Next-generation sequence characterization of complex genome structural variation

Can Alkan
Department of Genome Sciences,
University of Washington
Seattle, WA, USA
Size range of genetic variation

- Single nucleotide (SNPs)
- Up to ~50bp (small indels, microsatellites)
- >50bp to several megabases (structural variants):
  - Deletions
  - Insertions
    - Novel sequence
    - Mobile elements (Alu, L1, SVA, etc.)
  - Segmental Duplications
    - Duplications of size ≥ 1 kbp and sequence similarity ≥ 90%
      - Tandem or Interspersed
  - Inversions
  - Translocations
- Chromosomal changes
Human genome structural variation

Duplications and CNV hotspots

(Bailey et al., Science, 2002)
CNVs and disease

- **17q21.31 Deletion** (Sharp et al. *Nat Gen* 2006)
  - ~1% mental retardation in European populations
- **15q24.1 Deletion** (Sharp et al. *HMG* 2007)
  - autism spectrum disorder/MR/growth deficiency
- **17q12 Deletion** (Mefford et al. *AJHG* 2007)
  - 20% of patients with renal disease due to hypodysplastic kidneys and 36% of children with MODY-5
- **1q21.1 Deletion** (Mefford et al. *NEJM* 2008)
  - 0.4% of mental retardation and 0.3% schizophrenia.
## Multi-copy CNP and disease

<table>
<thead>
<tr>
<th>CNP</th>
<th>Variation</th>
<th>Disease/Effect</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4A/C4B</td>
<td>decrease</td>
<td>lupus</td>
<td>Yang, 2007</td>
</tr>
<tr>
<td>CCL3L1</td>
<td>decrease</td>
<td>HIV susceptibility</td>
<td>Gonzalez, 2005</td>
</tr>
<tr>
<td>FCGR3B</td>
<td>decrease</td>
<td>glomerulonephritis</td>
<td>Aitman, 2006; Fanciulli, 2008</td>
</tr>
<tr>
<td>IRGM</td>
<td>deletion, deletion</td>
<td>Crohn's disease, alternative splicing</td>
<td>McCarroll, 2008; Bekpen, 2009</td>
</tr>
</tbody>
</table>

** correspond to more ancient primate segmental duplications

- Previous work largely array based
  - Biased against duplicated regions due to signal saturation
  - Limited to CNVs (no balanced events)
  - Give only (relative) copy number
    - No information on variant sequence or organization

- **Goal:** systematically discover and characterize genomic structural variation and *copy, content, and structure* of segmental duplications.
# Genome-wide SV Discovery Approaches

## Hybridization-based
- Iafrate et al., 2004, Sebat et al., 2004
- SNP microarrays: McCarroll et al., 2008, Cooper et al., 2008, Itsara et al., 2009
- Array CGH: Redon et al. 2006, Conrad et al., 2010, Park et al., 2010, WTCCC, 2010

## Sequencing-based
- Read-depth: Bailey et al., 2002
- Sanger sequencing: Mills et al., 2006
  - 1000 Genomes Project

## Single molecule analysis
- Optical mapping: Teague et al., 2010

## Hybridization-based

## Sequencing-based
Detection diversity

Gains & Losses > 5 Kbp in the same 5 individuals

Fosmid clone End-sequence pair Kidd et al., 2008 (N = 1,206)

Ultra-dense tiling array CGH Conrad et al., 2010 (N = 1,128)

Affymetrix 6.0 SNP microarray McCarroll et al., 2008 (N = 236)

Kidd et al. Cell, 2010
Resequencing Genomes

Test genome

Random shearing and Size-selection

Paired-end sequencing

Read mapping

Reference Genome (HGP)
Sequence signatures of structural variation

- Read pair analysis
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size
- Read depth analysis
  - Deletions and duplications only
  - Relatively poor breakpoint resolution
- Split read analysis
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution
- Local and de novo assembly
  - SV in unique segments
  - 1bp breakpoint resolution
SV callers in the 1000 GP

Read Pair
U. Washington: VariationHunter (Illumina)
WUGSC: BreakDancer (Illumina)
ABI/LifeTech: Corona Light (SOLiD)
Yale: PEMer (Roche/454)
Wellcome-Trust: (unnamed) (Illumina)
BGI: (unnamed) (Illumina)

Read Depth
U. Washington: WSSD (Illumina)
Yale: CNVnator (Illumina)
UCSD/CSHL: EWT/RDXplorer (Illumina)
AECOM: (unnamed) (Illumina)

