Visualisation tools for next-generation sequencing

Simon Anders

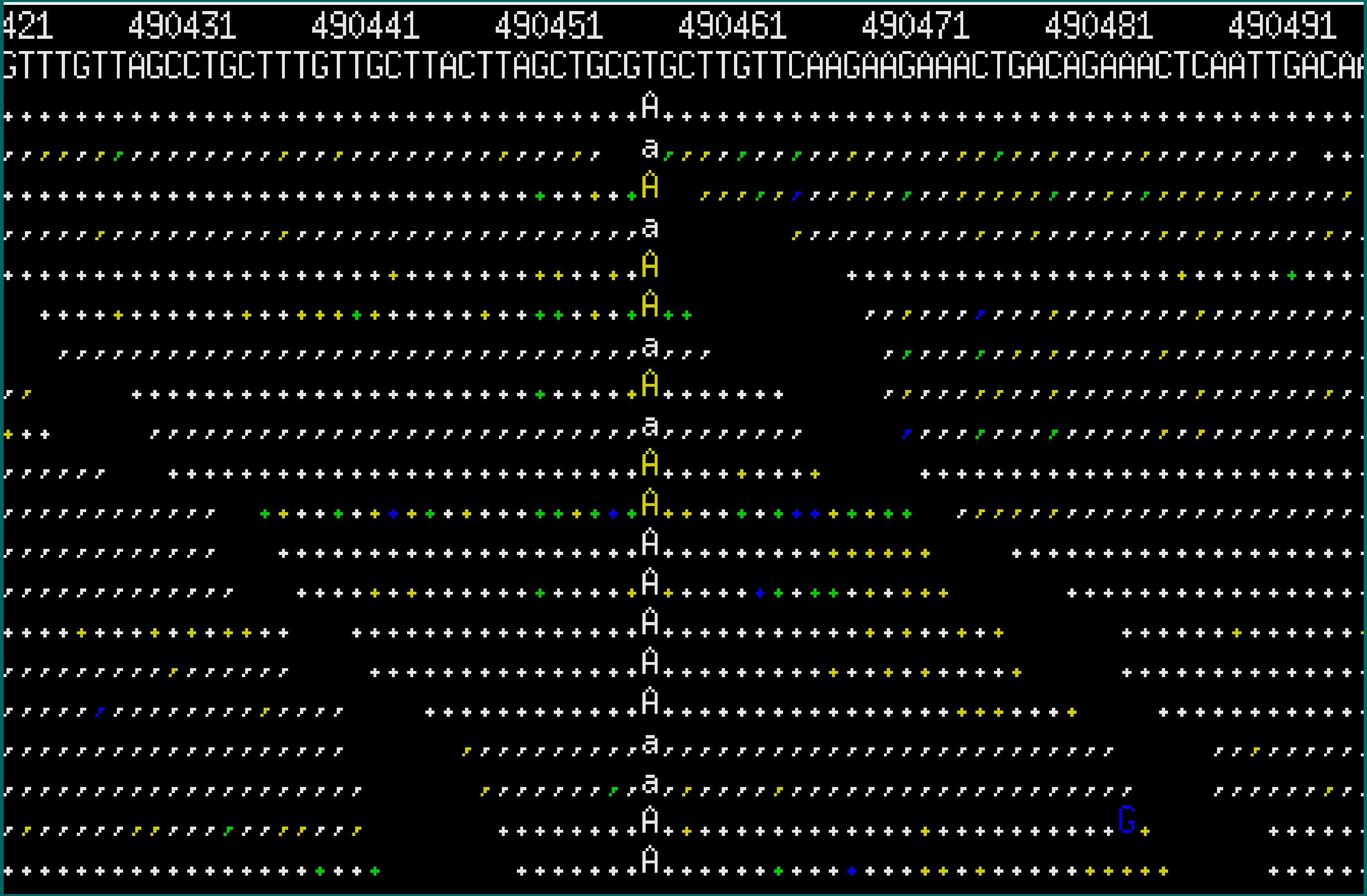
EMBL-EBI

Outline

- Exploring and checking alignment with alignment viewers
- Using genome browsers
- Getting an overview over the whole data with Hilbert curve visualization
- Displaying peaks alongside feature annotation with the "GenomeGraph" package

EMBL-EBI

Alignment Viewers: SAMtools tview

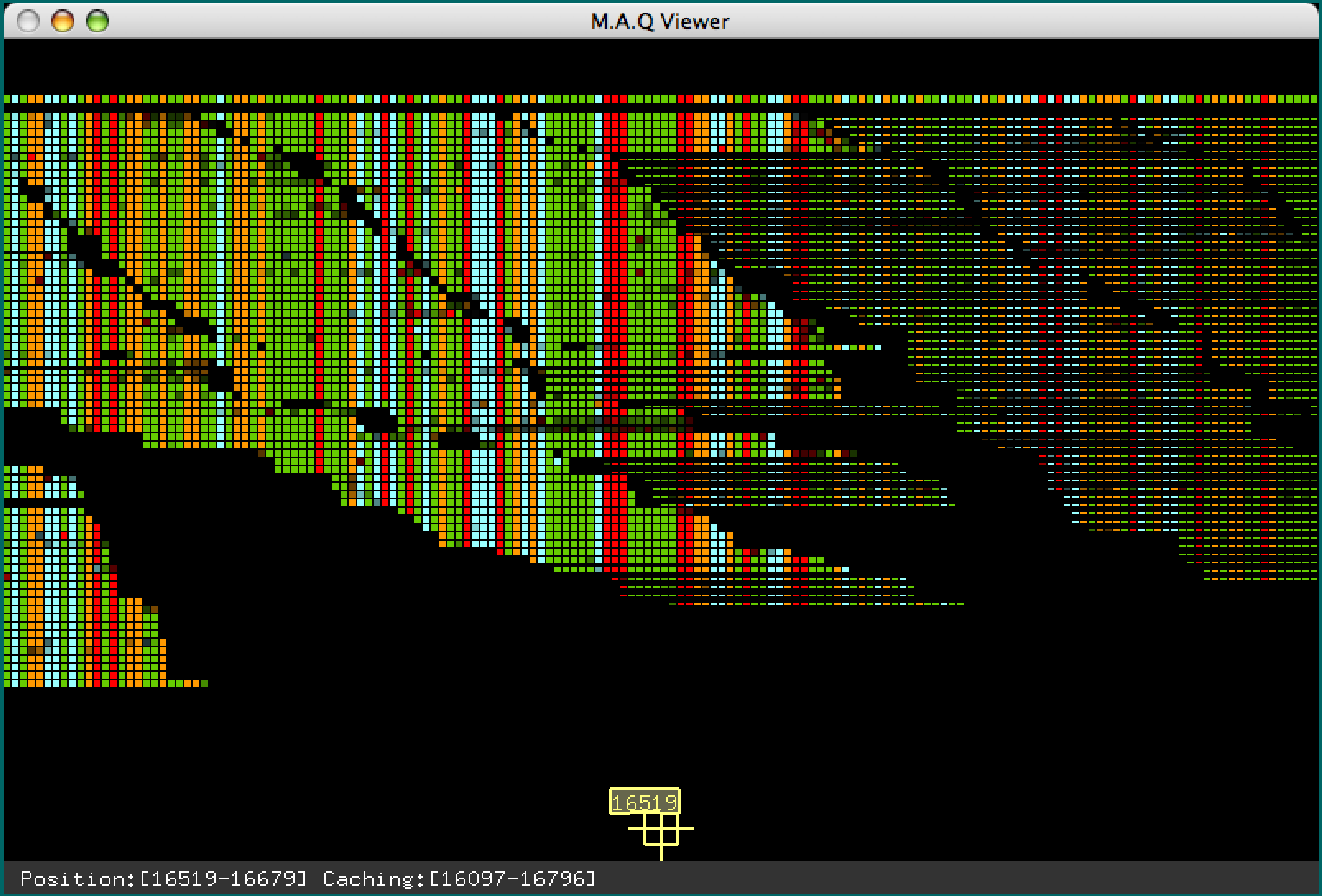


```
421    490431    490441    490451    490461    490471    490481    490491
GTTTGTTAGCCTGCTTTGTTGCTTACTTAGCTGCGTGCTTGTTCAAGAAGAAACTGACAGAAACTCAATTGACAA
...................................A.......................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ...
...................................A  ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, 
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a       ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
...................................A          .............................
  .................................A..         ,,,,,,,,,,,,,,,,,,,,,,,,,,,,
   ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,         ,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,     ............................A.......     ,,,,,,,,,,,,,,,,,,,,,,,,,,,
...     ,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,     ,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,   ..........................A.........     .........................
,,,,,,,,,,,,  .....................A..............  ,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,   ....................A...............    ....................
,,,,,,,,,,,,,   ...................A................      .................
................   ................A...................      ..............
,,,,,,,,,,,,,,,,    ...............A....................     ..............
,,,,,,,,,,,,,,,,,,,    ............A.......................    ............
,,,,,,,,,,,,,,,,,,,      ,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,     ,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,      ,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,    ,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,       ........A.........................G.      ......
.....................       .......A............................     ......
```

Heng Li (Sanger Institute) et al.

EMBL-EBI

Alignment viewers: MaqView

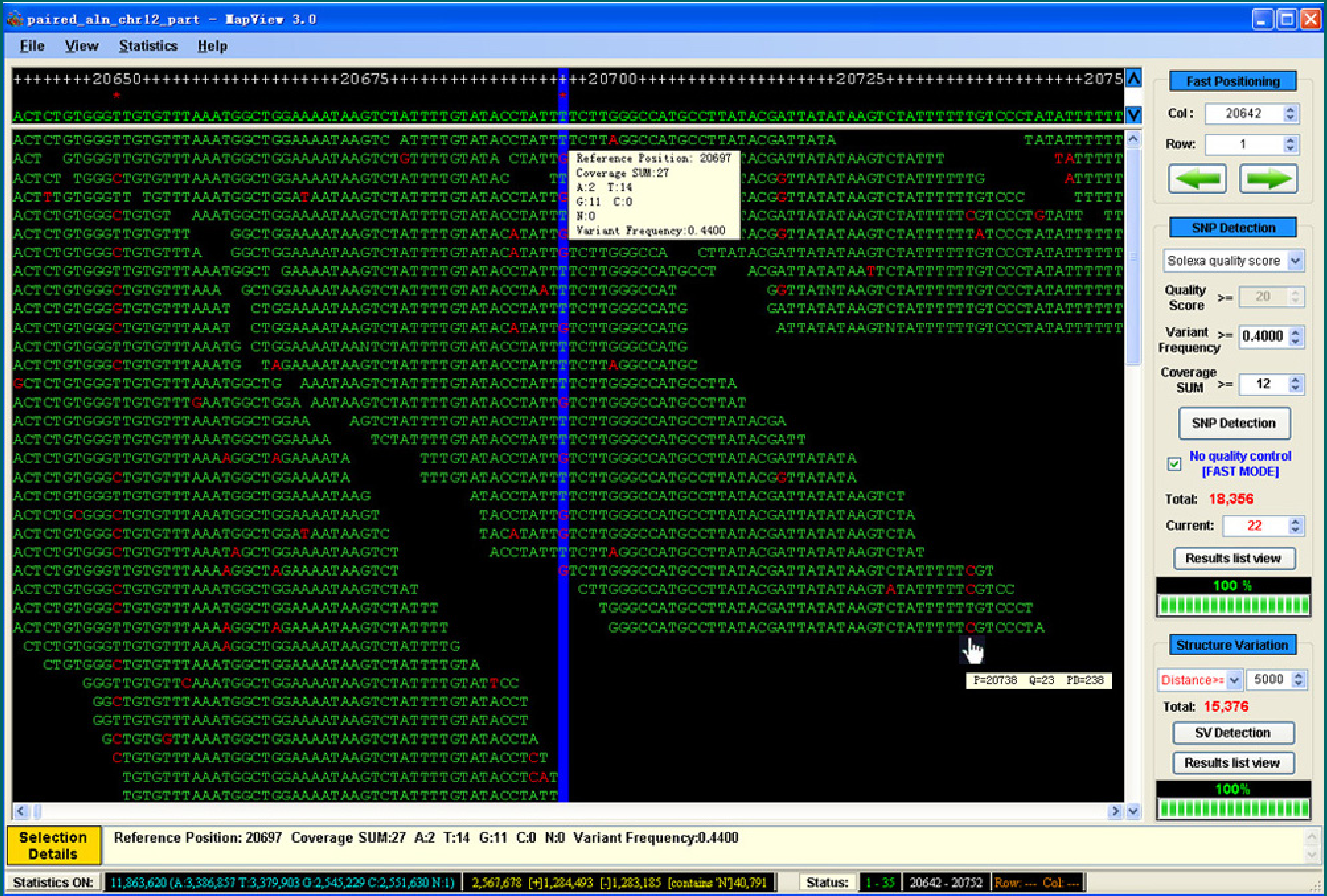


```
Position:[16519-16679] Caching:[16097-16796]
```

Jue Ruan (Beijing Genomics Institute) et al.

