CNV Detection and Interpretation in Genomic Data

Benjamin W. Darbro, M.D., Ph.D.
Assistant Professor of Pediatrics
Director of the Shivanand R. Patil Cytogenetics and Molecular Laboratory
Overview

- What are CNVs and why are they important?
- How can you find CNVs in your genomic datasets?
- How can you interpret the significance of CNVs discovered in your genomic datasets?
DNA Variations

- Single nucleotide variations (SNVs) and polymorphisms (SNPs)
- Insertions/Deletions (Indels)
- Structural Variations
- Short and Variable Repeats
- Copy-Number Variations (CNVs)
  - Duplications (insertions)
  - Deletions

Helpful Clarifications

• **Indel**: Gain or loss of segment of DNA typically of a small scale (<1kb)

• **CNV**: Segment of DNA with different copy numbers across individuals (>1kb classically but now >500bp)

• **CNP**: Copy number polymorphism is a CNV shared by >1% of a population

• **CNA**: Copy number aberration is what CNVs in cancer studies are often called including SCNA (somatic copy number aberration)

• **Segmental Duplication (SegDup)**: Segment of DNA >1kb that occurs in two or more copies in a haploid genome (share > 90% sequence identity)
Segmental Duplications (SDs)

- Low copy repeat (LCR) regions (2-50 copies)
  - Approximately 5-6% of the human genome
  - Common in pericentromeric/subtelomeric regions (majority are interstitial)
  - Majority are intrachromosomal (many interchromosomal)
  - Can mediate the creation of additional CNVs
How CNVs Are Formed

- Nonallelic homologous recombination (NAHR)
  - Unequal crossing over (LCR)
  - Reciprocal deletion and duplication
  - Deletion only
  - Translocations and inversions

- Nonallelic homologous end-joining (NHEJ)
  - Double stranded DNA break and subsequent repair pathways
  - Diverse repair products possible

How CNVs Are Formed

• Fork stalling and template switching (FoSTeS)
  - Nick causes fork collapse
  - 3’ overhang invades leading strand guided by microhomology

• LINE1 Retrotransposition (not shown): Only active class of retrotransposons in humans and responsible for ~90% of small insertions in the human genome

How Prevalent are CNVs?

- HapMap project suggests that ~12% of the human genome is copy number variable

- CNVs cover ~360Mb of the human genome (far more nucleotide content per genome than SNPs)

- Disease causing CNVs are typically rare, highly penetrant lesions (often de novo)
Well Known Disease CNVs

• Deletion of 22q in DiGeorge/Velocardiofacial Syndrome

• Trisomy and monosomy (numerous examples)

• HER-2 amplification in breast cancer

• N-myc amplification in neuroblastoma

• Deletion of 13q14.3 in favorable prognosis CLL
CNVs and Disease

- First Line Clinical Diagnostic Test
  - Non-syndromic intellectual disability
  - Multiple congenital anomalies
  - Autism spectrum disorders

- Other Conditions in which CNVs are Implicated
  - Congenital cardiac disease
  - Epilepsy
  - Schizophrenia
  - Spina bifida, BOR, CLP, obesity, and a variety of different cancers

What Do They Do?

• Alter chromosomal structure and gene expression through many different mechanisms
  – Changing gene dosage
  – Influence gene expression through positional effects (loss or gain of regulatory sequence)
  – Create gene fusion events
  – Predispose to deleterious genetic changes (structural variations)
Haploinsufficiency and Disease

• Virtually any disease caused by haploinsufficiency of a gene could theoretically be caused by a deletion CNV (duplication and over-expression is less clear cut)
Overview

• What are CNVs and why are they important?

• How can you find CNVs in your genomic datasets?

• How can you interpret the significance of CNVs discovered in your genomic datasets?
How do (did?) we currently detect these aberrations?
• Karyotype: Full genome but low *resolution*

• FISH: Improved *resolution* but locus specific

*Photos courtesy of the Shivanand R. Patil Cytogenetics and Molecular Laboratory*
CNV Detection in Genomic Data

• Chromosomal Microarrays
  – Comparative genomic hybridization (CGH) microarrays
  – Single nucleotide polymorphism (SNP) microarrays
  – Combination of CGH+SNP arrays

• High Throughput Sequencing
  – Custom targeted capture
  – Whole exome sequencing
  – Whole genome sequencing

Photos courtesy of the Shivanand R. Patil Cytogenetics and Molecular Laboratory
What is the difference between these techniques when it comes to CNV detection?

