CS4220: Knowledge Discovery Methods for Bioinformatics
Unit 7: Pathway Perturbations in a Disease Context

Niranjan Nagarajan
Identifying causal genes for a disease

- Genetic polymorphism and mutation calling
- Basic association analysis
- Causal genes in Cancer
- Challenges in identifying causal genes in Cancer
- Pathways and Integrated Analysis

Genetic Polymorphism and Mutation Calling
• DNA – Sequence of As, Cs, Gs and Ts (~3 billion in human genome)
• Reference Genome – Reference sequence for a species
Polymorphism

Vs

Somatic Mutation

Human genome is Diploid
Heterozygous vs Homozygous
Classes of Genomic Variations (GVs)

Single Nucleotide Variations (SNVs)

Short Insertions and Deletions (Indels) <100bp

Copy Number Variations (CNVs) >100bp

Structural Variations (SVs)
Ways to detect GVs

• **Microarrays**
  – Hybridization Based
  – SNP array, CGH array
  – Probes for common polymorphisms, >1 million
  – Distributed “evenly” over the genome

• **Sequencing**
  – Directly “read” DNA sequence
  – Compare to reference genome
    Or Reconstruct sample genome *de novo*
Affymetrix GeneChip Array
Using Arrays to calls SNPs

- **SNPs** = Single Nucleotide Polymorphisms i.e. common variant positions with two “alleles”

- **Information:** A allele and B allele intensity

- **Normalize Data**

- **Assign to Cluster**

Image Source: Lamy et al, NAR 2006
Using arrays to call CNVs

- **BAF** = “B Allele Frequency” i.e. normalized measure of relative signal of the B and A alleles
- **RR** = “Log R Ratio” i.e. total intensity (normalized)

# Sequencing Technologies

<table>
<thead>
<tr>
<th></th>
<th>Read Length</th>
<th>Time/Run</th>
<th>Cost/Mbp</th>
<th>Sequence (Mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger</td>
<td>~700 bp</td>
<td>&lt;1 day</td>
<td>$1000</td>
<td>2</td>
</tr>
<tr>
<td>454</td>
<td>~500 bp</td>
<td>&lt;1 day</td>
<td>&lt;$100</td>
<td>~500</td>
</tr>
<tr>
<td>Illumina</td>
<td>~100 bp</td>
<td>~1 week</td>
<td>&lt;$3</td>
<td>&gt;100,000</td>
</tr>
<tr>
<td>SOLiD</td>
<td>~50 bp</td>
<td>~1 week</td>
<td>&lt;$5</td>
<td>&gt;100,000</td>
</tr>
</tbody>
</table>

Sanger Sequencing (1977) - Gel based, Chain Termination
454 Sequencing (2004) - Chip, Pyrosequencing
State of the art in Sequencing

http://www.youtube.com/watch?v=v8p4ph2MAvI
Workflow for calling SNVs

Read Mapping

Date Cleaning - Read Alignment, Removing Duplicates, Recalibration

SNV Calling

A = 7/10, C=3/10 ??
SNV Calling (General Idea)

Bayes Rule: Genotype = AA|AB|BB

\[
P(\text{Genotype}|\text{Data}) = \frac{P(\text{Data}|\text{Genotype}) \cdot P(\text{Genotype})}{\sum_{\text{Genotype}} P(\text{Data}|\text{Genotype}) \cdot P(\text{Genotype})}
\]

Maximize Posterior \hspace{2cm} Likelihood \hspace{2cm} Prior

Base Quality → Probability of sequencing error

Q_{20} \rightarrow \text{prob. of error} = 10^{(-20/10)} = 10^{-2} = 0.01
Q_{30} \rightarrow \text{prob. of error} = ??

Prior = SNV every ~1000 bases for humans
Rare SNV calling

• What if the variant has allele frequency << 0.5?

