Principles of genomic data analysis

Joaquín Dopazo

Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Functional Genomics Node, (INB), Bioinformatics Group (CIBERER) and Medical Genome Project, Spain.

http://bioinfo.cipf.es
http://www.medicalgenomeproject.com
http://www.babelomics.org
http://www.hpc4g.org
@xdopazo

MLPM Summer School, Paris, September 18th, 2014
Background

The road of excess leads to the palace of wisdom

(William Blake, 28 November 1757 – 12 August 1827, poet, painter, and printmaker)

Biology has become a data-driven discipline. Biology is now bigger than physics, as measured by the size of budgets, by the size of the workforce, or by the output of major discoveries; and biology is likely to remain the biggest part of science through the twenty-first century.

The introduction and popularisation of high-throughput techniques offer the possibility of interrogating biological systems with an unprecedented level of detail.

But do we ask the proper questions…?
Looking for disease genes using a conventional testing schema: finding variants significantly over-represented in cases.
Candidate gene studies using GWAS

The promise in the 90s: “In 10 years we will unravel the genetic bases of complex diseases!!” ... And it seemed easy to fulfil...

**Odds Ratio:** 3.6
95% CI = 1.3 to 10.4

**Association Studies**

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Type 1</th>
<th>Controls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA DR4</td>
<td>17</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>NON-HLA DR4</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 5.377 \]

p < 0.025
Published Genome-Wide Associations

By the time of the completion of the human genome sequence, in 2005, just a few genetic variants were known to be significantly associated to diseases. When the first exhaustive catalogue of GWAS was compiled, in 2008, only three years later, more than 500 single nucleotide polymorphisms (SNPs) were associated to traits. Today such catalog had collected more than 1,900 papers reporting 14,012 SNPs significantly associated to more than 1,500 traits.

NHGRI GWA Catalog
www.genome.gov/GWASudies
Lessons learned from GWAS

• Many loci/variants contribute to complex-trait variation

• There is evidence for pleiotropy, i.e., that the same loci/variants are associated with multiple traits.

• Much of the heritability of the trait cannot be explained by the individual loci/variants found associated to the trait.

Visscher et al, 2013 AJHG
Where did the heritability go?

The missing heritability problem: individual genes cannot explain the heritability of traits

How to explain all this?
Rare Variants, rare CNVs, epigenetics or.. epistatic effects?
If rare variants eluded detection because were under represented among the SNPs, genomic sequencing would reveal them.

http://www.genome.gov/sequencingcosts/
Exome sequencing has been systematically used to identify Mendelian disease genes.
The principle: comparison of patients to reference controls or segregation within families

Segregation within a pedigree
Variant/gene prioritization by successive filtering

- **Variant level**
  - Potential impact of the variant
  - Population frequencies

- **Experimental design level**
  - Family(es)
  - Trios
  - Case / control

- **Functional (system) level**
  - Gene set
  - Network analysis
  - Pathway analysis

Control of sequencing errors (missing values)

Testing strategies
3-Methylglutaconic aciduria (3-MGA-uria) is a heterogeneous group of syndromes characterized by an increased excretion of 3-methylglutaconic and 3-methylglutaric acids. WES with a consecutive filter approach is enough to detect the new mutation in this case.
Use known variants and their population frequencies to filter out.

- Typically dbSNP, 1000 genomes and the 6515 exomes from the ESP are used as sources of population frequencies.
- We sequenced **300 healthy controls** (rigorously phenotyped) to add an extra filtering step to the analysis pipeline.

How important do you think local information is to detect disease genes?

