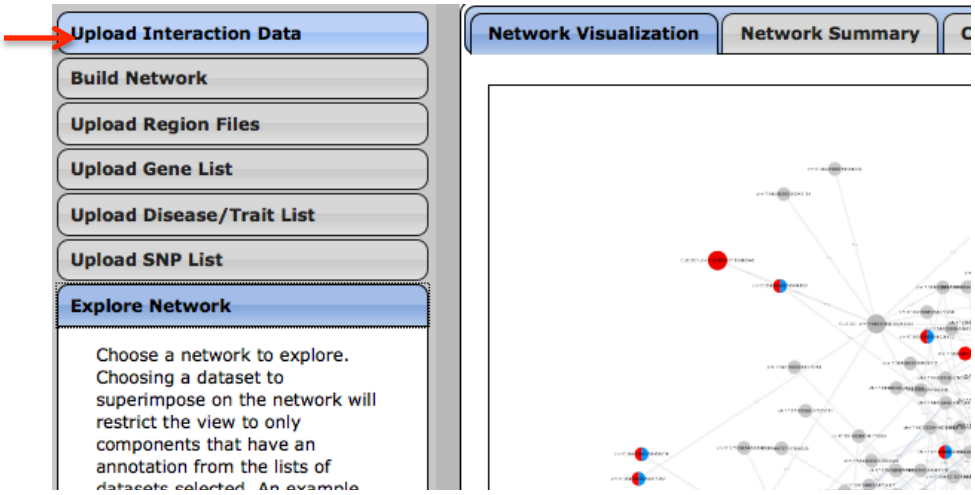
QuIN
Tutorial Manual

I. Uploading Data

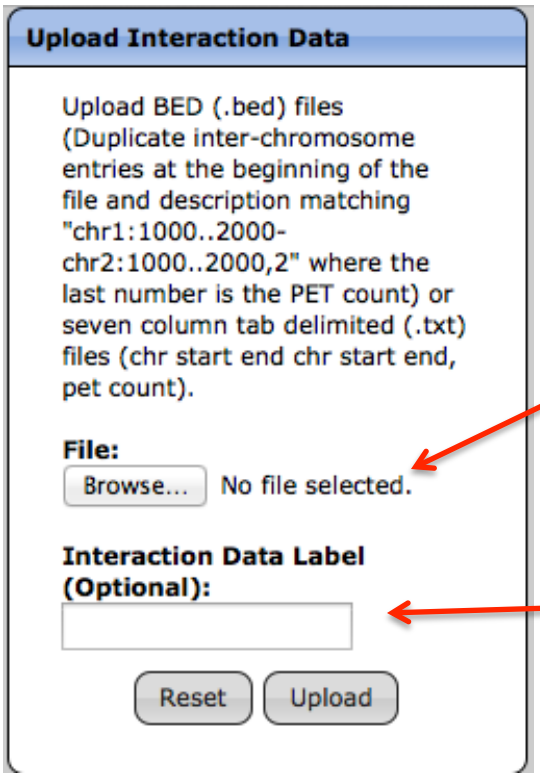
QuIN allows users to upload various sets of data for their interaction network analyses. Uploaded data will be stored as long as the user actively accesses the site, and will be removed after 30 days of inactivity. In order to keep data private while also maintaining a registration free web service, cookies are used for maintaining a user's identity and as a consequence, clearing cookies will remove access to the uploaded data. Additionally, QuIN is currently tailored for the hg19 reference genome and datasets should be uploaded with this in mind. Datasets for other genomes can be uploaded and analyzed, but information for SNPs, diseases, and promoters/genes will be misaligned. Building a network and/or superimposing data onto a network however are independent of the reference genome and these features can be used as long as the reference genome dependent features such as viewing the SNPs of a node are ignored. Currently there are five different types of data that can be uploaded to QuIN, and are further discussed below.

A. Interaction Data

Interaction data can be uploaded by first selecting the "Upload Interaction Data" option in the left panel.



After selecting, a menu will appear for selecting the file containing the interaction data and optionally labeling the data. If a label is not provided, the filename is used to refer to the data.



Clicking "Browse…" will open a file dialog to choose the .bed or .txt file containing the interaction data.

Optionally provide a label for the data.

Interaction data can be uploaded in two file formats, BED (.bed) and Text (.txt).

1. BED Format

The accepted BED format for interaction data is derived from the BED output files after interaction clustering for ChIA-PET data. This is the same format used to provide the ChIA-PET interaction clusters from ENCODE and is primarily meant for visualization in the UCSC Genome Browser. For this reason, QuIN exclusively uses the description of the interactions rather than the start and end position columns where the descriptions are expected to have the necessary information for each interaction. Below is an example of the interaction data for MCF-7 Replicate 3 from ENCODE.

Chromosome	Start	End	Description
chr9	132371263	132374233	chr9:132371263..132374233-chrX:12991997..12994687,2
chrX	12991997	12994687	chr9:132371263..132374233-chrX:12991997..12994687,2
chr1	757601	764067	chr1:757601..759455-chr1:761281..764067,2
chr1	832869	841188	chr1:832869..835690-chr1:838672..841188,2
chr1	836854	842379	chr1:836854..839705-chr1:840122..842379,2
chr1	839148	958325	chr1:839148..840925-chr1:956344..958325,2
chr1	851440	936146	chr1:851440..855732-chr1:932773..936146,4
chr1	853100	932621	chr1:853100..854849-chr1:931120..932621,2
chr1	855300	1011209	chr1:855300..857973-chr1:1009350..1011209,2
chr1	857967	1245066	chr1:857967..862189-chr1:1243062..1245066,4
chr1	858741	999731	chr1:858741..860915-chr1:998053..999731,2
chr1	858861	1247362	chr1:858861..860639-chr1:1245678..1247362,2

As shown above, the inter-chromosome interactions are listed first where each anchor or endpoint of these interactions are

listed as one interaction. For usage in the UCSC genome browser, this is required as each track can only show one chromosome and therefore it is expected that inter-chromosome interactions are at the beginning of the BED file and are duplicated. The rest of the file is expected to be only intra-chromosome interactions and one entry per interaction.

As mentioned before, it is expected that each description contains the interaction information using the following format as seen in the above example:

```
<chr>:<start>..<end>-
<chr>:<start>..<end>,<PET count>
```

Here, each interaction is defined as two genomic loci represented by a chromosome, start, end, and a PET count representing the number of pairs of reads showing an interaction between the two regions. Uploading BED formatted interaction data is primarily offered as a convenience and it is suggested that the TEXT format be used instead when converting interaction data to a format that is compatible with QuIN.

2. Text Format

The TEXT format is a tab-delimited file containing seven data elements per line:

1. **Chromosome** of the first interaction anchor (ex: Chr1, ChrY)
2. **Start** position of the first interaction anchor in bp
3. **End** position of the first interaction anchor in bp
4. **Chromosome** of the second interaction anchor (ex: Chr1, ChrY)
5. **Start** position of the second interaction anchor in bp
6. **End** position of the second interaction anchor in bp
7. **PET-Count/Read Count** of the number of reads showing an interaction between the two anchor regions. If a PET-Count/Read Count is not applicable, then a positive integer representation of a **score** (preferably "higher is better") should be used instead. Alternatively, if no read count or score is available setting this column to any integer value will suffice.

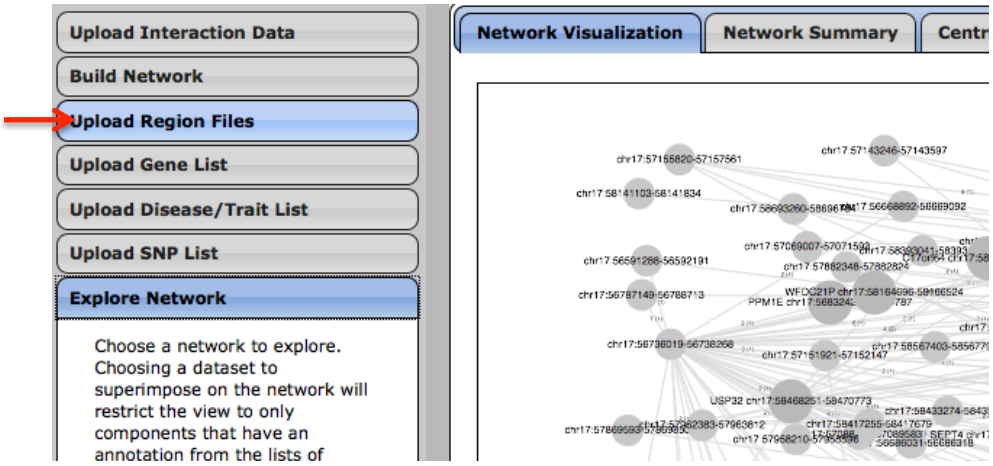
Below is an example of the format:

Chromosome A1	Start A1	End A1	Chromosome A2	Start A2	End A2	PET/Read Count
chr1	713971	714495	chr1	937010	937794	2
chr1	714125	714639	chr1	762369	762940	2
chr1	757365	758009	chr8	182518	183019	2
chr1	761369	762199	chr8	182999	183804	2
chr1	767238	768045	chr8	276760	277262	2
chr1	774410	775007	chr8	270215	270861	2
chr1	792044	792910	chr8	249210	250154	2

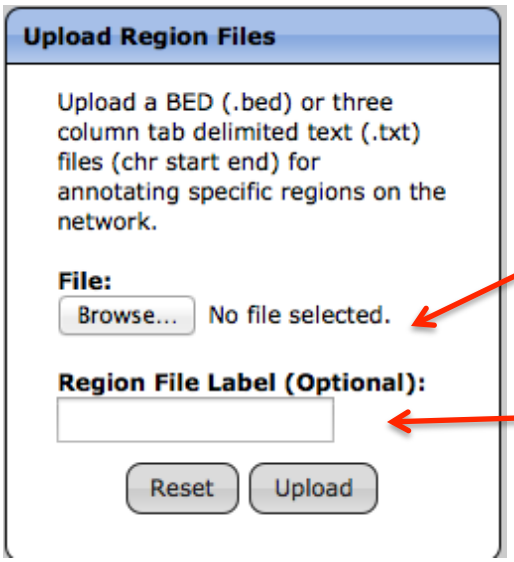
A1 and A2 represent anchors or end points 1 and 2 respectively of the interaction.

B. Genomic Regions

Genomic regions of interest from for example ChIP-Seq or DNASE-Seq datasets can be uploaded under the "Upload Region Files" option of the left panel:



After selecting, a menu will appear for selecting the file and optionally providing a label for the data.



Clicking "Browse…" will open a file dialog to choose the .bed or .txt file containing the region data.

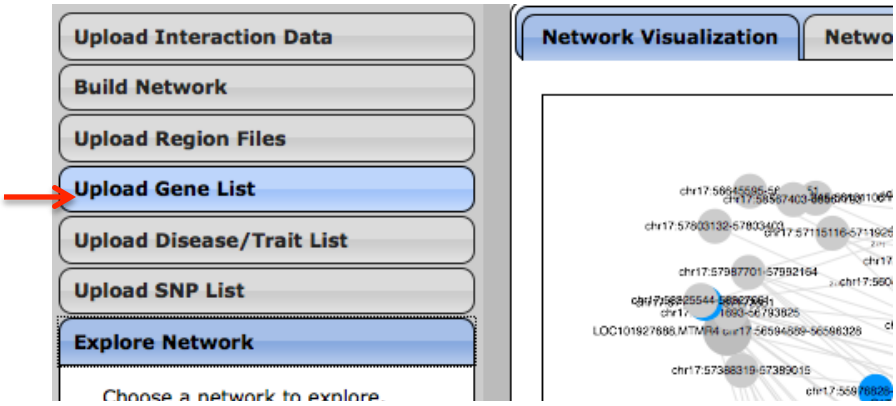
Optionally provide a label for the data.

Region data can be uploaded in two formats: BED (.bed) and Text (.txt) where in both file formats, the data is expected to be tab-delimited and have the first three elements per line be the chromosome, start, and end position of the genomic region. Additional columns can be included in the data but are ignored. An example of the expected data is shown below:

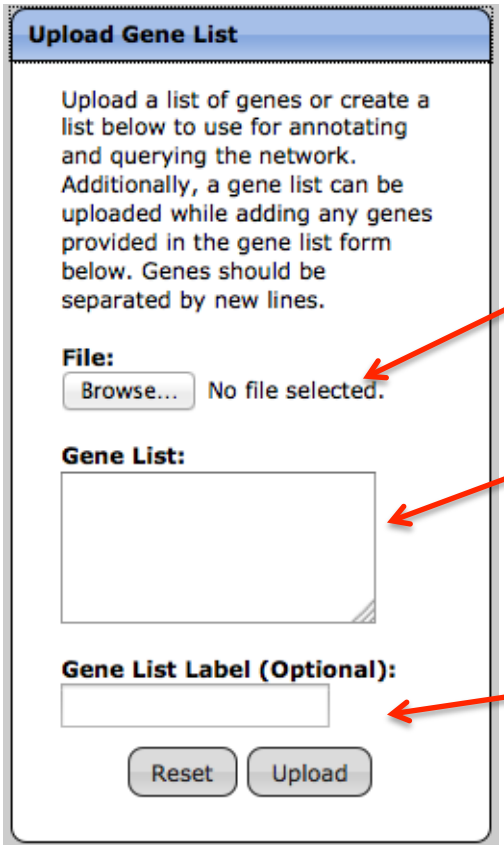
Chromosome	Start	End
chr1	713876	714575
chr1	725156	725395
chr1	752707	753462
chr1	762148	763148
chr1	777992	778502

C. Genes

Gene lists (using official gene symbols) can be uploaded by selecting the "Upload Gene List" option in the left panel.



Selecting this option will provide three forms: one for uploading a Text (.txt) file where each gene is its own line, one for providing a list of genes in a textbox where each gene is also its own line, and one for providing a label for the list of genes.



Clicking "Browse…" will open a file dialog to choose the .txt file containing the gene list.

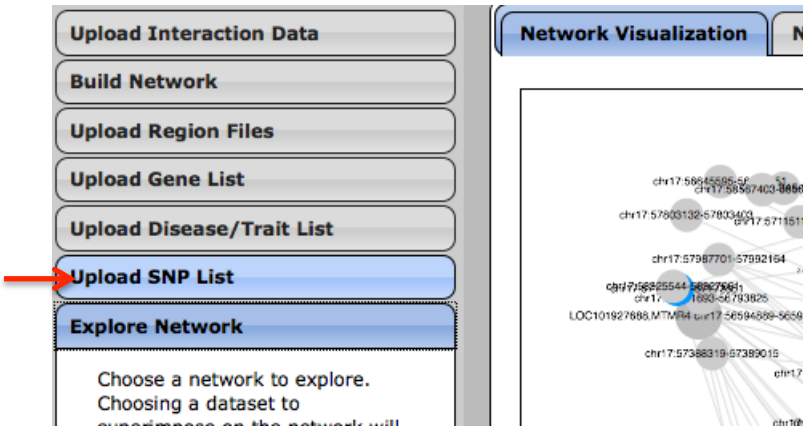
Provide a list of genes also included along with the list of genes from the file specified (if provided)

Optionally provide a label for the data.

Only one of the options (File or Gene List) needs to be provided to upload a list of genes. If both are provided, then the union of both sources is used as the list of genes.

