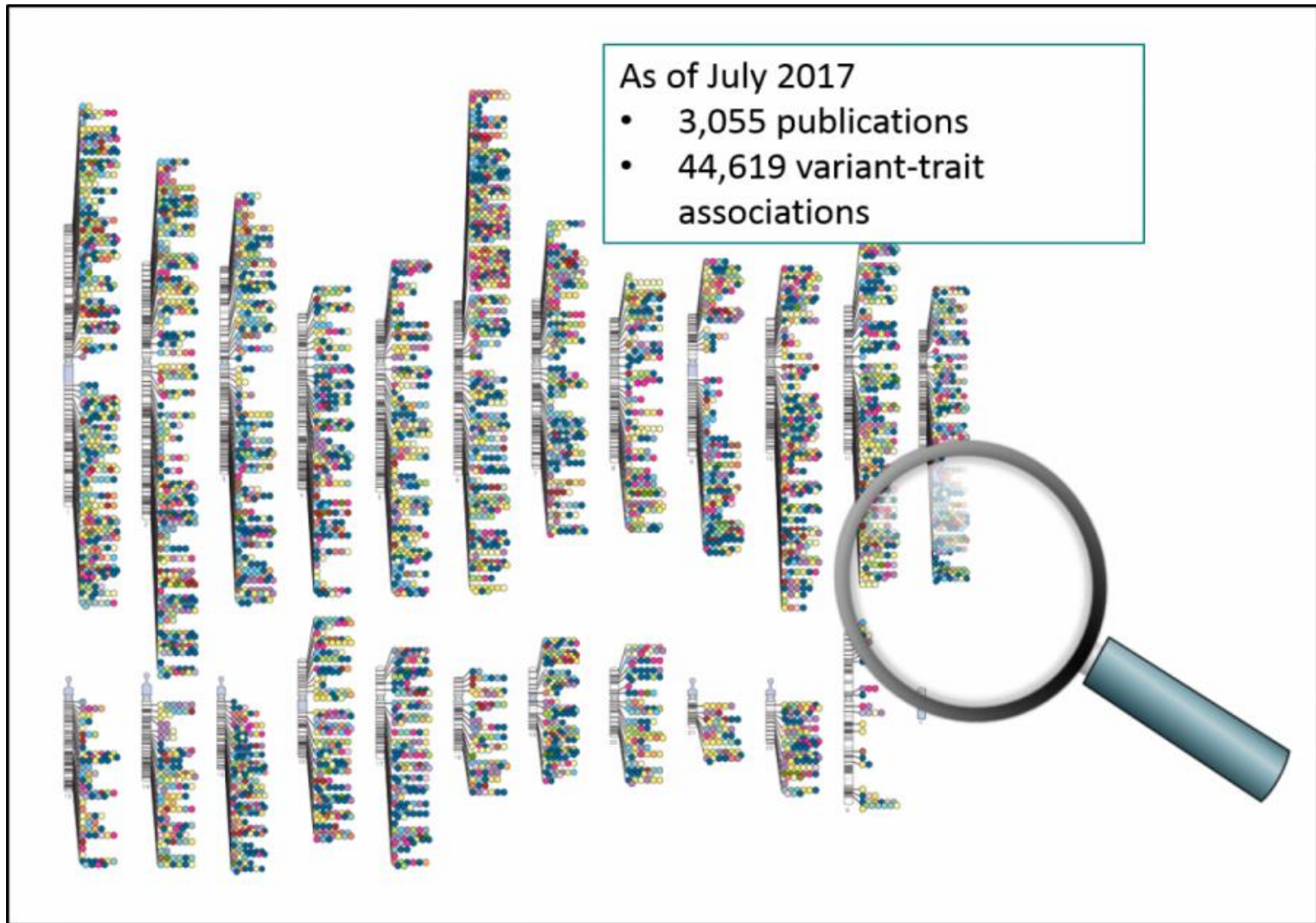


# **Gene-Gene /SNP-SNP Interaction: BIOFILTER**

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**

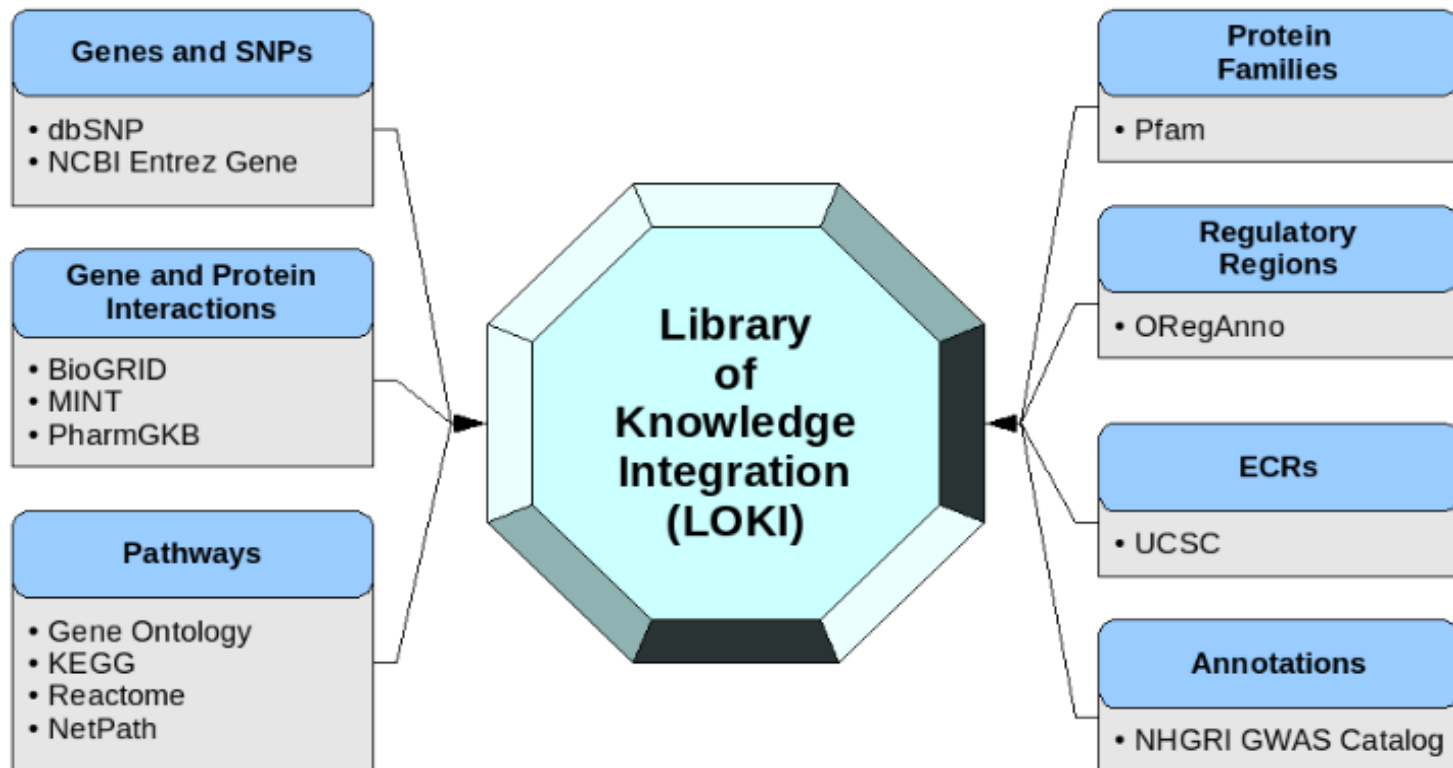
Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA

