



# BIOTECH Basics

## Large-scale Genome Sequencing

### What you need to know:

- Technological advances have made it possible to sequence large portions of an individual's genome to uncover disease-causing mutations.
- This approach is used extensively in research settings and is becoming more common in the clinic to help guide patient care.
- Currently, most projects sequence the exome - the gene-containing parts of the genome.
- A multistep process identifies DNA changes and determines their clinical significance.
- Only certain types of diseases are good candidates for sequencing. Disease-causing mutations will not be found in all patients and finding a mutation does not guarantee treatment availability.
- Several technological and educational challenges must be overcome as sequencing gains widespread use in patient care.

In the fall 2011 issue of *Through the Microscope* we introduced readers to Nicholas Volker, a young boy with an unusual, severe and unexplained gastrointestinal disorder. His condition progressively worsened and in 2009, after exhausting other testing options, his medical team turned to the novel approach of exome sequencing. This involved looking at the DNA sequence of Nicholas' ~23,000 genes to try and find the disorder's cause. A rare mutation in a gene involved in the immune response was identified, leading to a stem cell transplant that alleviated Nicholas' symptoms.

Since that time, exome sequencing has become an established part of many research projects focused on identifying disease-causing genetic changes. The National Institutes of Health estimated more than 70,000 research subjects had their exomes or entire genomes sequenced by the end of 2012. The technique has also gained a place in patient care, especially when the constellation of symptoms suggests the presence of a genetic syndrome but previous genetic tests have been negative. Exome sequencing casts a much broader testing net because it simultaneously examines every gene, rather than just a handful of potential candidates. This is the type of approach HudsonAlpha is initiating

in collaboration with the North Alabama Children's Specialists in Huntsville (see associated story on page one). In this edition of *Biotech Basics*, we revisit genome sequencing for research and healthcare purposes and explore the current state of the science.

### An introduction to sequence analysis

#### Sequencing

Historically, DNA fragments were sequenced and analyzed individually, one stretch at a time. This approach was used to produce the reference sequence for the Human Genome Project during the 1990s and early 2000s. Today's next-generation analysis systems simultaneously sequence billions of overlapping DNA fragments, making it possible to obtain the DNA information from the entire genome in a few or even a single experiment. This results in an enormous text file containing billions of As, Ts, Cs, and Gs. A series of analyses provides meaning to this string of nucleotides, identifying those DNA changes that are important for medical care (see figure).

#### Alignment

The sequence of each DNA fragment is first aligned to its corresponding location on

If you want to know more:

[genomebiology.com/2013/14/3/304](http://genomebiology.com/2013/14/3/304)

This link includes the electronic version of a report on the 6th annual Future of Genomic Medicine conference held this past March. While a bit technical in places, it offers a high-level overview of the coming impact of genomics on patient care.

[phenomena.nationalgeographic.com/2013/03/11/we-gained-hope-the-story-of-lilly-grossmans-genome/](http://phenomena.nationalgeographic.com/2013/03/11/we-gained-hope-the-story-of-lilly-grossmans-genome/)

Another example where genome sequencing identifies a disease-causing mutation, this time in the Grossman family, where daughter Lilly is affected with tremors and muscle weakness.

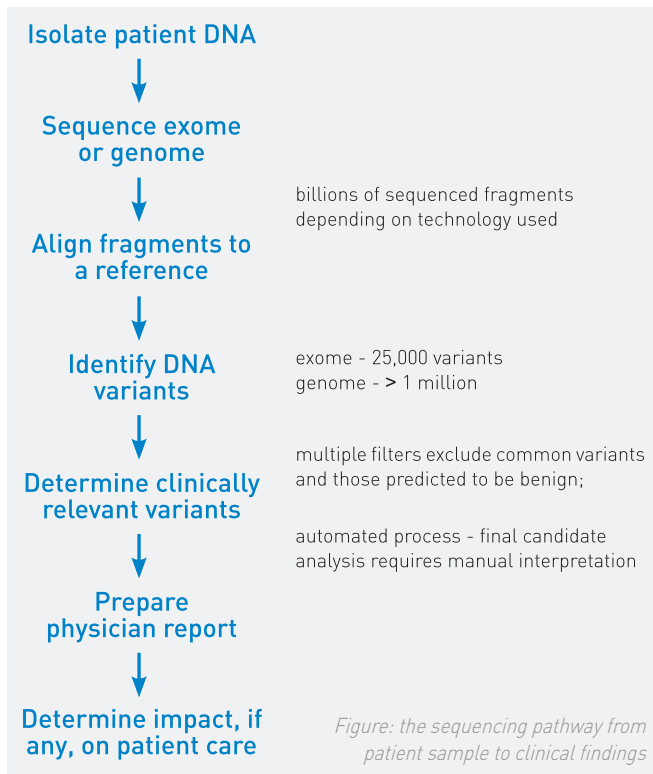
[www.the-scientist.com/?articles.view/articleNo/34730/title/Opinion--Genomics-in-the-Clinic/](http://www.the-scientist.com/?articles.view/articleNo/34730/title/Opinion--Genomics-in-the-Clinic/)

An opinion piece written for *The Scientist* magazine by Richard Resnick, the CEO of a genome software analysis company, highlighting the current use of next-generation sequencing in clinical settings.

