# A "big data" strategy relevant to transcriptional regulatory networks

VJ Carey, Ph.D.

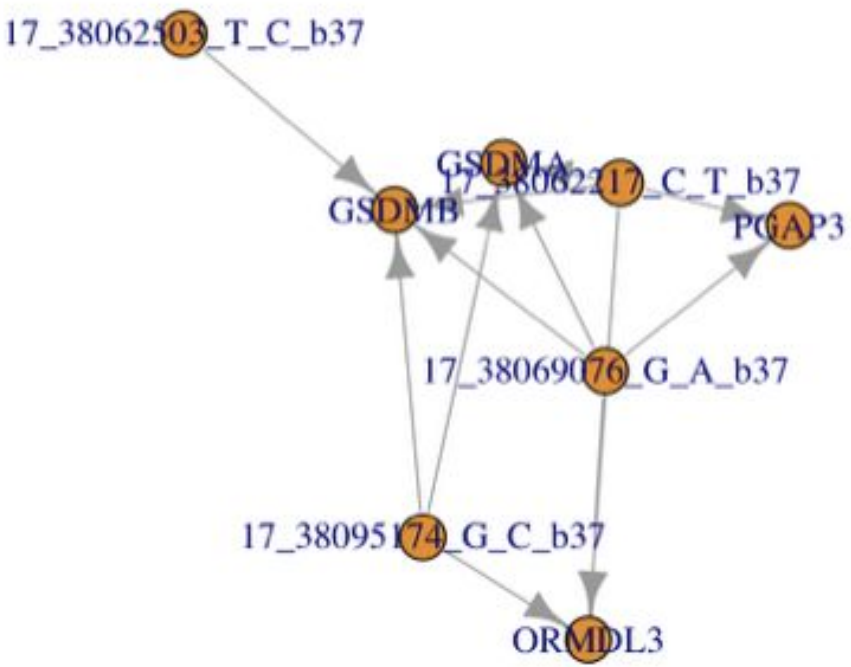
CSAMA 2019

# Road map

- A view of the task
- mongolite package to resolve queries
- TxRegInfra package to organize interfaces

# Organizing TFs, TFBS, tissue-specific eQTL results (SNP:gene associations declared in GTEx v6)

limited to
chr17 38e6-38.1e6




this is fairly primitive!
want to know more about edges, variant characteristics, etc.

how to scratch the surface even to this extent is not clear/shared
… or … neo4j?
graphQL?

# Some guiding principles

- Objective is to smooth the path from large genomic reference resources or experimental results to exploratory and confirmatory analysis activities -- leading to a new resource or interface
- Principle 1: record provenance and bind it to the resource
- Principle 2: minimize modification to original resource
- Principle 3: consider how the X[G, S] idiom can be supported
  - G refers to 'features', S refers to 'samples'
  - The underlying structure need not be 'rectangular'
- Principle 4: produce a useful 'colData'
- Principle 5: support subsetByOverlaps

# Principles 2+4 for eQTL+FP+GWAS

Show 10 entries

Search:

| | base | type | mid |
|---|---|---|---|
| Spleen_allpairs_v7_eQTL | Spleen | eQTL | allpairs_v7 |
| Stomach_allpairs_v7_eQTL | Stomach | eQTL | allpairs_v7 |
| Testis_allpairs_v7_eQTL | Testis | eQTL | allpairs_v7 |
| Thyroid_allpairs_v7_eQTL | Thyroid | eQTL | allpairs_v7 |
| Uterus_allpairs_v7_eQTL | Uterus | eQTL | allpairs_v7 |
| Vagina_allpairs_v7_eQTL | Vagina | eQTL | allpairs_v7 |
| Whole_Blood_allpairs_v7_eQTL | Whole | eQTL | Blood_allpairs_v7 |
| CD14_DS17215_hg19_FP | CD14 | FP | DS17215_hg19 |
| CD19_DS17186_hg19_FP | CD19 | FP | DS17186_hg19 |
| CD34_DS12274_hg19_FP | CD34 | FP | DS12274_hg19 |

Showing 41 to 50 of 2,640 entries

Previous 1 … 4 5 6 … 264 Next

# searchability is valuable

Show 10 entries Search: lung

| | base | type | mid |
|---|---|---|---|
| Lung_allpairs_v7_eQTL | Lung | eQTL | allpairs_v7 |
| fLung_DS14724_hg19_FP | fLung | FP | DS14724_hg19 |

