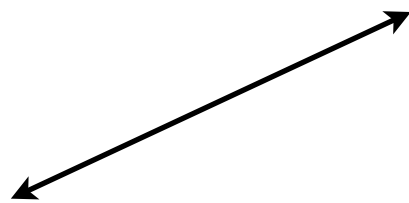
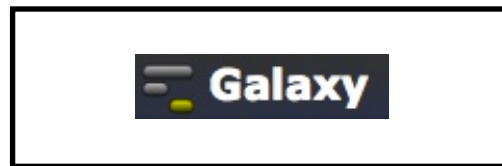
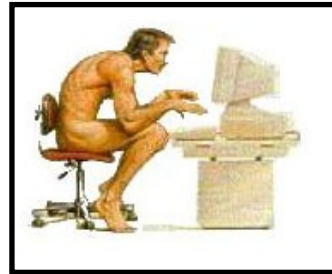


Overview of Galaxy

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

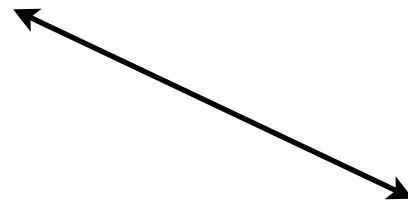
Galaxy: overview



Command line tools



Scripts
(Python, R, Perl...)



Databases

<http://usegalaxy.org>

Galaxy: tools

[Get Data](#)
[Send Data](#)
[ENCODE Tools](#)
[Lift-Over](#)
[Text Manipulation](#)
[Convert Formats](#)
[FASTA manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Get Genomic Scores](#)
[Operate on Genomic Intervals](#)
[Statistics](#)
[Graph/Display Data](#)
[Regional Variation](#)
[Multiple regression](#)
[Multivariate Analysis](#)
[Evolution](#)
[Motif Tools](#)
[Multiple Alignments](#)
[Metagenomic analyses](#)
[Human Genome Variation](#)
[Genome Diversity](#)
[EMBOSS](#)

- Includes a large range of tools
(these are just tool categories,
each contains from 1-30 tools)

NGS TOOLBOX BETA
[NGS: QC and manipulation](#)
[NGS: Mapping](#)
[NGS: SAM Tools](#)
[NGS: Indel Analysis](#)
[NGS: Peak Calling](#)
[NGS: RNA Analysis](#)
[NGS: Picard \(beta\)](#)

RGENETICS
[SNP/WGA: Data; Filters](#)
[SNP/WGA: QC; LD; Plots](#)
[SNP/WGA: Statistical Models](#)
[NGS: GATK Tools \(beta\)](#)
[NGS: Variant Detection](#)