- Chromosomally Microarray
- Massively Parallel Sequencing
- Fluorescence In Situ Hybridization
- Chromosome Analysis (Karyotype)
What is the difference between these techniques when it comes to CNV detection?

Throughput
Resolution
Precision
Target Space
Comparing Genomes

Like MegaTouch’s Photo Hunt
Comparing Genomes

Patient Genome  Reference Genome

Comparing patient to reference genome and finding gains/losses
Comparing Genomes

Patient Genome

Reference Genome

Several “deletions” and “duplications” present in patient sample
Method Comparisons

Near “Perfect” Resolution: Entire genome, every base pair “seen”
Method Comparisons

Patient Genome  Reference Genome

Resolution: ~4-10Mb

Conventional Cytogenetic Karyotyping
Method Comparisons

Fluorescence In Situ Hybridization (FISH)

Reference Genome

Patient Genome

Resolution: ~50-100kb
Method Comparisons

Patient Genome

Reference Genome

Resolution: ~10kb

Chromosomal Microarray Resolution
Method Comparisons

Whole Genome Sequencing

Resolution: ~1bp
(but incomplete)

Patient Genome

Reference Genome

Whole Genome Sequencing
CNV Detection Algorithms

Chromosomal Microarray VS. Massively Parallel Sequencing

(See the last several slides of this presentation)

(See the next several slides of this presentation)
Detecting CNVs in MPS Data

- Paired-End
- Split-Read
- Read-Depth
- Sequence Assembly
- Combination

http://bioinfo.mc.vanderbilt.edu/CNVannotator/links.cgi
No method alone is capable of detecting all types of CNVs.

The 1000 Genomes Project used 19 algorithms to independently identify CNVs in 185 human genomes.

http://bioinfo.mc.vanderbilt.edu/CNVannotator/links.cgi
## Advantages & Disadvantages

<table>
<thead>
<tr>
<th></th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Paired-End</strong></td>
<td>Precise Balanced (Inversions) Low Coverage</td>
<td>Bad for Targeted Projects Resolution ~ Insert Dependent</td>
</tr>
<tr>
<td><strong>Split-Read</strong></td>
<td>Good Breakpoint Resolution</td>
<td>Lose Precision in Targeted Projects</td>
</tr>
<tr>
<td><strong>Read-Depth</strong></td>
<td>Most Like Array Analysis Any Unique Sequence Good for Custom Targets</td>
<td>Bias (Normalize or Control Sample(s)) Only Imbalanced Need High Coverage</td>
</tr>
<tr>
<td><strong>Sequence Assembly</strong></td>
<td>Not Insert Dependent Find Novel Insertions</td>
<td>Intensive for Large Genomes (Humans) Bad with Repeats</td>
</tr>
</tbody>
</table>
Which Algorithm Do I Choose?
Which Algorithm Do I Choose?

Depends on Your Data!!

...and a few other things.
Which Algorithm Do I Choose?

Depends on Your Data!!

…and a few other things.