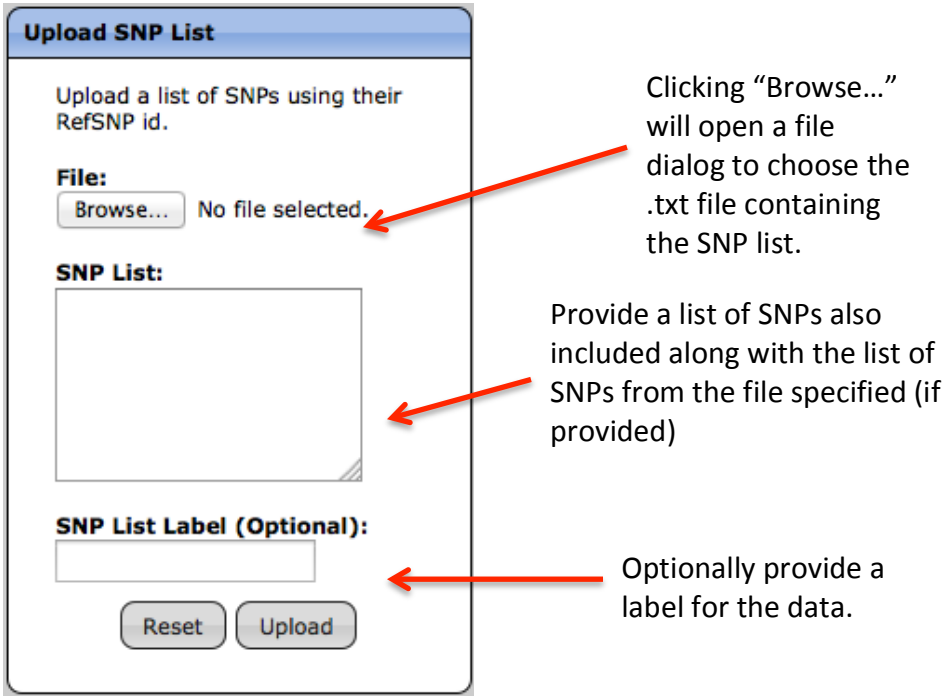
D. SNPs

SNP lists (using RefSNP ids ex: rs123456) can be uploaded by selecting the "Upload SNP List" option in the left panel.



Selecting the option will provide three forms: one for uploading a Text (.txt) file where each SNP is its own line, one for providing a list of SNPs in

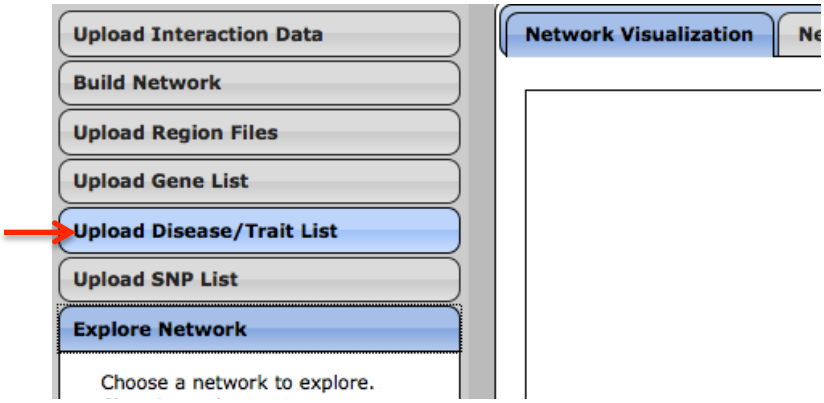
a textbox where each SNP is also its own line, and one for providing a label for the list of SNPs.



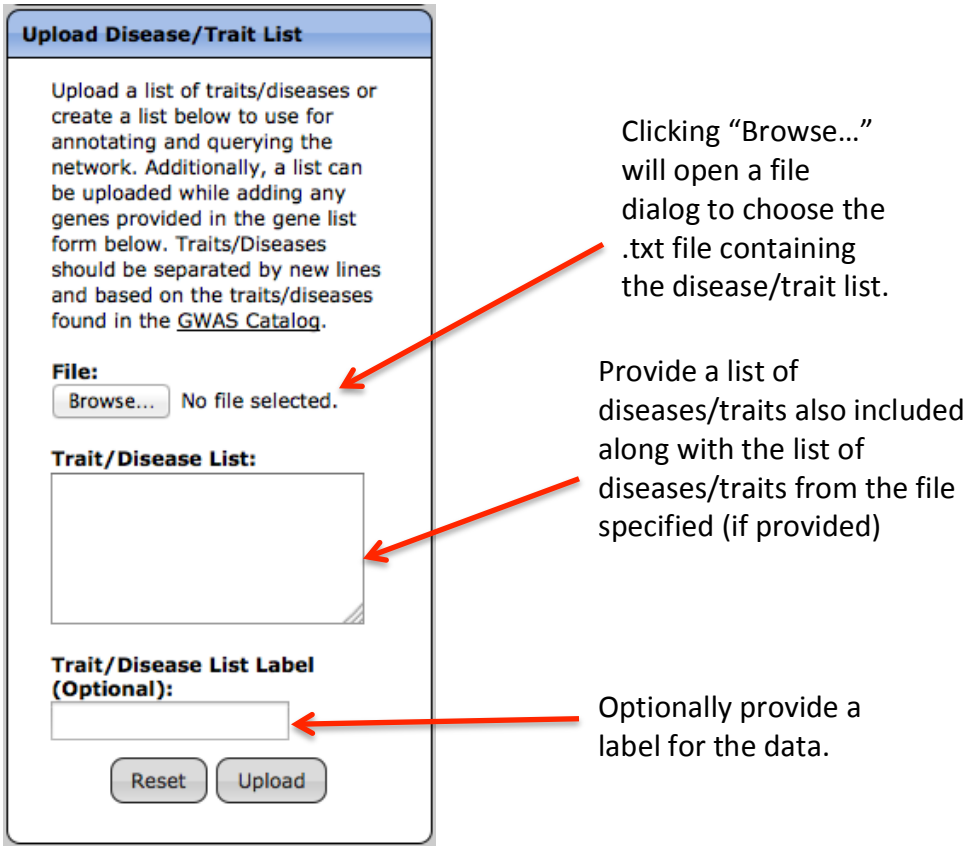
Only one of the options (File or SNP List) needs to be provided to upload a list of genes. If both are provided, then the union of both sources is used as the list of SNPs.

E. Diseases & Traits

QuIN provides the option of uploading a set of diseases or traits, where each disease or trait is derived from the GWAS Catalog, (http://www.genome.gov/gwastudies/) by selecting the "Upload Disease/Trait List" option in the left panel.



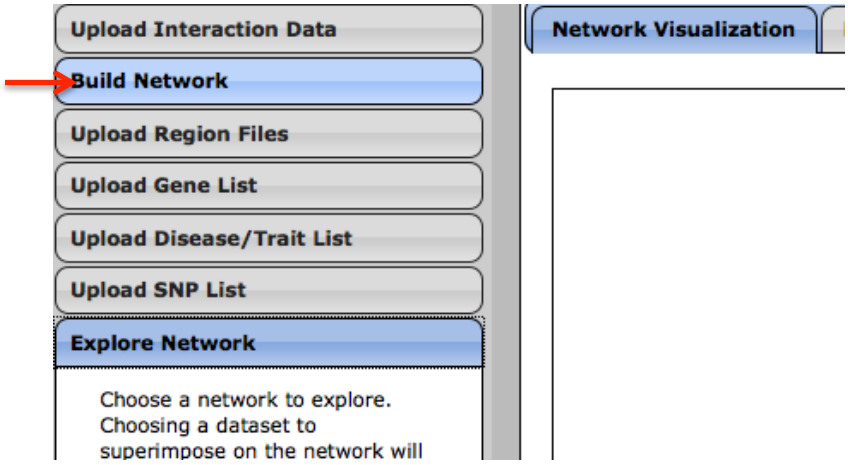
Selecting the option will provide three forms: one for uploading a Text (.txt) file where each disease/trait is its own line, one for providing a list of diseases/traits in a textbox where each disease/trait is also its own line, and one for providing a label for the list of diseases/traits.



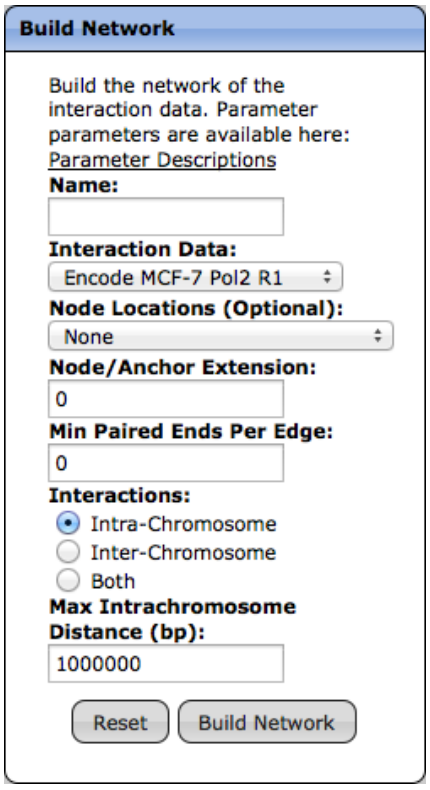
Only one of the options (File or Disease/Trait List) needs to be provided to upload a list of genes. If both are provided, then the union of both sources is used as the list of diseases/traits. Unlike the other similar upload options, the form for providing a disease/trait list will attempt to auto-complete diseases/traits recognized in the database. As diseases and traits are matched exactly, it is recommended to use this feature if typing a list of features. For example, "Type 2 diabetes" is different from "Type 2 diabetes and other traits". This also means that when querying for a particular disease/trait in the network, all variations may need to be provided.

II. Building Networks

Networks can be built from uploaded interaction data by using the "Build Network" option in the left panel.



Selecting the "Build Network" option will display the menu shown below.



The first field asks to provide a name of the network, which will show up in the list of available networks for visualization. The remaining fields are discussed in more detail in the following sections.

A. Interaction Data

"Interaction Data" determines which interaction data to use when building the network. A dropdown menu lists all available uploaded interaction data that can be used in constructing the network.

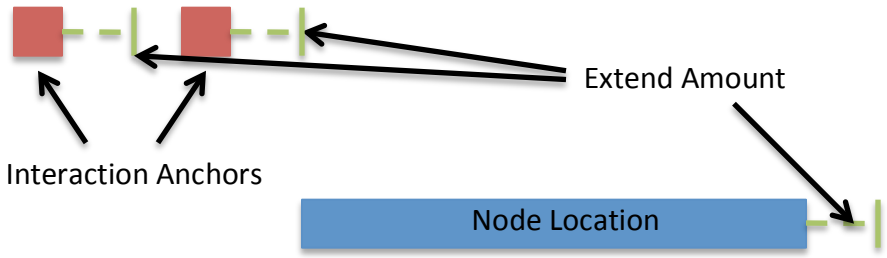
B. Node Locations

The "Node Locations" field is an optional field that allows for using the genomic regions/loci of uploaded non-interaction data as the location for the nodes in the network. Specifying the set of node locations changes the method used for constructing the network such that the node locations are fixed, where as in the case of not providing a set of node locations, the nodes are inferred from the anchors of the interactions. Appendix A describes both of these methods in detail. This feature is useful for defining nodes using open chromatin sites defined by for example DNASE-Seq or ATAC-Seq datasets.

C. Node/Anchor Extension

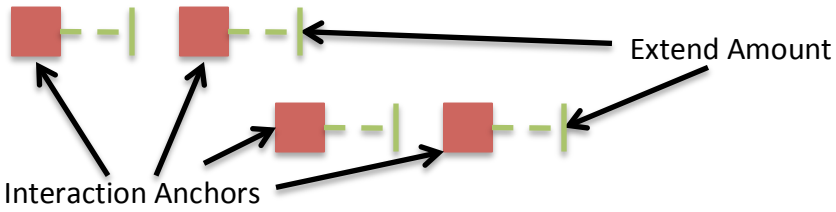
The "Node/Anchor Extension" parameter is used slightly differently depending on the method used. The purpose of this parameter is to provide some flexibility in determining if two regions interact with each other or if two anchors are close enough to be grouped together as one Node.

In the case when node locations are defined, the Extend parameter extends both the anchors of the interaction and the specified node locations at their end points (END+Extend) and uses these extended locations to determine whether they overlap.



The result is the same as extending each node location in both directions by the amount specified.

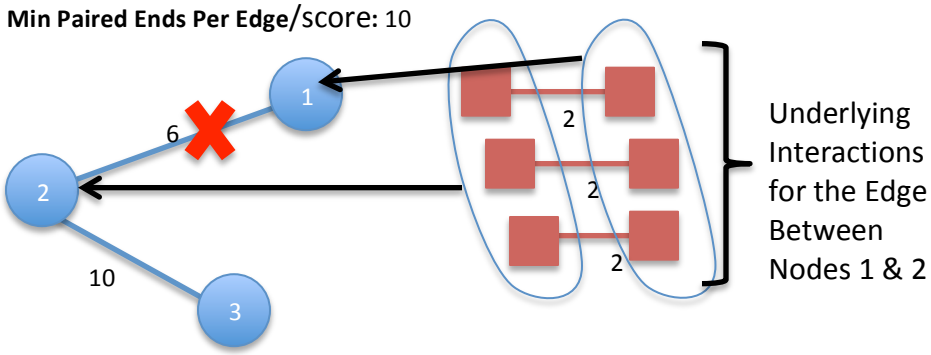
When the nodes are not specified, the extension is used to extend each anchor of an interaction to determine whether the anchors should be in the same node. The method used only extends each anchor at its end position as before, but achieves similar results as extending each anchor by an amount at both ends.



Since anchors are compared to other anchors, extending in one direction is not exactly the same as extending by the same amount in both directions. In the both-directions case, each anchor is only extended by half of the amount. For this reason, the extend parameter provided is doubled in order to achieve the same effect as extending each end of each anchor by the amount specified.

D. Min Paired Ends Per Edge

The "Min Paired Ends Per Edge" field represents the minimum number of PETs/Reads that an edge requires to be present in the network. The PET count for an edge is specified as the summation of the PET Counts/Reads for each interaction represented in the edge. For example, if this parameter is set to 10 and an edge contains three interactions with each interaction having a PET count/Read count of 2, (making the total count of the edge 6), the edge will not be included in the network.



