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

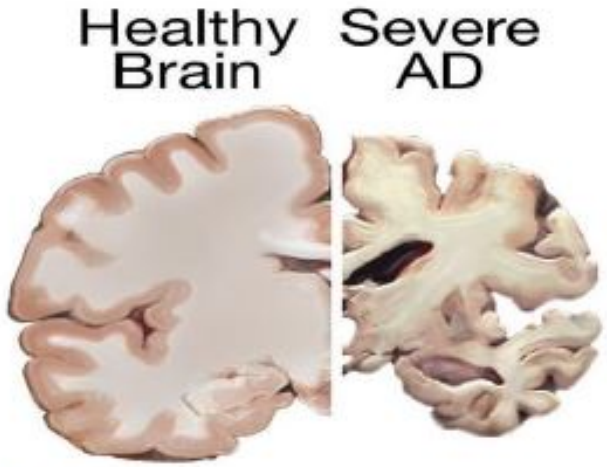
- genomic
- transcriptomic
- methylomic

What is *omic* data ?

What is *omic* data ?

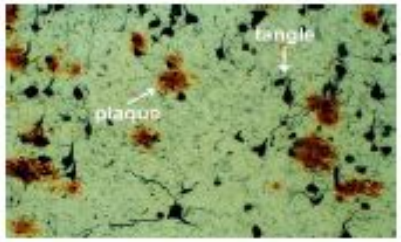
- Big data in biology
- High dimensional data (a lot of features) collected at different biological domains.

Biological Levels (orders of magnitude)



1m (1e 0m)

1cm (e-2m)

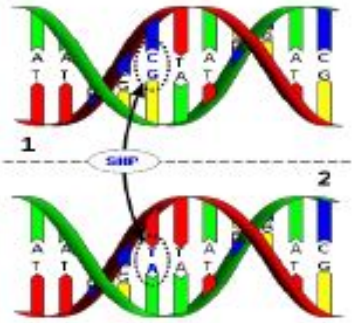


0.1mm (e-4m)

1micro (e-6m)

1nm (e-9m)

700pc (e-10m)

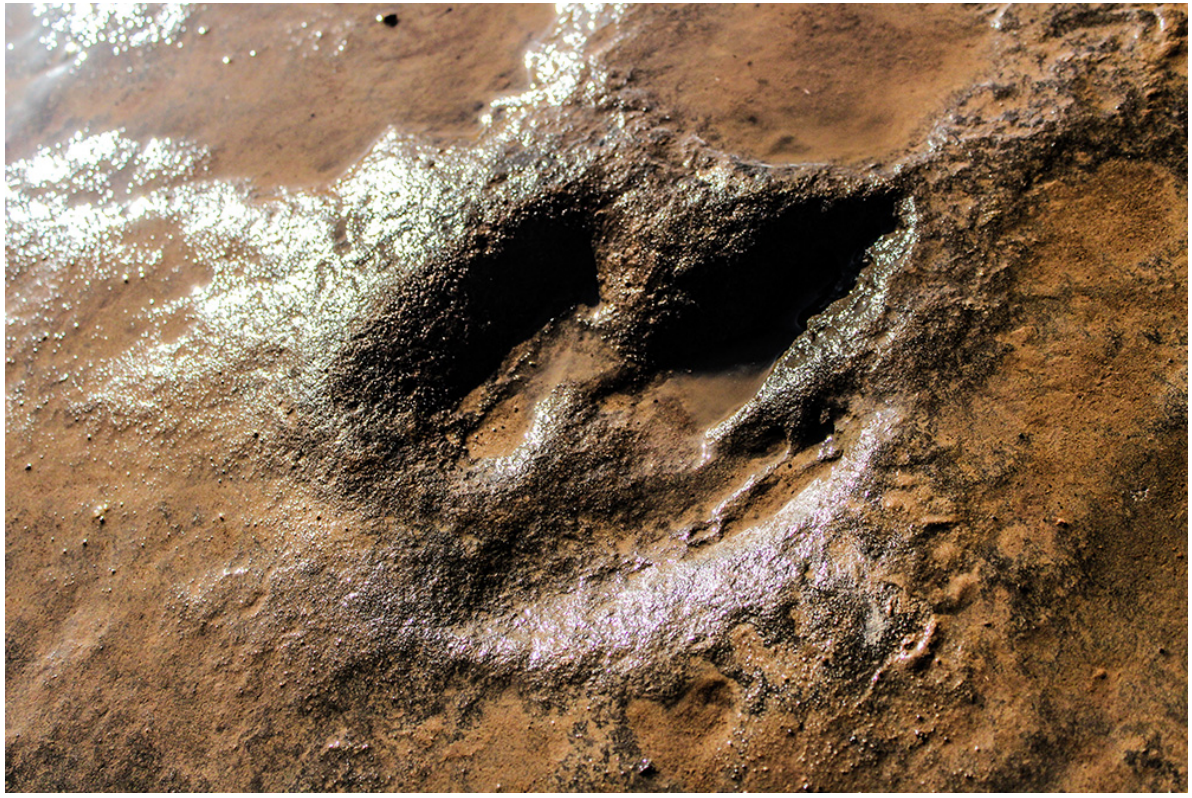


Omic data

Ome refers to the totality of elements in a biological domain
gen*ome*, proteome*ome*, ... interact*ome*, phen*ome*, exposome*ome*

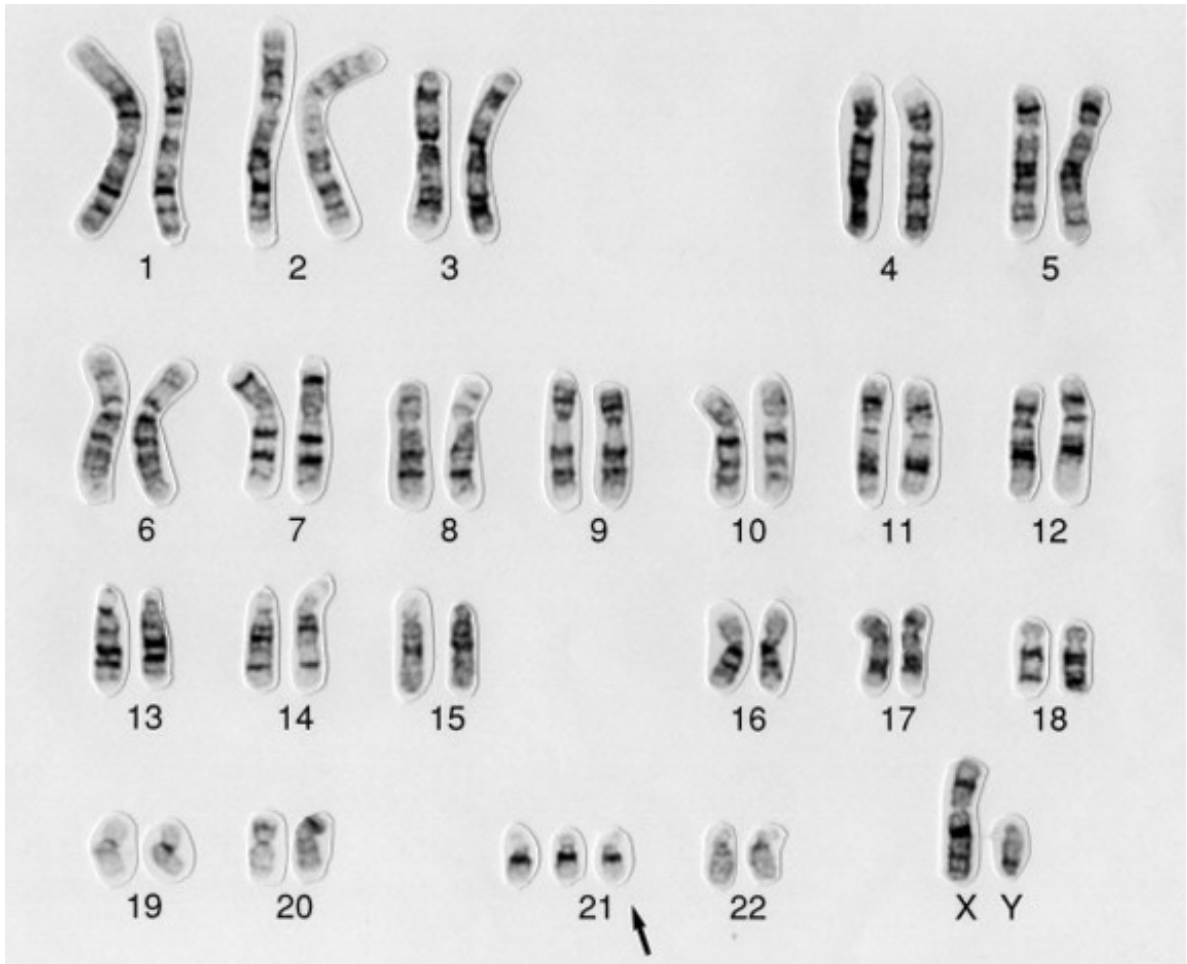
Operational definition:

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: **gen**omic 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

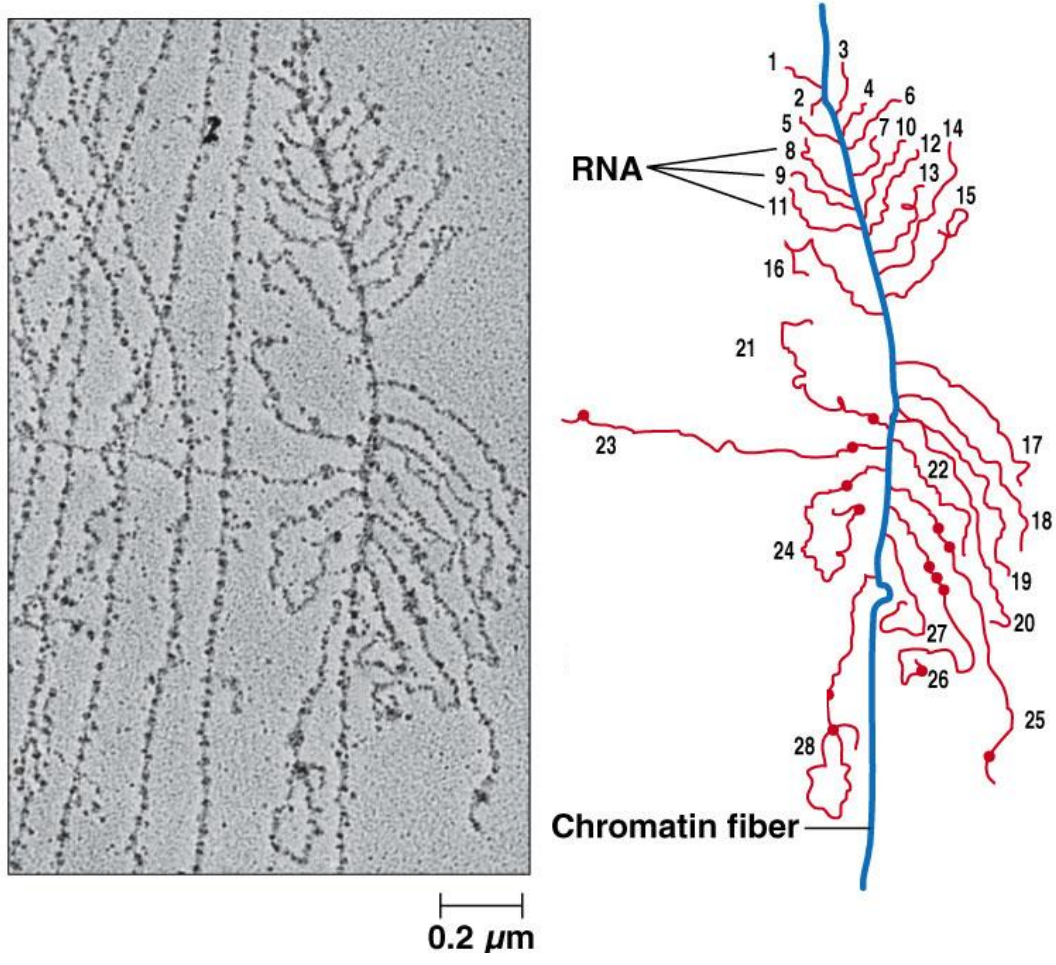
Property:

