Canadian Bioinformatics Workshops

bioinformatics.ca
Somatic copy number alterations in cancer

Sohrab Shah
Bioinformatics for Cancer Genomics
May 27, 2015

Cancer is an evolutionary process

Module 5 bioinformatics.ca
Clonal evolution leads to populations of cells with distinct genotypes

- Do clonal genotypes drive different phenotypic behaviours?
- How does this relate to treatment response, progression, metastasis?
- How can relative fitness be measured?

Diverse cellular constituents of tumours confound genomic measurements

Computational approaches must account for confounding factors
Genome-wide copy number profiling from bulk tissues


Module 5 bioinformatics.ca

Copy number alterations in cancer

Module 5 bioinformatics.ca
A normal human karyotype

Boveri’s hypothesis

• “We start with the assumption that the qualities of malignant cells have their origin in a defect that exist within them.” - Boveri (1914)

• Observations of ‘multipolar mitoses’ in sea urchin cells led Boveri to believe that abnormal distributions of chromosomes in cells were the culprit behind initiation of malignancies

• He was proven correct in 1960 with the discovery of the Philadelphia chromosome in CML
Cancers exhibit disrupted karyotypes

Copy number alterations (CNA) disrupt normal cellular behaviour

- CNAs are segments of a chromosome ~1Kb to whole chromosomes where genetic material is lost or gained
- CNAs are a hallmark of tumour genomes
- CNAs can lead to adverse expression changes of targeted genes
- Task: find CNAs for diagnostics/prognostics, gene-disease association, targets for therapeutics

Module 5

bioinformatics.ca
Classes of copy number alteration

- Single copy gain
- Diploid deletion
- Hemizygous deletion
- Amplification
- Copy neutral loss of heterozygosity

Example: high level amplification of \textit{ERBB2}
Fluorescence in situ hybridization of ERBB2 amplification

Example: homozygous deletion of PTEN
Module 5

Spearman rho = 0.557554

Driver CNAs alter expression of the genes they harbour

17q11.2-q12

11q,13

Hemizygous deletion leads to bi-allelic inactivation of tumour suppressor genes with mutation

12 tumour types, >2700 tumours, >500,000 mutations,

30 novel cancer gene candidates with tumour suppressor LOF mutation patterns
89 novel cancer gene candidates associated with disrupted pathway expression

Ding et al (Nature Communications 2015)
Genes known to be affected by somatic CNAs

- **Amplifications**
  - ERBB2, EGFR, MYC, PIK3CA, IGF1R, FGFR1/2, KRAS, CDK4, CDK6

- **Deletions**
  - RB1, PTEN, CDKN2A/B, ARID1A, MAP2K4, NF1, SMAD4, BRCA1/2, MSH2/6, CDH1

- **Recent high resolution interrogations of somatic copy number landscapes**
  - Ciriello et al – TCGA Pan Cancer Nature 2013
### Actionable gene-based copy number alterations

