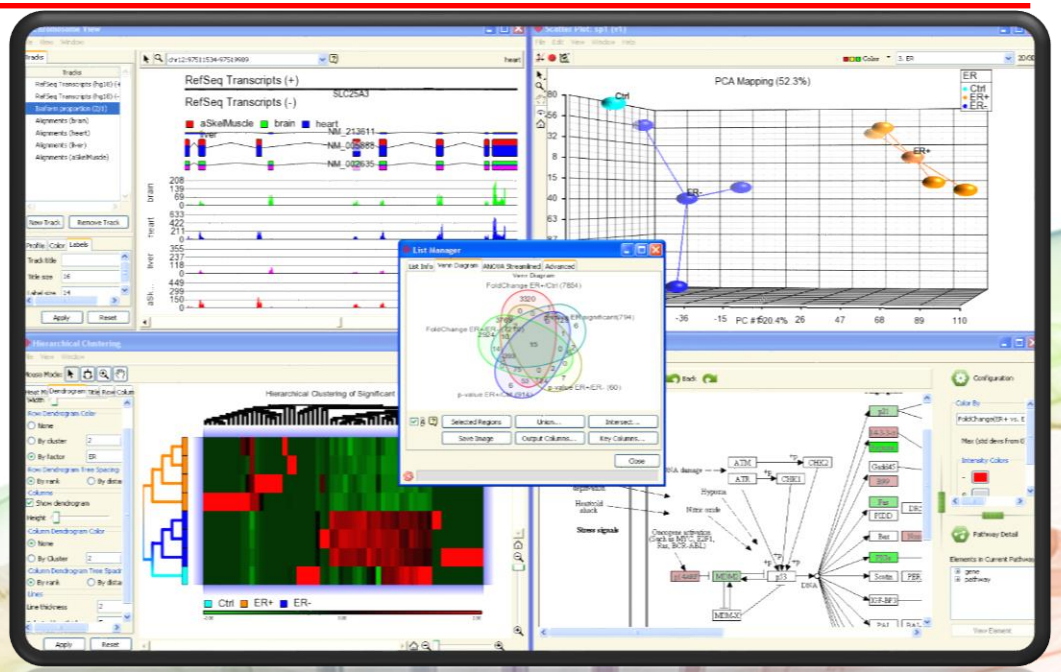
Data Analysis Using Partek® Software Packages

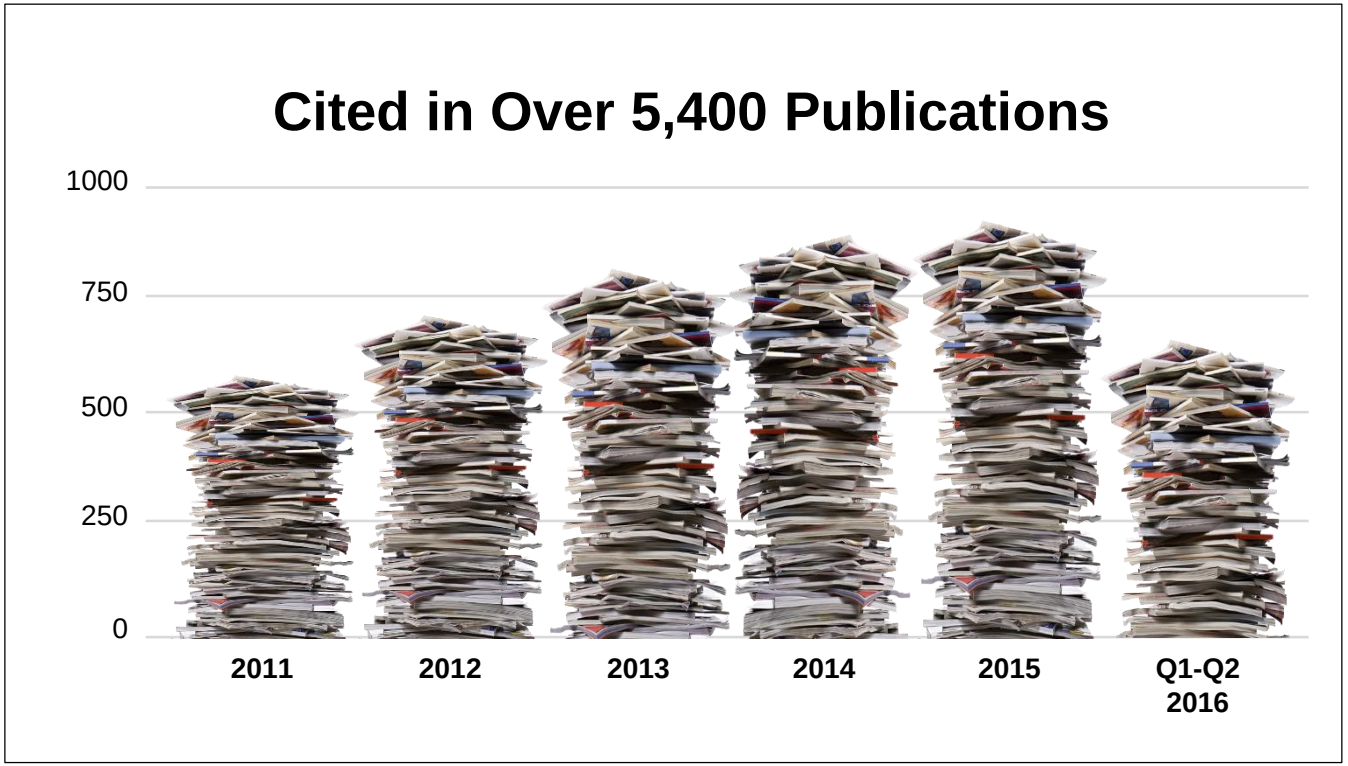
Eric Seiser, PhD
Field Application Scientist
Partek Inc.



