

utpnet: variant-transcription factor-phenotype networks

VJ Carey

April 27, 2020

1 Introduction

In a wide-ranging paper (PMID 22955828 Maurano et al. (2012)), Maurano and colleagues illustrate the concept of “common networks for common diseases” with a bipartite graph. One class of nodes is a set of autoimmune disorders, the other class is a set of transcription factors (TFs). In this graph, an edge exists between a disorder node and a TF node if a SNP that is significantly associated with the risk of the disorder lies in a genomic region possessing a strong match to the binding motif of the TF. This package defines tools to investigate the construction and statistical interpretation of such bipartite graphs, which we will denote VTP (variant-transcription factor-phenotype) networks.

2 Illustrative example of an unpruned VTP

The following code uses the `graphNEL` class to construct an approximation to the complete bipartite graph underlying Figure 4A of the Maurano paper; Figure 1 illustrates an arbitrary complete subgraph. The elements of `diseaseTags` are formatted to allow multiline rendering of the strings in node displays. It will be useful to distinguish a display token type and an analysis token type to simplify programming.

```
> #  
> # tags formatted for display  
> #  
> diseaseTags = c("Ankylosing\\nspondylitis", "Asthma",  
+ "Celiac\\ndisease", "Crohn's\\ndisease",  
+ "Multiple\\nsclerosis", "Primary\\nbiliary\\ncirrhosis",  
+ "Psoriasis", "Rheumatoid\\narthrititis",  
+ "Systemic\\nlupus\\nerythematosus",  
+ "Systemic\\nsclerosis", "Type 1\\ndiabetes",
```

```

+       "Ulcerative\\nocolitis"
+ )
> TFtags = c("ELF3", "MEF2A", "TCF3", "PAX4", "STAT3",
+   "ESR1", "POU2F1", "STAT1", "YY1", "SP1", "CDC5L",
+   "NR3C1", "EGR1", "PPARG", "HNF4A", "REST", "PPARA",
+   "AR", "NFKB1", "HNF1A", "TFAP2A")
> # define adjacency matrix
> adjm = matrix(1, nr=length(diseaseTags), nc=length(TFtags))
> dimnames(adjm) = list(diseaseTags, TFtags)
> library(graph)
> cvtp = ugraph(aM2bpG(adjm)) # complete (V)TP network; variants not involved yet

```

3 Data on GWAS variants: their associated phenotype, locations, and other characteristics

We will use the GWAS data provided at <https://www.sciencemag.org/content/suppl/2012/09/04/science.1222794.DC1/1222794-Maurano-tableS2.txt>, which was manually imported to a GRanges instance in hg19 origin-1 coordinates.

```

> library(vtpnet)
> data(maurGWAS)
> length(maurGWAS)

[1] 5654

> names(values(maurGWAS))

[1] "name"                "disease_trait"
[3] "disease_class"       "internally_replicated"
[5] "independently_replicated" "In_DHS"
[7] "fetal_origin"        "X.LOG.P."
[9] "sample_size"

```

4 Data on transcription factor binding sites

We have included the result of using FIMO Grant et al. (2011) to scan for motif matches for TF PAX4 as modeled in the Bioconductor *MotifDb* collection. The `-max-stored-scores` parameter was set to 10000000 so that p of up to 10^{-4} are retained.

```

> data(pax4)
> length(pax4)

```

```

> library(Rgraphviz)
> #flat = function(x, g) c(x, edges(g)[[x]])
> #sub = subGraph(unique(c(flat("Crohn's\\ndisease", cvtp),
> #   flat("Ulcerative\\ncolitis", cvtp))), cvtp)
> sub = subGraph(unique(c(diseaseTags[1:4], TFtags[1:6])), cvtp)
> plot(sub, attrs=list(node=list(shape="box", fixedsize=FALSE)))
> #plot(cvtp, attrs=list(graph=list(margin=c(.5,.5), size=c(4.1,4.1)),
> #   node=list(shape="box", fixedsize=FALSE, height=1)))