If average error-rate is 1% can variants at 1% frequency be discovered?
Ruling out Sequencing Error

• Null Hypothesis – variant bases are from sequencing errors

• Null Model
  – If all base qualities are the same?
  – With different base qualities?
    • Binomial(n, p) =>$\rightarrow$ Poisson-Binomial(n, $<p_1, \ldots, p_n>$)

• P-value = probability k variant bases or more under null model

$$P_n(X=k) = P_{n-1}(X=k)(1-p_n) + P_{n-1}(X=k-1) p_n$$

Sensitivity is only limited by Quality and Coverage
Performance improves for high frequency SNVs as well

Real data, mixed *in silico*

SNVer and Breseq do not fully exploit base quality values
Rare variants can be experimentally validated

Fluidigm Digital PCR

LoFreq – 9/9
Breseq – 7/9
SNVer – 2/9
Calling Somatic Mutations

- Ad hoc
- Joint
- Rare mutations

Cancer

\[ \begin{align*}
\text{A} & \quad \text{C} \\
\text{A} & \\
\end{align*} \]

Normal

\[ \begin{align*}
\text{A} & \\
\text{A} & \\
\end{align*} \]

\[ \begin{align*}
\text{A} & = 7/10, \quad \text{C} = 3/10 \\
\end{align*} \]

Cancer

\[ \begin{align*}
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{C} & \\
\text{A} & \\
\text{A} & \\
\text{C} & \\
\text{A} & \\
\text{C} & \\
\text{A} & \\
\end{align*} \]

\[ \begin{align*}
\text{A} & = 7/10, \quad \text{C} = 3/10 \\
\end{align*} \]

Normal

\[ \begin{align*}
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\text{A} & \\
\end{align*} \]

\[ \begin{align*}
\text{A} & = 7/10, \quad \text{C} = 1/10 \\
\end{align*} \]
Ad hoc

1. Call SNVs in Cancer
2. Call SNVs in Normal
3. Filter Cancer list using Normal list
4. Remove SNVs where Normal has >1 base of that kind

Software: JointSNVMix

- Hypothesis: simultaneous analysis will result in better detection of shared signals (SNPs or technical noise) and weak signals for somatic mutations

<table>
<thead>
<tr>
<th>Reference Genome</th>
<th>Joint Genotype Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTCCGTCGGAACGAATGCCACG</td>
<td>( P(G_t^j</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( g_N \backslash g_T )</th>
<th>AA</th>
<th>AB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.01</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>AB</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>BB</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

- LOH = Loss of Heterozygosity
Joint SNVmix Model

Model parameters and latent variables trained using EM

## Results

<table>
<thead>
<tr>
<th>Caller</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>F-meas</th>
<th>MCC</th>
<th>FP Germlines</th>
<th>FP Wild-types</th>
</tr>
</thead>
<tbody>
<tr>
<td>JointSNVMix1 (Trained)</td>
<td>140</td>
<td>13</td>
<td>999788</td>
<td>59</td>
<td>0.795</td>
<td>0.802</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>JointSNVMix1</td>
<td>153</td>
<td>50</td>
<td>999751</td>
<td>46</td>
<td>0.761</td>
<td>0.761</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>SNVMix1 (Trained)</td>
<td>190</td>
<td>823</td>
<td>998978</td>
<td>9</td>
<td>0.314</td>
<td>0.423</td>
<td>743</td>
<td>70</td>
</tr>
<tr>
<td>SNVMix1</td>
<td>178</td>
<td>1653</td>
<td>998148</td>
<td>21</td>
<td>0.175</td>
<td>0.295</td>
<td>1632</td>
<td>0</td>
</tr>
</tbody>
</table>
Software: MuTect

1. Cancers can have a heterogeneous mixture of cells

2. Sample might also have normal cells

=> Mutations need not have 50% frequency if they are heterozygous
MuTect Algorithm

1. Call SNVs in Cancer aggressively
2. Filter artifacts
3. Filter potential germline SNVs aggressively

\[ L(M^m_f) = \text{Likelihood of having a mutation at frequency } f \]
\[ L(M_0) = \text{Same as above with } f = 0 \text{ i.e. no mutation} \]
\[ P(m,f) = \text{Probability of a mutation} \]
• Call SNVs in Cancer aggressively

\[
L(M_f^m) = P(\{b_i\} | \{e_i\}, r, m, f) = \prod_{i=1}^{d} P(b_i | e_i, r, m, f) = \begin{cases} 
  f^{e_i/3} + (1 - f)(1 - e_i) & \text{if } b_i = r \\
  f(1 - e_i) + (1 - f)^{e_i/3} & \text{if } b_i = m \\
  e_i/3 & \text{otherwise}
\end{cases}
\]