Partek®
turning data into discovery®

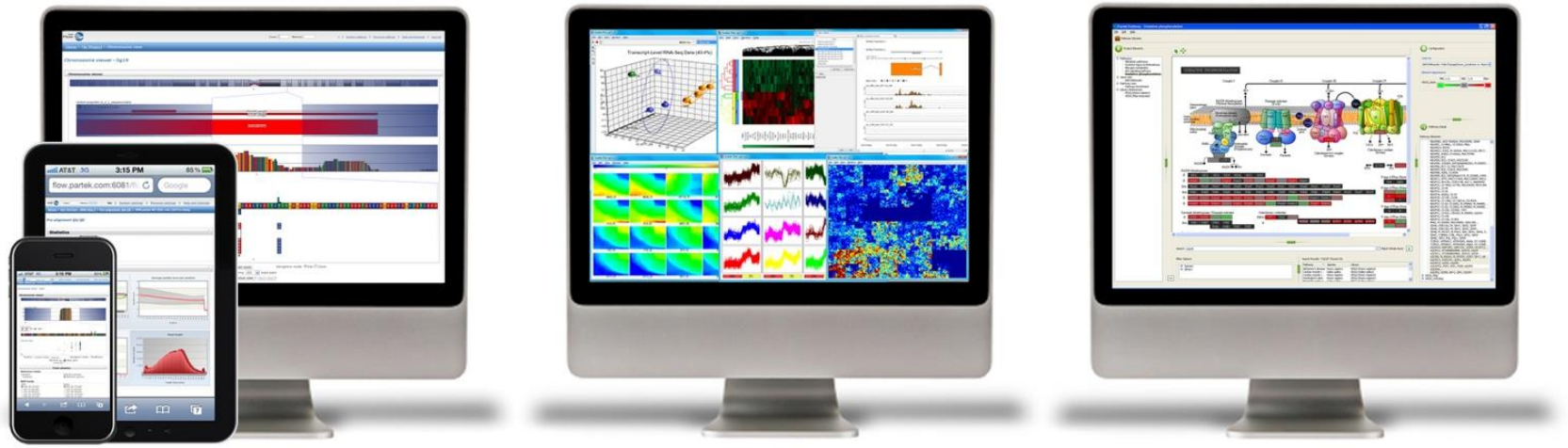
Statistics and Visualization Software Researchers Trust