### **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

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 ID\(rs\)](#)
  - [Reference SNP ID\(rs\)](#)

SNP



homo sapiens

[Create alert](#) [Advanced](#)

Display Settings: ▼ Summary, 20 per page, Sorted by SNP\_ID

[Send to](#)**Search results****Items: 1 to 20 of 336845724**[First](#) [< Prev](#) Page  of 16842287 [Next >](#) [Last](#) [rs248](#) [*Homo sapiens*]

1.

ATTTTCTTTTTCTTCCAAAGGAGGA [A/G] TTAACTACCCTCTGGACAAATGTCC

Chromosome: 8:19953315

Gene: LPL ([GeneView](#))

Functional Consequence: synonymous codon

Clinical significance: Likely benign

Validated: by 1000G,by cluster,by frequency,by hapmap,by submitter

Global MAF: A=0.0387/194

HGVS: NC\_000008.10:g.19810826G&gt;A, NC\_000008.11:g.19953315G&gt;A,

NG\_008855.1:g.19245G&gt;A, NM\_000237.2:c.435G&gt;A, NP\_000228.1:p.Glu145

[PubMed](#) [View](#) [rs268](#) [*Homo sapiens*]

2.

TGCAACAATCTGGGCTATGAGATCA [A/G] TAAAGTCAGAGCCAAAAGAAGCAGC

Chromosome: 8:19956018

Gene: LPL ([GeneView](#))

Functional Consequence: missense

Allele Origin: A(germline)/G(germline)

Clinical significance: Pathogenic

Validated: by 1000G,by cluster,by frequency,by hapmap

Global MAF: G=0.0052/26

HGVS: NC\_000008.10:g.19813529A&gt;G, NC\_000008.11:g.19956018A&gt;G,

NG\_008855.1:a.21948A&gt;G, NM\_000237.2:c.953A&gt;G, NP\_000228.1:p.Asn318Ser

# LOKI DB : KEGG database

<http://www.genome.jp/kegg/pathway.html>



## KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

hsa for “human”

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


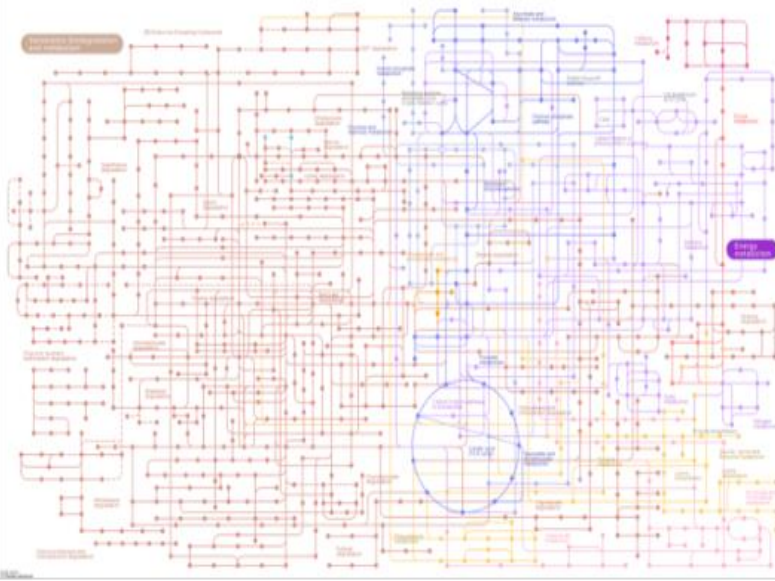
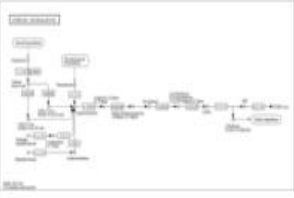
Each pathway map is identified by the combination of 2-4 letter prefix code and 5 digit number (see [KEGG Identifier](#)). The prefix has the following meaning:

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:

011	global map (lines linked to KOs)
-----	----------------------------------



map01100		Metabolic pathways	<p>...15983 (kshB), 1.14.13.142, R09860 R09885 K16047 (<i>hsaA</i>), K16048 (<i>hsaB</i>), 1.14.14.12, R09819 K16049 (hs...</p>	Neomycin, kanamycin and gentamicin biosynthesis Glycosaminoglycan biosynthesis - chondroitin sulfa...
map01120		Microbial metabolism in diverse environments	<p>...99.5, R00295 3.12.1.1, 3.12.1.1, R01930 K08352 (<i>phsA</i>), K08353 (<i>phsB</i>), K08354 (<i>phsC</i>), 1.8.5.5, R10149...</p>	Vitamine B6 metabolism Xylene degradation Glyoxylate and dicarboxylate metabolism Aminobenzoate ...
map00984		Steroid degradation	<p>...125A), 1.14.13.141, R11357 R09885 R09885 K16047 (<i>hsaA</i>), K16048 (<i>hsaB</i>), 1.14.14.12, R09819 K16049 (hs...</p>	STEROID DEGRADATION Cholest-4-en-3-one 1.1.3.6 1.14.13.141 (25S)-3-Oxo- cholest-4-en-26-oate 9alpha-...



# LOKI DB : BioGRID Database

BioGRID 3.4

[home](#) [help](#) [wiki](#) [tools](#) [contribute](#) [stats](#) [downloads](#) [partners](#) [about us](#)



## Welcome to the Biological General Repository for Interaction Datasets

BioGRID is an interaction repository with data compiled through comprehensive curation efforts. Our current index is version **3.4.155** and searches **63,959** publications for **1,507,991** protein and genetic interactions, **27,785** chemical associations and **38,559** post translational modifications from major model organism species. All data are **freely** provided via our search index and available for download in standardized formats.

[INTERACTION STATISTICS](#)

[LATEST DOWNLOADS](#)

## Search the BioGRID

Search by identifiers, keywords, and gene names...

p53

Homo sapiens

[SUBMIT GENE SEARCH Q](#)



[Advanced Search](#)



[Search Tips](#)



[Featured Datasets](#)

By Gene

By Publication

## AREAS OF INTEREST TO HELP YOU GET STARTED



### Build and Download Interaction Datasets

Create custom interaction datasets by protein or by publication. You can also download our entire dataset in a wide variety of standard formats.



### Online Tools and Resources

We've developed tools that make use of BioGRID data. Check out the list of tools to see if we can help you work with our data.



### Link To Us or Submit Interactions

Send us your datasets or link to the BioGRID directly from your own website or database. Full details on how to contribute are available here.



### View Our Interaction Statistics

Find out how many organisms, proteins, publications, and interactions are available in the current release of the BioGRID.

## BIOGRID FUNDING AND PARTNERS



[more partners](#)



# Result Summary

Gene / Identifier Search

p53

Homo sapiens

GO

## TP53

*Homo sapiens*

BCC7, LFS1, P53, TRP53

tumor protein p53

UBI NEDD FAT10 SUMO

GO Process (61)

GO Function (25)

GO Component (14)

### EXTERNAL DATABASE LINKOUTS

[HGNC](#) | [OMIM](#) | [VEGA](#) | [Entrez Gene](#) | [RefSeq](#) | [UniprotKB](#) | [Ensembl](#) | [HPRD](#)

Download 3000 Published Interactions For This Protein

### Stats & Options

#### Current Statistics

Publications: 1101

High Throughput

Low Throughput

514 (18%)

2877 Physical Interactions

2363 (82%)

104 (81%)

128 Genetic Interactions

24 (19%)

#### Search Filters

Customize how your results are displayed...

No Filter: Show All Associations



Switch View: **Interactors (1034)** Interactions (3005) Network Chemicals (2) PTM Sites (4)

Displaying 1 - 300 of 1034 total unique interactors

< Previous | **1** 2 3 4 | Next >

Sort By: **[Evidence]** [Alphabetical]

**MDM2** | ACTFS, HDMX, hdm2

MDM2 proto-oncogene, E3 ubiquitin protein ligase

UBI NEDD FAT10 SUMO

413 1

[details]

**EP300** | RP1-85F18.1, KAT3B, RSTS2, p300

E1A binding protein p300

UBI SUMO

85 1

[details]

# Use of Biofilter software (1)

- We can annotate genomic location or region based data, such as results from association studies, or CNV analyses, with relevant biological knowledge for deeper interpretation.**
  
- We can filter genomic location or region based data on biological criteria, such as filtering a series SNPs to retain only SNPs present in specific genes within specific pathways of interest.**