Really checking if \(\log(L(M_f^m)/L(M_0)) \geq 6.3\)

• Filter potential germline SNVs aggressively

Remove positions with bases in normal having \(\geq 2\) observations of the alternate allele or \(\geq 3\%\) of the reads and sum of their quality scores being \(> 20\)
High Sensitivity

Low False Positive Rate?
Comparison to other Methods

- MuTect is more sensitive for rare mutations
Summary

• GVs vary in size and impact on the genome
• Microarrays and Sequencing can be used to detect GVs with corresponding tradeoffs
• Model-based approaches are extremely effective at calling SNPs and somatic mutations from sequencing data
• Rare somatic mutations can be called without sacrificing precision
Must Read


Good to Read


Basic Association Analysis
Basic Idea behind Association Analysis

Odds Ratio =
Frequency in Case / Frequency in Control
(e.g. 52.6/44.6 = 1.17)

Statistical Test
\(\chi^2\) test (1-degree of freedom)

Correct for multiple-hypothesis testing

WTCCC, Nature 2007

Genome-wide Association Study (GWAS)

Manhattan Plot

Nitric oxide

K+ Channel Q

Na+ Channel

K+ Channel H

K+ Channel E
### GWAS CENTRAL

#### Search studies using keywords

Enter study ID, authors, pubmed ID, other terms | p-value threshold | Markers per Study
--- | --- | ---
| | -log p ≥ 0 | All

**Example:** HGVSST176, Todd JA, 21738484, metabolic

**Search**

---

### 1,008 Studies in the database containing data with -log p ≥ 0 (showing 1-20)

**Pages:** 1 2 3 4 5 6 7 8 9 10 >>

<table>
<thead>
<tr>
<th>Order results by</th>
<th>No. per page</th>
<th>Export these results as</th>
<th>Related data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date Created (descending)</td>
<td>20</td>
<td>--choose a format--</td>
<td>Related data</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Name</th>
<th>Phenotype(s)</th>
<th>Total p-values</th>
<th>Related citations</th>
<th>Add data sets to Browser</th>
<th>Related data</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGVSST1012</td>
<td>GWAS of bipolar disorder</td>
<td>Bipolar disorder (dominant) &lt;br&gt; Bipolar disorder (recessive) &lt;br&gt; Bipolar disorder (additive)</td>
<td>11</td>
<td>Jiang Y et al. Hindorff LA et al.</td>
<td><a href="#">Add</a></td>
<td>Results Chromosomes</td>
</tr>
<tr>
<td>HGVSST1013</td>
<td>GWAS of prion diseases</td>
<td>Prion diseases</td>
<td>11</td>
<td>Mead S et al. Hindorff LA et al.</td>
<td><a href="#">Add</a></td>
<td>Results Chromosomes</td>
</tr>
<tr>
<td>HGVSST1014</td>
<td>GWAS of atopic dermatitis</td>
<td>Atopic dermatitis</td>
<td>6</td>
<td>Paternoster L et al. Hindorff LA et al.</td>
<td><a href="#">Add</a></td>
<td>Results Chromosomes</td>
</tr>
<tr>
<td>HGVSST1003</td>
<td>GWAS of ankle-brachial index</td>
<td>Ankle-brachial index</td>
<td>2</td>
<td>Murabito JM et al. Hindorff LA et al.</td>
<td><a href="#">Add</a></td>
<td>Results Chromosomes</td>
</tr>
<tr>
<td>HGVSST1004</td>
<td>GWAS of mean platelet volume</td>
<td>Platelet counts &lt;br&gt; Mean platelet volume</td>
<td>36</td>
<td>Gieger C et al. Hindorff LA et al.</td>
<td><a href="#">Add</a></td>
<td>Results Chromosomes</td>
</tr>
<tr>
<td>HGVSST1005</td>
<td>GWAS of IgA nephropathy</td>
<td>IgA nephropathy</td>
<td>9</td>
<td>Yu XQ et al. Hindorff LA et al.</td>
<td><a href="#">Add</a></td>
<td>Results Chromosomes</td>
</tr>
</tbody>
</table>

