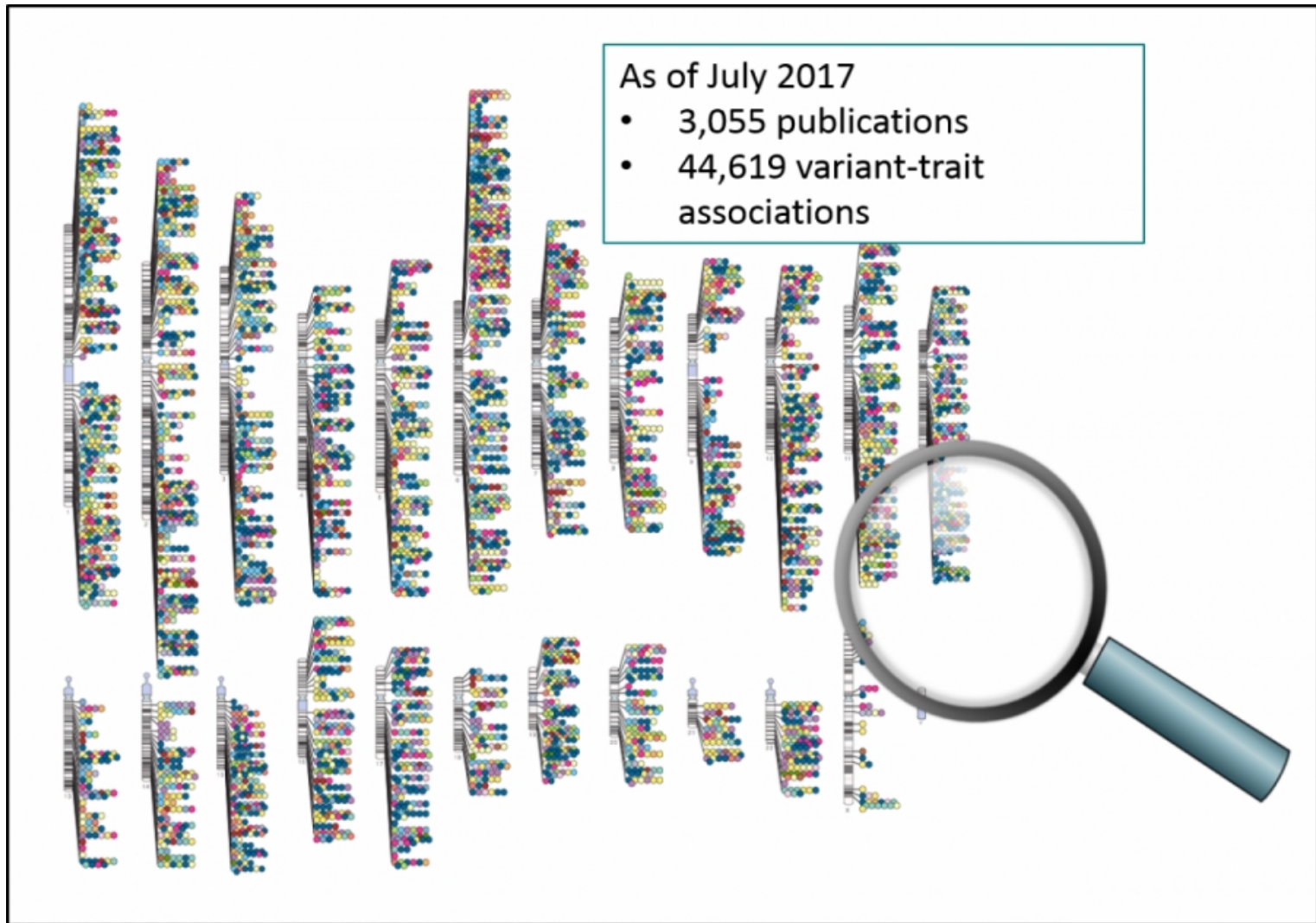


Gene-Gene /SNP-SNP Interaction: BIOFILTER

GBIO0002

Archana Bhardwaj
University of Liege



The combinatorial problem of jointly analyzing the millions of genetic variations accessible by high-throughput genotyping technologies is a difficult challenge.



NIH Public Access

Author Manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2010 April 26.

Published in final edited form as:

Pac Symp Biocomput. 2009 ; : 368–379.

Biofilter: A Knowledge-Integration System for the Multi-Locus Analysis of Genome-Wide Association Studies*

