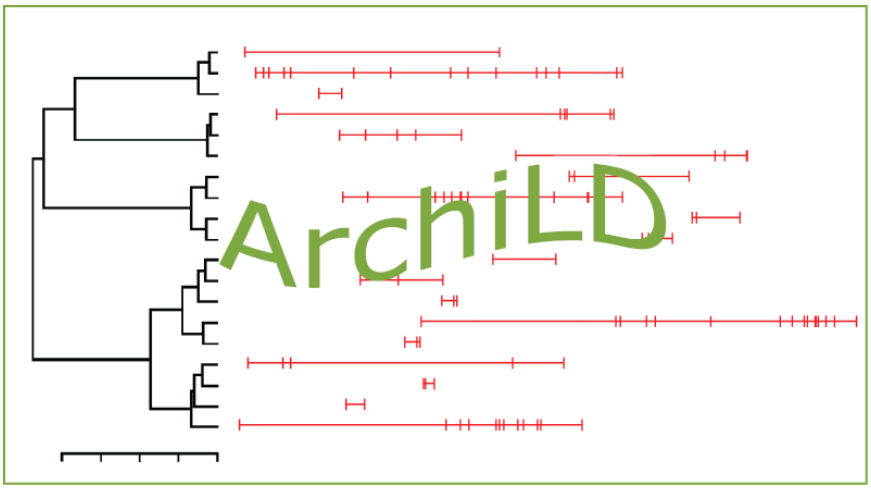
ArchiLD



Contents

1 Introduction

This software provides tools to compute and visualize LD architectures, clusters of SNPs in perfect linkage disequilibrium (LD) according to the metric r^2. Visualization of clusters is achieved through an integrated instance of the UCSC Genome Browser. LD patterns for four distinct populations (CEU,CHB+JPT,YRI) are available for visualization. LD was computed using Haploview on datasets provided by the 1000 Genome Pilot Project. Custom genotypic dataset can also be uploaded to analyze LD across different populations.

2 ArchiLD1k

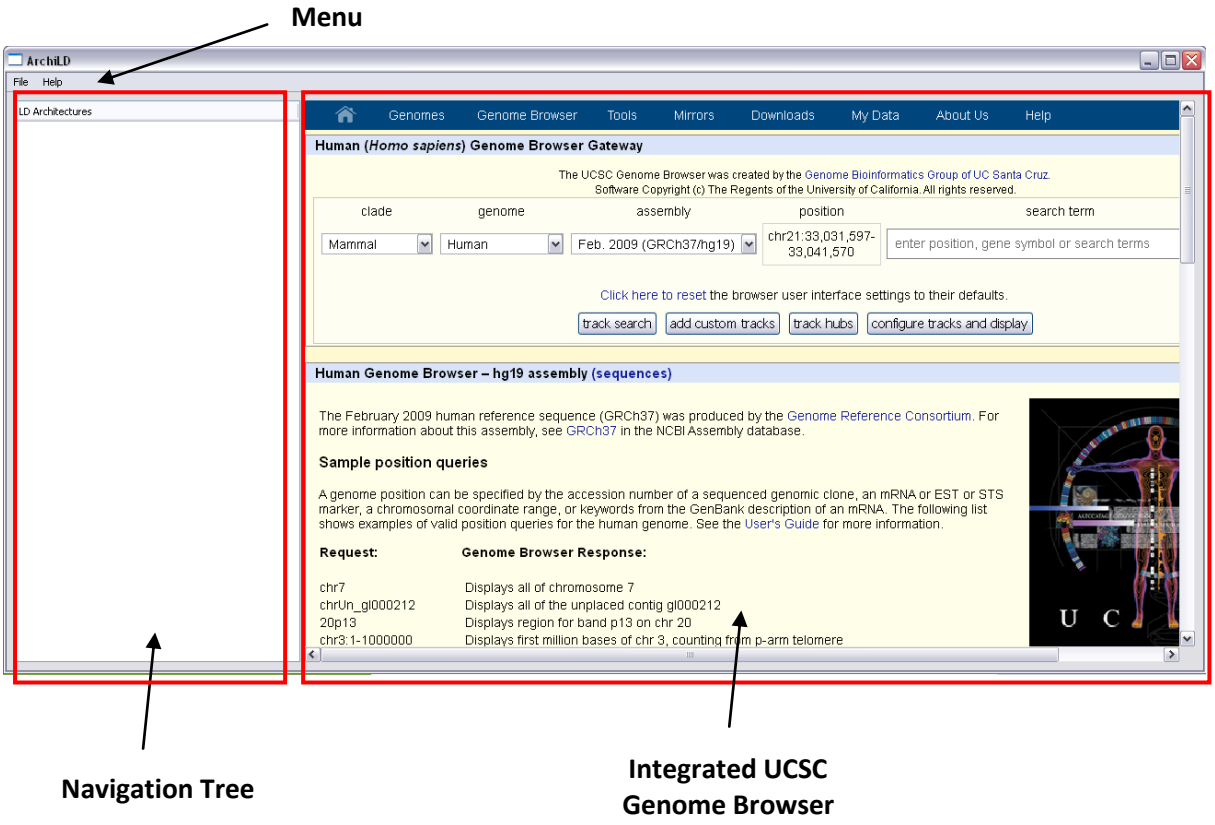
2.1 Requirements

The server version of ArchiLD is compatible with Windows XP, Windows 7, Linux and Mac OS. It requires Java Runtime Environment 1.6+.

2.2 Starting ArchiLD1k using the jar file (Windows/Linux/Mac)

1. Open http://archild.sign.a-star.edu.sg
2. Click on **Download ArchiLD1k** (choose the right platform)
3. Run the software
   - ⇨ **To start the software under Windows**
     1. Double click on the jar file
   - ⇨ **To start the software under Linux (command line)**
     Linux 64 bits requires glibc 2.9
     1. Browse to the directory containing the jar file
     2. Type java –jar jarfilename
   - ⇨ **To start the software under Mac (command line)**
     1. Browse to the directory containing the jar file
     2. Type java -XstartOnFirstThread –d32 –jar jarfilename

2.3 Description of the Graphical Interface



2.4 Create a new LD architecture

There are three distinct ways of creating an LD architecture: by gene, by SNP and by chromosome. The gene-centered LD architecture considers only clusters with at least a SNP inside the selected gene or in a small region spanning the gene (the size of this region can be adjusted through the parameter **max distance**). Transcription start sites and transcription end sites are defined as the smallest TSS and the largest TSE across all entries in UCSC associated to the Entrez ID of the transcript chosen, independently of the gene format selected for the analysis. All gene IDs are converted to Entrez IDs before computing clusters. The SNP-centered analysis builds an LD architecture from a reference SNP. The algorithm selects all SNP in LD with the reference SNP (only SNPs with an r^2 of at least 0.5 with the reference SNP are considered). The chromosome-centered analysis computes all the clusters located on a particular chromosome.

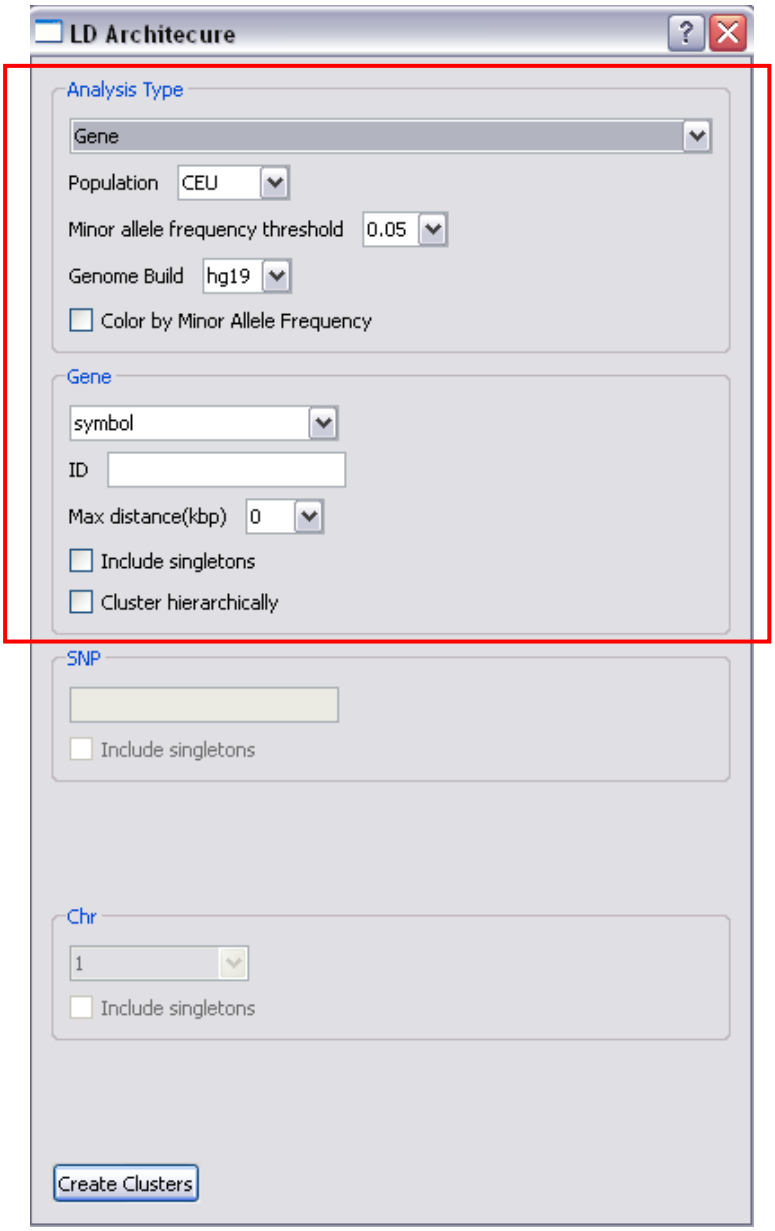
1. Click on **File>New**

⇨ **To create a new LD architecture starting from a gene**

1. Select **"Gene"** from the drop-down list **Analysis Type**
2. Choose the population of interest (CEU/CHBJPT/YRI)
3. Choose a minor allele frequency threshold for the clusters (0/0.01/0.05/0.1)
4. Choose a genome build (hg18/hg19)
5. Select Color by Minor Allele Frequency to color each SNP/block according to its minor allele frequency
6. Choose the format of the gene ID
   a. **entrez_geneid – Entrez gene ID**
   b. **mrna_accession – RefSeq mRNA ID**

