

Andrey Kartashov
Dr. Barski lab.

What do we need

- store experiment description, all the data and analysis steps in a formalized way
- automated data downloading
- automated quality control and preliminary analysis
- ability to use published data with our pipelines
- share our own data with co-workers
- sophisticated analysis
- high quality plots and UCSC genome browser visualization
- easy way to add novel types of analysis
- professor friendly interface
- all of the above have to be integrated

Off the shelf products are not comprehensive enough

- LabGuru, iLabber, PerkinElmer (electronic lab notebooks) - do not integrated with analyses
- Illumina basespace, Galaxy (online bioinformatics analysis) - you have to upload data, have basic knowledge about tools, problems with personal genomic data sharing, difficult to restore analysis steps
- GeneSpring, GoldenHelix commercial desktop/server tools - do not have NGS epigenetic analysis
- Ensemble and UCSC Genome Browsers - there are no integrated analyses, not enough description

EMS life cycle

Preliminary analysis fully automated

Advanced analysis project dependent

comprehensive
description

RNA-Seq/ChIP-Seq data

from core

external source

corresponding
analysis (pipeline)

genome browser

preliminary results

quality control

project
designer

Types of analysis

DESeq / DESeq2

MANorm/Diffbind

Average Tag
Density

...

sophisticated results

Consider an example

- Data
 - CD4 Naive Cells 0m RNA-Seq
 - CD4 Naive Cells 150m RNA-Seq
 - CD4 Naive Cells H3K4Me3 ChIP-Seq
 - CD4 Effector Cells H3K4Me3 ChIP-Seq
- Task
 - Compare RNA-Seq (find differentially expressed genes)
 - Compare epigenetic modifications (differentially enriched regions)
 - Compare epigenetic modifications for specifically expressed genes

Experiment Description

Flexible and convenient way
to describe an experiment

The screenshot displays two windows from the NCI Energy department experiments management software.

Top Window: Experiment basic data

This window contains fields for experiment metadata:

- Belongs to:** Kartashov, Andrey
- General info** tab is selected.
- Cells:** CD4 T cells Naive
- Conditions:** 0h
- Genome Type:** Human
- Experiment Type:** RNA-Seq
- Fragmentation:** Sonication
- Experiment date:** 07/16/2011
- Spikeins pool:** N/A
- Spikeins:** 0
- Western blot:** Select an image
- Crosslink:** N/A
- Antibody:** H3K4me3
- Genome browser** section: Name for browser: CD4 T cells Naive 0h; Browser group name: NIH Project; Share data online? checked.
- Downloading** section: Template for file from Core: NIH_Naive_0h; Direct link to the file: [empty field].

Bottom Window: Laboratory data

This window shows a list of laboratory data entries:

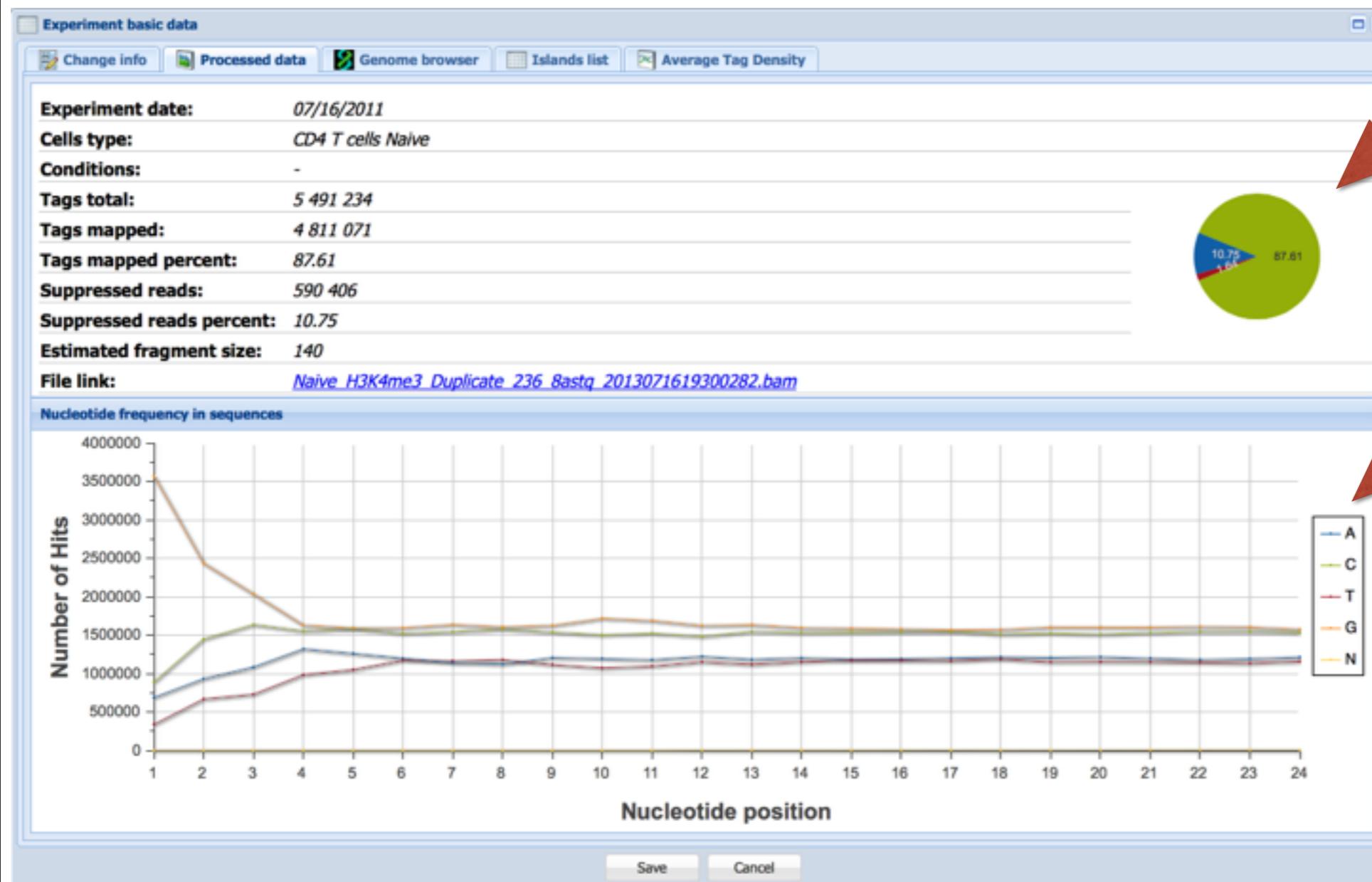
Record ID	Genome	Type	Cells
1	Human	CD4 T cells	Naive
2	Human	CD4 T cells	Naive
3	Human	CD4 T cells	Naive

Red arrows point to the top-left corner of the main window and the Laboratory data list window, highlighting specific features or sections of the interface.

Preliminary Automated Analysis

- RNA and ChIP
 - download data
 - quality control
 - upload to local UCSC genome browser
- RNA
 - mapping data by tophat
 - calculating RPKMs
- RNA With Spike In
 - Linear regression, shows scatter plot
- ChIP
 - mapping data by bowtie
 - identification of islands by MACS (I/2)
 - Average Tag Density
 - Islands distribution

Quality Control (DNA)



Mapping statistics

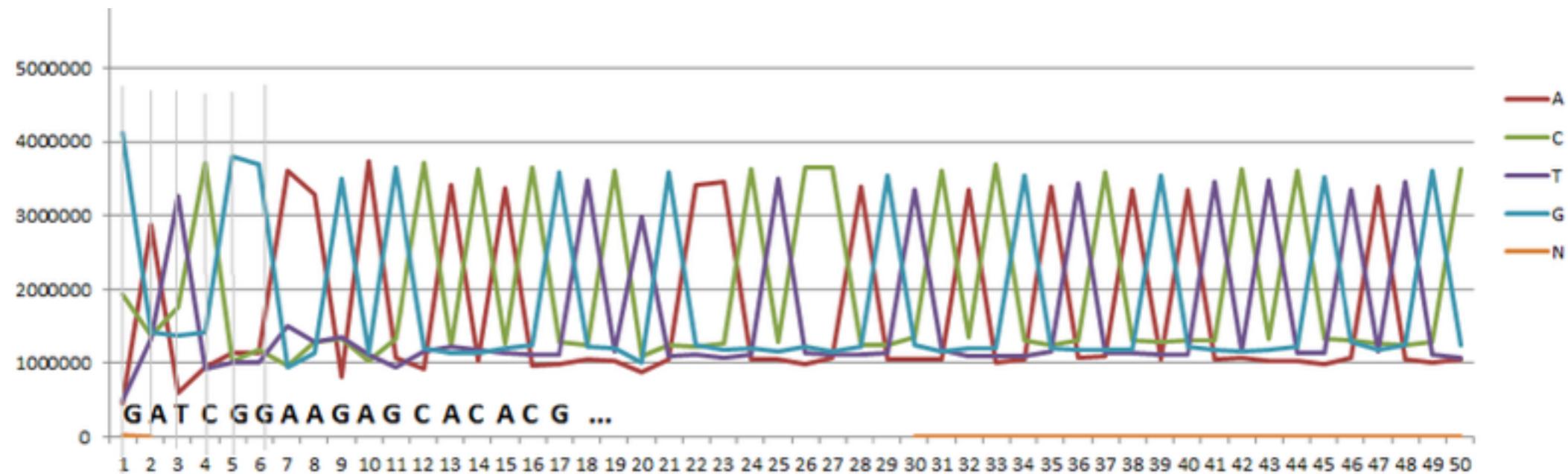
- Total tags mapped
- Suppressed reads
- Unmapped reads

Base call frequency

Nucleotide domination at position

Base Call Frequency RNA/DNA

Perfect example of an adaptor contamination



In cases of adapter contamination, peaks reflect the sequence of adapters.

Quality Control (RNA)

Mapping statistics

- Total tags mapped
- Ribosomal contamination
- Non-uniquely mapped and unmapped reads



Base call frequency
Nucleotide domination
at position

UCSC genome browser

- Data is uploaded directly to a local mirror of UCSC genome browser and available as a track
- Data is normalized to a total number of mapped reads
- Data is represented as a coverage
 - DNA:
 - for a pair-end libraries original fragment size is used
 - for a single-read estimated fragment size is used
 - RNA: read size is used

UCSC genome browser

Coverage

DNA: Estimated fragment size for single reads, original fragment size for pair-end reads

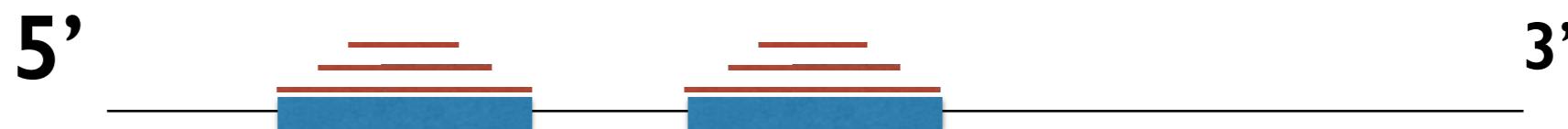
RNA: read length



UCSC genome browser Coverage

DNA: Estimated fragment size for single reads, original fragment size for pair-end reads

RNA: read length



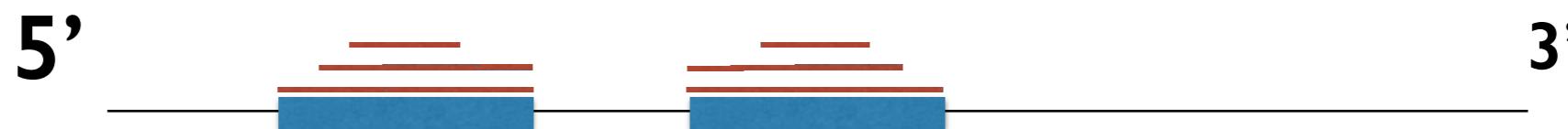
UCSC genome browser

Coverage

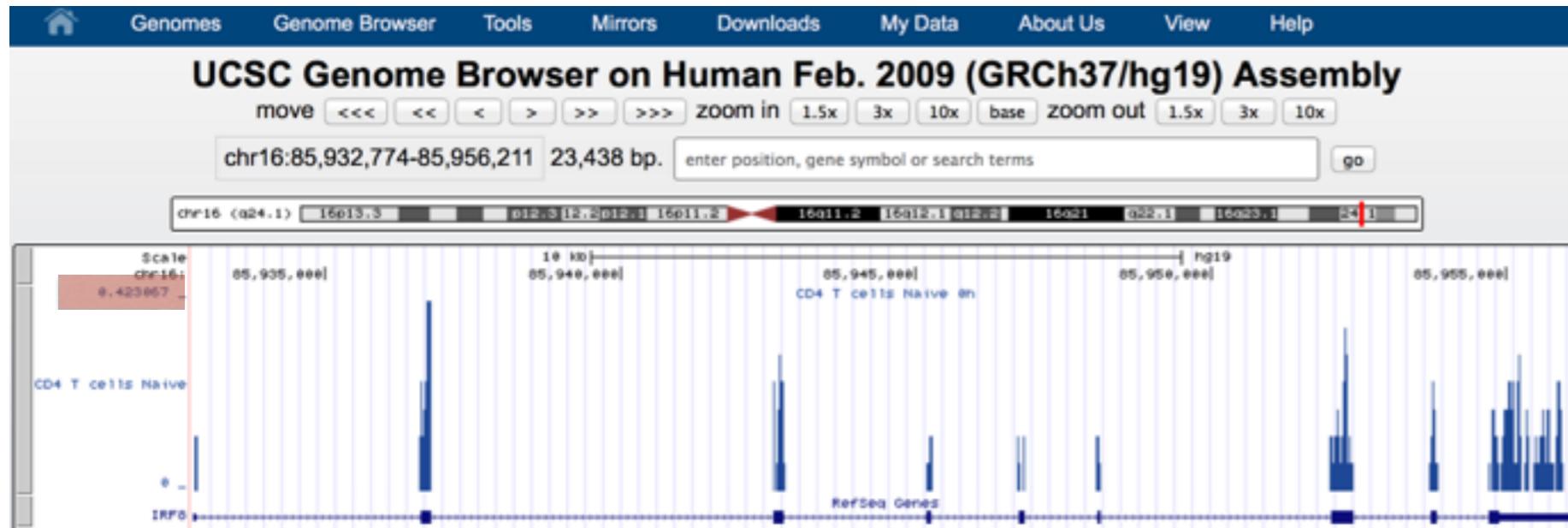
DNA: Estimated fragment size for single reads, original fragment size for pair-end reads



