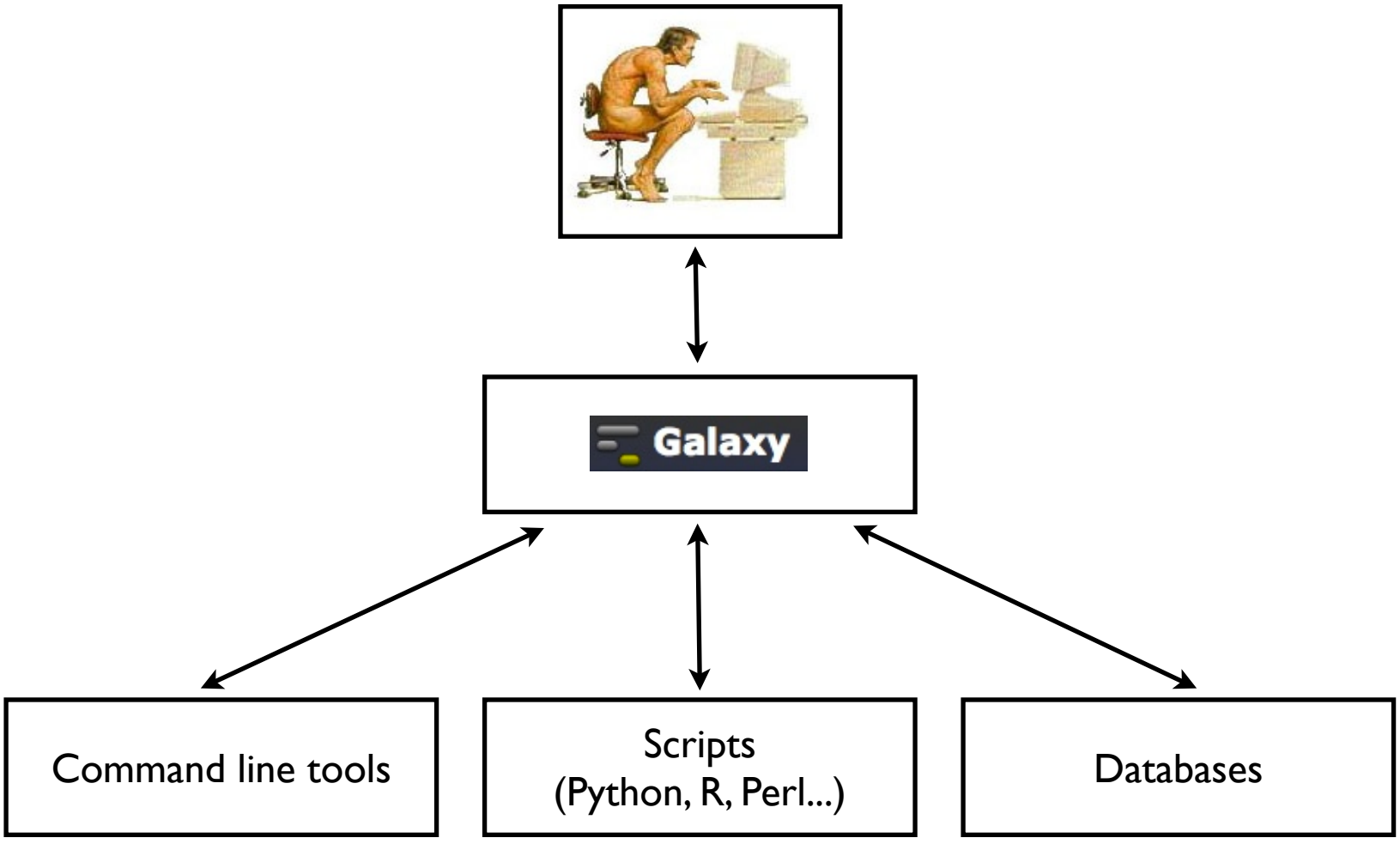
Overview of Galaxy

Galaxy vs the Genomic HyperBrowser

- Galaxy is a general research framework developed at Penn State University
- The Genomic HyperBrowser is a system focused on statistical analysis of genomic data. It is developed locally, in a collaboration between the Norwegian Radium Hospital, Statistics for Innovation (at the Norwegian Computing Center) and the University of Oslo.
- The Genomic HyperBrowser is built on top of Galaxy