Novembre et al., 2008. Genes mirror geography within Europe. Nature

Comparison of MGP controls to 1000g
Filtering with or without local variants

Number of genes as a function of individuals in the study of a dominant disease
Retinitis Pigmentosa autosomal dominant

The use of local variants makes an enormous difference
The CIBERER Exome Server (CES): the first repository of variability of the Spanish population

Only another similar initiative exists: the GoNL http://www.nlgenome.nl/

http://ciberer.es/bier/exome-server/
Information provided

- Genomic coordinates, variation, and gene.
- Genotypes in the different reference populations
- SNPId if any
Occurrence of pathological variants in "normal" population

Reference genome is mutated

Nine carriers in 1000 genomes

One affected and 73 carriers in EVS
Exome sequencing clearly finds mutations. Low rate of false negatives. An example with MTC

Dominant:
Heterozygotic in A and D
Homozygotic reference allele in B and C
Homozygotic reference allele in controls

The codon 634 mutation
Another example: RP X-linked. Gene RPGR
BiERapp: interactive web-based tool for easy candidate prioritization by successive filtering

SEQUENCING CENTER

- FASTQ
- No-SQL (Mongo) VCF indexing
- VCF
- BiERapp filters
- Population frequencies
- Consequence types
- Experimental design
- BAM viewer and Genomic context

Data preprocessing

BAM

Genome Maps
BiERapp: the interactive filtering tool for easy candidate prioritization

http://bierapp.babelomics.org
However, WES finds more variants than you’d like. High rate of false positives.

The promising variant (a frameshift present in all patients but not detected in controls) was nor real. It was not properly covered by reads in controls.
Consecutive filters approach. Exome sequencing produces many false positives

Average values obtained per exome (>800)

<table>
<thead>
<tr>
<th>Variant type</th>
<th>Mean number of variants (± sd) in African Americans</th>
<th>Mean number of variants (± sd) in European Americans</th>
<th>SNVs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Novel variants</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missense</td>
<td>303 (± 32)</td>
<td>192 (± 21)</td>
<td>60,000</td>
</tr>
<tr>
<td>Nonsense</td>
<td>5 (± 2)</td>
<td>5 (± 2)</td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>209 (± 26)</td>
<td>109 (± 16)</td>
<td></td>
</tr>
<tr>
<td>Splice</td>
<td>2 (± 1)</td>
<td>2 (± 1)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>520 (± 53)</td>
<td>307 (± 33)</td>
<td></td>
</tr>
<tr>
<td><strong>Non-novel variants</strong></td>
<td></td>
<td></td>
<td>30,000</td>
</tr>
<tr>
<td>Missense</td>
<td>10,828 (± 342)</td>
<td>9,319 (± 233)</td>
<td></td>
</tr>
<tr>
<td>Nonsense</td>
<td>98 (± 8)</td>
<td>89 (± 6)</td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>12,567 (± 416)</td>
<td>10,536 (± 280)</td>
<td></td>
</tr>
<tr>
<td>Splice</td>
<td>36 (± 4)</td>
<td>32 (± 3)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>23,529 (± 751)</td>
<td>19,976 (± 505)</td>
<td>5,000</td>
</tr>
<tr>
<td><strong>Total variants</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missense</td>
<td>11,131 (± 364)</td>
<td>9,511 (± 244)</td>
<td>150-300</td>
</tr>
<tr>
<td>Nonsense</td>
<td>103 (± 5)</td>
<td>93 (± 6)</td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>12,776 (± 434)</td>
<td>10,645 (± 286)</td>
<td></td>
</tr>
<tr>
<td>Splice</td>
<td>38 (± 5)</td>
<td>34 (± 4)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>24,049 (± 791)</td>
<td>20,283 (± 523)</td>
<td>&lt; 100</td>
</tr>
</tbody>
</table>

The table lists the mean number (± standard deviation) of novel and non-novel coding single nucleotide variants from 100 sampled African Americans and 100 European Americans. Non-novel variants refer to those found in dbSNP131 or in 200 other control exomes. Capture was performed using the Nimblegen V2 target. The analysis pipeline consisted of: alignment using the Burrows-Wheeler alignment tool; recalibration; realignment around insertion-deletions and merging with the Genome Analysis Toolkit (GATK)\(^1\); and removal of duplicates with PICARD. Variants were called using the following parameters: quality score > 50; allele balance ratio < 0.75; homopolymer run > 3; and quality by depth < 5. Variants were called from a RefSeq37.2 target (35,804,408 bp).