- Science
- Nature
- Cell
- PNAS
- NEJM
- JCI
- Nature Medicine



Partek®

Comprehensive Solution for Data Analysis



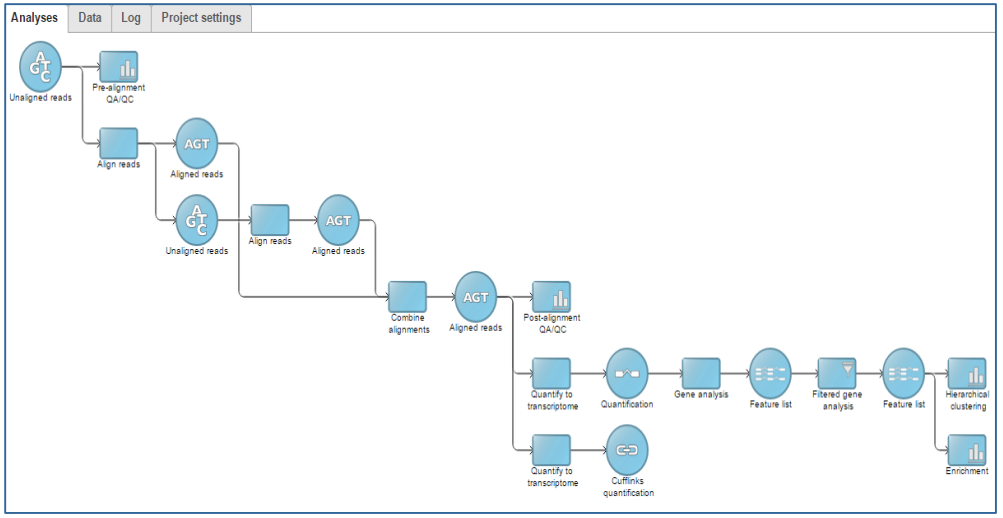
Partek Flow | Partek Genomics Suite | Partek Pathway

Raw Data → QA/QC → Normalization → Statistics → Visualization → Genomic Integration → Biological Interpretation

Start to Finish Analysis for NGS, Microarray and Other Platforms

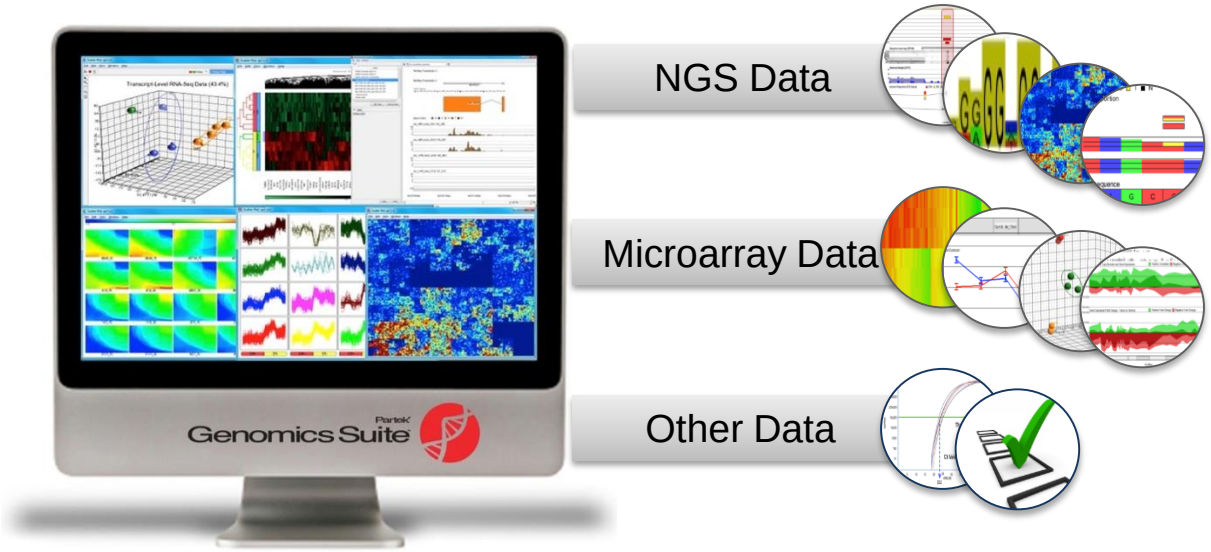
Partek® Flow®

- Web based application
- Flexible data storage
- Visual analysis pipelines
- Guidance on next analysis steps
- Broad choice of public domain tools
- Comprehensive statistics and visualization