EMBL-EBI

Alignment Viewer: MapView



```
++++++++20650++++++++++++++++++++++20675++++++++++++++++++++20700++++++++++++++++++++20725++++++++++++++++++++2075
ACTCTGTGGGTTGTGTTTAAATGGCTGGAAAATAAGTCTATTTTGTATACCTATTTTCTTGGGCCATGCCTTATACGATTATATAAGTCTATTTTTTGTCCCTATATTTTTT
```

```
Reference Position: 20697
Coverage SUM:27
A:2  T:14
G:11  C:0
N:0
Variant Frequency:0.4400
```

```
P=20738  Q=23  PD=238
```

Col: 20642
Row: 1

Solexa quality score
Quality Score >= 20
Variant Frequency >= 0.4000
Coverage SUM >= 12
☑ No quality control [FAST MODE]
Total: 18,356
Current: 22
100 %

Distance>= 5000
Total: 15,376
100%

Reference Position: 20697 Coverage SUM:27 A:2 T:14 G:11 C:0 N:0 Variant Frequency:0.4400

Statistics ON: 11,863,620 (A:3,386,857 T:3,379,903 G:2,545,229 C:2,551,630 N:1) | 2,567,678 [+]1,284,493 [-]1,283,185 [contains 'N']40,791 | Status: 1 - 35 | 20642 - 20752 | Row: --- Col: ---

EMBL-EBI

Hua Bao (Sun Yat-Sen University, Guangzhou) et al.

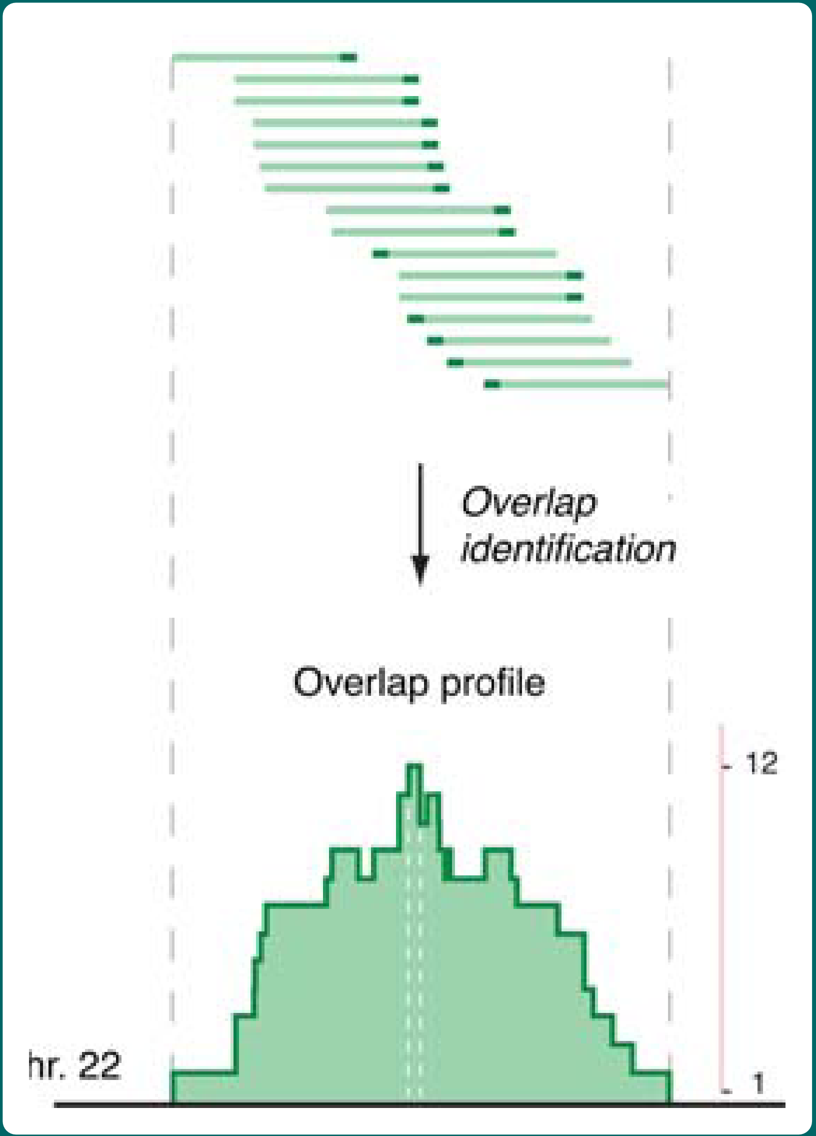
SAMtools pileup format

```
I      25514  G      G      42     0      25     5      ....^:.                 CCCCC
I      25515  T      T      42     0      25     5      .....                   CC?CC
I      25516  A      G      48     48     25     7      GGGGG^:G^:g             CCCCCC5
I      25517  G      G      51     0      25     8      .......^:,              CCCCCC1?
I      25518  T      T      60     0      25     11     .......,,^:.^:,^:,      CCCCCC3A<:;
I      25519  T      T      60     0      25     11     .......,,.,,            CCCCCC>A@AA
I      25520  G      G      60     0      25     11     .......,,.,,            CCCACC>A@<A
I      25521  T      T      60     0      25     11     .......,,.,,            CCCCCC?ACAA
I      25522  A      A      60     0      25     11     .......,,.,,            CCCCCC>ACAA
I      25523  A      A      72     0      25     15     .......,,.,,^:.^:,^:,^:.   CCCCCC;ACAAC??C
I      25524  C      C      72     0      25     15     .......,,.,,.,,.        CCCCCC6<<A?C=9C
I      25525  C      C      56     0      24     18     .......,,.,,.,,.^:,^!.^:T  CCCCCC>ACA?C=AC<
I      25526  A      A      81     0      24     18     .......,,.,,.,,.,..     CCCCCC>ACAACAACA
I      25527  A      A      56     0      24     18     .......,,.,,.,,.,.G     CCCCCC?ACAA@A?CA
```

Fields: chromosome, position, reference base, consensus base, consensus quality, SNP quality, maximum mapping quality, coverage, base pile-up, base quality pile-up

EMBL-EBI

Coverage vectors



<-- Solexa reads,
aligned to genome

<-- coverage vector

Figure taken from Zhang et al., PLoS Comp. Biol. 2008

EMBL-EBI

Coverage vectors

- A coverage vector (or "pile-up" vector) is an integer vector with on element per base pair in a chromosome, tallying the number of reads (or fragments) mapping onto each base pair.
- It is the essential intermediate data type in assays like ChIP-Seq or RNA-Seq
- Visualising coverage vectors is non-trivial, but essential for
  - quality control
  - hypothesis forming
  - etc.

EMBL-EBI

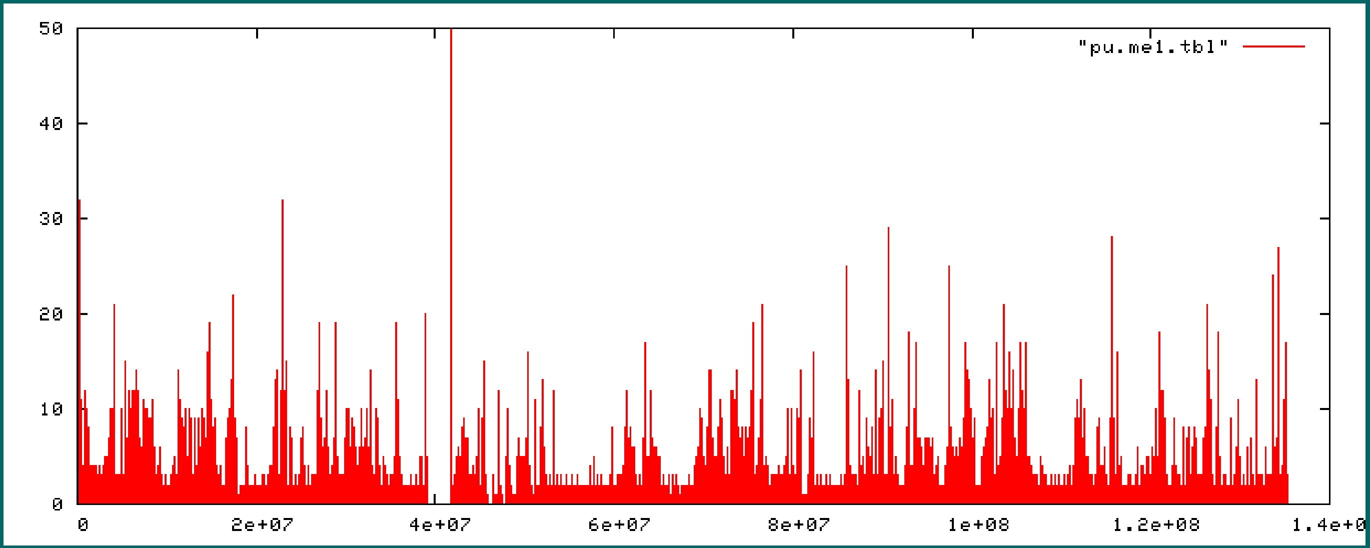
Example: Histone modifications

- Barski et al. (Cell, 2007) have studied histone modification in the human genome with ChIP-Seq
- I use their data for H3K4me1 and H3K4me3 as example data.

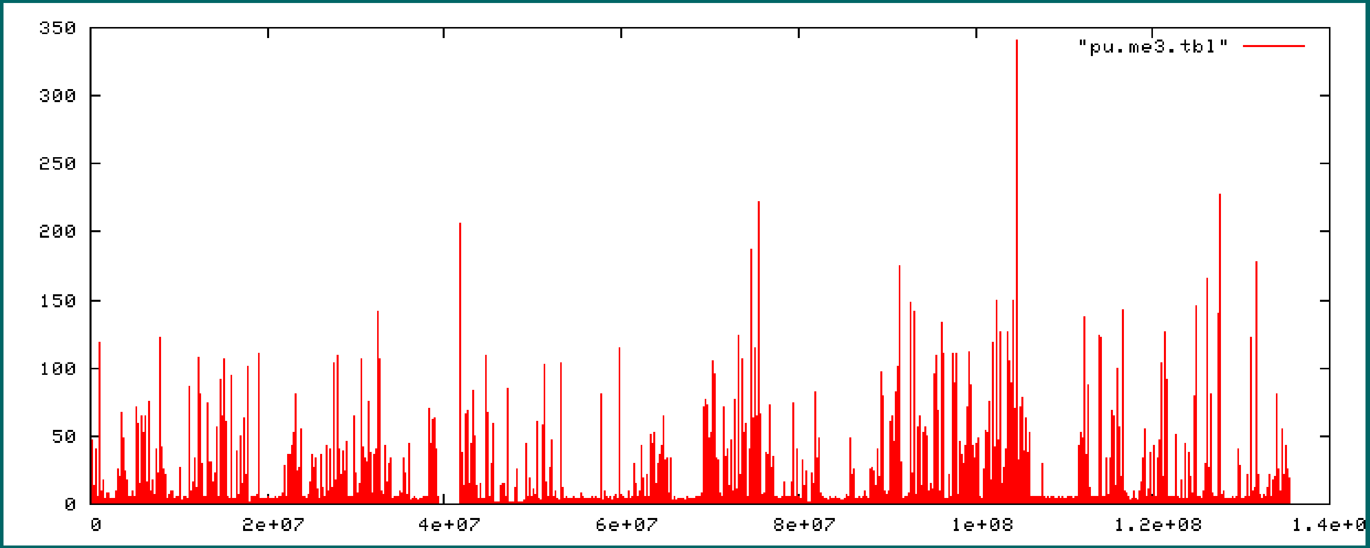
- (Each data set is from two or three Solexa lanes)

Coverage vector for a full chromosome (chr10)

