Correspondence to Nature Methods

Exploring long-range genome interaction data using the WashU Epigenome Browser

Xin Zhou, Rebecca F. Lowdon, Daofeng Li, Heather A Lawson, Pamela A.F. Madden, Joseph Costello, Ting Wang

Supplementary Material

Supplementary Notes

Availability

To access the public site, please visit http://epigenomegateway.wustl.edu/.

For general questions regarding the Browser, please contact user support via email at epigenome@lists.genetics.wustl.edu. For the latest development news from WashU Epigenome Browser, users can follow us on Twitter (@washugbrowser) and our blog (http://washugb.blogspot.com/).

A video tutorial is available at: http://epigenomegateway.wustl.edu/info/video/long_range_video.avi

Implementation and system requirements

The WashU Epigenome Browser is best viewed with open source web browsers including Firefox and Chromium. At this stage Microsoft IE is not supported.

WashU Epigenome Browser continues to be built upon open source technologies, including GNU C, GNU Linux, MySQL, Tabix, and NCBI-BLAST. The source code of WashU Epigenome Browser is freely available at http://epigenomegateway.wustl.edu/info/source/.

Current computational systems for long-range genomic interaction data

Visualizing long-range genome interactions has long been a challenge in the design and implementation of Genome Browsers. Public resources capable of providing such services are almost non-existent at the time of writing. A few research groups excelling in long-range interaction assays have developed their own computational tools, for both data analysis and visualization. The Dekker lab (http://my5c.umassmed.edu) developed the My5C[1] (http://my5c.umassmed.edu/my5Cprimers/5C.php) which is a sophisticated web-based platform. My5C designs primers for the target genomic region of the 5C experimental assay. It allows a user to upload raw 5C assay results, conduct a series of analyses, and visualize the results as heatmaps. Later, the group published the first Hi-C assay, which interrogated chromatin interactions genome-wide[2]. Along with it a simple web interface was created to visualize the Hi-C data as heatmaps (http://hic.umassmed.edu/heatmap/heatmap.php).

The Wei/Ruan group developed the ChIA-PET assay, which captures spatially proximal protein binding sites at genome-wide scale that are indicative of chromatin looping. The group developed the ChIA-PET Tool[3] (http://chiapet.gis.a-star.edu.sg/downloads/chia-pet-tools), which is a Java package for analyzing ChIA-PET data. It visualizes assay results via a built-in function and GBrowse (http://gmod.org/wiki/GBrowse). The group has published studies on chromatin interaction using the ChIA-PET assay[4], and data were visualized by Circos (http://circos.ca/) and the UCSC Genome Browser (http://genome.ucsc.edu/).

The Ren group (http://bioinformatics-renlab.ucsd.edu/rentrac/) has recently published a study using the Hi-C assay[5]. Along with the publication, a web interface was created to visualize the data (http://chromosome.sdsc.edu/mouse/hi-c/database.php). It represents interaction data as triangular-shaped heatmaps, which are coupled with the UCSC Genome Browser so that both interaction data and genomic annotation data can be displayed simultaneously.

No existing tools provide general solutions for long-range interaction data visualization. They either display the long-range interaction data as a generic track in a genome browser, or are hard-coded for specific projects and cannot be expanded or adapted easily. The genome browser tools including GBrowse and the UCSC Genome Browser are excellent for visualizing genomic annotations, but by displaying only one genomic region at a time, it becomes difficult to display interacting loci over large chromosomal distances, and a linear genome representation cannot properly display the inter-chromosomal interaction data. My5C's 2-dimensional heatmap representation is good solution, as it can set arbitrary genomic coordinates to the axes of the heatmap, making the display unbiased and intuitive. But it is difficult for a heatmap display to accommodate genomic annotation data, thus it remains difficult to perceive the essential biological context for understanding the interaction events.