<table>
<thead>
<tr>
<th>Category of Genomic Alteration</th>
<th>Exemplary Cancer Gene</th>
<th>Type of Cancer</th>
<th>Targeted Therapeutic Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translocation</td>
<td>BCR/ABL</td>
<td>Chronic myelogenous leukemia</td>
<td>Imatinib</td>
</tr>
<tr>
<td></td>
<td>PML-RARA</td>
<td>Acute promyelocytic leukemia</td>
<td>All-trans retinoic acid</td>
</tr>
<tr>
<td></td>
<td>EML-AALK</td>
<td>Breast, colorectal, lung</td>
<td>ALK Inhibitor</td>
</tr>
<tr>
<td></td>
<td>ETS gene fusions</td>
<td>Prostate</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>CDK11</td>
<td>Leukemia, lymphoma, meningioma</td>
<td>—</td>
</tr>
<tr>
<td>Amplification</td>
<td>EGFR</td>
<td>Lung, colorectal, glioblastoma, pancreatic</td>
<td>Cetuximab, gefitinib, erlotinib, panitumumab, lapatinib</td>
</tr>
<tr>
<td></td>
<td>ERBB2</td>
<td>Breast, ovarian</td>
<td>Trastuzumab, lapatinib</td>
</tr>
<tr>
<td></td>
<td>KIT, PDGFR</td>
<td>GISTs, glioma, HCC, RCC, OML</td>
<td>Imatinib, nilotinib, suinib, sunitinib, sorafenib</td>
</tr>
<tr>
<td></td>
<td>NRAS</td>
<td>Brain, colon, leukemia, lung</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>SRC</td>
<td>Sarcoma, OML, ALL</td>
<td>Dasatinib</td>
</tr>
<tr>
<td></td>
<td>PRKCA</td>
<td>Breast, ovarian, colorectal, endometrial</td>
<td>PI3 kinase inhibitors</td>
</tr>
<tr>
<td>Deletions</td>
<td>KIT, PDGFR</td>
<td>GISTs, glioma, HCC, RCC, OML</td>
<td>Imatinib, nilotinib, suinib, sunitinib, sorafenib</td>
</tr>
<tr>
<td></td>
<td>PRKCA</td>
<td>Breast, ovarian, colorectal, endometrial</td>
<td>PI3 kinase inhibitors</td>
</tr>
<tr>
<td></td>
<td>Braf</td>
<td>Melanoma, pediatric acute myeloid</td>
<td>RAF inhibitor</td>
</tr>
<tr>
<td></td>
<td>KRAS</td>
<td>Colorectal, pancreatic, GI tract, lung</td>
<td>Resistance to erlotinib, cetuximab (colorectal)</td>
</tr>
</tbody>
</table>

Abbreviations: ALK, anaplastic lymphoma kinase; GIST, gastrointestinal stromal tumor; HCC, hepatocellular carcinoma; RCC, renal cell carcinoma; OML, chronic myelogenous leukemia; ALL, acute lymphoblastic leukemia; PI3, phosphoinositide 3-kinase.

---

### Patient population stratification by copy number alterations
Copy number profiles indicate compromised DNA repair mechanisms, which in turn can be used to stratify cancers.

Ciriello et al Nature Genetics, 2013

Recent progress in breast cancer

ARTICLE

doi:10.1038/nature10963

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups
High resolution combined with expression sharply focuses the driver landscape

Clustering of combined genomic architecture/expression landscapes reveals novel subgroups
Subgroups of breast cancer with contrasting copy number profiles

Curtis et al. Nature 2012

Subgroups split pre-existing subtypes into groups with different prognostic signatures

Discovery set

Logrank $P = 1.2 \times 10^{-14}$

Months
Major conclusions from METABRIC

- Recurrent copy number profiles can be used to stratify patients and identify molecular subgroups
- The subgroups co-segregated with prognostic profiles
- Typically, ‘driver’ alterations will be focal and (high) or (low) amplitude

Copy number profiles stratify endometrial cancer

Copy number profiles stratify endometrial cancer

Global properties of CNAs across the genome for companion diagnostics
PARP inhibitors induce synthetic lethality in BRCA deficient cells

<table>
<thead>
<tr>
<th>Gene X</th>
<th>Gene Y</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>No effect</td>
</tr>
<tr>
<td>−</td>
<td>+</td>
<td>No effect</td>
</tr>
<tr>
<td>+</td>
<td>−</td>
<td>No effect</td>
</tr>
<tr>
<td>−</td>
<td>−</td>
<td>Death</td>
</tr>
</tbody>
</table>

Alan Ashworth J Clin Oncol 26:3785-3790.

BRCA testing as a reflection of compromised homologous recombination
Patterns of LOH predict response to PARP inhibitors

Measurement technologies and computational methods
Measurement technologies for copy number analysis

- Fluorescence in situ hybridization, BAC arrays, genotyping arrays, whole genome shotgun sequencing

<table>
<thead>
<tr>
<th>Tech:</th>
<th>FISH</th>
<th>Array CGH</th>
<th>Genotype arrays</th>
<th>WGSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>&lt;10</td>
<td>30-100K</td>
<td>100K-2M</td>
<td>3G!</td>
</tr>
</tbody>
</table>

Resolution

The challenge of statistical inference of biological events from cancer samples

- Normal contamination
  - dilution of tumour-specific signals
- Intra-tumoural heterogeneity
  - clonal populations of cells with different genomes
  - most experimental designs consist of a single sample from a tumour
- Inference of somatic aberrations as distinct from germline polymorphisms
- Tumour ploidy influences allele-specific signals