Split Read
Leiden/WTSI: Pindel (Illumina)
Yale: (unnamed) (Roche/454)

Read Pair + Depth
Boston College: Spanner (Illumina)
Broad: Genome STRiP (Illumina)

Assembly
U. Washington: NovelSeq (Illumina)
BGI: SOAPdenovo (Illumina)
EBI: Cortex (Illumina)

Breakpoint assembly
WUGSC: TIGRA (Illumina)

Upcoming from 1000Genomes: Delly/Invy (EMBL Heidelberg) + updates to the above
Non-1000G: MoDIL, MOGuL, CommonLAW, CNVer + GASV + HYDRA + many more
See Alkan et al., Nat Rev Genet, 2011, Mills et al. Nature 2011, and
Medvedev et al., Nat Methods, 2009
Challenges and algorithms for NGS

- Short read mapping artifacts:
  - Discard non-unique: lose sensitivity and detection power.
  - Try to utilize all information: higher sensitivity, less specificity.
- *micro-read Fast Alignment Search Tools*

```
Sequence reads

mrFAST
mrsFAST
drFAST
```

```
Read Depth
Read Pair
Local and *de novo* assembly

WSSD
Segmental duplications
Large Deletions

VariationHunter
Deletions, small novel insertions,
mobile element insertions,
inversions

NovelSeq
Novel sequence insertions
> 200 bp
```
[mr(s) | dr]FAST algorithms

- *mr(s)*FAST: Designed for reads generated with the Illumina platform
  - *mr*FAST: Small indels up to 4 bp are supported
  - *mrs*FAST: Mismatch-only version for increased mapping speed
- *dr*FAST: Di-base (color-space) read version for the SOLiD platform (Hamming distance only)
- Guaranteed sensitivity within user-specified edit (Hamming for mrsFAST and drFAST) distance threshold $d$:
  - $d \leq (\text{read\_length} / k) - 1$ ; $[k=12]$
- **Iterative search**: All locations and underlying sequence variation are returned
  - “best” mapping option for single-location mapping available; minimize edit distance & paired-end span size closest to median

http://mrfast.sourceforge.net
http://drfast.sourceforge.net

Multiple vs. unique mapping

Modified from Chiang & McCarroll, Nat Biotech, 2009
1) Read depth - copy number correlation

R = 0.92

Alkan et al., Nature Genetics, 2009
Personal duplication maps

- Next-gen specific error correction (GC% normalization, resolving repeats)

- Copy numbers predicted in 1kb non-overlapping windows
- 3.8% (662/17601) genes show a difference of at least one copy
  - 113/662 validated with arrayCGH
  - Most others are in high-copy regions, biased against aCGH
- Absolute copy number counts and characterization of the content of duplications made possible for the first time

<table>
<thead>
<tr>
<th>Individual</th>
<th>#WGS Reads</th>
<th>Read Length</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jim Watson</td>
<td>74,198,831</td>
<td>266bp*</td>
<td>454 FLX Wheeler 2008</td>
</tr>
<tr>
<td>HapMap NA18507 (YRI)</td>
<td>1,776,928,308</td>
<td>36bp</td>
<td>Illumina Bentley 2008</td>
</tr>
<tr>
<td>YH (Han Chinese)</td>
<td>1,315,249,404</td>
<td>35bp</td>
<td>Illumina Wang 2008</td>
</tr>
</tbody>
</table>

* Rendered into 509,667,772 reads of length 36bp

http://mrcanavar.sourceforge.net

Alkan et al., Nature Genetics, 2009
Personal duplication maps

- Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child

http://mrcanavar.sourceforge.net

Alkan et al., Nature Genetics, 2009
FISH: defensin

Associated with psoriasis and Crohn’s disease

Alkan et al., Nature Genetics, 2009
FISH validation: NPEPFS

Alkan et al., Nature Genetics, 2009
FISH: Morpheus gene family

Primate-specific duplication

Alkan et al., Nature Genetics, 2009
Scaling up: 1000 Genomes and more

Individuals sequenced in Pilot 1

- Yoruba 56
- CEPH 48
- Japanese 27
- Han 29

Histogram of Pilot 1 Illumina effective coverage

Individuals sequenced in Pilot 2

<table>
<thead>
<tr>
<th>ID</th>
<th>Effective Coverage</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA19240</td>
<td>24</td>
<td>YORUBA</td>
</tr>
<tr>
<td>NA19239</td>
<td>19</td>
<td>YORUBA</td>
</tr>
<tr>
<td>NA19238</td>
<td>13</td>
<td>YORUBA</td>
</tr>
<tr>
<td>NA12891</td>
<td>21</td>
<td>CEPH</td>
</tr>
<tr>
<td>NA12892</td>
<td>18</td>
<td>CEPH</td>
</tr>
<tr>
<td>NA12878</td>
<td>22</td>
<td>CEPH</td>
</tr>
</tbody>
</table>

