Anaquin TransQuin Report

Ted Wong

8 July 2015

This is an automatic generation by Anaquin.

Introduction

An accurate quantification of variability in RNA sequencing analysis is critical for bioinformatics analysis. To control for these sources of variabiltiy, Garvan Institute of Medical Research has developed a set of spiked-in synthetic sequins known as *TransQuin*. The sequins emulate a diverse range of genomic features and they are combined together across a range of concentration to formulate a mixture to emulate quantitative features.

The sequins align to artificial gene loci encoded within the *in silico* chromosome. These artificial gene loci are designed to emulate human gene loci including the range of sizes, sequence elements and splicing complexity. The alignment of sequenced reads to these artificial gene loci thereby enables the assessment of RNA sequencing performance and analysis, including spliced-read alignment, isoform assembly, gene discovery and quantitiative gene expression profiling.

Anaquin is a statistical software toolkit built for TransQuin. It is designed to analyze outputs generated by the NGS tools, and accommodates the analysis with detailed statistics on the sequins. Please refer to www.sequin.xyz for more details.

Data

The datasets included in this study were obtained from two different Illumina sequencing machines, differing in their read length and overall throughput but sharing the same sequencing technology that takes place on a glass slide called a 'flow cell'. A flow cell is made up of eight independent sequencing areas, or 'lanes'. Libraries are deposited on these lanes in order to be sequenced. A library contains cDNAs representative of the RNA molecules that are extracted form a given culture or tissue and are pre-processed in order to be adapted to the sequencing procedire. Simularly to microarrays, the library composition reflects the RNA repertoire expressed in the corresponding culture or tissue. The 'library size' refers to the number of mapped short reads obtained from the sequencing process of the library. In this study, a single library was sequenced in each lane.

Methods

K562 was provided by Gillian Lehrbach (Tissue Culture Facility, The Kinghorn Cancer Centre & Cancer Research Program) and *GM12878* was provided by Madhavi Maddugoda (Epigenetics Research Group, Garvan Institute of Medical Research). K562 and GM12878 cells were cultured according to Coriell Cell Repositories growth protocols and standards. Briefly, K562 and GM12878 were cultured in RPMI 1640 medium (Gibco®) supplemented with 10% fetal bovine serum (FBS) at 37°C under 5% CO2.

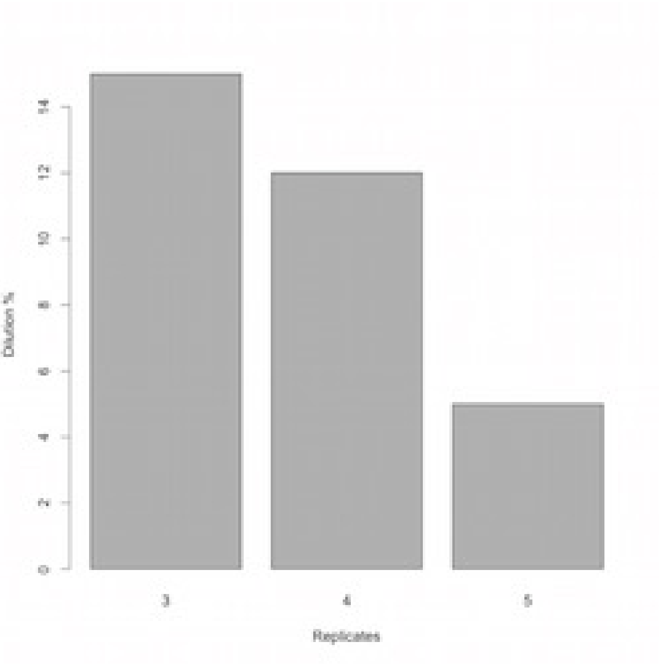
Total RNA was extracted from K562, GM12878 and mouse using TRIzol (Invitrogen) according to the manufacturer's instruction. DNAse treatment was subsequently performed on each sample with TURBO DNase (Life Technologies) followed by a cleanup with the RNA Clean and Concentrator-25 Kit (Zymo Research). Total RNA was run on an Agilent Bioanalyzer 2100 to assess intactness and both the Nanodrop

(Thermo Scientific) and Qubit (Life Technologies) were used to determine the concentration. Only RNA with a RNA integrity number (RIN) > 8.0 was used for library preparation.