Partek ® Genomics Suite™

- Guided workflows for major assays and platforms
- Flexible spreadsheet format for any tabular data
- Tools for exploratory analysis and inferential statistics
- Comprehensive visualizations
- Integration of different omics data



Desktop software on Windows, Linux, Mac

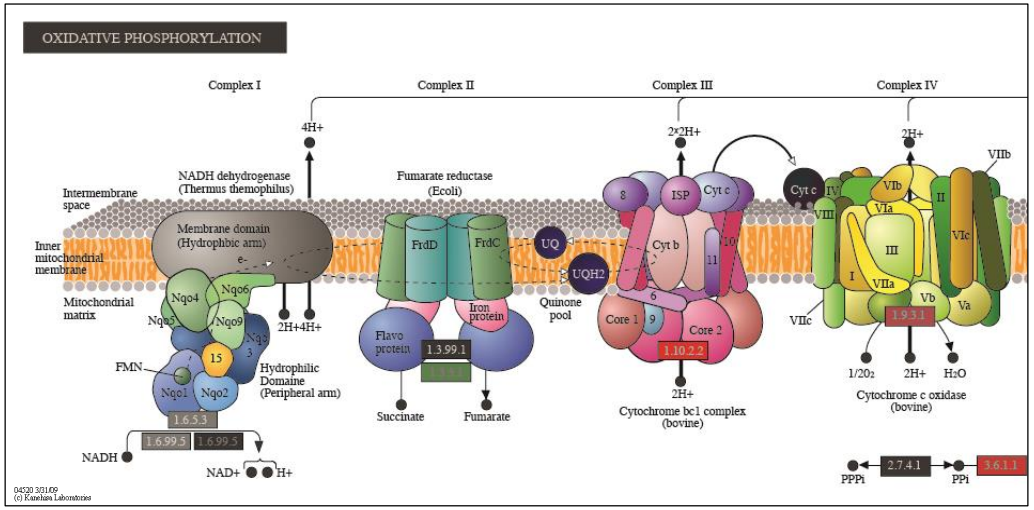
Partek® Pathway™

- Seamlessly integrated with Genomics Suite
- Support 2000+ species in KEGG database
- Find enriched pathways
- Detect differentially expressed pathways
- Visualize gene relationships