E. Interaction Type

The "Interactions" field allows for selecting between intra-chromosome, inter-chromosome or both types of interactions that will be present in the network. Generally, only the intra-chromosome interactions are used, but if needed, the options to view both or only inter-chromosome interactions are available. Including both types of interactions is not recommended as inter-chromosome interactions tend to connect many

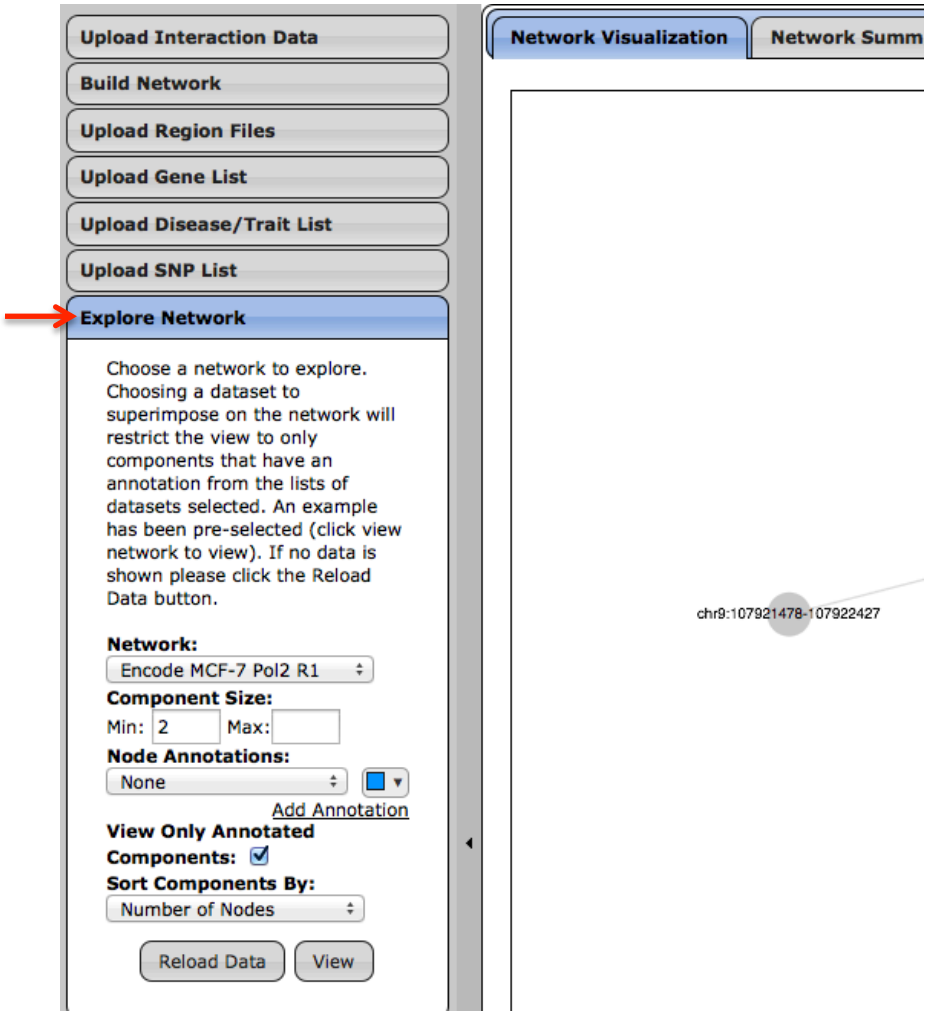
components together, resulting in very large networks which can cause the web browser to stop responding when viewed.

F. Max Intra-Chromosome Distance

As an additional filter for intra-chromosome interactions, the “Max Intra-Chromosome Distance” specifies the maximum distance between two nodes on the same chromosome where edges are removed if the distance between the two nodes that it connects is greater than the amount specified.

III. Network Exploration & Visualization

Once a network is built, it can be visualized by selecting the "Explore Network" option in the left panel.



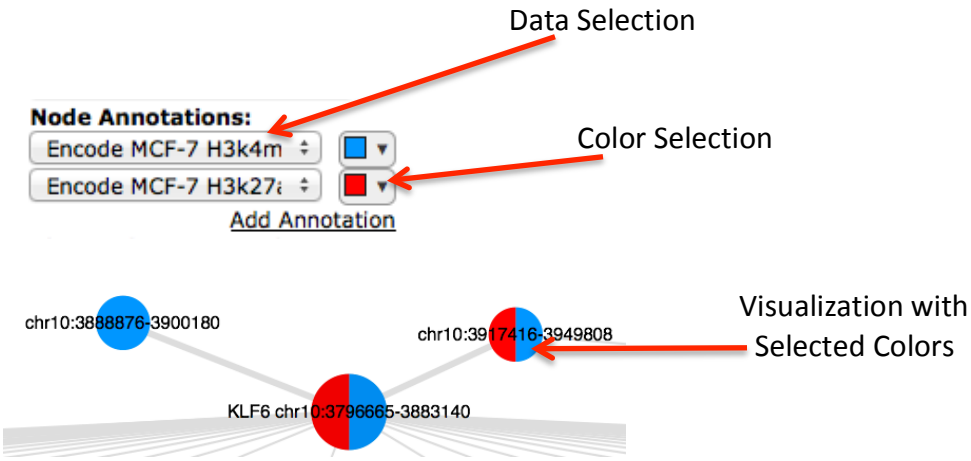
When selected, a menu will appear with fields for selecting the network to visualize, selecting the annotations to super-impose on the network, and selecting options for how to sort the components in the network.

A. Component Size Selection

The component size field allows for selecting the size of the components that will be visualized and used for further analysis of the network where the total number of nodes in a component reflects the size of the component. Setting the minimum and maximum size of the components will result in only components that have at least the specified minimum and at most the specified maximum number of nodes.

B. Adding Annotations

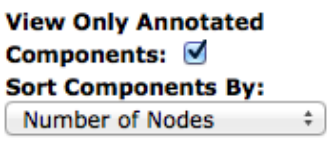
An annotation in the network (optional) can be any dataset uploaded other than the interaction data. To add an annotation, select a dataset from the drop-down menu and the nodes in the network will be colored in the visualization if they overlap with the annotations provided. Additional annotations can be added by selecting the "Add Annotation" link below the data and color selections. If multiple annotations are present for a single node, the node will be colored by a pie chart, showing all annotations present.



The above figure shows an example of specifying two annotations and how these annotations are represented in the visualization of the network. Adding annotations takes time to initially superimpose the data onto the network. Therefore, the more annotations superimposed on the network for the first time, the longer it will take to load the network. Once the network is annotated however, loading the annotations again on the same network will not take as long.

C. Component Sorting

Since QuIN visualizes the network by viewing one component of the network at a time, the final two options for visualization determine how the components are sorted, and which ones to display.



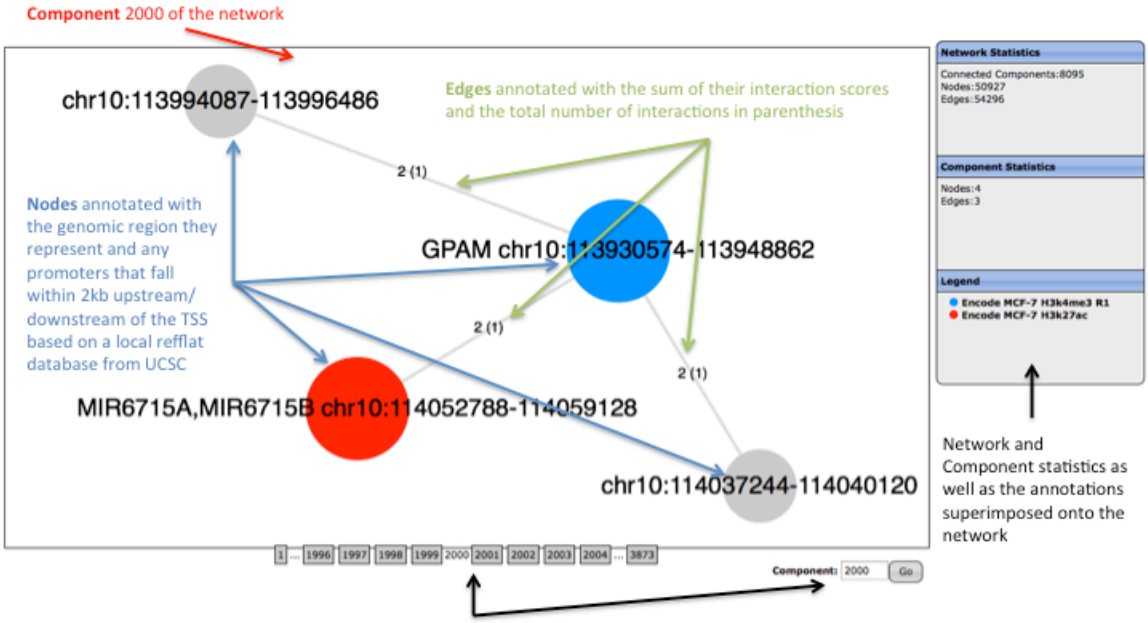
The first option is a checkbox to determine whether or not to display components that do not have any annotations. Checking this option will remove all components without at least one of the specified annotations

(if provided). The second option sorts the components in one of three ways in descending order:

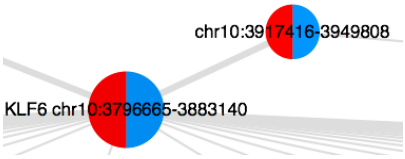
1. Number of Nodes
2. Number of Edges
3. Number of Annotations

D. Visualization & Interpretation

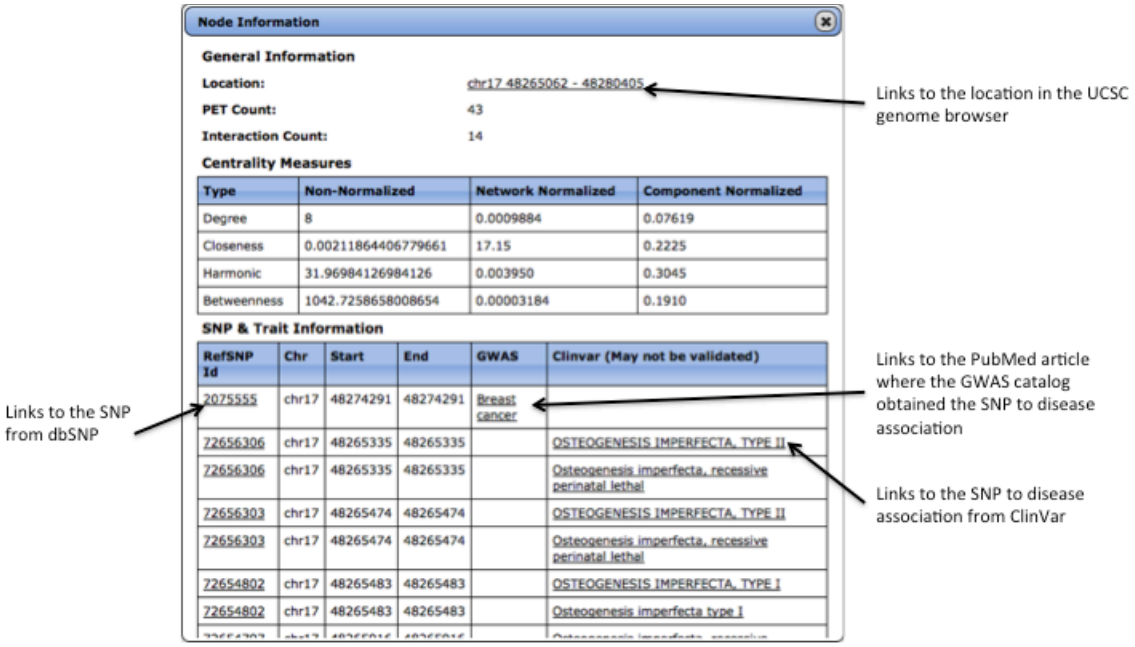
The visualization displays a single component of the network as shown below.



An interaction network is composed of several connected components where a connected component is a subnetwork in which for any two nodes, they are connected by some path in the subnetwork. In addition to displaying the nodes and edges of each connected component, nodes are also annotated with genes if they fall within 2kb upstream or downstream of a TSS. Nodes with a promoter/gene annotation are represented with a slightly larger size than other nodes. Nodes with different colors refer to the specific annotations shown in the legend on the right. If a node has multiple annotations then all annotations are shown using a pie chart of all of the colors where each portion of the pie chart is the reciprocal of the total number of annotations. An example is shown in the next figure:

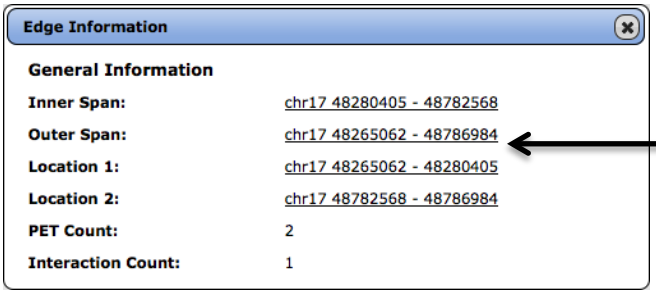


"Right Clicking" or "Double Tapping" a node or edge will reveal more information about that particular node or edge. For nodes, various information including centrality measures, SNPs and diseases associated with the genomic region is provided.



The centrality measures include degree (connectivity degree) which is a measure of how many connections the node has, closeness which measures how connected the node is to all of the other nodes in the component, harmonic which is similar to closeness but takes into consideration the size of the component, and betweenness which measures how often a node is on the shortest path between every other pair of nodes. For each centrality measure, a raw score, a score normalized with respect to the entire network and a score normalized to the size of the component the node is in is calculated and depicted.

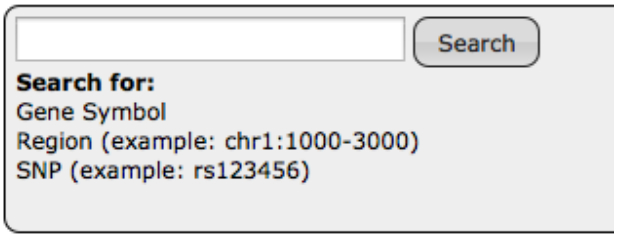
More information on specific edges in the network provides links to the UCSC genome browser for various regions of interest associated with the edge for example, the node locations it is connected to, the inner span (the region not including the node regions), and the outer span (the region including the node region). Inner and outer span links are only provided for intra-chromosome interactions.



Links to the UCSC genome browser

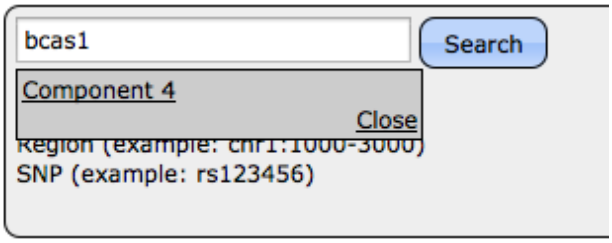
E. Searching

Just below the visualization in the Network tab, the search feature is available:

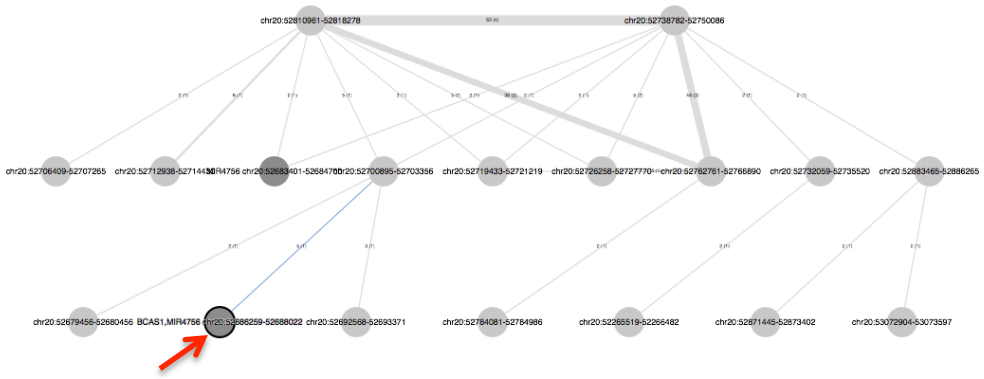


Using this feature, one can query the components to locate genes (using their official gene symbol), regions of interest specified by <chr>:<start>-<end> (ex: chr1:1000-3000), or SNPs based on their RefSNP id (ex: rs123456).