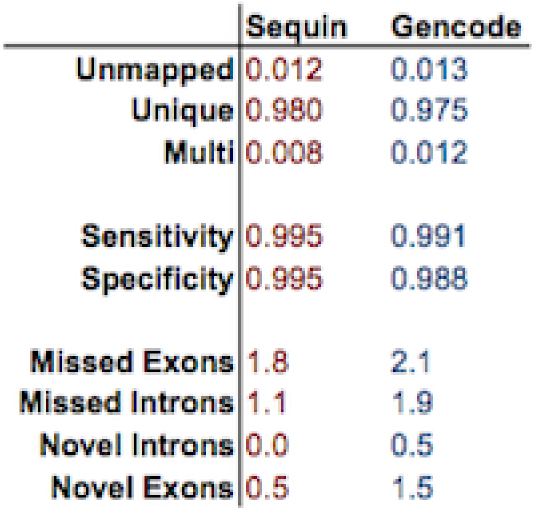
All 164 RNA transcripts' concentration was measured on a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The RNA standards mixtures are manufactured in two formulations, RNA Mixture A and RNA Mixture B, each containing the full complement of 164 transcripts. The 164 RNA transcripts in each mixture were pooled using an epMotion 5070 epBlue™ software program to make the final mixtures robotically spanning a 106 -fold concentration range. The transcripts in Mix A and Mix B are present at defined Mix A:Mix B molar concentration ratios, described by 9 subgroups at the gene level and 19 subgroups at the isoform level.

Reads Alignment

There are 500,456,345 total reads for the six replicates, in which 98% of the reads are mapped. **Figure 1** shows the dilution for each replicate, the average dilution is 2.25%.

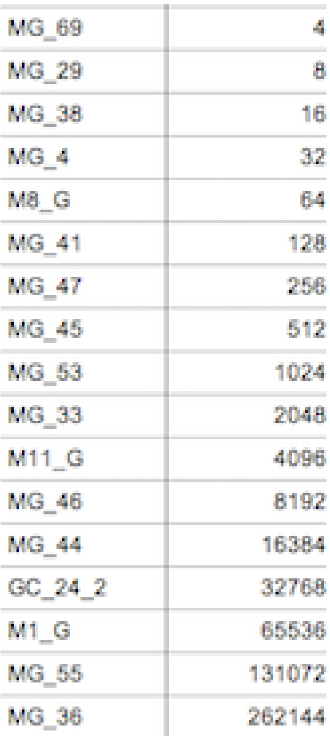


The performance for the alignments at the base, exon and intron level is tablulated in **Figure 2**. An alignment is considered as a true-positive at the exon level if is is aligned within an exon, intron is similar. In the base level, an aligned nucleotide is a true-positive if it is part of an exon, otherwise it is a false-positive.



	Sequin	Gencode
Unmapped	0.012	0.013
Unique	0.980	0.975
Multi	0.008	0.012
Sensitivity	0.995	0.991
Specificity	0.995	0.988
Missed Exons	1.8	2.1
Missed Introns	1.1	1.9
Novel Introns	0.0	0.5
Novel Exons	0.5	1.5

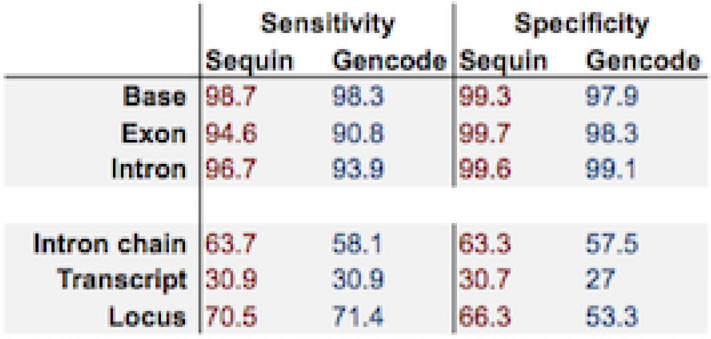
Figure 3 shows the undetected sequins.



MG_69	4
MG_29	8
MG_38	16
MG_4	32
M8_G	64
MG_41	128
MG_47	256
MG_45	512
MG_53	1024
MG_33	2048
M11_G	4096
MG_46	8192
MG_44	16384
GC_24_2	32768
M1_G	65536
MG_55	131072
MG_36	262144

Assembly

The performance for the assembly is compared to the reference TransQuin annotation. This is done at both the isoform and gene level.

