ggbio: visualization tool kit for genomic data.

Tengfei Yin

Iowa State University

July 31, 2012

Outline

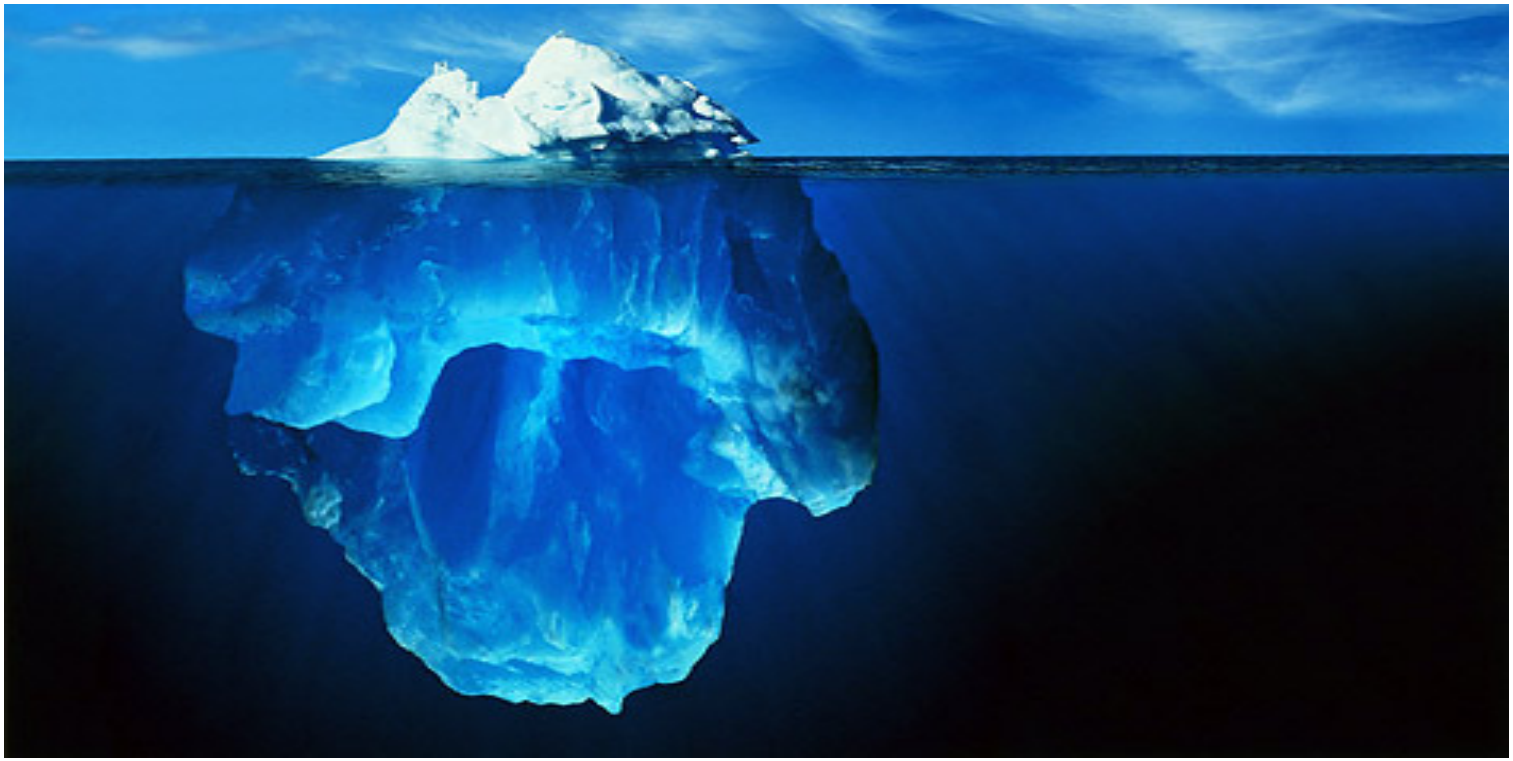
Outline

Outline

Tip of iceberg

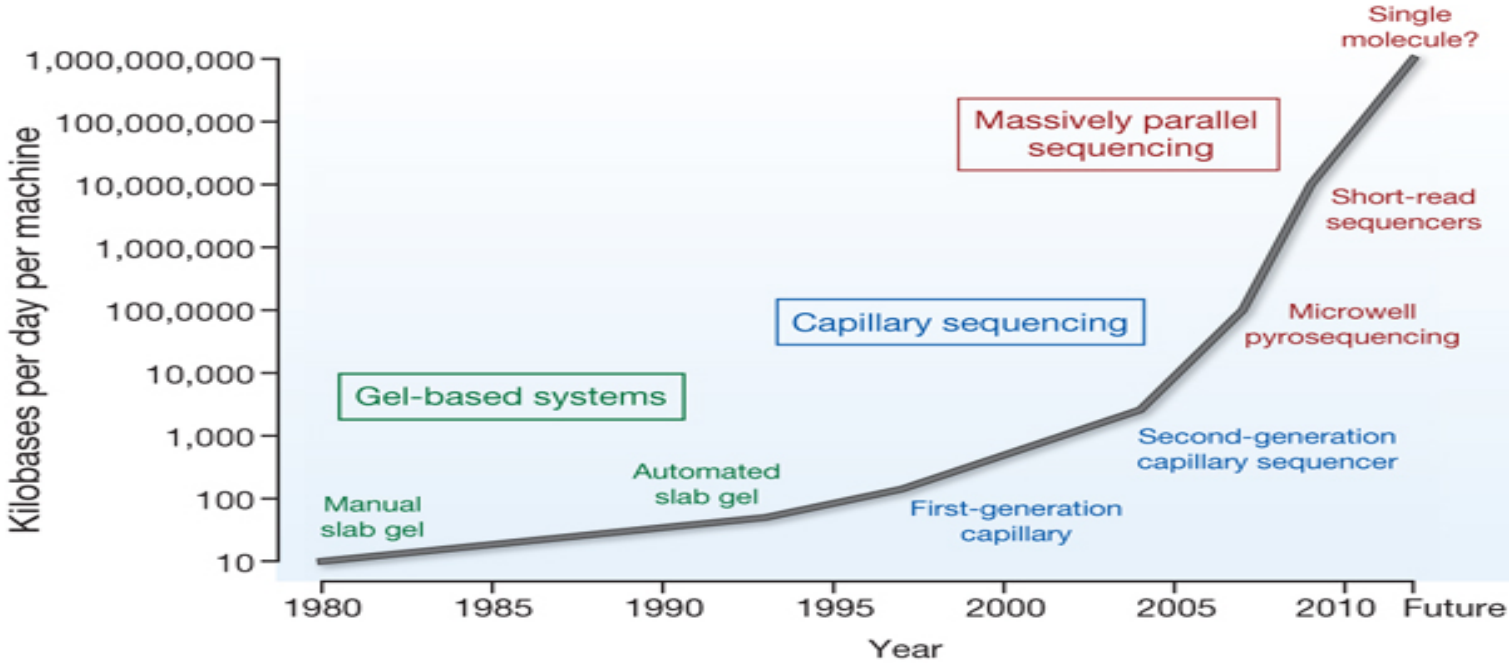
Definition

big data is a loosely-defined term used to describe data sets so large and complex that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage,search, sharing, analysis, and visualization(*wiki*).



Sequencing data: the iceberg.

- Next generation sequencing data is huge, and keep growing.
  - Raw image over TB.
  - Over GB results data per run.
- Analysis to find *tip* of that "Ice berg".
  - Analytical tool kits.
  - Visualization tool kits.



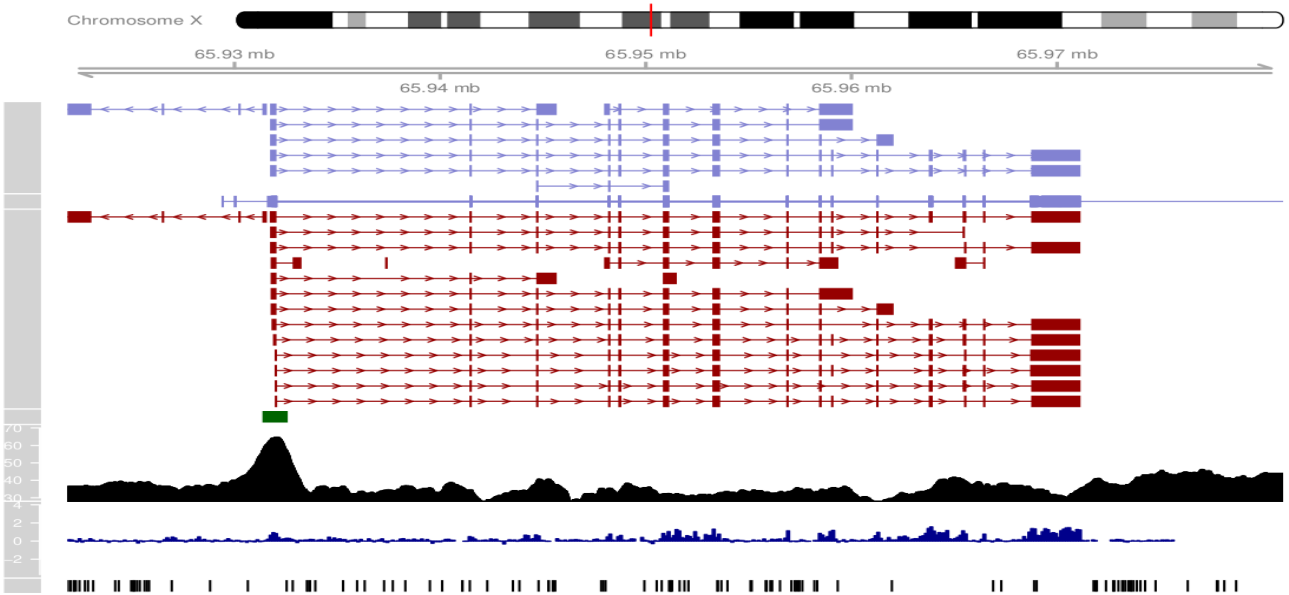
MR Stratton Nature 485

Review of current tools

- GenomeGraphs, Gviz(in R).
- Interactive Desktop-app: IGB/IGV(Java).
- Web based: UCSC genome browser, DNAnexus.
- Specialized: Circos.

