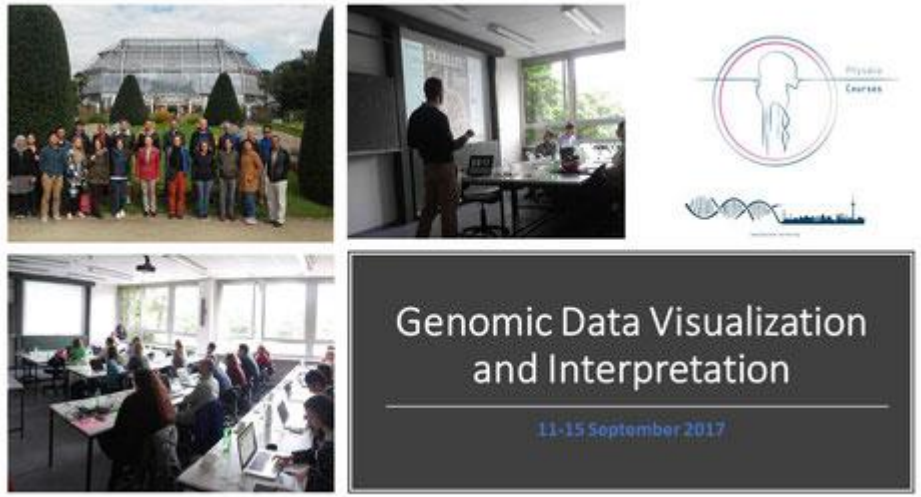
Short overview of the

Genomic Data Visualisation and Interpretation

workshop



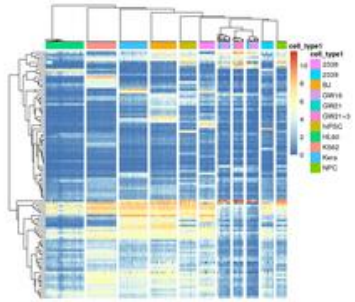
https://www.physalia-courses.org

2018

Bioinformatics with R and Bioconductor

22-26 January 2018

Instructors:
Dr. Levi Waldron
Dr. Ludwig Geistlinger



Analysis of single cell RNA-seq data

5-9 February 2018

Instructors:
Dr. Vladimir Kiselev
Dr. Tallulah Andrews

Assembly and Annotation of genomes

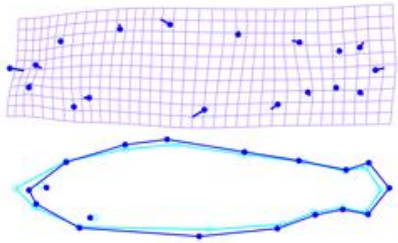
12-16 February 2018

Instructor:
Dr. Thomas D. Otto

Eukaryotic-Metabarcoding

26 February-2 March 2018

Instructors:
Dr. Owen S. Wangensteen
Dr. Vasco Elbrecht



Geometric Morphometrics

5-9 March 2018

Instructor:
Dr. Carmelo Fruciano

Introduction to Linux for biologists

12-16 March 2018

Instructor:
Dr. Martin Jones

BELGIAN NETWORK FOR DNA BARCODING

MITOCHONDRIAL METAGENOMICS: THEORY AND PRACTICE

11-12 DECEMBER 2017

ROYAL MUSEUM FOR CENTRAL AFRICA
LEUVENSESTEENWEG 13, 3080 TERVUREN, BELGIUM

INSTRUCTORS:

PROF. ALFRIED P. VOGLER

(IMPERIAL COLLEGE LONDON, UK)

DR. THOMAS CREEDY

(IMPERIAL COLLEGE LONDON, UK)

DR. MASSIMILIANO BABBUCCI

(UNIVERSITY OF PADOVA, ITALY)

PROGRAMME

MONDAY 11TH

PLENARY LECTURE "MITOCHONDRIAL METAGENOMICS"

TUESDAY 12TH

HANDS-ON SESSIONS:

SESSION 1A:
DE NOVO ASSEMBLY OF A WHOLE MITOCHONDRIAL GENOME FROM ILLUMINA PAIRED SEQUENCING READS

- INTRODUCTION TO ASSEMBLY AND THE LINUX COMMAND LINE
- QUALITY CONTROL, ADAPTER TRIMMING AND READ PAIRING
- MITOGENOME ASSEMBLY USING TWO DIFFERENT TYPES OF ASSEMBLERS

SESSION 1B:
ASSEMBLY INTERROGATION AND ANNOTATION

- VIEWING AND ASSESSING ASSEMBLY QUALITY
- THE EFFECTS OF VARIATION OF ASSEMBLY PARAMETERS
- ANNOTATING ASSEMBLY USING LOCAL AND WEB TOOLS

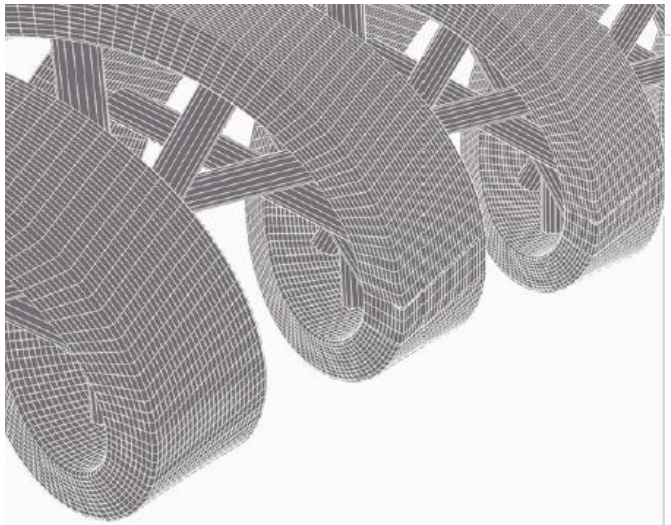
SESSION 2A:
ASSEMBLING WITH SPARSE AND LOW QUALITY DATA

- SELECTION OF REFERENCE SEQUENCES
- ASSEMBLY TO A REFERENCE SEQUENCE
- VIEWING AND ANNOTATING INCOMPLETE MITOGENOMES

SESSION 2B:
MITOCHONDRIAL METAGENOMICS

- ASSEMBLY OF MIXED SEQUENCE LIBRARIES USING MULTIPLE ASSEMBLERS
- META-ASSEMBLY OF MULTIPLE RESULTS
- CONTIG IDENTIFICATION

http://genviz.org/course/



The Elizabeth H.
and James S. McDonnell III
McDonnell Genome Institute
at Washington University