Searching for one of the elements will provide a list of components with the search term present:



Selecting the component will navigate the visualization to the specified component, highlighting the node that contains the term.

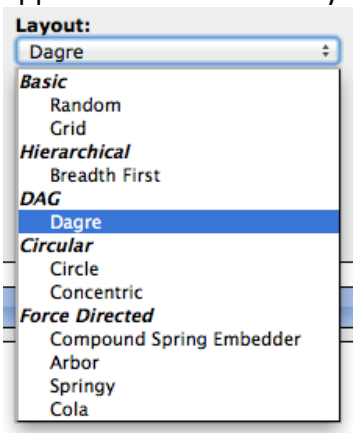


Node With The Search Term BCAS1

F. Additional Features

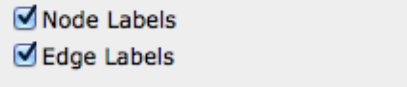
Found below the network are features relating to the network labels, layout and position.

Under “Layout”, there are a variety of different layouts that can be applied to the network by selecting the dropdown menu.

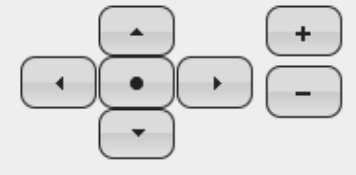


Descriptions of each layout are omitted here but more information can be obtained from Cytoscape JS (http://js.cytoscape.org). Caution is placed on using the layouts under the “Force Directed” category as these layouts are computationally intensive and may freeze the browser for some time if used on large components. How large the components can be before seeing this issue varies based on the computer and browser used to visualize the network.

Beneath the layout drop-down menu are two checkboxes that toggle the node and edge labels on the network. Checking or Unchecking these boxes will add or remove the node or edge labels respectively.



The arranged set of buttons on the right provides options for panning and zooming in and out of the network.



Clicking the buttons with arrows will pan the entire network in the direction selected. Selecting "+" or "-" will zoom the network in or out respectively. Finally, the center button will reset the network to be in close proximity if the network is somehow lost when panning or zooming. Resetting the network may not make the network visible at first depending on the layout used, but zooming out after resetting should bring the network back into view.

IV. Network Analysis

A. Network Summary

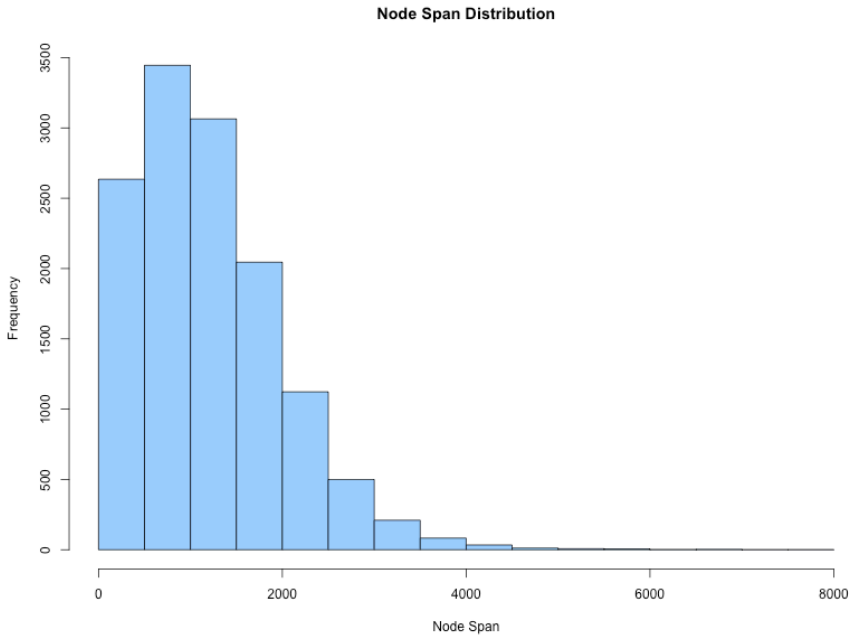
Navigating to the "Network Summary" tab will provide some general information about the network. At the top are two tables, which reveal statistics for the network as well as the parameters used for building and visualizing the network.

MCF7 - 5 Or More	
Number of Components	928
Number of Nodes	13,171
Number of Edges	16,763
Average PET per Edge	6.01

Parameters	
Extend	0
Minimum PET per Edge	0
Interaction Type	Intra-Chromosome
Max Intrachromosome Distance	1000000
Minimum Component Size	5
Maximum Component Size	

To get a sense of the size (i.e., bps spanned by the node) of each node in the network, a histogram is provided. This histogram is useful in assessing the impact of parameters used to construct the network on the

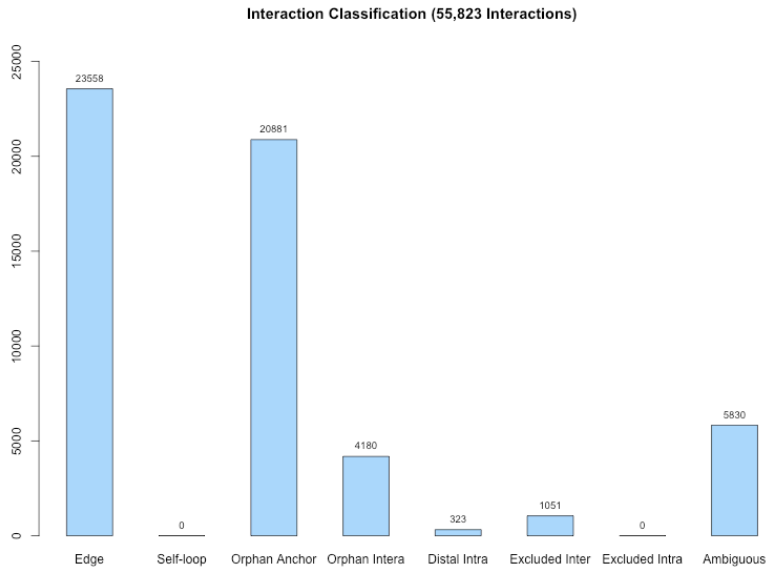
final node definitions and hence the network connectivity. For example setting the extend parameter high such as 2kb, may increase the span of the nodes as much as 20kb, due to the merging of nearby anchors.



Below the node span distribution, a bar plot is generated categorizing the interactions based on their inclusion or exclusion in the network. There are in total eight different categories:

1. **Edge** – If the interaction is included as an edge in the network
2. **Self-loop** – If the interaction's anchors exist in the same node, i.e., self-loop
3. **Orphan Node**– Only one anchor meets pre-defined criteria, such as overlapping with a DNAse-seq peak (Relevant only if nodes are filtered out using a second source of data, such as open chromatin peaks).
4. **Orphan Interaction** – Neither of the interaction's anchors meet a predefined criteria, such as overlapping with a DNAse-seq peak (Relevant only if nodes are filtered out using a second source of data, such as open chromatin peaks).
5. **Distal Intrachrom** – If the interaction is an intra-chromosome interaction and filtered out due to the intra-chromosome distance parameter
6. **Excluded Inter-chromosomal** – If the interaction is inter-chromosomal and is excluded since the network is constructed only from intra-chromosome interactions

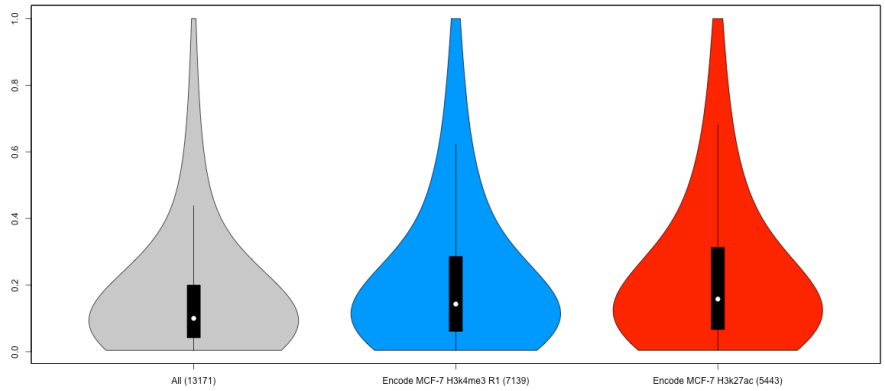
7. **Excluded Intra-chromosomal** – If the interaction is intra-chromosomal and is excluded because the network is constructed only from inter-chromosome interactions.
8. **Ambiguous** – If a genomic region list is provided as the set of nodes in the network, an interaction is considered ambiguous if one of its anchors overlaps with multiple nodes. (Relevant only if nodes are determined using a second source of data, such as open chromatin peaks)



B. Centrality Measures

The centrality measures tab depicts general centrality measures calculated for the network as well as any specific annotations. Within this tab, there is a button to load a series of four violin plots and tables for the four centrality measures: degree (connectivity degree), closeness, harmonic, and betweenness. To briefly describe these measures, degree is a measure of how many edges the node has, closeness measures how connected the node is to all of the other nodes in the component, harmonic is similar to closeness but takes into consideration the size of the component, and betweenness measures how often a node is on the shortest path between every other pair of nodes in the connected component.

Degree:



	Encode MCF-7 H3k4me3 R1	Encode MCF-7 H3k27ac
Encode MCF-7 H3k4me3 R1	1.0000e+0	3.4346e-92
Encode MCF-7 H3k27ac	3.4346e-92	1.0000e+0

The violin plot above shows the component normalized degree of the H3K4me3 and H3k27ac histone marks. Underneath the violin plot is a table containing Mann-Whitney-Wilcoxon p-values to quantify the network measure differences between differently annotated nodes of the network.

C. GO Analysis

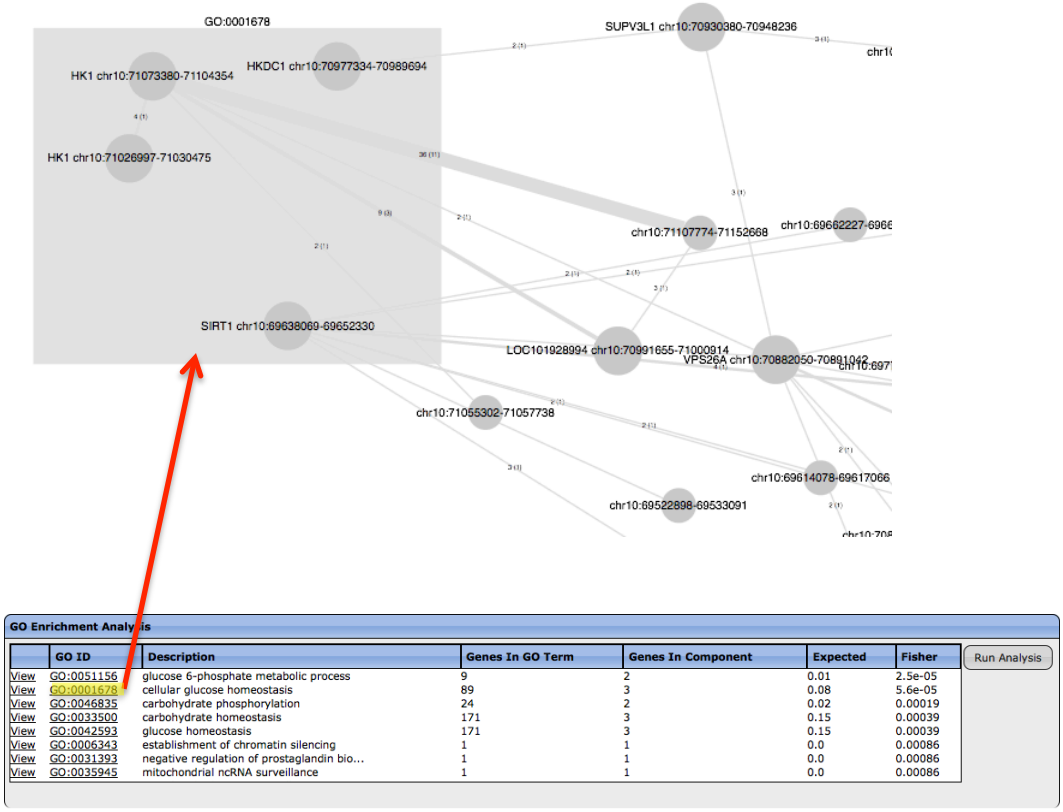
GO analysis via topGO can be performed on the current component by selecting “Run Analysis” under “GO Enrichment Analysis”.

GO Enrichment Analysis

	GO ID	Description	Genes In GO Term	Genes In Component	Expected	Fisher
View	GO:0051156	glucose 6-phosphate metabolic process	9	2	0.01	2.5e-05
View	GO:0001678	cellular glucose homeostasis	89	3	0.08	5.6e-05
View	GO:0046835	carbohydrate phosphorylation	24	2	0.02	0.00019
View	GO:0033500	carbohydrate homeostasis	171	3	0.15	0.00039
View	GO:0042593	glucose homeostasis	171	3	0.15	0.00039
View	GO:0006343	establishment of chromatin silencing	1	1	0.0	0.00086
View	GO:0031393	negative regulation of prostaglandin bio...	1	1	0.0	0.00086
View	GO:0035945	mitochondrial ncRNA surveillance	1	1	0.0	0.00086

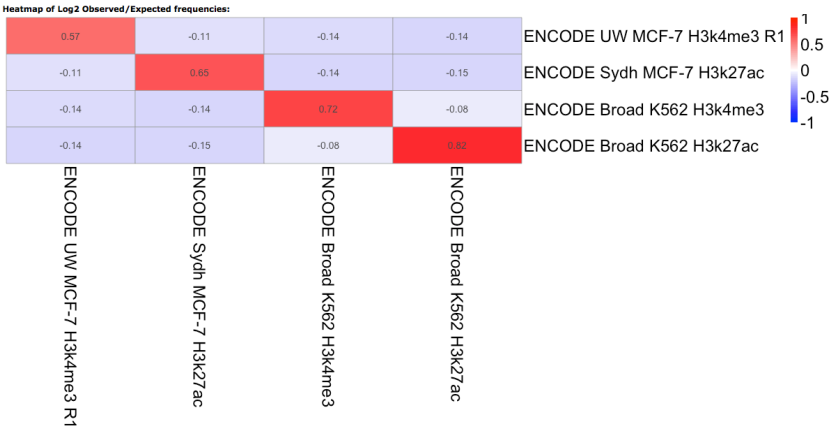
Run Analysis

Running the analysis will list the top 25 GO terms providing the GO ID which links to GO term in Amigo, a brief description of the GO term, the number of genes in the GO term, the number of genes present in the component, the expected representation of the GO term, and the P-Value from Fisher’s exact test. Selecting “View” will update the network visualizing and group together all of the genes for the selected GO term.



D. Annotation Interaction Enrichment

If multiple annotations are superimposed onto the network, the Annotation Interaction Enrichment tab provides an analysis for understanding if two annotations are interacting more or less frequently than expected. Clicking "Run Annotation Interaction Enrichment" will provide a heatmap and three tables.



