Genomic set enrichment analysis enhanced through integration of chromatin long-range interactions

Michael M. Hoffman @michaelhoffman

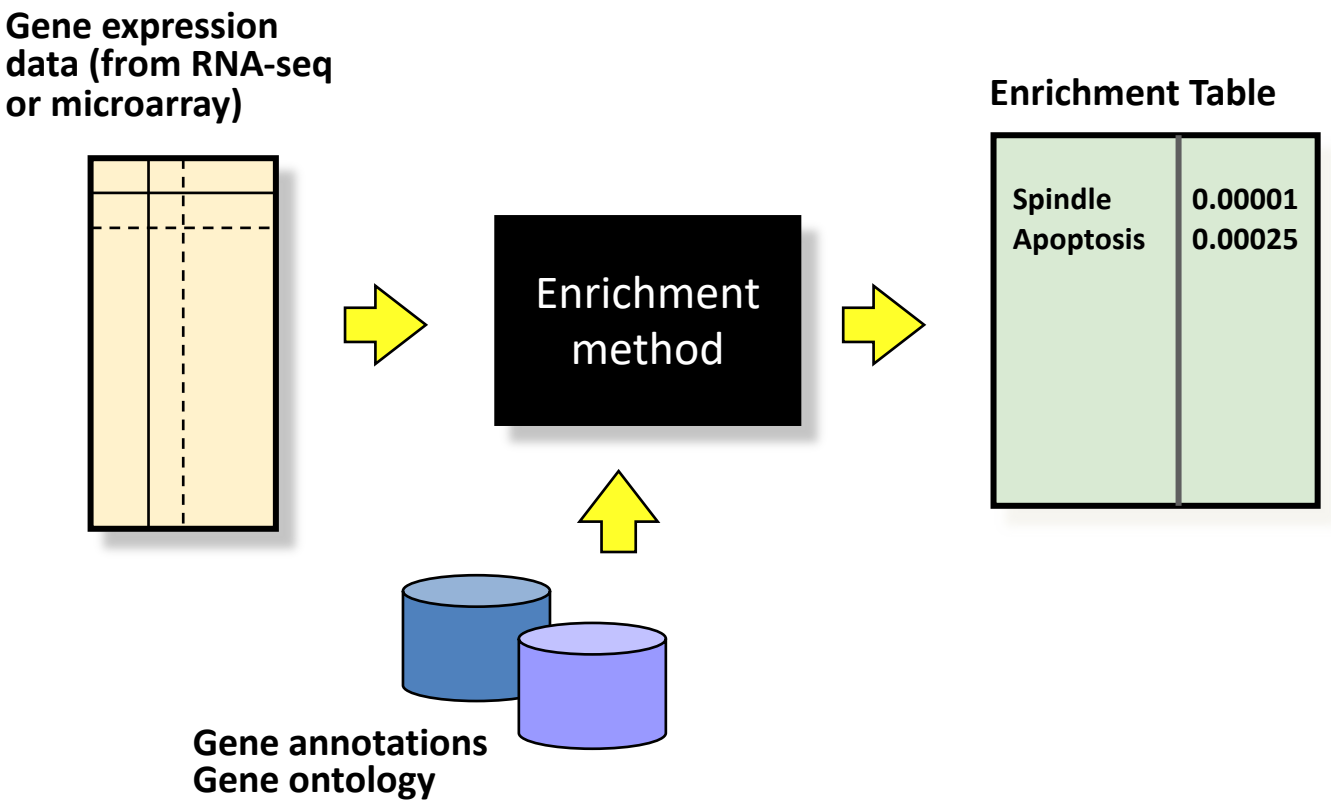
Annie Lu
@zhiyuanlu_annie

Linh Huynh
@vietlinh_huynh

Davide Chicco
@davidechicco_it

Sarah Bi
@h_s_b_i

Gene set enrichment



Adapted from Quaid Morris

Enrichment for arbitrary genomic regions

Genomic region (including noncoding) list

chr1:32169580-32169730
chr1:20656800-20656950
chr1:20992300-20992450
chr1:21011700-21011850
chr1:21103160-21103310
chr1:21900720-21900870
…

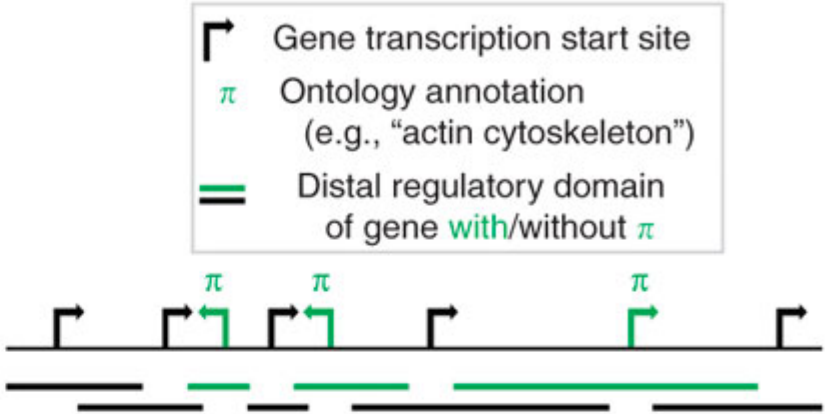
Gene list

DDX5
KLHL41
MEF2C
MEF2D
MYF6
MYOD1
MYOG
PAX3
PAX7
SOX8

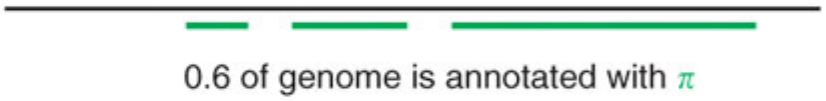
Enrichment in a non-gene context

Genomic Regions Enrichment of Annotations Tool (GREAT)

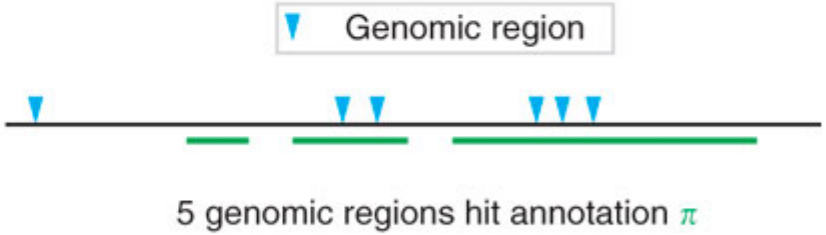
Step 1: Infer distal gene regulatory domains