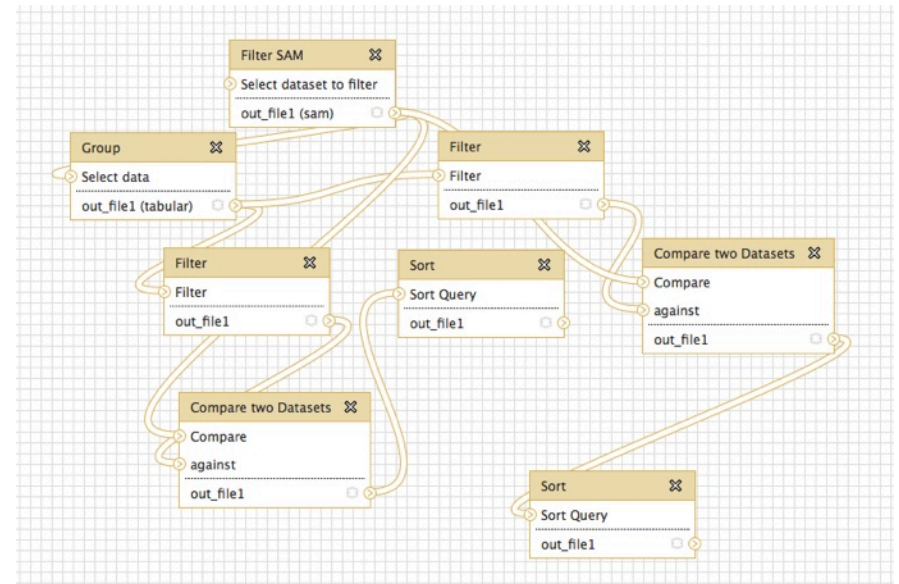
Galaxy: history

- 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

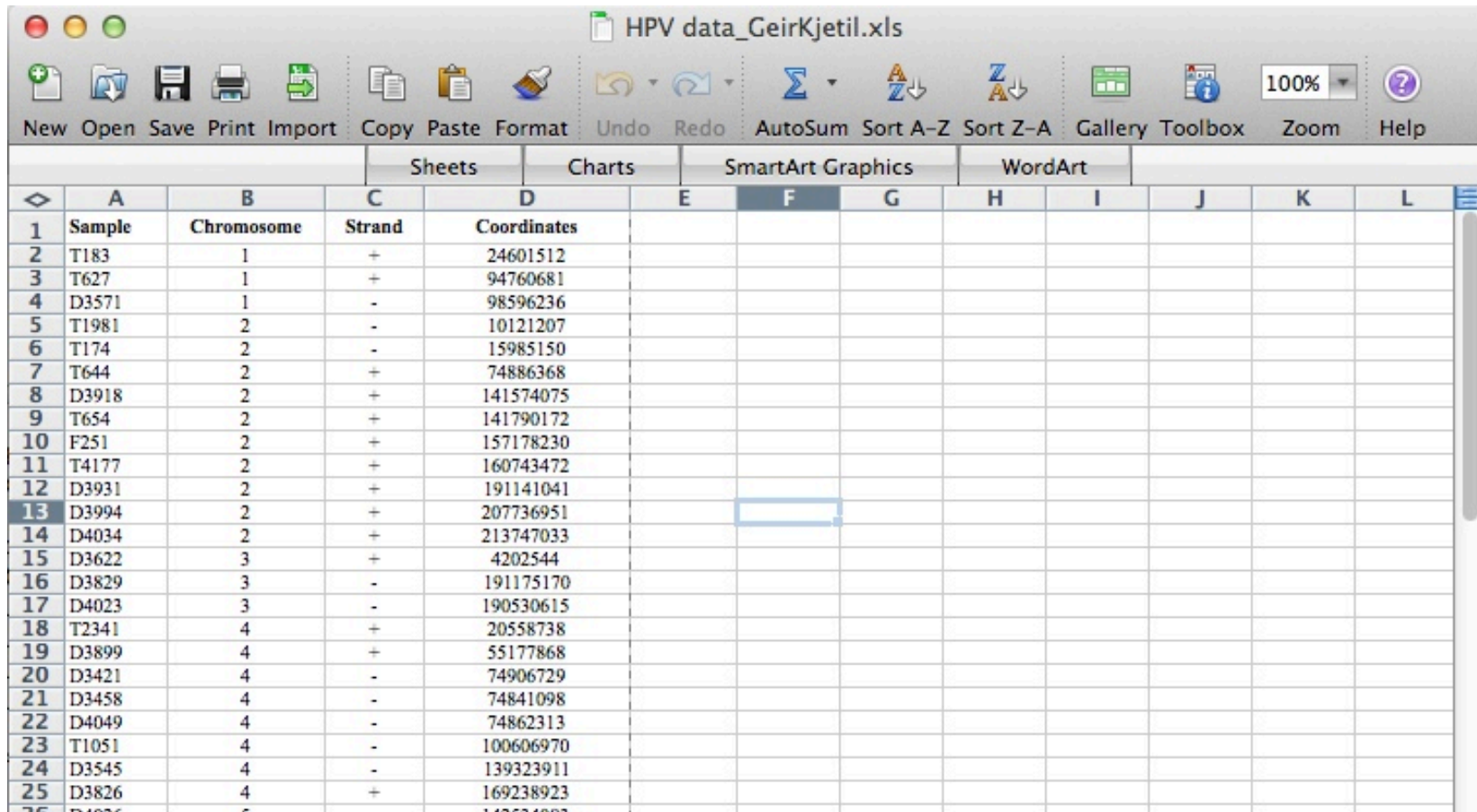
History 	
  Interesting investigation   53 bytes	
<u>17: Cut on data 3</u>   	
<u>8: Perform analysis</u>   	
<u>7: Perform analysis</u>   	
<u>6: Perform analysis</u>   	
<u>5: Perform analysis</u>   	
<u>4: Split BED file</u>   	
<u>3: Extract track</u>   	
<u>2: Extract track</u>   	
<u>1: Perform analysis</u>   	