Galaxy: overview



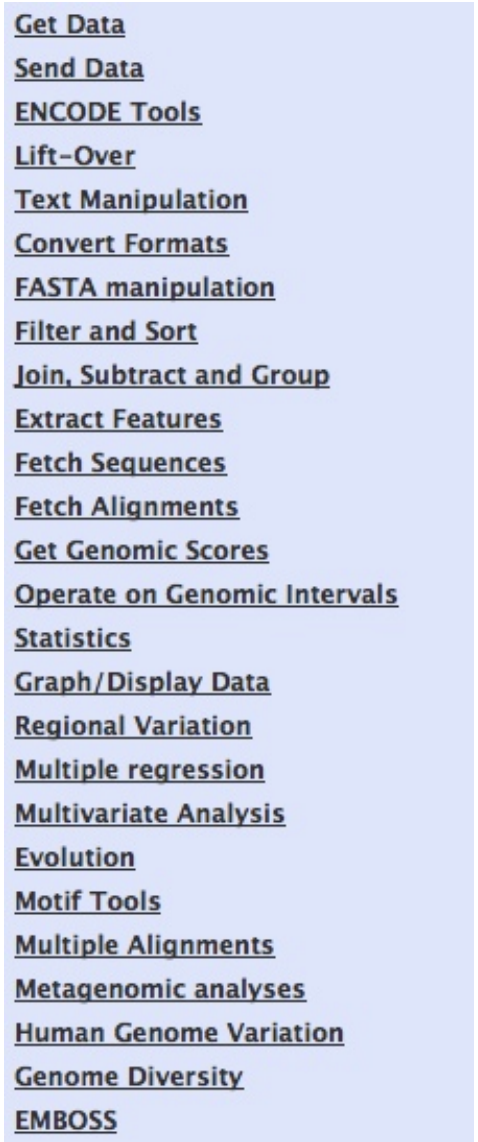
http://usegalaxy.org

Galaxy: tools

- Includes a large range of tools

  (these are just tool categories,

  each contains from 1-30 tools)