Step 2: Calculate annotated fraction of genome

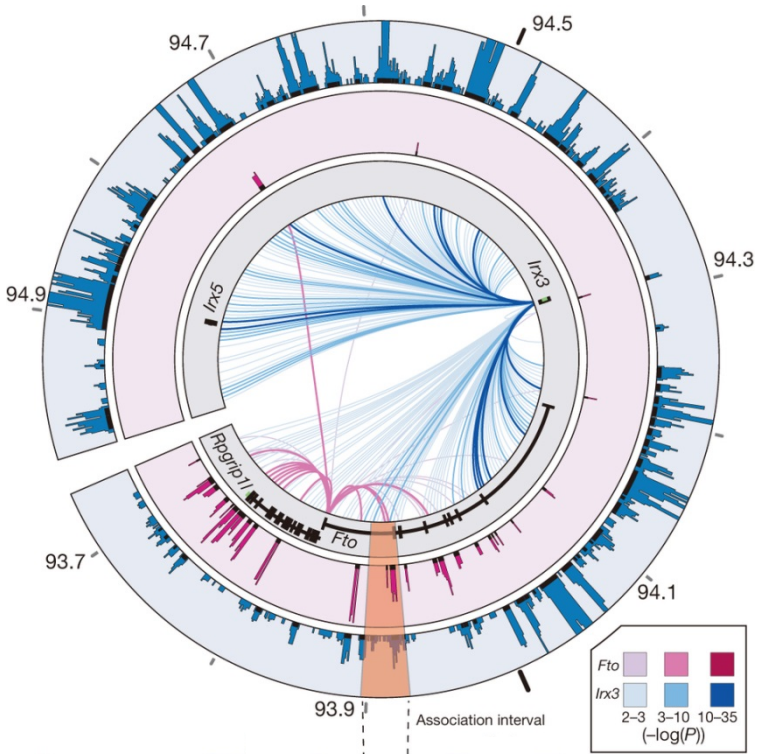


Step 3: Count genomic regions associated with the annotation



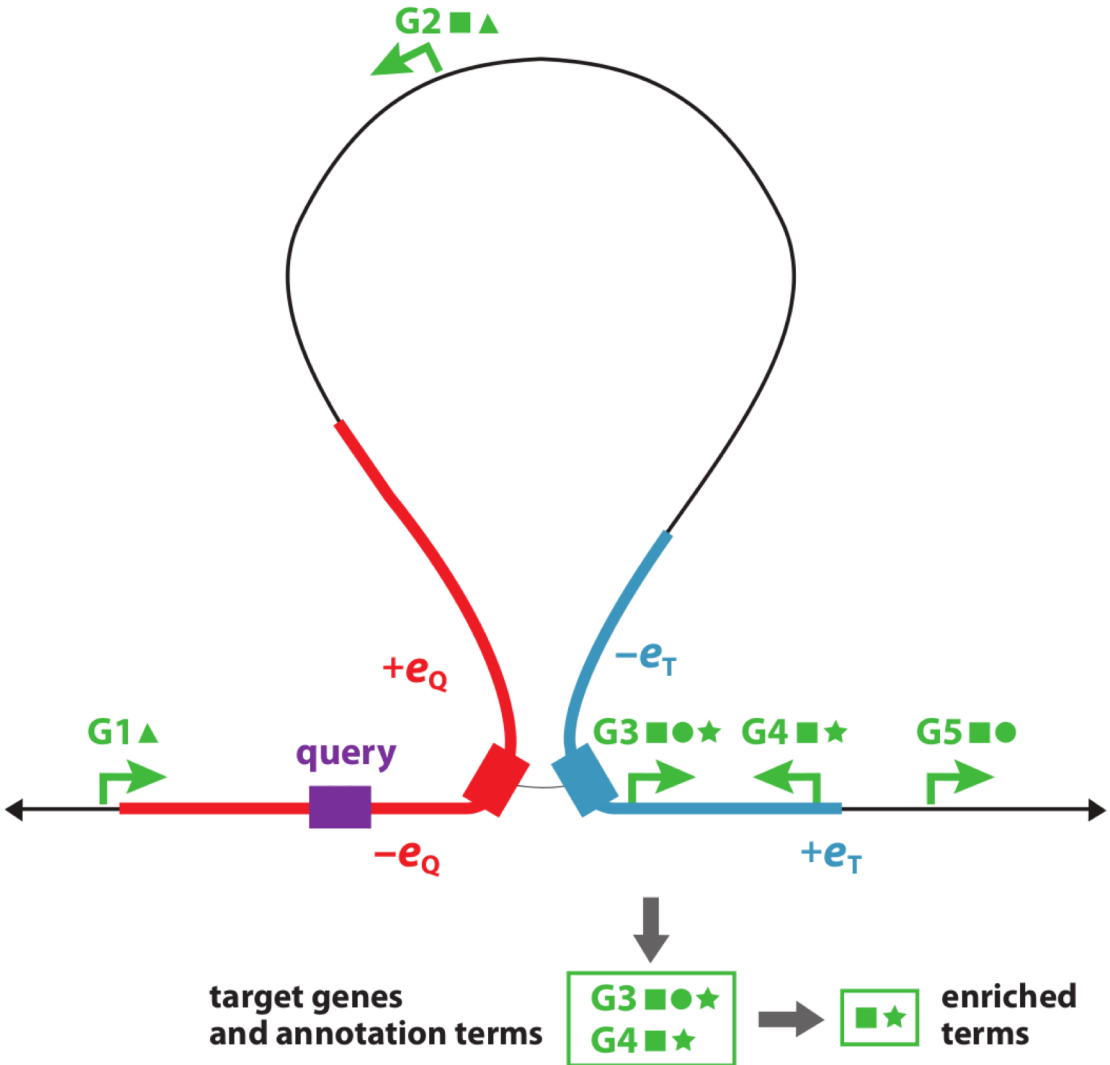
McLean et al. 2010. *Nat Biotechnol* 28:495.

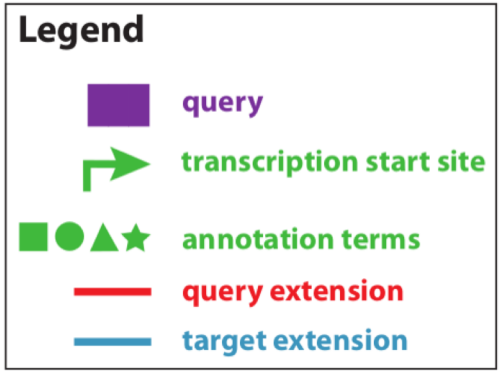
Regulatory elements & adjacent genes



Genome-wide association studies (GWAS) have reproducibly associated variants within introns of *FTO* with increased risk for obesity and type 2 diabetes (T2D)[1–3]. Although the molecular mechanisms linking these noncoding variants with obesity are not immediately obvious, subsequent studies in mice demonstrated that *FTO* expression levels influence body mass and composition phenotypes[4–6]. However, no direct connection between the obesity-associated variants and *FTO* expression or function has been made[7–9]. Here we show that the obesity-associated noncoding sequences within *FTO* are functionally connected, at megabase distances, with the homeobox gene *IRX3*. The obesity-associated *FTO* region directly interacts with the promoters of *IRX3* as well as *FTO* in the human, mouse and zebrafish genomes. Furthermore, long-range enhancers within this region recapitulate aspects of *IRX3* expression, suggesting that the obesity-associated interval belongs to the regulatory landscape of *IRX3*. Consistent with this, obesity-associated single nucleotide polymorphisms are associated with expression of *IRX3*, but not *FTO*, in human brains. A direct link between *IRX3* expression and regula-

Smemo et al. 2014. *Nature* 507:371.


G2
+e_Q
−e_T
G1
query
G3
G4
G5
−e_Q
+e_T
target genes
and annotation terms
G3
G4
enriched
terms


Legend
query
transcription start site
annotation terms
query extension
target extension

