Analysis of mosaicism for sequence and copy number variants in a diverse set of hereditary disorders in a large clinical cohort

DANIEL E. PINEDA-ALVAREZ, M.D, FACMG
MEDICAL DIRECTOR, ONCOLOGY
Declaration of conflict of interest

<table>
<thead>
<tr>
<th>Type</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment full time</td>
<td>Invitae Corporation</td>
</tr>
<tr>
<td>Research Grant (P.I., collaborator or consultant; pending and received grants)</td>
<td>None</td>
</tr>
<tr>
<td>Other research support</td>
<td>None</td>
</tr>
<tr>
<td>Speakers Bureau / Honoraria</td>
<td>None</td>
</tr>
<tr>
<td>Ownership interest (stock, stock-options, patent or intellectual property)</td>
<td>Invitae Corporation</td>
</tr>
<tr>
<td>Consultant / advisory board</td>
<td>None</td>
</tr>
</tbody>
</table>
Background

- Rare variation arising post-zygotically within Mendelian disease genes can lead to mosaicism and contribute to the pathogenesis of hereditary disorders.
- Because we routinely use high depth of coverage sequencing, we had the opportunity to study a clinical cohort of nearly half a million people to understand the prevalence of mosaicism in hereditary disease.
- Investigation of mosaicism detection by NGS has to consider a variety of parameters and can be done through “genome-mixing” experiments.
Validation study

- Establishing the sensitivity and specificity of mosaicism detection by NGS requires attention to both the chemistry and the bioinformatics pipeline
- We used “genome-mixing” to investigate mosaicism detection by NGS

<table>
<thead>
<tr>
<th>Percent of mixture comprised of Genome A</th>
<th>80</th>
<th>70</th>
<th>60</th>
<th>50</th>
<th>40</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>5</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected allele balance for het’s in Genome A</td>
<td>40</td>
<td>35</td>
<td>30</td>
<td>25</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>2.5</td>
<td>1</td>
</tr>
</tbody>
</table>

- In-silico downsampling to simulate lower depth of coverage
- For heterozygous variants expected to be at 0.5 allele balance in an un-mixed sample, determine the following:
  - Observed allele balance in mixed samples
  - Observed depth of coverage
  - Absence of the variant
Low allele balance sensitivity

Sensitivity = \( \frac{\text{variants observed in titration}}{\text{variants in unmixed genome}} \)

- Computed for each
  - Titration level
  - Observed depth of coverage
  - Variant type / length
  - Genomic contexts

- High sensitivity
  - > 10% allele balance
  - > 200x
Distinguishing mosaic and non-mosaic variants

Observed allele balance distribution of heterozygous variants in unmixed genomes
Distinguishing mosaic and non-mosaic variants

Observed allele balance distribution of variants in 20% allele balance titration

Threshold

Observed allele balance distribution of heterozygous variants in unmixed genomes
Choosing a threshold to minimize FP and FN mosaics

Observed allele balance distribution of heterozygous variants in unmixed genomes

Threshold

Observed allele balance distribution of variants in 40% allele balance titration
Choosing a threshold to minimize FP and FN mosaics

Observed allele balance distribution of variants in 40% allele balance titration
Choosing a threshold to minimize FP and FN mosaics

We detect mosaic variants and distinguish them from non-mosaic variants with high positive predictive value and sensitivity.
Validation conclusions

- Sensitivity can be reduced at both high and low allele balance
- Sensitivity is reduced at lower depth of coverage
- The appropriate threshold for calling mosaic and non-mosaic variants needs to balance FP and FN mosaic calls
- Our bioinformatics approach provides high sensitivity for mosaic variants present at ~10% - 30% allele balance
- For CNVs, the sensitivity is limited to ~ 20% - 30% allele balance
Clinical data indicating mosaicism at Invitae

- Investigated genetic testing results in 472,991 individuals
  - 1606 different genes represented
  - Equivalent to 20 million single-gene tests
- 2459 mosaic variants found
  - 2107 SNVs
  - 282 small indels
  - 70 CNVs (del/dup)
- Observed allele balance 7% - 40%
  - Up to 90% allele balance in X-linked genes in males
Estimating the prevalence of mosaicism in hereditary disorders

- 286 genes
- ~1% of positive test results
- 70% in genes associated with autosomal dominant disorders
  - 93% if AD/AR included
- 40% classified as LP/P
  - 41 LP/P in X-linked genes
  - 19 in females
  - 22 in males
Prevalence of mosaicism by gene and clinical area

- Genes with >10 LP/P mosaic variants or prevalence of >10% are shown
- Gene counts are for \((\text{genes with LP/P mosaic variants}) / (\text{all genes})\)
Additional evidence of mosaicism

- Mosaic variants were typically not present in a second tissue when it was available to test
Additional evidence of mosaicism

- Mosaic variants were typically not present in a second tissue when it was available to test.
- Among cases in which both parents were tested, 35/36 mosaic variants were found to be de novo.
- Among individuals with a mosaic variant, 4% of their children carried the variant.
Mosaicism enriched in older individuals with cancer