Learning objectives of the course

- **Module 1: Introduction to genomic data visualization and interpretation**
- Module 2: Using R for genomic data visualization and interpretation
- Module 3: Introduction to GenVisR
- Module 4: Expression profiling, visualization, and interpretation
- Module 5: Variant annotation and interpretation
- Module 6: Q & A, discussion, integrated assignments, and working with your own data

- Tutorials
  - Provide working examples of data visualization and interpretation
  - Self contained, self explanatory, portable

Malachi Griffith, Obi Griffith, Zachary Skidmore
Genomic Data Visualization and Interpretation
September 11-15, 2017
Berlin

1. Online tools
2. GenVisR
3. ggplot
4. shiny
5. markdown

http://genviz.org/course/

Why do we create visualizations of genomic data?

- Data exploration and interpretation of results
  - QC analysis
  - Understanding whether/how an experiment worked
  - Discovery

- Communication
  - Slides for presentations
    - e.g. Keynote, Powerpoint, etc.
  - Figures for publications
    - e.g. PDFs, PNGs, etc.

Online tools

http://ensemblgenomes.org/

EnsemblMetazoa ▾ HMMER | BLAST | BioMart | Tools | Downloads | Documentation | Website help

Login/Re

Search Ensembl Metazoa…

Find a Species

Show All entries | Show/hide columns | Filter

	Name	Classification	Taxon ID	Assembly	Accession	Variation database	Regulation database	Whole genome alignments	Other alignments	In peptide compara	In pan-taxonom compar
	Acyrthosiphon pisum	Hemiptera	7029	Acyr_2.0	GCA_000142985.2	-	-	-	✓	✓	-
	Adineta vaga	Rotifera	104782	AMS_PRJEB1171_v1	GCA_000513175.1	-	-	-	✓	✓	-
	Aedes aegypti	Diptera	7159	AaegL3	GCA_000004015.1	✓	✓	✓	✓	✓	-
	Amphimedon queenslandica	Porifera	400682	Aqu1	GCA_000090795.1	-	-	-	✓	✓	✓
	Anopheles darlingi	Diptera	43151	AdarC3	GCA_000211455.3	-	-	✓	✓	✓	-
	Anopheles gambiae	Diptera	7165	AgamP4	GCA_000005575.1	✓	✓	✓	✓	✓	✓
	Anoplophora glabripennis	Coleoptera	217634	Agla_1.0	GCA_000390285.1	-	-	-	✓	✓	-
	Apis mellifera	Hymenoptera	7460	Amel_4.5	GCA_000002195.1	-	-	-	✓	✓	✓

Gene comparisons across species

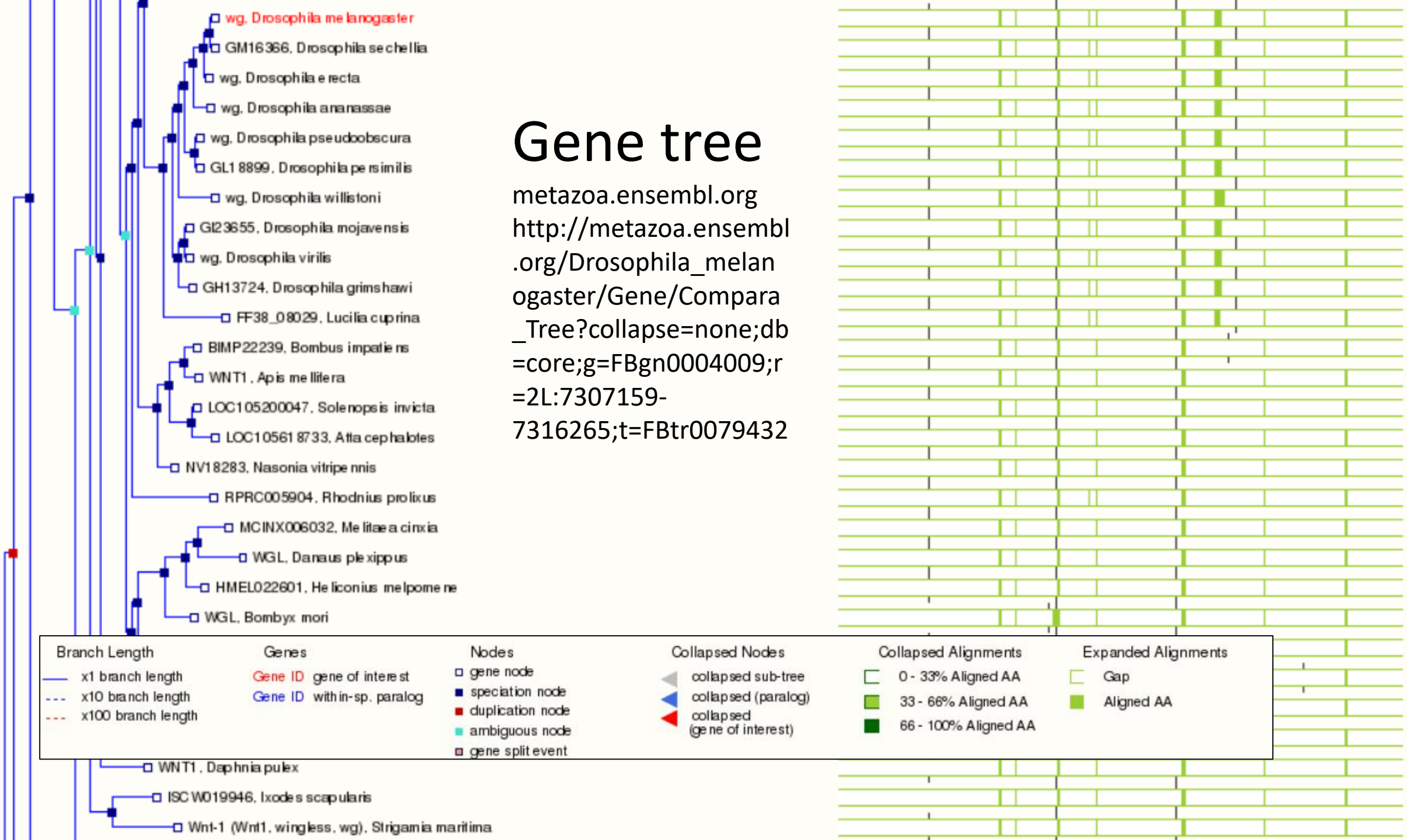
http://metazoa.ensembl.org/Multi/Search/Results?species=all;idx=;q=wingless;site=ensemblunit

