Estudi d’associació de mutacions genòmiques rares en la predisposició al càncer a partir de tècniques de seqüenciació de nova generació

Geòrgia Escaramís Babiano
Grup de Genòmica i Malaltia
Departament de Bioinformàtica i Genòmica
Centre de Regulació Genòmica

19 de Desembre de 2016
• Variation in Human Genome
• Next Generation Sequencing
• Cancer Predisposition Genes
• Rare Variant Association Study (RVAS) Pipeline/Methods
• RVAS in Chronic Lymphocytic Leukemia Project
• RVAS in Pan-Cancer of Whole Genomes Project
Variation in the Human Genome

- Human Genome sequence of billions of bases (letters) A, G, T, C

```
TAGTCATTAAATAACTCCTTTATTTCCGTTCCCTCTCCCCCTCAAATGGCTCATGTCACATCAAAGGCAAGGAAACATCTATGACCCCAACTATGAACATAGAAGCAGCTAATTCTGACTACATTGAACCCATGCTCAAAACCTTTATGC
```

- Distributed over 24 chromosomes, each of which contains between 45 and 280MB
Variation in the Human Genome

- **Chromosomal variants**
  Substantial changes in chromosome structure with strong effects on phenotype

- Changes in the number of chromosomes
  - **Polyploidy**: the presence of three or more complete sets of chromosomes
  - **Aneuploidy**: the presence of additional chromosomes or missing individual chromosomes

- Changes in the structure of the chromosomes
  - **Translocations**: a segment of one chromosome becomes attached to a non homologous chromosome
Variation in the Human Genome

- **Structural variants**
  - Copy Number Variant (CNV)
    - **Deletion**
    - Insertion
    - Duplication
  - Inversion
  - Usually complex regions with segmental duplications
Variation in the Human Genome

- **Structural variants**
  - Copy Number Variant (CNV)
    - Deletion
    - Insertion
    - **Duplication**
  - Inversion
  - Usually complex regions with segmental duplications
Variation in the Human Genome

- **Structural variants**
  - Copy Number Variant (CNV)
    - Deletion
    - Insertion
    - Duplication
  - Inversion
    - Usually complex regions with segmental duplications
 Variation in the Human Genome

- Relatively short variants
  - Single Nucleotide Polymorphism (SNP)
  - Insertions and deletions (indel)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Subject 2</th>
<th>Subject 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homozygote CC</td>
<td>Heterozygote CT</td>
<td>Homozygote TT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Maternal chromosome</th>
<th>Paternal chromosome</th>
<th>Maternal chromosome</th>
<th>Paternal chromosome</th>
</tr>
</thead>
</table>

Genotypic frequencies
(+ strand)                Genotypic frequencies
(- strand)                Allelic frequencies
(+ strand)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Allelic frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>60 indiv (60%)</td>
<td>C = (60*2 +30)/200 = 75% = 0.75</td>
</tr>
<tr>
<td>CT</td>
<td>30 indiv (30%)</td>
<td>T = (30 + 10*2)/200 = 25% = 0.25</td>
</tr>
<tr>
<td>TT</td>
<td>10 indiv (10%)</td>
<td>Minor allele frequency = MAF</td>
</tr>
</tbody>
</table>
Next Generation Sequencing

(a) DNA fragments can only be sequenced inwards from each end.

(b) Genome is assembled from overlapping, matching sequences.

1 kb = 1000 base pairs of DNA

10 kb

gap assembled region

http://www.nimr.mrc.ac.uk/mill-hill-essays/bringing-it-all-back-home-next-generation-sequencing-technology-and-you
Next Generation Sequencing

Genome of individual re-sequenced by aligning short reads against the reference genome.

Individual is homozygous ‘T’ at this A/T polymorphism.

Individual is heterozygous at this G/A polymorphism.
• Definition of cancer predisposition genes (CPG)

“genes in which rare mutations confer high or moderate risks of cancer (greater than two fold relative risks) and to those for which at least 5% of individuals with the relevant mutations develop cancer”
Rare variants have larger effect sizes than common variants

Cancer Predisposition Genes - Background

Manolio et al., Nature 2009
Cancer Predisposition Genes - Background

- Cancer Predisposition Genes

Rahman et al., Nature, 2014
Overlapping of Somatic and Germline Cancer Genes

- COSMIC database-468 somatically mutated genes
- Mutual integration of somatically and germline mutated genes, an useful approach for identification of new cancer genes.