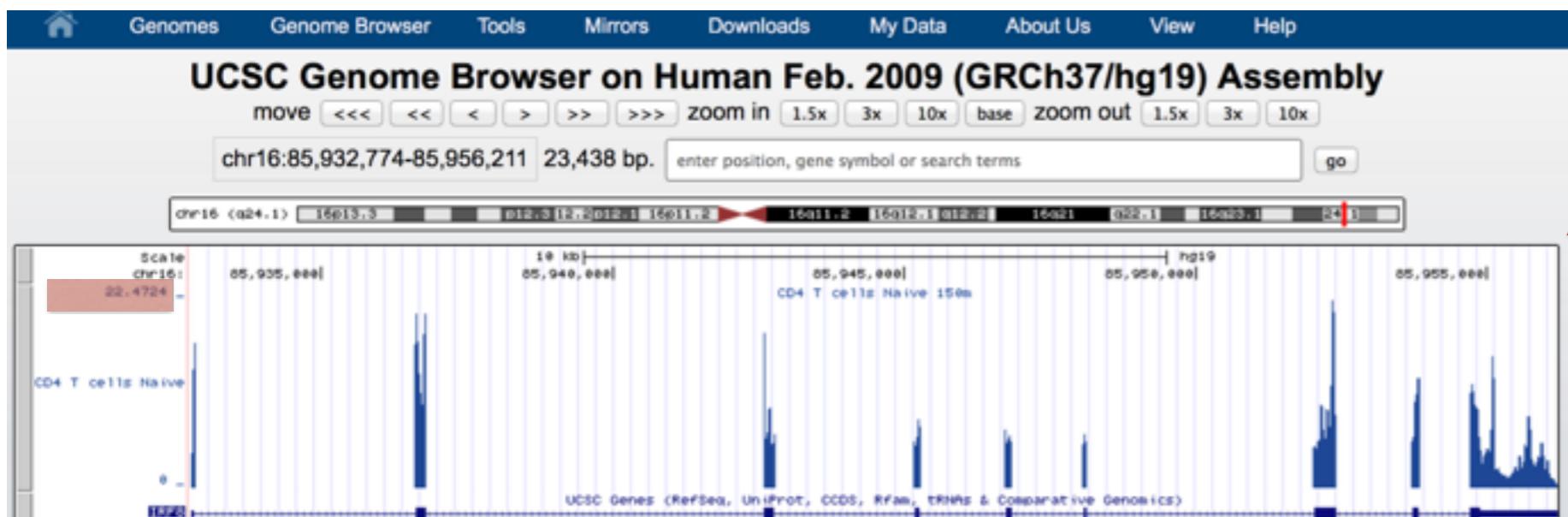
RNA: read length



UCSC genome browser RNA

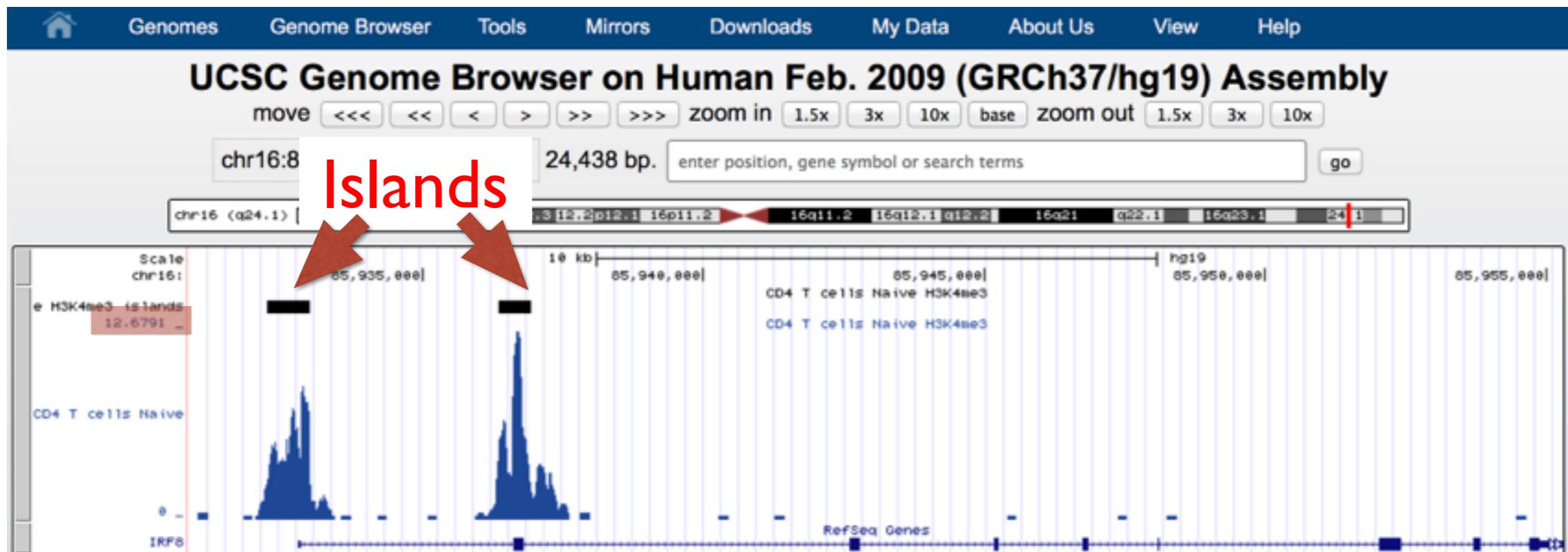


IRF8
0 time point
0.42 Normalized lvl
2.271 RPKM



IRF8
150m time point
22.47 Normalized lvl
149.363 RPKM

UCSC genome browser DNA



IRF8
0 time point
12.6791 Normalized lvl
2 Islands were found

Calculating RPKMs

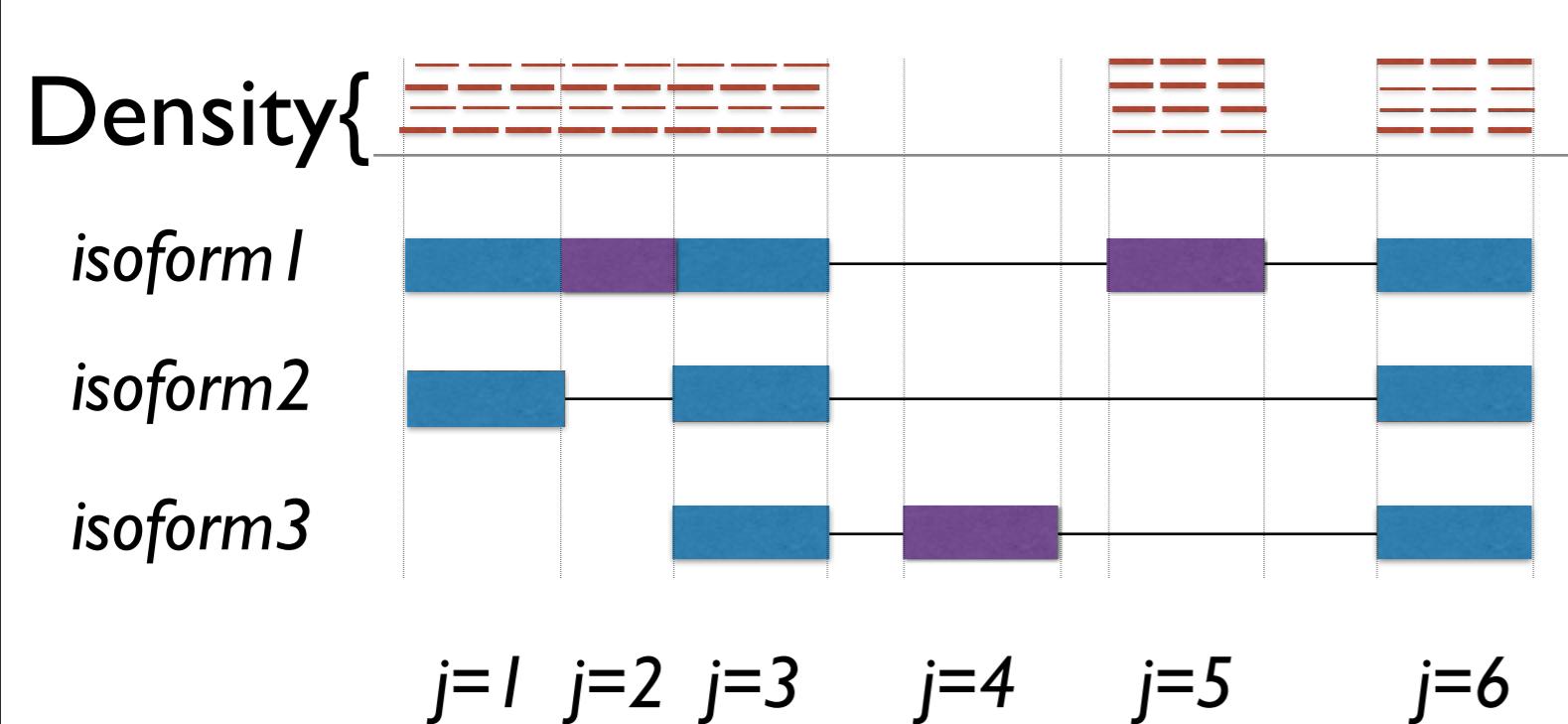
RPKM = Reads Per Kilobase of transcript length per Millionth of mapped reads

Existing tools:

- HTSeq-count
- Cufflinks
- Genespring
- IQSeq
- ...

Our objective is to estimate RPKMs with respect to the current annotation and to quantify expression precisely as possible.

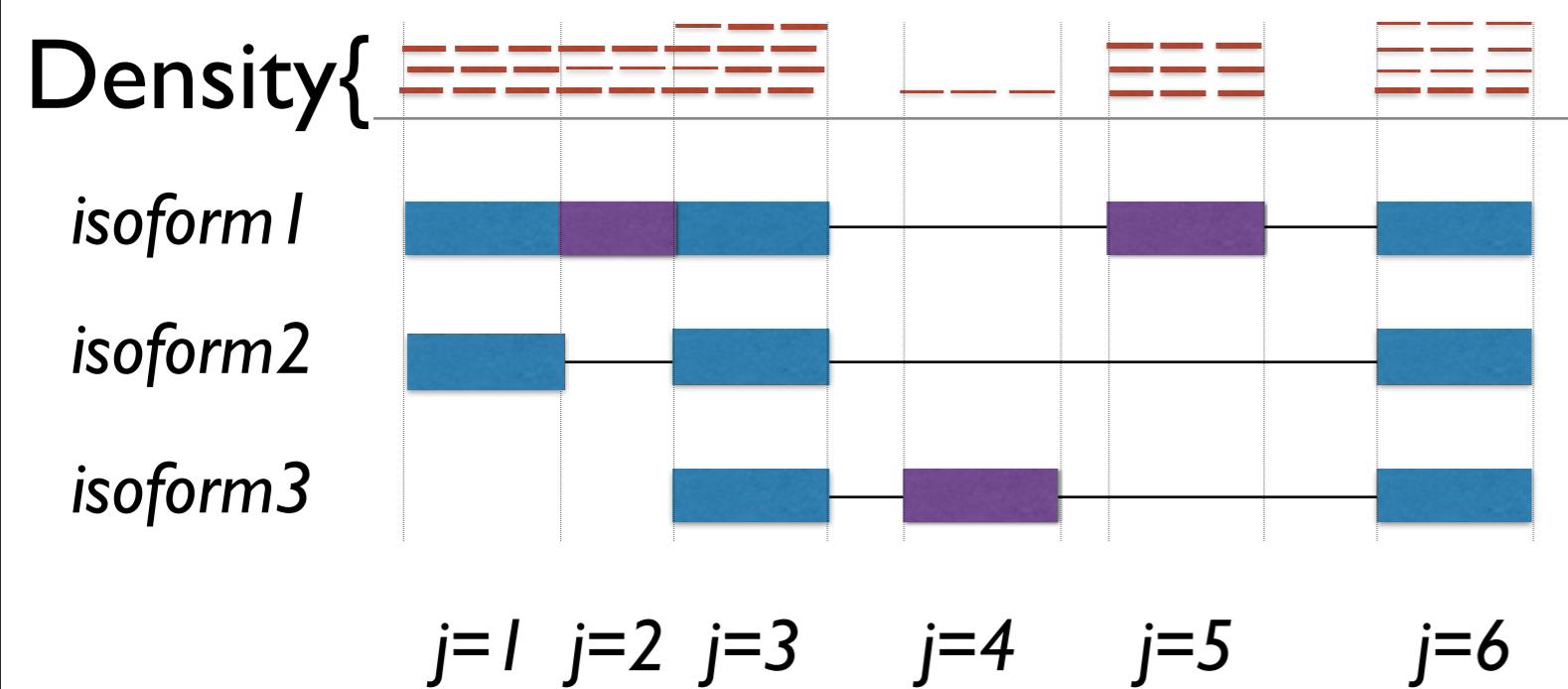
Calculating RPKMs



1. Densities across exons within the same isoform have to be the same
2. Sum of exon densities for the same region should be equal to the total density for that region

density $d_{i_j} = \frac{N_{i_j}}{L_j}$, where N_{i_j} is a number of reads in a region J of isoform I and L is the length of the region J

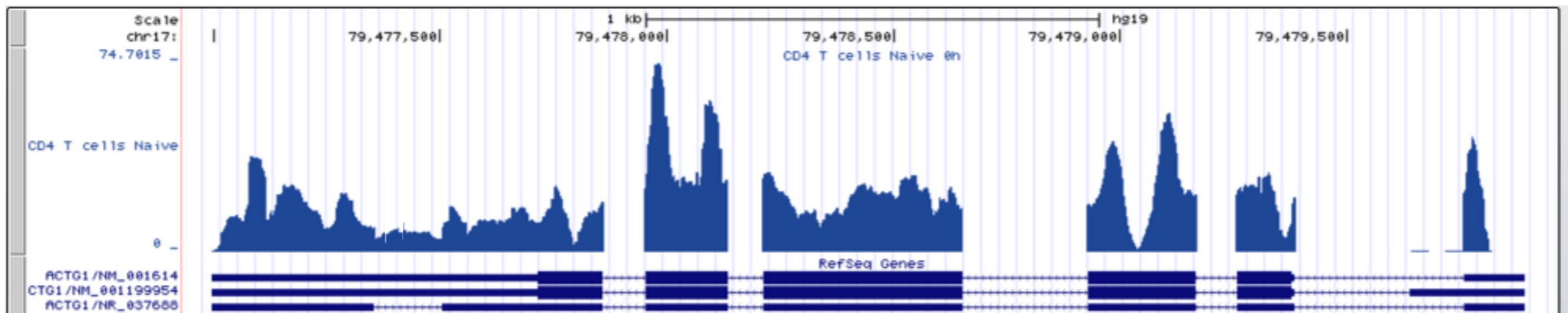
Calculating RPKMs



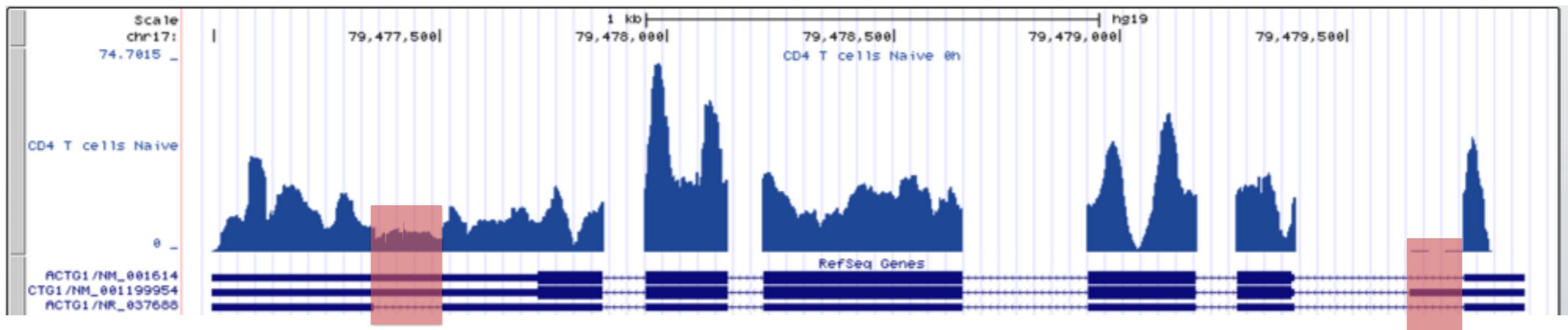
1. Densities across exons within the same isoform have to be the same
2. Sum of exon densities for the same region should be equal to the total density for that region

density $d_{i_j} = \frac{N_{i_j}}{L_j}$, where N_{i_j} is a number of reads in a region J of isoform I and L is the length of the region J