Here is the List…
# PEM, SR, AS Based Tools

<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Language</th>
<th>Input</th>
<th>Comments</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEM-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BreakDancer</td>
<td><a href="http://breakdancer.sourceforge.net/">http://breakdancer.sourceforge.net/</a></td>
<td>Perl, C++</td>
<td>Alignment files</td>
<td>Predicting insertions, deletions, inversions, inter- and intra-chromosomal translocations</td>
<td>[1]</td>
</tr>
<tr>
<td>PEMer</td>
<td><a href="http://sv.gersteinlab.org/pemer/">http://sv.gersteinlab.org/pemer/</a></td>
<td>Perl, Python</td>
<td>FASTA</td>
<td>Using simulation-based error models to call SVs</td>
<td>[2]</td>
</tr>
<tr>
<td>VariationHunter</td>
<td><a href="http://compbio.cs.fsu.edu/strvar.htm">http://compbio.cs.fsu.edu/strvar.htm</a></td>
<td>C</td>
<td>DIVET³</td>
<td>Detecting insertions, deletions and inversions</td>
<td>[3]</td>
</tr>
<tr>
<td>commonLAW</td>
<td><a href="http://compbio.cs.fsu.edu/strvar.htm">http://compbio.cs.fsu.edu/strvar.htm</a></td>
<td>C++</td>
<td>Alignment files</td>
<td>Aligning multiple samples simultaneously to gain accurate SVs using maximum parsimony model</td>
<td>[4]</td>
</tr>
<tr>
<td>Spanner</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Using PEM to detect tandem duplications</td>
<td>[6]</td>
</tr>
<tr>
<td>SR-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pindel</td>
<td><a href="http://www.ebi.ac.uk/~kyle/pindel/">http://www.ebi.ac.uk/~kyle/pindel/</a></td>
<td>C++</td>
<td>BAM / FASTQ</td>
<td>Using a pattern growth approach to identify breakpoints of various SVs</td>
<td>[8]</td>
</tr>
<tr>
<td>SLOPE</td>
<td><a href="http://www-genepi.med.utah.edu/suppl/SLOPE">http://www-genepi.med.utah.edu/suppl/SLOPE</a></td>
<td>C++</td>
<td>SAM/ FASTQ/ MAQᵇ</td>
<td>Locating SVs from targeted sequencing data</td>
<td>[9]</td>
</tr>
<tr>
<td>SRIC</td>
<td>N/A</td>
<td>N/A</td>
<td>BLAT output</td>
<td>CalibratingSV calling using realistic error models</td>
<td>[10]</td>
</tr>
<tr>
<td>AS-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnolya</td>
<td><a href="http://sourceforge.net/projects/magnolya/">http://sourceforge.net/projects/magnolya/</a></td>
<td>Python</td>
<td>FASTA</td>
<td>Calling CNV from co-assembled genomes and estimating copy number with Poisson mixture model</td>
<td>[11]</td>
</tr>
<tr>
<td>Cortex assembler</td>
<td><a href="http://cortexassembler.sourceforge.net/">http://cortexassembler.sourceforge.net/</a></td>
<td>C</td>
<td>FASTQ / FASTA</td>
<td>Using alignment of de novo assembled genome to build de Bruijn graph to detect SVs</td>
<td>[12]</td>
</tr>
<tr>
<td>TIGRA-SV</td>
<td><a href="http://gmt.genome.wustl.edu/tigra-sv/">http://gmt.genome.wustl.edu/tigra-sv/</a></td>
<td>C</td>
<td>SV calls⁵ + BAM</td>
<td>Local assembly of SVs using the iterative graph routing assembly (TIGRA) algorithm</td>
<td>N/A</td>
</tr>
</tbody>
</table>

---

http://bioinfo.mc.vanderbilt.edu/CNVannotator/links.cgi

---

*The specific input format for VariationHunter, including the reads with multiple alignments.

*bFile format from MAQ mapview.

*The file including the detected structure variations using other tools.
## RD Based Tools for WGS

