The Genomic HyperBrowser

Exploring the borders of the galaxy

Geir Kjetil Sandve,
University of Oslo,
26.may 2011

The Genomic HyperBrowser

Exploring the borders of the galaxy

Since 2007..

Outline

- The Genomic HyperBrowser
- Life at the borders of the Galaxy
- Why we like Galaxy

Outline

- The Genomic HyperBrowser
- Life at the borders of the Galaxy
- Why we like Galaxy

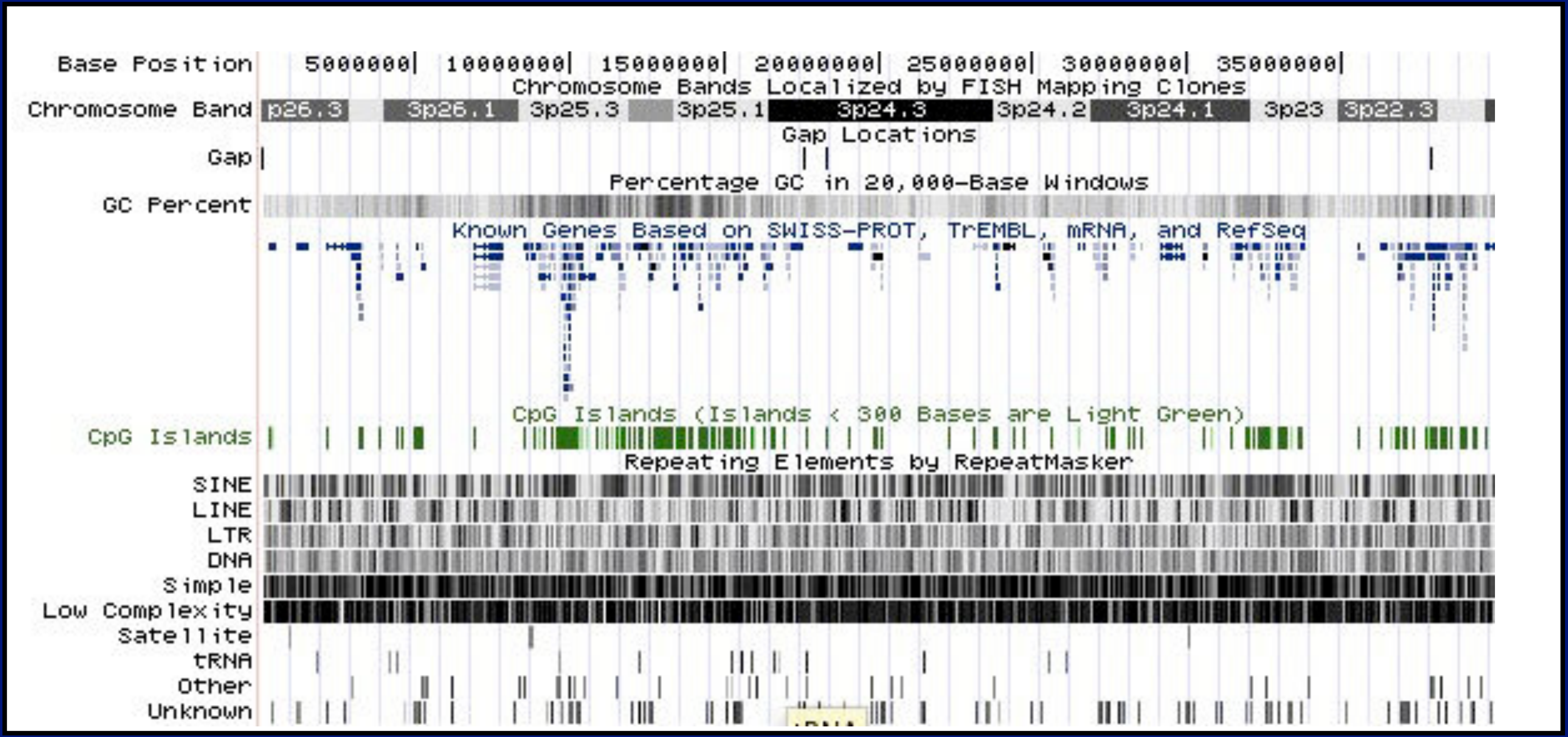
The Genomic HyperBrowser

Why?