wg [FBgn0004009]

Description	***wingless*** [Source:FlyBase;Acc:FBgn0004009]
Gene ID	FBgn0004009
Species	*Drosophila melanogaster*
Location	2L:7307159-7316265
Synonyms	Br Bristled CG4889 Complementation group I DWint 1 DWnt DWnt 1 Dint 1 Dm 1 Dm Wg FBgn0001109 FBgn0003467 FBgn0003469 FBgn0011783 Fg Flag Gla Glazed I Int 1 L(2)02657 L(2)rO727 L(2)wg Sp Spade Spd Sternopleural Wg Wgl ***Wingless*** Wnt Wnt 1 Wnt/Wg Wnt1

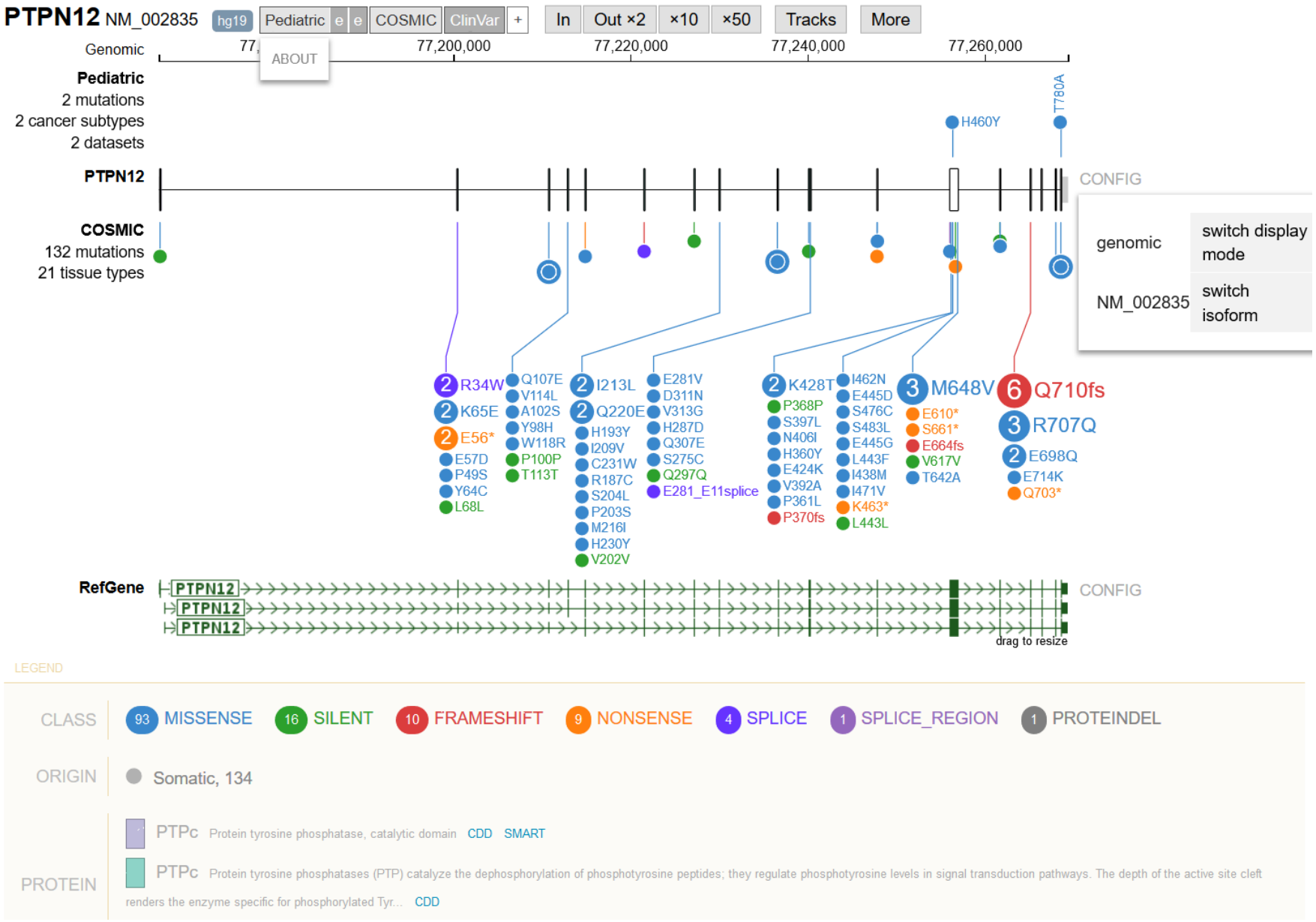
Gene tree

metazoa.ensembl.org
http://metazoa.ensembl.org/Drosophila_melanogaster/Gene/Compara_Tree?collapse=none;db=core;g=FBgn0004009;r=2L:7307159-7316265;t=FBtr0079432



