The Breakage Fusion Bridge and other complex SVs: Combinatorics & Cancer Genomics

Vineet Bafna

Computer Science & Engineering

UCSD
Variation

• Genomes vary within a population
• The mutational landscape includes small nucleotide variation, but also large structural variation.
  – This is particularly true for a population of tumor cell genomes.
• This talk is on the mechanisms for these variations.
(A) Aberrant peptides in ANP32B, with frame-shifts, and deletion events

(B) Aberrant peptides in ANP32B, with frame-shifts, and deletion events
Detecting SVs with next generation sequencing

Evidence
1. Discordant Paired-end
2. Depth of coverage
3. Loss of heterozygosity
4. Split reads
Mechanisms for SVs.

A) Retrotransposon Insertion

B) Non-classical insertion

C) NAHR mediated insertion/deletion

D) NHEJ mediated deletion

Xing J et al. Genome Res. 2009;19:1516-1526
Can these mechanisms explain complex rearrangements?

Zhao, Cancer Res. 2004

Kitada, 2008
Somatic rearrangements in BACs. (A) Chromosome 17q21 amplicon in HCC1954 including ERBB2; (B) chimeric amplicon in HCC1954 including MYC; (C) chimeric amplicon in NCI-H2171 including MYC; (D) chromosome 2 amplicon in NCI-H1770 including MYCN. The color of t...

...The architecture of rearrangements in this amplicon recapitulates remarkably well the structural features predicted by a classical breakage–fusion–bridge cycle involving sister chromatids, as originally proposed by Barbara McClintock (1941) (see Supplemental Fig. 4)....

....Therefore, the ERBB2 amplicon in HCC1954 bears all the architectural hallmarks at the sequence level of a classical sister chromatid breakage–fusion–bridge process, the first time that this has been demonstrated in human cancer....

Bignell G R et al. Genome Res. 2007;17:1296-1303
This talk

• We formally define the BFB detection problem, provide a poly (linear time) solution, and identify BFB events on existing genomic data.

• We re-examine the evidence for Chromothriipsis, and suggest that it may be more rare(*) than suspected.
A second model of complex genomic rearrangement
Chromothripsis

• The null hypothesis:
  – “Cancer is driven by somatically acquired point mutations and chromosomal rearrangements, conventionally thought to accumulate gradually over time.” – Stephens et al.

• In Chromothripsis:
  – “tens to hundreds of genomic rearrangements occur in a one-off cellular crisis. Rearrangements involving one or a few chromosomes crisscross back and forth across involved regions, generating frequent oscillations between two copy number states.”
This talk

• We formally define the BFB detection problem, provide a poly (linear time) solution, and identify BFB events on existing genomic data.

• Extraordinary hypotheses require extraordinary proofs: We re-examine the evidence for Chromothripsis, and suggest that it may be less frequent(*) than suspected.
SO, WHAT IS BFB?
Telomere Loss

THE BEHAVIOR IN SUCCESSIVE NUCLEAR DIVISIONS OF A CHROMOSOME BROKEN AT MEIOSIS

BY BARBARA MCCINTOCK
DEPARTMENT OF BOTANY, UNIVERSITY OF MISSOURI
Communicated July 7, 1939
THE BEHAVIOR IN SUCCESSIVE NUCLEAR DIVISIONS OF A CHROMOSOME BROKEN AT MEIOSIS

BY BARBARA McCINTOCK

DEPARTMENT OF BOTANY, UNIVERSITY OF MISSOURI

Communicated July 7, 1939

PNAS 1939
THE BEHAVIOR IN SUCCESSIVE NUCLEAR DIVISIONS OF A CHROMOSOME BROKEN AT MEIOSIS

BY BARBARA McCINTOCK

DEPARTMENT OF BOTANY, UNIVERSITY OF MISSOURI

Communicated July 7, 1939

PNAS 1939
Telomere fusions in breast carcinoma

Tanaka, PNAS 2012
Detecting BFB

• Laborious experiments can possibly detect BFB in action
  – Telomere fusions, di-centric chromosomes, repeated sub-structures,…

• However, molecular data is the most likely source
  – Deep genomic sequencing with (short) PE reads
  – Genomic hybridization
The BFB-count-vector problem

While we cannot sequence and assemble the genome, we can measure copy counts of individual segments.
The BFB reconstruction problem

Given a segmentation of the genome, and counts of copy numbers for each segment, are the copy numbers consistent with a BFB sequence?
BFB Trees

Kinsella, Jnl Comp. Bio., 2012
Properties of BFB trees

- Properties
  - Long ends
  - Node symmetry
  - Pair Symmetry

- A traversal on a tree with these 3 properties is a BFB sequence.