Individuals with mosaic variants in hereditary cancer genes were considerably older than those with non-mosaic variants (p < 0.001 Student’s t-test)
Mosaicism and symptom severity

Cohort:
Individuals suspected to have diseases that have **highly specific diagnostic criteria** (e.g., NF1, TSC1, NIPBL)

Findings:
Individuals with a **non-mosaic variant** were more likely to meet **diagnostic criteria** than were Individuals with a **mosaic variant** (Fisher’s exact test p<0.001)
Conclusions

● We can detect mosaic variants and distinguish them from non-mosaic variants with **high sensitivity** and **high positive predictive value**
  ○ The appropriate allele balance thresholds to maximize positive and negative predictive values should be adjusted for prior probability of mosaicism in the gene in question

● Mosaic variants **contribute to ~1% of positive test results** in a cohort of nearly half a million people with a diverse set of hereditary disorders

● Occurrence of mosaicism can correlate with clinical observations
  ○ Individuals with mosaic variants in hereditary cancer genes were considerably older than those with constitutional variants
  ○ Mosaic variants were found to be transmitted to offspring 4% of the time
ACKNOWLEDGEMENTS

Rebecca Truty, Tina Hambuch, Curtis Kautzer, Michael Kennemer, Jennifer Rhees, Amanda Stafford, Robert L. Nussbaum, Swaroop Aradhya
Backup
Validation study

- Establishing the sensitivity and specificity of mosaicism detection by NGS pipeline requires attention to both the chemistry and the bioinformatics pipeline

<table>
<thead>
<tr>
<th>Genome A</th>
<th>Genome B</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA24385</td>
<td>NA12878</td>
</tr>
<tr>
<td>NA24385</td>
<td>NA24631</td>
</tr>
<tr>
<td>NA12878</td>
<td>NA24631</td>
</tr>
</tbody>
</table>

- We used a “genome-mixing” approach to investigate mosaicism detection by NGS

<table>
<thead>
<tr>
<th>Percent of mixture comprised of Genome A</th>
<th>80</th>
<th>70</th>
<th>60</th>
<th>50</th>
<th>40</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>5</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected allele balance for het’s in Genome A</td>
<td>40</td>
<td>35</td>
<td>30</td>
<td>25</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>2.5</td>
<td>1</td>
</tr>
</tbody>
</table>
Small variant validation
Distinguishing mosaic and non-mosaic variants

When there is a high prior for mosaics:
- $3\sigma$ threshold reduces mosaic FN
- High NPV

When there is a low prior for mosaics:
- $6\sigma$ threshold reduces mosaic FP
- High PPV
Small variant validation
Distinguishing mosaic and non-mosaic variants

Genes with high prevalence of mosaicism (e.g. GATA1, TP53, WDR45, PITX3, ACTB)
- Use 3σ threshold
- High PPV for high prior
- High NPV

Genes with low prevalence of mosaicism
- Use 6σ threshold
- High PPV for all priors
- Moderate sensitivity
CNV validation

Titration experiment using clinical samples with CNVs

Partial gene del ALG1

~25% allele balance - confident copy number = 2

~30% allele balance - low quality calls

~35% allele balance - confident copy number = 1
Limitations and Open Questions

● Limitations to our Analysis
  ○ Technical artifacts due to difficult to sequence / map regions of the genome
  ○ Statistical fluctuations in read depth at a given allele

● Additional biological evidence is needed to elucidate the mechanism for low allele balance variants:
  ○ Mosaicism arising early in embryogenesis
  ○ Hematologic malignancy or clonal hematopoiesis
  ○ Complex structural variation
  ○ Maternal engraftment
  ○ Mosaicism due to reversion of a germline mutation
## CNV validation

<table>
<thead>
<tr>
<th></th>
<th>Confidently called as copy number = 2</th>
<th>Not confidently called requires manual inspection may be called as mosaic</th>
<th>Confidently called as integer ploidy CNV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole gene dup NIPA1</td>
<td>9.3% 10.2% 10.3% 13.1% 17.2% 22.0% 28.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial gene dup AARS</td>
<td>14.5% 19.8% 20.4% 25.1% 30.2% 34.5% 39.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole gene del NPHP1</td>
<td>13.2% 18.7% 19.1% 24.1% 28.9% 33.8% 38.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial gene del ALG1</td>
<td>13.2% 18.7% 19.1% 24.1% 28.9% 33.8% 38.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single exon del CTNNA3</td>
<td>14.5% 19.8% 20.4% 25.1% 30.2% 34.5% 39.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Observed allele balance (from SNVs)
CNV validation: Clinical example

Mosaic deletion
RYR1 Exons 48-49

Array confi
Higher allele balance correlates with early onset

- Blue: mosaic variants in early onset genes in children (<18 years old)
- Green: mosaic variants in late onset genes (no age restriction)