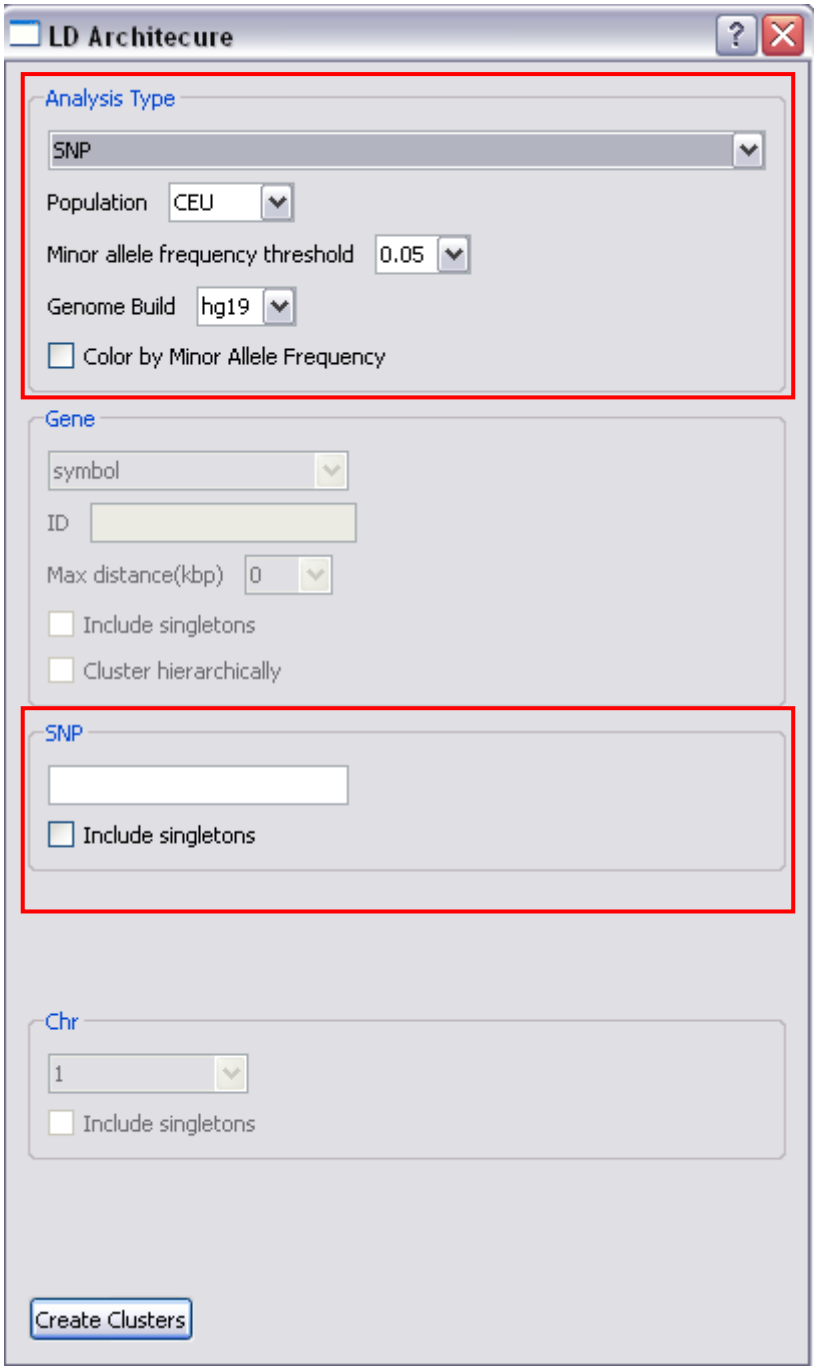
c. **symbol – Hugo gene ID**

7. Input the corresponding **ID**
8. Choose **max distance** (kbp) from the corresponding drop-down list (0kpb/10kbp/25kpb/50kbp/100kbp/250kbp/500kbp)
    a. Only SNPs located in the selected region upstream of the TSS, inside the gene and downstream of the TSE of the selected gene will be used to compute clusters
9. Tick **Include singletons** to include SNPs not belonging to any LD cluster
    a. If this option is selected it will not be possible to perform hierarchical clustering
10. Tick Cluster hierarchically to generate a hierarchical clustering of all the clusters associated to the gene
11. Click on **Create Clusters**



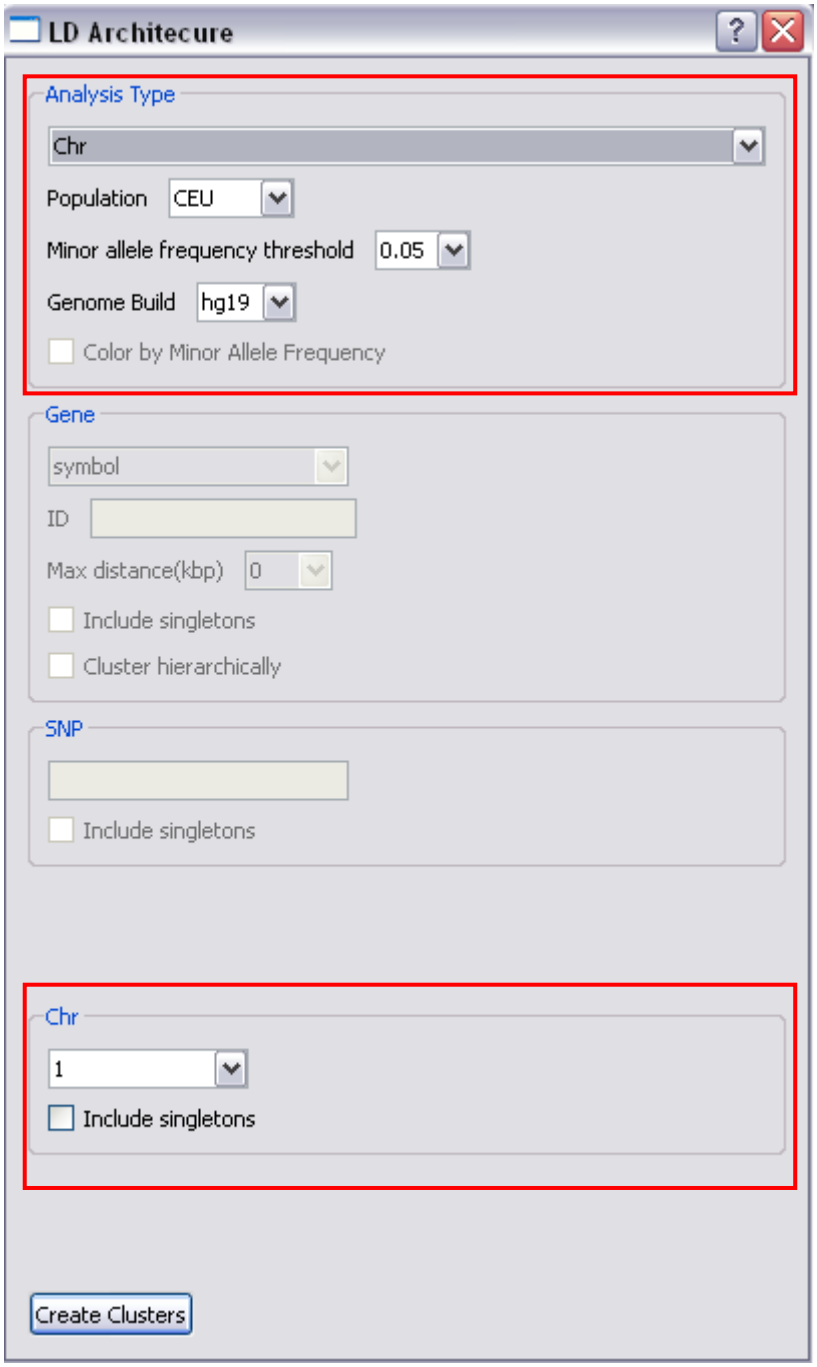
⇨ **To create a new LD architecture from a SNP**

1. Select **"SNP"** from the drop-down list **Analysis Type**
2. Choose the population of interest (CEU/CHBJPT/YRI)
3. Choose a minor allele frequency threshold for the clusters (0/0.01/0.05/0.1)
4. Choose a genome build (hg18/hg19)
5. Select Color by Minor Allele Frequency to color each SNP/block according to its minor allele frequency
6. Input the **ID** of the SNP of interest
7. Tick **Include singletons** if you want to include SNPs not belonging to any LD cluster
8. Click on **Create Clusters**

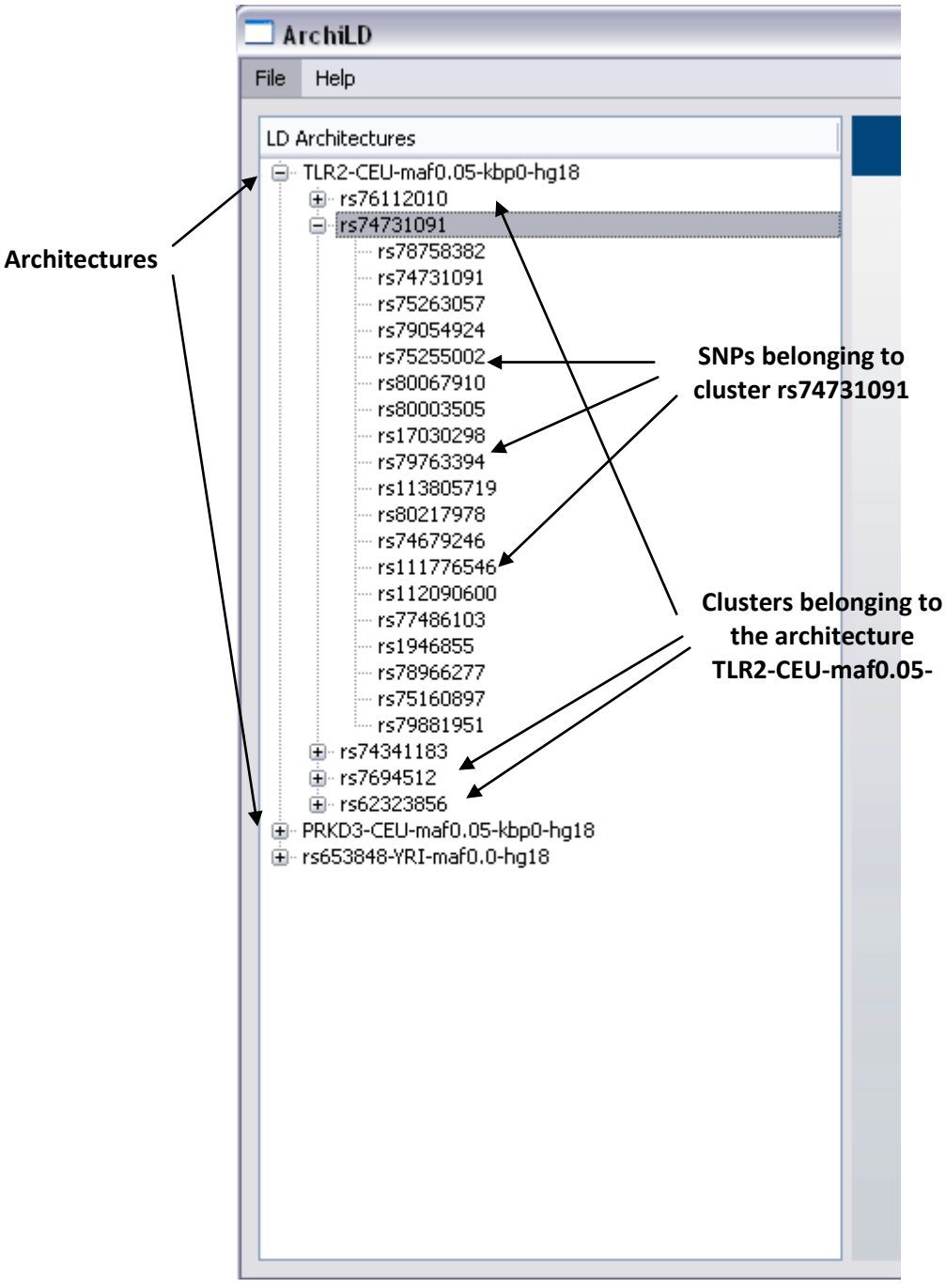


⇨ **To create a new LD architecture from a chromosome**

1. Select **"Chr"** from the drop-down list **Analysis Type**
2. Choose the population of interest (CEU/CHBJPT/YRI)
3. Choose a minor allele frequency threshold for the clusters (0/0.01/0.05/0.1)
4. Choose a genome build (hg18/hg19)
5. Choose the chromosome of interest
6. Tick **Include singletons** if you want to include SNPs not belonging to any LD cluster
7. Click on **Create Clusters**



Newly generated architectures are added to the navigation tree. For gene-centered and SNP centered architectures a list of cluster belonging to the architecture is also added to the tree. To see this list it is sufficient to expand the corresponding architecture item in the tree. It is also possible to see the list of SNPs belonging to a cluster by expanding the corresponding item in the tree.



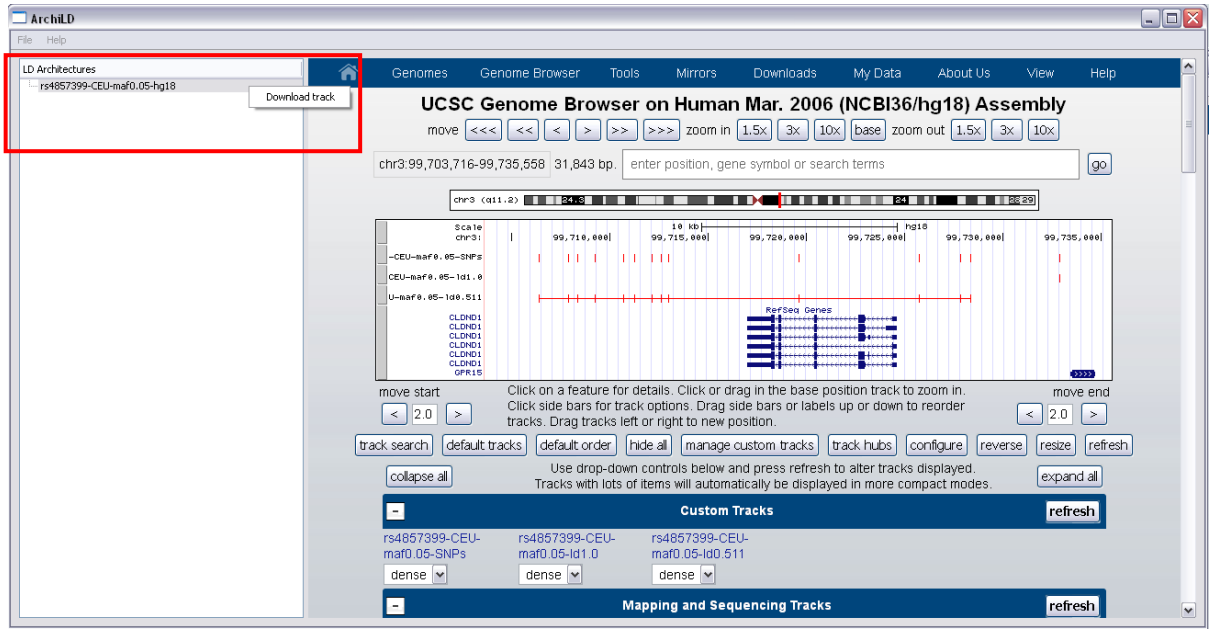
2.5 Export a UCSC bed track

This option allows the user to export a UCSC bed track.