<table>
<thead>
<tr>
<th>Tool</th>
<th>URL</th>
<th>Language</th>
<th>Input</th>
<th>Comments</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV-seqa</td>
<td><a href="http://tiger.dbs.nus.edu.sg/cnv-seq/">http://tiger.dbs.nus.edu.sg/cnv-seq/</a></td>
<td>Perl, R</td>
<td>Aligned read positions</td>
<td>Identifying CNVs using the difference of observed copy number ratios</td>
<td>[14]</td>
</tr>
<tr>
<td>RDXplorerb</td>
<td><a href="http://rdexplorer.sourceforge.net/">http://rdexplorer.sourceforge.net/</a></td>
<td>Python, Shell</td>
<td>BAM</td>
<td>Detecting CNVs through event-wise testing algorithm on normalized read depth of coverage</td>
<td>[15]</td>
</tr>
<tr>
<td>BIC-seqa</td>
<td><a href="http://compbio.med.harvard.edu/Supplements/PNAS11.html">http://compbio.med.harvard.edu/Supplements/PNAS11.html</a></td>
<td>Perl, R</td>
<td>BAM</td>
<td>Using the Bayesian information criterion to detect CNVs based on uniquely mapped reads</td>
<td>[16]</td>
</tr>
<tr>
<td>CNAsseg</td>
<td><a href="http://www.compbio.group.cam.ac.uk/software/cnaseg">http://www.compbio.group.cam.ac.uk/software/cnaseg</a></td>
<td>R</td>
<td>BAM</td>
<td>Using flowcell-to-flowcell variability in cancer and control samples to reduce false positives</td>
<td>[17]</td>
</tr>
<tr>
<td>cn.MOPSc</td>
<td><a href="http://www.bioinf.jku.at/software/cnmos/">http://www.bioinf.jku.at/software/cnmos/</a></td>
<td>R</td>
<td>BAM/read count matrices</td>
<td>Modelling of read depths across samples at each genomic position using mixture Poisson model</td>
<td>[18]</td>
</tr>
<tr>
<td>ReadDepth</td>
<td><a href="http://code.google.com/p/readdepth/">http://code.google.com/p/readdepth/</a></td>
<td>R</td>
<td>BED files</td>
<td>Using breakpoints to increase the resolution of CNV detection from low-coverage reads</td>
<td>[20]</td>
</tr>
<tr>
<td>rSW-seqa</td>
<td><a href="http://compbio.med.harvard.edu/Supplements/BMCBioInfo10-2.html">http://compbio.med.harvard.edu/Supplements/BMCBioInfo10-2.html</a></td>
<td>C</td>
<td>Aligned read positions</td>
<td>Identifying CNVs by comparing matched tumor and control sample</td>
<td>[21]</td>
</tr>
<tr>
<td>CNVnator</td>
<td><a href="http://sv.gersteinlab.org/">http://sv.gersteinlab.org/</a></td>
<td>C++</td>
<td>BAM</td>
<td>Using mean-shift approach and performing multiple-bendwidth partitioning and GC correction</td>
<td>[22]</td>
</tr>
<tr>
<td>CNVnormb</td>
<td><a href="http://www.precancer.leeds.ac.uk/cnanorm">http://www.precancer.leeds.ac.uk/cnanorm</a></td>
<td>R</td>
<td>Aligned read positions</td>
<td>Identifying contamination level with normal cells</td>
<td>[23]</td>
</tr>
<tr>
<td>CMDSc</td>
<td><a href="https://dsgweb.wustl.edu/qunyuan/software/cmds">https://dsgweb.wustl.edu/qunyuan/software/cmds</a></td>
<td>C, R</td>
<td>Aligned read positions</td>
<td>Discovering CNVs from multiple samples</td>
<td>[24]</td>
</tr>
<tr>
<td>mrCaNaVar</td>
<td><a href="http://mrcanavar.sourceforge.net/">http://mrcanavar.sourceforge.net/</a></td>
<td>C</td>
<td>SAM</td>
<td>A tool to detect large segmental duplications and insertions</td>
<td>[25]</td>
</tr>
<tr>
<td>CNVeM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Predicting CNV breakpoints in base-pair resolution</td>
<td>[26]</td>
</tr>
<tr>
<td>cnvHMM</td>
<td><a href="http://genome.wustl.edu/software/cnvHmm">http://genome.wustl.edu/software/cnvHmm</a></td>
<td>C</td>
<td>consensus sequence from SAMtools</td>
<td>Using HMM to detect CNV</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*a* Tools require matched case-control sample as input.  
*b* Tools use multiple samples as input.

http://bioinfo.mc.vanderbilt.edu/CNVannotator/links.cgi
Tools for Exome Sequencing

Table 3 - Summary of bioinformatics tools for CNV detection using exome sequencing data