```

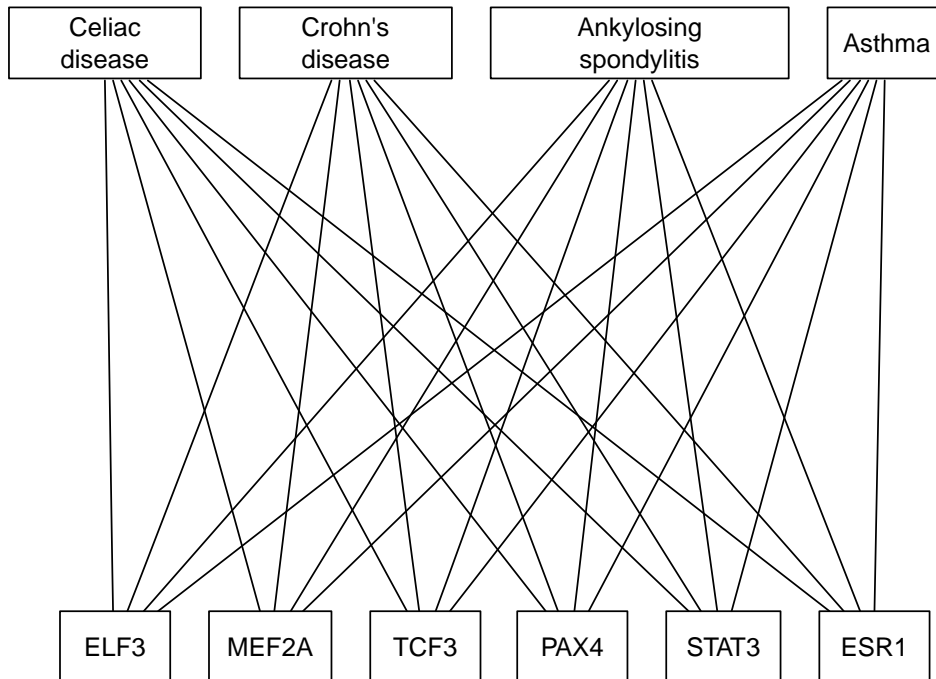


Figure 1: A complete bipartite graph for arbitrarily selected subsets of the autoimmune disorders and TFs found in Figure 4A of Maurano et al.

```
[1] 1862156
```

```
> pax4[1:4]
```

GRanges object with 4 ranges and 8 metadata columns:

	seqnames <Rle>	ranges <IRanges>	strand <Rle>	source <factor>	type <factor>	score <numeric>
[1]	chr1	10273-10302	+	fimo	nucleotide_motif	999.917
[2]	chr1	10279-10308	+	fimo	nucleotide_motif	999.962
[3]	chr1	11703-11732	-	fimo	nucleotide_motif	999.999
[4]	chr1	11704-11733	-	fimo	nucleotide_motif	999.955

	phase <integer>	Name <character>	pvalue <character>	qvalue <character>
[1]	<NA>	+Mmusculus-JASPAR_CORE-Pax4-MA0068.1	8.35e-05	0.396
[2]	<NA>	+Mmusculus-JASPAR_CORE-Pax4-MA0068.1	3.79e-05	0.361
[3]	<NA>	-Mmusculus-JASPAR_CORE-Pax4-MA0068.1	8.04e-07	0.194
[4]	<NA>	-Mmusculus-JASPAR_CORE-Pax4-MA0068.1	4.46e-05	0.368

	sequence <character>
[1]	TAACCCTAACCCCTAACCCCAACCCCAACCC
[2]	TAACCCTAACCCCAACCCCAACCCCAACCC
[3]	AAAAAAATACACATGGCCAGGCCCCAGCCC
[4]	TAAAAAAATACACATGGCCAGGCCCCAGCC

