Sorting out sequencing data

Monya Baker

The toughest work is not sequencing a genome: it is finding the mutations that matter.

Alexis Beery literally breathes easier after having her genome sequenced. She and her twin brother, Noah, were diagnosed with dopa-responsive dystonia at age five, but when Alexis developed a life-threatening cough six years later, doctors thought it was an unrelated condition. Luckily for them, their father, chief information officer at Life Technologies, got the twins into a study as a part of which their genomes were sequenced. Results revealed mutations in a gene that encodes an enzyme crucial for making the neurotransmitters serotonin and dopamine[1]. Although the twins were already taking a precursor of dopamine for the dystonia, the sequencing results indicated that a precursor for serotonin could help, too. It did. Now Alexis's breathing problems are gone, and her brother's health has also improved[2].

Alexis and Noah are just two of thousands of people whose genomes have been sequenced, and that number is growing. The first full human genome sequence required over a decade and roughly three billion dollars. Now, a high-coverage sequence can be had in less than ten weeks, at a raw cost of around $4,000.

The result is a surfeit of data on human genetic variation, with researchers struggling to work out the best ways to extract useful information from the data. The focus has shifted from generating sequences to hunting for the critical sequence differences that distinguish individuals from each other and health from disease. This requires a multistep analysis and is much less certain than identifying a string of nucleotides.

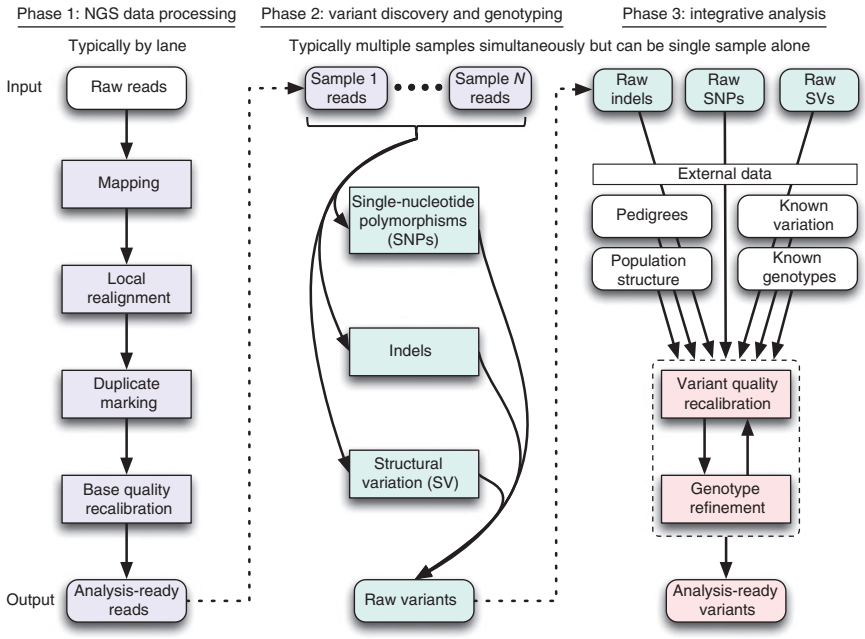
"The excitement about the ability to sequence human genomes is justified," says David Goldstein at Duke University. "But it's also true that one of the frustrating things is that almost everything you do affects the sequence that comes out the other side. This is a big, complicated and dynamic enterprise, figuring out sequencing data. And it's changing every month."

Few can expect benefits as dramatic as the Beery twins'. Right now, in fact, few people can expect to benefit at all. For one thing, finding the cause of a disease is usually easier than finding a treatment. For another, determining what causes a disease is rarely so straightforward. If personal sequencing data are to become more useful, researchers not only need more data from more genomes but also need to get better at identifying variants and figuring out their biological relevance.

Getting to variants

Some variants are unique to a particular individual; other variants will be shared by others with a similar trait or disease. Finding variants in genomes is like assembling and comparing incomplete jigsaw puzzles of two nearly identical pictures. It can be hard to tell whether differences are from how the puzzle pieces are cut or from the pictures themselves.

The process of identifying variants requires highly specialized software. Next-generation sequencing technologies break the three-billion base pair human genome into pieces; depending on the machine being used, sequenced fragments, called 'reads', can be as short as 50 base pairs or longer than 1,000 base pairs. These reads are aligned to and compared with a reference genome. The number of variants, or places at which an individual genome differs from the (acknowledged to be



Pipelines bring together algorithms to identify genetic variation from sequencing data. Reprinted from reference 3. NGS, next-generation sequencing.