William S. Bush, Scott M. Dudek, and Marylyn D. Ritchie

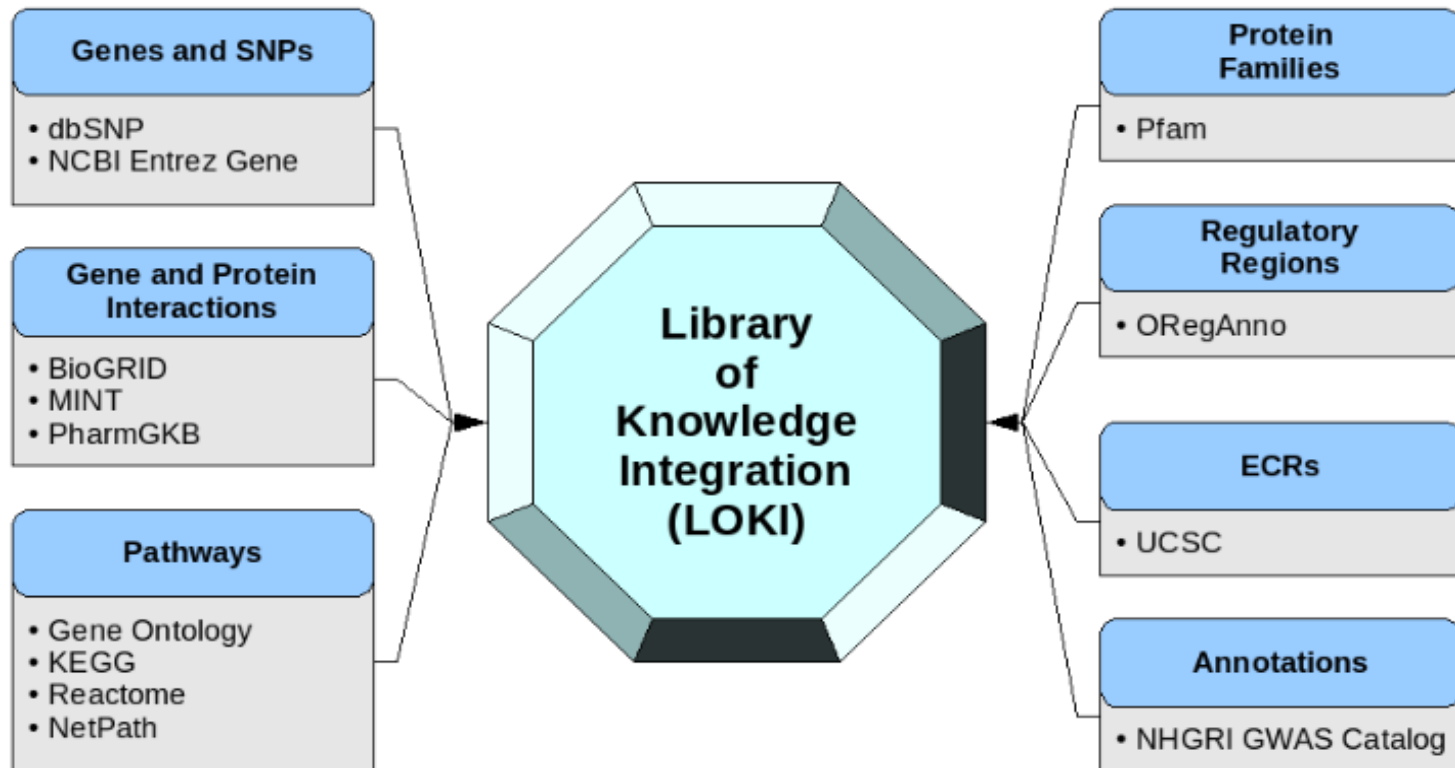
Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA

Abstract

Genome-wide association studies provide an unprecedented opportunity to identify combinations of genetic variants that contribute to disease susceptibility. The combinatorial problem of jointly analyzing the millions of genetic variations accessible by high-throughput genotyping technologies is a difficult challenge. One approach to reducing the search space of this variable selection problem is to assess specific combinations of genetic variations based on prior statistical and biological knowledge. In this work, we provide a systematic approach to integrate multiple public databases of gene groupings and sets of disease-related genes to produce multi-SNP models that have an established biological foundation. This approach yields a collection of models which can be tested statistically in genome-wide data, along with an ordinal quantity describing the number of data sources that support any given model. Using this knowledge-driven approach reduces the computational and statistical burden of large-scale interaction analysis while simultaneously providing a biological foundation for the relevance of any significant statistical result that is found.

□ Biofilter uses publicly available databases to establish relationships between gene-products

LOKI: Library of Knowledge Integration



LOKI DB : dbSNP

The screenshot shows the NCBI dbSNP website. At the top left is the NCBI logo. The main title is "dbSNP Short Genetic Variations" with a 3D protein structure image to the right. Below the title is a navigation bar with tabs for dbVar, ClinVar, GaP, PubMed, Nucleotide, and Protein. A search bar is present with the text "Search small variations in dbSNP or large structural variations in dbVar". Below the search bar is a dropdown menu set to "dbSNP" and a "Go" button. On the left side, there is a sidebar with a "Have a question about dbSNP? Try searching the SNP FAQ Archive!" section and a "Go" button. Below that is a "GENERAL" section with links for RSS Feed, Contact Us, Organism Data, dbSNP Homepage, NCBI Variation Resources, Announcements, dbSNP Summary, and FTP Download. Further down are sections for "SNP SUBMISSION", "DOCUMENTATION", "SEARCH", and "RELATED SITES". The main content area features a yellow "ANNOUNCEMENT" banner stating that dbSNP and dbVar no longer accept submissions for non-human organism data. Below this is a "Search by IDs on All Assemblies" section with a note that rs# and ss# must be prefixed with "rs" or "ss", respectively. It includes a search form with an "ID:" input field, a "Reference cluster ID(rs#)" dropdown, and "Search" and "Reset" buttons. The "Submission Information" section lists links for "By Submitter", "New Submitted Batches", "Method", "Population", and "Publication". The "Batch" section lists "Enter List" with sub-links for "NCBI Assay ID(ss)" and "Reference SNP IDs".

NCBI

dbSNP

Short Genetic Variations

dbVar ClinVar GaP PubMed Nucleotide Protein

Search small variations in dbSNP or large structural variations in dbVar

Search Entrez dbSNP for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

ANNOUNCEMENT

dbSNP and dbVar no longer accept submissions for non-human organism data. Please read more [here](#).

GENERAL

RSS Feed

Contact Us

Organism Data

dbSNP Homepage

NCBI Variation Resources

Announcements

dbSNP Summary

FTP Download

SNP SUBMISSION

DOCUMENTATION

SEARCH

RELATED SITES

Search by IDs on All Assemblies

Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

ID: Reference cluster ID(rs#)

Search Reset

Submission Information

- [By Submitter](#)
- [New Submitted Batches](#)
- [Method](#)
- [Population](#)
- [Publication](#)

Batch

- Enter List
 - [NCBI Assay ID\(ss\)](#)
 - [Reference SNP IDs](#)

LOKI DB : KEGG database



KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

Menu PATHWAY BRITE MODULE KO GENES LIGAND NETWORK DISEASE DRUG DBGET

Select prefix

map

Organism

Enter keywords

hsa

Go

Help

[[New pathway maps](#) | [Update history](#)]

Pathway Maps

KEGG PATHWAY is a collection of manually drawn [pathway maps](#) representing our knowledge on the molecular interaction, reaction and relation networks for:

1. Metabolism

[Global/overview](#) [Carbohydrate](#) [Energy](#) [Lipid](#) [Nucleotide](#) [Amino acid](#) [Other amino](#) [Glycan](#)
[Cofactor/vitamin](#) [Terpenoid/PK](#) [Other secondary metabolite](#) [Xenobiotics](#) [Chemical structure](#)

2. Genetic Information Processing

3. Environmental Information Processing

4. Cellular Processes

5. Organismal Systems

6. Human Diseases

7. Drug Development

KEGG PATHWAY is a reference database for [Pathway Mapping](#).