⇨ **To export the bed track associated to a LD Architecture**

1. Right click on the name of the LD Architecture to export in the navigation tree

2. Select **Download track**
3. Specify a filename and a location where to save the track
4. The track exported can be imported in any instance of the UCSC genome browser

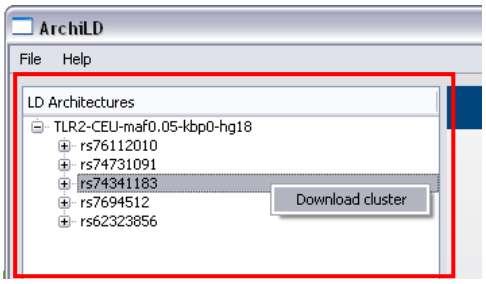


2.6 Export a cluster

This option allows the user to export the list of SNPs belonging to a cluster.

⇨ **To export a text file containing all SNPs belonging to a cluster**

1. Right click on the name of the cluster to export in the navigation tree
2. Select **Download cluster**
3. Specify a filename and a location where to save the text file

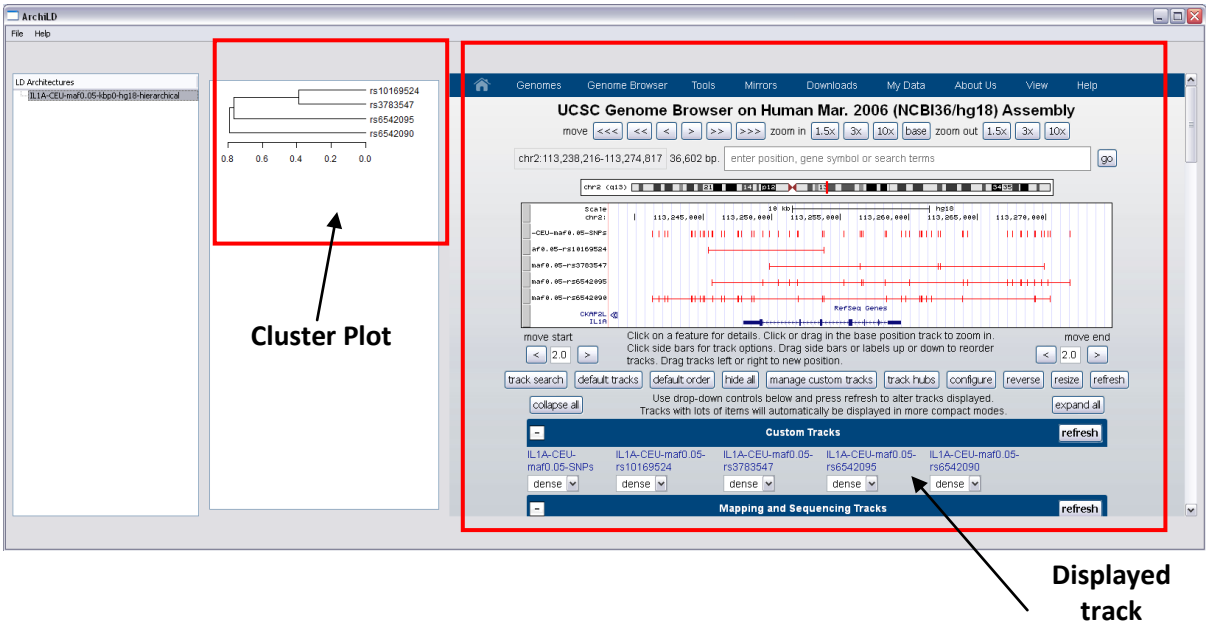


2.7 Display a gene/SNP UCSC bed track in the integrated browser

Once an LD Architecture has been created the corresponding link appears in the navigation tree located on the left side of the application window.

⇨ **To display a gene/SNP UCSC bed track in the integrated browser**

1. Click on the item corresponding to the track to visualize in the navigation tree
2. The corresponding track will be loaded in the integrated browser
    a. If the **Cluster hierarchically** option was selected a plot with the hierarchical cluster tree will be shown

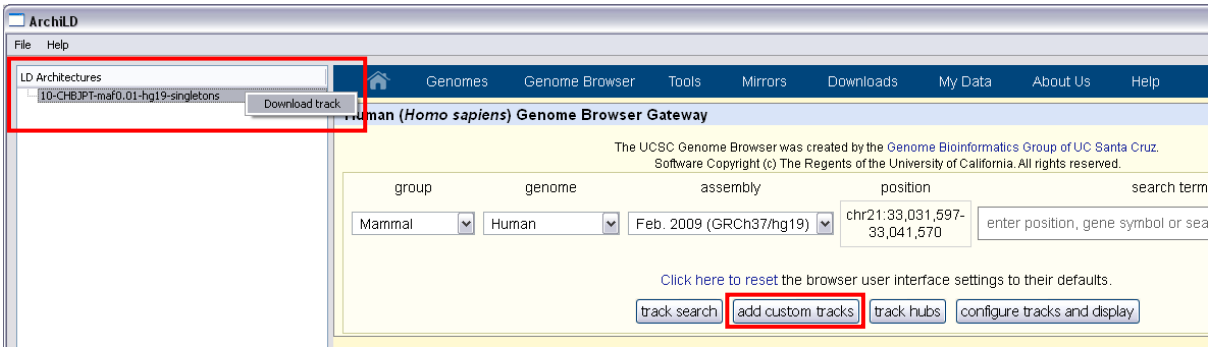


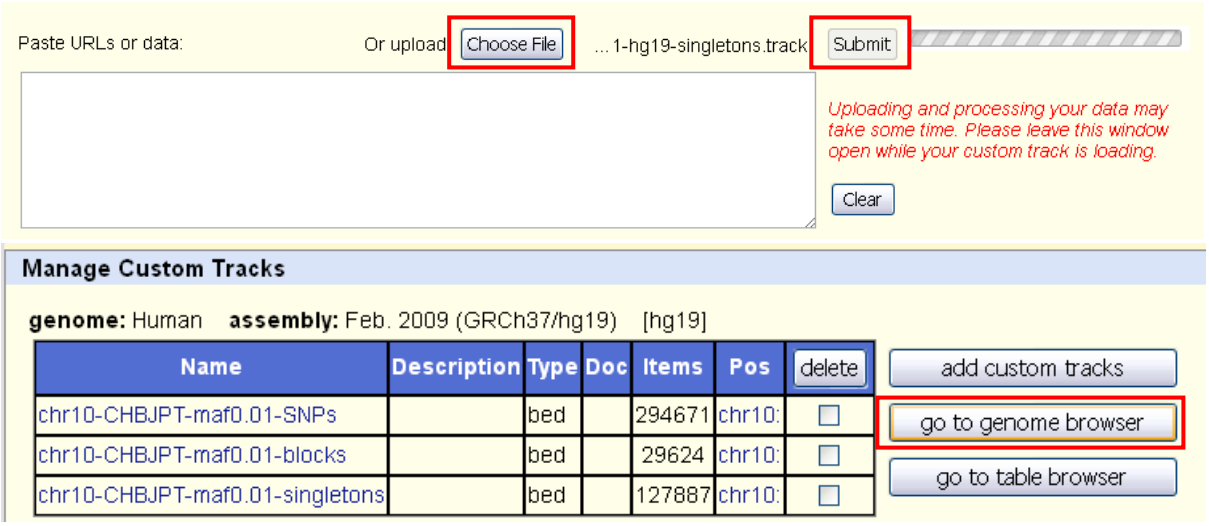
2.8 Display a chromosome UCSC bed track in the integrated browser

Once an LD architecture has been created the corresponding link appears in the navigation tree located on the left side of the application window.

⇨ **To display a chromosome UCSC bed track in the integrated browser**

1. Click on the item corresponding to the chromosome track
2. Export the track as described above
3. Import the track manually into the UCSC browser by clicking on **add custom track**
4. Click on **Choose file**, select the exported track file and click on **submit**
    a. Importing a chromosome file can take up to 5 minutes
5. Click on **go to genome browser**
6. The corresponding track will be loaded in the integrated browser





For more information about UCSC custom tracks please visit the UCSC website at http://genome.ucsc.edu/goldenPath/help/customTrack.html

3 ArchiLDCustom

3.1 Requirements

The local version of ArchiLD is compatible with Windows XP, Windows 7, Linux and Mac OS. It requires Java Runtime Environment 1.6+, a MySQL server, Haploview 4.2 and R.

3.1.1 MySQL installation

MySQL Cluster can be downloaded from: http://dev.mysql.com/downloads/mysql/. Users are advised to create a database to use exclusively for ArchiLD. Annotation tables are included in the zip file corresponding to the OS of choice. Annotation tables needs to be imported in the MySQL database of choice before ArchiLD can be used. The following mysql command can be used to import the tables:

mysql -u username -p databasename < path/table.sql

3.1.2 Haploview installation

Haploview (JAR Archive) can be downloaded from: http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/downloads#JAR. The software does not require any additional step to be installed. The installation of Haploview is optional and required only to upload datasets using genotypic data in the linkage format.

3.1.3 R installation

R can be downloaded from: http://www.r-project.org/.

3.2 Starting ArchiLDCustom using the jar file (Windows/Linux/Mac)

1. Open http://archild.sign.a-star.edu.sg
2. Click on **Download ArchiLDCustom** (choose the right platform)
3. Run the software
   - ⇨ **To start the software under Windows**
     1. Double click on the jar file