Partek® Pathway™

KEGG Organisms: Complete Genomes

Eukaryotes: 333 Bacteria: 3746 Archaea: 229

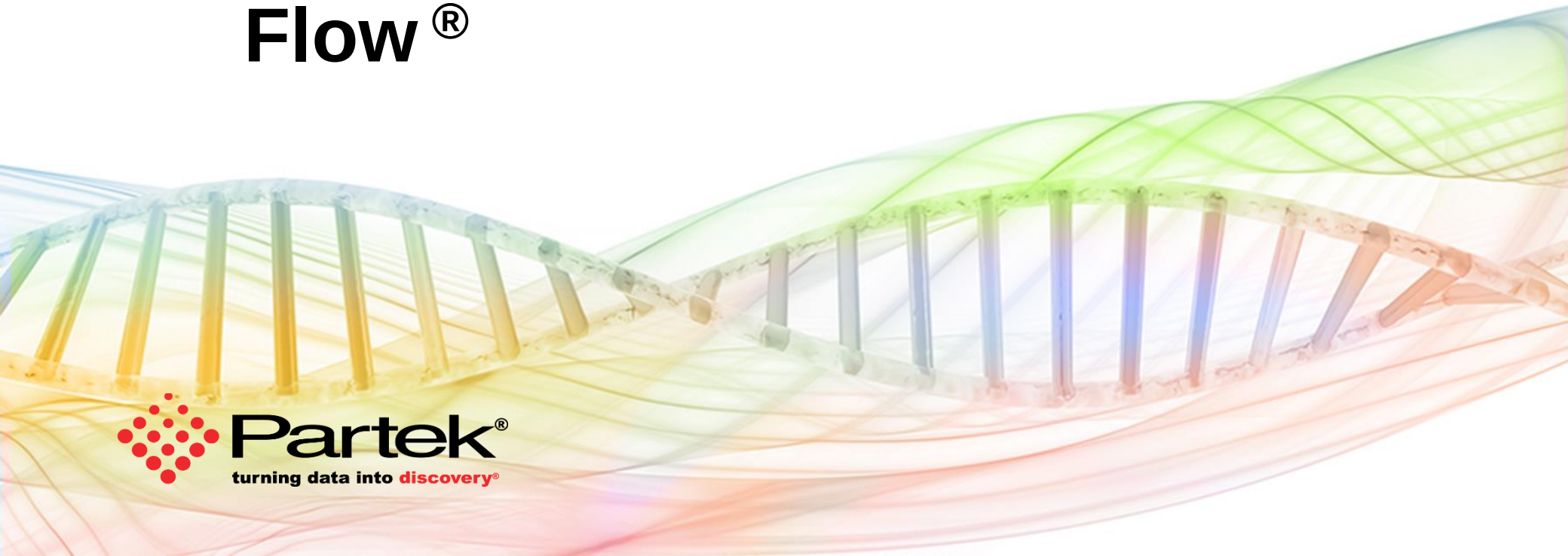


Compatible with All Major Genomics Formats and Assays

RNA | Noncoding RNA | DNA | ChIP | Methylation | Copy Number

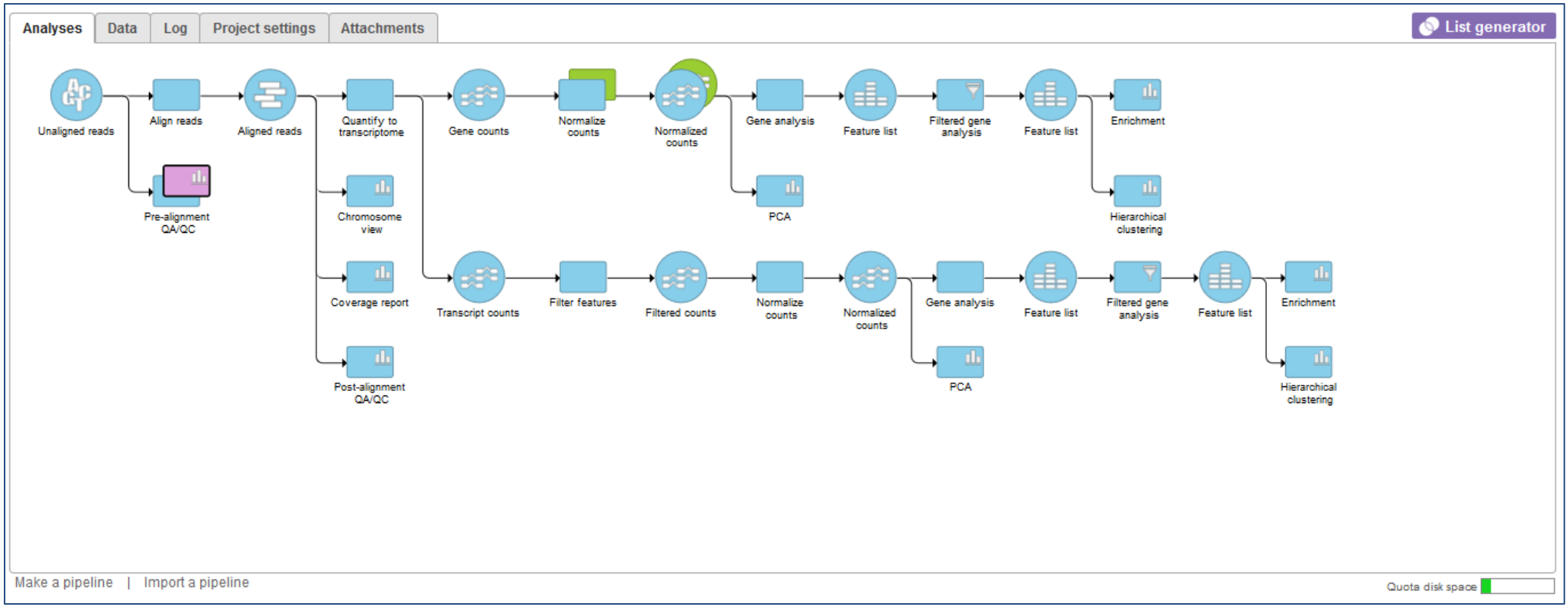
Microarray | Next Generation Sequencing | qPCR