The heatmap visualizes the log base 2 observed/expected frequencies where the expected frequencies are calculated as:

Different Annotations (a & b): $\mathbf{E}(\mathbf{N}_a/\mathbf{N})(\mathbf{N}_b/(\mathbf{N}-1))+ (\mathbf{N}_a/\mathbf{N})(\mathbf{N}_b/(\mathbf{N}-1))$
Same Annotations (a): $\mathbf{E}(\mathbf{N}_a/\mathbf{N})(\mathbf{N}_a-1/(\mathbf{N}-1))$

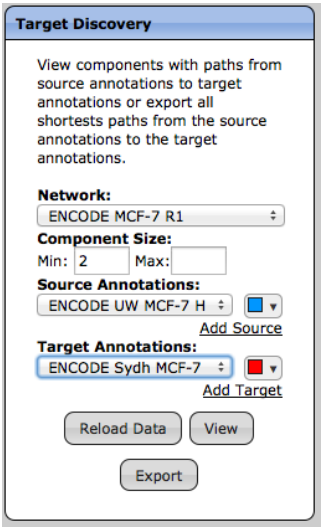
Where **E** represents the total number of edges, **N** represents the total number of nodes, and $\mathbf{N}_a$ & $\mathbf{N}_b$ represent the total number of nodes with the respective annotations.

The first table below the heatmap provides P-Values based on the binomial test. P-Values are obtained from both less than and greater than hypothesis and the more significant P-Value is kept. Positive P-Values are from the greater than hypothesis where as negative P-Values are from the less than hypothesis, providing directionality.

The remaining tables provide the observed count of edges with the annotations and the expected number of edges calculated using the formula discussed above.

V. Target Discovery

The target discovery in the left menu provides users the ability to visualize and export paths from a set if source annotations to a set of target annotations.



Selecting a network and source and target annotations and selecting view will select all components which have at least one source and at least one target annotation within them to quickly visualize and query among all components which have such interactions.

Selecting “Export” will prompt the user to download a zip file which contains two files: A file of node ids and genomic regions and a 25 column tab delimited

shortest path file which provides data for all of the shortest paths from the source annotations to the target annotations.

The 25 columns contain the following information:

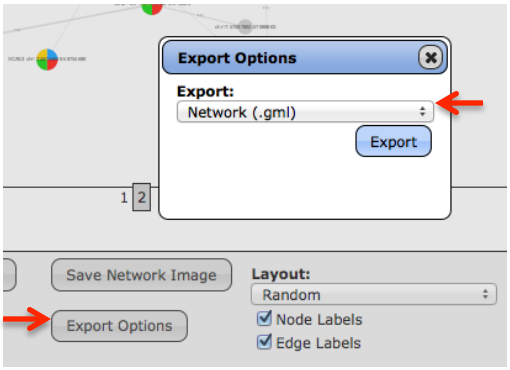
1. **Source Dataset** – The dataset of the source annotation.
2. **Source Term** – The term of the annotation, if for example the annotation is a gene list this will be the gene provided in the list.
3. **Source Chr** – The chromosome of the source annotation.
4. **Source Start** – The start position of the source annotation.
5. **Source End** – The end position of the source annotation.
6. **Source Nearest TSS** – The Official Gene Symbol of the nearest TSS to the source node's location on the genome.
7. **Source Nearest TSS Distance** – The distance in bp to the TSS from the source node.
8. **Hop Count** – The minimum number of hops/edges required to get to the target node from the source node.
9. **Distance** – The genomic distance between the annotated node and the target node.
10. **Target Dataset** – The dataset of the target annotation.
11. **Target Term** – The term of the annotation, if for example the annotation is a gene list this will be the gene provided in the list.
12. **Target Chr** – The chromosome of the target annotation.
13. **Target Start** – The start position of the target annotation.
14. **Target End** – The end position of the target annotation.
15. **Target Nearest TSS** – The Official Gene Symbol of the nearest TSS to the target node's location on the genome.
16. **Target Nearest TSS Distance** – The distance in bp to the TSS from the target node.
17. **AVG PET/Read Count** – The average PET/Read Count of the interactions on the shortest path from the source to the target.
18. **Min PET/Read Count** – The minimum PET/Read Count of the interactions from the annotation to the target.
19. **Max PET/Read Count** – The maximum PET/Read Count of the interactions from the annotation to the target.
20. **AVG Interactions** – The average number of interactions for each edge on the path from the annotation to the target. Note: Edges can be representing multiple interactions.
21. **Min Interactions** – The minimum interactions for each edge on the path from the annotation to the target
22. **Max Interactions** – The maximum interactions for each edge on the path from the annotation to the target.
23. **Total Nodes In Component** – The total number of nodes in the component of the annotation/target nodes.

24. **Total Edges In Component** – The total number of edges in the component of the annotation/target nodes.
25. **Path** – A pipe delimited string where each number between the | represent a node id. This represent the node's traversed in order to go form the annotated starting node to the target node.

VI. Network Export

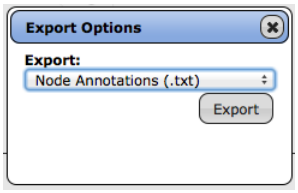
A. Network Export

QuIN currently exports the network in the Graph Modeling Language (.gml) format. To export, select "Export Options" under the visualization, select Network in the dropdown menu, and click the export button.



B. Node Annotations

To export nodes and annotations superimposed on them, select the "Node Annotations" option from the select menu.



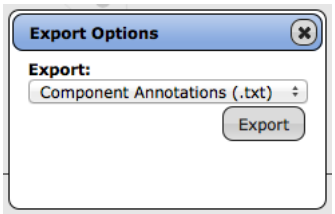
A tab-delimited file will be exported providing the following data:

1. **Node Id** – The Id of the node
2. **Chr** – The chromosome of the node's position
3. **Start** – The start position of the node
4. **End** – The end position of the node
5. **Degree** – The connectivity degree of the node
6. **Closeness** – The closeness centrality of the node
7. **Harmonic** – The haromic centrality of the node
8. **Betweenness** – The betweenness centrality of the node
9. **Normalized Degree** – Degree Normalized with respect to the component
10. **Normalized Closeness** – Closeness centrality normalized with respect to the component
11. **Normalized Harmonic** – Harmonic centrality normalized with respect to the component

12. **Normalized Betweeness** – Betweenness centrality normalized with respect to the component
13. The remaining columns provide a count of how many annotations corresponding with the column label overlap with the node

C. Component Annotations

The component annotations file provides annotation information on every component in the network and can be obtained by selecting the "Component Annotations" option.



Selecting this option exports a (4+number of annotations)-column tab delimited file where each column contains the following:

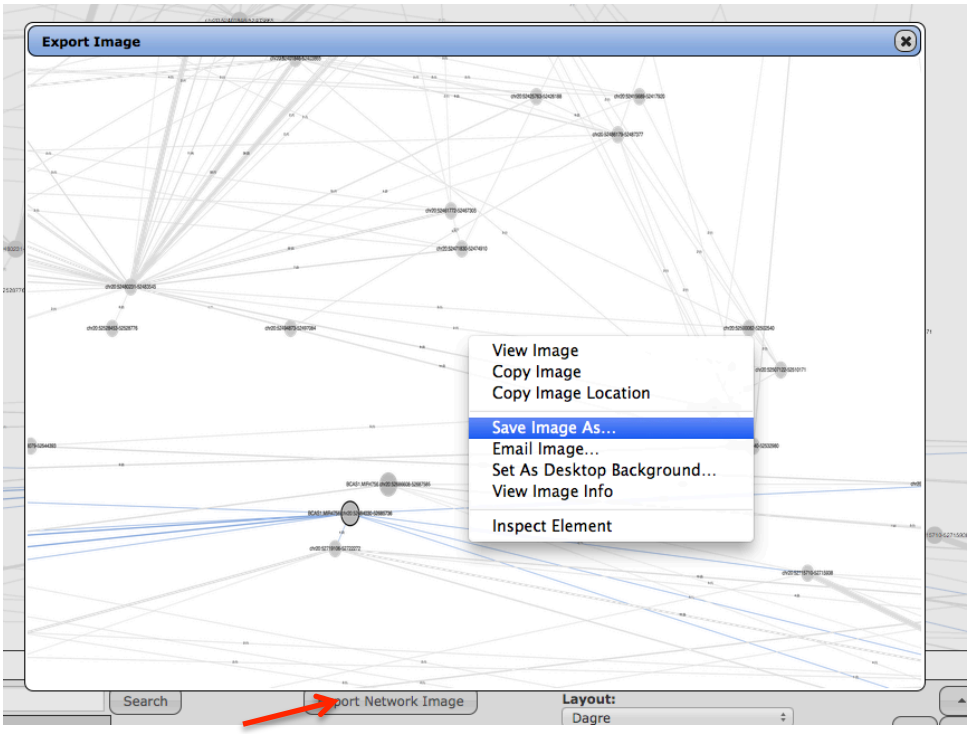
1. Component – The id of the component in the network.
2. Nodes in Component – The total number of nodes in the component.
3. Promoters in Component – The total number of promoters in the component based on 2kb upstream and downstream of the TSS from genes in the refflat UCSC database.
4. The number of columns after this point is dependent on the number of annotations provided for the network. For each annotation, each column represents the number of nodes with that annotation in the network.
5. The last column represents the comma delimited gene symbols corresponding the promoter nodes in the component.

	A	B	C	D	E	F	G
1	Component	Nodes in Cor	Promoters in	Encode MCF	Encode MCF	Promoter Genes	
2	2	52	25	31	34	LOC100130417,SAMI	
3	4	2	1	0	1	LOC100130417	
4	7	2	1	1	1	SDF4,B3GALT6	
5	12	4	3	4	3	SSU72,C1orf233,MIB	
6	14	5	3	3	3	NADK,TMEM52	
7	18	2	0	0	0	null	
8	19	5	2	5	5	C1orf86,SKI	
9	22	2	0	0	0	null	
10	24	6	2	3	3	PEX10,PANK4	

D. Image Export

To save a quick snapshot of the current view of the network, choose the "Save Network Image" option below the current visualization. Once

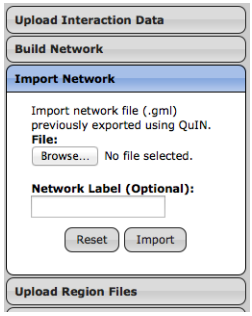
selected, a dialog box will pop up with an image of the current network view. Right click this image and select “Save Image As” to save it.



For resolution purposes after saving, the actual image size is increased and will appear much larger after saving.

VII. Import Network

Networks constructed in QuIN and exported using the export tool can be imported back into QuIN by navigating to the "Import Network" menu in the left panel:



Select the GML file and click the Import button to import a network saved previously. The GML file provided by QuIN encodes the necessary information for this feature to import correctly and therefore other GML files will not be capable of being importing into QuIN in this way. Additionally, annotation information is not imported with this process and will need to be uploaded again if they are needed.

VIII. Appendix A (Methods)

A. Network Construction with Regions from a Secondary Data Source

Networks are created with three main steps in the following order: Node creation, Edge creation and Component creation.

Node Creation:

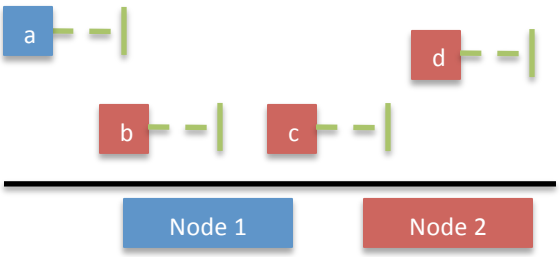
As the nodes are already defined, the node creation step only needs to assign anchors to the nodes for the edge creation step. Assigning the anchors to the nodes begins by separating both the list of nodes and the list of anchors from the interactions by chromosome into different groups. Each group for both nodes and anchors is then sorted by their start position.

For each chromosome, the start position sorted groups of nodes and anchors are iterated simultaneously as follows: Select the first node in the sorted list and iterate through all anchors until the next anchor's start position is greater than the node's end position after extending. While assigning anchors, the position of the next node is also checked to see if the anchor also overlaps with it. If an anchor overlaps two nodes after extension, the algorithm first determines if the anchor overlaps just one of the two nodes without extending where the anchor is assigned to the

node that it overlaps. In the case that they both overlap, an assignment is made based upon if more than 50% of the anchor overlaps with one of them. In the case neither of them overlap the algorithm tries to make an assignment if after extending the node's position, more than 50% of the anchor overlaps with one node while less than 50% overlaps with the other. If no assignment is made, the interaction is considered ambiguous and not used in the network.

Example of Assigning Anchors to Nodes:

1: Node 1 & Anchor a

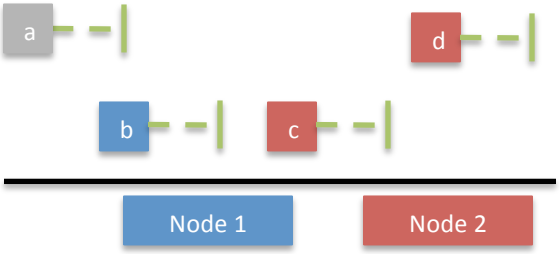


Anchor List

Node 1 Anchors
a

Node 2 Anchors

2: Node 1 & Anchor b

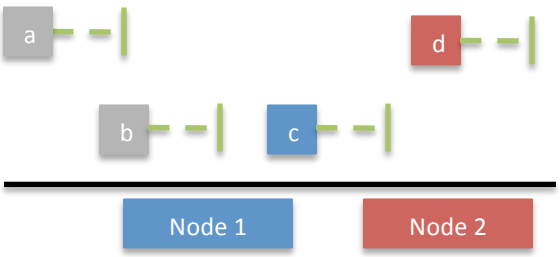


Anchor List

Node 1 Anchors
a,b

Node 2 Anchors

3: Node 1 & Anchor c

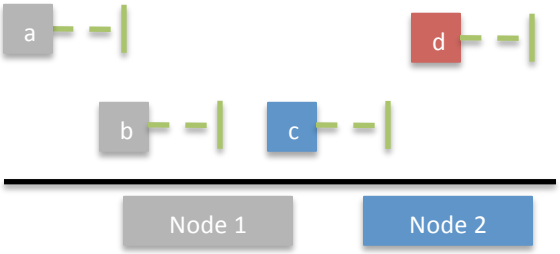


Anchor List
c

Node 1 Anchors
a,b

Node 2 Anchors

4: Transition to Node 2, Check the anchor list between nodes and make final assignment to them.

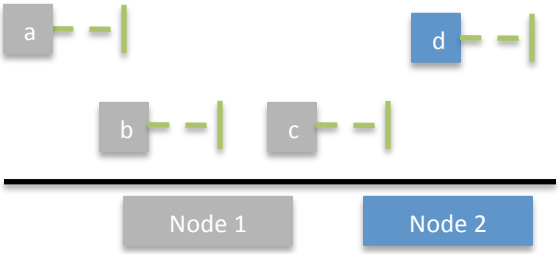


Anchor List

Node 1 Anchors
a,b,c

Node 2 Anchors

5: Node 2 & Anchor d

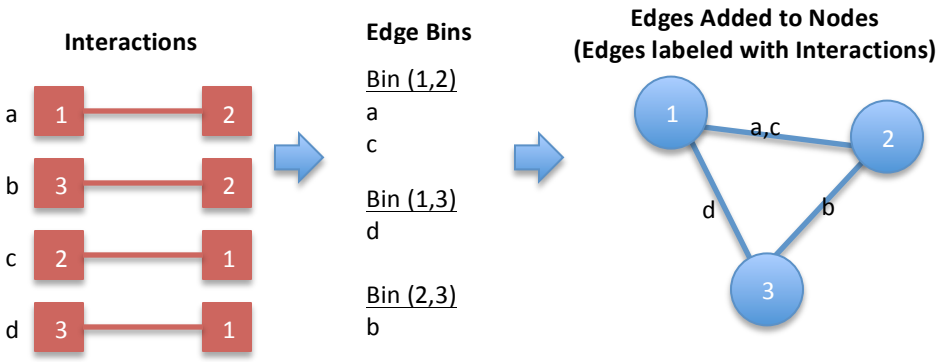


Anchor List

Node 1 Anchors
a,b,c

Node 2 Anchors
d

Edge Creation:
When reading the interaction data, the interactions are created such that they maintain references to their anchors. At the same time, anchors, when added to a node, also maintain references to their respective nodes. These two references make the edge creation step easy as all that needs to be done is to iterate through each interaction and to place them into a sorted set of bins where each bin is defined by the ids of the two nodes the interaction connects which is easily obtained by referencing the two anchors of each interaction and then referencing the node and the node id of the anchors. Once the interactions are binned based on their respective node ids, then edges can be created between the respective nodes while maintain the list of interactions associated with them.



