New Genomics Technology:

Copy Number Variation Analysis Methods

Annelise Mah

Genomics and Medicine Fall 2008

Final Paper 12/01/08

Professor Doug Brutlag

Stanford University

What are Copy Number Variations?

CNVs, or sometimes CNPs (Copy Number Polymorphisms), are changes in the number of appearances a certain pattern makes in the genome. Copy Number Variable Regions (CNVRs) are regions of the genome that are copied, deleted, or varied in number in some way. Normally these regions are defined as a kilobase (Kb, 10^3) to several megabases (Mb, 10^6) in size. These CNVRs make up around 12% of the human genome, cause disease, affect gene expression, and alter the organism's phenotype. A total of 1447 CNVRs spanning 360 Mb and associated with over 3000 genes has been discovered. (1) So far, it has been estimated that human individuals differ from each other by anywhere from 4 to 20 Mb (2,3).

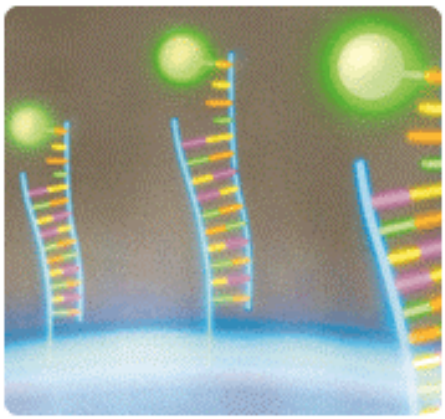
CNVs have in the past thought to be much rarer; several researchers discovered their ubiquity within the last few years and, since then, many new studies have been conducted. This paper will cover several techniques that have been used to discover CNVs.

Methods for Finding CNVs

ROMA (Representational Oligonucleotide Microarray Analysis)

ROMA was one of the first methods designed to randomly search for CNVs. For this method, scientists choose selective DNA fragments (representations) and design oligonucleotide (around 50-100 base pairs) probes for them. These are laid on a microarray with different wells containing many copies each probe. More recently, microarray chips are made, either by printing on glass or synthesizing onto silica by laser photochemistry, with many copies of a single probe comprising each dot. Two sets of genes are then tagged with different fluorescent dyes, usually green or red, following the common microarray method. Often, a reference gene and a tumor gene are used in order to compare the CNVs of the two. These genes are then hybridized to an array with those genes on it. If a dot glows yellow, then the gene is present equally in both

samples. A mostly-red or a mostly-green dot will indicate a deletion or replication in the gene. Depending on the color's intensity, an estimation of the number of copies/deletions can be made. A hidden markov model or other algorithm is used to accurately predict the difference in copy number. (4) One problem with the ROMA technique is that its probes do not cover very much of the genome.



(Fanciful illustration of oligonucleotide probes with tagged DNA attached, Illumina)

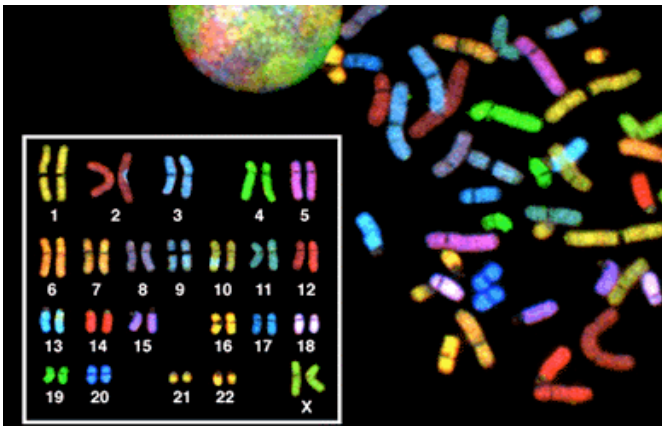
Fosmid Paired-End Sequencing

In 2005, Tuzen et al. used a novel comparative method to locate deletions, insertions, and even inversions in the human genome that could denote CNVs. The group sequenced the ends of fosmids (bacteria vectors with inserted human DNA) containing a human genome. These ends were a known distance apart, around 40 kb. The sequences of these end-pairs were then compared to a reference genome. If the sequences that made up the ends fell much shorter or farther apart on the reference genome than predicted by the fosmid size, then an insertion or deletion must have occurred in the test genome. If the sequences were unexpectedly flipped, then an inversion could have occurred. They also checked the insertions they found by designing PCR primers at both ends and the middle of the change. If the shorter segments that started in the inserted DNA showed up as well as the whole sequence, then DNA was inserted. If only the

