Tech News

Copy Number Variants: Mapping the Genome's 'Land Mines'

A full decade after the first human genome draft sequences were published, much remains unclear about our genetic blueprint, and even its very structure. While single nucleotide polymorphisms (SNPs) have been extensively mapped and studied through efforts such as the International HapMap project, a larger group of genetic variants—insertions, deletions, duplications, and other copy-number structural differences that vary from individual to individual—are another matter entirely.

To identify and decipher structural variation, researchers need a reference against which to measure differences. Yet such comparative maps are sparse. Researchers still don't know the breadth of human genome variation, which makes it difficult to assess the novelty and thereby significance of any given copy-number disparities. Single-copy genes are, of course, well charted. But the genome is also riddled with repeated sequences. From highly repetitive LINE and Alu elements to some high–copy number protein-coding genes, the number, position, and orientation of these sequences are difficult to precisely pin down. But according to Evan Eichler, professor of genome sciences and Howard Hughes Medical Institute Investigator at the University of Washington School of Medicine, these repeated regions are also where much of the action is. Regions with high sequence homology—those with long tandem or interspersed duplications, for instance—are more prone to unequal crossover, and thus are hotspots of instability and disease, says Eichler; he calls these regions "land mines," and they dot and shape the genomic landscape. For instance, over the past five years Eichler's group has used these genome features to discover nearly a dozen regions associated with autism, developmental delay and epilepsy, including a 520-kb segment on the short arm of chromosome 16 associated with severe developmental delay (1).

Ironically, the same next-generation sequencing technologies that have been driving today's genomics explosion and SNP mapping efforts struggle to map and ascribe biology to these land mines. Sequencers tend to spit out billions upon billions of reads, but since they're all too short to span (and thereby unambiguously map) the variants, this large group has been excluded from many genomic analyses in the past.

Today, though, the genomics world is changing. Armed with new experimental tools and computational approaches, researchers are beginning—albeit slowly—to tackle copy-number variation.

Certain genomic regions—those with long tandem or interspersed duplications, for instance—are hotspots of structural instability and disease. Evan Eichler, a genome scientist at the University of Washington, calls these regions "land mines." Photo credit: Clare McLean, University of Washington.

Variation in definitions

The first question to consider when thinking about structural variants in the genome is superficially simplistic: what is a copy number variation (CNV)? The term is so broad that in theory, a CNV could be any sequence difference that's larger than one nucleotide. "The grossest example is trisomy," says Rafael Irizarry, professor of biostatistics at the Bloomberg School of Public Health at Johns Hopkins University. A common example is trisomy 21, whose characteristic three copies of chromosome 21 cause Down's syndrome.

Most variants are much smaller than a whole chromosome, though. The Database of Genomic Variants (http://projects.tcag.ca/variation), with some 102,000 entries, defines structural variation as "genomic alterations that involve segments of DNA that are larger than 1 kb," and InDels (insertions/deletions) as variations in the 100- to 1,000-bp range. That's the working definition used by Lars Feuk, associate professor of immunology, genetics and pathology at Uppsala University. He estimates that there are on order of 500 to 1000 CNVs between any two individuals' genomes.

Jan Korbel of the European Molecular Biology Laboratory in Heidelberg, Germany used a >50-bp working definition of structural variants in his and colleagues' survey of 1000 Genomes Project sequence data on 185 individuals. They found some 28,000 CNVs (median size 729 bp) comprising deletions (22,025), duplications (501), mobile element insertions (5371), and novel sequence insertions (128) (2).

Traditionally, techniques including fluorescence in situ hybridization and G-banded karyotype analysis have been used to ferret out these variations. But these were coarse-grained analyses, capable of detecting changes on the order of megabases or more. More recently, though, researchers have applied the finer tools of the genomics trade, including chromosomal microarrays and even next-generation sequencing to catalogue variants.

Arrays for the masses

The chromosomal microarrays used for CNV analysis come in two basic formats: SNP arrays fleshed out with nonpolymorphic content, and array-comparative genome hybridization (aCGH) microarrays, whose oligonucleotide probes create a mostly unbiased "tiling path" across the genome, with some concentration on known CNV hotspots thrown in for good measure.

These arrays offer several advantages, including speed and lower cost, which make them ideal for analyzing large patient populations. Another key benefit of microarrays experiments is that the data analysis pipelines are mature, says Feuk.

