The Genomics Task Force, GTrack, and Galaxy warping

Swiss Galaxy Workshop, Oct. 3, 2012

Sveinung Gundersen, PhD student

The Norwegian Radium Hospital, Oslo University Hospital

Overview

- The Genomics Task Force
  - statistical analysis of genomic tracks, employing:

The Genomic
HyperBrowser

- GTrack
  - generic file format for all types of genomic tracks
- Warping the Galaxy
  - Example: custom web tool definition framework

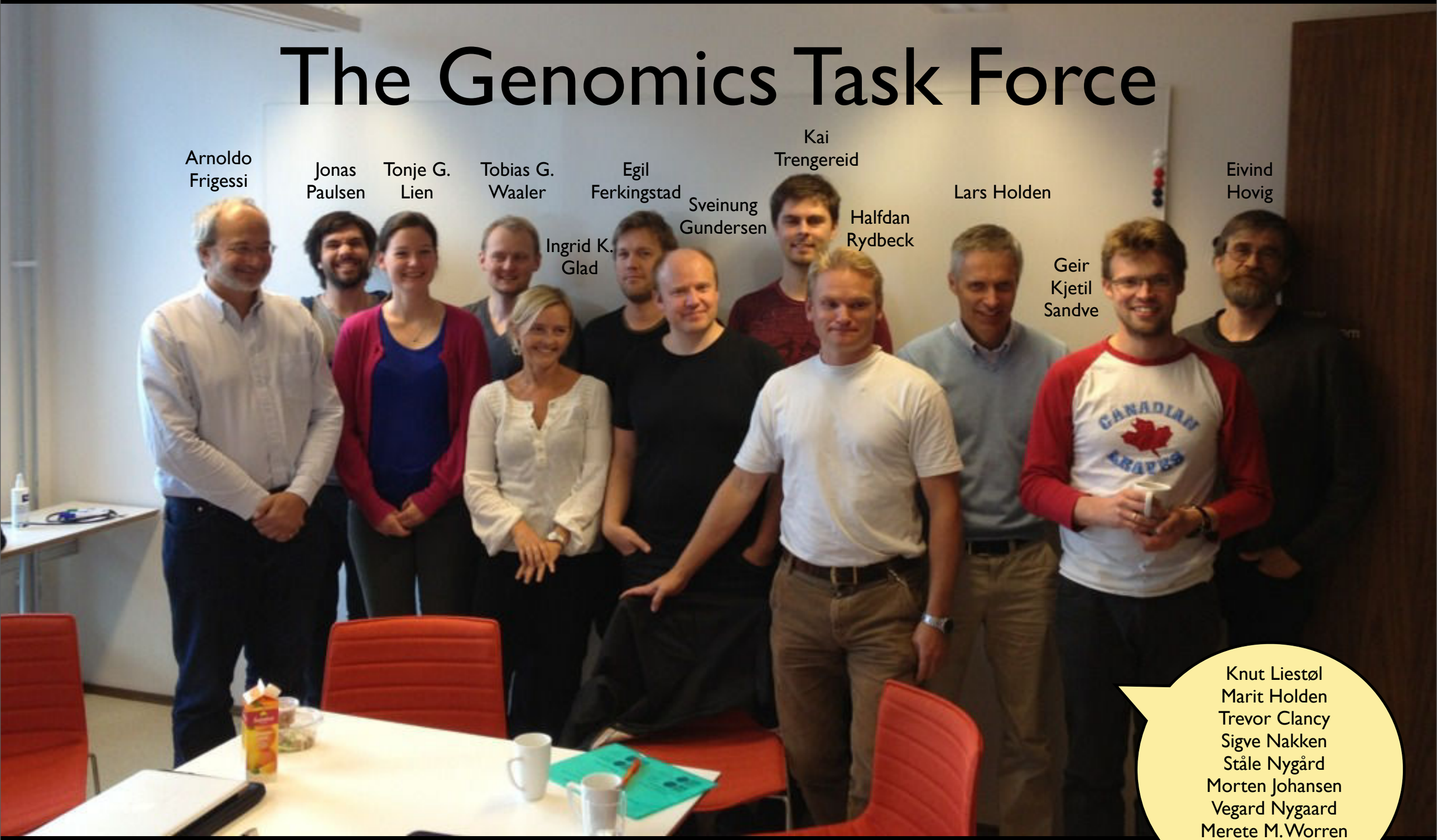
Overview

- The Genomics Task Force
  - statistical analysis of genomic tracks, employing:

The Genomic
HyperBrowser

- GTrack
  - generic file format for all types of genomic tracks
- Warping the Galaxy
  - Example: custom web tool definition framework

The Genomics Task Force



Knut Liestøl
Marit Holden
Trevor Clancy
Sigve Nakken
Ståle Nygård
Morten Johansen
Vegard Nygaard
Merete M. Worren
Daniel Vodák
...

The Genomics Task Force

- 4 professors of Biology/Bioinformatics/Statistics (via Statistics for Innovation)
- 3 researchers of Statistics from Norwegian Computing Center
- 4 postdocs of Biology/Bioinformatics/Statistics
- 4 PhD students of Bioinformatics/Statistics
- 5 bioinformaticians connected to the core facility
- 5 master students of (Bio)informatics

The Norwegian Radium Hospital
& University of Oslo

The Genomic HyperBrowser

- Implemented on top of Galaxy
- Launched in Dec 2010:

  Sandve, G. K., Gundersen, S., Rydbeck, H., Glad, I. K., et al. The Genomic HyperBrowser: inferential genomics at the sequence level. Genome biology 11, R121 (2010).

- The Genomic HyperBrowser provides hypothesis testing and descriptive statistics on any:
  - *single genomic track*
  - *pair of genomic tracks*
  - *more than two genomic tracks*
- Available at http://hyperbrowser.uio.no

Genomic track:
Anything that can be positioned on a reference genome

Example analysis

- Track 1:
  - 57 genomic regions associated with multiple sclerosis (MS) from a recent GWAS study:
    - *Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214-219 (2011)*
  - One BED-file with 57 regions:

```
chr1   2396586     2827369
chr1   92014277    93306499
chr1   100983315   101455310
chr1   116832232   116909542
chr1   190733439   190814781
chr1   199128354   199336605
...
```

Example analysis (cont.)

- Track 2:
  - Genome-wide chromatin profiling from another recent Nature paper:
    - *Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49 (2011).*
  - We want to compare active promoter regions in B-lymphoblastoid cells (GM12878) with those in normal epidermal keratinocytes (NHEK).
  - These are two separate datasets, already included in the track repository (> 200 000 tracks) in the HyperBrowser.
  - We want to combine them into one case/control track. The HyperBrowser has a webtool for this (around 90 custom Galaxy tools)

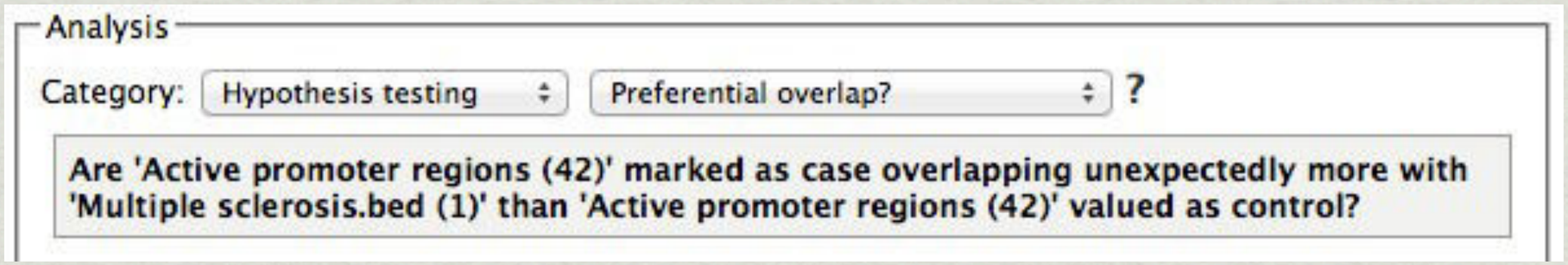
Example analysis (cont.)