Genomic information is becoming plentiful

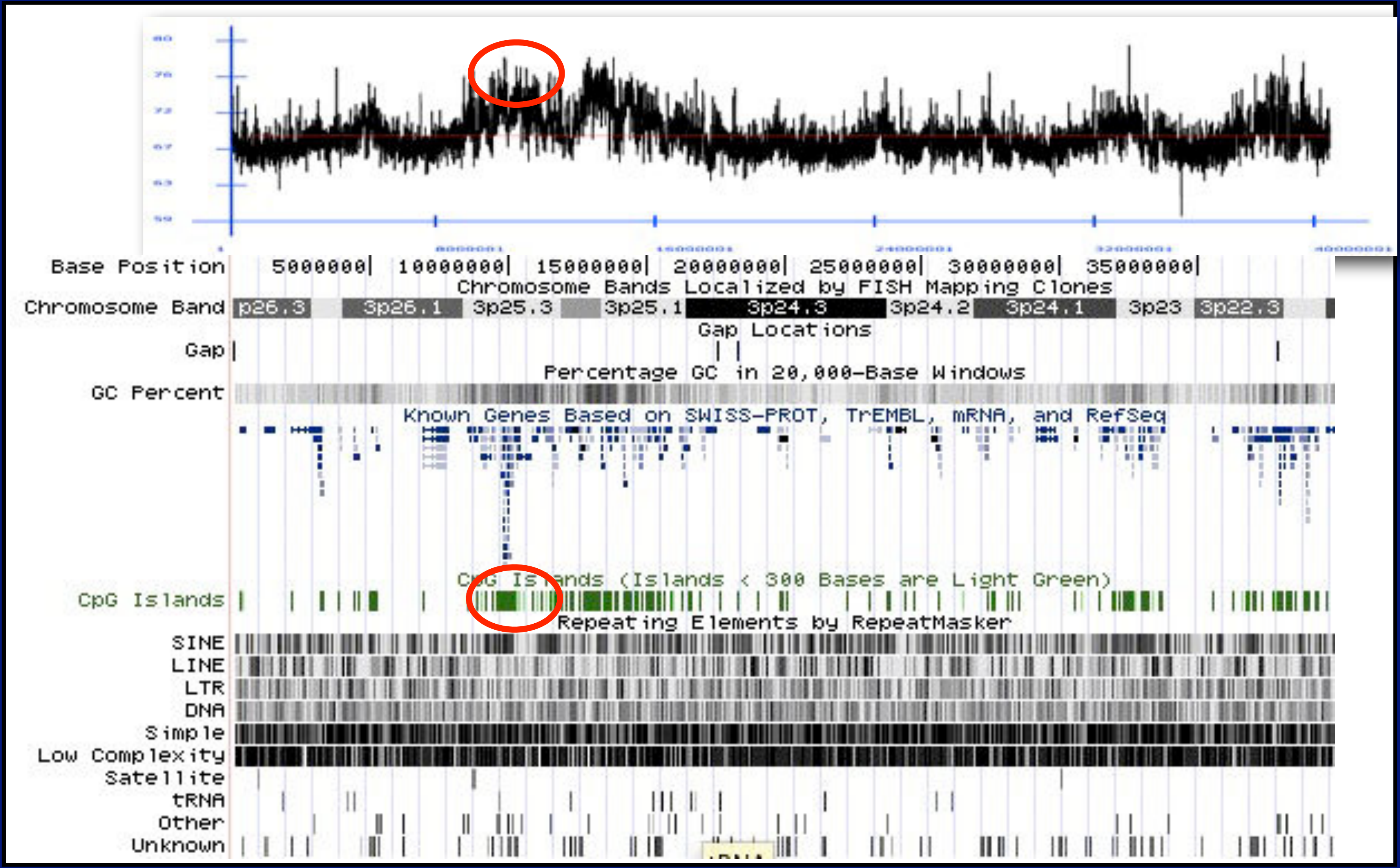
chr13 (q12.3)
p13
p12
p11.2
q13.3
14.11
q21.1
21.33
13q31.1
q31.3
13q34

40 N



Trac
anno

But analysis lags behind

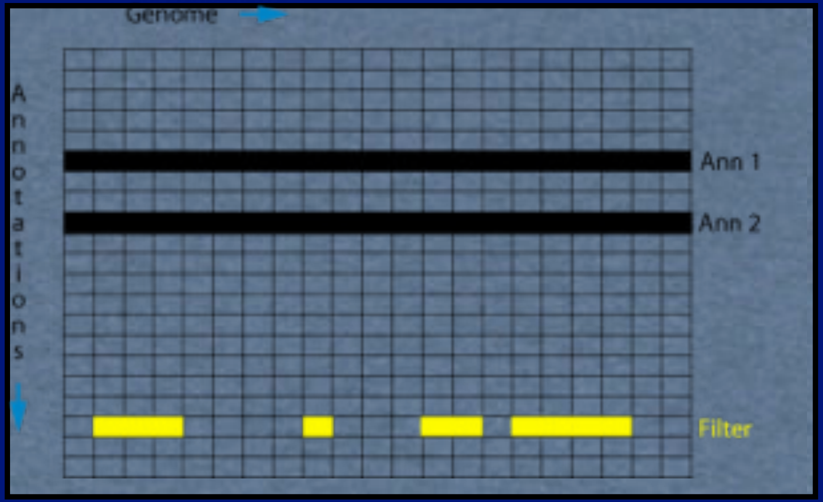


40 M

Trac
anno

Enter:

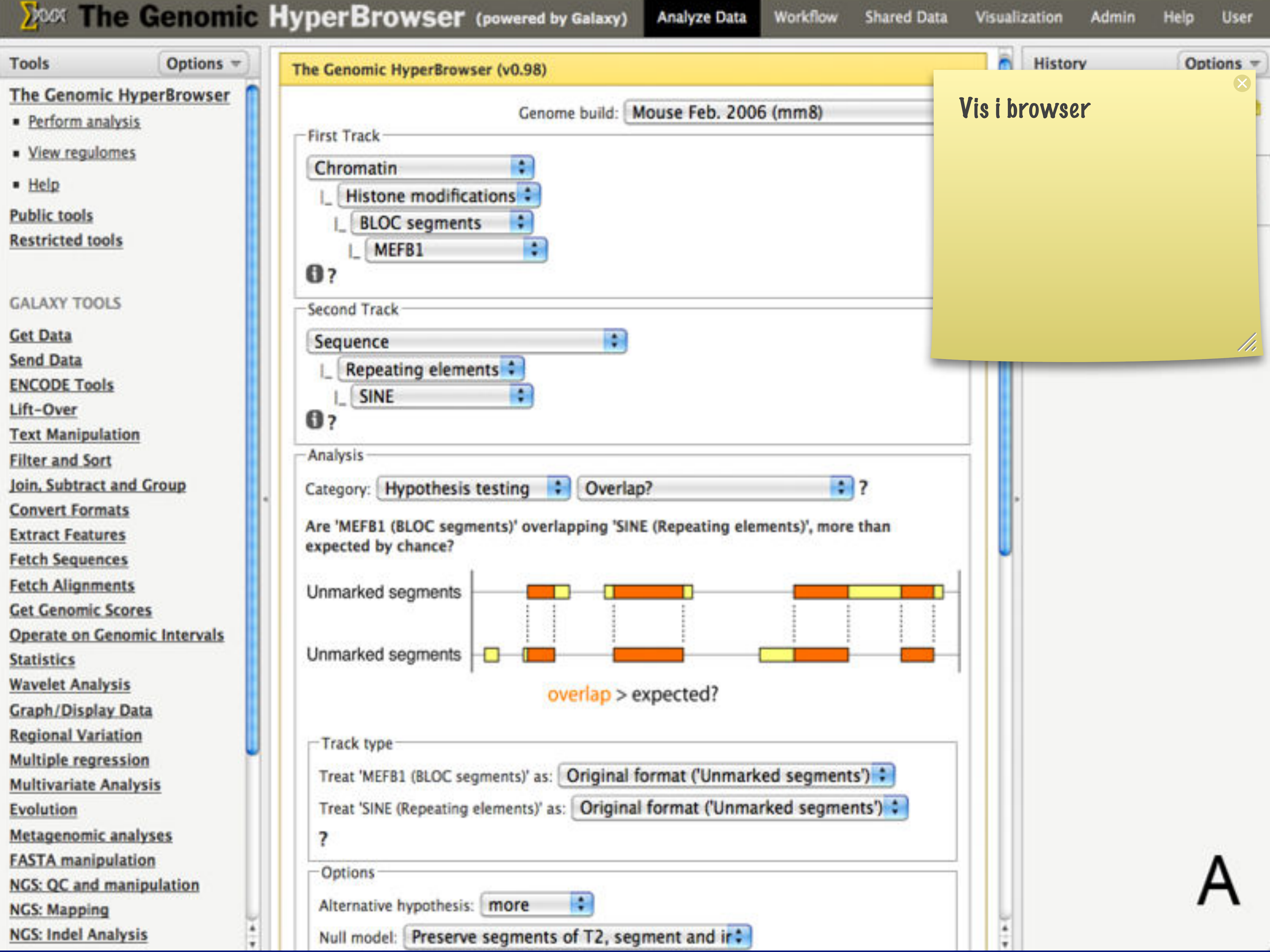
The Genomic hyperbrowser



- Do two tracks at a time, but robustly and comprehensively
- Emphasize local analysis
- Each result is in fact a new track

The Genomic HyperBrowser

What?



The Genomic HyperBrowser
- Perform analysis
- View regulomes
- Help

Public tools
Restricted tools

GALAXY TOOLS

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: Indel Analysis

The Genomic HyperBrowser (v0.98)

Genome build: Mouse Feb. 2006 (mm8)

First Track

Chromatin
|_ Histone modifications
|_ BLOC segments
|_ MEFB1

?

Second Track

Sequence
|_ Repeating elements
|_ SINE

?

Analysis

Category: Hypothesis testing | Overlap? ?

Are 'MEFB1 (BLOC segments)' overlapping 'SINE (Repeating elements)', more than expected by chance?

Track type

Treat 'MEFB1 (BLOC segments)' as: Original format ('Unmarked segments')

Treat 'SINE (Repeating elements)' as: Original format ('Unmarked segments')

?