⇨ **To start the software under Linux (command line)**

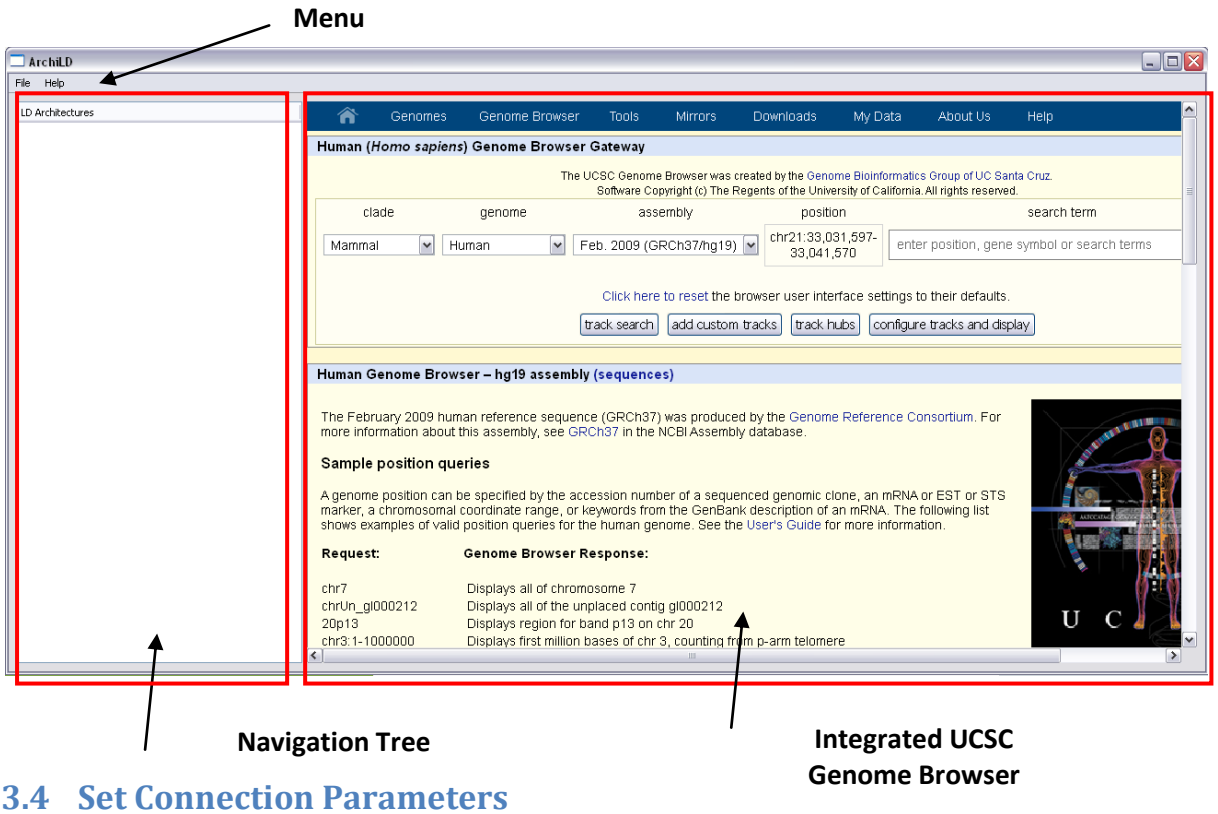
Linux 64 bits requires glibc 2.9

1. Browse to the directory containing the jar file
2. Type java –jar jarfilename

⇨ **To start the software under Mac (command line)**

1. Browse to the directory containing the jar file
2. Type java -XstartOnFirstThread –d32 –jar jarfilename

3.3 Description of the Graphical Interface

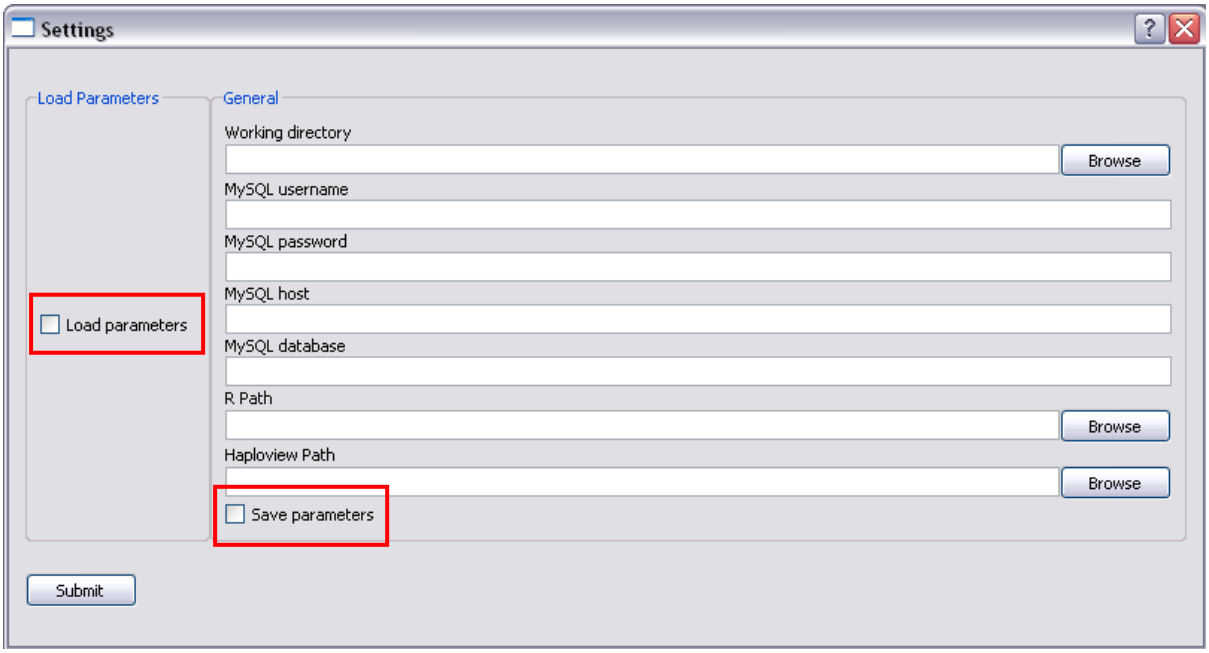


3.4 Set Connection Parameters

Before any analysis can be performed connection parameters need to be loaded. Connection parameters contain information about which MySQL server to use and about paths to software that needs to be installed for the analysis.

⇨ **To set parameters**

1. **Click on File>Settings**



Here is a description of the parameters that needs to be set:

Working directory: directory in which to store all temporary files created by the software. It is preferable to use a dedicated directory for ArchiLD.
MySQL username: username to use to connect to the MySQL server
MySQL password: password to use to connect to the MySQL server
MySQL host: host where the MySQL server resides
MySQL database: database containing all annotation tables where tables created by the software will be stored
R Path: complete path to the R executable
Haploview Path (optional): complete path to the Haploview jar file.

- ⇨ **To save parameters for future use**
  1. Fill up all fields
  2. Click on the **Save parameters** checkbox
  3. Choose a directory and a filename
  4. Click on **Submit**
- ⇨ **To load parameters**
  1. Click on the **Load parameters** checkbox
  2. Select a previously saved parameter file
  3. Click on **Submit**

THE COMPLETE PATH OF THE WORKING DIRECTORY MUST NOT CONTAIN ANY SPACE.

3.5 Load a dataset

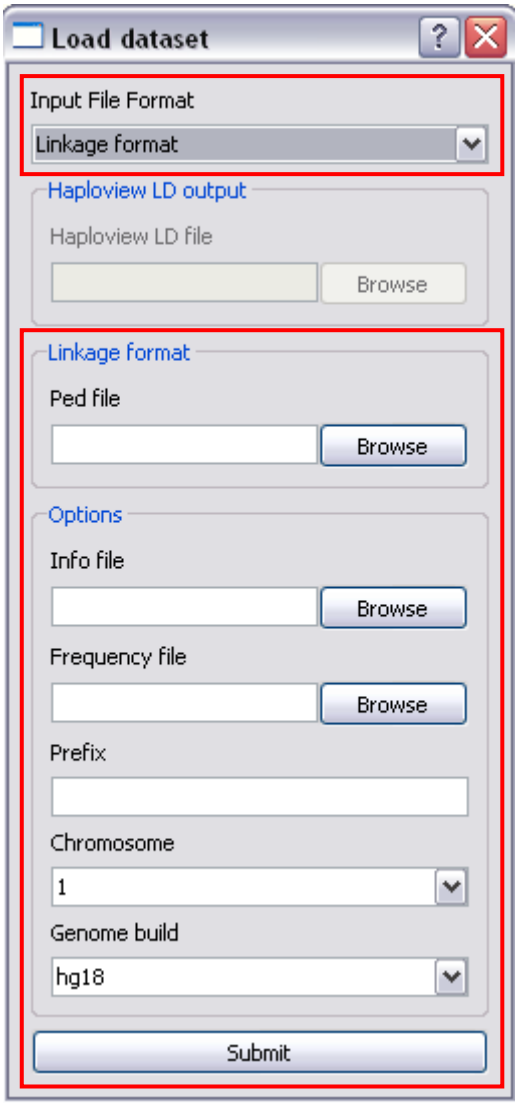
Parameters need to be set before this option becomes available.

- ⇨ **To load a new dataset using a genotype file (Linkage format)**

1. Click on **File>Load dataset**
2. Select **Linkage format** as the **Input File Format**
3. Select a **PED** file
4. Select a **INFO** file
5. Select a **frequency** file
6. Select a **prefix** (all temporary files created will have this prefix, the dataset will appear in the list with that name)
   Prefix should not contain any special character.
7. Select the **chromosome** corresponding to the INFO file imported
8. Select the **genome build** corresponding to the INFO file imported
9. Click on Submit

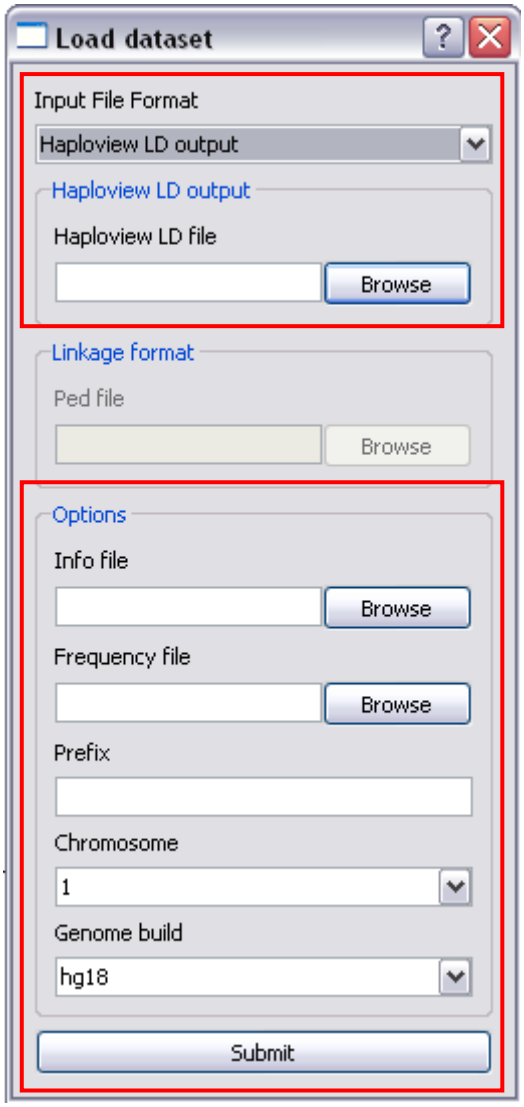
LD values will be computed using Haploview and all the necessary tables will be automatically created in the database. The path to Haploview needs to be set in the parameters for this option to become available.

PED AND INFO FILENAMES MUST NOT CONTAIN ANY SPACE CHARACTER NOR BE LOCATED IN A DIRECTORY WHOSE COMPLETE PATH CONTAINS SPACE CHARACTERS.



⇨ **To load a new dataset using a pre-computed LD Haploview file**

1. Click on **File>Load dataset**
2. Select **Haploview LD output** as the **Input File Format**
3. Select an LD file generated by Haploview
4. Select a **INFO** file
5. Select a **frequency** file
   Select a **prefix** (all temporary files created will have this prefix, the dataset will appear in the list with that name)
   Prefix should not contain any special character.
6. Select the **chromosome** corresponding to the INFO file imported
7. Select the **genome build** corresponding to the INFO file imported
8. Click on Submit



⇨ **To re-load a previously loaded dataset**

1. Previously loaded datasets are imported automatically

3.5.1 PED file description

A PED file contains information about the genotypes of the samples included in the file. It is organized as follows:

Family ID: Family ID of the sample (not used by the software)
Individual ID: ID of the sample
Paternal ID: paternal ID of the sample (not used by the software)
Maternal ID: maternal ID of the sample (not used by the software)
Sex (1=male; 2=female; other=unknown): sex of the sample (not used by the software)
Phenotype (1=not affected;2=affected;0=unknown): phenotype of the sample (not used by the software)
The following columns contain information about the genotype of each sample. Alleles are separated by a space. All markers should be bi-allelic.

ALL GENOTYPES CONTAINED IN THE PED FILE NEEDS TO REFER TO THE SAME CHROMOSOME. SEX CHROMOSOMES CANNOT BE USED IN ARCHILD.

Example

```
Family1  Sample1  0  0  1  0  A A  T C  C G  A A
Family2  Sample2  0  0  1  0  A G  C C  C G  A A
Family3  Sample3  0  0  2  0  A G  T C  G G  A T
Family4  Sample4  0  0  2  0  A A  T T  C G  A T
Family5  Sample5  0  0  1  0  G G  T T  C G  A T
```

Missing phenotypes must be encoded as 0. The file should not contain any header.

3.5.2 INFO file description

An INFO file contains information about the markers included in the genotype file. The order of the markers in the INFO file needs to correspond to the order of the genotypes in the PED file. It is organized in two columns:

SNP: SNP ID
Position: base-pair position on the chromosome

All markers need to reside on the same chromosome. Sex chromosomes cannot be used in ArchiLD. The file should not contain any header.

Example
The INFO file corresponding to the previous PED file could be:

```
Marker1  847849393  (this marker has two possible alleles A/G)
Marker2  847888393  (this marker has two possible alleles C/T)
Marker3  878600000  (this marker has two possible alleles C/G)
Marker4  878603564  (this marker has two possible alleles A/T)
```

3.5.3 Frequency file description

A frequency file contains 5 columns:

CHR: chromosome
SNP: SNP ID
A1: minor allele
A2: major allele
MAF: minor allele frequency
NCHROBS: non-missing allele count

This file can be easily generated using the software PLINK with the command –freq on PED and MAP files. MAP files are very similar to INFO files but contain two additional columns:
CHR: chromosome
SNP: SNP ID
Genetic distance: genetic distance (not used by ArchiLD, the value can be set to 0)
Position: base-pair position on the chromosome

3.5.4 Haploview LD output description

A Haploview LD output contains all pairwise LD metrics for the set of SNPs described in the INFO file. It includes 9 columns:

L1: first SNP identifier
L2: second SNP identifier
D': D' prime between the two SNPs
LOD: log of the likelihood odds ratio
r^2: correlation coefficient between the two SNPs
CIlow: 95% confidence interval lower bound on D'
CIhi: 95% confidence interval upper bound on D'
Dist: distance between the two SNPs in base pairs
T-int: statistic used by the HapMap project as a measure of completeness of information

More details about this file type can be found at the following page:
http://www.broadinstitute.org/science/programs/medical-and-population-genetics/haploview/output-file-formats

Example datasets containing all four files here described can be found in the zip file containing the software under the directory Datasets.