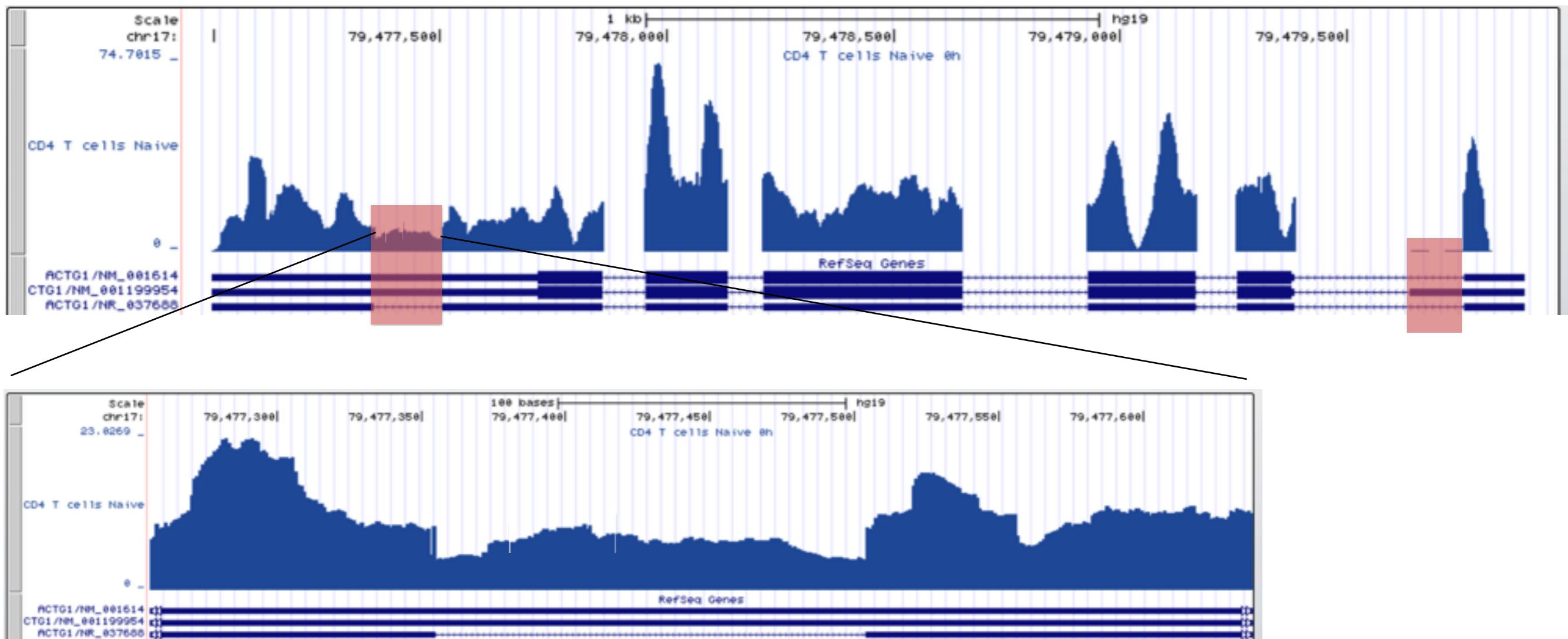
Calculating RPKMs



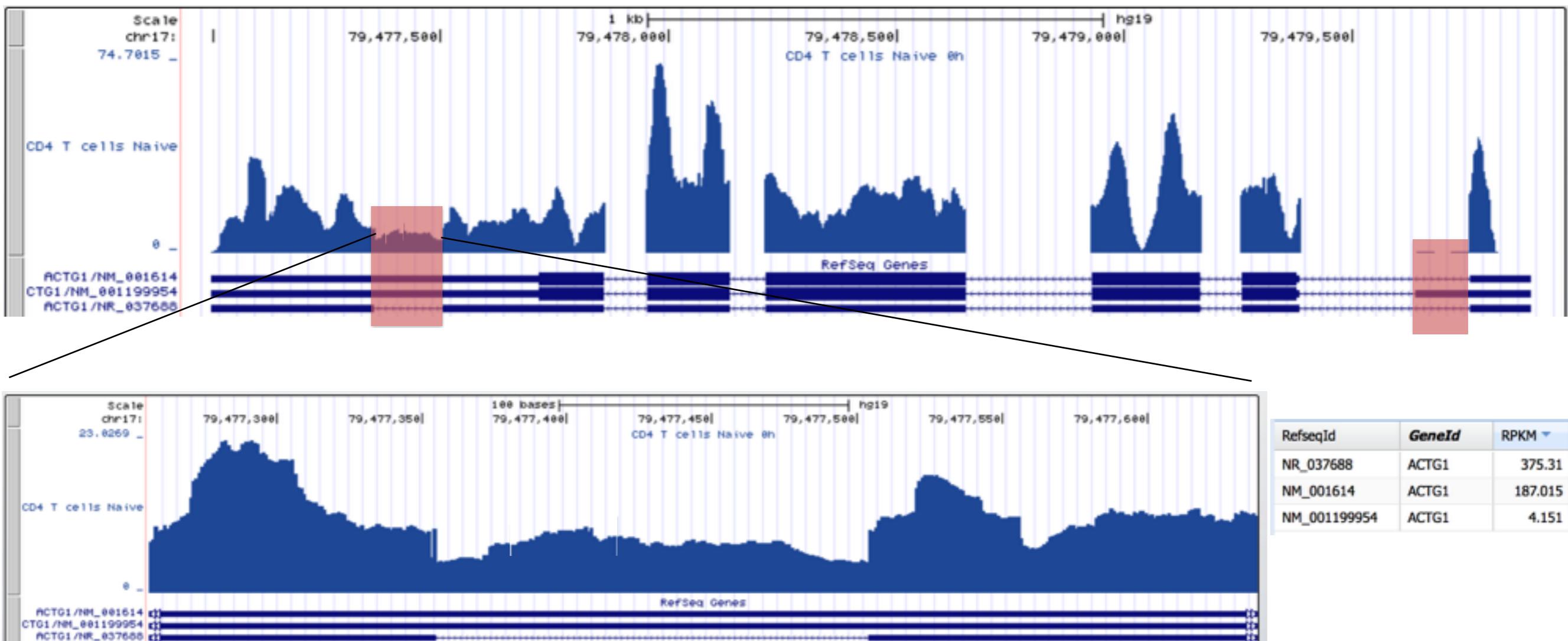
Calculating RPKMs



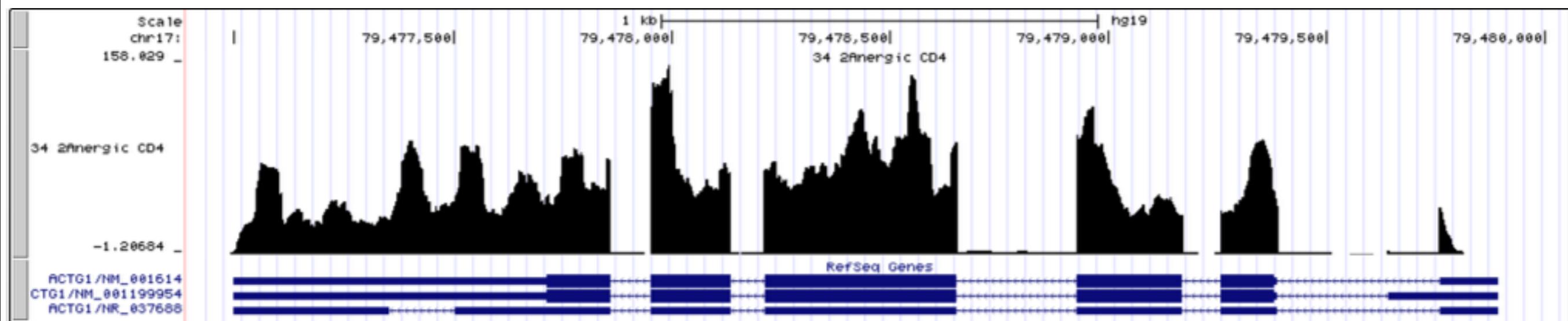
Calculating RPKMs



Calculating RPKMs

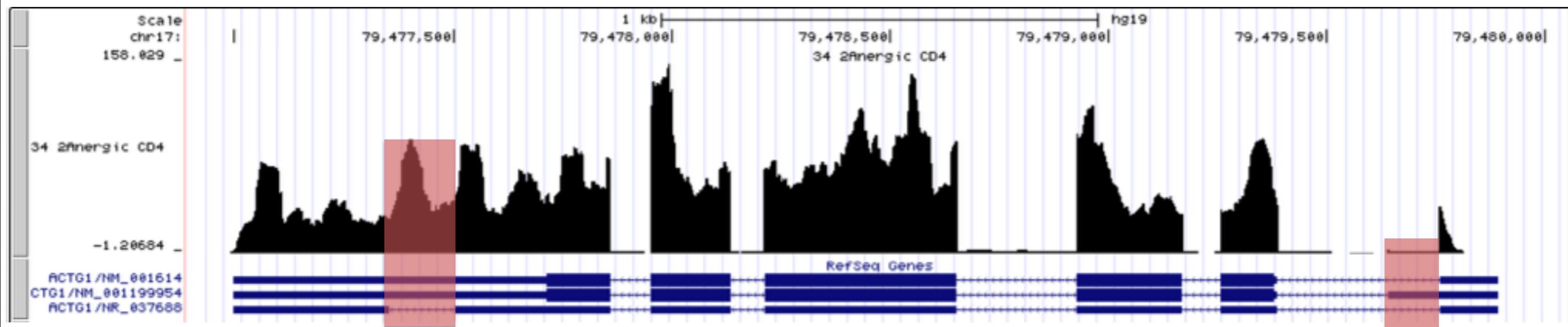


Calculating RPKMs



RefseqId	GeneId	RPKM ▾
NM_001614	ACTG1	994.948
NR_037688	ACTG1	118.744
NM_001199954	ACTG1	14.3

Calculating RPKMs

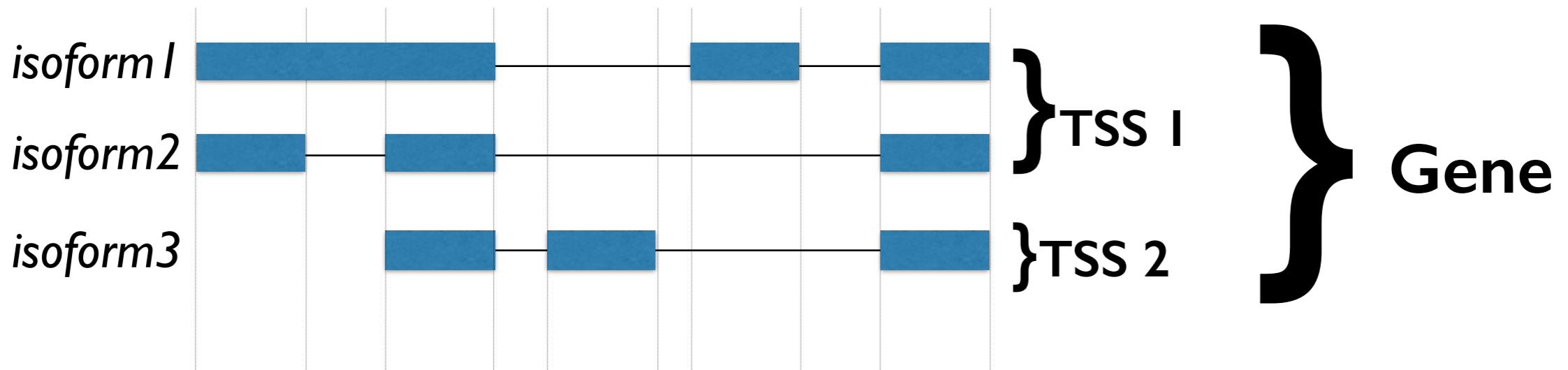


RefseqId	GeneId	RPKM ▾
NM_001614	ACTG1	994.948
NR_037688	ACTG1	118.744
NM_001199954	ACTG1	14.3

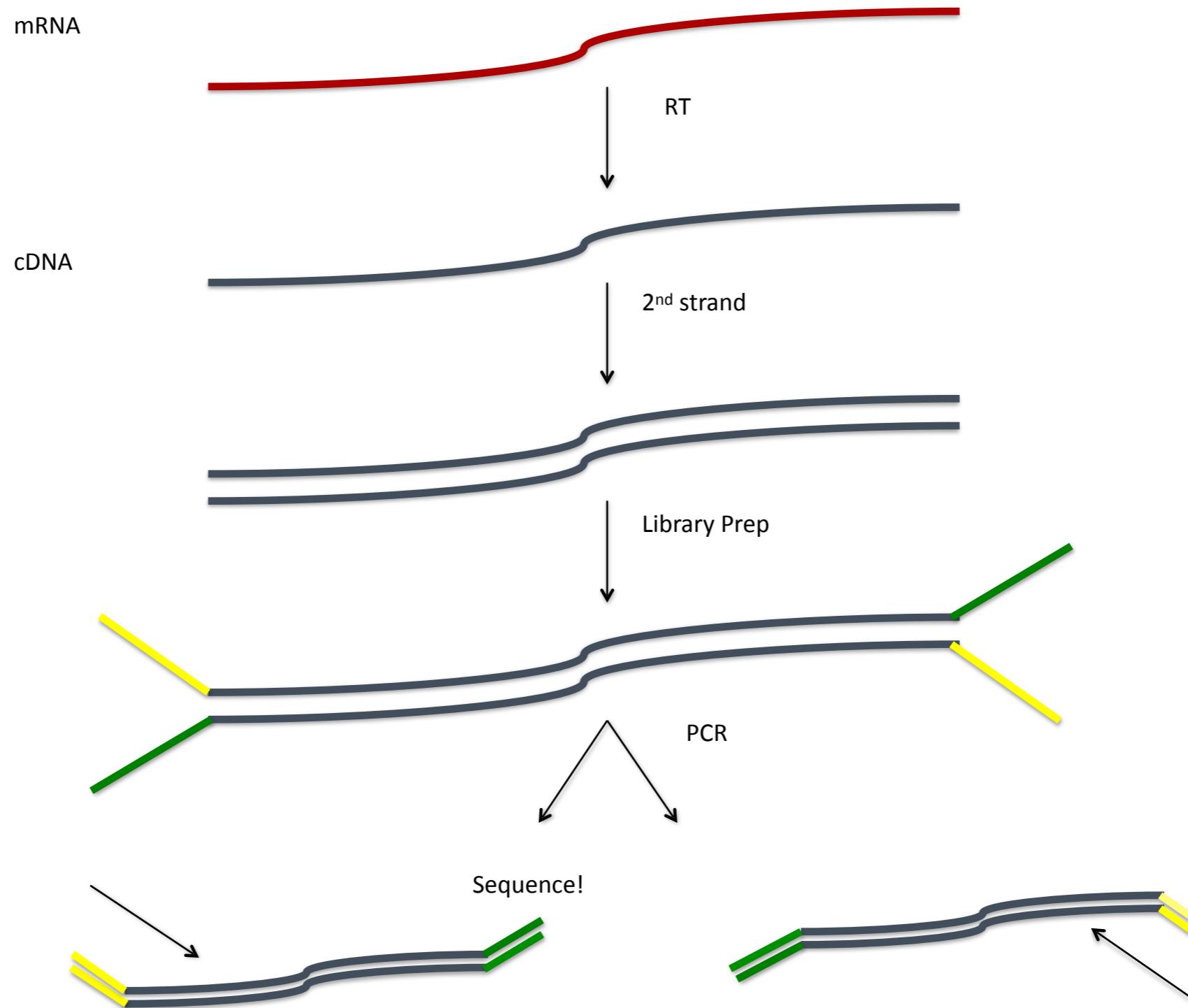
Different RPKM grouping

With respect to annotation, RPKMs can be grouped by

- Isoforms
- Gene name, sum up all isoform's RPKMs for the same gene name
- Transcription Start Site, sum up all isoform's RPKMs for common TSS

