SCALPEL
MICRO-ASSEMBLY APPROACH TO DETECT INDELS WITHIN EXOME-CAPTURE DATA

Giuseppe Narzisi, PhD
Schatz Lab

Cold Spring Harbor Laboratory
Outline

1. Scalpel micro-assembly pipeline
2. Large-scale validation experiment
3. De novo/Transmitted mutations in Autism
SCALPEL

Micro-assembly pipeline
The detection challenge

**Irregularity** in capture efficiency near the edges of the coding region

**SNP or Indel?**

```
..TTGAATTAGCCTTGGAATTGAGCCTT...TTTAG-------AGTGC...

GAATGAGCC

GAAT-AGGCC

TTTAGAATAGGC

ATAGGCCAGTGC
```
Scalpel

- Novel DNA sequence **micro-assembly** pipeline to detect mutations within exome-capture data.

<table>
<thead>
<tr>
<th>Whole-Genome assembly</th>
<th>Micro-assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large scale genome structure</td>
<td>Detect genome variations</td>
</tr>
<tr>
<td>Genotypic</td>
<td>Haplotypic (Hom/Het state)</td>
</tr>
<tr>
<td>Heuristics to optimize resources (Time and Space)</td>
<td>Feasible to perform exhaustive search</td>
</tr>
</tbody>
</table>

- **Features:**
  1. **Self-tuning** k-mer.
  2. **On-the-fly** repeat composition analysis.
  3. Family pedigree: **joint analysis** of family members to detect de novo and **transmitted** mutations.
Micro-Assembly Approach to detect INDELs

1. Extract reads
2. Build de Bruijn graph
3. Remove low coverage nodes, dead-ends and compress
4. Mark Source and Sink
5. If cycle or near-perfect repeat in any path
   - yes
   - no
   - Traverse graph and enumerate haplotype paths
   - Align to reference

K = K+1

Reference

Source
Sink
Deletion
Insertion
Walking along the exome

- Extraction, assembly, alignment and INDEL detection performed in **overlapping windows** along the exon.

1. Localized assembly (smaller graph).
2. Minimize problem with coverage drops.
3. Distributed approach.

![Graph showing coverage across genome locations for different relatives](image_url)
LARGE SCALE EXPERIMENT
Re-sequencing of 1000 INDELs
INDELs in one Exome

Individual affected by Attention Deficit/Hyperactivity Disorder (ADHD)

Captured using Agilent SureSelect v.2 and sequenced on the Illumina platform.

80% of the target at >20x coverage.

Micro-Assembly Approach to detect INDELs

Hard to judge the quality of INDELs specific to each pipeline. Superior sensitivity or poor specificity??

1000 INDELs for validation:

- 200 Scalpel
- 200 Haplotype Caller
- 200 SOAPindel
- 200 within intersection
- 200 long INDELs (>30bp)
Focus on size distribution

- **Bias** towards deletions (for HaplotypeCaller) or insertion (for SOAPindel).
- Scalpel instead shows a **well-balanced** distribution between insertions and deletions.
Validated INDELs
specific to each pipeline

- INDELs not passing validation correlate well with size bias.
• INDELs not passing validation correlate well with size bias.
DE NOVO MUTATIONS IN AUTISM
Simons Simplex Collection
Simons Simplex Collection

- ~2700 families.
- Quad: two parents, one affected child and one unaffected child.
- NimbleGen SeqCap EZ Exome v2.0 (36 Mb).
- Illumina HiSeq: ~93bp reads after removing barcodes.

Three major studies reporting strong enrichment for de novo gene killing mutations in autistic kids:
INDELs in 593 families

Database with > 3 million INDELs

Increased power to detect insertions.

Subdivide by annotation category.

**Goal:** discover significant biology that was impossible to measure a few year ago
De novo INDELs in Autism
593 families: 343 CSHL, 200 StateLab, and 50 EichlerLab

<table>
<thead>
<tr>
<th>INDEL effect</th>
<th>Aut</th>
<th>Sib</th>
<th>Aut M</th>
<th>Aut F</th>
<th>Sib M</th>
<th>Sib F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame shift</td>
<td>35</td>
<td>16</td>
<td>25</td>
<td>10</td>
<td>12</td>
<td>4</td>
<td>51</td>
</tr>
<tr>
<td>Intron</td>
<td>13</td>
<td>16</td>
<td>11</td>
<td>2</td>
<td>6</td>
<td>10</td>
<td>29</td>
</tr>
<tr>
<td>Intergenic</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>No frame shift</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Splice-site</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>UTR</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>39</td>
<td>46</td>
<td>12</td>
<td>19</td>
<td>20</td>
<td>97</td>
</tr>
</tbody>
</table>

*De novo INDELs that are likely to severely disrupt the encoded protein are significantly more abundant in affected children than in unaffected siblings*
CONCLUSION
Conclusions

- **Scalpel**: highly accurate tool to detect de novo, transmitted, and somatic INDELs.

- Errors of current detection software explained by a large-scale (1000 INDELs) re-sequencing experiment.

- Population wide analysis: de novo INDELs in Autism.
Acknowledgment

Michael C. Schatz

ADHD project
  • Jason O’Rawe
  • Yiyang Wu

Michael Wigler

Autism project
  • Dan Levy
  • Michael Ronemus
  • Yoonha Lee
  • Zihua Wang
  • Ewa Grabowska
  • Peter Andrews
  • Mitchell Bekritsky
  • Jude Kendall

Gholson J. Lyon

Ivan Iossifov
THANK YOU

Email: gnarzisi@cshl.edu