Rahman et al., Nature, 2014
Approaches to identify cancer predisposition genes

- Candidate gene
- Genome-wide linkage/association studies
- Exome/Genome Sequencing

Rahman et al., Nature, 2014
Epidemiological Design

- Genes enriched in rare mutations in cases vs controls

![Diagram showing genetic comparison between cases and controls with functional impact highlighted.](chart.png)
Pipeline

Local Controls

Split is repeated N times (N permutations)

Dataset for AF estimates

CONTROLS

CASES

QC and Impact Filter

Outliers

Variant Classification

CADD impact factor score

Rare variant association study

Burden test (KBAC*)

Sequence Kernel Association Test* (SKAT-O)

Mixed Effects Model+ (Mist)

Susceptibility Genes

+ Sun et al. Genet Epidemiol. 2013
Association Methods

• **Common variant analysis**
  • Common variant: Minor allele frequencies (MAF) $\geq 5$
  • Using linkage disequilibrium (LD)

• **Rare variant analysis**
  • Rare variant: MAF $< 1\%$ (or 5%)
  • High allelic heterogeneity: collectively by multiple rare variants with moderate to high penetrance's
  • Associations through LD would not be suitable
Association Methods

**Burden test:** based on collapsing or summarizing the rare variants within a region by a single value, which is then tested for association with the trait of interest. Assumption: all variants have the similar effect magnitudes and direction.

**Heterogeneity effects test:** allow different variants within a region to have different directions and magnitude of effects, including no effects. Therefore enriched and depleted variants are well accommodated.

**Mixed-effects models:** model a set of variants within a region as a function of variant characteristics while allowing for variant-specific effect (heterogeneity). Unifies burden and heterogeneity in a single procedure.
Association Methods

AR4 (strong selection; only MAF<1% causal)
VE per gene = 1%

Moutsianas et al., PLoS Genetics, 2015
Association Methods (Kernel Based Adaptive Cluster)

- A burden test method
- $M+1$ mutation patterns ($G_0, G_1, \ldots, G_M$) across $k$ variants
- $n_{i1}$ and $n_{i0}$ cases and controls with mutation pattern $G_i$
- Test statistic:

\[
T = \left[ \sum_{i=1}^{M} \left( \frac{n_{i1}}{n_1} - \frac{n_{i0}}{n_0} \right) w_i \right]^2
\]

where $w_i$ is a kernel-based weight on each mutation pattern $G_i$

- Allows to account for mutation interaction

Leu & Leal, PLoS Genetics, 2010
Association Methods (Variance Component Tests)

- **Sequence Kernel Association Test Optimized (SKAT-O):**
  - $\beta_j$ random effects with mean 0 and variance $w_j \tau^2$
  - $H_0$: $\tau^2 = 0$

- **Mixed-effects Test (MiST):**
  - Two step hierarchical approach
  - $\delta_j$ random effects with mean 0 and variance $w_j \tau^2$
  - $H_0$: $\pi = 0$, $\tau^2 = 0$

\[
\log \text{it}(Y_i) = X_i^t \alpha + G_i^t \beta
\]

confounders sample-based matrix

\[
\beta_j = Z_j^t \pi + \delta_j
\]

confounders variant-based vector for $j$-th mutation effect
Association Methods (Our proposal)

- Generalized Linear Mixed Effects Model
  \[ \log it(Y_i) = X_i^\top \alpha + (G_i^\top Z)\pi + G_i^\top \delta \]
  - genotype vector for \( i \)-th individual
  - confounders variant-based vector for \( j \)-th mutation effect

- Estimated using Integrated Nested Laplace Approximation (INLA)
  - \( \delta_j \) random effects with mean 0 and variance \( w_j \tau^2 \)
  - \( H_0: \pi = 0, \tau^2 = 0 \)
RVAS analysis in Chronic Lymphocytic Leukemia (CLL) project
CLL – Dataset Description

- 437 CLL samples (Whole Exome Seq.)
  - Spanish population samples
  - 2 Kits
  - 3 subtypes (CLL, SLL, MBL)

- 780 Controls (Whole Exome Seq.)
  - Spanish population samples
  - 3 Kits
  - Samples from ~18 different projects
    (none of them is cancer study)

- Multisample call is done together on both, cases (CLL samples) and controls
CLL – QC Filtration

Number of mutations per patient

Count

Number of mutations per patient

Count
CLL – QC Filtration PCA
CLL – QC Filtration PCA
Mutations in genes (per 100 patients)

CLL – QC Filtration Check
A) 95 quantile of p-values < 0.05

B) 95 quantile of adjusted p-values < 0.05

CLL – Results for 3 methods
• Interacts directly with other 6 cancer risk genes
• It is involved in cell division cycle
• Has one publication showing association between CDC27 and risk in breast cancer

Association between polymorphisms in cdc27 and breast cancer in a Chinese population.

Guo H^1, Chen W, Ming J, Zhong R, Yi P, Zhu B, Miao X, Huang T.