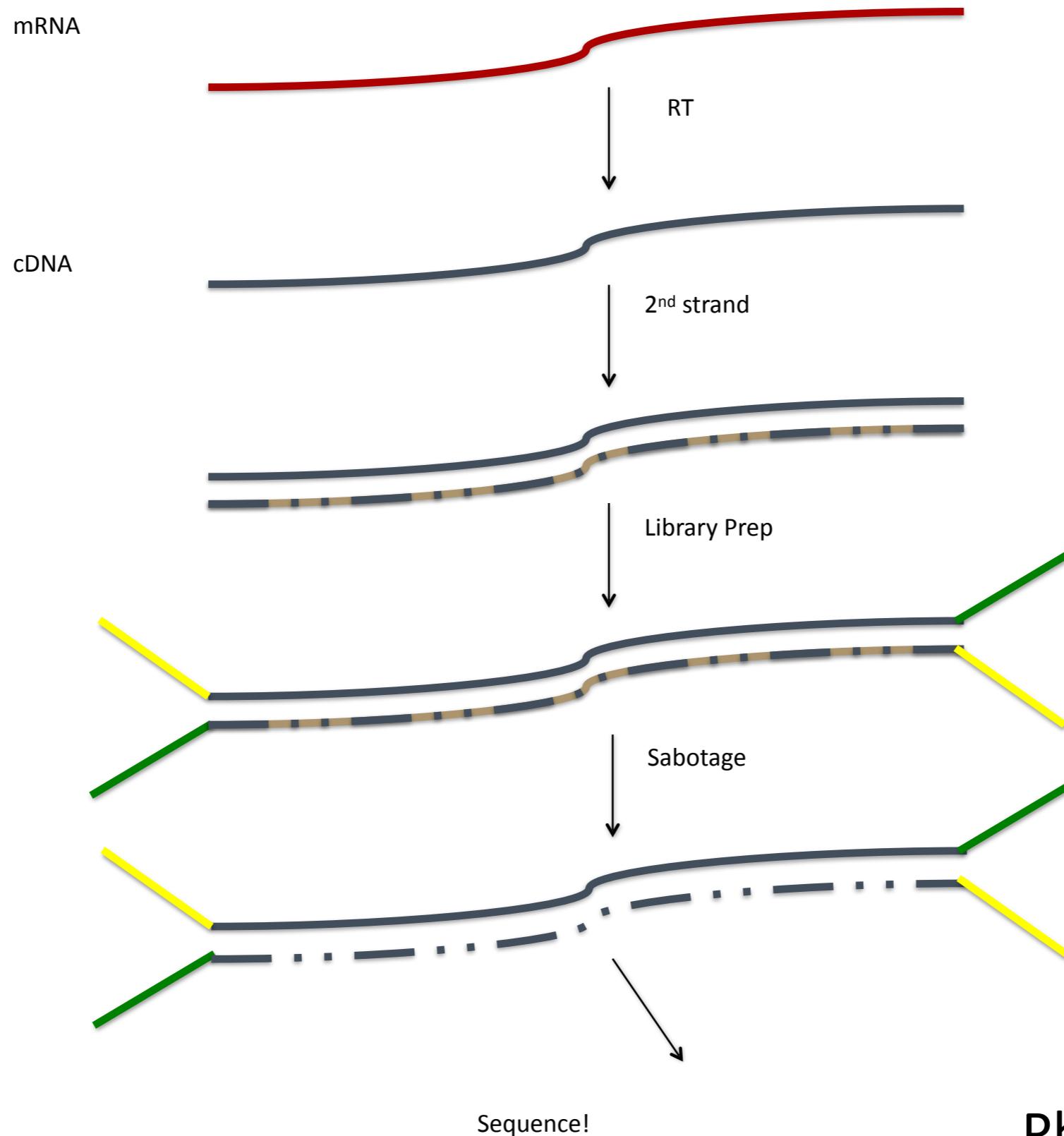


Mapping dUTP RNA-Seq libraries



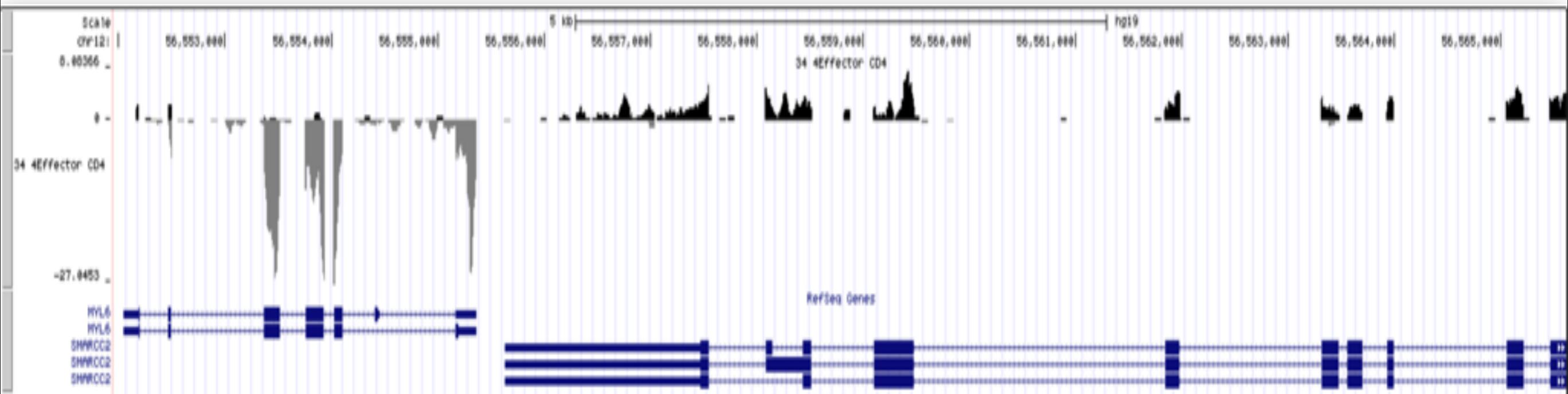
Phillip Dexheimer

Mapping dUTP RNA-Seq libraries



Phillip Dexheimer

Mapping dUTP RNA-Seq libraries



Distinguishing reads between positive and negative strand helps to increase accuracy.

RPKMs List

Experiment basic data

Change info Processed data Genome browser RPKM list

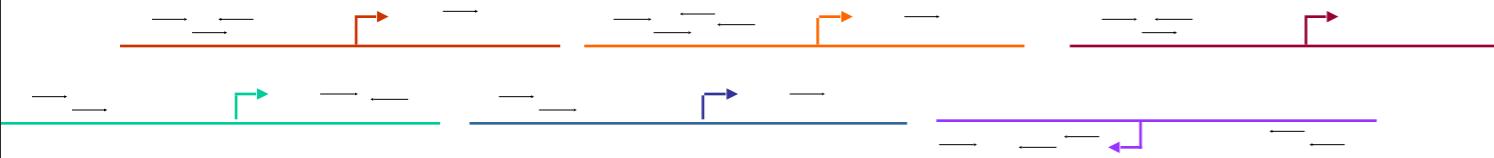
Page 1 of 283 Displaying 1 - 100 of 28287 Common Tss jump save

RefseqId	GeneId	chrom	txStart	txEnd	strand	RPKM	
NM_001030	RPS27	chr1	153963239	153964631	+	10983.7	
NM_001035267,NM_021104	RPL41	chr12	56510374	56511616	+	6310.114	
NM_001402	EEF1A1	chr6	74225473	74230755	-	6214.92	
NM_003295	TPT1	chr13	45911304	45915297	-	5689.13	
NR_026712	RPL13AP5	chr19	49990865	49995096	+	5254.85	
NM_001015	RPS11	chr19	49999622	50002969	+	5191.6	
NR_039666	MIR4461	chr5	134263729	134263802	+	5002.48	
NM_000980	RPL18A	chr19	17970687	17974133	+	4583.93	
NM_001004	RPLP2	chr11	809936	812876	+	4492.28	
NM_001020	RPS16	chr19	39923847	39926618	-	4132.23	
NM_000989	RPL30	chr8	99053938	99057818	-	4040.19	
NM_001003,NM_213725	RPLP1	chr15	69745159	69747884	+	3894.541	
NM_001028	RPS25	chr11	118886422	118889057	-	3763.92	
NM_000991,NM_001136134,NM_0...	RPL28	chr19	55897300	55903451	+	3709.856	
NM_000967,NM_001033853	RPL3	chr22	39708887	39715670	-	3661.28	
NM_000998	RPL37A	chr2	217363520	217366188	+	3632.94	
NM_022551	RPS18	chr6	33239852	33244281	+	3587.15	
NM_000975,NM_001199802	RPL11	chr1	24018269	24022915	+	3561.91	
NM_002952	RPS2	chr16	2012062	2014827	-	3529.34	
NM_000981	RPL19	chr17	37356536	37360980	+	3391.06	
NM_000999,NM_001035258	RPL38	chr17	72199795	72206019	+	3320.06	
NM_001016	RPS12	chr6	133135708	133138703	+	3230.98	
NM_004048	B2M	chr15	45003685	45010357	+	3149.52	
NM_001010	RPS6	chr9	19376254	19380235	-	3080.08	
NM_001031	RPS28	chr19	8386384	8387280	+	2808.47	

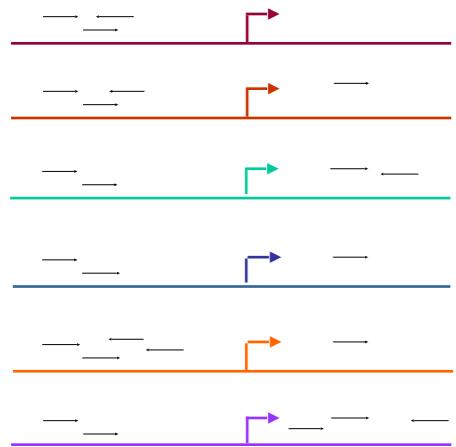
Save

Cancel

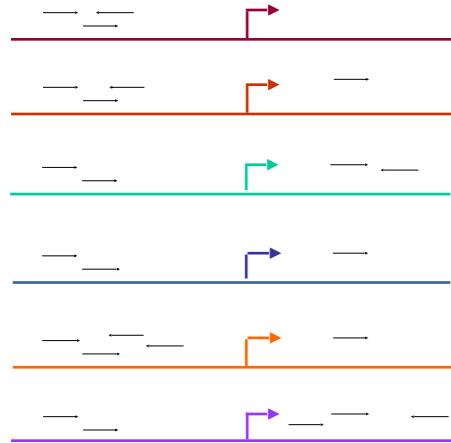
Average tag density profiles as part of the quality control



