

vtpNet: variant-transcription factor-phenotype networks

VJ Carey

July 22, 2014

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\\ncolitis"
+ )
> 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)))

```

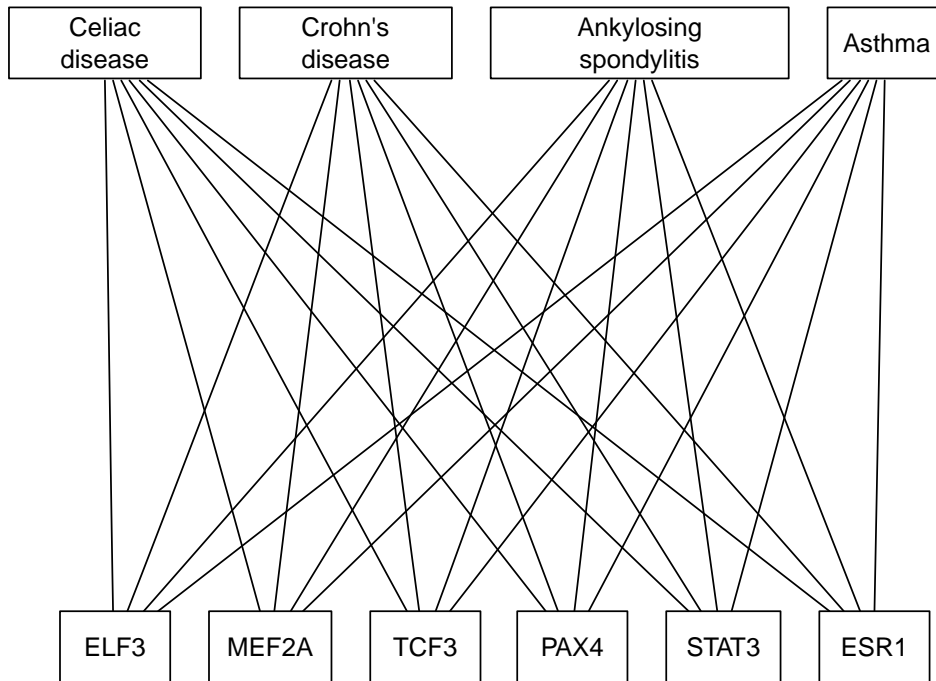


Figure 1: A complete bipartite graph for arbitrarily selected subsets of the autoimmune disorders and TFs found in Figure 4A of Maurano et al.

```
[1] 1862156
```

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

```
GRanges with 4 ranges and 8 metadata columns:
```

	seqnames	ranges	strand	source	type	score
	<Rle>	<IRanges>	<Rle>	<factor>	<factor>	<numeric>
[1]	chr1	[10273, 10302]	+	fimo	nucleotide_motif	999.9165
[2]	chr1	[10279, 10308]	+	fimo	nucleotide_motif	999.9621
[3]	chr1	[11703, 11732]	-	fimo	nucleotide_motif	999.9992
[4]	chr1	[11704, 11733]	-	fimo	nucleotide_motif	999.9554
	phase	Name	pvalue	qvalue		
	<integer>	<character>	<character>	<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]	AAAAAAAATACACATGGCCAGGCCCCAGCCC					
[4]	TAAAAAAAATACACATGGCCAGGCCCCAGCCC					

```
---
```

```
seqlengths:
```

chr1	chr10 ...	chrY
NA	NA ...	NA

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)
> cangw = filterGWASbyMap( gwrngs, cancerMap )
> getOneHits( pax4, cangw, "fimo" )
```

GRanges with 8 ranges and 41 metadata columns:

	seqnames	ranges	strand	Date.Added.to.Catalog	PUBMEDID
	<Rle>	<IRanges>	<Rle>	<character>	<integer>
[1]	chr8	[129194641, 129194641]	*	09/12/2013	23535729
[2]	chr11	[65583066, 65583066]	*	09/12/2013	23535729
[3]	chr2	[26526419, 26526419]	*	01/25/2013	23144319
[4]	chr6	[143943314, 143943314]	*	01/15/2013	23108145
[5]	chr20	[32588095, 32588095]	*	11/30/2012	22976474
[6]	chrX	[37854727, 37854727]	*	11/15/2010	20932654
[7]	chr12	[14653867, 14653867]	*	07/12/2010	20543847
[8]	chr10	[63752159, 63752159]	*	09/04/2009	19684604