Review of current tools

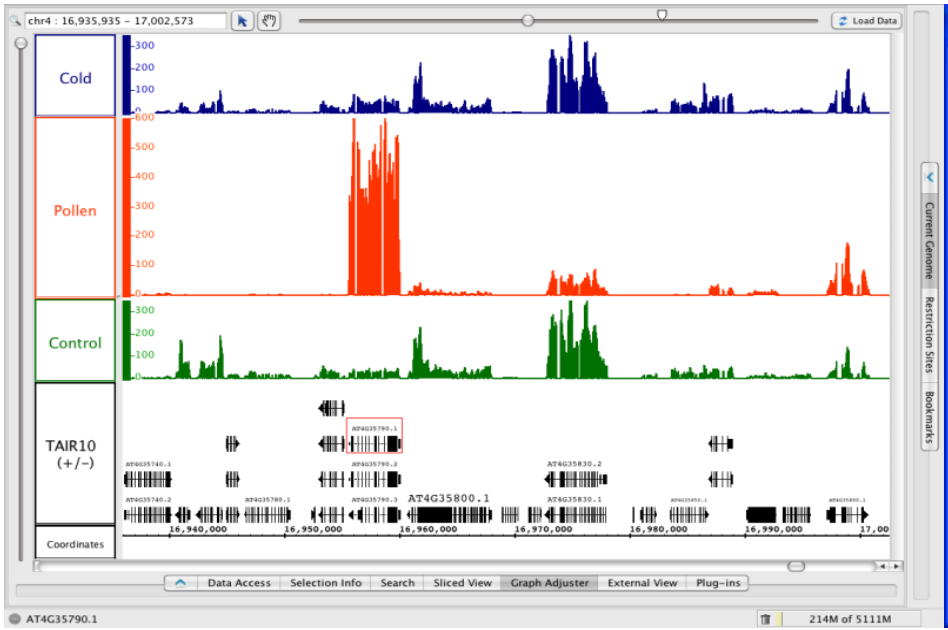
- GenomeGraphs, Gviz(in R).
- Interactive Desktop-app: IGB/IGV(Java).
- Web based: UCSC genome browser, DNAnexus.
- Specialized: Circos.

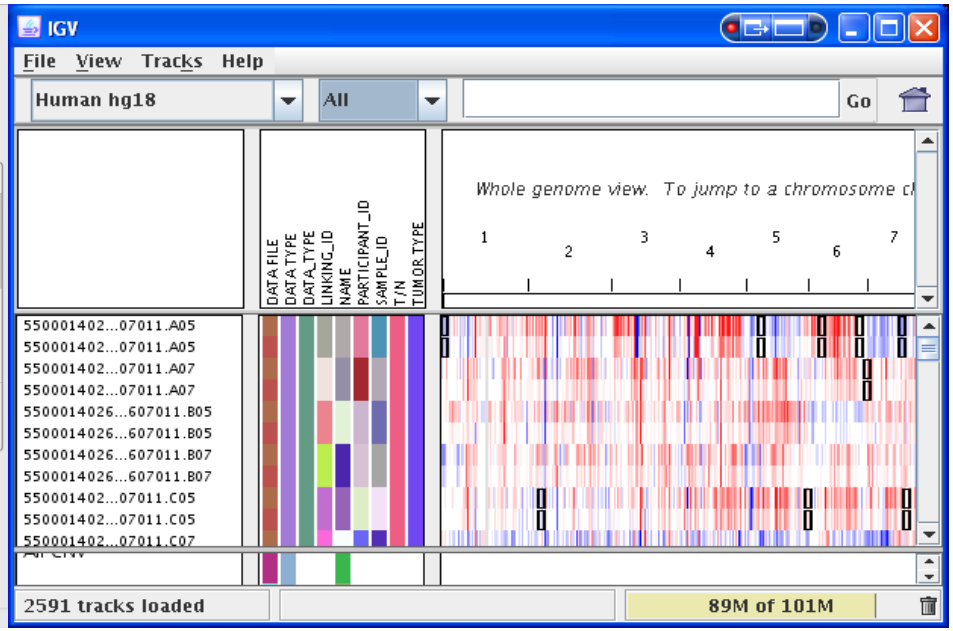


gviz package vignette.

Review of current tools

- GenomeGraphs, Gviz(in R).
- Interactive Desktop-app: IGB/IGV(Java).
- Web based: UCSC genome browser, DNAnexus.
- Specialized: Circos.

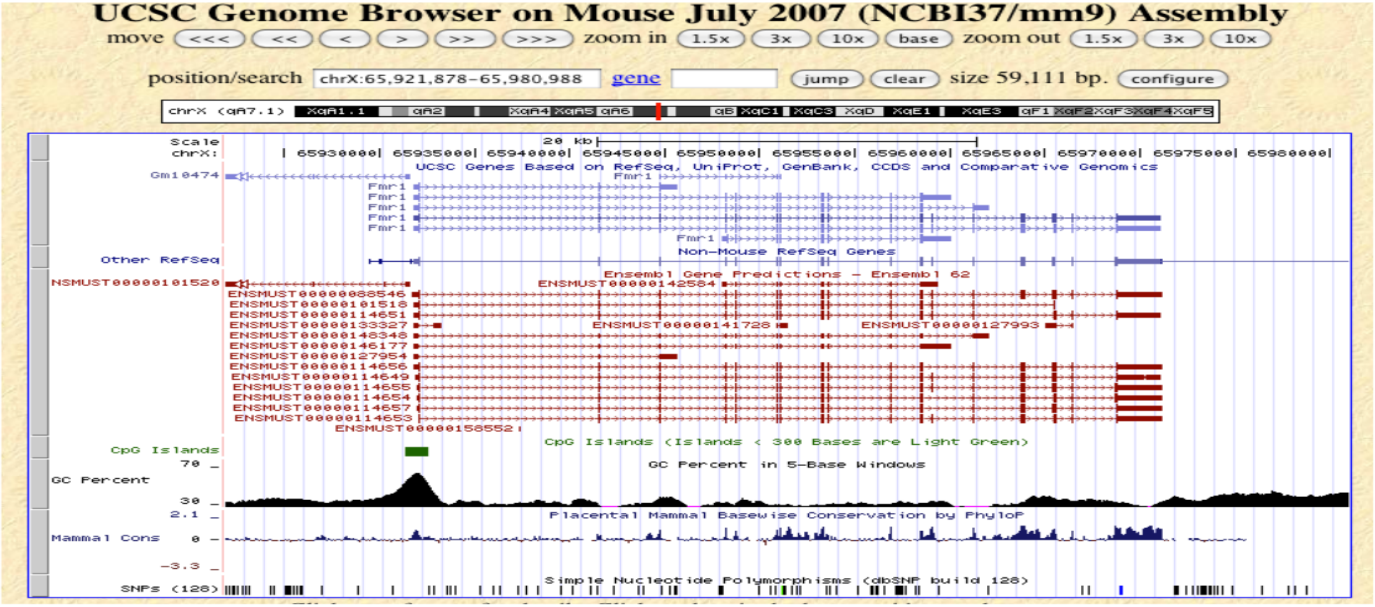




IGB(Integrated Genome Browser)/IGV(Integrative Genome Viewer) website

Review of current tools

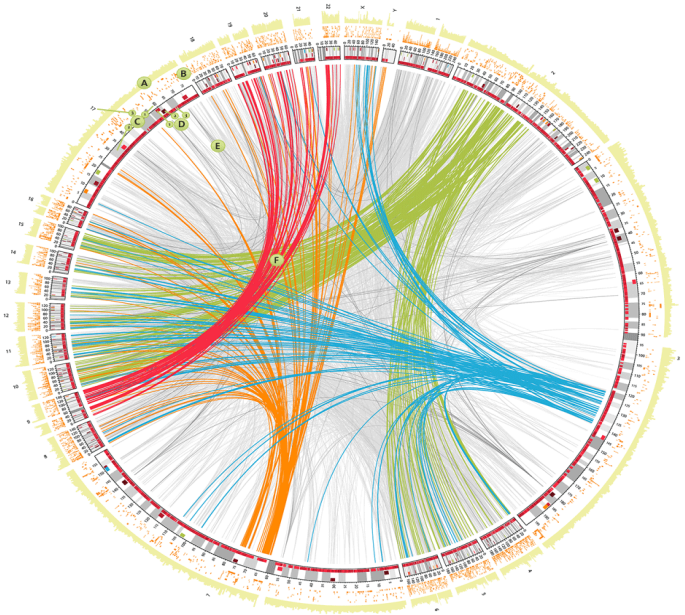
- GenomeGraphs, Gviz(in R).
- Interactive Desktop-app: IGB/IGV(Java).
- Web based: UCSC genome browser, DNAnexus.
- Specialized: Circos.



UCSC genome browser website

Review of current tools

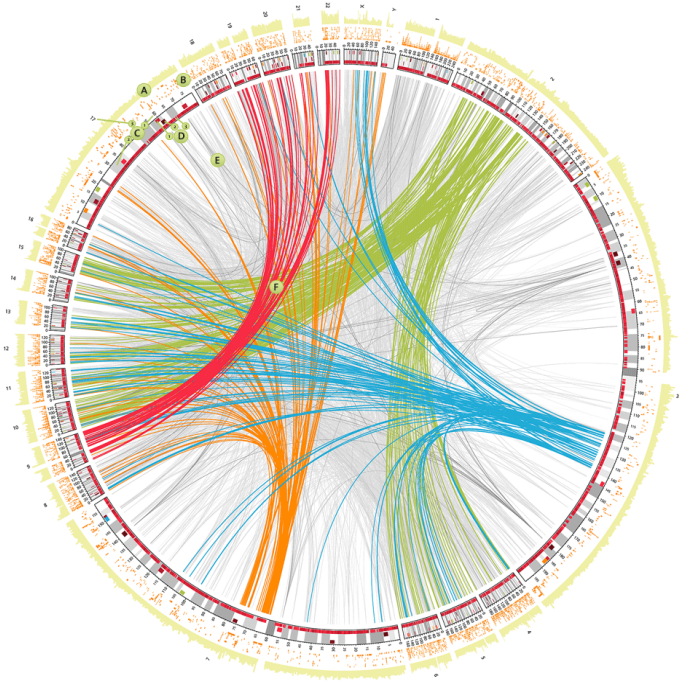
- GenomeGraphs, Gviz(in R).
- Interactive Desktop-app: IGB/IGV(Java).
- Web based: UCSC genome browser, DNAnexus.
- Specialized: Circos.

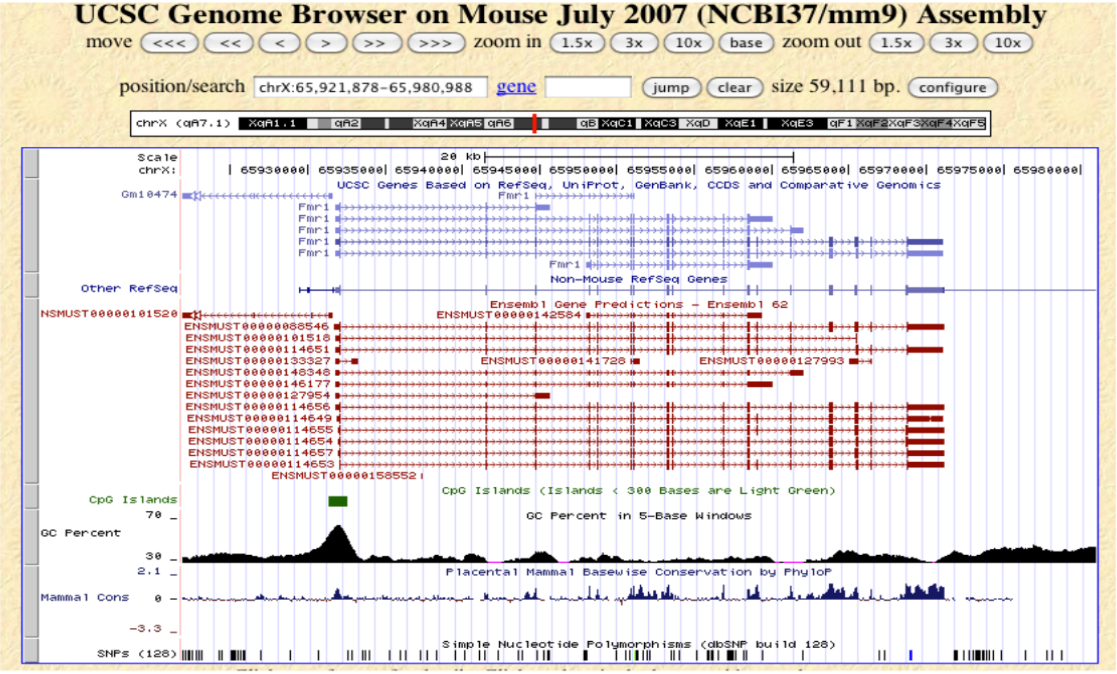


circos project.