Jonathan Sebat, associate professor of psychiatry at University of California–San Diego, has made extensive use of aCGH arrays in his research of genomic regions associated with neuropsychiatric disease. In one 2007 study, his lab surveyed 264 families with autism spectrum disorders with an 85,000-probe array offering 35-kb resolution. The work identified de novo CNVs (that is, mutations present in affected individuals but not in unaffected parents) in 12 of 118 simplex families and 2 of 77 multiplex families. The CNVs ranged in size from about 100 kb to 12 Mb, including one 1.1 million–bp variant in a patient with Asperger's syndrome that deleted some 23 genes, including oxytocin (3). More recently, Sebat's team identified a new gene associated with schizophrenia, VIPR2, based on a 362-kb microduplication on the long arm of chromosome 7. The variant was present in 29 of 8,290 patients (0.35%) and just 0.03% of controls. Compelling on its own, the work is also indicates the need for large samples sizes when examining some structural variants.

An analysis led by Jan Korbel of data from the 1000 Genomes Project identified 28,000 CNVs—mostly deletions—in 185 individuals. Image courtesy of Jan Korbel.

Sometimes, though, clearly associating a disease with a structural variant can be difficult. The Genetics Diagnostic Laboratory at Children's Hospital Boston runs 200–300 genomic variant assays each month using the Agilent platform, according to assistant lab director David Miller, and has run probably 7,000 overall since 2006. "We get reliable data," he says of microarrays, but says that chief among the challenges faced is clinical significance—that is, assessing whether or not a particular CNV is actually causative of disease. "There are areas that have gains or losses that don't have many genes, or genes that are not known to be associated with disease," Miller explains. If a CNV falls in such a region, "it can be difficult to be decisive about whether that gain or loss is truly related to the clinical symptoms."

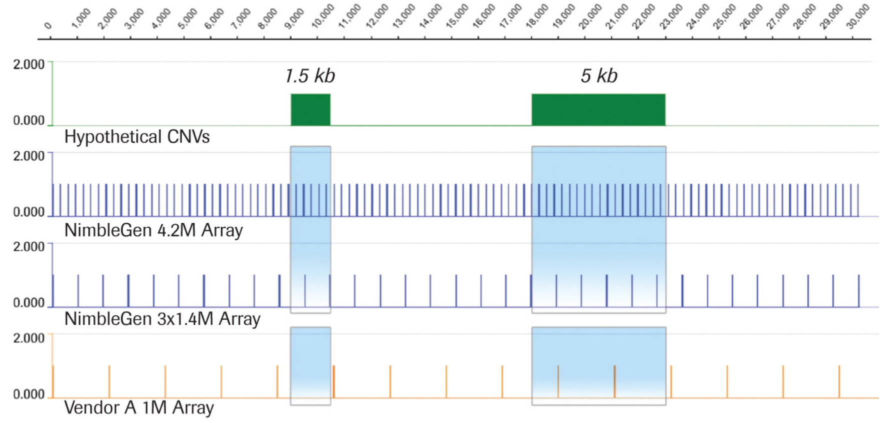
Time for a test

"One of the most common questions you get if you work a lot with arrays is, 'what array should I use and what analysis should I use?' ," says Feuk.

To address that question, Feuk and Irizarry independently put different array platforms and data analysis algorithms to the test, measuring interexperiment, interlab, and interplatform reproducibility, sensitivity, and accuracy on a series of control samples. Feuk's team tested two arrays each from Agilent, NimbleGen, and Affymetrix, four from Illumina, and a bacterial artificial chromosome array developed by the Wellcome Trust Sanger Institute. Irizarry's group tested six arrays — two from Affymetrix, two from NimbleGen, and one each from Illumina and Agilent — and used spike-in controls to address accuracy in measuring absolute copy number, as well.

The results suggest that, despite the maturity of the field, no one platform and analysis algorithm can do everything. "In our hands, the NimbleGen 2.1M was the array that worked best, but only when using our own analysis pipeline," Irizarry says of his study (Halper-Stromberg, E. et al. 2011. Bioinformatics *27*:1052-1060). When using manufacturer-recommended software, Affymetrix and Illumina arrays won out. But lab-to-lab and technician-to-technician differences, not to mention systemic artifacts such as "wave effects" and "batch effects," all cause data fluctuations. "It's possible that if we sent these samples to other labs, we might have gotten slightly different results," he says.

Feuk's study (Pinto, D. et al. Nat. Biotechnol. *29*:512-520) stressed the importance of using multiple platforms and analysis tools, he says. The report doesn't declare an actual winner—Feuk uses an Affymetrix array with the Genotyping Console, iPattern, and Birdsuite analysis packages in his own work—yet the conclusion, he says, is quite clear: "Although they are getting better and better, there's still a lot of disparity between different arrays." -JP



Greater microarray probe density provides higher-resolution CNV analysis. The green track represents hypothetical CNVs at 1.5 kb and 5 kb. Subsequent tracks represent average probe spacing across the genome for the NimbleGen CGH 4.2M and 3 × 1.4 M arrays (blue) and a competing vendor's 1 M CGH array (orange). When requiring five consecutive probes to make CNV calls, only the highest-density array can detect a 1.5-kb CNV. Image courtesy of Roche/NimbleGen.

