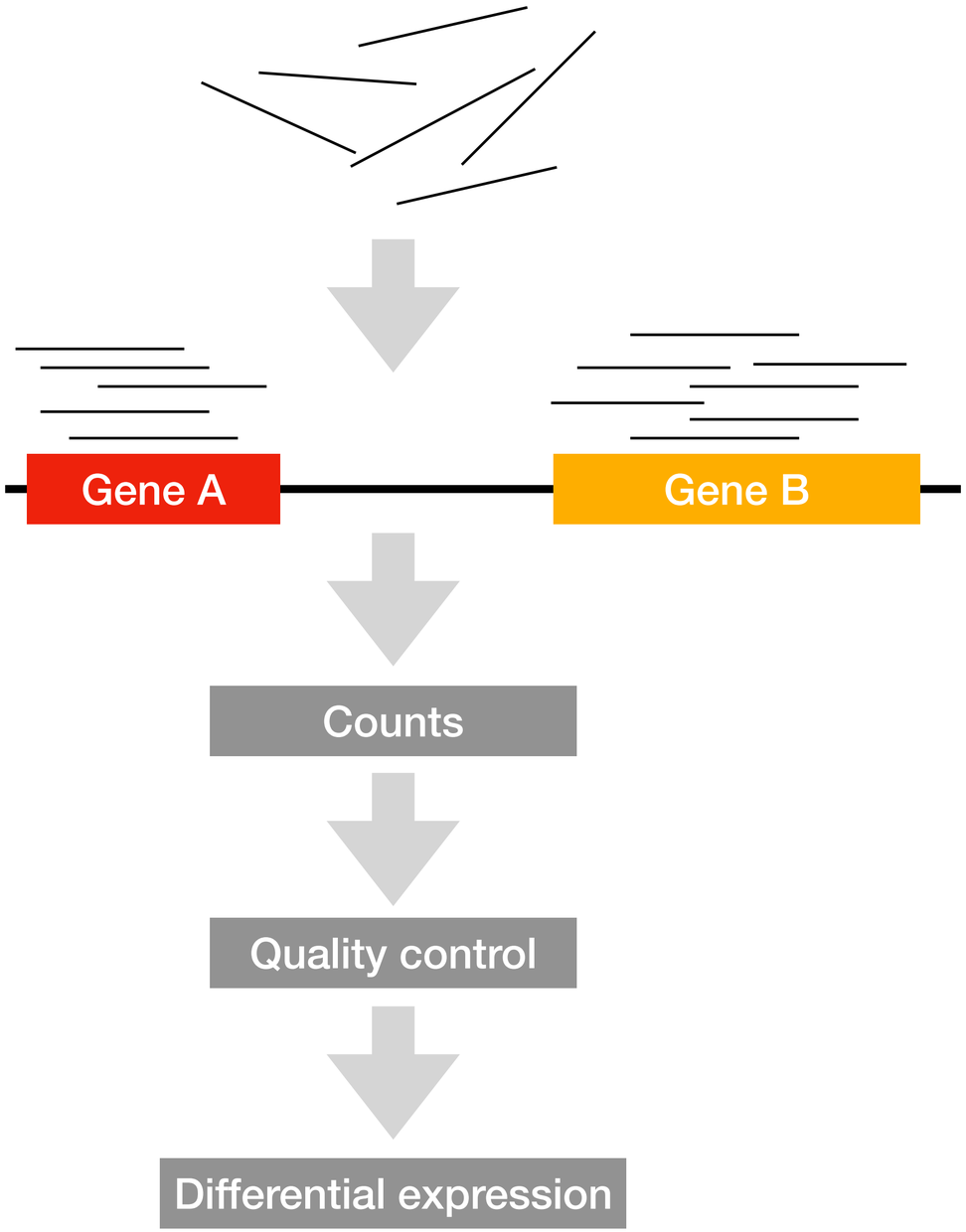
Intro to R for Biologists

IBiS Special Topics, Fall 2021
Class 16: Nov. 15, 2021

Erik Andersen and Shelby Blythe

RNA-seq analysis



Sample quality control (FastQC)

