Structural variation: the genome's hidden architecture

Monya Baker

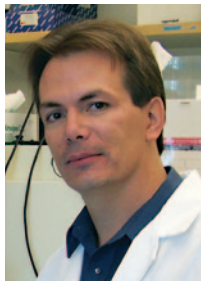
Next-generation sequencing is uncovering more variants than ever before, but it also faces limitations.

The Austrian monk Gregor Mendel may have founded the science of genetics, but his ideas now limit genomic studies, according to Jim Lupski, a molecular geneticist at Baylor College of Medicine in Texas. "Mendelian thinking is to genetics as Newtonian thinking is to physics. We saw a whole new world when Einstein came along."

According to Mendelian principles, individuals inherit exactly two copies of each gene—one from each parent. Genes on sex chromosomes have long been recognized as exceptions, but genetic deletions and duplications also break the rules, and in ways that are much harder to track.

In the early 1990s, Lupski and colleagues found that the hereditary neuropathy Charcot-Marie-Tooth disease was associated not with a flawed version of a particular genetic region but instead with the presence of an extra copy of the normal version. "We all believed Mendel for many, many years, but when you have a duplication, you are triallelic rather than biallelic," he says. Experts were so unwilling to accept the idea of a disease caused by a gene-dosage effect that both *Science* and *Nature* declined even to send his paper out for review, he recalls. When Lupski did publish, in *Cell*, his lab used four independent methods to substantiate his conclusions within a skeptical scientific community[1].

Despite additional findings like this, and the sequencing of the human genome, studies are still built on Mendelian assumptions, says Lupski. The concept of paired homologous chromosomes is taught in high school biology; the term 'paralogous', which refers to clusters of identical or near-identical sequences at different chromosomal locations within the same genome, is much more obscure. Genome-wide association studies, which find correlations between a disease population and specific genetic variation, tend to take diploidy as the default. Even microarrays designed to assess copy number generally assume that most individuals carry exactly two copies of any particular region, which can throw off some calculations.



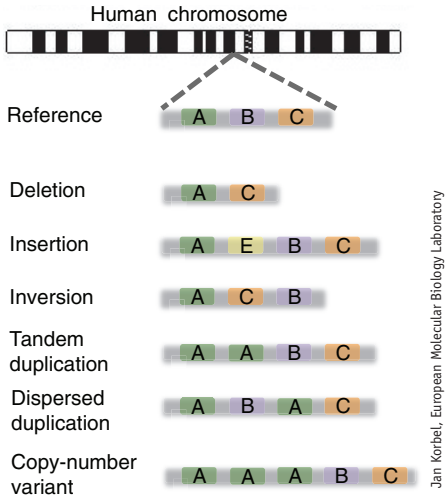
Next-generation sequencing is revealing new variation, but it won't be able to find everything, says Evan Eichler at the University of Washington, Seattle.

With the exceptions of very large variants that can be caught under a microscope, finding copy-number variations has been difficult, says Evan Eichler, who studies genome sciences at the University of Washington in Seattle. "Every technology developed in genomics up to about 2007 is really biased toward typing unique tags in the genome," he says.

And copy-number variation is just one type of a class of previously overlooked differences now collectively known as structural variation. This catch-all category includes insertions, duplications, deletions, inversions, recurring mobile elements, and other rearrangements, now usually defined as those covering 50 or more base pairs (**Fig. 1**). (The number is arbitrary; earlier definitions set the number at 1,000 base pairs until sequencing technologies capable of detecting smaller variants drove it down.)

There is a growing recognition that structural variation is pervasive and important. Since the publication of seminal papers[2,3] in 2004, the numbers of references to structural variation in the scientific literature and entries in the curated archive Database of Genomic Variants have soared (**Fig. 2**). It is now recognized that, in terms of the number of nucleotides, structural variation accounts for more differences between human genomes than the more extensively studied single-nucleotide differences. A 2010 study estimated that such "non-SNP variation" totaled about 50 megabases per human genome[4].

Conditions including autism, schizophrenia and Crohn's disease have all been associated with structural variation. And uncovering structural variation will be essential for understanding heterogeneity within tumors, says Jan Korbel, who studies structural variation at the European Molecular Biology Laboratory in Heidelberg. "There will be mechanisms that lead to cancer that no one would have thought of a year ago," he predicts. Last year, researchers at the Broad Institute of Harvard and MIT in Cambridge, Massachusetts, sequenced the entire genomes of both normal and cancer-



Jan Korbel, European Molecular Biology Laboratory

Figure 1 | Structural variation occurs in all forms and sizes. Genome structural variation encompasses polymorphic rearrangements 50 base pairs to hundreds of kilobases in size and affects about 0.5% of the genome of a given individual.

ous tissue taken from seven men with prostate cancer: point mutations were relatively infrequent, whereas chromosome rearrangements were much more common[5].