	First.Author	Date	Journal
	<character>	<character>	<character>
[1]	Michailidou K	04/01/2013	Nat Genet
[2]	Michailidou K	04/01/2013	Nat Genet
[3]	Lee Y	11/08/2012	Carcinogenesis
[4]	Wang LE	10/29/2012	Cancer Res
[5]	Siddiq A	09/13/2012	Hum Mol Genet
[6]	Kerns SL	10/05/2010	Int J Radiat Oncol Biol Phys
[7]	Turnbull C	06/13/2010	Nat Genet
[8]	Papaemmanuil E	08/16/2009	Nat Genet

	Link
	<character>
[1]	http://www.ncbi.nlm.nih.gov/pubmed/23535729
[2]	http://www.ncbi.nlm.nih.gov/pubmed/23535729
[3]	http://www.ncbi.nlm.nih.gov/pubmed/23144319
[4]	http://www.ncbi.nlm.nih.gov/pubmed/23108145
[5]	http://www.ncbi.nlm.nih.gov/pubmed/22976474
[6]	http://www.ncbi.nlm.nih.gov/pubmed/20932654
[7]	http://www.ncbi.nlm.nih.gov/pubmed/20543847
[8]	http://www.ncbi.nlm.nih.gov/pubmed/19684604

[1]
 [2]
 [3]
 [4]
 [5] A meta-a
 [6] Genome-wide association study to identify single nucleotide polymorphisms (SNPs)
 [7]
 [8]

Disease.Trait
 <character>
 [1] Breast cancer
 [2] Breast cancer
 [3] Non-small cell lung cancer
 [4] Lung Cancer (DNA repair capacity)
 [5] Breast cancer
 [6] Erectile dysfunction and prostate cancer treatment
 [7] Testicular germ cell cancer
 [8] Acute lymphoblastic leukemia (childhood)

[1] 10,052 European and
 [2] 10,052 European and
 [3]
 [4] 914 European ancestry non-small cell lung
 [5] 3,666 European ancestry cases, 28,864 European ancestry controls, 1,004 African A
 [6] 27 Africa
 [7] 979 European an
 [8] 503 European ancestry pediatric cas

[1]
 [2]
 [3]
 [4]
 [5] 562 European ancestry cases, 6,410 European ancestry controls, 84 Japanese ancest
 [6]
 [7]
 [8]

Region Chr_id Chr_pos Reported.Gene.s.
 <character> <character> <numeric> <character>

[1]	8q24.21	8	129194641	MIR1208, MYC
[2]	11q13.1	11	65583066	DKFZp761E198, OVOL1, SNX32, CFL1, MUS81
[3]	2p23.3	2	26526419	GPR113
[4]	6q24.2	6	143943314	PHACTR2
[5]	20q11.22	20	32588095	RALY, EIF2S2, ASIP
[6]	Xp11.4	23	37854727	SYTL5
[7]	12p13.1	12	14653867	ATF7IP
[8]	10q21.2	10	63752159	ARID5B

	Mapped_gene	Upstream_gene_id	Downstream_gene_id	Snps_gene_ids
	<character>	<character>	<character>	<character>
[1]	MIR1208 - MIR3686	100302281	100500839	
[2]	OVOL1 - SNX32	5017	254122	
[3]	HADHB - GPR113	3032	165082	
[4]	PHACTR2	<NA>	<NA>	9749
[5]	RALY	<NA>	<NA>	22913
[6]	CXorf27 - MIR548AJ2	25763	100616252	
[7]	ATF7IP - PLBD1	55729	79887	
[8]	ARID5B	<NA>	<NA>	84159

	Upstream_gene_distance	Downstream_gene_distance	Strongest.SNP.Risk.Allele
	<character>	<character>	<character>
[1]	32.21	1301.66	rs11780156-T
[2]	18.38	18.34	rs3903072-G
[3]	13.09	4.62	rs6753473-G
[4]	<NA>	<NA>	rs9390123-A
[5]	<NA>	<NA>	rs2284378-T
[6]	4.16	28.42	rs872690-?
[7]	2.17	2.73	rs2900333-C
[8]	<NA>	<NA>	rs7089424-C

	SNPs	Merged	Snps_id_current	Context	Intergenic
	<character>	<character>	<character>	<character>	<character>
[1]	rs11780156	0	11780156	Intergenic	1
[2]	rs3903072	0	3903072	Intergenic	1
[3]	rs6753473	0	6753473	Intergenic	1
[4]	rs9390123	0	9390123	intron	0
[5]	rs2284378	0	2284378	intron	0
[6]	rs872690	0	872690	Intergenic	1
[7]	rs2900333	0	2900333	Intergenic	1
[8]	rs7089424	0	7089424	intron	0

