Structural Variation in Human and Cancer Genomes

Ben Raphael
CS29650-C Lecture 6
Oct. 6, 2009

DNA Sequencing

Human Genome Project

Sources: Rob Carlson

454
Illumina
ABI SOLiD
Genome Sequencing and Comparison

Comparative genomics
Differences between species?

Personal genomics
Genetic basis for traits of individuals?

Cancer genomics
Which somatic mutations lead to cancer?

Genome Variation

Copy number variants

Structural variants
An 900kb inversion with frequency of 20% in European population. Subject to positive selection.
Structural Variation

Change gene structure
Create novel fusion genes

Alter gene regulation

Somatic

Structural Variation

Breast

Leukemia

Somatic
Structural Variation

Measurement techniques + Computation = Biology

- Whole genome sequencing
- Microarray: array comparative hybridization
- Paired-end sequencing/mapping

Paired-End Sequencing/Mapping

1) Fragments of test genome
2) Sequence ends of fragments.
3) Map end sequences to reference genome.

Fragment from test genome

End sequence pair (ES pair) (x, y) in reference genome.

** Not 1-1 correspondence: repeats in reference genome.
1) Fragments of test genome

2) Sequence ends of fragments.

3) Map end sequences to reference genome.

**Concordant** ES pairs
- Observed length: $L = y - x$
- $L_{min} \leq L \leq L_{max}$ min/max size of clone.
- Convergent* orientation.

* varies with sequencing technology

**Discordant** ES pairs: putative structural variant in test genome
- Insertions/deletions: $L$ too short/long
- Inversions: Parallel orientations
- Translocations: Different chromosomes
Paired-End Sequencing/Mapping

1) Fragments of test genome
2) Sequence ends of fragments.
3) Map end sequences to reference genome.

<table>
<thead>
<tr>
<th>Sequencing Technology</th>
<th>Normal</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clone-based</td>
<td>40kb fosmids</td>
<td>150kb BACs</td>
</tr>
<tr>
<td>Reads: 500-1000bp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next-gen.</td>
<td>3kb 454 [Korbel et al. (2007)]</td>
<td>200bp Illumina</td>
</tr>
<tr>
<td>Fragments: 200-3000bp</td>
<td>200bp Illumina [1000Genomes]</td>
<td>[Campbell et al. (2008)]</td>
</tr>
<tr>
<td>Reads: 35-100bp</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example: Deletion

L = Fragment Length
Where’s the Deletion?

True Fragment size: $L$
Observed Size: $y_1 - x_1$

Where’s the Deletion?

`breakpoints (a,b)`

$L = (a - x_1) + (b - y_1)$
Where’s the Deletion?

breakpoints \((a, b)\) \[ L_{\text{min}} \leq (a - x_1) + (b - y_1) \leq L_{\text{max}} \]

Example: Inversion

Test Genome

Reference Genome
**Ambiguity in Structural Variants**

ES point toward *breakpoints* \((a,b)\)

**Deletion**

\[
\begin{align*}
(x_1, y_1) & \\
\bullet (a,b) & \\
(\text{Genome coordinate})
\end{align*}
\]

**Inversion**

\[
\begin{align*}
(x_2, y_2) & \\
\bullet (a,b) & \\
(\text{Genome coordinate})
\end{align*}
\]

**Geometric Analysis of Structural Variants (GASV)**

- Classify and compare structural variants *with* associated measurement uncertainty.
- Use computational geometry algorithm to efficiently compute polygon intersections.

Computational Problem

• Given a set $B_1, B_2, \ldots B_n$ of breakpoint regions, identify groups of breakpoint regions that indicate the same variant.

• All (Maximal) Intersections of Breakpoint Regions:
  Given a set $B = \{B_1, B_2, \ldots B_n\}$ identify and label all (maximal) intersections of subsets of $B$.

Solution: Plane Sweep

Line $L$ with slope parallel to trapezoids.
Move $L$ through plane and compute intersections.
Sweep-Line Algorithm

• Problem: Given n line segments, report all intersections.