H3K4me1



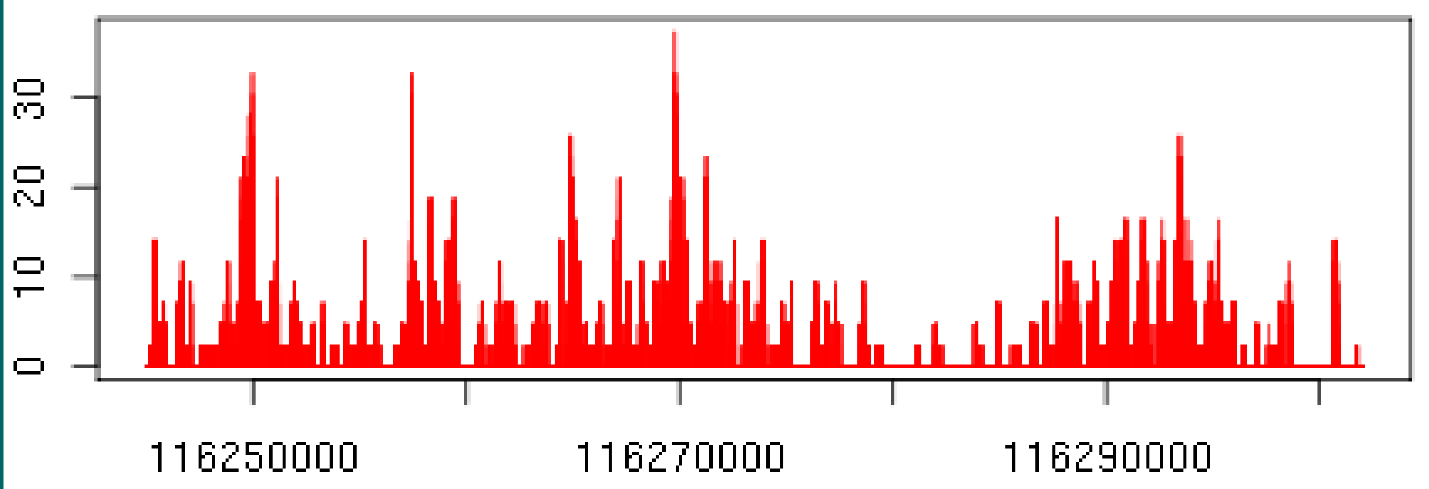
H3K4me3



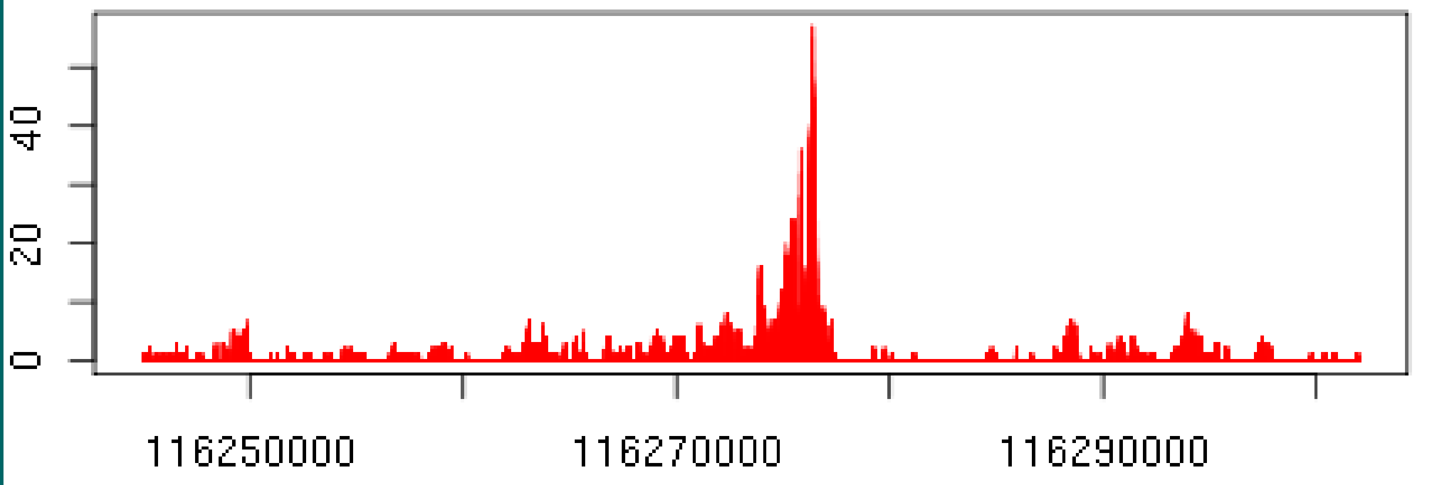
EMBL-EBI

Zoom in

H3K4me1



H3K4me3



EMBL-EBI

Genome browser tracks

Tracks may contain

- Features (intervals with name)
  - without score
  - with score
- vectors (continuously varying score)

Standard formats for genome browser tracks

- BED
- GFF
- Wiggle fixedStep and variableStep

EMBL-EBI

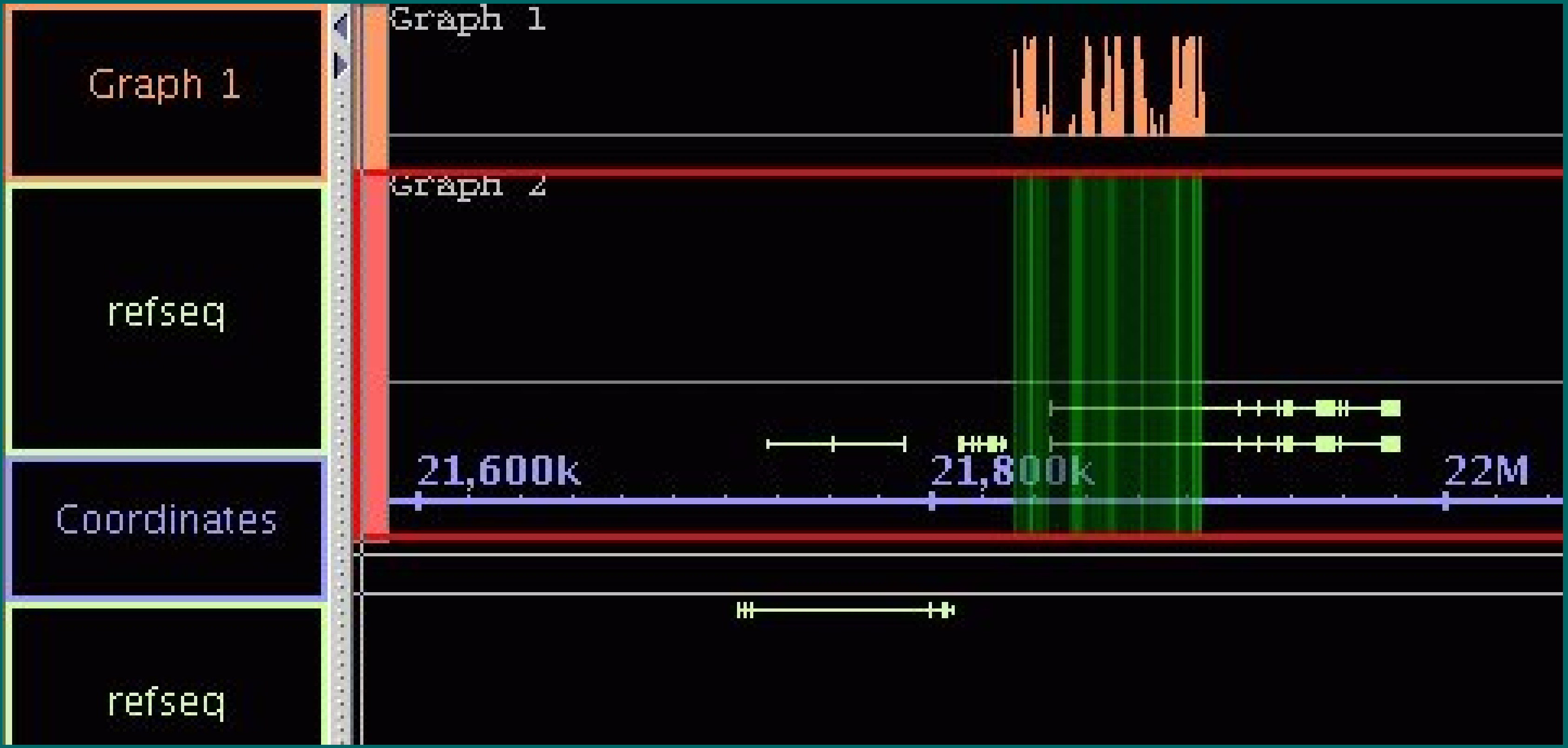
Displaying tracks alongside annotation

- Either, upload your track file to a web-base browser
  - UCSC genome browse
  - Ensembl genome browser
- or use a stand-alone browser on your desktop computer
  - Integrated Genome Browser (IGB) [Genoviz]
  - Argo Genome Browser [Broad Institute]
  - Artemis [Sanger Institute]

Displaying large amounts of data requires patience and lots of RAM. Not all tools handle it well.

EMBL-EBI

IGB


Graph 1
refseq
Coordinates
refseq
Graph 1
Graph 2
21,600k
21,800k
22M

EMBL-EBI

rtracklayer

rtracklayer: Bioconductor package by M. Lawrence (FHCRC)

- import and export BED, Wiggle, and GFF files
- manipulate track data and get sub-views
- directly interact with a genome browser (UCSC or Argo) to drive displaying of track data

Difference between the track formats

- Formats for feature-by-feature data:
    - BED
    - GFF
- Formats for base-by-base scores
    - Wiggle
    - BedGraph

- Wiggle has three sub-types:
    - [BED-like]
    - variableStep
    - fixedStep

Wiggle format: variableStep and fixedStep

```
browser position chr19:59304200-59310700
browser hide all
track type=wiggle_0 name="varStepTrack" description="varStep example" \
    visibility=full autoScale=off viewLimits=0.0:25.0 color=50,150,255 \
    yLineMark=11.76 yLineOnOff=on priority=10
variableStep chrom=chr19 span=150
59304701 10.0
59304901 12.5
59305401 15.0
59305601 17.5
59305901 20.0
59306081 17.5
59306301 15.0
59307871 10.0
track type=wiggle_0 name="fixedStepTrack" description="fixedStep examle"
fixedStep chrom=chr19 start=59307401 step=300 span=200
1000
 900
 800
 700
 600
 500
 400
 300
 200
 100
```

All coordinates 1-based!

EMBL-EBI

bedGraph format

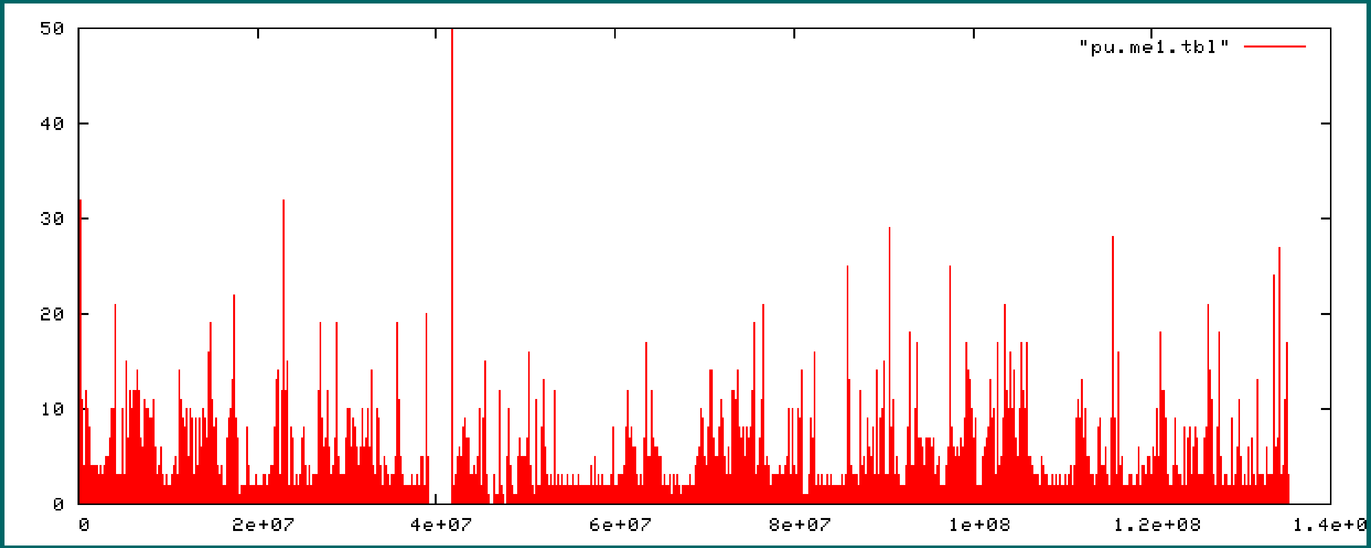
```
track type=bedGraph name="BedGraph Track"
chr19 59302000 59302300 -1.0
chr19 59302300 59302600 -0.75
chr19 59302600 59302900 -0.50
chr19 59302900 59303200 -0.25
chr19 59303200 59303500 0.0
chr19 59303500 59303800 0.25
chr19 59303800 59304100 0.50
chr19 59304100 59304400 0.75
chr19 59304400 59304700 1.00
```

All coordinates 0-based, half-open!