3.6 Example datasets

Two example datasets are included with ArchiLDCustom.

CEU.chr7.first.500kbp contains all the SNPs (1137) sequenced by the 1000 Genomes Pilot Project in 60 CEU samples located in the first 500kbp of chromosome 7. The SNPs described in the INFO files are annotated using the genome build hg18.

CHBJPT.chr12.first.500kbp contains all the SNPs (1356) sequenced by the 1000 Genomes Pilot Project in 60 CHB+JPT samples located in the first 500kbp of chromosome 12. The SNPs described in the INFO files are annotated using the genome build hg18.

Files can be imported for analysis as described in 3.5. Loading the dataset can take up to 2min.

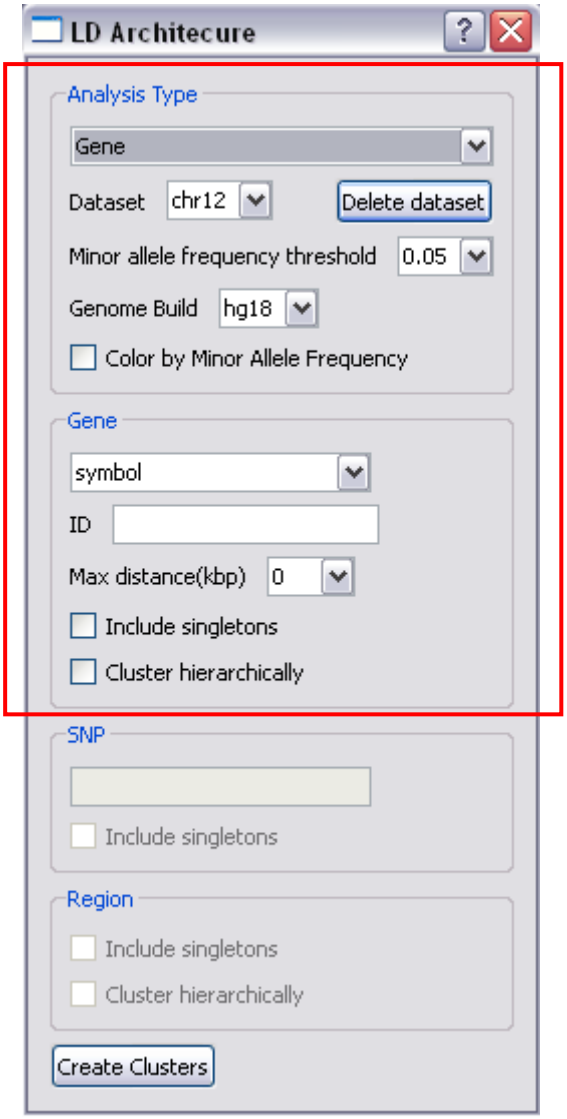
3.7 Create a new LD architecture

There are three distinct ways of creating an LD architecture: by gene, by SNP and by region. The gene-centered LD architecture considers only clusters with at least a SNP inside the selected gene or in a small region spanning the gene (the size of this region can be adjusted through the parameter **max distance**). Transcription start sites and transcription end sites are defined as the smallest TSS and the largest TSE across all entries in UCSC associated to the Entrez ID of the transcript chosen, independently of the gene format selected for the analysis. All gene IDs are converted to Entrez IDs before computing clusters. The SNP-centered analysis builds an LD architecture from a reference SNP. The algorithm selects all SNP in LD with the reference SNP (only SNPs with an r^2 of at least 0.5 with the reference SNP are considered). The region-centered analysis computes all the clusters in the region described by the selected dataset.

1. Click on **File>New**

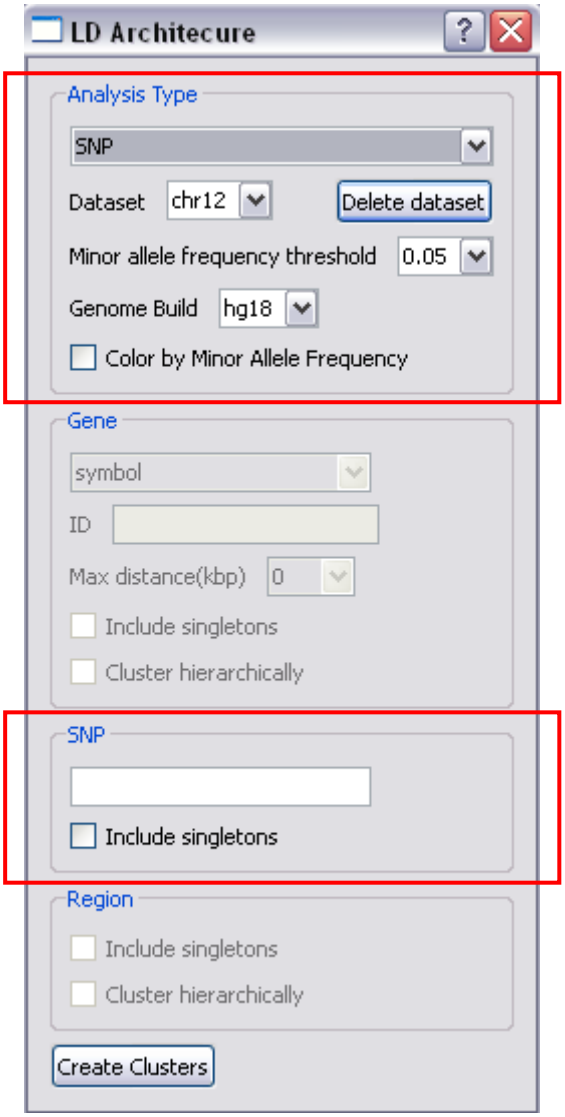
⇨ **To create a new LD architecture starting from a gene**

1. Select **"Gene"** from the drop-down list **Analysis Type**
2. Choose the **dataset** to use
    a. All previously loaded datasets are made available for the analysis
    b. The field genome build contains the genome build of the selected dataset
3. Choose a minor allele frequency threshold for the clusters (0/0.01/0.05/0.1)
4. Choose the format of the gene ID
    a. **entrez_geneid – Entrez gene ID**
    b. **mrna_accession – RefSeq mRNA ID**
    c. **symbol – Hugo gene ID**
5. Input the corresponding **ID**
6. Choose **max distance** (kbp) from the corresponding drop-down list (0kpb/10kbp/25kpb/50kbp/100kbp/250kbp/500kbp)
    a. Only SNPs located in the selected region upstream of the TSS, inside the gene and downstream of the TSE of the selected gene will be used to compute clusters
7. Select Color by Minor Allele Frequency to color each SNP/block according to its minor allele frequency
8. Tick **Include singletons** to include SNPs not belonging to any LD cluster
    a. If this option is selected it will not be possible to perform hierarchical clustering
9. Tick Cluster hierarchically to generate a hierarchical clustering of all the clusters associated to the gene
10. Click on **Create Clusters**



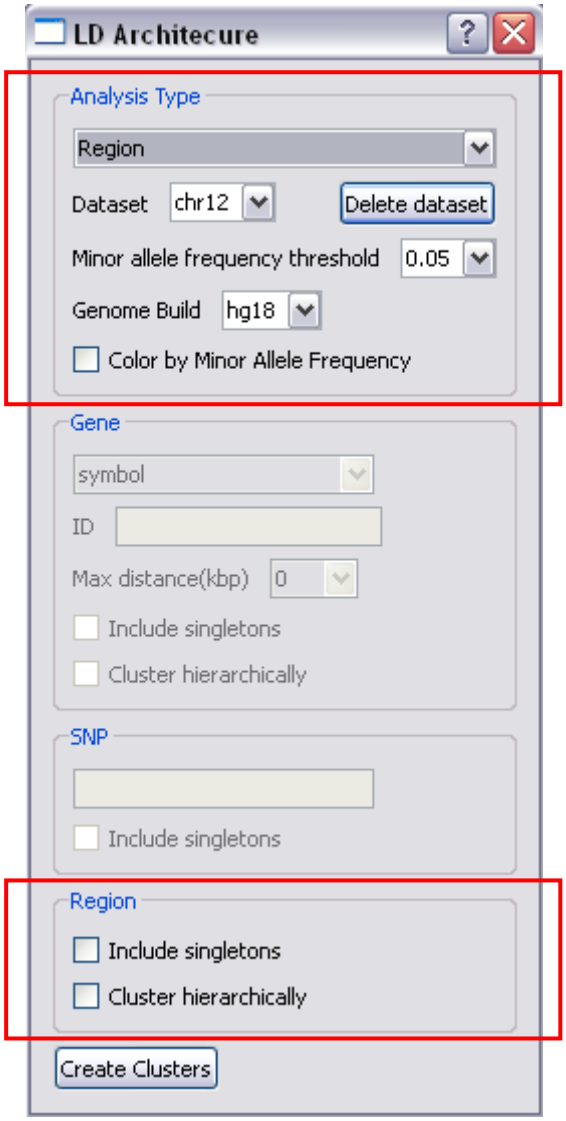
⇨ **To create a new LD architecture from a SNP**

1. Select **"SNP"** from the drop-down list **Analysis Type**
2. Choose the **dataset** to use
    a. All previously loaded datasets are made available for the analysis
    b. The field genome build contains the genome build of the selected dataset
3. Choose a minor allele frequency threshold for the clusters (0/0.01/0.05/0.1)
4. Select Color by Minor Allele Frequency to color each SNP/block according to its minor allele frequency
5. Input the **ID** of the SNP of interest
6. Tick **Include singletons** if you want to include SNPs not belonging to any LD cluster
7. Click on **Create Clusters**

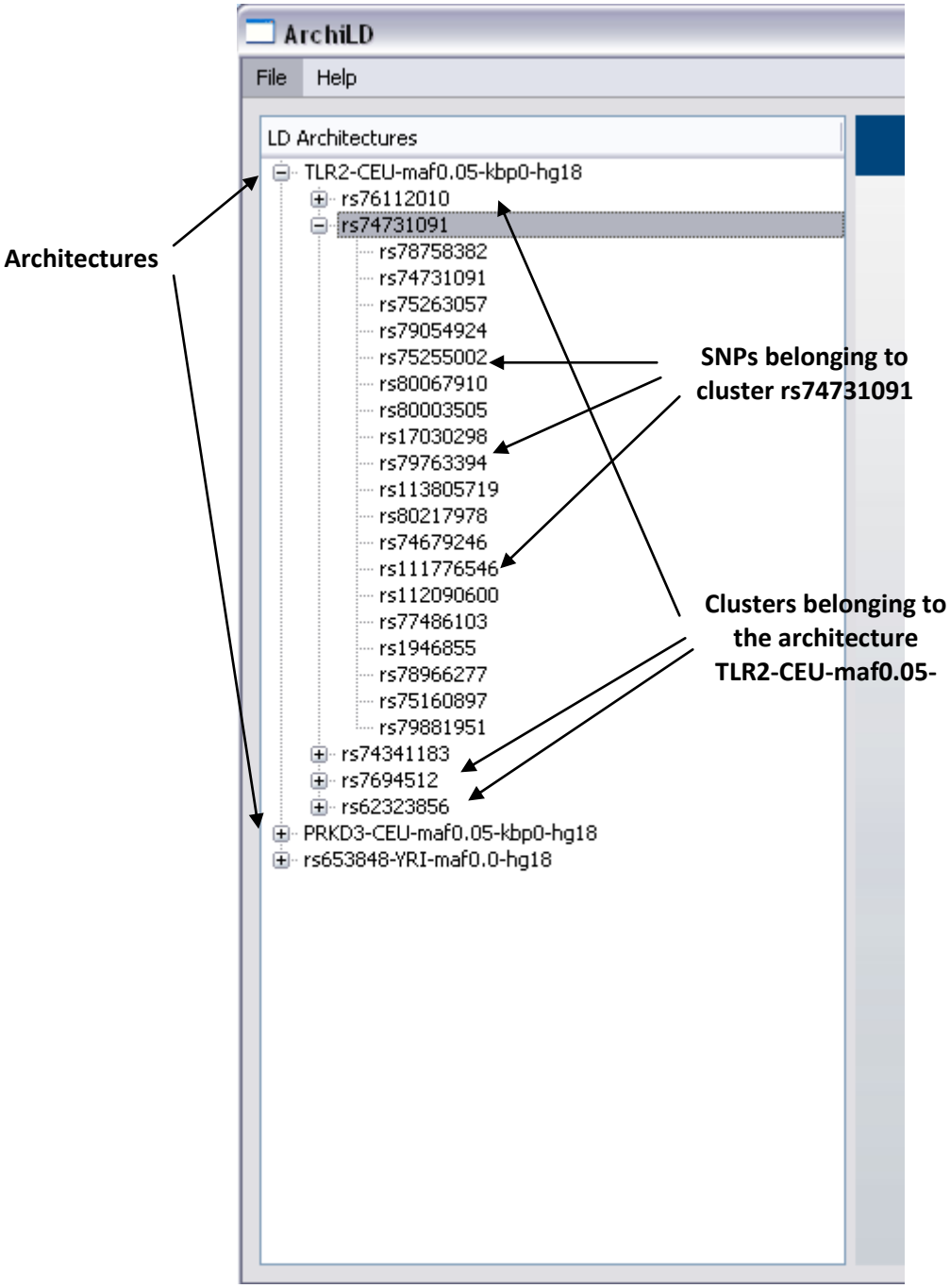


⇨ **To create a new LD architecture from a region**

1. Select **"Region"** from the drop-down list **Analysis Type**
2. Choose the **dataset** to use
    a. All previously loaded datasets are made available for the analysis
    b. The field genome build contains the genome build of the selected dataset
3. Choose a minor allele frequency threshold for the clusters (0/0.01/0.05/0.1)
4. Select Color by Minor Allele Frequency to color each SNP/block according to its minor allele frequency
5. Tick **Include singletons** if you want to include SNPs not belonging to any LD cluster
6. Tick Cluster hierarchically to generate a hierarchical clustering of all the clusters associated to the gene
7. Click on **Create Clusters**