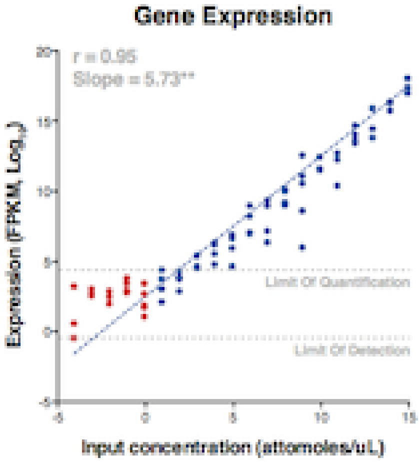


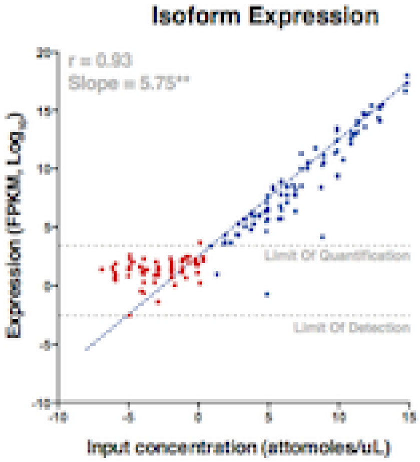
	Sensitivity		Specificity	
	Sequin	Gencode	Sequin	Gencode
Base	98.7	98.3	99.3	97.9
Exon	94.6	90.8	99.7	98.3
Intron	96.7	93.9	99.6	99.1
Intron chain	63.7	58.1	63.3	57.5
Transcript	30.9	30.9	30.7	27
Locus	70.5	71.4	66.3	53.3

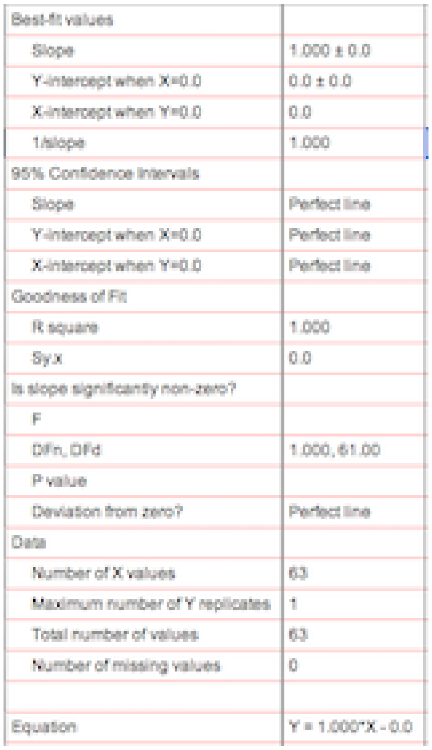
Expression

Gene expression levels and alternative splicing are dynamically regulated during development and between different tissue- or-cell types. By adding alternative RNA sequin mixtures to multiple samples enables a closer comparison between samples, and assessment of differential gene expression and splicing. We repeat the analysis for the synthetic isoforms and genes, as shown in **Figure 3, 4**.

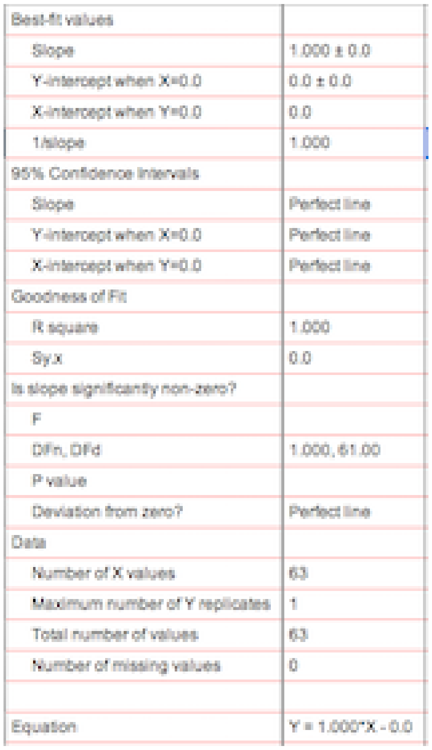
Out of `160` sequins, `150` of them are detected. The undetected sequins are not included in the analysis.





