Robust Detection and Identification of Sparse Segments in Ultra-High Dimensional Data Analysis

Hongzhe Li

hongzhe@upenn.edu, http://statgene.med.upenn.edu

University of Pennsylvania Perelman School of Medicine

Joint work with Jessie Jeng and Tony Cai
Genetic variations and complex diseases

Commonly observed genetic variations:
- Single nucleotide variants (SNVs).
- Small insertions/deletions (InDels).
- Structure variations, including the copy number variations (CNVs).

All are associated with risk of complex diseases.
Copy number variants (CNVs)

Many CNVs have functional consequences: alter gene dosage, disrupt genes, or uncover deleterious alleles.

NEJM 2007, 356:1169-1171
CNV Associations

Relative Impact of Nucleotide and Number Variation on Gene Expression

Strong Association of De Novo Copy Number Mutations with Autism

Increase in GSK3β Gene Copy Number Variation in Bipolar Disorder

Copy Number Variation in the Human Genome and Its Implications for Cardiovascular Disease

Robust Segment Identification - 4 of 34-1
Data available for CNV Analysis, literature

Two types of data can be used for CNV analysis for germline DNA.

- **SNP chip data for GWAS** - high dimensional continuous data. Sebat et al. 2004; McCarrol and Altshuler 2007; Wellcome trust (Nature 2010).

- **Next generation sequencing data (NGS)** - ultra high dimensional discrete data; Medvedev et al., 2009 (Nat. Meth).

Methods available:

- GWAS SNP data: circular binary segmentation (CBS) (Olshen et al, 2004); HMM based method (PennCNV, Wang et al 2007); scanning-based methods (Zhang and Siegmund 2010); Likelihood ratio selection (Jeng, Cai and Li, 2010).

- NGS data: Most methods are computational.
CNV analysis based on SNP Chip data

Visualization of CNVs

- 4 copies: BBBB, ABBB, AABB, AAAB, AAAA
- 3 copies: BBB, ABB, AAB, AAA
- Normal: BB, AB, AA

Robust Segment Identification - 6 of 34-1
CNV Analysis and Next Generation Sequence Data

Sequence Read Depth Analysis

Individual sequence

Reads

Mapping

Reference genome

Counting mapped reads

Read depth signal

Zero level
CNV Analysis and Next Generation Sequence Data
Statistical challenges

$Y_i$: # of read counts covering location $i$, $i = 1, \cdots, n$; or $Y_i$: # of read counts in 100bp intervals.

- $n$ is ultra-high, computational challenge.

- $Y_i$ usually does not follow a normal distribution, outliers. Existing methods do not work well when noise distribution is non-Gaussian and hard to be estimated.

Cauchy distribution: data, LRS, RSI.
Statistical model for read depth data - one sample

For a given individual, observe read counts \( \{Y_i, i = 1, \ldots, n\} \) with

\[
Y_i = \mu_1 1_{\{i \in I_1\}} + \ldots + \mu_q 1_{\{i \in I_q\}} + \xi_i, \quad 1 \leq i \leq n. \tag{1}
\]

\( n \): length of genome (billions);

\( q = q_n \): unknown number of the signal segments;

\( \Pi = \{I_1, \ldots, I_q\} \): disjoint intervals representing signal segments with unknown locations;

\( \mu_1, \ldots, \mu_q \) are unknown means

\( \xi_i \) is symmetric at 0 (median(\( \xi_i \))=0) and density function \( h \) s.t.

\[
h(0) > 0, \quad |h(y) - h(0)| \leq Cy^2 \text{ in an open nbhd of 0.}
\]

We want to
(a) (detection) test \( H_0 : \Pi = \emptyset \) against \( H_1 : \Pi \neq \emptyset \),
(b) (identification) if the alternative is true, identify each \( I_j \in \Pi \).
Methods Assuming Gaussian Noise (Jeng et al., 2010)

Methods for detecting the presence of segments assuming Gaussian noise: Arias-Castro, Donoho and Huo (2005)

Identification methods assuming Gaussian noise:
likelihood ratio selector (LRS) (Jeng, Cai and Li 2010 JASA).

Key of the LRS:
(1) For any given interval $\tilde{I} \subseteq \{1, 2, ..., n\}$, define its likelihood ratio statistic as

$$Y(\tilde{I}) = \sum_{i \in \tilde{I}} Y_i / \sqrt{|\tilde{I}|}.$$ 

(2) Scan the genome with intervals of length $\leq L$, threshold $\sigma \sqrt{2 \log(nL)}$.
(3) Identify local maximums by removing the overlapping intervals.