Generalization possible?

Do they really look so different?

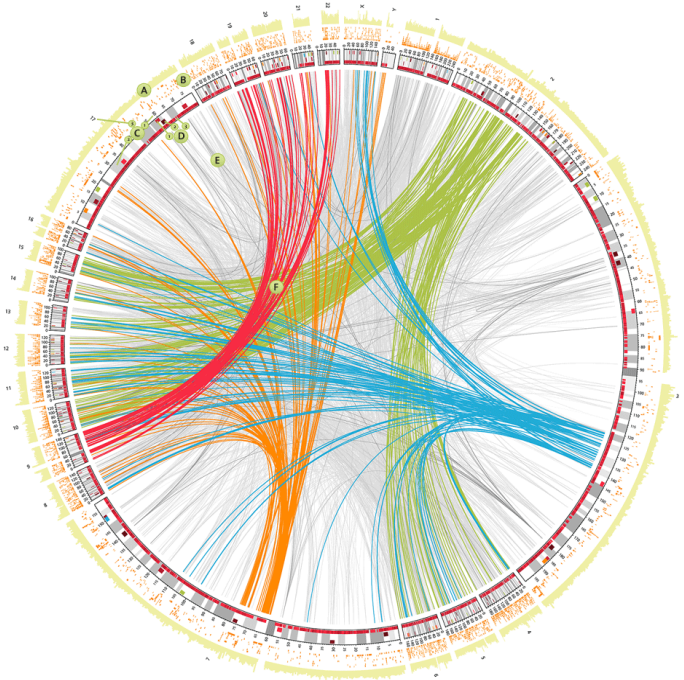


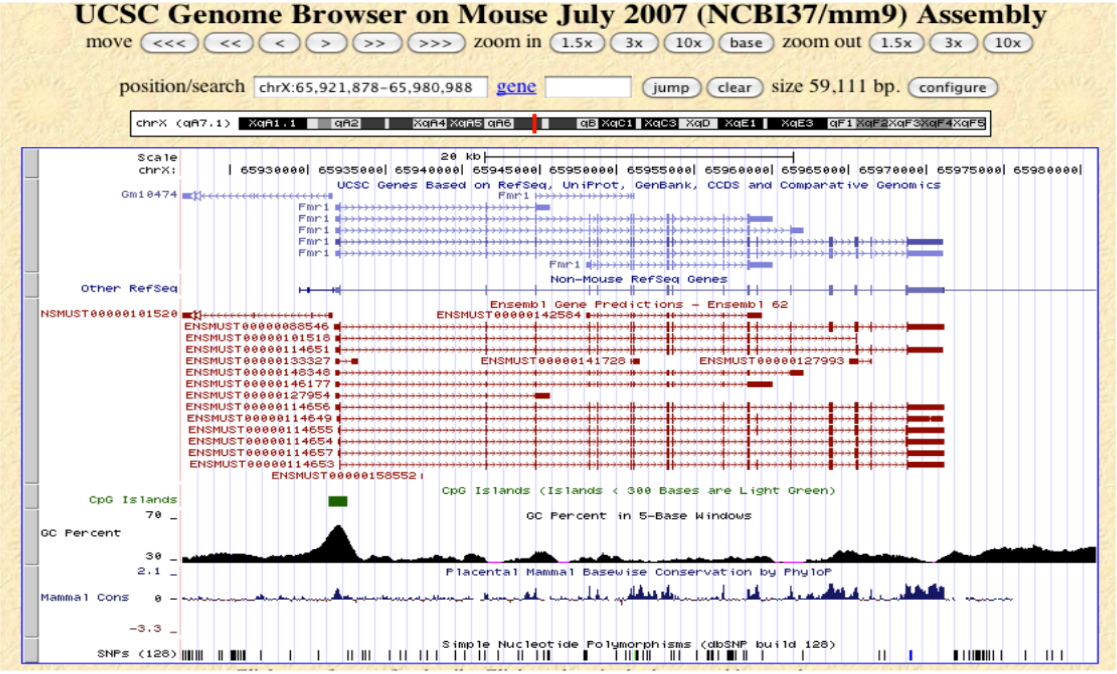


circos project/UCSC Genome browser.

Generalization possible?

Possible to generalize!





circos project/UCSC Genome browser.

Outline

Motivation

- Explore the data in different ways.
- Generalize genomic visualization frame work.
- Elegant graphics for publications.
- Modularize components to facilitate construction of high level graphics.

Design basic

- A powerful computational platform.
  - R: statistical platform, numerous model.
- A general data model.
  - Bioconductor: For I/O, data model, and some analytical.
- A general graphic model and grammar.
  - *ggplot2*: A grammar of graphics in R.

Design basic

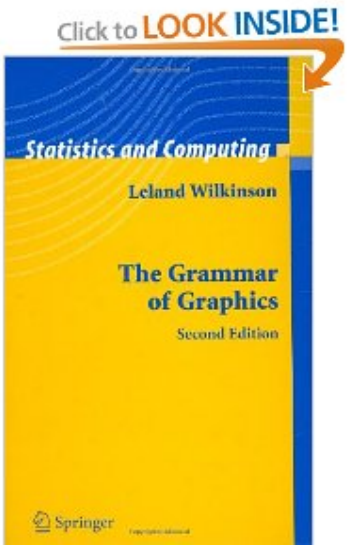
- A powerful computational platform.
  - R: statistical platform, numerous model.
- A general data model.
  - Bioconductor: For I/O, data model, and some analytical.
- A general graphic model and grammar.
  - *ggplot2*: A grammar of graphics in R.

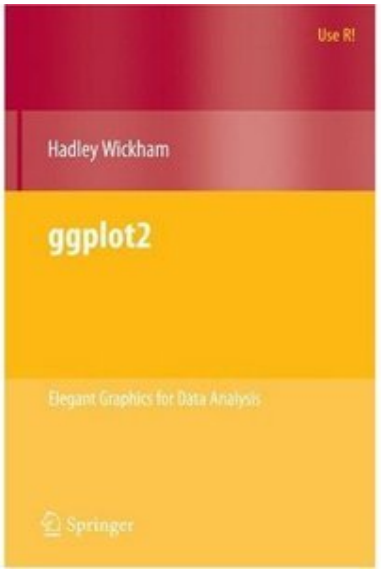
Design basic

- A powerful computational platform.
  - R: statistical platform, numerous model.
- A general data model.
  - Bioconductor: For I/O, data model, and some analytical.
- A general graphic model and grammar.
  - *ggplot2*: A grammar of graphics in R.

Design basic

- A powerful computational platform.
  - R: statistical platform, numerous model.
- A general data model.
  - Bioconductor: For I/O, data model, and some analytical.
- A general graphic model and grammar.
  - *ggplot2*: A grammar of graphics in R.

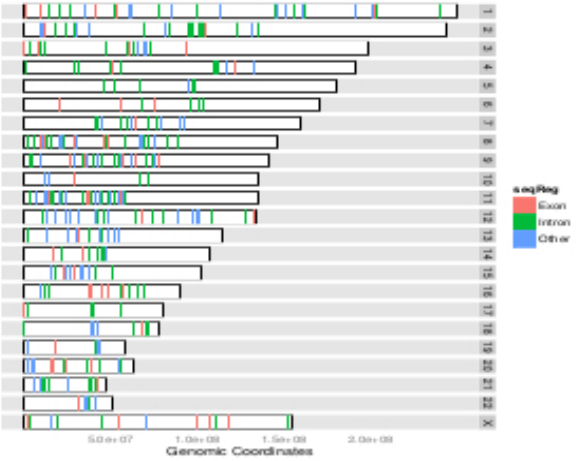


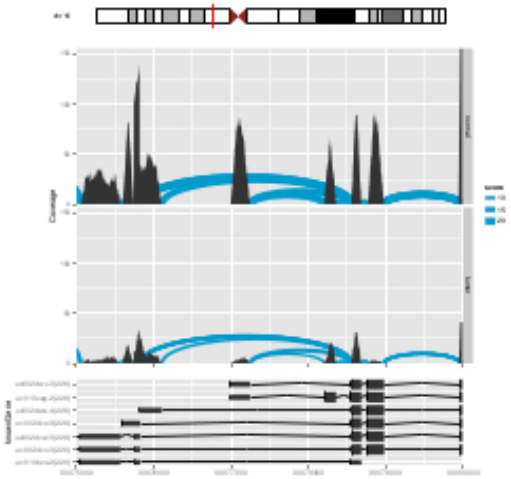


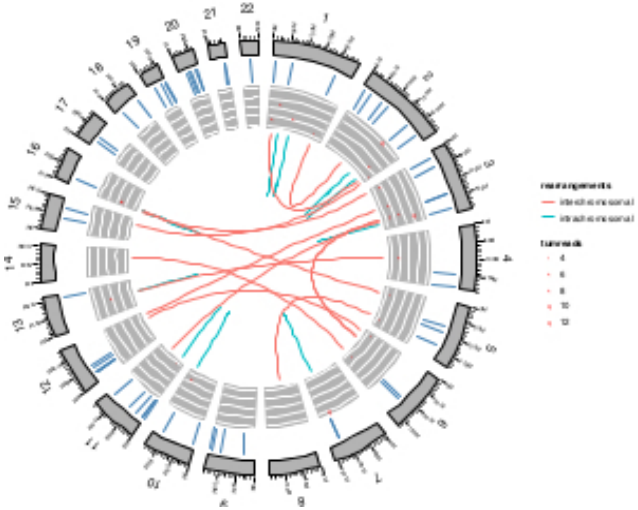
Outline

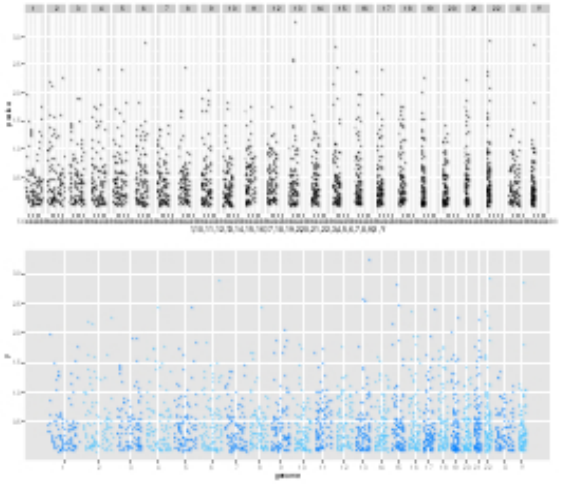
Outline

What can ggbio do?