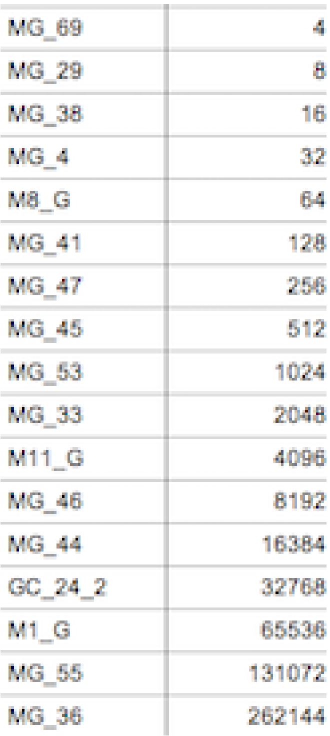


Best-fit values	
Slope	1.000 ± 0.0
Y-intercept when X=0.0	0.0 ± 0.0
X-intercept when Y=0.0	0.0
1/slope	1.000
95% Confidence Intervals	
Slope	Perfect line
Y-intercept when X=0.0	Perfect line
X-intercept when Y=0.0	Perfect line
Goodness of Fit	
R square	1.000
Sy.x	0.0
Is slope significantly non-zero?	
F	
DFn, DFd	1.000, 61.00
P value	
Deviation from zero?	Perfect line
Data	
Number of X values	63
Maximum number of Y replicates	1
Total number of values	63
Number of missing values	0
Equation	Y = 1.000*X - 0.0



Best-fit values	
Slope	1.000 ± 0.0
Y-intercept when X=0.0	0.0 ± 0.0
X-intercept when Y=0.0	0.0
1/slope	1.000
95% Confidence Intervals	
Slope	Perfect line
Y-intercept when X=0.0	Perfect line
X-intercept when Y=0.0	Perfect line
Goodness of Fit	
R square	1.000
Sy.x	0.0
Is slope significantly non-zero?	
F	
DFn, DFd	1.000, 61.00
P value	
Deviation from zero?	Perfect line
Data	
Number of X values	63
Maximum number of Y replicates	1
Total number of values	63
Number of missing values	0
Equation	Y = 1.000*X - 0.0

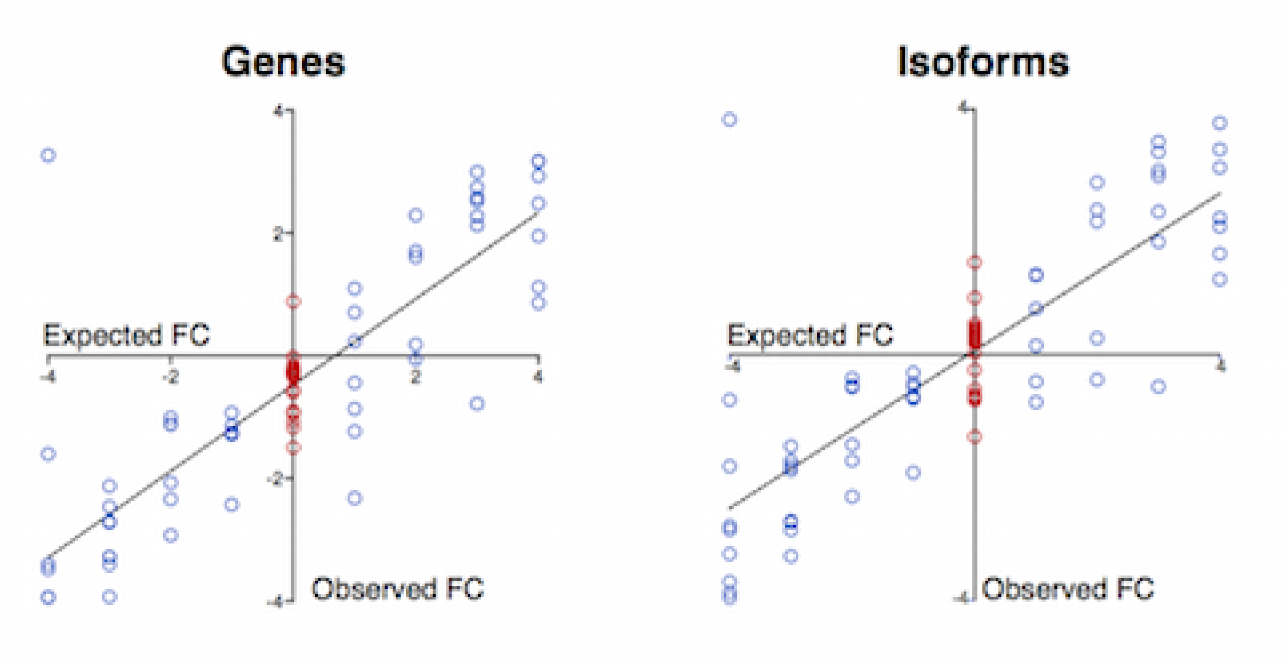
Figure 5 shows the undetected sequins.