Component Creation:
After creating all of the edges between each node, components are created by preforming breadth first search on every node not yet visited such that all of the nodes and edges reached after performing one breadth first search are part of the same component.

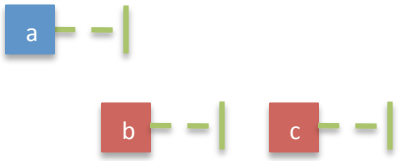
B. Network Construction Only from the Interaction Data

This method differs from the previous method in only the node creation step where the remaining procedures are the same as those when constructing a network with a provided list of regions. Instead of using a predefined set of nodes, nodes are defined from the interaction anchors where the anchors are iterated in increasing order and merged if the next anchors fall within the previous anchor's region after extending by the **extend** parameter after correcting the parameter (doubling the value) to be the equivalent of extending each end of the anchor by the initially provided value.

To describe the procedure in more detail, anchors are first separated into different groups by chromosome and are then sorted by the anchor's start position. Sorting allows the clustering of anchors to be done by simply iterating through each chromosome group once in ascending order. While iterating a group of anchors, a list of anchors is maintained as well as the maximum end point currently seen by the anchors. If the next anchor's start position is less than or equal to the endpoint after extending by the corrected **extend** parameter, the anchor is added to the current list of anchors. If the next anchor's start position is greater, then the current, then a new node is created with the current list of anchors and a new list is started with the next anchor.

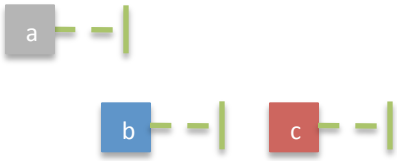
Example: Iterating from anchors **a** to **c**

1: Initialize anchor **a**



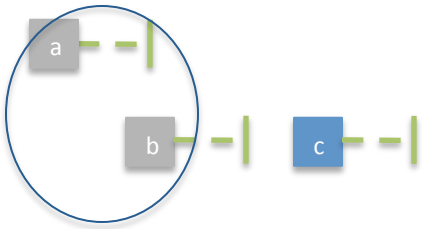
Anchor List	**Node List**
a	

2: Check the next anchor: **b**, where anchor **b** does overlap with the previous endpoint.



Anchor List	**Node List**
a	
b	

3: Check the next anchor: **c**, where anchor **c**, where anchor **c** does not overlap with the previous endpoint.



Anchor List	**Node List**
c	(a, b)

Finally, the node positions are defined as the minimum starting point to the maximum ending point of all the anchors they are composed of.

IX. Appendix B (Case Study)

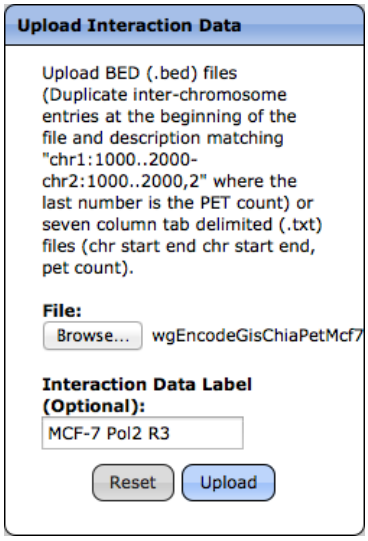
In this example, we will show how to use QuIN to easily identify the targets of SNPs from interaction data.

The list of SNPs used in this example is the list of 71 SNPs and their nearest gene targets used by Rhie et al. (2013)[1] (Table S1).

The interaction data used in this example is the ChIA-PET interaction clusters of MCF-7 Pol2 (Replicate 3), downloaded from ENCODE repository: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGisChiaPet/ (wgEncodeGisChiaPetMcf7Pol2InteractionsRep3.bed.gz)

Step 1

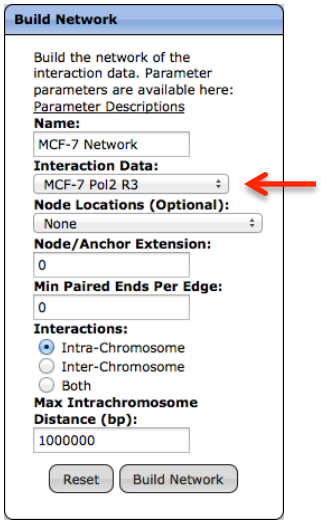
First, after downloading and unzipping the bed file of MCF-7 Pol2 Interactions, the interactions are loaded into the tool under "Upload Interaction Data".



[1] Rhie SK, Coetzee SG, Noushmehr H, et al. Comprehensive Functional Annotation of Seventy-One Breast Cancer Risk Loci. Zhao Z, ed. *PLoS ONE*. 2013;8(5):e63925. doi:10.1371/journal.pone.0063925.

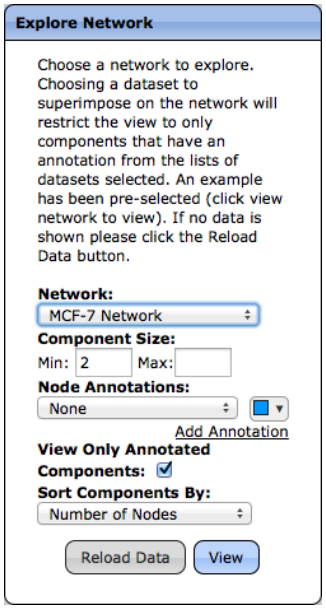
Step 2

Once the interaction data is uploaded, it should now appear in the select menu under “Interaction Data” in the “Build Network” menu.

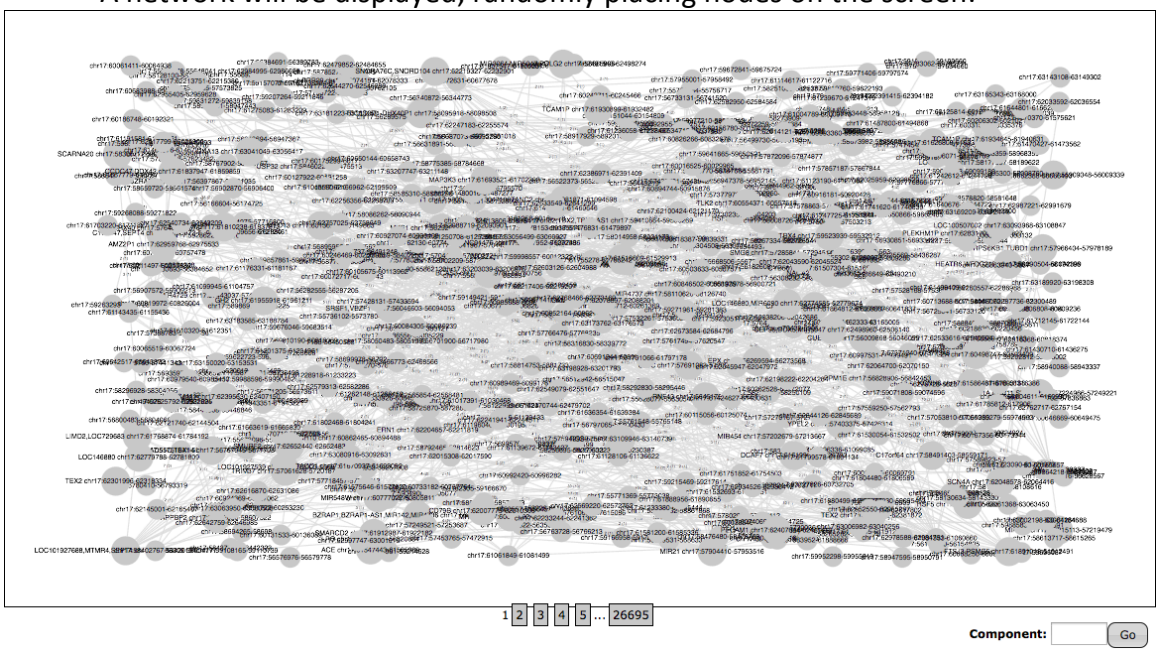


In this example, we build the network only from the ChIA-PET data (extending by 100bp) without providing a list of Node Locations from a secondary data source. To pre-define the nodes in the network, the user needs to upload the list of regions and then select that list from the select menu under “Node Locations”. For this particular example, open chromatin sites for MCF-7 via DNASE-Seq from ENCODE would be an excellent choice for defining the nodes in the final network.

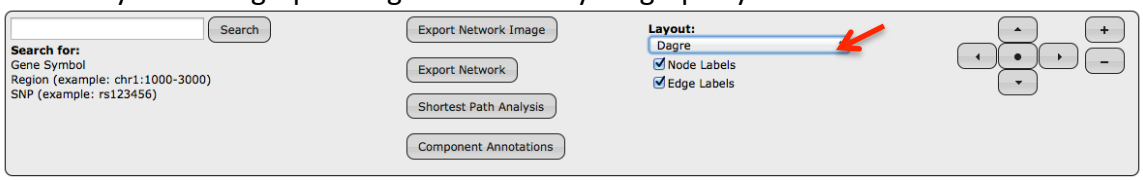
To verify that the network has been successfully built, we simply go to “Explore Network”, select our network and click “View”



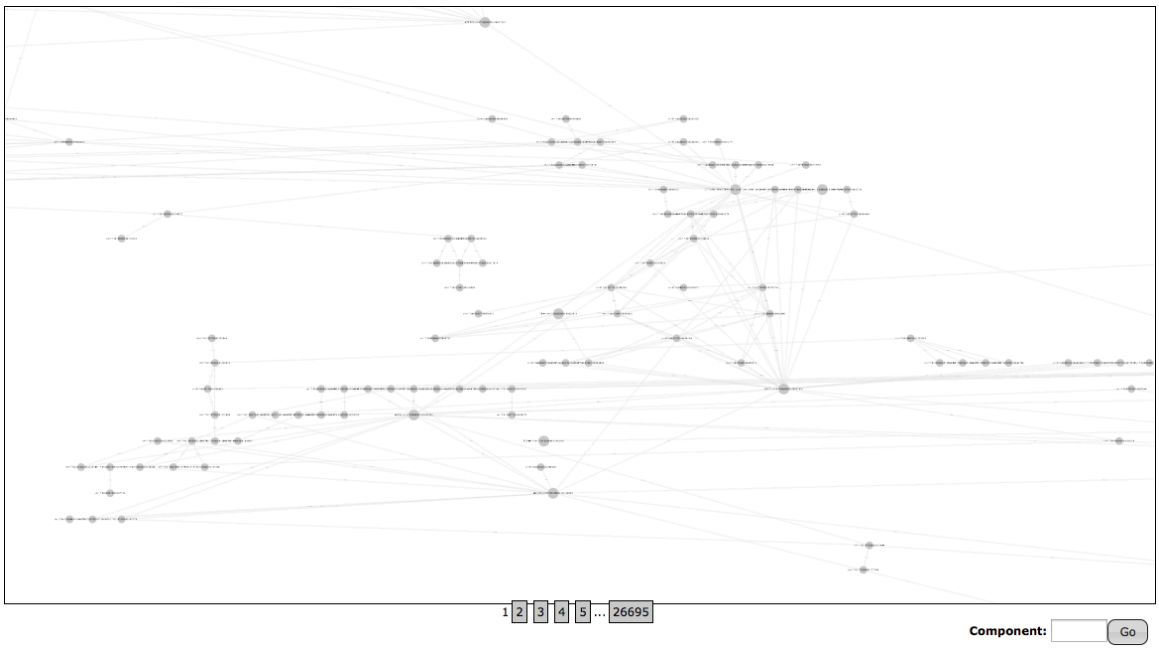
A network will be displayed, randomly placing nodes on the screen:



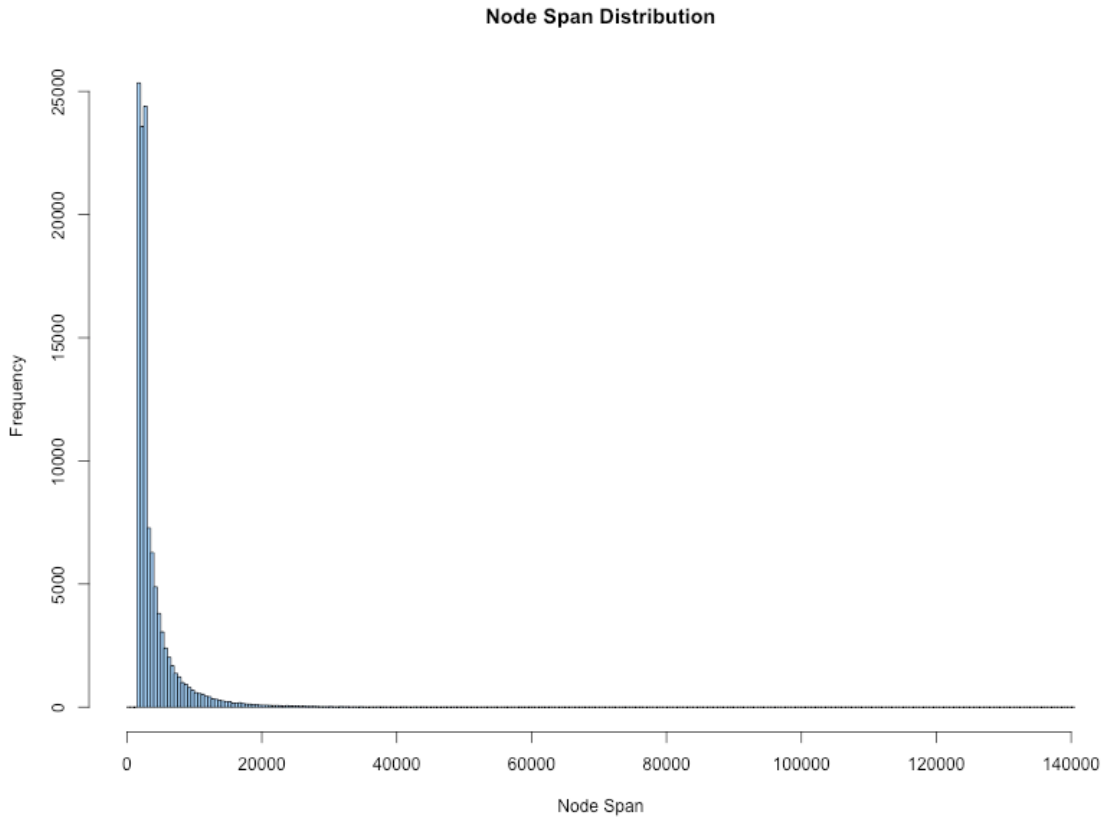
To get a better placement of the nodes, select a different layout under the "Layout" option below the network. In this example, Dagre is selected which lays out the graph using a directed acyclic graph system.



