A Grammar of Graphics for Genomics

The *ggbio* Package

Michael Lawrence

Genentech

August 29, 2012

Outline

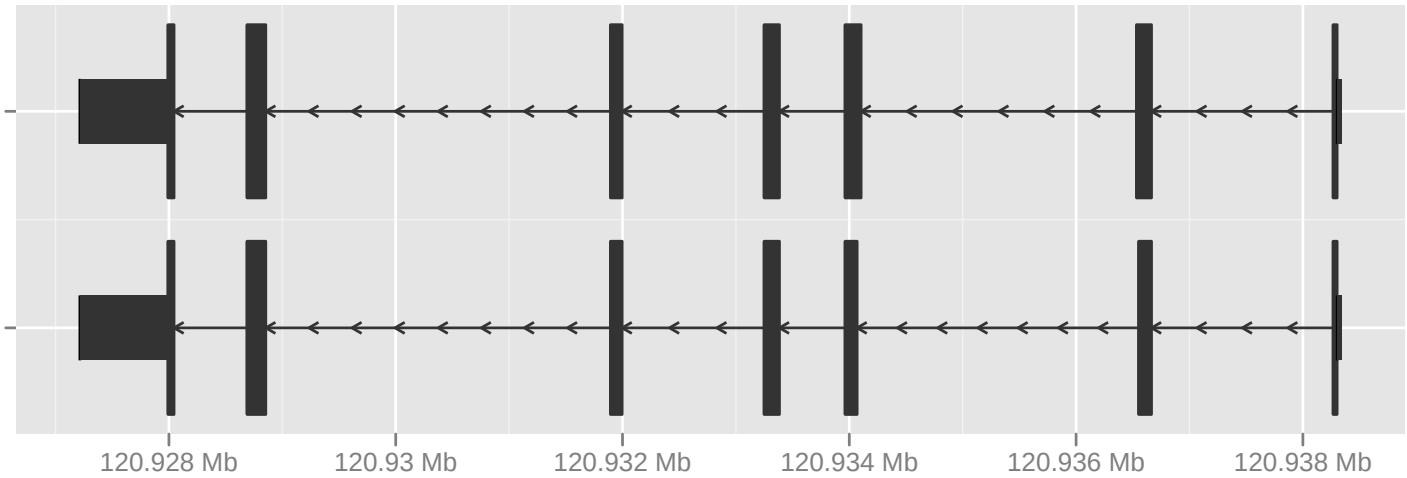
Outline

Data on the Genome

- Comes in two flavors:
  - Annotations (genes, TF binding sites, ...)
  - Experimental measurements (sequence reads)
- Both types are tied to genomic coordinates, providing a common axis that permits cross-dataset comparison and inference
- Typically stored as a table, with the range as a fundamental variable type, plus metadata

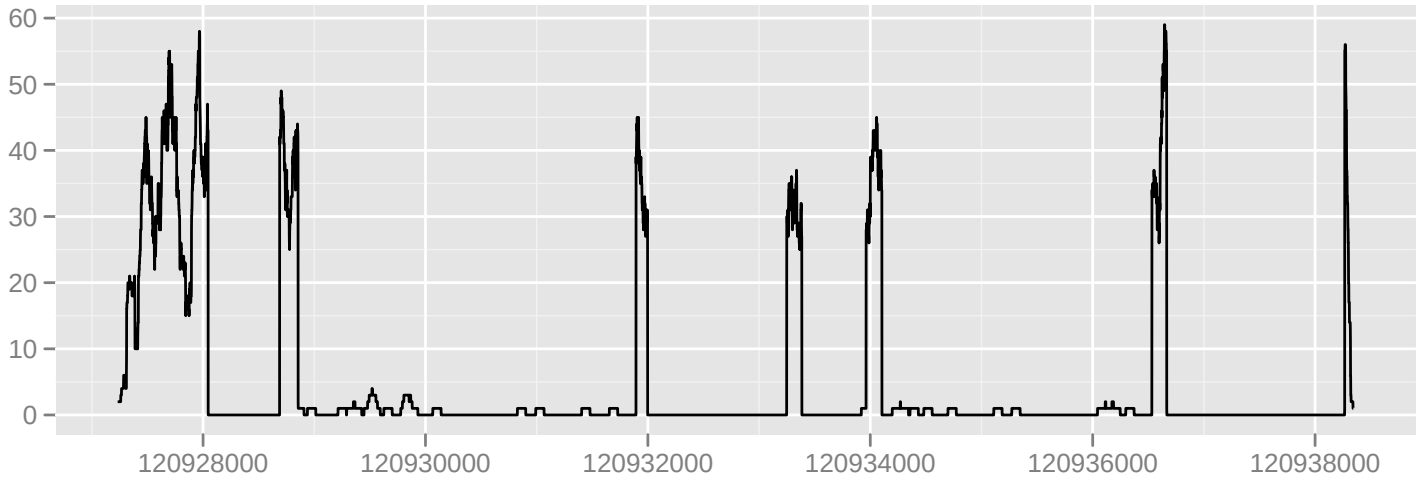
Data on the Genome

- Comes in two flavors:
  - Annotations (genes, TF binding sites, ...)
  - Experimental measurements (sequence reads)
- Both types are tied to genomic coordinates, providing a common axis that permits cross-dataset comparison and inference
- Typically stored as a table, with the range as a fundamental variable type, plus metadata



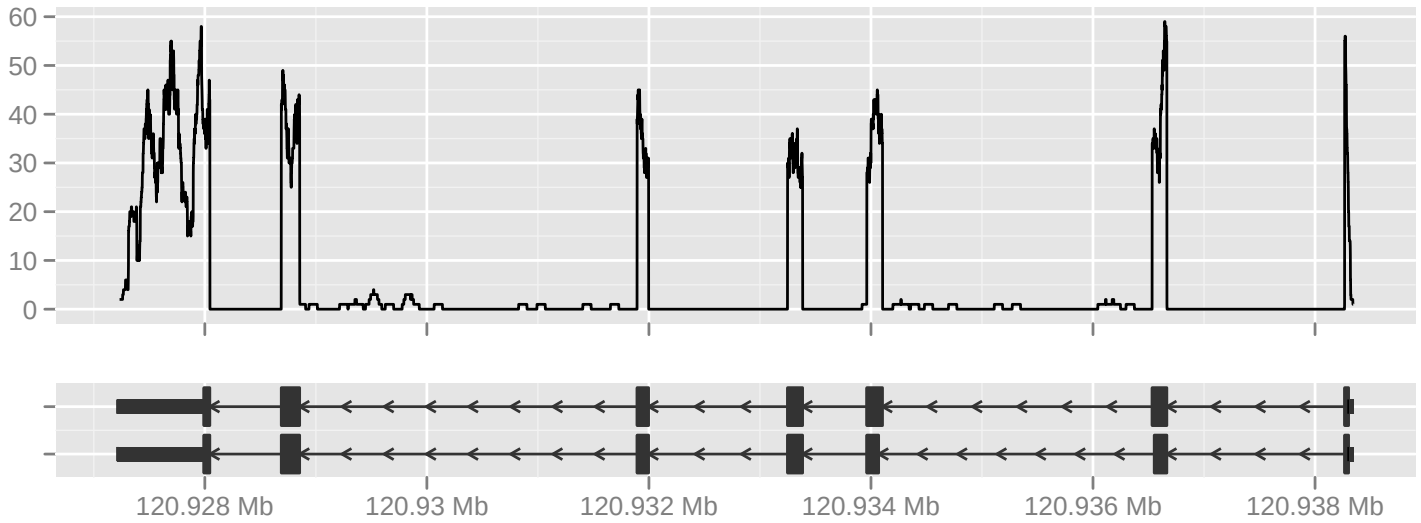
Data on the Genome

- Comes in two flavors:
  - Annotations (genes, TF binding sites, ...)
  - Experimental measurements (sequence reads)
- Both types are tied to genomic coordinates, providing a common axis that permits cross-dataset comparison and inference
- Typically stored as a table, with the range as a fundamental variable type, plus metadata



Data on the Genome

- Comes in two flavors:
  - Annotations (genes, TF binding sites, ...)
  - Experimental measurements (sequence reads)
- Both types are tied to genomic coordinates, providing a common axis that permits cross-dataset comparison and inference
- Typically stored as a table, with the range as a fundamental variable type, plus metadata



Data on the Genome

- Comes in two flavors:
  - Annotations (genes, TF binding sites, ...)
  - Experimental measurements (sequence reads)
- Both types are tied to genomic coordinates, providing a common axis that permits cross-dataset comparison and inference
- Typically stored as a table, with the range as a fundamental variable type, plus metadata

seqnames	start	end	strand	exon_id	tx_id
10	120927215	120928045	-	129230	14886,14887
10	120928689	120928854	-	129229	14886,14887
10	120931894	120931997	-	129228	14886,14887
10	120933249	120933384	-	129227	14886,14887
10	120933963	120934069	-	129226	14886
10	120933963	120934104	-	119757	14887
10	120936533	120936665	-	119756	14887
10	120936552	120936665	-	129225	14886
10	120938267	120938345	-	129224	14886,14887

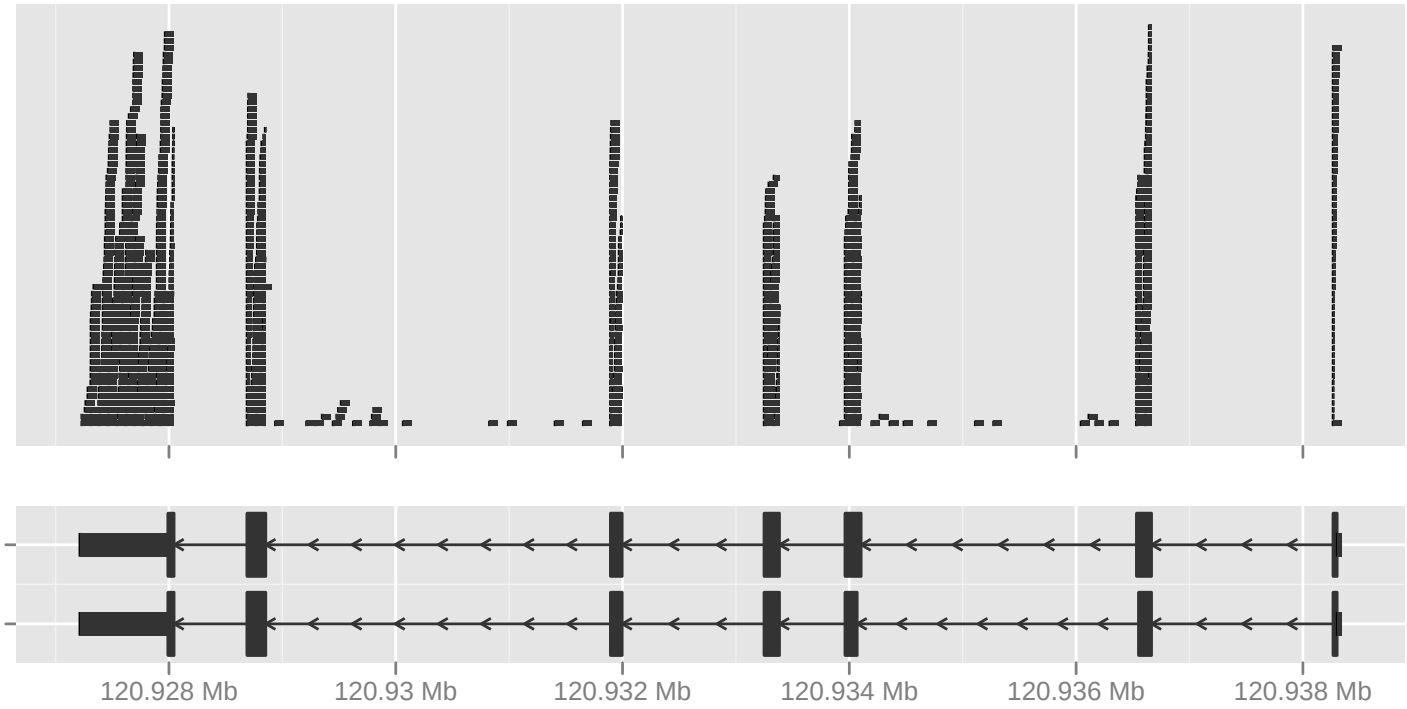
Challenges

Big data, wide spaces

- Need summaries that are efficiently computed, communicate more with less and expose the most interesting aspects of the data
- Need different ways of viewing the data, depending on the density and scale, from whole genome to single basepair

Challenges

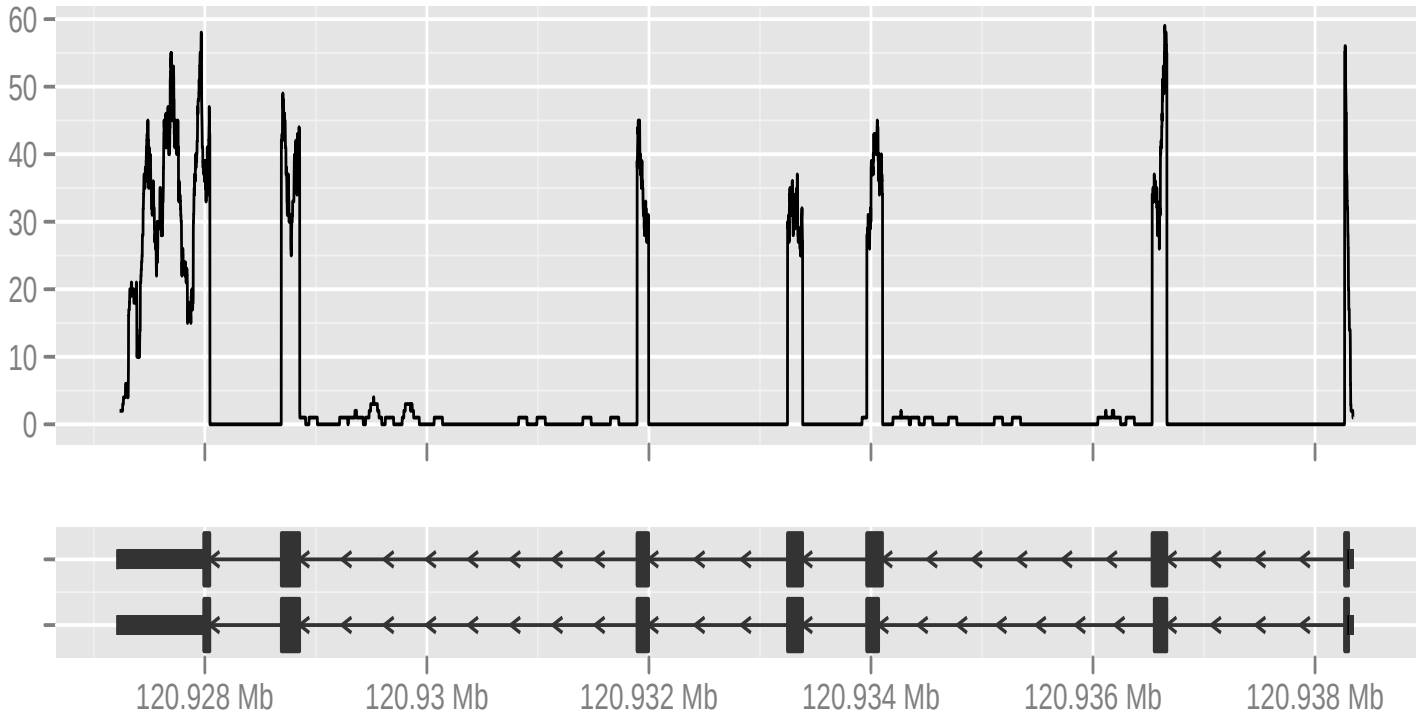
Big data, wide spaces



- Need summaries that are efficiently computed, communicate more with less and expose the most interesting aspects of the data
- Need different ways of viewing the data, depending on the density and scale, from whole genome to single basepair

Challenges

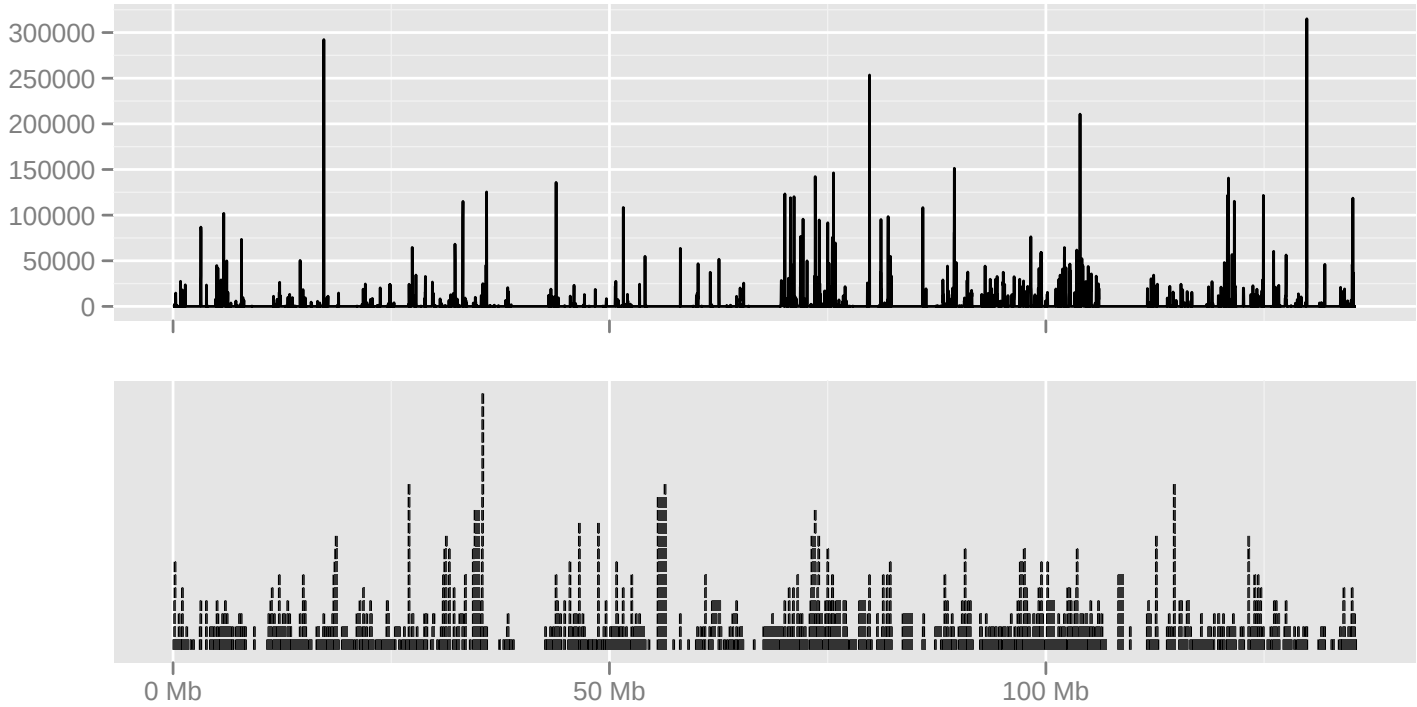
Big data, wide spaces



- Need summaries that are efficiently computed, communicate more with less and expose the most interesting aspects of the data
- Need different ways of viewing the data, depending on the density and scale, from whole genome to single basepair