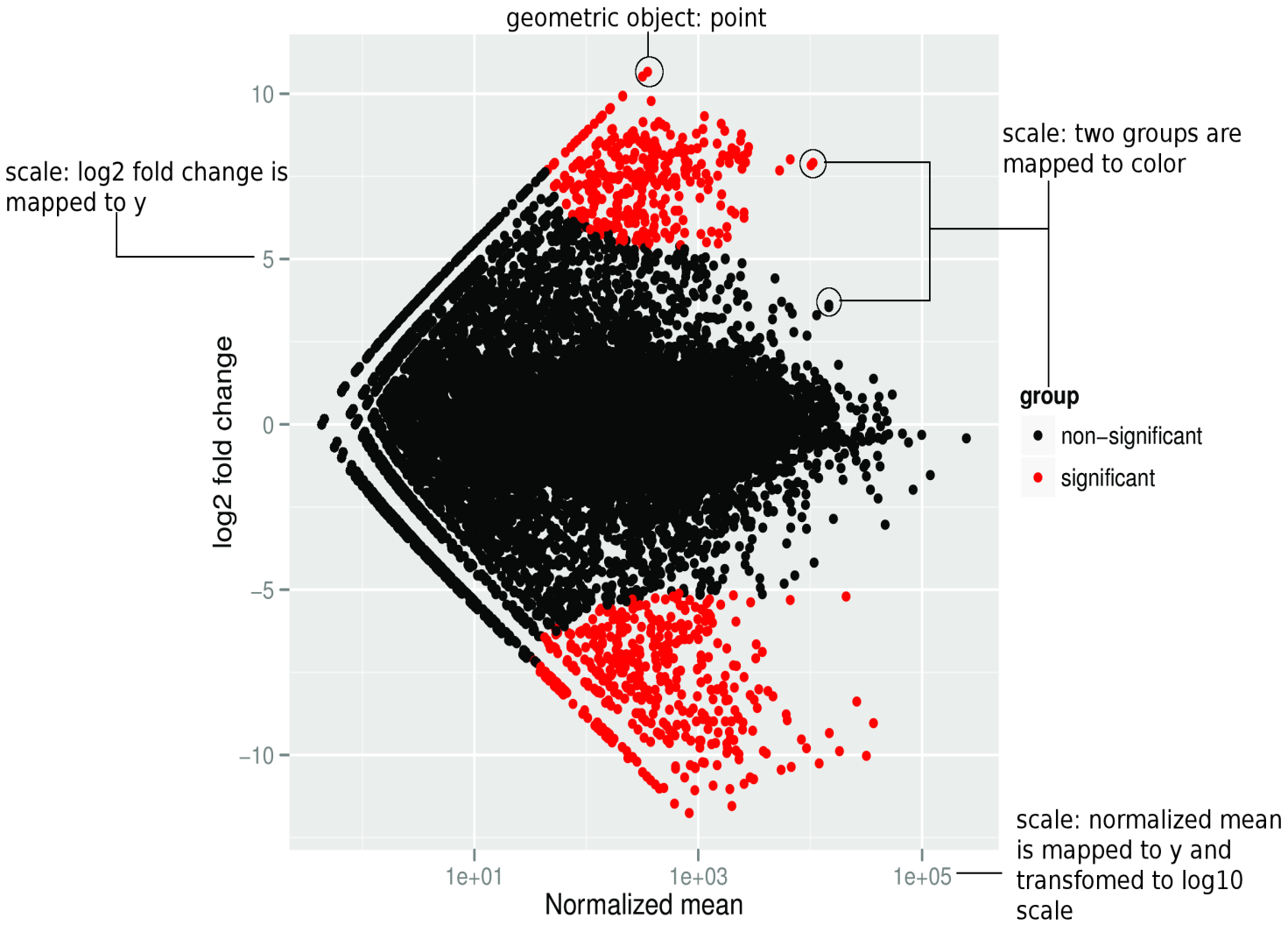






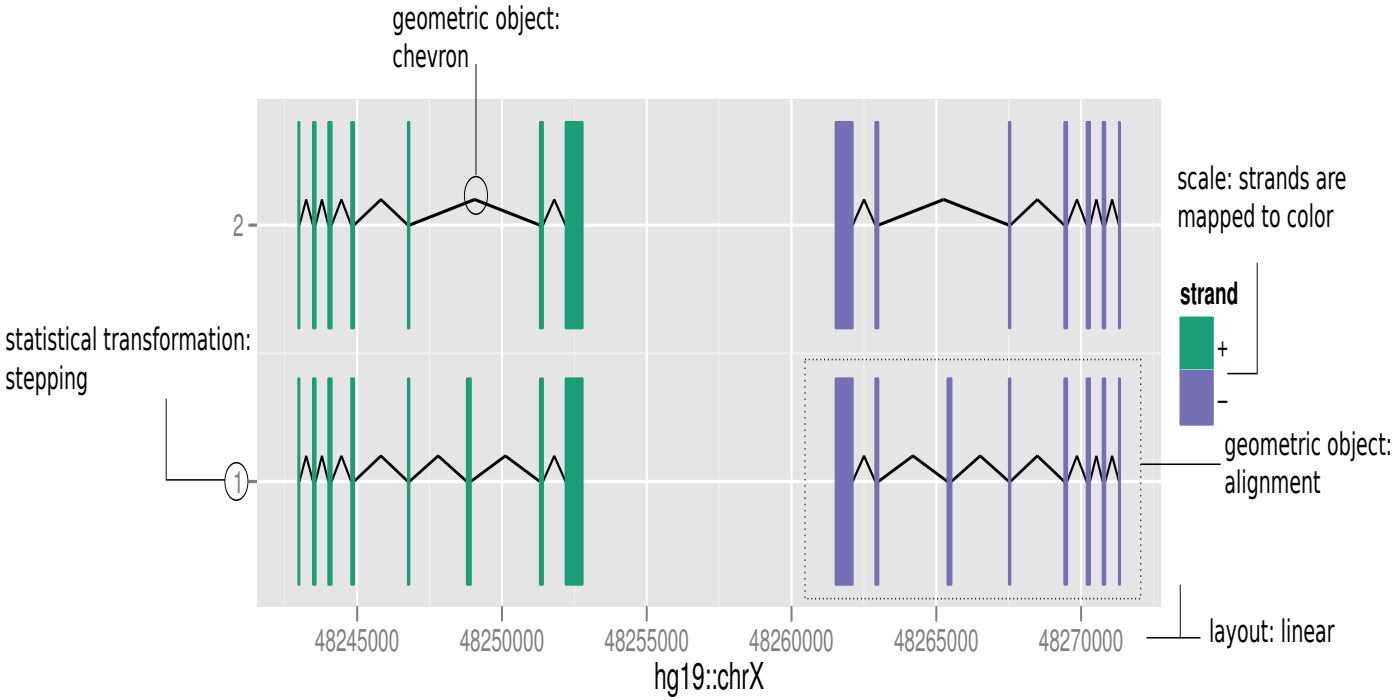


What is grammar of graphics(GoG)

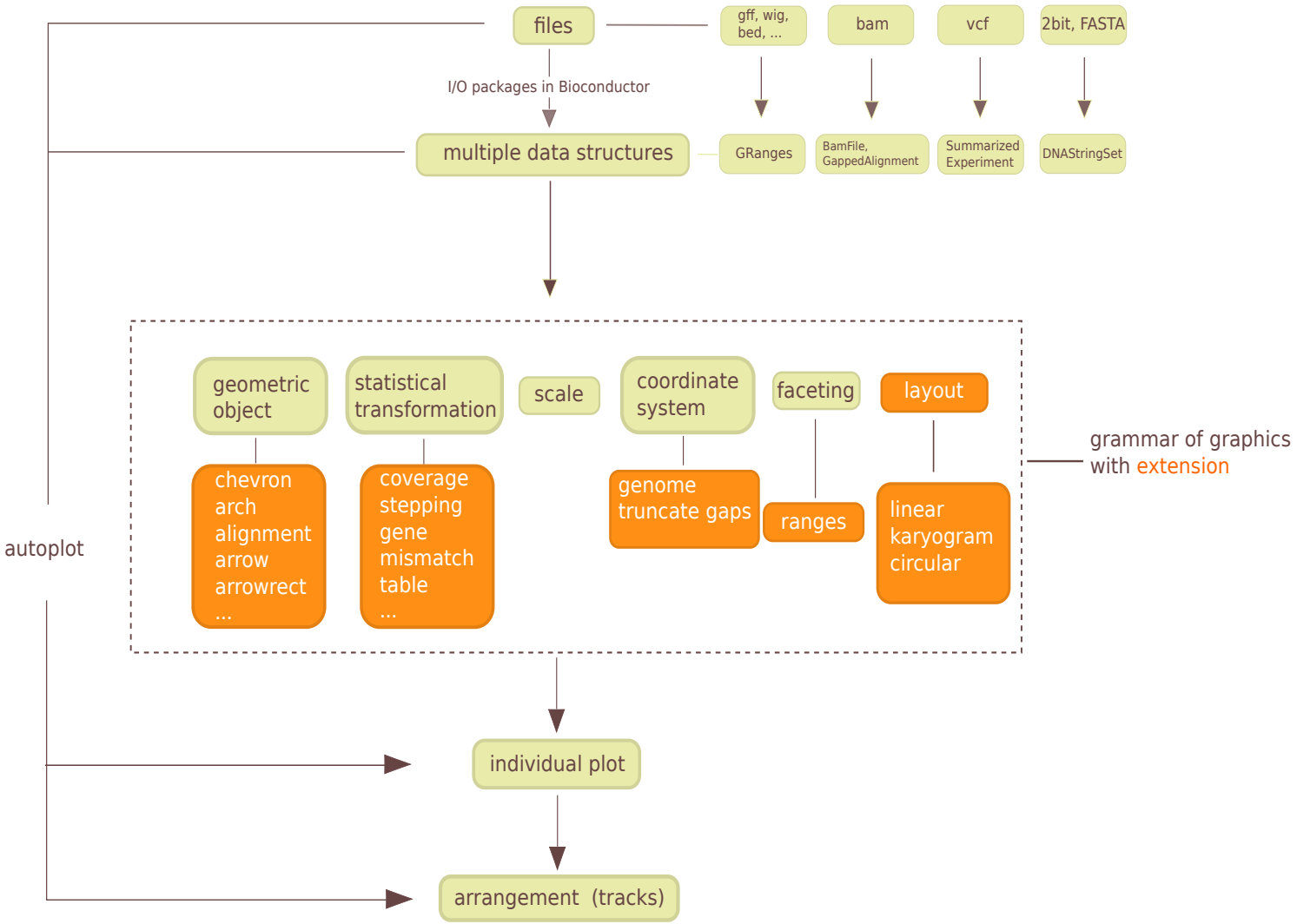


Outline

GoG in Genomic World.



