MAPPING COPY NUMBER VARIATION BY POPULATION-SCALE GENOME SEQUENCING (1000 GENOME PROJECT)

Published in Nature, 3 February 2011

BioInformaatikute TeadusArtiklite Lugemise Klubi ettekanne
Maido Remm, 19.09.2011
Mapping copy number variation by population-scale genome sequencing

Genomic structural variants (SVs) are abundant in humans, differing from other forms of variation in extent, origin and functional impact. Despite progress in SV characterization, the nucleotide resolution architecture of most SVs remains unknown. We constructed a map of unbalanced SVs (that is, copy number variants) based on whole genome DNA sequencing data from 185 human genomes, integrating evidence from complementary SV discovery approaches with extensive experimental validations. Our map encompassed 22,025 deletions and 6,000 additional SVs, including insertions and tandem duplications. Most SVs (53%) were mapped to nucleotide resolution, which facilitated analysing their origin and functional impact. We examined numerous whole and partial gene deletions with a genotyping approach and observed a depletion of gene disruptions amongst high frequency deletions. Furthermore, we observed differences in the size spectra of SVs originating from distinct formation mechanisms, and constructed a map of SV hotspots formed by common mechanisms. Our analytical framework and SV map serves as a resource for sequencing-based association studies.
OBJECTIVES OF THE STUDY:

1. To compare performance of different methods and algorithms for discovery of structural variants (SV) from sequencing data.

2. To create a list of all SVs of 50 bp and larger in size within studied individuals for further reference.

Initial focus was on **deletions**. Less focus was placed on insertions and duplications. The balanced variations (inversions and chromosomal rearrangements) were not studied.
DATA:

- High-coverage sequences (42x coverage)
  - 1 parent-offspring trio from CEU
  - 1 parent-offspring trio from YRI

- Low-coverage sequences (3.6x coverage)
  - 60 CEU
  - 60 JPT+CHB
  - 59 YRI
4 ALGORITHMS, 19 METHODS

- 6 methods using Read-Pair (RP)
- 4 methods using Read-Depth (RD)
- 4 methods using Split-Read (SR)
- 3 methods using local Sequence Assembly (AS)
- 2 methods using combination of RP and RD (PD)

Color-coding:

- AS
- RP
- SR
- RL
- PD
- RD
19 methods were applied separately to low-coverage and high-coverage data and deletions and insertions were collected into separate datasets (callsets).

Altogether 36 callsets: 15 callsets for low-coverage data and 21 callsets for high-coverage data (trios).
## LOW-COVERAGE CALLSETS