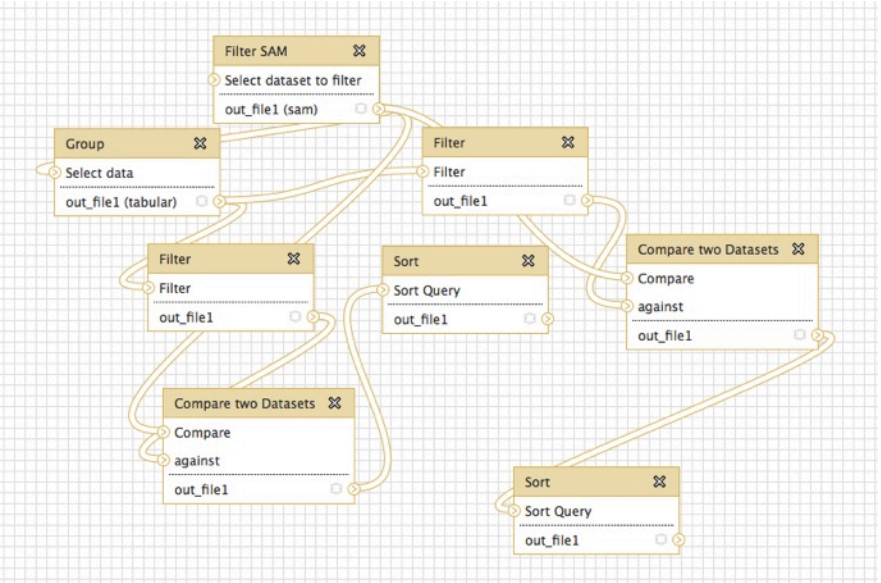


Galaxy: history

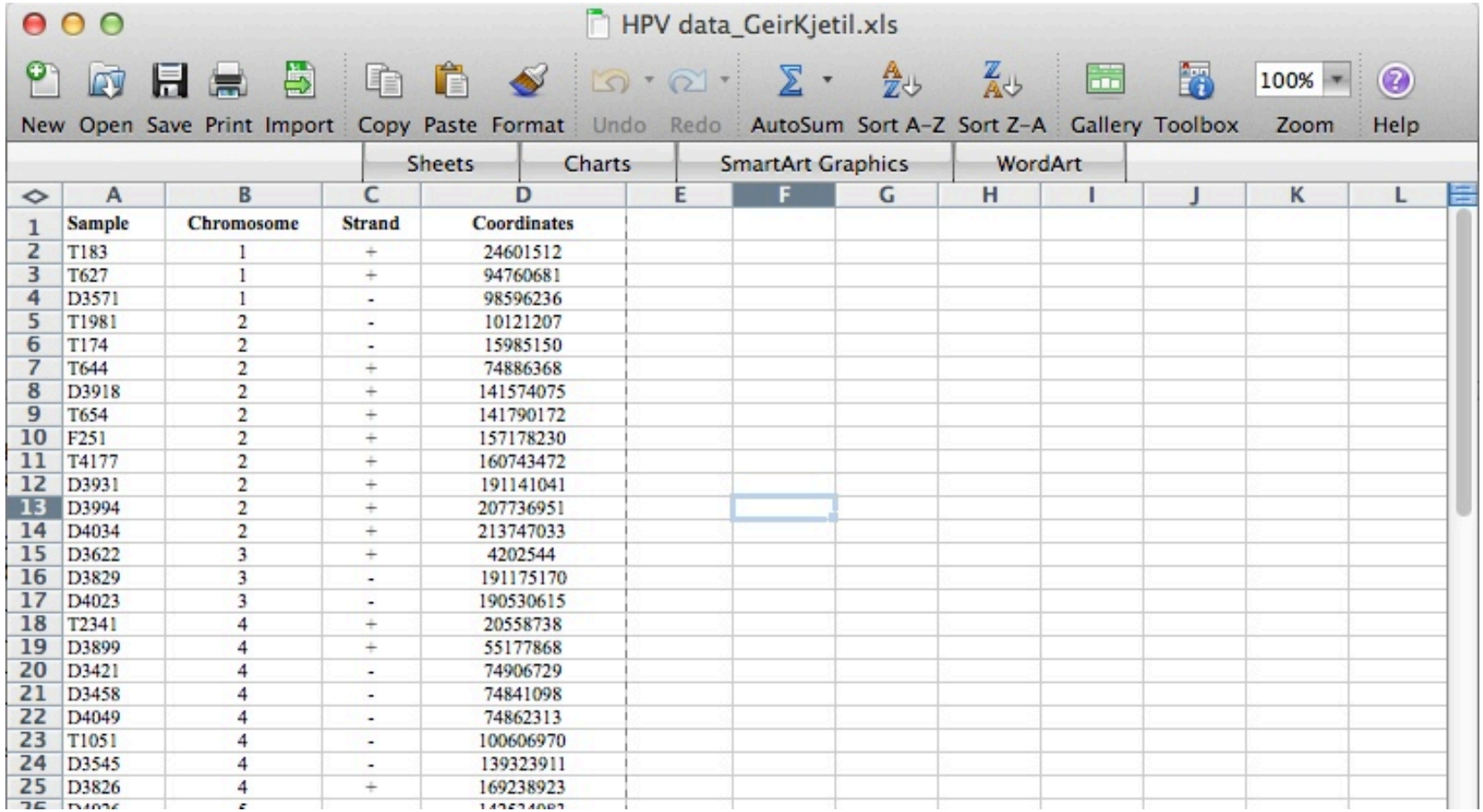
- Input files/datasets
- Intermediate files/datasets
- Analysis output
- Every tool takes input from the history and generates output to the history
- One may create a history for each investigation
- Histories may be shared with others, privately or publicly

Galaxy: workflows

- Tools can be arranged into workflows
- Output and input data are routed with lines
- Other parameters of the tools may be specified
- The complete analyses is then started with a single click
- Workflows can be shared



Galaxy:
data upload (in the real world)