- 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



Transcritomic data: unbiased scan of RNA sequence

Definition:

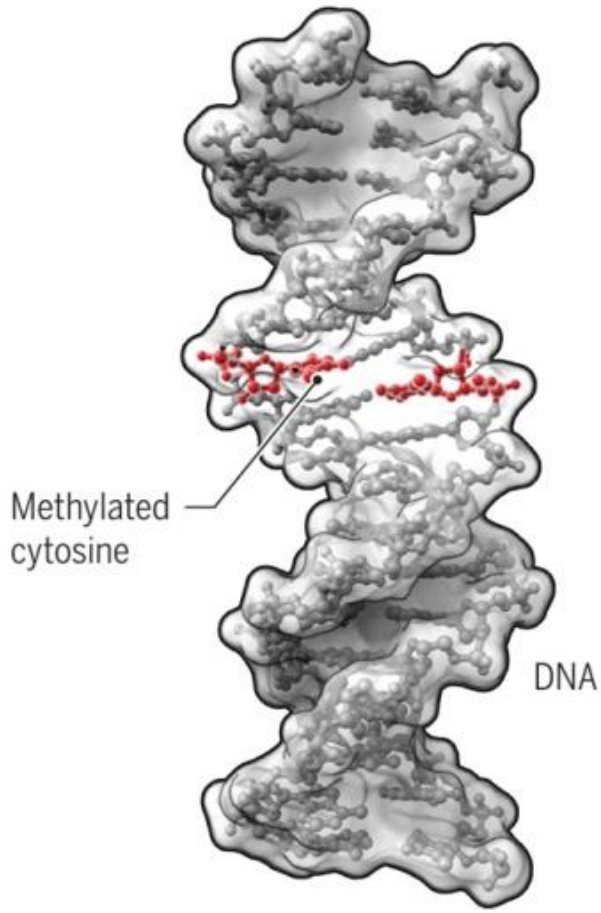
- A **Transcript variable** is the **amount** of a given RNA sequence from reference values (reference genome, hybridization probes)

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:

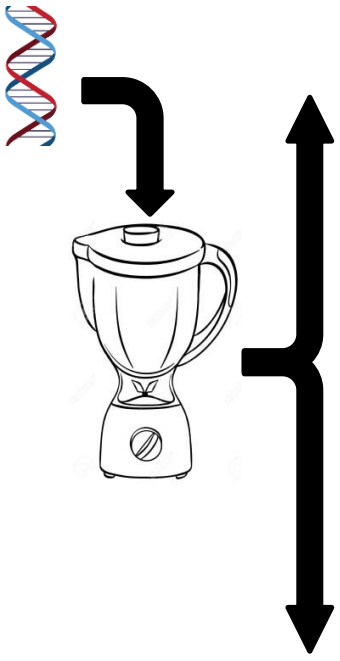
- A **Methylomic variable** is the **average state** of methylation at a given DNA site for the cells in a sample

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



```
Sense      5′-CCTCTCAACATTGAGGTCCCCAAAATCAGCCTCCACAGCCTCATTCTCGACTTTTCAGCAGTGTCCTTTCTTGATGTTTCTTCAGTGAGGGGCCTTAAA-3′
Antisense  3′-GGAGAGTTGTAACTCCAGGGGTTTTAGTCGGAGGTGTCGGAGTAAGAGCTGAAAAGTCGTCACAGGAAAGAACTACAAAGAAGTCACTCCCCGGAATTT-5′
                                                          *
           3′-GGAGCGTTGTAACTCCAGGGGTTTTAGTCGGAGGTGTCGGAGTAAGAGTT-5′
                3′-GTTGTAACTCCAGGGTTTTTAGTCGGAGGTGTCGGAGTAAGAGTTGAAAA-5′
                     3′-AACTCCAGGGTTTTTCGTCGGAGGGGTCGGAGTAAGAGTTGAAAAGTCGT-5′
                       5′-ctccaggggttttagtcggaggtgtcggagtaagagttgaaaagtcgtca-3′
                         3′-CCAGGGGTTTTAGTCGGAGGTGTCGGAGTAAGAGTTGAAAAGTCGTCACA-5′
                            5′-ggggttttagtcggaggtgtcggagtaagagttgaaaagtcgtcacagga-3′
                               3′-TTTTTGGTGGGAGGTGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAG-5′
                                  3′-TTTAGTCGGAGGTGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAA-5′
                                      3′-GTCGGAGGCGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAACTAC-5′
                                        5′-cggaggtgtcggagtaagagttgaaaagtcgtcacaggaaagaactacaa-3′
                                         3′-GGGGGGGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAACTACAAA-5′
                                          5′-gaggtgtcggagtaagagatgaaaagtcgtcacaggaaagaactacaaag-3′
                                             3′-GGGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAACTACAAAGAAG-5′
                                                5′-tcggagtaagagttgaaaagtcgtcacaggaaagaactacaaagaagtca-3′
                                                   3′-GAGTAAGAGTAGAAAAGTCGTCACAGGAAAGAACTACAAAGAAGTCACTC-5′
                                                      5′-agagttgaaaagtcgtcacaggaaagaactacaaagaagtcactccccgg-3′
                                                         3′-GTTGAAAAGTCGTCACAGGAAAGAACTACAAAGAAGTCACTCCCCGGAAT-5′
```

Hybridization

```
Sense      5′-CCTCTCAACATTGAGGTCCCCAAAATCAGCCTCCACAGCCTCATTCTCG
Antisense  3′-GGAGAGTTGTAACTCCAGGGGTTTTAGTCGGAGGTGTCGGAGTAAGAGC —●
```

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

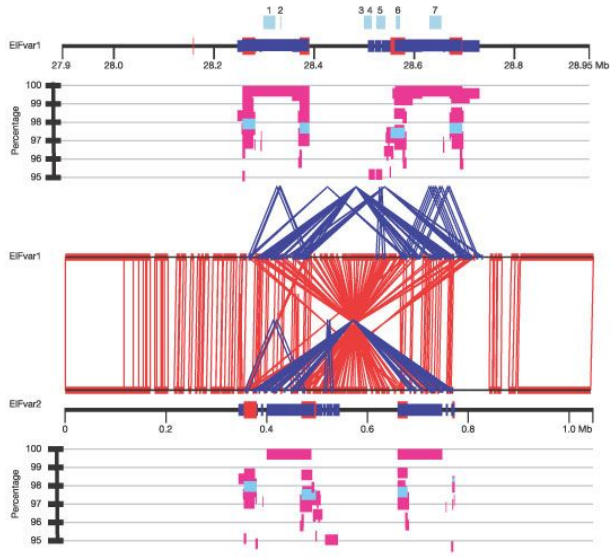
Examples

multi **omic** data integration

- Lost of chromosome Y, male susceptibility to disease.

Example 1
(hidden structure)

Studying inversion polymorphisms

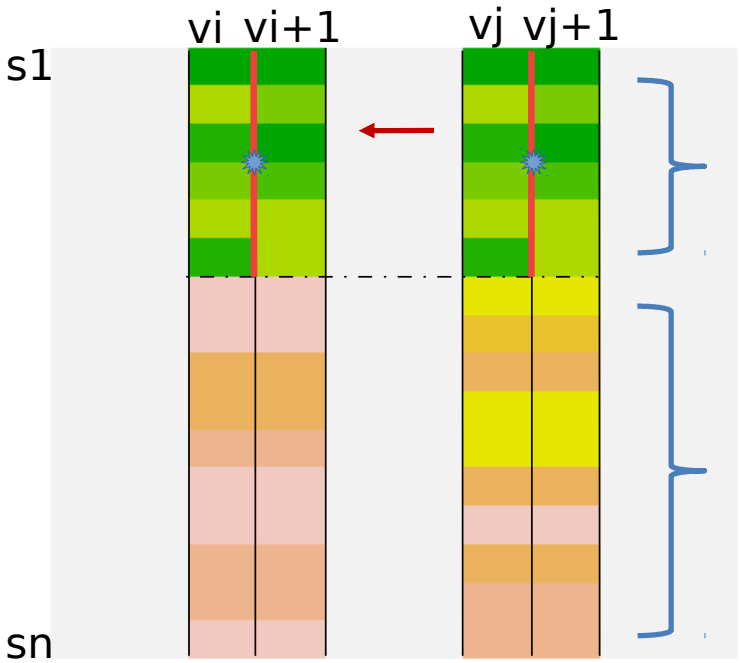


Inversions are DNA sequences that run in the opposite direction of a reference sequence.

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

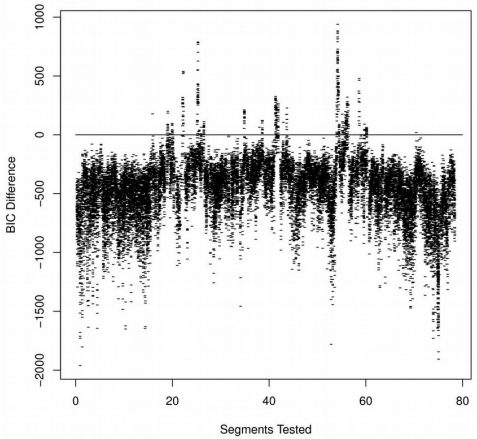
inversion imprint in genomic data

Genomic Data



Where?

