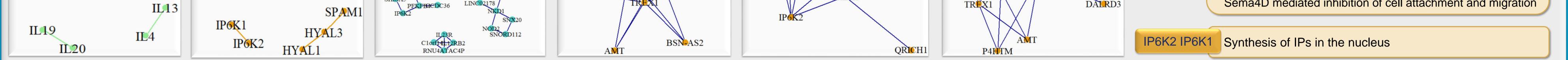
# **Improving efficiency in epistasis detection with a gene-based analysis using functional filters** Diane Duroux<sup>1,©</sup>, Héctor Climente-González<sup>2,3,4,©</sup>, Aldo Camargo<sup>1</sup>, Lars Wienbrandt<sup>5</sup>, David Ellinghaus<sup>5</sup>, Chloe-Agathe Azencott<sup>4,2,3</sup>, Kristel Van Steen<sup>1</sup>

<sup>©</sup> Equally contributing authors, <sup>1</sup> BIO3 - GIGA-R Medical Genomics, Liege, Belgium; <sup>2</sup> Institut Curie, PSL Research University, F-75005 Paris, France; <sup>3</sup> INSERM, U900, F-75005 Paris, France; <sup>4</sup> MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France; <sup>5</sup> Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany

Motivation	Methodology						
	Prior biological knowledge						
Epistasis analysis can complement oversimplified GWAS to understand complex traits.	Prior knowledge with FUMA <sup>1</sup> SNP to gene	Distance       Chromatine       eQTL         UMA_Dist={SNPs}, such as the NP was within the boundaries of the gene ± 10kb       FUMA_Chrom= {SNPs}, a contact had been observed between the SNP and the region around the gene's promoter (250bp upstream and 500bp downstream from the TSS)       FUMA_eQTL= {SNPs}, SNP had been linked to the expression of a gene in any tissue in GTEx       NA					
Available biological knowledge is growing. Many choices can be made to use one of the sources or a combinations of several ones. There is a need to study	Knowledge integration with Biofilter <sup>2</sup> Gene-gene interactions	$Biofilter_{models} = \{G_k, G_l\} \text{ such as the pair of genes has evidence of co-function across 2 publicly available biological databases. It includes genomic locations of SNPs and genes, and known relationships among genes and proteins such as interaction pairs, pathways and ontological categories. No use of trait information.}$			NA	NA	
the associated benefits. Genes are the natural units of analysis to interpret findings as they can indicate relevant pathways and biological processes. However in the context of epistasis, gene-based methods are underinvestigated.	Selection of SNP pairs	Dist= $(SNP_i, SNP_j)$ , such $SNP_i \in G_{ik}$ (FUMA_Dist $SNP_j \in G_{jl}$ (FUMA_Dist $(G_k, G_l) \in Biofilter_{model}$ 4,576,310 SNP-pairs	t), st), $SNP_i \in G_{ik}(FUMA\_Chron SNP_j \in G_{jl}(FUMA\_Chron SNP_j \in G_{jl})$	$ \begin{array}{l} & G_{ik}(FUMA\_Chrom), \\ & \in G_{jl}(FUMA\_Chrom), \\ & G_{l}) \in Biofilter_{models} \end{array} \end{array} \begin{array}{l} & SNP_i \in G_{ik}(FUMA\_eQTL), \\ & SNP_j \in G_{jl}(FUMA\_eQTL), \\ & (G_k, G_l) \in Biofilter_{models} \end{array} \end{array} $			
				SNP-based epistasis and	alysis		
	<ul> <li>QC + LD + adjustement for population structure</li> <li>Epistasis detection</li> <li>Post controls</li> </ul>	Epistasis linear regression: Y = β <sub>0</sub> + β <sub>1</sub> g <sub>A</sub> + β <sub>2</sub> g <sub>B</sub> + β <sub>3</sub> g <sub>A</sub> g <sub>B</sub> , to evaluate β <sub>3</sub> =0 using PLINK Remove pairs with two SNPs in HLA region and in LD r <sup>2</sup> >0.75 Bonferroni correction					
	<ul> <li>Multiple testing</li> </ul>	0 SNP-pair	8 SNP-pairs	25 SNP-pairs	· · · · · · · · · · · · · · · · · · ·	7 SNP-pairs	
Data				Gene-based epistasis an	<b>IAIYSIS</b> $(SNP_i, SNP_j)$ , such as $(SNP_i, SNP_j)$ , such as	(SNP <sub>i</sub> , SNP <sub>i</sub> ), such as	
Inflammatory bowel disease is characterized by chronic inflammation of the gastrointestinal tract. Cronh's disease and ulcerative colitis are the principal types of IBD.	Selection of SNP pairs	$(SNP_i, SNP_j) \in Dist$ 0 SNP-pair	$(SNP_i, SNP_j) \in Chrom$ 8 SNP-pairs	$(SNP_i, SNP_j) \in eQTL$	$SNP_{i} \in G_{ik}(FUMA\_Dist)$ $SNP_{j} \in G_{jl}(FUMA\_Dist)$ $SNP_{j} \in G_{jl}(FUMA\_Dist)$ $SNP_{j} \in G_{jl}(FUMA\_Chrom)$ $S$	$SNP_i \in G_{ik}(FUM1\_eQTL)$ $SNP_j \in G_{jl} (FUMA\_eQTL).$ Colon Pancreas Stomach	
	Significance	0 SNP-pair       8 SNP-pairs       2 SNP-pairs       0 SNP-pairs       3 SNP-pairs       2 SNP-pairs       2 SNP-pairs         Adaptative Truncated Product Method (ATPM <sup>3</sup> ) to derive empirical p-values for gene pairs         Bonferroni correction         0 gene-pair       0 gene-pairs       0 gene-pairs       0 gene-pairs       30 gene-pairs       17 gene-pairs       18 gene-pairs					
IBD data from the IBD consortium: 66,280 samples,		0 gene-pair	4 gene-pairs	Pathway and Network an		So gene-pairs Tr gene-pairs To gene-pairs	
corresponding to 32,622 cases and 33,658 controls and	Pathway significance	ATDM using KECC. Departures DIED and norther interaction databases					
130,071 variants .	Networks	Visualisation and analysis of degrees largest component and communities					
Results							
SLC22A5 SLC22A4 SLC22A4 SLC22A4 SLC22A5	ical mapping and sto	biological filter mach eQTL mapping	No biological filter and colon eQTL mapping	No biological filter and pancreas eQTL mapping QRICHI	MST1R MST1 a6b1 and a6b4 Integrin signaling Signaling by MST1 Axon guidance mediated by G alpha (12/13) signalling e	r semaphorin	
ILIS SLC22A4 RHOA HI33 SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA SLC22A4 RHOA	LINC02168 CYLD LOC101927272 LINC02178	ZNESSERPS18AP1 CCDC36	HYALS ZNE 589REX1	CDC25A	SLIT2:ROBO1 increases RI Sema4D induced cell migra		