We can detect the disease mutation(s)… along with many other unrelated variants
And there are many real variants with potential phenotypic effect

Findings:
20,000 total variants
1,000 new variants
300-500 LOF variants (>50 homozygous)
100 known variants associated to disease

My first exome...

List of variants

<table>
<thead>
<tr>
<th>Variants</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRKCZ, LOC10</td>
<td>Amyotrophic Lateral Sclerosis (ALS)</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Parkinson’s disease</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Rheumatoid Arthritis</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>common polymorphism</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Multiple complex diseases—Crohn’s disease, combined control dataset</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Alzheimer’s Disease</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>LDL cholesterol</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Skin pigmentation</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Type 1 diabetes</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Coronary Artery Disease</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>in allele DQB1<em>0501 and allele DQB1</em>0502</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Multiple complex diseases—Bipolar disorder</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Multiple complex diseases—Coronary Artery Disease, gender differentiated</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Multiple complex diseases—Crohn’s disease, combined control dataset, gender differentiated</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Multiple complex diseases—Type 1 Diabetes, combined control dataset</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Multiple complex diseases—Type II Diabetes Mellitus, combined control dataset</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Systemic Lupus Erythematosus (SLE), gender differentiated in women</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Triglycerides</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>Type 1 Diabetes</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>893Ser—expressing (ABCB1:2677C&gt;T (A1a893Ser)) cells showed</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>47% lower intracellular digo...</td>
</tr>
<tr>
<td>PRKCZ, LOC10</td>
<td>A study in 336 recipients of hematopoietic—cell transplants...</td>
</tr>
</tbody>
</table>
A high level of deleterious variability exists in the human genome

- Variants predicted to severely affect the function of human protein coding genes known as loss-of-function (LOF) variants were thought:
  - To have a potential deleterious effect
  - To be associated to severe Mendelian disease

- However, an unexpectedly large number of LOF variants have been found in the genomes of apparently healthy individuals: 281-515 missense substitutions per individual, 40-85 of them in homozygous state and predicted to be highly damaging.

- A similar proportion was observed in miRNAs and possibly affect to any functional element in the genome.

Such apparently deleterious mutation must be first detected and then distinguished from real pathological mutations.
Moreover, even Mendelian genes can be elusive.

Intuitive belief: multiple family information should help

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Observation: this is not always true, not even in cases of Mendelian diseases
Is the single-gene approach realistic? Can we easily detect disease-related variants?

There are several problems:

a) Interrogating 60Mb sites (3000 Mb in genomes) produces too many variants. A large number of these segregating with our experimental design

b) There is a non-negligible amount of *apparently deleterious* variants that (apparently) has no pathologic effect

c) In many cases we are not targeting rare but *common* variants (which occur in normal population)

d) In many cases only one variant does not explain the disease but rather a *combination* of them (epistasis)

e) Consequently, the few individual variants found associated to the disease usually account for a *small portion* of the trait heritability
Is the heritability missing or are we looking at the wrong place?

How to explain missing heritability? Rare Variants, rare CNVs, epigenetics or.. *epistatic effects*?

At the end, most of the heritability was there...

---

**Table 1 | Estimates of heritability and number of loci for several complex traits**

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of loci</th>
<th>Proportion of heritability explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-related macular degeneration</td>
<td>5</td>
<td>50%</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>32</td>
<td>20%</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>6</td>
<td>15%</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>18</td>
<td>6%</td>
</tr>
<tr>
<td>HDL-cholesterol</td>
<td>7</td>
<td>5%</td>
</tr>
<tr>
<td>Height</td>
<td>40</td>
<td>5%</td>
</tr>
<tr>
<td>Early onset myocardial infarction</td>
<td>9</td>
<td>3%</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>4</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

*Residual is after adjustment for age, gender, diabetes.