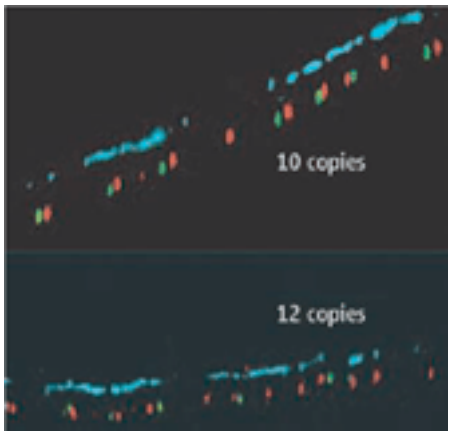
entire sequence was copied, then no DNA was inserted. (5) This method is now considered limited in scope, but it showed around 300 CNVRs, including many that were related to disease.

SKY (Spectral Karyotyping) or FISH (Fluorescence In Situ Hybridization)

In situ, or in place, means that the reaction takes place between *in vivo* and *in vitro,* in its natural tissue but not in the organism (Merriam-Webster, Wikipedia). To detect CNVs, probes of target genes are fluorescently tagged and hybridized to chromosome or sections of DNA to see where they land. (6) These methods are not that accurate discovery or quantification tools, and are mostly used to validate finds and map them to a certain location.



(SKY coloring of chromosomes, Medical University of South Carolina)



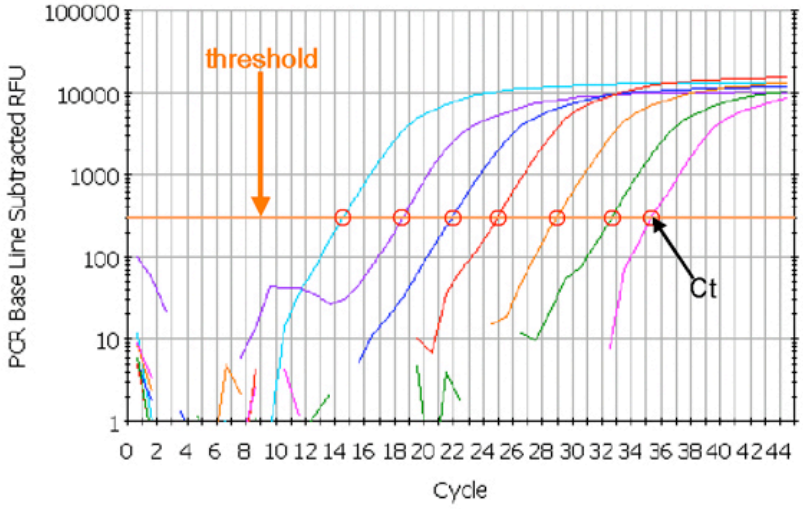
(CNV seen using FISH)

CGH (Comparative Genomic Hybridization)

For a general look at CNV, scientists can isolate metaphase chromosomes and hybridize both tagged tumor genes and tagged normal genes to the normal chromosome. Different areas of the chromosome will glow different colors based on the variations in copy number in the tumor and reference gene. However, this strategy gives a global view, and really only operates at the Mb level (7). It is considered too low-resolution, unable to detect smaller changes.