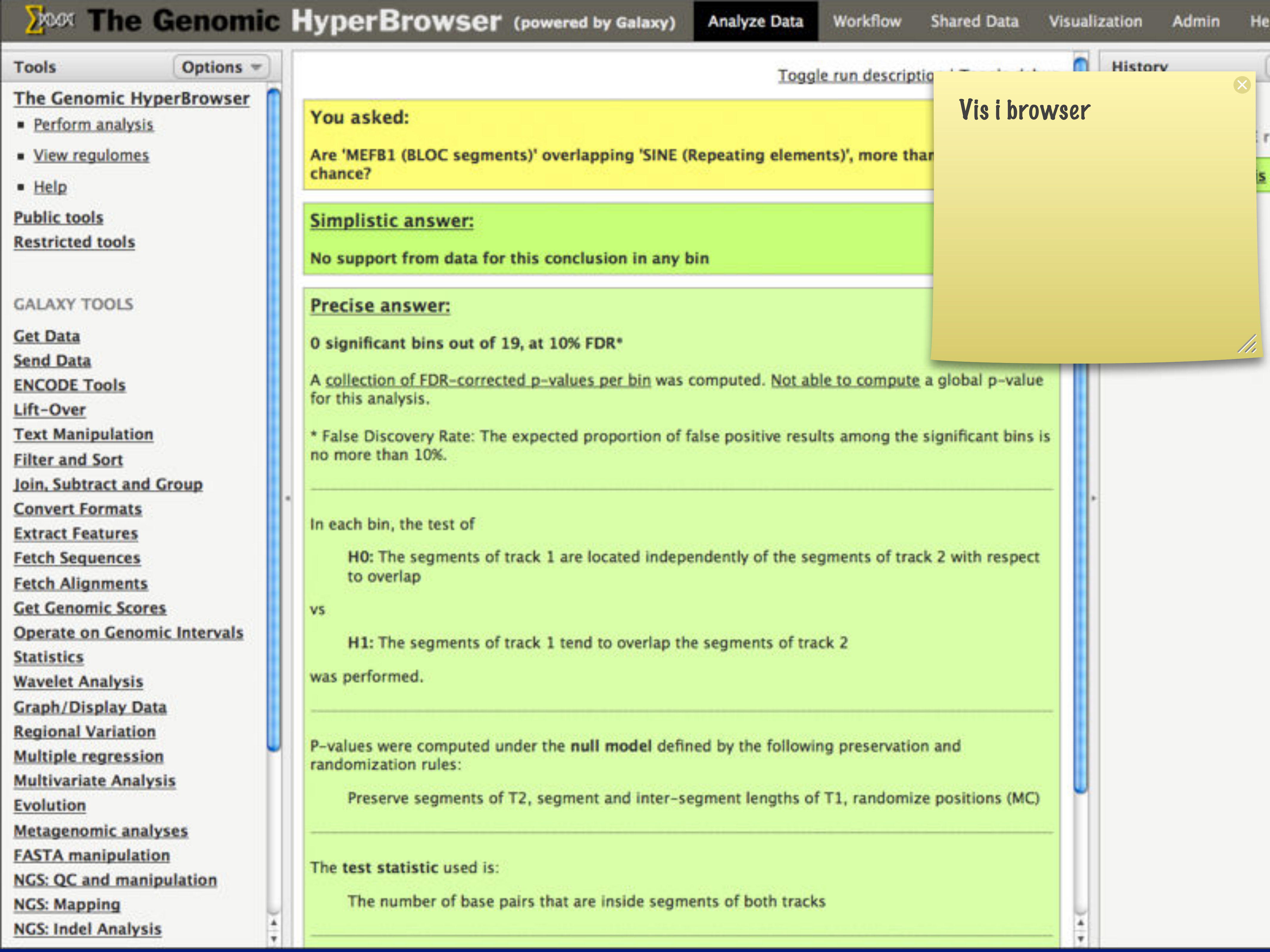
Options

Alternative hypothesis: more

Null model: Preserve segments of T2, segment and ir

Vis i browser

A



Vis i browser

You asked:

Are 'MEFB1 (BLOC segments)' overlapping 'SINE (Repeating elements)', more thar chance?

Simplistic answer:

No support from data for this conclusion in any bin

Precise answer:

0 significant bins out of 19, at 10% FDR*

A collection of FDR-corrected p-values per bin was computed. Not able to compute a global p-value for this analysis.

* False Discovery Rate: The expected proportion of false positive results among the significant bins is no more than 10%.

In each bin, the test of

> **H0:** The segments of track 1 are located independently of the segments of track 2 with respect to overlap

vs

> **H1:** The segments of track 1 tend to overlap the segments of track 2

was performed.