- Assumptions in most statistical software packages ignore at least one of these issues
- Software and statistical models designed for analysis of normal genomes do not generalise to the cancer setting
- Specialised tools for cancer are needed
A review of statistical considerations

• Reference:
P. Neuviel, H. Bengtsson and T. P. Speed, Statistical analysis of Single Nucleotide Polymorphism microarrays in cancer studies, HHS Lu, B Schölkopf, H Zhao (Eds.), Handbook of Statistical Bioinformatics, Springer, 2011 (pp. 225--255). ISBN: 978-3-642-16344-9 (Print) 978-3-642-16345-6 (Online)

Genotypes and alleles

<table>
<thead>
<tr>
<th>State($K$)</th>
<th>Total copy number ($c$)</th>
<th>Genotype ($G$)</th>
<th>Zygosity Status ($Z/S$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$K_2$</td>
<td>A/AA</td>
<td>LOH</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>AB</td>
<td>HET</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>B/BB</td>
<td>LOH</td>
</tr>
<tr>
<td>4</td>
<td>$K_3$</td>
<td>AAA</td>
<td>LOH</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>AAB</td>
<td>HET</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>ABB</td>
<td>HET</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>BBB</td>
<td>LOH</td>
</tr>
<tr>
<td>8</td>
<td>$K_4$</td>
<td>AAAA</td>
<td>LOH</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>AAAB</td>
<td>ASCNA</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>AABB</td>
<td>HET</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>ABBB</td>
<td>ASCNA</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>BBBB</td>
<td>LOH</td>
</tr>
<tr>
<td>13</td>
<td>$K_5$</td>
<td>AAAAA</td>
<td>LOH</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>AAAAB</td>
<td>ASCNA</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>AAABB</td>
<td>HET</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>ABBBB</td>
<td>HET</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>ABBBB</td>
<td>ASCNA</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>BBBBB</td>
<td>LOH</td>
</tr>
</tbody>
</table>

Gavin Ha et al Genome Research 2012
Workflow for high density genotyping array analysis

- Data generation (CEL files)
- Preprocessing and normalisation
- Total copy number extraction
- B allele extraction
- Segmentation for breakpoints and CNAs
- Segmentation for LOH and ASCNA
- Gene and pathway analysis/clinical correlations

Affymetrix SNP 6.0

- Specifications
  - 25-mer oligonucleotide probes
  - 900K SNP probes
  - 900K CNV probes
  - hybridization intensities
  - Chip definition file:
General preprocessing

- Normalisation is required to remove platform-induced artefacts
- Method of choice for Affy SNP 6.0: *aroma.affymetrix*
  
  - [http://www.aroma-project.org/vignettes/CRMAv2](http://www.aroma-project.org/vignettes/CRMAv2)
  - Outperforms commercial software and is transparent and open source
  - Outputs allele-specific and total copy number real-valued data

Inference of genomic features

- Total copy number (CNA)
- Loss of heterozygosity (LOH)
- Allele specific copy number (ASCNA)
### From signal processing to allelic imbalance

![Images of signal processing and allelic imbalance](image1.png)

Neuvial, Bengsston, Speed BMC Bioinformatics (2010)

### Summary of methods

<table>
<thead>
<tr>
<th>Name</th>
<th>Settings</th>
<th>Detection method</th>
<th>Ploidy</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP [8]</td>
<td>unpaired</td>
<td>HMM on $\delta$</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Gardina [11]</td>
<td>unpaired</td>
<td>HMM on &quot;genotypes&quot;</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>BAFsegmentation [6]</td>
<td>paired or unpaired</td>
<td>segmentation of $\delta$</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>SOMATICs [7]</td>
<td>unpaired</td>
<td>segmentation of $\delta$</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>AsCNAR/CNAG [36]</td>
<td>unpaired</td>
<td>HMM on $\delta$</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>OverUnder [65]</td>
<td>unpaired</td>
<td>$2 \times 1d$ smoothing</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>PSCBS [66]</td>
<td>paired</td>
<td>two-way segmentation</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>GAP [67]</td>
<td>unpaired</td>
<td>$2 \times 1d$ segmentation</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Lamy [68]</td>
<td>paired</td>
<td>HMM on $(\gamma, \delta)$</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>PSCN [69]</td>
<td>unpaired</td>
<td>HMM on $(\gamma, \delta)$</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>PICNIC [43]</td>
<td>unpaired</td>
<td>HMM on $(\gamma, \delta)$</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>genoCNA [70]</td>
<td>unpaired</td>
<td>HMM on $(\gamma, \delta)$</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>OncoSNP</td>
<td>paired/unpaired</td>
<td>HMM</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Neuvial, Bengsston and Speed, 2011