RT/Q-PCR (Real Time Quantitative PCR)

RT-PCR or Q-PCR is another strategy that is mainly used to confirm or validate finds. PCR, Polymerase Chain Reaction, is a technique used to quickly amplify a sample of DNA. The elements of PCR are fluorescently tagged and the intensity of the glow is monitored between cycles. Comparatively, a single gene with more copies will increase sooner than one with fewer copies when the fluorescent glow is measured on a logarithmic scale. A more dilute sample will increase later (graphically, shifted to the right). This technique is not perfect because of uncontrollable environmental factors, but it can give a comparative quantification of copy numbers (8).

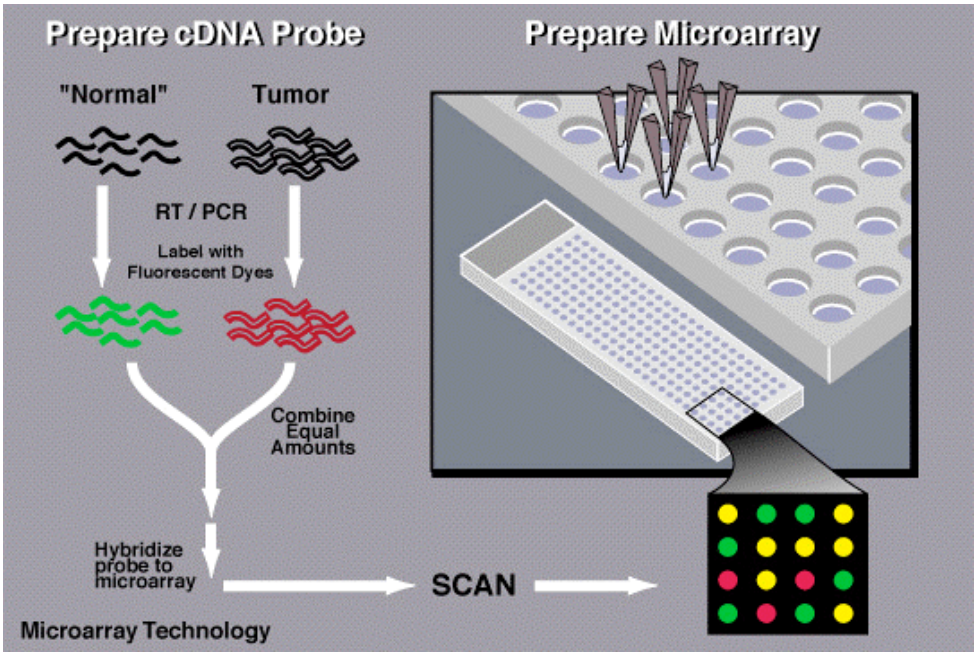


(Logarithmic graph of fluorescent intensity by time, 8)

aCGH (Array Comparative Genomic Hybridization)

Array Comparative Genomic Hybridization (CGH) has become one of two new powerful array-based methods. Rather than picking certain representative probes, scientists using aCGH can use most of the genome. Genes of interest are inserted into a BAC library (Bacterial Artificial Chromosome, usually *e. coli* holding human DNA), for concentration and replication. Normal microarray procedure is followed—these genes are spliced into fragments, labeled with

fluorescent tags, and hybridized onto a microarray containing the 1genes of interest. Based on the color of each dot, the color's intensity, and a complicated algorithm, the amount of copying or deletion can be estimated. (9) The data is then "smoothed" so that the gene is deleted or copied in integer fashion. These models show areas of deletion or copying, as well as "breakpoints" in the genome. These breakpoints are fragile areas of the chromosome where copy number variations and other chromosomal abnormalities occur. (7) Although the name implies whole-genome testing, the need for more specific tests has produced technology like the NimbleGen chip (Roche Company) that can probe either whole genome or specific chromosomes/loci.