<table>
<thead>
<tr>
<th>Approach</th>
<th>Callset Origin</th>
<th>Discovery Algorithm Name and Reference*</th>
<th>Platform</th>
<th>Mapping Algorithm</th>
<th>Genomes Analyzed</th>
<th>SV Type</th>
<th>Algorithm Parameters Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD</td>
<td>AE</td>
<td>N/A&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MAQ</td>
<td>8</td>
<td>DEL</td>
<td>window size (&lt;500bp); p-value (&lt;10^-5)</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>Event-wise testing&lt;sup&gt;15,#&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MAQ</td>
<td>162</td>
<td>DEL</td>
<td>read mapping quality(&lt;Q30); window size (100bp); cluster size with merged events of same type (500bp); read depth (4.75 and a 1.25 mean read depth); significance level (&lt;1e-10); event size (1 kb); absolute difference between median read counts (&lt;0.5)</td>
</tr>
<tr>
<td></td>
<td>YL</td>
<td>CNVnator&lt;sup&gt;10&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MAQ</td>
<td>65</td>
<td>DEL</td>
<td>NA</td>
</tr>
<tr>
<td>RP</td>
<td>BC</td>
<td>Spanner&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MOSAIK</td>
<td>138</td>
<td>DEL</td>
<td>maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwith (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100); mapping distance (P-value&lt;0.99); minimum read-pairs (4, 2 from each side); map distance to annotated loci (&lt;400bp); gap between the F and R clusters (&lt;30 bp &lt; gap &lt; 500 bp)</td>
</tr>
<tr>
<td></td>
<td>SI</td>
<td>N/A&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MAQ</td>
<td>144</td>
<td>INS</td>
<td>maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwith (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100); MAX mapping quality (&gt;20); read-pairs in a cluster (&gt;2), start/end distance (&lt;10 x median absolute deviation of the insert size distribution); event size (&lt;1Mb) span-size (within 15% deviation from the median of span-size)</td>
</tr>
<tr>
<td></td>
<td>YL</td>
<td>PEMer&lt;sup&gt;16,23&lt;/sup&gt;</td>
<td>SOLiD</td>
<td>CORONA</td>
<td>25</td>
<td>DEL</td>
<td>RMAQ mapping quality (&gt;35); outer distance (&gt; mean + 4stdev of the insert size)</td>
</tr>
<tr>
<td></td>
<td>WU</td>
<td>BreakDancer</td>
<td>Illumina</td>
<td>MAQ</td>
<td>138</td>
<td>DEL</td>
<td>hash size (15bp); mismatch bases in alignments (&gt;5%); match bases aligned to one of the mobile element consensus sequences (40bp); gap length (&lt;60bp); alignment quality score (&lt;40); mobile element alignement length (&lt;60bp); distance from annotated mobile elements (&lt;100bp)</td>
</tr>
<tr>
<td>SR</td>
<td>BC</td>
<td>Mosaik&lt;sup&gt;13&lt;/sup&gt;</td>
<td>454</td>
<td>MOSAIK</td>
<td>22</td>
<td>INS</td>
<td>MAQ mapping quality (&gt;0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>Pindel&lt;sup&gt;17&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MAQ</td>
<td>145</td>
<td>DEL</td>
<td>MAQ mapping quality (&gt;0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>Pindel&lt;sup&gt;17&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MAQ</td>
<td>145</td>
<td>INS</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>YL</td>
<td>N/A&lt;sup&gt;13&lt;/sup&gt;</td>
<td>454</td>
<td>BLAT</td>
<td>5</td>
<td>DEL</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>YL</td>
<td>N/A&lt;sup&gt;13&lt;/sup&gt;</td>
<td>454</td>
<td>BLAT</td>
<td>5</td>
<td>INS</td>
<td>NA</td>
</tr>
<tr>
<td>PD</td>
<td>BC</td>
<td>Spanner&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MOSAIK</td>
<td>138</td>
<td>TDUP</td>
<td>mapping quality values of read pairs (&gt;30); mapping distance between the pairs (p-value&lt;0.02%); number of supporting read pairs (&gt;3); minimum deletion size (50bp); &quot;Alignability&quot; in the clustered regions (&gt;0.01); Net read coverage over all samples (&lt;2.5 x the expected coverage); event length (&lt;250bp); copy number (&gt;2.2) clusters of paired-ends (N &gt; 2); apparent insert size (&gt; the median of the insert size distribution + 10 x the median absolute deviation of insert size from the median for that lane/library)</td>
</tr>
<tr>
<td></td>
<td>BI</td>
<td>Genome STRIP&lt;sup&gt;18&lt;/sup&gt;</td>
<td>Illumina</td>
<td>MAQ</td>
<td>168</td>
<td>DEL</td>
<td>NA</td>
</tr>
</tbody>
</table>
## High-Coverage Callsets