---

Yau et al Genome Biology 2010
Next Generation Sequencing

Point mutation • Indel • Homozygous deletion • Hemi-zygous deletion • Gain • Translocation breakpoint

Copy number alterations


Analysis of NGS data

• Library construction methods introduce bias

Lai and Shah, unpublished
Effect of preprocessing

![Graph showing Effect of preprocessing]

HMMCopy, Bioconductor

Segmentation of NGS data

![Graph showing Segmentation of NGS data]

HMMCopy: http://compbio.bccrc.ca/software/hmmcopy
In the case of LOH, since B-alleles are not always on the same coverage, and the CNV test can be performed as described above. Of all exons in a segment constitutes the segment's depth. Reads are independent of each other, the sum of depth-to-search for larger CNV and LOH. In each segment, composed of arbitrary number of individual exons, we extended our method above to call CNV/LOH as sample size and the probability can be modeled by a binomial distribution with depth-coverage ratio and the frequency (BAF) of polymorphic positions in the sequenced exome sequence data using EMM copyWGSS copy number segments.

Module 5

LOH analysis from NGS

OncoSNP
SNP8 array data

APOLLOH
+SC +CN +SP

HMMcopy
WGSS copy number segments

Model Features
SC = Spatial Correlation
CN = Copy Number Aware
SP = Stromal Parameter

Module 5

CN analysis from exomes?

Deletion 0.97 0.86
Duplication 0.92 0.88
LOH 0.88 0.68

ExomeCNV:
Sathirapongsasuti et al Bioinformatics 2011
CN analysis from exomes?

Control FreeC: Boeva et al Bioinformatics 2011

From measurements to inference
Lower prevalence events emit weaker statistical signals

Ha et al Genome Research (2014) doi:10.1101/gr.180281.114

TITAN workflow

Ha et al Genome Research (2014) doi:10.1101/gr.180281.114
**Benchmark dataset**

Ha et al. Genome Research (2014)
doi:10.1101/gr.180281.114

---

**TITAN: comparing performance**

Ha et al. Genome Research (2014)
doi:10.1101/gr.180281.114
Copy number profiles to identify clonally mixed samples

Inferring clonal genotypes from bulk: deconvolution of mixtures

McPherson, Roth et al. Nature Genetics (2016)
Chromothripsis

- Chromosome ‘shattering’ followed by non-homologous end-joining

Properties of the genome in addition to traditional gene-based biomarkers may be clinically useful
Emergent technologies: single cell sequencing

Summary

- The genome architecture is a fundamentally important aspect of studying the cancer genome
- Somatic copy number alterations change gene dosage and can drive expression of oncogenes and/or tumour suppressors
- Copy number alterations can be measured using array based hybridization and/or next generation sequencing
- Properties of the genome revealed through the copy number profile can indicate important phenotypic characteristics of cancers
- Copy number profiles can be used as clonal marks to reveal distinct clonal populations within a bulk sample
Tools

- CRMAv2: http://www.aroma-project.org/vignettes/CRMAv2
- HMM-Dosage: http://compbio.bccrc.ca/software/hmm-dosage/
- PICNIC: http://www.sanger.ac.uk/genetics/CGP/Software/PICNIC/
- OncoSNP: https://sites.google.com/site/oncosnp/
- HMMCopy: http://compbio.bccrc.ca/software/hmmcopy/
- Apolloh: http://compbio.bccrc.ca/software/apolloh/
- Control Free-C: http://bioinfo-out.curie.fr/projects/freec/