The long-range chromatin interaction visualization function of WashU Epigenome Browser

The WashU Epigenome Browser encodes long-range interaction data in the format of two separate records for the two loci in an interaction (see next section on file format). In this way, both intra- and inter-chromosomal interaction data can be properly recorded. A locus from an interaction can be displayed on the Browser as a genomic feature, but when both interacting loci are shown, they will be connected and represented by a glyph. Five types of glyph are available:

- Thin (**Supplementary Figure 3**): thin lines that extend from one locus to the other are used to represent the interaction. No name or score is printed. It is a compact way to represent interaction data.
- Full (**Supplementary Figure 4**): a thick box is used to indicate a locus that interacts with another locus. If its mate (interacting locus) is not visible, the coordinate of the mate will be printed on the left side of the box. If the mate is visible, the mate will also be represented as a box and a straight line will connect the two boxes. The interaction score will be printed on the left of the joined pair.
- Arc (**Figure 1b** and **1c**): an arc can be drawn to connect the interacting loci when both of them are visible. All arcs have the same radian ($\pi/2$) and width (1 pixel), and the radii are determined by the on-screen distance between the two loci.
- Heatmap (**Figure 1d**): squares or rectangles are drawn at the tip of an isosceles right triangle whose longest side is the genomic distance between the two interacting loci (plus the loci themselves). If the loci have equal length, the shape of the heatmap cell will be a square, otherwise it will be a rectangle.
- Circlet (**Supplementary Figure 2**): the entire chromosomes curl into a "circlet" and arcs are drawn inside the circle to link pairs of interacting loci.

The arc and heatmap are the most representative glyphs to visualize the interactions. The arc is suitable for sparse interactions and the heatmap is suitable for dense interactions, as is usually the case for the Hi-C assays. But the heatmap plot differs from the heatmap used by My5C, as the latter requires two axes. Circlet View is especially useful at showing long-range interaction data over whole chromosomes or entire genome.

The Hi-C data can be extraordinarily dense depending on how the data is analyzed. The genome is typically divided into equally sized bins, and each bin can interact with thousands of bins across the genome. When a Hi-C track is shown using the Browser, an enormous amount of data is transmitted from the server to the client-side. The user should take caution in viewing data over large genomic regions to avoid stalling the web browser.

To switch between the glyph types, right click on the track image and press a glyph type button from the context menu. In all cases, the color of the boxes, arcs, and heatmap cells are determined by their scores with respect to a user defined maximum cutoff score. The user can interact with the track by clicking on the glyphs to get the details of the interactions. Please refer to the **Tutorial** for detailed instructions.

As mentioned above, previous genome browsers are limited to displaying only one genomic region at a time, which seriously restricts their application in long-range interaction data display. The WashU Epigenome Browser distinguishes itself with the extraordinary versatility of genomic coordinate arrangement. It is able to show data across multiple chromosomes in arbitrary order, and juxtapose data to focus the view on a subset of the genome. This provides an ideal environment for displaying long-range interaction tracks. **Figure 1** combines the long-range interaction tracks with Gene Set View to compare interaction pattern between different genomic regions. **Supplementary Figure 5** shows a Hi-C track on human K562 cells[2]. By placing chr22 next to chr9, the interaction pattern between the two chromosomes becomes visible.

File format of the long-range genome interaction data

The long-range genome interaction data is stored in a compressed tabular text file and is accessed by the server, in the same way as all other genomic annotation tracks for the WashU Epigenome Browser. Tabix[6] is used to compress, index, and query the file with fast speed and minimum system footprint. It can also be used to access remote files through the network. The user can convert his or her own long-range interaction data into a tabix-formatted file, place the file on a web server, and visualize it on the WashU Epigenome Browser. This procedure does not require file upload, and it is highly efficient and secure.

The tabular text file has a format similar as the BED format, albeit with variations. It has following six required fields:

1. Chromosome name.
2. Start coordinate of the locus.
3. Stop coordinate of the locus.
4. The coordinate of this locus' interacting mate and the score of the interaction, in the format of "chr1:567-890,3.14", where "chr1:567-890" is the mate's coordinate, and "3.14" is the score. They are joined by a comma.
5. Unique integer.
6. Strand indicator. If the mate is on a different chromosome, use dot '.'. If the mate is downstream of the locus, use '+', else '-'.

Two records are needed to represent a pair of interacting loci. A unique integer must be assigned for each record in the fifth field. This format is versatile enough to encode arbitrary pairwise interaction data (intra-chromosomal or inter-chromosomal). While the nature of long-range interactions usually involves many loci, current assay methods all yield pairwise interaction data so this encoding approach works fine. In future developments, the format could be expanded by including additional loci in the fourth field, so that new generations of interaction data can be encoded.