# Use of Biofilter software (2)

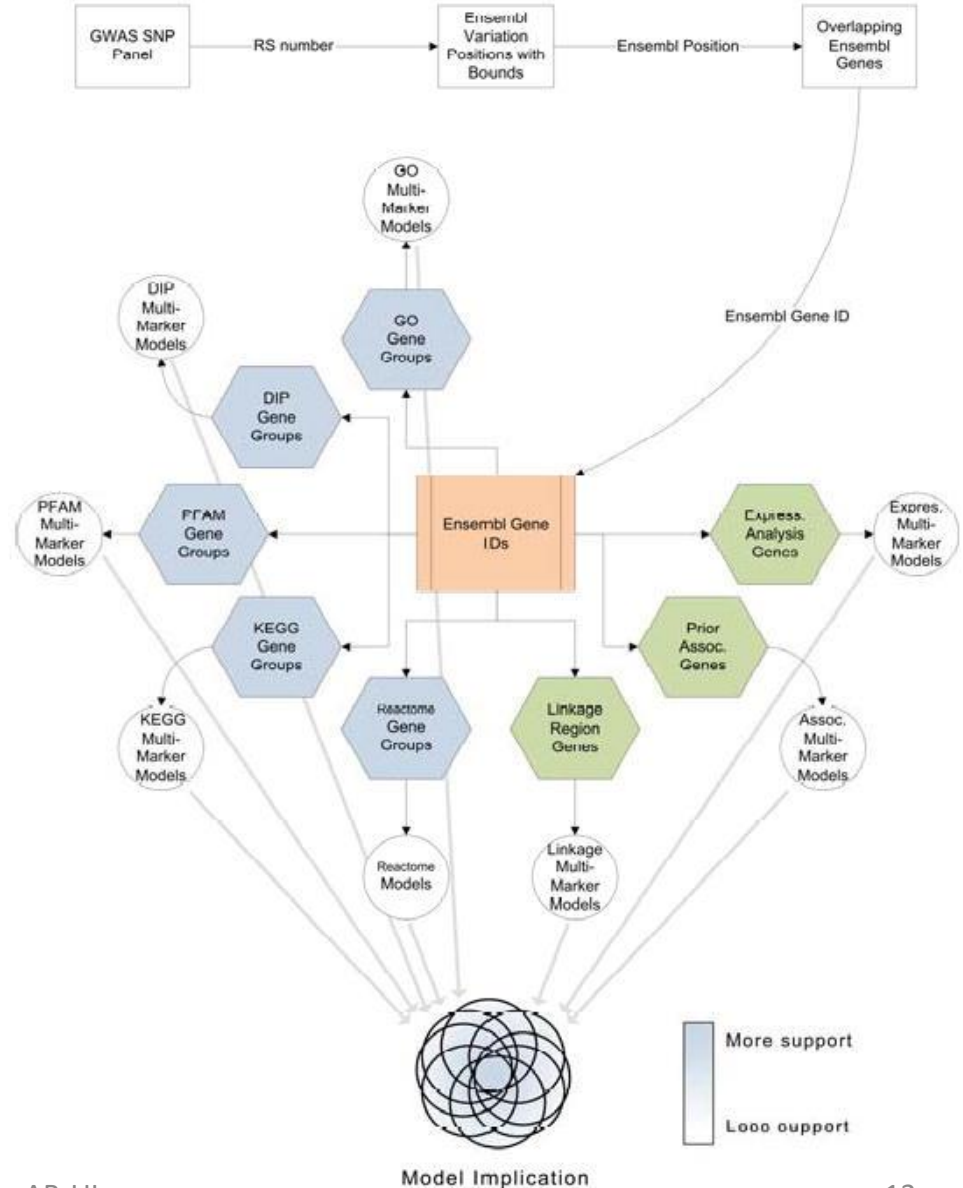
- **Biofilter allows researchers to annotate and/or filter data as well generate gene-gene interaction models based on existing biological knowledge.**
- **We can generate Predictive Models for gene-gene, SNP-SNP, or CNV-CNV interactions based on biological information, with priority for models to be tested based on biological relevance, thus narrowing the search space and reducing multiple hypothesis-testing.**

# Biofilter : Overview

□ GWAS platform SNPs are mapped to Ensembl gene Ids.

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

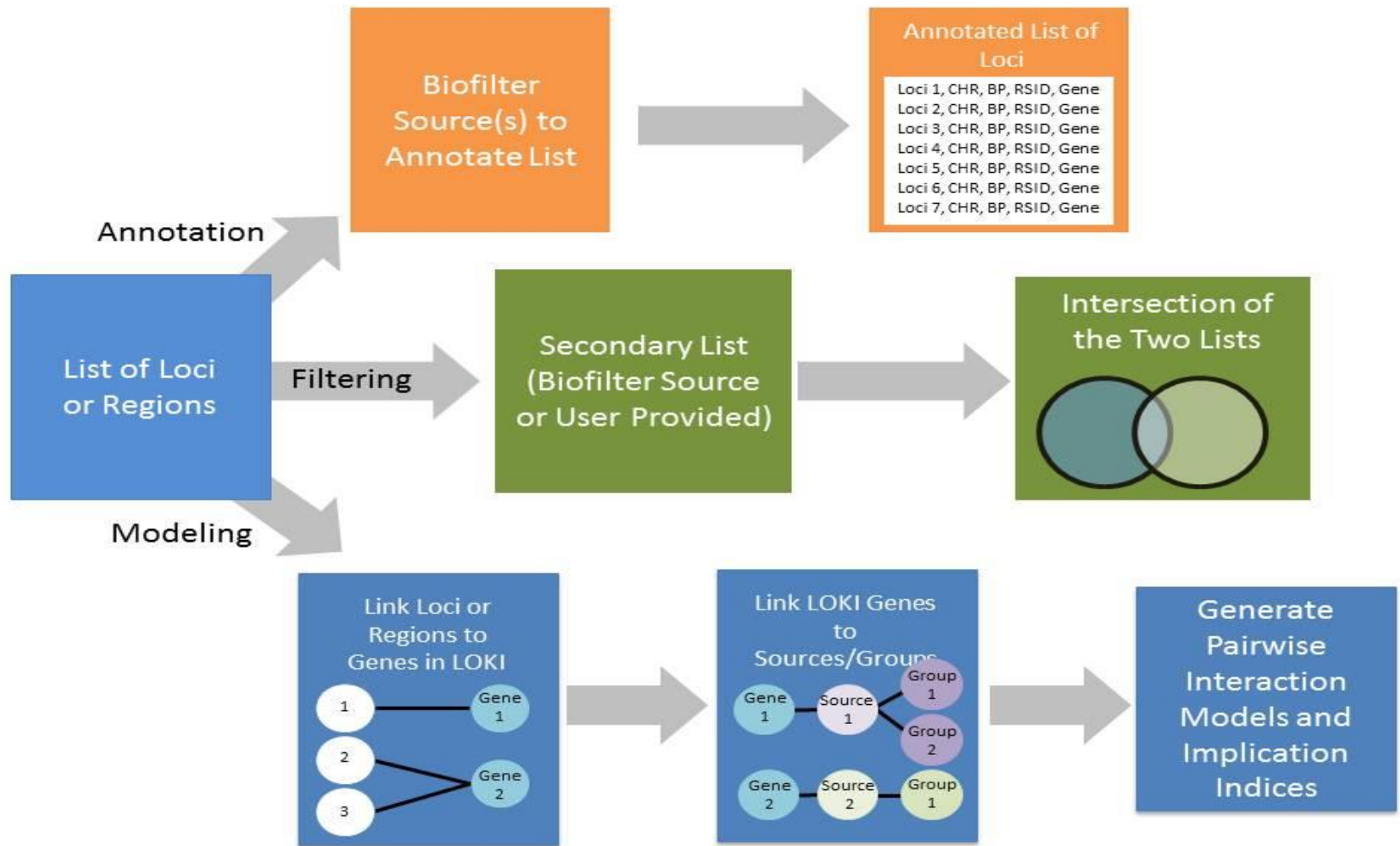
□ Derived models are overlaid to assess overall model implication.