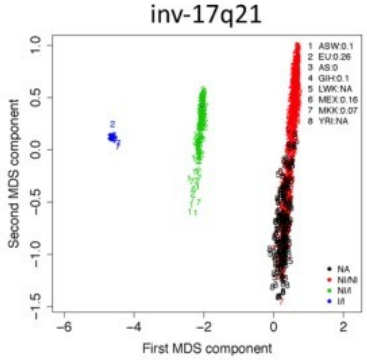
inveRsion



(Caceres et al BMC Bioinformatics , 2012)

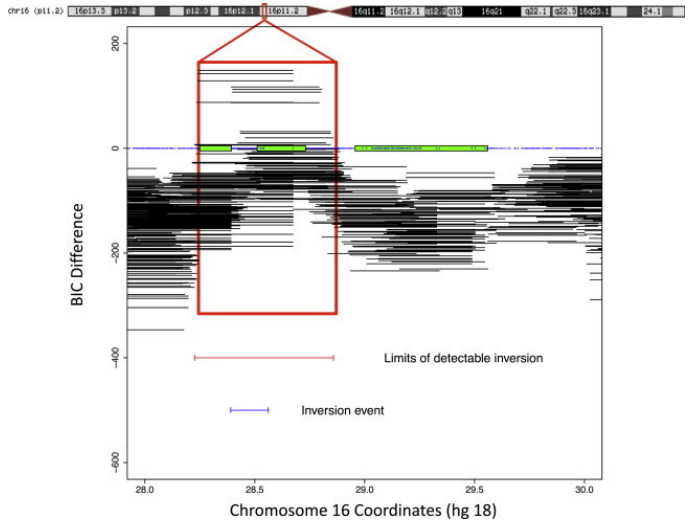
invClust

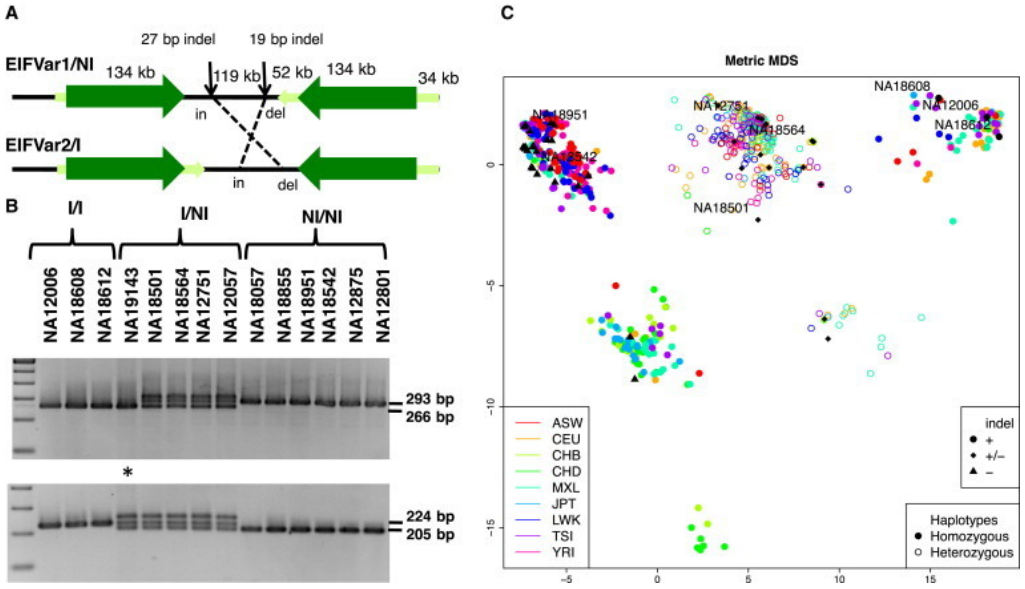
who?



(Caceres et al NAR, 2015)

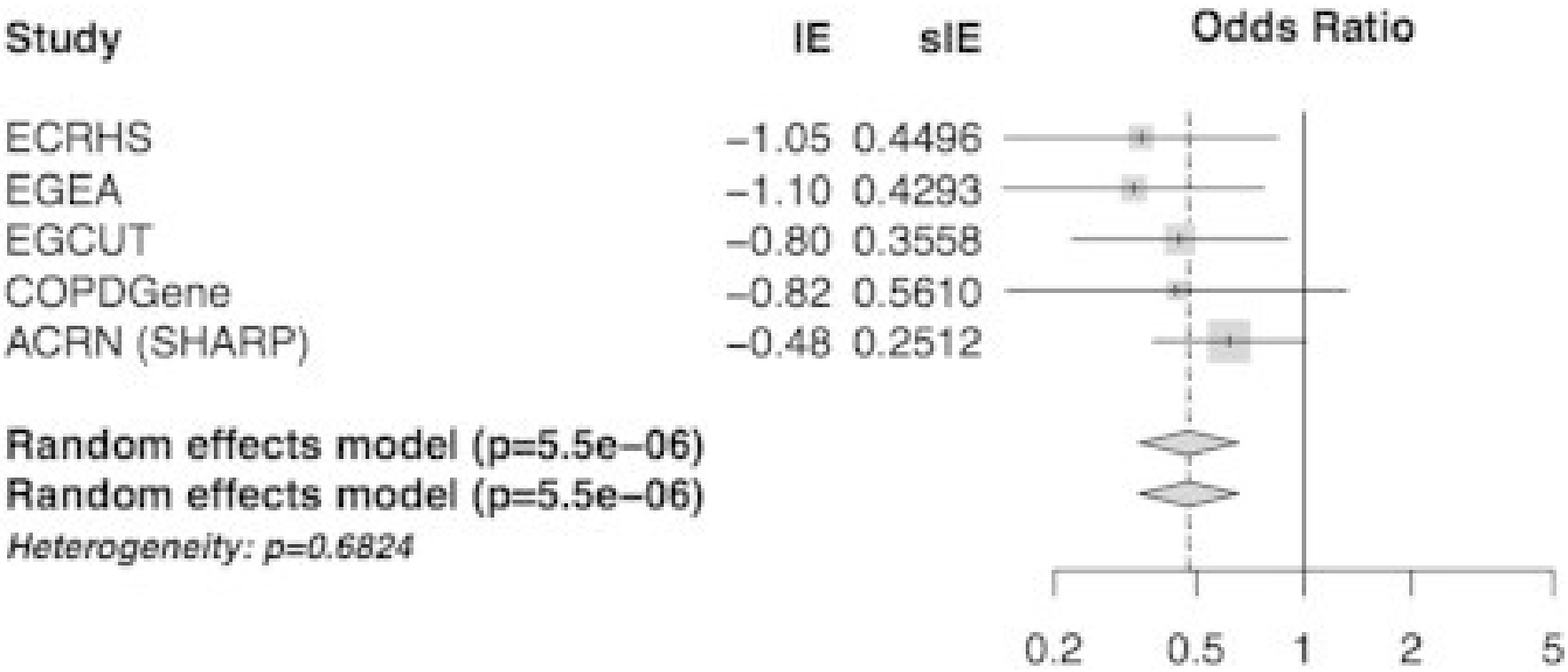
Detection and genotyping of inv-16p11





inversion 16p11

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



(Gonzalez*, Caceres*, et al. AJHG, 2014, *fisrt joint author)

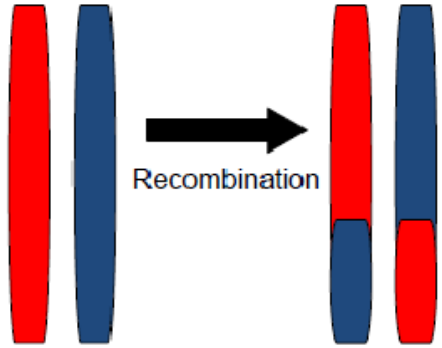
studying inversions with genomic data

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

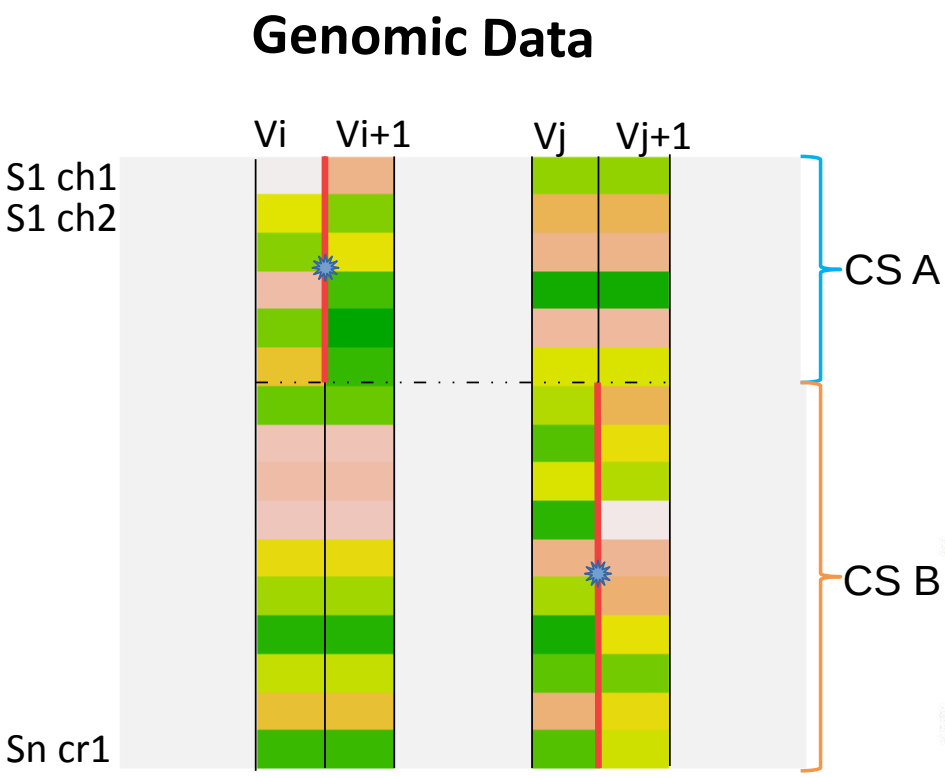


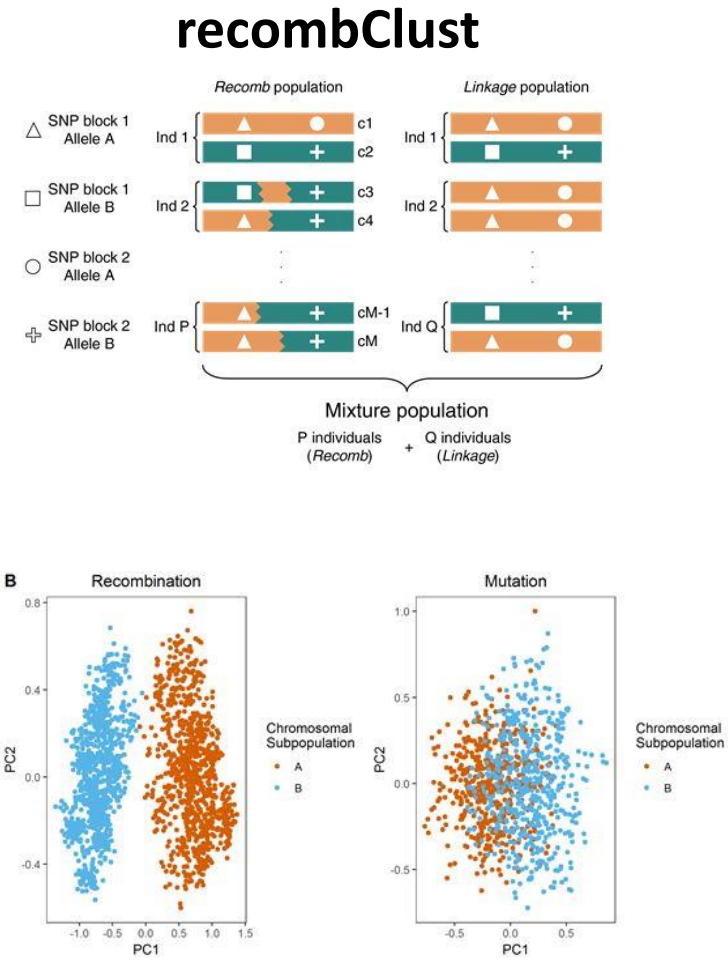
- increases genetic diversity
- different ancestries have different recombination patterns