- Track 2 (cont.):
  - We end up with a GTrack file (which contains headers specifying that this is a case/control track). Here 1 (case) is GM12878 and 0 (control) is NHEK:

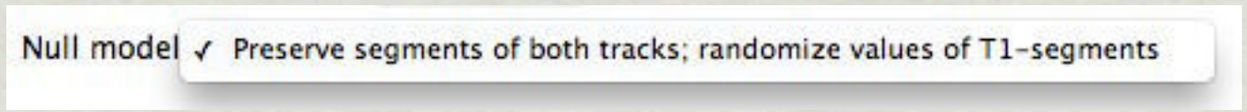
```
##gtrack version: 1.0
##track type: valued segments
##value type: binary
###seqid  start   end     value
chr1      29737   29937   0
chr1      715337  715537  0
chr1      715537  715737  1
chr1      752737  753137  0
chr1      761737  761937  0
chr1      894537  894737  1
...
```

Example analysis (cont.)

- We want to find whether MS-related regions overlap with active promoter regions more in B- lymphoblastoid cells (GM12878) than in normal epidermal keratinocytes (NHEK)
- In the Genomic HyperBrowser, we select tracks 1 and 2 as specified before, and the following analysis (out of around 100 different analyses):



Null model selection



- A null model is a model of the observed system where the effect under study has been "nullified" in some way
- Defined via:
  - Preservation rule
  - Stochastic process of non-preserved elements
- The null model should reflect biological realism, but also allow sufficient variation to permit the construction of tests.
- The results of a hypothesis test is highly dependent on the null model
- The most difficult choice when calculating p-values. Often overlooked
- In this example, only one null model has been implemented

Answer

You asked:

Are 'Active promoter regions' marked as case overlapping unexpectedly more with 'Multiple sclerosis.bed' than 'Active promoter regions' valued as control?

Simplistic answer:

Yes - the data suggests this (p-value: 0.004975)

Precise answer:

The p-value is 0.004975.

Low p-values are evidence against H0.

The test was also performed for each bin separately, resulting in 0 significant bins out of 26, at 10% FDR* (17 bins excluded from FDR-analysis due to lacking p-values).

Please note that both the effect size and the p-value should be considered in order to assess the practical significance of a result.

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

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

Preserve segments of both tracks; randomize values of T1-segments

---

The **test statistic** used is:

Main result of analysis

The value of the test statistic is 105 737.

The p-values may be subject to further parameter choices, which are listed in the run description. See note for a more complete description of the test.

See full details of the results in table form.

Answer (local results)

Region	P-value	FDR-adjusted p-values	Test statistic: Main result	Mean of null distribution	Median of null distribution	Standard deviation of null distribution	Difference from mean	Number of Monte Carlo samples
chr1:1-121535434 [chr1p]	0.09967	0.3396	5 200	1 549.	1600.0	2 782.	3 651.	300
chr1:142535435-249250621 [chr1q]	1.0	1.0	0	0.0	0.0	0.0	0.0	100
chr2:1-90545103 [chr2p]	0.3168	0.6865	1 000	380.0	600.0	923.3	620.0	100
chr2:95326172-243199373 [chr2q]	0.1045	0.3396	3 400	788.0	600.0	1 842.	2612.0	200
chr3:1-90504854 [chr3p]	1.0	1.0	0	0.0	0.0	0.0	0.0	100
chr3:93504855-198022430 [chr3q]	1.0	1.0	-4 600	680.0	1000.0	2 061.	-5280.0	100
chr4:1-49660117 [chr4p]	None	nan	0	0.0	0.0	0.0	0.0	100
chr4:52660118-191154276 [chr4q]	0.04790	0.3311	4 000	248.0	400.0	2 129.	3752.0	500
chr5:1-46405641 [chr5p]	0.9901	1.0	-4 000	920.0	3200.0	3 543.	-4920.0	100
chr5:49405642-180915260 [chr5q]	0.07481	0.3311	10 072	1 933.	3128.0	5 972.	8 139.	400
chr6:1-58830166 [chr6p]	None	nan	0	0.0	0.0	0.0	0.0	100
chr6:61830167-171115067 [chr6q]	0.2277	0.5383	7 800	1488.0	1000.0	6 248.	6312.0	100
chr7:1-58054331 [chr7p]	None	nan	0	0.0	0.0	0.0	0.0	100
chr7:61967158-159138663 [chr7q]	0.7030	1.0	600	1052.0	1000.0	1 481.	-452.0	100
chr8:1-43838887 [chr8p]	None	nan	0	0.0	0.0	0.0	0.0	100
chr8:46838888-146364022 [chr8q]	0.7624	1.0	200	728.0	1000.0	1 037.	-528.0	100