# 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:

<b>SNP</b>		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.
<b>Position</b>		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.
<b>Region</b>		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.
<b>Gene</b>		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.
<b>Group</b>		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.
<b>Source</b>		Specified by name, i.e. "GO". Used to refer to a specific external data source.

# Biofilter : Filtering mode

□ Given any combination of input data, Biofilter can cross-reference the input data using the relationships stored in the knowledge database to generate a filtered dataset of any supported type (or types).

□ For example, a user can provide a list of SNPs (such as those covered by a genotyping platform) and a list of genes (such as those thought to be related to a particular phenotype) and request a filtered set of SNPs. Biofilter will use LOKI's knowledge of SNP positions and gene regions to filter the provided

□ SNP list, removing all those that are not located within any of the provided genes.

# Biofilter : Annotation mode

❑ The annotations are based on the relationships stored in the knowledge database; unlike filtering, any data which cannot be annotated as requested (such as a SNP which is not located within any gene) will still be included in the output, with the annotation columns of the output simply left blank.

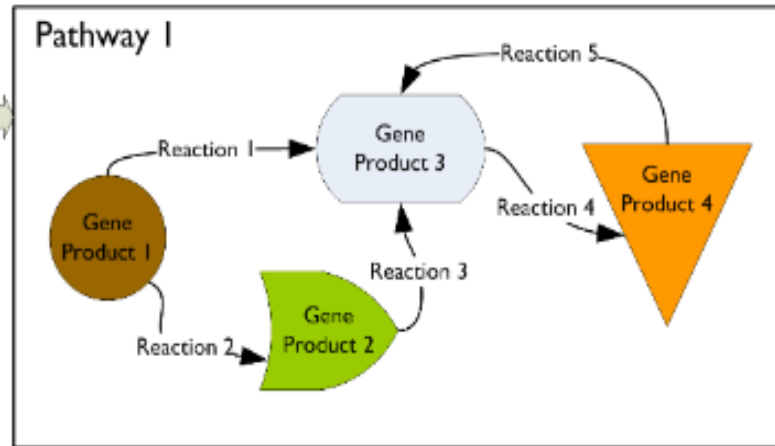
❑ For example, a list of SNPs can be annotated with positions to generate a new list of all the same SNPs, but with extra columns containing the chromosome and genomic position for each SNP (if any). Any SNP with multiple known positions will be repeated, and any SNP with no known position will have blanks in the added columns.

# Biofilter : Annotation mode

## Single Locus Statistical Results

SNP 1, Rs101841,  $p = 0.000163$   
SNP 2, Rs182645,  $p = 0.000268$   
SNP 3, Rs23876,  $p = 0.00324$   
SNP 4, Rs378645,  $p = 0.004354$   
SNP 5, Rs37564,  $p = 0.02341$   
SNP 6, Rs8751,  $p = 0.03412$   
SNP 7, Rs86745,  $p = 0.03685$   
SNP 8, Rs41254,  $p = 0.04675$

## Biofilter Analysis



## Annotated Statistical Results

### Results in the Same Gene

SNP 3, Rs23876,  $p = 0.00324$   
SNP 6, Rs8751,  $p = 0.03412$

### Results in the Same Pathway

SNP 2, Rs182645,  $p = 0.000268$   
SNP 3, Rs23876,  $p = 0.00324$   
SNP 6, Rs8751,  $p = 0.03412$   
SNP 7, Rs86745,  $p = 0.03685$

### Results with Biological Interaction

SNP 3, Rs23876,  $p = 0.00324$   
SNP 6, Rs8751,  $p = 0.03412$   
SNP 7, Rs86745,  $p = 0.03685$



# Biofilter : Model analysis mode(1)

**□The last of Biofilter's primary analysis modes is a little different from filtering and annotation.**

**□In addition to simply cross-referencing any given data with the other available prior knowledge, Biofilter can also search for repeated patterns within the prior knowledge which might indicate the potential for important interactions between SNPs or genes.**