---

**Common SNPs explain a large proportion of the heritability for human height**

Jian Yang1, Bebén Benyamin1, Brian P McEvoy1, Scott Gordon1, Anjali K Henders1, Dale R Nyholt1, Pamela A Madden2, Andrew C Heath3, Nicholas G Martin3, Grant W Montgomery3, Michael E Goddard3 & Peter M Visscher1

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the missing heritability? We estimated the proportion of variance in human height explained by 294,831 SNPs genotyped in 3,925 unrelated individuals using a linear mixed model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously, thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

GWASs in human populations have discovered hundreds of SNPs of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium with the genotyped SNPs. This has led to the hypothesis that the heritability of complex traits is accounted for at only a small fraction of the loci detected by GWASs, whereas the remaining heritability has been estimated to be 'missing' (e.g., 2-5%). Minor variants that cause extreme short or tall stature have been found14,15, but these do not explain much of the variation in the general population. Recent GWASs on tens of thousands of individuals have detected ~50 variants that are associated with height in the population, but these in total account for only ~5% of phenotypic variance16–19. Data from a GWAS that are collected to detect statistical associations between SNPs and complex traits are usually analyzed by testing each...
At the crossroad: how detection power of genomic technologies can be increased?

There are two (non mutually exclusive) ways:

**Scaling up:** by increasing sample size.
It is known that larger size allows detecting more individual gene (biomarker) associations.

**Changing the perspective:** systems approach to understand variation
Interactions, multigenicity can be better detected and the role of variants understood in the context of disease mechanism.

**Limitations:** Budget, patients availability and the own nature of the disease.
Modular nature of human genetic diseases

- With the development of **systems biology**, studies have shown that phenotypically similar diseases are often caused by **functionally related genes**, being referred to as the **modular nature of human genetic diseases** (Oti and Brunner, 2007; Oti et al, 2008).

- This modularity suggests that **causative genes** for the same or phenotypically similar diseases may generally reside in the same **biological module**, either a **protein complex** (Lage et al, 2007), a **sub-network** of protein interactions (Lim et al, 2006), or a **pathway** (Wood et al, 2007).
Initiatives to explore beyond single gene biomarkers

LINCS aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents, and by using computational tools to integrate this diverse information into a comprehensive view of normal and disease states that can be applied for the development of new biomarkers and therapeutics.

And the connectivity map, a pioneer initiative
An approach inspired on systems biology can help in detecting causal genes

Affected cases in complex diseases will be a heterogeneous population with different mutations (or combinations).

Many cases and controls are needed to obtain significant associations.

The only common element is the (know or unknown) pathway affected.

Disease understood as the failure of a functional module
Gene Ontology are labels to genes that describe, by means of a controlled vocabulary (ontology), the functional role(s) played by the genes in the cell. A set of genes sharing a GO annotation can be considered a functional module.
An example of GWAS

GWAS in Breast Cancer.

The CGEMS initiative. (Hunter et al. Nat Genet 2007)

1145 cases 1142 controls. Affy 500K

Conventional association test reports only 4 SNPs significantly mapping only on one gene: FGFR2

Conclusions: conventional SNP-based or gene-based tests are not providing much resolution.
Breast Cancer

CGEMS initiative. (Hunter et al. Nat Genet 2007)

1145 cases 1142 controls. Affy 500K

Only 4 SNPs were significantly associated, mapping only in one gene: 
FGFR2

PBA reveals 19 GO categories including regulation of signal transduction (FDR-adjusted p-value=4.45x10^{-03}) in which FGFR2 is included.

Bonifaci et al., BMC Medical Genomics 2008; Medina et al., 2009 NAR
GO processes significantly associated to breast cancer

- Metastasis
- Chromosomal instability
- Rho pathway
From gene-based to function-based perspective

SNPs, Gene expression

<table>
<thead>
<tr>
<th></th>
<th>SNPs, gene exp.</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection power</td>
<td>Low (only very prevalent genes)</td>
<td>high</td>
</tr>
<tr>
<td>Annotations available</td>
<td>many</td>
<td>many</td>
</tr>
<tr>
<td>Use</td>
<td>Biomarker</td>
<td>Illustrative, give hints</td>
</tr>
</tbody>
</table>
Can the interactome help to find disease mutations?