	A	B	C	D
1	Sample	Chromosome	Strand	Coordinates
2	T183	1	+	24601512
3	T627	1	+	94760681
4	D3571	1	-	98596236
5	T1981	2	-	10121207
6	T174	2	-	15985150
7	T644	2	+	74886368
8	D3918	2	+	141574075
9	T654	2	+	141790172
10	F251	2	+	157178230
11	T4177	2	+	160743472
12	D3931	2	+	191141041
13	D3994	2	+	207736951
14	D4034	2	+	213747033
15	D3622	3	+	4202544
16	D3829	3	-	191175170
17	D4023	3	-	190530615
18	T2341	4	+	20558738
19	D3899	4	+	55177868
20	D3421	4	-	74906729
21	D3458	4	-	74841098
22	D4049	4	-	74862313
23	T1051	4	-	100606970
24	D3545	4	-	139323911
25	D3826	4	+	169238923

Commonly used format: Excel!

Galaxy:
what we want

```
chr1  24601511   24601512   T183  0   +
chr1  94760680   94760681   T627  0   +
chr1  98596235   98596236   D3571 0   -
chr2  10121206   10121207   T1981 0   -
chr2  15985149   15985150   T174  0   -
chr2  74886367   74886368   T644  0   +
chr2  141574074  141574075  D3918 0   +
chr2  141790171  141790172  T654  0   +
chr2  157178229  157178230  F251  0   +
chr2  160743471  160743472  T4177 0   +
chr2  191141040  191141041  D3931 0   +
```

BED file

Galaxy:
what we want

seqid	start	end	name	score	strand
chr1	24601511	24601512	T183	0	+
chr1	94760680	94760681	T627	0	+
chr1	98596235	98596236	D3571	0	-
chr2	10121206	10121207	T1981	0	-
chr2	15985149	15985150	T174	0	-
chr2	74886367	74886368	T644	0	+
chr2	141574074	141574075	D3918	0	+
chr2	141790171	141790172	T654	0	+
chr2	157178229	157178230	F251	0	+
chr2	160743471	160743472	T4177	0	+
chr2	191141040	191141041	D3931	0	+

BED file

Demo 1

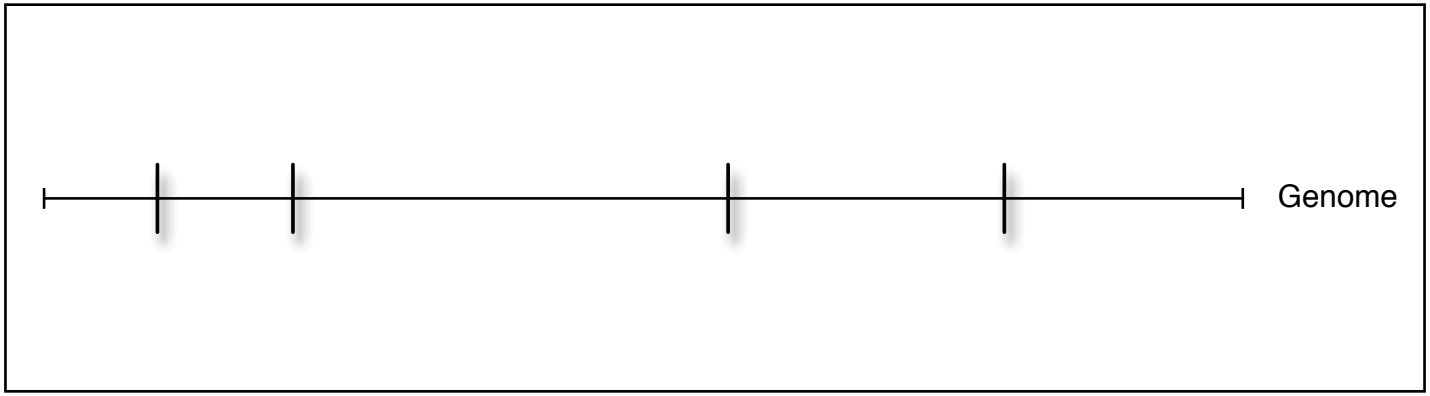
- Convert from Excel document to BED file using Galaxy tools