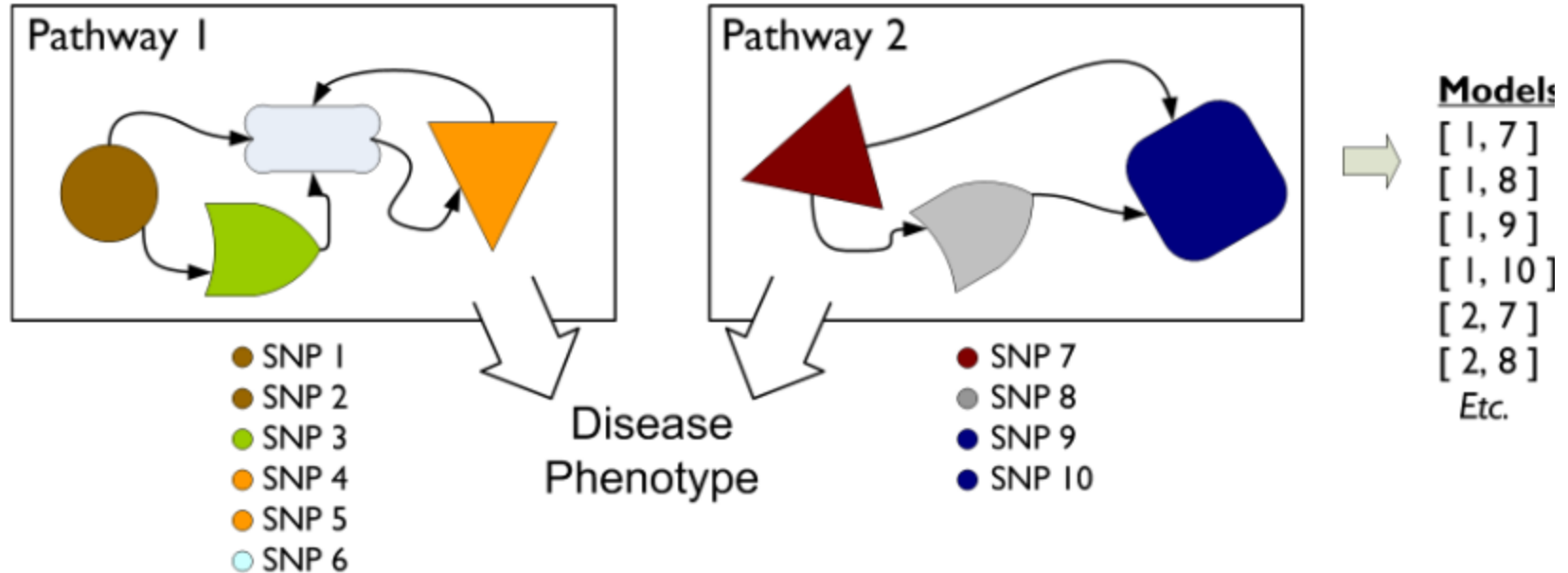
# Biofilter : Model analysis mode(2)

□ The key idea behind this analysis is that if the same two genes appear together in more than one grouping, they're likely to have an important biological relationship; if they appear in multiple groups from several independent sources, then they're even more likely to be biologically related in some way.



# Biofilter : Model analysis mode (3)

□ Biofilter has access to thousands of such groupings and can analyze all of them to identify the pairs of genes or SNPs appearing together in the greatest number of groupings and the widest array of original data sources. These pairs can then be tested for significance within a research dataset, avoiding the prohibitive computational and multiple testing Burden of an exhaustive pairwise analysis.



# Compiling Prior Knowledge: Loki.db

❑ The LOKI prior knowledge database must be generated before Biofilter can be used. This is done with the “loki-build.py” script which was installed along with Biofilter. There are several options for this utility which are detailed below, but to get started, you just need “--knowledge” and “--update”:

*loki-build.py --verbose --knowledge loki.db --update*

❑ This will download and process the bulk data files from all supported knowledge sources, storing the result in the file “loki.db” (which we recommend naming after the current date, such as “loki20140521.db”).

# Updating Prior Knowledge: Loki.db (1)

❑ ***--update Arguments: [source] [...] Default: all***

Instructs the build script to process the bulk data from the specified sources and update their representation in the knowledge database. If no sources are specified, all supported sources will be updated.

❑ ***--update-except Arguments: [source] [...] Default: none***

Similar to “--update” but with the opposite meaning for the specified sources: all supported sources will be updated **except for the ones specified. If no sources are specified, none are excluded, and all supported sources are updated.**

❑ ***--option Arguments: <source> <options> Default: none***

Passes additional options to the specified source loader module. The options string must be of the form “option1=value,option2=value” for any number of options and values. Supported options and values for each source can be shown with “12/5/2017 ***--list-sources***”.

# Updating Prior Knowledge : Loki.db (2)

□ *--force-update Argument: none*

The build script will normally only update from a sources if it detects that an update is necessary, either because new data files have been downloaded from the source or because the source's loader module code has been updated. With this option, the build script will update all specified sources, even if it believes no update is necessary.

# 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 : Configuration file

Input files:

<b>input1</b>	<b>input2</b>
#snp	#snp
rs9	rs14
rs11	rs15
rs12	rs16
rs13	rs17
rs14	rs18
rs15	rs19
rs16	

Configuration:

```
KNOWLEDGE test.db
SNP_FILE input1
SNP_FILE input2
FILTER snp
```

▪ ***biofilter.py test.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”.

# Configuration Options

## ❑ ***--help / HELP***

Displays the program usage and immediately exits.

## ❑ ***--version / VERSION***

Displays the software versions and immediately exits. Note that Biofilter is built upon LOKI and SQLite, each of which will also report their own software versions.

## ❑ ***--report-configuration / REPORT\_CONFIGURATION***

Argument: [yes/no]     Default: no

Generates a Biofilter configuration file which specifies the current effective value of all program options, including any default options which were not overridden.