<table>
<thead>
<tr>
<th>Tool</th>
<th>URL</th>
<th>Language</th>
<th>Input</th>
<th>Comments</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control-FREEC</td>
<td><a href="http://bioinfo-out.curie.fr/projects/freec/">http://bioinfo-out.curie.fr/projects/freec/</a></td>
<td>C++</td>
<td>SAM/BAM/pileup, Eland, BED, SCAP, arachne, psi (BLAT) and Bovtke formats</td>
<td>Correcting copy number using matched case-control samples or GC contents</td>
<td>[27]</td>
</tr>
<tr>
<td>CoNIFER</td>
<td><a href="http://conifer.sf.net/">http://conifer.sf.net/</a></td>
<td>Python</td>
<td>BAM</td>
<td>Using singular value decomposition to normalize copy number and avoiding batch bias by integrating multiple samples</td>
<td>[28]</td>
</tr>
<tr>
<td>XHMM</td>
<td><a href="http://atgu.mgh.harvard.edu/xhmm/">http://atgu.mgh.harvard.edu/xhmm/</a></td>
<td>C++</td>
<td>BAM</td>
<td>Uses principal component analysis to normalize copy number and HMM to detect CNVs</td>
<td>[29]</td>
</tr>
<tr>
<td>ExomeCNV</td>
<td><a href="http://cran.r-project.org/web/packages/ExomeCNV">http://cran.r-project.org/web/packages/ExomeCNV</a></td>
<td>R</td>
<td>BAM/pileup</td>
<td>Using read depth and B-allele frequencies from exome sequencing data to detect CNVs and LOHs</td>
<td>[30]</td>
</tr>
<tr>
<td>CONTRA</td>
<td><a href="http://contra-cnv.sourceforge.net/">http://contra-cnv.sourceforge.net/</a></td>
<td>Python</td>
<td>SAM/BAM</td>
<td>Comparing base-level log-ratios calculated from read depth between case and control samples</td>
<td>[31]</td>
</tr>
<tr>
<td>CONDEX</td>
<td><a href="http://code.google.com/p/condex/">http://code.google.com/p/condex/</a></td>
<td>Java</td>
<td>Sorted BED files</td>
<td>Using HMM to identify CNVs</td>
<td>[32]</td>
</tr>
<tr>
<td>SeqGene</td>
<td><a href="http://seqgene.sourceforge.net">http://seqgene.sourceforge.net</a></td>
<td>Python, R</td>
<td>SAM/pileup</td>
<td>Calling variants, including CNVs, from exome sequencing data</td>
<td>[33]</td>
</tr>
<tr>
<td>PropSeq</td>
<td><a href="http://bioinformatics.nki.nl/ocs/">http://bioinformatics.nki.nl/ocs/</a></td>
<td>R, C</td>
<td>N/A</td>
<td>Using the read depth of the case sample as a linear function of that of control sample to detect CNVs</td>
<td>[34]</td>
</tr>
<tr>
<td>VarScan2</td>
<td><a href="http://genome.wustl.edu/software/varscan">http://genome.wustl.edu/software/varscan</a></td>
<td>Java</td>
<td>BAM/pileup</td>
<td>Using pairwise comparisons of the normalized read depth at each position to estimate CNV</td>
<td>[35]</td>
</tr>
<tr>
<td>ExoCNVTest</td>
<td><a href="http://www1.imperial.ac.uk/medicine/people/l.colly/">http://www1.imperial.ac.uk/medicine/people/l.colly/</a></td>
<td>Java, R</td>
<td>BAM</td>
<td>Identifying and genotyping common CNVs associated with complex disease</td>
<td>[36]</td>
</tr>
<tr>
<td>ExomeDepth</td>
<td><a href="http://cran.r-project.org/web/packages/ExomeDepth/index.html">http://cran.r-project.org/web/packages/ExomeDepth/index.html</a></td>
<td>R</td>
<td>BAM</td>
<td>Using beta-binomial model to fit read depth of WES data</td>
<td>[37]</td>
</tr>
</tbody>
</table>

*a* Control-FREEC accepts either matched case-control samples or single sample as input.

*b* Tools use multiple samples as input.

*c* Tools require matched case-control samples as input.

http://bioinfo.mc.vanderbilt.edu/CNVannotator/links.cgi
Combination Tools

Which One Should I Use!? 