Gene structure

https://pecan.stjude.org/proteinpaint/PTPN12



Other browsers

UCSC
https://genome-euro.ucsc.edu/cgi-bin/hgGateway?hgsid=225854038_Mb7rRsXnMer5S6Lzkrev8u8LmLhA

Ensembl BioMart
http://www.ensembl.org/biomart/martview/3724710e7c1b89444cd99efd738c4c49

GenVisR

http://genviz.org/module%203/0003/01/01/Introduction_to_GenVisR/

Description of the GenVisR package

1. GenVisR is **built upon ggplot2** and thus allows the user to leverage the many existing graphical functions of that package (as well as the information we've learned in previous modules).
2. GenVisR is intended to be flexible, **supporting multiple common genomic file formats**, species, etc.
3. GenVisR attempts to make popular, but very complex genomic visualizations much simpler to produce. It essentially offers **convenience functions** that allow sophisticated plots to be made from common genomic data types with just a handful of lines of code.
4. GenVisR is **relatively popular** (in the top 20% of bioconductor downloads) and therefore benefits from a large community of users, many published examples, and a number of online tutorials.
5. GenVisR is maintained by the griffithlab and is **regularly updated** with improvements, bug fixes, etc.
6. We recommend installing GenVisR from bioconductor in order to ensure the most stable version

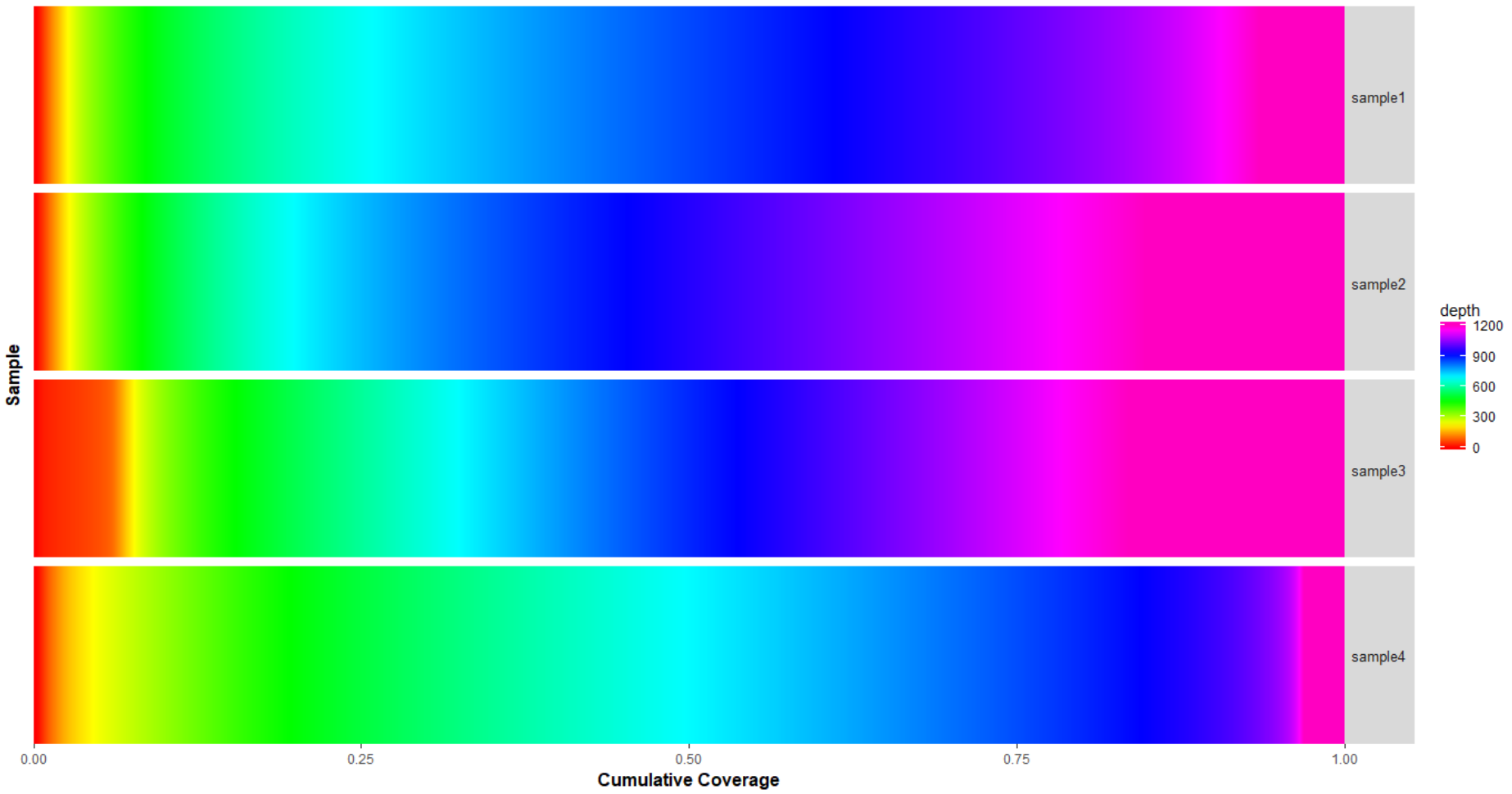
- Part of Bioconductor: https://www.bioconductor.org/packages/3.3/bioc/html/GenVisR.html
- Active development on github: https://github.com/griffithlab/GenVisR

```
#Install and load GenVisR
source("https://bioconductor.org/biocLite.R") #not on CRAN (The Comprehensive R Archive Network)
biocLite("GenVisR")
library(GenVisR)
```