Publishable results

- *Disanto, G., Sandve, G. K., Berlanga-Taylor, A. J., Morahan, J. M., et al. Genomic regions associated with multiple sclerosis are active in B cells. PLoS One 7, e32281 (2012).*

- From the same authors (In total 3 publications using the HyperBrowser in half a year, 2 more coming):

It is possible to publish based on public datasets (all 3).

2 of these required some custom extensions that we helped with, the third used the HyperBrowser without modifications

Of course powerful to analyse own datasets

Disanto, G, G K Sandve, A J Berlanga-Taylor, G Ragnedda, J M Morahan, C T Watson, G Giovannoni, G C Ebers, and S V Ramagopalan. "Vitamin D Receptor Binding, Chromatin States and Association with Multiple Sclerosis." Human molecular genetics 21, no. 16 (2012): 3575-3586.

Watson, C. T., Disanto, G., Sandve, G. K., Breden, F., et al. Age-Associated Hyper-Methylated Regions in the Human Brain Overlap with Bivalent Chromatin Domains. PLOS ONE ***7,*** *e43840 (2012).*

The Genomics Task Force

- High competence in bioinformatics and statistics
- Can implement custom-tailored hypothesis tests, analyses, result plots and advanced Galaxy tools
- Everything will be available from a simple user interface in the Genomic HyperBrowser (http://hyperbrowser.uio.no)
- We are interested in collaborations (from simple extensions of existing functionality to new research collaborations).
- E-mail: hyperbrowser-requests@usit.uio.no

Overview

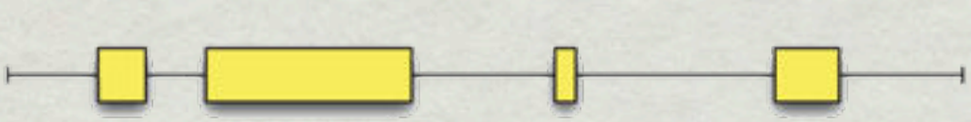
- The Genomics Task Force
  - statistical analysis of genomic tracks, employing:

The Genomic HyperBrowser

- GTrack
  - generic file format for all types of genomic tracks
- Warping the Galaxy
  - Example: custom web tool definition framework

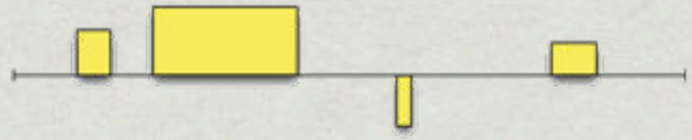
Track types

- Developed in:
  - *Gundersen, S., Kalas, M., Abul, O., Frigessi, A., et al. Identifying elemental genomic track types and representing them uniformly. BMC bioinformatics 12, 494 (2011).*



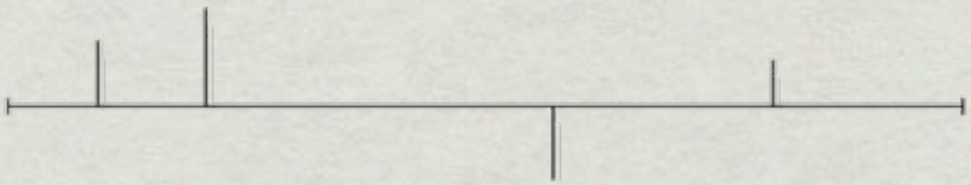
Segments (S)

BED



Valued Segments (VS)

BedGraph

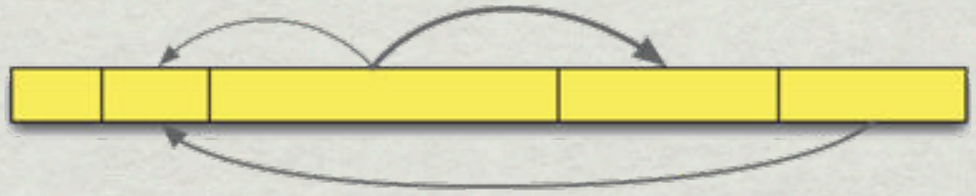


Valued Points (VP)