# Prior Knowledge Options

***--knowledge / KNOWLEDGE***

Argument: <file>     Default: *none*

***--report-genome-build / REPORT\_GENOME\_BUILD***

Argument: [yes/no]     Default: *yes*

***--report-gene-name-stats / REPORT\_GENE\_NAME\_STATS***

Argument: [yes/no]     Default: *no*

***--report-group-name-stats / REPORT\_GROUP\_NAME\_STATS***

Argument: [yes/no]     Default: *no*

***--allow-unvalidated-snp-positions /  
ALLOW\_UNVALIDATED\_SNP\_POSITIONS***

Argument: [yes/no]     Default: *yes*

***--allow-ambiguous-snps / ALLOW\_AMBIGUOUS\_SNPS***

# Primary Input Data Options

## ❑ ***--snp / SNP***

Arguments: <snp> [snp] [...] Default: *none*

## ❑ ***--snp-file / SNP\_FILE***

Arguments: <file> [file] [...] Default: *none*

## ❑ ***--position / POSITION***

Arguments: <position> [position] [...] Default: *none*

## ❑ ***--position-file / POSITION\_FILE***

Arguments: <file> [file] [...] Default: *none*

## ❑ ***-region / REGION***

Arguments: <region> [region] [...] Default: *none*

## ❑ ***--region-file / REGION\_FILE***

Arguments: <file> [file] [...] Default: *none*

# 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*

# Filter mode : search SNPs that correspond to a list of genes

input1

input2

#snp

#gene

rs11

A

rs12

C

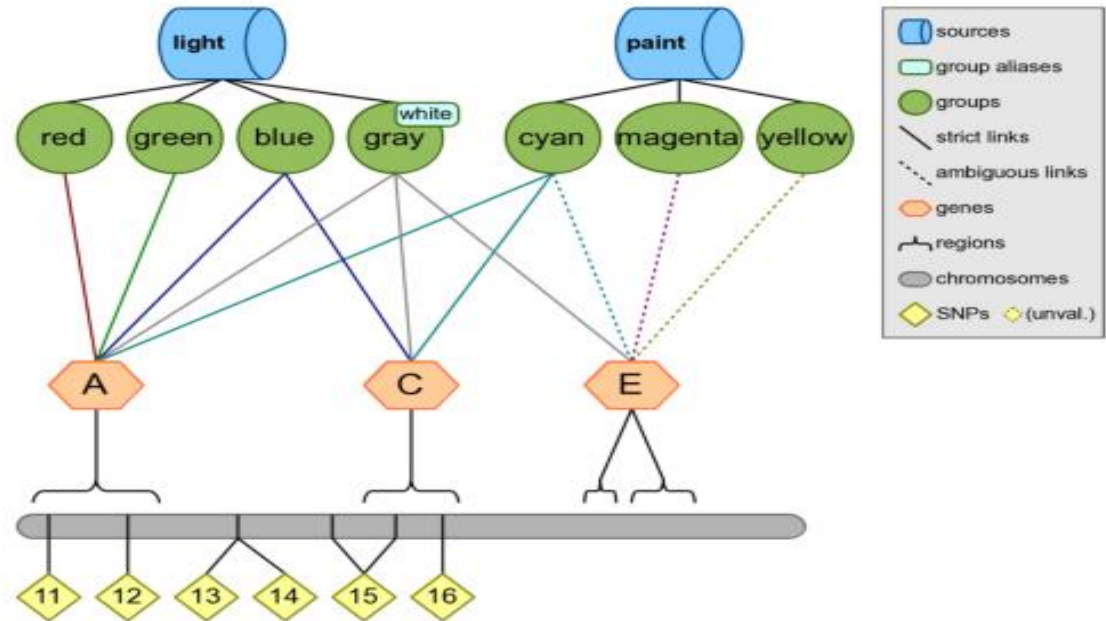
rs13

E

rs14

rs15

rs16



Test.config

KNOWLEDGE test.db

SNP\_FILE input1

GENE\_FILE input2

FILTER snp

run " biofilter.py Test.config"