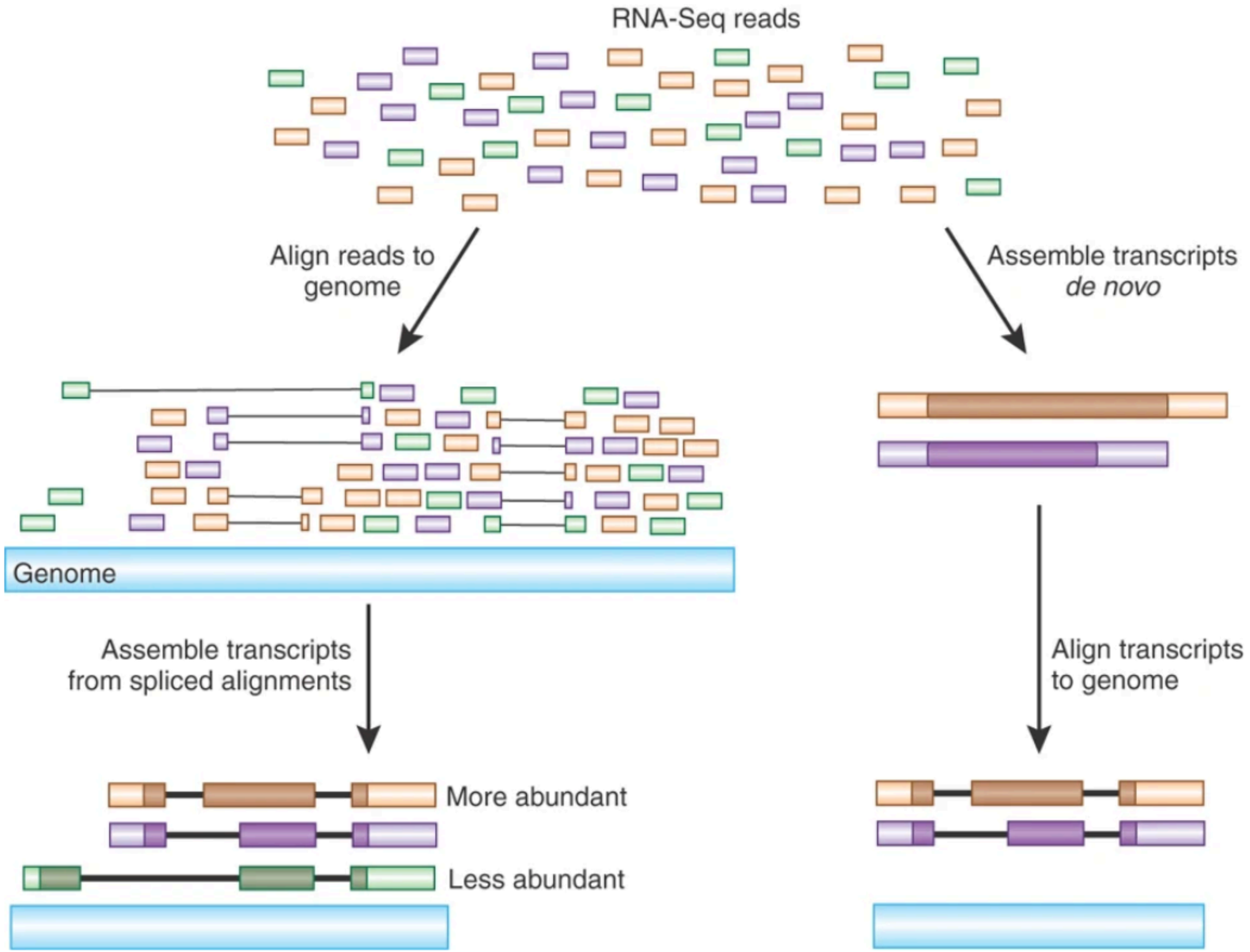
Alignment/mapping (bowtie2, HISAT2, STAR)

Quantification (DESeq2, EdgeR, sleuth)

Data quality control (PCA)

DGE (DESeq2, EdgeR, sleuth)

Two general flavors of alignment methods



Haas and Zody *Nature Biotech.*

Read aligners that use reference genomes and gene models

bowtie2

Ultrafast, memory-efficient tool for reads up to thousands of bases, indexes the reference genome

Langmead lab at Johns Hopkins University

HISAT2

Fast and sensitive graph-based aligner, creates graphs of the reference genome

Langmead and Salzberg labs at Johns Hopkins University

STAR

Ultrafast and sensitive aligner that uses models of spliced transcripts in the reference transcriptome

Dobin lab at Cornell University

Read aligners that don’t use reference genomes

kallisto

Ultrafast, memory-efficient tool that quantifies transcript abundances using pseudo alignment of reads to a graph transcriptome

Pachter lab at CalTech

salmon

Fast tool to build an indexed transcriptome and quantify reads

Patro lab at Stony Brook University

Which analysis workflow is “better”?