CORRECTED AFTER PRINT 28 OCTOBER 2011

imperfect) reference genome, runs into the low millions. Compiling these long lists is even more complicated than it sounds: every step is riddled with errors. But by learning the patterns of errors that machines make, researchers are getting better at detecting artifacts. "The thing that we really assume is that real genetic variation doesn't look like machine error," says Mark DePristo, who leads the group developing algorithms for next-generation DNA sequencers at the Broad Institute.

Some of the first error-catching filters are for redundant and faulty reads. These can be identified in many ways. One means of error correction relies on the fact that sequenced fragments incorporate mistakes faster as they get longer. This knowledge can be used by researchers to 'trim' sequences back, or to choose which sequence is more likely to be correct when judging discrepancies between overlapping reads.

An ongoing challenge is that machine-generated variation keeps changing. New sequencing machines are constantly being introduced. Reads get longer and more accurate. Commercial kits are also getting better for amplifying and sequencing exomes, the protein-coding portions of the genome. Although these improvements make for reads that are easier to align, they also force a constant redesign of error-catching filters.

DePristo and colleagues recently published a description of the genome analysis toolkit (GATK), which runs sequencing data through a series of analysis and error-correcting algorithms to produce a list of variants[3]. It can make sense of very different data sets: exome sequencing, very high, (60×) coverage sequencing (meaning that each base is read an average of 60 times) and very low (4×) coverage sequencing. The 2011 publication used sequencing data produced by machines sold by Illumina, Life Technologies and 454 Roche, but DePristo believes that the toolkit can work across more

BOX 1 DESIGNING A SEQUENCING PIPELINE FOR A CLINICAL SETTING

Many rare, inherited diseases have still not been pinned to a particular mutation. Even when the genes behind a heritable disease are known, sequencing can be used to find genetic variants that established tests miss. Although some clinical genetic testing services may screen as many as 100 genes, these tests generally look for previously observed mutations rather than potential unknown errors. In an effort to diagnose rare diseases in very young individuals, Stephen Kingsmore and colleagues have taken their sequencing analysis expertise, developed at the US National Center for Genome Resources, and are applying it at the Center for Pediatric Genomic Medicine at Children's Mercy Hospital. They plan to use sequencing to look for new variants in a set of nearly 600 genes associated with childhood diseases, with follow-up studies expanding the search to exome sequences. Even though important mutations certainly exist outside of that set, says Kingsmore, the goal is to focus on mutations that will be easier to find and interpret.

Stephen Kingsmore at Children's Mercy Hospital is developing a pipeline for finding variants that cause rare diseases. If clinicians are to make sense of the data, he says, "you need something highly standardized and authoritative."

Such a clinically focused project puts special demands on variant-calling pipelines. For neonatal testing, speed is paramount, a constraint that makes matching reads to each other for paired-end analysis impractical, says Neil Miller, deputy director of informatics at the center. Instead, the algorithms rely on the first 100 letters of a typical 120-nucleotide read, discarding the most error-prone regions.

In addition to being fast, the pipeline must be simple, with minimal manual tweaking of parameters. Most importantly, it must reduce the number of false calls. Focusing on just a few hundred genes means that the project will catch less of the variation, but for this kind of diagnostic approach, the trade-off is worth it, says Kingsmore. "We have to produce information that's not just going into a manuscript, it's going into a hospital chart."

Mark DePristo of the Broad Institute says algorithms are getting better for finding variants from sequencing data, but tools so far are much better at finding single-nucleotide variants.

platforms. In fact, the team is about to publish results using instruments from Pacific Biosciences, which began shipping its machines only in April 2011. "We have a model for true genetic variation, and we can apply that to derive what are real variations and machine artifacts," says DePristo. "We are learning the error process of each individual run of the machine."

Software tools to analyze sequencing data are expanding rapidly, so many bioinformatics teams are looking for ways to combine these algorithms into straightforward workflows. These efforts result in 'pipelines', or assemblies of algorithms that often combine well over 100 components.

GATK is probably the most widely used pipeline: a cloud-based logging service records 100,000 runs of the program a day, says DePristo. SAMtools, an open-access program produced collaboratively by the Broad and Sanger institutes, is also popular[4]. Fee-for-service sequencing offerings from BGI, Complete Genomics and Illumina have their own variant-calling pipelines for the sequences they produce. Informatics company Accelrys offers Pipeline Pilot, a suite of visual programming tools for integrating open-source, proprietary and third-party sequencing algorithms. It handles data from 454 Roche, Illumina and Life Technologies, and it will be shipped with Oxford Nanopore when that company launches its single-molecule sequencing system.

