# *vtpnet*: variant-transcription factor-phenotype networks

VJ Carey

October 14, 2013

## 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\\\narthritis",
+       "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)))
```

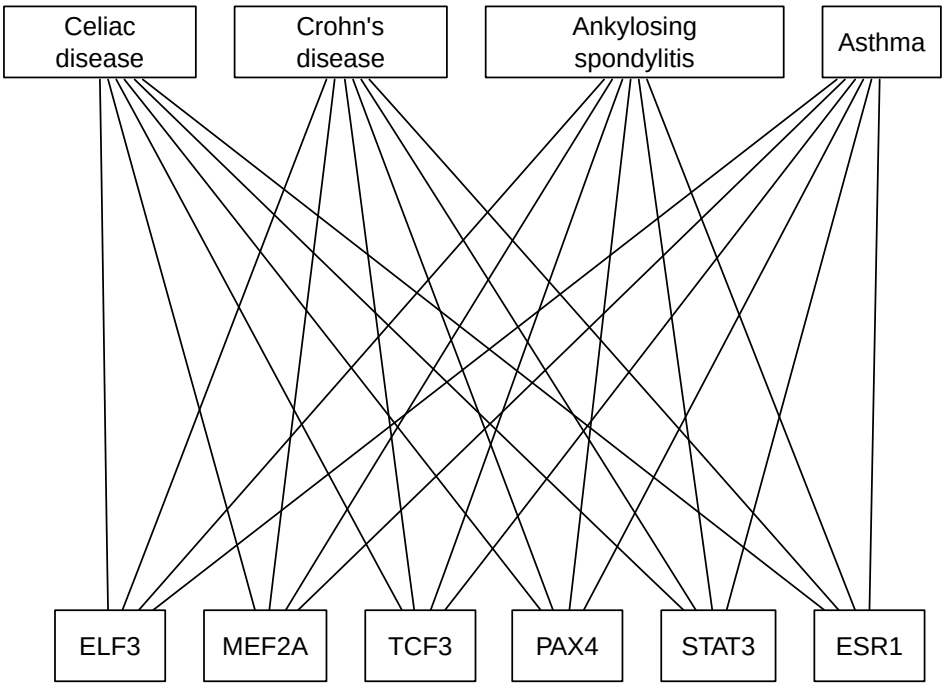


Figure 1: A complete bipartite graph for arbitrarily selected subsets of the autoimmune disorders and TFs found in Figure 4A of Maurano et al.

```
[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] TAACCCTAACCCTAACCCCAACCCCAACCC
  [2] TAACCCTAACCCCAACCCCAACCCCAACCC
  [3] AAAAAAATACACATGGCCAGGCCCCAGCCC
  [4] TAAAAAAATACACATGGCCAGGCCCCAGCC
  ---
  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 3 ranges and 41 metadata columns:
      seqnames               ranges strand | Date.Added.to.Catalog  PUBMEDID
         <Rle>            <IRanges>  <Rle> |              <factor> <integer>
  [1]     chrX [37854727, 37854727]      * |            11/15/2010  20932654
  [2]    chr12 [14653867, 14653867]      * |            07/12/2010  20543847
  [3]    chr10 [63752159, 63752159]      * |            09/04/2009  19684604
       First.Author       Date                       Journal
           <factor>   <factor>                      <factor>
  [1]       Kerns SL 10/05/2010 Int J Radiat Oncol Biol Phys
  [2]     Turnbull C 06/13/2010                    Nat Genet
  [3] Papaemmanuil E 08/16/2009                    Nat Genet
                                          Link
                                       <factor>
  [1] http://www.ncbi.nlm.nih.gov/pubmed/20932654
  [2] http://www.ncbi.nlm.nih.gov/pubmed/20543847
  [3] http://www.ncbi.nlm.nih.gov/pubmed/19684604


  [1] Genome-wide association study to identify single nucleotide polymorphisms (SNPs)
  [2]
  [3]
                                          Disease.Trait
                                               <factor>
  [1] Erectile dysfunction and prostate cancer treatment
  [2]                          Testicular germ cell cancer
  [3]           Acute lymphoblastic leukemia (childhood)
                                                                  Initial.Sample.Size
                                                                            <factor>
  [1]                            27 African American cases, 52 African American controls
  [2]                       979 European ancestry cases, 4,947 European ancestry controls
  [3] 503 European ancestry pediatric cases, 1,438 European ancestry pediatric controls
```