Challenges

Big data, wide spaces



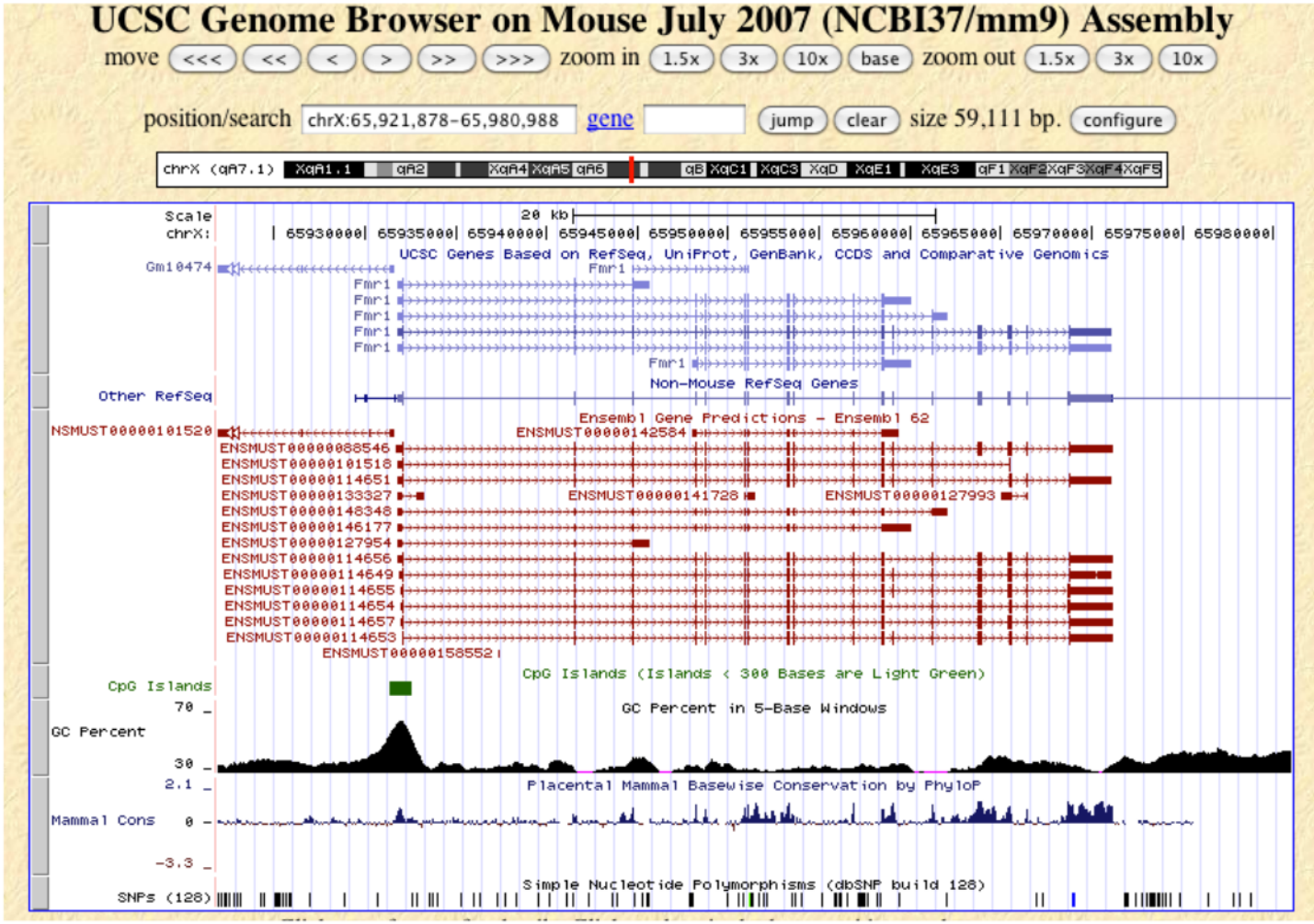
- Need summaries that are efficiently computed, communicate more with less and expose the most interesting aspects of the data
- Need different ways of viewing the data, depending on the density and scale, from whole genome to single basepair