[www.youtube.com/watch?v=PMIF6zUeKko](http://www.youtube.com/watch?v=PMIF6zUeKko)

Next-Generation Sequencing Technologies – A talk by Dr. Elaine Mardis, co-director of The Genome Institute at Washington University, given as part of the **Current Topics in Genome Analysis 2012** lecture series at NHGRI. This is a very detailed hour-long overview of different approaches in sequencing and the pros and cons of the various platforms. If you really want to dig into the specifics of next-generation sequencing, this is an excellent resource.





a reference sequence (produced by the Human Genome Project). This is essentially an enormous matching activity, placing billions of fragments in the correct order along each chromosome. Computer programs that control this alignment process must be tolerant enough to allow mismatches that represent genetic variation but exclude fragments that belong elsewhere in the genome. Each position in the DNA letter sequenced several times, providing greater confidence in the accuracy of the sequence. In a clinical setting this depth of coverage may be 50X or more, which provides confirmation that an identified variant represents a true DNA change.

### Variant detection

Additional software programs detect DNA variants - changes in the sequence relative to the reference. These include single letter variants, as well as insertions or deletions of varying size. Roughly 25,000 variants will be identified in each human exome. The vast majority is benign and has no clinical implications.

### Variant interpretation

Further computational analysis identifies a set of candidate "clinically relevant" DNA changes from within the population of variants. This step involves a series of filtering tools to exclude harmless variants commonly found in the population. These tools are somewhat crude as the field is still maturing, and this step is often the bottleneck in applying genomic information in the clinic. When

the list of potential candidates has been created, software algorithms are replaced by manual interpretation, usually performed by a clinician or clinical scientist. Each variant must be evaluated in light of the patient profile. If a candidate disease-causing variant is not identified at this point, the process begins again with a less stringent set of conditions.

In practice, genetic variants can be classified into one of four major categories: 1) those with clear clinical implications for the disease being studied, 2) those with a potential connection to disease but lacking conclusive evidence, 3) those with no known disease association and 4) variants of unknown significance. This last category is particularly challenging because these changes may be associated with disease, but haven't been previously identified or assessed.

### Making the diagnosis

In many cases, large-scale exome or genome sequencing will uncover a definitive (or at least plausible) biological explanation for a patient's disorder. Early estimates from laboratories that routinely perform these analyses suggest around a quarter of patients who undergo exome sequencing receive some sort of genetic diagnosis. Note that only a fraction of unexplained disorders are good candidates for exome sequencing. Disorders with multiple genetic and environmental risk factors, such as diabetes or cardiovascular disease, are difficult to diagnose by this process. For these conditions, each genetic factor impacts risk only slightly, and many genetic risk factors are still unknown. Consequently, it is hard to convincingly determine that a specific genetic change truly contributes to complex disease. For similar reasons, exome sequencing to screen for future disease risk in healthy individuals is not yet practical.

Discovering the genetic cause of a disorder does not guarantee the existence of treatment options. In many cases, our understanding of the biology behind the disease is incomplete and links to therapy have not been made. Even still, providing information about causation to patients and families answers the often-asked question of "Why?". It also brings a welcome halt to the parade of tests and screens they have encountered during

their diagnostic odyssey.

### Handling unexpected findings

In addition to the causative DNA change, exome or genome sequencing may also identify so-called incidental or secondary findings - variants that predispose to diseases unrelated to the initial reason for undergoing genomic sequencing. Every genome contains millions of variants, ranging in frequency from rare to common and carrying functional impacts from inconsequential to harmful. Potentially damaging, incidental findings include mutations that lead to adult-onset cancer, those that confer carrier status for recessive disorders or variants that uncover non-paternity within a family. Whether researchers and clinicians have a responsibility to actively search for and disclose secondary findings and if so, what specific findings should be assessed, is still under discussion.

Our understanding of the genome is evolving. As additional genomes are sequenced, the way we classify a given variant may change. What was initially thought to cause disease may later be reclassified as benign. In the short term, patient exomes and genomes will need to be periodically re-analyzed as the clinical implication of certain variants becomes clearer.

### Looking ahead

While exome sequencing has become an important tool for both research and clinical settings, the technology still faces several challenges. Even with the steep decline in costs, sequencing is still relatively expensive and involves significant computational and manual analysis. Our ability to interpret genetic variants and determine their clinical impact is still rudimentary. Genomic information is poorly integrated into patient medical records. In addition, most healthcare providers lack a solid understanding of genomics and there are few real-time tools that use genomic data to guide physician decisions. These hurdles will hopefully be overcome as the technology continues to mature and genomics assumes a larger role in patient care.

As costs continue to drop, an analysis of the entire genome will likely replace exome-based methods, uncovering new disease-associated DNA changes, especially those that regulate how genes behave. This approach will also increase the number of variants of unidentified significance, requiring an even greater level of interpretation.

- Neil Lamb, Ph.D.

director of educational outreach  
HudsonAlpha Institute for Biotechnology