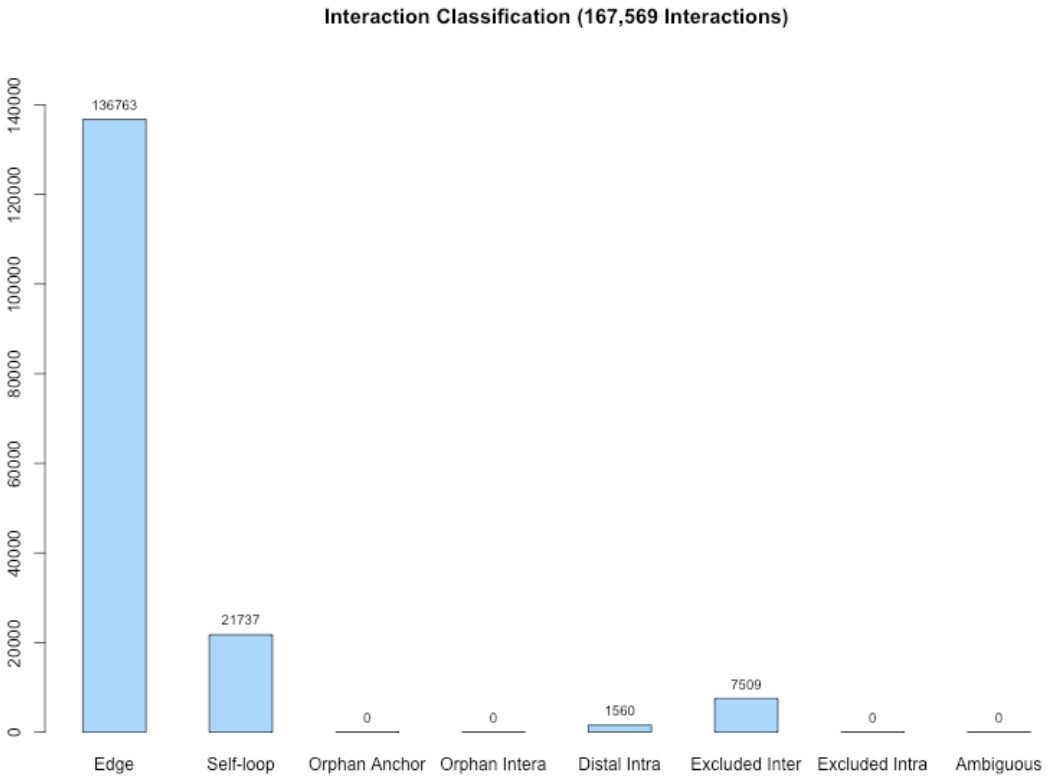
Applying this layout results in the layout shown below:



The “Network Summary” tab provides a histogram of the node span distribution and a bar plot categorizing the interactions. These plots are informative in assessing the properties of the constructed network. If needed, the user can re-generate the network using different parameter settings.



Looking at the node span distribution, the majority of the nodes spans less than 20kb, and reveals a maximum node span of 140k. These node spans are quite large and will lead to merging nearby regulatory regions. Pre-defining the nodes using additional data such as DNASE-Seq, is useful in such cases, and will limit node spans to reasonable values (i.e., between 2kb-4kb).

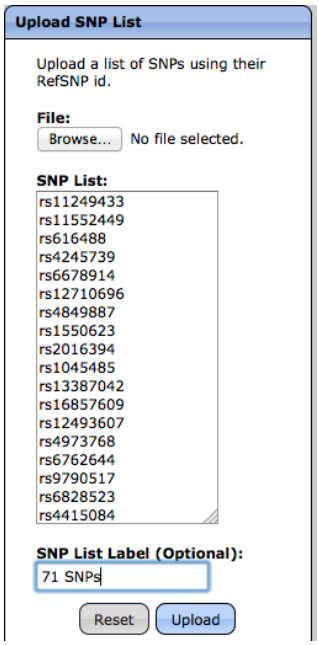


The interaction classification bar plot shows more information on types of interactions in the input data and whether they are included in the final network. This above plot reveals that the majority of interactions are represented as edges in the network, 21,737 of these interactions have anchors within the same node, 1,560 interactions were filtered out because the distance between the nodes is greater than 1mb (defined by the default parameter for "Max Intra-chromosome distance), and 7,509 interactions were filtered out due to being inter-chromosome interactions.

Step 3

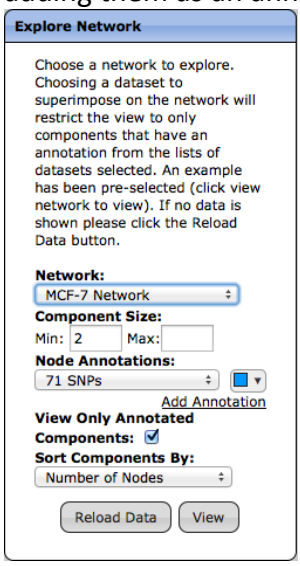
Moving on to our SNP analysis, we simply need to upload the list of 71 SNPs used by Rhie et al. (2013). To do this we first navigate to the Upload SNPs menu in the left panel where we have two options for uploading SNPs. In this example, we

can simply copy the list of SNPs from their table and paste them into the “SNP List” form provided.

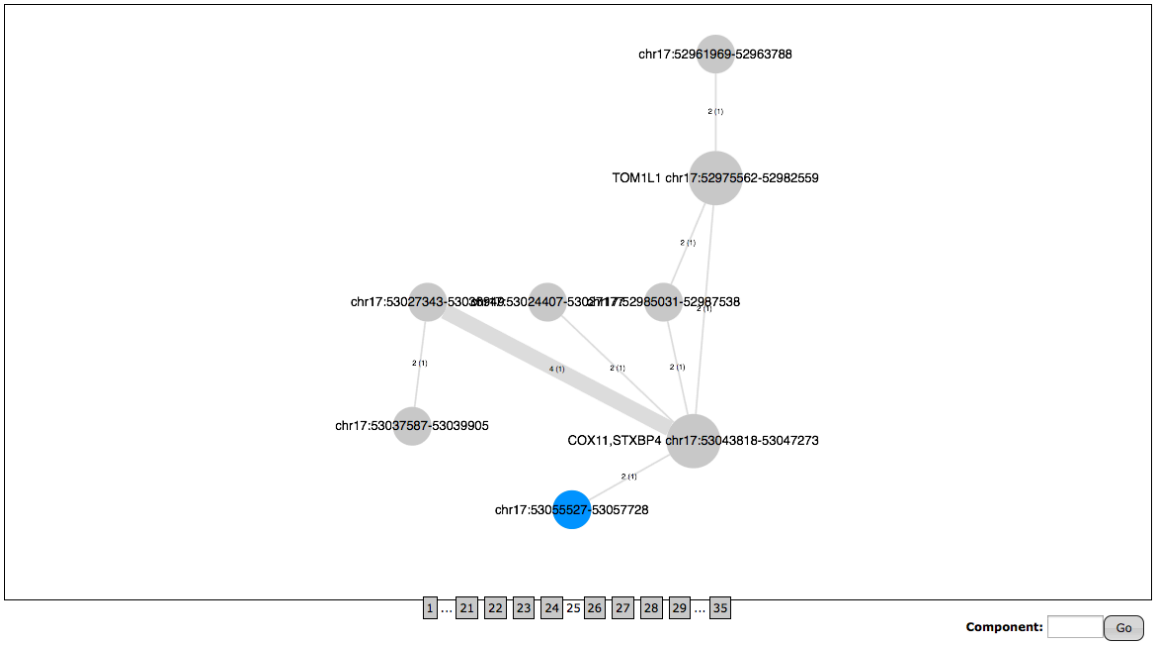


Step 4:

Having uploaded the SNPs, they can then be superimposed onto the network by adding them as an annotation in the “Explore Network” menu.



Viewing the network will now display the network where nodes containing the SNPs uploaded are colored in blue. Since we have “View Only Annotated Components” checked, only the components that have a node with one of the SNPs of interest are shown.



Here we have 35 components that contain the 71 SNPs. In the above screenshot, a SNP is shown to be interacting with the promoters of COX11/STXBP4. Right clicking or double tapping the node containing the SNP reveals more information about the node as well as all of the SNPs defined in dbSNP within the region.

Node Information

General Information

Location: chr17 53055527 - 53057728

PET Count: 2

Interaction Count: 1

Centrality Measures

Type	Non-Normalized	Network Normalized	Component Normalized
Degree	1	0.00003746	0.1429
Closeness	0.06666666666666667	1780	0.4667
Harmonic	3.666666666666667	0.0001374	0.5238
Betweenness	0	0.000	0.000

SNP & Trait Information

RefSNP Id	Chr	Start	End	GWAS	Clinvar (May not be validated)
6504950	chr17	53056471	53056471	Breast cancer	
182178005	chr17	53055658	53055658		
145086249	chr17	53055665	53055665		
143484109	chr17	53055703	53055703		
373402420	chr17	53055897	53055897		
185525318	chr17	53055929	53055929		
35087154	chr17	53056019	53056019		
180870738	chr17	53056318	53056318		

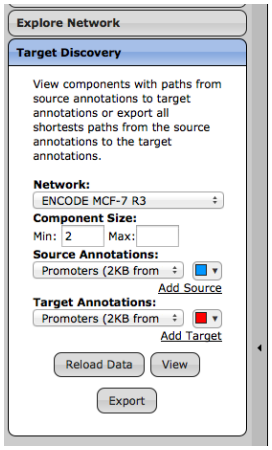
Here the SNP rs6504950 is the SNP from the list, and it is also associated with Breast Cancer from the GWAS catalog, which happens to be the source used by

Rhie et al. (2013) for obtaining their list of 71 SNPs. It is also worth pointing out that the nearest gene provided by Rhie et al. (2013) for this SNP is STXBP4, which is the promoter that the node with this SNP is interacting with.

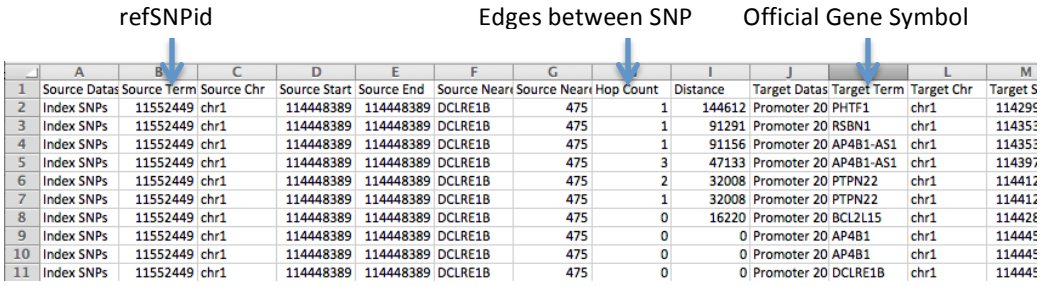
To summarize, in a quick few steps, we were able to upload and construct a network from MCF7 Pol2 ChIA-PET interaction data, visualize the resulting network while understanding some properties of the network, upload a list of SNPs associated with Breast Cancer and annotate the network with this SNP list, and identify regulatory targets of these disease-associated SNPs in a breast cancer cell line.

Target Discovery:

The target discovery analysis can be used to list the genes that are directly or indirectly (via multiple edges) interacting with the 71 SNPs. To perform this analysis, first select the "Target Discovery" option in the left menu.



From here, select the network and choose the index SNPs as the source annotation and Promoters (2KB from TSS) as the target and select export. This will download a zip file containing two files where the shortestpathanalysis.txt file will provide the information necessary to get the SNP to gene interactions. In this file, the "Source Term" column contains the SNP id, the "Hop Count" contains the number of hops or edges between the SNP and the promoter, and the "Target Term"column contains the official gene symbol.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Source Datas	Source Term	Source Chr	Source Start	Source End	Source Near	Source Near	Hop Count	Distance	Target Datas	Target Term	Target Chr	Target S
2	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	1	144612	Promoter 20	PHTF1	chr1	114295
3	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	1	91291	Promoter 20	RSBN1	chr1	114353
4	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	1	91156	Promoter 20	AP4B1-AS1	chr1	114353
5	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	3	47133	Promoter 20	AP4B1-AS1	chr1	114397
6	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	2	32008	Promoter 20	PTPN22	chr1	114412
7	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	1	32008	Promoter 20	PTPN22	chr1	114412
8	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	0	16220	Promoter 20	BCL2L15	chr1	114428
9	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	0	0	Promoter 20	AP4B1	chr1	114445
10	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	0	0	Promoter 20	AP4B1	chr1	114445
11	Index SNPs	11552449	chr1	114448389	114448389	DCLRE1B	475	0	0	Promoter 20	DCLRE1B	chr1	114445

The rows of interest in this data are the rows where the “hop count” is 0 or 1, meaning the SNP is on the same node or one edge away from the node overlapping a promoter. Below is a table of SNPs to genes.

