Estimating tumor copy number profiles from DNA sequencing

Venkatraman E. Seshan
MSKCC

Nov 2, 2015, iBright 2015, MD Anderson

Collaborator: Ronglai Shen

Adam Olshen
Nick Socci, Barry Taylor
Arshi Arora, Charlotte Ng, Britta Weigelt, Jorge Reis-Filho, Paul Paik
Somatic Mutations in Cancer

Cancer is characterized by somatic changes to DNA

- p53 (or tumor protein 53)
- EGFR, BRAF, KIT/PDGFR mutations targetable
- KRAS mutation associated with non-response
- ERBB2 (Her2/neu) amplification

High throughput assay to screen simultaneously

- CGH and SNP arrays for copy number profiling
- Sequenom multiplex assay for (known) mutations
- Next Generation Sequencing of DNA
Next Generation Sequencing

Library of DNA fragments to be “read”

- Whole genome — library is unselected
- Targeted — capture technology to select DNA fragments
- Whole exome — all exons (coding part) of the genome targeted region cover ∼50 million bases
- Cancer gene panel — MSK-IMPACT ∼410 genes and < 1 million bases

Fundamental unit of data is a DNA fragment
Analysis of NGS Data

- Tumor and normal tissue sequenced
- Reads mapped to the genome
- Tumor and normal reads compared
  - nucleotide mismatch for point mutations
  - gapped alignment for indels (insertions/deletions)
  - abundance (depth) ratio for copy number
  - account for technical artifacts (e.g. GC dependence)

Focus of this talk is copy number profiling
Available Data

Data are counts of reference and alternate alleles

- Whole exome sequencing $\sim 50$ million bases in $200k$ intervals
- Expanded intervals by $50$ bases on each size (target overhang)
- dbSNP provides $1.9$ million polymorphic positions among them
- Use only reads that meet base and mapping quality thresholds
- Select positions that meet depth threshold in normal sample
  say at least $25$ in $50\times$ coverage experiment
- Avoid serial correlation by spacing SNP positions

Convert to (250k) microarray data: R Irizarry/J Leek

Fewer SNP loci (10k) in cancer gene panels; still (often) sufficient.
6 snp locations one of which is heterozygous in sample

<table>
<thead>
<tr>
<th>chr</th>
<th>pos</th>
<th>ref</th>
<th>depth</th>
<th>A</th>
<th>a</th>
<th>C</th>
<th>c</th>
<th>G</th>
<th>g</th>
<th>T</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6705180</td>
<td>T</td>
<td>129</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>112</td>
</tr>
<tr>
<td>1</td>
<td>6705184</td>
<td>C</td>
<td>124</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6705793</td>
<td>G</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>47</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6705828</td>
<td>G</td>
<td>88</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6705874</td>
<td>G</td>
<td>123</td>
<td>1</td>
<td>1</td>
<td>38</td>
<td>23</td>
<td>38</td>
<td>22</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6705875</td>
<td>T</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>61</td>
<td>47</td>
</tr>
</tbody>
</table>

One snp from rows 1-2 and 3-6 make it to the pseudo-array.

Read depth ratio gives information on (relative) total copy number

Allele depth (het snps) informs of changes in parental chromosome
Estimate What?

$D_{rn}, D_{an}, D_{rt}, D_{at}$ : ref and alt read depths from tumor and normal

- Normal cell — one chrom from each parent (call M & P)
- Tumor has $m$ copies from M and $p$ copies from P
- Purity of tumor sample is $\phi$

Need to estimate $m, p, \phi$ — identifiability issues.

Are $m^* = 1 + (m - 1)\phi, \quad p^* = 1 + (p - 1)\phi$ estimable?

$$\log(D_{t}/D_{n}) = \log\{(m^* + p^*)/2\} + w(\cdot) + \lambda + \epsilon$$

$w$ is the systematic bias (GC artifact)

$\lambda$ constant for relative to absolute copy number conversion

Venkatraman E. Seshan  MSKCC  CN profile from DNAseq, iBright2015
Allelic Imbalance

- B or alt allele frequency informs on $m^*/(m^* + p^*)$
- Does not account for ref-alt bias ($r$) in sequencing

