Benchmark Dataset Construction | GSA Tools Comparison

Shaurya Jauhari, PhD

Mora Lab.

shauryajauhari@gzhmu.edu.cn

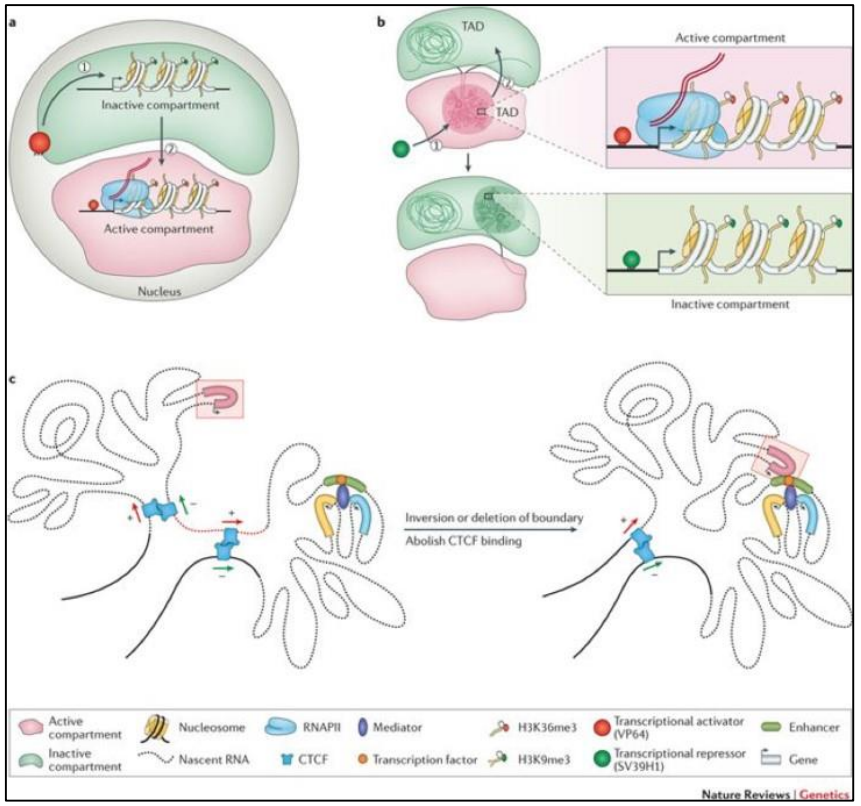
2019/01/15

Outline

- Core Idea
- Benchmarking Data | “Gold Standard” dataset
- Testing
- Caveat(s) hitherto
- Files/ Snapshots

Core Idea

- GSA Pipeline: Sequences/ Peaks/ Genomic Regions -> Genes (Gene List) -> Pathways (Gene Sets)
- BED files are the major input form that manifest genomic regions for variegated interactions.
- Genome organization is a complex structuring in 3D space; loops, domains, sub-compartments aren't just plainly to accommodate ~ 2 meter long DNA into the nucleus, but also have biological meaning, i.e. to enact the genomic code in an organism.



Organization and function of the 3D genome, Boyan Bonev and Giacomo Cavalli, *Nature Reviews Genetics* **volume 17**, pages 661–678 (2016)

Core Idea … Contd.

- Current suite of GSA tools typically overlook this standard of 3D genome organization. They assume a linear window around the gene(TSS) to be the ideal regulatory region.
- GREAT, CompGO, Seq2pathway, Enrichr, Chipenrich, Broadenrich, *Polyenrich* are the contenders.

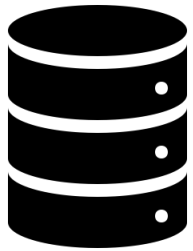
```
$ head ../data/Input_tags.bed
chr1 233604    233639    0    2    -
chr1 559767    559802    0    3    +
chr1 742600    742635    0    2    +
chr1 742600    742635    0    0    +
chr1 744231    744266    0    0    +
chr1 744307    744342    0    2    -
chr1 746885    746920    0    2    +
chr1 746958    746993    0    1    +
chr1 748226    748261    0    2    +
chr1 748357    748392    0    0    -
```

Coding Regions

- Straightforward mapping to the known "working" definitions of genes.

Non-Coding Regions

- "Speculation" on gene association.