WIG

Function (F)

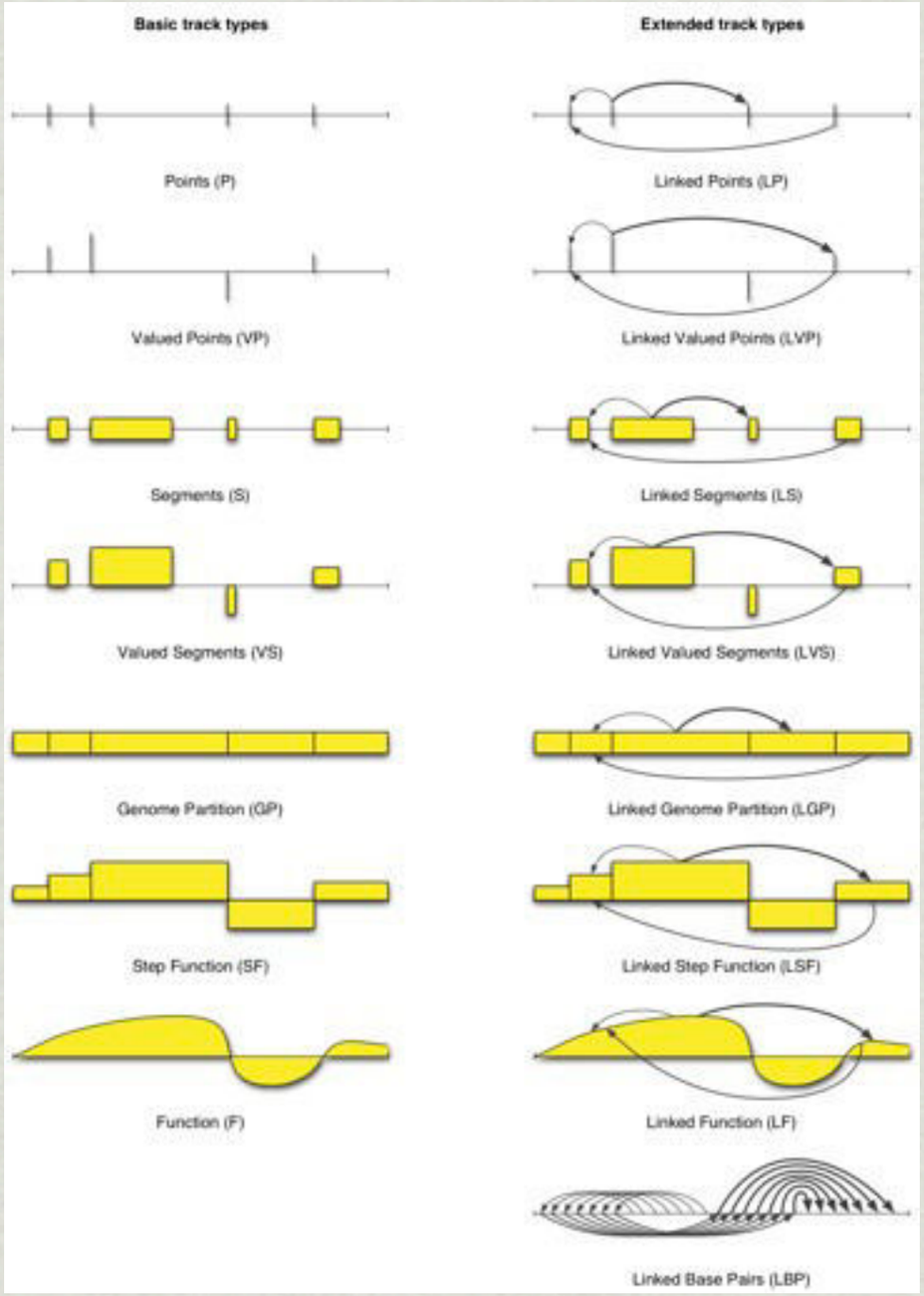
WIG



Linked Genome Partition (LGP)