Disease genes are close in the interactome
Goh 2007 PNAS

Cancer genes are central.
Hernandez, 2007 BMC Genomics

Deleterious mutations in 1000g (up) and somatic CLL deleterious mutations (down)
Garcia-Alonso 2014 Mol Syst Biol
The role of interactome in buffering the deleteriousness of LoF mutations

Comparison of the interactome damage between real and random individuals after removing the nodes corresponding to proteins containing deleterious variants in both alleles (homozygote). Two different scenarios are simulated: Simulated populations with uniform probability, where proteins are randomly removed, and Simulated populations with observed frequencies, where proteins are removed with a probability proportional to the frequency of variation in the 1000 genomes population.

Garcia-Alonso 2014 Mol Syst Biol
From gene-based to function-based perspective

Using protein interaction networks as a scaffold to interpret the genomic data in a functionally-derived context.

Gene integrity

AND/OR

Gene activity

What part of the interactome is active and/or is damaged
Example with Inherited Retinal Dystrophies

- Prevalence 1 in 3000
- Clinically and genetically very heterogeneous
- 190 GENES account for approx. 50% of IRDs.

Novel variants and genes remain to be discovered
Network analysis.
Is genetic overlapping among IRDs related to protein interaction?

LCA - Leber Congenital Amaurosis
CORD/COD - Cone and cone-rod dystro.
CVD - Colour Vision Defects
MD - Macular Degeneration
ERVR/EVR - Erosive and Exudative Vitreoretinopathies
USH - Usher Syndrome
RP - Retinitis Pigmentosa
NB - Night Blindness
BBS - Bardet-Biedl Syndrome
Connectivity among IRDs explain genetic overlap

Significant Clustering coefficient, p-value=0.0103

SNOW Tool. Minguez et al., NAR 2009 Implemented in Babelomics (http://www.babelomics.org)
Network enrichment methods


Network analysis in one of its first proposals.

An objective function (discrimination potential) is used to evaluate a growing network built up around a seed (known genes, most differentially expressed genes, etc.)
Mutation network analysis

This approach prioritizes well-known, infrequently mutated genes, which are shown to interact with highly recurrently mutated genes yet have been ignored by conventional single-gene-based approaches.

These observations suggest cancer genes could mutate in a broad range of frequency spectrums, making it difficult for the frequency-based filtering approach to be effective.


BRAF mutated in 76 samples

MAP2K2 mutated in 4 samples

BRAF and MAP2K2 interact in 80 samples
Network analysis helps to find disease genes in complex diseases

SNPs validated in independent cohorts

CHRNA7 (rs2175886 p = 0.000607)
IQGAP2 (rs950643 p = 0.0003585)
DLC1 (rs1454947 p = 0.007526)
RASGEF1A* (rs1254964 p = 3.856x10^{-05})

*no interactions known (yet)
Programs to detect the most significant sub-network within a list of genes

Network analysis and the interactive, web-based network viewer CellMaps

http://cellmaps.babelomics.org/
From gene-based to function-based perspective

<table>
<thead>
<tr>
<th></th>
<th>SNPs, gene expression, etc.</th>
<th>GO</th>
<th>Protein interaction networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection power</td>
<td>Low (only very prevalent genes)</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Information coverage</td>
<td>Almost all</td>
<td>Almost all</td>
<td>Less (~9000 genes in human)</td>
</tr>
<tr>
<td>Use</td>
<td>Biomarker</td>
<td>Illustrative, give hints</td>
<td>Biomarker*</td>
</tr>
</tbody>
</table>

*Need of extra information (e.g. GO) to provide functional insights in the findings
From gene-based to mechanism-based perspective

Transforming gene expression values into another value that accounts for a function. Easiest example of modeling function: signaling pathways. Function: transmission of a signal from a receptor to an effector.

Activations and repressions occur.