Table 4 - Combinatorial bioinformatics tools for CNV detection using NGS data

<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Language</th>
<th>Input</th>
<th>Combination</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NovelSeq</td>
<td><a href="http://compbio.cs.sfu.ca/strvar.htm">http://compbio.cs.sfu.ca/strvar.htm</a></td>
<td>C</td>
<td>FASTA/SAM</td>
<td>PEM+AS</td>
<td>[38]</td>
</tr>
<tr>
<td>HYDRA</td>
<td><a href="http://code.google.com/p/hydra-sv/">http://code.google.com/p/hydra-sv/</a></td>
<td>Python</td>
<td>discordant paired-end</td>
<td>PEM+AS</td>
<td>[39]</td>
</tr>
<tr>
<td>CNVer</td>
<td><a href="http://compbio.cs.toronto.edu/CNVer/">http://compbio.cs.toronto.edu/CNVer/</a></td>
<td>Perl, C++</td>
<td>BAM/ aligned positions</td>
<td>PEM+RD</td>
<td>[40]</td>
</tr>
<tr>
<td>GASVPro</td>
<td><a href="http://code.google.com/p/gasv/">http://code.google.com/p/gasv/</a></td>
<td>C++</td>
<td>BAM</td>
<td>PEM+RD</td>
<td>[41]</td>
</tr>
<tr>
<td>Genome STRiP</td>
<td><a href="http://www.broadinstitute.org/software/">http://www.broadinstitute.org/software/</a></td>
<td>Java, R</td>
<td>BAM</td>
<td>PEM+RD</td>
<td>[42]</td>
</tr>
<tr>
<td>SVDetect</td>
<td><a href="http://svdetect.sourceforge.net/">http://svdetect.sourceforge.net/</a></td>
<td>Perl</td>
<td>SAM/BAM/ ELAND</td>
<td>PEM+RD</td>
<td>[43]</td>
</tr>
<tr>
<td>inGAP-sv</td>
<td><a href="http://ingap.sourceforge.net/">http://ingap.sourceforge.net/</a></td>
<td>Java</td>
<td>SAM</td>
<td>PEM+RD</td>
<td>[44]</td>
</tr>
<tr>
<td>Nord et al.</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>RD+SR</td>
<td>[46]</td>
</tr>
</tbody>
</table>

^RD: read depth-based approach; PEM: paired-end mapping approach; SR: split read approach; AS: de novo assembly approach.
Ask Yourself These Questions

• What did I sequence?
  – Whole genome, exome, custom
  – Human, mouse, fly, plant, other (genome size)

• How much data do I have?
  – Just a few individuals or many
  – Several tools require several samples for effective normalization

• What type of data do I have?
  – Cases only?
  – Cases and Controls? Tumor – Normal Pairs?
Next Consider

- You will want to select more than one method or a combination algorithm.
- You will want to test the algorithms yourself to ensure they are working correctly:
  - Is it easy to implement? Sample data? Does it run effective on your computer/cluster?
  - Trio inheritance (median transmission rate from parent to child converges to 50%)
  - How much of your data does it “throw out”?
  - Does it calculate QC metrics and/or provide filtering recommendations?
  - Biologic criteria (previous publications)
Calling CNVs – Tailor to the Array

- FASST, RANK, SegMNT segmentation algorithms used to create CNV sets
- Chose those CNVRs that overlapped across all three calling algorithms
- Deletions > Duplications
- Even distribution of unique vs. poly CNVRs
- Appropriate gene/exon content of CNVRs

Overview

• What are CNVs and why are they important?

• How can you find CNVs in your genomic datasets?

• How can you interpret the significance of CNVs discovered in your genomic datasets?
Interpretation of CNVs

Who are these gentlemen??

http://dailyheadlines.uark.edu/  http://content.usatoday.com/topics/photo
Interpretation of CNVs

James Watson

Craig Venter

http://dailyheadlines.uark.edu/

http://content.usatoday.com/topics/photo
Interpretation of CNVs

James Watson
Craig Venter
First sequenced diploid genomes
Interpretation of CNVs

23 CNVs
26kb-1.6Mb
9 duplications
14 deletions

62 CNVs
30 duplications
32 deletions
How Prevalent are CNVs?

- HapMap project suggests that ~12% of the human genome is copy number variable

- CNVs cover ~360Mb of the human genome (far more nucleotide content per genome than SNPs)

- Disease causing CNVs are typically rare, highly penetrant lesions (often *de novo*)
What is a Benign CNV?