Another possible solution is the SAM format[7], which is widely used to store paired-end read data from DNA sequencing experiments. Pairs of interacting loci could be treated as "paired reads" and thus be encoded in a SAM format file. However, compared with the custom format used by WashU Epigenome Browser, the disadvantage of SAM format is that it identifies paired reads only by their shared identifier, and it does not store locations of both reads in either one of the reads. This potentially requires the program to scan through the entire file (in the worst case) to obtain the location of the mate locus, or else this information would be missing. Besides, it is not possible to extend the SAM format for reads that have more than two segments.

We have produced a blog article describing the details of how to create a custom long-range interaction track and display it on Wash U Epigenome Browser. The link to the article is:

http://washugb.blogspot.com/2012/09/prepare-custom-long-range-interaction.html

This post also included a link to a sample Python script to convert a publicly available ChIA-PET data file into Wash U Epigenome Browser long-range interaction track format.

Supplementary Methods

Here we provide additional information about **Figure 1**. The **Tutorial** provides a guide on how to recreate this figure. Following is the information on all the histone modification tracks in the genome heatmap. The displayed data is aligned to human reference genome hg19/GRCh37. Track name, track color, and GEO accession are given for all tracks:

1. H3K4me3 of IMR90, green, GSM469970
2. H3K4me3 of IMR90, green, GSM521901
3. H3K4me1 of IMR90, teal, GSM521895
4. H3K4me1 of IMR90, teal, GSM521897
5. H3K9me3 of IMR90, red, GSM521909
6. H3K9me3 of IMR90, red, GSM521913
7. H3K27me3 of IMR90, fuchsia, GSM469968
8. H3K27me3 of IMR90, fuchsia, GSM521889
9. H3K4me3 of K562, green, GSM608165
10. H3K4me3 of K562, green, GSM945297
11. H3K4me1 of K562, teal, GSM788085
12. H3K4me1 of K562, teal, GSM733692
13. H3K9me3 of K562, red, GSM607494
14. H3K9me3 of K562, red, GSM733776
15. H3K27me3 of K562, fuchsia, GSM945228 (replication 1)
16. H3K27me3 of K562, fuchsia, GSM945228 (replication 2)

The gene model track is RefSeq genes, and the data were obtained from UCSC Genome Browser website (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/).

The two ChIA-PET tracks were generated by ENCODE Consortium and can be downloaded from here: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGisChiaPet/. The name of the first track is “ENCODE GIS ChIA-PET on K562, POL2, replicate 1”, the second is “ENCODE GIS ChIA-PET on K562, CTCF, replicate 1”.

The Hi-C track is described by Dixon, J.R. et al.[5] and can be downloaded here: http://chromosome.sdsc.edu/mouse/hi-c/download.html. The bin size is 40 kilobyte, the restriction enzyme is HindIII, with two biological replicates combined.

Acknowledgement

We would like to thank the laboratories of Bing Ren, Chia-Lin Wei, Yijun Ruan, Job Dekker, the UCSC Genome Browser team and the ENCODE Consortium for providing long-range interaction data sets, and the Roadmap Epigenomics Consortium for providing epigenomics data sets. We thank many collaborators in the Roadmap Epigenomics Consortium who participated in software testing and provided helpful comments. We acknowledge support from NIH Roadmap Epigenomics Program, sponsored by the National Institute on Drug Abuse (NIDA) and the National Institute of Environmental Health Sciences (NIEHS). J.F.C., M.H. and T.W. are supported by NIH grant 5U01ES017154. X.Z. is supported by NIDA’s R25 program DA027995. T.W. is supported in part by the March of Dimes Foundation, the Edward Jr. Mallinckrodt Foundation, P50CA134254 and a generous start-up package provided by Washington University School of Medicine.