<table>
<thead>
<tr>
<th></th>
<th>ProtH</th>
<th>ProtN</th>
<th>ProtR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProtA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ProtI</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ProtQ</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

function: Action A, Action B, Action C
Inferring the functional effect of gene expression changes in signaling pathways

Patricia Sebastián-León¹, José Carbonell¹, Francisco Salavert¹,², Rubén Sanchez³,
Ignacio Medina¹ and Joaquín Dópazo¹,²,⁴,*

¹Department of Computational Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012, Spain, ²CIBER de Enfermedades Paras (CIBERER), Valencia 46012, Spain, ³Genometra S.L., Valencia, Spain and ⁴Functional Genomics Node (INB) at CIPF, Valencia 46012, Spain

Received March 3, 2013; Revised April 18, 2013; Accepted May 2, 2013

Activation

\[
P(A \rightarrow G \text{ activated}) = P(A)P(B)P(D)P(F)P(G) + P(A)P(C)P(E)P(F)P(G) - P(A)P(F)P(G)P(B)P(C)P(D)P(E)
\]

Prob. = P(A activated)P(B activated)

Inhibition

\[
\text{Prob.} = [1 - P(A \text{ activated})]P(B \text{ activated})
\]

Sub-pathway

P(A→G activated) = P(A)P(B)P(D)P(F)P(G) + P(A)P(C)P(E)P(F)P(G) - P(A)P(F)P(G)P(B)P(C)P(D)P(E)
Modeling pathways

We only need to estimate the probabilities of gene activation and then calculate the probability for each circuit of being active.

Gene expression

A large dataset of affymetrix microarrays (10,000) is used to adjust a mixture of distributions of gene activity for all the genes.

Then, the activation state of any gene from a new microarray can be calculated as a probability:

\[
P(x, \text{Activated}) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}
\]

And the probability for each circuit of the pathway of being active can be calculated as well:

\[
P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} P\left(\bigcap_{k=1}^{n} A_k\right)
\]
Obtaining probability distributions for ALL the probes in the microarray

A large dataset of affymetrix microarrays (10,000) is used to adjust a mixture of distributions of gene activity for all the probes.
Using probability distributions to estimate gene activation probabilities

Then, the activation state of any probe from a new microarray can be calculated as a probability:

$$P_{\text{probe}} = \begin{array}{cccc}
0.001 & 0.89 & 0.5 & \ldots & 0.4 \\
\end{array}$$

Finally, gene activation probabilities are summarized from their corresponding probes as the 90% percentil value (to avoid outliers)
Gene activation probabilities are transformed into signal transduction probabilities

And the probability of being **active** for each circuit of each pathway can be calculated as well:

\[
P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} P\left(\bigcap_{k=1}^{n} A_k\right)
\]

We have transformed a physical genomic measure (gene expression) into a value that accounts for cell functionality
Sensitivity and specificity of model results

Simulated datasets with equal random gene expression values with noise added

Absence of false positives (low type I error)

Circuit activities used for disease class prediction. Accuracy is evaluated by 10-fold cross-validation. A good prediction is a proxy for sensitivity (low type II error)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>KNN</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.86-0.89</td>
<td>0.90-0.92</td>
<td>0.99</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>MCC</td>
<td>RMSC</td>
<td>AUC</td>
</tr>
<tr>
<td>MCC</td>
<td>0.74-0.79</td>
<td>0.81-0.83</td>
<td>0.98</td>
</tr>
<tr>
<td>RMSC</td>
<td>0.29-0.32</td>
<td>0.31</td>
<td>0.04</td>
</tr>
<tr>
<td>AUC</td>
<td>0.97-0.98</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.88-0.89</td>
<td>0.92-0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>AML</td>
<td>MCC</td>
<td>RMSC</td>
<td>AUC</td>
</tr>
<tr>
<td>MCC</td>
<td>0.76-0.78</td>
<td>0.84-0.90</td>
<td>0.92-0.93</td>
</tr>
<tr>
<td>RMSC</td>
<td>0.27-0.30</td>
<td>0.31</td>
<td>0.10-0.11</td>
</tr>
<tr>
<td>AUC</td>
<td>0.94</td>
<td>0.98</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Real dataset of pediatric acute myeloid leukemia with gene expression microarray data of 237 samples
What would you predict about the consequences of gene activity changes in the apoptosis pathway in a case control experiment of colorectal cancer?