Ontology Databases

Benchmarking Data | "Gold Standard" dataset

- Tabulation of Genomic Regions (DNA) – Associated TF/ Histone Mark (Protein) interactions, pertaining to a physiological state.
- Data chosen on the grounds of:
  - Study/ Publication impact (Journal, Year, Citations)
  - Coverage (KEGG, GO, Cistrome)
- Omitting of Input/Normal assays.
- Experiment (GSE) – Samples (GSM) form the key denominations of the benchmark dataset.
- In compliance to Bioconductor standards, a major criterion is to make use of existing packages to scale reproducibility. **GRanges** has been used as a proforma to store BED files for the R data objects.
- To apply to the tools, the samples (data objects) are again converted to BED files via **export.bed()**.

Granges object

```
## GRanges object with 3 ranges and 0 metadata columns:
##       seqnames    ranges strand
##          <Rle> <IRanges>  <Rle>
##   [1]     chr1    [1, 3]      +
##   [2]     chr2    [3, 5]      -
##   [3]     chr1    [5, 7]      +
##   -------
##   seqinfo: 2 sequences from hg19 genome
```

Testing

- The tracking of the comparative analysis is being maintained as a R markdown document (code and output).
- Currently underway with GREAT (Web interface) and Seq2pathway (R package)
  - Still some samples remaining
- GO terms for Biological Processes , and Disease Ontology terms have been considered from GREAT.
- The standard window size has been derived from Seq2pathway's default value, i.e. **150 kBP**. This shall be maintained as a basis of comparison, for now.

Caveat(s) Hitherto

- **MEMORY INSUFFICIENCY**
  - Samples over 100 MB disk file size fail to run
- The MAC system has **32 GB RAM**.
- The **BioInfoServer** (IP Address: 10.168.122.47) has internet connectivity issues that impede package installation and system upgradation.
- The **NAS** in the lab cannot be a substitute for extended system storage; it's an auxiliary storage device.

R Markdown

```
## Loading Benchmark dataset
```

```{r sourcing the compiled gold standard benchmark dataset }

install.packages("~/Desktop/Labwork/GSAChIPSeqComparitiveAnalysis/GSAChIPSeqGold_0.1.0.tar.gz", repos = NULL, type =
"source")

```

```
* installing *source* package ‘GSAChIPSeqGold’ ...
** data
*** moving datasets to lazyload DB
** help
No man pages found in package  ‘GSAChIPSeqGold’
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (GSAChIPSeqGold)
```

```{r Iterative variable naming from file names}

for (i in 1:length(grl)){
  export.bed(grl[i],format="bed",con=paste(eval(parse(text='names(grl[i])')),"data.bed",sep=""))
  paste(eval(parse(text='names(grl[i])')),"data",sep="") =
read_bed(paste(eval(parse(text='names(grl[i])')),"data.bed",sep=""))
 paste(eval(parse(text='names(grl[i])')),"seq2pathway",sep="_")=runseq2pathway(paste(eval(parse(text='names(grl[i])')
),"data",sep=""))
}

```

```
## Testing individual data in the benchmark dataset with
## each GSA tool package
## Seq2Pathway
```

```{r}
library(GSAChIPSeqGold)
library(seq2pathway)
library(rtracklayer)
library(chipenrich)

export.bed(GSM1847178,format="bed",con="./regen/GSM1847178.bed")
GSM1847178data <- read_bed("./regen/GSM1847178.bed")
GSM1847178_seq2pathway=runseq2pathway(GSM1847178data,genome = "hg19")

export.bed(GSM2058015,format="bed",con="./regen/GSM2058015.bed")
GSM2058015data <- read_bed("./regen/GSM2058015.bed")
GSM2058015_seq2pathway=runseq2pathway(GSM2058015data, genome = "hg19")
```

Loading Benchmark dataset R Markdown

GREAT

Species Assembly

- (•) Human: GRCh37 (UCSC hg19, Feb/2009)
- ( ) Mouse: NCBI build 37 (UCSC mm9, Jul/2007)
- ( ) Mouse: NCBI build 38 (UCSC mm10, Dec/2011)
- ( ) Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010) Zebrafish CNE set

Can I use a different species or assembly?

Test regions

- (•) BED file: Browse... GSM2058022.bed
- ( ) BED data:

What should my test regions file contain?
How can I create a test set from a UCSC Genome Browser annotation track?

Background regions

- (•) Whole genome
- ( ) BED file: Browse... No file selected.
- ( ) BED data:

When should I use a background set?
What should my background regions file contain?

Association rule settings

Hide settings »

Associating genomic regions with genes

GREAT calculates statistics by associating genomic regions with nearby genes and applying the gene annotations to the regions. Association is a two step process. First, every gene is assigned a regulatory domain. Then, each genomic region is associated with all genes whose regulatory domain it overlaps.