Other Genomes

<table>
<thead>
<tr>
<th>ID</th>
<th>Effective Coverage</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>YH-1</td>
<td>22</td>
<td>HAN CHINESE°</td>
</tr>
<tr>
<td>NA18507</td>
<td>29</td>
<td>YORUBA ‡</td>
</tr>
<tr>
<td>NA18506</td>
<td>30</td>
<td>YORUBA *</td>
</tr>
<tr>
<td>NA18508</td>
<td>25</td>
<td>YORUBA *</td>
</tr>
<tr>
<td>KOREAN</td>
<td>12</td>
<td>KOREAN †</td>
</tr>
</tbody>
</table>

Sudmant, Kitzman, et al., Science, 2010
Increased \textit{CCL3L1} CN in Africans

Associated with HIV susceptibility

\textit{Sudmant, Kitzman, et al., Science, 2010}
Smaller *LPA* structure in Africans

More copies: protective against coronary heart disease

Differentiating Paralogous Genes

Associated with psoriasis and Crohn’s disease

CFHR

Associated with color blindness

opsin

Alkan et al., Nature Genetics, 2009
Singly Unique Identifiers (SUNs)

Copy 1  ATACTAGGCAATATAATATCCGAAGATATACATATAATGTTAG
Copy 2  ATGCTAGGCATGTAATATCCGAAGACGACATACATACATACATGTTAG
Copy 3  ATACTAGGCAATATACATCCGAAGATATACATACATACATGTTAG
Copy 4  ATGCTAAGGCATATACATATCCACGATATACATACATACATGTTAG
Copy 5  ATGCTAAGGCATATAATATCCGAAGATATACATACATACATGTTAG
Copy 6  ATACTAGGCAATGTAATATCCGAAGATATACATACATACATGTTAG
2) Read pair analysis: VariationHunter

- VariationHunter: Maximum parsimony approach; using all discordant map locations; finds an optimal set of SVs through a combinatorial (greedy) algorithm based on approximation to set-cover.

http://variationhunter.sourceforge.net (coming soon)

Hormozdiari, Alkan, et al, Genome Res. 2009
Definitions

Paired-end read

$\text{PE} := (\text{PE}_L, \text{PE}_R)$

PE-Alignment

$(\text{PE}, L(\text{PE}), R(\text{PE}), O(\text{PE}))$

$O(\text{PE})$: mapping orientation:

- “+/-”: normal
- “+/+” or “-/-”: inversion
- “-/+”: tandem duplication

$\text{SV} = (P_L, P_R, L_{\text{min}}, L_{\text{max}})$

F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009
Mathematical model

Let $L_{\text{min}}$, $L_{\text{max}}$ be minimum and maximum size of the predicted variant.

A Structural Variation is defined by event:

$$SV = (P_L, P_R, L_{\text{min}}, L_{\text{max}})$$

A PE-Alignment $APE=(PE, L(PE), R(PE), O(PE))$ supports an insertion $SV = (P_L, P_R, L_{\text{min}}, L_{\text{max}})$ if:

$$L(PE) \leq P_L$$
$$R(PE) \geq P_R$$
$$L_{\text{min}} \geq \Delta_{\text{min}} - (R(PE) - L(PE))$$
$$L_{\text{max}} \leq \Delta_{\text{max}} - (R(PE) - L(PE))$$

F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster $C$ is a valid cluster supporting insertions if:

$\exists \text{loc}, \forall APE \in C: L(APE) < \text{loc} < R(APE)$

$\exists \text{InsLen}, \forall APE \in C: \Delta_{\text{min}} - (R(APE') - L(APE)) < \text{InsLen} < \Delta_{\text{max}} - (R(APE) - L(APE))$

\[ F \text{Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009 } \]
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster C is a valid cluster supporting insertions if:

\[ \exists \text{loc}, \forall \text{APE} \in C : L(\text{APE}) < \text{loc} < R(\text{APE}) \]

\[ \exists \text{InsLen}, \forall \text{APE} \in C : \Delta_{\text{min}} - R(\text{APE}) + L(\text{APE}) < \text{InsLen} < \Delta_{\text{max}} - R(\text{APE}) + L(\text{APE}) \]
Maximal Valid Clusters for Insertions

A Maximal Valid Cluster is a valid cluster that no additional APE can be added without violating the validity of the cluster.