To address this challenge, the Children's Hospital's lab looks to the past, comparing its CNVs against resources like the Database of Genomic Variants, DECIPHER, and the ISCA database, all of which catalog structural variants. Still, even using this approach, every so often a new variant pops up that they've never seen before.

David Ledbetter, chief scientific officer of Geisinger Health System (and former Director of the Division of Medical Genetics at Emory University), who coauthored with Miller a "consensus statement" regarding the use of chromosomal microarrays in clinical labs, estimates that only about 2–3% of patient samples that came through his lab at Emory exhibited so-called "variants of unknown clinical significance." These biologically nebulous CNVs lead to inconclusive results, he says. By comparison, the technique has a "diagnostic yield" of 15–20%, according to the consensus statement (compared to 3% for karyotyping). The remaining 80% or so exhibit no apparently pathogenic CNVs at all.

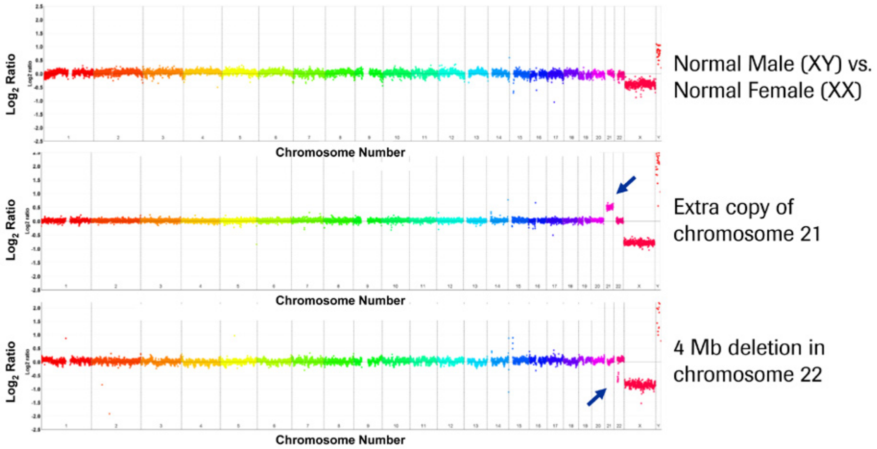
The novelty of sequencing

When it comes to CNV analysis, different microarray platforms are now being tested to determine which, if any, is most effective (see "Time for a test," below). But all have their limitations.

Microarrays cannot detect events such as novel insertions and balanced translocations, nor can they reveal CNV breakpoints with anything approaching nucleotide accuracy; instead, their resolution is a function of the number and distribution of their probes. For instance, Roche NimbleGen's newest 4.2 million–probe array has a median probe spacing of 284 bases, using 50- to 75-bp oligos. Requiring signal changes on at least five probes to confidently call a CNV, that means the chip can detect variants as small as 1.4 kb, says Rob Brazas, the company's international senior product manager for clinical products. These smaller variants often are associated with disease or serve as markers, says Brazas. "The higher-resolution your array is, the better you can detect smaller variants, or the rare events," he says. Importantly, arrays cannot indicate where a replicated segment of DNA lies: in other words, if array data indicate a region has been duplicated, is that duplication tandem, or did the DNA copy to another chromosome entirely?

Whereas microarrays are inherently limited by what is on the array itself, sequence data is unbiased. In theory, it captures all variation—assuming it can be mapped—and can therefore better uncover the genetic loci underlying many complex phenotypes and diseases. "A much bigger piece of the iceberg is now captured with whole-genome sequencing," explains Sebat, "and a significant chunk of the missing heritability is sure to emerge."

In practice, sequencing and mapping have biases of their own, detecting far more deletions than insertions or duplications, for instance. Yet, researchers can use sequencing reads to map breakpoints to the base pair, enabling them to better understand the functional consequence of a given CNV (like whether it encroaches upon a coding or regulatory region) and



Sensitive and accurate detection of copy number changes on CGH microarrays. Data are displayed as whole-genome "rainbow" plots where each chromosome is denoted by a different color. Shown are normal samples, a gain of an entire chromosome (associated with trisomy 21), and a large (~4-Mb) deletion in chromosome 22 associated with velocardiofacial syndrome. All research samples were referenced against normal genomic DNA. Credit: Roche/NimbleGen

also probe the molecular mechanisms underlying those variants. In one 2010 study, researchers led by Matthew Hurles at the Wellcome Trust Sanger Institute sequenced 324 CNV breakpoints from three individuals using sequence capture and 454 sequencing. Their data suggest that several distinct molecular mechanisms underlie deletion events: of 315 deletion breakpoints, 70% included short regions of "microhomology," and 33% included inserted sequences. Just 10% of breakpoints included both elements, "suggesting that there are at least two different mutational mechanisms," the authors wrote (4).