```
                                                                Replication.Sample.Size
                                                                            <character>
[1]                                                                                 NR
[2]                                  664 European ancestry cases, 3,456 European ancestry controls
[3] 404 European ancestry pediatric cases, 960 European ancestry pediatric controls
      Region      Chr_id   Chr_pos Reported.Gene.s.     Mapped_gene
    <factor> <character> <numeric>      <character>        <factor>
[1]   Xp11.4          23  37854727            SYTL5 CXorf27 - SYTL5
[2]  12p13.1          12  14653867           ATF7IP  ATF7IP - PLBD1
[3]  10q21.2          10  63752159           ARID5B          ARID5B
    Upstream_gene_id Downstream_gene_id Snp_gene_ids Upstream_gene_distance
         <character>        <character>     <factor>            <character>
[1]            25763              94122                                4.16
[2]            55729              79887                                2.17
[3]             <NA>               <NA>        84159                    <NA>
    Downstream_gene_distance Strongest.SNP.Risk.Allele      SNPs      Merged
                 <character>               <character>  <factor> <character>
[1]                    38.42                rs872690-?  rs872690           0
[2]                     2.73               rs2900333-C rs2900333           0
[3]                     <NA>               rs7089424-C rs7089424           0
    Snp_id_current    Context  Intergenic Risk.Allele.Frequency   p.Value
       <character>   <factor> <character>              <factor> <numeric>
[1]         872690 Intergenic           2                  0.03     9e-06
[2]        2900333 Intergenic           2                  0.62     6e-10
[3]        7089424     intron           1                  0.34     7e-19
    Pvalue_mlog p.Value..text. OR.or.beta X95..CI..text.
      <numeric>       <factor>  <numeric>    <character>
[1]    5.045757                    11.78           [NR]
[2]    9.221849                     1.27   [1.12-1.44]
[3]   18.154902                     1.65   [1.54-1.76]
    Platform..SNPs.passing.QC.      CNV num.Risk.Allele.Frequency      dclass
                      <factor> <factor>                 <numeric> <character>
[1]       Affymetrix [512,497]        N                      0.03    Prostate
[2]         Illumina [298,782]        N                      0.62  Testicular
[3]         Illumina [291,473]        N                      0.34   ALL (ped)
       score   tfstart     tfend    pvalue    qvalue
    <numeric> <integer> <integer> <numeric> <numeric>
[1]  999.9028  37854721  37854750  9.72e-05     0.403
[2]  999.9895  14653848  14653877  1.05e-05     0.301
[3]  999.9621  63752142  63752171  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(chtag = "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", chtag)
+     if (length(chk)>0) {
+       warning("clearing prefix ch or chr from chtag")
+       chtag = gsub("ch|chr", "", chtag)
+       }
+     snpgettag = paste0("ch", chtag)
+     ggettag = paste0("chr", chtag)
+     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 3.0.2 (2013-09-25)
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] splines   parallel  grid      stats     graphics  grDevices utils
 [8] datasets  methods   base

other attached packages:
 [1] TxDb.Hsapiens.UCSC.hg19.knownGene_2.10.1
 [2] GenomicFeatures_1.14.0
 [3] AnnotationDbi_1.24.0
 [4] Biobase_2.22.0
 [5] gwascat_1.6.0
 [6] snpStats_1.12.0
 [7] Matrix_1.0-14
 [8] lattice_0.20-24
 [9] survival_2.37-4
[10] vtpnet_0.2.0
[11] GenomicRanges_1.14.0
[12] XVector_0.2.0
[13] IRanges_1.20.0
[14] BiocGenerics_0.8.0
[15] Rgraphviz_2.6.0
[16] graph_1.40.0

loaded via a namespace (and not attached):
 [1] BSgenome_1.30.0    Biostrings_2.30.0  DBI_0.2-7          RCurl_1.95-4.1
 [5] RSQLite_0.11.4     Rsamtools_1.14.0   XML_3.98-1.1       biomaRt_2.18.0
```

```
 [9] bitops_1.0-6       rtracklayer_1.22.0 stats4_3.0.2       tools_3.0.2
[13] zlibbioc_1.8.0
```

# 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 Kutyavin, 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.