---

**www.gwascentral.org**

> 1000 studies!
Challenges

1. What if the SNVs are not “common”?
2. What if the association is not to a SNV?
3. What if the impact of the SNV (“effect size”, odds ratio) is small?
4. Are the controls appropriate?
Frequency and Effect Size

• **SNP is not on array?**
  – Typically in “linkage” with a SNP that is...
Are the controls appropriate?

Do cases and controls have some obvious differences that could explain things?

Alternately: Is the SNP associated with other confounding factors?

- Sex
- Age
- Geographical or historical populations (Population Stratification)
Genes mirror Geography

Software: Eigenstrat

Based on PCA

1. Properties?
2. Adjust for variation
   \[ g_{ij} \leftarrow g_{ij} - \gamma_i a_j \]
   \[ \gamma_i = \sum_j a_j g_{ij} \]
3. Do association analysis with adjusted data

Example for Genotype Adjustment with Eigenstrat

• Suppose principal component 1 (PC1) is perfectly correlated with $g_i$ s.t. $g_{ij}=0$ if $a_j=0.1$ and $g_{ij}=1$ if $a_j=-0.1$

• Let $a_j=0.1$ for 50 out of 100 samples and -0.1 otherwise

• Then $\gamma_i = 50*0.1*0+50*-0.1*1 = -5$

• For $a_j=0.1$, adjusted $g_{ij}= 0-0.1*-5 = 0.5$

• For $a_j=-0.1$, adjusted $g_{ij}= 1-(-0.1*-5) = 0.5$

Thus the impact of ancestry is cancelled out from the genotypes values!
Pathways and GWAS

- Same idea as expression analysis
  - Pathways can help identify meaningful collections of genes

- [Wang et al 2007] Modified GSEA algorithm based on using $\chi^2$ scores with a Kolmogorov-Smirnov statistic

<table>
<thead>
<tr>
<th>Gene Set or Pathway</th>
<th>Set Size</th>
<th>Rank of Most Significant SNP</th>
<th>Nominal $P$</th>
<th>FDR $P$</th>
<th>FWER $P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDP-glycotransferase activity</td>
<td>89</td>
<td>59</td>
<td>$&lt;.001$</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>O-glycan biosynthesis</td>
<td>23</td>
<td>69</td>
<td>$&lt;.001$</td>
<td>.011</td>
<td>.018</td>
</tr>
</tbody>
</table>

NOTE.—Two pathways (database identifiers: GO accession number 0008194 and KEGG accession number hsa00512) demonstrate statistical significance after adjustment for multiple testing, but the SNPs and genes within these pathways cannot be detected by the most-significant SNPs/genes approach.

PD = Parkinson’s Disease
Good To Read


Glossary

Allele – An alternative form of the gene
Diploid – Carries 2 copies of each chromosome
Germline – In cells that can give rise to offspring
Heterozygous – Alleles are different
Homozygous – Alleles are same
Mutation – Change in nucleotide sequence
Polymorphism – Common variant of a gene
SNV – Single-nucleotide Variant
Somatic – Not in germline cells
Variant – Differs from the reference genome
Causal Genes in Cancer
Cancer is **not** a Single Disease

- **Classification**
  - Typically by the type of cells and the presumed origin of the cancer
    - Lung (small-cell, non-small-cell)
    - Breast (ductal, lobular)
    - Leukemia (acute, chronic, lymphoblastic, mylogenous)
  - Perturbed Pathways

- **Staging**
  - I, II, III, IV …
Scale of Genomic Changes in Cancer

- >10,000 point mutations and indels
- 100s of CNVs
- Merging, splitting of chromosomes

Image Source: Nagarajan et al, Genome Biology (2012)
Oncogenes and Tumor Suppressor Genes

