ANALYSIS OF STRUCTURAL VARIATION

PART II

SEQUENCE ANALYSIS WORKSHOP
HYUN MIN KANG
Many CNVs are contributed by Non-Allelic Homologous Recombination (NAHR)
**Shotgun Sequence Reads**

- Typical short read might be <25-100 bp long and not very informative on its own

- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences
Paired End Sequencing

Population of DNA fragments of known size (mean + stdev)
Paired end sequences
Paired End Sequencing

Paired Reads

Initial alignment to the reference genome

Paired end resolution
Distance Between Pair Ends

- The graph shows distance between paired end reads
- Data summarized across 24 samples
- Courtesy: Xiaowei Zhan, University of Michigan DNA Sequencing Core
EVIDENCE FOR A DELETION WITHIN A SINGLE INDIVIDUAL

- Split Reads
- Read Pair Separation
- Read Depth

Figure from Handsaker et al (2011)
Multi-sample CNV (Deletions) Calling

- CNV events (HMM on depth)
- SV junction events (read mapping)
- Merged Candidate Intervals
- GC-Corrected Coverage
- Multi-sample Clustering
Detecting Copy Number Variation Based on Read Depths

• Focus on a particular feature of the data
  – e.g., read depth

• Normalize depth for each individual
  – e.g., adjust for total read count
  – e.g., adjust for GC content specific read count

• Model data as a mixture of distributions, characterized using maximum likelihood
Where

\[ d_i \sim p_0 N(\mu_0, \sigma_0^2) + p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2) \]

\[ \begin{align*}
  d_i & \text{ is the depth for individual } i \\
p_j & \text{ is the frequency of individuals with } j \text{ deletions (assuming Hardy Weinberg Equilibrium)} \\
\mu_j \text{ and } \sigma_j^2 & \text{ are the mean and variance of adjusted read depth distribution for deletion count } j
\end{align*} \]
To estimate a deletion model, maximize

\[ L(d_i) = \sum_j p_j (2\pi)^{-\frac{1}{2}} \sigma_j^{-1} \exp \left[ -\frac{(d_j - \mu_j)^2}{2\sigma_j^2} \right] \]

To keep number of parameters modest, we use HWE for modeling (one parameter for three frequencies) and can impose additional structure on means and variances.
WELL SEPARATED REGION
MODERATELY SEPARATED REGION

Frequency vs. Normalized Read Counts (left)

Frequency vs. Normalized Read Counts (right)

Sara Rashkin
HARD TO CALL REGION

Sara Rashkin
Cluster Evaluation - # of Components

- Bayesian Information Criterion (BIC)
  \[ \text{BIC} = -2 \ \text{LLK} + k \ \ln (n) \]
  - LLK : Log-likelihood
  - k : number of clusters
  - n : number of samples

- Fit Gaussian mixtures with 1, 2 and 3 components, and compare BIC of 1, 2, 3 component models to decide number of clusters.
BIC Example

BIC chooses 1 component

BIC chooses 2 components

Singleton
Cluster Evaluation – Overlap

• Evaluation of cluster separation
  – Unavoidable error: overlap between two distributions
  – Bayes error rate ~ Bhattacharyya coefficient
  – For two Gaussian distributions,
    \[ D = \frac{(\mu_1 - \mu_2)^2}{8 \sigma_{\text{avg}}^2} + \frac{1}{2} \log \left[ \frac{\sigma_{\text{avg}}}{\sqrt{\sigma_a \sigma_b}} \right] \]
    \[ P(\text{Overlap}) = \exp(-D) \]
P(Overlap) Example

max exp(-D) = 0.933

max exp(-D) = 0.0032
SENSITIVITY/ERROR VS P(OVERLAP)

Quality Metrics with Max. Overlap

- Error rate
- HET/REF
- HET/ALT
- # Variants
### 1000G CNV Default CG vs Multi-sample

<table>
<thead>
<tr>
<th></th>
<th>CG Default</th>
<th>Multi-sample</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td># Large Deletions</td>
<td>2,374</td>
<td>8,321</td>
<td>-</td>
</tr>
<tr>
<td>Call Rate (%)</td>
<td>95.2</td>
<td>99.9</td>
<td>96.73</td>
</tr>
<tr>
<td>Merlin-estimated Error Rate (%)</td>
<td>&gt;10</td>
<td>0.078</td>
<td>0.494</td>
</tr>
<tr>
<td>Trio HET/HomREF</td>
<td>0.750</td>
<td>1.012</td>
<td>0.962</td>
</tr>
<tr>
<td>Trio HET/HomALT</td>
<td>1.055</td>
<td>1.014</td>
<td>1.001</td>
</tr>
</tbody>
</table>
CHALLENGES IN READ DEPTH BASED CALLING

- Ideal if number of reads per region is large
- As technologies improve and reads get longer ...
- ... read depth based calling becomes harder
- Important to integrate different types of signal!
EVIDENCE AT THE POPULATION LEVEL

• Allele Shared Between Multiple Individuals
  – Multiple individuals show cluster of reads with unusual separation in the same location

• Evidence for Deletion Recurs in the Same Individuals
  – Individuals with one unusually separated pair of reads, likely to show additional nearby read pairs with unusual separation

• Evidence for Reference Allele Decreases as Evidence for Deletion Increases
  – When the number of reads with unusual separation increases, the number of nearby reads with expected separation decreases

• Deletions Segregate on Specific Haplotypes
**Refined Algorithm**

- Build list of candidate variants by finding read pairs with abnormal separation
- Focus on regions supported by multiple pairs
- Check whether highly separated pairs are evenly distributed across individuals (*why?*)
- Evaluate read depth distribution
- Search for split reads spanning breakpoint
- Combine with haplotype based hidden Markov model analysis

Handsaker et al (2011)
Search for Abnormal Read Pairs

• Search for read pairs where separation >10x the individual specific standard deviation

• Even if we require multiple supporting events, the number of potential copy number changes is ~10x larger than expected

• This is because of experimental limitation in preparing read pair libraries and of shortcomings in read mapping

• A major challenge is to reduce list of candidates

Handsaker et al (2011)
“Heterogeneity”

- Is rate at which widely separated read pairs occur constant among individuals?

- Calculated expected number of widely separated pairs using sequencing depth, average pair separation

\[ \text{Handsaker et al (2011)} \]
The approach of Handsaker et al. requires that we calculate, for each individual, the expected number of widely separated read pairs.

To do this, Handsaker et al. (2011) calculate the distance between every mapped pair of reads.

They then assume that the number of read pairs separated by >x bp is proportional to the number of reads (across the genome) for which this distance exceeds x.
“Allelic Substitution”

• If we see evidence for deletion, based on read pair separation ...
• Expect to see reduced evidence for reference based on read depth

![Graphs showing read depth comparison](image-url)
“Allelic Substitution”

- If we see evidence for deletion, based on read pair separation ...
- Expect to see reduced evidence for reference based on read depth

Handsaker et al (2011)
**Sizing the Deletion**

- If we know the distribution of read pair distances for one individual...

- Observing an abnormal read pair suggests a specific deletion size, but with low confidence

- Observing many abnormal read pairs gradually suggests more specific deletion sizes and locations
Combining Information Across Individuals is Key

Chr. 5: 107.226–107.227 Mb

Actual deletion length = 837 bp

Likelihood estimate from:
- Individual read pairs
- All genomes together
Conclusions

• Combining information across individuals improves the power of deletion analyses

• Combining different sources of information within each individual also provides increased resolution

• Avoiding experimental artifacts is a major challenge in analysis of copy number
Recommended Reading