SNP	Gene
rs132390	EMID1
rs204247	NOL7
rs204247	RANBP9
rs204247	MCUR1
rs614367	MYEOV
rs614367	LINC01488
rs614367	CCND1
rs616488	DFFA
rs616488	PEX14
rs704010	ZMIZ1-AS1
rs704010	ZMIZ1
rs865686	KLF4
rs889312	MAP3K1
rs941764	CCDC88C
rs999737	ZFP36L1
rs1353747	PDE4D
rs2016394	SLC25A12
rs2016394	DLX2
rs2016394	DLX2-AS1
rs2046210	ZBTB2
rs2046210	RMND1
rs2046210	ARMT1
rs2046210	CCDC170
rs2046210	ESR1
rs2236007	PAX9
rs2236007	SLC25A21-AS1
rs2236007	SLC25A21
rs2363956	ANKLE1
rs2380205	GDI2
rs2588809	ZFP36L1
rs2981582	FGFR2

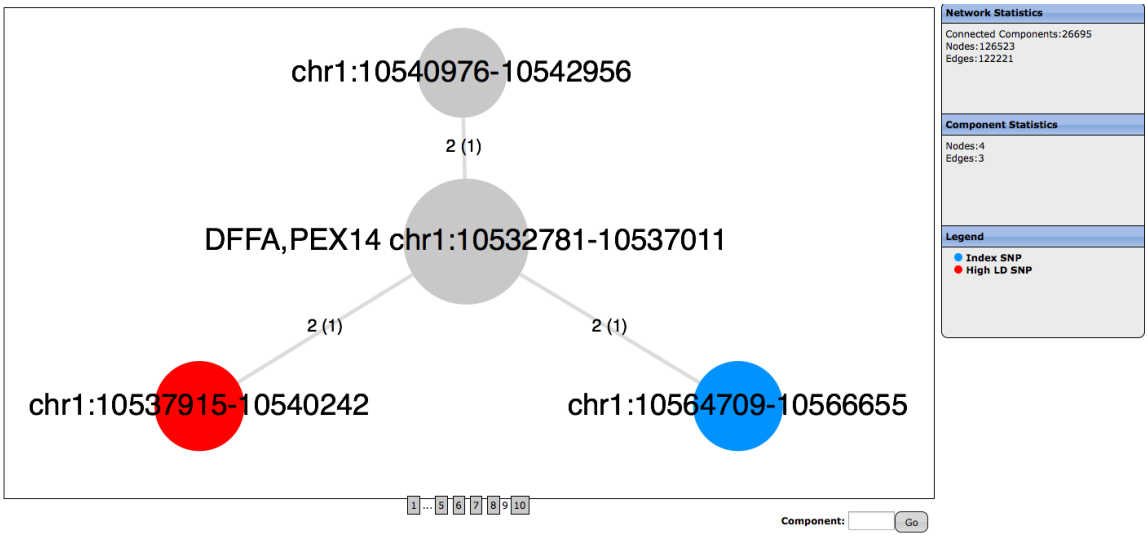
rs3760982	LYPD3
rs3760982	SMG9
rs3760982	KCNN4
rs3760982	LYPD5
rs3903072	KAT5
rs3903072	RNASEH2C
rs3903072	AP5B1
rs3903072	MIR1234
rs3903072	OVOL1
rs3903072	OVOL1-AS1
rs3903072	CFL1
rs3903072	MUS81
rs3903072	EFEMP2
rs4245739	MDM4
rs4849887	INHBB
rs6504950	COX11
rs6504950	STXBP4
rs7072776	MIR1915
rs7072776	CASC10
rs7072776	SKIDA1
rs11552449	PHTF1
rs11552449	RSBN1
rs11552449	AP4B1-AS1
rs11552449	PTPN22
rs11552449	BCL2L15
rs11552449	AP4B1
rs11552449	DCLRE1B
rs11552449	HIPK1-AS1
rs11552449	HIPK1
rs11552449	TRIM33
rs11780156	CASC11
rs11780156	MYC
rs11780156	PVT1
rs11780156	MIR1204
rs11780156	MIR1208
rs12422552	ATF7IP
rs13329835	DYNLRB2
rs13329835	CDYL2

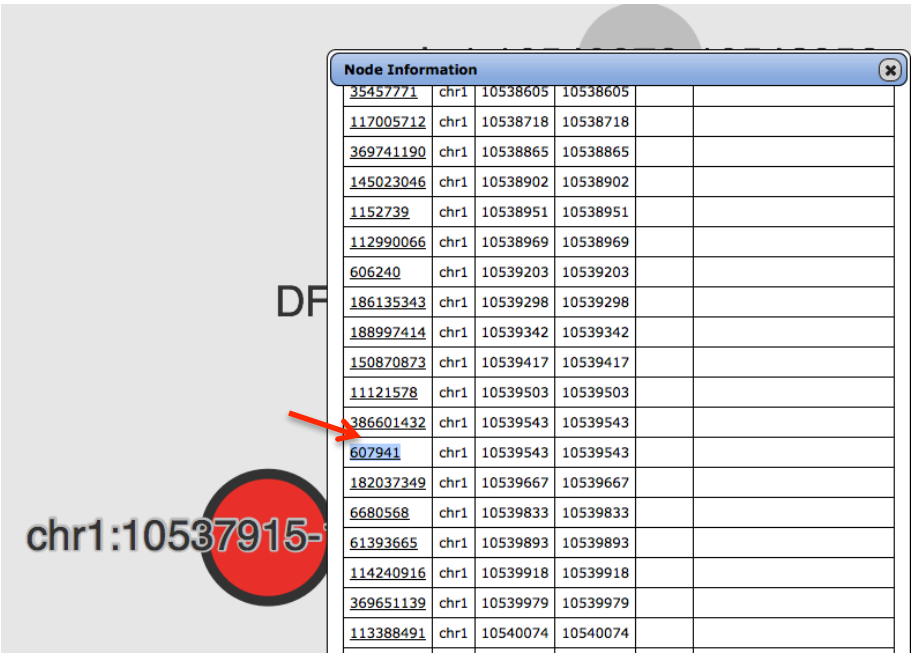
After performing this analysis, we see that 70 SNPs have interactions with genes specified showing that with the help of chromatin interaction networks, we can

identify true regulatory targets of disease-causing SNPs, which is informative in understanding the regulatory mechanisms disrupted by these sequence variants.

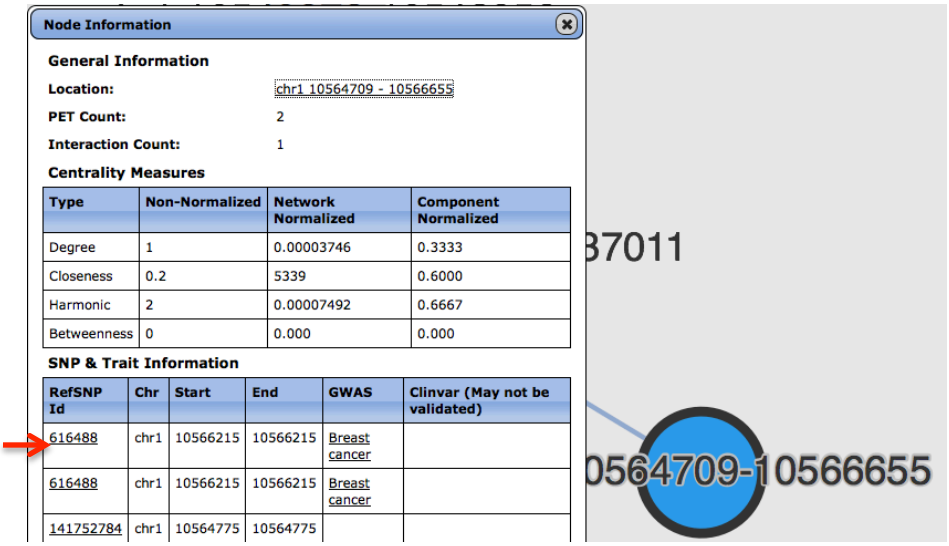
Multiple Annotations using SNPs in Linkage disequilibrium (LD buddies)
Rhie et al. (2013) also provide a list of SNPs in LD (LD buddies). Uploading each list of SNPs (Index SNP, and High LD SNP) the relationship between these SNPs can be further explored.

After uploading the SNP lists as before, and adding the annotations for both lists, below shows an example of rs616488 and rs607941 (R^2 value of 0.71) interacting with PEX14. With this analysis, we can capture the regulatory targets of not only the index SNPs but also the other SNPs that are in LD with the index SNP, which might be the real causal variant for the disease.

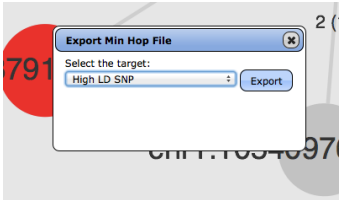




35457771	chr1	10538605	10538605		
117005712	chr1	10538718	10538718		
369741190	chr1	10538865	10538865		
145023046	chr1	10538902	10538902		
1152739	chr1	10538951	10538951		
112990066	chr1	10538969	10538969		
606240	chr1	10539203	10539203		
186135343	chr1	10539298	10539298		
188997414	chr1	10539342	10539342		
150870873	chr1	10539417	10539417		
11121578	chr1	10539503	10539503		
386601432	chr1	10539543	10539543		
607941	chr1	10539543	10539543		
182037349	chr1	10539667	10539667		
6680568	chr1	10539833	10539833		
61393665	chr1	10539893	10539893		
114240916	chr1	10539918	10539918		
369651139	chr1	10539979	10539979		
113388491	chr1	10540074	10540074		

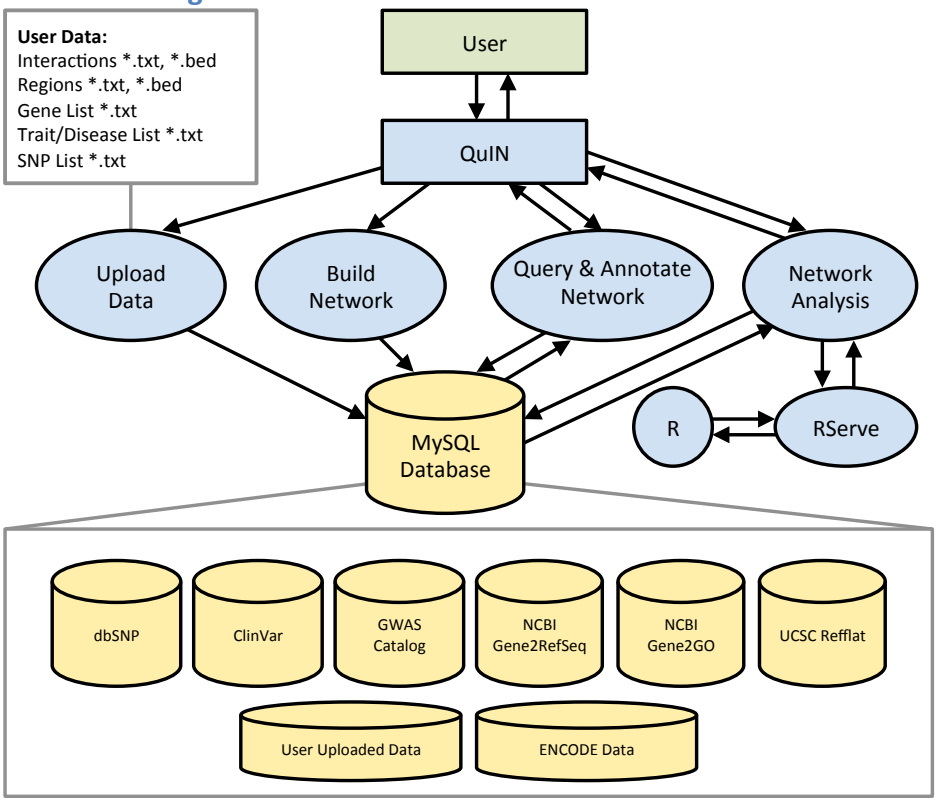


Additionally a target discovery analysis can be performed as before by choosing the target to be the LD buddy SNPs providing a way of globally analyzing the interactions between index SNPs and LD buddy SNPs. Rather than manually inspecting the interactions, the exported data can be further processed to determine which LD buddy SNPs have interactions with their respective index SNPs.



X. Appendix C (Software & Databases)

A. Data Flow Diagram



B. Software

Apache Tomcat - http://tomcat.apache.org/

Cytoscape JS - http://js.cytoscape.org/

Google GSON - https://github.com/google/gson

jQuery - https://jquery.com/

JQuery UI - http://jqueryui.com/

MySQL - https://www.mysql.com/

R - http://www.r-project.org/

RServe - https://rforge.net/Rserve/

Spectrum (JQuery Plugin) - http://bgrins.github.io/spectrum/

topGO - http://www.bioconductor.org/packages/release/bioc/html/topGO.html

Vioplot - http://cran.r-project.org/web/packages/vioplot/index.html

C. Databases

ClinVar - http://www.ncbi.nlm.nih.gov/clinvar/

dbSNP - http://www.ncbi.nlm.nih.gov/SNP/

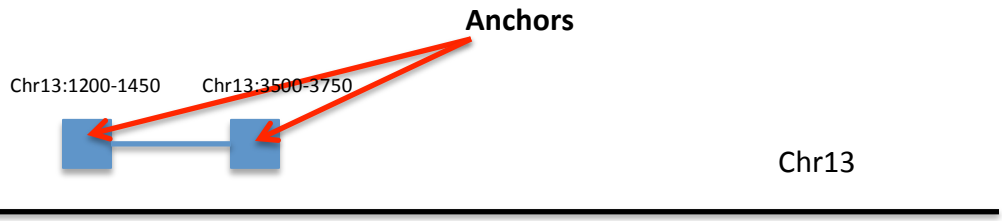
GWAS Catalog - http://www.genome.gov/gwastudies/

NCBI Gene - http://www.ncbi.nlm.nih.gov/gene/

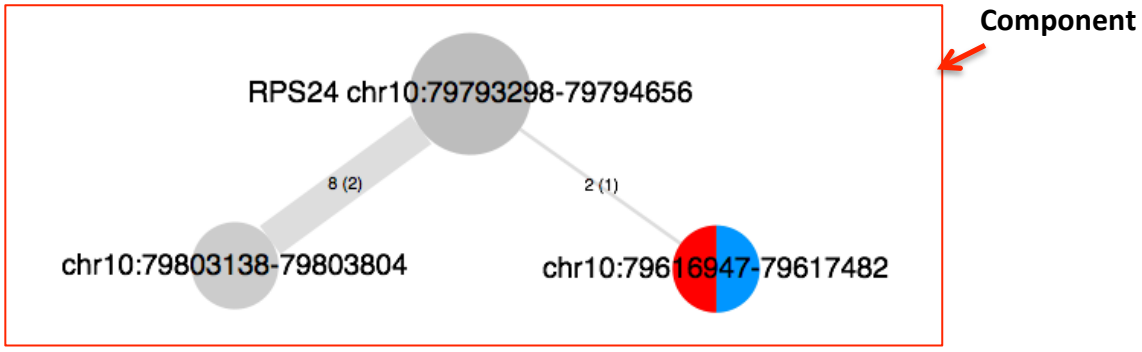
UCSC Refflat - http://genome.ucsc.edu/

XI. Appendix D (Interaction & Network Terminology)

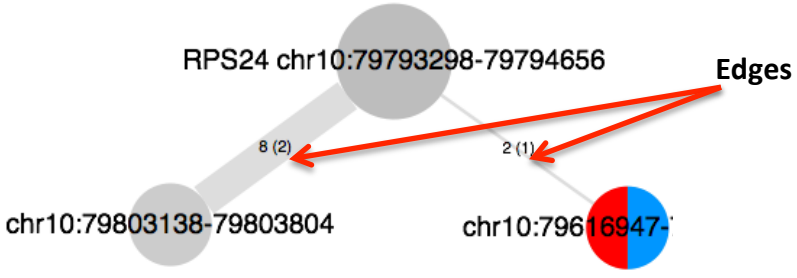
Anchor – One of the two genomic regions that are in close proximity. The chromosome, start position, and end position on the genome represents a single anchor and is paired with another anchor in an interaction.



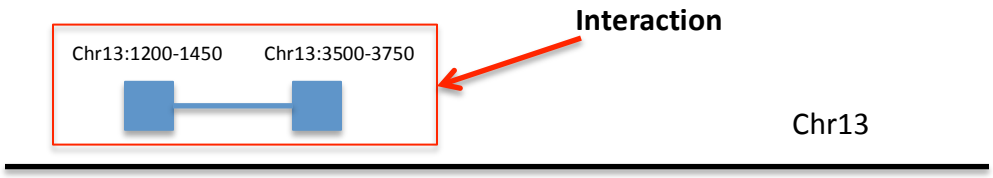
Component – A component is the subset of nodes that are connected to each other. Formally, it is the subset of nodes such that if there exists a path of edges from a node in the component to another node, the other node must also be in the component.



Edge – Edges define the interactions between two nodes and in the network visualization are represented as lines connected nodes. Edges are labeled based on the summation of the interaction counts and the number of interactions supporting the edge in parenthesis.



Interaction – An interaction is defined as two anchors that represent the genomic regions that are in close proximity with each other. Two anchors represent an interaction.



Node – Nodes define the genomic regions in the network obtained from either clusters of anchors or by predefined lists of regions. In the network visualization, nodes are represented as filled circles, which may be colored differently depending on the annotations provided and are labeled by the genomic region they represent. Additionally, if nodes are within 2kb upstream or downstream of a transcription start site of a gene from the UCSC "refflat" database, the node will also be labeled with that gene symbol.