Despite the rich data, CNV analysis by next-generation sequencing isn't easy. Says Sebat: "The methods for calling CNVs are still not mature." Even in the absence of CNVs, some segments of DNA sequence better than others, which is a sequence bias that plagues all genome-sequencing efforts. Plus, when it comes to nucleotide-level CNV analysis, the devil is in the details. As Eichler explains, it isn't enough to know, for example, that a region is duplicated; what matters is its precise structure. Once sequences duplicate, each copy can evolve independently. The biological consequence of three tandem duplicated genes of the form A–A–A differs, for instance, from A–A'–A''. And both are different than A–inverted A'–A''.

To solve that problem, Eichler says, researchers require high-quality, long-read sequence data, similar to what was generated as part of the Structural Variation Genome Sequencing Project. This kind of clone-based work used to be the norm with Sanger sequencing, but is becoming harder to do using next-generation sequencing platforms. But even Sanger sequencing, whose reads approach a kilobase in size, may not be long enough according to Eichler. He would like to see reads in the hundreds of kilobases, larger than the longest repeats and beyond the ranges promised by existing third-generation sequencing technologies. "The ability to get outside a duplicated region gives you the ability to anchor and build a robust assembly," he says.

In the absence of such reads, CNV researchers today have devised a variety of different strategies to find the signal in the sequencing noise. As Eichler explained in a 2011 review of array and sequencing based strategies for CNVs, CNV-calling algorithms can be grouped into four basic strategies. For instance, "read-pair" approaches use paired-end reads to assess whether the distance between the paired ends differs from that of the reference, while "read-depth" approaches exploit the absolute number of reads over a given region to identify variants (5).

Often, researchers will apply a cross-section of these approaches to hedge their bets. When the 1000 Genomes Project team mapped CNVs in a set of 185 genomes, it used 19 different calling algorithms to find them.

Of course, as in all things 'omic, arrays and sequencing really are complementary, at least at the moment. Case in point: a 2010 study by Eichler's team compared structural variants in five individuals detected by sequencing, aCGH, and an Affymetrix SNP microarray. "A comparison of the three studies shows that 11–65% of discovered variants are unique to a single study and corresponding experimental platform," the authors wrote (6).

And, as in all things 'omic, sequencing will likely ultimately become the dominant technology. "Sequencing technology has left Moore's Law in the dust," says Sebat, "it's Moore's Law times two." Already, the 1000 Genomes Project has generated terabases of human genomics data. Yet that isn't nearly enough to cover the breadth of human sequence variation, says Sebat. To identify rare variants and get a feel for their prevalence in the general population, thousands upon thousands of genomes will have be analyzed, he says.

"It seems like a very daunting proposition to be sequencing tens of thousands of genomes at 30× coverage, but it will happen, and my lab will never be the same again," he says. "We are swimming in data right now, and it's not going to stop."

References

1. **Girirajan, S., J.A. Rosenfeld, G.M. Cooper, F. Antonacci, P. Siswara, A. Itsara, L. Vives, T. Walsh, et al.** 2010. A recurrent 16p12.1 microdeletion suggests a two-hit model for severe developmental delay. Nat. Genet. *42*:203-209.
2. **Mills, R.E., K. Walter, C. Stewart, R.E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S.C. Yoon, et al.** 2011. Mapping copy number variation by population-scale genome sequencing. Nature *470*: doi:10.1038/nature09708.
3. **Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, et al.** 2007. Strong association of de novo copy number mutations with autism. Science *316*:445-449.
4. **Conrad, D.F., C. Bird, B. Blackburne, S. Lindsay, L. Mamanova, C. Lee, D.J. Turner, and M.E. Hurles.** 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat. Genet., *42*:385-391.
5. **Alkan, C., B.P. Coe, and E.E. Eichler.** 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*:363-375.
6. **Kidd, J.M., T. Graves, T.L. Newman, R. Fulton, H.S. Hayden, M. Malig, J. Kallicki, and R. Kaul.** 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*:837-847.

Written by Jeffrey M. Perkel, Ph.D.

BioTechniques 51:21-24 (July 2011)
doi 10.2144/000113699

To purchase reprints of this article, contact: biotechniques@fosterprinting.com