Detection boundaries and optimality results are established.
Non-normal Data - Local median transformation

- Equally divide the $n$ observations (e.g., counts at each bp) into $T = T_n$ groups with $m = m_n$ observations in each group.

- Define $k$th interval $J_k = \{ i : (k - 1)m + 1 \leq i \leq km \}$ and take median:
  
  $$X_k = \text{median}(Y_i : i \in J_k), \quad \eta_k = \text{median}\{\xi_i : i \in J_k\}, \quad 1 \leq k \leq T.$$

- We have
  
  $$X_k = \theta_k + \eta_k, \quad 1 \leq k \leq T,$$

  
  
  $$\theta_k \begin{cases} 
  = \mu_j, & J_k \subseteq I_j \text{ for some } I_j, \\
  \in [0, \mu_j], & J_k \cap I_j \neq \emptyset \text{ for some } I_j \text{ and } J_k \not\subseteq I_j, \\
  = 0, & \text{otherwise}.
  \end{cases}$$

  Key point: $\sqrt{m}\eta_k = \frac{1}{2h(0)}Z_k + \zeta_k, \quad Z_k \sim N(0, 1), \quad \zeta_k \rightarrow D 0$ fast

  $\Rightarrow \eta_k \sim N(0, 1/(4h^2(0)m))$.

  (Brown, Cai and Zhou: AoS 08).
Robust Segment Detection (RSD)

Segment detection: test $H_0 : \mathbb{I} = \emptyset$ vs $H_1 : \mathbb{I} \neq \emptyset$.

For any interval $\tilde{I}$, define

$$X(\tilde{I}) = \sum_{k \in \tilde{I}} X_k / \sqrt{|\tilde{I}|},$$

and threshold

$$\lambda_n = \sqrt{2 \log n / (2h(0)\sqrt{m})}.$$  

The RSD rejects $H_0$ when $\max_{\tilde{I} \in \mathbb{J}_T} X(\tilde{I}) > \lambda_n$, where $\mathbb{J}_T$ is the collection of all possible intervals in $\{1, \ldots, T\}$.

Note the effect of $h(0)$. 
Under the assumed model and median transformation with $m = \log^{1+b} n$ for some $b > 0$.

**Type 1 error**: For the collection $\mathbb{J}_T$ of all the possible intervals in $\{1, \ldots, T\}$,

$$P_{H_0}(\max_{\tilde{I} \in \mathbb{J}_T} X(\tilde{I}) > \lambda_n) \leq \frac{C}{\sqrt{\log T}} \to 0, \quad T \to \infty.$$

**Power**: If there exists some segment $I_j \in \mathbb{I}$ that satisfies

$$|I_j|/m \to \infty$$

and

$$\mu_j \sqrt{|I_j|} \geq \sqrt{2(1 + \epsilon) \log n/(2h(0))}$$

for some $\epsilon > 0$, then RSD has the sum of the probabilities of type I and type II errors going to 0.
Robust Segment Identifier (RSI)

- Perform local median transformation with bin size $m$, get
  \[ X_k = \theta_k + \eta_k, \ 1 \leq k \leq T. \]

- Set data-driven threshold at
  \[ \lambda_n^* = \hat{\sigma} \sqrt{2 \log n}, \quad \hat{\sigma}^2 : \text{estimate of } \Var(\eta_k) (e.g., \text{MAD}) \]

- Apply LRS (Jeng, Cai and Li, JASA 2010) on $X_k$:
  - select intervals with their likelihood ratio statistics $> \lambda_n^*$ and
    achieve local maximums (removing overlapping intervals).
  - only consider short intervals with length $\leq L/m$, $L$: max CNV size.