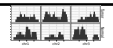
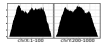
Pipeline in ggbio



Extended grammar of graphics table I

Comp	name	usage	icon
geom	geom_rect	rectangle	
	geom_segment	segment	
	geom_chevron	chevron	
	geom_arrow	arrow	
	geom_arch	arches	
	geom_bar	bar	
	geom_alignment	alignment (gene)	
stat	stat_coverage	coverage (of reads)	
	stat_mismatch	mismatch pileup for alignments	
	stat_aggregate	aggregate in sliding window	
	stat_stepping	avoid overplotting	
	stat_gene	consider gene structure	
	stat_table	tabulate ranges	
	stat_identity	no change	

Extended grammar of graphics table II

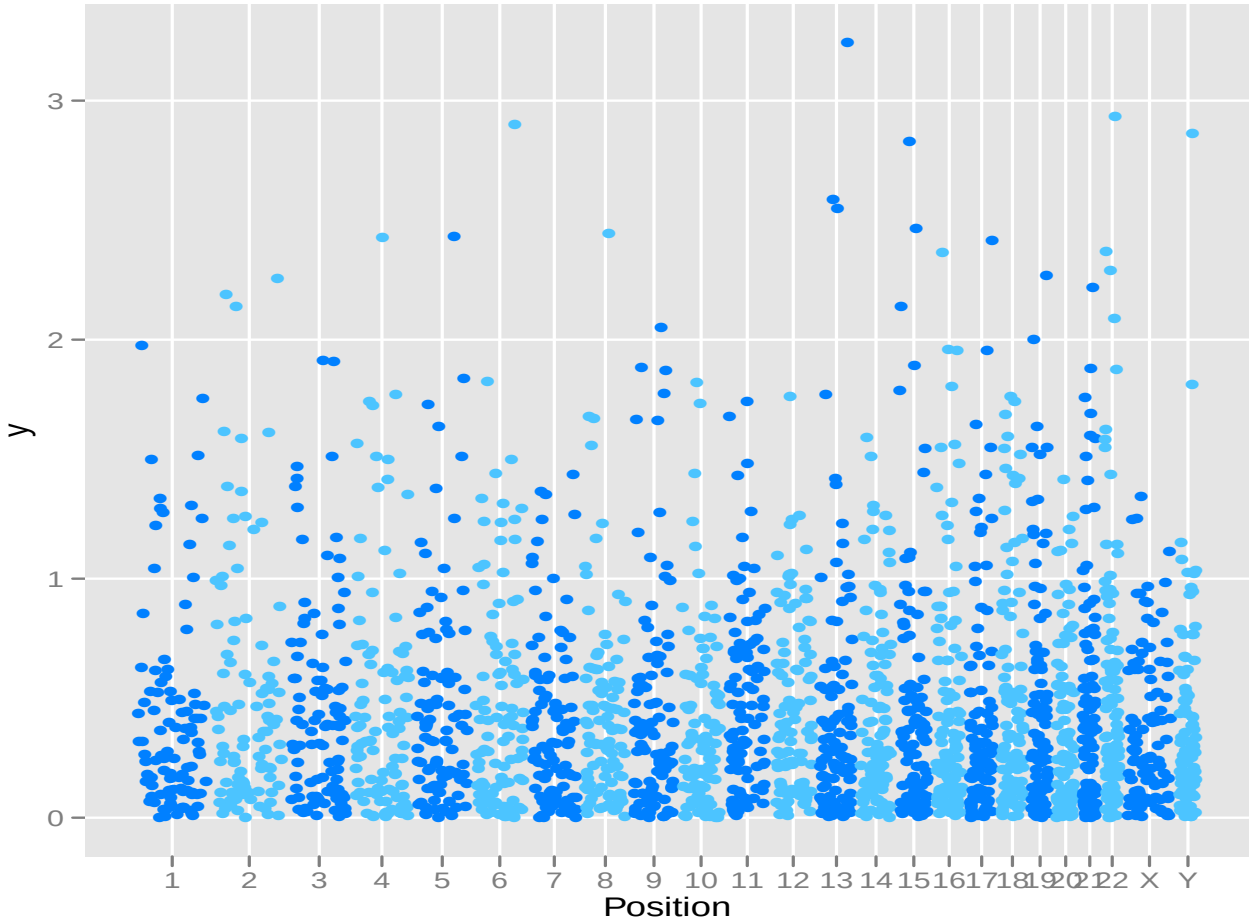
coord	linear	ggplot2 linear but facet by chromosome	
	genome	put everything on genomic coordinates	
	truncate gaps	compact view by shrinking gaps	
layout	track	stacked tracks	
	karyogram	karyogram display	
	circle	circular	
faceting	formula	facet by formula	
	ranges	facet by ranges	
scale	not extended	*ggplot2*default	

Extended grammar of graphics table III

Outline

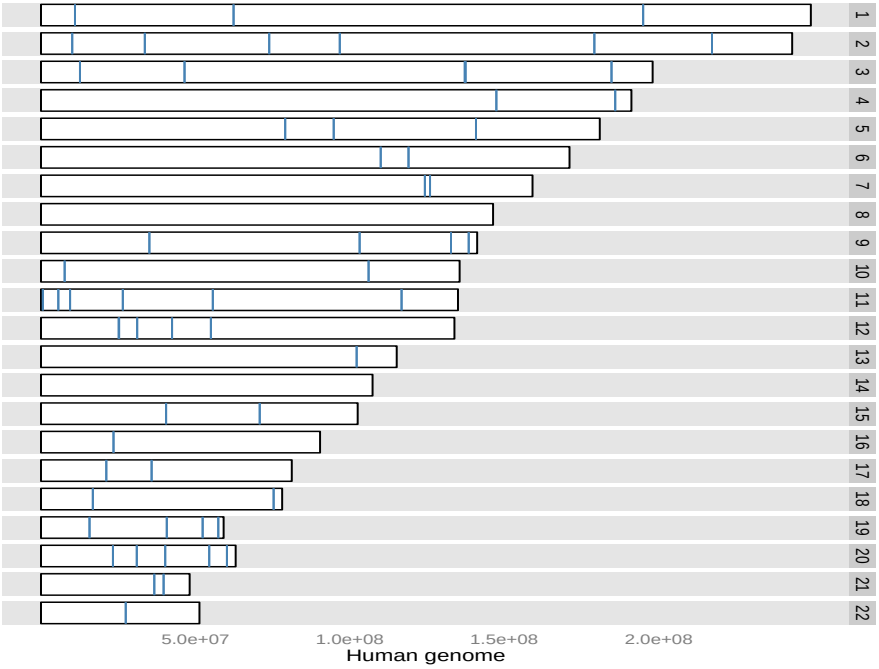
Manhattan plot(Grandlinear)

```
> autoplot(obj, coord = "genome", ...)
```



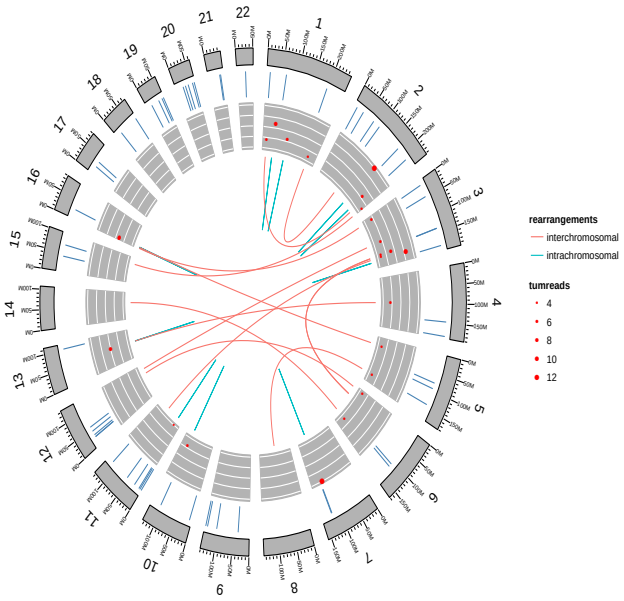
Karyogram layout

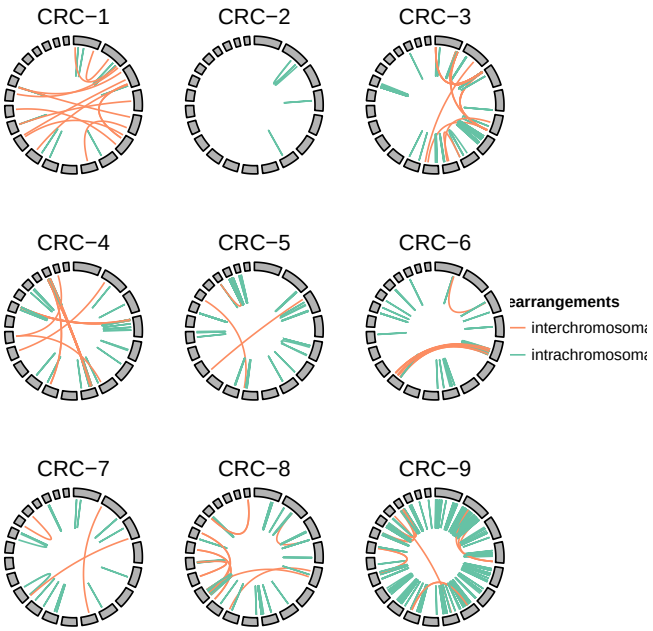
```
@
> autoplot(obj, layout = "karyogram", ...)
```



Circular layout

```
> autoplot(obj, layout = "circle", ...)
```





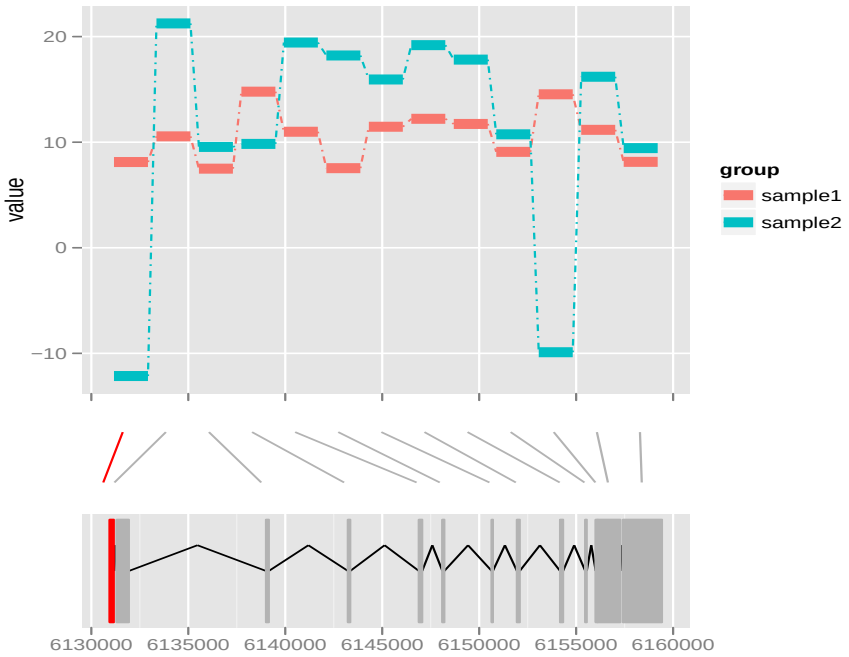
Outline

Specialized plots.

- Higher level: for convenience usage.
- Complex than prototypes, usually integrated with tracks.