BEST

Biological Enrichment of Sequence Targets

BEHST

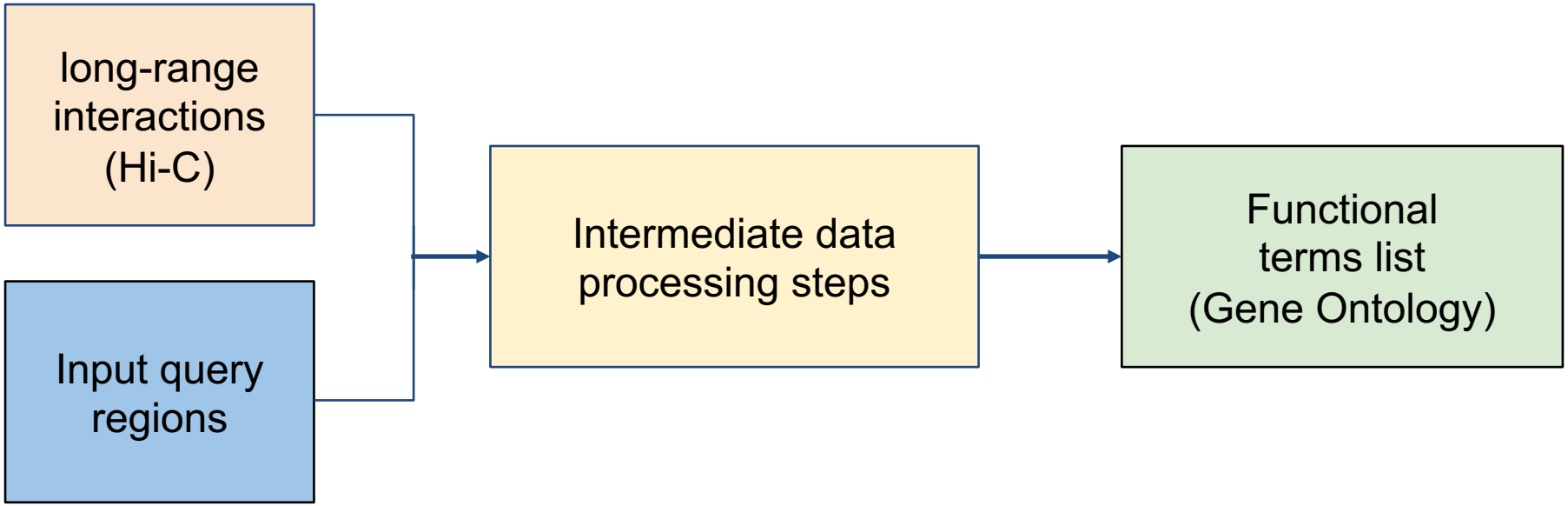
Biological Enrichment of Hidden Sequence Targets

BEHST

Biological Enrichment of Hidden Sequence Targets

Made in Canada! 🍁

BEHST workflow

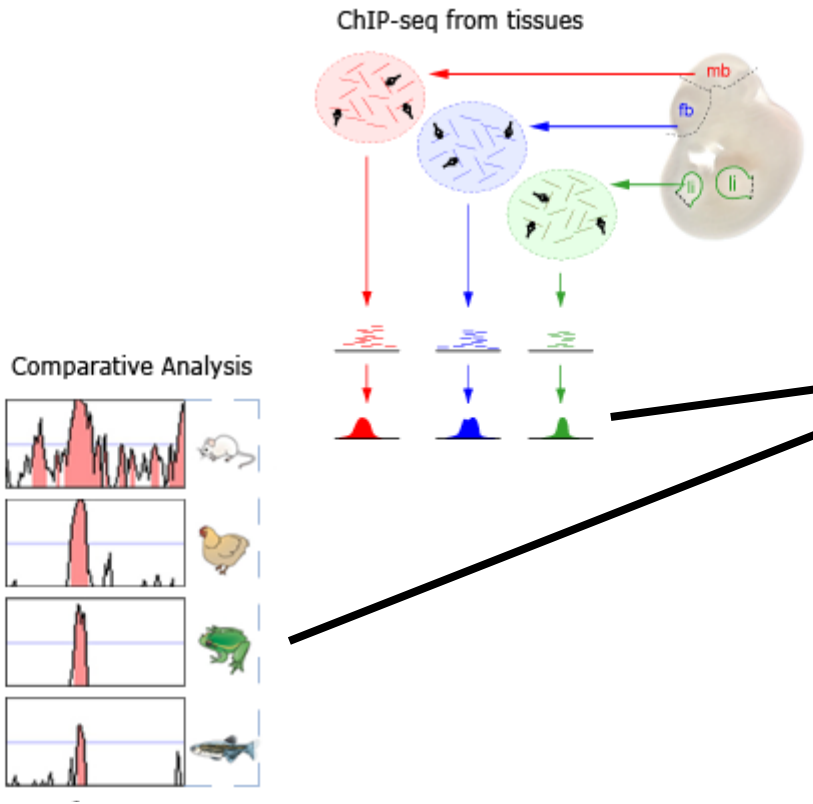


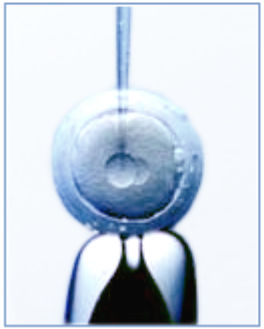
Hi-C datasets

cell type	description	# Hi-C interactions	mean interaction resolution (bp)
GM12878	B-lymphocyte lymphoblastoid	9 448	1 173 831
HeLa-S3	epithelioid cervical carcinoma	3 094	1 435 018
HMEC	mammary epithelial cell	5 152	215 167
HUVEC	umbilical vein endothelial cells	3 865	389 545
IMR90	fetal lung fibroblasts	8 040	416 673
K562	immortalized myelogenous leukemia	6 057	656 974
KBM7	chronic myelogenous leukemia	2 634	487 749
NHEK	normal epidermal keratinocytes	4 929	434 663
Union	**union of 8 cell types, excluding duplicates**	**34 367**	**742 691**