For heterozygous snp do not know allele-parent map

<table>
<thead>
<tr>
<th></th>
<th>counts</th>
<th>ref on M</th>
<th>ref on P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ref alt</td>
<td>ref alt</td>
<td>ref alt</td>
</tr>
<tr>
<td>normal</td>
<td>$D_{rn}$ $D_{an}$</td>
<td>1 $r$</td>
<td>1 $r$</td>
</tr>
<tr>
<td>tumor</td>
<td>$D_{rt}$ $D_{at}$</td>
<td>$m^<em>$ $p^</em>$</td>
<td>$p^<em>$ $r m^</em>$</td>
</tr>
</tbody>
</table>

log-odds ratio  \( \text{logOR} = \log\{(D_{rt}D_{an})/(D_{rn}D_{at})\} \)

Squared logOR measure of allelic imbalance \( \log^2(m^*/p^*) \)
Whole exome sequencing of breast tumor

248166 snp loci, 34244 heterozygous in germline
How do we estimate $m^*, p^*$ and then $m, p, \phi$ 

- Changes affect contiguous loci
- Segment the genome into regions of constant copy number and allelic imbalance
- Allelically balanced regions are possibly “normal” diploid
- Use it to estimate the log-ratio for 2 copy state
- Copy number per cell is integer valued
  - use it to estimate copy number and tumor fraction
- Borrow strength from similar regions for a better estimate
Let $Z_1, Z_2, \ldots, Z_n$ be the data ordered by an indexing set.
If $Z_1, \ldots, Z_\nu \sim F_0$ and $Z_{\nu+1}, \ldots, Z_n \sim F_1$,
then $\nu$ is a change-point (Page, 1954).

For the copy number problem
- the data are the log-ratio measurements
- ordered by the location of a probe on a chromosome
- a change-point corresponds to where the copy number changed on a chromosome
- There may be multiple changes.
Circular Binary Segmentation

View the data as if on a circle and segment into two arcs hence circular binary segmentation (CBS).

Partial sums: \( S_i = Z_1 + \cdots + Z_i, \quad i = 1, \ldots, n \)

Test statistic: \( T = \max_{1 \leq i < j \leq N} |T_{ij}|, \) where

\[
T_{ij} = \frac{\hat{Z}_{ij} - \bar{Z}_{ij}}{s \sqrt{(j - i)^{-1} + (i + n - j)^{-1}}},
\]

\[
\hat{Z}_{ij} = \frac{(S_j - S_i)}{(j - i)}
\]

and \( \bar{Z}_{ij} = \frac{(S_i + (S_n - S_j))}{(i + n - j)} \)

Either one change-point \( (j = n) \) or two \( (j < n) \).

(Olshen et al Biostatistics (2004); Venkatraman & Olshen Bioinformatics (2007))
$O(n \log(n))$ Computation

- $T_{ij}$ computed for all $i, j$, $(n^2)$ can some be eliminated?
- Divide $1, \ldots, n$ into $\sqrt{n}$ bins of $\sqrt{n}$ indices
- Leads to $n$ blocks of $i, j$ combinations

- $|S_j - S_i|$ for block $b_{kl}$ bounded by difference of local max and min for bins $k$ and $l$ of index $i$ and $j$
- $T_{ij}$ bound for blocks
- Maximize $T_{ij}$ by blocks starting from block with largest bound
- Most blocks ruled out and never searched
Joint Segmentation

PSCBS (Olshen et al, 2011)

- Extended to SNP array and parent specific copy number
- Two stage segmentation — total copy number (log-ratio) followed by (mirrored) B-allele fraction
- Not the most efficient use of data

Joint segmentation statistic: \( T = \max_{1 \leq i < j \leq N} \left( T_{1ij}^2 + c T_{2ij}^2 \right) \)

where \( T_{1}^2 \) is the statistic from copy number log-ratio
and \( T_{2}^2 \) is the statistic from the variant allele log-odds-ratio