P-values were computed under the **null model** defined by the following preservation and randomization rules:

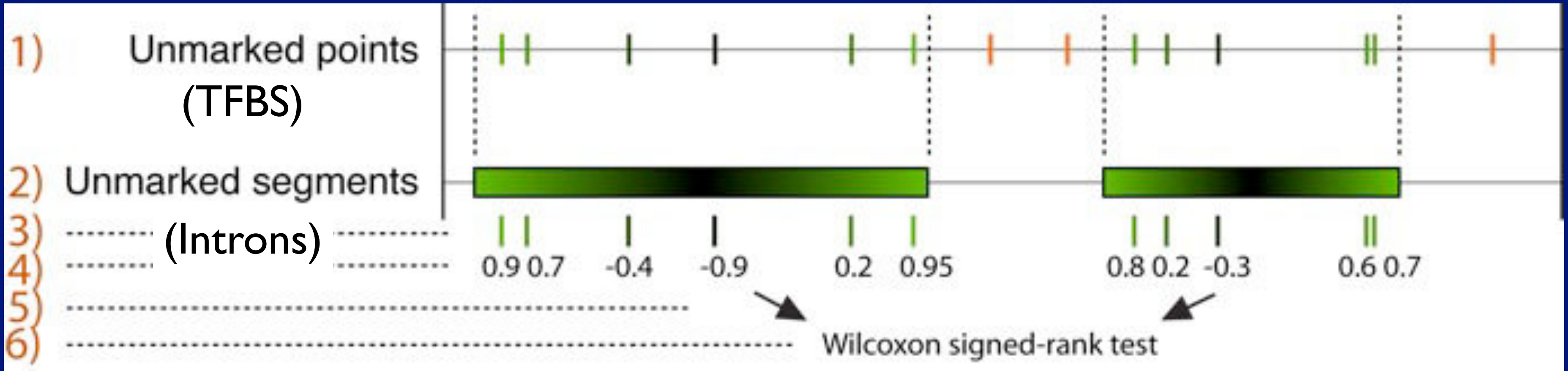
> Preserve segments of T2, segment and inter-segment lengths of T1, randomize positions (MC)

The **test statistic** used is:

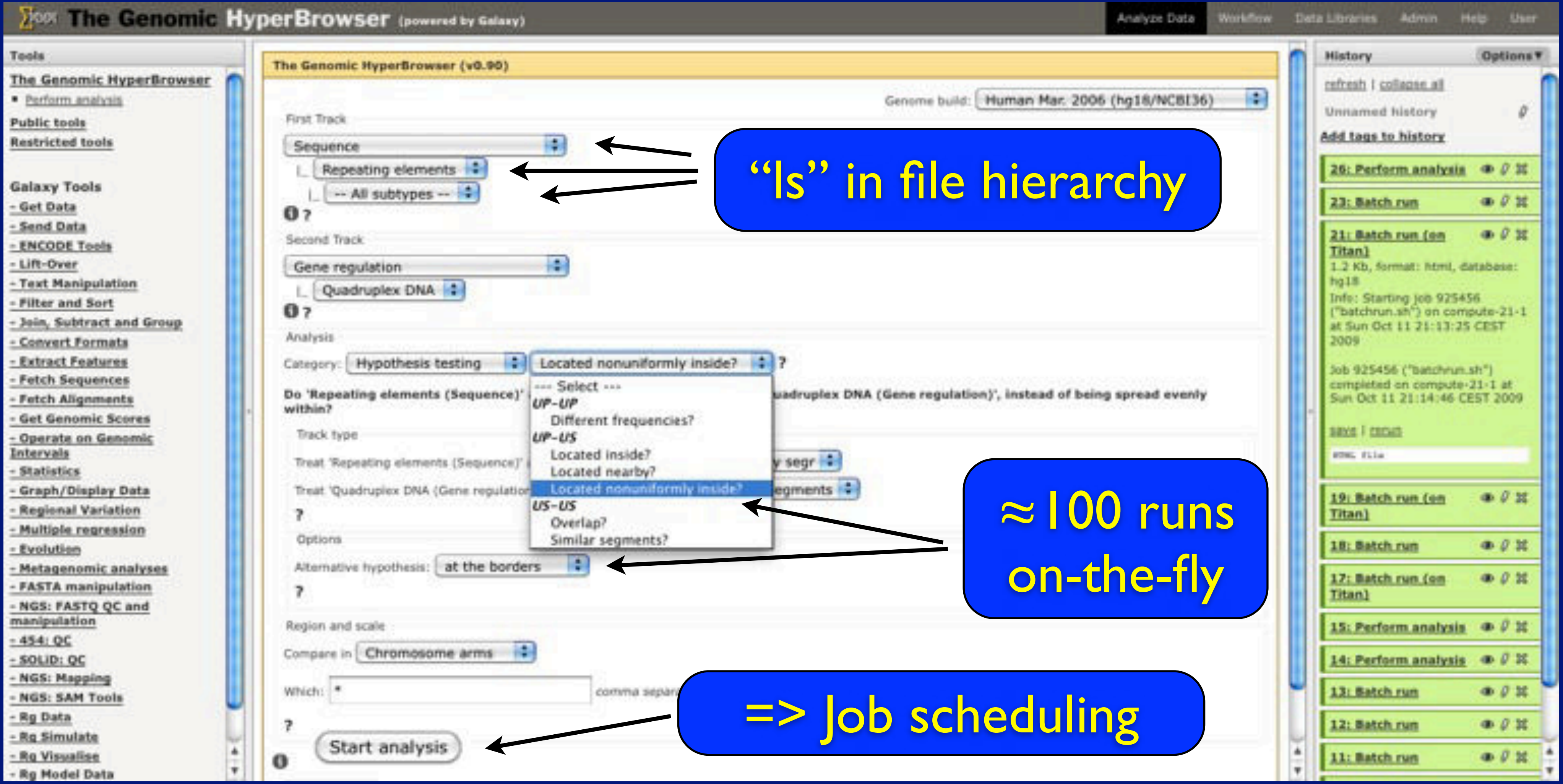
> The number of base pairs that are inside segments of both tracks