Pathway Identifiers

Each pathway map is identified by the combination of 2-4 letter prefix code and 5 digit number (see [KEGG Identifier](#)). The prefix has the following meaning:

| | |
|-------|---|
| map | manually drawn reference pathway |
| ko | reference pathway highlighting KOs |
| ec | reference metabolic pathway highlighting EC numbers |
| rn | reference metabolic pathway highlighting reactions |
| <org> | organism-specific pathway generated by converting KOs to gene identifiers |

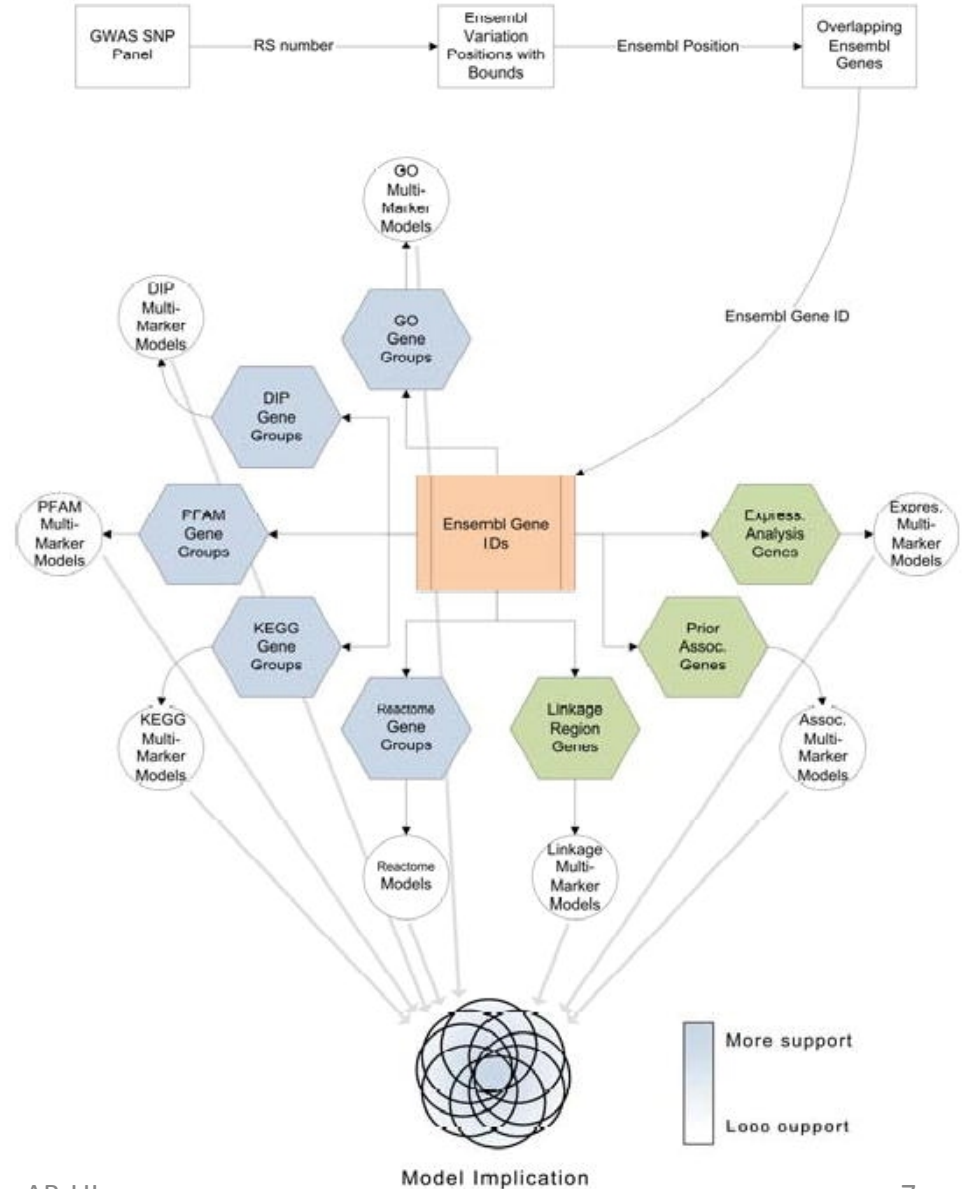
and the numbers starting with the following:

Biofilter : Overview

□ GWAS platform SNPs are mapped to Ensembl gene Ids.