• Solutions:
  – Trivial Solution: Check all pairs of line segments $O(n^2)$
  – Sweep Line Algorithm: $O((n+p) \log n)$, $p = \#$ of intersection points

Sweep-Line Algorithm

Maintain a sorted list of line-segments intersecting the sweep-line $L$. 
Sweep-Line Algorithm

Add Event: New line segment intersects the line L.

Sweep-Line Algorithm

Remove Event: Line segment ends.
Sweep-Line Algorithm

Add & Remove events are determined by the endpoints.

Sweep-Line Algorithm

**Intersect** Event: If two line segments intersect, they must have been adjacent along the sweep line.
Sweep-Line Algorithm

Addition/Removal Events are known in advance.

Intersection Events are computed during the sweep.
Sweep-Line Algorithm

For $n$ line segments with a total of $p$ points of intersection there are $(2n + p)$ events each taking $\log n$ time.

$O( (n+p) \log n)$

Applications

Sample
Individual Genomes
Cancer Genomes

Measurement
Paired-end sequencing
Array Comparative Genomic Hybridization (aCGH)

Geometric Analysis (GASV)

Multiple sample/platform integration

Human Genome Structural Variation Project
GASV Advantages

1. Visualize breakpoint regions.

2. Distinguish simple variants from complex variants.

3. Precise localization of breakpoints of variants.

4. Integration of data from different individuals, measurement platforms.
Distinguish Simple from Complex Structural Variants

Simple Variant: Cluster
There exists a single breakpoint \((a,b)\) consistent with all fragments.

Complex Variant: Non-Cluster
No single breakpoint consistent with all fragments.

Structural Variation

Kidd et al. (2008)

8 HapMap individuals + 1 additional

Paired-end sequencing (fosmid clones 40kb)

747 deletions
724 insertions
224 inversions
### Clustering of Single Samples

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Cluster</th>
<th>Non-Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inversions</td>
<td>926</td>
<td>860</td>
<td>66</td>
</tr>
<tr>
<td>Deletions</td>
<td>1,810</td>
<td>1,746</td>
<td>64</td>
</tr>
</tbody>
</table>

**Inversion**
- 6 ES pairs

**Deletion**
- 11 ES pairs
Geometric Localization of Breakpoints

Deletion: 5 ES pairs
Localization: 9,361 bp

Incorrect prediction of structural variant

106/131 sequenced fosmids support our localization.

Next-generation sequencing

Two individuals also sequenced with Illumina paired ends (36bp reads, ≈200bp fragments).

- NA12878 (CEU): 1000 Genomes Project
Next-gen vs. Fosmids

\[ \approx 59\text{-}68\% \text{ of validated deletions (Kidd et al. 2008) detected by Illumina fragments} \]

NA18507 (YRI)
- 3431 total GASV predictions
- 54 deletions
- 38 deletions

NA12878 (CEU)
- 2215 total GASV predictions
- 106 deletions
- 51 deletions

Comparison to other Programs

**GASV** (Sindi, et al. 2009)
- 3,486 predictions
- 54 detected

**VariationHunterWeighted** (Hormozdiari, et al. 2009)
- 8,959 predictions
- 60 detected

**BreakDancer** (Chen, et al. 2009)
- 27,092 predictions
- 55 detected

**GASV**: Comparable sensitivity, higher specificity.
Other Variants

**Insertions:**
One breakpoint on reference, straightforward analysis.

**Translocations/interchromosomal variations:**
Not thought to be common genetic polymorphisms.
Not reported in structural variation studies.

NA12878:
18,395 predicted translocations!
(2321 predicted deletions.)
False positives!

Somatic Structural Variants in Cancer Genomes

**Leukemia**

**Breast**

Fusion gene in >50% prostate cancer patients

(Tomlins et al. *Science* 2005)
Somatic Structural Variant Prediction

Hundreds/thousands of structural variant candidates in cancer sample. Filter variants found in matched normal using GASV

Somatic structural variant candidates.