Rao et al. 2014. *Cell* 159:1665

Use case: E11.5 mouse enhancers



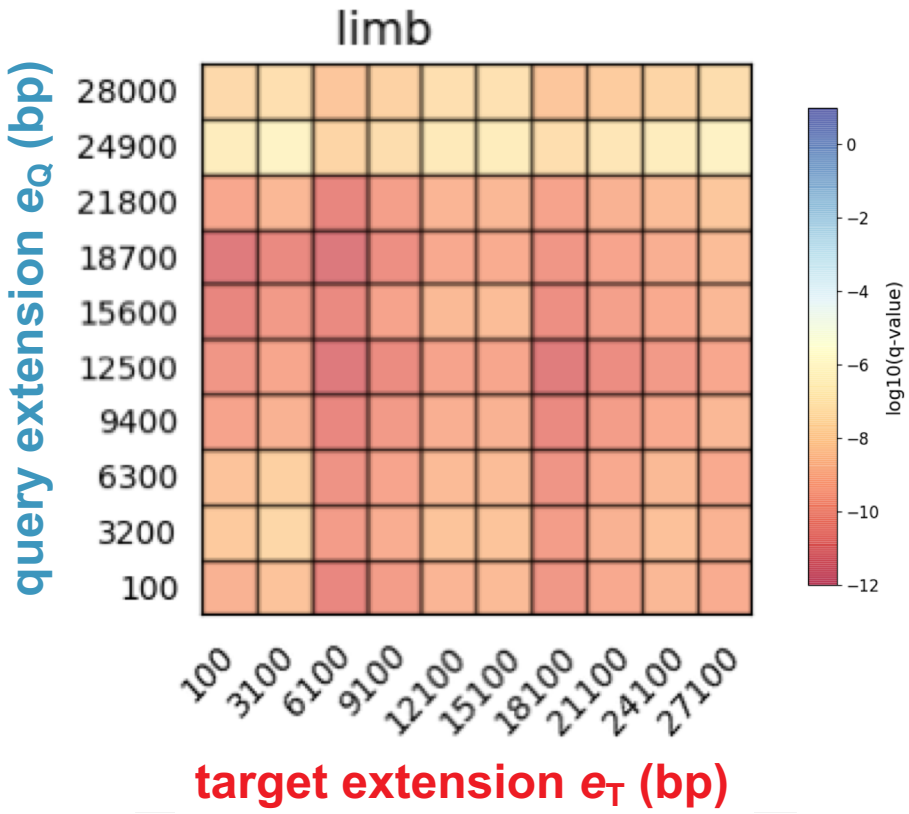


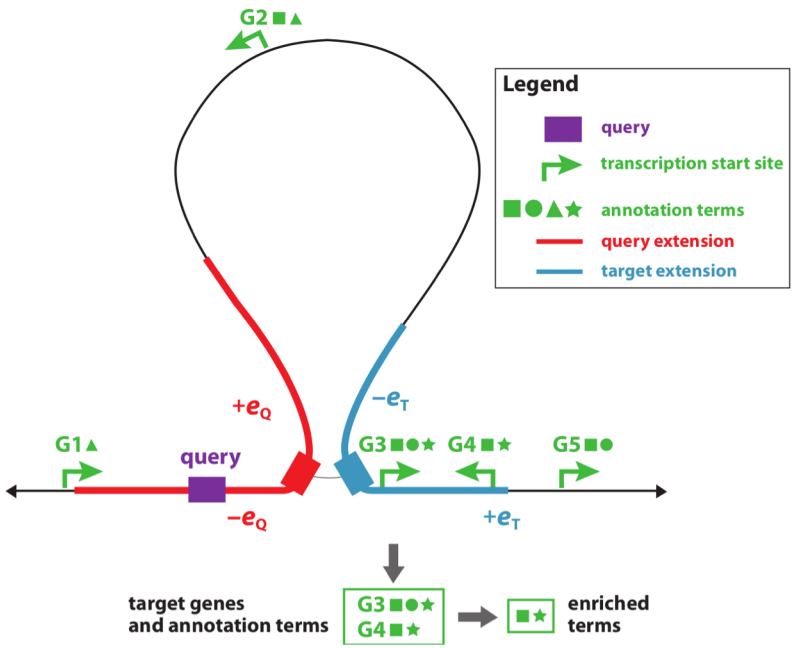
Mouse Egg Microinjection

E11.5 Reporter Staining

https://enhancer.lbl.gov/

Grid search of extension parameters



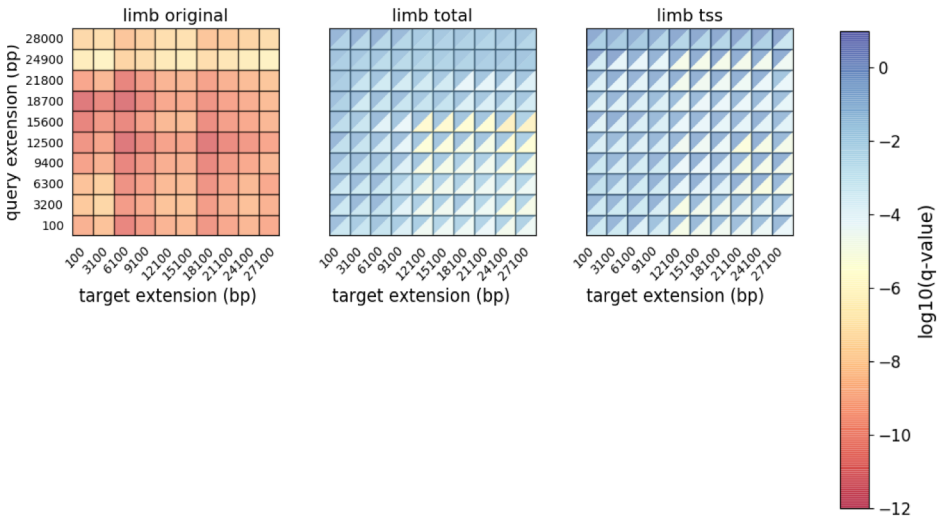


Shuffled controls

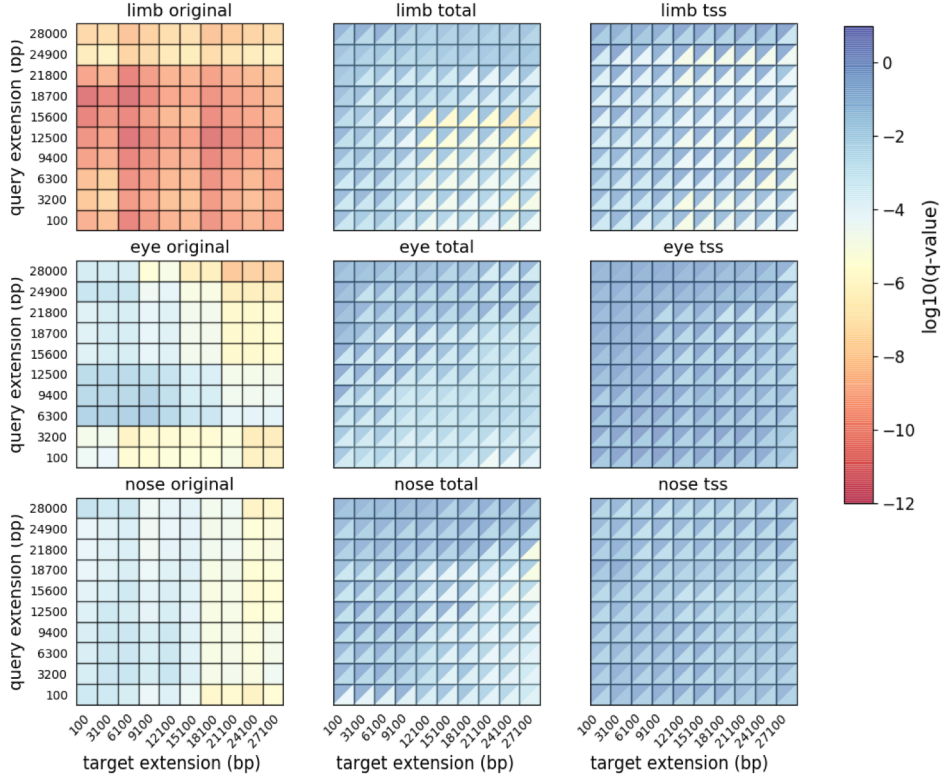
- Expectation: BEHST outputs more significant enrichment from original data than random data
- Applied BEHST to 7 sets of VISTA enhancers
- Compared with two shuffled negative controls:
  1. **Total shuffle**: randomly shuffle the enhancers across the whole genome
  2. **TSS shuffle**: shuffle in a way that preserved distance to the nearest transcription start site (TSS)

