Understanding disease with *omic* data

Alejandro Cáceres
ISGlobal, Barcelona
Talk’s aim

How we can use omic data for understanding biological processes that are involved in disease

....in six examples
Types of omic data

- genomic
- transcriptomic
- methylomic
What is *omic* data?
What is *omic data*?

- Big data in biology
- High dimensional data (a lot of features) collected at different biological domains.
Biological Levels (orders of magnitude)
Omic data

Ome refers to the totality of elements in a biological domain: genome, proteome, ... interactome, phenome, exposome

Operational definition:

Omic data is an unbiased scan of variables that cover a given biological range.
Genomic data: unbiased scan of DNA sequence

At the level of chromosome molecules: genomic data
Genomic data: unbiased scan of DNA sequence

Definition:

- A **genomic variable** is the **presence** of a given DNA sequence from reference values (reference genome, hybridization probes)

**SNP:**

- ref: A T G C T G
- chr1: A T G C T T T

Property:

- The values are (almost) **stable** throughout an individual’s cells and life span
Transcritomic data: unbiased scan of RNA sequence

At the level of RNA molecules: transcriptomic data
Transcriptomic data: unbiased scan of RNA sequence

Definition:

- A **Transcript variable** is the **amount** of a given RNA sequence from reference values (reference genome, hybridization probes)

Property:

- The values are **highly dynamic** in time and are different for each **cell type** - snapshot of the cell at work in the nucleus
Methylomic data: unbiased scan of DNA methylated sites

At the level of DNA sequence: methylomic data
Methylomic data: unbiased scan of DNA methylated sites

Definition:
- A **Methylomic variable** is the **average state** of methylation at a given DNA site for the cells in a sample

Property:
- The values change in time according to the **individual’s development/age** and are different for each cell within a **cell type**
Measuring omic data

Sequencing + mapping

Hybridization
Omic data from different methods

Omic data based from sequencing:
+ collects all the possible information on an individual (maximum coverage)
+ is useful to detect rare variables (large effects)
– is computationally demanding

Omic data based from microarrays:
+ is highly scalable (100,000s of individuals)
+ is useful to detect small effects of variables on phenotypes
– is not unbiased
Understanding disease with *omic data
Method

1. Study how a biological process is imprinted on a given **omic** data
2. Develop a **method** to mine the **omic** data
3. Understand the role of the **biological process** in human **disease**
Examples

hidden structure in **omic** data
- inversion polymorphisms, asthma and obesity
- recombination substructure, breast cancer
Examples

interaction between **omic** variables

- epistasis, Alzheimer’s disease
- reliability of co-expression networks, evaluating networks across different tissues
- cosplicing, predicting genes’ physiological interactions
Examples

multi **omic** data integration

- Lost of chromosome Y, male susceptibility to disease.
Example 1
(hidden structure)
Studying inversion polymorphisms

Inversions are DNA sequences that run in the opposite direction of a reference sequence.

- important structural variants involved in adaptation and chromosomal evolution (chr Y)
- little studied in humans because they are difficult to measure in large cohorts
inversion imprint in genomic data

Genomic Data

vi vi+1
vj vj+1

s1
sn

Where?

(Caceres et al BMC Bioinformatics, 2012)

inveRsion

invClust

(Caceres et al NAR, 2015)

who?

inv-17q21

(Caceres et al NAR, 2015)
Detection and genotyping of inv-16p11
inversion 16p11

inv-16p11 is a risks factor for the cooccurrence of asthma and obesity (OMIM #615835)

(Gonzalez*, Caceres*, et al. AJHG, 2014, *fisrt joint author)
studying inversions with genomic data

Significance

- First hypothesis for the joint susceptibility to asthma and obesity
- Study of inversions in human populations using large cohorts
Example 2
(hidden structure)
Studying recombination

- increases genetic diversity
- different ancestries have different recombination patterns

Detection of **population substructure** is commonly based on **mutation** differences not on **allele combination** differences

can we detect allele combination substructure?
Recombination differences in genomic data

(Ruiz*, Caceres* et al submitted NAR, *first joint author)
Recombination substructure in 1q21.1

The recombination substructure at 1q21.1 associates with the risk of breast cancer
Studying recombination substructure with genomic data

Significance

- The causal variant in the susceptibility locus 1q21.1 to breast cancer may be a structural variant or process that suppressed recombination of the risk chromosomes with others.