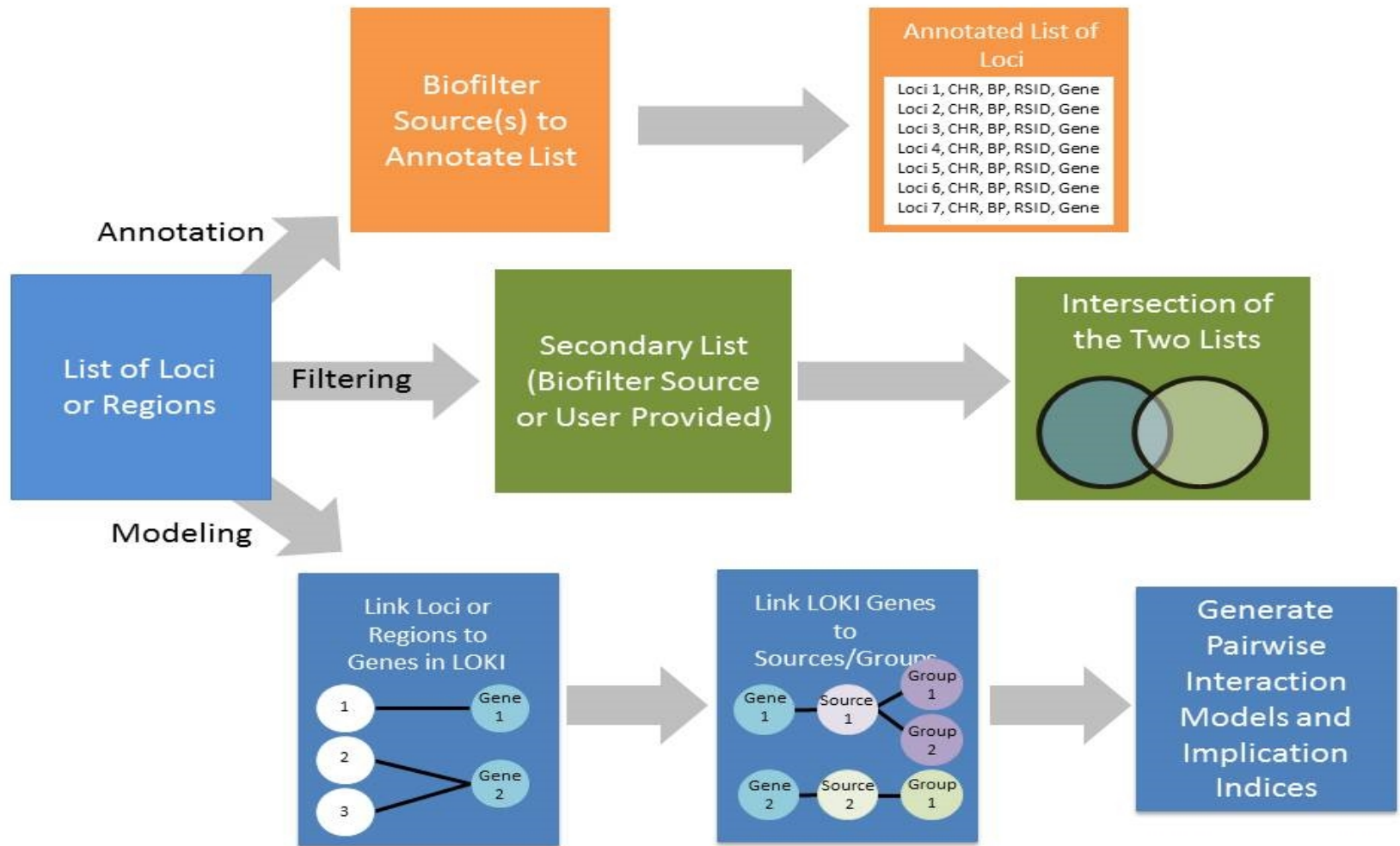
□ Multi-marker models are generated from SNPs within knowledge-related genes.

□ Derived models are overlaid to assess overall model implication.



Biofilter : Three Analysis mode







Biofilter has three primary analysis modes and uses the available biological knowledge in slightly different ways.



Biofilter Data types

Data Types

Biofilter can work with and understand the relationships between six basic types of data:

- | | | |
|-----------------|---|---|
| SNP |  | Specified by an RS number, i.e. "rs1234". Used to refer to a known and documented SNP whose position can be retrieved from the knowledge database. |
| Position |  | Specified by a chromosome and basepair location, i.e. "chr1:234". Used to refer to any single genomic location, such as a single nucleotide polymorphism (SNP), single nucleotide variation (SNV), rare variant, or any other position of interest. |
| Region |  | Specified by a chromosome and basepair range, i.e. "chr1:234-567". Used to refer to any genomic region, such as a copy number variation (CNV), insertion/deletion (indel), gene coding region, evolutionarily conserved region (ECR), functional region, regulatory region, or any other region of interest. |
| Gene |  | Specified by a name or other identifier, i.e. "A1BG" or "ENSG00000121410". Used to refer to a known and documented gene, whose genomic region and associations with any pathways, interactions or other groups can be retrieved from the knowledge database. |
| Group |  | Specified by a name or other identifier, i.e. "lipid metabolic process" or "GO:0006629". Used to refer to a known and documented pathway, ontological group, protein interaction, protein family, or any other grouping of genes, proteins or genomic regions that was provided by one of the external data sources. |
| Source |  | Specified by name, i.e. "GO". Used to refer to a specific external data source. |

LD Profiles : GWAS information

- Biofilter and LOKI allow for gene regions to be adjusted by the linkage disequilibrium (LD) patterns in a given population.
- When comparing a known gene region to any other region or position (such as CNVs or SNPs), areas in high LD with a gene can be considered part of the gene, even if the region lies outside of the gene's canonical boundaries.
- This step require use of additional tool

Biofilter : Command lines vs Configuration

- ❑ Biofilter can be run from a command-line terminal by executing

biofilter.py or python biofilter.py

- ❑ All options can either be provided directly on the command line

biofilter.py --option-name

- ❑ configuration files could be given as input such as

biofilter.py analysis.config

Biofilter : Command lines vs Configuration

- ❑ Options on the command line are lower-case, start with two dashes and may contain single dashes to separate words (such as “--snp-file”),
- ❑ while in a configuration file the same option would be in upper-case, contain no dashes and instead use underscores to separate words (i.e. “SNP_FILE”).
- ❑ Many command line options also have alternative shorthand versions of one or a few letters, such as “-s” for “--snp-file” and “--aag” for “--allow-ambiguous-genes”.