seqinfo: 92 sequences from an unspecified genome; no seqlengths

We can also generate our own motif-match ranges. Here is an example of a parallelized search against hg19 using `matchPWM`.

```
> library(foreach)
> library(doParallel)
> registerDoParallel(cores=12)
> library(BSgenome.Hsapiens.UCSC.hg19)
> library(MotifDb)
> sn = seqnames(Hsapiens)[1:24]
> pax4 = query(MotifDb, "pax4")[[1]]
> ans = foreach(i=1:24) %dopar% {
+   cat(i)
+   subj = Hsapiens[[ sn[i] ]]
+   matchPWM( pax4, subj, "75%" )
+ }
> pax4_75 =
+ do.call(c, lapply(1:length(ans), function(x)
```

```
+ {GRanges(sn[x], as(ans[[x]], "IRanges"))})}
> save(pax4_75, file="pax4_75.rda")
```

Results of such searches retaining matches at scores of 85% and 75% of the maximum achievable score have been stored with this package.

5 Building a VTP network: one edge per phenotype

5.1 Raw matches

We can survey the entire GWAS catalog for intersection with putative PAX4 binding sites. First the two Bioconductor internal binding site sets.

```
> data(pax4_85)
> vp_pax4_85 = maurGWAS[ overlapsAny(maurGWAS, pax4_85) ]
> length(vp_pax4_85)
```

```
[1] 0
```

```
> data(pax4_75)
> vp_pax4_75 = maurGWAS[ overlapsAny(maurGWAS, pax4_75) ]
> length(vp_pax4_75)
```

```
[1] 54
```

Then the FIMO-based set.

```
> vp_pax4_fimo = maurGWAS[ overlapsAny(maurGWAS, pax4) ]
> length(vp_pax4_fimo)
```

```
[1] 67
```

The lengths reported here are the numbers of phenotypes linked to PAX4 in a VTP according to various motif matching schemes. For the two non-null results, we have

```
> u75 = unique(vp_pax4_75$disease_trait)
> ufimo = unique(vp_pax4_fimo$disease_trait)
> length(setdiff(u75, ufimo))
```

```
[1] 23
```

```
> length(setdiff(ufimo, u75))
```

```
[1] 28
```

Clearly the identification of TP links is sensitive to the approach used to locate binding sites. However, as noted in the Maurano paper, the use of matching to the reference genome without SNP injection is potentially problematic.

5.2 Filtering

It is useful to restrict the phenotypes of interest, and to map them to broader classes, and to include TFBS matching scores for the purpose of filtering edges. Here we will use the NHGRI GWAS catalog against FIMO-based (reference genome matching only) PAX4 calls.

```
> data(cancerMap)
> library(gwascat)
> data(gwrngs19)
> cangw = filterGWASbyMap( gwrngs19, cancerMap )
> getOneHits( pax4, cangw, "fimo" )
```

GRanges object with 8 ranges and 41 metadata columns:

	seqnames	ranges	strand	Date.Added.to.Catalog	PUBMEDID
	<Rle>	<IRanges>	<Rle>	<character>	<integer>
3475	chr8	129194641	*	09/12/2013	23535729
3480	chr11	65583066	*	09/12/2013	23535729
6963	chr2	26526419	*	01/25/2013	23144319
7155	chr6	143943314	*	01/15/2013	23108145
7480	chr20	32588095	*	11/30/2012	22976474
12585	chrX	37854727	*	11/15/2010	20932654
13650	chr12	14653867	*	07/12/2010	20543847
15145	chr10	63752159	*	09/04/2009	19684604

	First.Author	Date	Journal
	<character>	<character>	<character>
3475	Michailidou K	04/01/2013	Nat Genet
3480	Michailidou K	04/01/2013	Nat Genet
6963	Lee Y	11/08/2012	Carcinogenesis
7155	Wang LE	10/29/2012	Cancer Res
7480	Siddiq A	09/13/2012	Hum Mol Genet
12585	Kerns SL	10/05/2010	Int J Radiat Oncol Biol Phys
13650	Turnbull C	06/13/2010	Nat Genet
15145	Papaemmanuil E	08/16/2009	Nat Genet