Use ranks instead of original data
Segment Summaries

- Segments are the units of analysis
- Let \( m, p \) be the parental copy numbers and \( \phi \) the tumor purity
- Log-ratio function of \( \log(m^* + p^*) \)
  (up to a centering constant)
- Log-ratio (logR) summarized using median
- Log-odds-ratio of \( \log(m^*/p^*) \)
  (up to sign since only allele and not parent known)
- Square of log-odds-ratio (logOR) is non-central chi-square

\[
\log \text{ACR}^2 = \frac{\sum \{ (\log OR)^2 - \text{var(\log OR)} \} / \text{var(\log OR)} }{\sum 1/\text{var(\log OR)}}
\]

where sum is over snps in a segment.
Notice several regions of copy neutral LOH (chrom 3p, 5 etc.)
Segments with Allelic Balance

Need diploid logR level for estimating integer CN

- logACR is the logarithm of allelic copy number ratio
- Close to zero when alleles are balanced \( i.e. \ m = p \)
- Normal diploid state is \( m = p = 1 \)
- Homozygous deletion \( m = p = 0 \) on small (focal) segments
- Genome doubling when \( m = p = 2 \) (cell division event)
- Assumption: large segments of \( m = p > 2 \) is not very likely.

logR representing diploid state estimated using balanced segs
Example of Balanced Segments

Diploid level estimated to be -0.2
Segments on chromosomes 11, 12, and 17 above the diploid line
8 segments on chromosome 1

<table>
<thead>
<tr>
<th>seg</th>
<th>num.mark</th>
<th>nhet</th>
<th>cnlr.median</th>
<th>logACR</th>
<th>ocn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5037</td>
<td>812</td>
<td>-0.7977</td>
<td>1.1942</td>
<td>1.32</td>
</tr>
<tr>
<td>2</td>
<td>227</td>
<td>34</td>
<td>-0.0695</td>
<td>0.0769</td>
<td>2.19</td>
</tr>
<tr>
<td>3</td>
<td>283</td>
<td>30</td>
<td>0.1998</td>
<td>0.2607</td>
<td>2.64</td>
</tr>
<tr>
<td>4</td>
<td>8206</td>
<td>1000</td>
<td>-0.1996</td>
<td>0.0135</td>
<td>2.00</td>
</tr>
<tr>
<td>5</td>
<td>519</td>
<td>56</td>
<td>0.1998</td>
<td>3.5574</td>
<td>2.64</td>
</tr>
<tr>
<td>6</td>
<td>78</td>
<td>7</td>
<td>0.5590</td>
<td>NA</td>
<td>3.38</td>
</tr>
<tr>
<td>7</td>
<td>10191</td>
<td>1378</td>
<td>0.1998</td>
<td>3.7957</td>
<td>2.64</td>
</tr>
<tr>
<td>8</td>
<td>704</td>
<td>81</td>
<td>0.5590</td>
<td>4.8399</td>
<td>3.38</td>
</tr>
</tbody>
</table>

Quick and dirty tumor fraction estimate is 65%
No Normal Diploid Segment

Balanced alleles at 0.295 (chr 1, 5 etc.) and −0.081 (2p)

Diploid level estimated to be −0.633
Estimating Allelic Copy Numbers

- Observed copy number a function tumor fraction $\phi$
  and parental copy numbers are $m$ and $p$
- $m$ and $p$ not known; instead infer order statistics
- without loss of generality $m \leq p$
- logACR estimates $\log(p^*/m^*)$
  
  $$m^* = 1 + (m - 1)\phi, \quad p^* = 1 + (p - 1)\phi$$

- Use observed copy number estimate parental copy numbers

Venkatraman E. Seshan  MSKCC
CN profile from DNAseq, iBright2015
\( \phi \) ranges from 0.1 to 0.9; copy numbers mostly identifiable.

---

Venkatraman E. Seshan  MSKCC  
CN profile from DNAseq, iBright2015
Examples Revisited

Venkatraman E. Seshan  MSKCC  CN profile from DNAseq, iBright2015
Software available soon
Fraction and Allele-specific Copy number Estimates from Tumor/normal Sequencing

http://biostats.bepress.com/mskccbiostat/paper29/
Other methods (apologies to the missed ones)

- Copy numbers - GLAD, CGHflasso, CNVnator, CNVseq, SeqSeq, seqCBS
- Tumor composition - ABSOLUTE, ASCAT, cloneHD, PyLOH, SomatiCA, TITAN

Ongoing work

- Hyper-fragmentation of profiles
- Cancer gene panels works sometimes - wave artifacts