Association of Denovo Copy Number Variants with Oral Clefts

Rob Scharpf

June 24, 2012
Cleft palate, cleft lip and palate

- Incomplete cleft palate
- Unilateral complete lip and palate
- Bilateral complete lip and palate
To identify structural variants contributing to oral cleft:

- denovo CNV discovery (by-sample)
- compare denovo frequency among oral cleft offspring to similar controls
Sample size

<table>
<thead>
<tr>
<th></th>
<th>Cleft</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td>2082</td>
<td>752</td>
</tr>
<tr>
<td>WGA</td>
<td>120</td>
<td>99</td>
</tr>
<tr>
<td>MAD &gt; 0.3</td>
<td>212</td>
<td>138</td>
</tr>
<tr>
<td>non-EA</td>
<td>1090</td>
<td>-</td>
</tr>
<tr>
<td>n</td>
<td>660</td>
<td>515</td>
</tr>
</tbody>
</table>

- cell frequencies are the number of Father-Mother-Offspring trios
- The controls are trios with dental caries.
Low level summaries

B allele frequencies

log$_2$ R ratios
Data for case-parent trio: false positive

\[ \log_2(R \text{ ratios}) \quad \text{B allele frequencies} \]

PennCNV joint HMM: ‘332’ is denovo hemizygous deletion
False positives

\[
\log_2(R \text{ ratios}) \quad \text{B allele frequencies}
\]

chr2: 23054_03@100849 (father)

chr2: 23054_02@100849 (mother)

chr2: 23054_01@100849 (offspring)

physical position (Mb)
Computational characteristics

- The joint HMM is $\approx 3$ hours per trio
- $3 \times 2082$ trios $\approx 6,246$ CPU hours

$\$ qstat -u hschwend

\begin{verbatim}
job-ID  name      user    state submit/start at
371289  Top10.sh  hschwend  r  10/30/2009 00:12:21
371291  reasons.sh hschwend  r  10/30/2009 00:14:39
371292   for.sh    hschwend  r  10/30/2009 00:14:39
371293  working.sh hschwend  r  10/30/2009 00:15:09
371294    late.sh  hschwend  r  10/30/2009 00:15:09
371295    and.sh   hschwend  r  10/30/2009 00:15:09
371296   not.sh    hschwend  r  10/30/2009 00:15:09
371297  going.sh   hschwend  r  10/30/2009 00:15:25
371298    to.sh    hschwend  r  10/30/2009 00:15:25
371299   bed.sh    hschwend  r  10/30/2009 00:15:25
\end{verbatim}
The minimum distance

\[
\text{minimum distance} \equiv (r_O - r_F) \times I[|r_O - r_F| > |r_O - r_M|] + (r_O - r_M) \times I[|r_O - r_F| \leq |r_O - r_M|]
\]
False positives

$\log_2(R \text{ ratios})$  B allele frequencies

chr2: 23054_03@100849 (father)

chr2: 23054_02@100849 (mother)

chr2: 23054_01@100849 (offspring)

chr2: md

physical position (Mb)
Algorithm

1. Calculate the minimum distance
2. Segment the minimum distance
3. Posterior classification
Computational characteristics

- 591 CPU hours (versus 6,246)
- When multiple CPUs are detected, the MinimumDistance R package parallelizes by chromosome
  - system time $\approx$ 1 day
Denovo CNV discovery from arrays

\[ \log_2(R \text{ ratios}) \]

\[ B \text{ allele frequencies} \]

\[ \log_2(4/2) \]

\[ \log_2(3/2) \]

\[ \log R \text{ ratios} \]

\[ \log_2(4/2) \]

\[ \log_2(3/2) \]

\[ \log_2(4/2) \]

\[ \log_2(3/2) \]

\[ \text{physical position (Mb)} \]

\[ \log Pr(332 | .) = 2725.0 \]

\[ \log Pr(333 | .) = 2302.8 \]

\[ \text{log ratio: 422.3} \]
Denovo CNV discovery from arrays

\[ \log_2(\text{R ratios}) \quad \text{B allele frequencies} \]

- `chr22: 16142_03@100849 (father)`
- `chr22: 16142_02@100849 (mother)`
- `chr22: 16142_01@100849 (offspring)`
- `chr22: min dist`

Physical position (Mb):

- States:
  - State 332

Log probabilities:

\[ \log \text{Pr}(332 \mid .) = 3071.5 \]
\[ \log \text{Pr}(333 \mid .) = 2837.7 \]

Log ratio: 233.8
Denovo CNV discovery from arrays

log₂(R ratios) B allele frequencies

chr22: 22048_03@007029 (father)

chr22: 22048_02@007029 (mother)

chr22: 22048_01@007029 (offspring)

chr22: min dist

State 335

log Pr(335 | .) = 9232.9
log Pr(333 | .) = 6045.5
log ratio: 3187.4
Denovo CNV discovery from arrays

log$_2$(R ratios)  B allele frequencies

chr22: 21207_03@100842 (father)

chr22: 21207_02@100842 (mother)

chr22: 21207_01@100842 (offspring)

chr22: min dist

log Pr(335 | .) = 2769.8
log Pr(333 | .) = 1836.5
log ratio: 933.3
Denovo CNV in the DiGeorge region
## Sample size

<table>
<thead>
<tr>
<th></th>
<th>Cleft</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td>2082</td>
<td>752</td>
</tr>
<tr>
<td>WGA</td>
<td>120</td>
<td>99</td>
</tr>
<tr>
<td>MAD &gt; 0.3</td>
<td>212</td>
<td>138</td>
</tr>
<tr>
<td>non-EA</td>
<td>1090</td>
<td>-</td>
</tr>
<tr>
<td><strong>n</strong></td>
<td>660</td>
<td>515</td>
</tr>
</tbody>
</table>

- Cell frequencies are the number of Father-Mother-Offspring trios.
- The controls are trios with dental caries.
Frequency of de novo deletions

Cleft GWAS

June 24, 2012
Size of deletions

Size of deletions (kb)

Control
Cleft

Rob Scharpf ()
Cleft GWAS
June 24, 2012 23 / 41
Distance to nearest gene

![Distance to nearest gene (kb)](image)

Control

Cleft GWAS

Cleft

June 24, 2012 24 / 41
Association of de novo deletions with oral cleft

Rob Scharpf

Cleft GWAS

June 24, 2012

26 / 41
Association of de novo deletions with oral cleft
MinimumDistance
Association of cleft with denovo deletions

PennCNV
What about the DiGeorge region?
What about the DiGeorge region?

Only 2 de novo deletions among the offspring with European ancestry
European ancestry
Chromosome 7

PennCNV only
Chromosome 7

PennCNV only
Conclusions

- False positives are problematic using standard approaches for de novo CNV discovery
- Oral cleft offspring tend to have slightly bigger de novo deletions than in the controls, and the deletions tended to be closer to coding regions of the genome
- The association of the chromosome 7 region was statistically significant after adjusting for multiple testing.
- The association study used a small fraction of the available oral cleft data.
Acknowledgements

- Samuel Younkin
- Ingo Ruczinski
- Terri Beaty
- Holger Schwendl
- Alan Scott
- Jackie Hetmanski