Showing 1 to 2 of 2 entries (filtered from 2,640 total entries) Previous 1 Next

Classes are useful -- this is not quite mature, 'RaggedExperiment' extension … accommodate irregular 'assays'

```
> rme
class: RaggedMongoExpt
dim: 2640 2640
assays(0):
rownames: NULL
colnames(2640): Adipose_Subcutaneous_allpairs_v7_eQTL
  Adipose_Visceral_Omentum_allpairs_v7_eQTL ...
  iPS_19_11_DS15153_hg19_FP vHMEC_DS18406_hg19_FP
colData names(6): base type ... type mid
```

# MongoDB+mongolite for each 'file' as 'document' rme = f(URL, db, colData)... URL here is to AWS

```
> rme@con
<Mongo collection> 'test'
 $aggregate(pipeline = "{}", options = "{\"allowDiskUse\":true}", handler = NULL, pagesize = 1000, it
erate = FALSE)
 $count(query = "{}")
 $disconnect(gc = TRUE)
 $distinct(key, query = "{}")
 $drop()
 $export(con = stdout(), bson = FALSE, query = "{}", fields = "{}", sort = "{\"_id\":1}")
 $find(query = "{}", fields = "{\"_id\":0}", sort = "{}", skip = 0, limit = 0, handler = NULL, pagesi
ze = 1000)
 $import(con, bson = FALSE)
 $index(add = NULL, remove = NULL)
 $info()
 $insert(data, pagesize = 1000, stop_on_error = TRUE, ...)
 $iterate(query = "{}", fields = "{\"_id\":0}", sort = "{}", skip = 0, limit = 0)
 $mapreduce(map, reduce, query = "{}", sort = "{}", limit = 0, out = NULL, scope = NULL)
 $remove(query, just_one = FALSE)
 $rename(name, db = NULL)
 $replace(query, update = "{}", upsert = FALSE)
 $run(command = "{\"ping\": 1}", simplify = TRUE)
 $update(query, update = "{\"$set\":{}}", filters = NULL, upsert = FALSE, multiple = FALSE)
```

# Task: use the resource to learn about eQTLs in two tissues

- Interested in subcutaneous adipose and lung
- We'll look in a region on chr17

```
> tiss = c("Adipose_Subcutaneous_allpairs_v7_eQTL",
+   "Lung_allpairs_v7_eQTL")
> qrange = GRanges("17", IRanges(38e6,width=1e6))
> ans = lapply(tiss, function(x) sbov(rme[,x], qrange))
..Warning messages:
1: In sbov(rme[, x], qrange) : genome is not set for for query GRanges
2: In sbov(rme[, x], qrange) : genome is not set for for query GRanges
> sapply(ans,length)
[1] 2644 2049
> names(ans) = tiss
> sapply(ans,length)
Adipose_Subcutaneous_allpairs_v7_eQTL                Lung_allpairs_v7_eQTL
                                 2644                                 2049
```