	Link
	<character>
3475	http://www.ncbi.nlm.nih.gov/pubmed/23535729
3480	http://www.ncbi.nlm.nih.gov/pubmed/23535729
6963	http://www.ncbi.nlm.nih.gov/pubmed/23144319
7155	http://www.ncbi.nlm.nih.gov/pubmed/23108145
7480	http://www.ncbi.nlm.nih.gov/pubmed/22976474
12585	http://www.ncbi.nlm.nih.gov/pubmed/20932654
13650	http://www.ncbi.nlm.nih.gov/pubmed/20543847

15145 <http://www.ncbi.nlm.nih.gov/pubmed/19684604>

3475		
3480		
6963		
7155		
7480		A meta
12585	Genome-wide association study to identify single nucleotide polymorphisms (SNPs)	
13650		
15145		
	Disease.Trait	
	<character>	
3475	Breast cancer	
3480	Breast cancer	
6963	Non-small cell lung cancer	
7155	Lung Cancer (DNA repair capacity)	
7480	Breast cancer	
12585	Erectile dysfunction and prostate cancer treatment	
13650	Testicular germ cell cancer	
15145	Acute lymphoblastic leukemia (childhood)	
3475		10,052 European a
3480		10,052 European a
6963		
7155		914 European ancestry non-small cell l
7480	3,666 European ancestry cases, 28,864 European ancestry controls, 1,004 African	
12585		27 Afri
13650		979 European
15145		907 European
3475		
3480		
6963		
7155		679 European
7480	562 European ancestry cases, 6,410 European ancestry controls, 84 Japanese ance	
12585		
13650		
15145		

Region Chr_id Chr_pos.hg38

	<character>	<character>	<numeric>	
3475	8q24.21	8	128182395	
3480	11q13.1	11	65815595	
6963	2p23.3	2	26303551	
7155	6q24.2	6	143622177	
7480	20q11.22	20	34000289	
12585	Xp11.4	23	37995474	
13650	12p13.1	12	14500933	
15145	10q21.2	10	61992400	
	Reported.Gene.s.		Mapped_gene	
	<character>		<character>	
3475	MIR1208, MYC		MIR1208 - LINC01263	
3480	DKFZp761E198, OVOL1, SNX32, CFL1, MUS81		OVOL1-AS1 - SNX32	
6963	GPR113		HADHB - GPR113	
7155	PHACTR2		PHACTR2	
7480	RALY, EIF2S2, ASIP		RALY	
12585	SYTL5		CXorf27 - SYTL5	
13650	ATF7IP		ATF7IP	
15145	ARID5B		ARID5B	
	Upstream_gene_id	Downstream_gene_id	Snp_gene_ids	Upstream_gene_distance
	<character>	<character>	<character>	<character>
3475	100302281	101927774		32.21
3480	101927828	254122		24.73
6963	3032	165082		13.09
7155	<NA>	<NA>	9749	<NA>
7480	<NA>	<NA>	22913	<NA>
12585	25763	94122		4.16
13650	<NA>	<NA>	55729	<NA>
15145	<NA>	<NA>	84159	<NA>
	Downstream_gene_distance	Strongest.SNP.Risk.Allele	SNPs	
	<character>		<character>	<character>
3475	222.87	rs11780156-T	rs11780156	
3480	18.24	rs3903072-G	rs3903072	
6963	4.62	rs6753473-G	rs6753473	
7155	<NA>	rs9390123-A	rs9390123	
7480	<NA>	rs2284378-T	rs2284378	
12585	11.11	rs872690-?	rs872690	
13650	<NA>	rs2900333-C	rs2900333	
15145	<NA>	rs7089424-C	rs7089424	
	Merged Snp_id_current	Context	Intergenic	
	<character>	<character>	<character>	<character>
3475	0	11780156	Intergenic	1