Researchers have only recently begun mining short-read sequencing data for structural variation. "Next-generation technologies are opening up a new zone of discovery," says Eichler, adding that sequencing still produces many artifacts that can send researchers on wild goose chases, and it overlooks many variants. "We pat ourselves on the back when we find new structural variants, but we're missing 50% of the variants out there because of the limitation of our methods and our technology," he says.

Uncovering variants

The extent of copy-number variation was first demonstrated using microarrays. These can analyze the greatest numbers of samples at the lowest cost, essential for achieving the population sizes necessary to associate rare variants with disease. One study that tied schizophrenia risk to duplications in a neuropeptide gene began with a genome-wide hunt for copy-number variations in 802 patients and 742 controls[6]. Such sample sizes would not have been feasible with sequencing, and nor would the follow-up analysis of 114 'regions of interest' assessed in an additional 14,177 subjects.

However, the information that arrays can reveal about structural variation is limited. Arrays for comparative genomic hybridization and SNP arrays are both used in structural variation studies, but they can detect only sequences that match the oligonucleotide probes used to make them, and these probes are usually biased against 'difficult', highly repetitive regions. Custom-made sets of arrays with tens of millions of probes may find variants as small as 500 bases, but the use of such arrays is not feasible for huge numbers of samples. For the most commonly used arrays, the limit of detection is usually much larger, generally on the order of 5 kilobases, and greater for highly repetitive sequences. And although arrays can detect that a sample has more or fewer copies of a region compared to a reference genome, they generally cannot determine an absolute number.

Another advantage that sequencing offers is in the ability to find 'breakpoints', the sequence boundaries where a structural variant begins and ends, says Korbel. Without knowing the breakpoints, it is hard to track a variant across a population and even more difficult to understand the functional impact a variant might have or the mechanisms that produced it.

But sequencing still misses considerable variation. In next-generation sequencing, genomic DNA is shredded into fragments of manageable size. These fragments are partially sequenced as 'reads', usually around 80–150 bp long.

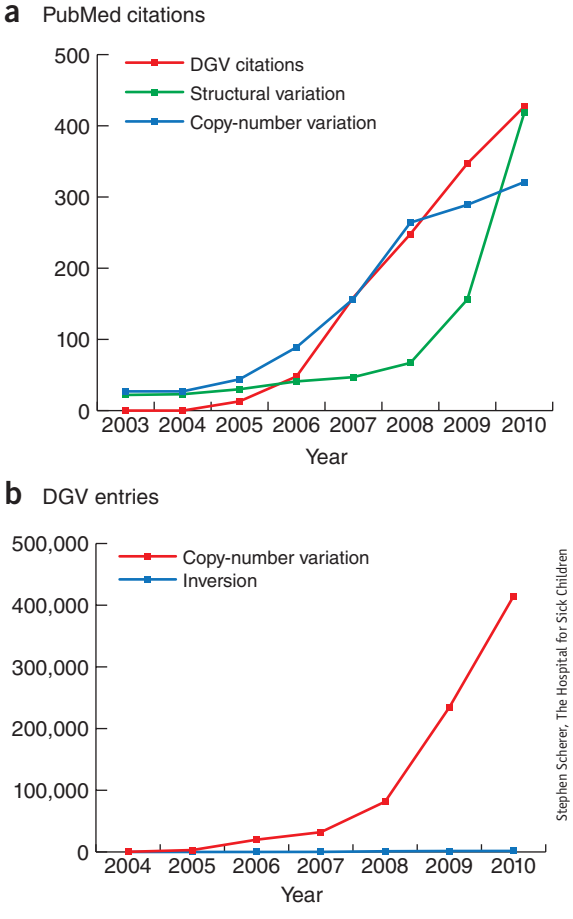
Reads are then aligned to the reference genome to check for differences. When reads match precisely to unique spots, alterations in a handful of nucleotides are readily apparent, revealing single-nucleotide variations. Finding structural variation is more complex; analysis must make sense of partially aligned reads and sort out repetitive sequences.