Ease of installation, maintained (check GitHub), used by others

Most perform well

Simulate and decide for yourself

Use known data and decide for yourself (always reverse lists)

Ease of “gut checks” (sleuth has beautiful shiny app)

sleuth | overview | analyses | maps | summaries | diagnostics | settings

transcript view

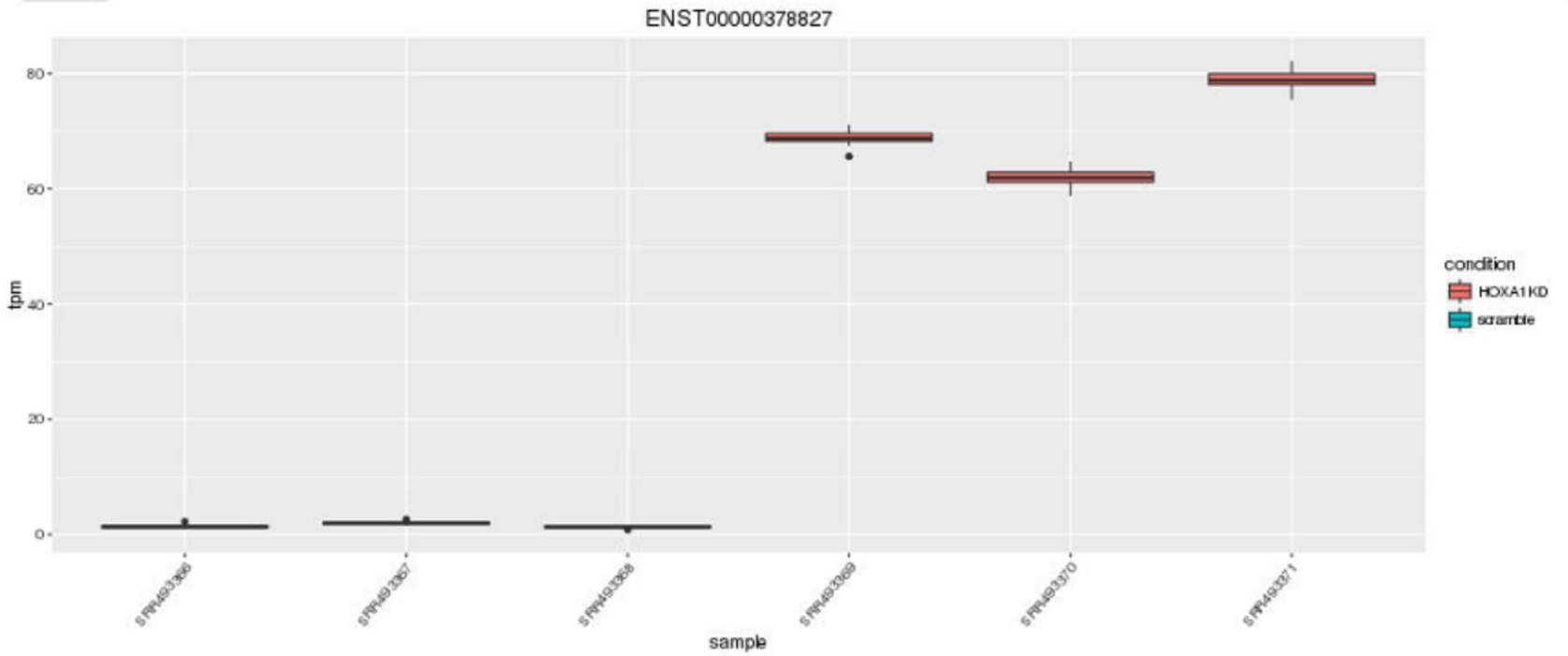
Boxplots of transcript abundances showing technical variation in each sample.