Output Options : Mode of analysis

--filter / FILTER

Argument: <type> [type] [...] Default: *none*

Perform a filtering analysis which outputs the specified type

--annotate / ANNOTATE

Argument: <type> [type] [...] [:] <type> [type] [...] Default: *none*

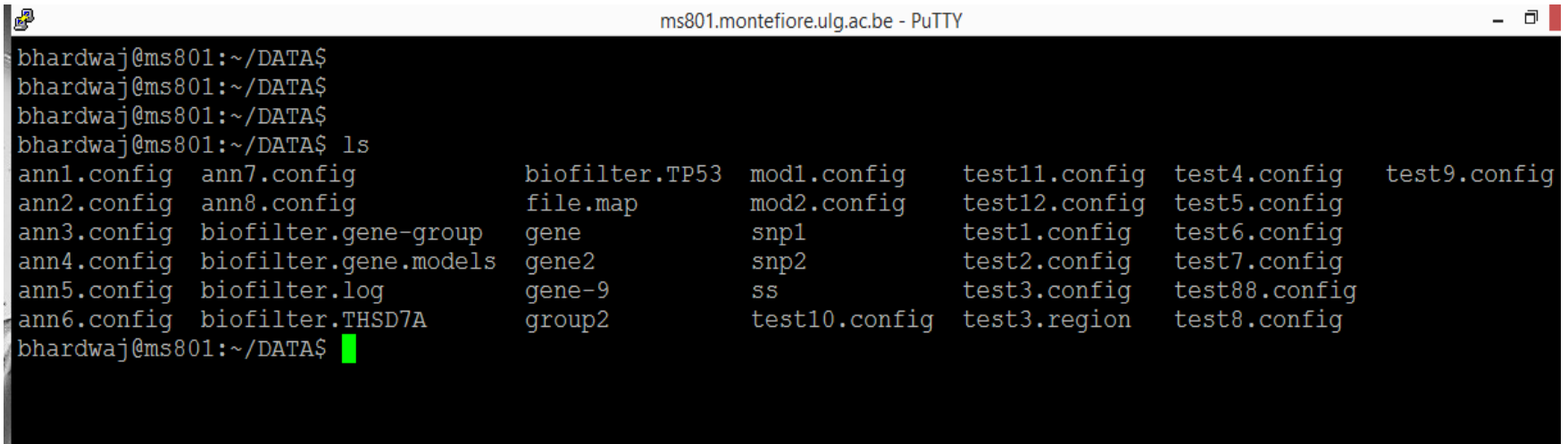
--model / MODEL

Argument: <type> [type] [...] [:] [type] [...] Default: *none*

Server Connectivity

❑ Connect to ms801 machine

❑ Use your login and password



```
ms801.montefiore.ulg.ac.be - PuTTY
bhardwaj@ms801:~/DATA$
bhardwaj@ms801:~/DATA$
bhardwaj@ms801:~/DATA$
bhardwaj@ms801:~/DATA$ ls
ann1.config  ann7.config  biofilter.TP53  mod1.config  test11.config  test4.config  test9.config
ann2.config  ann8.config  file.map        mod2.config  test12.config  test5.config
ann3.config  biofilter.gene-group  gene           snp1          test1.config   test6.config
ann4.config  biofilter.gene.models  gene2          snp2          test2.config   test7.config
ann5.config  biofilter.log          gene-9         ss            test3.config   test88.config
ann6.config  biofilter.THSD7A      group2         test10.config  test3.region   test8.config
bhardwaj@ms801:~/DATA$
```

Get the genes statistic of loki.db

❑ Open terminal and type

```
biofilter.py --knowledge loki.db --report-gene-name-stats yes
```

❑ This indicates number of genes belongs to each category

| #type | names | unique | ambiguous |
|--------------|--------|--------|------------|
| symbol | 117857 | 115238 | 2619 |
| entrez_gid | | 81664 | 81664 0 |
| uniprot_pid | | 32983 | 32668 315 |
| ensembl_gid | | 75453 | 75369 84 |
| pharmgkb_gid | | 26650 | 26650 0 |
| refseq_gid | | 189201 | 189201 0 |
| refseq_pid | | 117448 | 117448 0 |
| ensembl_pid | | 41732 | 41732 0 |
| hgnc_id | 41163 | 41163 | 0 |
| mim_id | 17130 | 17130 | 0 |
| vega_id | 19138 | 19123 | 15 |
| mirbase_id | | 1879 | 1879 0 |
| unigene_gid | | 29152 | 27997 1155 |

Get the groups statistic of loki.db

Open terminal and type

- ❑ `biofilter.py --knowledge loki.db --report-group-name-stats`
`yes`
- ❑ This indicates no of entries represented by each group

| #type | names | unique | ambiguous |
|---------------|-------|--------|-----------|
| biogrid_id | | 371104 | 371104 0 |
| go_id | 44957 | 44957 | 0 |
| ontology | 44957 | 44957 | 0 |
| kegg_id | 323 | 323 | 0 |
| pathway | 2605 | 2588 | 17 |
| netpath_id | | 28 | 28 0 |
| oreganno | 23393 | 23393 | 0 |
| pfam_id | 16718 | 16718 | 0 |
| proteinfamily | | 32501 | 32164 337 |
| pharmgkb_id | | 108 | 108 0 |
| reactome_id | | 2163 | 2163 0 |
| ucsc_ecr | 77858 | 77858 | 0 |

Comparison of Two SNPs list

- ❑ Download snp1 and snp2 from course website
- ❑ Create file name test1.config

```
KNOWLEDGE loki.db
SNP_FILE snp1
SNP_FILE snp2
FILTER snp
```

- ❑ Run biofilter.py test1.config
- ❑ As a result , you will get two output files named as biofilter1.log , biofilter1.snp