Existing Tools

UCSC	IGB	IGV	Circos	GViz

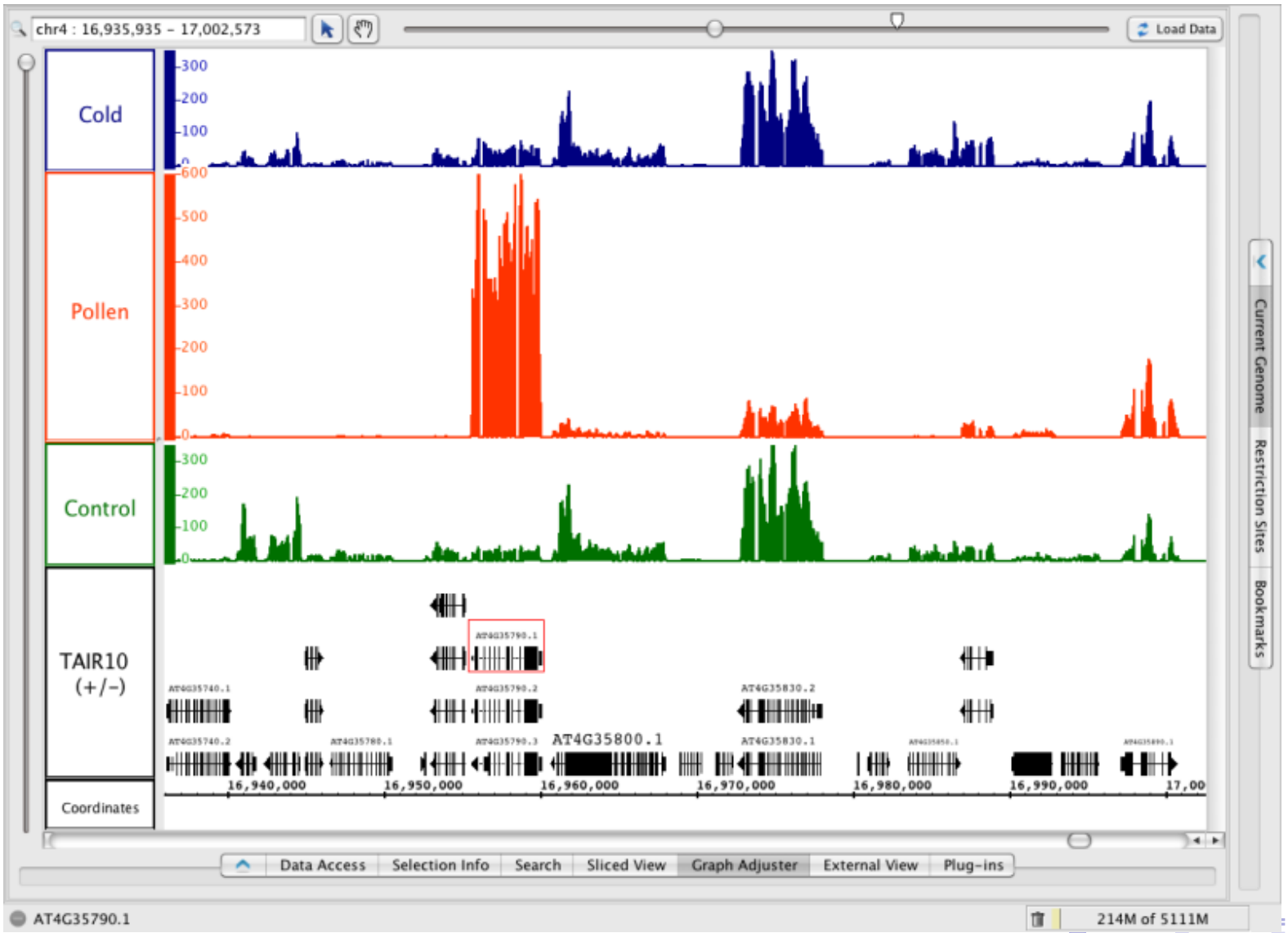
Existing Tools

UCSC IGB IGV Circos GViz



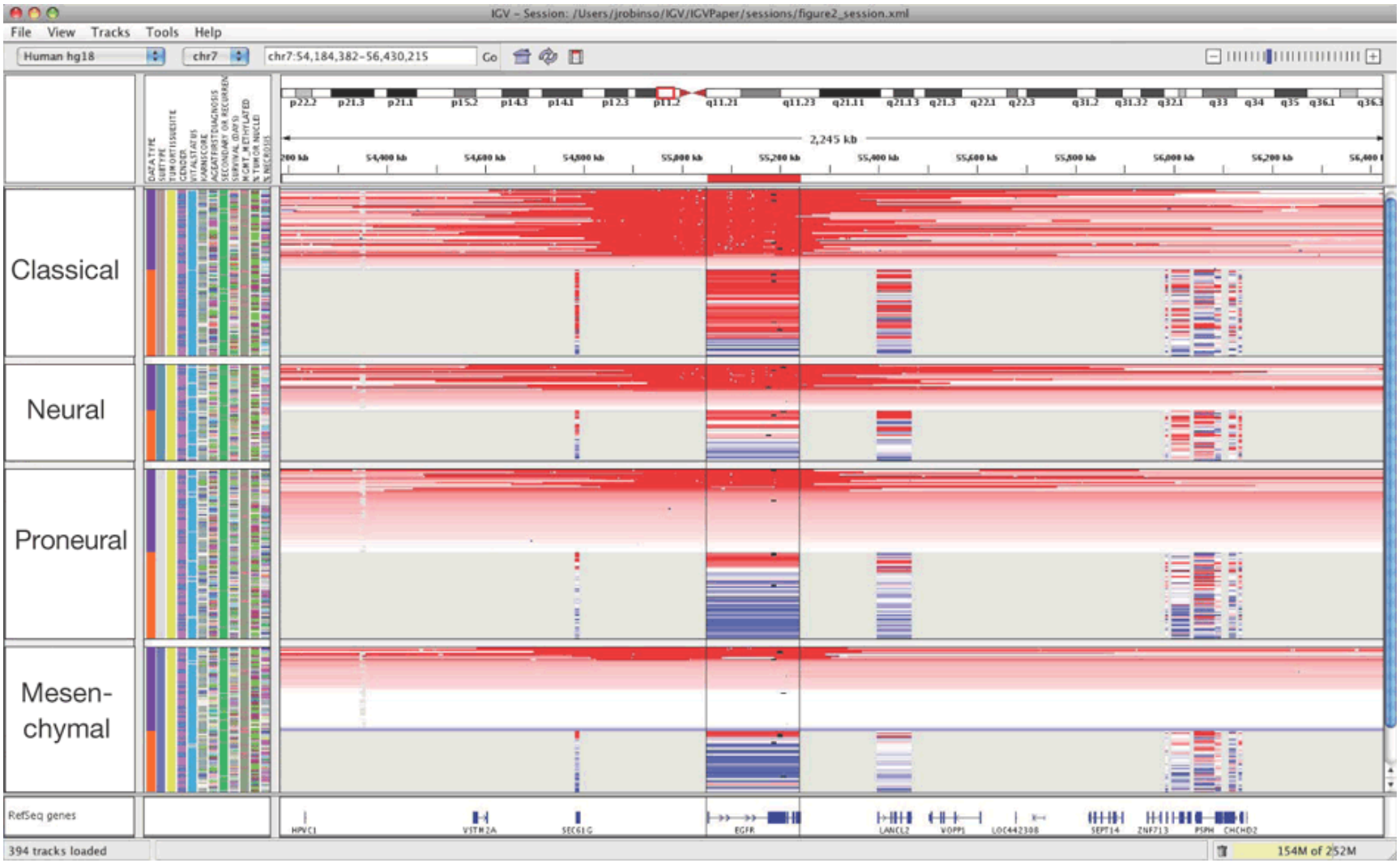
Existing Tools

UCSC IGB IGV Circos GViz



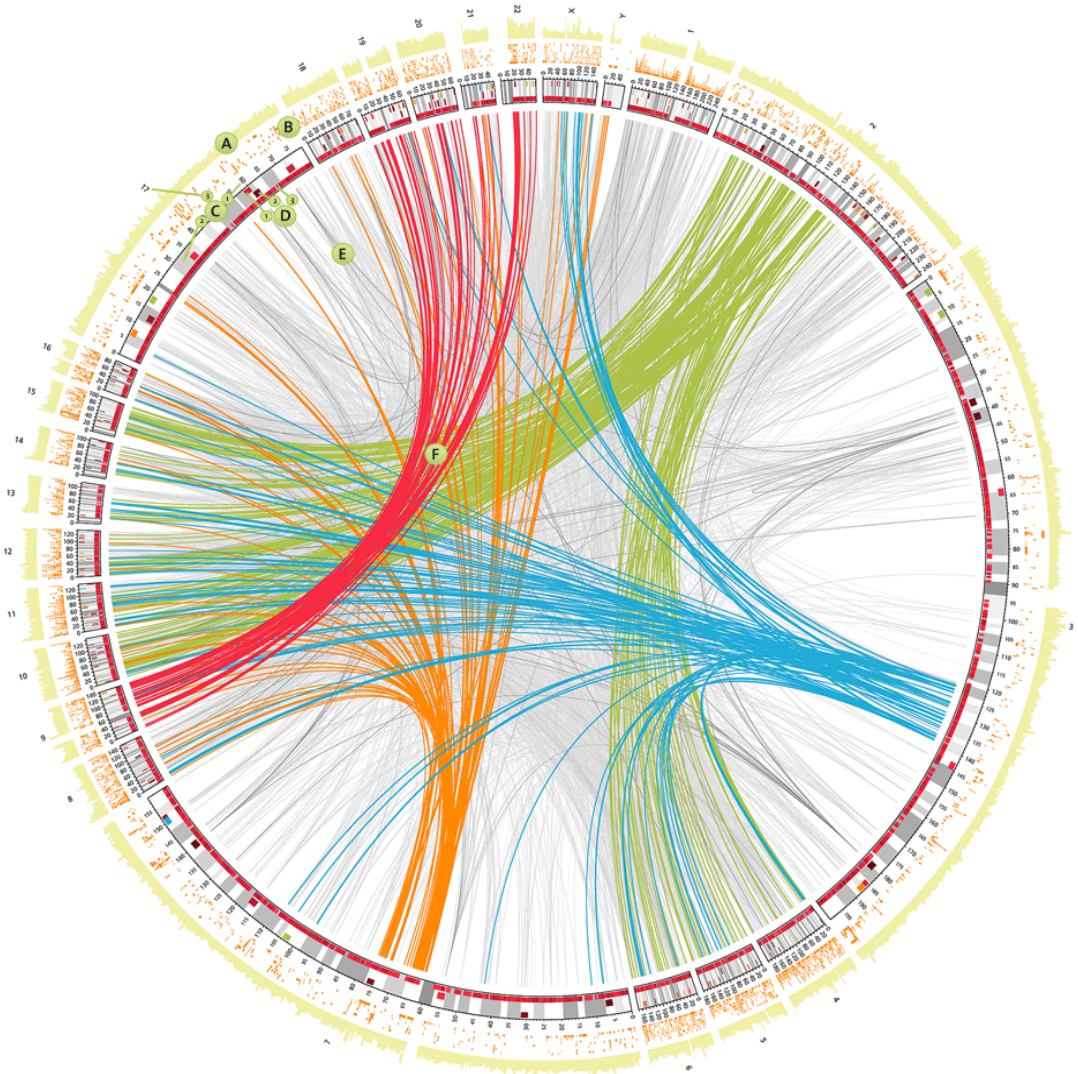
Existing Tools

UCSC IGB **IGV** Circos GViz



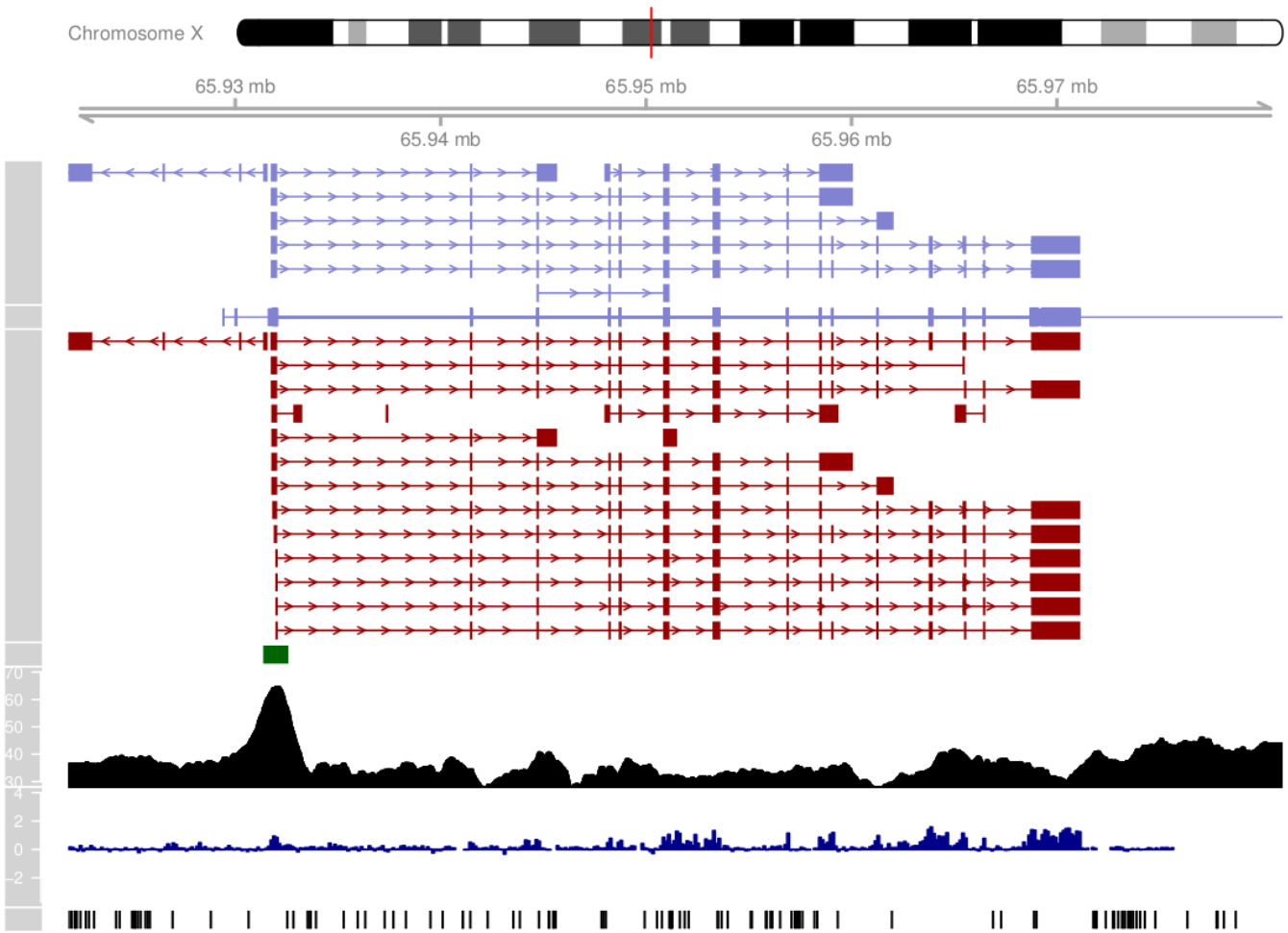
Existing Tools

UCSC IGB IGV Circos GViz



Existing Tools

UCSC IGB IGV Circos GViz



Existing Tools

UCSC	IGB	IGV	Circos	GViz

Limitations

- Limited to one type of view (linear or circular)
- Not tightly integrated with an analysis environment through standard, abstract data structures (except GViz)
- No low-level toolkit for prototyping new types of graphics

Grammars of Graphics

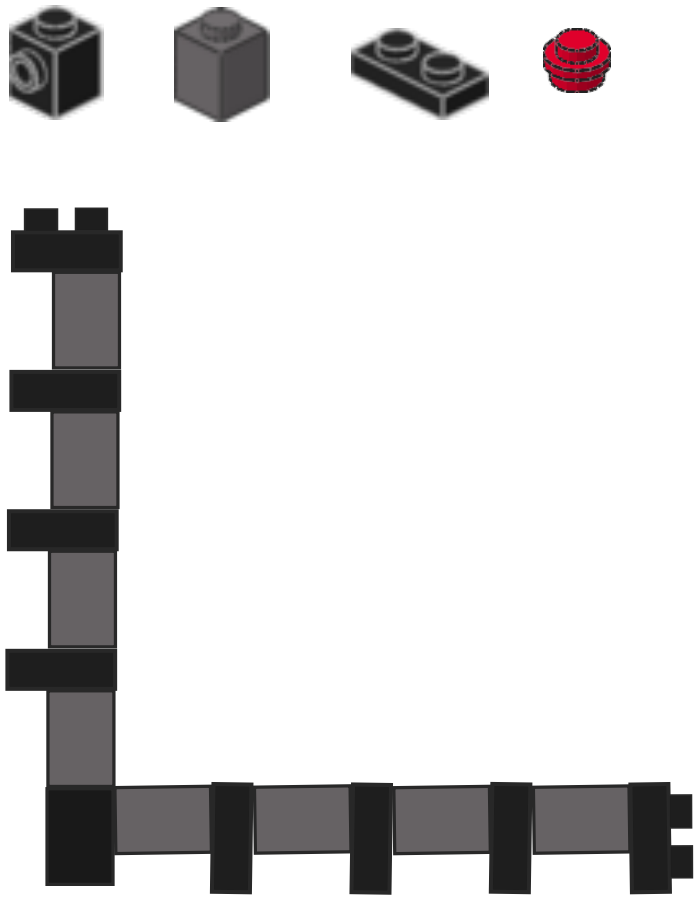
- A grammar of graphics is a language for expressing plots
- Graphics are constructed through the combination of various types of primitives; like legos for graphics
- The most prominent grammar was introduced by Wilkinson's book *The Grammar of Graphics*
- Wilkinson's grammar was extended by Wickham and the *ggplot2* package

Grammars of Graphics

- A grammar of graphics is a language for expressing plots
- Graphics are constructed through the combination of various types of primitives; like legos for graphics
- The most prominent grammar was introduced by Wilkinson's book *The Grammar of Graphics*
- Wilkinson's grammar was extended by Wickham and the *ggplot2* package

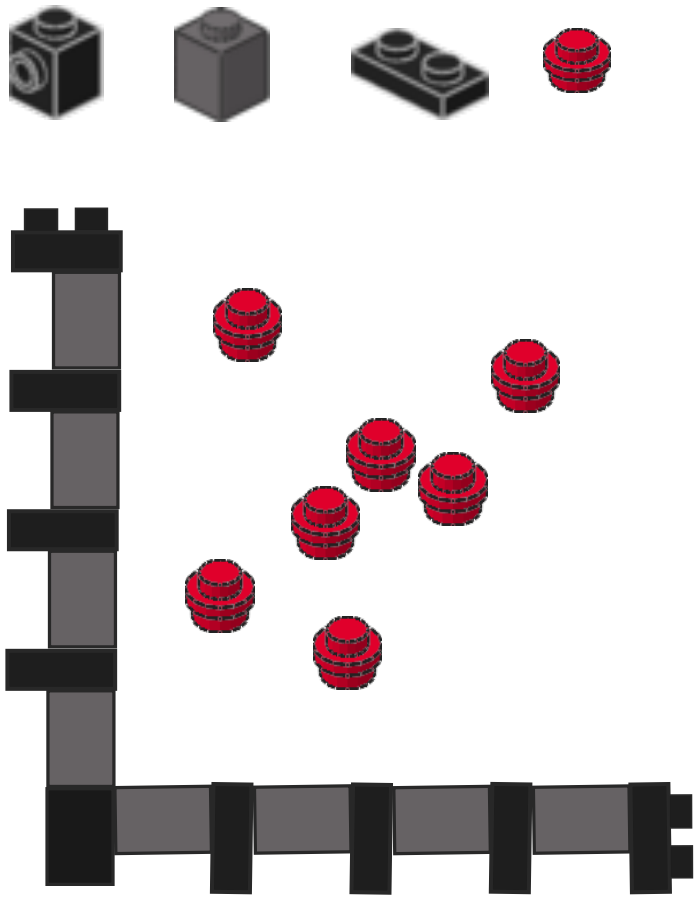
Grammars of Graphics

- A grammar of graphics is a language for expressing plots
- Graphics are constructed through the combination of various types of primitives; like legos for graphics
- The most prominent grammar was introduced by Wilkinson's book *The Grammar of Graphics*
- Wilkinson's grammar was extended by Wickham and the *ggplot2* package



Grammars of Graphics

- A grammar of graphics is a language for expressing plots
- Graphics are constructed through the combination of various types of primitives; like legos for graphics
- The most prominent grammar was introduced by Wilkinson's book *The Grammar of Graphics*
- Wilkinson's grammar was extended by Wickham and the *ggplot2* package



The *ggbio* Package

- An R/Bioconductor package that extends the Wilkinson/Wickham grammar for applications in genomics
- Integrated with Bioconductor
  - Operates on standard, abstract genomic data structures
  - Leverages efficient range-based algorithms
- Programming interface has two levels of abstraction:

  `autoplot` Maps Bioconductor data structures to plots

  `grammar` Mix and match to create custom plots

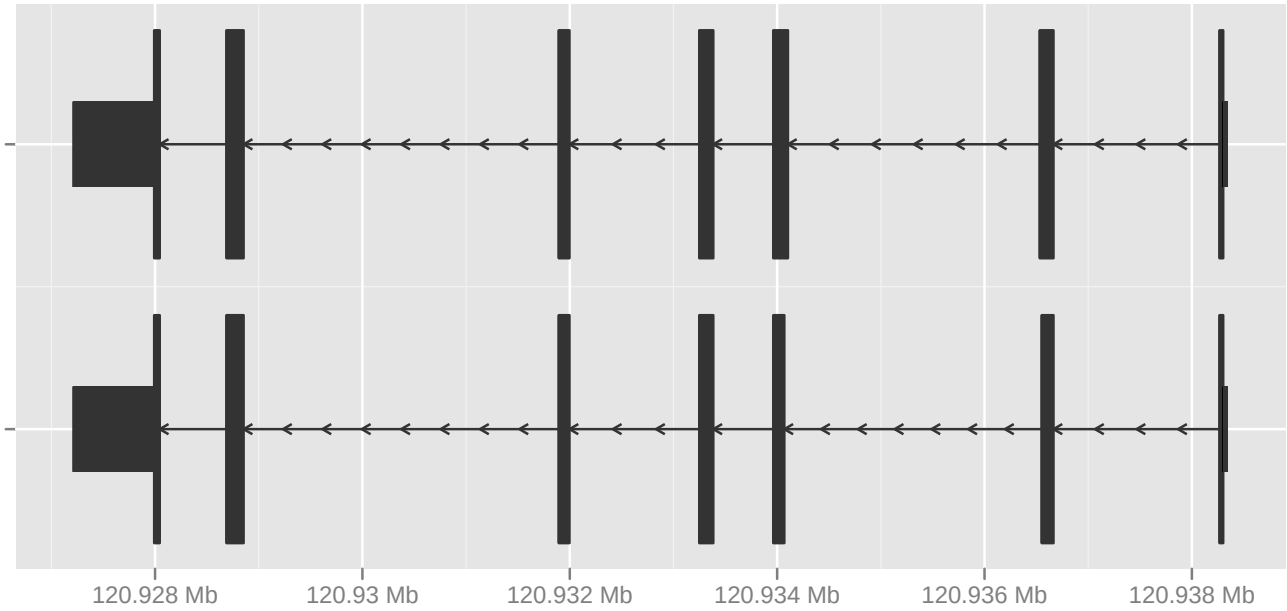
Outline

Basic Plots

Gene Structures | Read Alignments | Sequence | Multiple

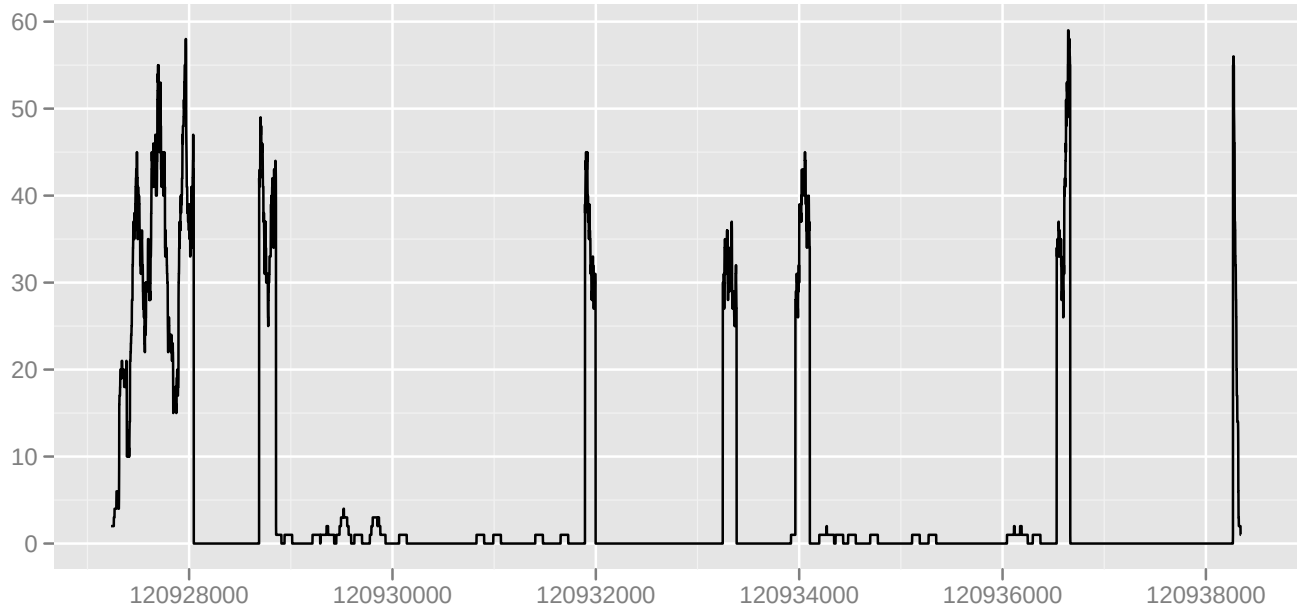
Basic Plots

Gene Structures | Read Alignments | Sequence | Multiple



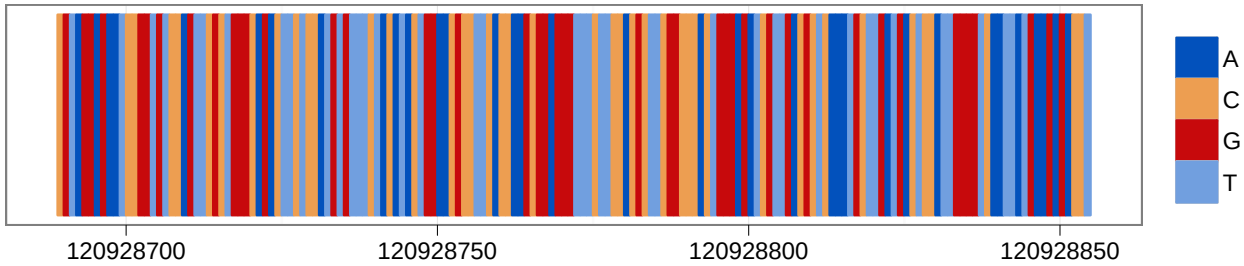
Basic Plots

Gene Structures | Read Alignments | Sequence | Multiple



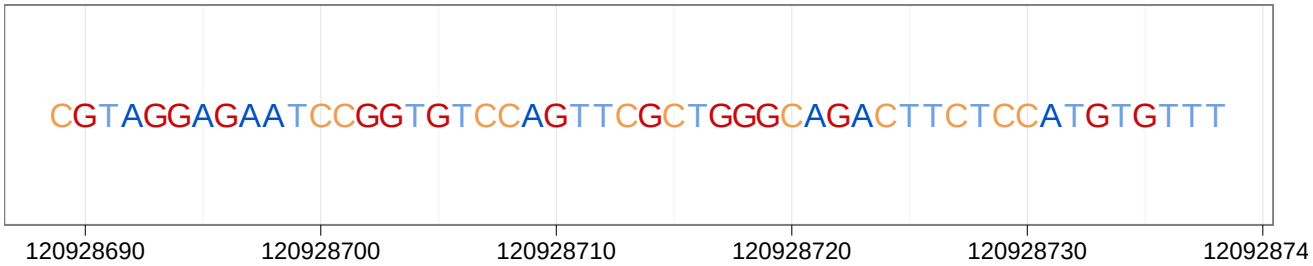
Basic Plots

Gene Structures | Read Alignments | Sequence | Multiple



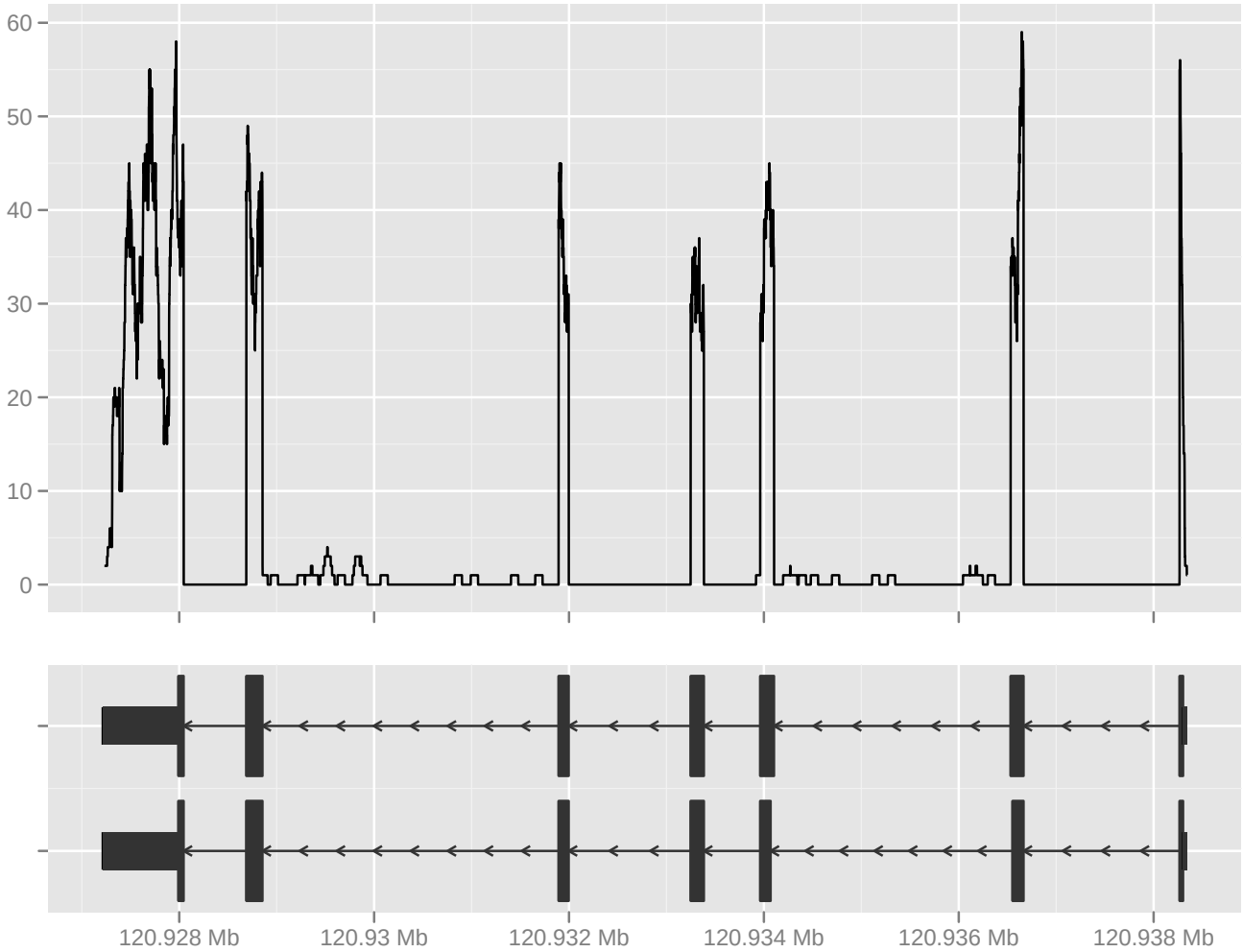
Basic Plots

Gene Structures | Read Alignments | Sequence | Multiple



Basic Plots

Gene Structures | Read Alignments | Sequence | Multiple

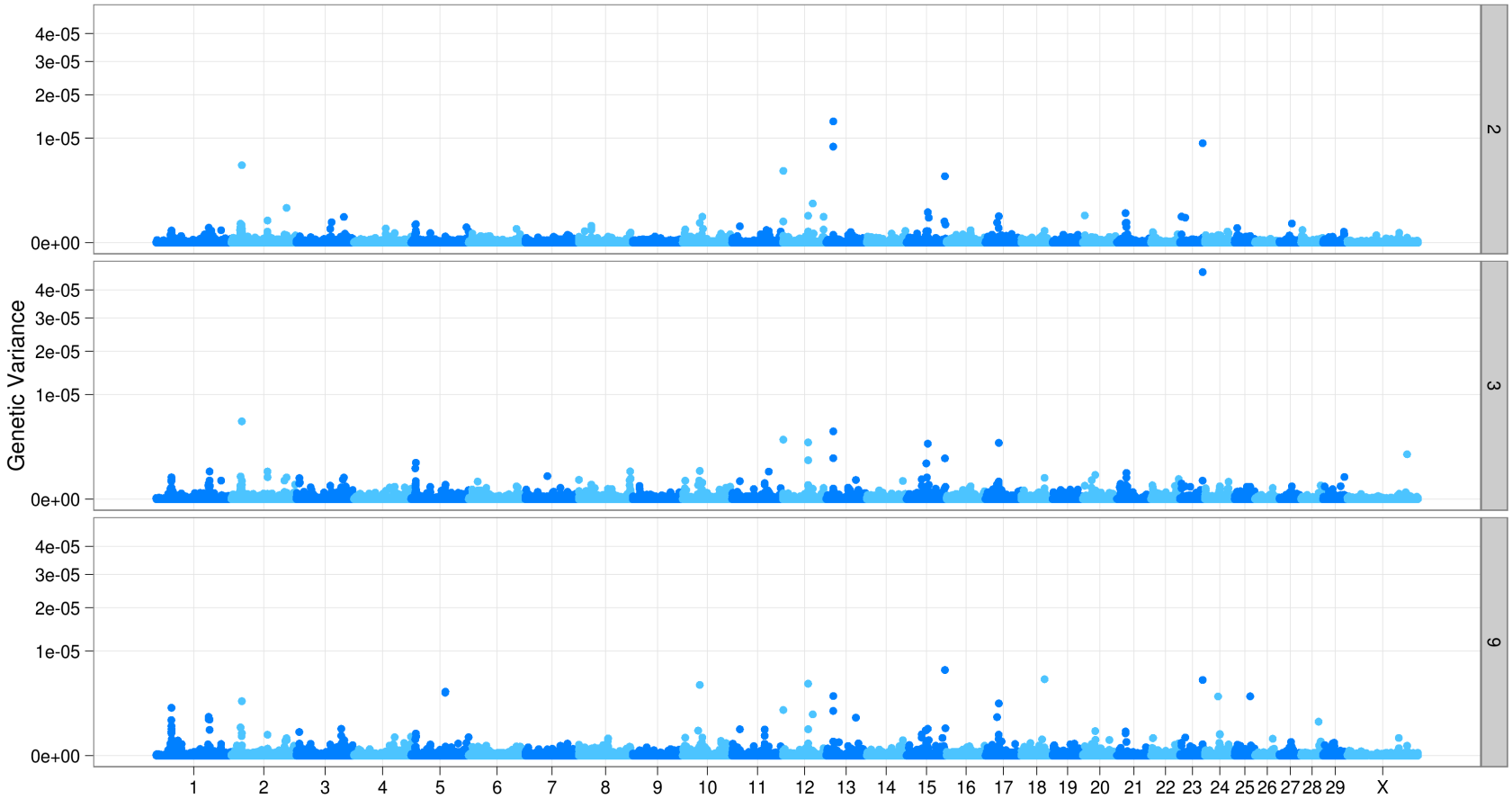


Overview Plots

Grand Linear | Karyogram | Circular

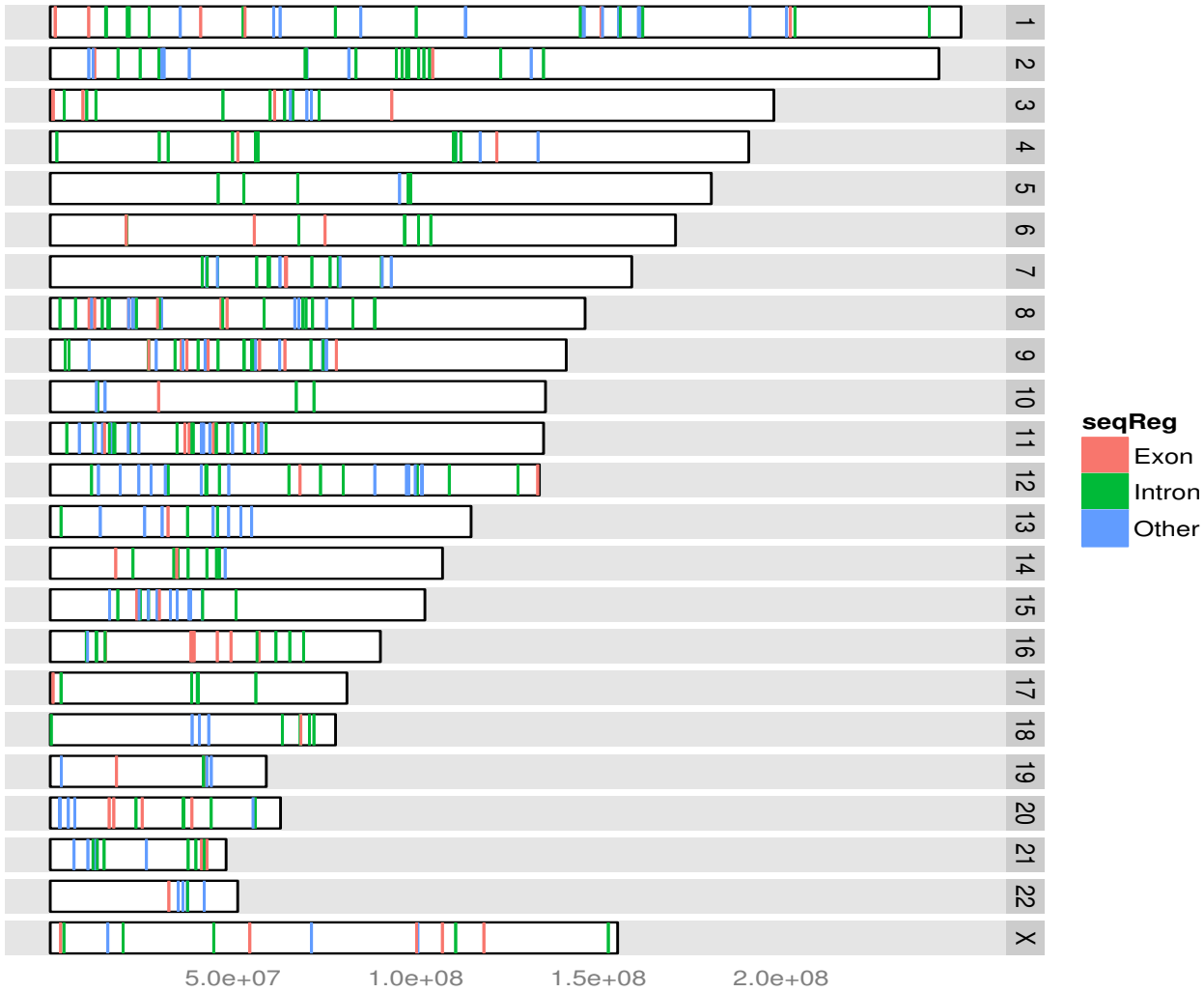
Overview Plots

Grand Linear | Karyogram | Circular



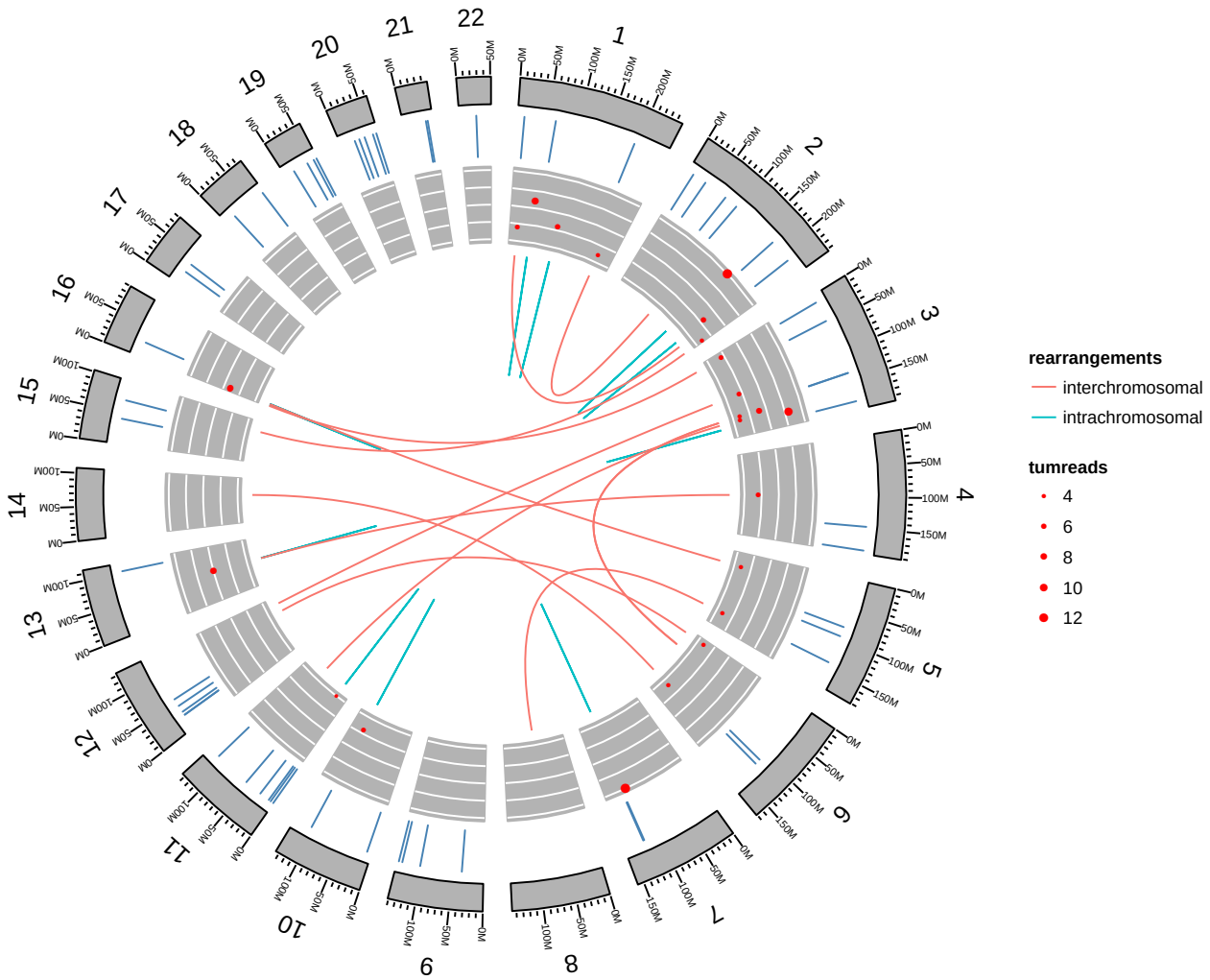
Overview Plots

Grand Linear | Karyogram | Circular



Overview Plots

Grand Linear | Karyogram | Circular



Specialized Plots