Kinsella, Jnl Comp. Bio., 2012
Finding Longest Palindromes

\[ 2 \cdot n_1 + 2 \cdot n_2 + 1 \cdot n_3 = 6 \]

\[ 2 \cdot n_1 + 4 \cdot n_2 = x \]
BFB-Tree Has Better Worst-Case Performance

BFB-Tree for \([20, 3, 19, 19, 19]\) took 12 seconds.

BFB-Pivot for \([20, 18, 20, 20, 6]\) took 22,251 seconds.

Kinsella, Jnl Comp. Bio., 2012
Most Count Vectors Do Not Admit BFB

- Expanded analysis to 64,000,000 count vectors of length $\leq 6$ and $n_i \leq 20$. Of those count vectors, only 7.2% can be achieved by BFB.

- Finding a count vector achievable by BFB is strong evidence BFB occurred?
Experimental Count Vectors Are Imprecise

Table 1. Copy numbers of the segments in chromosome 7amp

<table>
<thead>
<tr>
<th>Segment</th>
<th>Size (Mb)</th>
<th>Fold changes&lt;sup&gt;b&lt;/sup&gt; (detected by array CGH)</th>
<th>Copy numbers&lt;sup&gt;c&lt;/sup&gt; per cell (estimated from array CGH data)</th>
<th>Copy numbers per cell (observed in I-FISH)</th>
<th>Copy numbers in Chr 7amp&lt;sup&gt;f&lt;/sup&gt; (observed in M-FISH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>p whole arm plus q-proximal region 1.0 ± 0.1</td>
<td>3.6 ± 0.6</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>2.2</td>
<td>1.7 ± 0.2</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>c</td>
<td>16.6</td>
<td>2.1 ± 1.3</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>d</td>
<td>2.3</td>
<td>2.7 ± 0.9</td>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>e</td>
<td>2.6</td>
<td>11.0 ± 3.1</td>
<td>33</td>
<td>Many</td>
<td>Many</td>
</tr>
<tr>
<td>f</td>
<td>2.7</td>
<td>0.7 ± 0.1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Lung cancer cell-line PTX250

Observed: [3 3 5 6 8 33]

BFB: [3 3 4 6 34]
Many CVs close to a BFB sequence

\[ d(x, y) = \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|} \]

<table>
<thead>
<tr>
<th>Distance (%ile)</th>
<th>Count Vector</th>
<th>Nearest Admitting BFB</th>
</tr>
</thead>
<tbody>
<tr>
<td>.097 (50)</td>
<td>(14,7,18,16,9,12)</td>
<td>(14,7,19,17,9,13)</td>
</tr>
<tr>
<td>.129 (60)</td>
<td>(7,13,6,4,19,12)</td>
<td>(8,14,6,4,20,12)</td>
</tr>
<tr>
<td>.192 (70)</td>
<td>(9,7,7,8,2,14)</td>
<td>(8,8,8,8,2,14)</td>
</tr>
<tr>
<td>.362 (80)</td>
<td>(16,17,14,18,1,19)</td>
<td>(16,18,14,18,2,19)</td>
</tr>
<tr>
<td>.458 (90)</td>
<td>(20,1,7,3,10,6)</td>
<td>(20,2,7,3,11,7)</td>
</tr>
<tr>
<td>.566 (95)</td>
<td>(15,8,8,1,15,2)</td>
<td>(16,8,8,2,15,3)</td>
</tr>
<tr>
<td>.889 (99)</td>
<td>(9,1,5,9,1,15)</td>
<td>(10,2,5,9,3,15)</td>
</tr>
</tbody>
</table>
A non-BFB count vector