The figure shows the gene up-regulations (red) and down-regulations (blue).
Apoptosis inhibition is not obvious from gene expression

Two of the three possible sub-pathways leading to apoptosis are inhibited in colorectal cancer. Upper panel shows the inhibited sub-pathways in blue. Lower panel shows the actual gene up-regulations (red) and down-regulations (blue) that justify this change in the activity of the sub-pathways.
Different pathways cross-talk to deregulate programmed death in Fanconi anemia

FA is a rare chromosome instability syndrome characterized by aplastic anemia and cancer and leukemia susceptibility. It has been proposed that disruption of the apoptotic control, a hallmark of FA, accounts for part of the phenotype of the disease.
Mechanisms-based biomarkers can also be used to predict features

Mechanism-based biomarkers (signaling circuits here) can be used to predict phenotypes that can be discrete (e.g. disease subtype) or even continuous (e.g. the IC50 of drugs, model animals from cell lines, etc.).
Prediction of IC$_{50}$ values from the activity of signaling circuits

\[ y = 2.3 + 0.8 \cdot x, \quad \text{adj.} r^2 = 0.481, \quad p - \text{value} = 1.6326\times10^{-99} \]
Mechanism-based biomarkers have meaning by themselves

Unlike in single gene biomarkers, the selected mechanism-based biomarkers (probability of circuit activation) have meaning by itself. Survival activation in the apoptosis pathway is one of the predictive mechanism-based biomarkers of breast cancer.
From gene-based to function-based perspective

<table>
<thead>
<tr>
<th></th>
<th>SNPs, gene expression, etc.</th>
<th>GO</th>
<th>Protein interaction networks</th>
<th>Models of cellular functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Detection power</strong></td>
<td>Low (only very prevalent genes)</td>
<td>High</td>
<td>High</td>
<td>Very high</td>
</tr>
<tr>
<td><strong>Information coverage</strong></td>
<td>Almost all</td>
<td>Almost all</td>
<td>Low (~9000 genes in human)</td>
<td>Low (~6700 genes in human)*</td>
</tr>
<tr>
<td><strong>Use</strong></td>
<td>Biomarker</td>
<td>Illustrative, give hints</td>
<td>Biomarker</td>
<td>Biomarker that explain disease mechanism</td>
</tr>
</tbody>
</table>

*Only ~800 genes in human signaling pathways*
Is individualized treatment a realistic option today?

An integrated view

Patient’s omic data
- Genomics and transcriptomics
- Epigenomics
- Proteomics
- Metabolomics

Systems biology computational models
- Network drugs
- Best combination

Biological knowledge
- Regulation
- Interaction
- Function
- Diagnostic biomarkers
- Personalized medicine

Cell culture → Drug treatment
- Personalized therapy
- Xenograft model

Therapeutic targets

Dopazo, 2003, Drug Discovery Today
In silico KOs or over-expressions

If we can calculate the effect of gene expression over signaling, we can simulate KOs (or over-expressions)

Colorectal cancer activates a signaling circuit of VEGF pathway that produces PGI2.
Virtual KO of COX2 interrupts the circuit (known therapeutic inhibitor in CGR)
Different virtual KOs predict different effects

JAK1 KO

CISH KO

STAT1 KO
**Distribution of Human Genetic Diseases**

Pablo Lapunzina, Personal communication

All genetic/genomic or epigenetic diseases with known cause:
~ 5000 disorders

<table>
<thead>
<tr>
<th>GENOMICS</th>
<th>GENETICS</th>
<th>EPIGENETICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Kb- ? Mb</td>
<td>1 bp- 200 bp</td>
<td>No dosage changes</td>
</tr>
<tr>
<td>8-12%</td>
<td>82-87%</td>
<td>2-3%</td>
</tr>
</tbody>
</table>