<table>
<thead>
<tr>
<th>Approach</th>
<th>Callset Origin</th>
<th>Discovery Algorithm Name Reference</th>
<th>Platform</th>
<th>Mapping Algorithm</th>
<th>Genomes analyzed</th>
<th>SV Type</th>
<th>Algorithm Parameters Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td></td>
<td>Event-wise testing</td>
<td>Illumina</td>
<td>MAQ</td>
<td>6</td>
<td>DEL</td>
<td>read mapping quality (q20); window size (1000bp); cluster size with merged events of same type (500bp); read depth (4×7.5× and 8×1.2× mean read depth); significance level P&lt;10^-6; event size (¼ INDEL); absolute difference between median read counts (¼ INDEL); median read depth (¼ INDEL); common deletion regions (¼ INDEL); RepeatMasker (on human reference genome build 35, with the sensitivity option “x5” enabled); Tandem Repeats Finder (search tandem repeats ≤50bp); edit distance (¼ INDEL); unique P(extend 5kbp of unmapped sequence); windows (¼ INDEL; copy number ¼ INDEL)</td>
</tr>
<tr>
<td>UW</td>
<td></td>
<td>mrFAST</td>
<td>Illumina</td>
<td>mrFAST</td>
<td>6</td>
<td>DEL</td>
<td>read mapping quality (q20); window size (1000bp); cluster size with merged events of same type (500bp); read depth (4×7.5× and 8×1.2× mean read depth); significance level P&lt;10^-6; event size (¼ INDEL); absolute difference between median read counts (¼ INDEL); median read depth (¼ INDEL); common deletion regions (¼ INDEL); RepeatMasker (on human reference genome build 35, with the sensitivity option “x5” enabled); Tandem Repeats Finder (search tandem repeats ≤50bp); edit distance (¼ INDEL); unique P(extend 5kbp of unmapped sequence); windows (¼ INDEL; copy number ¼ INDEL)</td>
</tr>
<tr>
<td>YL</td>
<td></td>
<td>CNVnator</td>
<td>Illumina</td>
<td>MAQ</td>
<td>6</td>
<td>DEL</td>
<td>read mapping quality (q20); window size (1000bp); cluster size with merged events of same type (500bp); read depth (4×7.5× and 8×1.2× mean read depth); significance level P&lt;10^-6; event size (¼ INDEL); absolute difference between median read counts (¼ INDEL); median read depth (¼ INDEL); common deletion regions (¼ INDEL); RepeatMasker (on human reference genome build 35, with the sensitivity option “x5” enabled); Tandem Repeats Finder (search tandem repeats ≤50bp); edit distance (¼ INDEL); unique P(extend 5kbp of unmapped sequence); windows (¼ INDEL; copy number ¼ INDEL)</td>
</tr>
<tr>
<td>AB</td>
<td></td>
<td>AB large indel tool</td>
<td>SOLID</td>
<td>MAPREADS</td>
<td>1</td>
<td>DEL</td>
<td>read mapping quality (q20); read pairs in a cluster (¼ INDEL); start/stop distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td>Spanner</td>
<td>Illumina</td>
<td>MOSAIK</td>
<td>6</td>
<td>DEL</td>
<td>maximum mismatch threshold (4 for 35-43mer reads, 6 for 44-53mers, and ½ for 64mers and longer); hash size (¼ INDEL); Smith-Waterman bandwidth (¼ INDEL); alignment candidate threshold (¼ INDEL); local alignment search radius (¼ INDEL); hash position threshold (¼ INDEL); maximum read count (¼ INDEL); gap between the F and P clusters (¼ INDEL); gap ≤50 bp; gap ≤100 bp; maximum mismatch threshold (¼ INDEL); alignment candidate threshold (¼ INDEL); local alignment search radius (¼ INDEL); hash position threshold (¼ INDEL)</td>
</tr>
<tr>
<td>SI</td>
<td></td>
<td>N/A</td>
<td>Illumina</td>
<td>MAQ</td>
<td>6</td>
<td>DEL</td>
<td>maximum mismatch quality (q20); read pairs in a cluster (¼ INDEL); start/stop distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>UW</td>
<td></td>
<td>Variation Hunter</td>