What is expected output ??? Can you make inference by looking

figure

# Annotation mode : a SNP with gene region information

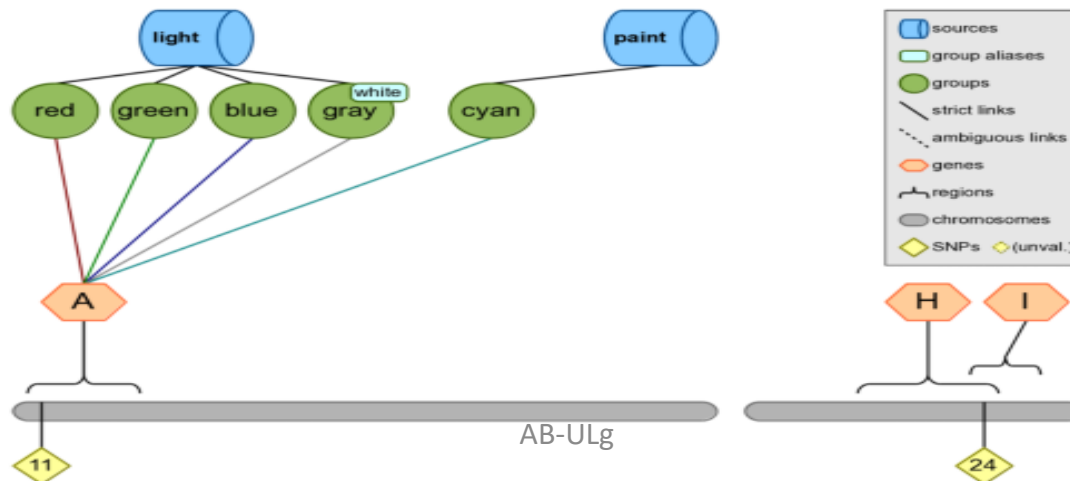
## ❑ Test.config

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

## ❑ Biofilter.py test.config

## ❑ Output

#snp	chr	region	start	stop
rs11	1	A	8	22
rs24	2	H	22	42
rs24	2	I	38	48
rs99				





# 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 : Pair wise Gene-Gene and SNP-SNP interaction

- ❑ we will use all of the SNPs on the first chromosome.
- ❑ Test.config

KNOWLEDGE test.db

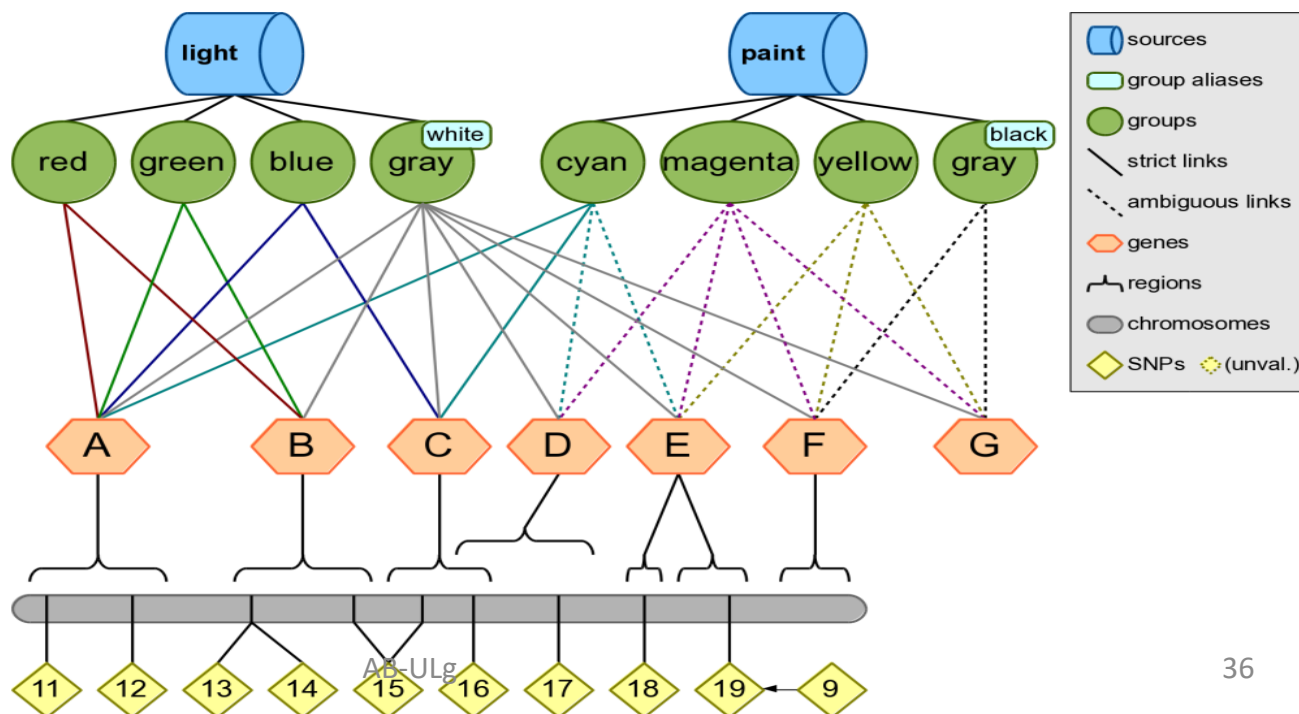
SNP 11 12 13 14 15 16 17 18 19

FILTER gene

❑ Output:

#gene

A  
B  
C  
D  
E



# Step 2 : Connect, pairwise, the genes that contain SNPs in the input list of SNPs.

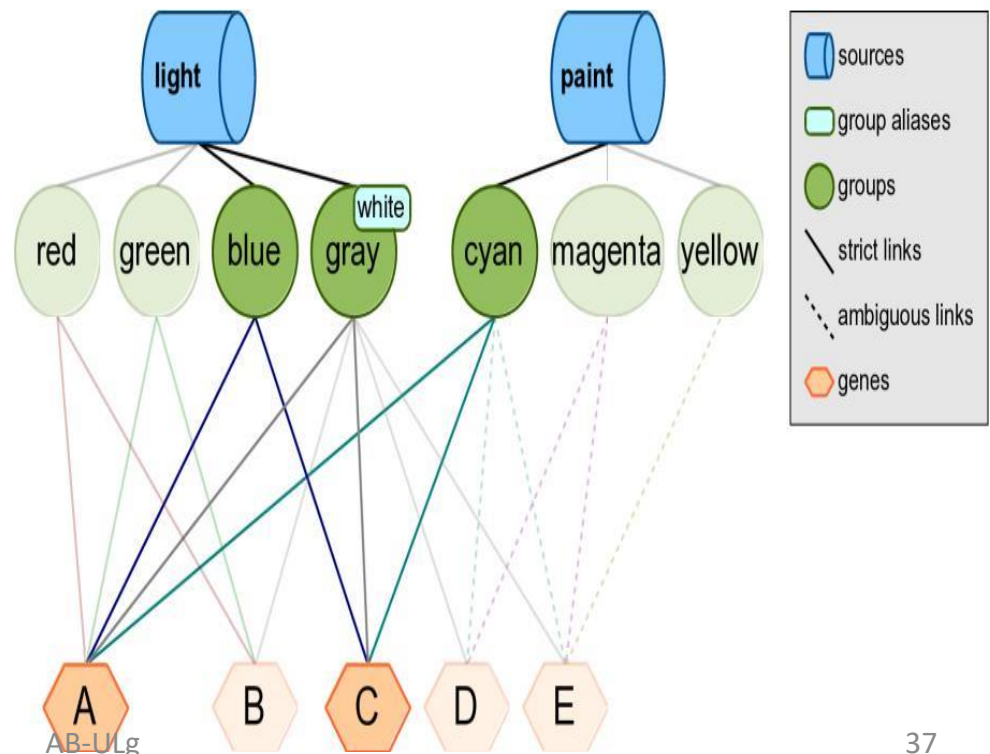
❑ Test.config

```
KNOWLEDGE test.db
GENE A B C D E
MODEL gene
```

❑ biofilter.py test.config

❑ output

#gene1	#gene2	score
A	C	2-3



# Step 3 : Break down the gene-gene models into all pairwise combinations of SNPs

Configuration:

`biofilter.py test.config`

```
KNOWLEDGE test.db
SOURCE light paint
MODEL snp
```

Output:

```
#snp1 snp2 score(src-grp)
rs11 rs15 2-3
rs11 rs16 2-3
rs12 rs15 2-3
rs12 rs16 2-3
```