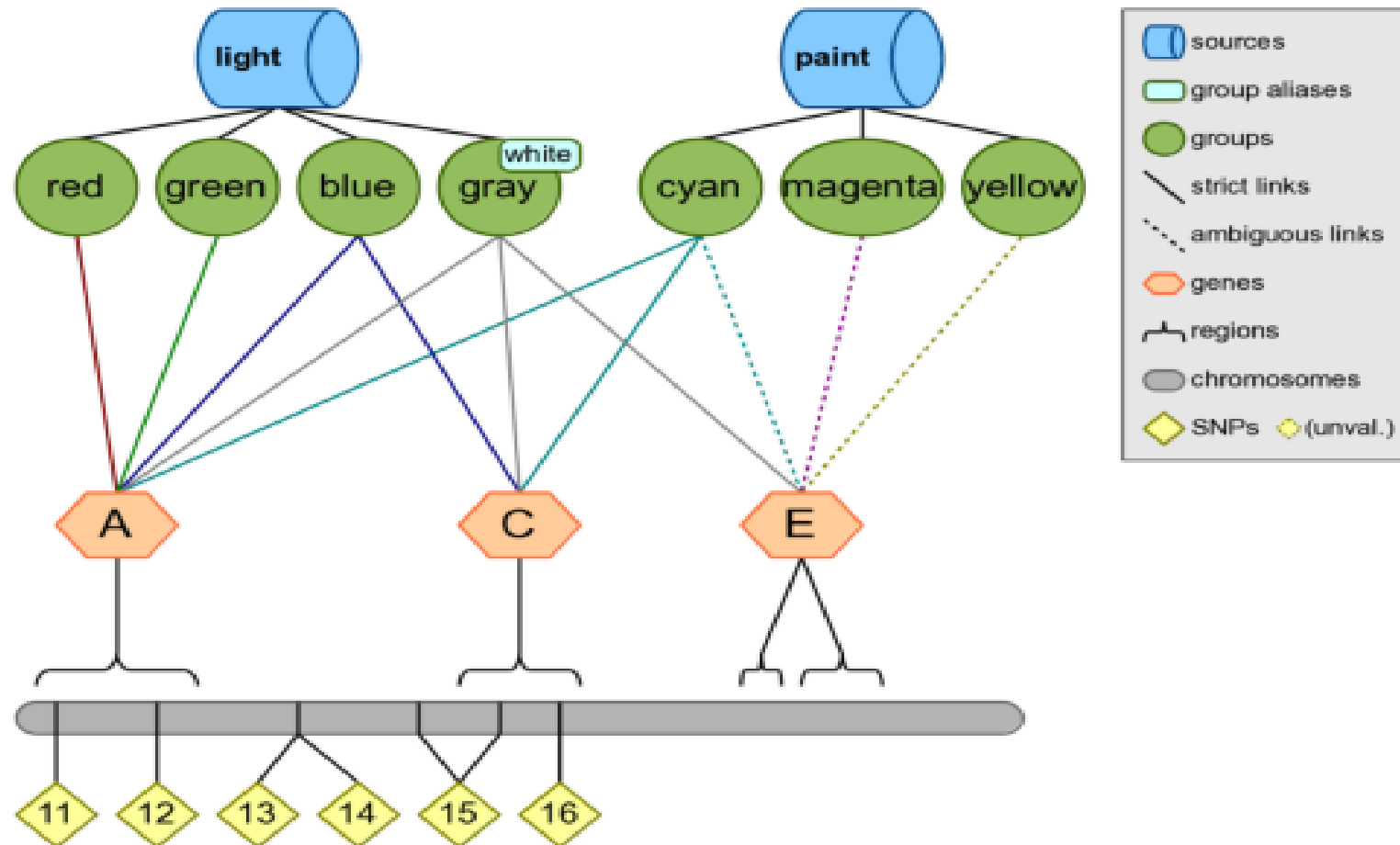
biofilter1.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main SNP filter ...  
... OK: added 4 SNPs (1 RS#s merged)  
reducing main SNP filter ...  
... OK: kept 1 SNPs (3 dropped, 0 RS#s merged)  
writing 'snp' filter to 'biofilter.snp' ...  
.. OK: 1 results
```

biofilter1.snp

```
#snp  
rs62653571
```

Let us find the SNPs falling on genes (1)



Let us find the SNPs falling on genes (2)

- ❑ Create file name test2.config , define snp1 as input

```
KNOWLEDGE loki.db
SNP_FILE snp1
GENE_FILE gene
FILTER snp
```

← Type of filter

- ❑ Run as `biofilter.py test2.config`

- ❑ As a result you will get two files :
`biofilter.log`
`biofilter.snp`

biofilter.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main SNP filter ...  
... OK: added 4 SNPs (1 RS#s merged)  
adding to main gene filter ...  
... OK: added 2 genes  
writing 'snp' filter to 'biofilter.snp' ...  
... OK: 4 results
```

biofilter.snp

```
#snp  
rs62653571  
rs2071569  
rs2075596  
rs6533526
```



Let us find the groups contains specific “regions”

- ❑ create `regions.config` which contain group information “R-HSA-5083635”

```
KNOWLEDGE loki.db
GROUP R-HSA-5083635
FILTER region
```

- ❑ Run as `biofilter.py regions.config`
- ❑ As a result you will get two files :
 - `biofilter.log`
 - `biofilter.region`

biofilter.region :It will consist of all regions belongs to group

| #chr | region | start | stop |
|------|----------|-----------|---------------------|
| X | CFP | 47624213 | 47630305 |
| 15 | THBS1 | 39581079 | 39598918 |
| 6 | THBS2 | 169215780 | 169254114 |
| 5 | SEMA5A | 9035026 | 9546121 |
| 1 | ADAMTS4 | 161189725 | 161199080 |
| 4 | ADAMTS3 | 72280969 | 72568799 |
| 5 | ADAMTS2 | 179110851 | 179345430 |
| 21 | ADAMTS1 | 26836287 | 26845409 |
| 9 | ADAMTSL2 | | 133532164 133575519 |
| 4 | SPON2 | 1166932 | 1208962 |
| 11 | SPON1 | 13962637 | 14268133 |
| 9 | ADAMTS13 | | 133414339 133459403 |
| 11 | ADAMTS8 | 130404923 | 130428993 |
| 21 | ADAMTS5 | 26917912 | 26967120 |
| 15 | ADAMTS7 | 78759203 | 78811464 |
| 5 | ADAMTS6 | 65148736 | 65482014 |
| 13 | B3GLCT | 31199975 | 31332276 |
| 7 | SSPO | 149776042 | 149833965 |

Output a list of all genes within a data source

- ❑ Create `group.config` which contain source information

```
KNOWLEDGE loki.db
SOURCE biogrid pfam
FILTER gene
```

- ❑ Run as `biofilter.py group.config`
- ❑ As a result you will get two files :
 - `biofilter.log`
 - `biofilter.gene`

biofilter.gene : All genes belong to group PFAM and biogrid

```
#gene  
TRIM54  
HDGF  
EXOSC10  
JMJD6  
MC4R  
NEDD8  
COPS7A  
COG1  
SIAH1 ..... SO ON
```

Can you count the number of genes belong to “PFAM” only ????