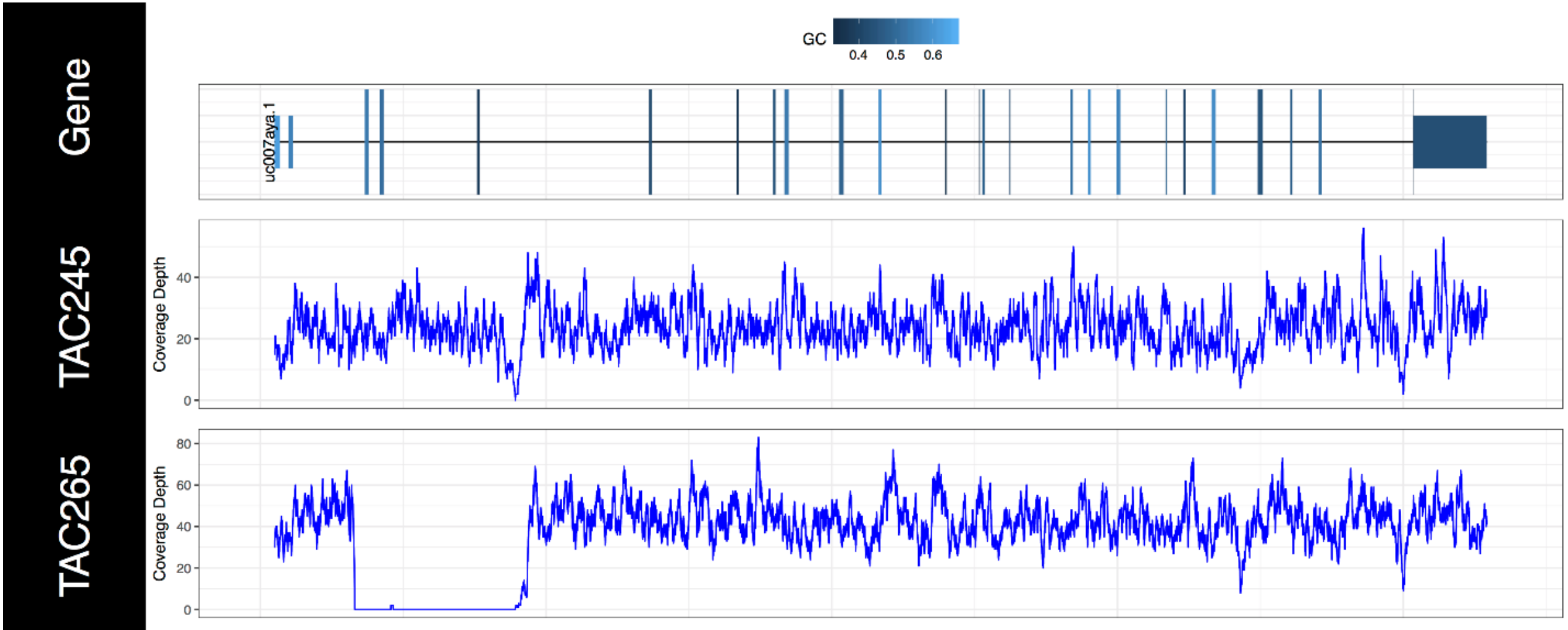
Sequencing coverage plots

Function covBars()
Data from samtools_depth
Manipulated in R using function “count” of package plyr

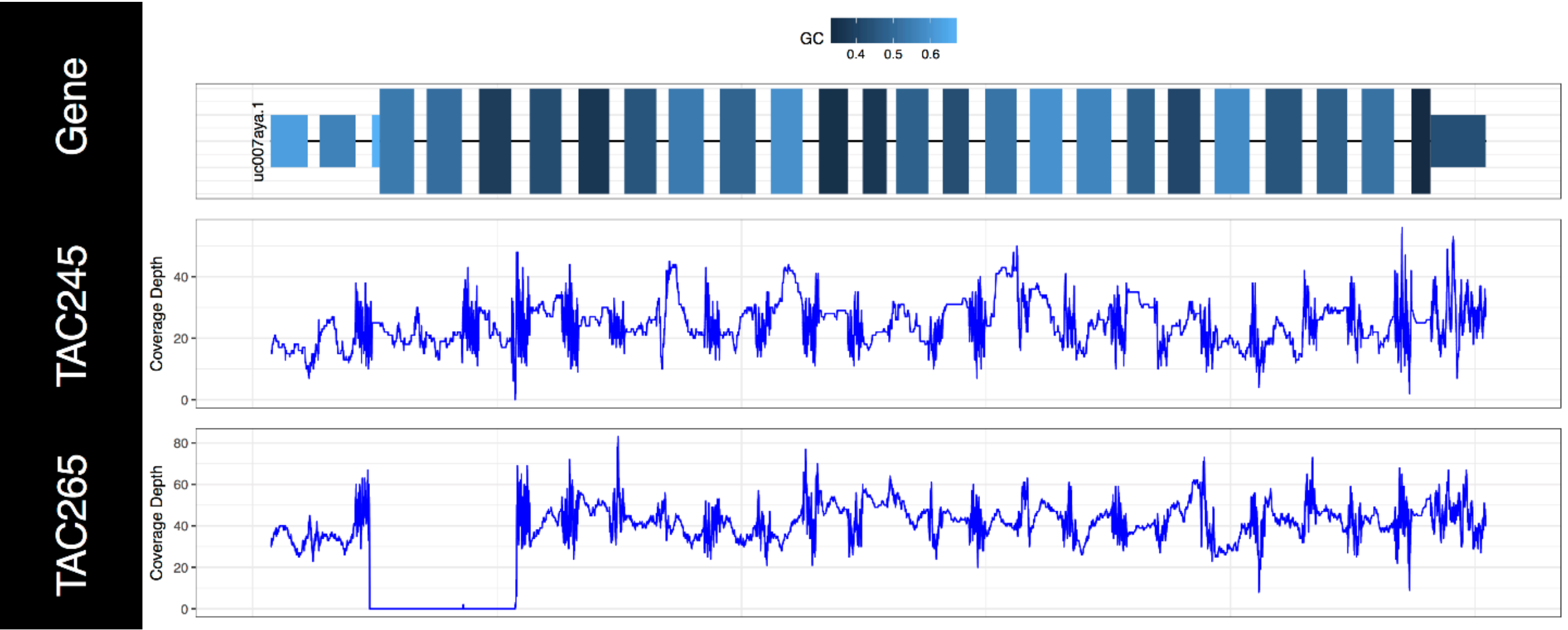
depth	sample1	sample2	sample3	sample4
0	1111	1289	1296	1743
1	1171	908	917	725
2	746	545	785	432
3	431	371	404	424
4	492	277	417	347
5	344	295	480	319
6	362	302	344	351
7	308	229	371	359
8	343	290	330	287
9	264	303	359	364
10	141	229	288	292
11	175	216	362	230
12	157	194	280	313
13	157	208	302	203
14	158	195	316	174
15	155	199	362	156
16	170	170	368	171
17	159	140	409	188
18	161	118	442	188
19	138	150	453	190
20	…			