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

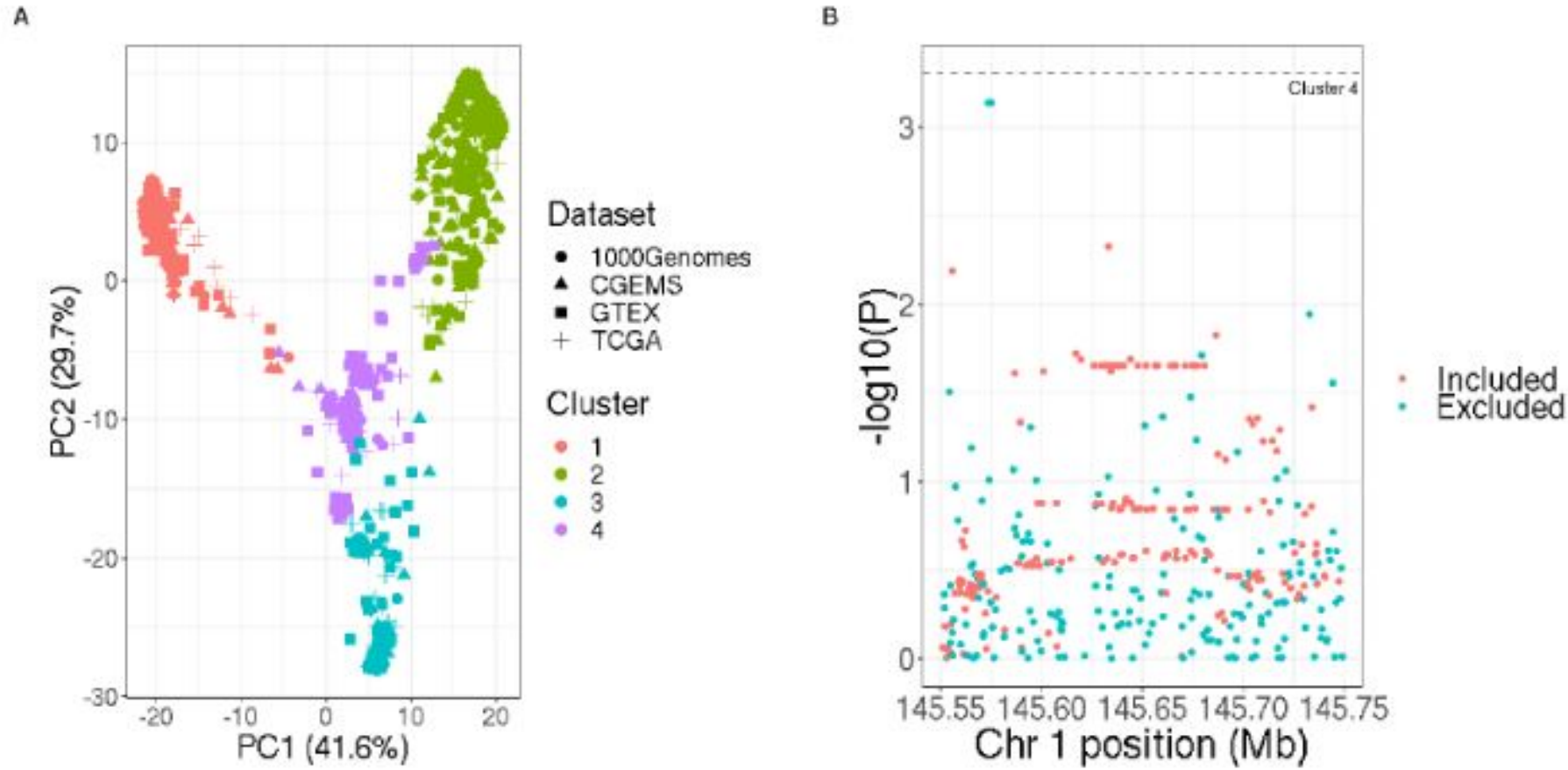




(Ruiz*, Caceres* et al submitted NAR, *first joint author)

Recombination substructure in 1q21.1

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



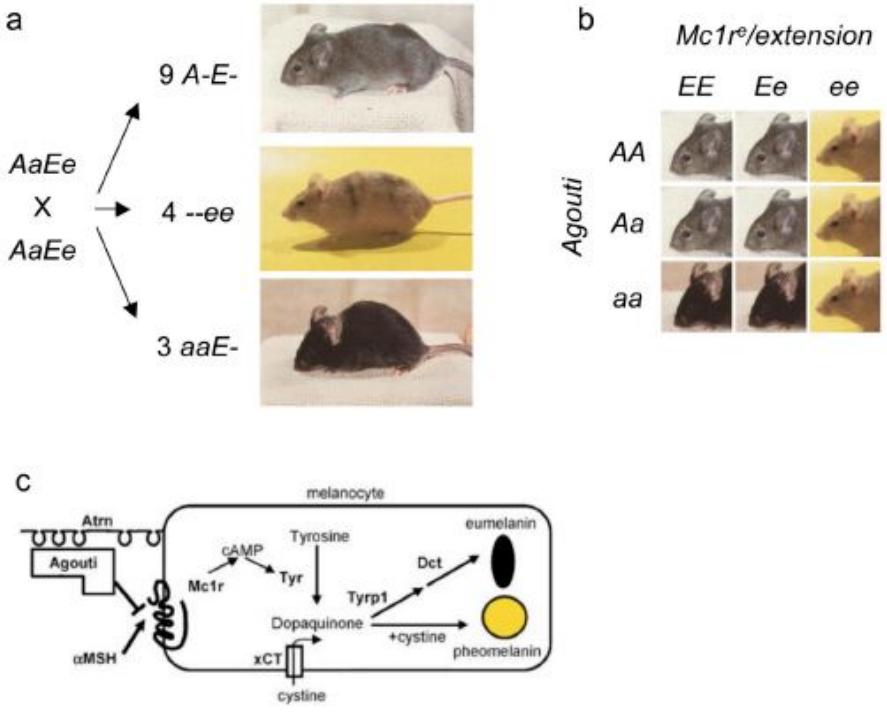
Studying recombination substructure with genomic data

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

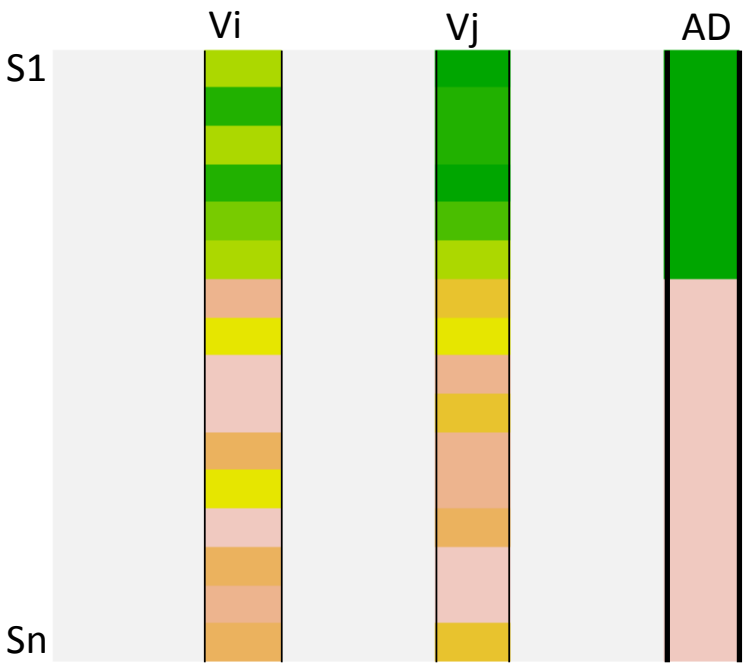


- complex traits are likely to emerge from the interaction between genomic variables
- there are too many to test ($\sim 10^{13}$ possibilities)

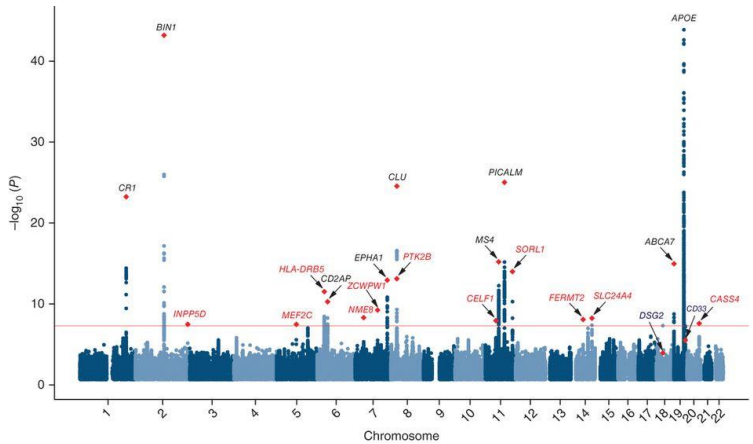
Do the interactions of validated risk SNPs **overlap**?