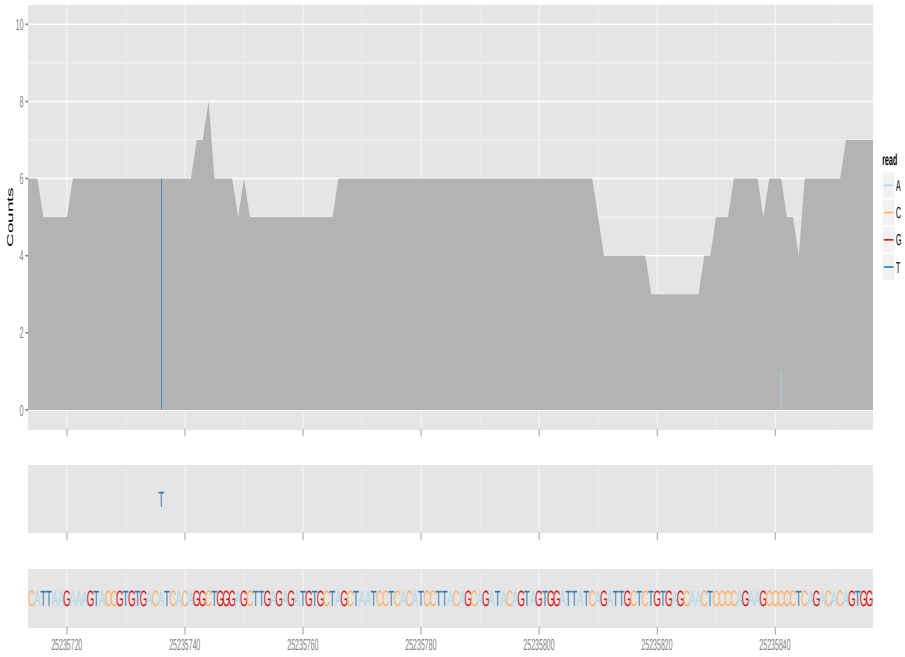
'ranges-linked-to-data' plot.

```
> plotRangesLinkedToData(obj, ...)
```



Mismatch summary plot

```
> autoplot(obj, stat = 'mismatch', ...)
```

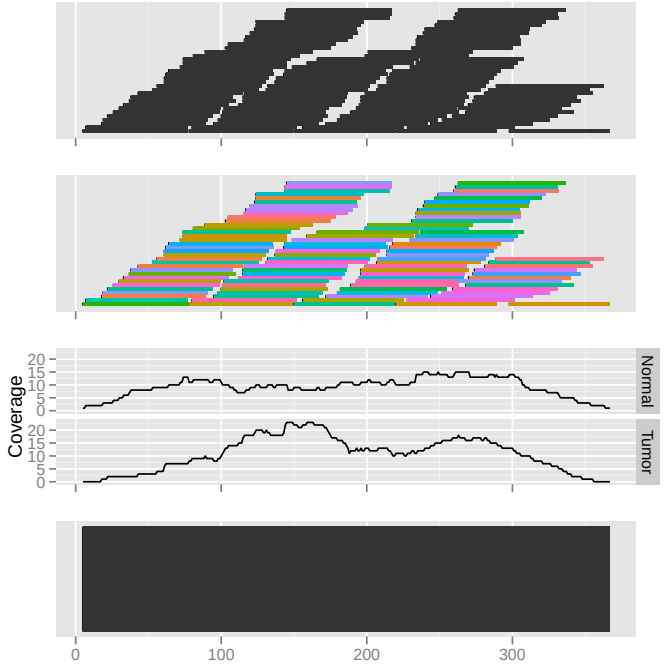


Object supported(overview).

GRanges				

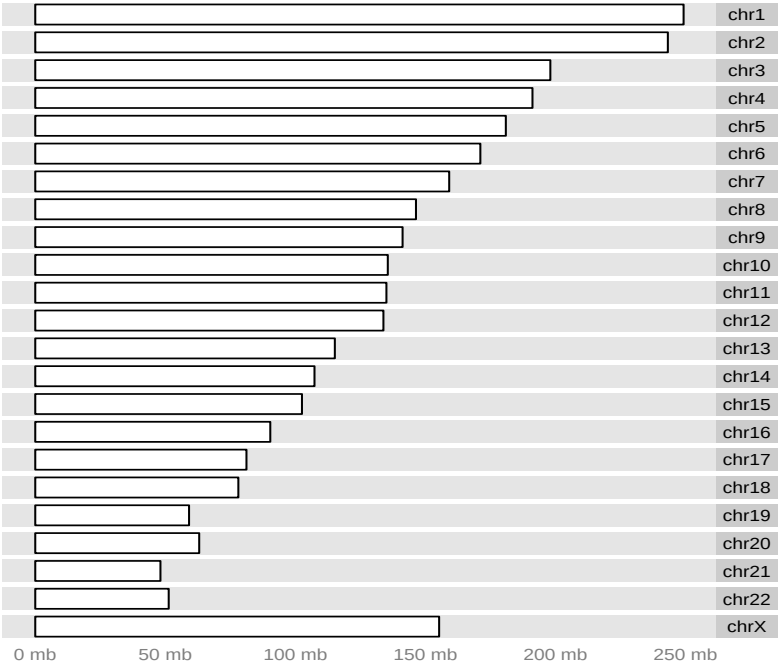
Object supported(overview).

GRanges	*IRanges*			



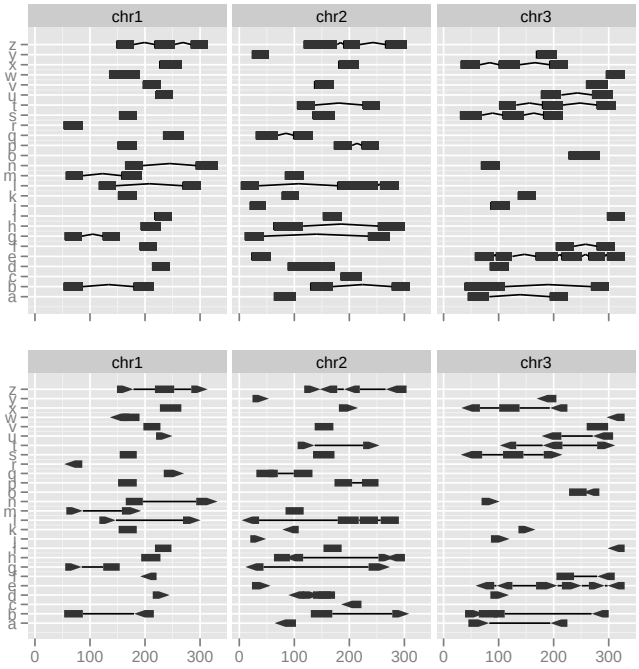
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*		



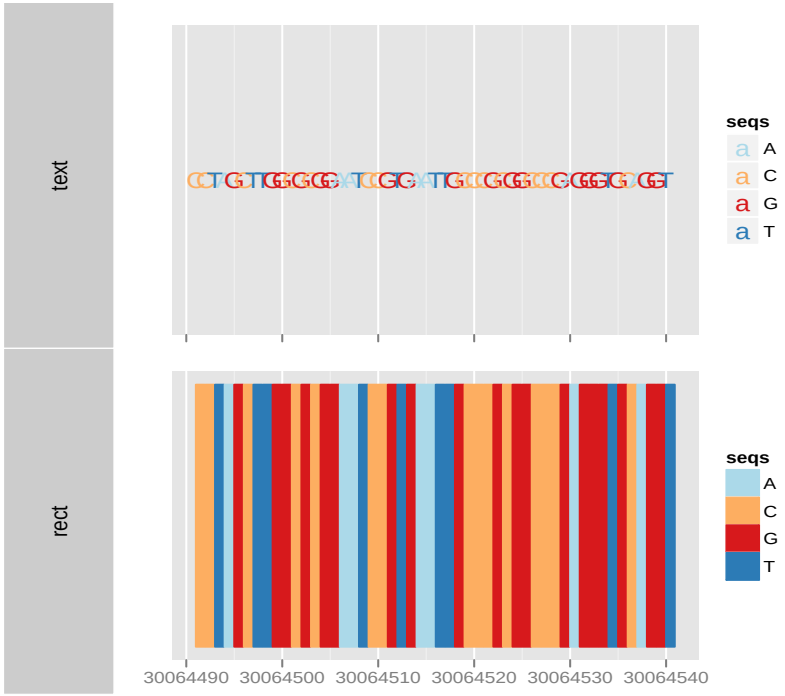
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	



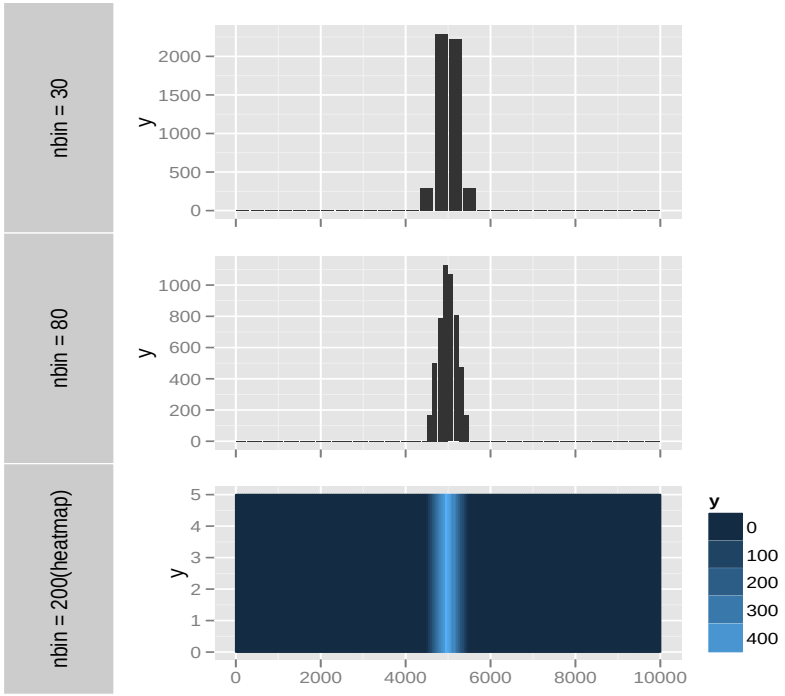
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*



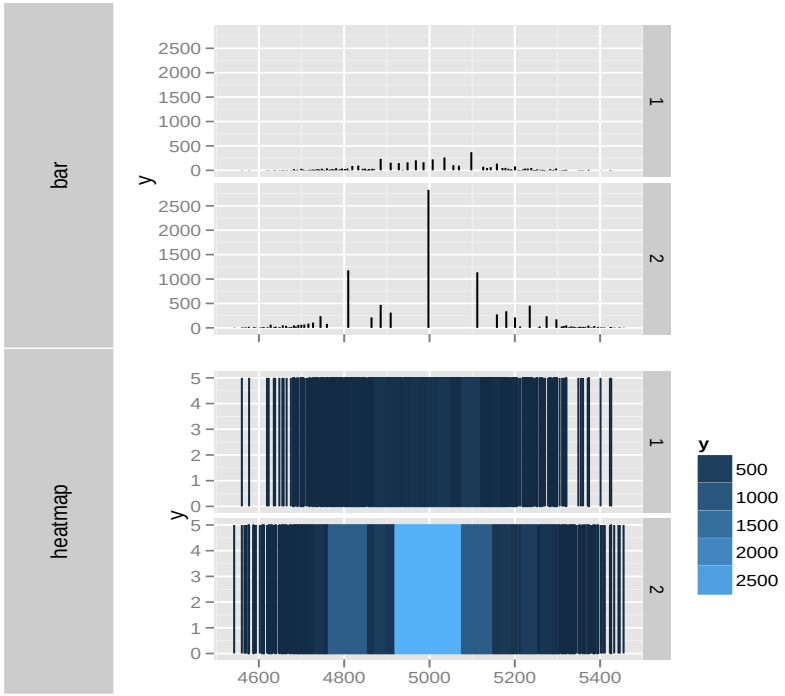
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle				