Demo 2

- Convert from Excel document to BED file using GTrack
- GTrack: Universal file format for (almost) any kind of genomic tracks (created by us)

Track type: Points

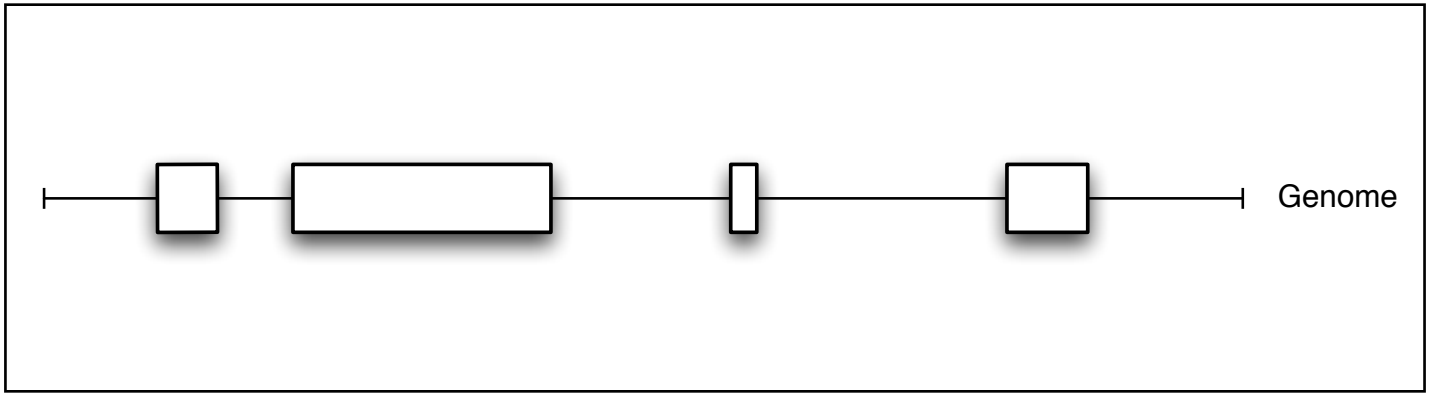
- HPVs visualized genome-wide:



- The only information here is the positions

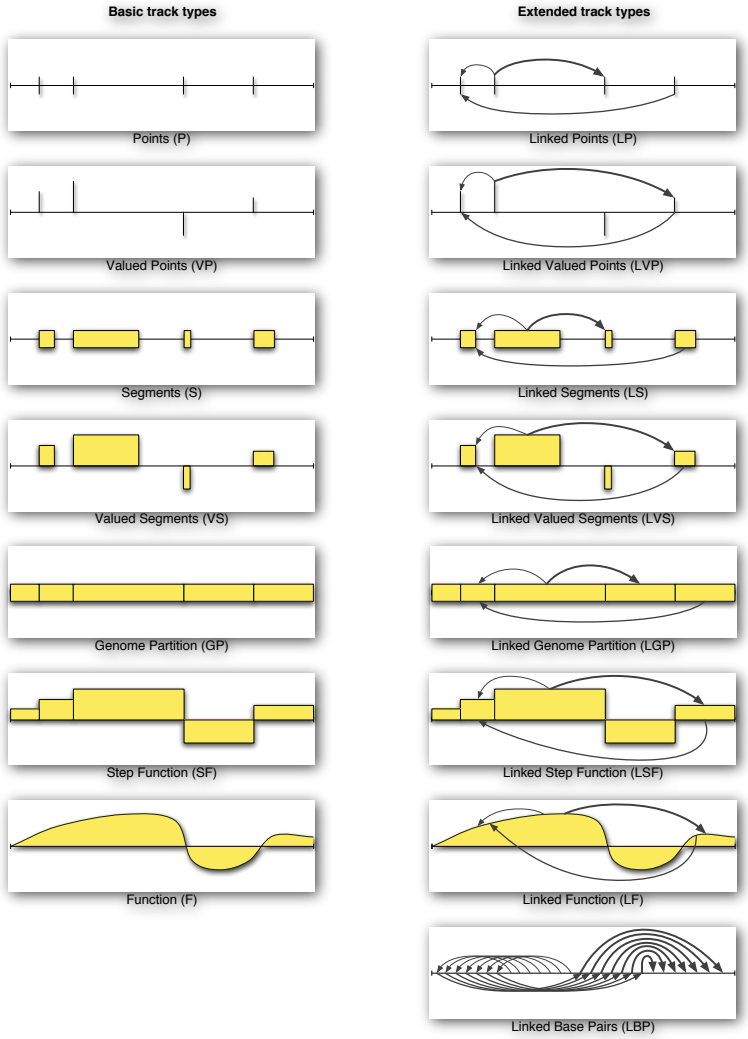
Track type: Segments

- Genes visualized genome-wide:



- The information is positions & lengths

Track types: all of them



- We have defined 15 track types
- All track types are supported by the HyperBrowser
- Each *single track type* defines a set of analyses appropriate for that track type (e.g. counting, coverage)
- Each *pair of track types* defines another set of relational analyses (e.g. overlap, correlation...) specific to that combination

Which track types do existing formats support?

BED

Format	Ref.	Data	Repr.	P	S	VP	VS	GP	SF	F	L	Strand	#Cols	Value type
GFF3/GTF	[2]	General	Tab.	√(1)	√	√(1)	√				(2)	√	9	Float (3)
BED/bigBed	[4]	General	Tab./ Bin.	√(1)	√	√(1)	√				(2)	√	3-12	Int(0-1000) /string (4)
BED15	[4]	Microarray	Tab.			√(1)	√				(2)	√	15	List of floats (5)
bedGraph	[4]	General	Tab.			√(1)	√						4	Float
WIG/bigWig (fixedStep)	[8]	General	Tab./ Bin.			√	√		√	√			1	Float
WIG/bigWig (variableStep)	[8]	General	Tab./ Bin.			√	√						2	Float
CNT	[36]	Copy number	Tab.			√							4	Float
Personal Genome SNP	[4]	Variation	Tab.			√(1)	√						7	String (6)
VCF	[37]	Variation	Tab.			√	√						≥ 8	String (6) (3)
GVF	[6]	General/ Variation	Tab.	√(1)	√	√(1)	√				(2)	√	9	Float (3)
PSL	[4]	Alignment	Tab.		√		√					√	21	Int (7)
SAM/BAM	[38]	Alignment	Tab./ Bin.		√		√					√	11	Int /string (8)
BioHDF	[39]	Alignment	Bin.		√		√					√	11	Int /string (8)
MAF	[4]	Multiple Alignment	Tab.		√		√				(9)	√	2-7	Float /string (8)
FASTA	[40]	Sequence	Text							√			N/A	Char
DAS XML	[12]	General	XML	√(1)	√	√(1)	√				(2)	√	N/A	Float
BioXSD 1.0	[16]	General	XML	√(10)	√(10)	√(10)	√(10)				√(11)	√	N/A	Float (12)
USeq	[19]	General	Bin.	√	√	√	√					√	N/A	Int/float/string
Genomedata	[41]	General	Bin.			√	√		√	√			N/A	Int/float/char

GTrack format

- All track types are supported by GTrack (may replace most of the formats of the last slide)
- Supports a variable number of columns
- Fully supported by the Genomic HyperBrowser
- HyperBrowser includes 7 specific GTrack tools, including a tool for converting between GTrack and other file formats
- http://www.gtrack.no

GTrack example (simple)

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

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
```

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     .
```

Galaxy:
other tutorials

- For more tutorials and exercises, check out:

  http://wiki.g2.bx.psu.edu/Learn

UiO : University of Oslo

Lifeportal

- Galaxy installation at UiO, running on the Abel cluster
- Replacing the existing Bioportalen, with the same tools
- Additional NGS tools and others
- Should be in production Oct 1.
- Grand opening Oct. 9.

UiO : University of Oslo

Lifeportal

- Support:
  - Installing tools, technical issues: USIT
  - Biological investigations: Elixir.no
    - National project for life science infrastructure, aimed to be part of the european Elixir project
    - http://www.bioinfo.no/elixir