<td>Illumina</td>
<td>mrFAST</td>
<td>6</td>
<td>DEL</td>
<td>high-quality reads (average phred score ≥30); add distance (¼ INDEL); unique P(extend 5kbp of unmapped sequence); windows (¼ INDEL; copy number ¼ INDEL)</td>
</tr>
<tr>
<td>WU</td>
<td></td>
<td>BreakDancer</td>
<td>Illumina</td>
<td>MAQ</td>
<td>6</td>
<td>DEL</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>YL</td>
<td></td>
<td>PEMer</td>
<td>454</td>
<td>PEM</td>
<td>1</td>
<td>INS</td>
<td>read mapping quality (q20); window size (1000bp); cluster size with merged events of same type (500bp); read depth (4×7.5× and 8×1.2× mean read depth); significance level P&lt;10^-6; event size (¼ INDEL); absolute difference between median read counts (¼ INDEL); median read depth (¼ INDEL); common deletion regions (¼ INDEL); RepeatMasker (on human reference genome build 35, with the sensitivity option “x5” enabled); Tandem Repeats Finder (search tandem repeats ≤50bp); edit distance (¼ INDEL); unique P(extend 5kbp of unmapped sequence); windows (¼ INDEL; copy number ¼ INDEL)</td>
</tr>
<tr>
<td>YL</td>
<td></td>
<td>PEMer</td>
<td>454</td>
<td>PEM</td>
<td>1</td>
<td>INS</td>
<td>read mapping quality (q20); window size (1000bp); cluster size with merged events of same type (500bp); read depth (4×7.5× and 8×1.2× mean read depth); significance level P&lt;10^-6; event size (¼ INDEL); absolute difference between median read counts (¼ INDEL); median read depth (¼ INDEL); common deletion regions (¼ INDEL); RepeatMasker (on human reference genome build 35, with the sensitivity option “x5” enabled); Tandem Repeats Finder (search tandem repeats ≤50bp); edit distance (¼ INDEL); unique P(extend 5kbp of unmapped sequence); windows (¼ INDEL; copy number ¼ INDEL)</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td>Mosaik</td>
<td>454</td>
<td>MOSAIK</td>
<td>2</td>
<td>INS</td>
<td>hash size (¼ INDEL); maximum mismatch quality (¼ INDEL); hash size (¼ INDEL); maximum deletion size (¼ INDEL); number of fragments for unmapped reads (¼ INDEL; for deletion and ¼ INDEL for short insertions)</td>
</tr>
<tr>
<td>LN</td>
<td></td>
<td>Pindol</td>
<td>Illumina</td>
<td>MAQ</td>
<td>6</td>
<td>DEL</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>YL</td>
<td></td>
<td>N/A</td>
<td>454</td>
<td>BLAT</td>
<td>6</td>
<td>INS</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>YL</td>
<td></td>
<td>N/A</td>
<td>454</td>
<td>BLAT</td>
<td>6</td>
<td>INS</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>BG</td>
<td></td>
<td>SOAPdenovo</td>
<td>Illumina</td>
<td>SOAP</td>
<td>6</td>
<td>DEL</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>BG</td>
<td></td>
<td>SOAPdenovo</td>
<td>Illumina</td>
<td>SOAP</td>
<td>6</td>
<td>DEL</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>OX</td>
<td></td>
<td>Cortex</td>
<td>Illumina</td>
<td>CORTEX</td>
<td>1</td>
<td>INS</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>OX</td>
<td></td>
<td>Cortex</td>
<td>Illumina</td>
<td>CORTEX</td>
<td>1</td>
<td>INS</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>UW</td>
<td></td>
<td>NovelSeq</td>
<td>Illumina</td>
<td>MAQ</td>
<td>6</td>
<td>INS</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td>Spanner</td>
<td>Illumina</td>
<td>MOSAIK</td>
<td>6</td>
<td>TDUP</td>
<td>maximum mapping quality (¼ INDEL); outer distance (¼ INDEL; median absolute deviation of the insert size distribution); event size (¼ INDEL)</td>
</tr>
</tbody>
</table>
## Sensitivity and FDR