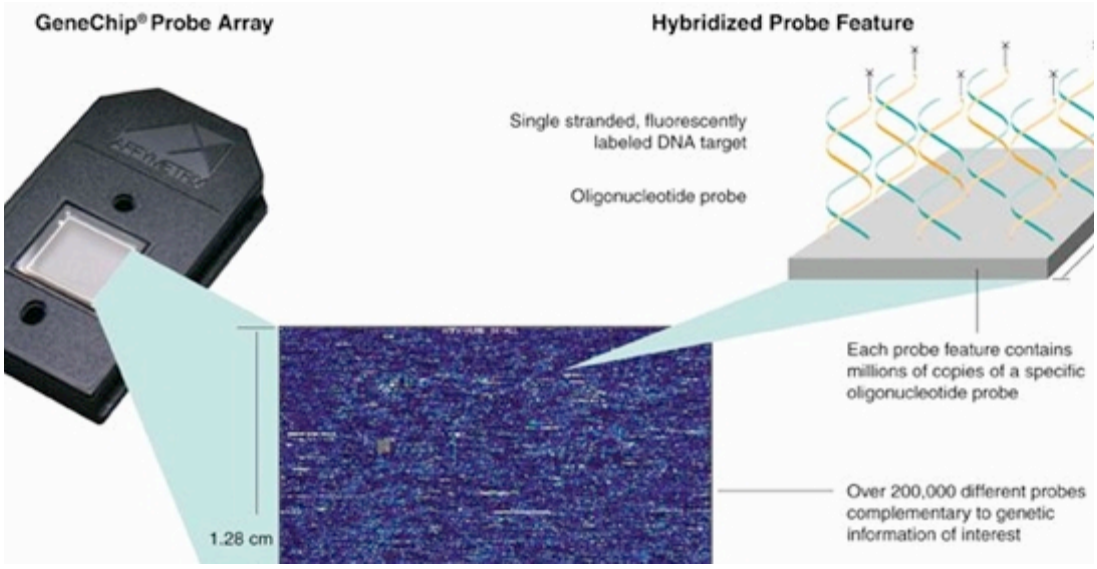


(Microarray technology, National Human Genome Research Institute)

High-density Whole-genome SNP Microarrays

SNP analysis has lately emerged as both a reliable way to differentiate genomes and a tool to analyze disease risk. Its popularity is in part because SNP data is abundant and well-spaced throughout the genome (10). SNPs, or Single Nucleotide Polymorphisms, are single bases in certain locations that differ between individuals. Oligonucleotides containing SNPs are laid

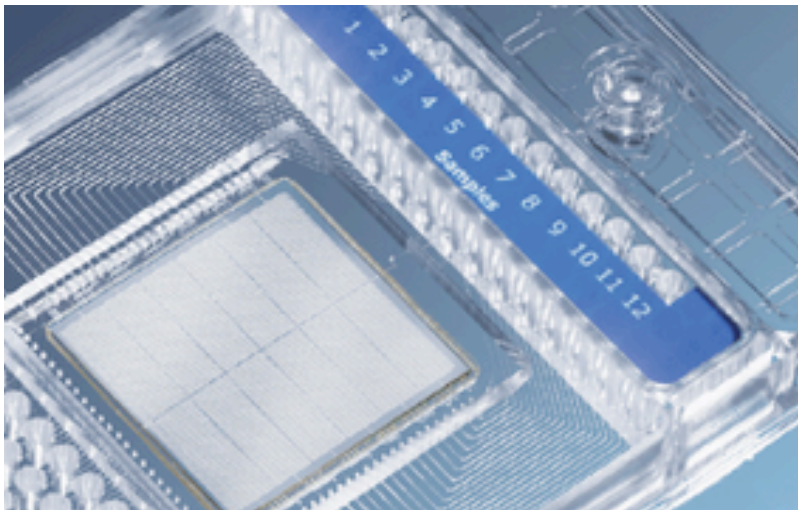
out on an array/chip. Makers such as Affymetrix have designed chips that can contain more than 900,000 SNPs from throughout the human genome (2). Gene samples will either hybridize perfectly or with one nucleotide off. If the ratio and intensity of perfect matches to mismatches of a reference gene is compared to a test gene, the frequencies will show if certain SNPs are in higher concentration in the reference or test gene. Alternately, different spots on the chip contain different SNPs. When a SNP's frequency does not match with predicted results, the location of those SNPs will give the general area of the CNVR. (11) For example, if a SNP of "A" (rather than "T") is within a region that is duplicated 3-fold in a test genome, the SNP ratio of "A" to "T" will be three times higher in the test than in the reference. This method has a good resolution (new chips offer markers every 700 bps (2)), especially since SNP technology is advancing quickly. However, some areas of the genome don't have usable SNPs. This had led chip makers such as Affymetrix to offer nearly 950,000 other probes specifically for CNVs in unSNPable regions. Illumina's bead-array technology, which consists of 3-micron oligonucleotide-covered beads set in wells, also offers CNV markers for some 3000 regions (12).