- Conditions on $m$ and $L$:
  \[ m = \log^{1+b} n, \quad \bar{s} \leq L < d, \]
  \[ b > 0, \quad \bar{s} = \text{length of the longest segment, } d = \text{shortest distance between two adjacent segments.} \]
Assume the general condition on the background noise $\xi_i$ and some sparsity conditions on the signal segments. Define $\underline{s} = \min_{I_j \in \mathbb{I}} |I_j|$, $\bar{s} = \max_{I_j \in \mathbb{I}} |I_j|$, and $d = \min_{I_j \in \mathbb{I}} \{\text{distance between } I_j \text{ and } I_{j+1}\}$, assume

$$\underline{s} \geq \log^2 n \quad \text{and} \quad \log \bar{s} = o(\log n) \quad \text{and} \quad \log q = o(\log n).$$

If all $I_j \in \mathbb{I}$ satisfies $|I_j|/m \to \infty$ and

$$\mu_j \sqrt{|I_j|} \geq \sqrt{2(1 + \epsilon) \log n/(2h(0))}$$

for some $\epsilon > 0$, then the RSI with $m = \log^{1+b} n$ for $b > 0$ and $\bar{s} \leq L < d$, is consistent for $\mathbb{I}$, i.e., for some $\delta_n = o(1)$,

$$P_{H_0}(|\hat{\mathbb{I}}| > 0) + P_{H_1}(\max_{I_j \in \mathbb{I}} \min_{\hat{I}_j \in \hat{\mathbb{I}}} D(\hat{I}_j, I_j) > \delta_n) \to 0$$

**Dissimilarity:** $D(\hat{I}, I) = 1 - |\hat{I} \cap I|/\sqrt{||\hat{I}||I||}$
If for all \( I_j \in \mathbb{I} \),

\[
\mu_j \sqrt{|I_j|} \leq \sqrt{2(1 - \epsilon) \log n / (2h(0))},
\]

then no method constructed on \( X_k \) with \( m \to \infty \) is consistent.
Comparison with Gaussian noises

- Compare to the case with Gaussian noise:
  - Assume $\xi_i \sim N(0, 1)$, then the original GLRT based on $Y_i$ is optimal.
  - Further, if $\exists I_j \in \mathbb{I}$ s.t.
    \[
    \mu_j \sqrt{|I_j|} \geq \sqrt{2(1 + \epsilon_n) \log n},
    \]
    then the original GLRT is consistent.

Possible price for robustness:
\[
\sqrt{2(1 + \epsilon_n) \log n} / (2h(0)) \approx 1.25 \times \sqrt{2(1 + \epsilon_n) \log n}
\]
Simulation studies

\[ n = 5 \times 10^4, |\Pi| = 3, |I_1| = 100, |I_2| = 40 \text{ and } |I_3| = 20, \]
the signal mean for all segments at \( \mu = 1.0, 1.5, \text{ and } 2.0. \)

Noise is generated from \( \text{Cauchy}(t(1)), t(3), t(30), \) median transformation \( m = 20. \)
Simulation results - robustness

\( n = 5 \times 10^4, |\| = 3 \). Noise is generated from \( t(1), t(3), t(30) \).

Estimation error for \( I_j \): \( D_j = \min_{\hat{I}_k \in \hat{I}} \left\{ 1 - |I_j \cap \hat{I}_k|/\sqrt{|I_j||\hat{I}_k|} \right\} \in [0, 1] \).

Number of over-selections: \#O = \#\{\hat{I} \in \hat{I} : \hat{I} \cap I_j = \emptyset, \forall j = 1, \ldots, q\}.

Medians of \( D_j \) and \#O for RSI with \( m = 20, L = 120, 100 \) replications.

|       | \( D_1(|I_1| = 100) \) | \( D_2(|I_2| = 40) \) | \( D_3(|I_3| = 20) \) | \#O     |
|-------|-----------------|-----------------|-----------------|--------|
| \( t(1) \) |                 |                 |                 |        |
| \( \mu = 1.0 \) | 0.080(0.015) | 1.000(0.026) | 1.000(0.000) | 2(0.33) |
| \( \mu = 1.5 \) | 0.087(0.003) | 0.184(0.017) | 1.000(0.000) | 2(0.26) |
| \( \mu = 2.0 \) | 0.087(0.009) | 0.150(0.020) | 0.423(0.220) | 2(0.14) |
| \( t(3) \) |                 |                 |                 |        |
| \( \mu = 1.0 \) | 0.087(0.005) | 1.000(0.270) | 1.000(0.000) | 0(0.00) |
| \( \mu = 1.5 \) | 0.060(0.009) | 0.175(0.029) | 1.000(0.000) | 0(0.00) |
| \( \mu = 2.0 \) | 0.050(0.008) | 0.150(0.016) | 0.293(0.019) | 0(0.00) |
| \( t(30) \)  |                 |                 |                 |        |
| \( \mu = 1.0 \) | 0.070(0.014) | 1.000(0.320) | 1.000(0.000) | 0(0.00) |
| \( \mu = 1.5 \) | 0.065(0.012) | 0.175(0.021) | 1.000(0.245) | 0(0.00) |
| \( \mu = 2.0 \) | 0.050(0.010) | 0.175(0.019) | 0.250(0.028) | 0(0.00) |
Simulation results - comparison with LRS and CBS