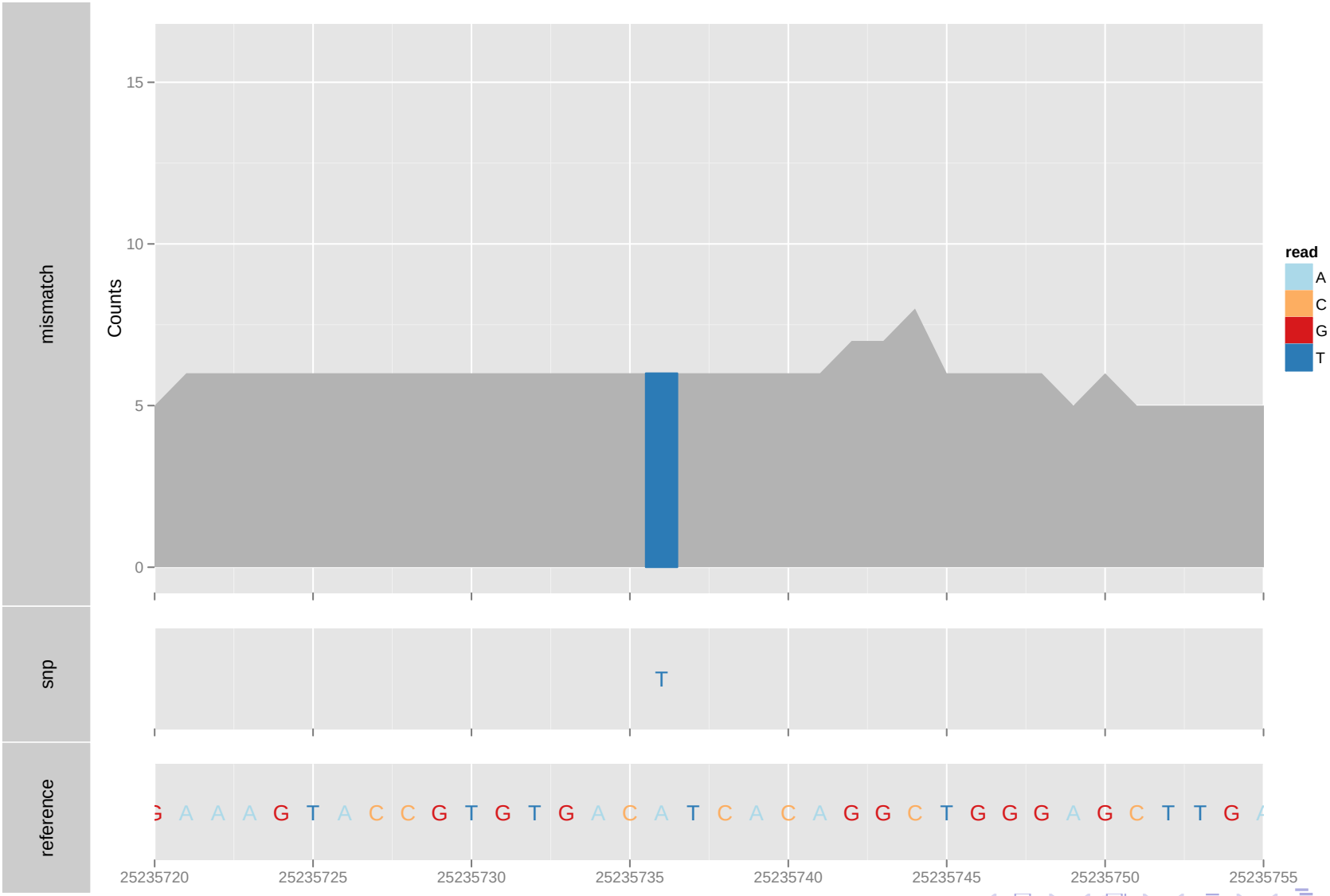
Mismatch summary + VCF

Edge-linked Intervals

Specialized Plots

Mismatch summary + VCF

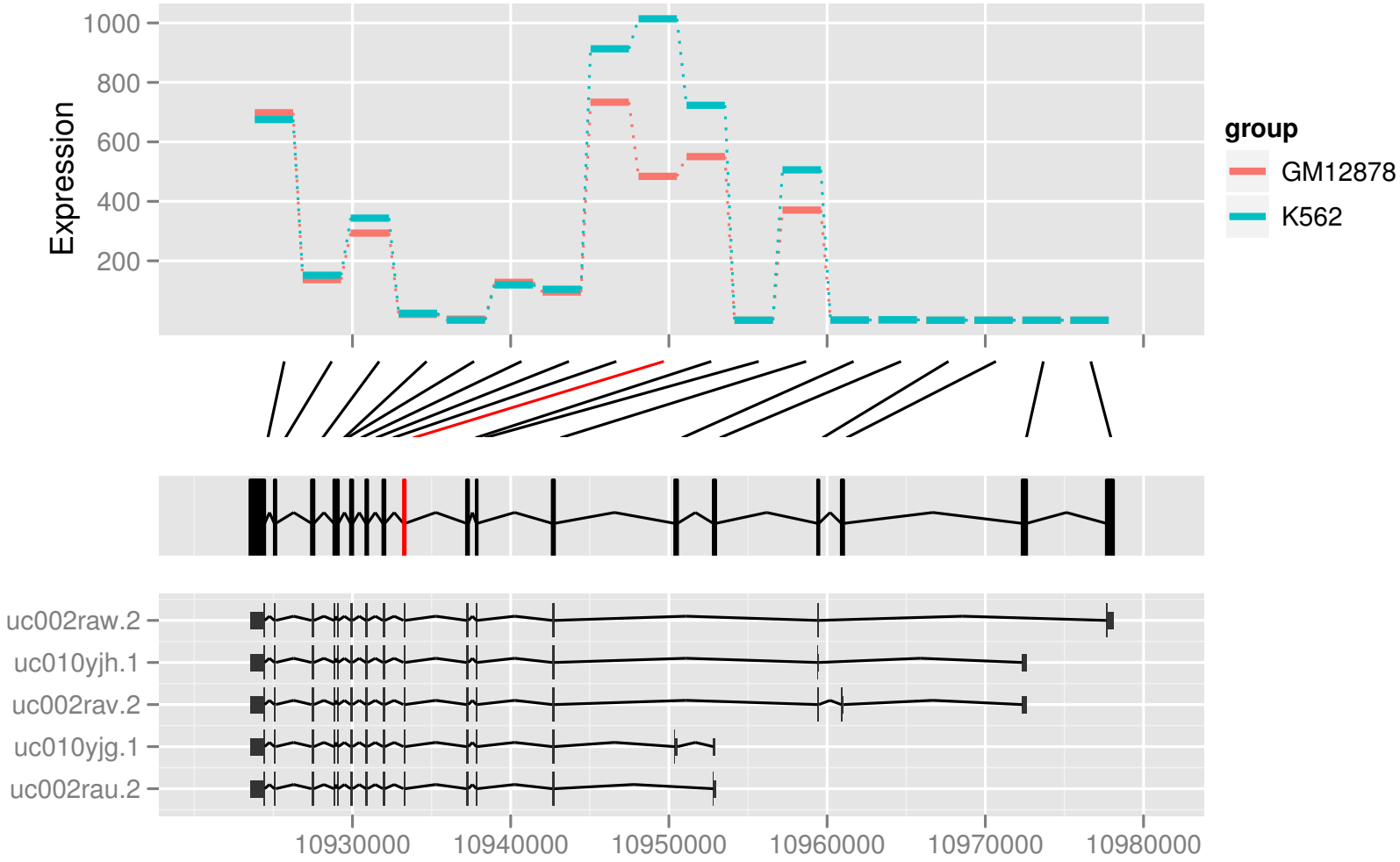
Edge-linked Intervals



Specialized Plots

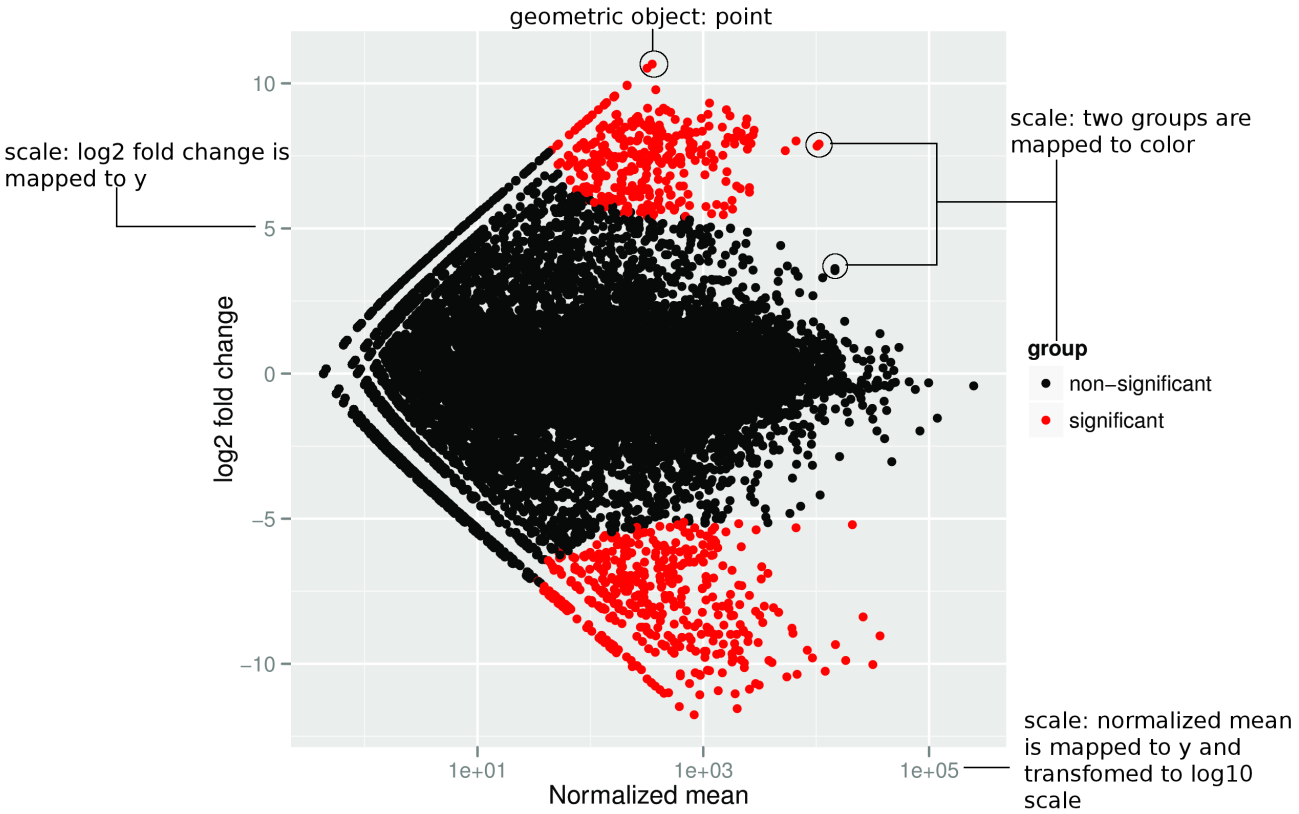
Mismatch summary + VCF

Edge-linked Intervals



Outline

The Wilkinson/Wickham Grammar of Graphics



Geom The shape used for drawing the data

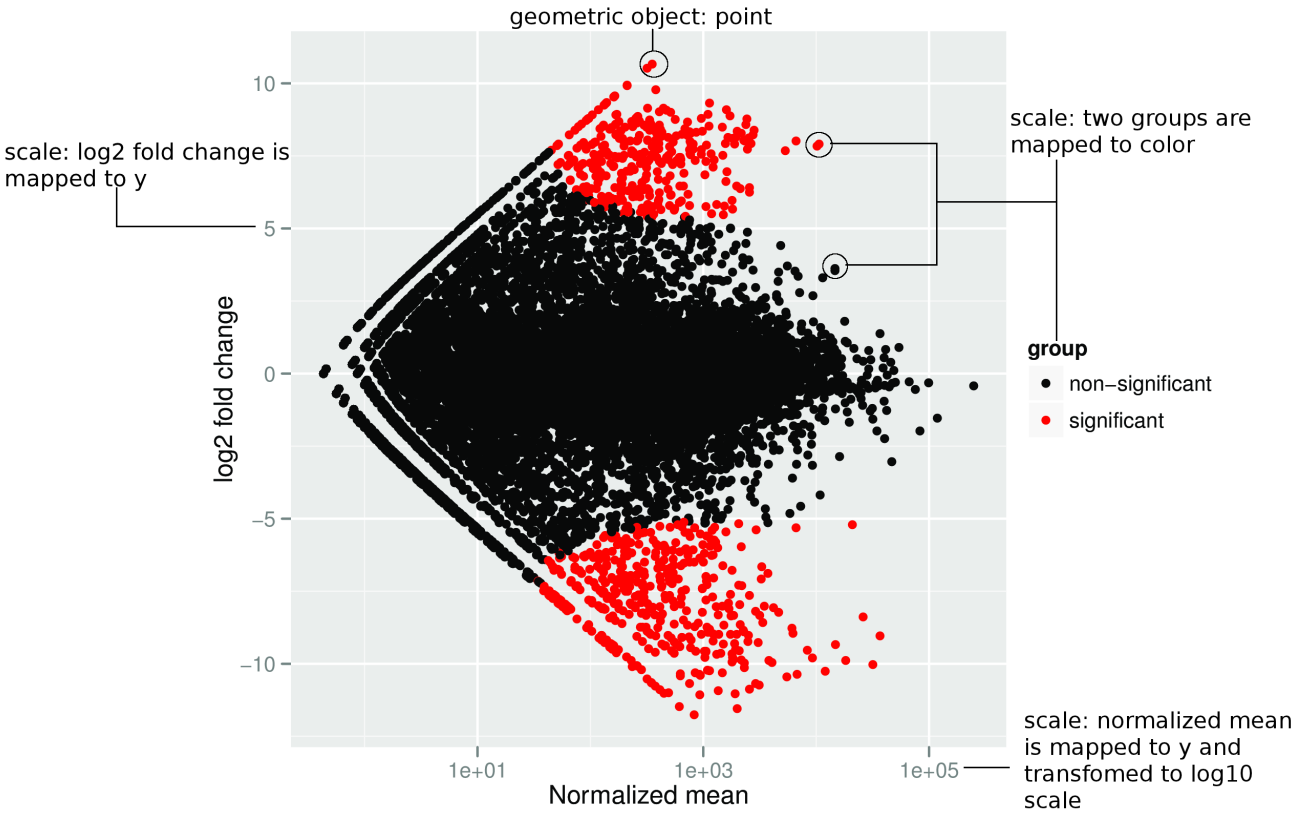
Stat Transforms the data before plotting

Scale Maps data to geom aesthetics, guides like legends and axes

Coord Maps from geom space to device space

Facet Small multiples of data subsets (trellis)

The Wilkinson/Wickham Grammar of Graphics



Geom The shape used for drawing the data

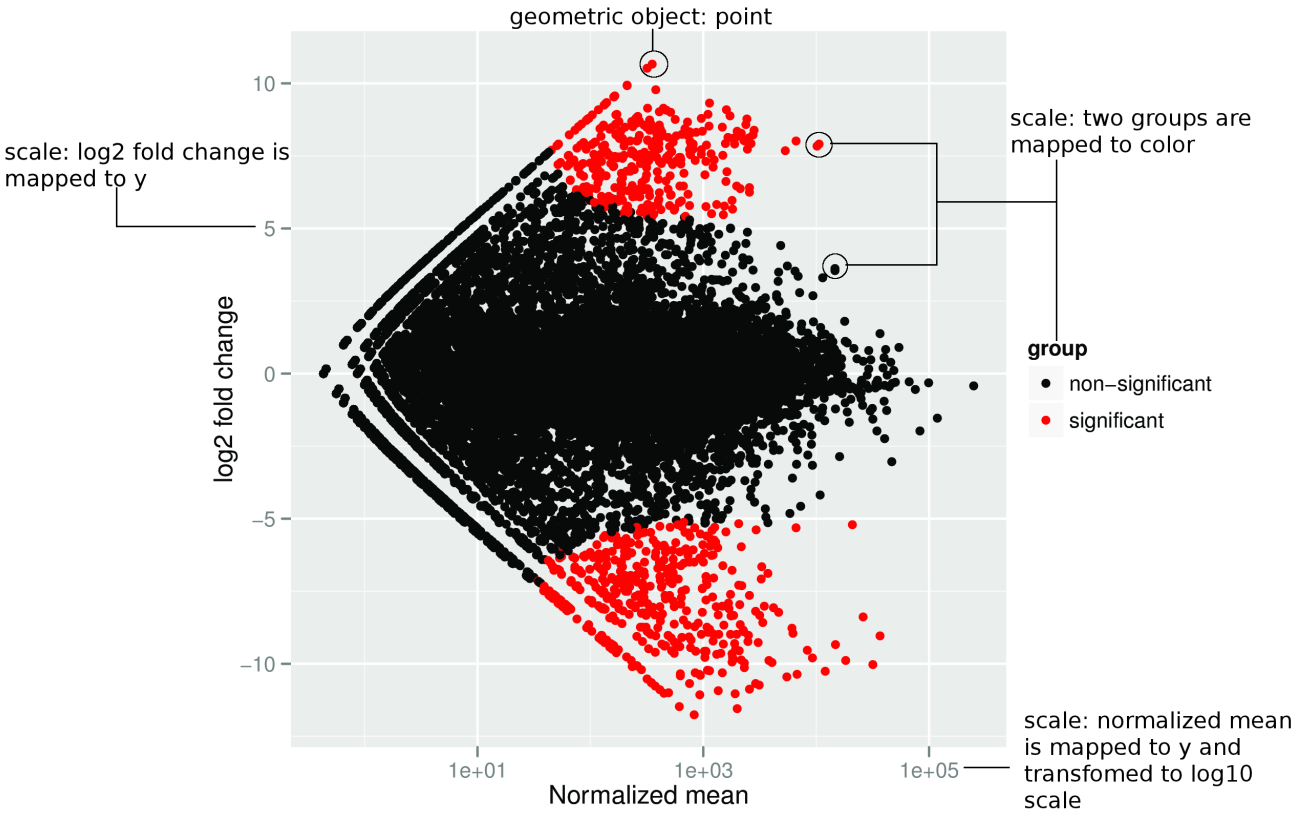
Stat Transforms the data before plotting

Scale Maps data to geom aesthetics, guides like legends and axes

Coord Maps from geom space to device space

Facet Small multiples of data subsets (trellis)

The Wilkinson/Wickham Grammar of Graphics



Geom The shape used for drawing the data

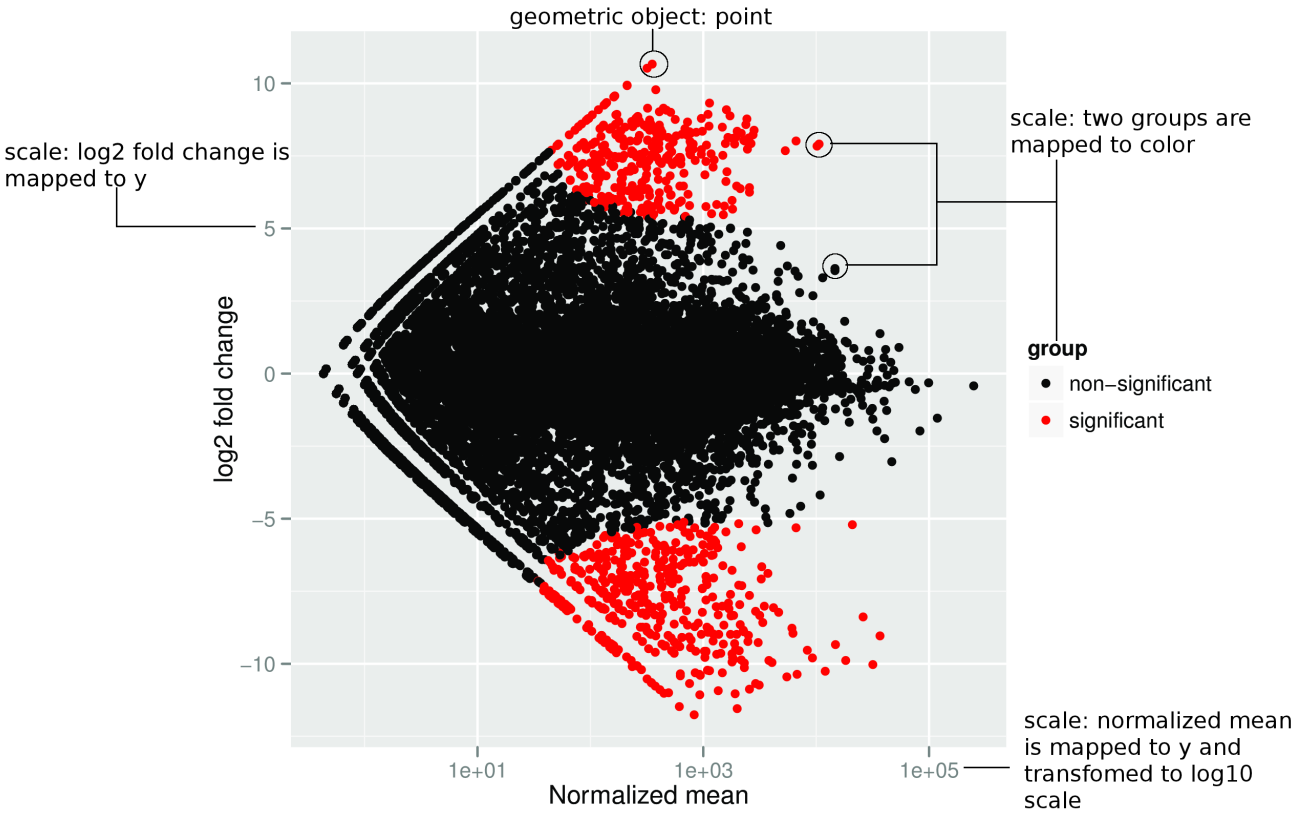
Stat Transforms the data before plotting