Let us find genes associated with a pathway or group

- ❑ Create `tt.config` which contains the Genes detail

```
KNOWLEDGE loki.db
GENE THSD7A COG8 UBC
FILTER gene snp region group source
```

- ❑ Run as `biofilter.py tt.config`
- ❑ As a result , you will get two files
 - `biofilter.log`
 - `biofilter.snp.region.group.source`

biofilter.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main gene filter ...  
... OK: added 3 genes  
writing 'gene snp region group source' filter to 'biofilter.gene-snp-  
region-group-source' ...  
... OK: 1668612 result
```



Open file and check information types

biofilter.snp.region.group.source : It consist of following entries (top 15 lines)

| #gene | snp | chr | region | start | stop | group | source |
|--------|--------------|-----|--------|----------|----------|----------------|---------|
| THSD7A | rs983143041 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs1015982900 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs962840243 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs558399301 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs539304291 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs974293917 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs921514204 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs571860735 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs932952501 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs750203807 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs945700395 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs1042762941 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs114612380 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs370567942 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs912527472 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |
| THSD7A | rs934667503 | 7 | THSD7A | 11370435 | 11832198 | biogrid:612143 | biogrid |

Let us find a list of genes falling within a group.

- ❑ Let us create `ge-gr.config` file which contains the list of genes and group

KNOWLEDGE loki.db

GENE HIST1H3A KIAA2013 PQBP1 DCAF8

GROUP R-HSA-5173214

FILTER gene group

- ❑ Run as `biofilter.py test.config`

- ❑ output : `biofilter.log` , `biofilter.gene-group`

✓ biofilter.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main gene filter ...  
... OK: added 4 genes  
adding to main group filter ...  
... OK: added 1 groups  
writing 'gene group' filter to 'biofilter.gene-group' ...  
... OK: 0 results
```



Open file and check information types

Annotating a SNP with gene region information

- ❑ Let us create `annotate.config` file which contains the list of snps

```
KNOWLEDGE loki.db
SNP rs11 rs24 rs99
ANNOTATE snp region
```

- ❑ Run as `biofilter.py annotate.config`
- ❑ output : `biofilter.log` , `biofilter.gene-group`

❑ biofilter.log ,

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main SNP filter ...  
... OK: added 3 SNPs (0 RS#s merged)  
writing 'snp : region' annotation to 'biofilter.snp.region' ...  
... OK: 3 results
```

❑ biofilter.snp-region

Open file and check information types

Annotating SNPs with location information

- ❑ Let us create `annotate2.config` file which contains the list of snps

KNOWLEDGE loki.db

SNP rs11 rs24 rs99

ANNOTATE snp position

- ❑ Run as `biofilter.py annotate2.config`

- ❑ output : `biofilter.log` , `biofilter.snp-position`

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main SNP filter ...  
... OK: added 3 SNPs (0 RS#s merged)  
writing 'snp : position' annotation to 'biofilter.snp.position' ...  
... OK: 3 results
```

biofilter.snp-region

- Open file and check chromosomal number where SNPs are present and also define position .
- Can you calculate distance among SNPs (in base pairs)

Map a SNP to the groups and sources where the SNP is present

- ❑ Let us create `annotate3.config` file which contains the list of snps

```
KNOWLEDGE loki.db
SNP rs11 rs24 rs99
ANNOTATE snp group source
```

- ❑ Run as `biofilter.py annotate3.config`
- ❑ output : `biofilter.log` , `biofilter.snp-group-source`

biofilter.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...
... OK
knowledge database genome build: GRCh38 / UCSC hg38
adding to main SNP filter ...
... OK: added 3 SNPs (0 RS#s merged)
writing 'snp : group source' annotation to 'biofilter.snp.group-source' ...
... OK: 10 results
```

Open file and check information types

Annotating a base pair region with the list of SNPs in that region.

- ❑ Let us create `annotate4.config` file which contains genome position and chromosomal number

```
KNOWLEDGE loki.db  
REGION 1:30000:40000  
ANNOTATE snp region
```

- ❑ Run as `biofilter.py annotate4.config`
- ❑ output : `biofilter.log` , `biofilter.snp-regions`

biofilter.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
WARNING: UCSC hg# build version was not specified for region input;  
assuming it matches the knowledge database  
adding to main region filter ...  
... OK: added 1 regions  
writing 'snp : region' annotation to 'biofilter.snp.region' ...  
  calculating main region zone coverage ... OK  
... OK: 88 results
```

88 SNPs falling in that regions

□ biofilter.snp-regions

```
#snp chr region start stop  
rs534702355 1 chr1:30000-40000 30000 40000  
rs867282737 1 chr1:30000-40000 30000 40000  
rs62028215 1 chr1:30000-40000 30000 40000  
rs778316262 1 chr1:30000-40000 30000 40000  
rs28688489 1 chr1:30000-40000 30000 40000  
rs28628742 1 chr1:30000-40000 30000 40000  
rs28594168 1 chr1:30000-40000 30000 40000  
rs558169846 1 chr1:30000-40000 30000 40000
```

..... SO ON

Count SNPs falling from 6000-10000 genomic regions of chromosome 1, 2, 3, 4 and 5

Create 5 different config files.

Analyse result

Develop the bar graph based on SNPs count .

Mapping regions to genes based on percent of overlap.

- ❑ Let us create `annotate5.config` file which contains genome position and chromosomal number

```
KNOWLEDGE loki.db
REGION 1:3000:40000
REGION_MATCH_PERCENT 50
FILTER gene
```

- ❑ Run as `biofilter.py annotate5.config`

- ❑ output : `biofilter.log` , `biofilter.gene`

□ biofilter.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
WARNING: UCSC hg# build version was not specified for region input;  
assuming it matches the knowledge database  
adding to main region filter ...  
... OK: added 1 regions  
writing 'gene' filter to 'biofilter.gene' ...  
  calculating main region zone coverage ... OK  
... OK: 6 results
```

□ biofilter.gene

```
#gene  
DDX11L1  
MIR1302-2  
MIR6859-1  
MIR1302-2HG  
WASH7P  
FAM138A
```

Mapping regions to genes based on base pair overlap

- ❑ Let us create `annotate6.config` file which contains genome position and chromosomal number