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)



The image shows a screenshot of a Microsoft Excel spreadsheet titled "HPV data_GeirKjetil.xls". The spreadsheet contains a table with 25 rows and 5 columns. The columns are labeled "Sample", "Chromosome", "Strand", and "Coordinates". The data represents HPV samples with their corresponding chromosome, strand orientation, and genomic coordinates.

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

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

<u>seqid</u>	<u>start</u>	<u>end</u>	<u>name</u>	<u>score</u>	<u>strand</u>
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 I

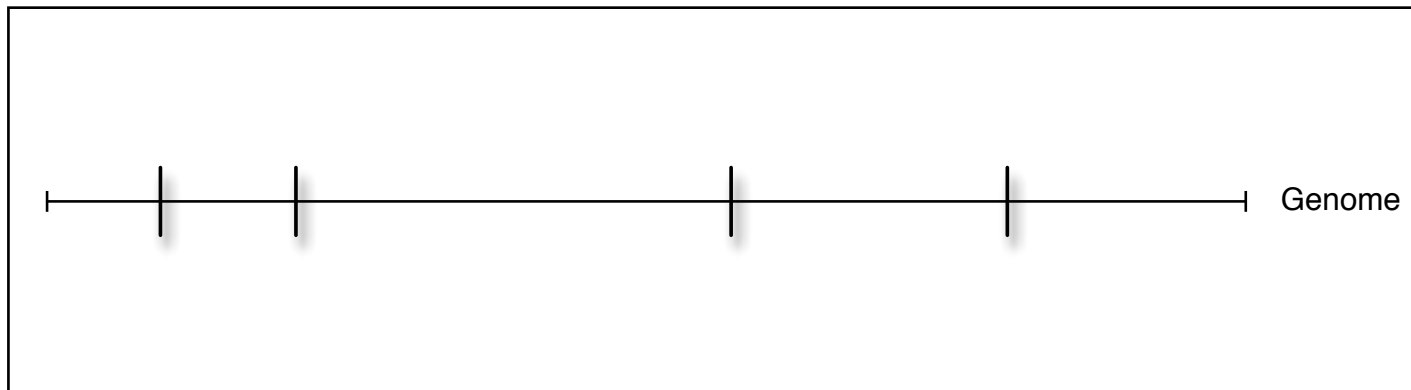
- Convert from Excel document to BED file using Galaxy tools