Scale Maps data to geom aesthetics, guides like legends and axes

Coord Maps from geom space to device space

Facet Small multiples of data subsets (trellis)

The Wilkinson/Wickham Grammar of Graphics



Geom The shape used for drawing the data

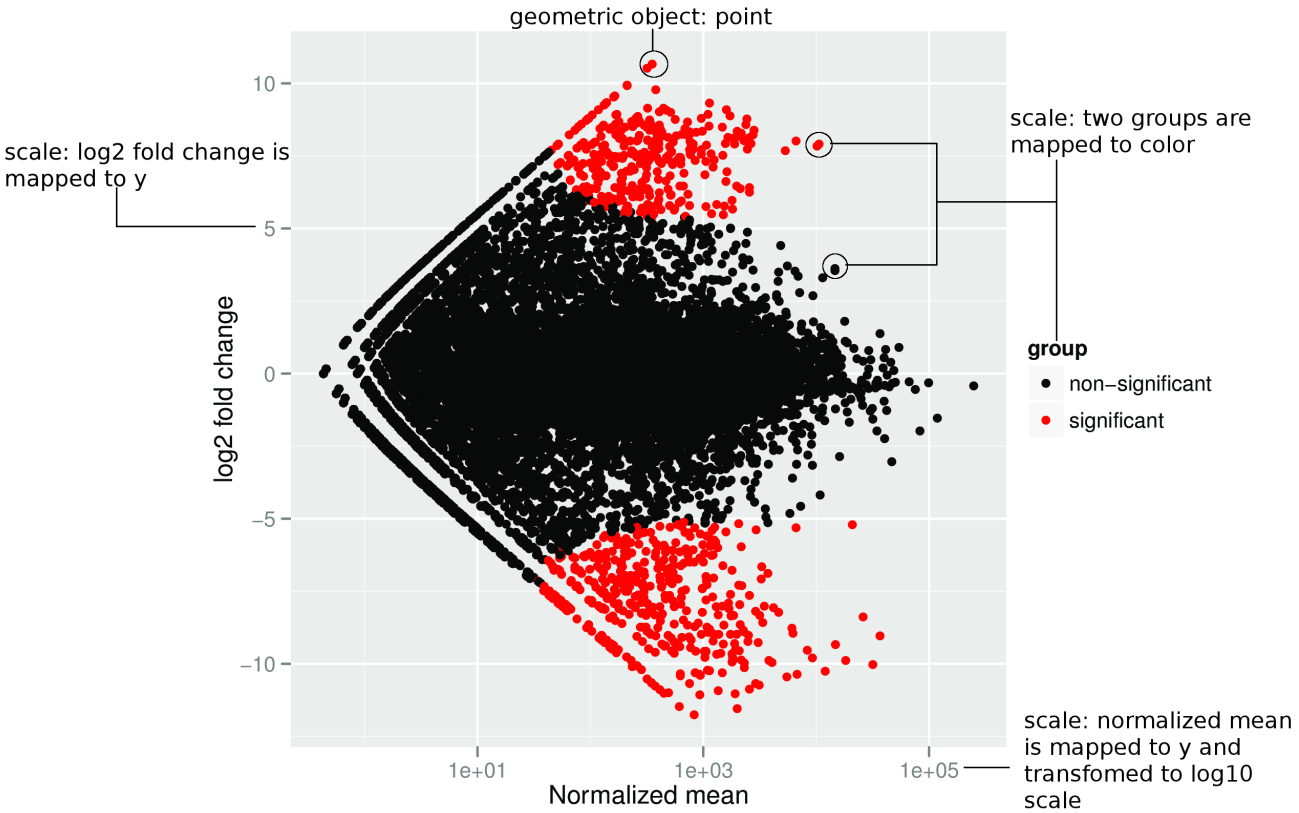
Stat Transforms the data before plotting

Scale Maps data to geom aesthetics, guides like legends and axes

Coord Maps from geom space to device space

Facet Small multiples of data subsets (trellis)

The Wilkinson/Wickham Grammar of Graphics



Geom The shape used for drawing the data

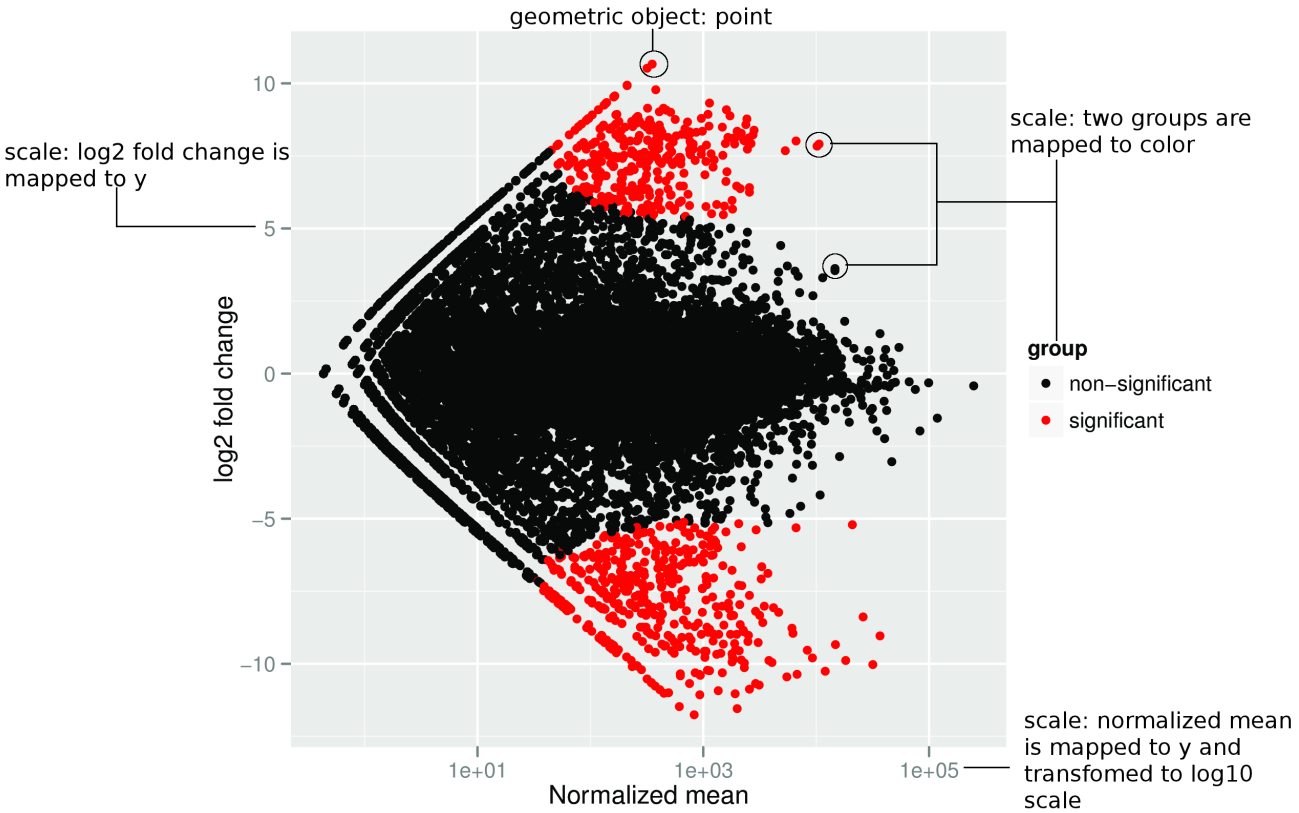
Stat Transforms the data before plotting

Scale Maps data to geom aesthetics, guides like legends and axes

Coord Maps from geom space to device space

Facet Small multiples of data subsets (trellis)

The Wilkinson/Wickham Grammar of Graphics



Geom The shape used for drawing the data

Stat Transforms the data before plotting

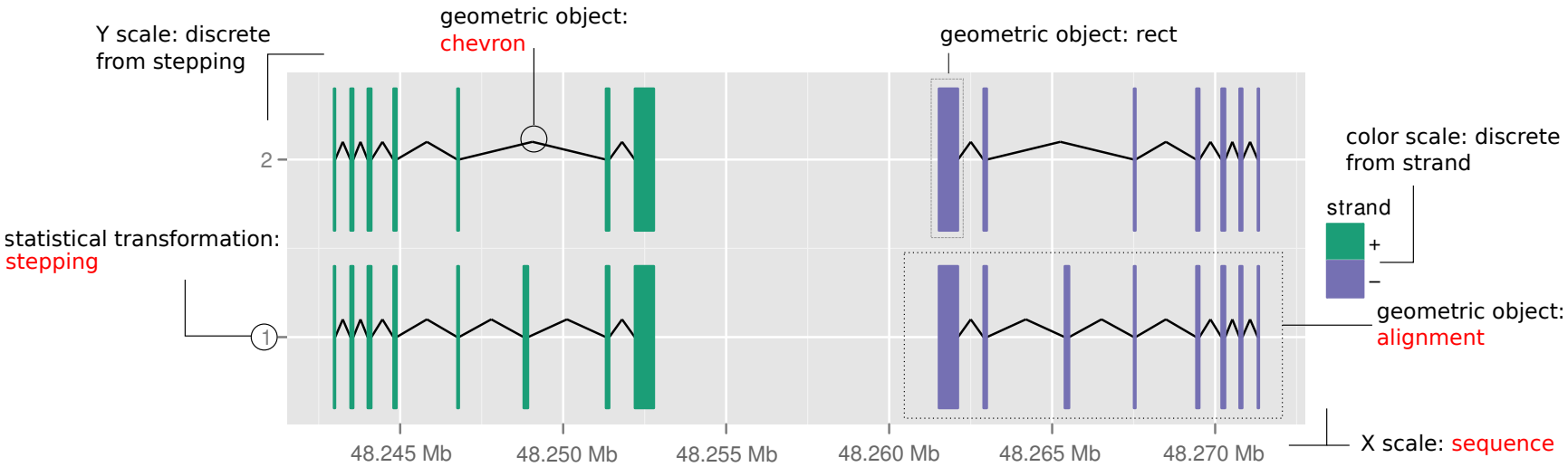
Scale Maps data to geom aesthetics, guides like legends and axes

Coord Maps from geom space to device space

Facet Small multiples of data subsets (trellis)

A Grammar of Graphics for Genomics

Extensions are marked in red

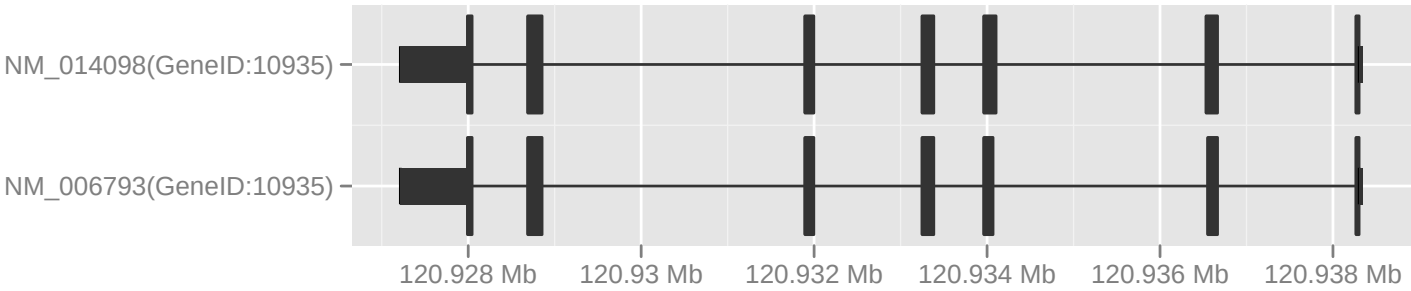


Components of the Genomic Grammar

Geom:	alignment chevron arch arrow arrowrect
Stat:	gene reduce stepping coverage mismatch table
Scale:	sequence genome fold-change giemsa
Coord:	truncate-gaps
Layout:	tracks range-facet

Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet



Components of the Genomic Grammar

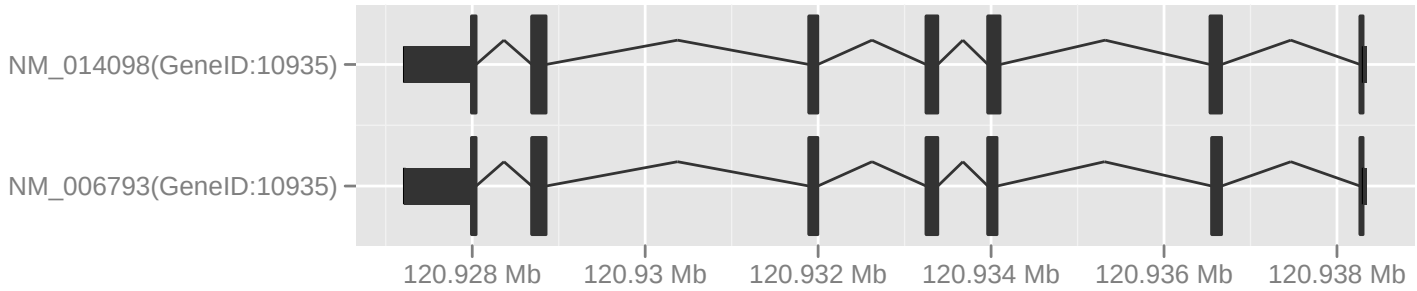
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

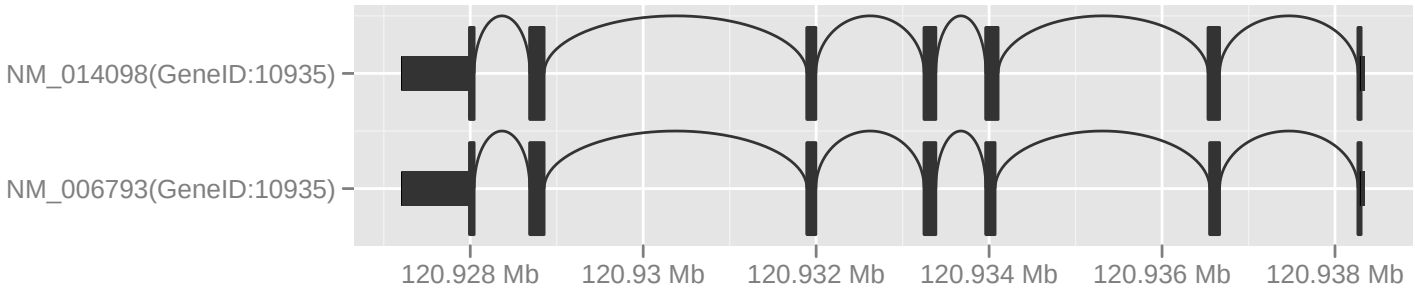
Coord: truncate-gaps

Layout: tracks range-facet



Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet



Components of the Genomic Grammar

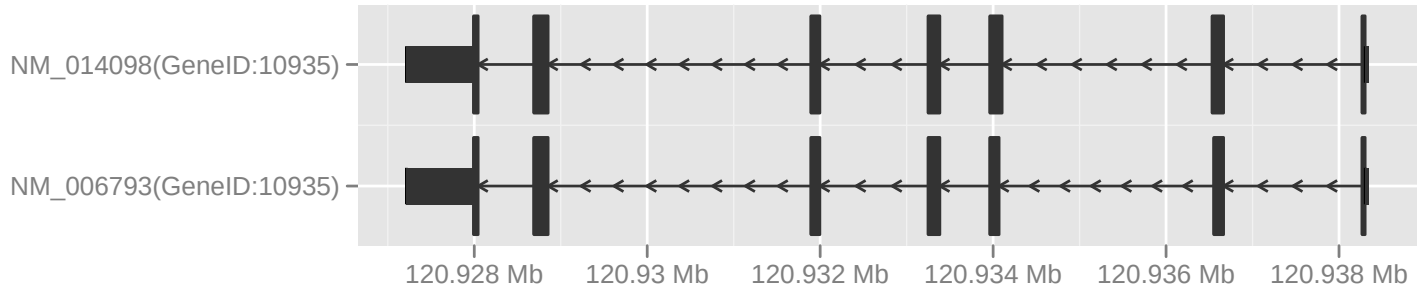
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

Coord: truncate-gaps

Layout: tracks range-facet



Components of the Genomic Grammar

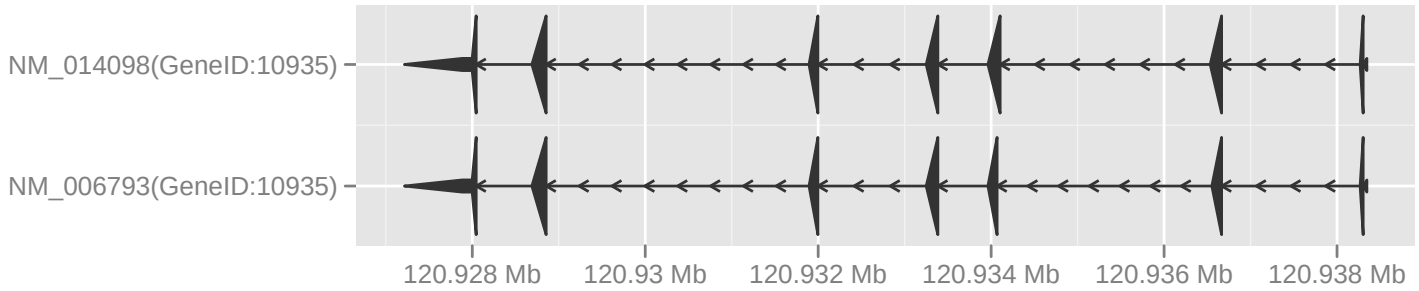
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