(Gene chip, Affymetrix)

Digital Array

A newer innovation is the use of a digital array on a nanofluidic chip, which holds PCR reactions in tiny individual chambers. Normal Digital PCR, used to quantify samples of DNA, is carried out through increasing dilutions to isolate DNA molecules. On a nanofluidic chip, the molecules can be separated rather than diluted. Target DNA is spliced and undergoes PCR within the chip. Then, the DNA is distributed so that each chamber gets one molecule or none. If it gets more than one, an algorithm can be used to correct the data. The number of "positive" chambers, i.e. chambers with molecules in them, will give the number of copies. Based on the number of times PCR has been run, if splicing accurately cut each gene sample into discrete copies, they can calculate almost exactly how many copies the gene had (13). They were also able to run this experiment using multiplex-PCR (more than one gene) by fluorescently tagging different genes and then tracking the color of the chamber.

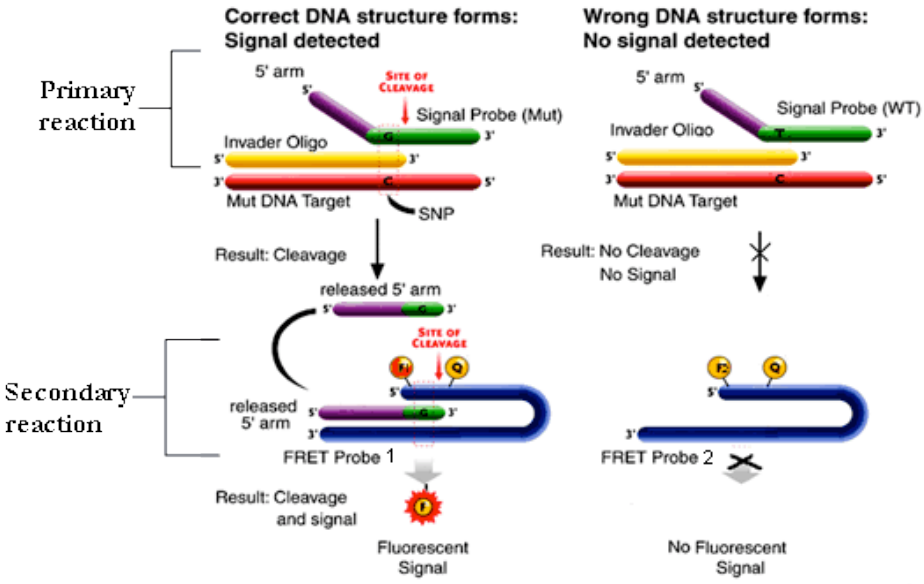


(Nanofluidic chip with tiny reaction chambers, 13)

PCR-RETINA (Real Time Invader Array)

One group, Hosono et al, modified RT-PCR to test for copy numbers and homozygosity. They used "Invader" probes to cleave DNA at a SNP and replace the DNA with a fluorescent label at that SNP depending on which base it was. They took fluorescence readings while this

reaction happened. During the reaction, they could see whether the individual being tested was homozygous for the SNP (all one type of tag), heterozygous (both tags included), or heterozygous with different CNVs, based on how fast the intensity of the fluorescence increased. Using this technique, they were able to tell the ratio of one tag to the other. They were able to use multiplex-PCR (many genes at once) with similar success. This combination of techniques allowed them to accurately analyze copy number variations (14).