annotation from GTEx

```
> head(ans[[1]])
GRanges object with 6 ranges and 16 metadata columns:
      seqnames    ranges strand |            gene_id              variant_id
         <Rle> <IRanges>  <Rle> |           <factor>                <factor>
  [1]       17  38001559      * | ENSG00000073605.14     17_38001559_A_G_b37
  [2]       17  38001571      * |  ENSG00000141744.3     17_38001571_G_A_b37
  [3]       17  38001659      * | ENSG00000073605.14    17_38001659_T_TC_b37
  [4]       17  38004929      * | ENSG00000073605.14 17_38004929_GATTT_G_b37
  [5]       17  38004929      * |  ENSG00000172057.5 17_38004929_GATTT_G_b37
  [6]       17  38004929      * |  ENSG00000264968.1 17_38004929_GATTT_G_b37
      tss_distance ma_samples  ma_count       maf pval_nominal     slope
         <integer>  <integer> <integer> <numeric>    <numeric> <numeric>
  [1]       -73344         50        51 0.0662338   0.00147706 -0.271482
  [2]       177337         16        17 0.0221354  0.000270805  0.359381
  [3]       -73244         50        51 0.0662338   0.00147706 -0.271482
  [4]       -69974        260       330  0.428571   0.00111841  0.122441
  [5]       -78925        260       330  0.428571  1.34357e-05  0.106083
  [6]       -79066        260       330  0.428571  1.45895e-06  0.248438
       slope_se              qvalue       chr   snp_pos       A1       A2
      <numeric>           <numeric> <integer> <integer> <factor> <factor>
  [1] 0.0846494   0.0831222693011164        17  38001559        A        G
  [2] 0.0975787   0.0221545451076158        17  38001571        G        A
  [3] 0.0846494   0.0831222693011164        17  38001659        T       TC
  [4] 0.0372283   0.0677038185910862        17  38004929    GATTT        G
  [5] 0.0239893  0.00170738349980929        17  38004929    GATTT        G
  [6] 0.0505987 0.000236491245248636        17  38004929    GATTT        G
         build                                origin
      <factor>                             <character>
  [1]     b37 Adipose_Subcutaneous_allpairs_v7_eQTL
  [2]     b37 Adipose_Subcutaneous_allpairs_v7_eQTL
  [3]     b37 Adipose_Subcutaneous_allpairs_v7_eQTL
  [4]     b37 Adipose_Subcutaneous_allpairs_v7_eQTL
  [5]     b37 Adipose_Subcutaneous_allpairs_v7_eQTL
  [6]     b37 Adipose_Subcutaneous_allpairs_v7_eQTL
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

# Are there tissue-specific SNP-associated genes?

```
> names(ans) = c("subcut", "lung")
> setdiff(ans$subcut$gene_id, ans$lung$gene_id)
 [1] "ENSG00000126368.5"  "ENSG00000108306.7"  "ENSG00000214546.3"
 [4] "ENSG00000173991.5"  "ENSG00000229028.2"  "ENSG00000131746.8"
 [7] "ENSG00000221852.4"  "ENSG00000126337.9"  "ENSG00000131759.13"
[10] "ENSG00000186832.4"  "ENSG00000128422.11" "ENSG00000173812.6"
> setdiff(ans$lung$gene_id, ans$subcut$gene_id)
[1] "ENSG00000108298.5" "ENSG00000266753.2" "ENSG00000265799.1"
[4] "ENSG00000204889.6" "ENSG00000270145.1" "ENSG00000196859.3"
[7] "ENSG00000265666.1" "ENSG00000186847.5" "ENSG00000171346.9"
```

primitive approach to acquiring GO mappings for eGenes unique to lung in this region

```
                                                                RNA binding
> ulung = setdiff(ans$lung$gene_id, ans$subcut$gene_id)
> ulung
[1] "ENSG00000108298.5" "ENSG00000266753.2" "ENSG00000265799.1"
[4] "ENSG00000204889.6" "ENSG00000270145.1" "ENSG00000196859.3"
[7] "ENSG00000265666.1" "ENSG00000186847.5" "ENSG00000171346.9"
> ul = gsub("\\..*", "", ulung)
> ul
[1] "ENSG00000108298" "ENSG00000266753" "ENSG00000265799" "ENSG0000020
[5] "ENSG00000270145" "ENSG00000196859" "ENSG00000265666" "ENSG0000018
[9] "ENSG00000171346"
> na.omit(mapIds(org.Hs.eg.db, keys=ul, keytype="ENSEMBL", column="GO"
'select()' returned 1:many mapping between keys and columns
ENSG00000108298 ENSG00000204889 ENSG00000196859 ENSG00000186847 ENSG00
   "GO:0000184"    "GO:0005198"    "GO:0005198"    "GO:0005200"    "GO
attr(,"na.action")
ENSG00000266753 ENSG00000265799 ENSG00000270145 ENSG00000265666
              2               3               5               7
attr(,"class")
[1] "omit"
> mapIds(GO.db, keys=.Last.value, keytype="GOID", column="TERM")
'select()' returned 1:1 mapping between keys and columns
                                                              GO:0000184
"nuclear-transcribed mRNA catabolic process, nonsense-mediated decay"
                                                              GO:0005198
                                          "structural molecule activity"
                                                              GO:0005198
                                          "structural molecule activity"
                                                              GO:0005200
                                 "structural constituent of cytoskeleton"
                                                              GO:0005200
                                 "structural constituent of cytoskeleton"
```

# For TFBS in this interval

See the TFutils package  fimo_granges()

# If time permits

indexedGF for GWAS summaries

vjcitn.shinyapps.io/ca43k