Specs: See UCSC Genome Browser web site

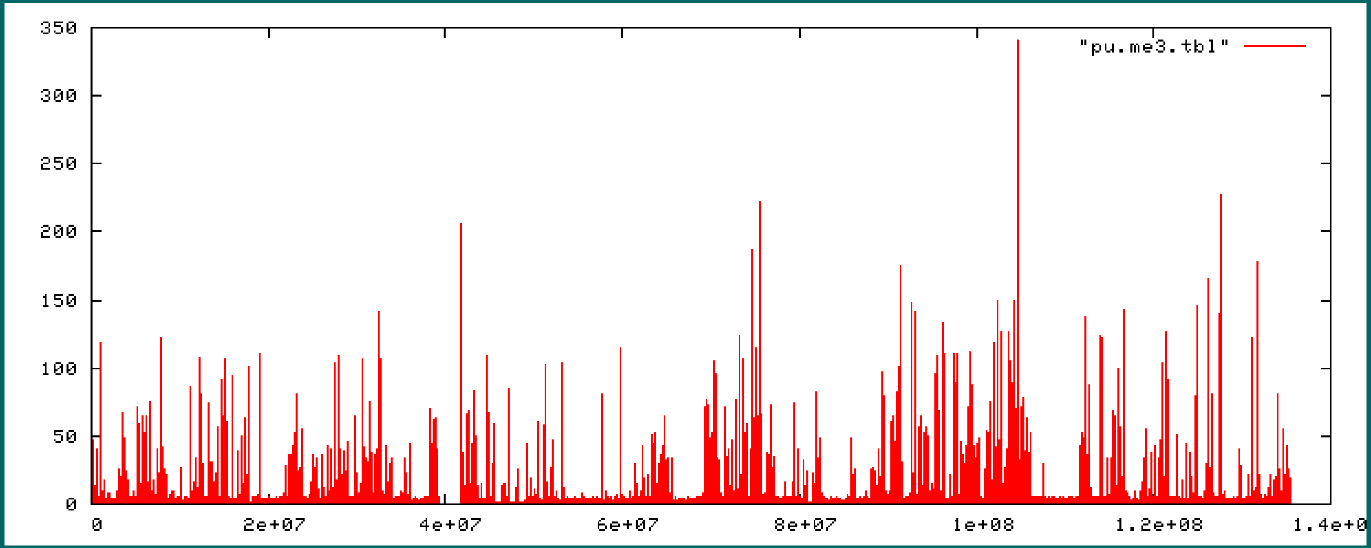
Back to the bird's eyes view

H3K4me1



H3K4me
1

H3K4me3



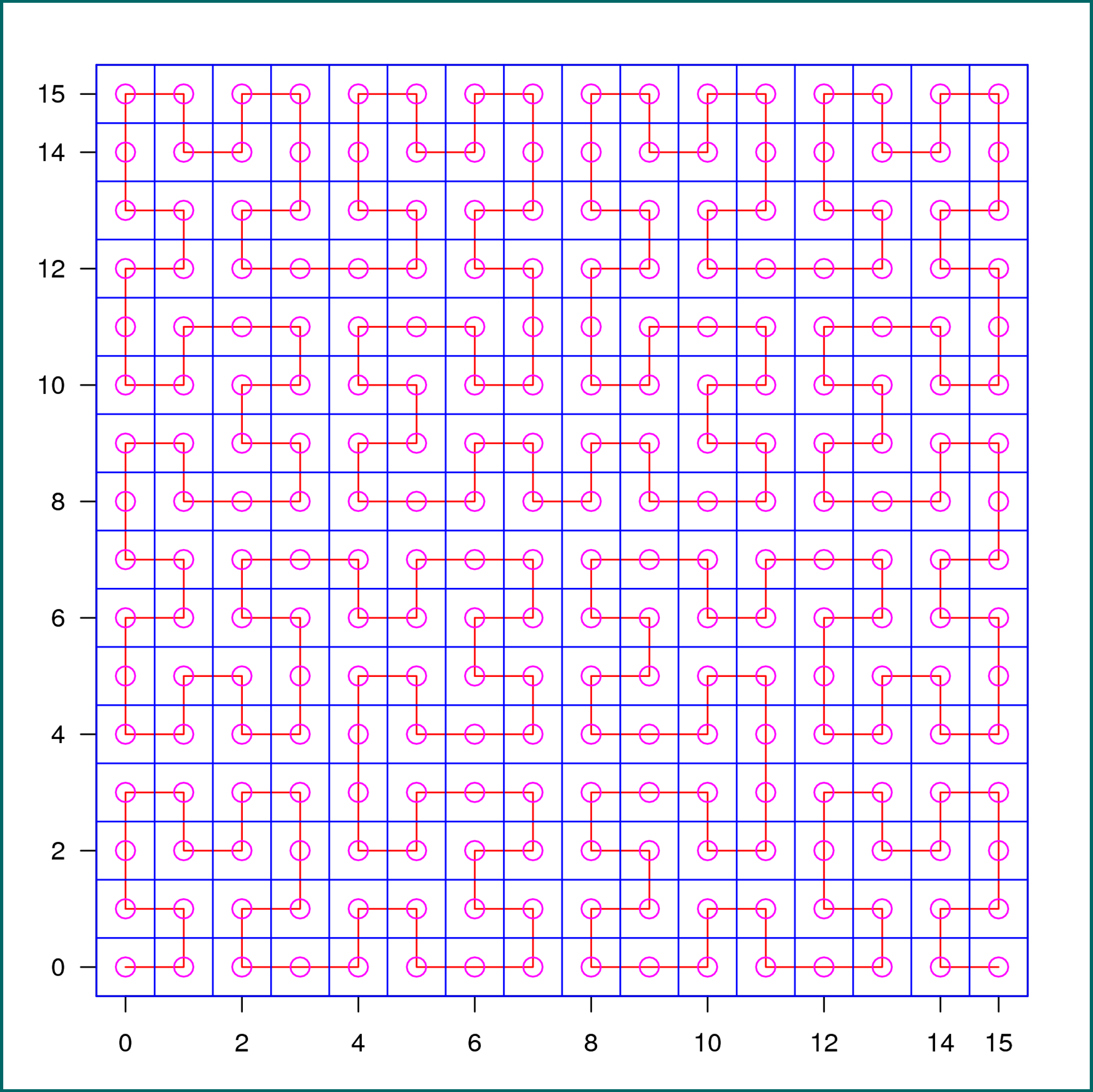
EMBL-EBI

We need a way to get a general overview on the data without either not seeing any details not getting lost in them.

A possible solution: Hilbert curve visualisation

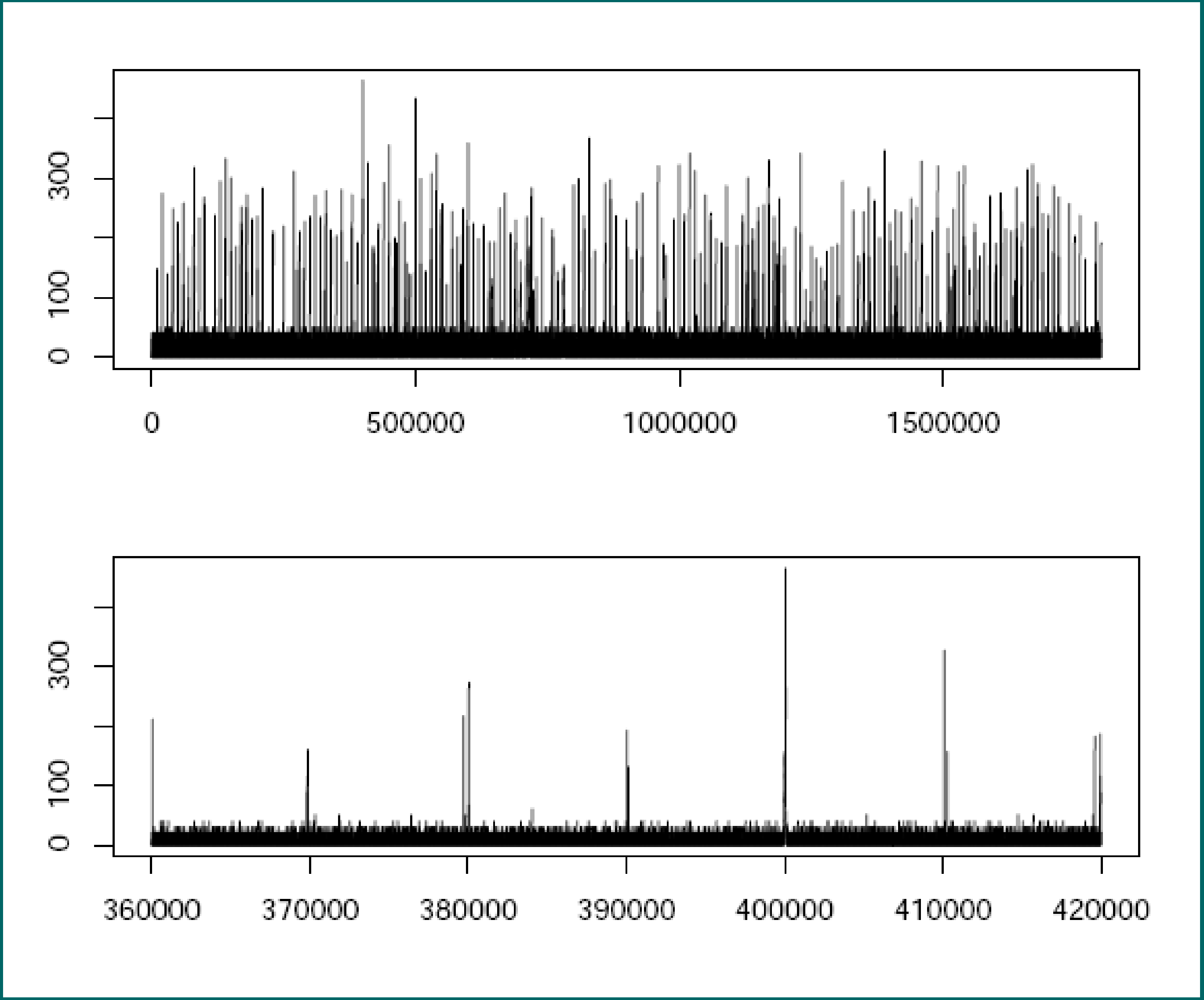
S. An.: “Visualisation of genomic data with the Hilbert curve”, Bioinformatics, Vol. 25 (2009) pp. 1231-1235

The Hilbert curve



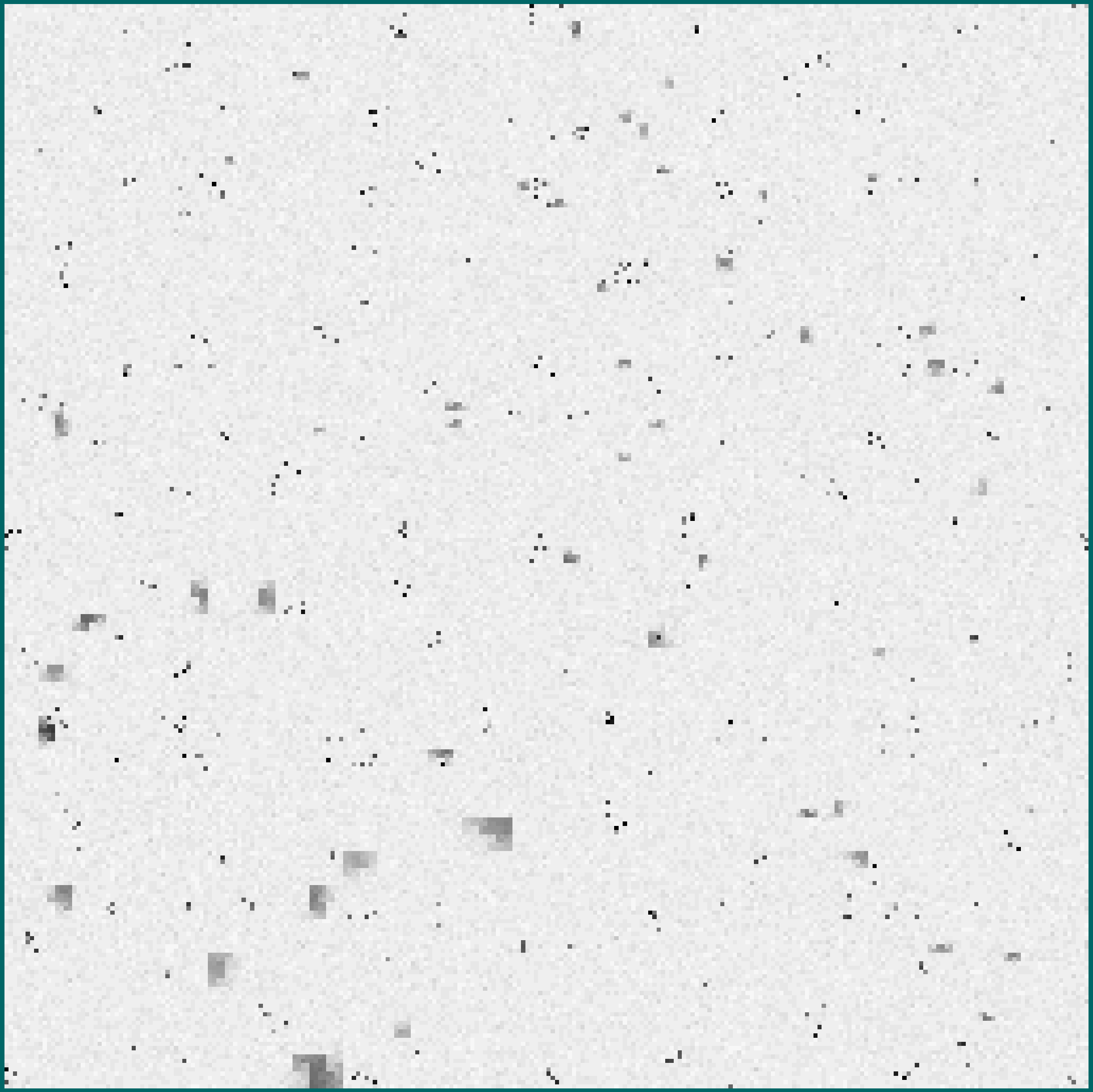
EMBL-EBI

What is hidden in here?