Table 1: Both homogeneous and heterogeneous noises are considered. Homogenous noise is generated from the $t$-distribution with degrees of freedom 1, 3, and 30. Heterogeneous noise is generated from a mixture of $N(0, 1)$ and $N(0, \sigma^2)$, where $\sigma \sim Gamma(2, \tau)$. $\mu$ is fixed at 2.0.

<table>
<thead>
<tr>
<th></th>
<th>RSI</th>
<th></th>
<th>LRS</th>
<th></th>
<th>CBS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_2(</td>
<td>I_2</td>
<td>= 40)$</td>
<td>#O</td>
<td>$D_2(</td>
<td>I_2</td>
</tr>
<tr>
<td>$t(1)$</td>
<td>0.163(0.024)</td>
<td>2(0.2)</td>
<td>0.340(0.054)</td>
<td>3882(7)</td>
<td>1.000(0.000)</td>
<td>0(0.0)</td>
</tr>
<tr>
<td>$t(3)$</td>
<td>0.125(0.028)</td>
<td>0(0.0)</td>
<td>0.025(0.006)</td>
<td>467(4)</td>
<td>1.000(0.000)</td>
<td>0(0.0)</td>
</tr>
<tr>
<td>$t(30)$</td>
<td>0.125(0.018)</td>
<td>0(0.0)</td>
<td>0.000(0.001)</td>
<td>2(0)</td>
<td>0.006(0.006)</td>
<td>0(0.0)</td>
</tr>
<tr>
<td>$\tau = 0.5$</td>
<td>0.125(0.015)</td>
<td>2(0.4)</td>
<td>0.013(0.005)</td>
<td>37(3)</td>
<td>0.180(0.006)</td>
<td>4(0.6)</td>
</tr>
<tr>
<td>$\tau = 1.0$</td>
<td>0.113(0.022)</td>
<td>12(0.6)</td>
<td>0.000(0.006)</td>
<td>227(6)</td>
<td>1.000(0.010)</td>
<td>10(1.1)</td>
</tr>
<tr>
<td>$\tau = 1.5$</td>
<td>0.125(0.016)</td>
<td>26(0.8)</td>
<td>0.000(0.006)</td>
<td>461(11)</td>
<td>1.000(0.000)</td>
<td>8(1.1)</td>
</tr>
</tbody>
</table>
Simulation results - effect of m

Table 2: Effect of bin size $m$ on the performance of RSI. $\mu$ is fixed at 2.

| $t$     | $m$  | $D_1(|I_1|=100)$ | $D_2(|I_2|=40)$ | $D_3(|I_3|=20)$ | #O  |
|---------|------|------------------|------------------|------------------|-----|
| $t(1)$  | $m=10$ | 0.035(0.009)     | 0.10(0.018)      | 0.184(0.033)     | 19(0.85) |
|         | $m=20$ | 0.087(0.009)     | 0.15(0.020)      | 0.423(0.220)     | 2(0.14)  |
|         | $m=40$ | 0.101(0.006)     | 0.25(0.056)      | 1.000(0.024)     | 0(0.00)  |
| $t(3)$  | $m=10$ | 0.030(0.004)     | 0.088(0.015)     | 0.150(0.033)     | 1(0.22)  |
|         | $m=20$ | 0.050(0.008)     | 0.150(0.016)     | 0.293(0.019)     | 0(0.00)  |
|         | $m=40$ | 0.087(0.006)     | 0.293(0.041)     | 1.000(0.250)     | 0(0.00)  |
| $t(30)$ | $m=10$ | 0.020(0.007)     | 0.075(0.008)     | 0.150(0.018)     | 0(0.00)  |
|         | $m=20$ | 0.050(0.010)     | 0.175(0.019)     | 0.250(0.028)     | 0(0.00)  |
|         | $m=40$ | 0.105(0.008)     | 0.293(0.035)     | 1.000(0.094)     | 0(0.00)  |
NA19240: an International HapMap project Yoruban daughter sample and parents, lymphoblastoid cell lines (Nature 2010).
Ave 42x, SOLiD, map to the human reference genome (BAM file), ave 2.36 Gb accessed.

\[ n = 54,361,060 \] read counts for Chr 19.
Apply RSI with \( m = 400, L = 60,000 \).
Identified 106 deletions and 63 duplications.
Take less than 3 mins.