- Recombination substructure (differential allele combinations) may help to explain additional heritability of complex diseases.
Example 3
(variable interaction)
Studying epistasis

- complex traits are likely to emerge from the interaction between genomic variables
- there are too many to test ($\sim 10^{13}$ possibilities)

Do the interactions of validated risk SNPs overlap?
Genome-wide association studies (Alzheimer’s Disease)

GWAS Genomic Data

Validated associations

APOE’s rs4420638
PICALM’s rs536841
MS4A6A’s rs610932
BIN1’s rs610932
...

Genomic Data

GWAS

Validated associations
Epistasis in genomic data

Genomic Data

Genome Wide Interaction

S1

Vi

Vj

AD

Vx*Vi

Vx*Vj

Sn

SNP × risk locus_1

SNP × risk locus_2

SNP × risk locus_3

...
Enrichment of epistatic effects

Pathway B is enriched in interactions with risk SNPs 1, 2, and 3

(Caceres et al, 2017 Alzheimer’s and Dementia)
Enrichment of epistatic effects in AD

Gonadotropin signaling is enriched in interactions with *APOE* and *MS4A6A*’s polymorphisms

<table>
<thead>
<tr>
<th>Risk Locus</th>
<th>Pathway</th>
<th>combined uncorrected p-value</th>
<th>combined corrected p-value</th>
<th>GENADA</th>
<th>NIA</th>
<th>ADG12</th>
<th>ADG31</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs429358 x</td>
<td>KEGG: GNRH SIGNALING PATHWAY</td>
<td>3.7e-5</td>
<td>0.01</td>
<td>0.025</td>
<td>0.033</td>
<td>0.001</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>KEGG: LONG TERM POTENTIATION</td>
<td>1.6e-5</td>
<td>0.02</td>
<td>0.145</td>
<td>0.001</td>
<td>0.001</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td>KEGG: ARRHYTHMOGENIC RIGHT VENTRICULAR CARDIOMYOPATHY</td>
<td>2.9e-5</td>
<td>0.04</td>
<td>0.001</td>
<td>0.146</td>
<td>0.001</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>KEGG: CALCIUM SIGNALING PATHWAY</td>
<td>1.1e-4</td>
<td>0.05</td>
<td>0.162</td>
<td>0.001</td>
<td>0.01</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td>KEGG: VASCULAR SMOOTH MUSCLE CONTRACTION</td>
<td>9.6e-5</td>
<td>0.08</td>
<td>0.222</td>
<td>0.002</td>
<td>0.001</td>
<td>0.265</td>
</tr>
<tr>
<td>rs610932 x</td>
<td>KEGG: PHOSPHATIDYLINOSITOL SIGNALING SYSTEM</td>
<td>1.0e-4</td>
<td>0.008</td>
<td>0.027</td>
<td>0.356</td>
<td>0.013</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>KEGG: DILATED CARDIOMYOPATHY</td>
<td>3.1e-7</td>
<td>0.01</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>BioCarta: HDAC PATHWAY</td>
<td>6.4e-6</td>
<td>0.02</td>
<td>0.002</td>
<td>0.016</td>
<td>0.147</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>KEGG: VASCULAR SMOOTH MUSCLE CONTRACTION</td>
<td>2.4e-7</td>
<td>0.03</td>
<td>0.001</td>
<td>0.001</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>KEGG: GNRH SIGNALING PATHWAY</td>
<td>5.6e-6</td>
<td>0.05</td>
<td>0.01</td>
<td>0.004</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>KEGG: HYPERTROPHIC CARDIOMYOPATHY HCM</td>
<td>1.0e-4</td>
<td>0.08</td>
<td>0.248</td>
<td>0.001</td>
<td>0.004</td>
<td>0.131</td>
</tr>
<tr>
<td></td>
<td>BioCarta: NKT PATHWAY</td>
<td>1.7e-4</td>
<td>0.11</td>
<td>0.022</td>
<td>0.001</td>
<td>0.201</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>KEGG: GALACTOSE METABOLISM</td>
<td>8.6e-5</td>
<td>0.12</td>
<td>0.789</td>
<td>0.002</td>
<td>0.005</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>BioCarta: PGC1A PATHWAY</td>
<td>1.6e-4</td>
<td>0.14</td>
<td>0.017</td>
<td>0.001</td>
<td>0.414</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>KEGG: LONG TERM DEPRESSION</td>
<td>1.4e-4</td>
<td>0.15</td>
<td>0.024</td>
<td>0.072</td>
<td>0.022</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Studying epistasis of risk variants with genomic data

Significance

- Clinical trials targeting the gonadotropin pathway should test *APOE* and *MS4A6A*'s polymorphisms for **response to treatment**.
- Epistasis helps to **link** risk SNPs by their interactions with common **biological processes** (join the dots of GWAS)
Example 4
(variable interaction)
Co-expression networks