1. Find all the Maximal sets of overlapping paired-end alignments
2. For each maximal set $S_k$ found in Step 1, find all the maximal subsets $s_i$ in $S_k$ that the insertion size ($InsLen$) they suggest is overlapping
3. Among all the sets $s_i$ found in Step 2, remove any set which is a proper subset of another chosen set

Problem: Among all the maximal valid clusters, which ones are correct?
Aim: Assign a single PE-Alignment to all paired-end reads

Maximum Parsimony Structural Variation

- Find a minimum number of SVs such that all the paired-end reads are covered
  - Similar to SET-COVER problem
  - Greedy algorithm. Approximation factor $O(\log(n))$

F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009
Case Study: NA18507

NA18507: Yoruba male, sequenced with the Illumina Genome Analyzer

- 42X sequence coverage
- Read length: 36bp; average insert size: 208bp, standard deviation: 8.25bp

<table>
<thead>
<tr>
<th></th>
<th>VariationHunter-SC</th>
<th>VariationHunter-Pr</th>
<th>Bentley 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted</td>
<td>Overlap</td>
<td>Predicted</td>
</tr>
<tr>
<td>Deletion</td>
<td>143</td>
<td>8,959</td>
<td>85</td>
</tr>
<tr>
<td>Inversion</td>
<td>82</td>
<td>504</td>
<td>23</td>
</tr>
<tr>
<td>Insertion</td>
<td>NA</td>
<td>5,575</td>
<td>NA</td>
</tr>
</tbody>
</table>

* predicted insertions < 100bp

Hormozdiari, Alkan, et al, Genome Res. 2009
Case Study: NA18507

Hormozdiari, Alkan, et al, Genome Res. 2009
1000 Genomes: Validation

YRI trio: NA19238, NA19239, NA19240
Deletion calls with Illumina read pair analysis
Validation: arrayCGH and PCR

- Higher sensitivity than unique-location based methods
- Higher false discovery rate
  - Assumes complete reference genome
  - Paralogous variants that do not exist in the reference genome will be invalidated
VariationHunter for *Alu* insertions

- 8 high-depth genomes
  - One trio (YRI: NA18506, NA18507, NA18508)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NA18506</td>
<td>YRI</td>
<td>40.1x</td>
<td>255x</td>
<td>1,720</td>
<td>1,280</td>
</tr>
<tr>
<td>NA18507</td>
<td>YRI</td>
<td>27.1x</td>
<td>157x</td>
<td>1,579</td>
<td>1,144</td>
</tr>
<tr>
<td>NA18508</td>
<td>YRI</td>
<td>37x</td>
<td>214x</td>
<td>1,744</td>
<td>1,293</td>
</tr>
<tr>
<td>NA10851</td>
<td>CEU</td>
<td>22x</td>
<td>160x</td>
<td>1,282</td>
<td>781</td>
</tr>
<tr>
<td>AK1</td>
<td>Korean</td>
<td>22.5x</td>
<td>49x</td>
<td>909</td>
<td>582</td>
</tr>
<tr>
<td>YH</td>
<td>Han Chinese</td>
<td>11.4x</td>
<td>27x</td>
<td>1,160</td>
<td>698</td>
</tr>
<tr>
<td>KB1</td>
<td>Khoisan</td>
<td>21x</td>
<td>25x</td>
<td>457</td>
<td>313</td>
</tr>
<tr>
<td>HGDP01029</td>
<td>Khoisan</td>
<td>4x</td>
<td>12x</td>
<td>307</td>
<td>214</td>
</tr>
<tr>
<td><strong>Non Redundant Total</strong></td>
<td></td>
<td></td>
<td></td>
<td>4,342</td>
<td>3,432</td>
</tr>
</tbody>
</table>

- 63/64 (98%) sites tested with PCR are validated.
- 1,437/4,342 (33.1%) map within genes (RefSeq May 2010)