Genome-wide association studies (Alzheimer's Disease)

Genomic Data



GWAS

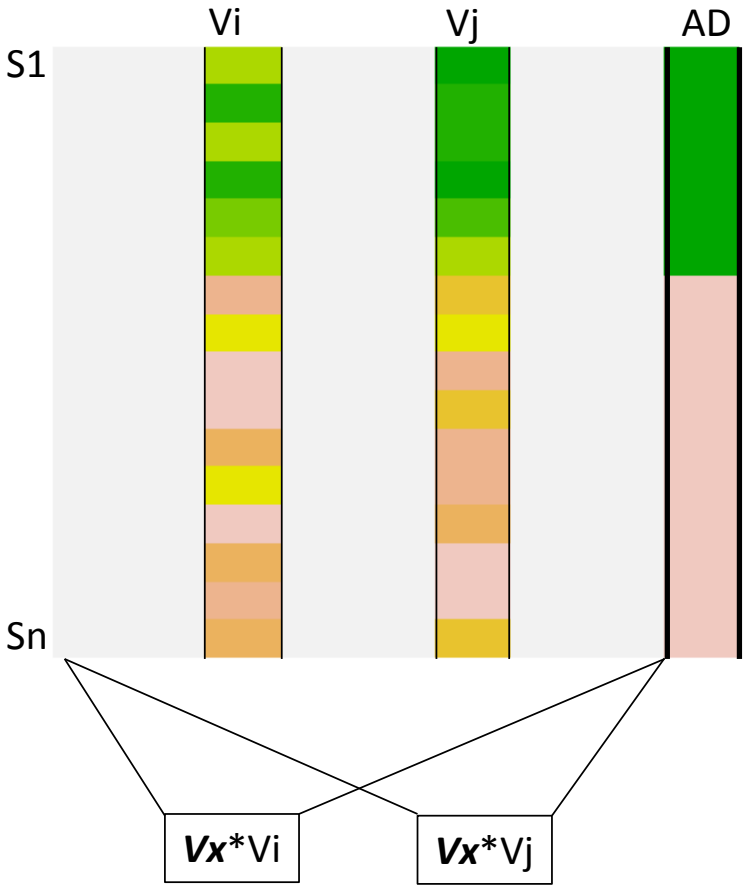


Validated associations

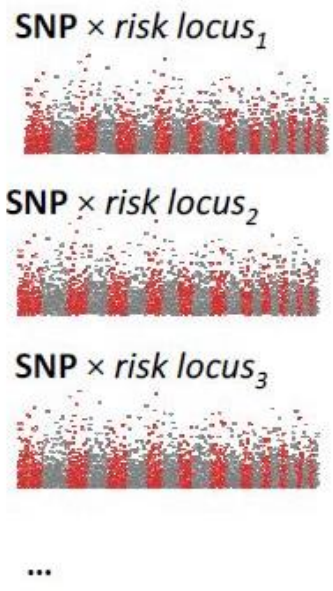
APOE's rs4420638
PICALM's rs536841
MS4A6A's rs610932
BIN1's rs610932
...

Epistasis in genomic data

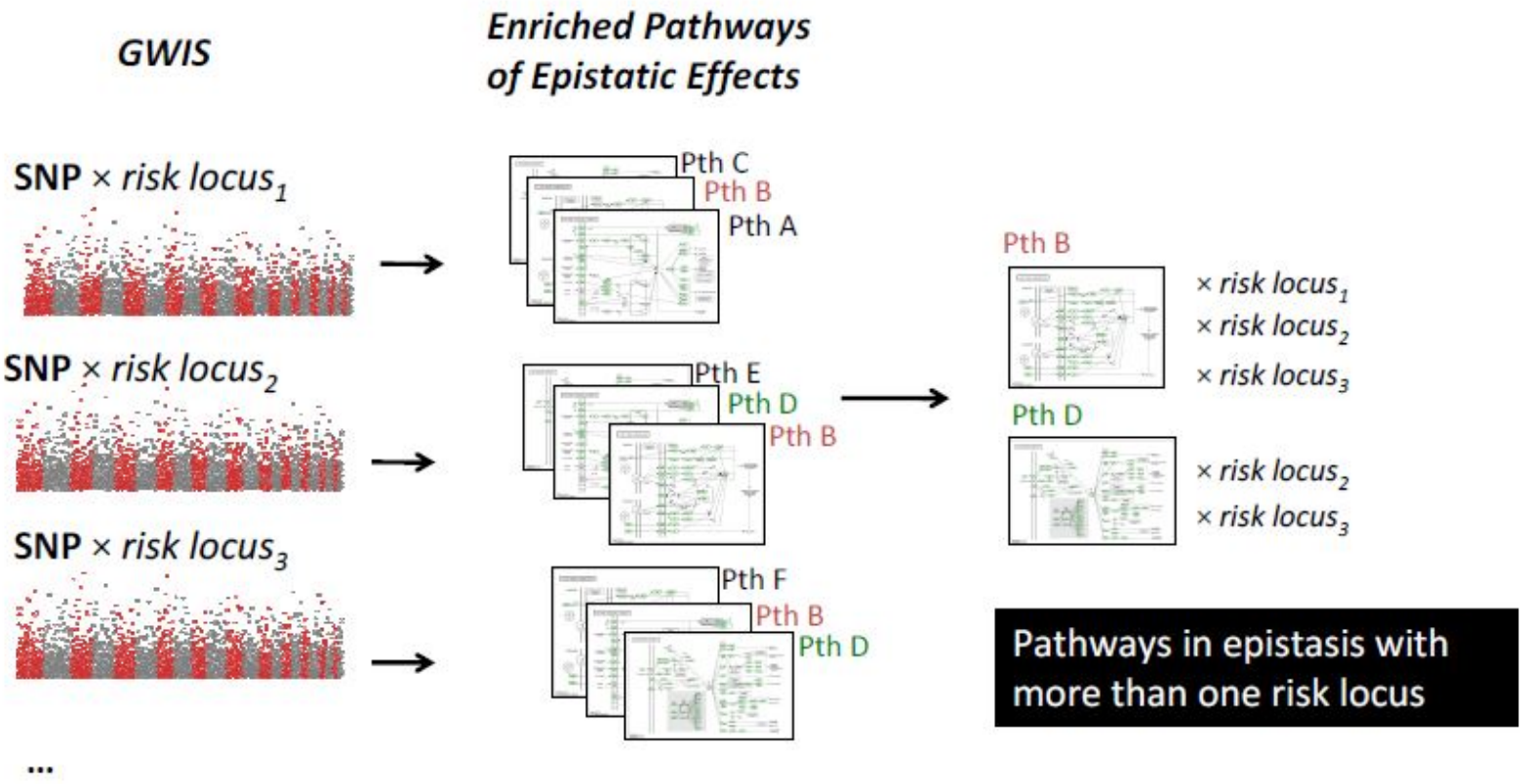
Genomic Data



Genome Wide Interaction



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

Gonodatropin signaling is enriched in interactions with *APOE* and *MS4A6A*'s polymorphisms

Risk Locus	Pathway	combined uncorrected p-value	combined corrected p-value	GENADA	NIA	ADG12	ADG31
rs429358 ×	*KEGG: GNRH SIGNALING PATHWAY*	*3.7e-5*	*0.01*	*0.025*	*0.033*	*0.001*	*0.046*
	KEGG: LONG TERM POTENTIATION	*1.6e-5*	*0.02*	0.145	0.001	0.001	0.096
	KEGG: ARRHYTHMOGENIC RIGHT VENTRICULAR CARDIOMYOPATHY	*2.9e-5*	*0.04*	0.001	0.146	0.001	0.196
	KEGG: CALCIUM SIGNALING PATHWAY	*1.1e-4*	*0.05*	0.162	0.001	0.01	0.087
	KEGG: VASCULAR SMOOTH MUSCLE CONTRACTION	*9.6e-5*	0.06	0.222	0.002	0.001	0.265
rs610932 ×	KEGG: PHOSPHATIDYLINOSITOL SIGNALING SYSTEM	*1.0e-4*	*0.008*	0.027	0.356	0.013	0.001
	KEGG: DILATED CARDIOMYOPATHY	*3.1e-7*	*0.01*	*0.01*	*0.001*	*0.001*	*0.014*
	BioCarta: HDAC PATHWAY	*6.4e-6*	*0.02*	0.002	0.016	0.147	0.001
	KEGG: VASCULAR SMOOTH MUSCLE CONTRACTION	*2.4e-7*	*0.03*	*0.001*	*0.001*	*0.013*	*0.008*
	KEGG: GNRH SIGNALING PATHWAY	*5.6e-6*	*0.05*	*0.01*	*0.004*	*0.011*	*0.009*
	KEGG: HYPERTROPHIC CARDIOMYOPATHY HCM	*1.0e-4*	0.06	0.248	0.001	0.004	0.131
	BioCarta: NKT PATHWAY	*1.7e-4*	0.11	0.022	0.001	0.201	0.055
	KEGG: GALACTOSE METABOLISM	*8.6e-5*	0.12	0.789	0.002	0.005	0.013
	BioCarta: PGC1A PATHWAY	*1.6e-4*	0.14	0.017	0.001	0.414	0.033
	KEGG: LONG TERM DEPRESSION	*1.4e-4*	0.15	0.024	0.072	0.022	0.005

Studying epistasis of risk variants with genomic data

Significance

- Clinical trials targeting the gonodatropin 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)