Average tag density profiles as part of the quality control

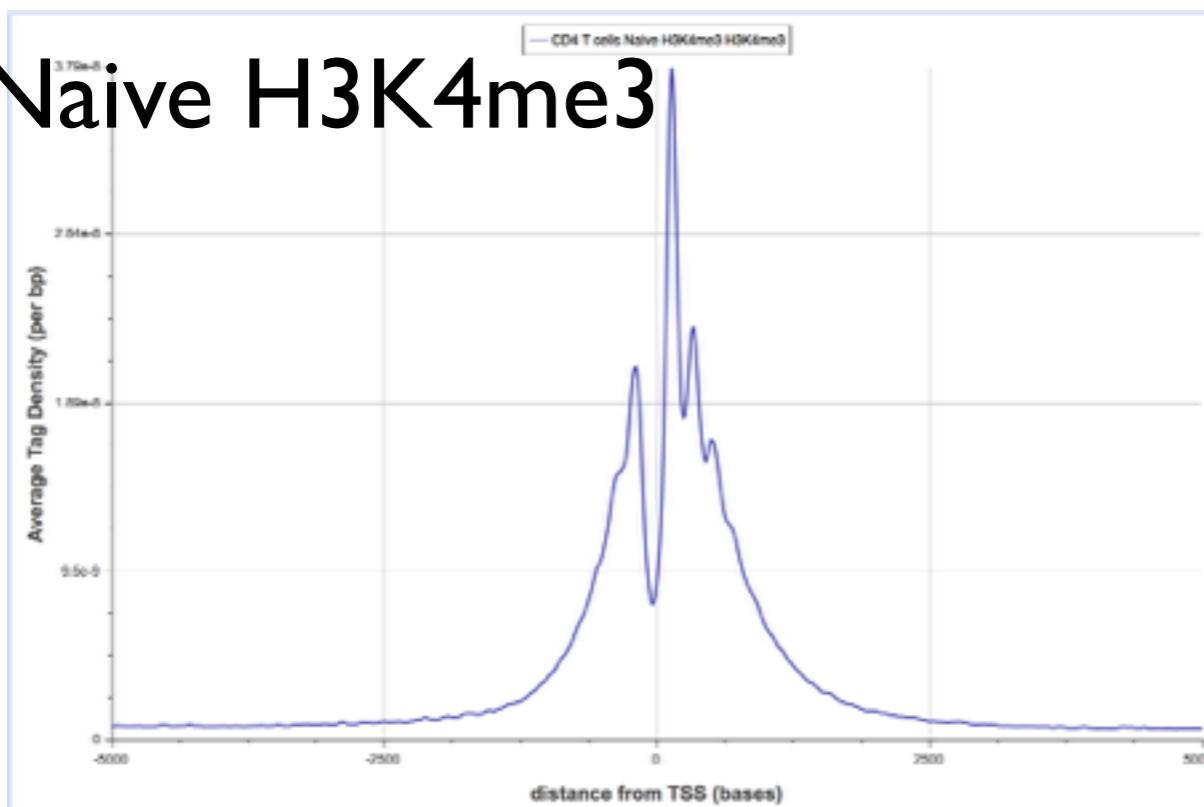


Average tag density profiles as part of the quality control

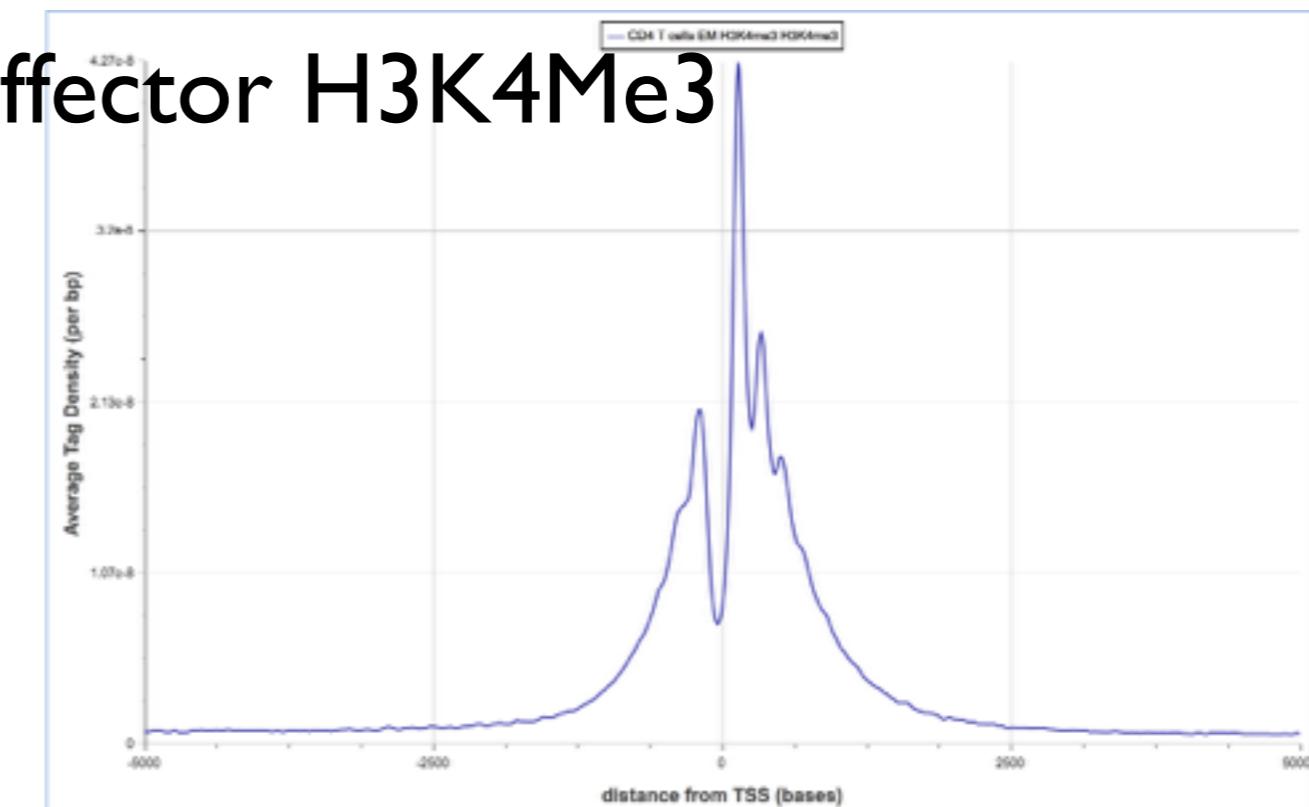


Same lvl of enrichment,
allows direct comparison

Naive H3K4me3



Effector H3K4Me3



Islands: MACS output

Experiment basic data

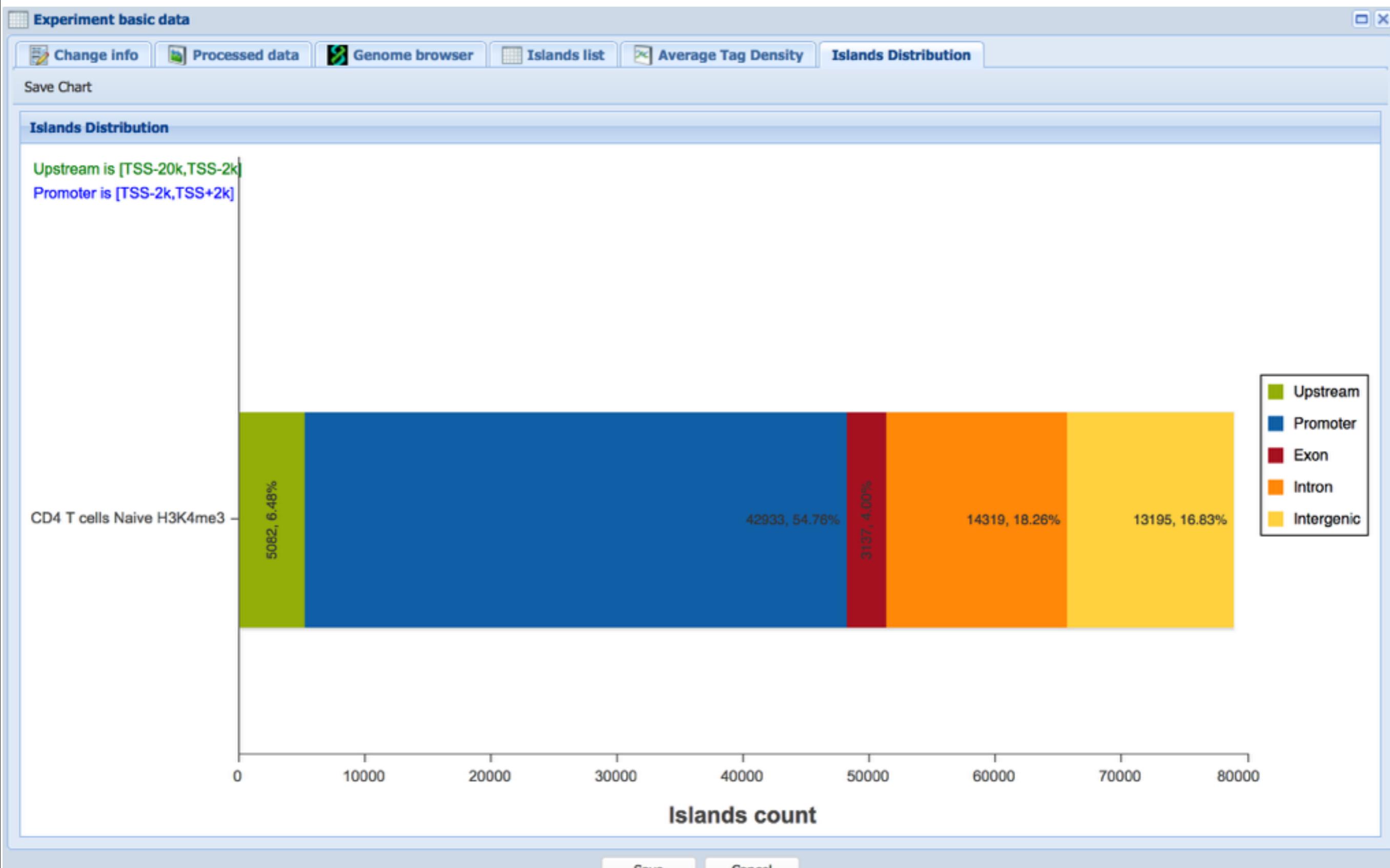
Change info Processed data Genome browser Islands list Average Tag Density Islands Distribution

Page 1 of 785 | Displaying 1 - 100 of 78403 | jump save

chrom	start	end	length	pileup	abssummit	log10p	foldenrich	log10q	
chr8	103540671	103541261	591	90	103540903	110.86	28.918	101.671	
chr6	27356237	27357176	940	88	27356587	90.386	20.231	82.753	
chr1	100231906	100232260	355	67	100232075	87.554	28.618	80.037	
chr7	66385785	66386133	349	91	66385955	87.471	18.036	80.014	
chr12	4714005	4714507	503	71	4714129	81.013	22.981	73.867	
chr17	19266076	19266320	245	61	19266195	79.12	27.032	72.1	
chr3	113775628	113776193	566	82	113775736	78.121	17.298	71.145	
chr11	31531338	31532072	735	82	31531464	75.95	16.449	69.113	
chr3	143566514	143567324	811	80	143566755	75.937	17.077	69.104	
chr1	39491775	39492761	987	77	39492150	75.863	18.07	69.058	
chr2	122732810	122733078	269	53	122732936	75.826	29.357	69.03	
chr17	48449923	48450555	633	64	48450425	75.474	23.421	68.697	
chr3	9290167	9291349	1183	77	9291044	75.196	17.787	68.448	
chr6	144415939	144416715	777	77	144416572	74.804	17.621	68.085	
chr1	85086328	85086758	431	58	85086623	74.367	25.879	67.683	
chr2	16789987	16790447	461	61	16790332	74.274	24.266	67.593	
chr13	96130960	96131384	425	57	96131085	74.066	26.232	67.411	
chr18	53144545	53145368	824	71	53144818	73.916	19.474	67.271	
chr11	62521392	62521691	300	60	62521489	73.605	24.4	66.984	
chr7	100888117	100888388	272	70	100888251	72.641	19.275	66.083	
chr19	3434996	3435253	258	62	3435125	72.56	22.814	66.008	
chr10	69609279	69609962	684	63	69609419	70.941	21.465	64.492	
chr19	41869785	41870112	328	62	41869925	70.561	21.731	64.151	
chr19	11639475	11639946	472	54	11639813	70.135	25.511	63.744	
chr12	121453800	121454272	473	57	121454131	69.484	23.59	63.107	

Save Cancel

Islands distribution



Advanced Analysis

- RNA
- DESeq / DESeq2
- RNA and ChIP
- ChIP
- MANorm / DiffBind
- Average Tag Density

Project Designer

Project designer

Projects

Type project name, press enter to add

Owned
Project1
Shared

To add a new project type project name in a textfield, which is in the left top corner of the window and then press enter. New project name (with folder icon) will be shown in a panel under the textfield. Select it by mouse click and available analysis will appear. At first you have to create lists with which you will working it can be done by "Genes Lists" or "DESeq" analysis, to do it just click on the corresponded panel.

Genes Lists

This function allows you to organize and manage genes lists (grouping, filtering) for future analysis. All lists can be saved in excel like format. If you dont know where to start, start from here.

DESeq

To produce differentially expressed gene lists use this function. You can use it to compare groups of treated and untreated experiments and also when you need differences in series of experiments.

ATP & filter

ATP is Average Tag Density Profile plot which shows modification level (enrichment) for particular gene list. You can combine all gene list created in "Genes Lists" or "DESeq" analysis and all DNA-Seq experiments in one plot.

DESeq2

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

MANorm

MANorm is a simple and effective method, for quantitative comparison of ChIP-Seq data sets describing transcription factor binding sites and epigenetic modifications.

Details

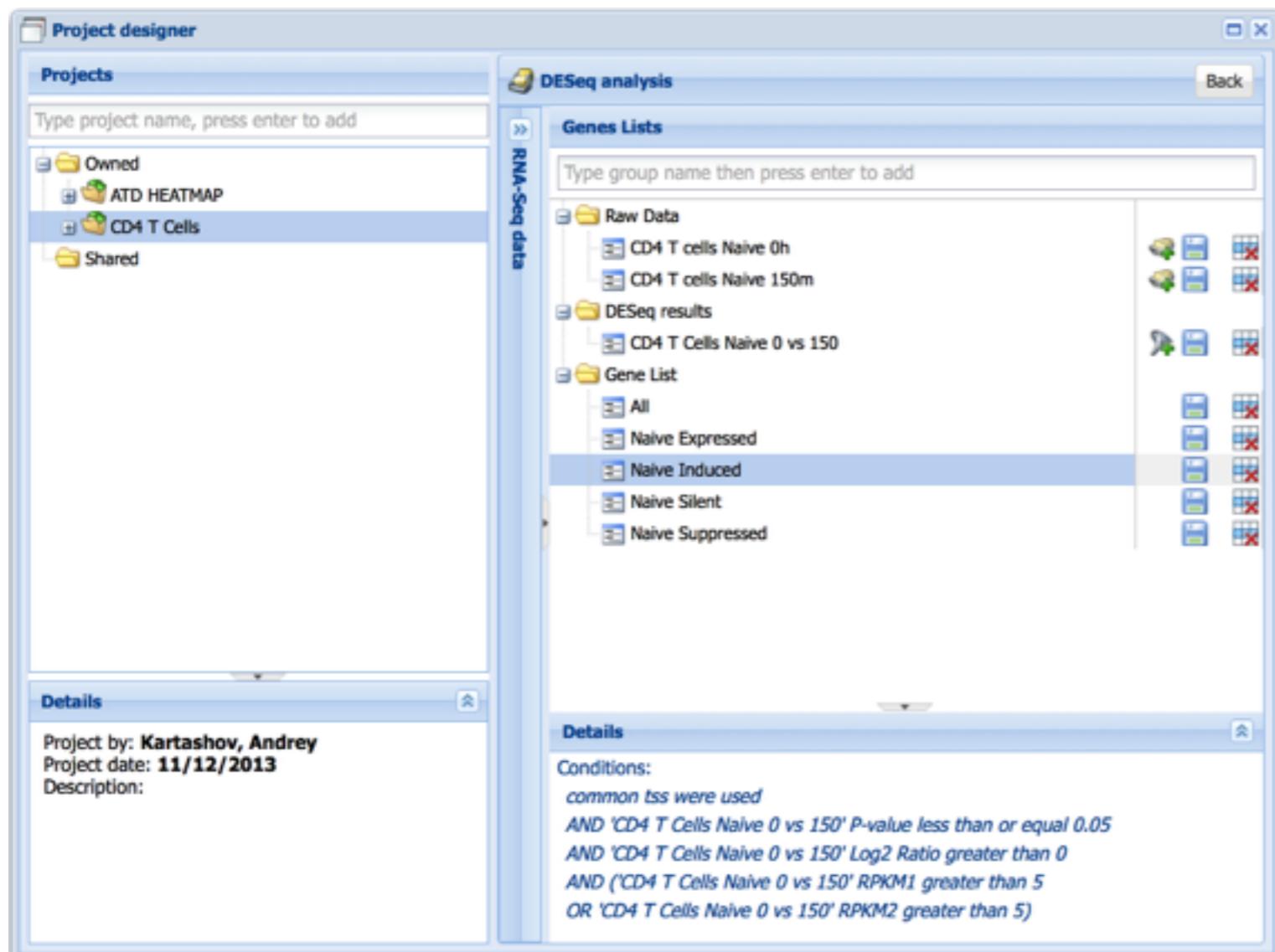
Project by: Kartashov, Andrey
Project date: 10/10/2013
Description:

Defined lists of genes

- All - all genes grouped by TSS
- Silent are genes which expression level less then 2 RPKM at point 0
- Expressed are genes which expression level greater then 8 RPKM at point 0
- Induced are genes which expression level has been significantly increased during 150m treatment, but were silent at 0 time point
- Suppressed are genes which expression level has been significantly decreased during 150m treatment and were expressed at 0 time point

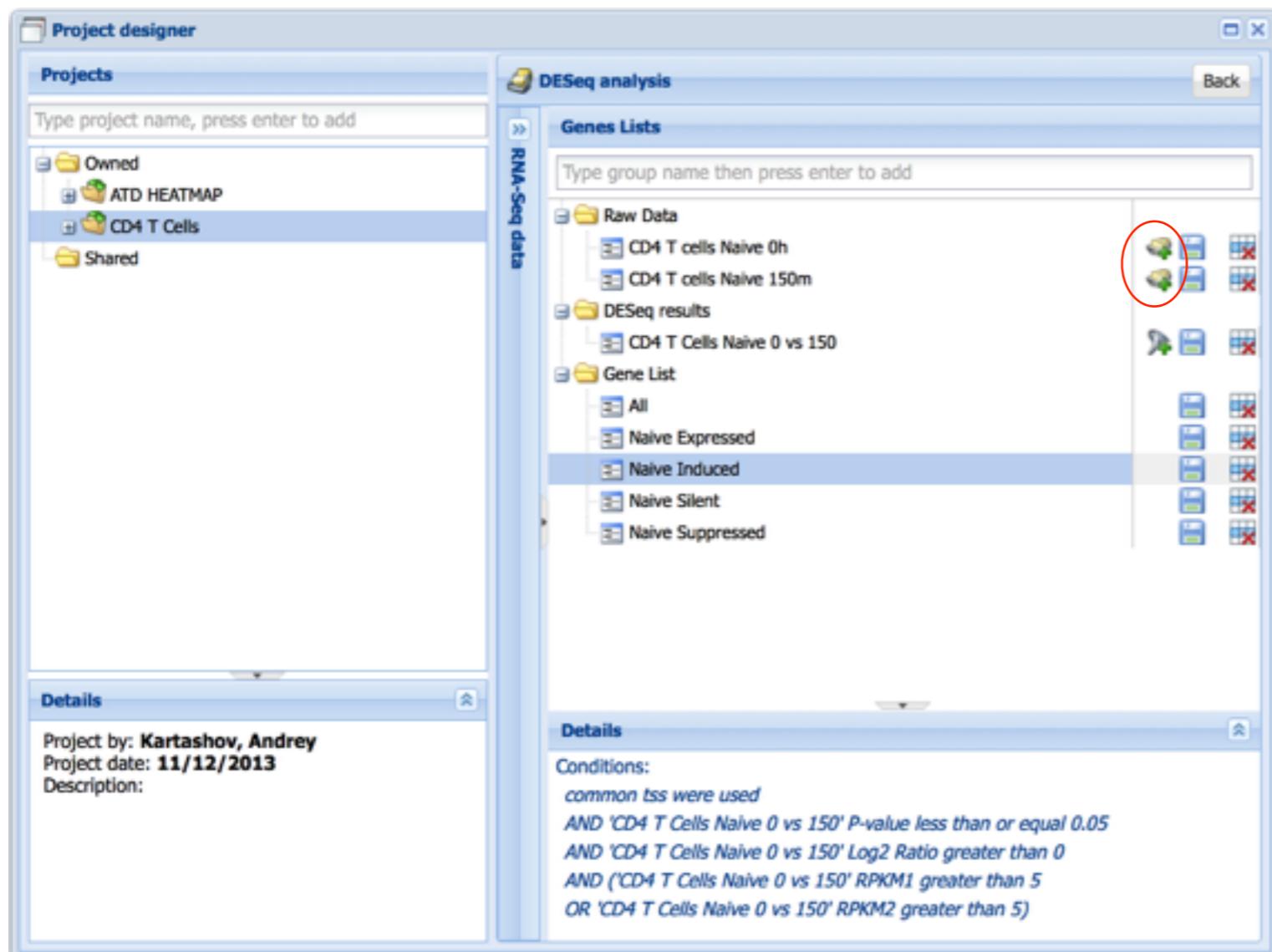
DESeq/DESeq2

(Differential Gene Expression Analysis)