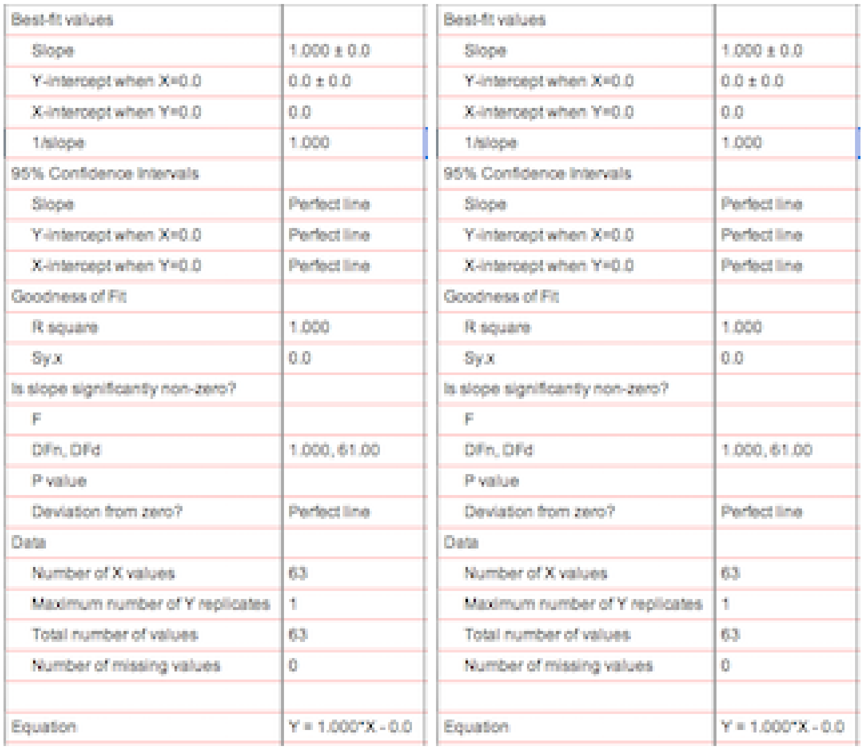


MG_69	4
MG_29	8
MG_38	16
MG_4	32
M8_G	64
MG_41	128
MG_47	256
MG_45	512
MG_53	1024
MG_33	2048
M11_G	4096
MG_46	8192
MG_44	16384
GC_24_2	32768
M1_G	65536
MG_55	131072
MG_36	262144

Differential Analysis

The observed TransQuin known-measured relationship is intended for quantitative assessment in differential analysis. A linear regression and it's 95% confidence interval is also plotted.