Abstract
Cdc27, as a core component of anaphase-promoting complex (APC), is a cell cycle regulator, which participates in control of mitotic checkpoint and surveys the mitotic spindle to maintain chromosomal integrity. It was hypothesized that polymorphisms in cdc27 gene might contribute to the susceptibility of breast cancer (BC) through influencing the mitotic progression of cells. Therefore, a hospital-based case-control study with 463 BC patients and 536 controls was implemented to investigate the association of six single-nucleotide polymorphisms (SNPs) in cdc27 and BC risk in a Chinese Han population. Among the six SNPs, two SNPs of rs11570443 and rs12601027 were positively correlated with BC risk. Individuals carrying rs11570443-CT or CC genotypes showed a higher BC risk with the OR of 1.75 (95% confidence interval (CI) = 1.13-2.69), compared with those carrying rs11570443- TT genotype. For rs12601027, an increased BC risk was significantly associated with homozygote TT genotype (odds ratio (OR) = 1.49, 95% CI = 1.12-1.98) compared with homozygote CC and heterozygote CT genotypes. In addition, a significant interaction effect of these two SNPs was found. The rs12601027- TT in combination with rs11570443-CT/CC genotypes showed a strongly elevated risk of BC compared with rs12601027-CC/CT and rs11570443- TT genotype (OR = 1.95, 95% CI = 1.06-3.59). These findings suggested that polymorphisms in cdc27 may contribute to the susceptibility of BC though functional studies are needed to further elucidate the underlying mechanisms of the associations.
RVAS analysis in Pan-Cancer Analysis of Whole Genomes (PCAWG) project
PCAWG – Project Description

- 2818 **whole genomes** (normal tissues) of cancer patients across 47 worldwide cancer projects
• Identify cancer specific germline rare mutations

1. Selected regions (GWAS hits, Rahman genes, DNA repair genes) coding and regulatory variants above functional score threshold, grouped by gene.

2. Genome wide Rare Variant Association (coding and regulatory regions)
PCs identified on ~115k SNP set (MAF>1%)
PCAWG – QC Check in European Samples
PCAWG – QC Check in European Samples
<table>
<thead>
<tr>
<th>Tissue</th>
<th>Number</th>
<th>vs</th>
<th>Number</th>
<th>Tissue Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>141</td>
<td>vs</td>
<td>1738</td>
<td>non-ECTODERM</td>
</tr>
<tr>
<td>Medulloblastoma</td>
<td>174</td>
<td>vs</td>
<td>1551</td>
<td>non-NEURAL-CREST</td>
</tr>
<tr>
<td>Skin cancer</td>
<td>104</td>
<td>vs</td>
<td>1551</td>
<td>non-NEURAL-CREST</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>81</td>
<td>vs</td>
<td>1123</td>
<td>non-ENDODERM</td>
</tr>
<tr>
<td>Pancreatic cancer</td>
<td>275</td>
<td>vs</td>
<td>1123</td>
<td>non-ENDODERM</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>188</td>
<td>vs</td>
<td>1123</td>
<td>non-ENDODERM</td>
</tr>
<tr>
<td>CLL</td>
<td>95</td>
<td>vs</td>
<td>1225</td>
<td>non-MESODERM</td>
</tr>
<tr>
<td>Lymphoid</td>
<td>99</td>
<td>vs</td>
<td>1225</td>
<td>non-MESODERM</td>
</tr>
<tr>
<td>Kidney cancer</td>
<td>157</td>
<td>vs</td>
<td>1225</td>
<td>non-MESODERM</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>97</td>
<td>vs</td>
<td>1225</td>
<td>non-MESODERM</td>
</tr>
</tbody>
</table>
PCAWG – Results by Cancer Type in Europeans

Genes Protein Truncating Variants

Breast
$\lambda=0.96$

Ovary
$\lambda=0.83$
Genes with Missense Variants

Medulloblastoma
\[ \lambda = 1.28 \]

CLL
\[ \lambda = 1.16 \]
Genes with promoter variants using Funseq2 damaging score

PCAWG – Results by Cancer Type in Europeans
Genes with promoter variants using MDS damaging score

- **CLL**: $\lambda = 1.48$
- **Medulloblastoma**: $\lambda = 1.15$
- **Pancreas**: $\lambda = 1.23$
Population substructure: driven by variants more frequent in latino vs european (EXAC / gnomad database)
A Subway Map for Cancer Pathways

By Claudia Blentley
## Acknowledgments

### Genomics and Disease Group
- Estivill, Xavier
- Rabionet, Kelly
- Domenech, Laura
- Prasad, Aparna
- Holik, Aliaksei

### Genomics and Epigenomic Variation in Disease Group
- Ossowski, Stephan
- Susak, Hana
- Bosio, Mattia
- Muyas, Francesc