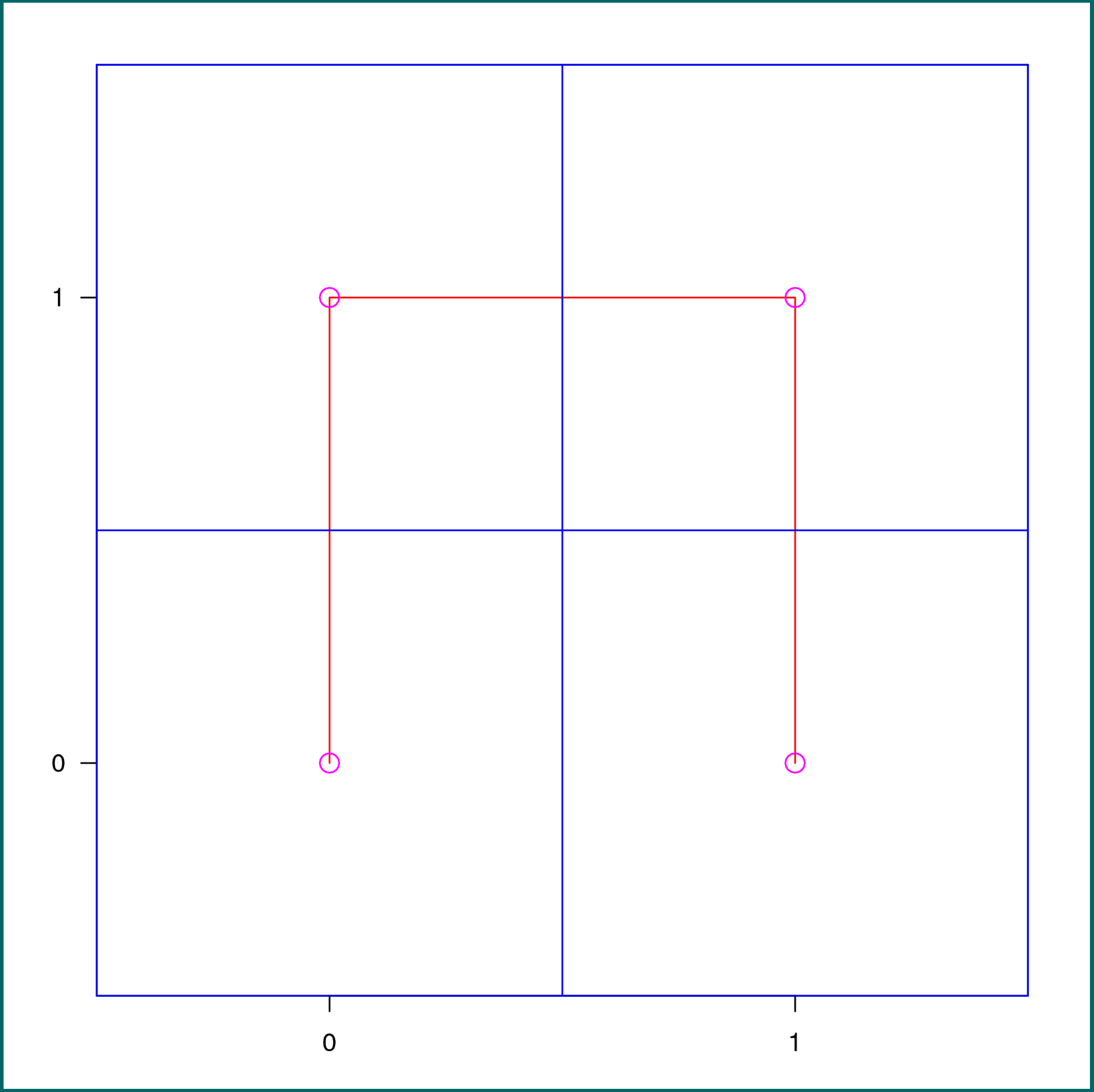
EMBL-EBI

Hilbert plot of the constructed example vector



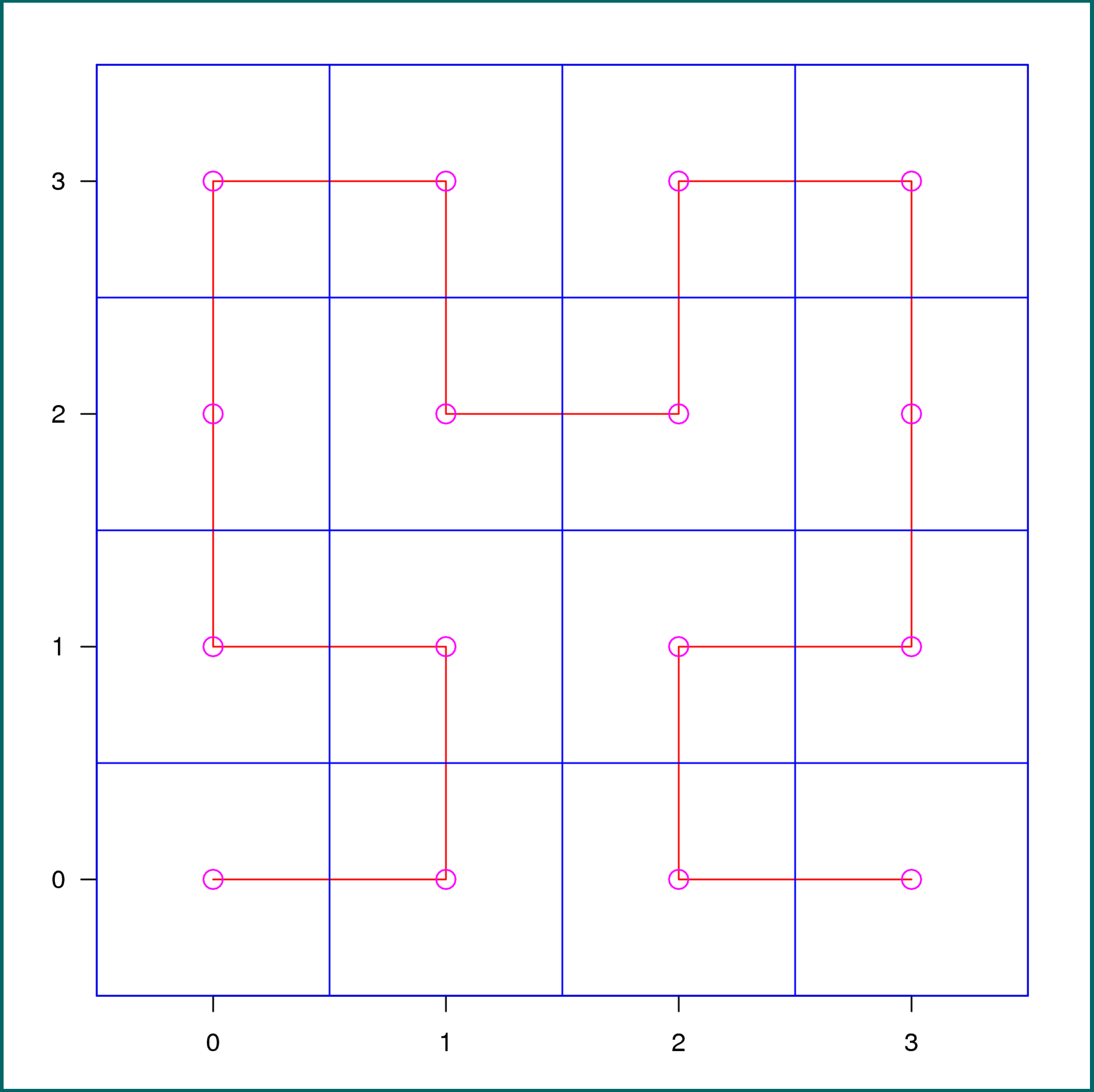
EMBL-EBI

Construction of the Hilbert curve: Level 1



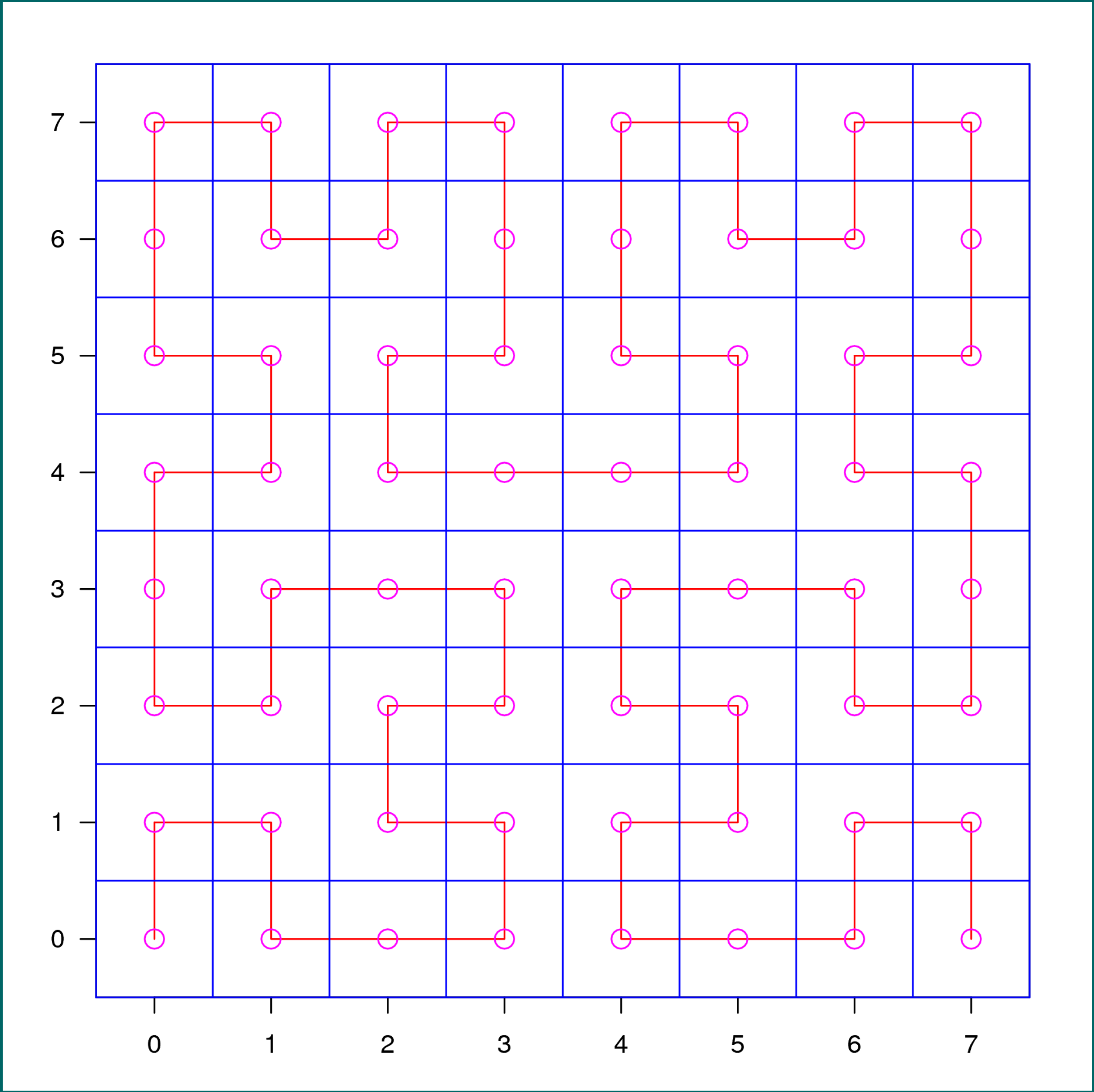
EMBL-EBI

Construction of the Hilbert curve: Level 2



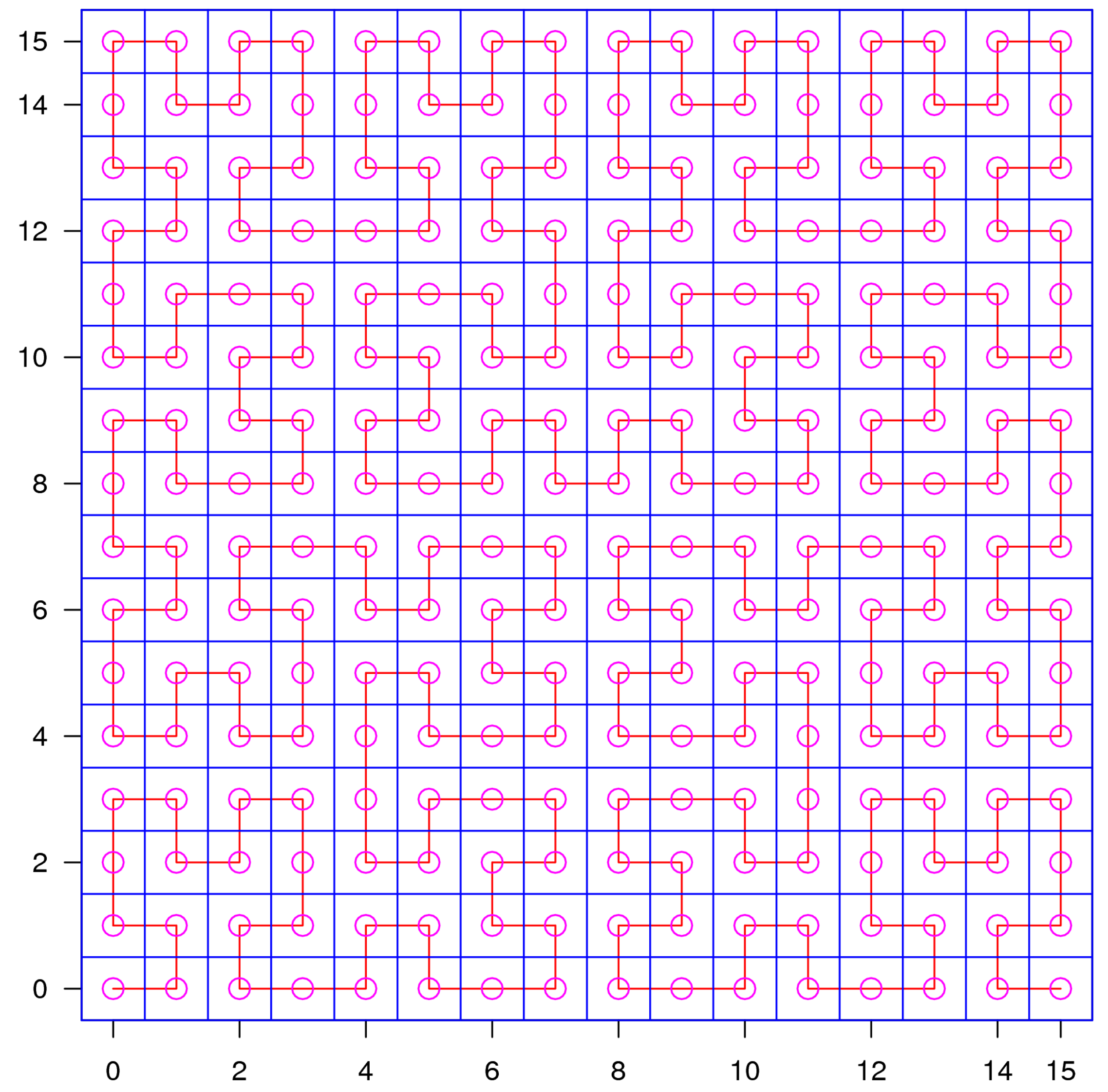