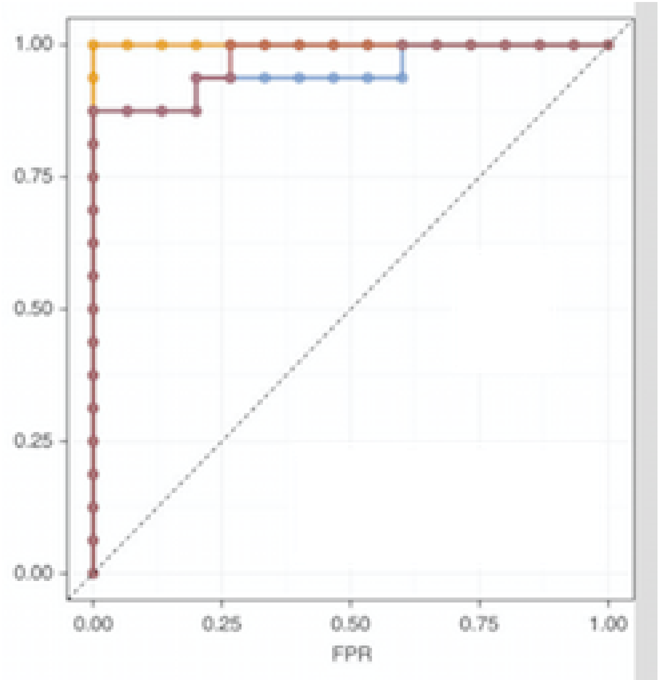




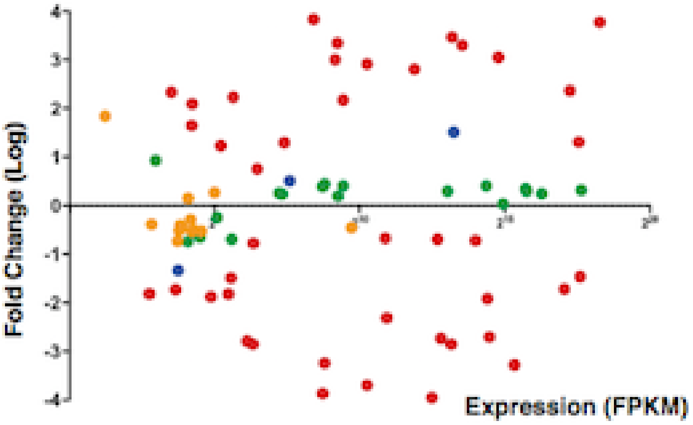
Best-fit values		Best-fit values	
Slope	1.000 ± 0.0	Slope	1.000 ± 0.0
Y-intercept when X=0.0	0.0 ± 0.0	Y-intercept when X=0.0	0.0 ± 0.0
X-intercept when Y=0.0	0.0	X-intercept when Y=0.0	0.0
1/slope	1.000	1/slope	1.000
95% Confidence Intervals		95% Confidence Intervals	
Slope	Perfect line	Slope	Perfect line
Y-intercept when X=0.0	Perfect line	Y-intercept when X=0.0	Perfect line
X-intercept when Y=0.0	Perfect line	X-intercept when Y=0.0	Perfect line
Goodness of Fit		Goodness of Fit	
R square	1.000	R square	1.000
Sy.x	0.0	Sy.x	0.0
Is slope significantly non-zero?		Is slope significantly non-zero?	
F		F	
DFn, DFd	1.000, 61.00	DFn, DFd	1.000, 61.00
P value		P value	
Deviation from zero?	Perfect line	Deviation from zero?	Perfect line
Data		Data	
Number of X values	63	Number of X values	63
Maximum number of Y replicates	1	Maximum number of Y replicates	1
Total number of values	63	Total number of values	63
Number of missing values	0	Number of missing values	0
Equation	Y = 1.000*X - 0.0	Equation	Y = 1.000*X - 0.0

For RNA-Seq experiments, differential expression testing of sequins and endogenous genes was performed with DESeq2, to generate P-values for all endogenous and ERCC features. To construct the ROC curves, the 1:1 subpool P-values were the true-negative group for each differential ratio ROC curve.