References

1. Lajoie, B.R. et al. Nat Methods 6(10), 690-619 (2009).
2. Aiden, E.L. et al. Science 326, 289-293 (2009).
3. Li, G. et al. Genome Biology 11:R22 (2010).
4. Handoko, L. et al. Nature Genetics 43, 630-638 (2011).
5. Dixon, J.R. et al. Nature 485, 376-380 (2012).
6. Li, H. Bioinformatics 27, 718-719 (2011)
7. Li, H. et al. Bioinformatics 25, 2078-9 (2009)

Supplementary Figure Legends

Supplementary Figure 1: The long-range interaction companion panel. A companion panel displays data over a locus (chr9:123603325-123608566, marked by light yellow background) that interacts with another locus displayed in the main Browser panel (chr9:132644414-132651364, marked by light blue background). Their connection was predicted in K562 cells by a ChIA-PET experiment (track name is "ENCODE GIS ChIA-PET K562 POL2 rep1", shown in "full" mode). A schematic representation of the two loci on the chromosome can be found in the top part of the companion panel. The order and settings of the tracks in the interaction panel are identical with those of the main panel. Through this interaction, the promoters of *PSMD5* and *LOC253039* are in contact with the 3' end of *USP20* and *FNBP1*. In the genome heatmap there are histone modification tracks and RNA-Seq tracks on the K562 cells. The histone tracks are the same as those displayed in **Figure 1**. In the metadata color map, histone mark types from top to bottom are: H3K4me3, H3K4me1, H3K9me3, H3K27me3. The RNA-Seq tracks contain strand-specific data. The top 6 tracks (in blue color) are transcription from the reverse strand, the bottom 6 tracks (in indigo color) are from forward strand, and they come from these GEO accessions: GSM765393, GSM767849, GSM765392, GSM765405, GSM765390, GSM758577. Comparison of track data patterns of the interaction panel with the main panel reveals that the two loci are both enriched in active marks and contain medium or low level of repressive marks. Consistent with their epigenetic status, genes that are involved in the interaction including *PSMD5*, *LOC253039*, and *USP20* are actively transcribed (high RNA-Seq signal over the gene body regions). *PSMD5* (proteasome 26S subunit) and *USP20* (ubiquitin specific peptidase 20) are functionally related, while *LOC253039* is uncharacterized.

Supplementary Figure 2: The Circlet View. Circlet View shows the Hi-C data (K562 cells, Pearson correlation of 1 MB bins)[2] over five genes on the whole-genome scale. Arcs connect interacting loci. Arcs in magenta represent positively correlated loci (Pearson correlation coefficient >0), and those in blue represents negatively correlated loci (Pearson correlation coefficient <0). The gene names are added to the screen shot to mark the position of the genes. It shows that *CYP4Z1* (chr1), *CYP4V2* (chr4), *CYP2C19* (chr10), and *CYP26B1* (chr2) are each abundant with positive intra-chromosomal correlations, and are generally negatively correlated with external chromosomes. This pattern is not observed with *CYP3A4* (chr7), which shows a distinct mixture of positive and negative correlation, both within and between chromosomes.

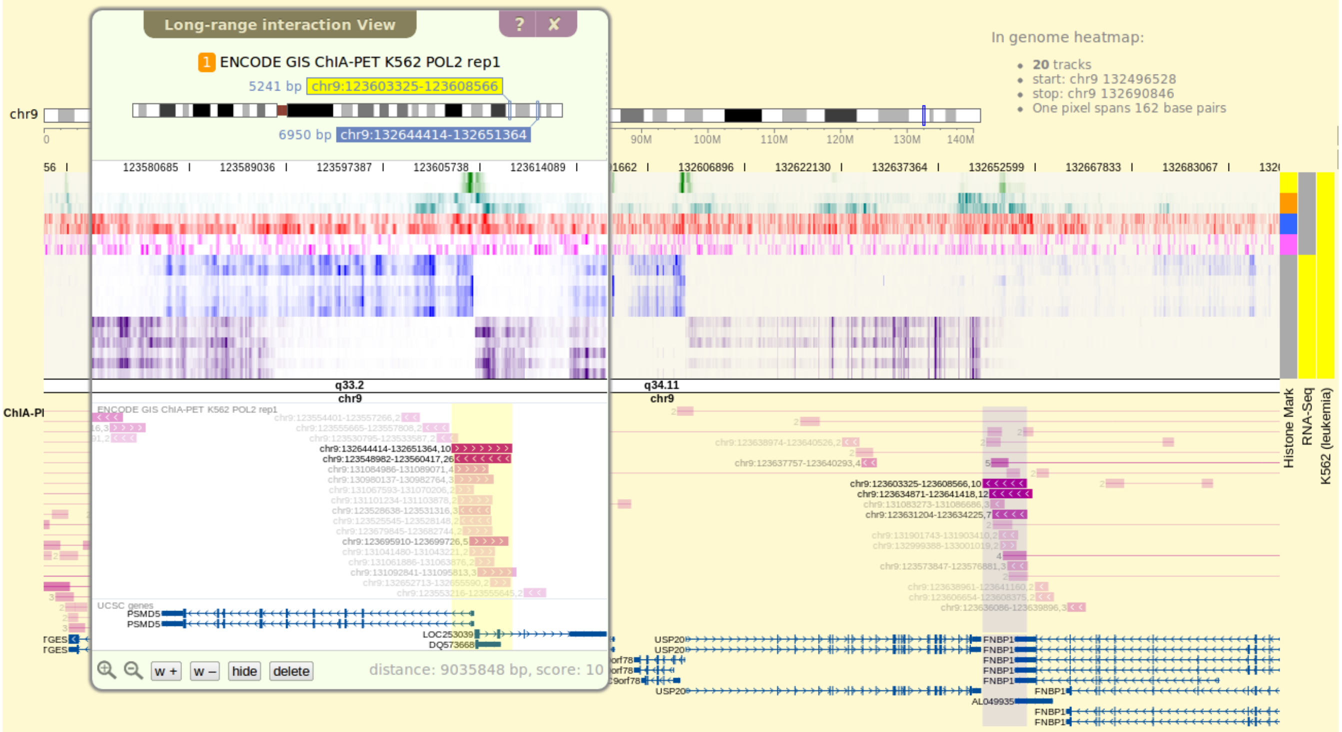
Supplementary Figure 3: "Thin" glyph style for long-range interaction track display. This screen shot replicates **Figure 1**, with only one long-range interaction track (**Figure 1b**) to show the "thin" glyph style.