3480	0	3903072	Intergenic	1	
6963	0	6753473	Intergenic	1	
7155	0	9390123	intron	0	
7480	0	2284378	intron	0	
12585	0	872690	Intergenic	1	
13650	0	2900333	UTR-3	0	
15145	0	7089424	intron	0	
	Risk.Allele.Frequency	p.Value	Pvalue_mlog	p.Value..text.	OR.or.beta
	<character>	<numeric>	<numeric>	<character>	<numeric>
3475	0.16	3e-11	10.52288		1.07
3480	0.53	9e-12	11.04576		1.05
6963	0.052	4e-06	5.39794	(Additive model)	NA
7155	0.3957	7e-06	5.15490		NA
7480	0.31	1e-08	8.00000		1.16
12585	0.03	9e-06	5.04576		11.78
13650	0.62	6e-10	9.22185		1.27
15145	0.34	7e-19	18.15490		1.65
	X95..CI..text.		Platform..SNPs.passing.QC.		
	<character>		<character>		
3475	[1.04-1.10]	Illumina & Affymetrix	[~2.6 million]	(Imputed)	
3480	[1.04-1.08]	Illumina & Affymetrix	[~2.6 million]	(Imputed)	
6963	NR		Affymetrix	[271,817]	
7155	NR		Illumina	[303,669]	
7480	[1.10-1.22]		Illumina	[2,608,509]	(imputed)
12585	[NR]		Affymetrix	[512,497]	
13650	[1.12-1.44]		Illumina	[298,782]	
15145	[1.54-1.76]		Illumina	[291,473]	
	CNV	num.Risk.Allele.Frequency	dclass	score	tfstart
	<character>	<numeric>	<character>	<numeric>	<integer>
3475	N	0.1600	Breast	999.985	129194621
3480	N	0.5300	Breast	999.952	65583065
6963	N	0.0520	Lung	999.987	26526415
7155	N	0.3957	Lung	999.939	143943292
7480	N	0.3100	Breast	999.928	32588075
12585	N	0.0300	Prostate	999.903	37854721
13650	N	0.6200	Testicular	999.990	14653848
15145	N	0.3400	ALL (ped)	999.962	63752142
	tfend	pvalue	qvalue		
	<integer>	<numeric>	<numeric>		
3475	129194650	1.49e-05	0.318		
3480	65583094	4.83e-05	0.373		
6963	26526444	1.25e-05	0.310		

7155	143943321	6.13e-05	0.383
7480	32588104	7.16e-05	0.388
12585	37854750	9.72e-05	0.403
13650	14653877	1.05e-05	0.301
15145	63752171	3.79e-05	0.361

seqinfo: 23 sequences from hg19 genome

6 Appendix: generating the ALT-injected genome image

```
> altize = function(htag = "21",
+ #
+ # from sketch by Herve Pages, May 2013
+ #
+   slpack="SNPlocs.Hsapiens.dbSNP.20120608",
+   hgpack = "BSgenome.Hsapiens.UCSC.hg19",
+   faElFun = function(x) sub("%%TAG%", x, "alt%%TAG%%chr"),
+   faTargFun = function(x)
+     sub("%%TAG%", x, "alt%%TAG%%_hg19.fa")) {
+   require(slpack, character.only=TRUE)
+   require(hgpack, character.only=TRUE)
+   require("ShortRead", character.only=TRUE)
+   chk = grep("ch|chr", htag)
+   if (length(chk)>0) {
+     warning("clearing prefix ch or chr from htag")
+     htag = gsub("ch|chr", "", htag)
+   }
+   snpgettag = paste0("ch", htag)
+   ggettag = paste0("chr", htag)
+   cursnps = getSNPlocs(snpgettag, as.GRanges=TRUE)
+   curgenome = unmasked(Hsapiens[[ggettag]])
+   ref_allele =
+     strsplit(as.character(curgenome[start(cursnps)]),
+       NULL, fixed=TRUE)[[1L]]
+   all_alleles = IUPAC_CODE_MAP[cursnps$alleles_as_ambig]
+   alt_alleles = mapply( function(ref,all)
+     sub(ref, "", all, fixed=TRUE),
+     ref_allele, all_alleles, USE.NAMES=FALSE)
+   cursnps$ref_allele = ref_allele
+   cursnps$alt_alleles = alt_alleles
+   cursnps$one_alt = substr(cursnps$alt_alleles, 1, 1)
+   altg = list(replaceLetterAt(curgenome, start(cursnps),
```

```
+      cursnps$one_alt))
+      names(altg) = faElFun(chtag)
+      writeFasta(DNAStringSet(altg), file=faTargFun(chtag))
+ }
```