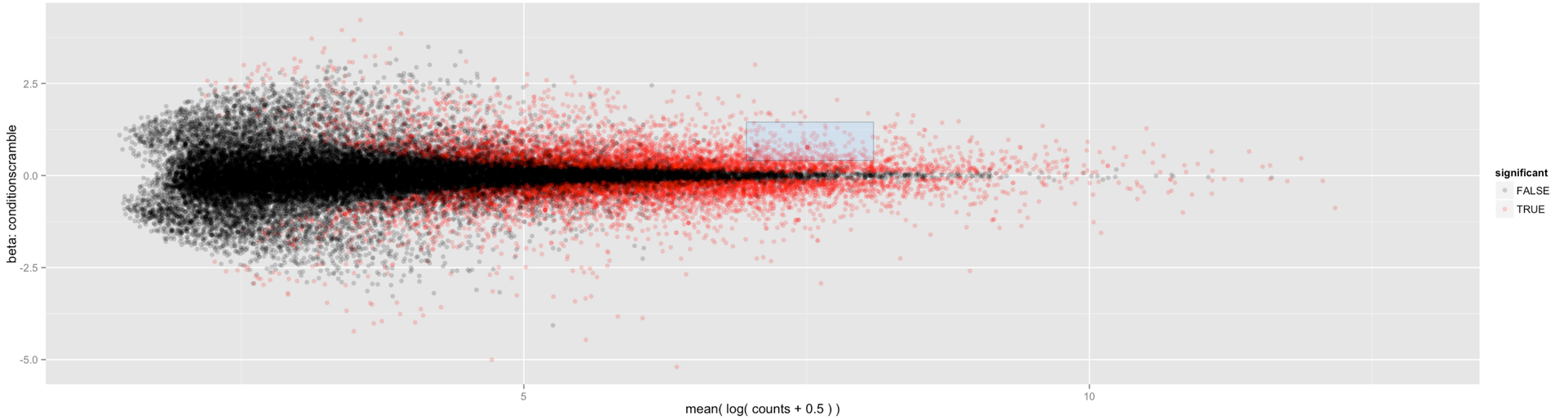
transcript: ENST00000378827

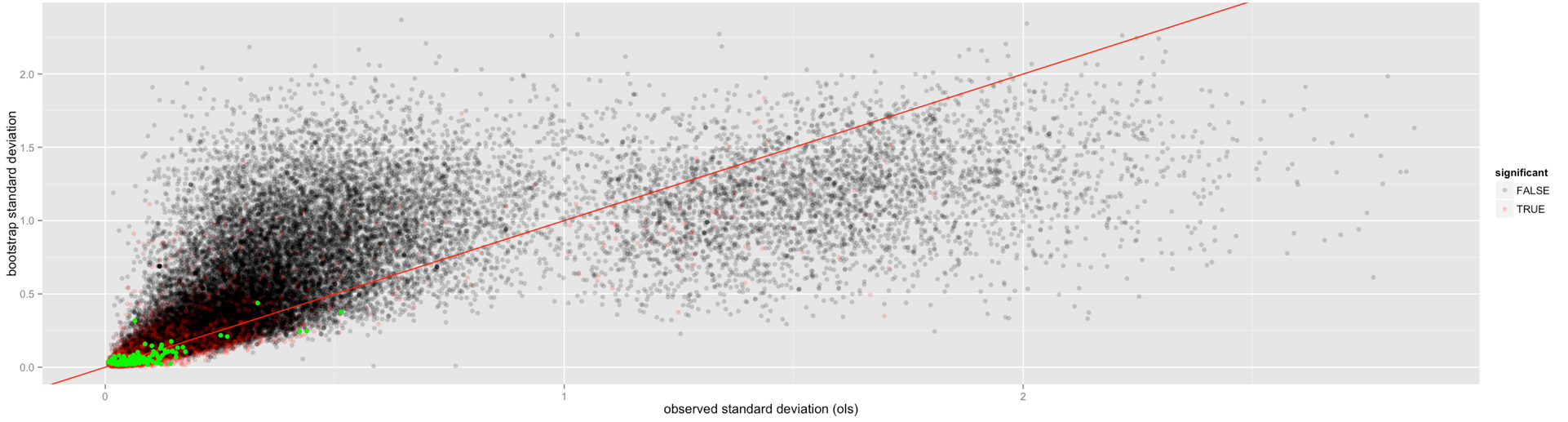
color by: condition

units: tpm

view







Show 25 entries

Search:

target_id	mean_obs	var_obs	sigma_sq	sigma_q_sq	smooth_sigma_sq	smooth_sigma_sq_pmax	b	se_b	pval	qval
ENST00000162749	7.464250	0.08356886	3.267697e-04	0.0009204916	0.001371049	0.001371049	0.5246302	0.03908573	4.460816e-41	2.630413e-39
ENST00000166139	7.061289	0.11405143	1.594565e-03	0.0009706678	0.001761722	0.001761722	0.6110080	0.04268013	1.738825e-46	1.213661e-44
ENST00000171887	7.873805	0.25226174	1.214537e-02	0.0008291675	0.001105609	0.012145367	0.8979275	0.09300371	4.693066e-22	1.359183e-20
ENST00000206423	7.907508	0.30737918	1.959143e-02	0.0005657831	0.001088449	0.019591431	0.9853145	0.11592300	1.900284e-17	4.293729e-16
ENST00000217426	7.144683	0.16147807	1.619497e-03	0.0009332947	0.001668757	0.001668757	0.7290081	0.04164975	1.351986e-68	1.684207e-66

sleuth | overview | analyses | maps | summaries | diagnostics | settings

volcano plot

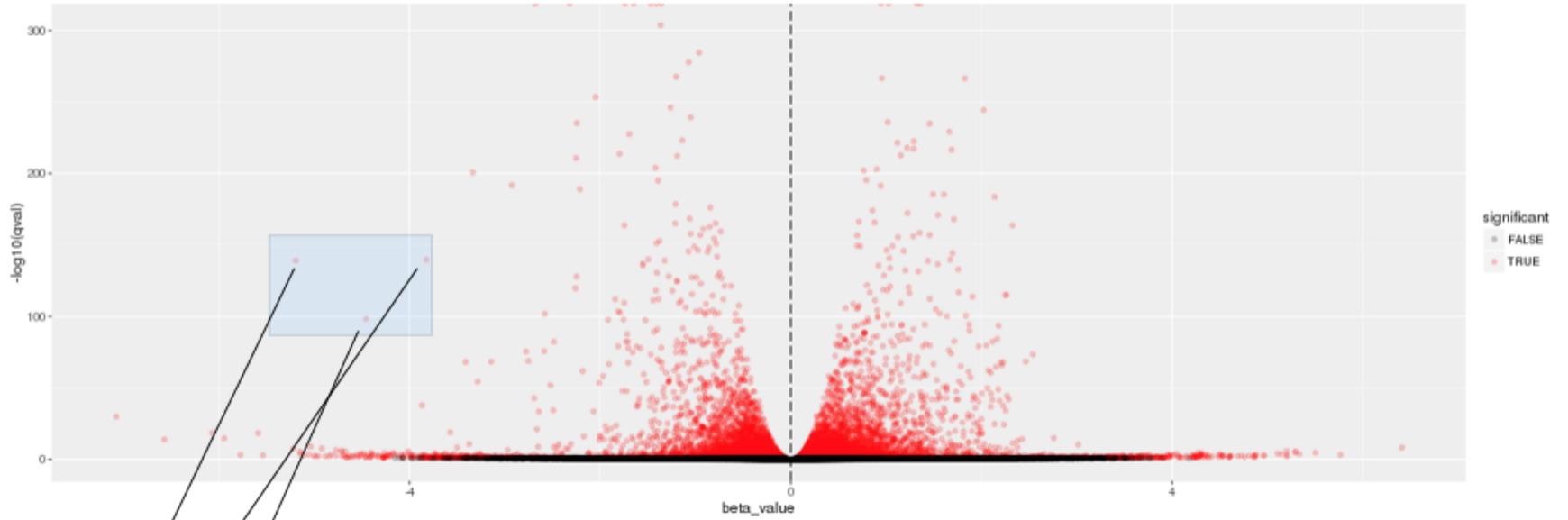
Plot of beta value (regression) versus log of significance. Select a set of transcripts to explore their variance across samples.

max Fdr: 0.1

fit: full

beta: conditionscramble

opacity: 0.2



Download Plot

Show 25 entries

Search:

target_id	pval	qval	b	se_b	mean	var	tech_var	sigma_sq	smooth_sigma_sq	final_sigma_sq
ENST00000378827	4.054993e-143	2.264781e-140	-3.820229	0.1499791	5.849599	4.405238	0.01195494	0.02178567	0.003088144	0.02178567
ENST00000395748	1.143183e-142	6.315463e-140	-5.190873	0.2041153	6.371001	8.133545	0.02090126	0.04159333	0.001856930	0.04159333
ENST00000367468	1.825833e-101	5.395229e-99	-4.456199	0.2083755	5.567478	6.009418	0.03389078	0.03123971	0.004358388	0.03123971
target_id	pval	qval	b	se_b	mean	var	tech_var	sigma_sq	smooth_sigma_sq	final_sigma_sq