More structural variation is being uncovered than ever before, says Jan Korbel at the European Molecular Biology Laboratory in Heidelberg.

In the past few years, scientists have produced an alphabet soup of algorithms to hunt out structural variation from sequencing data. These typically use one of a handful of strategies (**Fig. 3**). 'Read depth' refers to the number of sequenced fragments mapped to a particular part of the genome and can indicate how many copies of a region are present. In 'split reads', a single sequenced fragment maps to two parts of the reference genome that are far away from each other. That means that those pieces are next to each other in the donor genome, a situation that may indicate a deletion, insertion or inversion.

More complex but arguably more powerful are paired reads, which work as a kind of molecular calipers. First, the genome is precisely divided into molecules of known size: these might be 500-bp fragments, 3-kb fragments and 40-kb plasmids. Rather than sequencing the entire stretch of DNA, a task impossible for current high-throughput sequencing machines, reads are taken only from the ends. If these appear too close together (for example, the ends of a 3-kb fragment map to sequences on the reference genome that are 2 kb apart), then the newly sequenced genome may harbor an insertion. If the ends appear too far apart, the genome may have a deletion.

None of these measures reveals everything about a variant. Read depth can indicate how many copies are present, but not where the copies occur, for example. And the shorter the read is, the less likely it will be to reveal breakpoints, says Michael Brudno, a computer scientist at the University of Toronto: "Once the break is in a repeat and the short read cannot span that repeat, you don't know where the breakpoint really is." In general, the more repetitive a region is, the harder it is to analyze, says Brudno. "If a read matches in two separate places in the genome, you don't know which is right."



Stephen Scherer, The Hospital for Sick Children

Figure 2 | Entries for structural variation are increasing in the scientific literature (**a**) and in the Database of Genomic Variants (DGV; **b**), which posts curated data from peer-reviewed studies on human samples. It draws from two other databases (DGVa and dbVAR) that accept open submissions for data.

Another strategy uses aberrations from the reference genome to identify loci where structural variants might be, and then assembles reads just for that area. This does eliminate some biases introduced by the reference genome, says Ira Hall, a molecular geneticist at the University of Virginia School of Medicine, but it is still far from perfect. In particular, he says, assembly approaches tend not to deal well with heterozygosity, when one variant occurs on only one of a pair of homologous chromosomes.

Because individual algorithms tend to specialize in finding variants of particular sizes and types, researchers often use several algorithms together. Almost exactly a year ago, the 1000 Genomes Project described an effort to find structural variations using next-generation sequencing data from185 human genomes[7]. It used 19 algorithms to identify over 22,000 deletions plus 6,000 other variants such as insertions or duplications.

Combining algorithms effectively does not mean pooling results indiscriminately, says Korbel, one of the leaders of the study. The 1000 Genomes Project, for example, did not use all variants from all calling algorithms—instead it evaluated each algorithm's rates of false positives experimentally and included variants identified by less specific algorithms only if they were verified with microarrays or PCR.