(•) Basal plus extension

Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 150 kb

Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

( ) Two nearest genes

within 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the nearest gene's TSS but no more than the maximum extension in one direction.

( ) Single nearest gene

within 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no more than the maximum extension in one direction.

Gene Transcription Start Site (TSS)

[x] Include curated regulatory domains *What are curated regulatory domains?*

Submit Reset

GREAT . . .Contd.

GO Biological Process (20+ terms) Global controls

Table controls: ✓ Export
Shown ontology data as HTML
Shown ontology data as .tsv
All ontology data as .tsv

top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
organophosphate biosynthetic process	1	**0.0000**	**0.0000**	2.0243	3,691	4.17%	189	**8.1400e-6**	1.0866	405	428	2.58%
actin cytoskeleton organization	1	**0.0000**	**0.0000**	2.2466	3,916	4.42%	199	**1.5490e-5**	1.0941	323	339	2.06%
actin filament-based process	1	**0.0000**	**0.0000**	2.2123	4,259	4.81%	308	**1.1940e-3**	1.0732	357	382	2.27%
phospholipid metabolic process	216	**2.5418e-298**	**1.2285e-296**	2.2179	2,701	3.05%	254	**1.2768e-4**	1.0946	265	278	1.69%
apoptotic signaling pathway	230	**3.0831e-280**	**1.3995e-278**	2.1517	2,715	3.07%	126	**3.4024e-8**	1.1199	276	283	1.76%
glycerolipid metabolic process	258	**3.6824e-250**	**1.4901e-248**	2.0707	2,653	3.00%	217	**4.3636e-5**	1.0970	278	291	1.77%
phospholipid biosynthetic process	261	**1.1165e-247**	**4.4661e-246**	2.3070	2,066	2.33%	237	**7.8934e-5**	1.1097	201	208	1.28%
glycerophospholipid metabolic process	287	**1.1144e-226**	**4.0539e-225**	2.1773	2,142	2.42%	262	**1.8168e-4**	1.1024	216	225	1.37%
glycerolipid biosynthetic process	320	**6.0038e-206**	**1.9587e-204**	2.1421	2,020	2.28%	266	**2.2515e-4**	1.1046	202	210	1.29%
glycerophospholipid biosynthetic process	335	**2.1986e-198**	**6.8516e-197**	2.2164	1,803	2.04%	293	**7.1106e-4**	1.1049	178	185	1.13%
regulation of protein complex assembly	373	**3.4870e-177**	**9.7599e-176**	2.0056	2,037	2.30%	461	**1.7616e-2**	1.0808	192	204	1.22%
cellular response to external stimulus	421	**1.2189e-149**	**3.0226e-148**	2.0236	1,682	1.90%	477	**2.0245e-2**	1.0845	170	180	1.08%
intrinsic apoptotic signaling pathway	422	**1.3132e-149**	**3.2487e-148**	2.2391	1,329	1.50%	190	**8.4336e-6**	1.1398	134	135	0.85%
viral life cycle	461	**2.2567e-134**	**5.1105e-133**	2.3164	1,110	1.25%	373	**5.4159e-3**	1.1024	144	150	0.92%
establishment of protein localization to membrane	488	**1.5821e-125**	**3.3847e-124**	2.0476	1,369	1.55%	226	**5.4793e-5**	1.1171	179	184	1.14%
Ras protein signal transduction	500	**2.6677e-120**	**5.5702e-119**	2.0035	1,385	1.56%	511	**3.2678e-2**	1.0909	133	140	0.85%
cellular response to extracellular stimulus	512	**1.7237e-113**	**3.5148e-112**	2.0701	1,204	1.36%	545	**4.2173e-2**	1.0932	119	125	0.76%
transforming growth factor beta receptor signaling pathway	524	**5.1026e-111**	**1.0166e-109**	2.0086	1,269	1.43%	445	**1.4538e-2**	1.1027	121	126	0.77%
nuclear-transcribed mRNA catabolic process	529	**1.8898e-110**	**3.7295e-109**	2.0689	1,173	1.32%	185	**5.7836e-6**	1.1285	171	174	1.09%
positive regulation of I-kappaB kinase/NF-kappaB cascade	542	**6.9372e-107**	**1.3363e-105**	2.0153	1,211	1.37%	290	**6.2156e-4**	1.1168	142	146	0.90%

The test set of 88,556 genomic regions picked 15,711 (87%) of all 18,041 genes.
GO Biological Process has 10,440 terms covering 15,441 (86%) of all 18,041 genes, and 950,065 term - gene associations.
10,440 ontology terms (100%) were tested using an annotation count range of [1, Inf].