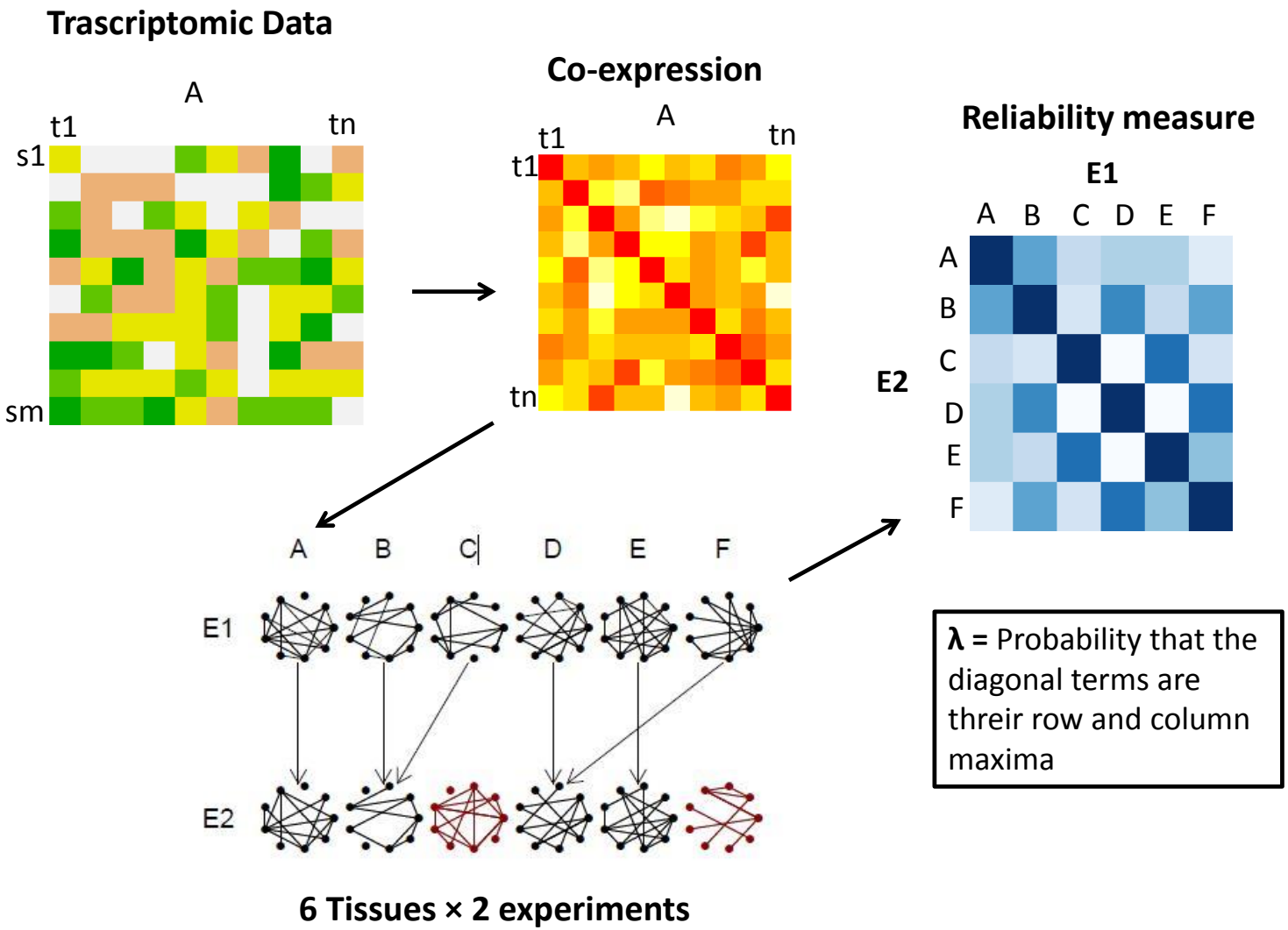
Studying co-expression networks

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



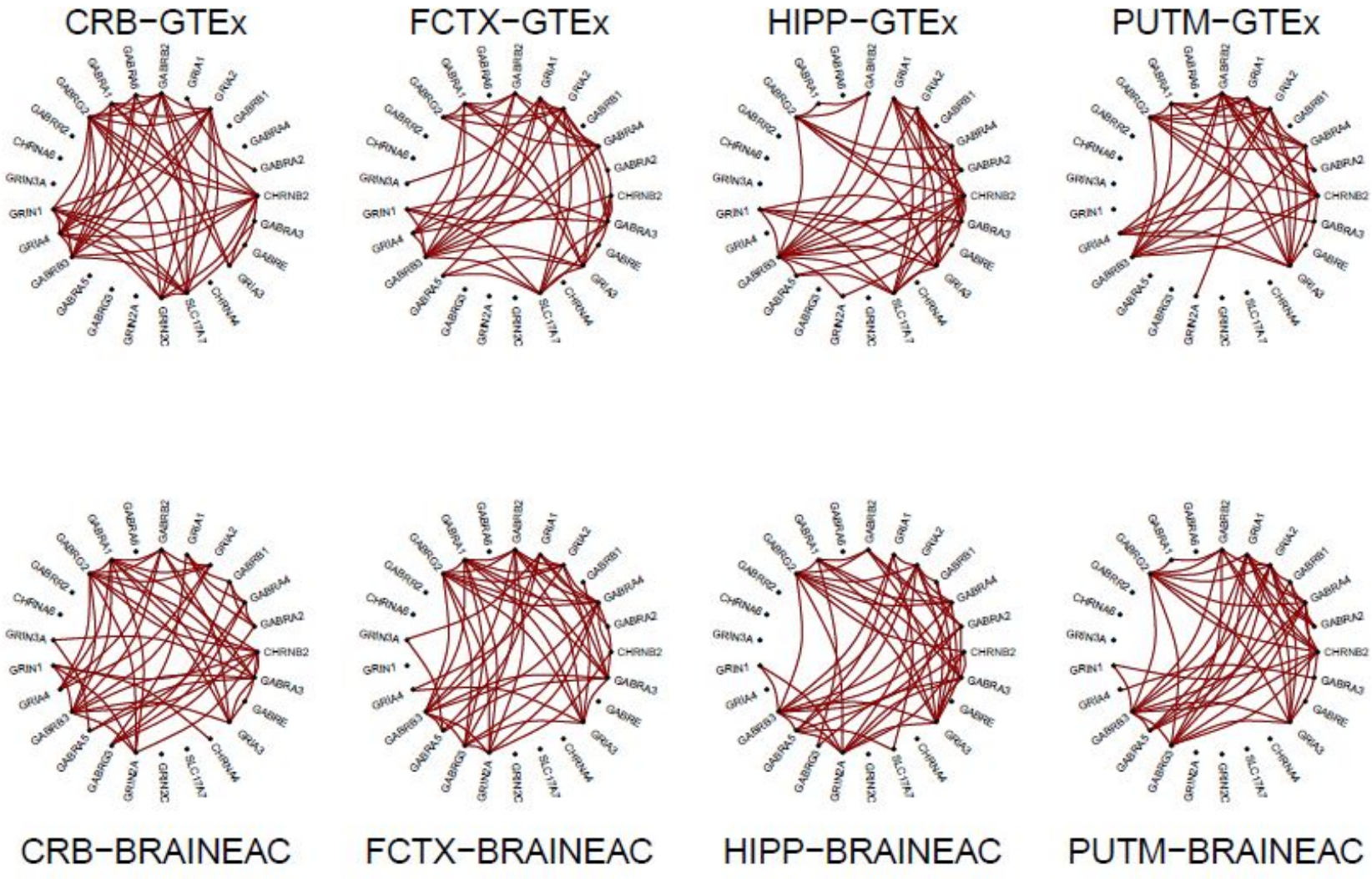
(Caceres et al. BMC genomics, under revision)

Inter-study reliability of networks across multiple tissues

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

λ	σ	Ref	Description
0.68	0.02	hsa05033	Nicotine addiction
0.67	0.04	hsa04720	Long-term potentiation
0.58	0.04	hsa05206	MicroRNAs in cancer
0.55	0.01	hsa04080	Neuroactive ligand-receptor int.
0.53	0.03	hsa04020	Calcium signaling pathway
0.52	0.03	hsa04261	Adrenergic sig. in cardiom.
0.51	0.02	hsa04912	GnRH signaling pathway

Nicotine addiction pathway across 4 brain regions



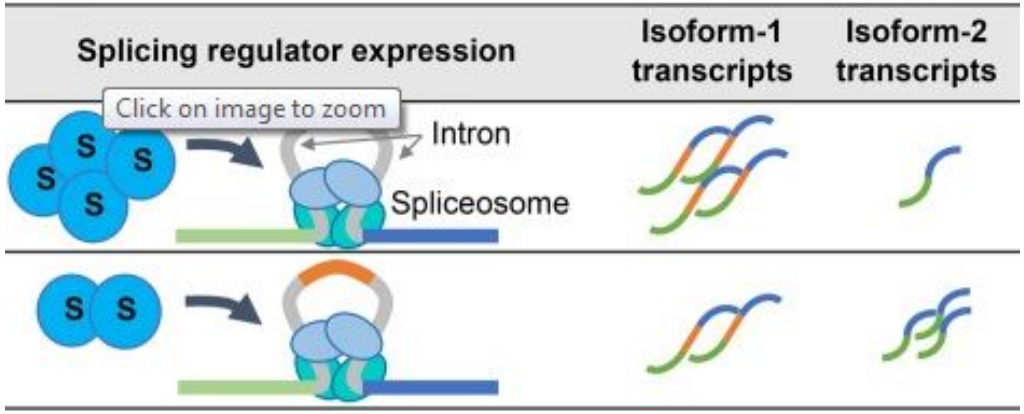
Studying network reliability with transcritomic data

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

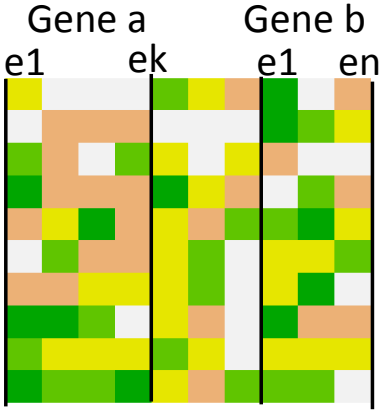


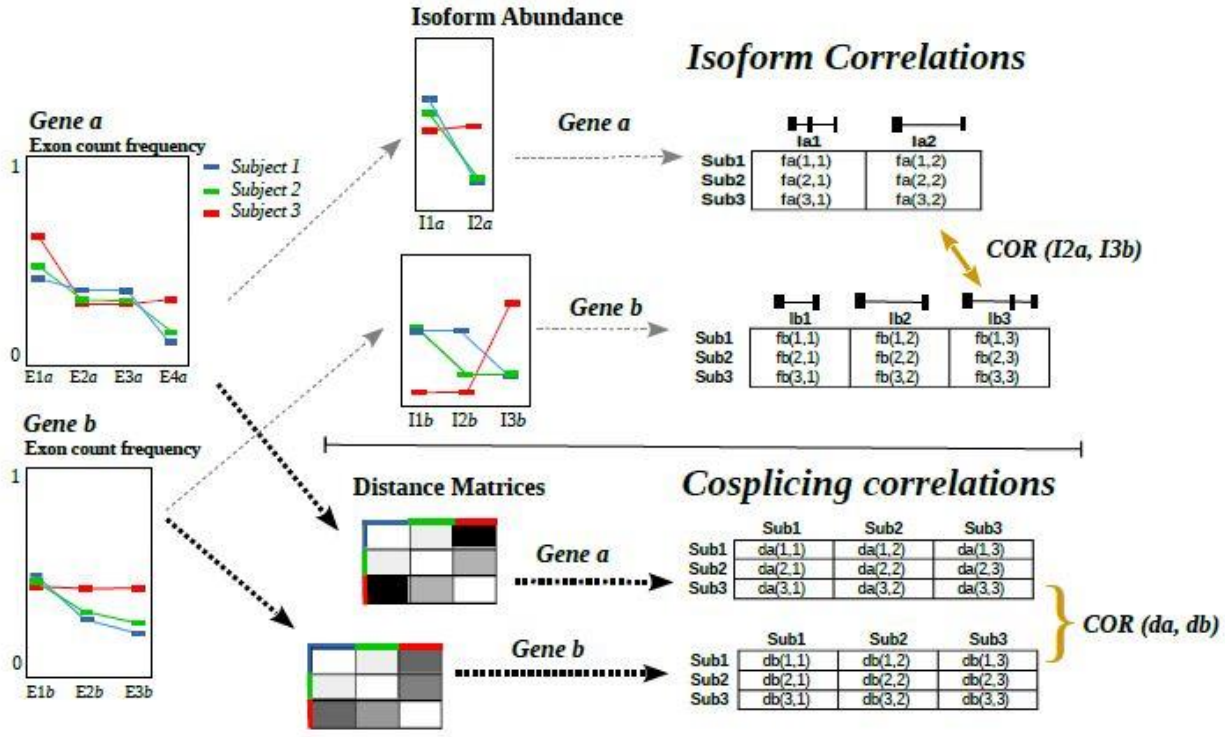
- Isoform ratios can correlate between two genes, across subjects

To which extent co-regulation of splicing can predict gene function?