• To determine which CNVs may be contributing to disease it is important to ascertain which CNVs are unlikely to cause disease.

• Determine which CNVs are present in “normal” or more accurately “control” populations:
  - Presumably do not have significant congenital disease.
  - Very common variation is probably safe to rule out but what about rare variation?
What is a Benign CNV?

- To determine which CNVs may be contributing to disease it is important to ascertain which CNVs are unlikely to cause disease.
Conrad et. al. 2010

Any VALIDATED CNV in the same direction (gain or loss) as well as ALL autosomal SegDups

>= 1% + 3 independent events OR >=0.5% and 5 independent events in the same direction (>30 controls, array studies only)

Cooper et. al. 2011

>= 12 control samples in “non-significant genes” in the same direction

>= 8 “Benign” or “VUS, LB” Interpretations in the same direction

UIHC Cytogenetics Local Database

>= 2% in the same direction
CNV Prioritization

~5,000 CNVs in our cohort

Benign (~98%)

Diagnostic

Involve > 0 RefSeq Genes

Known Pathogenic or Susceptibility Locus

Variant of Unclear Significance

Abnormal

Conrad et. al. 2009

Cooper et. al. 2011

Classification as Benign

• Typical overlap with a “benign” CNV necessary to consider a detected CNV as benign varies
  – >50% overlap commonly used
  – >90% is stringent (clinical use)

• Best Rules to Use?
  – Still an area of active investigation and development
  – Collaborative project between us and the Bioinformatics Core
  – Coming soon to the Galaxy Tool Shed
It’s Not Benign – Now What?

- Is it a known pathogenic or disease susceptibility CNV?

- Are there others like it in clinical databases?
  - If so, are the phenotypes similar?

- What is the gene content of the CNV?
  - Known disease genes (OMIM Morbid Map) or genomic loci? Inheritance and CNV copy number?
  - Candidate genes? Interaction with known disease gene product?
  - Predicted haploinsufficient genes?
  - Parent testing to determine inheritance (inherited or de novo)?
Haploinsufficiency and Disease

- Virtually any disease caused by haploinsufficiency of a gene could theoretically be caused by a deletion CNV (duplication and over-expression is less clear cut)

- What if you find a small deletion CNV involving a potential disease gene?
  - Predict haploinsufficiency
  - Huang et. al. PLOS Genetics. 2010. 6(10): e1001154
Overview

• What are CNVs and why are they important?

• How can you find CNVs in your genomic datasets?

• How can you interpret the significance of CNVs discovered in your genomic datasets?
Acknowledgements

Shivanand R. Patil
Cytogenetics and Molecular Laboratory

Agshin Taghiyev, Ph.D.

John Manak, Ph.D.

Pat Brophy, M.D.

Richard Smith, M.D.

Members of the Manak, Brophy and Smith Laboratories
The End
### Examples of algorithms for the detection of structural variants from array data.