Compare with the CNV map from 1000 Genomes Project based on 185 samples (Mills et al. 2011), 76 overlap with the reported deletions based on 185 low-coverage samples and three methods (438, 332 and 615 CNVs).
Concordant of the Yoruba trio - top ranked CNVs

CNVs are inheritable - concordant rates, ranked by $\mu \sqrt{|\hat{I}|}$, apply to 3 samples separately.
Possible error in reference genome: multicopy sequences which have been incorrectly assembled and collapsed into a single copy.
Chr 19 sequence data - NA19240
Chr 19 sequence data - NA19240

Robust Segment Identification - 28 of 34-1
Alternative Approach - Negative Binomial Counts

NB model:

\[ Y_i \sim \text{Negative Binomial}(r, p_i), \quad p_i = p_0 + \sum_{j=1}^{q} d_j 1(i \in I_j). \]

\[ \mu_i = \frac{rp_i}{1 - p_i}, \quad \sigma_i^2 = \frac{\mu_i^2}{r} + \mu_i. \]

Data transformation - mean-matching variance stabilizing transformation (VST) to turn into Gaussian noise problem: divide the \( n \) obs into \( T = T_n \) groups of \( m = m_n \) obs, for the \( k \)th interval, define

\[ X_k = 2\sqrt{\hat{r}} \ln \left( \sqrt{\frac{\sum_{i \in J_k} Y_i + 1/4}{m\hat{r} - 1/2}} + \sqrt{1 + \frac{\sum_{i \in J_k} Y_i + 1/4}{m\hat{r} - 1/2}} \right), \quad 1 \leq k \leq T, \]
Transformed Data - normally distributed

Model for transformed data:

\[ X_k = 2 \ln(\sqrt{\theta_k} + \sqrt{r + \theta_k}) + \epsilon_k + m^{-1/2}Z_k + \xi_k, \]

where

\[ \theta_k \begin{cases} = r(p_0 + d_j)/(1 - p_0 - d_j), & J_k \subseteq I_j \text{ for some } I_j, \\ \in [rp_0/(1 - p_0), \ r(p_0 + d_j)/(1 - p_0 - d_j)], & J_k \cap I_j \neq \emptyset \text{ for some } I_j \text{ and } J_k \nsubseteq I_j, \\ = rp_0/(1 - p_0), & \text{otherwise}, \end{cases} \]

\( \epsilon_k \) and \( \xi_k \) are stochastically small, \( Z_k \sim N(0, 1) \).

Apply LRS to \( X_k \) assuming the baseline distribution as

\[ N \left( 2 \ln\left( \frac{rp_0}{1 - p_0} + \sqrt{\frac{r}{1 - p_0}} \right), 1/m \right). \]
Concordant of the Yoruba trio - top ranked CNVs

CNVs are inheritable - concordant rates, ranked by $\mu \sqrt{|\hat{I}|}$, apply to 3 samples separately.

![Bar chart showing concordance percentages for top20, top50, top100, top200, and top300 CNVs.](chart.png)
Comments and extensions

Cai, Jeng and Li (2012): JRSS(B), in press.

Many other complicated factors: repeated regions, complex rearrangements, highly repetitive elements.

There is a relationship between GC content and coverage, but this effect is small for false CNVs.

Read depths data: difficulty in finding high repetitive CNVs (LINE, SINE), uncertain in CNV location, but can ba applied to paired-end, single-end and mixed data;

Paired-end whole genome sequencing data: statistical modeling of anomalous read pairs, can detect highly repetitive CNVs (LINE and SINE), precise location of CNVs; but span distances have effects on resolution.
Software - RobustCNV

C++ software available, with some post-processing steps:

- CG content adjustment;
- refining boundaries;
- filtering CNVs based on read mapping qualities;
- plots of the CNVs.
THANKS!

Collaborators:
Jessie Jeng - Postdoc
Yinghua Wu - Postdoc
Tony Cai - Statistics Dept
John Maris - CHOP.

NIH grants support.