You know what, we know how

Do 'TF binding sites' accumulate more towards the borders of 'Introns'?



Requirements for interactivity



The Genomic HyperBrowser (powered by Galaxy)

Tools

The Genomic HyperBrowser
- Perform analysis

Public tools

Restricted tools

Galaxy Tools
- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: FASTQ QC and manipulation
- 454: QC
- SOLiD: QC
- NGS: Mapping
- NGS: SAM Tools
- Rg Data
- Rg Simulate
- Rg Visualise
- Rg Model Data

The Genomic HyperBrowser (v0.90)

Genome build: Human Mar. 2006 (hg18/NCBI36)

First Track

Sequence
Repeating elements
-- All subtypes --

Second Track

Gene regulation
Quadruplex DNA

Analysis

Category: Hypothesis testing | Located nonuniformly inside?

--- Select ---
UP-UP
Different frequencies?
UP-US
Located inside?
Located nearby?
Located nonuniformly inside?
US-US
Overlap?
Similar segments?

Do 'Repeating elements (Sequence)' … uadruplex DNA (Gene regulation)', instead of being spread evenly within?

Track type

Treat 'Repeating elements (Sequence)' … y segr

Treat 'Quadruplex DNA (Gene regulatio… egments

Options

Alternative hypothesis: at the borders

Region and scale

Compare in: Chromosome arms

Which: *   comma separ…

Start analysis

History   Options

refresh | collapse all

Unnamed history

Add tags to history

26: Perform analysis

23: Batch run

21: Batch run (on Titan)
1.2 Kb, format: html, database: hg18
Info: Starting job 925456 ("batchrun.sh") on compute-21-1 at Sun Oct 11 21:13:25 CEST 2009

Job 925456 ("batchrun.sh") completed on compute-21-1 at Sun Oct 11 21:14:46 CEST 2009

save | rerun

19: Batch run (on Titan)

18: Batch run

17: Batch run (on Titan)

15: Perform analysis

14: Perform analysis

13: Batch run

12: Batch run

11: Batch run

"ls" in file hierarchy

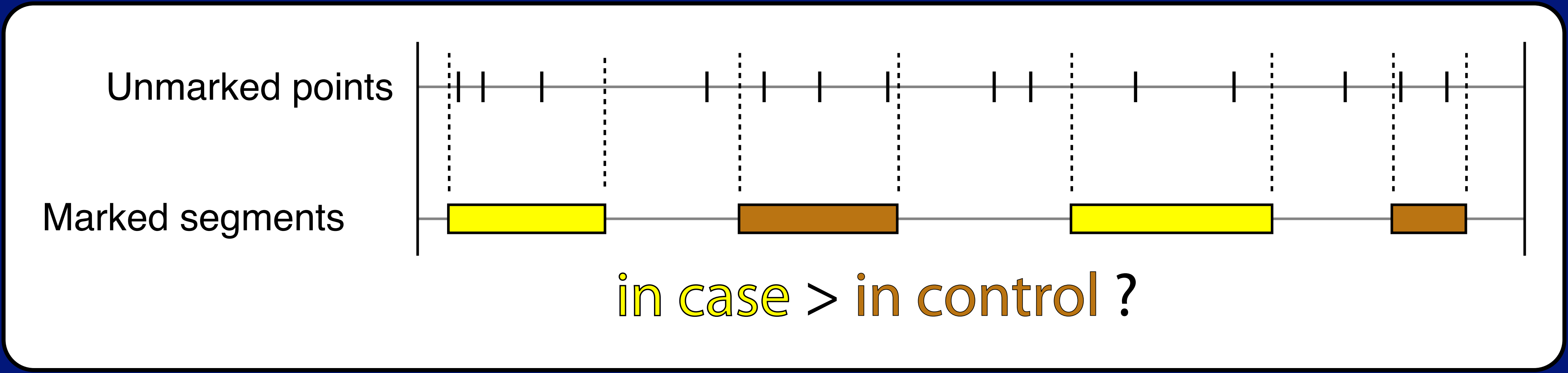
≈100 runs on-the-fly

=> Job scheduling

The Genomic HyperBrowser

The power of for-loops

AP2A vs Melanoma



All TFs vs Melanoma

AP2ALPHA
E2F
CREBP1
CREB
NFKAPPAB
CREB
ZIC2
CREBP1CJUN
CREL
....