Bioinformatics teams in drug companies and academic centers alike are struggling with a veritable plethora of parameters. Hunting for acquired cancer mutations requires comparing an individual's tumor cells to healthy cells, for example, and so is very different from scanning for rare, inherited variants that raise susceptibility to disease, explains Thomas Barber, genetics group leader at Eli Lilly. "Different pipelines give different results, and even within a pipeline, tweaking parameters gives different results, and that's not always a bad thing." Though time and computing cost prohibit doing so on every genome, he says, it is prudent to run the same genome through several pipelines at the beginning of a project to determine what sort of analysis works best for a particular biological question.

Exomes versus whole genomes

Researchers sometimes winnow the amount of sequencing and analysis they need to perform by focusing only on exomes. This represents about 1% of the total sequence and costs about a tenth of the price of sequencing a whole genome. Commercially available kits from companies including Agilent and Roche NimbleGen selectively copy these regions from individuals' DNA. In a typical

experiment, sequences from closely related subjects with and without a disease are examined. Variants linked with disease are identified and then further verified with focused sequencing in additional individuals and biological follow-up experiments.

The approach has had dramatic success. Articles published in just the first two weeks of August 2011 used exome sequencing to link rare mutations with aciduria, ataxia, polymicrogyria (a severe neuronal development disorder) and retinitis pigmentosa (an eye disease that leads to blindness). Sequencing experts are beginning to pair with clinicians in hopes of finding solutions to puzzling diseases (**Box 1**). In March of 2011, researchers at the Medical College of Wisconsin described the diagnosis of a 15-month-old boy with severe, intractable inflammatory bowel disease. Exome sequencing identified over 16,000 variants; subsequent analysis linked the boy's condition to one: a previously uncharacterized mutation in a gene that inhibits normal programmed cell death. The discovery led to successful treatment with a bone marrow transplant[5].

Advantages of exome sequencing include lower costs, fewer information technology infrastructure requirements and simplified analysis pipelines; it also generates data on the mutations that are easiest to understand and act on. But it is very incomplete. The approach may capture as many as 90% of protein-coding genes, but it neglects the other approximately 99% of the genome. Even for protein-coding genes, exome capture approaches are unable to detect copy number variants and larger structural differences. For diseases rife with such mutations, whole-genome sequencing makes the most sense, says Elaine Mardis, co-director of technology development at the Genome Institute at Washington University in St. Louis. "From a discovery standpoint in cancer, we've resisted the urge to go after just the exomes."

In the past couple of years, researchers studying all sorts of diseases have begun turning to whole genomes, says Radoje Drmanac, chief scientific officer of Complete Genomics. Not only has the price of sequencing come down, he says, but also researchers are learning how to make sense of noncoding portions of the genome. "The cost [of whole genomes] is still higher [than exomes]," he says. "But the value is higher still."

Sequencing and computational analysis are getting faster and cheaper, but finding biological meaning remains difficult, says Elaine Mardis of Washington University in St. Louis.

Digging deep or wandering wide

Even if researchers opt for whole-genome sequencing, they still need to decide how deep to go in the sequencing and analysis. Deeper analysis allows more variants to be identified with greater confidence, but shallower analysis allows more genomes to be studied, providing the sample sizes necessary to confirm that a variant with small effects is significantly associated with disease. "You don't have to know all the variation in an individual," says DePristo, who is participating in studies to find variants that contribute to risk for autism, diabetes and heart conditions. "Unless you're trying to find really rare things, you are better off looking for common variants that are skewed between cases and controls."

Whether researchers should look for rare mutations with large effects or common ones with small effects is a matter of debate and personal preference. Michael Snyder, director of the Stanford Center for Genomics and Personalized Medicine, wants to find as many variants as possible, particularly because researchers are still figuring out how to make sense of the data. "The number-one problem, whether it's single-nucleotide polymorphisms or structural variations, is missing them," he says. In a few cases, Snyder has had the same genome sequenced twice with different technologies. (Illumina tends to find more variants but also has a higher rate of false calls, he says. Complete Genomics identifies fewer variants but gets more correct.)

Michael Snyder of Stanford University is working on a variant-calling pipeline that can easily incorporate new algorithms.

Finding variants depends on more than the sheer number of reads. Snyder, who is developing his own pipeline called

HUGESeq, sometimes uses different sequencing technologies and pipelines on the same genome in order to identify more variants. The lists identified by GATK and SAMtools can vary by as much as 5%, he says. His team uses multiple algorithms to comb the genome for duplications, deletions and other structural variants: paired-end sequencing finds transposons and translocations but works less well for large repetitive regions, scanning for changes in read depth (the density of overlapping reads within a certain region) often works well for identifying copy number variants, and split reads (cases in which the same read is split so that the two parts align to different places in the genome) can help identify duplications and deletions. This kind of information is essential to plan the best follow-up studies, he says. "Variant calling is critical because it leads to decision making down the road."