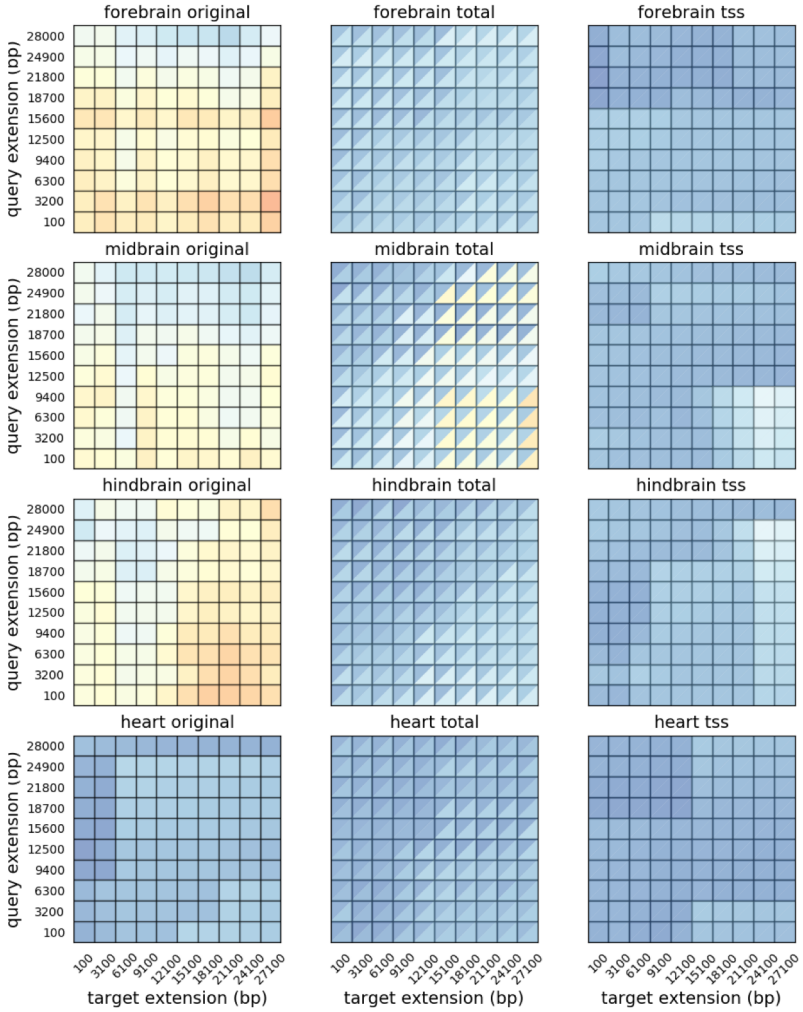
Comparing to shuffled controls

Comparing to shuffled controls



Comparing to shuffled controls





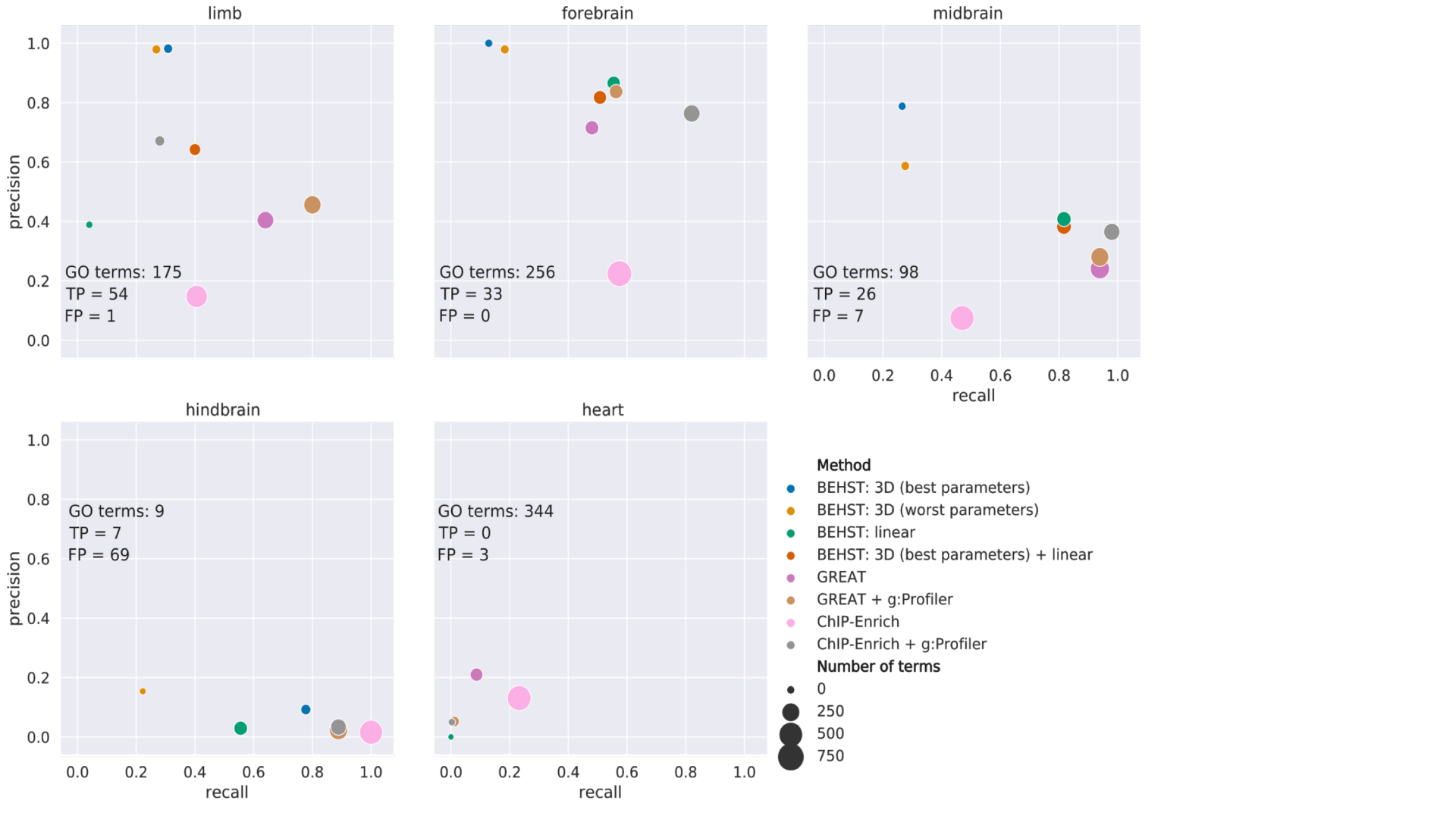
	sub-ontology	term ID	EF/UF	term name
8.21 [illegible] 14	MF	GO:0003700		sequence-specific DNA binding [illegible] iption factor activity
1.04×10 [illegible]	MF	GO:0001071		nucleic acid binding trans [illegible] factor activity
2.47×10^{-09}	[illegible]	GO:0072358	UF	cardiovascular syste [illegible] opment
3.00×10^{-09}		GO:0007507	UF	heart developmen [illegible]
1.06×10^{-08}	[illegible]	GO:0035108	EF	limb morphog [illegible]
1.07×10^{-08}	BP	[illegible]O:0060173	EF	limb develo [illegible]
1.21×10^{-08}	BP	[illegible]0045892		negative [illegible] on of transcription, DNA- [illegible] ent
1.27×10^{-08}	BP	G [illegible] 87		org [illegible] phogenesis
1.33×10^{-08}	BP	GO: [illegible]		[illegible] regulation of cellular macromolecule [illegible] nthetic process
2.88×10^{-08}	BP	GO:00512 [illegible]		[illegible] gative regulation of RNA metabolic process
3.31×10^{-08}	BP	GO:0035295		tube development
3.36×10^{-08}	BP	GO:0010629		negative regulation of gene expression
3.82×10^{-08}	BP	GO:001055 [illegible]		[illegible] gative regulation of macromolecule [illegible] nthetic process
7.97×10^{-08}	BP	GO: [illegible]	UF	[illegible] orphogenesis
9.81×10^{-08}	BP	G [illegible] 62		em [illegible] organ morphogenesis
1.42×10^{-07}	BP	[illegible] 0326	EF	embry [illegible] b morphogenesis
1.80×10^{-07}	BP	[illegible]0060562		epithelia [illegible] morphogenesis
2.19×10^{-07}	BP	[illegible]O:0035239		tube morph [illegible]
2.31×10^{-07}	M [illegible]	GO:0043565		sequence-speci [illegible] A binding
2.32×10^{-07}		GO:0060429		epithelium develo [illegible]
4.26×10^{-07}		GO:0000981		sequence-specific D [illegible] ling RNA polymerase II transcrip [illegible] tor activity
2.64×10 [illegible]	BP	GO:0048643	EF	regulation of skeletal mus [illegible] e development