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

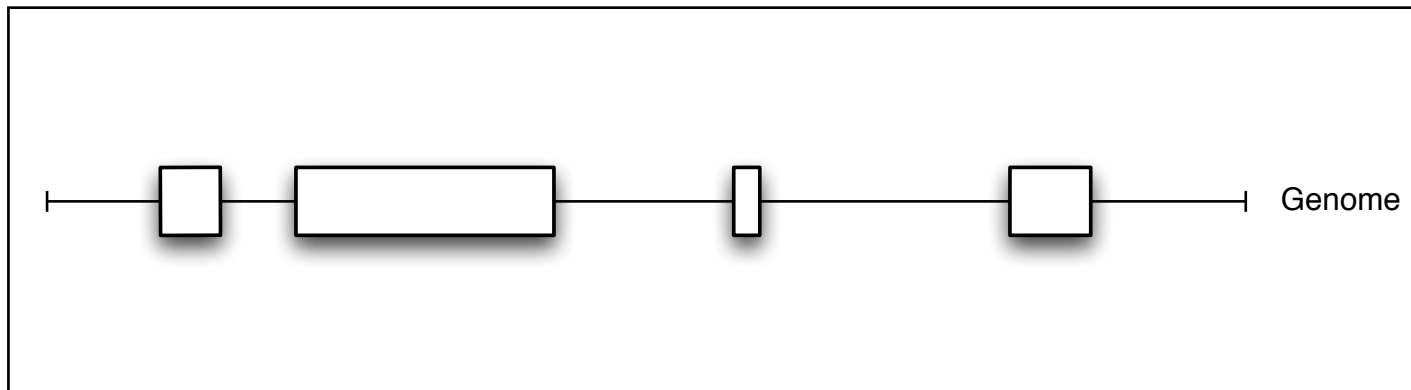
- HPV's visualized genome-wide:



- The only information here is the positions

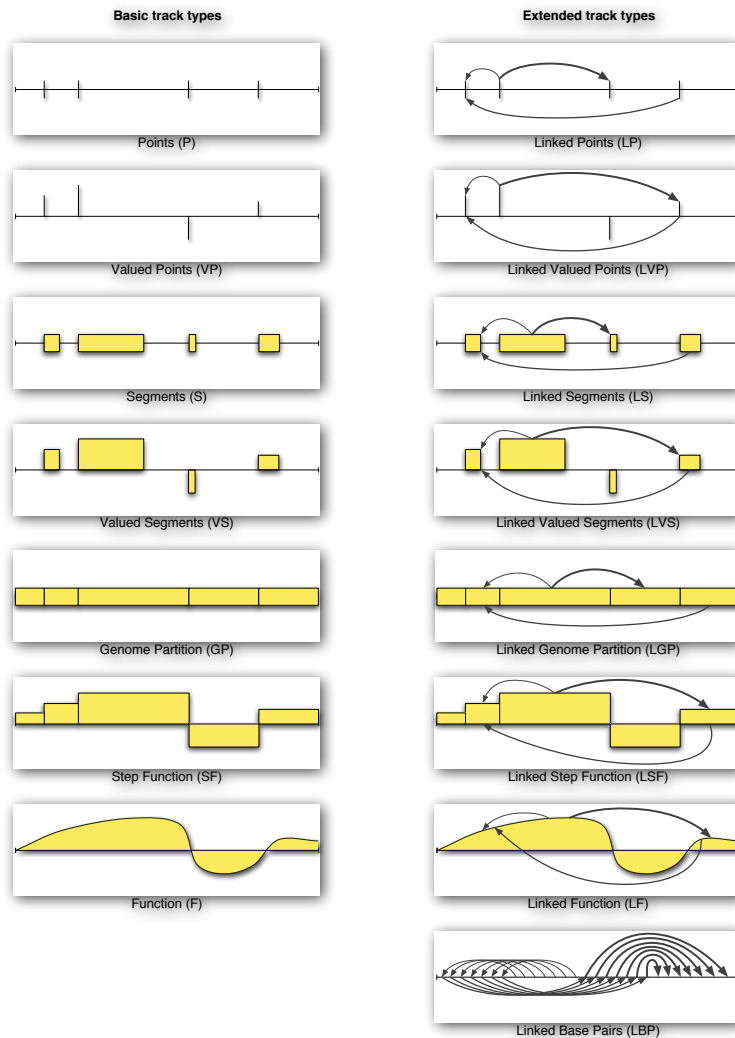
Track type: Segments

- 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>

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.

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>