[illegible]
E2F
[illegible]
[illegible]
[illegible]
[illegible]
ZIC2
[illegible]
[illegible]
[illegible]
[illegible]
ZIC3
[illegible]
[illegible]
[illegible]
NF1
[illegible]
AP1
[illegible]
USF
[illegible]
AP1
SP1
[illegible]
[illegible]
ATF
[illegible]
[illegible]
AHR
[illegible]
[illegible]
AP1
[illegible]
[illegible]

IK1
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
AP2
[illegible]
ZIC1
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
P53
[illegible]
YY1
[illegible]
IK2
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
E2F
[illegible]
AP1

[illegible]
CP2
GR
[illegible]
[illegible]
[illegible]
[illegible]
IRF2
[illegible]
[illegible]
ER
[illegible]
[illegible]
[illegible]
ZID
AP1
[illegible]
[illegible]
[illegible]
[illegible]
IK3
[illegible]
SRF
[illegible]
USF
E2F
[illegible]
USF
[illegible]
[illegible]
SP1
P53
[illegible]
[illegible]

[illegible]
HTF
[illegible]
[illegible]
AP4
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
NFY
[illegible]
E47
[illegible]
IRF7
NFY
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
SRF
[illegible]
[illegible]

GRE
[illegible]
ISRE
IRF1
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
NFY
AP4
[illegible]
[illegible]
YY1
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
E47
[illegible]
[illegible]
[illegible]
[illegible]
SRF
[illegible]
[illegible]
[illegible]
[illegible]
HLF
[illegible]

[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
AP4
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
SRY
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
OCT
[illegible]
[illegible]
[illegible]

[illegible]
[illegible]
EN1
[illegible]
USF
NCX
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
CDP
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]

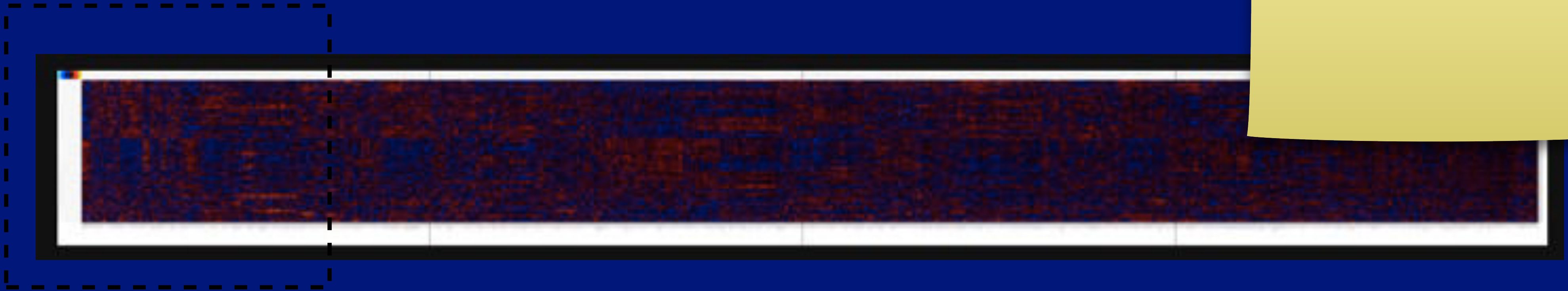
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
TBP
S8
[illegible]
CDP
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]