Gene coverage plots



Gene coverage plots

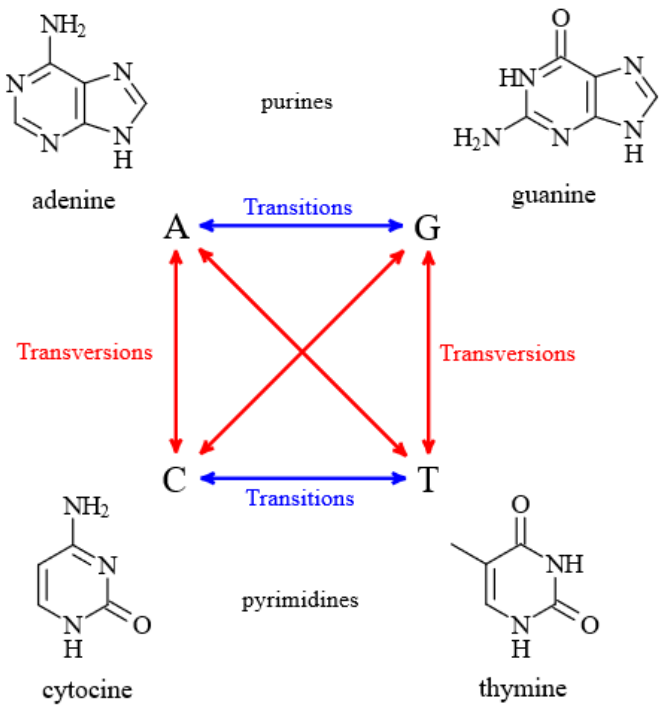


Gene coverage plots

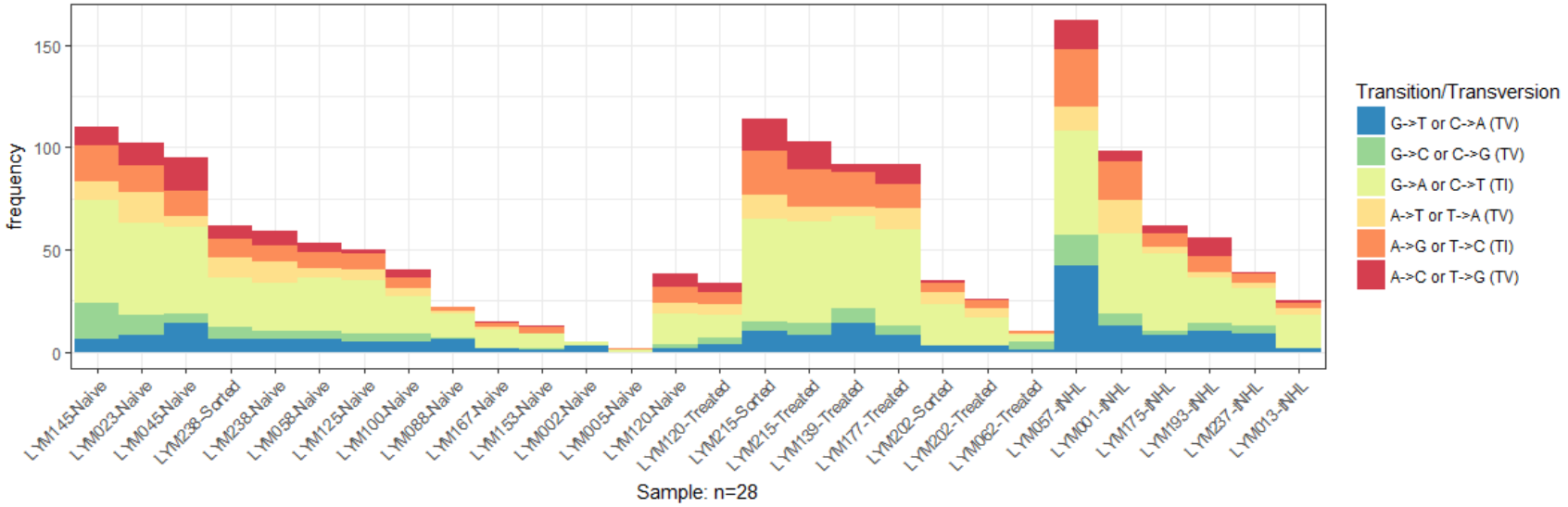
Function Gencov()
More flexible display than IGV
Data from bedtools multicov -bams M_CA-TAC245-TAC245_MEC.prod-refalign.bam -bed stat1.bed
Named list of data frames with list names corresponding to sample id's and column names "chromosome", "end", and "cov"
+ reference assembly (Bsgenome class) and annotations (e.g. TxDb object)