(Invader probe technology, National Genetics Reference Laboratory)

Conclusion

CNV-finding technology has greatly increased in resolution and accuracy over five years, giving an overlapping series of pictures of the locations and functions of CNVs. From literally whole-genome comparisons like karyotyping to hundreds of thousands of specific SNPs, these microarrays, chips, and tagging processes have evolved to find variants of many megabases long to ones comprised of just hundreds of base pairs. Big companies like Affymetrix, NimbleGen, and Illumina are shifting their considerable technological powers to investigate CNVs, providing

ordering and mass-production services. Many different algorithms and models have been created to analyze the data generated by these tests.

This technology has become so important because CNVs have much to offer science. As seen in the studies above, CNV detection is useful in comparing normal and tumor cells, showing the changes that can cause cancer (11). Copy number variations have also been shown to cause or affect many diseases, including big-name ones such as autism, HIV/AIDS, Parkinson's and Alzheimer's. Some even speculate that it makes up a considerable part of our phenotype, or traits (who?). Among those traits are reactions to medication and traits such as color-blindness and anatomical deformities. (15) Another interesting use of CNV technology is comparative biology: by comparing our DNA with apes, certain CNVs are evolutionarily human (16). Studies have also compared the different CNVs present in different populations of the world (1). CNVs help give humans their identities, evolutionarily and phenotypically. This research has a future in both understanding and curing disease, and will likely continue to explode in popularity as scientists refine their techniques and analyze the data.

Works Cited

(1) Redon et al. "Global variation in copy number in the human genome." Nature 44 (2006): 444-454.

(2) Affymetrix® Genome-Wide Human SNP Array 6.0 Data Sheet. Affymetrix, 2007.

(3) Kehrer-Sawatzki, Hildegard. "What a difference copy number makes." Biology 29.4 (2007): 311-313.

(4) Lucito et al. "Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation." Genome Research 13 (2003): 2291-2305

(5) Tuzun et al. "Fine-scale structural variation of the human genome." Nature Genetics 37 (2005): 727-732.

(6) "Fluorescence In Situ Hybridization (FISH)." National Human Genome Research Institute. 12 Sept. 2008. National Institutes of Health. 27 Nov. 2008. <http://www.genome.gov/10000206>

(7) Xing, Eric. 10-810, CMB lecture 11, "Introduction to array CGH analysis." School of Computer Science, Carnegie Mellon.

(8) Hunt, Margaret. "Real Time PCR." Microbiology and Immunology On-Line. 4 Sept. 2007. University of South Carolina School of Medicine. 28 Nov. 2008. <http://pathmicro.med.sc.edu/pcr/realtime-home.htm>

(9) Perry et al. "The Fine-Scale and Complex Architecture of Human Copy-Number Variation." The American Journal of Human Genetics 82 (2008): 685-695.

(10) Freeman et al. "Copy number variation: New insights in genomic diversity." Genome Research 16 (2006): 949-961.

(11) Huang et al. “Whole genome DNA copy number changes identified by high density oligonucleotide arrays.” Human Genomics 1 (2004): 287-299.

(12) “SNP Genotyping and CNV Analysis.” Illumina 100-101.

(13) Dube, Simant, Jian Qin, and Ramesh Ramakrishnan. “Mathematical Analysis of Copy Number Variation in a DNA Sample Using Digital PCR on a Nanofluidic Device.” PLoS ONE 3.8 (2006): e2876.

(14) Hosono et al. “Multiplex PCR-Based Real-Time Invader Assay (mPCR-RETINA): A Novel SNP-Based Method for Detecting Allelic Asymmetries Within Copy Number Variation Regions.” Human Mutation 29.1 (2008): 182-189

(15) Cohen, Jon. “Genomics: DNA Duplications and Deletions Help Determine Health.” Science 317 (2007): 1315-1317.

(16) Goidts et al. “Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome.” Human Genetics 120 (2006): 270-284.