Studying co-splicing with transcritomic data

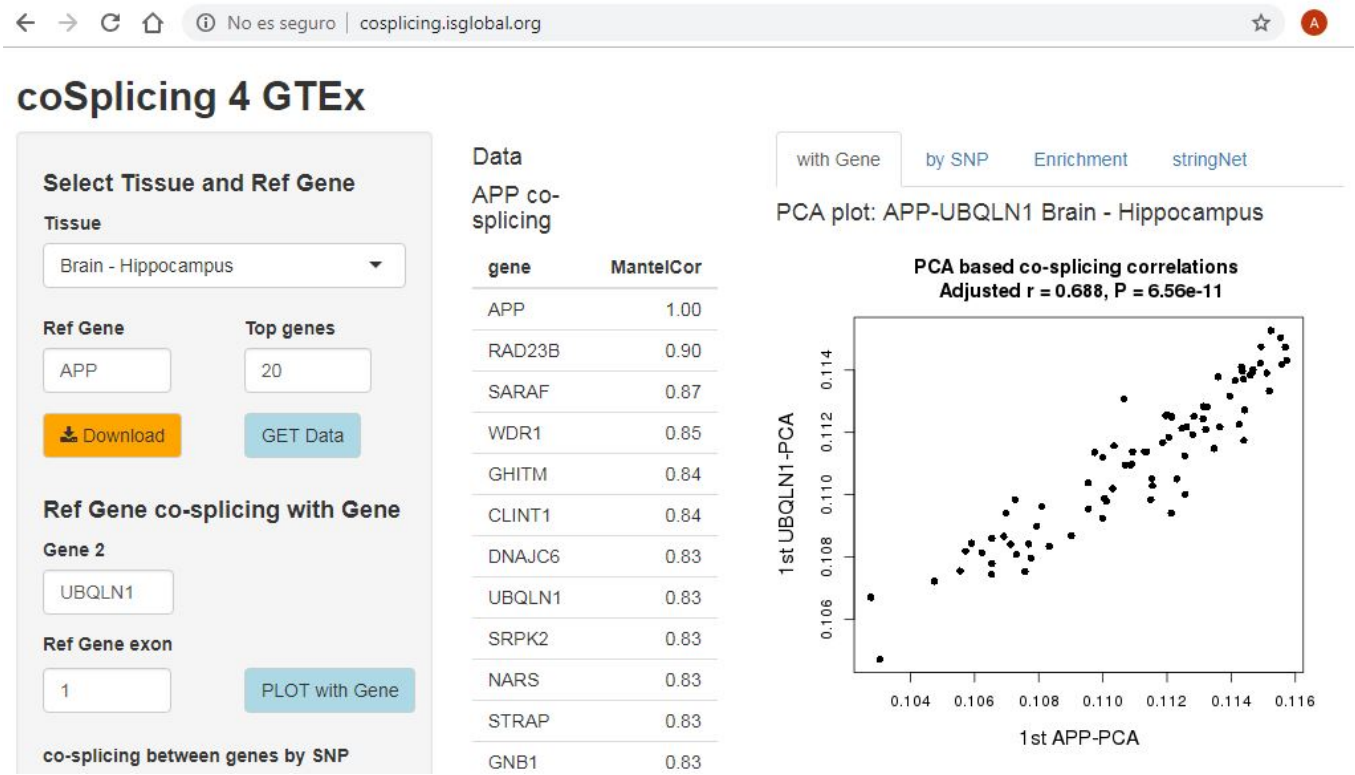
Trascriptomic Data





(Caceres et al, BMC genomics, 2018-accepted)

Physiological function of genes across multiple tissues



- work supported with computing hours from **RES**

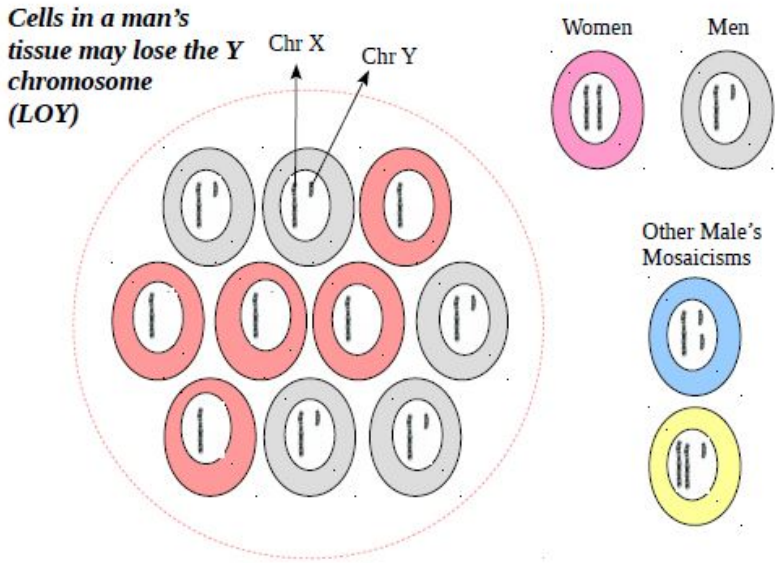
Studying co-splicing with transcriptomic data

Significance

- **APP** is physiologically liked 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)

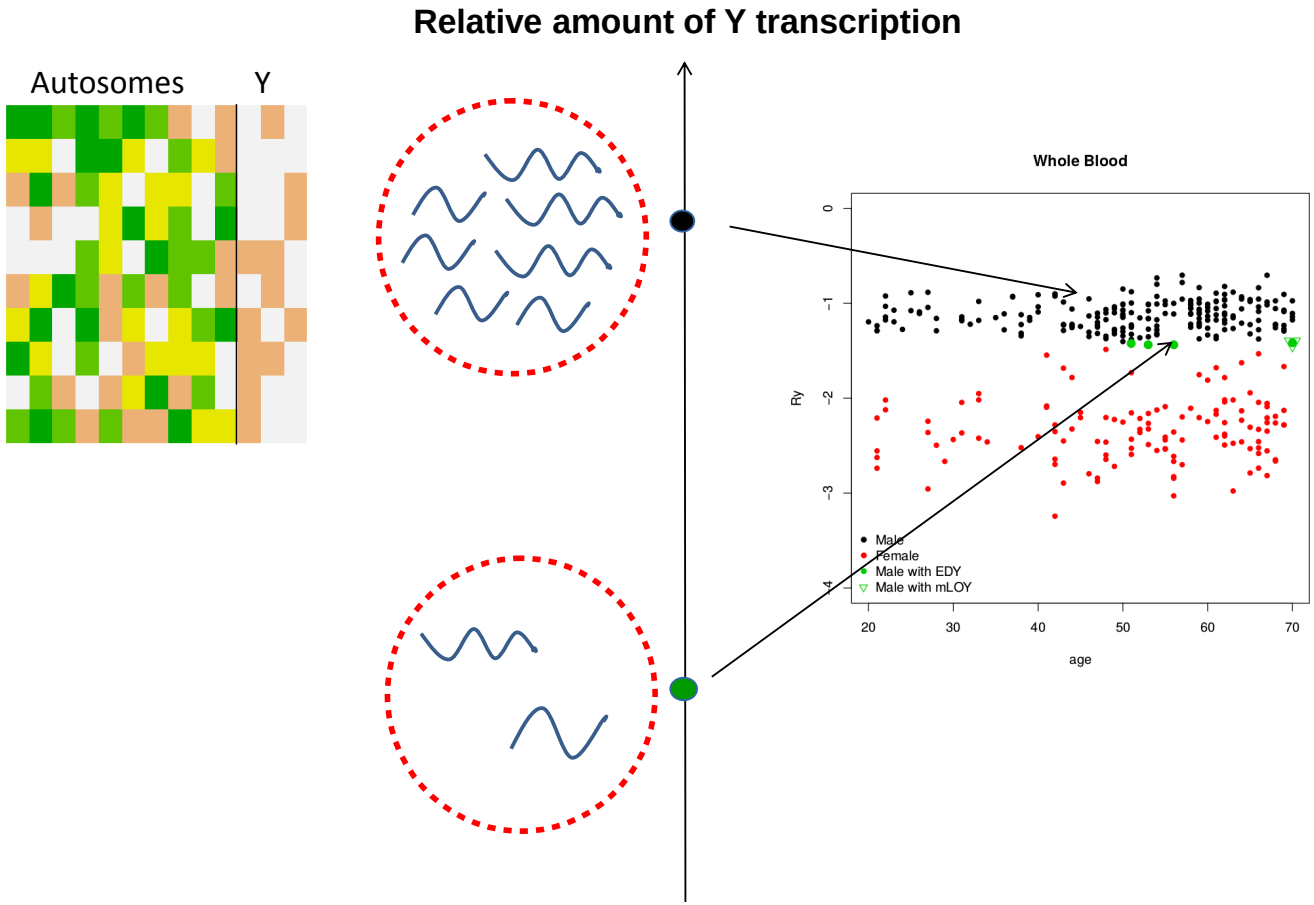
Studying loss of chromosome Y



- LOY associates with age and all-cause mortality in men (smoking, cancer and AD)
- We dont 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

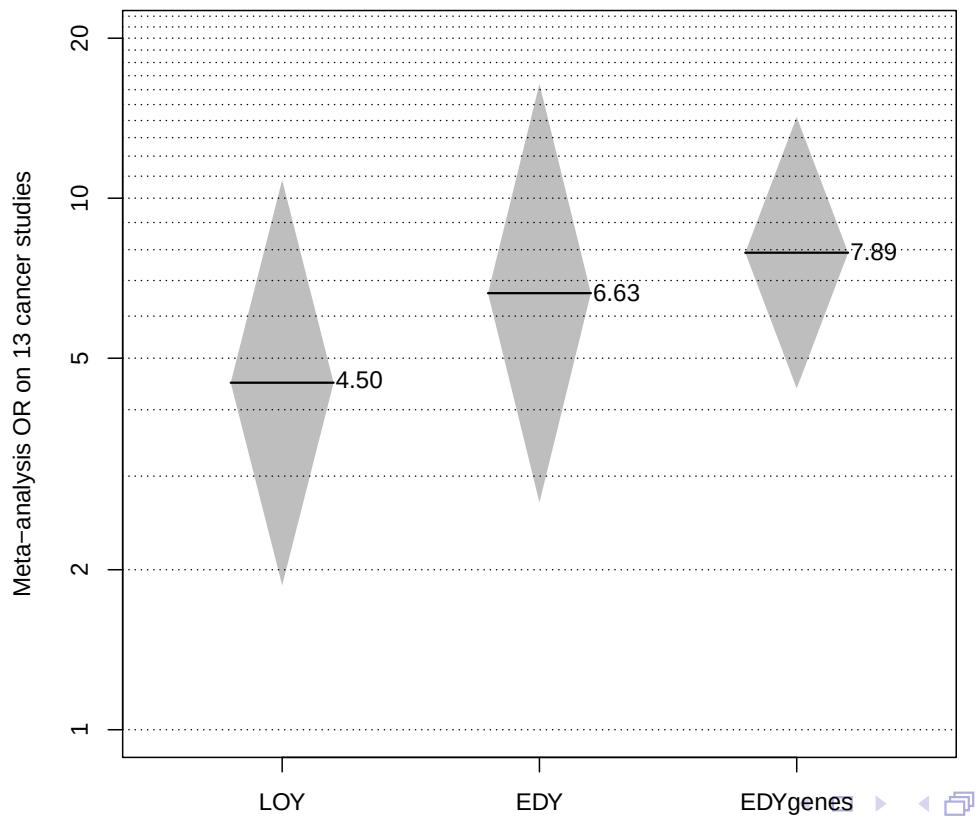


(Caceres et al, final draft ready!)