Oncogenes – potential to cause cancer when “activated” (e.g. WNT, MYC, RAS)

Tumor Suppressor Genes (TSGs) – “protects” a cell from cancer s.t. inactivation leads to cancer (e.g. TP53, PTEN, APC)

Image Credit: www.cancer.gov
Hallmarks of Cancer

Many different normal processes are hijacked and altered in Cancer

Complex Interactions

Image Source: KEGG
Experimental Approaches

1. **Transfect** a gene in to over-express it
2. **Knock a gene down**

**Drawbacks**
- Artificial cell-line specific information
- Time-consuming

Image Source: www.genegnews.com
### Frequently Mutated Genes (Gastric Cancer)

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Name</th>
<th>Length</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53</td>
<td>cellular tumor antigen p53 isoform b</td>
<td>1182</td>
<td>50%</td>
</tr>
<tr>
<td>PTEN</td>
<td>phosphatidylinositol-3,4,5-trisphosphate</td>
<td>1212</td>
<td>18%</td>
</tr>
<tr>
<td>AQP7</td>
<td>aquaporin-7</td>
<td>1029</td>
<td>10%</td>
</tr>
<tr>
<td>ACVR2A</td>
<td>activin receptor type-2A precursor</td>
<td>1542</td>
<td>10%</td>
</tr>
<tr>
<td>STA2</td>
<td>double-stranded RNA-binding protein Staufen</td>
<td>1713</td>
<td>10%</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>catenin beta-1</td>
<td>2346</td>
<td>10%</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>phosphatidylinositol-4,5-bisphosphate 3-kinase</td>
<td>3207</td>
<td>13%</td>
</tr>
<tr>
<td>TTK</td>
<td>dual specificity protein kinase TTK isoform 1</td>
<td>2574</td>
<td>10%</td>
</tr>
<tr>
<td>COPB2</td>
<td>coatamer subunit beta'</td>
<td>2721</td>
<td>10%</td>
</tr>
<tr>
<td>DHX36</td>
<td>probable ATP dependent RNA helicase DHX36</td>
<td>3027</td>
<td>10%</td>
</tr>
</tbody>
</table>

Nagarajan et al, Genome Biology (2012)
Challenges in Identifying Driver Genes in Cancer
Mutations are not Unbiased

- Mutations by different “mutagens” have different biases
- For e.g. in Gastric Cancer, C>T mutations are common in genes ...
- ... and specifically in CpG, GpC motifs

Nagarajan et al, Genome Biology (2012)
Cancer Subtypes

- Different drivers for each subtype
- Expression clustering to define subtypes

Breast Cancer
Patient-specific Drivers?

• Every patient has a unique complement of mutations

• Even a single tumor may have several different sub-populations …
Integrated Analysis
What have we learned?

- Cancers are **heterogeneous** in terms of mechanism of origin
- Driver changes often hide in a sea of “passenger” mutations
- Frequently mutated **genes** can provide hints for potential drivers
- Integration of genomic and transcriptional information is needed
Pathways and Integrated Analysis
Integration Approaches

1. Mutations in Network (HOTNET)
2. CNVs + Expression (CONEXIC)
3. Mutations/CNVs + Expression in Network (DriverNet)
HOTNET Idea

Find sub-networks that are frequently mutated

- Hubs affect connectivity of graph
- Compute “influence” between all pairs of nodes
  - A influences B if there are few and short paths between them
  - Modelled as a “diffusion process”

Finding a Sub-Network

Method I (Combinatorial)

i. Threshold on Influence

ii. Find sub-network of size K that maximizes # of mutated samples

Method II (Enhanced Influence)

i. Weight edges by Influence and # of mutated samples

ii. Threshold on weight and reported connected components
Statistical Significance

Null Model: Permute mutations or gene labels

Statistical Testing:

Problem - Large search-space of sub-networks $\Rightarrow$
Correction for multiple hypothesis testing might be too stringent

Solution - Two-step procedure

Step I: Is it possible to get r sub-networks of size K by chance?

Step II: If all r sub-networks are reported is the FDR low?