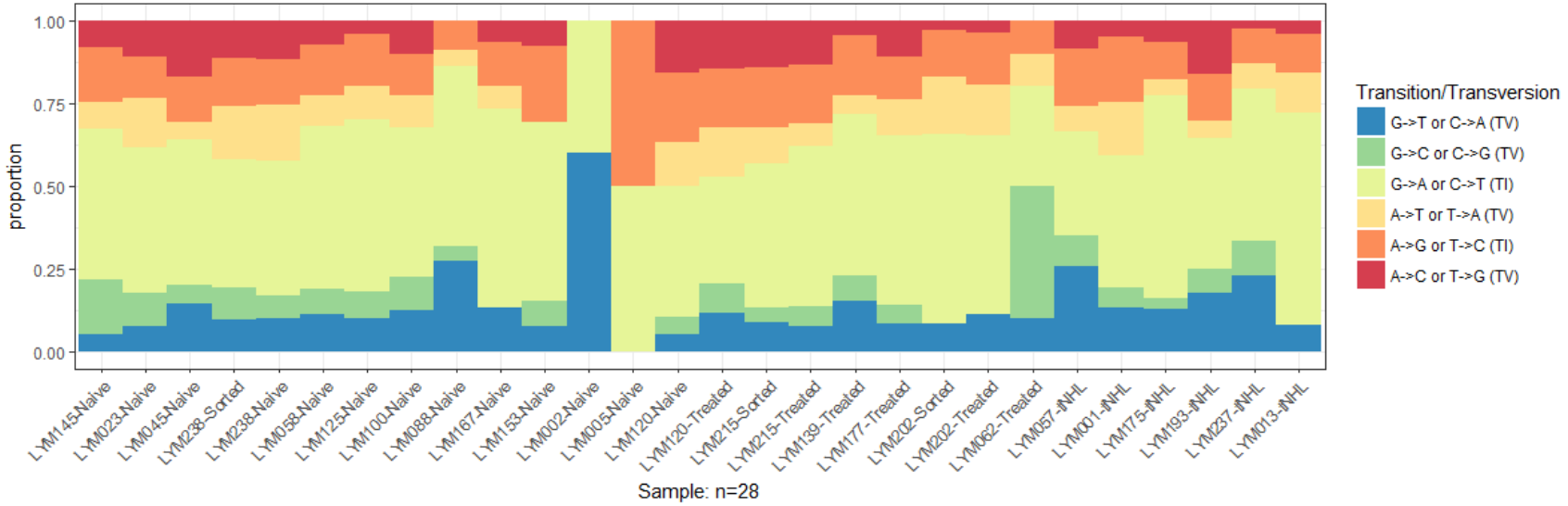
```
> str(covData)
List of 2
 $ TAC245:'data.frame':	42427 obs. of  3 variables:
  ..$ chromosome: int [1:42427] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ end: int [1:42427] 52176282 52176283 52176284 52176285 52176286 52176287 52176288 52176289 52176290 52176291 ...
  ..$ cov       : int [1:42427] 16 16 16 16 15 15 16 18 18 18 ...
 $ TAC265:'data.frame':	42427 obs. of  3 variables:
  ..$ chromosome: int [1:42427] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ end: int [1:42427] 52176282 52176283 52176284 52176285 52176286 52176287 52176288 52176289 52176290 52176291 ...
  ..$ cov: int [1:42427] 30 30 32 33 34 31 33 33 33 33 ...
```