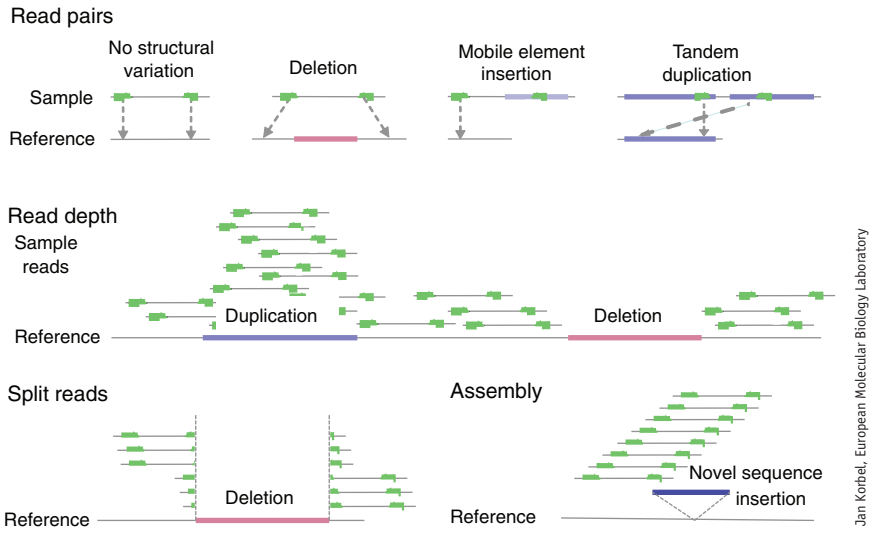


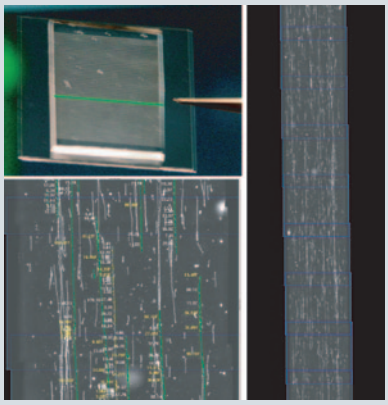
Figure 3 | Several analytic techniques are used to find structural variation.

Calling for more algorithms

Charles Lee, a cytogeneticist at Brigham and Women's Hospital in Boston and among the first researchers to associate gene duplications with disease, says that one of the take-home lessons of the study is just how

BOX 1 GETTING A BIGGER PICTURE

Short-read sequencing technologies are uncovering a wealth of new structural variants, but 150-base-pair reads can go only so far to build up an accurate picture of the six billion base pairs in a diploid human genome. Larger-scale analyses can show how many genes are present and in what orientation, as well as which versions occur together on the same chromosome.



David Schwartz, University of Wisconsin

Patterns in single stretched-out DNA molecules extending for hundreds of kilobases can identify structural variants across the genome.

One approach to such analyses is clever sample preparation followed by sequencing. Independent technologies from the laboratories of Jay Shendure at the University of Washington in Seattle and Stephen Quake at Stanford University in California may soon help to sort that out. Quake has a microfluidics system that physically separates all the chromosomes of a single cell; DNA on each chromosome can then be amplified in isolation[9]. (So far published work has only shown genotyping in SNPs, but Quake is working toward additional applications.) Shendure fragmented a genome into a very picky type of bacterial plasmid that accepts only about 40 kilobases at a time. These could then be grown up in pools and sequenced, and the resultant data could be built up into 400-kilobase regions, each derived from a single chromosome[10].

Other approaches get structural information from single molecules by directly analyzing extremely long stretches of DNA. One such approach is called optical mapping. This starts with a tube of long DNA molecules that have been isolated from a biological sample. These are stretched out onto a positively charged glass surface to which the negatively charged DNA molecules adhere firmly. The surface is then dipped into a solution of restriction enzymes; cleaved DNA recoils, leaving gaps that can be seen under a microscope. An entire human genome can be laid out on four patches, each about 1 centimeter square, says David C. Schwartz at the University of Wisconsin–Madison, who developed the technology. Next, the entire sample is meticulously photographed at high resolution, essentially creating a data set of a million single-molecule restriction maps. Sophisticated software assembles the fragments and then identifies inversions, translocations and duplications. "We know the size and position of each of those 325,000 base-pair fragments. We can tell you which have gotten bigger, which are smaller and which have been rearranged," explains Schwartz, who has performed optical mapping of bacterial genomes and several plant and animal genomes, including human[11]. However, the system is not something other researchers could easily set up on their own, he says. In October 2011, OpGen Inc., a company Schwartz founded, announced the launch of GenomeBuilder, a commercial system to use optical mapping for analyzing structural variation in large genomes.

Meanwhile, Schwartz, who is no longer affiliated with OpGen, is working on a new system called Nanocoding with the goal of having it be something that dedicated scientists can build in their own labs. With this technology, DNA is introduced into nanochannels that precisely control how much the DNA is stretched and so how many nucleotides there are per given length. The DNA is treated with 'nicking' restriction enzymes, which cut only one strand of DNA; the nicked strands can then be labeled with dye and imaged. The system makes it possible to explore variation from about 2 kilobases up to entire chromosomes, says Schwartz. (For people who don't want to build their own system, a company unaffiliated with Schwartz, BioNano Genomics, offers a product that works using similar principles.)

Ultimately, Schwartz says, studies of long DNA molecules can have even better resolution if combined with sequencing technologies performing alongside long, immobilized DNA strands. The reads produced are still short, he admits, but researchers will know where they came from.

Structural variation is more difficult to assess than single-nucleotide variation, says Charles Lee at Brigham and Women's Hospital.

far the hunt to catalog structural variation has to go. "Next-generation sequencing is a new area for everybody," he says. "There's a lot of work that has to happen." With very high-quality data (40× to 60× coverage), he estimates, sequencing data can find about 80% of known deletions but less than 20% of duplications. As for inversions and so-called 'copy-neutral variation', he says, the fraction detected so far is still insignificant.