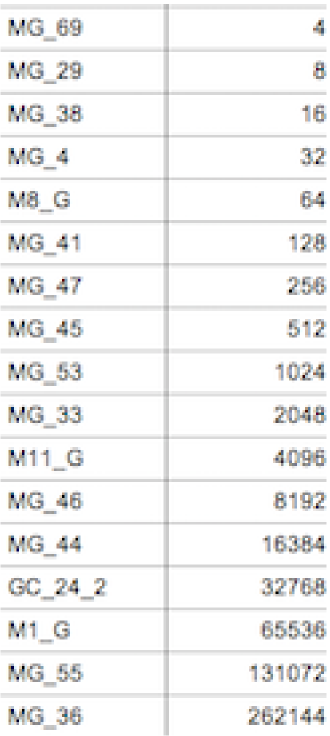
ROC Plot



MA Plot



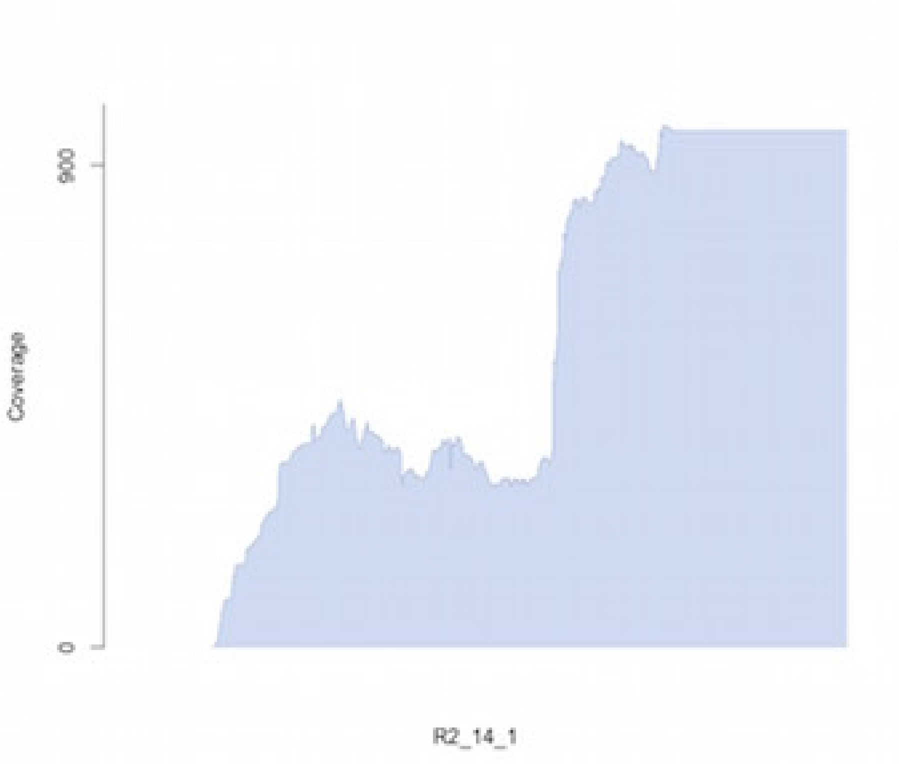
Undetected Sequins



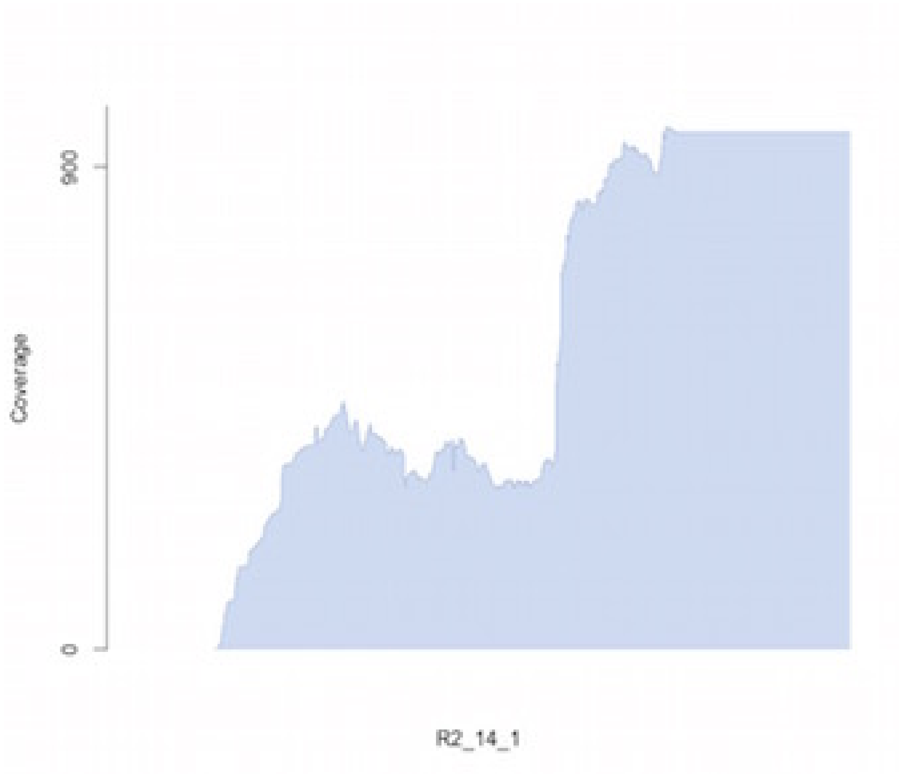
MG_69	4
MG_29	8
MG_38	16
MG_4	32
M8_G	64
MG_41	128
MG_47	256
MG_45	512
MG_53	1024
MG_33	2048
M11_G	4096
MG_46	8192
MG_44	16384
GC_24_2	32768
M1_G	65536
MG_55	131072
MG_36	262144