7 Session information

```
> sessionInfo()
```

```
R version 4.0.0 (2020-04-24)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Mojave 10.14.6
```

```
Matrix products: default
```

```
BLAS: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
```

```
locale:
```

```
[1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
attached base packages:
```

```
[1] stats4      grid      parallel  stats      graphics  grDevices  utils
[8] datasets   methods   base
```

```
other attached packages:
```

```
[1] vtpnet_0.28.0      doParallel_1.0.15  iterators_1.0.12
[4] foreach_1.5.0      gwascats_2.20.0    GenomicRanges_1.40.0
[7] GenomeInfoDb_1.24.0 IRanges_2.22.0     S4Vectors_0.26.0
[10] Rgraphviz_2.32.0   graph_1.66.0       BiocGenerics_0.34.0
```

```
loaded via a namespace (and not attached):
```

```
[1] Rcpp_1.0.4.6        lattice_0.20-41
[3] prettyunits_1.1.1   Rsamtools_2.4.0
[5] Biostrings_2.56.0   assertthat_0.2.1
[7] digest_0.6.25       BiocFileCache_1.12.0
[9] R6_2.4.1            RSQLite_2.2.0
[11] httr_1.4.1          ggplot2_3.3.0
[13] pillar_1.4.3        zlibbioc_1.34.0
[15] rlang_0.4.5         GenomicFeatures_1.40.0
[17] progress_1.2.2      curl_4.3
[19] blob_1.2.1          Matrix_1.2-18
[21] BiocParallel_1.22.0 stringr_1.4.0
```

[23] munsell_0.5.0	RCurl_1.98-1.2
[25] bit_1.1-15.2	biomaRt_2.44.0
[27] DelayedArray_0.14.0	compiler_4.0.0
[29] rtracklayer_1.47.0	pkgconfig_2.0.3
[31] askpass_1.1	openssl_1.4.1
[33] tidyselect_1.0.0	SummarizedExperiment_1.18.0
[35] tibble_3.0.1	GenomeInfoDbData_1.2.3
[37] codetools_0.2-16	matrixStats_0.56.0
[39] XML_3.99-0.3	crayon_1.3.4
[41] dplyr_0.8.5	dbplyr_1.4.3
[43] GenomicAlignments_1.24.0	bitops_1.0-6
[45] rappdirs_0.3.1	gtable_0.3.0
[47] lifecycle_0.2.0	DBI_1.1.0
[49] magrittr_1.5	scales_1.1.0
[51] stringi_1.4.6	XVector_0.28.0
[53] ellipsis_0.3.0	vctrs_0.2.4
[55] tools_4.0.0	bit64_0.9-7
[57] Biobase_2.48.0	glue_1.4.0
[59] purrr_0.3.4	hms_0.5.3
[61] colorspace_1.4-1	AnnotationDbi_1.50.0
[63] memoise_1.1.0	

8 Bibliography

References

- Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, 27(7):1017–8, Apr 2011. doi: 10.1093/bioinformatics/btr064.
- Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutayavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–5, Sep 2012. doi: 10.1126/science.1222794.