Newly generated architectures are added to the navigation tree. For gene-centered, SNP-centered and gene-centered architectures a list of cluster belonging to the architecture is also added to the tree. To see this list it is sufficient to expand the corresponding architecture item in the tree. It is also possible to see the list of SNPs belonging to a cluster by expanding the corresponding item in the tree.



3.8 Delete a dataset

This option allows the user to delete a dataset from the list of previously loaded datasets.

⇨ **To delete a dataset from the list**

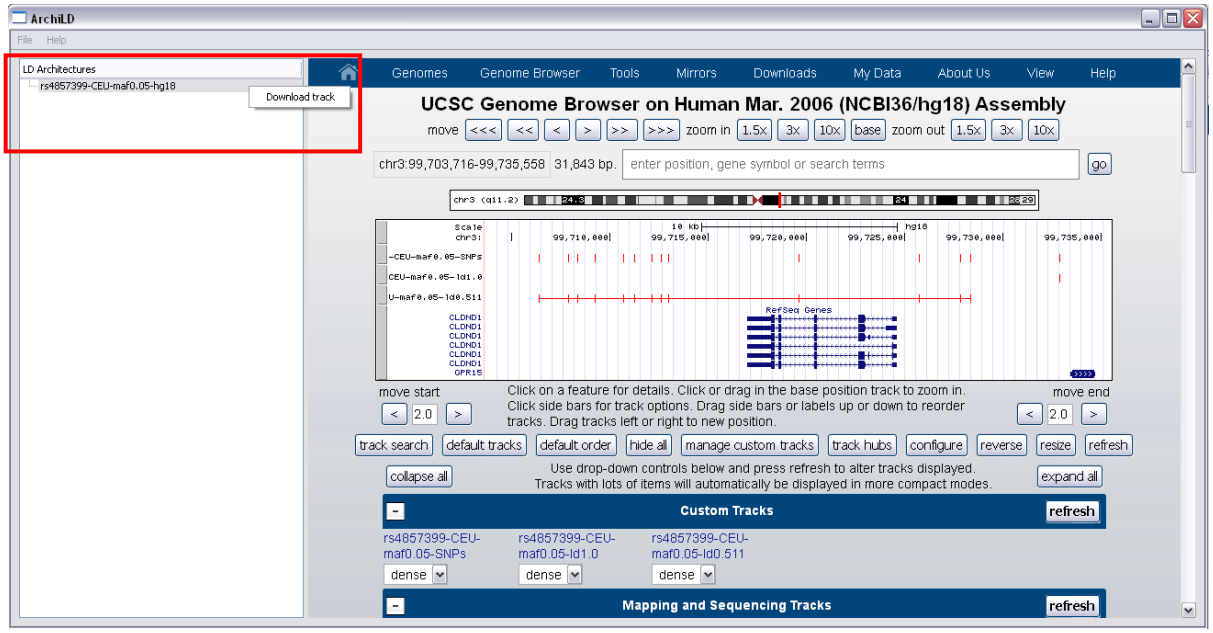
1. Click on **File>New**
2. Choose a dataset from the list **Dataset**
3. Click on **Delete dataset**

3.9 Export a UCSC bed track

This option allows the user to export a UCSC bed track.

⇨ **To export the bed track associated to a LD Architecture**

1. Right click on the name of the LD Architecture to export in the navigation tree
2. Select **Download track**
3. Specify a filename and a location where to save the track
4. The track exported can be imported in any instance of the UCSC genome browser

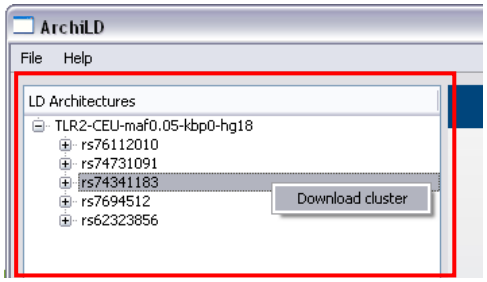


3.10 Export a cluster

This option allows the user to export a UCSC bed track.

⇨ **To export a text file containing all SNPs belonging to a cluster**

1. Right click on the name of the cluster to export in the navigation tree
2. Select **Download cluster**
3. Specify a filename and a location where to save the text file

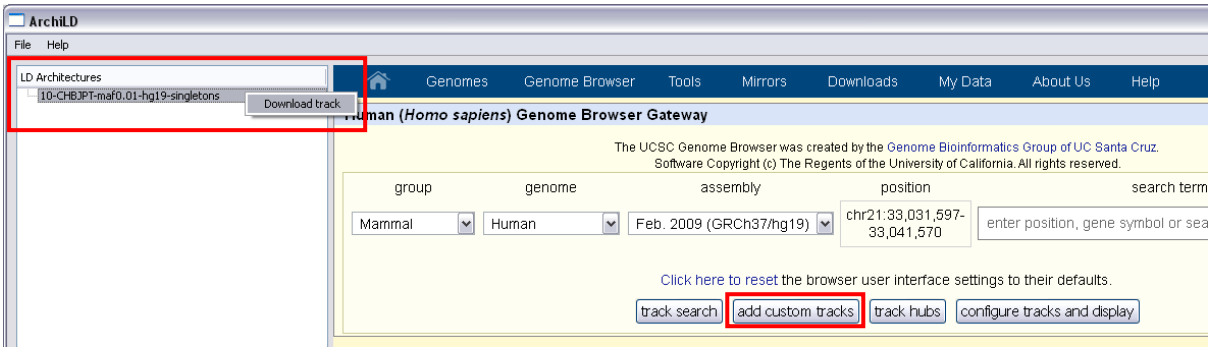


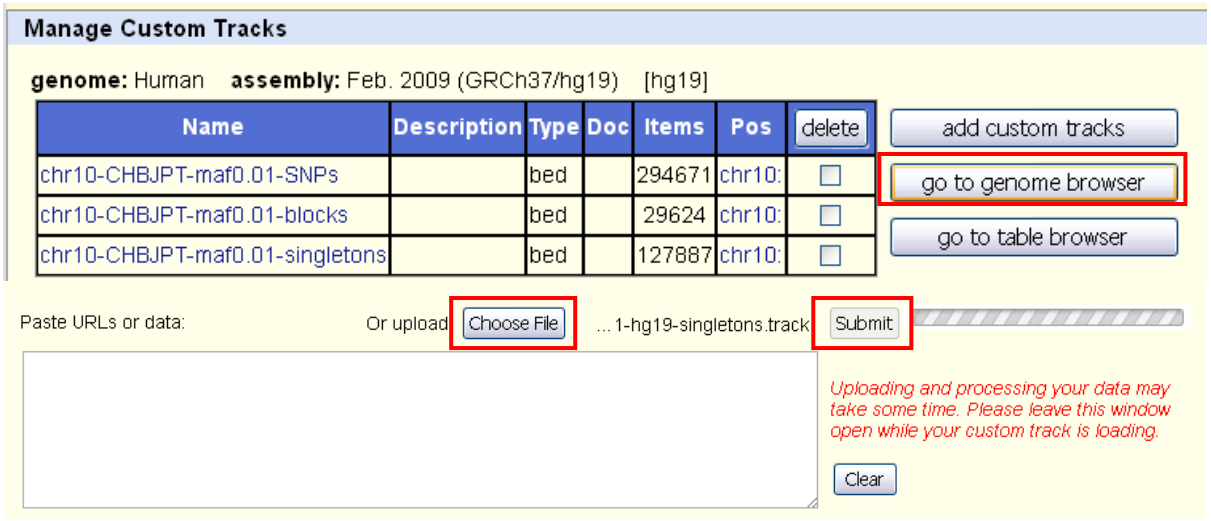
3.11 Display a gene/SNP/region UCSC bed track in the integrated browser

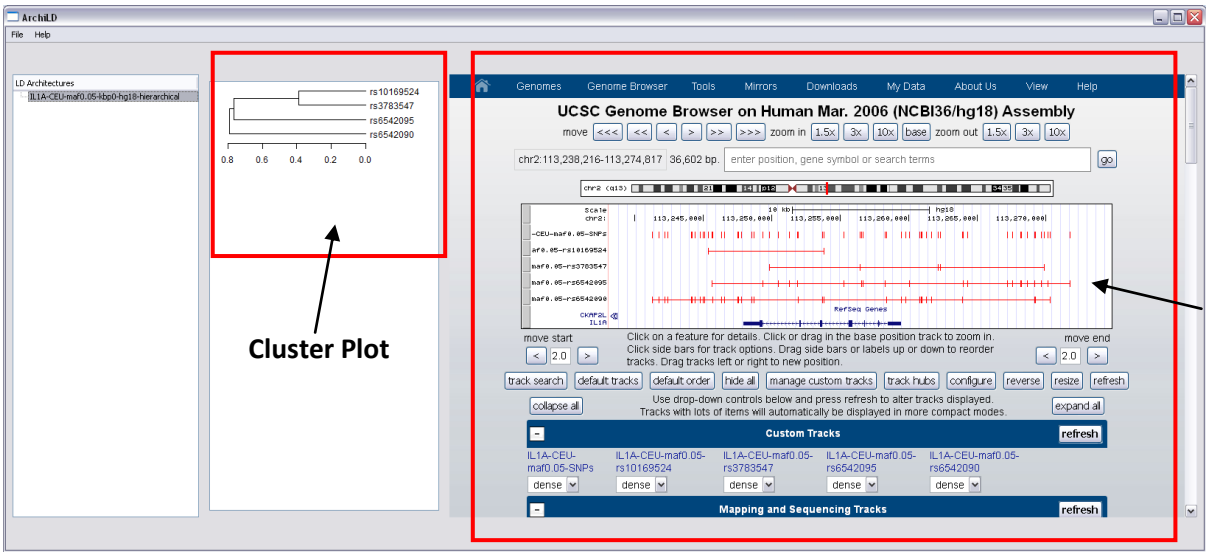
Once an LD Architecture has been created the corresponding link appears in the navigation tree located on the left side of the application window.