Making a list and checking it twice

In addition to producing lists of variants, pipelines create confidence scores, an estimate of whether a variant is a real or false call. Although current software uses various tricks to weed out false positives, it is still hard to be certain that a mutation detected by software actually exists. "When dealing with 100 billion pieces of data for a given genome, there will be errors," says Barber. "We need validation before we can rely on that data."

The first step is to look at the relevant reads for red flags. The next step is conducting additional experiments. "We take all the regions that we think are mutated, and we go back to the genome with a secondary method," says Mardis. Sequencing RNA from transcribed genes (a technique called RNA-seq) can be a quick way to verify that the expected variants are there. (Initially, RNA-seq was considered a potential alternative to whole-genome sequencing, because it would reveal only transcribed regions of the genome. In practice, though, the technique has too much background noise.) Mardis also designs PCR experiments to verify the breakpoints in a translocation.

The gold standard, however, is Sanger sequencing, the slow, labor-intensive technique that produced the first human genome in the years before faster, cheaper next-generation sequencing techniques appeared. Snyder turns to this method even if two independent sequencing platforms have identified the same variant. "If it's something you care about, you need to Sanger validate," he says.

In fact, as examples accumulate of sequencing data being used for clinical diagnosis, many clinicians and researchers are clamoring for guidelines to help them decide when information about variants should be acted upon by doctors or disclosed to individuals. Sequencing data routinely produces "artifactual discoveries," says Duke University's Goldstein. "You have to be really careful and not overinterpret." Clinical review boards at hospital centers across the globe are struggling to develop appropriate guidelines for assessing confidence in and acting on sequencing data when treating individuals.

Once variants are verified, patterns of errors can be used to improve variant-calling pipelines, but the real value is interpreting what a variant means biologically. "While we're very sophisticated about finding mutations, we're not very sophisticated in figuring out what these mutations mean for cell biology," Mardis explains. Understanding cancer is not just about finding the mutations. "What's really become important is an even more integrated analysis: gene expression, methylation status, microRNA expression and similar factors all have to be accounted for," she says. "We're not only looking at the sequence."

Emerging computational tools can help researchers decide which variants to focus on. Complete Genomics and other providers supply tools that can extract protein-coding and transcription factor–binding variants. Shareware programs such as SIFT and PolyPhen-2 predict how a given mutation might affect phenotype. Software company Omicia and scientists at the University of Utah recently released a program, variant annotation, analysis and search tool (called VAAST), which uses probabilities to identify potential disease-causing variants from harmless counterparts[6]. When presented with variant lists from surprisingly few individuals, VAAST quickly plucked out mutations known to be responsible for the rare genetic disease Miller syndrome. In a separate analysis, the program identified mutations for a formerly unrecognized neurodegenerative disease linked to a gene on the X chromosome. Researchers say such programs can be useful for prioritizing variants, but reliably predicting the functions of any specific variant is still far beyond a computer program.

Catalogs of previously discovered variants can also help, though existing databases are mainly restricted to easily identified variants like SNPs. Researchers hunting for the genes responsible for rare diseases, for example, discard candidates that are found too often in the general population. Those working on more complex diseases can use such catalogs to help verify their findings.

Ultimately, though, pinning a variant to a function is the part of science least amenable to high-throughput approaches, says DePristo. "You can easily generate data that now demands years' worth of work to digest."

1. Bainbridge, M.N. *et al. Sci. Transl. Med.* **3**, 87re3 (2011).
2. Check Hayden, E. *Nature* advance online publication, 15 June 2011 (doi:10.1038/news.2011.368).
3. DePristo, M.A. *et al. Nat. Genet.* **43**, 491–498 (2011).
4. Li, H. *et al. Bioinformatics* **25**, 2078–2079 (2009).
5. Worthey, E.A. *et al. Genet. Med.* **13**, 255–262 (2011).
6. Yandell, M. *et al. Genome Res.* **21**, 1529–1542 (2011).

Monya Baker is technology editor for *Nature* and *Nature Methods* (m.baker@us.nature.com).

Erratum: Sorting out sequencing data

Monya Baker

Nat. Methods **8, 799–803 (2011); corrected after print 28 October 2011.**

In the version of this article initially published, a figure was incorrectly attributed. It is reprinted from reference 3. The error has been corrected in the HTML and PDF versions of the article.