EMBL-EBI

Construction of the Hilbert curve: Level 3



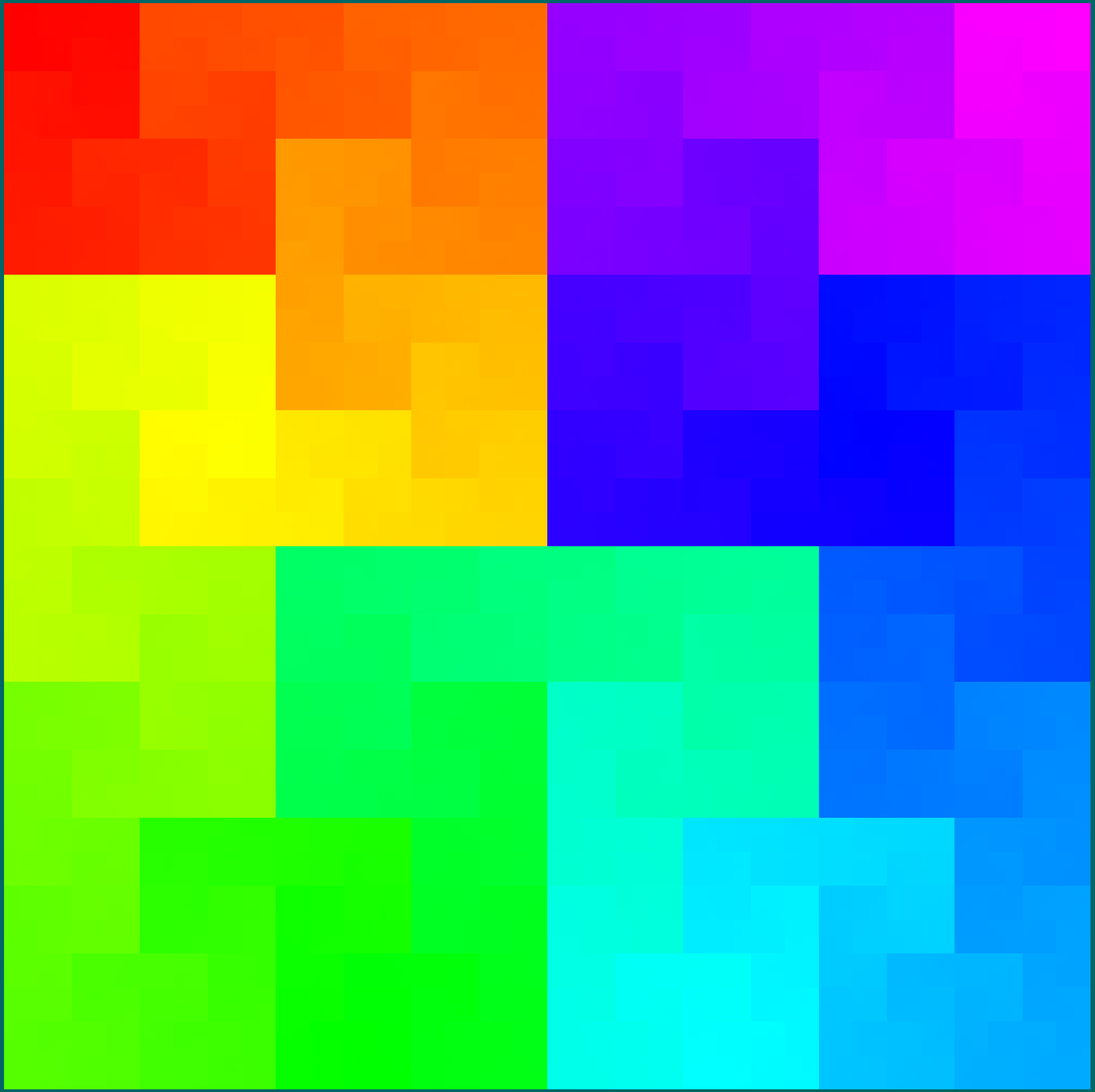
EMBL-EBI

Construction of the Hilbert curve: Level 4



EMBL-EBI

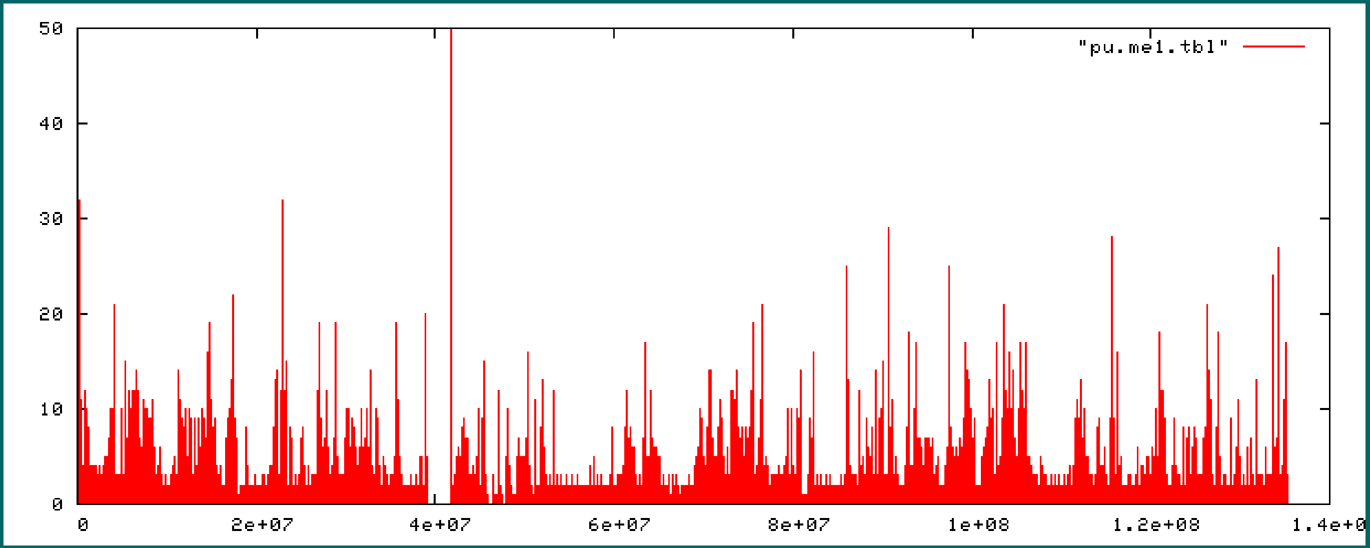
Hilbert curve: Approaching the limit



EMBL-EBI

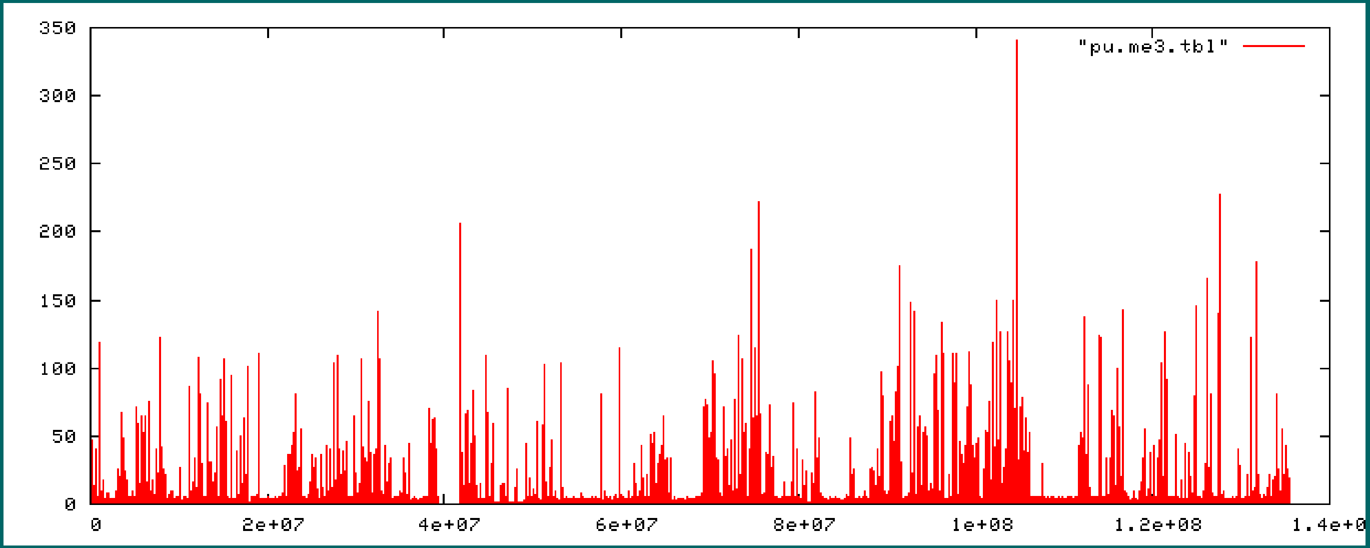
Coverage vector for a full chromosome (chr10)