### Real situation:
(can be tested by PCR or microarray)

<table>
<thead>
<tr>
<th>Positive (P)</th>
<th>Negative (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

### Software prediction results:

<table>
<thead>
<tr>
<th>Positive</th>
<th>True Positive (TP)</th>
<th>False Positive (FP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

- **Sensitivity:** $Sn = \frac{TP}{TP + FN}$
- **Specificity:** $Sp = \frac{TN}{FP + TN}$
- **Accuracy:** $ACC = \frac{(TP + TN)}{(P + N)}$
- **False Discovery Rate:** $FDR = \frac{FP}{TP + FP}$
Sensitivity in detecting deletions estimated for three gold standard sources, i.e., sets of published deletions (Conrad, 2010; McCarroll, 2008; Kidd, 2008; Mills, 2006). SVs in these publications were identified with capillary sequencing (median=0.2kb), tiling CGH microarrays (median=2kb), and fosmid sequencing (median=6kb).

Only 1bp overlap required for recording positive prediction!

Individual methods show sensitivity between 0% and 80%.

In final "release set" sensitivity was 69% (low-coverage set) to 82% (high-coverage set).

With more stringent sensitivity criterion (>50% overlap) the sensitivity was 51% (low-coverage) to 70% (high-coverage).
Findings in each callset were validated using PCR and CGH. PCR primers were designed for randomly chosen SV predictions from each callset. Custom array-CGH DNA Microarrays were used to validate deletions and duplications in the high coverage trios. Affy 6.0, Illumina 1.0 and NimbleGen 2.1M arrays were also used for some individuals.

\[
FDR = \frac{CGH_{invalidated}}{CGH_{validated} + CGH_{invalidated}} \times \frac{CGH_{validated} + CGH_{invalidated}}{N} + \frac{PCR_{invalidated}}{PCR_{validated} + PCR_{invalidated}} \times (1 - \frac{CGH_{validated} + CGH_{invalidated}}{N})
\]

Final FDR is weighted average from both experiments.
PCR and microarrays have only moderate agreement with each other on presence of structural variants.
Sensitivity and FDR on 2 individuals (low- and high coverage).

- CNVnator (Gerstein, YaleU)
- Spanner (Marth, BostonC)
- GenomeSTRiP (McCarroll, Broad Institute)
CONCLUSIONS

- None of the sequence-based methods is reliable for individual SV calling in inheritance studies or in medical diagnostics. For GWAS studies ??

- For example, one of the best methods Spanner has:
  - FDR ca 9% and
  - sensitivity ca 40%
  in high-coverage deletion callset.
RELEASE SET

- For final release only methods with overall FDR<10% were used + some experimentally validated SVs.

- These methods were:
  Spanner (from Marth group, Boston College)
  Mosaik (from Marth group, Boston College)
  GenomeSTRiP (from McCarroll group, Broad Institute)
RELEASE SET OF STRUCTURAL VARIATIONS:

- 28,000 structural variations described from given individuals (cell lines)
  + 22,000 deletions,
  + 5,400 mobile element insertions,
  + 500 duplications,
  + 100 insertions

Half of these were "novel" SVs, missing from dbVAR, DGV and from other sequenced genomes.
MAPPING OF BREAKPOINTS:

- Sequence data allows mapping of breakpoints with single nucleotide precision. This was done for ca 15000 SVs.

- Different methods have different precision
Sequence data allows mapping of breakpoints in single nucleotide precision. This was done for ca 15000 SVs.
Common SVs (MAF > 5%) were typically shared across populations, whereas rare alleles were frequently observed in only one population.

81% of deletions display linkage disequilibrium (LD) with SNPs at level $r^2 > 0.8$
11,000 nucleotide-level deletions were compared to primate genomes using BreakSeq classification approach (Nat. Biotechnology, 2010).

- Only 60% confirmed as deletions
- 23% are actually duplications
- 17% undetermined

\[
\frac{60}{60+23} = 28\% \\
\text{of determined deletions are NOT deletions. They only look like deletions because all comparisons are done wrt reference genome (single individual).}
\]
MECHANISM OF DELETION AND INSERTION

Deletion types

MEI: mobile element insertion
VNTR: variable number of tandem repeats (polymerase slippage)
NAHR: non-allelic homologous recombination (error of recombination)
NH: non-homologous end joining (DNA repair mechanisms)

Insertion types

MEI: mobile element insertion
VNTR: variable number of tandem repeats (polymerase slippage)
NAHR: non-allelic homologous recombination (error of recombination)
NH: non-homologous end joining (DNA repair mechanisms)
51 hotspots of SVs over the entire genome were detected, 6 of them are in regions of known genetic disorders previously associated with recurrent *de novo* deletions, including Miller-Dieker syndrome and Leri-Weill dyschondrosteosis.
CONCLUSIONS

- Sequencing-based methods are not yet reliable for most types of SV analyses. Even GWAS might be problematic.

- Reference genome is not representing ancestral state. Better to compare with ancestral genome.

- 28,000 SVs available from 1000GP webpage, majority of them are mapped to single nucleotide precision.