Transition/transversion plots



Transitions and Transversions by Krishnavedala with





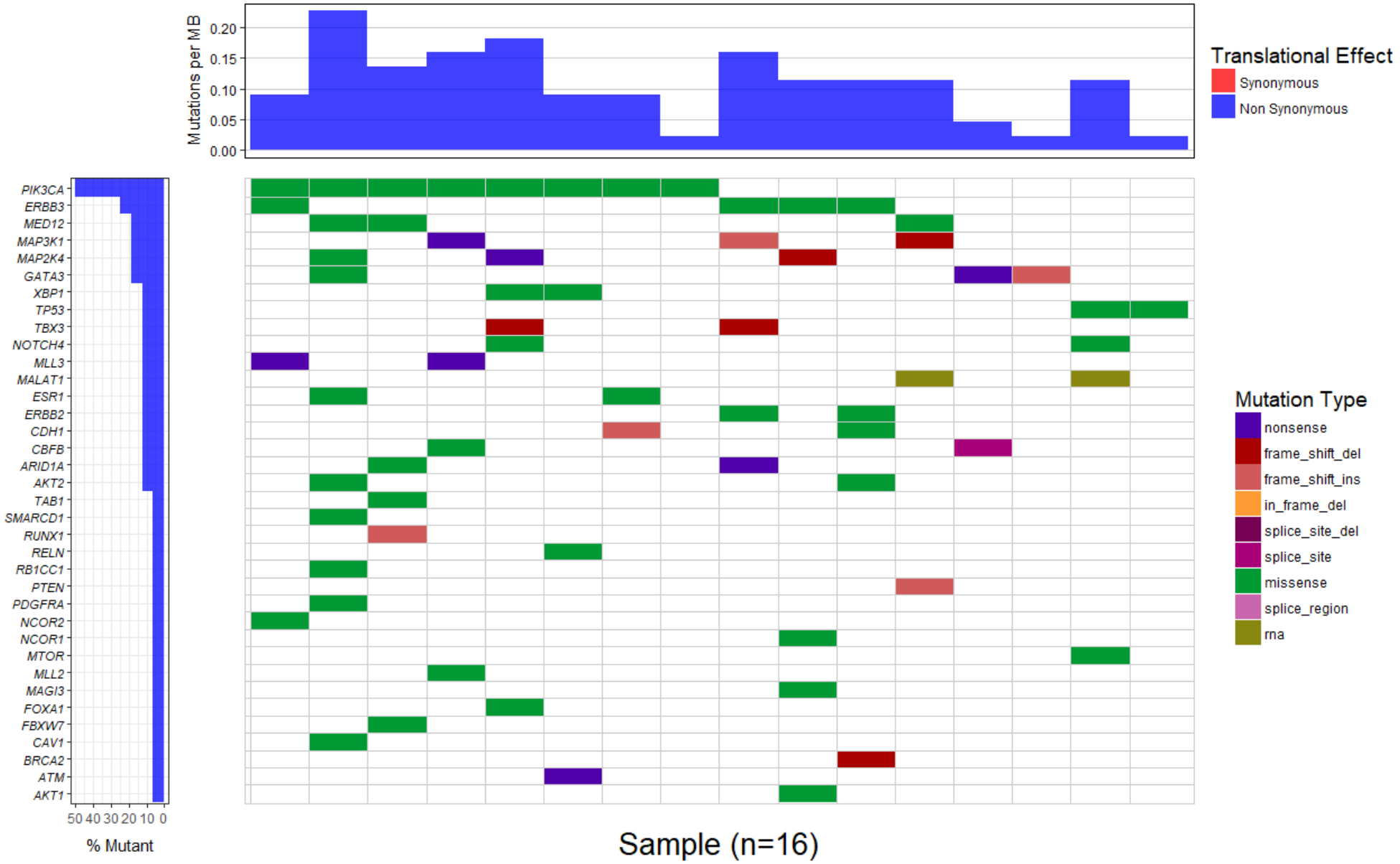
Transition/transversion plots

Function TvTi()
Input: data frame with column names “sample”, “reference”, and “variant” (fileType=“MGI“)

	sample	chrom	start	stop	reference	variant	type	gene_name
1	LYM001-tNHL	18	60985840	60985840	A	T	SNP	BCL2
2	LYM001-tNHL	1	2488123	2488123	G	A	SNP	TNFRSF14
3	LYM001-tNHL	19	16034777	16034777	G	A	SNP	CYP4F11
4	LYM001-tNHL	18	19021442	19021442	T	A	SNP	GREB1L
5	LYM001-tNHL	17	9631454	9631454	G	A	SNP	USP43

…

Waterfall plots



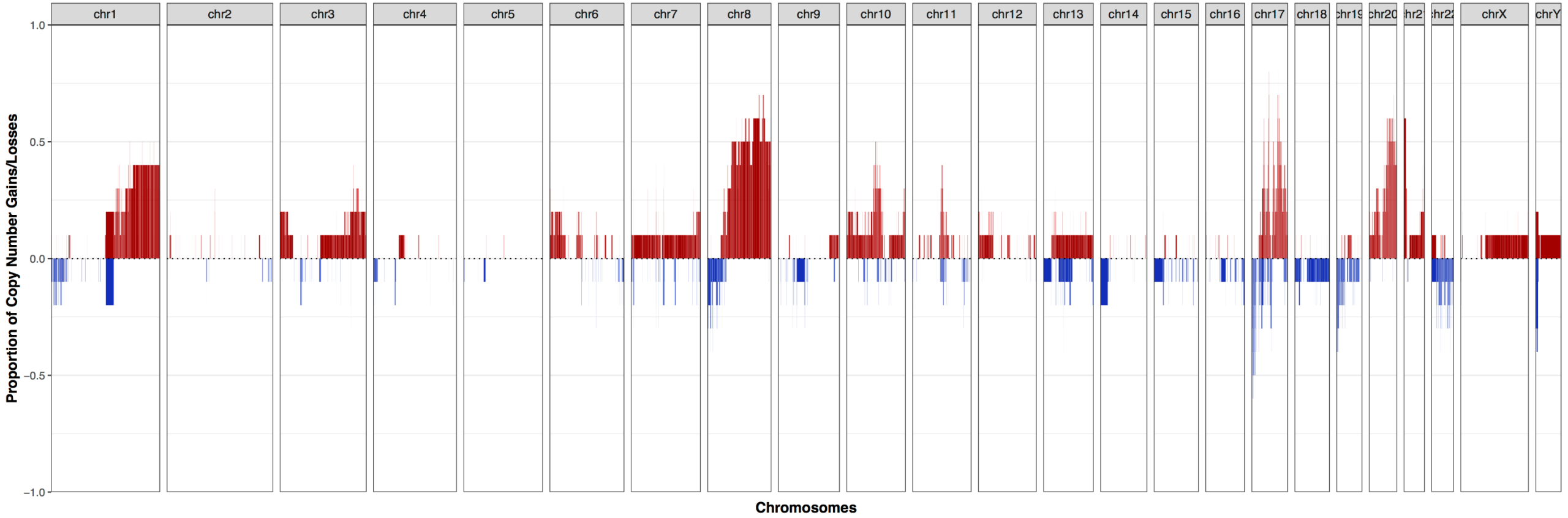
Waterfall plots

Input: data frames in Mutation Annotation Format (MAF)
BUT custom file types are possible (fileType="Custom“) as long as the column names “sample”, “gene”, and variant_class” are present.

sample	gene	variant_class	amino.acid.change
2	PTEN	splice_region	e7-3
2	PTEN	frame_shift_ins	p.P248fs
2	MAP3K1	frame_shift_del	p.S822fs
2	MALAT1	rna	NULL
2	MED12	missense	p.S1892F
…			

Copy number frequency plots

Function cnFreq()
Input = output of Copycat2 (file containing segmented copy number calls)
Cutoff: <1.5 ➔ loss (blue) ; >2.5 ➔ gain (red) of copy number



Other plots

Copy number frequency and spectrum
Loss of heterozygosity
…