Supplementary Figure 4: "Full" glyph style for long-range interaction track display. This screen shot replicates **Supplementary Figure 3**, and the long-range interaction track is shown with "full" glyphs.

Supplementary Figure 5: Intra- and inter-chromosomal interaction pattern as revealed by a permuted arrangement of chromosomes. A Hi-C track[2] (K562 cells, Pearson correlation of 1 MB bins) is displayed as a heatmap. Chromosomes 9 and 22 are placed side by side. Both chromosomes show strong intra-chromosomal contact. The inter-molecular contact between them is generally weak. A heatmap cell stands out with strong positive correlation between chr9 and chr22 (pointed by the arrow, the two bins are chr9:134010179-135010178 and chr22:22670000-23669999, correlation coefficient is 0.7). However it might be due to the genomic translocation in K562 cells instead of actual inter-chromosomal interaction.

Supplementary Figures

Supplementary Figure 1.



Long-range interaction View

1 ENCODE GIS ChIA-PET K562 POL2 rep1

5241 bp chr9:123603325-123608566

6950 bp chr9:132644414-132651364

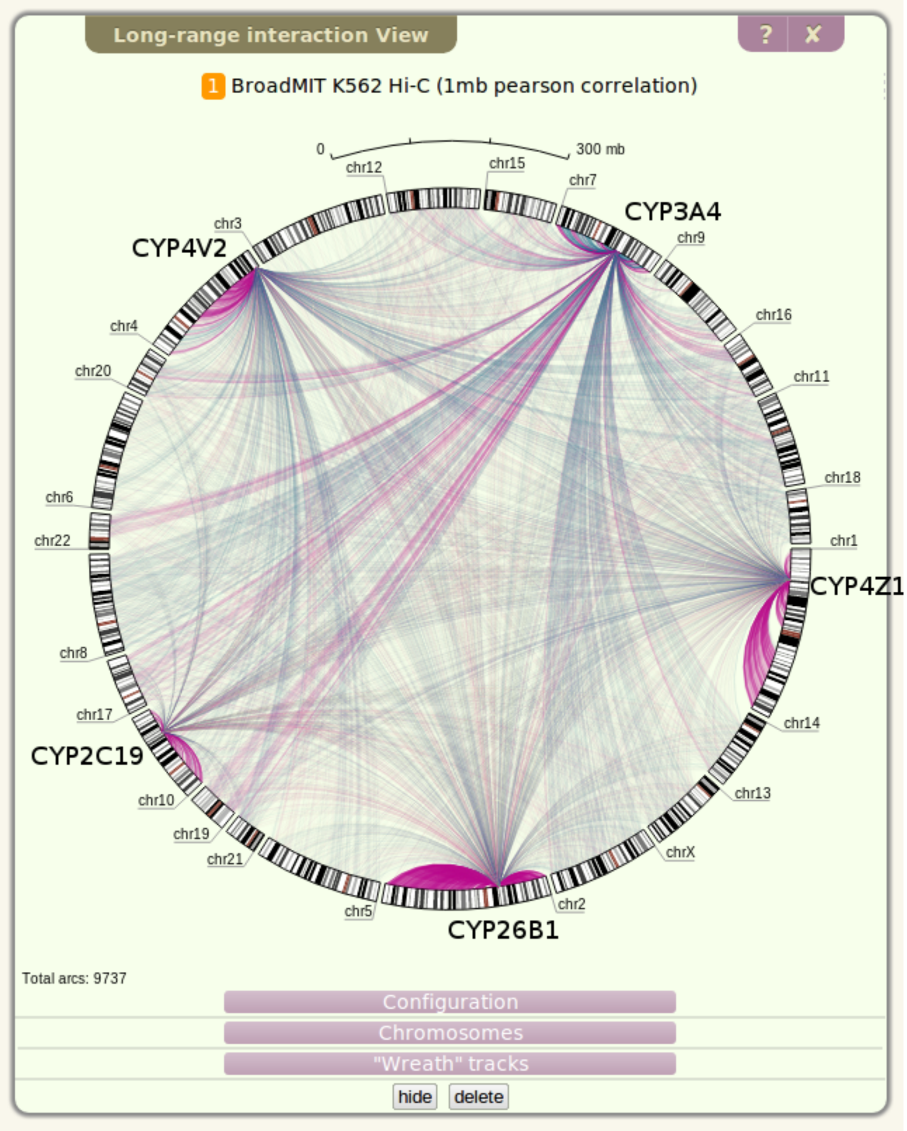
distance: 9035848 bp, score: 10

In genome heatmap:

- **20** tracks
- start: chr9 132496528
- stop: chr9 132690846
- One pixel spans 162 base pairs

?
X
chr9
0
90M
100M
110M
120M
130M
140M
123580685
123589036
123597387
123605738
123614089
132606896
132622130
132637364
132652599
132667833
132683067
q33.2
chr9
q34.11
chr9
ChIA-PE
ENCODE GIS ChIA-PET K562 POL2 rep1
chr9:132644414-132651364,10
chr9:123548982-123560417,26
chr9:131084986-131089071,4
chr9:130980137-130982764,3
chr9:123695910-123699726,5
chr9:131092841-131095813,3
chr9:123603325-123608566,10
chr9:123634871-123641418,12
chr9:123631204-123634225,7
chr9:123637757-123640293,4
chr9:123573847-123576881,3
UCSC genes
PSMD5
PSMD5
LOC253039
DQ573668
USP20
USP20
USP20
FNBP1
FNBP1
AL049935
Histone Mark
RNA-Seq
K562 (leukemia)
w +
w –
hide
delete

Supplementary Figure 2.