1. Click on the item corresponding to the chromosome track
2. Export the track as described above
3. Import the track manually into the UCSC browser by clicking on **add custom track**

4. Click on **Choose file**, select the exported track file and click on **submit**
5. Click on **go to genome browser**
6. The corresponding track will be loaded in the integrated browser
    a. If the **Cluster hierarchically** option was selected a plot with the hierarchical cluster tree will be shown





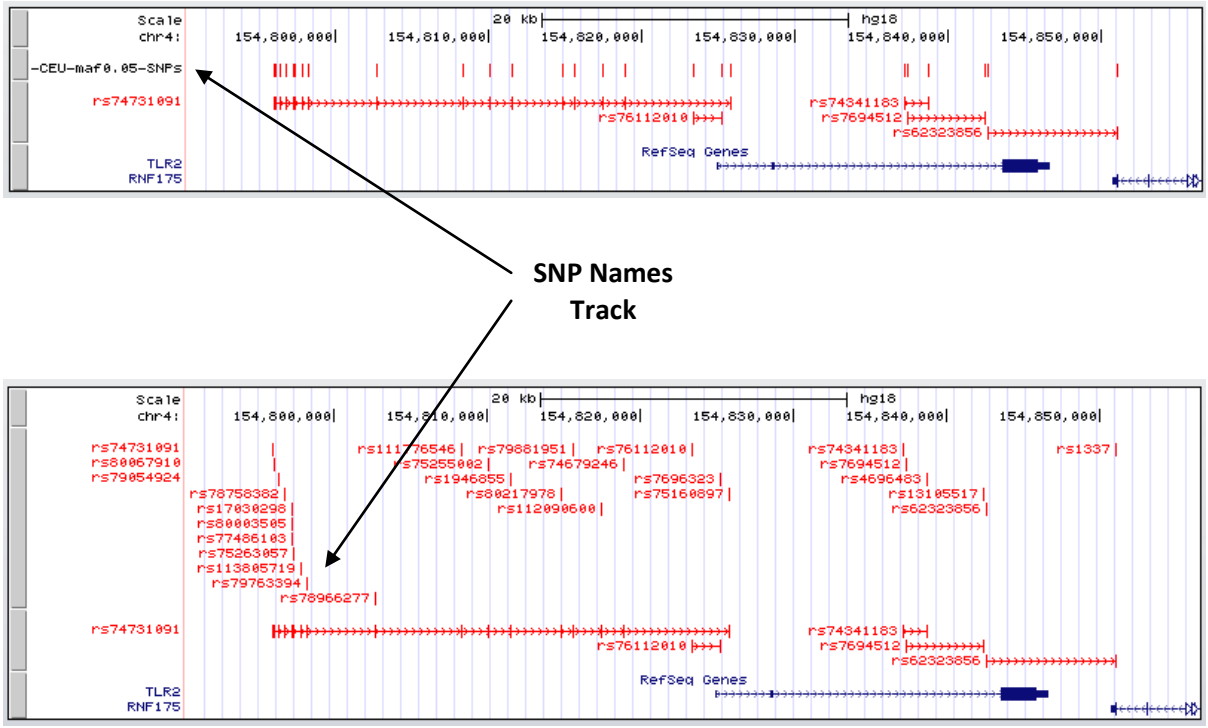


For more information about UCSC custom tracks please visit the UCSC website at http://genome.ucsc.edu/goldenPath/help/customTrack.html

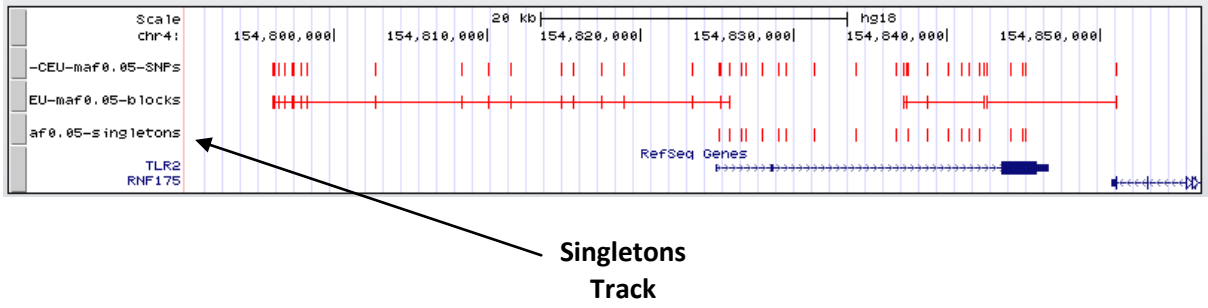
4 Interpret plots

- **Plot description for gene-centered/region-centered architectures (Cluster hierarchically option not selected)**

  Each vertical line represents a SNP. SNPs linked by a line are in perfect linkage. All the clusters are added to the same track. Every cluster is identified by the name of the first SNP in it. Individual SNP names can be obtained by expanding the first track.

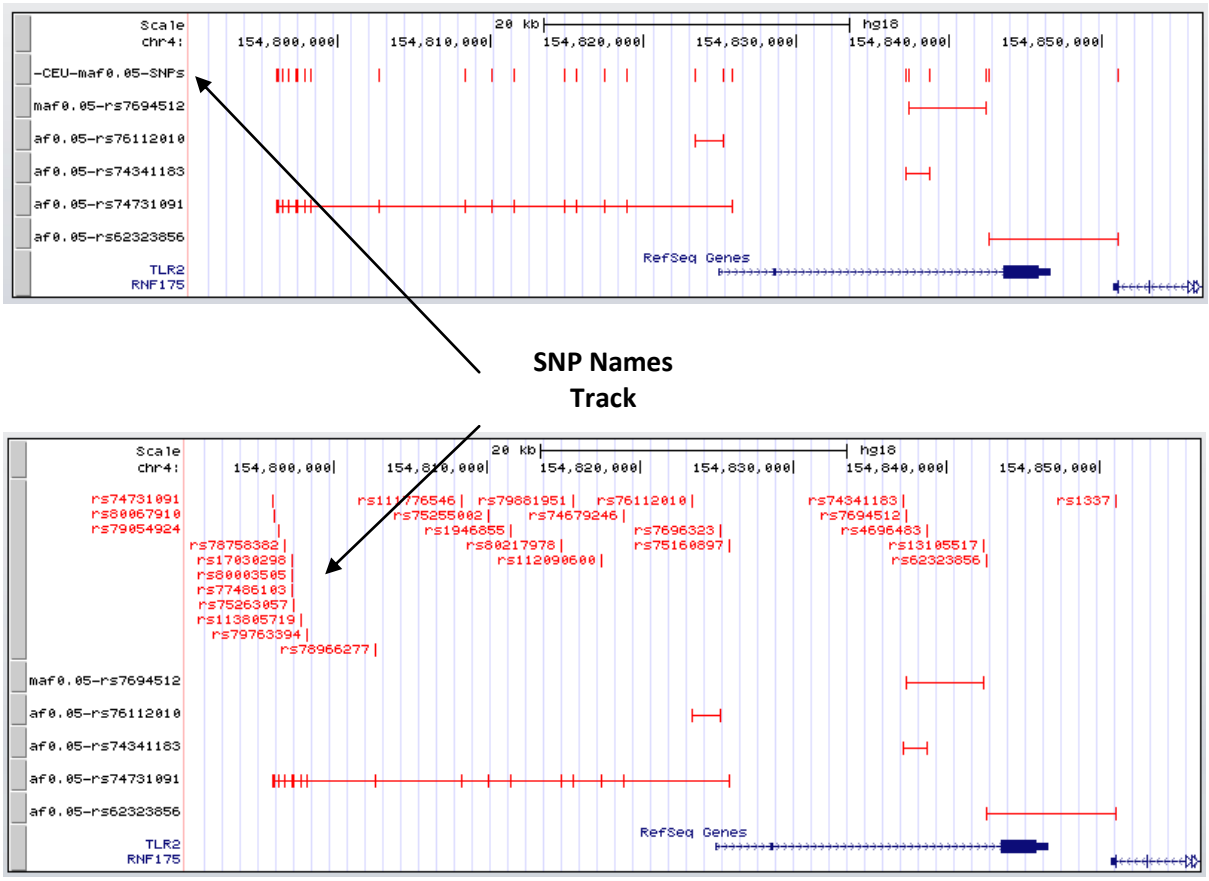


If the option Include Singletons has been selected singletons are included in a distinct track.

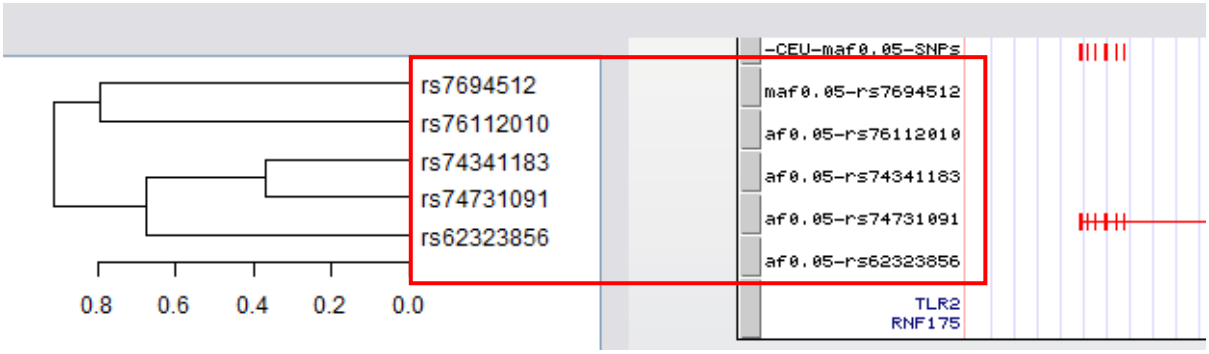


- **Plot description for gene-centered/region-centered architectures (Cluster hierarchically option selected)**

Each vertical line represents a SNP. SNPs linked by a line are in perfect linkage (r^2=1). Each track represents a cluster. Individual SNP names can be obtained by expanding the first track. To display the name of a cluster it is sufficient to change the visibility of the corresponding track to “pack”.



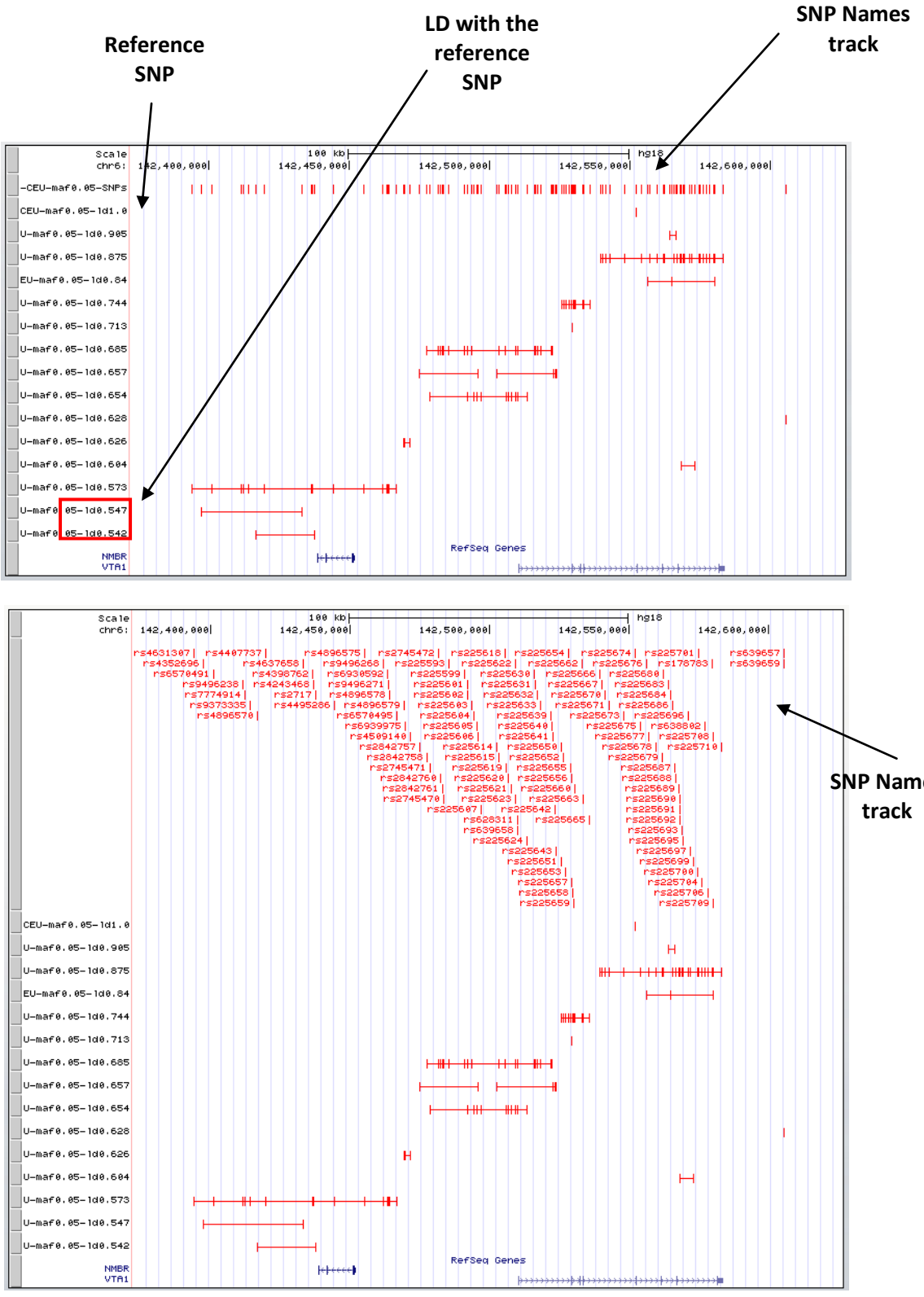
The order of the tracks follows the order of the SNPs in the clustering plot.