*Hormozdiari, Alkan et al., Genome Res. 2011*
MEI sequence signature

- Strand rules: MEI-mapping “+” reads and MEI mapping “-” reads should be in different orientations:
  - +/- and -/+ clusters; or +/+ and -/- clusters (inverted MEI)
- Span rules: A=(A1, A2); B=(B1, B2); C=(C1, C2); D=(D1, D2)
  - |A1-B1| ~ |A2-B2| and |C1-D1| ~ |C2-D2| (simplified; we have 8 rules)
- Location and 2-breakpoint rule:
  \[
  \exists \text{loc}, \forall \text{PE} : \text{RightMost}(+) < \text{loc} < \text{LeftMost}(-)
  \]
Alu subfamilies

Hormozdiari, Alkan et al., Genome Res. 2010
Familial transmission of *Alu*s

Hormozdiari, Alkan et al., Genome Res. 2010
Common Alu

Hormozdiari, Alkan et al., Genome Res. 2010
**Alu polymorphism**

- *in silico* genotyping with 1000GP:
  - 10.5% (21/201; chr1) are significantly stratified (Fst>0.2),
  - $\rightarrow$ ~400 markers for population genetics analyses
  - 18/21 show increased allele frequency in the YRI when compared to either the ASN or CEU populations

- PCR genotyping of 3 *Alu* insertions in 1,058 individuals from 52 populations (HGDP)

*Hormozdiari, Alkan et al., Genome Res. 2010*
Finding “novel” sequences

- DNA sequences that have no representation in the reference genome assembly
  - Excludes duplications & common repeats

- Two major NGS-based methods:
  - Whole genome de novo assembly
    - ALLPATHS-LG, SOAPdenovo, ABysS, Cortex, Velvet, Euler, etc.
    - Compute and memory intensive
  - Local de novo assembly using mapping information
    - Poor man’s method: Going through the trash that the mapper left
3) Local Assembly: NovelSeq

Reference genome → novel insertion → map with mrFAST → de novo assembly and contamination filter

Clustering → insertion breakpoint → OEA+ and OEA-

OEA+ and OEA- → local assembly

OEA contigs (length ~ L) → merging OEs & orphans

Novel sequences

http://compbio.cs.sfu.ca/strvar.htm

Hajirasouliha et al., Bioinformatics 2010
**NovelSeq: NA18507**

4,154 contigs (≥200bp)
Total 2.9 Mb sequence that is not in the reference assembly
N50 size: 955 bp

![Contig length histogram (AbySS)](image)

<table>
<thead>
<tr>
<th>Number of total contigs, and contigs map to various sequence databases with &gt;=99% identity</th>
<th>Total</th>
<th>Build36</th>
<th>Venter WGS</th>
<th>HuRef</th>
<th>NA18507 Fosmid WGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=1 kb</td>
<td>779</td>
<td>13</td>
<td>744</td>
<td>43</td>
<td>337</td>
</tr>
<tr>
<td>&gt;=200bp (&lt;1Kb)</td>
<td>3336</td>
<td>110</td>
<td>2957</td>
<td>69</td>
<td>654</td>
</tr>
</tbody>
</table>

**Hajirasouliha et al., Bioinformatics 2010**
No method is comprehensive

Summary

- Next-generation sequencing technologies
  - Promises to replace array based methods; but:
    - Entire spectrum of structural variation is not yet detected
    - Most current studies target only CNVs in relatively less complex areas of the genome
    - Different sequencing platforms present different error models
  - Need better methods to
    - Identify *inversions* and *translocations*
    - Discover SVs in repeat- and duplication-rich regions
    - Accurately characterize *copy, content, and structure* of structural variants
- Long term goal: **accurate de novo** assemblies to detect a broad range of variants
Acknowledgements

UW
Evan E. Eichler
Jeffrey M. Kidd
Tomas Marques-Bonet
Peter Sudmant
Jacob Kitzman
Gözde Aksay
Carl Baker
Francesca Antonacci
Santhosh Girirajan
Farhad Hormozdiari
Maika Malig
Emre Karakoç

SFU
S. Cenk Şahinalp
Fereydoun Hormozdiari
Iman Hajirasouliha
Faraz Hach

1000 Genomes Project
Structural Variation
Analysis Subgroup

EBI
Paul Flicek
Rasko Leinonen

Baylor College of Medicine
Richard A. Gibbs

NCBI
Martin Shumway

Carnegie Mellon University
Onur Mutlu