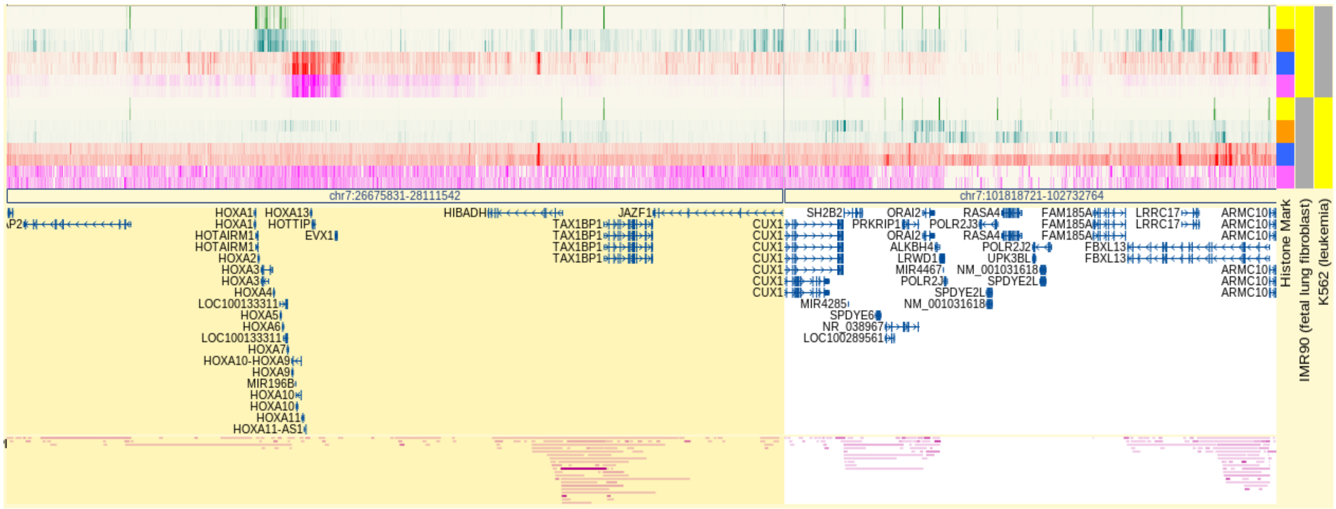


Long-range interaction View

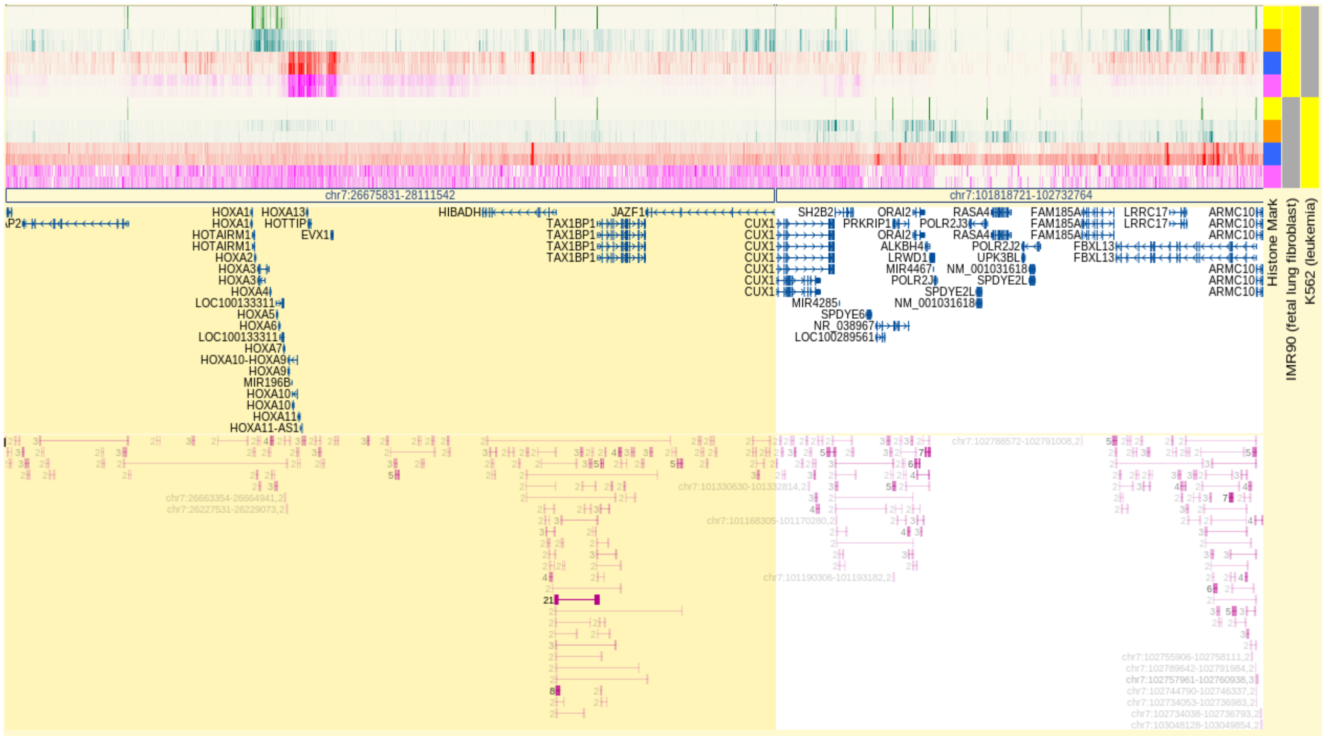
1 BroadMIT K562 Hi-C (1mb pearson correlation)