LOY → EDY → Male Disease

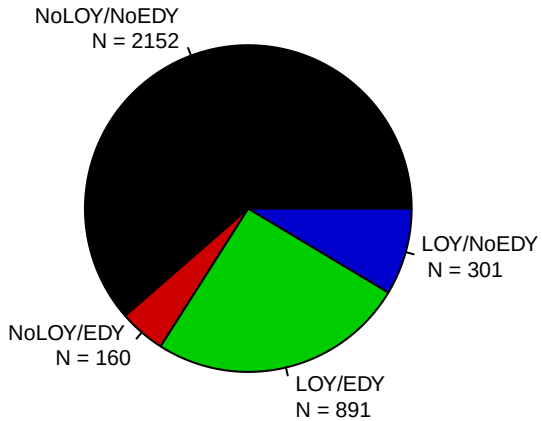
EDY:

- associates with LOY-associated conditions (age, AD, cancer)
- strongly correlates with LOY
- improves the effect of LOY with male disease

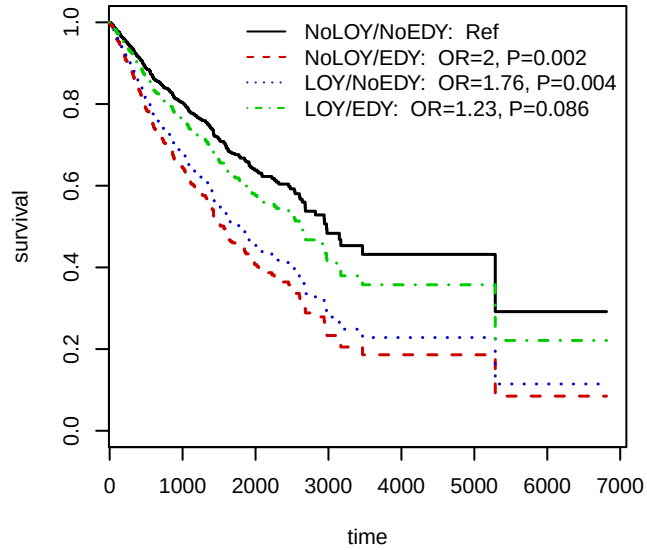


Studying EDY with multiple omic data

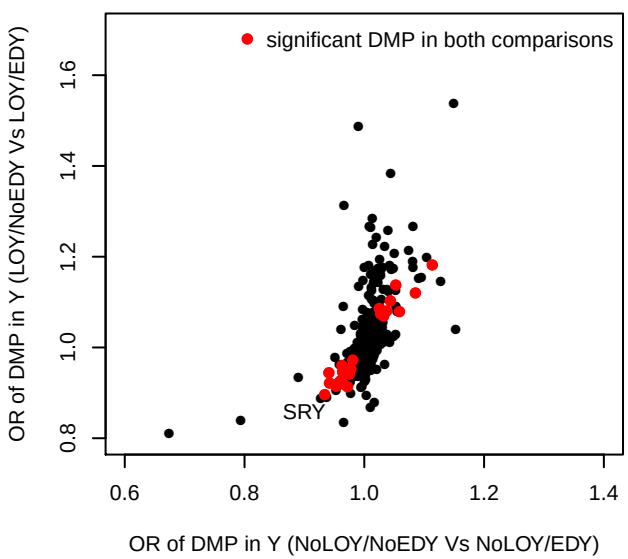
LOY−EDY status in cancer samples from TCGA



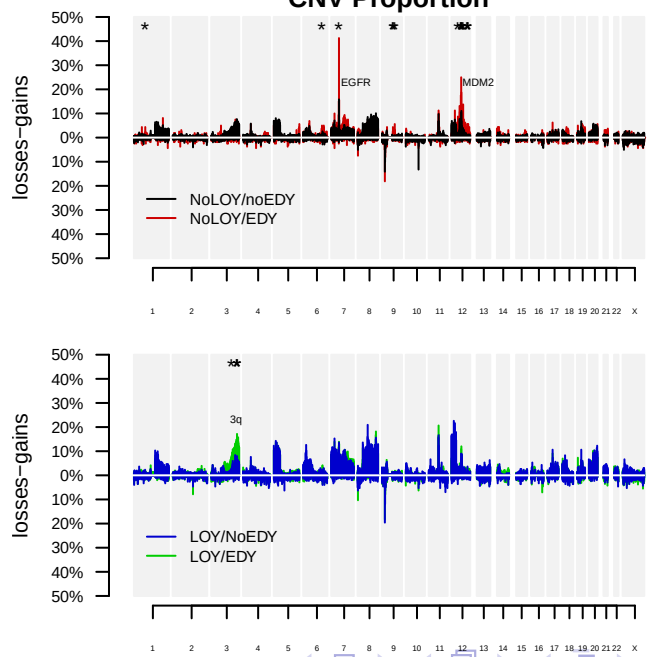
Survival of 13 different types of cancer



Differential Methylation across Y



CNV Proportion



Studying EDY with multiple omic data

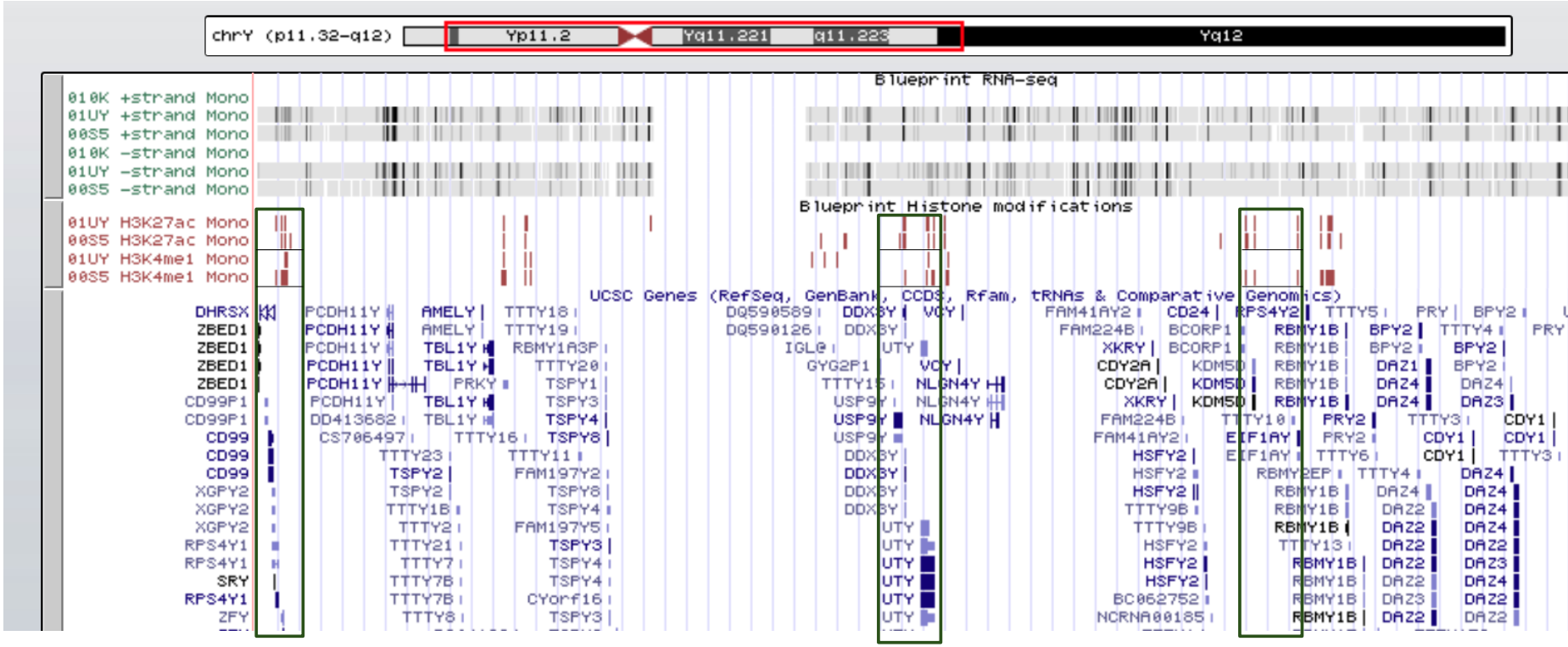
Significance

- 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?



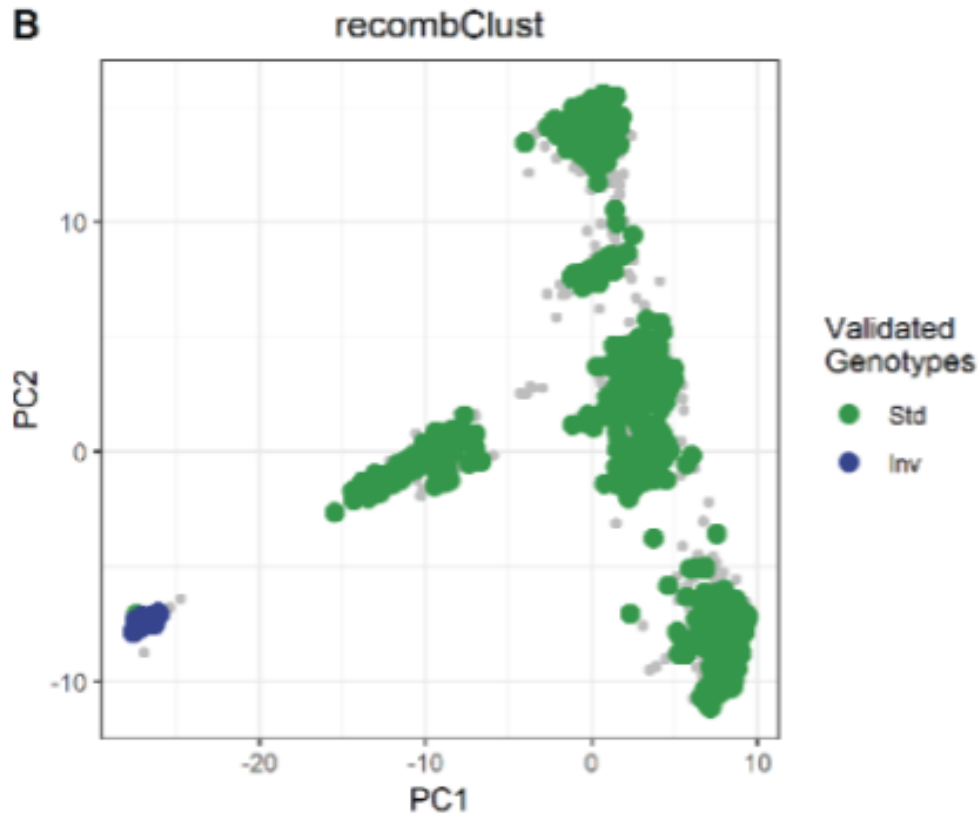
EDY is a protective factor for leukemia...
($controls = 3112, cases = 800, OR = 0.08, P = 5.3 \times 10^{-5}$)

Chromatine 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