Cancer Genome Rearrangements

• Glioblastoma multiforma and matched normal sample
• Illumina paired-end sequencing*
  – 200bp fragments, 30X coverage of tumor and normal

<table>
<thead>
<tr>
<th>Type</th>
<th>Cancer sample ≥10 paired reads</th>
<th>Not in normal</th>
<th>Not in normal ≥20 paired reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletions (&gt;5kb)</td>
<td>251</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>Inversions (&gt;10kb)</td>
<td>128</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Translocations</td>
<td>693</td>
<td>116</td>
<td>28</td>
</tr>
</tbody>
</table>

*Genome Center at Washington University R. Wilson, E. Mardis, L. Ding, et al.
Cancer Genome Rearrangements

Chr 16 Deletion (11kb)  
Localization: 148 bp  
Fragments: 25

Chr 7 inversion (370kb)  
Localization: 128bp  
Fragments: 20

Chr 1:7 Translocation  
Localization: 79 bp  
Fragments: 76

Translocation also found by BreakDancer  
K. Chen, et al.
DNA Basepairing

DNA Microarrays

RNA fragment hybridizes with DNA on GeneChip
Measuring Mutations in Cancer

Comparative Genomic Hybridization (CGH)

Chromosome CGH provides "cytogenetic" resolution ~ 10 Mb
Resolution of array CGH depends on spacing and length of clones

CGH Analysis (1)

- Divide genome into segments of equal copy number
CGH Analysis (1)

- Divide genome into segments of equal copy number

Input: $y_i = \log_2 T_i / R_i$, clone $i = 1, ..., N$
Output: Assignment $s(y_i) \in \{S_1, ..., S_K\}$

$S_i$ represent copy number states

An Approach to CGH Segmentation

- Circular Binary Segmentation (CBS), Olshen et al. 2004
- Use hypothesis test to compare means of two intervals using t-test
Microarray techniques

- Copy number variants identified by array comparative Genomic Hybridization (aCGH) give rise to rectangles.

TCGA Glioblastoma results

9/81 CGH breakpoints supported by ES pair (one end)

1 pair of CGH breakpoints supported by ES pair.
(Chr1:7 translocation).

Cancer ESP + aCGH

BAC (≈ 150kb fragments) paired-end sequencing
(Volik et al. 2003; Volik et al. 2006; Raphael et. al 2008)

Agilent 244K aCGH data

8Mb deletion on Chr17 in BT474
Cancer Paired-End Sequence

BAC paired-end sequencing (≈ 150kb fragments)
(Volik et al. 2003; Volik et al. 2006; Raphael et al. 2008)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Clusters</th>
<th>% Concordant w/ Kidd et. al inversions</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF7</td>
<td>379</td>
<td>5.5</td>
</tr>
<tr>
<td>BT474</td>
<td>151</td>
<td>9.9</td>
</tr>
<tr>
<td>SKBR3</td>
<td>148</td>
<td>9.4</td>
</tr>
<tr>
<td>Breast</td>
<td>162</td>
<td>14.8</td>
</tr>
<tr>
<td>Breast</td>
<td>96</td>
<td>20.8</td>
</tr>
<tr>
<td>Prostate</td>
<td>94</td>
<td>13.8</td>
</tr>
<tr>
<td>Ovary</td>
<td>84</td>
<td>15.4</td>
</tr>
<tr>
<td>Brain</td>
<td>67</td>
<td>7.5</td>
</tr>
</tbody>
</table>

5-20% of invalid clones in cancer ESP consistent with inversion polymorphisms.

Fusion Genes

Bashir, et al. (2008) PLOS Comp Biol

Used to predict fusion genes in SoLiD sequencing paper.
Summary

Evolutionary

Genetic

Somatic

1000 Genomes
A Deep Catalog of Human Genetic Variation

The Cancer Genome Atlas (TCGA)
Pilot Project
Charting a new course for prevention, diagnosis, and treatment of cancer