Comparison between BEHST and other tools

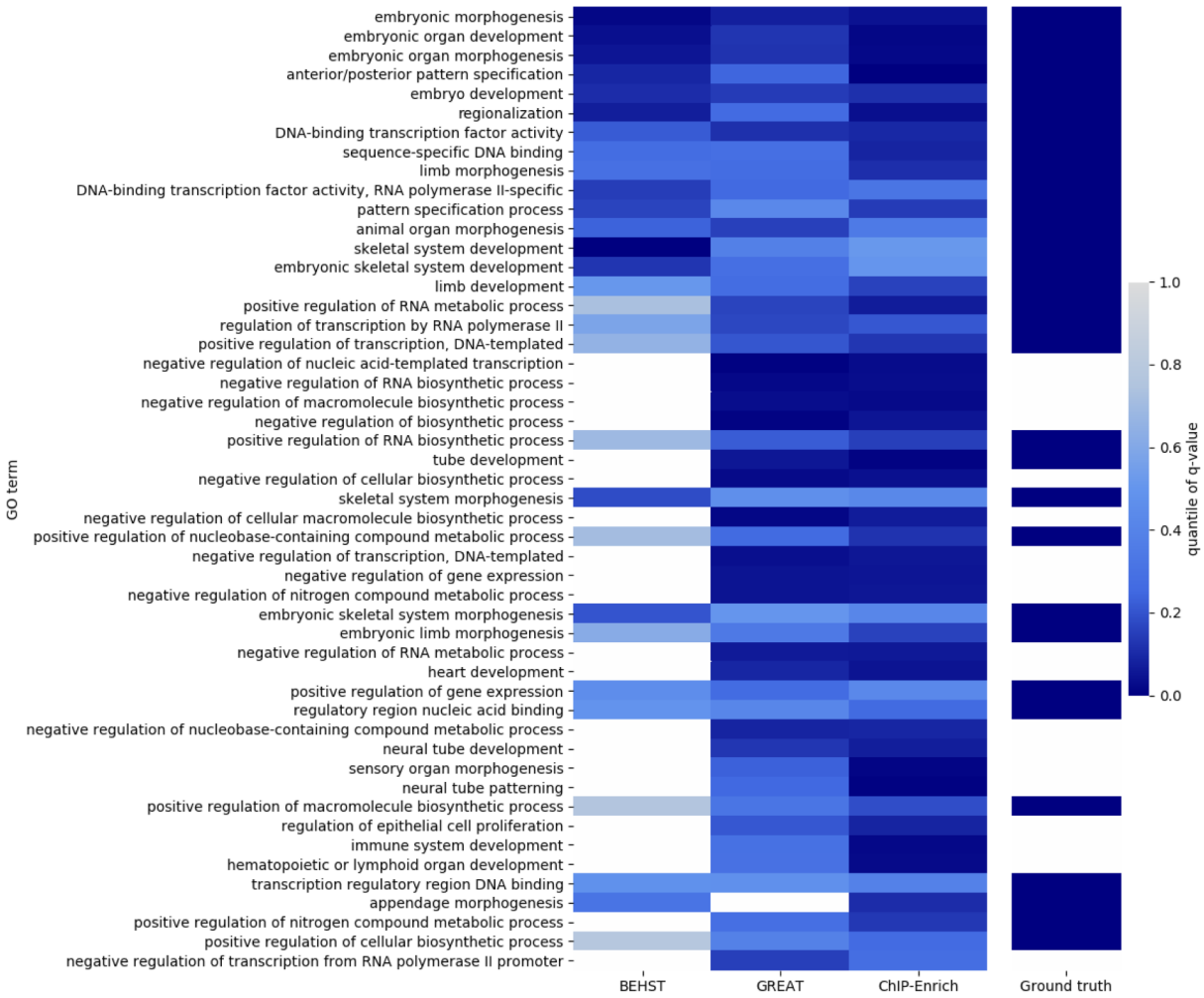
- Problem with old method: manually, biased, ad-hoc interpretation

Comparison between BEHST and other tools

- **New comparison**
  - Create a list of **ground-truth GO terms**
    - Choose tissue-specific genes from RNA-seq data
      - TPM > 1 and TPM > 5 $\langle TPM_{other} \rangle$
    - Run g:Profiler on these genes
  - Intersect the ground-truth GO term list with the GO terms from
    - BEHST
    - GREAT, GREAT-g:Profiler hybrid
    - ChIP-Enrich, ChIP-Enrich-g:Profiler hybrid
  - GO terms in both lists are true positive terms
  - GO terms only in output list but not ground-truth list are false positive terms


limb
forebrain
midbrain
hindbrain
heart
precision
recall
GO terms: 175
TP = 54
FP = 1
GO terms: 256
TP = 33
FP = 0
GO terms: 98
TP = 26
FP = 7
GO terms: 9
TP = 7
FP = 69
GO terms: 344
TP = 0
FP = 3
0.0
0.2
0.4
0.6
0.8
1.0
Method
BEHST: 3D (best parameters)
BEHST: 3D (worst parameters)
BEHST: linear
BEHST: 3D (best parameters) + linear
GREAT
GREAT + g:Profiler
ChIP-Enrich
ChIP-Enrich + g:Profiler
Number of terms
0
250
500
750

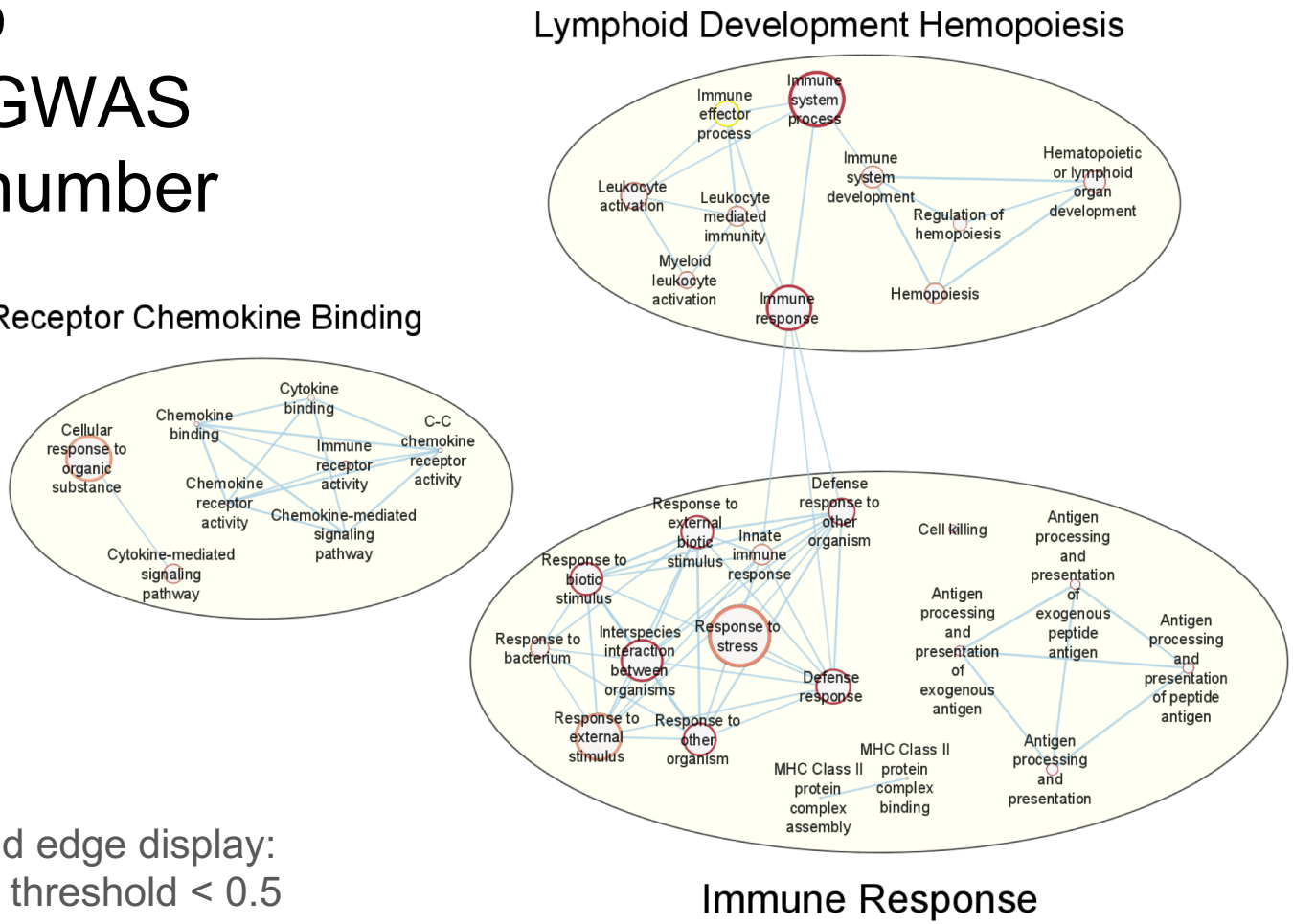
GO BP terms found by three methods



UK Biobank GWAS Data

- Get 17 anthropometric and blood-panel traits in the UK Biobank
- Select positions where p-value of beta-meta significance test < 10^-8
- Add eQ = 1000 bp to the single positions and run BEHST
- Find clusters of gene sets with Enrichment Map