The algorithms for calling structural variants from sequencing are just not accurate enough for results to be believed without additional confirmation, says Stephen Scherer at the University of Toronto, who has also linked duplications with disease. "If you want complete structural variation data now, you need to do whole-genome sequencing and also run microarrays."

Algorithms to find structural variants can be hard to evaluate individually. Because they generally target particular classes of variants, doing rigorous comparisons is difficult. "There hasn't been to my knowledge really good benchmarks of many of these algorithms—of course everyone's algorithm does the best when you read the paper [describing it]," says Ben Raphael, a computer scientist at Brown University in Rhode Island. Academic institutions and sequencing services companies, such as Illumina and Complete Genomics, are intent on improving their abilities to identify structural variation. So far, though, offerings are not yet mature. "The most powerful algorithms right now are all open source," says Korbel.

But scientists are learning how to make more sophisticated algorithms, in particular by combining multiple kinds of analysis—for example, considering paired ends in the context of split reads or read depth. "The types of variants that can be discovered by each method are different, so bringing all these together is the most promising," explains Brudno. Another emerging strategy is to write programs designed to analyze reads that support multiple variants. In repetitive regions, explains Raphael, reads can have many good alignments. Instead of making a call for each variant individually, alternative variants will be considered simultaneously.

"We want to push the limits of getting harder and harder variants," he says.

Even relying on a reference is a limitation. If read lengths were extended to 1,000 base pairs, that would still not be long enough to uniquely map regions characterized by very long repeated elements. What's more, the current reference genome, itself a compilation of several individuals, is incomplete, particularly around the centromeres and telomeres, regions known to be both highly repetitive and highly variable between individuals. If a genetic rearrangement involves DNA missing from the reference genome, there is no way to map it.

The obvious solution is to do away with the reference genome and piece together each newly sequenced genome from scratch. The sequencing company BGI recently described a *de novo* assembly of two human genomes and reported over a quarter-million variants ranging from 1 base pair to 23 kilobases, mostly insertions and deletions[8]. However, the analysis was designed to detect homozygous variation, and researchers are not convinced that current *de novo* assembly methods are ready to outstrip mapping techniques. In the latest study, about a quarter of the genomes was inaccessible to assembly, and the inaccessible regions are probably particularly rich in structural variation, says Brudno. "Until we have better reads with much better quality, you want to use the billions of dollars invested in the whole human genome," he says. (In fact, the interplay between read length, accuracy and cost is a topic of continuing discussion among experts, particularly as sequencing platforms continue to improve.) But those improvements are coming, along with better analysis techniques. "Everyone in the field agrees that we should be assembling genomes," says Hall. "What there is controversy about is how soon it will be before we can do that in a meaningful way."

Ultimately, the goal of finding structural variation will not be achieved by any existing technology, says Eichler, but through technological shifts to additional methods. Researchers need better ways to scan for large-scale variation and to tell which variants occur together on the same chromosome (**Box 1**). Most needed, though, are better ways to simultaneously analyze the full range of genetic differences that can occur in a single individual. "The strength will be when we can go in and integrate across all the variation, irrespective of class, type, variant and frequency," Eichler says. Structural variation is not a nascent field, he reflects. It's just another "type of variation that has been under-studied. I hope in ten years, people will just be studying the full spectrum of genetic variation."

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

1. Lupski, J.R. *et al. Cell* **66**, 219–232 (1991).
2. Iafrate, A.J. *et al. Nat. Genet.* **36**, 949–951 (2004).
3. Sebat, J. *et al. Science* **305**, 525–528 (2004).
4. Pang, A.W. *et al. Genome Biol.* **11**, R52 (2010).
5. Berger, M.F. *et al. Nature* **470**, 214–220 (2011).
6. Vacic, V. *et al. Nature* **471**, 499–503 (2011).
7. Mills, R.E. *et al. Nature* **470**, 59–65 (2011).
8. Li, Y. *et al. Nat. Biotechnol.* **29**, 723–730 (2011).
9. Fan, H. C. *et al. Nat. Biotechnol.* **29**, 51–57 (2011).
10. Kitzman, J.O. *et al. Nat. Biotechnol.* **29**, 59–63 (2011).
11. Teague, B. *et al. Proc. Natl. Acad. Sci. USA* **107**, 10848–10853 (2010).