GREAT . . .Contd.

Disease Ontology (4 terms) Global controls

Table controls: Export Shown top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
bile duct cancer	167	**1.2170e-90**	**1.6288e-89**	2.0149	1,023	1.47%	72	**2.8482e-2**	1.0924	128	133	0.81%
bile duct carcinoma	173	**1.8155e-88**	**2.3455e-87**	2.0032	1,013	1.45%	74	**3.0373e-2**	1.0921	127	132	0.80%
stomach carcinoma	207	**1.1912e-71**	**1.2861e-70**	2.0668	756	1.08%	65	**2.3513e-2**	1.1107	91	93	0.57%
gastric adenocarcinoma	211	**3.3796e-68**	**3.5798e-67**	2.0480	735	1.05%	75	**3.1259e-2**	1.1096	87	89	0.55%

The test set of 69,699 genomic regions picked 15,894 (88%) of all 18,041 genes.
Disease Ontology has 2,235 terms covering 7,886 (44%) of all 18,041 genes, and 232,324 term - gene associations.
2,235 ontology terms (100%) were tested using an annotation count range of [1, Inf].

Seq2pathway

4. more columns: Other custom defined information (optional)

`search_radius` A non-negative integer, with which the input genomic regions can be assigned not only to the matched or nearest gene, but also with all genes within a search radius for some genomic region type. This parameter works only when the parameter "SNP" is FALSE. Default is 150000.

`promoter_radius` A non-negative integer. Default is 200. Promoters are here defined as upstream regions of the transcription start sites (TSS). User can assign the promoter radius, a suggested value is between 200 to 2000.

`promoter_radius2` A non-negative integer. Default is 100. Promoters are here defined as downstream regions after the transcription start sites (TSS).

`genome` A character specifies the genome type. Currently, choice of "hg38", "hg19", "mm10", and "mm9" is supported.

`adjacent` A Boolean. Default is FALSE to search all genes within the search_radius. Using "TRUE" to find the adjacent genes only and ignore the parameters "SNP" and "search_radius".

`SNP` A Boolean specifies the input object type. FALSE by default to keep on searching for intron and neighboring genes. Otherwise, runseq2gene stops searching when the input genomic region is residing on exon of a coding gene.

`PromoterStop` A Boolean, "FALSE" by default to keep on searching neighboring genes using the parameter "search_radius". Otherwise, runseq2gene stops searching neighboring genes. This parameter has function only if an input genomic region maps to promoter of coding gene(s).

`NearestTwoDirection` A boolean, "TRUE" by default to output the closest left and closest right coding genes with directions. Otherwise, output only the nearest coding gene regardless of direction.

`UTR3` A boolean, "FALSE" by defalt to calculate the distance from genes' 5UTR. Otherwsie, calculate the distance from genes' 3UTR.

`DataBase` A character string assigns an R GSA.genesets object to define gene-set. User can call GSA.read.gmt to load customized gene-sets with a .gmt format. If not specified, a character "GOterm" by default, three categories of GO-defined gene sets (BP,MF,CC) will be used. Alternatively, user can specify a category by the choice of "BP","MF","CC".

`FAIMETest` A boolean values. By default is FALSE. When true, executes function of gene2pathway test using the FAIME method, which only functions when the fifth column of input file exsists and is a vector of scores or values.

`FisherTest` A Boolean value. By default is TRUE to excute the function of the Fisher's exact test. Otherwise, only excutes the function of gene2pathway test.

`collapsemethod` A character for determining which method to use when call the function collapseRows in package WGCNA. The function "collapsemethod" uses this paramter to call the collapseRows() function in package "WGCNA".

`alpha` A positive integer, 5 by default. This is a FAIME-specific parameter. A higher value puts more weights on the most highly-expressed ranks than the lower expressed ranks.

`logCheck` A Boolean value. By default is FALSE. When true, the function takes the log-transformed values of gene if the maximum value of sample profile is larger than 20.

Seq2pathway

- <variable_name>$seq2gene_result.FET$seq2gene_CodingGeneOnlyResult
- <variable_name>$seq2gene_result.FET$seq2gene_FullResult

- <variable_name>$gene2pathway_result.FET$GO.MF
- <variable_name>$gene2pathway_result.FET$GO.BP
- <variable_name>$gene2pathway_result.FET$GO.CC

- In R, the results can be directed to a dataframe format to extract useful heads, eg. Gene Names, Ensemble IDs, GO terms, etc.

谢谢