**Fact:** exons represent a comparatively small part of the complete genome

**Other fact:** there is still a lot of missing heritability
The ENCODE project suggests a functional role for a large fraction of the genome

Which percentage of the genome is occupied by:

- Coding genes: 2.4%
- TFBSs: 8.1%
- Open chromatin regions: 15.2%
- Different RNA types: 62.0%
- Total annotated elements: 80.4%

Exomes are only covering a small fraction of the potential functionality of the genome (2.4%).

Is the **missing heritability** hidden in the remaining 78%?
If so, what type of variant should be expect to discover? SNVs? SVs?
Future prospects

Future prospects

We need to efficiently query all the information contained in the genome, including all the epigenomic signatures.

This means data integration and “epistatic” queries

We need to prepare our health systems to deal with all the genomic data flood

<table>
<thead>
<tr>
<th>Information about variations</th>
<th>Processed</th>
<th>Raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome variant information (VCF)</td>
<td>150 MB</td>
<td>250 GB</td>
</tr>
<tr>
<td>Epigenome</td>
<td>150 MB</td>
<td>250 GB</td>
</tr>
<tr>
<td>Each transcriptome</td>
<td>20 MB</td>
<td>80 GB</td>
</tr>
<tr>
<td>Individual complete variability</td>
<td>400 MB</td>
<td>525 GB</td>
</tr>
<tr>
<td>Hospital (100.000 patients)</td>
<td>40 TB</td>
<td>50 PB</td>
</tr>
</tbody>
</table>

There are technical problems and conceptual problems on how genomic information is managed that must be addressed in the near future.
Software development

Babelomics is the third most cited tool for functional analysis. Includes more than 30 tools for advanced, systems-biology based data analysis.

More than 150,000 experiments were analyzed in our tools during the last year.

See interactive map of for the last 24h use http://bioinfo.cipf.es/toolsusage

Mapping
HPC on CPU, SSE4, GPUs on NGS data processing
Speedups up to 40X

Visualization
Ultrafast genome viewer with google technology

Variant annotation
Genome maps is now part of the ICGC data portal

CellBase
Knowledge database

Functional analysis
Percentage of the number of citations per tool (%)

Diagnostic
NGS panels

Signaling network

Regulatory network

Interaction network

Variant prioritization

Variant annotation

Mapping
HPC on CPU, SSE4, GPUs on NGS data processing
Speedups up to 40X

Visualization
Ultrafast genome viewer with google technology

Variant annotation
Genome maps is now part of the ICGC data portal

CellBase
Knowledge database
From data to knowledge
Contribution from the computational biology side
(from molecular biology to genomic/computational biology)

- **Technical contribution**: huge datasets need to be managed by computers. Technically useful, scientifically irrelevant.

- **Data analysis contribution**: bioinformatics is necessary to properly analyze *big data*. Of course p-values are necessary, but do not provide novelty. No *big data* vision, only analysis.

- **Science from the computational biology perspective**:
  - Do not use genomic data to simply scaling up conventional approaches. New ideas emerge when data are seen from the proper perspective.
  - *Big data*-driven questions and hypotheses can only be formulated when such data can be easily managed.
  - Niche disciplines: genomics and systems biology.
  - Most challenging problems: Complex diseases, disease mechanisms, biological processes (e.g. metabolism), drug action mechanisms, drug repositioning…

Scientific relevance
Babelomics in the Maniatis

The Babelomics suite of programs becomes a classic. Now is cited as a method in the last edition of *Molecular Cloning*, the popular *Maniatis*. The protocol 4 of chapter 8, Expression Profiling by Microarray and RNA-seq, contains a description on how to use Babelomics to analyze expression data.
The Computational Genomics Department at the Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, and...

...the INB, National Institute of Bioinformatics (Functional Genomics Node) and the CIBERER Network of Centers for Rare Diseases.