Hi-C (no format yet)

15 track types



GTrack

- "Genomic Track" or "Generic Track"
- Supports all fifteen track types (using practical representations)
- Humanly readable
- Reasonably simple to parse
- Extensible for future extensions (can add custom columns)
- Research groups or tool developers can define their own specific GTrack subtypes
- GTrack fully supported by the HyperBrowser
- http://www.gtrack.no
- In the process of creating standalone Python library

GTrack example (simple)

```
chr1        2396586     2827369
chr1        92014277    93306499
chr1        100983315   101455310
chr1        116832232   116909542
chr1        190733439   190814781
chr1        199128354   199336605
...
```

Basic three-column BED file is directly compatible with a GTrack file

GTrack example (intermediate)

```
##gtrack version: 1.0
##track type: valued segments
##value type: binary
###seqid  start       end         value
chr1      2396586     2827369     1
chr1      92014277    93306499    1
chr1      100983315   101455310   0
chr1      116832232   116909542   0
chr1      190733439   190814781   1
chr1      199128354   199336605   0
...
```

More than standard 3-column BED, needs column specification line, and other basic headers lines

GTrack example (advanced)

```
##gtrack version: 1.0
##track type: linked genome partition
##edge weights: true
###end    id    edges

####seqid=chr1; start=1000; end=2000
1015        a       c=1.3,d=0.1
1060        b       a=1.0
1154        c       a=1.3
1267        d       .
...
```

3D data set

Overview

- The Genomics Task Force
  - statistical analysis of genomic tracks, employing:

The Genomic HyperBrowser

- GTrack
  - generic file format for all types of genomic tracks
- Warping the Galaxy
  - Example: custom web tool definition framework

Warping the Galaxy

- custom modifications of Galaxy

- Custom web tool definition framework:
  - Based on our own mako code (genericTool.mako)
  - New web tools are generated based on new python classes
  - We have defined an interface (GeneralGuiTool.py), which we subclass for each tool
  - We have defined a standard set of reusable GUI elements (textbox, selection box, genome and track lists, etc.)
  - The contents of the GUI can be dynamically changed according to the python code, (almost) anything is possible

Custom Galaxy webtools (example)

Create GTrack file

Select input source: Tabular file from history

Select tabular file: 14 – Active promoter regions

Character to use to split lines into columns: Tab

Number of lines to skip (from front): 6

Column selection method: Select individual columns

Select the name for column #1: seqid

Select the name for column #2: start

Select the name for column #3: end

Select the name for column #4: value

Select a specific genome? Yes

Genome build: Human Feb. 2009 (hg19/GRCh37)

Dependency between choices

Table created from dataset in history

seqid	start	end	value
chr1	28537	29337	1
chr1	28537	29737	0
chr1	540737	540937	1
chr1	713337	715337	0
chr1	713537	715137	1
chr1	715537	715737	0
chr1	761937	763537	0
chr1	761937	763537	1
chr1	893937	894737	0
chr1	893937	894737	1
chr1	895737	895937	1

Current track type (based on the selected column names):

Valued Segments (VS)

Warping the Galaxy

- custom modifications of Galaxy

- Advantages:
  - No need for knowledge on GUI coding or Galaxy to create a web tool
  - Quick prototyping => Simplifies sharing of new functionality with others
  - Independent of Galaxy, and could in principle be connected to another system
  - We have full control and can expand as needed
- Disadvantages:
  - Minimal workflow support

Warping the Galaxy

- custom modifications of Galaxy

- Other tweaks:
  - Separate, monolithic codebase (850 files, 72500 lines of code)
  - Separate data collection
  - Separate result files (linked to by HTML)
  - Batch run interface (of all analyses and tools)
  - And more...

Support

RIKSHOSPITALET

EMBIO Styringsgruppen for Molecular Life Sciences ved UiO

UNIVERSITY OF OSLO

NR Norsk Regnesentral NORWEGIAN COMPUTING CENTER

FUGE Functional genomics

(sfi)² Statistics for Innovation

sfi Centre for Research-based Innovation

Questions/comments?