Zhao, Cancer Res. 2004 (lung cancer cell-line NCI-H2171)

\[\begin{align*}
[1, 2, 1, 2, 8, 1, 3, 1, 3] & \rightarrow [2, 2, 2, 2, 8, 3, 3, 3, 3, 3]
\end{align*}\]
A linear time algorithm for BFB

• Once you allow errors in counts, many count vectors can conceivably be explained by BFB, although there are a few that are conclusively not BFB

• What if we choose large numbers of segments (k=8,10,12)? The BFB-tree heuristic is much too slow to deal with large data sets.
Count vector canonization

\[ 2\vec{n}' = [2, 14, 18, 12, 8] \]

Zakov, 2013 PNAS
Layer 4

\[ 2\tilde{n}_4 = [0, 0, 0, 0, 8] \]

\[ \alpha_4 \]

- \( 4 \times DD \)

At the top-most level, the 8 Ds rearrange uniquely into 4 4-BFB blocks

Zakov, 2013 PNAS
Layer 3

$2n \downarrow 3 = [0, 0, 0, 12, 8]$

- There are 6 CC at level 3.
- Soln:
  - Add 2 $\varepsilon$ to the 4 DDs
  - **Wrap:** to get 4XCDDDC, 2XCC

Zakov, 2013 PNAS
Layer 2

We have 6 blocks (4XCDDDC, 2XCC) at level 2, but 9 at level 2.

**Soln:** Add 3 ε

**Wrap** to get 4XBCDDDDB, 2XBCCCB, 3XBB

Zakov, 2013 PNAS
Layer 1

- The 9 blocks at level 2 must be folded into 7 blocks.

Zakov, 2013 PNAS
Layer 1

9 blocks at level 2, 7 at Block 1

Fold $2X\{BCDDDCB-BCCB-BCDDDCB\}$ to get 5 blocks.

Add 2 $\varepsilon$ to get 7

Wrap to get $2X\{A-BCDDDCB-BCCB-BCDDDCB-A\}$, $3XABBA$, $2XAA$. 

$2n l1 = [0, 14, 18, 12, 8]$
Layer 0

\[ 2\tilde{n}_0 = [2, 14, 18, 12, 8] \]

- Fold all of the 7 blocks at level 1 to get a single block.
- Wrap the $ around the block.
Wrapping and folding

• The algorithm does a sequence of wrapping and folding.
• The folding must be done in a special way so that a greedy algorithm always folds correctly.
Algorithmic Result

- Linear time algorithm for construction $O(kn)$, and detection $O(k \log n)$ of BFB
- Additionally, a fast technique $O(kn^2)$ for identifying the nearest BFB sequence, and computing the distance to it.
Using BFB in practice

• For each window, we output the distance between the example, and the nearest BFB
• We apply a sliding window. At each position (and window-length), we identify a count-vector with minimum distance to a BFB.
• If the distance meets a threshold, we accept, else reject
Detecting BFB versus other rearrangements

• The BFB operations take place among other rearrangements that change the copy number
• The count of copies is inexact at best
• Negative examples: We simulate 15000 diploid chromosomes undergoing over 50 rearrangements (inversions, tandem duplications, deletions)
• Positive examples: 5000 cases with 50 rearrangements on diploid chromosomes, and one with 2-10 BFB operations
• In each case, a count-vector is created
• $\delta =$ min. distance of the count vector from a true BFB
• Long count vectors certainly help, but they might not be enough either.
• If 10% of the cases are BFB, a 10% FPR rate becomes 50%FP in naturally occurring samples.
• Working with 1% FPR reduces TP to 18%.
• Excess of fold-backs (e.g. –A, A) might provide other clues
Fold-backs

• Long count vectors certainly help, but they might not be enough either.
• If 10% of the cases are BFB, a 10% FPR rate becomes 50%FP in naturally occurring samples.
• Working with 1% FPR reduces TP to 18%.
• Excess of fold-backs (e.g. –A, A) might provide other clues
Detecting fold-backs