Theorem 3. Fix $\beta_1, \beta_2, \ldots, \beta_K$ such that $\sum_{i=1}^{K} \beta_i = \beta$. Let $s^*$ be the first $s$ such that $\tilde{r}_s \geq \frac{E[r_{s+1}]}{\beta_{s+1}}$. If we return as significant all connected components of size $\geq s^*$, then the FDR of the test is bounded by $\beta$. 
Alternate Proof for Theorem 3

- If \( \tilde{r}_s \geq \frac{E[r_s]}{\beta_s} \), \( \beta_s \geq \frac{E[r_s]}{\tilde{r}_s} \geq FDR \) for component of size \( s \).

- So overall FDR is bounded by \( \sum_s \beta_s = \beta \)
## Results I

### Model I – Combinatorial

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$k$</th>
<th>samples</th>
<th>$H_0^{sample}$ p-val</th>
<th>$H_0^{gene}$ p-val</th>
<th>Pathway enrichment p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM</td>
<td>10</td>
<td>67</td>
<td>$&lt; 10^{-10}$</td>
<td>$4 \times 10^{-3}$</td>
<td>3 $\times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>78</td>
<td>$&lt; 10^{-10}$</td>
<td>$&lt; 10^{-3}$</td>
<td>10$^{-5}$</td>
</tr>
<tr>
<td>Lung</td>
<td>10</td>
<td>140</td>
<td>$&lt; 10^{-10}$</td>
<td>0.02</td>
<td>8 $\times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>151</td>
<td>$&lt; 10^{-10}$</td>
<td>0.03</td>
<td>3 $\times 10^{-3}$</td>
</tr>
</tbody>
</table>

### Model II – Enhanced Influence

<table>
<thead>
<tr>
<th>$s$</th>
<th># c.c.</th>
<th>$H_0^{sample}$ $\mu$ p-val</th>
<th>$H_0^{gene}$ $\mu$ p-val</th>
<th>Pathway enrichment p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15</td>
<td>22.18 0.97</td>
<td>13.63 0.38</td>
<td>/</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>6.37   0.98</td>
<td>4.38 0.6</td>
<td>/</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>$&lt; 10^{-3}$ $&lt; 10^{-3}$</td>
<td>0.07 $&lt; 10^{-3}$</td>
<td>0.9</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>$&lt; 10^{-3}$ $&lt; 10^{-3}$</td>
<td>0.05 0.05</td>
<td>4 $\times 10^{-6}$</td>
</tr>
</tbody>
</table>
Results II

Combination of Network and Frequency Identifies Rare Pathways

MAPK pathway

Notch signaling pathway
CONEXIC Idea


The Genomic Signature of a Driver

Assumptions on driver mutations:

A. A driver mutation should occur in multiple tumors more often than would be expected by chance

B. A driver mutation may be associated with the expression of a group of genes that form a ‘module’

C. Copy number variations often influence the expression of genes in the module via changes in expression of the driver

Slide: Anja Kiesel
Algorithm I

1. Selection of Candidate Drivers:
   - GISTIC algorithm to identify genes that overlap CNV regions often
     - 513 peak genes in 27 amplified regions
     - 384 peak genes in 23 deleted regions

2. Expression Filtering:
   - Remove genes that are expressed at constant level or not expressed
     - Final set of 428 genes

3. Single Modulator Step (Initial Model):
   - Correlation between CNV and expression
     - 347 candidate drivers left

   - Associating target genes with driver gene
     - 78 modulators explaining behavior of 4018 genes (min. 20 genes per module)

Slide: Anja Kiesel
Algorithm II

4. *Network Learning Step (Iteratively alternating 2 tasks):*
   
a) Learn regulation program by choosing candidate driver, that best splits gene expression of the module genes into 2 distinct behavior
b) Re-assign each gene into the module the best models its behavior

64 modulators explaining behavior of 7896 genes (of **7981 genes in total**)

Slide: Anja Kiesel
CONEXIC Results

Many Modulators are Involved in Pathways Related to Melanoma

LitVAn = Literature Vector Analysis

Searches for overrepresented terms in papers associated with genes in a gene set (manually curated database - NCBI Gene)