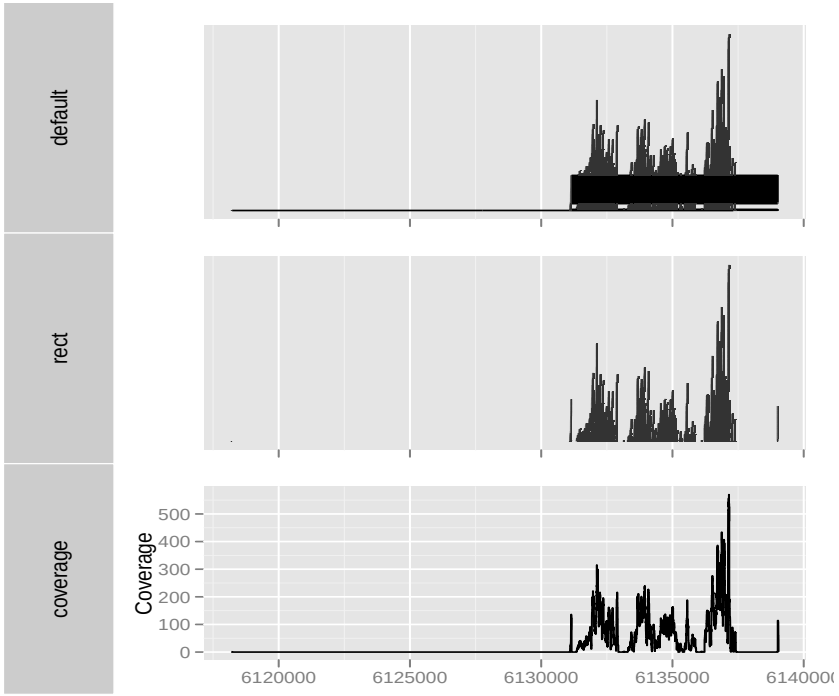
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*			



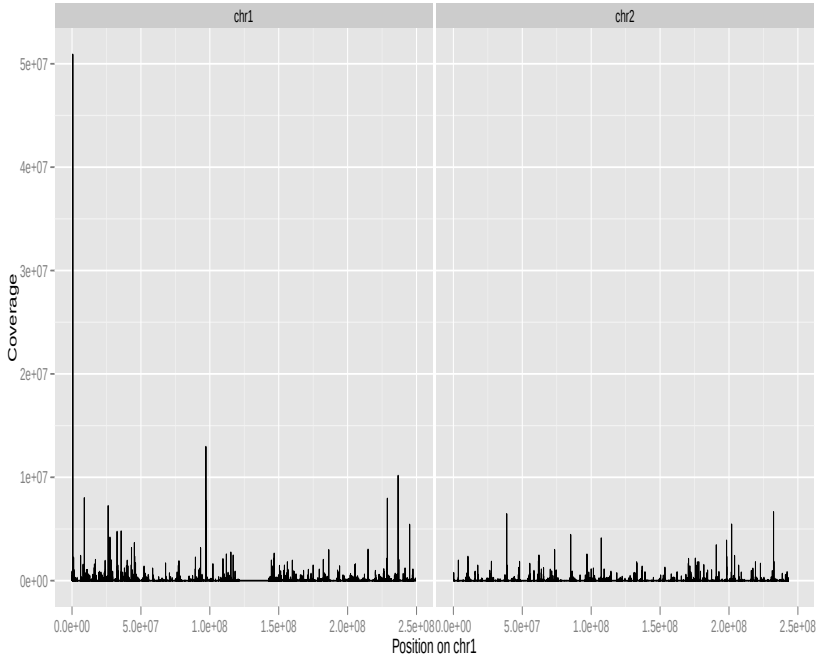
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*		



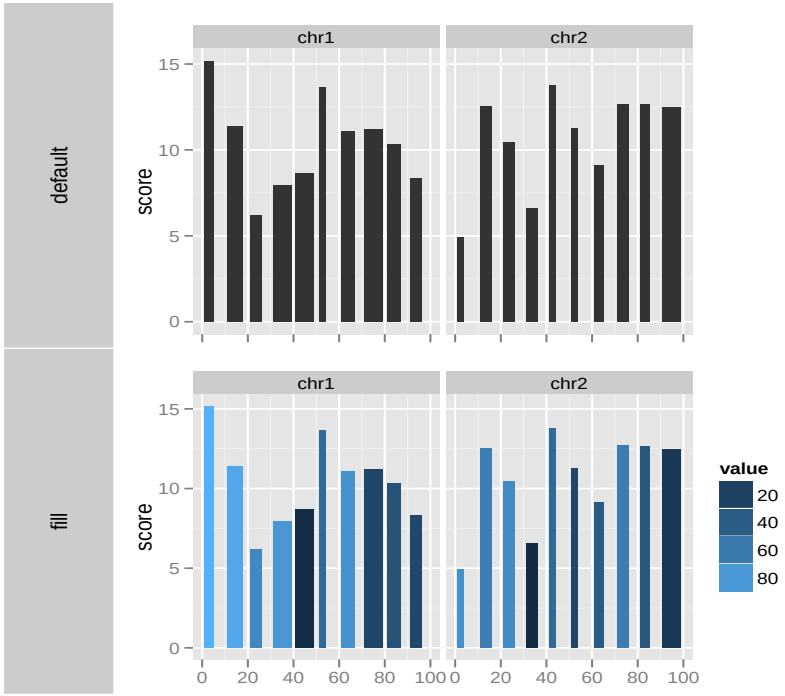
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*	*BamFile*	



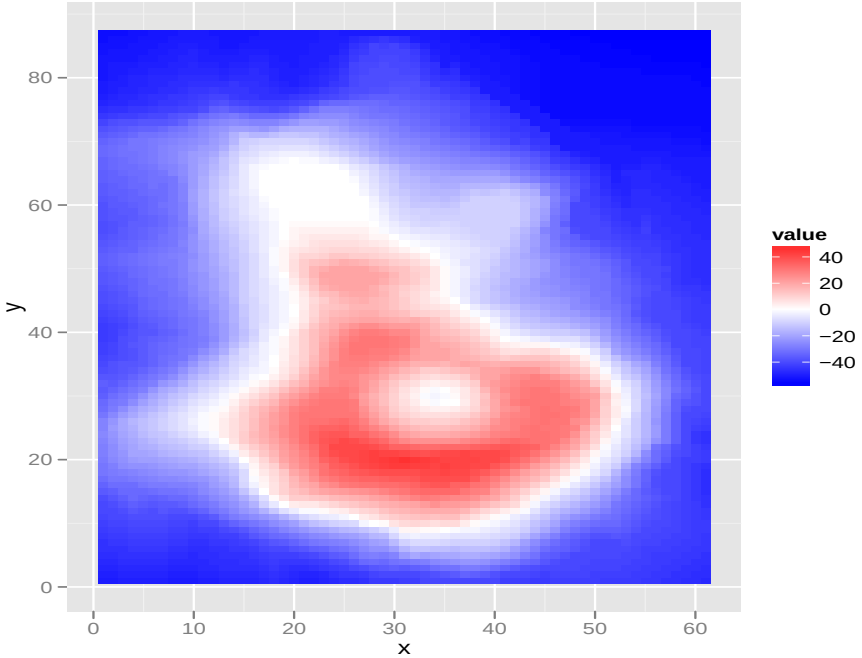
Object supported(overview).

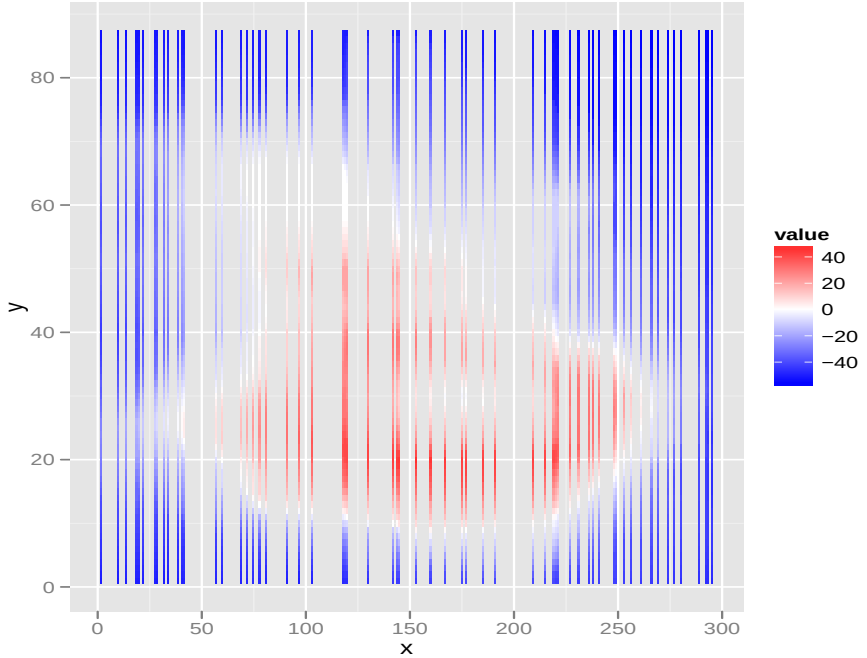
GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*	*BamFile*	*character*



Object supported(overview).