H3K4me1



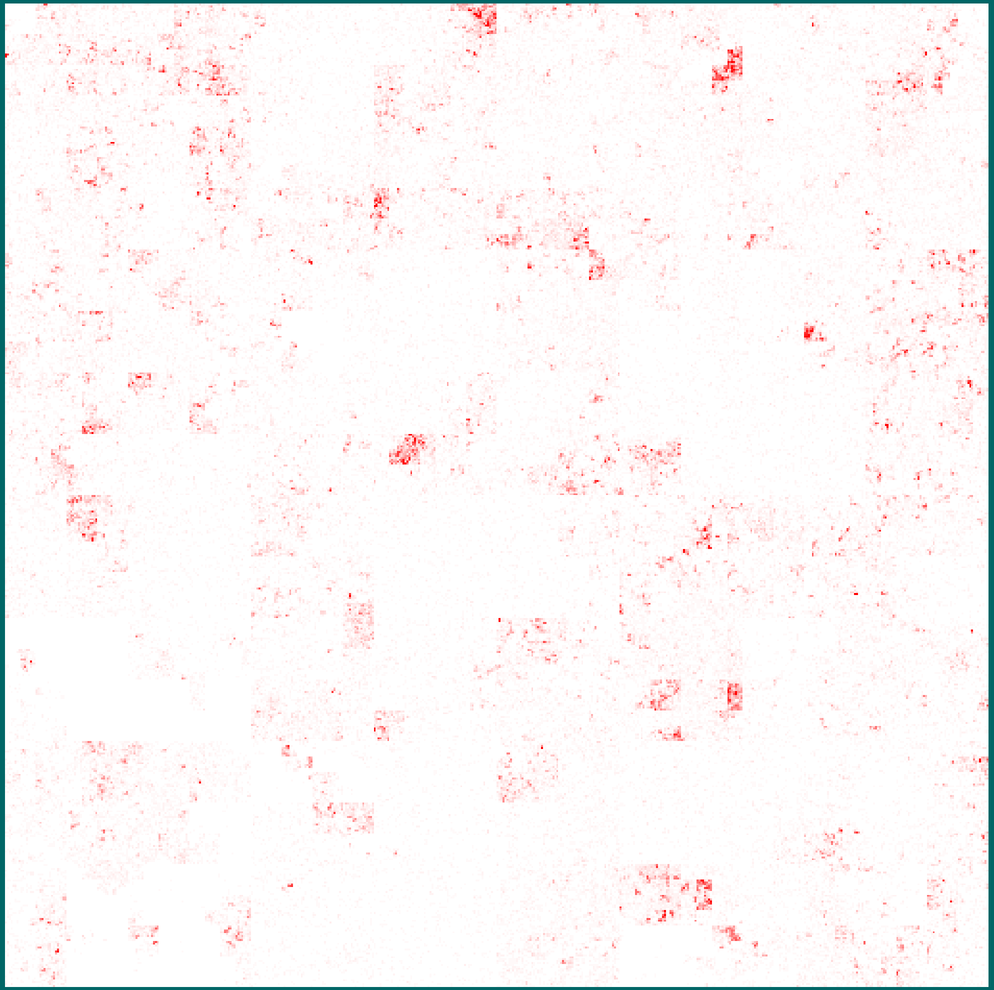
chrom. 10

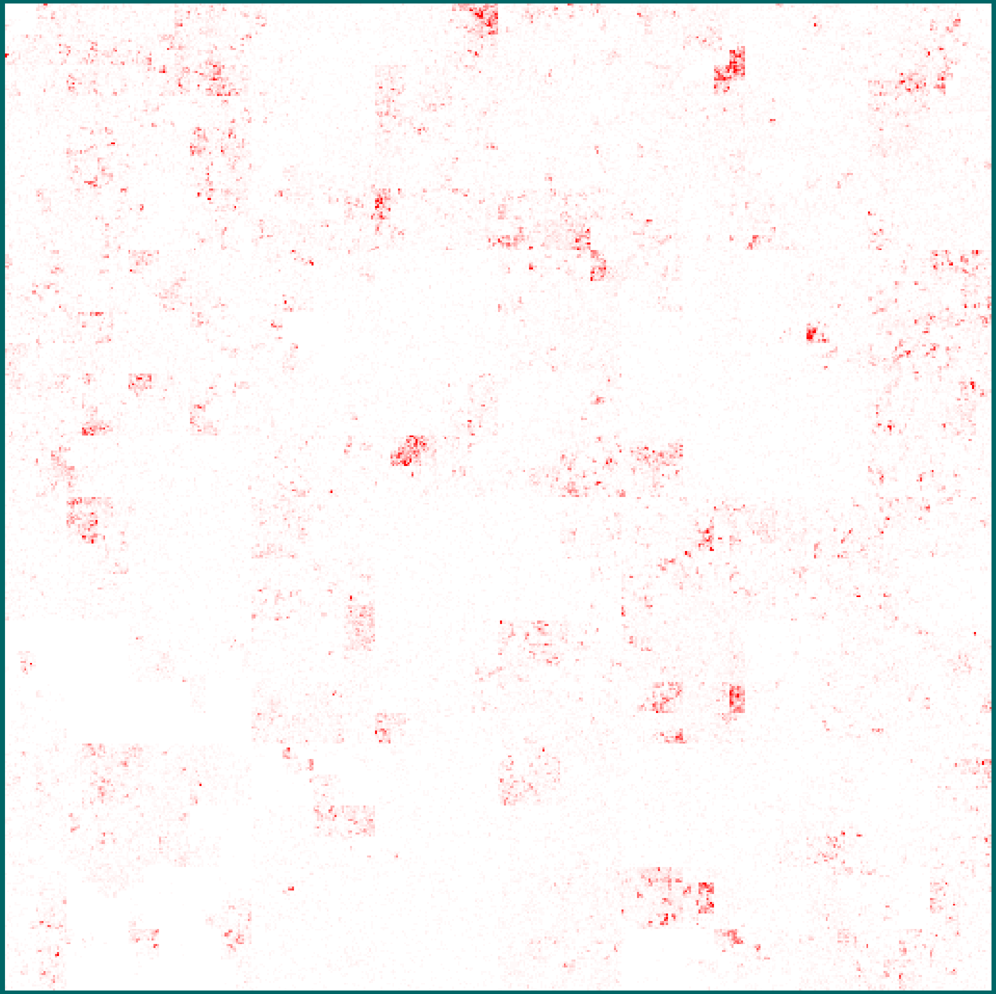
H3K4me3



EMBL-EBI

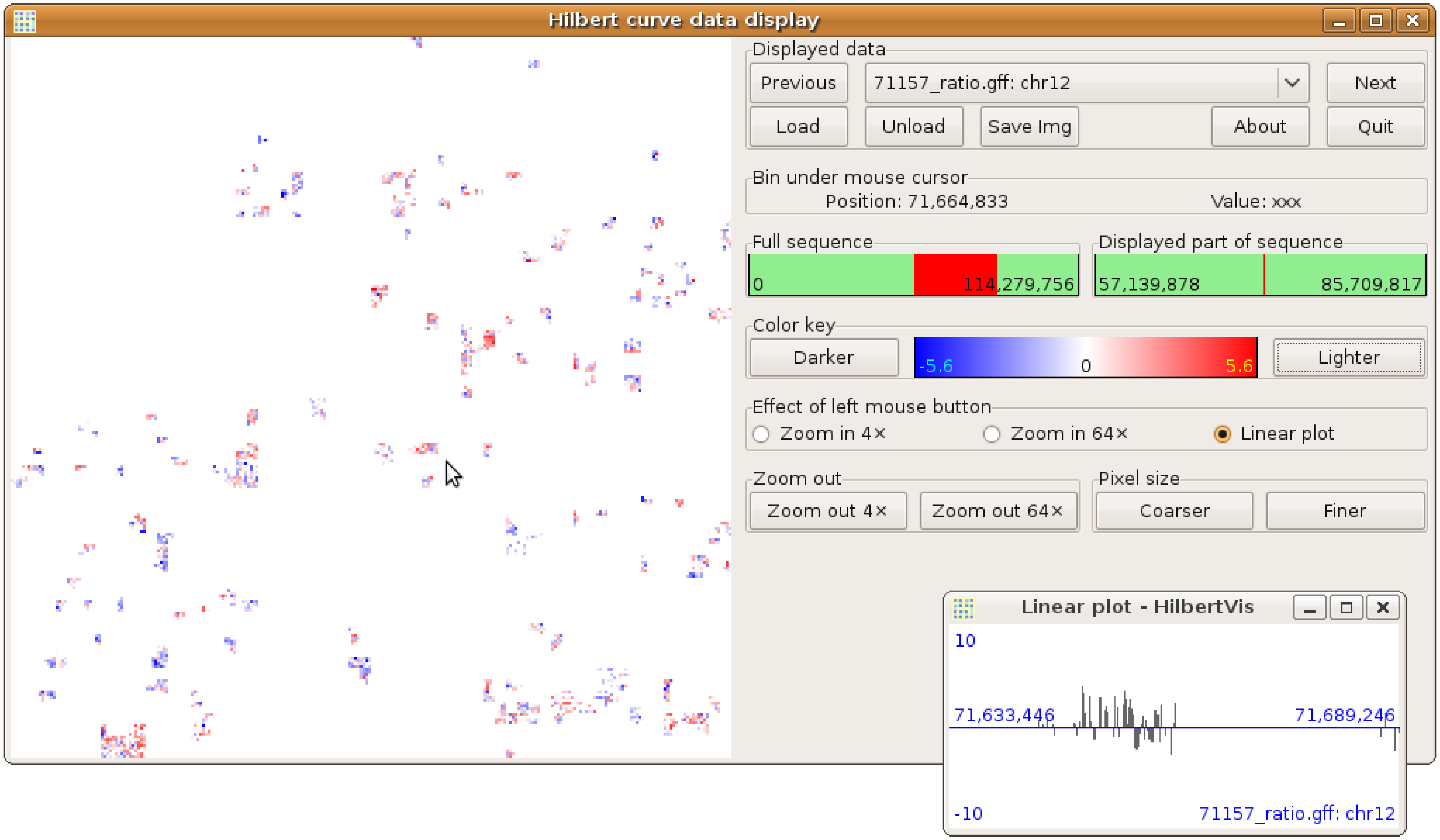
Hilbert plot of the coverage vectors


H3K4me1
(mono-methylation)


H3K4me3
(tri-methylation)

EMBL-EBI

HilbertVis


Hilbert curve data display
Displayed data
Previous
71157_ratio.gff: chr12
Next
Load
Unload
Save Img
About
Quit
Bin under mouse cursor
Position: 71,664,833
Value: xxx
Full sequence
0
114,279,756
Displayed part of sequence
57,139,878
85,709,817
Color key
Darker
-5.6
0
5.6
Lighter
Effect of left mouse button
Zoom in 4×
Zoom in 64×
[✓ Linear plot]
Zoom out
Zoom out 4×
Zoom out 64×
Pixel size
Coarser
Finer
Linear plot - HilbertVis
10
71,633,446
71,689,246
-10
71157_ratio.gff: chr12

HilbertVis

- stand-alone tool to display GFF, Wiggle, Maq map

  http://www.ebi.ac.uk/huber-srv/hilbert/

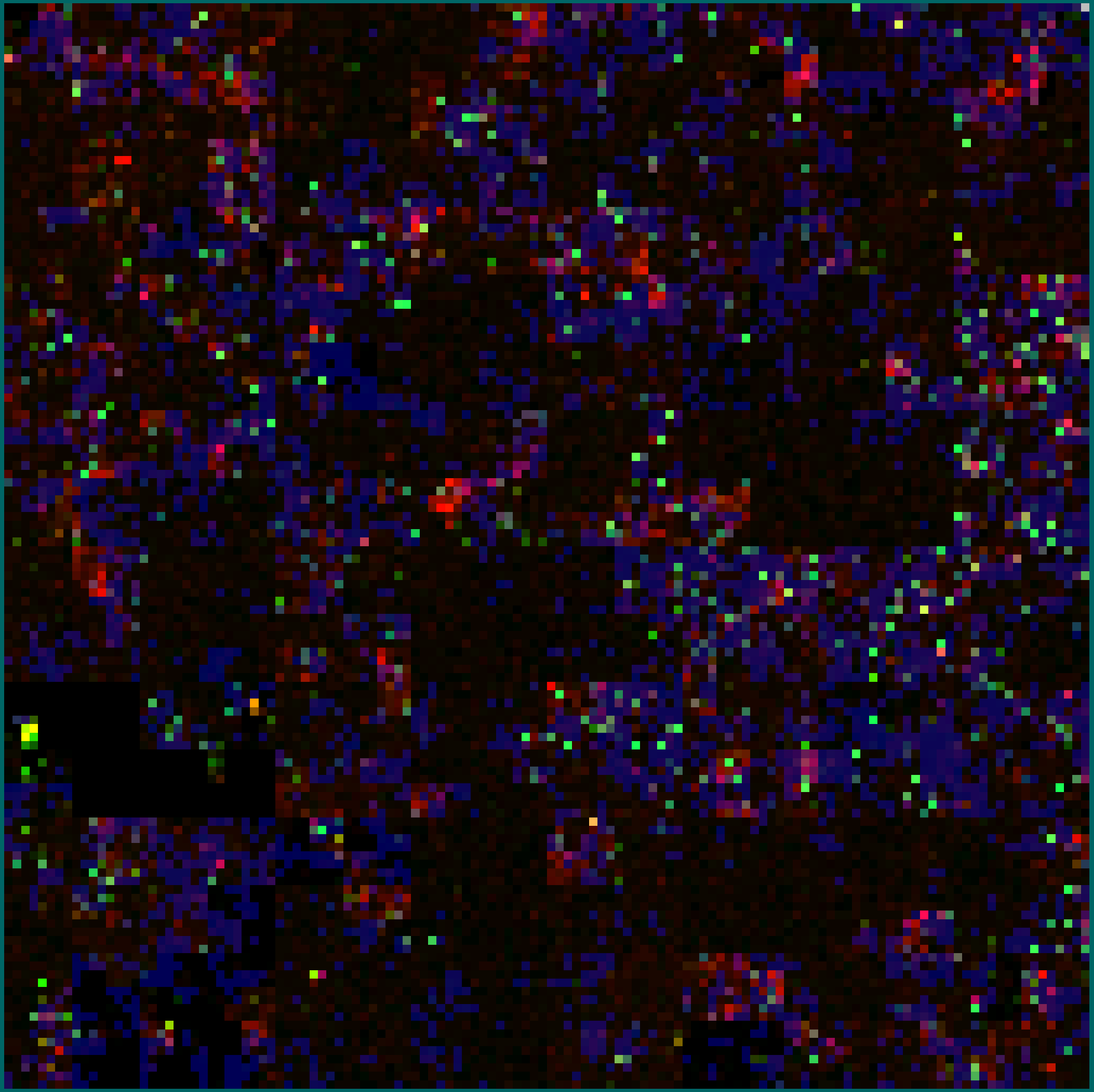
  (or Google for “hilbertvis”)

- R package to display any long R vector
  - either via commands for batch processing

    Bioconductor package “HilbertVis”
  - or via GUI for exploring

    Bioconductor package “HilbertVisGUI”