Total arcs: 9737

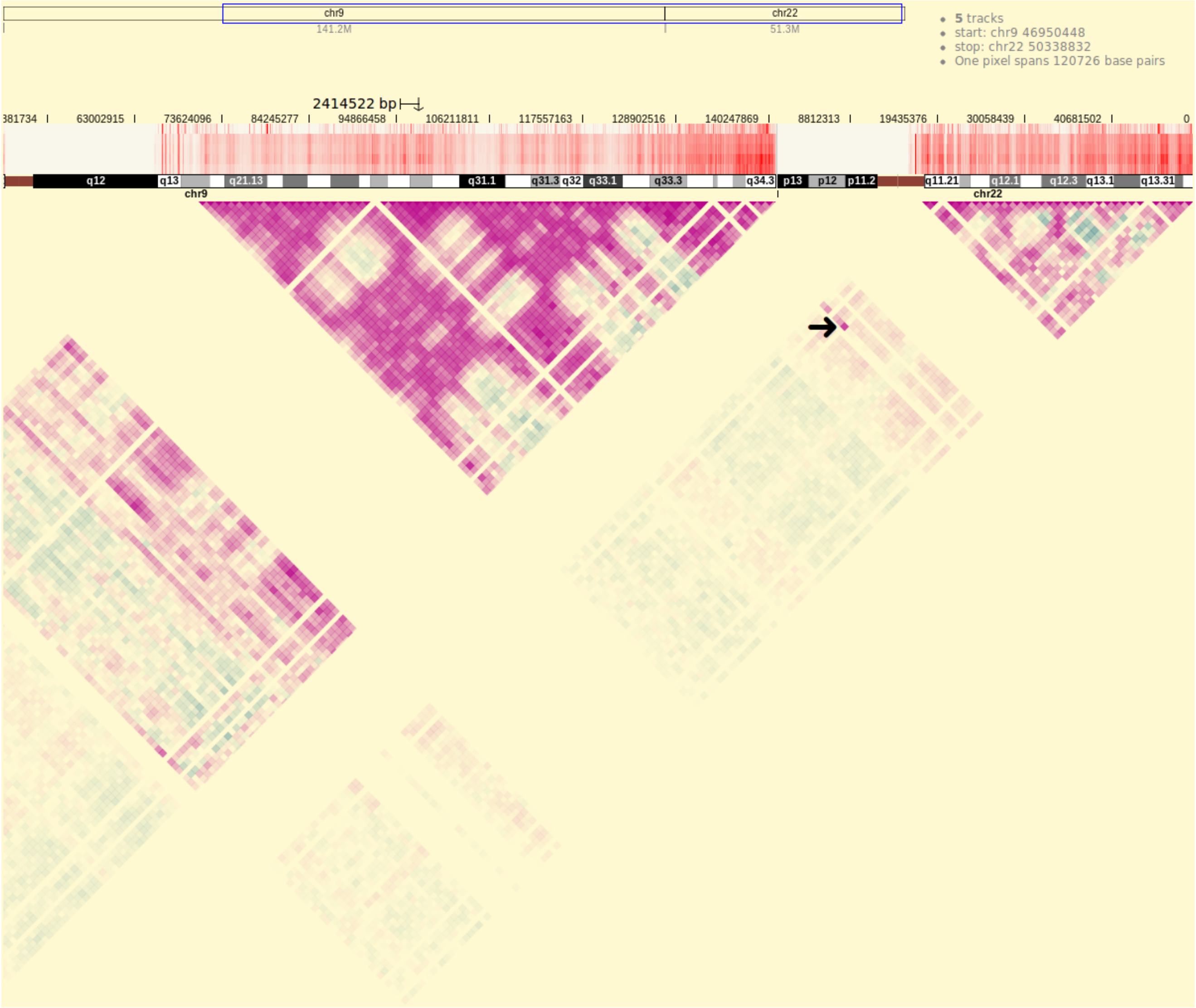
Supplementary Figure 3.



Supplementary Figure 4.



Supplementary Figure 5.



- 5 tracks
- start: chr9 46950448
- stop: chr22 50338832
- One pixel spans 120726 base pairs