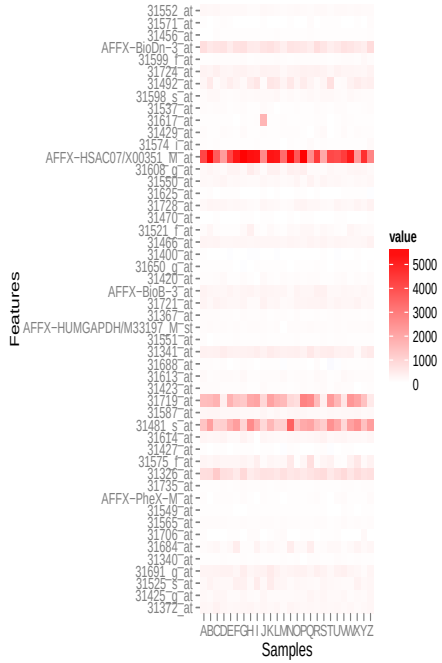
GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*	*BamFile*	*character*
matrix				

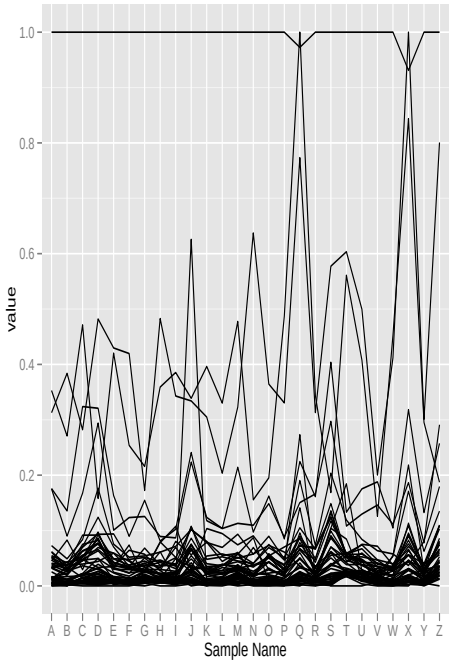


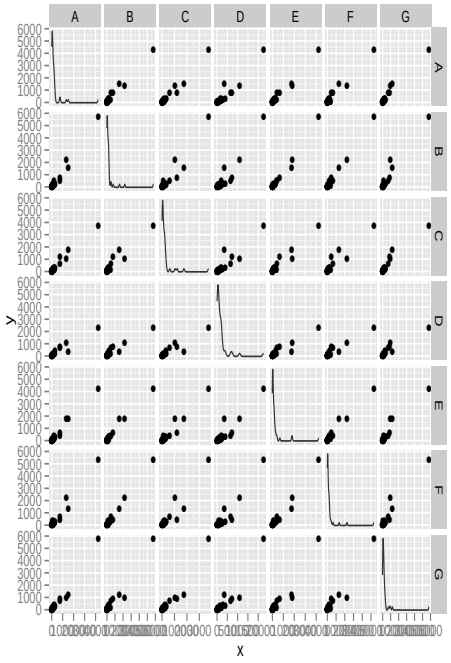


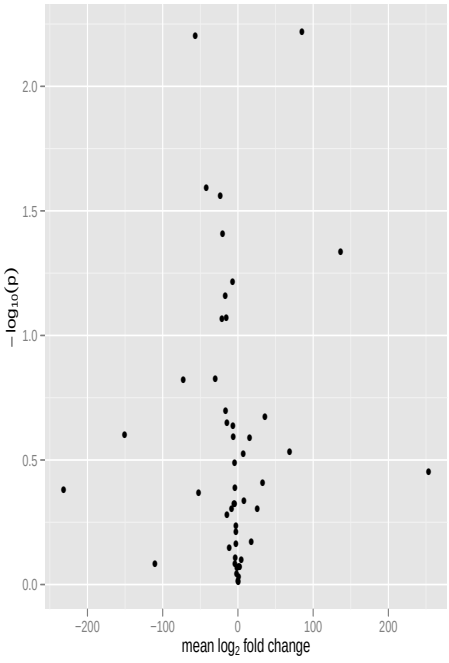
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*	*BamFile*	*character*
matrix	*ExpressionSet*			



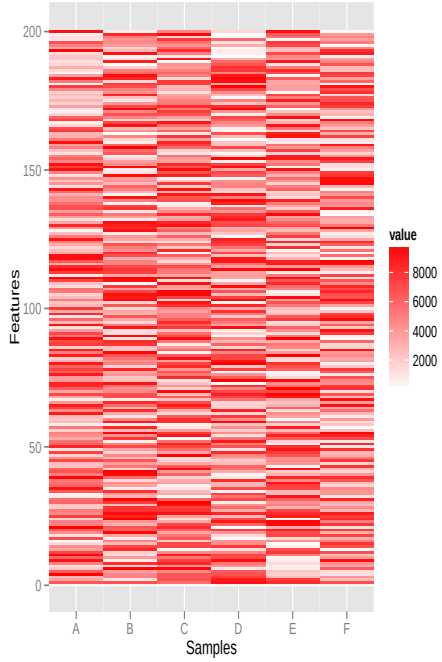


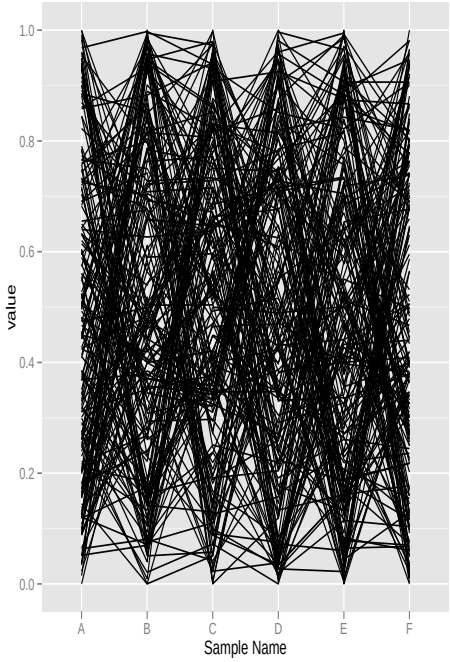


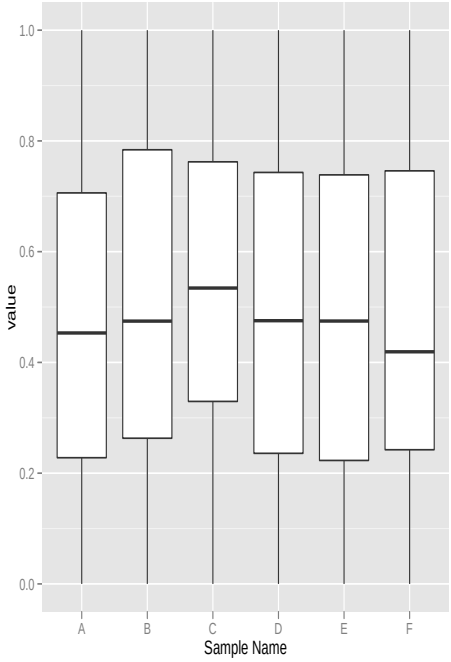


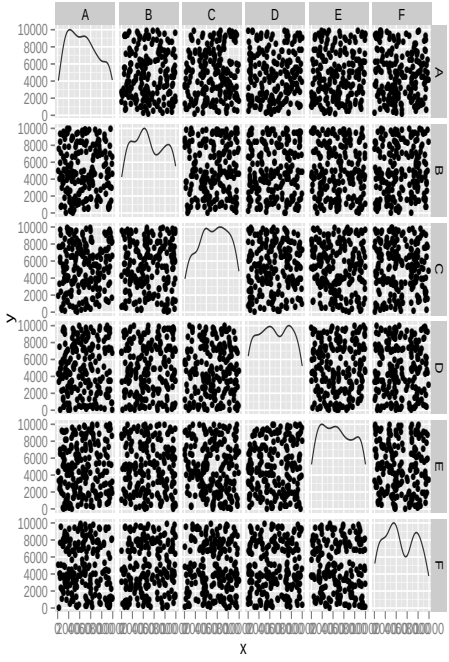
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*	*BamFile*	*character*
matrix	*ExpressionSet*	*SummarizedExperiment*		



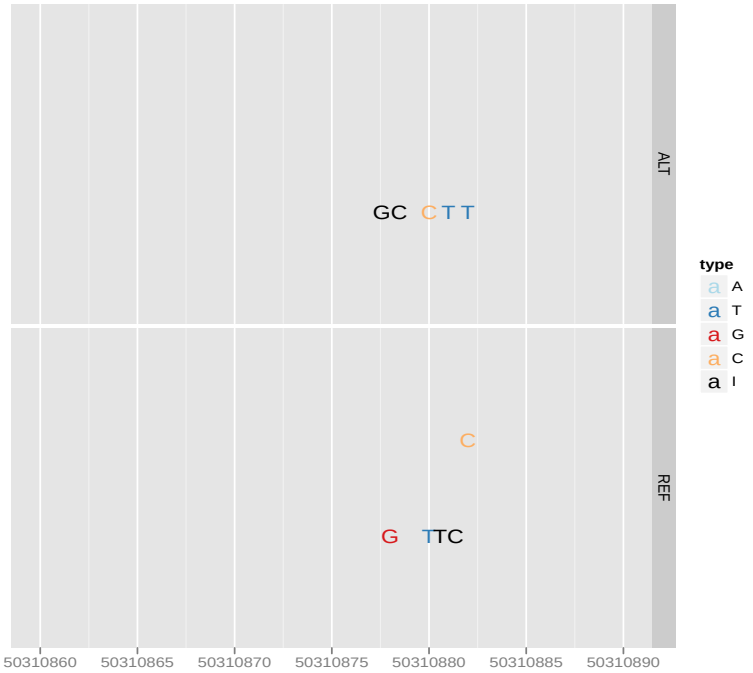






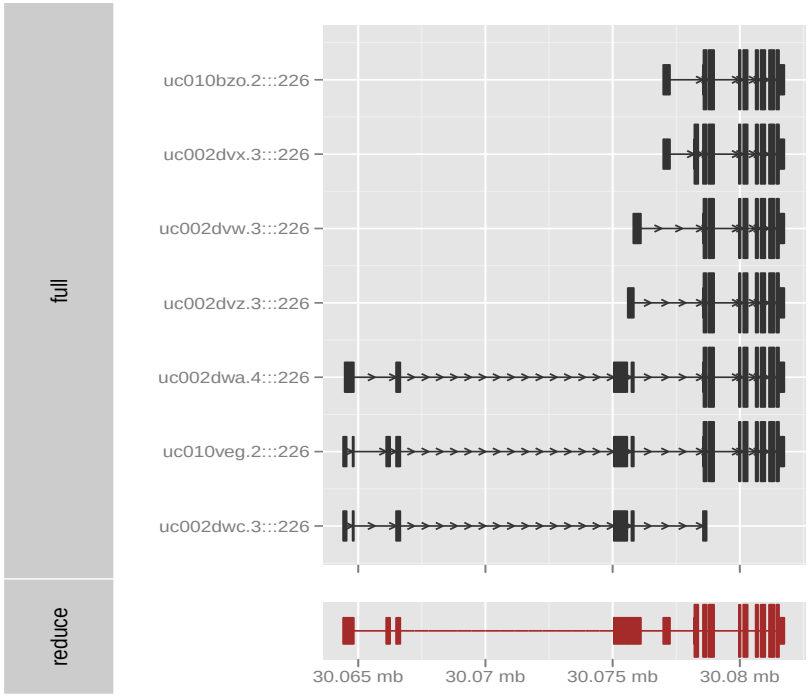
Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*	*BamFile*	*character*
matrix	*ExpressionSet*	*SummarizedExperiment*	*VCF*	



Object supported(overview).

GRanges	*IRanges*	*Seqinfo*	*GRangesList*	*BSgenome*
Rle	*RleList*	*GappedAlignment*	*BamFile*	*character*
matrix	*ExpressionSet*	*SummarizedExperiment*	*VCF*	*TranscriptDb*



Outline

Future study

- Support more core data model in Bioconductor.
- More elegant theme for tracks.
- Keep improving with new ggplot2 development
- More powerful tracks function, may accept lattice graphics.

Acknowledgment

- Michael Lawrence, Dianne Cook
- Genentech

Thank you !!!