- In this application, the chromatine filter does not seem well balanced in terms of number of tests and number of significant pairs detected because it requires more tests than eQTL and highlights fewer gene pairs.
- eQTL has a reduced number of tests, leads to more significant gene pairs than the hypothesis free analysis and identifies significant pathways.
- Hypothesis-free screening is more cumbersome but detects a higher number of significant SNP-pairs and gene-pairs.

#### **Discussion & Conclusion**

Gene-based epistasis analysis is recommended to understand the biological mechanism of diseases because it highlights biological processes and pathways. The hypothesis-free screening allows the identification of more gene-pairs and communities, whereas the eQTL filter is more suited to detect significant pathways.

The combination of epistasis networks and SNP aggregation enhances identification and interpretation of SNP-based epistasis findings via communities and pathway detection.

This pipeline not only allows to find high-dimensional interactions at the gene level, but also gives a global visualization with neighbors, and highly connected substructures.

# References

 <sup>1</sup> Watanabe, Kyoko, et al. "Functional mapping and annotation of genetic associations with FUMA." *Nature communications* 8.1 (2017): 1826
 <sup>2</sup> Pendergrass, Sarah A., et al. "Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development." *BioData mining* 6.1 (2013): 25.
 <sup>3</sup> Sheng, Xuguang, and Jingyun Yang. "An adaptive truncated product method for combining dependent p-values." *Economics letters* 119.2 (2013): 180-182.

### Acknowledgement

# ✓ We discovered novel susceptibility genes and gene-pairs for IBD. For instance and to our knowledge, HYAL3, LINC02168, and SNORD112 have never been associated to IBD before.

#### Permutation analysis validate type I error and current knowledge about IBD was largely retrieved.

Further simulations will be necessary to study the influence of the number of tests on the number of false positives.

Therefore, additional tests of this protocol with other analytic tools will be performed.



Contact: diane.duroux@uliege.be