```
KNOWLEDGE loki.db  
REGION 1:4000:10000  
REGION_MATCH_BASES 10  
FILTER gene
```

- ❑ Run as `biofilter.py annotate6.config`

- ❑ output : `biofilter.log` , `biofilter.gene`

Open `biofilter.gene` and check genes count

Pair wise Gene-Gene and SNP-SNP interaction

Step 1

Map the input list of SNPs to genes within Biofilter.

Step 2

Connect, pairwise, the genes that contain SNPs in the input list of SNPs.

Step 3

Break down the gene-gene models into all pairwise combinations of SNPs across the genes within sources

Step 1

❑ Let us create mod1.config file which contains snp list

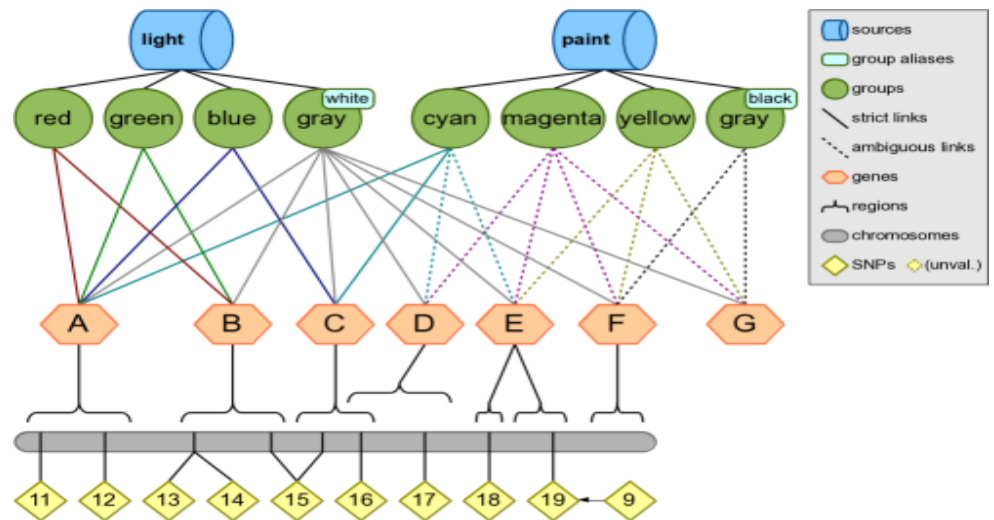
KNOWLEDGE loki.db

SNP rs983143041 rs101598290 rs962840243 rs558399301 rs539304291
rs974293917 rs921514204 rs571860735 rs932952501 rs750203807
rs945700395 rs1042762941 rs114612380

FILTER gene

❑ Run as biofilter.py
mod1.config

❑ output : biofilter.log ,
biofilter.gene



```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main SNP filter ...  
... OK: added 13 SNPs (0 RS#s merged)  
writing 'gene' filter to 'biofilter.gene' ...  
... OK: 2 results
```



2 SNPs falling in gene regions

```
#gene  
LOC105375153  
THSD7A
```

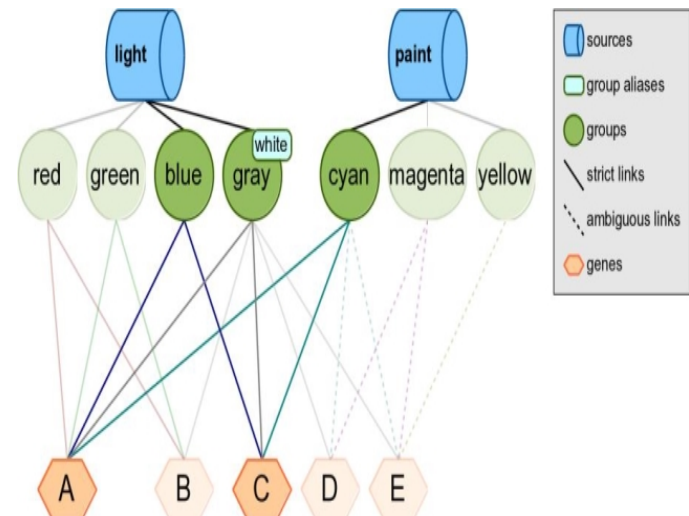
Step 2

❑ Let us create mod2 .config file which contains gene list observed in step 1

```
KNOWLEDGE loki.db
GENE LOC105375153 THSD7A
MODEL gene
```

❑ Run as `biofilter.py mod2.config`

❑ output :
`biofilter.log` ,
`biofilter.gene`



biofilter.log

```
loading knowledge database file '/usr/local/bin/loki.db' ...  
... OK  
knowledge database genome build: GRCh38 / UCSC hg38  
adding to main gene filter ...  
... OK: added 2 genes  
writing 'gene' models to 'biofilter.gene.models' ...  
  identifying main model candidates ... OK: 2 candidates  
  identifying candidate model groups ... OK: 161848 groups  
  calculating baseline models ... OK: 0 models  
... OK: 0 results
```

0 genes validated in modelling step .
These genes have no interaction .

Let us create mod11.config file which contains snp list

KNOWLEDGE loki.db

SNP rs268 rs316 rs326 rs328 rs333 rs334 rs544 rs551 rs567 rs662 rs669 rs671 rs683
rs684 rs688 rs689 rs690 rs693 rs694 rs695 rs696 rs698 rs699 rs700 rs703 rs705 rs712
rs715 rs835 rs868 rs900 rs910 rs958 rs1124 rs1164 rs1182 rs1183 rs1208 rs1303
rs1321 rs1421 rs1442 rs1506 rs1510 rs1545 rs1547 rs1590 rs1748 rs2506 rs2566
rs2688 rs2689 rs2765 rs2767 rs2942 rs2962

FILTER gene

Let us create mod22.config file which contains gene list from step 1 .

Check output . Did you get any output ? YES

| #gene1 | gene2 | score(src-grp) |
|--------|-------|----------------|
| LIPC | LPL | 4-12 |
| INS | INSR | 3-16 |
| APOB | LDLR | 3-6 |
| APOB | LPL | 3-5 |
| ADH1C | ALDH2 | 3-4 |
| APOB | LIPC | 3-4 |
| LDLR | LIPC | 3-4 |
| MKKS | CCT5 | 3-4 |
| SNRPN | SNURF | 2-8 |
| GH1 | INS | 2-3 |
| INS | PAX6 | 2-2 |
| INS | HNF1B | 2-2 |



These genes are interacting in pair wise manner

- ❑ Download SNP from NCBI (like 100 snps), check their gene information.**
- ❑ Analyse which genes are interacting and back trace the SNPs based on that interaction result.**
- ❑ Analyse their genomic position.**
- ❑ Calculate distance among SNPs.**