Apprendix

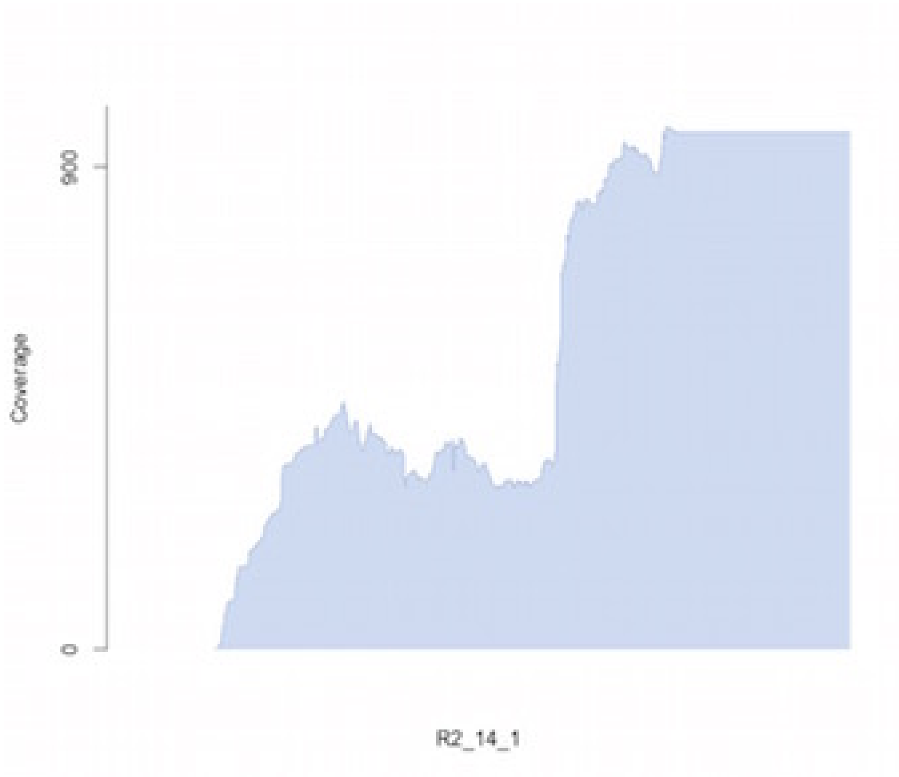
Density plot for



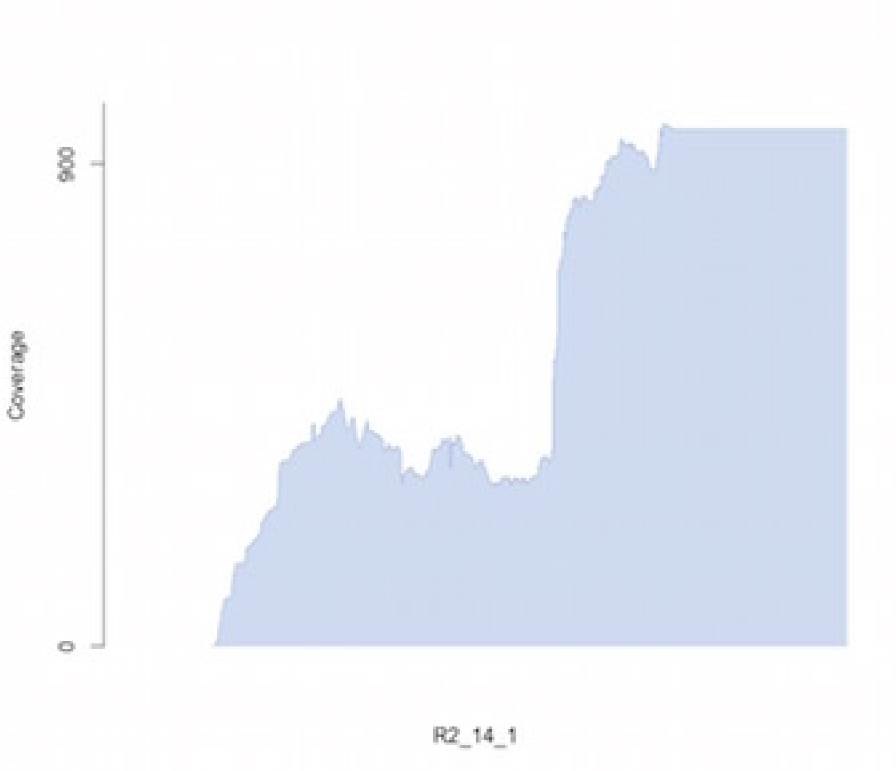
Density plot



Density plot



Density plot


900
0
Coverage
R2_14_1