Reference

Donor
A simple score function

- $f =$ fraction of breakpoints that are fold-backs (higher the better)
- $\delta =$ min. distance of the count vector from a true BFB
- $s = (1-\lambda)\delta + \lambda(1-f)$ is a generic score function (lower the better)
Campbell, Nature 2010. (Exemplar of BFB in patient with pancreatic cancer)

- Note the excess of Head-head fold-backs relative to others rearrangements
- We identify a count-vector of length 10 with a very low distance threshold that allows us to detect this as BFB with some confidence.
Chr 12: Novel BFB observed in primary tumor (pancreatic) sample

Zakov, 2013 PNAS
Detecting BFB: conclusions

• With errors in segment copy counts, it is easy to create short, false BFB-like count vectors.
• We cannot detect short BFB-cycles.
• With 8 or more BFB cycles, and correspondingly long count-vectors, we need advanced algorithms to identify BFBs.
• Long count vectors allow us to detect BFB operations with high confidence even amidst noise of other rearrangements.
• Other signals, such as fold-back operations help in further improving the signal.
Chromothripsis
Chromothripsis

• The null hypothesis:
  – “Cancer is driven by somatically acquired point mutations and chromosomal rearrangements, conventionally thought to accumulate gradually over time.” –Stephens et al.

• In Chromothripsis:
  – “tens to hundreds of genomic rearrangements occur in a one-off cellular crisis. Rearrangements involving one or a few chromosomes crisscross back and forth across involved regions, generating frequent oscillations between two copy number states.”
Observations

- Massive remodeling of a single chromosome
- Tens to hundreds of genomic rearrangements
- Very few copy number states (two or perhaps three).
- Retention of heterozygosity in regions with higher copy number.
- Breakpoints show significantly more clustering than observed by chance.

*Stephens, Cell 2011*
Re-examining the simulation: the input

Stephens, Cell 2011
Stephens’ simulation

- Do progressive simulations to match the breakpoints.
- As the number of breakpoints (defined as number of copy number changes) grows, the number of distinct copy number states also grows.
- This is in contrast to the proposed model
THREE STRIKES AGAINST CHROMOTHIRIPSIS

1. The proposed signal for Chromothriipsis is not as strong as it appears on first glance
Simulation results
Over-representing Inversions

Paired end sequencing

Microarray

Number of Breakpoints

Number of Copy Number States

99% Interval
Median

TK-10 8505C

TK-10 8505C
THREE STRIKES AGAINST CHROMOTHIRPISIS

2. A genome simulated with a progressive model of rearrangement also shows the Chromothriipsis signature.
A special progressively rearranged chromosome passes the Stephen’s test
THREE STRIKES AGAINST CHROMOTHRIPSIS

3. The Stephens et al. Chromosome SNU-C1 (colon cancer, 200+ breakpoints, 2 copy number states can be explained by progressive rearrangements
Progressive rearrangements (HP theory) can explain Stephens’ et al. chromosome

<table>
<thead>
<tr>
<th>Cell-line</th>
<th>Chr</th>
<th>Covered</th>
<th>Not</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK10</td>
<td>5</td>
<td>52</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>8505C</td>
<td>9</td>
<td>74</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>SNU-C1</td>
<td>15</td>
<td>228</td>
<td>11</td>
<td>75</td>
</tr>
</tbody>
</table>
Concluding thoughts: complex structural variations, or not?

- There are many lines of enquiry for complex structural variation
  - Given fragmentary data, can you use it to reconstruct the architecture of the sampled genome (AKA sequence assembly)
  - Given a set of operations, predict a plausible sequence of events (or the number) that takes a sampled genome to a reference target (AKA genome rearrangements)
  - Given a sampled genome, what kind of rearrangements are the most likely explanation? Is there evidence for more complex rearrangements (e.g., this talk)?
Conclusions

• Understanding structural variants from a mechanistic perspective is fundamental, but relatively understudied.

• We took two examples. For BFB, we claim that computation reveals a testable signature. For Chromothripsis, perhaps not
Acknowledgments

• Shay Zakov (BFB),
• Marcus Kinsella (BFB & Chromothriipsis)
• NIH/NSF