<table>
<thead>
<tr>
<th>Software</th>
<th>Affymetrix</th>
<th>Illumina</th>
<th>CGH Method</th>
<th>Use allelic intensities</th>
<th>Multi-sample analysis</th>
<th>Copy number output</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCAT (Van Loo et al., 2010)</td>
<td>X</td>
<td>X X X X</td>
<td></td>
<td>Allele-specific piecewise constant fitting</td>
<td>X</td>
<td>Allele-specific copy number (CN)</td>
<td><a href="http://heima.ifi.uio.no/bioinf/projects/ASCAT/">http://heima.ifi.uio.no/bioinf/projects/ASCAT/</a></td>
</tr>
<tr>
<td>PICNIC (Greenman et al., 2010)</td>
<td>X</td>
<td></td>
<td></td>
<td>Hidden–Markov model (HMM)</td>
<td>X</td>
<td>Continuous CN + CN genotypes</td>
<td><a href="http://www.sanger.ac.uk/genetics/CGP/software/PICNIC/">http://www.sanger.ac.uk/genetics/CGP/software/PICNIC/</a></td>
</tr>
<tr>
<td>PennCNV (Wang et al., 2007)</td>
<td>X X* X X X</td>
<td>X</td>
<td></td>
<td>HMM</td>
<td>X</td>
<td>Trfos only</td>
<td><a href="http://www.openbioinformatics.org/penncnv/">http://www.openbioinformatics.org/penncnv/</a></td>
</tr>
<tr>
<td>QuantiSNP (Colella et al., 2007)</td>
<td>X X* X X X</td>
<td>X</td>
<td></td>
<td>HMM</td>
<td>X</td>
<td>Discrete CN + CN genotypes</td>
<td><a href="http://www.sanger.ac.uk/QuantSNP/">http://www.sanger.ac.uk/QuantSNP/</a></td>
</tr>
<tr>
<td>Affymetrix.arena (Bengtsson et al., 2008)</td>
<td>X X</td>
<td></td>
<td></td>
<td>Copy number estimation using robust multichip analysis (CRMA)</td>
<td>X</td>
<td>Unclassified segments</td>
<td><a href="http://www.aroma-project.org/">http://www.aroma-project.org/</a></td>
</tr>
<tr>
<td>GADA (Pique-Regi, 2008)</td>
<td>X X X X X</td>
<td>X</td>
<td></td>
<td>Sparse Bayesian learning</td>
<td>X</td>
<td>Unclassified segments</td>
<td><a href="http://bion.usc.edu/~piqueg/GADA/">http://bion.usc.edu/~piqueg/GADA/</a></td>
</tr>
</tbody>
</table>


### CMA CNV Detection Algorithms

- Multiple algorithms for detecting CNVs in microarray data
  - Many based off Hidden Markov Models
  - CBS one of the earliest and most heavily used
  - ASCAT useful for determining tumor purity also
CMA CNV Detection Algorithms

### Table 1

<table>
<thead>
<tr>
<th>Platform type</th>
<th>Platform</th>
<th>No. of probes</th>
<th>Avg. probe length (bp)</th>
<th>Site</th>
<th>CNV analysis tool</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Birdsuite</td>
</tr>
<tr>
<td>CGH</td>
<td>Sanger WGTP</td>
<td>29,043</td>
<td>170,000</td>
<td>WTSI</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Agilent 244K</td>
<td>236,381</td>
<td>60</td>
<td>TCAG</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Agilent 2x244K</td>
<td>462,609</td>
<td>60</td>
<td>TCAG</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>NimbleGen 720K</td>
<td>720,412</td>
<td>60</td>
<td>WTSI</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>NimbleGen 2.1M</td>
<td>2,161,679</td>
<td>60</td>
<td>HMS, WTSI</td>
<td>X</td>
</tr>
<tr>
<td>SNP</td>
<td>Affymetrix 500K</td>
<td>500,568</td>
<td>25</td>
<td>TCAG</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Illumina 650Y</td>
<td>650,918</td>
<td>50</td>
<td>TCAG</td>
<td>X</td>
</tr>
<tr>
<td>SNP + CNV probes</td>
<td>Affymetrix 6.0</td>
<td>1,856,600</td>
<td>25</td>
<td>TCAG, WTSI</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Illumina 1M</td>
<td>1,072,820</td>
<td>50</td>
<td>HMS, TCAG</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Illumina 660W</td>
<td>657,366</td>
<td>50</td>
<td>HMS, TCAG</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Illumina Omni</td>
<td>1,140,419</td>
<td>50</td>
<td>HMS, TCAG</td>
<td>X</td>
</tr>
</tbody>
</table>

- Additional algorithms that can be used with microarray data to detect CNVs
- Reference below provides a comparison of the different algorithms and array types

Hidden Markov Model

Goal of the HMM is to find the best “state path” using the symbols (ACGT), states (E, 5, I), emission probabilities (top), and transition probabilities (arrows) defined for a specific problem.