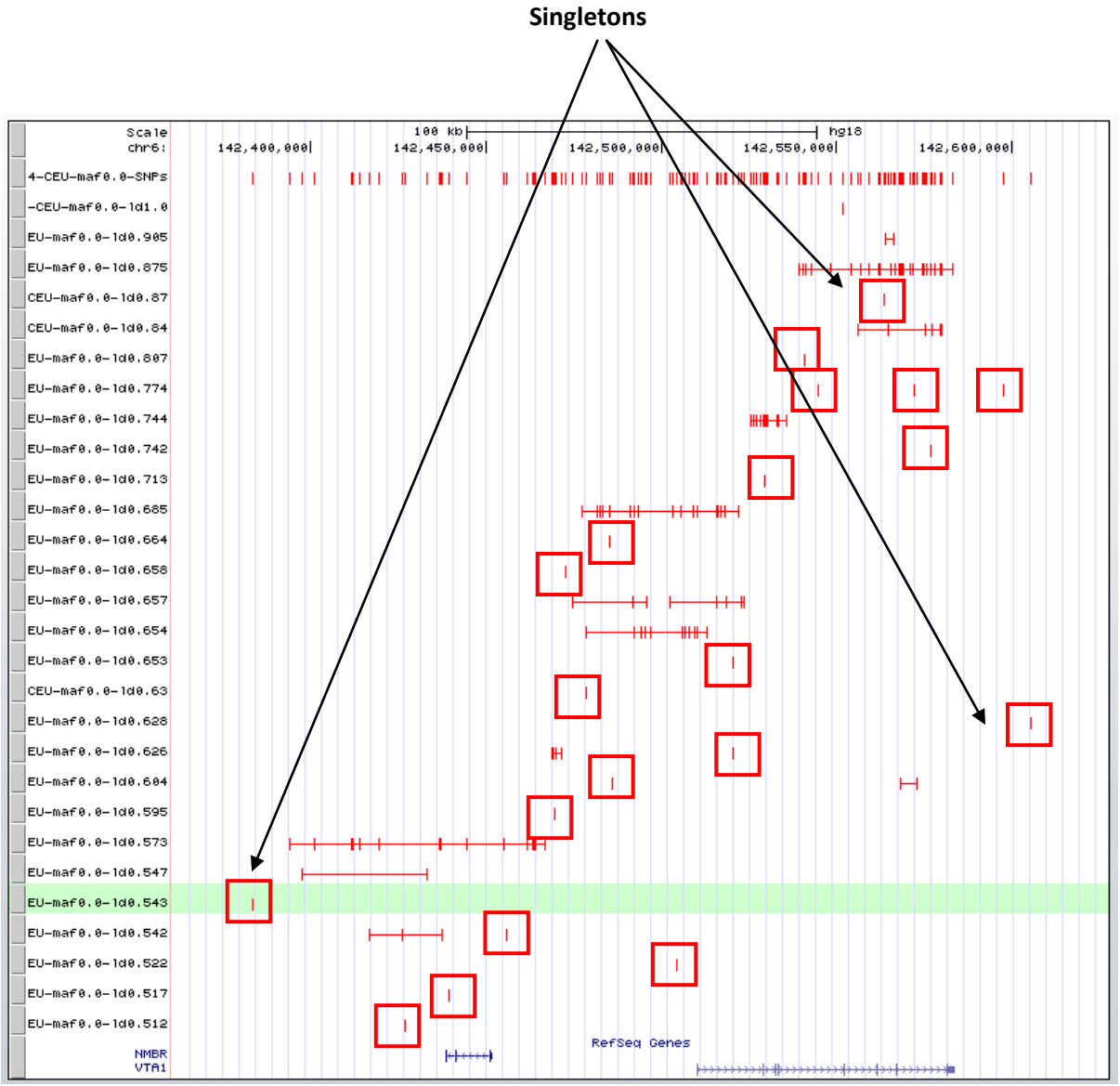
⇨ **Plot description for SNP-centered architectures**

Each vertical line represents a SNP. SNPs linked by a line are in perfect linkage (r^2=1). Each track represents a group of SNPs in LD with the reference SNP. A new track is created for each r^2 value. Clusters are ordered according to their LD with the reference SNP whose track is displayed just after the track containing SNP names. The LD between the selected track and the reference SNP is indicated in the name of the track. Individual SNP names can be

obtained by expanding the first track. To display the name of a cluster it is sufficient to change the visibility of the corresponding track to “pack”.

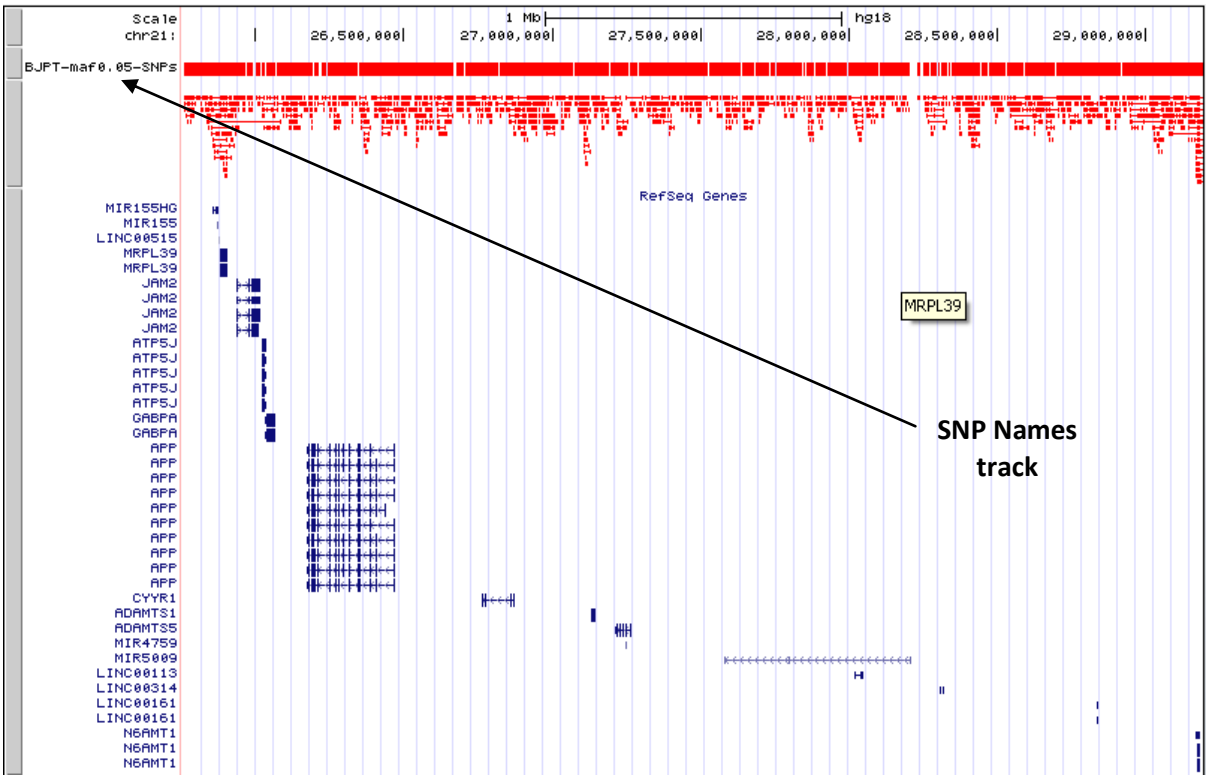


If the option **Include singletons** has been selected singletons are added to the corresponding r^2 track.

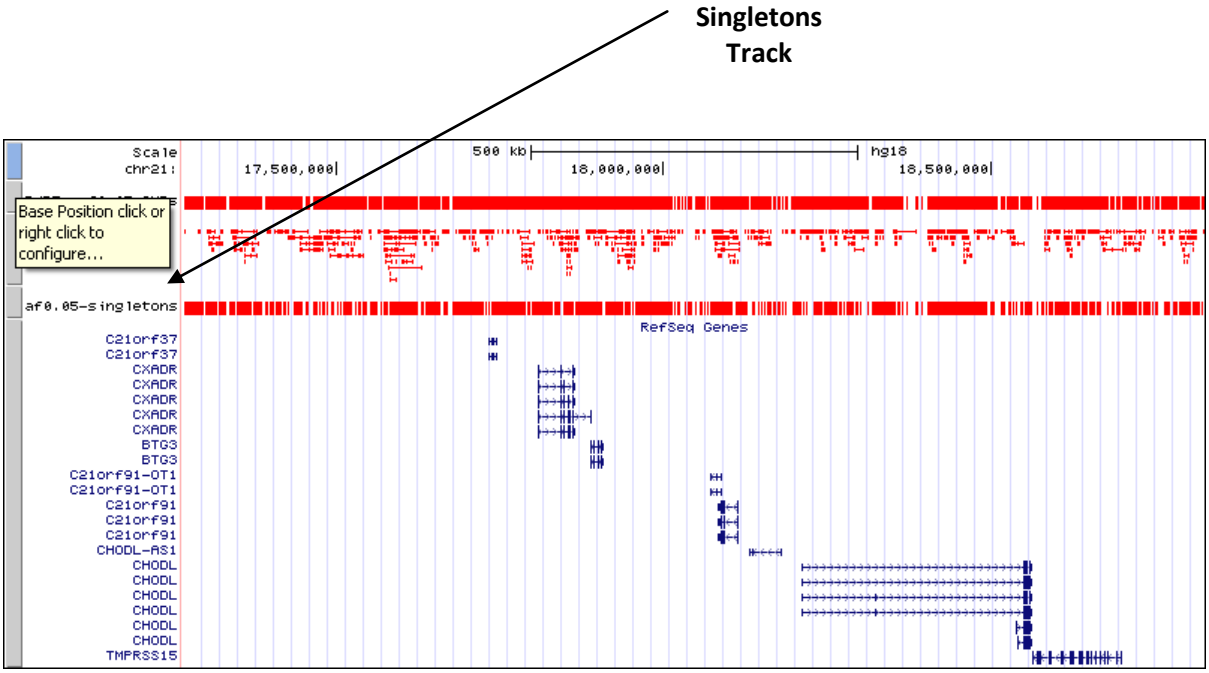


⇨ **Plot description for chromosome-centered architectures (ArchiLD1k only)**

Each vertical line represents a SNP. SNPs linked by a line are in perfect linkage. All the clusters are added to the same track. Every cluster is identified by the name of the first SNP in it. Individual SNP names can be obtained by expanding the first track.



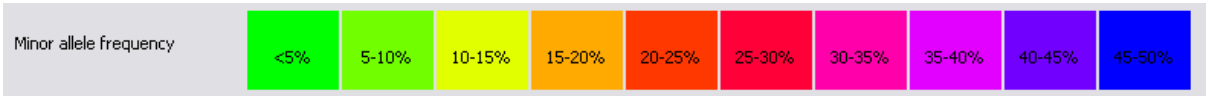
If the option Include Singletons has been selected singletons are included in a distinct track.

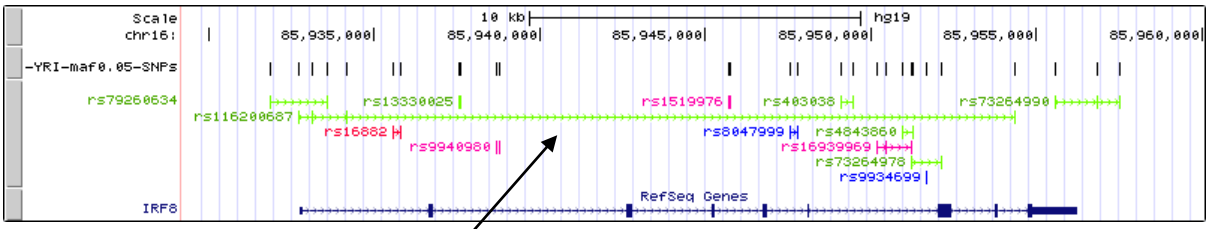


⇨ **Color by Minor Allele Frequency**

When this option is selected SNPs and LD blocks are colored according to their Minor Allele Frequency (MAF).

Below is the description of the color associated to each MAF interval





Blocks colored by Minor Allele Frequency