All TFs vs all diseases

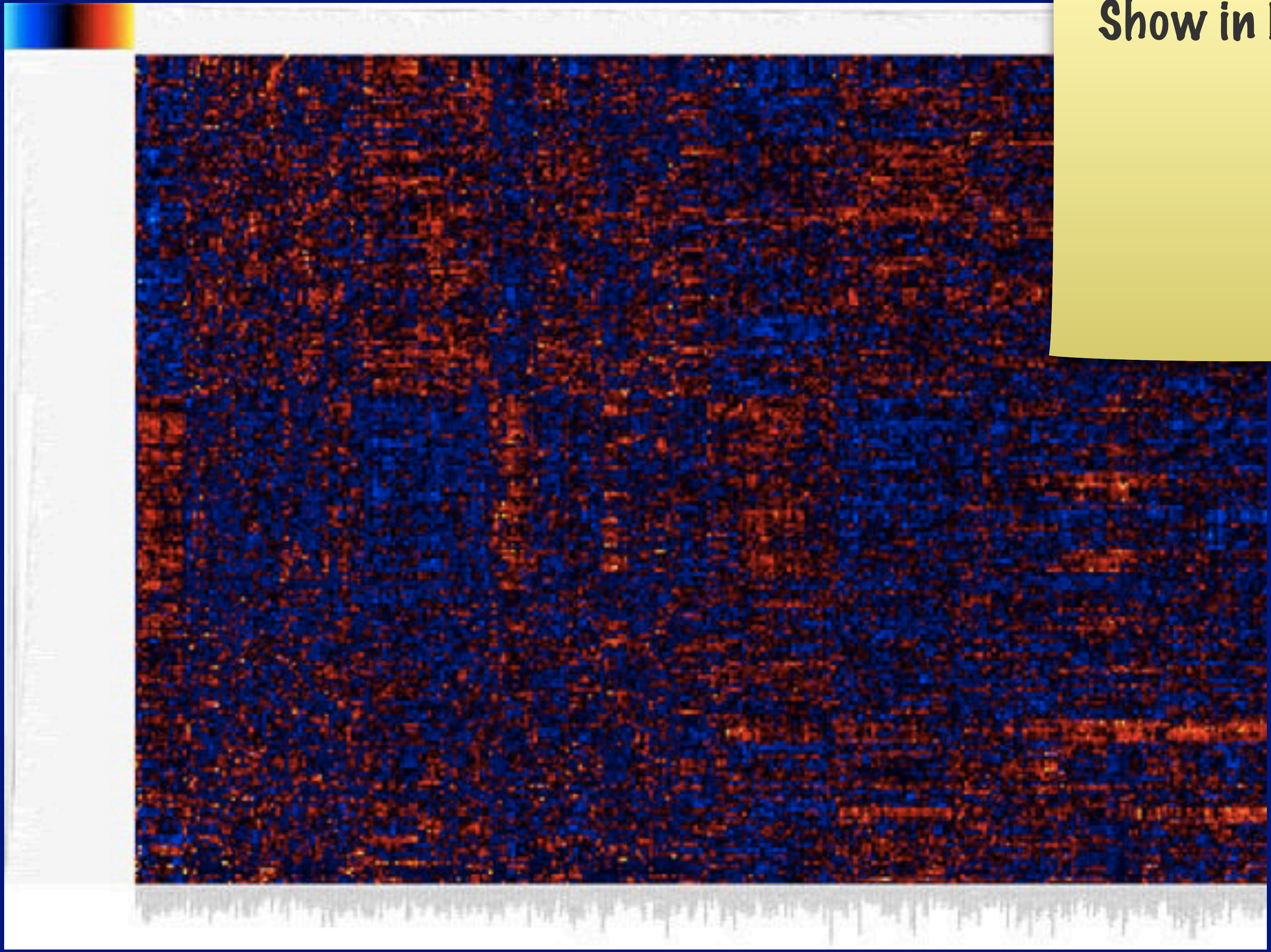
[multiply previous slide by 1068]

The disease regulome

Show in browser



The disease regulome



Show in browser

The disease regulome

- Generating hypotheses on the regulation of disease
- .. but also an interactive machine for generating such maps

Outline

- The Genomic HyperBrowser
- Life at the borders of the Galaxy
- Why we like Galaxy

Galaxy selling points (for developers)

- Stop wasting time writing interfaces
- Get your tools used by biologists

Galaxy selling points (for developers)

- ~~Stop wasting time writing interfaces~~
  - Already had GUI, and still partly external
- Get your tools used by biologists

Galaxy selling points (for developers)

- ~~Stop wasting time writing interfaces~~
  - Already had GUI, and still partly external
- ~~Get your tools used by biologists~~
  - Not distributed anyway (have our own server)

Life at the borders of the Galaxy

- Separate, monolithic codebase
- Separate GUI
- Separate data collection
- Separate results files

So, how come we still like Galaxy?

- Web server, user handling, job scheduling
- History is indeed powerful
- With time, we added 20 supporting tools..
- With time, we now consider distribution..

Looking beyond our ego..

- The HyperBrowser can't solve all problems
  - .. but Galaxy can!?
- Working towards active national installation
- Use webtools for internal code sharing

The team

Trevor Clancy
Fang Liu
Halfdan Rydbeck
Sveinung Gundersen
Ingrid K. Glad
Arnoldo Frigessi
Geir Kjetil Sandve
Lars Holden
Sigve Nakken
Knut Liestøl
Eivind Tøstesen
Eivind Hovig
Morten Johansen
Marit Holden
Vegard Nygaard
Egil Ferkingstad

Support

RIKSHOSPITALET

EMBIO Styringsgruppen for Molecular Life Sciences ved UiO

FUGE Functional genomics

UNIVERSITY OF OSLO

NR Norsk Regnesentral NORWEGIAN COMPUTING CENTER

(sfi)^2 Statistics for Innovation

sfi Centre for Research-based Innovation

Summary

- The Genomic HyperBrowser asks what, and solves how
- Tightly integrated with Galaxy, but expanding borders
- Google -> ‘hyperbrowser’

Conclusion

- Highly modifying a wheel is still better than reinventing it
- Galaxy passed our stress tests, and constantly adds value
- Google -> ‘hyperbrowser’