Three-colour Hilbert plot



Overlay of the previous plots and exon density

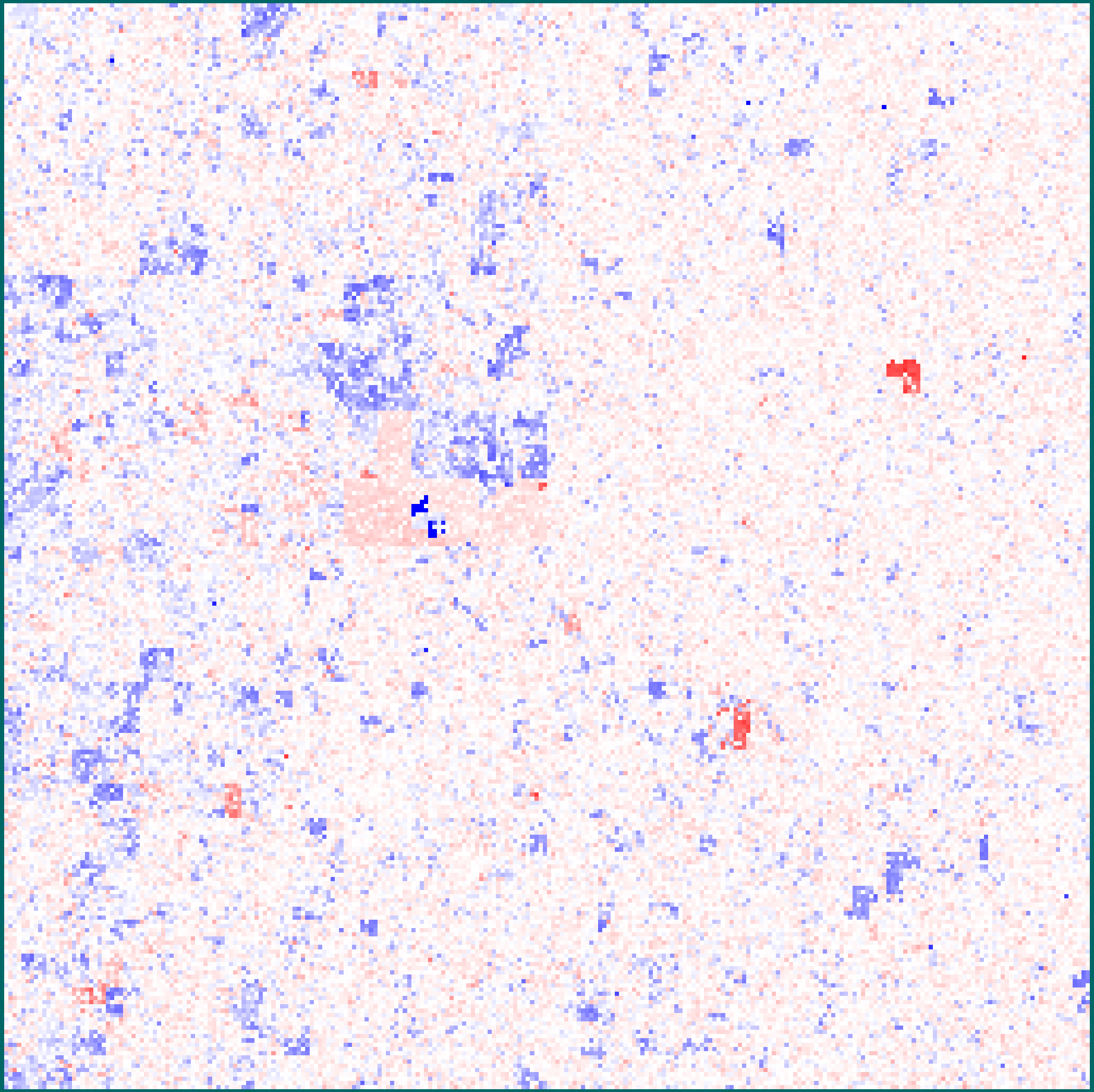
red:
mono-methylation

green:
tri-methylation

blue:
exons

EMBL-EBI

Other uses: Array-CGH

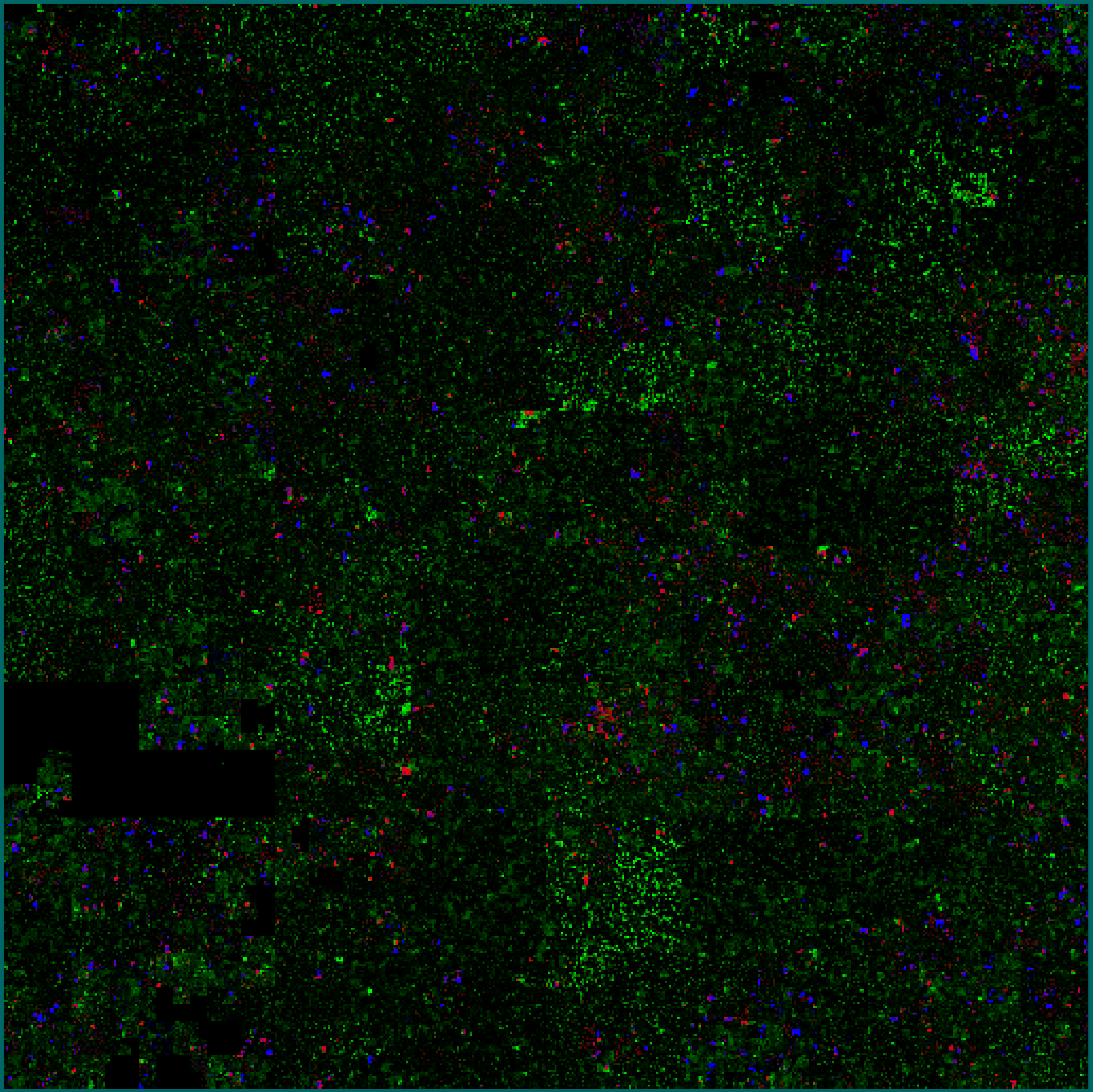


Log fold-changes between two *Arabidopsis* eco-types, chromosome 2

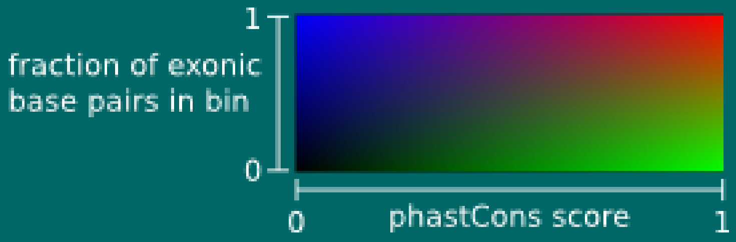
[Data courtesy of
M. Seiffert, IPK Gatersleben

EMBL-EBI

Other uses: Conservation scores



Human chromosome 10: Comparing phastCons conservation scores with exon density

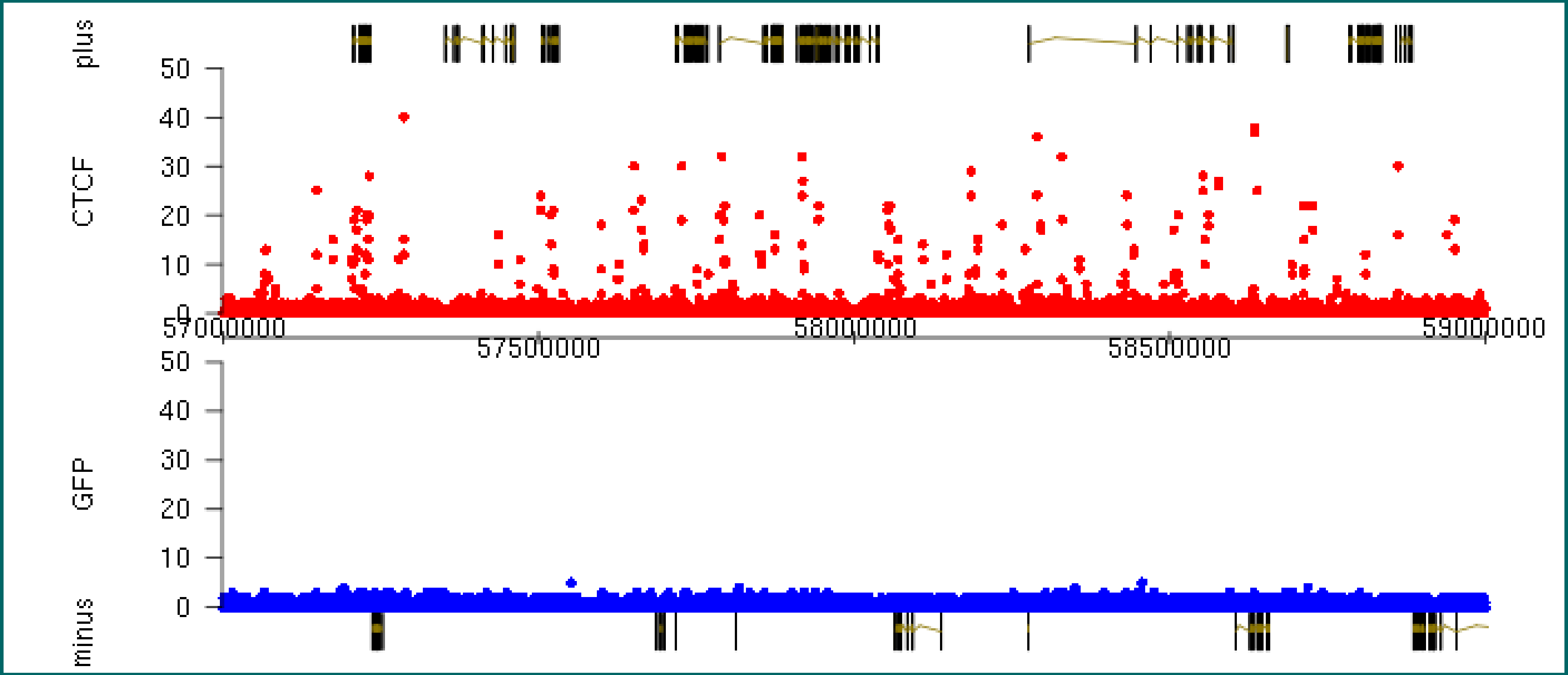


EMBL-EBI

GenomeGraphs

GenomeGraphs: Bioconductor package by S. Durrinck, UCB

- Load gene models from Ensembl via BiomaRt and plots them, alongside experimental data



GenomeGraphs: Code for sample plot

```
library(GenomeGraphs)
library(HilbertVis)

mart <- useMart("ensembl", dataset = "mmusculus_gene_ensembl")

start <- 57000000
end <-   59000000

plusStrand <- makeGeneRegion( chromosome = 10,
   start = start, end = end, strand = "+", biomart = mart )

minusStrand <- makeGeneRegion( chromosome = 10,
   start = start, end = end, strand = "-", biomart = mart )

genomeAxis <- makeGenomeAxis( )
```

EMBL-EBI

GenomeGraphs: Code for sample plot, cont'd

```
track.ctcf <- makeBaseTrack(
   base = seq( start, end, length.out = 10000 ),
   value = shrinkVector(
      as.vector( cov.ctcf$chr10[start:end] ), 10000 ),
   dp = DisplayPars( lwd = 0.5, color="red", ylim=c(0, 50) ) )

track.gfp <- makeBaseTrack(
   base = seq( start, end, length.out = 10000 ),
   value = shrinkVector(
      as.vector( cov.gfp$chr10[start:end] ), 10000 ),
   dp = DisplayPars( lwd = 0.5, color="blue", ylim=c(0, 50) ) )

gdPlot( list( `plus`=plusStrand, `CTCF`=track.ctcf,
   genomeAxis, `GFP`=track.gfp, `minus`=minusStrand ) )
```

EMBL-EBI

*

EMBL-EBI