Application to UK Biobank GWAS for Basophil number



Threshold for node and edge display:
p-value < 0.001, edge threshold < 0.5

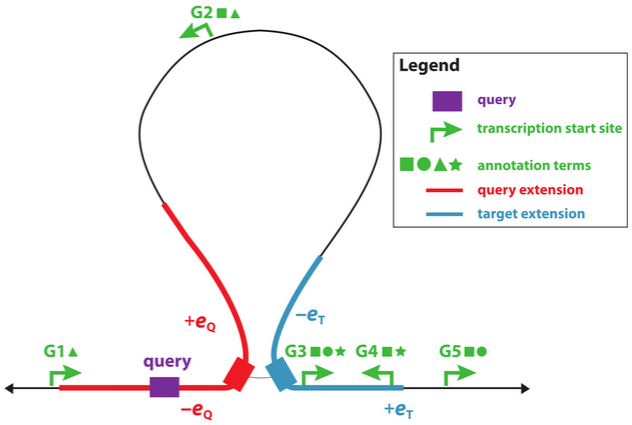
BEHST Biological Enrichment of Hidden Sequence Targets

Genomic set enrichment analysis enhanced through integration of chromatin long-range interactions

Chicco D, Bi HS, Reimand J, Hoffman MM. 2017. BEHST – Genomic set enrichment analysis enhanced through integration of chromatin long-range interactions. *In preparation.*

The free BEHST software package efficiently associates functional enriched Gene Ontology terms to input genomic regions

BEHST reads a dataset of genomic regions, and intersects them with the chromatin interactions available in the Hi-C dataset (Rao et al, Cell, 2014). Of these genomic regions, BEHST selects those that are present in the regulatory regions of genes a dataset of principal isoform annotations. We defined these cis-regulatory regions upon the position of their nearest transcription start site of the genes' principal transcripts, plus an upstream and downstream extension. Afterwards, BEHST selects the genes of the resulting partner loci found in gene regulatory regions, and inserts them into g:Profiler. BEHST, finally, produces the list of the most significant Gene Ontology terms detected by g:Profiler.



Installation

BEHST can run on any Linux and Mac computers. You can find the

behst.hoffmanlab.org/

behst.hoffmanlab.org

Upload your files

Query regions

File upload/ URL (.bed)

https://www.pmgenomics.ca/hoffmanlab/proj/behst/data/pressto_LUNG_enhancers.bed Upload file

Submit

Optional parameters and files:

Query extension (bp) 24100

Target extension (bp) 9400

Gene annotation (.gtf)

http://behst.hoffmanlab.org/data/gencode.v19.annotation_withproteinids.gtf Upload file

Chromosomal interactions (.hiccups)

http://behst.hoffmanlab.org/data/hic_8celltypes.hiccups Upload file

Principal transcripts (.bed)

http://behst.hoffmanlab.org/data/appris_data_principal.bed Upload file

Reset

g:Profiler - functional profiling o

biit.cs.ut.ee/gprofiler/index.cgi

Search

g:Profiler

- **g:GOSt** Gene Group Functional Profiling
- **g:Cocoa** Compact Compare of Annotations
- **g:Convert** Gene ID Converter
- **g:Sorter** Expression Similarity Search
- **g:Orth** Orthology search
- **g:SNPense** Convert rsID

Welcome! | Contact | FAQ | R / APIs | Beta | Archive

J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, J. Vilo: g:Profiler -- a web server for functional interpretation of gene lists (2016 update) Nucleic Acids Research 2016; doi: 10.1093/nar/gkw199 (PDF, more)

[?] Organism

Homo sapiens

[?] Query (genes, proteins, probes)

ENSG00000000938 ENSG00000003393
ENSG00000003400 ENSG00000003402
ENSG00000005844 ENSG00000006059
ENSG00000007933 ENSG00000013583
ENSG00000013588 ENSG00000019186
ENSG00000020181 ENSG00000022556
ENSG00000024422 ENSG00000025708
ENSG00000025770 ENSG00000026103
ENSG00000026751 ENSG00000028528

[?] or Term ID:

g:Profile! Clear

Example or random query

g:Profiler version r1732_e89_eg36. Version info

Options

- [?] ☑ Significant only
- [?] ☐ Ordered query
- [?] ☑ No electronic GO annotations
- [?] ☐ Chromosomal regions
- [?] ☐ Hierarchical sorting
- [?] Hierarchical filtering: Best per parent group (strong)
- [?] Output type: Graphical (PNG)

Show advanced options

- ☐ [?] Gene Ontology ☑ *Biological process* ☐ *Cellular component* ☐ *Molecular function*
  - Inferred from experiment [IDA, IPI, IMP, IGI, IEP]
  - Direct assay [IDA] / Mutant phenotype [IMP]
  - Genetic interaction [IGI] / Physical interaction [IPI]
  - Traceable author [TAS] / Non-traceable author [NAS] / Inferred by curator [IC]
  - Expression pattern [IEP] / Sequence or structural similarity [ISS] / Genomic context [IGC]
  - Biological aspect of ancestor [IBA] / Rapid divergence [IRD]
  - Reviewed computational analysis [RCA] / Electronic annotation [IEA]
  - No biological data [ND] / Not annotated or not in background [NA]
- Biological pathways ☐ KEGG ☐ Reactome
- [?] Regulatory motifs in DNA ☐ TRANSFAC TFBS ☐ miRBase microRNAs
- [?] Protein databases ☐ Human Protein Atlas ☐ CORUM protein complexes
- ☐ [?] Human Phenotype Ontology (sequence homologs in other species)
- ☐ [?] BioGRID protein-protein interactions

>> g:Convert	>> g:Orth	>> g:Sorter	>> g:Cocoa	>> Static URL
Gene ID Converter	Orthology Search	Expression Similarity Search	Compact Compare of Annotations	or generate compact link