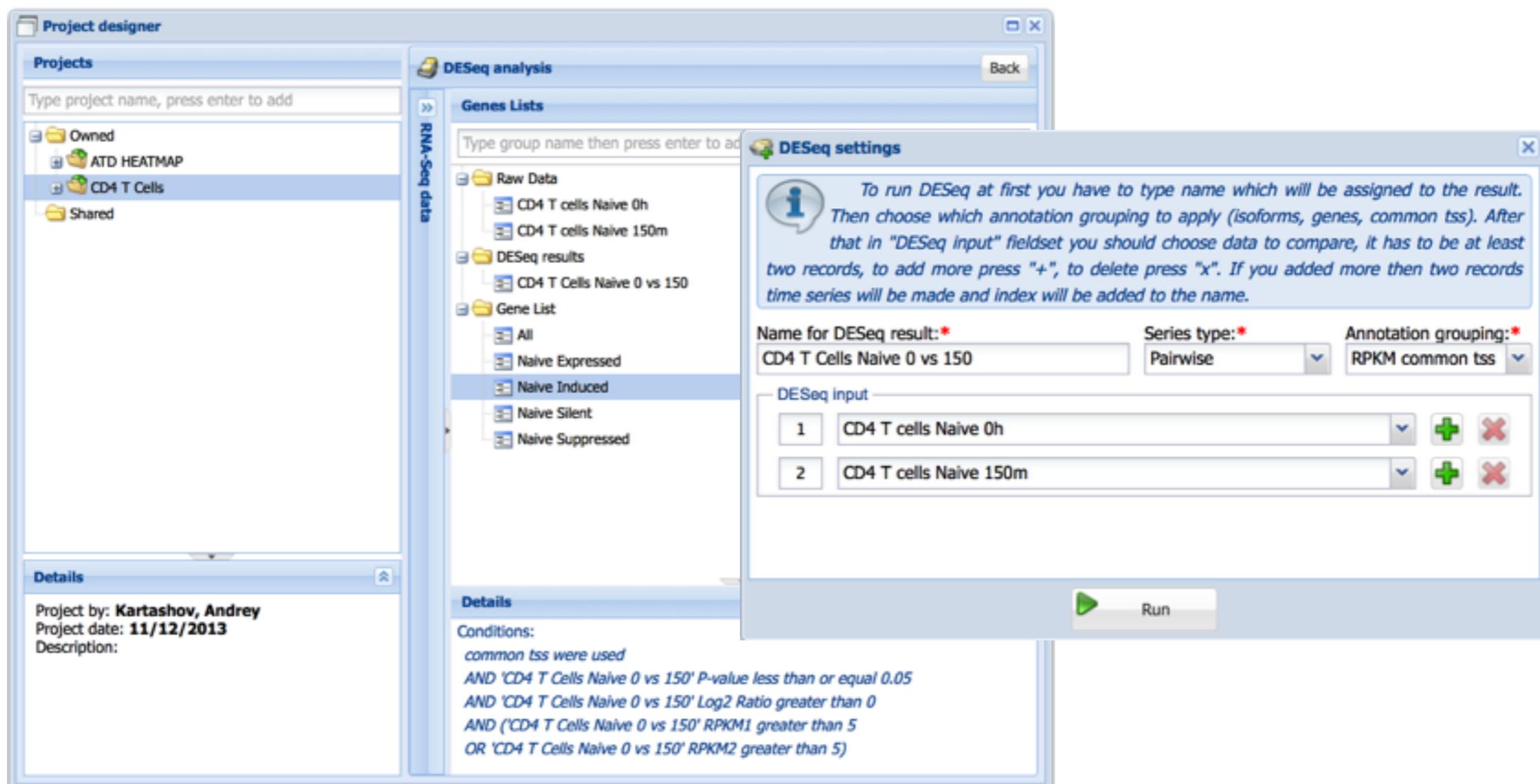
DESeq/DESeq2

(Differential Gene Expression Analysis)



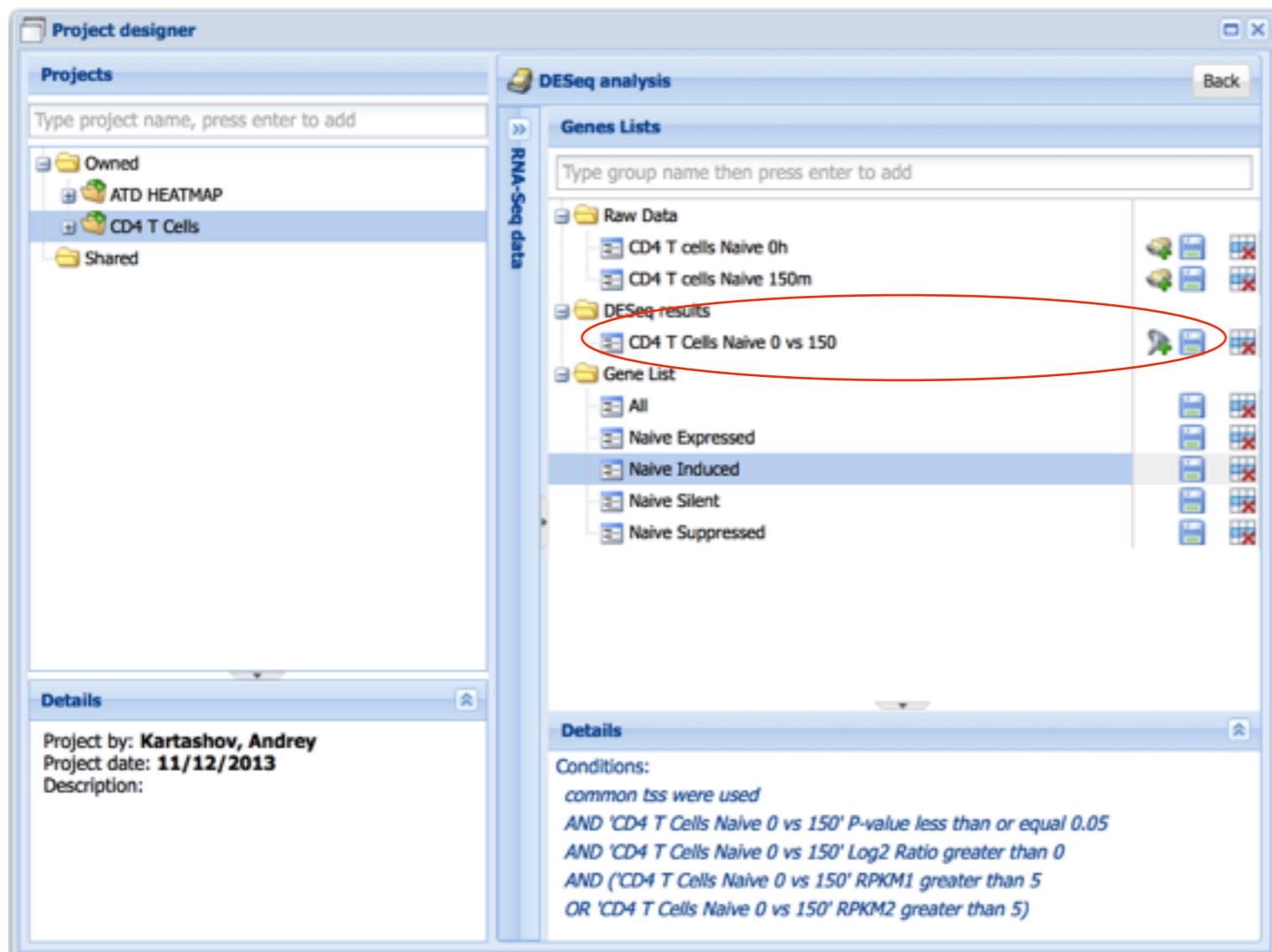
DESeq/DESeq2

(Differential Gene Expression Analysis)



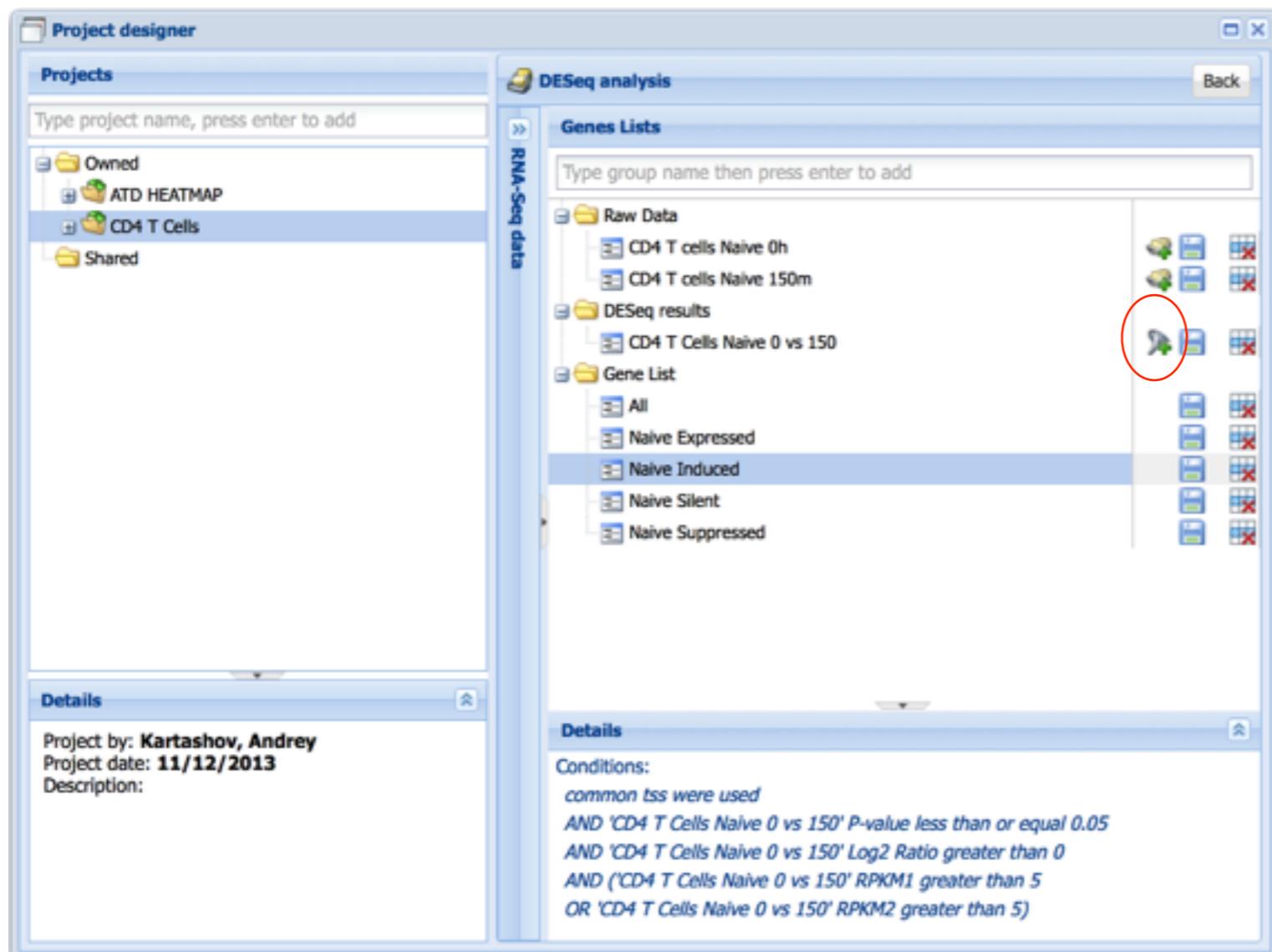
DESeq/DESeq2

(Differential Gene Expression Analysis)



DESeq/DESeq2

(Differential Gene Expression Analysis)



DESeq/DESeq2

(Differential Gene Expression Analysis)

Project designer

Projects

Type project name, press enter to add

- Owned
 - ATD HEATMAP
 - CD4 T Cells
- Shared

DESeq analysis

Genes Lists

Filter settings

To apply a filter at first you have to type filter name which will be assigned to the result. Then choose which annotation grouping to apply (isoforms, genes, common tss). After that in "Filter" fieldset you should choose data source and field which you want to filter; it has to be at least one record, to add more press "+", to delete press "x".

Filter name, will be saved with this name: * Naive Induced

Annotation grouping: * RPKM common tss Remove Non-coding RNA

Filter

AND	CD4 T Cells Naive 0 vs 150	P-value	less than or equal	0.05	<input type="button" value="+"/> <input type="button" value="X"/>
AND	CD4 T Cells Naive 0 vs 150	RPKM1	less than	2	<input type="button" value="+"/> <input type="button" value="X"/>
AND	CD4 T Cells Naive 0 vs 150	RPKM2	greater than	5	<input type="button" value="+"/> <input type="button" value="X"/>

Details

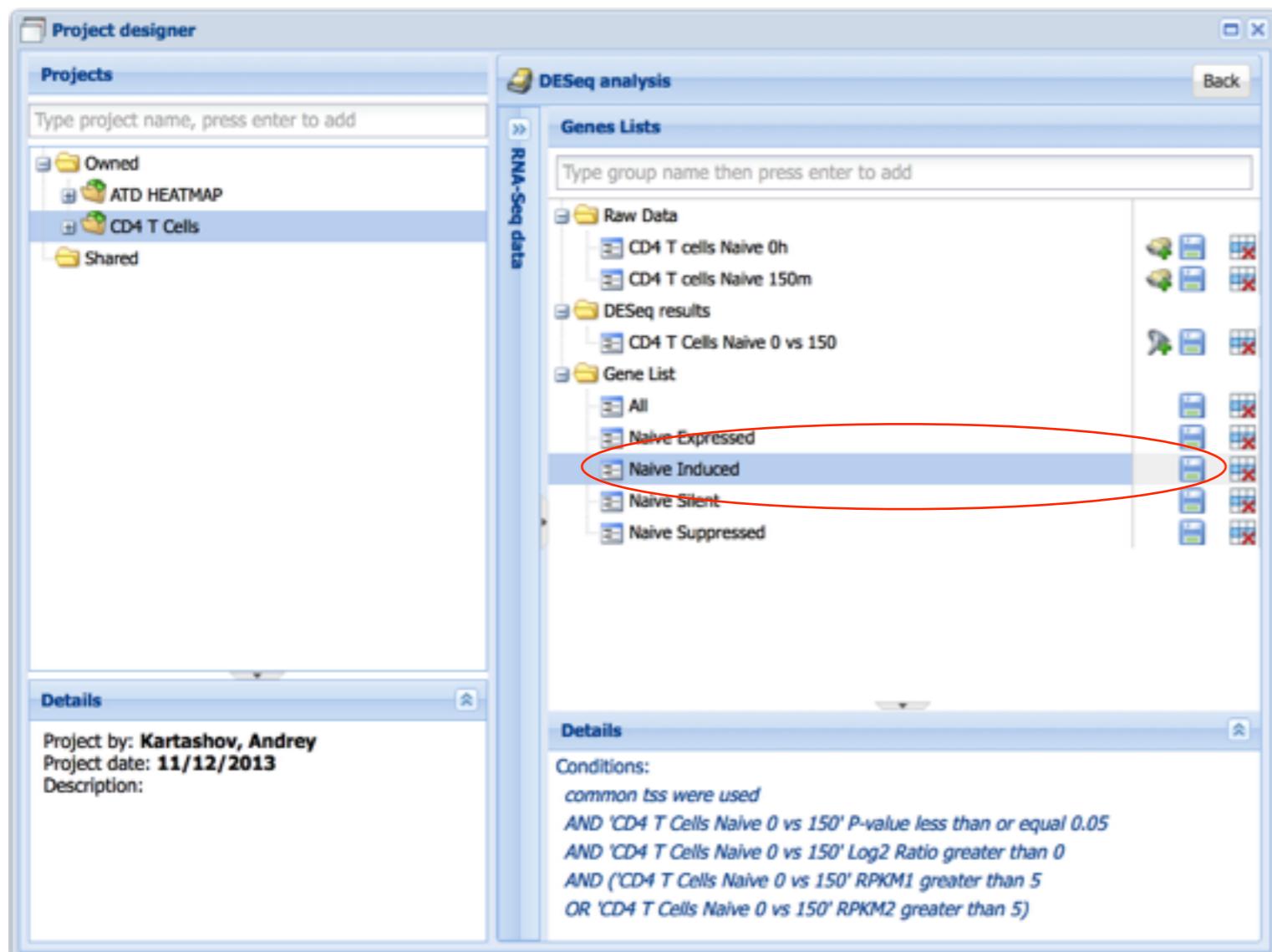
Project by: Kartashov, Andrey
Project date: 11/12/2013
Description:

Conditions:
common tss were used
AND 'CD4 T Cells Naive 0 vs 150' P-value less than or equal 0.05
AND 'CD4 T Cells Naive 0 vs 150' Log2 Ratio greater than 0
AND ('CD4 T Cells Naive 0 vs 150' RPKM1 greater than 5
OR 'CD4 T Cells Naive 0 vs 150' RPKM2 greater than 5)

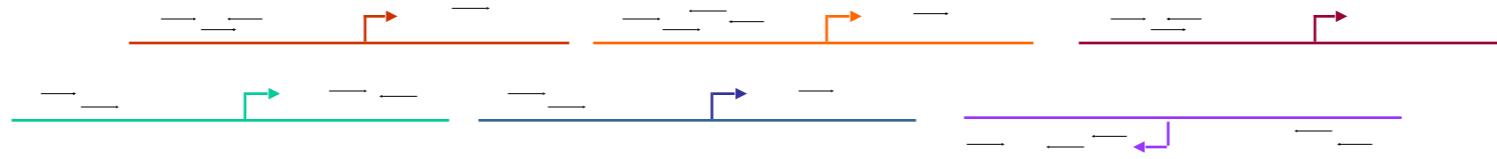
Set

DESeq/DESeq2

(Differential Gene Expression Analysis)



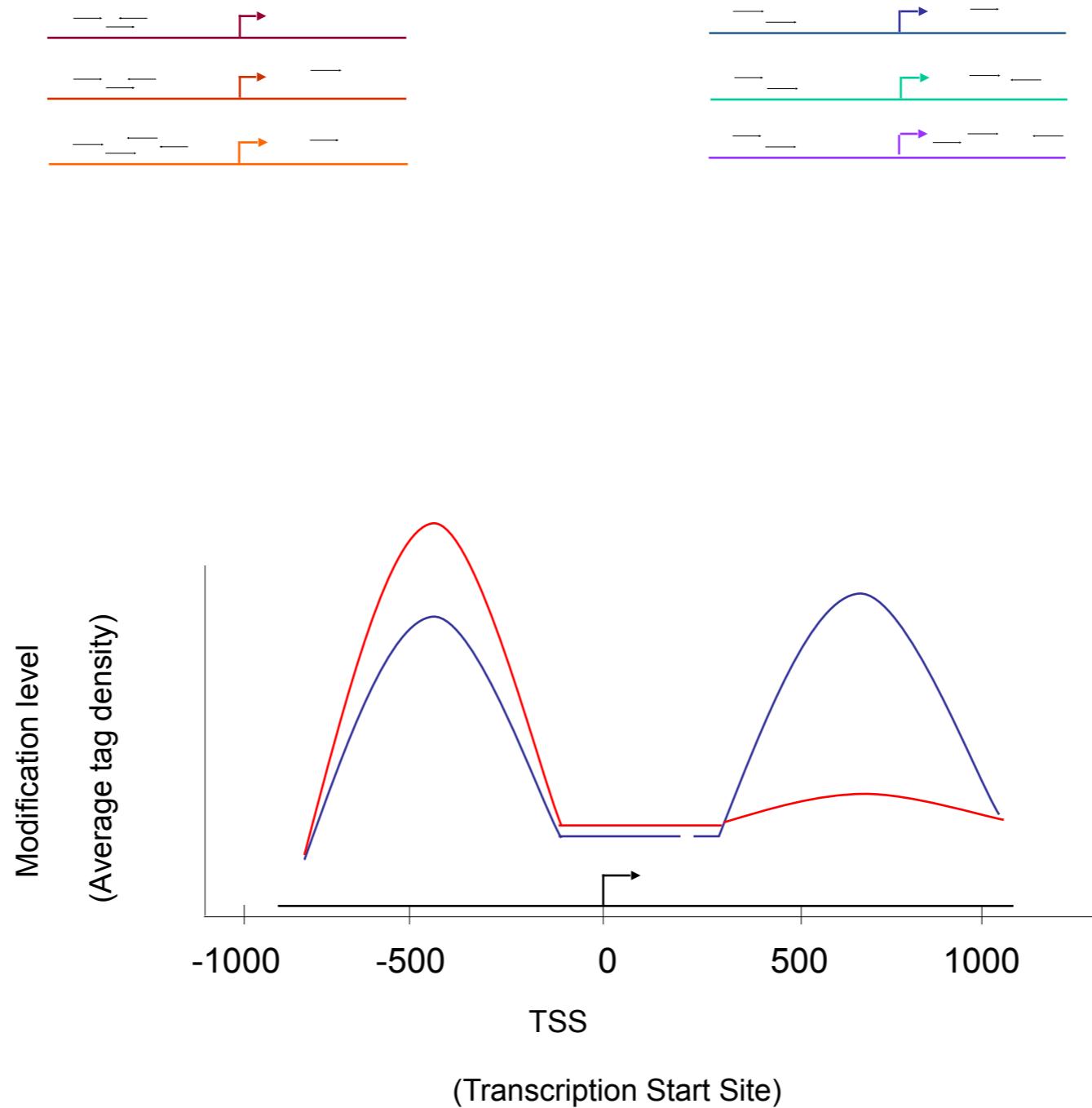
Average tag density profiles are used to study groups of genes



Average tag density profiles are used to study groups of genes

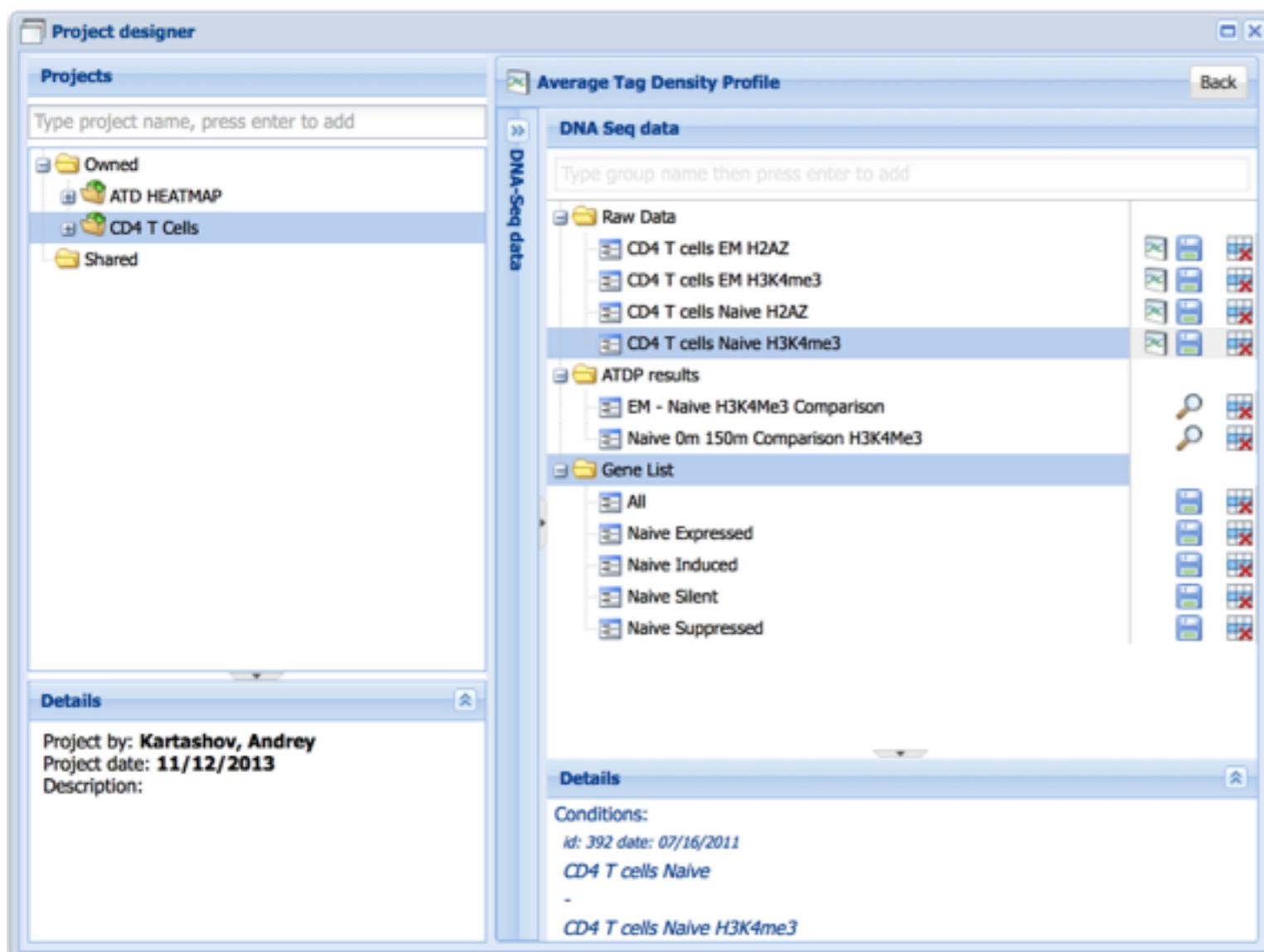


Average tag density profiles are used to study groups of genes



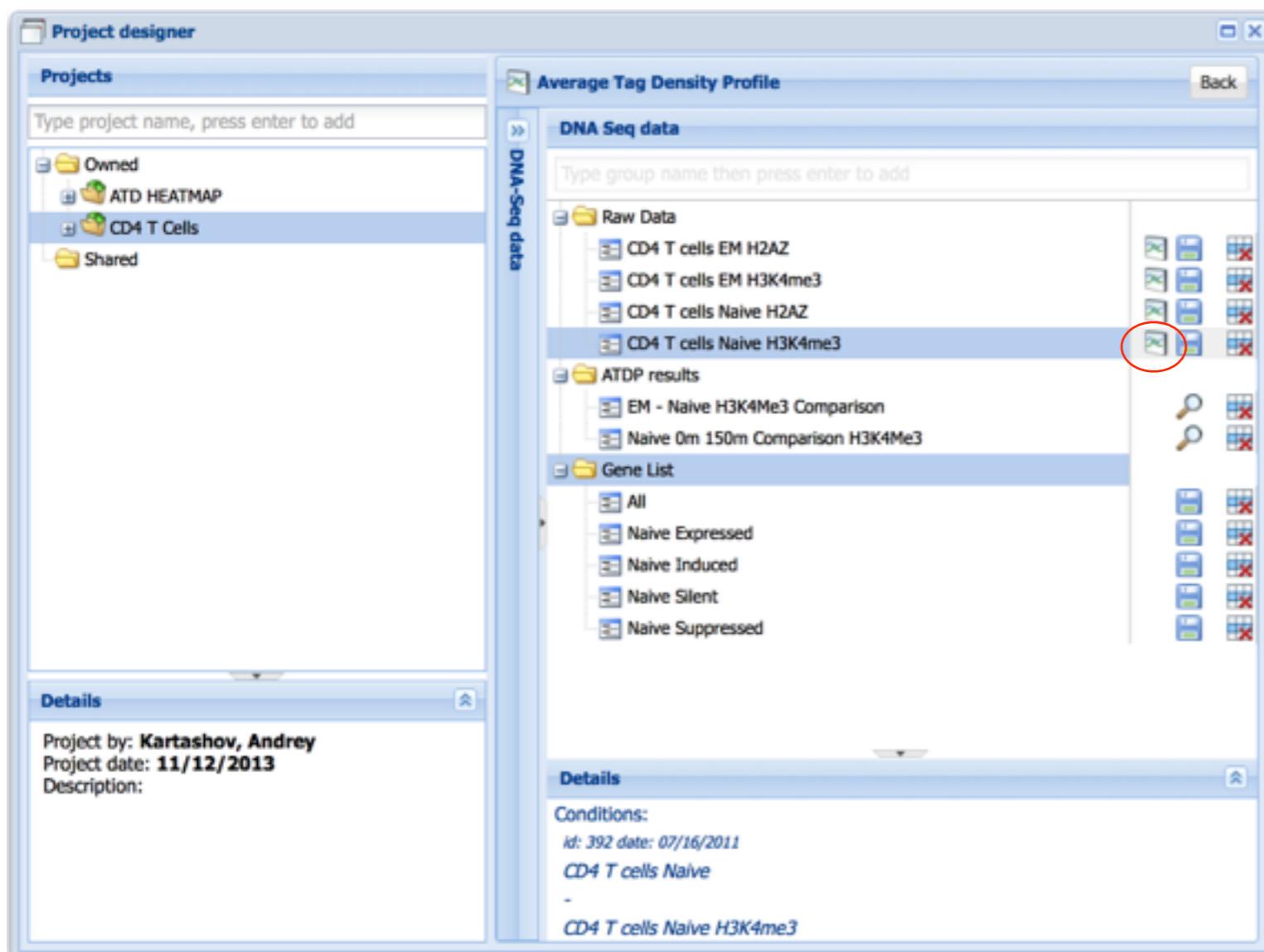
Average Tag Density Profile

(Average Tag Density Profiles are used to study groups of genes)



Average Tag Density Profile

(Average Tag Density Profiles are used to study groups of genes)