- inform which genes are co-regulated, functional related or work together in the same pathway
- must be reproducible

Can we identify the tissues for which a network is functional?
Co-expression networks across multiple tissues

Trascriptomic Data

Co-expression

Reliability measure

\[ \lambda = \text{Probability that the diagonal terms are their row and column maxima} \]

6 Tissues \(\times\) 2 experiments

(Caceres et al. BMC genomics, under revision)
Inter-study reliability of networks across multiple tissues

Top agreement between BRAINEAC and GTEx across 4 brain regions in 287 KEGG pathways

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\sigma$</th>
<th>Ref</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.68</td>
<td>0.02</td>
<td>hsa05033</td>
<td>Nicotine addiction</td>
</tr>
<tr>
<td>0.67</td>
<td>0.04</td>
<td>hsa04720</td>
<td>Long-term potentiation</td>
</tr>
<tr>
<td>0.58</td>
<td>0.04</td>
<td>hsa05206</td>
<td>MicroRNAs in cancer</td>
</tr>
<tr>
<td>0.55</td>
<td>0.01</td>
<td>hsa04080</td>
<td>Neuroactive ligand-receptor int.</td>
</tr>
<tr>
<td>0.53</td>
<td>0.03</td>
<td>hsa04020</td>
<td>Calcium signaling pathway</td>
</tr>
<tr>
<td>0.52</td>
<td>0.03</td>
<td>hsa04261</td>
<td>Adrenergic sig. in cardiom.</td>
</tr>
<tr>
<td>0.51</td>
<td>0.02</td>
<td>hsa04912</td>
<td>GnRH signaling pathway</td>
</tr>
</tbody>
</table>
Nicotine addiction pathway across 4 brain regions
Studying network reliability with transcriptomic data

Significance

▶ the changes in **nicotine addiction** pathway are consistent across four brain regions with dopaminergic projections

▶ **inter-study reliability** of pathway changes across tissues can inform on the fraction of tissues with **specific functional changes** in network structure.
Example 5
(variable interaction)
Studying co-splicing

- Isoform ratios can correlate between two genes, across subjects

To which extent co-regulation of splicing can predict gene function?
Studying co-splicing with transcriptomic data

(Caceres et al, BMC genomics, 2018-accepted)
Physiological function of genes across multiple tissues

- work supported with computing hours from RES
Studying co-splicing with transcriptomic data

Significance

- **APP** is physiologically linked with genes affected in Alzheimer’s disease, supporting the hypothesis that a **loss of function** of **APP** contributes to the disease.
- Co-splicing is a common phenomena and should be taken into account to predict **gene function**.
Example 6
(multi omic data)
Studying loss of chromosome Y

- LOY associates with age and all-cause mortality in men (smoking, cancer and AD)
- We don't know whether LOY causes disease or vice-versa.

Can we predict a consequence of LOY that is closer to disease?
Detecting extreme deregulation of chromosome Y

(Caceres et al, final draft ready!)
LOY → EDY → Male Disease

EDY:
- associates with LOY-associated conditions (age, AD, cancer)
- strongly correlates with LOY
- improves the effect of LOY with male disease
Studying EDY with multiple omic data

LOY-EDY status in cancer samples from TCGA

Survival of 13 different types of cancer

Differential Methylation across Y

CNV Proportion

NoLOY/NoEDY: Ref
NoLOY/EDY: OR=2, P=0.002
LOY/NoEDY: OR=1.76, P=0.004
LOY/EDY: OR=1.23, P=0.086

NoLOY/NoEDY: N = 2152
NoLOY/EDY: N = 160
LOY/NoEDY: N = 301
LOY/EDY: N = 891

significant DMP in both comparisons
Studying EDY with multiple omic data

Significance

- We give first evidence of a likely path from LOY to disease
- EDY is a novel biomarker for male disease which can be triggered by multiple mechanisms including LOY
Further questions
Histone modification of EDY

What are the histone marks of EDY?

EDY is a protective factor for leukemia...
(control = 3112, cases = 800, OR = 0.08, $P = 5.3 \times 10^{-5}$)
Chromatin modification of inversions

What are the histone marks of inversions?
Machine learning for recombination substructure

Can we train a neural network to detect recombination substructures?
Thanks:

- Juan R Gonzalez - ISGlobal
- Luis Perez Jurado - UPF
- Mario Caceres - UAB
- Suzzane Sindi - University of California Merced
- Carlos Ruiz - ISGlobal
- Mariona Bustamante - ISGlobal
- Tonu Esko - University of Tartu