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Pathway</th>
<th>Band</th>
<th>Genes in Region</th>
<th>Validation p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MITF</td>
<td>Melanoma</td>
<td>3p14.2-p14.1</td>
<td>1</td>
<td>&lt;10^{-9}</td>
</tr>
<tr>
<td>TBC1D16</td>
<td>Vesicular Trafficking</td>
<td>17q25.3</td>
<td>24</td>
<td>&lt;10^{-6}</td>
</tr>
<tr>
<td>ZFP106</td>
<td>Insulin/Ras</td>
<td>15q15.1</td>
<td>7</td>
<td>&lt;10^{-6}</td>
</tr>
<tr>
<td>DIXDC1</td>
<td>Wnt/JNK/Pi3K</td>
<td>11q23.1</td>
<td>17</td>
<td>0.0001</td>
</tr>
<tr>
<td>OIP5</td>
<td>Cell Cycle</td>
<td>15q15.1</td>
<td>13</td>
<td>&lt;10^{-6}</td>
</tr>
<tr>
<td>TTBK2</td>
<td></td>
<td>15q15.2</td>
<td>7</td>
<td>0.0383</td>
</tr>
<tr>
<td>TRAF3</td>
<td>NFkappaB/JNK</td>
<td>14q32.32</td>
<td>19</td>
<td>0.0121</td>
</tr>
<tr>
<td>RAB27A</td>
<td>Vesicular Trafficking</td>
<td>15q15-q21.1</td>
<td>33</td>
<td>&lt;10^{-9}</td>
</tr>
<tr>
<td>C12orf35</td>
<td></td>
<td>12p11.21</td>
<td>45</td>
<td>&lt;10^{-9}</td>
</tr>
<tr>
<td>WBP2</td>
<td></td>
<td>17q25</td>
<td>92</td>
<td>0.0275</td>
</tr>
<tr>
<td>MOCS3</td>
<td></td>
<td>20q13.13</td>
<td>16</td>
<td>&lt;10^{-6}</td>
</tr>
<tr>
<td>NDUFB2</td>
<td></td>
<td>7q34</td>
<td>10</td>
<td>&lt;10^{-6}</td>
</tr>
<tr>
<td>ST6GALNAC2</td>
<td></td>
<td>17q25.1</td>
<td>92</td>
<td>&lt;10^{-6}</td>
</tr>
<tr>
<td>GRB2</td>
<td>EGFR/Ras</td>
<td>17q24-q25</td>
<td>92</td>
<td>0.1373</td>
</tr>
<tr>
<td>ECM1</td>
<td></td>
<td>1q21</td>
<td>55</td>
<td>0.0083</td>
</tr>
<tr>
<td>KCNC1</td>
<td></td>
<td>20q13</td>
<td>16</td>
<td>0.202</td>
</tr>
<tr>
<td>DPM1</td>
<td></td>
<td>20p13.13</td>
<td>16</td>
<td>0.097</td>
</tr>
<tr>
<td>PFKP</td>
<td>Metabolism</td>
<td>10p15.3-p15.2</td>
<td>3</td>
<td>0.0801</td>
</tr>
<tr>
<td>KLF6</td>
<td>Cell cycle, c-JUN (JNK)</td>
<td>10p15</td>
<td>3</td>
<td>&lt;10^{-9}</td>
</tr>
<tr>
<td>TIMM8B</td>
<td>Mitochondria</td>
<td>11q23.1-q23.2</td>
<td>17</td>
<td>0.7622</td>
</tr>
</tbody>
</table>

Slide: Anja Kiesel
Results II

A known driver, MITF, is correctly associated with target genes

- MITF expression correlates with targets better than copy number (A,B)
- MITF correctly annotated with its known role in melanoma (C)

2 types of melanoma: **high** MITF expression => proliferation
**low** MITF expression => invasion

Slide: Anja Kiesel
DriverNet Idea


Rare drivers can be discovered by their “influence” on genes with outlying expression levels