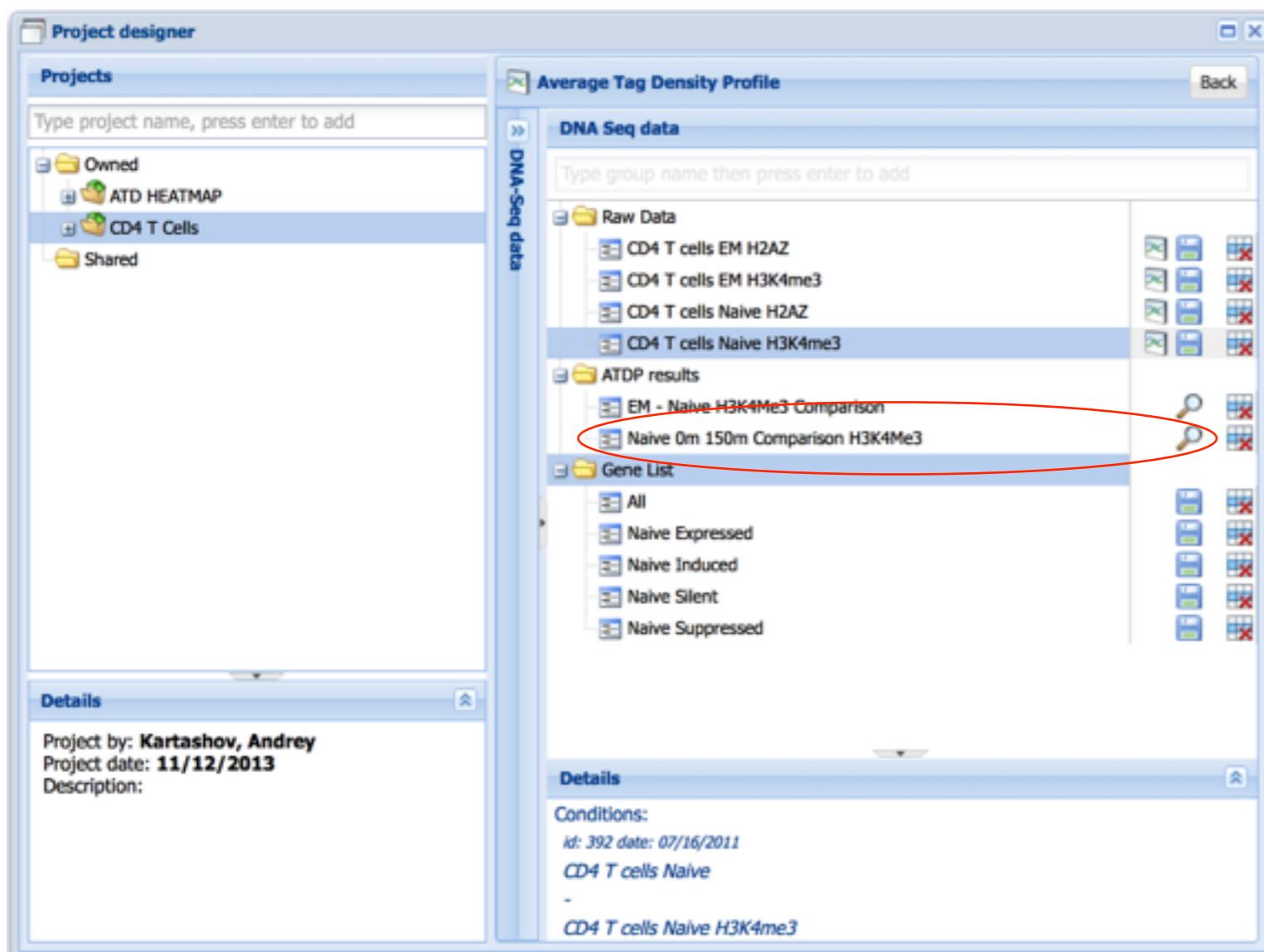
Average Tag Density Profile

(Average Tag Density Profiles are used to study groups of genes)

The screenshot shows a software application window titled "Project designer". On the left, there's a sidebar with "Projects" and "DNA-Seq data" sections. Under "DNA-Seq data", the "ATD HEATMAP" project is selected. In the main area, a sub-project "Average Tag Density Profile" is open, showing "DNA Seq data" with a tree view of "Raw Data", "ATDP results", and "Gene List". The "Gene List" node is expanded, showing categories like "All", "Naive Expressed", etc. A modal dialog box titled "Average Tag Density settings" is displayed in the center. It contains an informational message about creating ATDP figures, a text input field for the name ("Naive 0m 150m Comparison H3K4Me3"), and a grid for selecting ATDP inputs. The grid has three columns: "Naive All", "CD4 T cells Naive H3K4me3", and "All". Each row corresponds to a category from the gene list: "Naive Silent", "Naive Expressed", "Naive Induced", and "Naive Suppressed". Each row has a "Run" button at the bottom right. At the bottom of the dialog, there's a "Run" button.

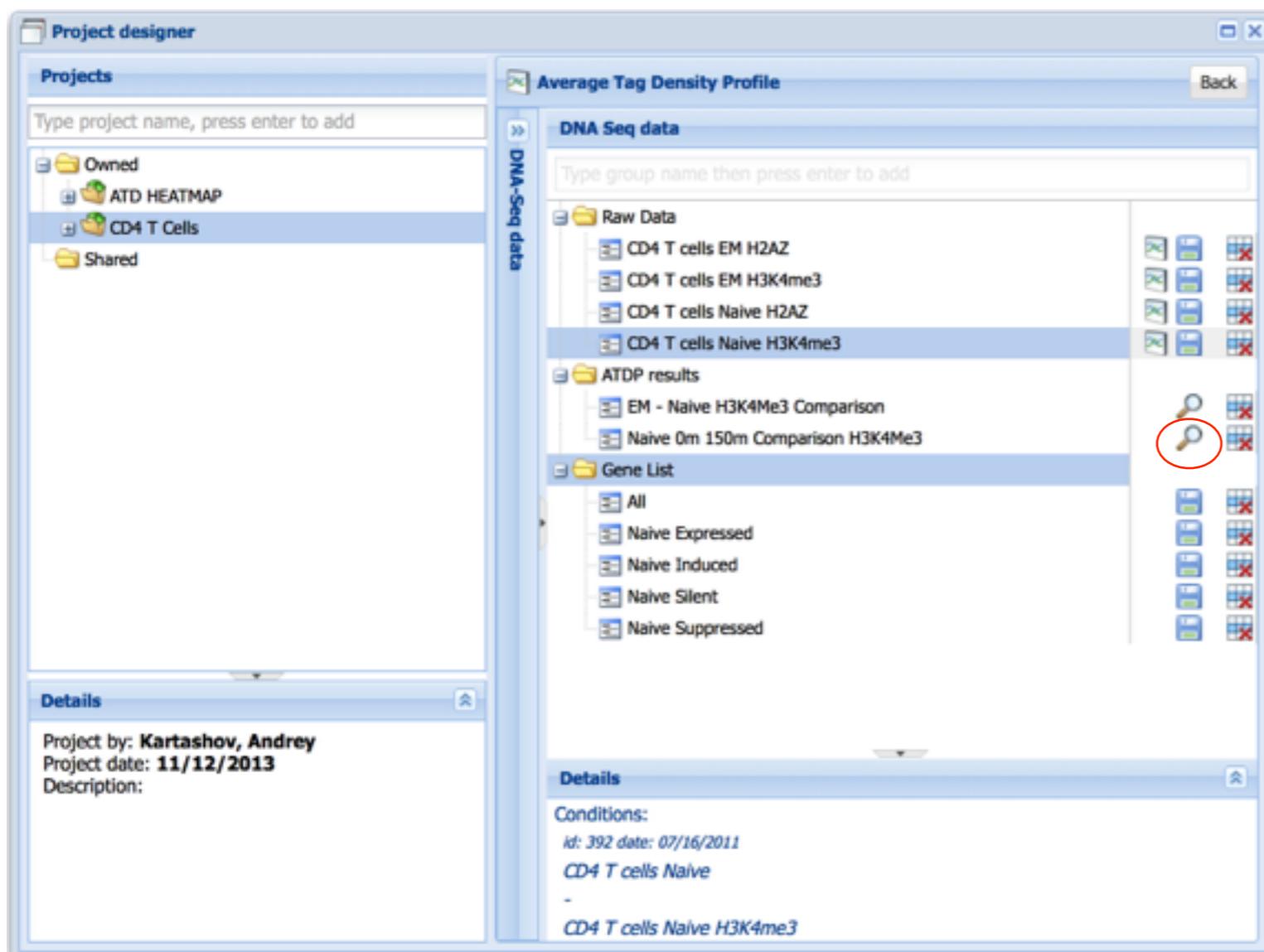
Average Tag Density Profile

(Average Tag Density Profiles are used to study groups of genes)



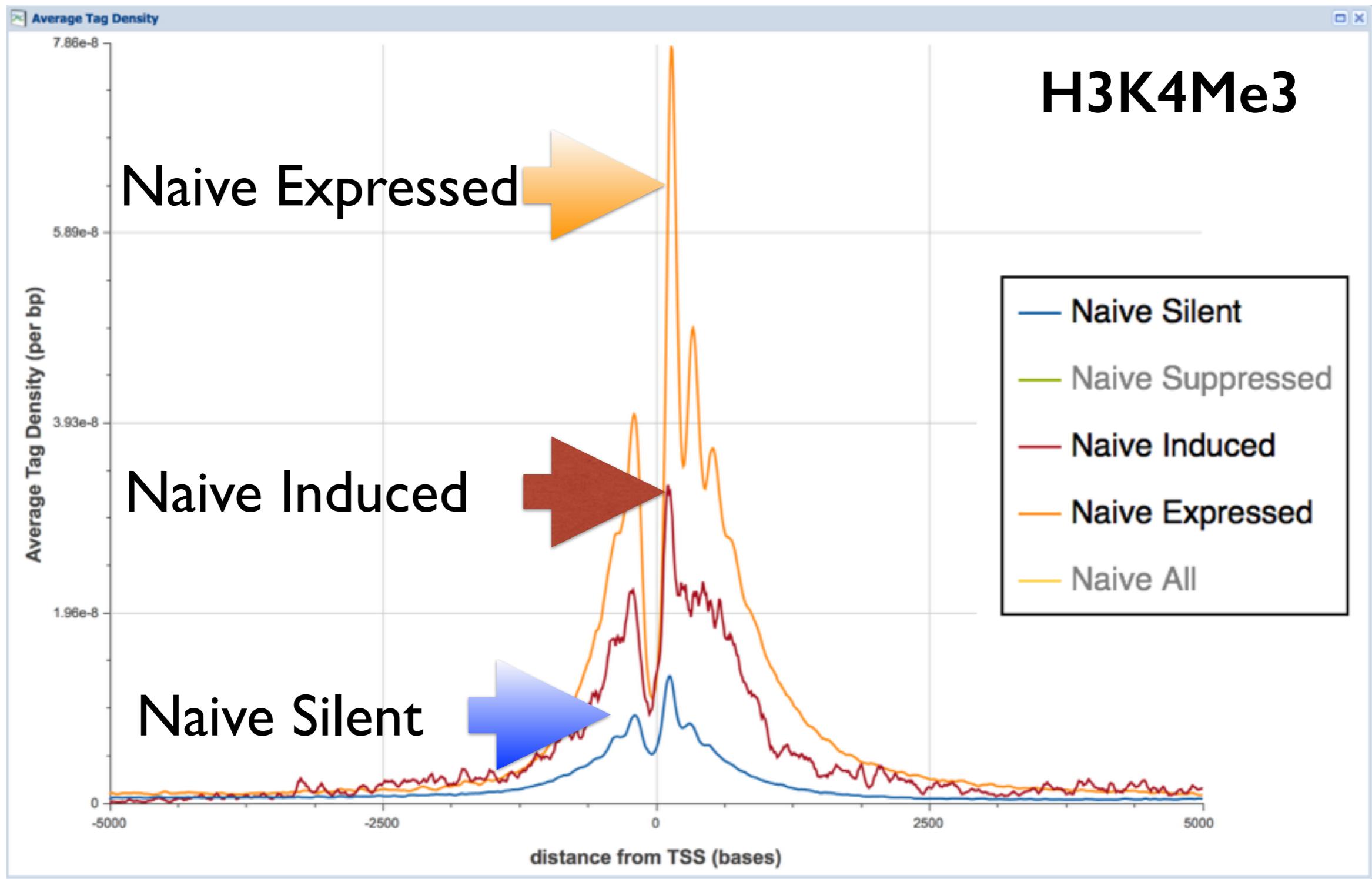
Average Tag Density Profile

(Average Tag Density Profiles are used to study groups of genes)



Average Tag Density Profile

(Average Tag Density Profiles are used to study groups of genes)



**To compare ChIP-Seq experiments
with the same modification we use
MANorm**

MANorm

The screenshot displays the MANorm software interface, which includes two main windows: "Project designer" and "DNA Seq data".

Project designer (Left Window):

- Projects:** A search bar labeled "Type project name, press enter to add".
- Owned:** A folder containing "ATD HEATMAP" and "CD4 T Cells".
- Shared:** An empty folder.

DNA Seq data (Right Window):

- MANorm:** A title bar with a "Back" button.
- DNA Seq data:** A search bar labeled "Type group name then press enter to add".
- Raw Data:** A folder containing "CD4 T cells EM H2AZ", "CD4 T cells EM H3K4me3", "CD4 T cells Naive H2AZ", and "CD4 T cells Naive H3K4me3".
- MANorm results:** A folder containing "EM vs Naive Islands".

Details (Bottom Left):

Project by: **Kartashov, Andrey**
Project date: **11/12/2013**
Description:

Details (Bottom Right):

Conditions:
MANorm were used for analysis .
Data from 'CD4 T cells EM H3K4me3' and 'CD4 T cells Naive H3K4me3' has been compared.

MANorm

View table data

Page 1 of 33 | Displaying 1 - 100 of 3265 | jump

chrom	start	end	description	raw_read1	raw_read2	M_value_rescaled	A_value_rescaled	log10_p_value	
chr15	70850864	70851764	merged_commo...	18	608	-5.0341625213...	6.7332172393799	152.06977844238	
chr19	8641148	8642301	merged_commo...	592	71	2.9776721000671	7.6587610244751	95.603507995605	
chr14	52734423	52735852	merged_commo...	433	37	3.45228099823	6.9740681648254	81.021423339844	
chr10	51501790	51502742	unique_peak2	16	314	-4.242486000061	6.1779651641846	70.83243560791	
chr12	6554077	6555005	merged_commo...	63	442	-2.8344254493...	7.3739500045776	70.400085449219	
chr20	1665867	1666685	unique_peak2	5	258	-5.4527497291...	5.2904334068298	67.336303710938	
chr2	100758267	100759097	merged_commo...	25	319	-3.656242609024	6.4938068389893	64.936973571777	
chr20	24929770	24931051	merged_commo...	258	7	4.9603414535522	5.4801707267761	63.424602508545	
chr13	100067035	100067608	unique_peak2	0	189	-7.57383537292...	3.7829377651215	54.784824371338	
chr19	6668725	6669258	unique_peak1	217	5	5.1283826828003	5.1491537094116	54.70911026001	
chr19	6668725	6669258	unique_peak1	217	5	5.1283826828003	5.1491537094116	54.70911026001	
chr4	6691267	6692235	unique_peak1	187	1	6.501148223877	4.2505741119385	52.395317077637	
chr2	75643547	75644467	unique_peak2	0	180	-7.50382566452...	3.7479329109192	52.096687316895	
chr2	75643547	75644467	unique_peak2	0	180	-7.50382566452...	3.7479329109192	52.096687316895	
chr2	75643547	75644467	unique_peak2	0	180	-7.50382566452...	3.7479329109192	52.096687316895	
chr19	51875246	51876115	unique_peak1	191	2	5.9463605880737	4.5581426620483	51.788089752197	
chr8	79716585	79717910	merged_commo...	408	75	2.3672471046448	7.4315509796143	51.5087890625	
chr1	25889952	25890728	unique_peak2	25	258	-3.3511228561...	6.3412470817566	48.783340454102	
chr1	25889952	25890728	unique_peak2	25	258	-3.3511228561...	6.3412470817566	48.783340454102	
chr9	139869720	139870746	unique_peak1	178	2	5.8458762168884	4.5079007148743	48.233657836914	
chr9	139869720	139870746	unique_peak1	178	2	5.8458762168884	4.5079007148743	48.233657836914	
chr19	4301693	4302614	unique_peak2	6	192	-4.8074622154...	5.1887259483337	47.043758392334	
chr19	4301693	4302614	unique_peak2	6	192	-4.8074622154...	5.1887259483337	47.043758392334	
chr17	2698896	2700668	merged_commo...	849	327	1.3060750961304	9.0105895996094	46.456298828125	
chr4	22516606	22517279	unique_peak2	3	170	-5.4349265098...	4.7003893852234	44.409236907959	
chr4	22516606	22517279	unique_peak2	3	170	-5.4349265098...	4.7003893852234	44.409236907959	
chr4	22516606	22517279	unique_peak2	3	170	-5.4349265098...	4.7003893852234	44.409236907959	

Close

MANorm