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000	ENSG0000

g:Profiler - functional profiling o

biit.cs.ut.ee/gprofiler/index.cgi

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value	ENSG00000000938	ENSG00000003393	ENSG00000003400	ENSG00000003402	ENSG00000005844	ENSG00000006059	ENSG00000007933	ENSG00000013583	ENSG00000013588	ENSG00000019186	ENSG00000020181	ENSG00000022556	ENSG00000024422	ENSG00000025708	ENSG00000025770	ENSG00000026103	ENSG00000026751	ENSG00000028528	ENSG00000035664	ENSG00000037241	ENSG00000039650	ENSG00000039987	ENSG00000042753	ENSG00000049089
BP	cornification	GO:0070268	107	621	39	1.17e-25						A																		
BP	keratinization	GO:0031424	195	621	42	9.97e-18						A																		
BP	keratinocyte differentiation	GO:0030216	254	621	45	1.41e-15						A																		
BP	epidermis development	GO:0008544	327	621	51	1.88e-15						A																		
BP	epidermal cell differentiation	GO:0009913	276	621	46	7.04e-15						A																		
BP	skin development	GO:0043588	301	621	47	4.41e-14						A																		
BP	epithelial cell differentiation	GO:0030855	507	621	61	2.95e-13						A																		
BP	epithelium development	GO:0060429	773	621	76	3.82e-12						A																	A	
BP	complement activation, classical pathway	GO:0006958	112	621	24	3.45e-09																								
BP	cell death	GO:0008219	1490	621	108	3.84e-09			A-	D-		A						D				D-			M-					
BP	programmed cell death	GO:0012501	1397	621	103	4.99e-09			A-	D-		A						D				D-			M-					
BP	humoral immune response mediated by circulating immunoglobulin	GO:0002455	115	621	24	6.33e-09																								
BP	complement activation	GO:0006956	162	621	28	1.25e-08																								
BP	immunoglobulin mediated immune response	GO:0016064	146	621	26	3.58e-08																								
BP	B cell mediated immunity	GO:0019724	148	621	26	4.93e-08																								
BP	tissue development	GO:0009888	1208	621	90	8.96e-08				S		A																	A	
BP	protein activation cascade	GO:0072376	184	621	28	2.85e-07																								
BP	regulation of humoral immune response	GO:0002920	112	621	20	7.61e-06																								
BP	lymphocyte mediated immunity	GO:0002449	234	621	29	1.79e-05																	a							
BP	immune response	GO:0006955	1768	621	109	4.07e-05	S-				M-											Ba	A-							
BP	animal organ development	GO:0048513	1812	621	111	4.13e-05				S		A																	A	
BP	regulation of complement activation	GO:0030449	103	621	18	6.59e-05																								
BP	regulation of protein activation cascade	GO:2000257	103	621	18	6.59e-05																								
BP	humoral immune response	GO:0006959	295	621	32	7.87e-05																								
BP	adaptive immune response based on somatic recombination of immune receptors bui ...	GO:0002460	227	621	27	1.41e-04																								
BP	phagocytosis, recognition	GO:0006910	58	621	13	2.68e-04																								
BP	immune system process	GO:0002376	2329	621	131	2.77e-04	S-				M-											Ba	A-		M				A	
BP	adaptive immune response	GO:0002250	265	621	29	2.86e-04																								
BP	regulation of acute inflammatory response	GO:0002673	127	621	19	3.61e-04																								
BP	regulation of immune system process	GO:0002682	1179	621	78	3.71e-04	S-				A												A		M					
BP	positive regulation of immune response	GO:0050778	670	621	52	4.70e-04	S-																							
BP	activation of immune response	GO:0002253	578	621	47	4.84e-04	A																							
BP	positive regulation of immune system process	GO:0002684	873	621	62	7.21e-04	S-																		M					
BP	phagocytosis	GO:0006909	263	621	28	8.35e-04	S-																							

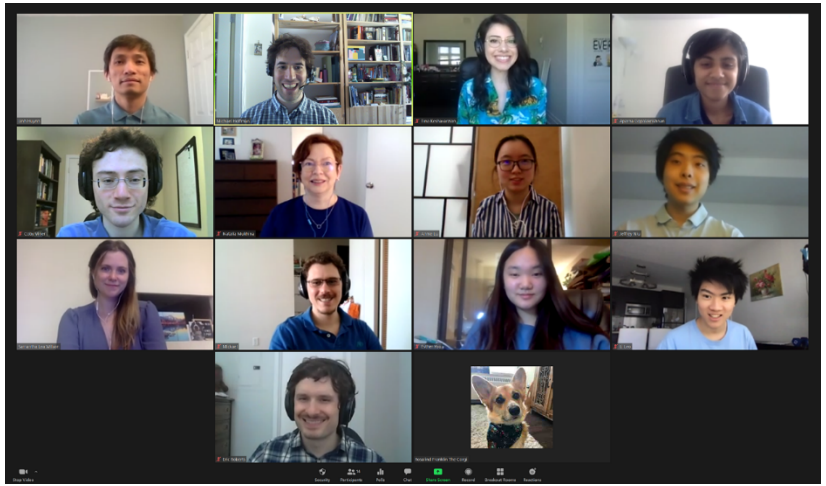
Concept and methodology in the preprint:
https://doi.org/fm2z

New evaluation procedure, GWAS applications:

Revised preprint coming soon!

Acknowledgments

The Hoffman Lab



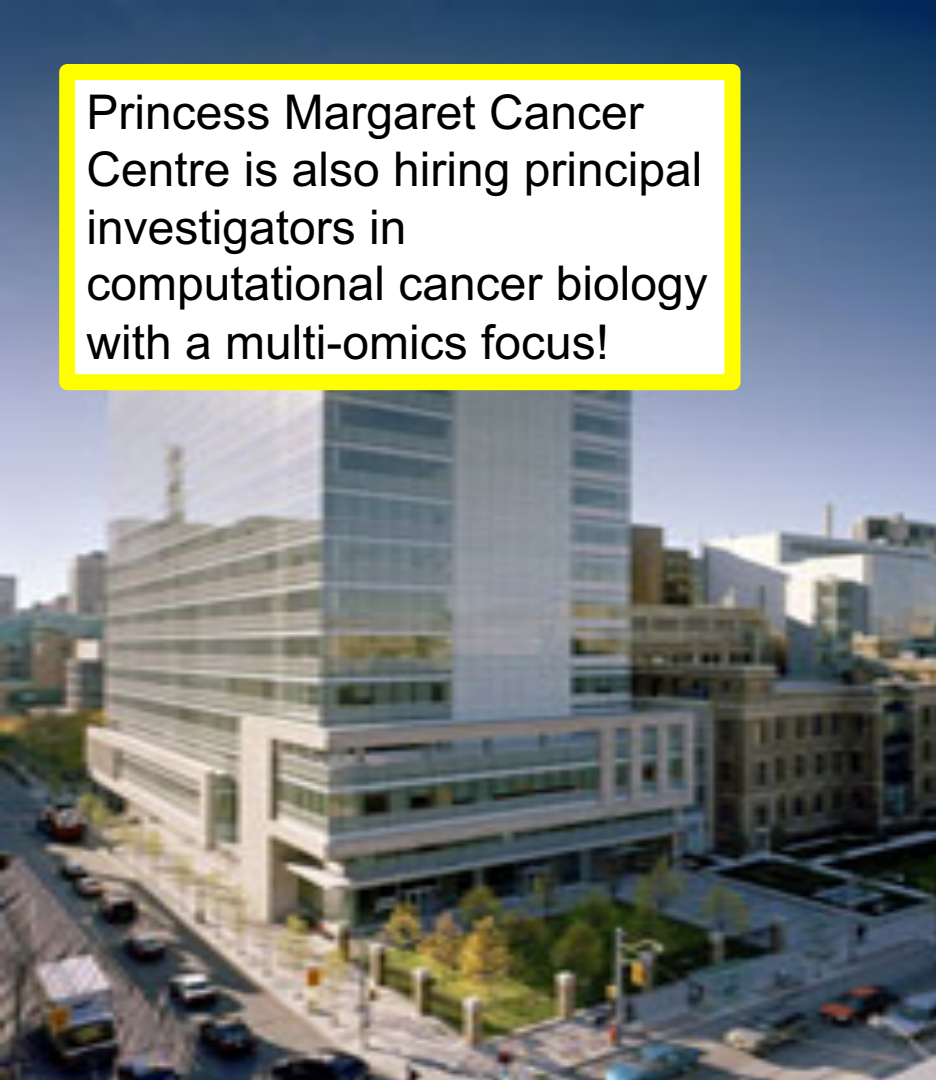
Samantha Wilson
Eric Roberts
Mickaël Mendez
Annie Lu
Leo Li
Linh Huynh
Coby Viner
Jeffrey Niu
Aparna Gopalakrishnan
Esther Yu

Natalia Mukhina

Davide Chicco
Sarah Bi
Jüri Reimand
Hae Kyung Im
Wail Ba-Alawi
Anna Narday
Zhibin Lu
Carl Virtanen

Funding

Canadian Institutes of Health Research; Princess Margaret Cancer Foundation; Natural Sciences and Engineering Research Council of Canada; Ontario Institute for Cancer Research; Ontario Ministry of Economic Development, Job Creation and Trade; Medicine by Design; McLaughlin Centre



Postdoctoral, MSc, PhD positions available in my research lab at the

Princess Margaret Cancer Centre

Dept of Medical Biophysics
Dept of Computer Science
University of Toronto

Please approach me for details.

Michael Hoffman
https://hoffmanlab.org/

michael.hoffman@utoronto.ca

@michaelhoffman