Overrepresented pathways of genes exhibiting outlying expression associated with EGFR high level amplification
- PI3K signaling: FOXO1, FOXO3, AKT1, AKT3, GNA10, SOS1, PIK3R5, PIK3R1, GSK3B, JAK1
- MAPK signaling: PDGFA, P53, AKT1, MAP2K7, AKT3, EGFR, BRAF, FLNA, JUN, PPM1A, KRAS, ELK4, SOS1, JUN, NFATC4, RASA2, TAO1, RPS6KA2, RAP1A, GACNA1H
- ErbB signaling pathway: BTC, STAT5B, AKT1, MAP2K7, AKT3, EGFR, BRAF, JUN, SRC, KRAS, SOS1, CAMK2D, PIK3R5, PIK3R1, PLCG1, GSK3B

A

B

FGF11

PIK3R1

PRKACB

Gene Expression
Algorithm

Min. number of mutated genes that explain the max. number of differentially expressed genes
Statistical Significance

Null Model

Permute entries of the mutation and differential expression matrices (not gene or sample based)

Statistical Testing

Is the number of covered genes observed rarely in 500 random instances?

Benjamini-Hochberg correction for multiple-hypothesis testing
Results I

Higher Concordance to Known Drivers

A

B

C

D

E

F

Concordance with Cancer Gene Census

Concordance with COSMIC

Top N Ranked Genes

Fisher-based Approach
Frequency-based Approach
DriverNet
# Results II

## Rare Drivers are Predicted and Meaningful

### Table 2 The predicted rare drivers

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gene</th>
<th>Gband</th>
<th>SNV/Indel</th>
<th>HLAMP</th>
<th>HOMD</th>
<th>Corrected P value</th>
<th>Percent altered</th>
</tr>
</thead>
<tbody>
<tr>
<td>METABRIC</td>
<td>PTEN</td>
<td>10q23.31</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>1.54</td>
</tr>
<tr>
<td>METABRIC</td>
<td>RB1</td>
<td>13q14.2</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>1.54</td>
</tr>
<tr>
<td>METABRIC</td>
<td>MDM2</td>
<td>12q15</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>1.06</td>
</tr>
<tr>
<td>METABRIC</td>
<td>MYC</td>
<td>8q24.21</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0.96</td>
</tr>
<tr>
<td>METABRIC</td>
<td>CDKN2A</td>
<td>9p21.3</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>1.54</td>
</tr>
<tr>
<td>METABRIC</td>
<td>CLTC</td>
<td>17q23.1</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1.54</td>
</tr>
<tr>
<td>METABRIC</td>
<td>CREBBP</td>
<td>16p13.3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.29</td>
</tr>
<tr>
<td>METABRIC</td>
<td>GNAS</td>
<td>20q13.32</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>METABRIC</td>
<td>EGFR</td>
<td>7p11.2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0.39</td>
</tr>
<tr>
<td>METABRIC</td>
<td>CDH1</td>
<td>16q22.1</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>1.54</td>
</tr>
<tr>
<td>METABRIC</td>
<td>CCNE1</td>
<td>19q12</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>METABRIC</td>
<td>EP300</td>
<td>22q13.2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0.39</td>
</tr>
<tr>
<td>METABRIC</td>
<td>CBL</td>
<td>11q23.3</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>1.25</td>
</tr>
<tr>
<td>METABRIC</td>
<td>PIK3R1</td>
<td>5q13.1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.00E-04</td>
<td>0.19</td>
</tr>
<tr>
<td>METABRIC</td>
<td>JAK2</td>
<td>9p24.1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>1.00E-04</td>
<td>0.67</td>
</tr>
<tr>
<td>METABRIC</td>
<td>TP53</td>
<td>17p13.1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>2.00E-04</td>
<td>0.67</td>
</tr>
<tr>
<td>METABRIC</td>
<td>NUP98</td>
<td>11p15.4</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0.0011</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Good To Read

- **[CONEXIC]** Akavia et al. “An Integrated Approach to Uncover Drivers of Cancer” *Cell 143*:1005-17, 2010
Acknowledgements

Denis Bertrand  Anja Kiesel

- Slides on association analysis were adapted from slides by Dr. Chiea Chuen Khor (Senior Research Scientist, Genome Institute of Singapore)