Most of the titin-truncating variants were located in the A-band region of the titin.

The large size of the titin gene and protein, and high frequency of mutations in DCM patients, suggests that the gene is a frequent target for mutational events. The clustering of DCM mutations in a subdomain of titin also suggests that there may be additional non-DCM phenotypes associated with mutations in other regions of the gene. Clearly, patients presenting with DCM should now have titin mutation studies as a front-line genetic test, and this is likely best done with TS.

Titin mutations have also been found through WES in patients with skeletal muscle disease. The first examples of using WES for titin mutations was in families with autosomal-dominant hereditary myopathy with early respiratory failure (OMIM #603689) by Ohlsson et al65 and Pfeffer et al.66 Both groups identified the same disease-causing mutation in titin (g.274375T>C; p.Cys30071Arg) in A-band titin and were able to provide statistical support for causality of this point mutation using family studies and linkage analyses. Interestingly, no cardiac abnormalities were reported in the affected patients.

All these studies show the power of next-gen sequencing to elucidate the role of the giant protein titin that can cause autosomal-dominant disorders of heterogeneous phenotypes with mutations situated in the same A-band region (DCM and hereditary myopathy with early respiratory failure). As pointed out above, the 10% drop-out rate of exons of titin when using whole exome sequence, compared with the 1% drop-out rate in TS, is an important technical distinction between the 2 approaches.

Figure 3. An overview of titin (also known as connectin) gene and protein. A. The gene for titin (TTN) is located on the long arm of chromosome 2 at position 2q31 with a huge size of 283 kb encoding 363 exons.54 B. The coding region of titin has been divided into different regions based on the protein’s location in the sarcomere.58,59 The N-terminal region of TTN contains the Z-band/disk region, which is anchored to the Z line of the sarcomere and comprises exons 1 to 28. L-band of titin roughly covers exons 28 to 251; this region undergoes extensive alternative splicing.10,11 The I-band consists of tandemly repeated Ig domains in the proline, glutamate, valine, and lysine domain that contains a repetitive motif of the amino acids: proline, glutamate, valine, and lysine.22 A-band of titin is encoded for by exons 251 to 358; exon 358 codes for the titin-kinase domain and is at the A/M junction. M-band titin is encoded by the last few exons 358 to 363 and is near the carboxy-terminal of titin attached to the M line of the sarcomere. C. A sarcomere is the contractile unit of the myofibril made up of myofilaments: actin and myosin, along with titin. The image depicts the structural arrangement of a single sarcomere along with the electron micrograph image of a human sarcomere. TTN filaments overlap at the M line, and adjacent TTN filaments overlap at the Z-disk.65,66
Earlier this year, Andreasen et al\textsuperscript{33} compared previously associated cardiomyopathy variants (missense and nonsense) with variants in the Exome Sequencing Project (exome variant server)\textsuperscript{41} database and found that a significant proportion of sequence variants published as pathogenic may in fact be variant of unknown significance or polymorphisms. Because phenotypic data from Exome Sequencing Project are not available, they defined cut-off values on the basis of estimated prevalence of DCM at 1:2500 and only picked variants above this cut off. Approximately 17\% (58 of 337) of variants known to be associated with DCM were present in the Exome Sequencing Project. Polyphen-2 analysis on these 58 variants predicted 33\% to be benign. The authors hypothesize that if these variants were to be disease causing, then a much higher prevalence of cardiomyopathy would be expected in the general population (1000-fold higher than expected). The conclusions from this study were that many cardiomyopathy-associated variants are wrongly classified as damaging and are probably false positives (benign polymorphisms). With more next-gen data being generated, it will be easier to identify benign polymorphisms among different populations resulting in fewer false positives.

Conclusions

As NGS technologies become more routine, the field of human genetics will receive an influx of information that may be in contrast with what was reported in the past. This represents a retrograde-storm before the calm situation where there might be initial unsettlement of established norms only to result in a greater understanding of human genetics. The future holds much excitement for uncovering new genotype–phenotype co-relations, development of bioinformatics tools to accelerate data analyses, and discovery of new genes related to disease.

Acknowledgments

We thank Dr Simon C Watkins for kindly providing us the electron micrograph scan of human sarcomere.

Sources of Funding

This work was supported, in part, by the National Institutes of Health (3R01 NS29525) and the Center for Genetic Medicine Research at Children’s National Medical Center.

Disclosures

None.

References


Punetha and Hoffman Next-gen and Cardiomyopathy 433


KEY WORDS: cardiomyopathies; cardiomyopathy, dilated; connec tin; exome; genetics; humans; Massively-Parallel Sequencing
Short Read (Next-Generation) Sequencing: A Tutorial With Cardiomyopathy Diagnostics
as an Exemplar
Jaya Punetha and Eric P. Hoffman

*Circ Cardiovasc Genet*. 2013;6:427-434; originally published online July 14, 2013;
doi: 10.1161/CIRCGENETICS.113.000085
*Circulation: Cardiovascular Genetics* is published by the American Heart Association, 7272 Greenville Avenue,
Dallas, TX 75231
Copyright © 2013 American Heart Association, Inc. All rights reserved.
Print ISSN: 1942-325X. Online ISSN: 1942-3268

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circgenetics.ahajournals.org/content/6/4/427

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published
in *Circulation: Cardiovascular Genetics* can be obtained via RightsLink, a service of the Copyright Clearance
Center, not the Editorial Office. Once the online version of the published article for which permission is being
requested is located, click Request Permissions in the middle column of the Web page under Services. Further
information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Circulation: Cardiovascular Genetics* is online at:
http://circgenetics.ahajournals.org//subscriptions/