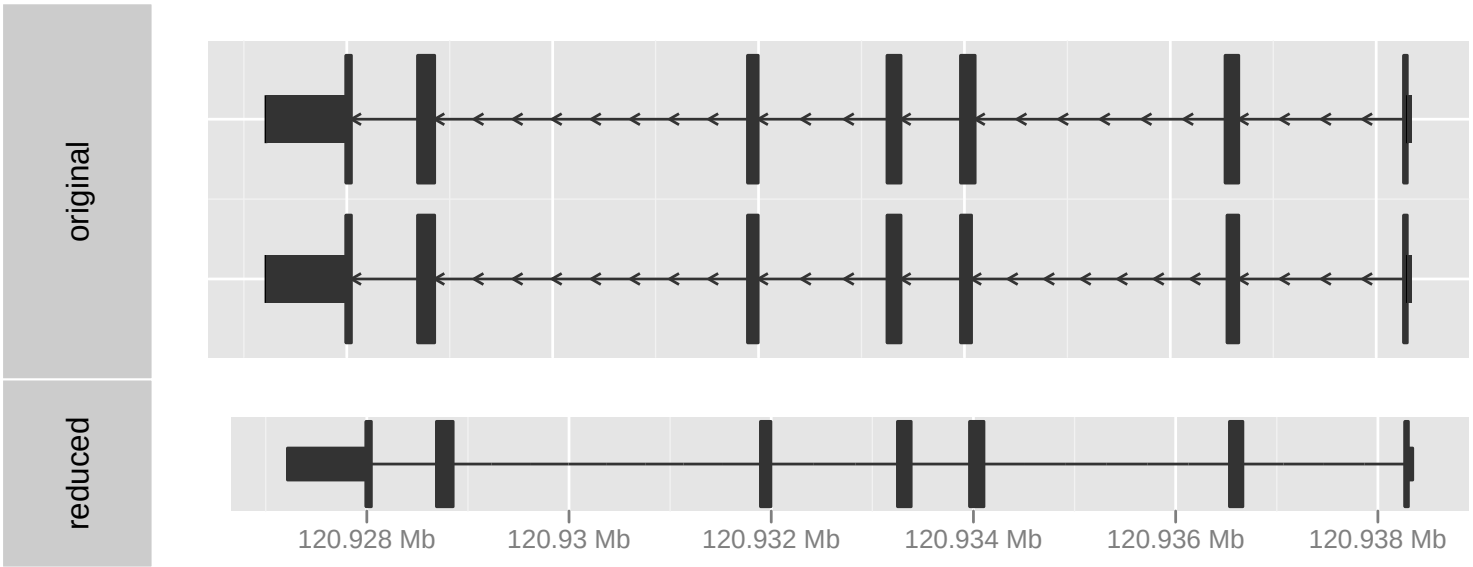
Coord: truncate-gaps

Layout: tracks range-facet



Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet



Components of the Genomic Grammar

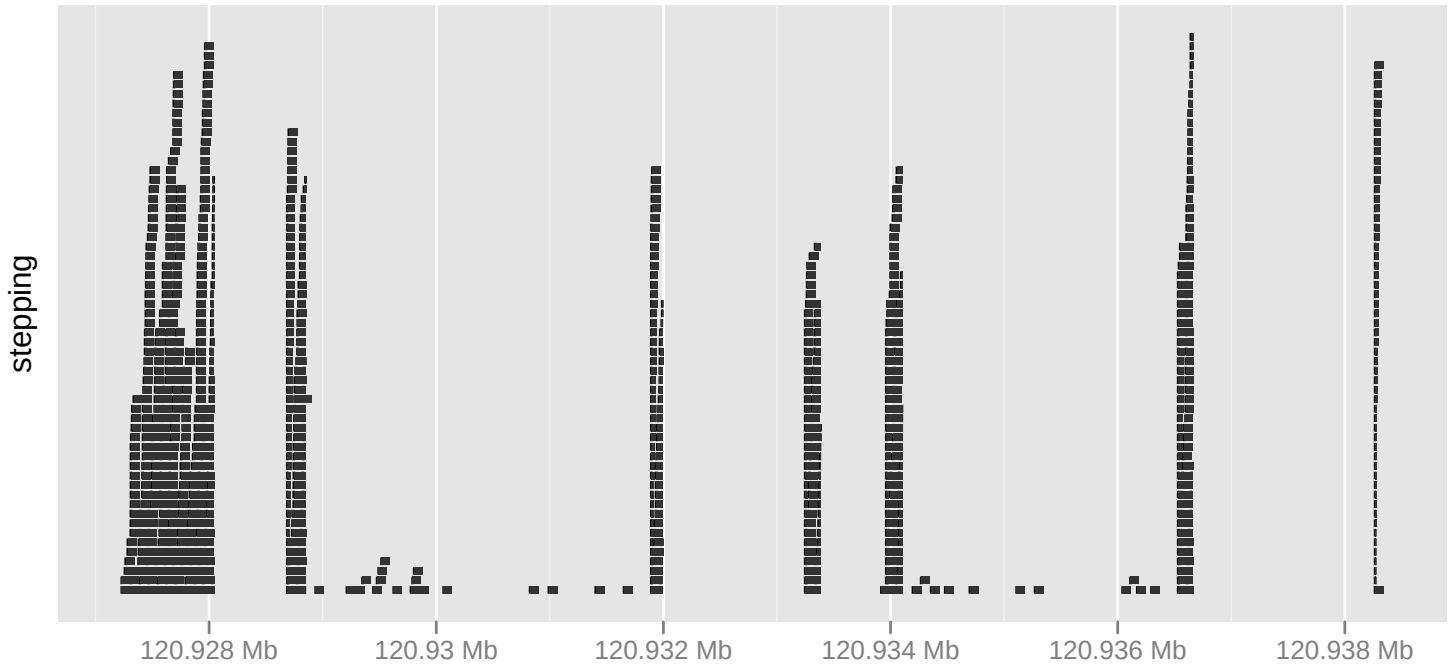
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

Coord: truncate-gaps

Layout: tracks range-facet



Components of the Genomic Grammar

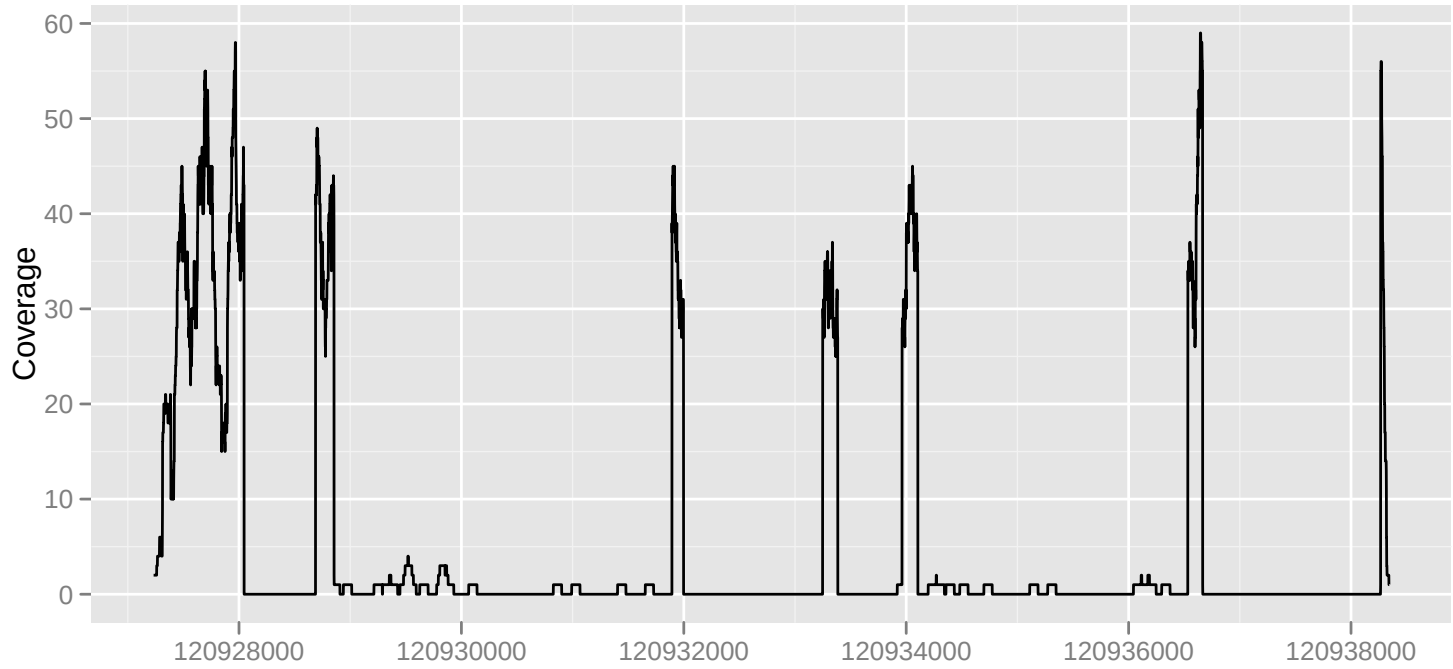
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

Coord: truncate-gaps

Layout: tracks range-facet



Components of the Genomic Grammar

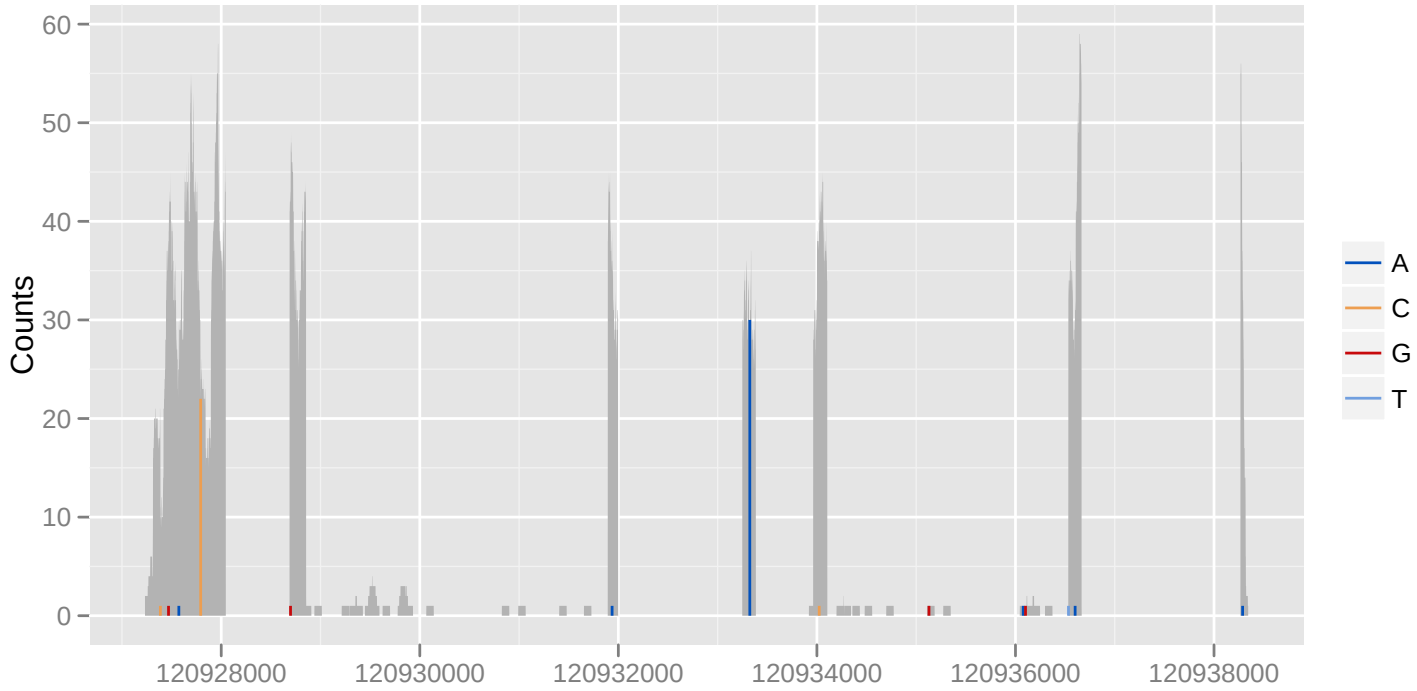
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

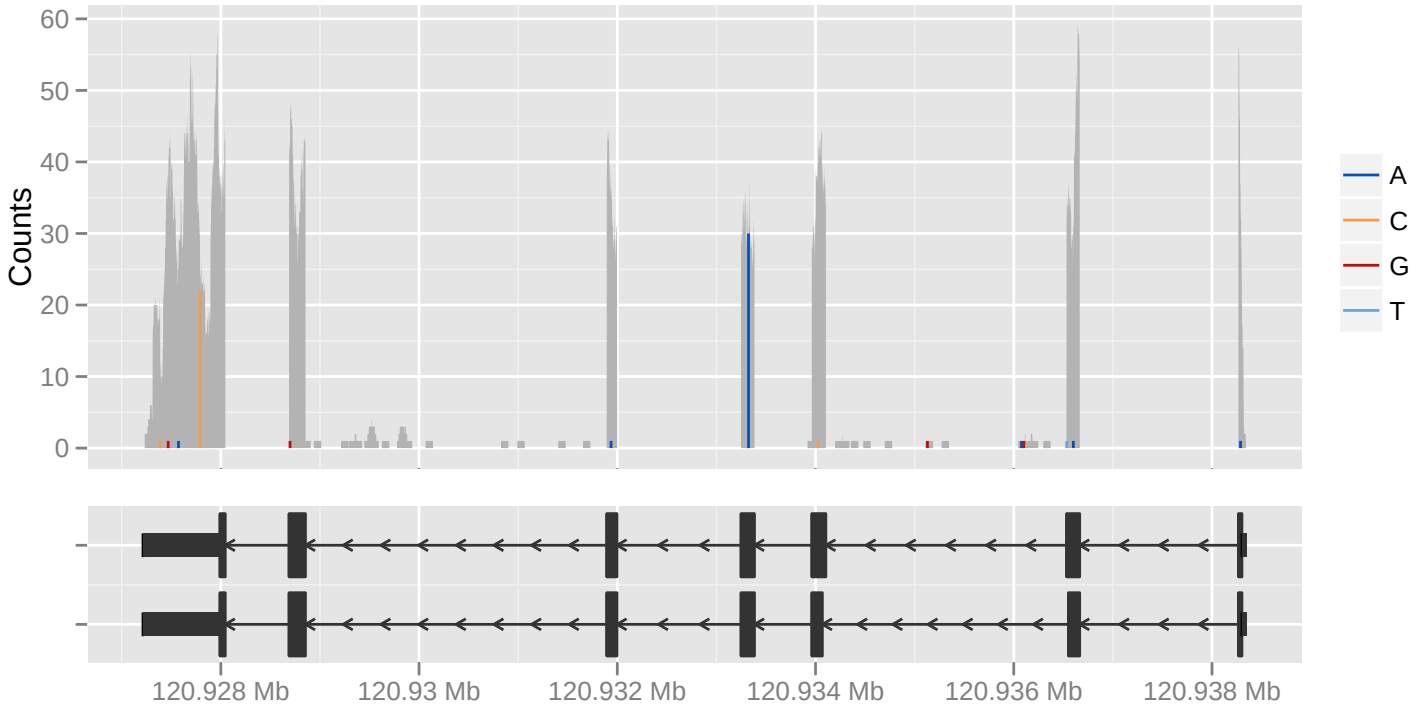
Coord: truncate-gaps

Layout: tracks range-facet



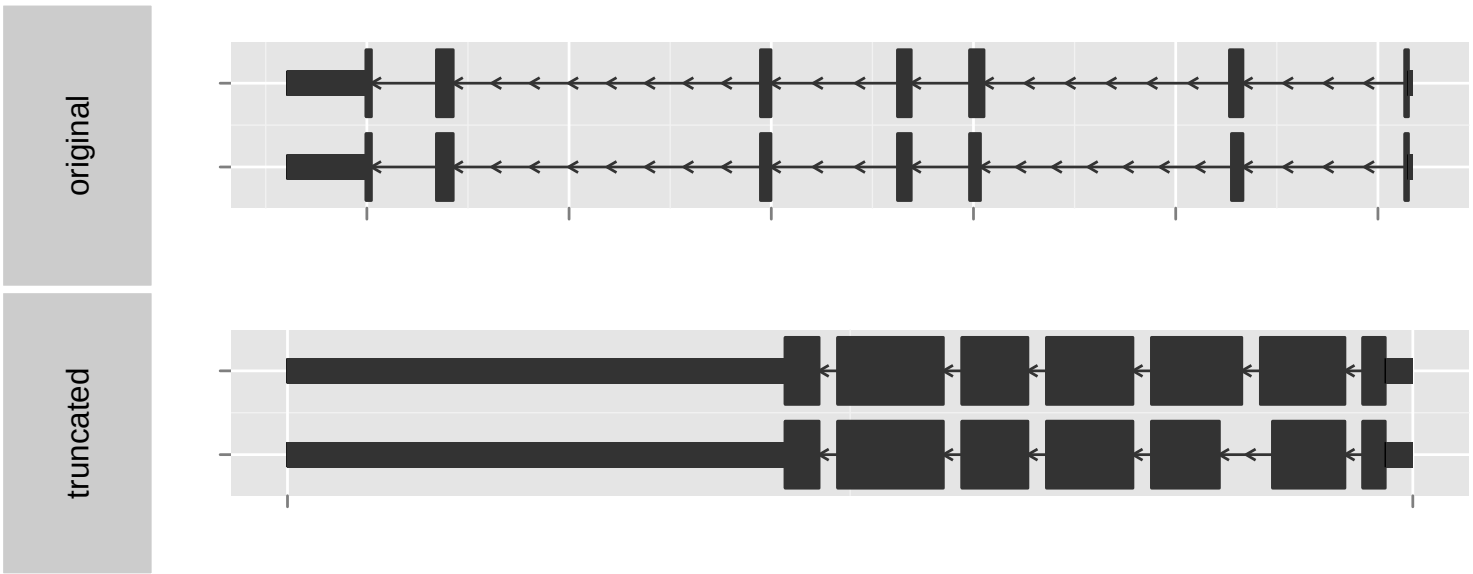
Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet



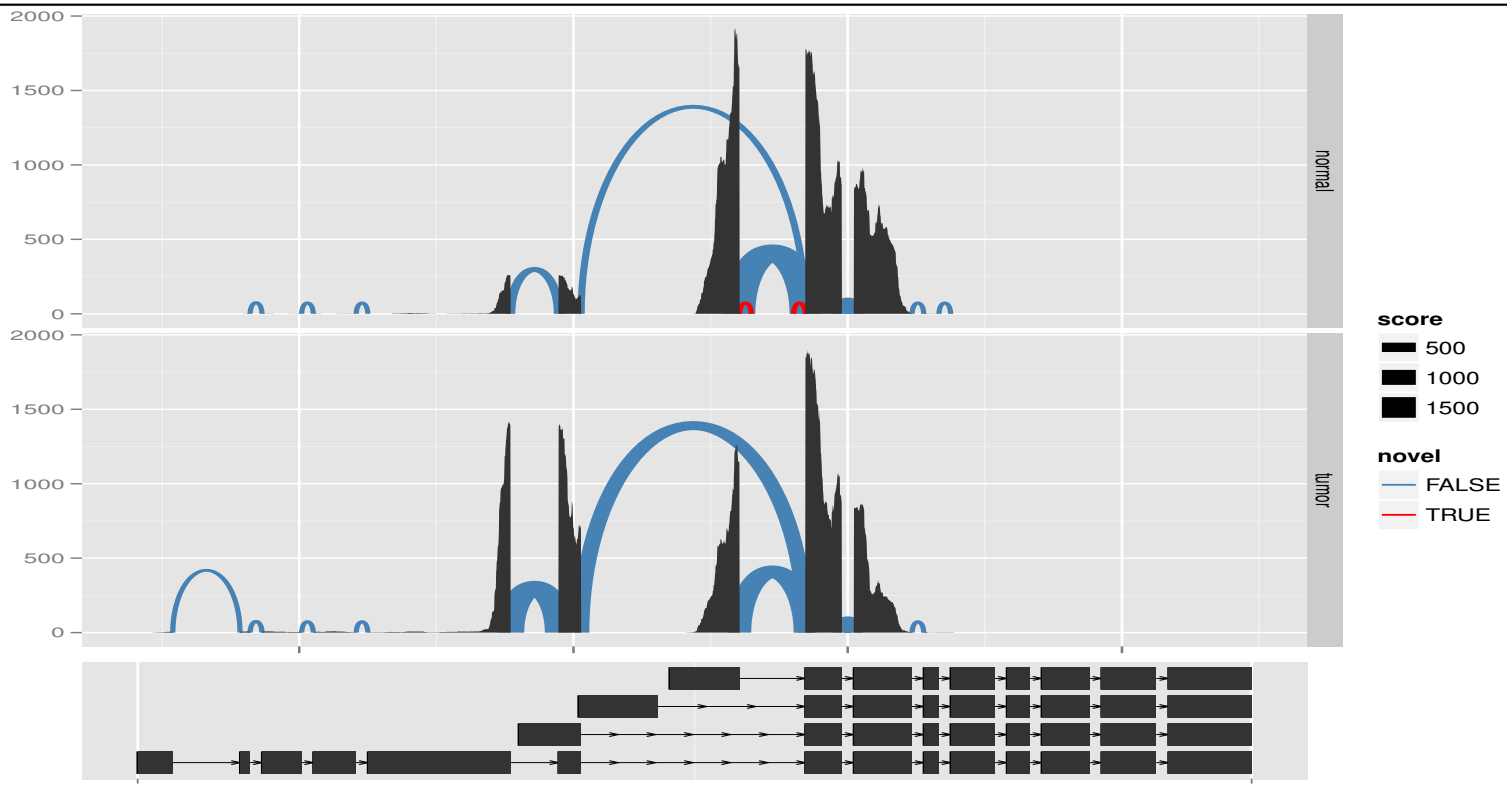
Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet



Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet



Components of the Genomic Grammar

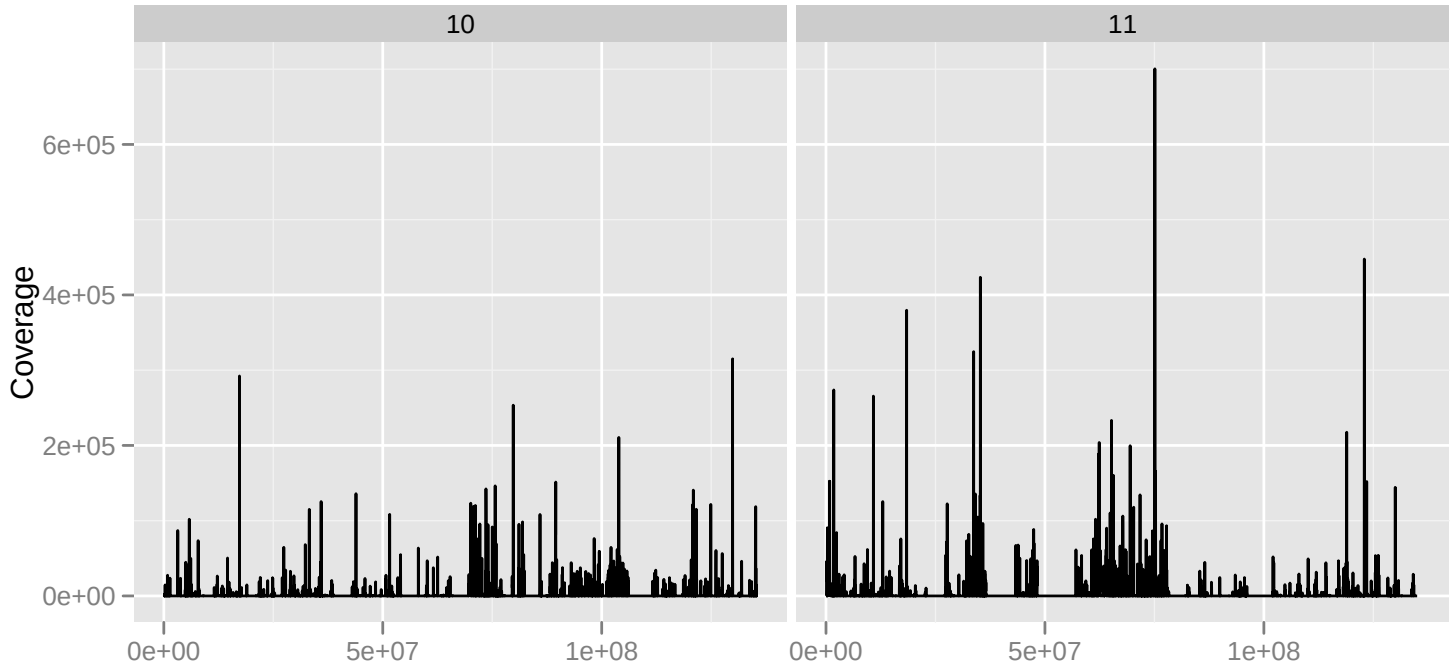
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

Coord: truncate-gaps

Layout: tracks range-facet



Components of the Genomic Grammar

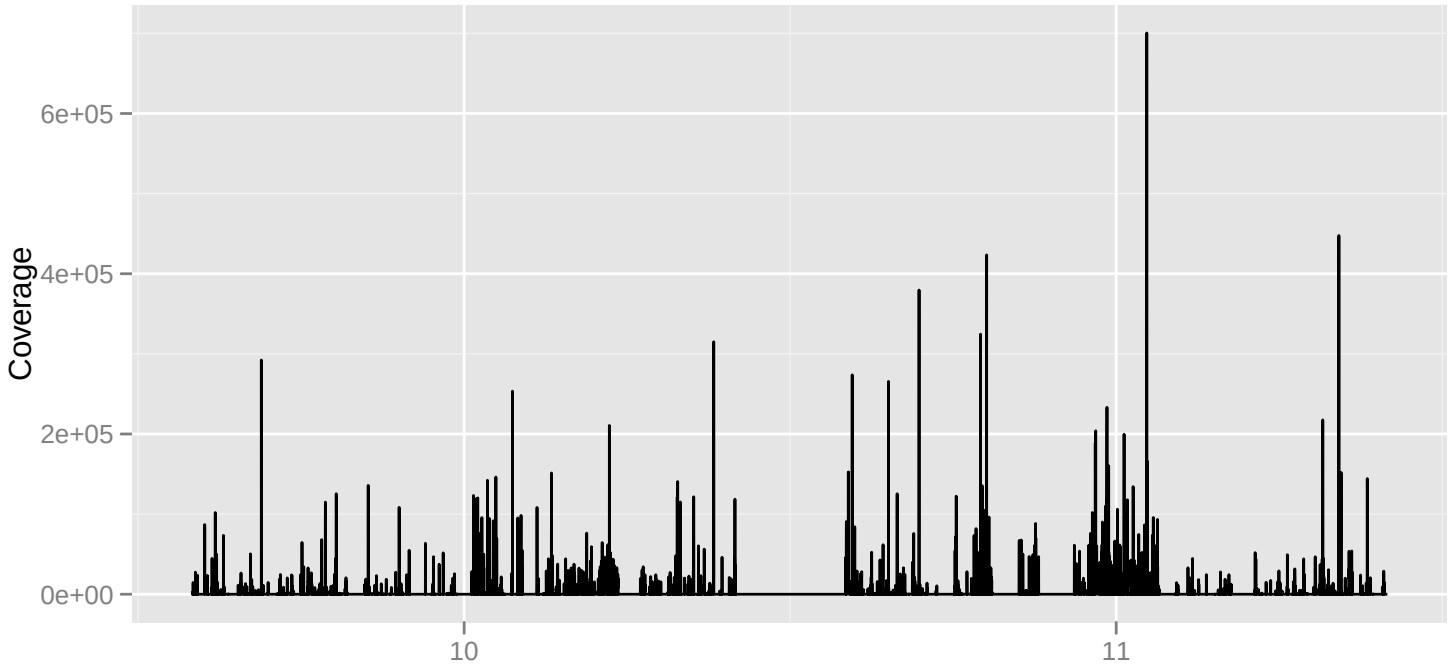
Geom: alignment chevron arch arrow arrowrect

Stat: gene reduce stepping coverage mismatch table

Scale: sequence genome fold-change giemsa

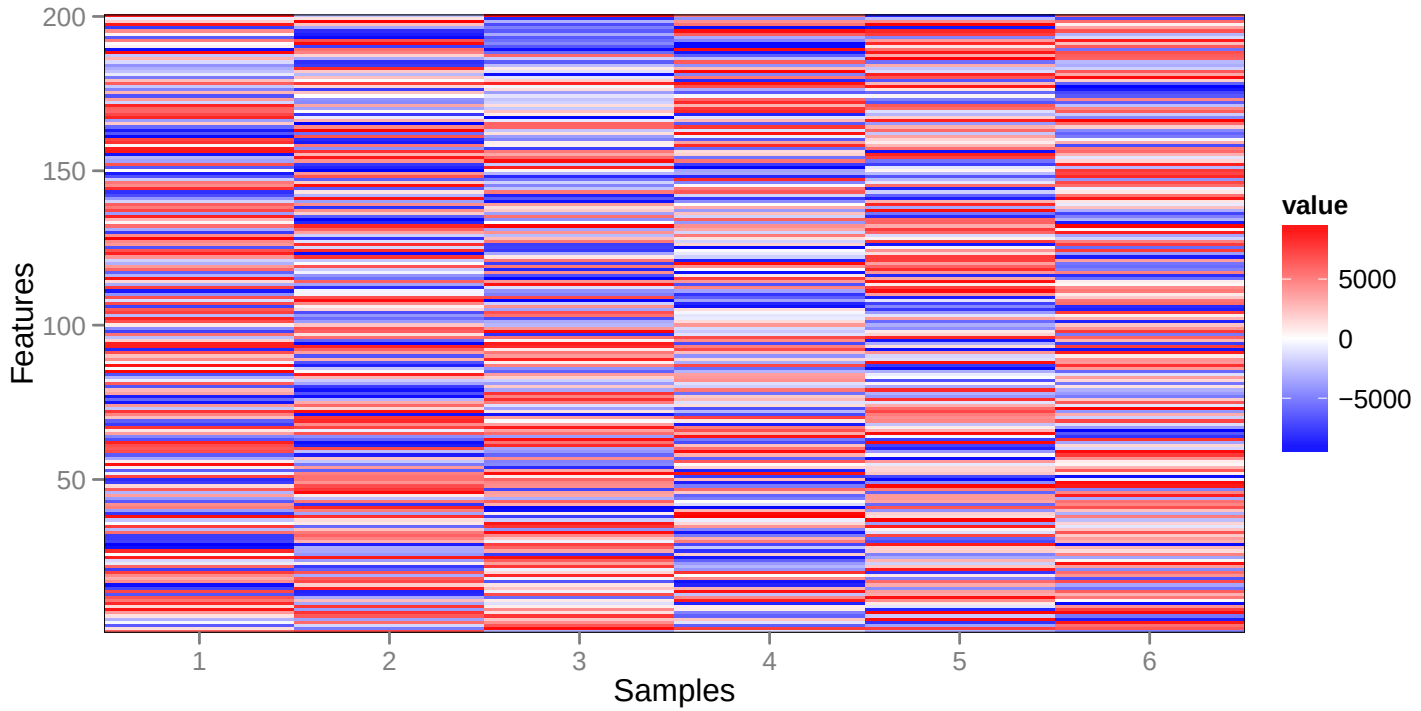
Coord: truncate-gaps

Layout: tracks range-facet



Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet

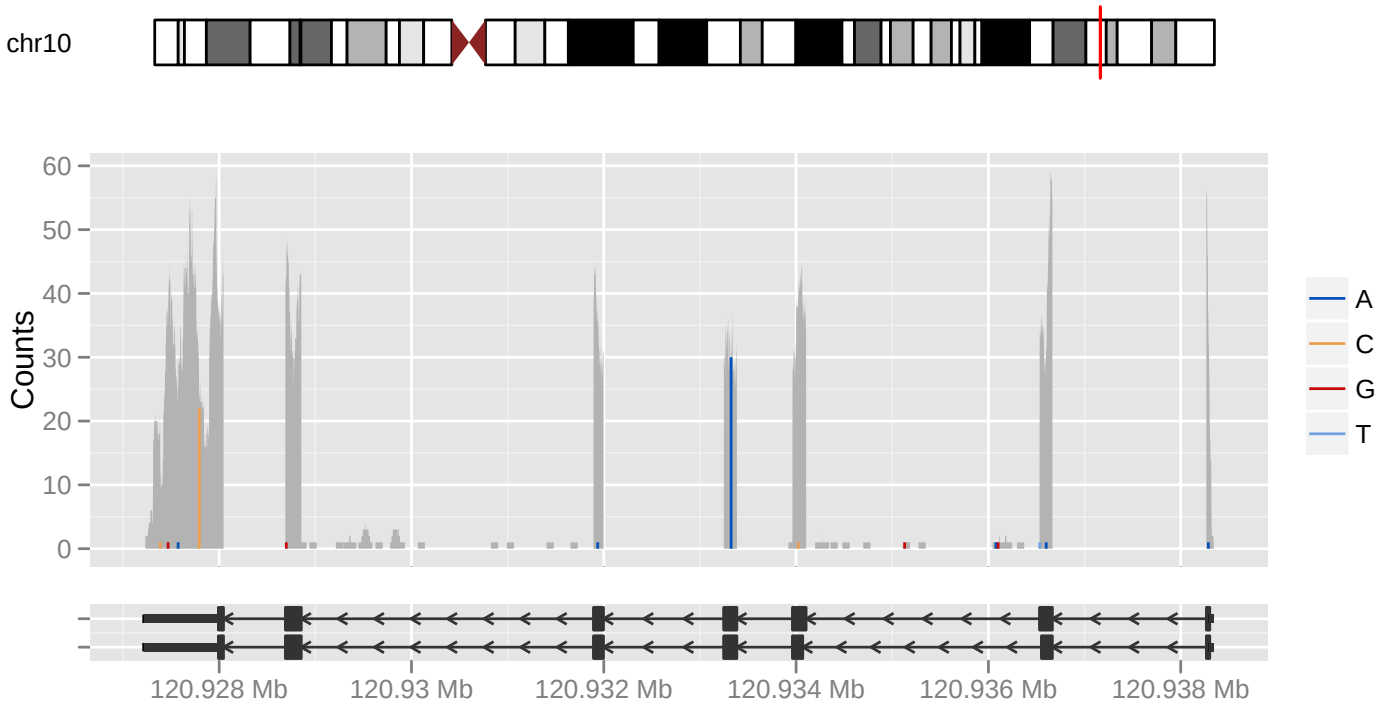


Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet

Components of the Genomic Grammar

Geom: alignment chevron arch arrow arrowrect
Stat: gene reduce stepping coverage mismatch table
Scale: sequence genome fold-change giemsa
Coord: truncate-gaps
Layout: tracks range-facet



Summary

- The *ggbio* package is a toolkit for plotting genomic data and annotations
- Available as part of the Bioconductor project
- Easy to use and flexible enough to handle the diverse use cases encountered in genomics
- Useful plots are automatically generated from Bioconductor data structures using reasonable defaults
- New types of plots can be constructed from grammar primitives specially designed for genomics

Acknowledgements

Tengfei Yin
Di Cook

Robert Gentleman