	Risk.Allele.Frequency	p.Value	Pvalue_mlog	p.Value..text.	OR.or.beta
	<character>	<numeric>	<numeric>	<character>	<numeric>
[1]	0.16	3e-11	10.522879		1.07
[2]	0.53	9e-12	11.045757		1.05

[3]	0.052	4e-06	5.397940 (Additive model)	<NA>
[4]	0.3957	7e-06	5.154902	<NA>
[5]	0.31	1e-08	8.000000	1.16
[6]	0.03	9e-06	5.045757	11.78
[7]	0.62	6e-10	9.221849	1.27
[8]	0.34	7e-19	18.154902	1.65

X95..CI..text.	Platform..SNPs.passing.QC.	CNV
----------------	----------------------------	-----

	<character>		<character>	<character>
[1]	[1.04-1.10]	Illumina & Affymetrix [~2.6 million]	(Imputed)	N
[2]	[1.04-1.08]	Illumina & Affymetrix [~2.6 million]	(Imputed)	N
[3]	NR	Affymetrix [271,817]		N
[4]	NR	Illumina [303,669]		N
[5]	[1.10-1.22]	Illumina [2,608,509]	(imputed)	N
[6]	[NR]	Affymetrix [512,497]		N
[7]	[1.12-1.44]	Illumina [298,782]		N
[8]	[1.54-1.76]	Illumina [291,473]		N

	num.Risk.Allele.Frequency	dclass	score	tfstart	tfend
	<numeric>	<character>	<numeric>	<integer>	<integer>
[1]	0.1600	Breast	999.9851	129194621	129194650
[2]	0.5300	Breast	999.9517	65583065	65583094
[3]	0.0520	Lung	999.9875	26526415	26526444
[4]	0.3957	Lung	999.9387	143943292	143943321
[5]	0.3100	Breast	999.9284	32588075	32588104
[6]	0.0300	Prostate	999.9028	37854721	37854750
[7]	0.6200	Testicular	999.9895	14653848	14653877
[8]	0.3400	ALL (ped)	999.9621	63752142	63752171

	pvalue	qvalue
	<numeric>	<numeric>
[1]	1.49e-05	0.318
[2]	4.83e-05	0.373
[3]	1.25e-05	0.310
[4]	6.13e-05	0.383
[5]	7.16e-05	0.388
[6]	9.72e-05	0.403
[7]	1.05e-05	0.301
[8]	3.79e-05	0.361

seqlengths:

chr1	chr2	chr3	chr4	...	chr21	chr22	chrX
249250621	243199373	198022430	191154276	...	48129895	51304566	155270560

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(htag)
+   writeFasta(DNAStringSet(altg), file=faTargFun(htag))
+ }
```

7 Session information

```
> sessionInfo()
```

```
R version 3.1.1 (2014-07-10)
```

```
Platform: x86_64-unknown-linux-gnu (64-bit)
```

```
locale:
```

```
[1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8      LC_COLLATE=C
[5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8     LC_NAME=C
[9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
```

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

```
other attached packages:
```

```
[1] gwascat_1.8.0      vtpnet_0.4.1      GenomicRanges_1.16.3
[4] GenomeInfoDb_1.0.2 IRanges_1.22.9    BiocGenerics_0.10.0
[7] Rgraphviz_2.8.1    graph_1.42.0
```

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

```
[1] BBmisc_1.7          BSGenome_1.32.0    BatchJobs_1.3
[4] BiocParallel_0.6.1 Biostrings_2.32.1  DBI_0.2-7
[7] GenomicAlignments_1.0.3 Matrix_1.1-4       RCurl_1.95-4.1
[10] RSQLite_0.11.4      Rsamtools_1.16.1  XML_3.98-1.1
[13] XVector_0.4.0       bitops_1.0-6       brew_1.0-6
[16] checkmate_1.2       codetools_0.2-8    digest_0.6.4
[19] fail_1.2            foreach_1.4.2      iterators_1.0.7
[22] lattice_0.20-29     rtracklayer_1.24.2 sendmailR_1.1-2
[25] snpStats_1.14.0     splines_3.1.1      stats4_3.1.1
[28] stringr_0.6.2       survival_2.37-7    tools_3.1.1
[31] zlibbioc_1.10.0
```

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.