NGS Data Analysis in Partek® Flow ®



Partek®
turning data into discovery®

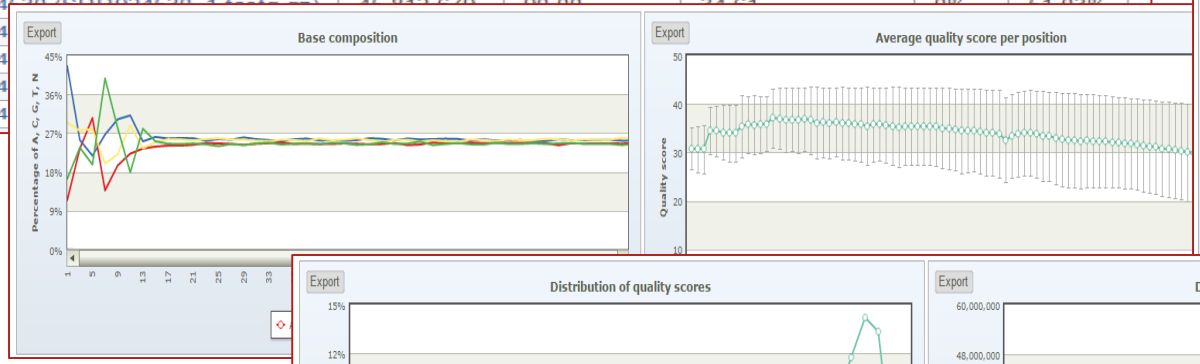
RNA-seq pipeline in Partek® Flow ®

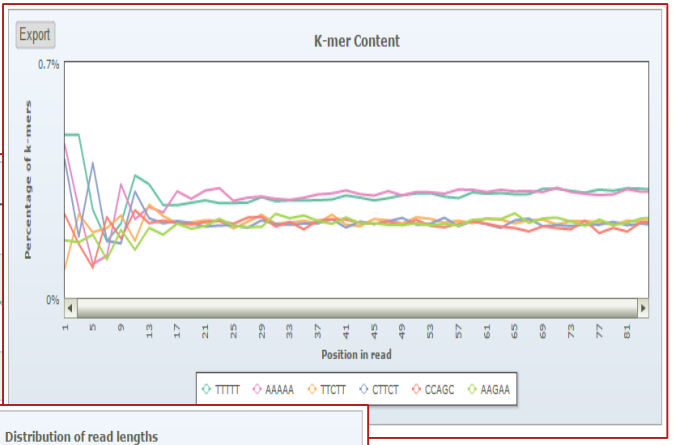


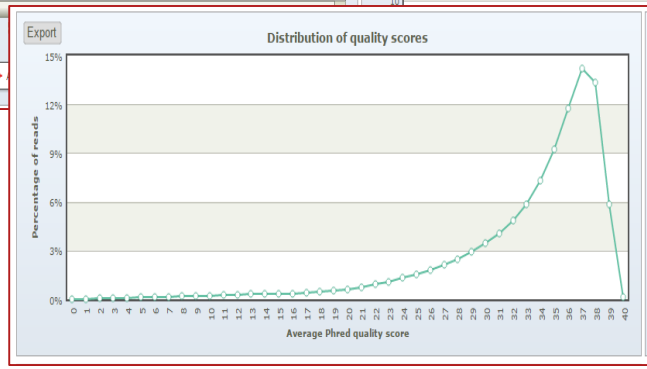
Pre-alignment QA/QC

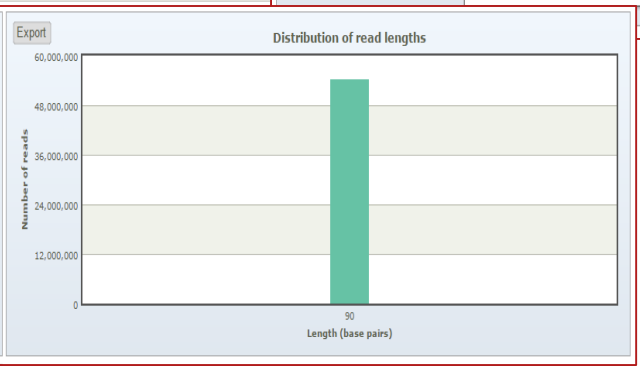
- Quality control of raw sequencing data is essential to identify technical issues and to ensure high quality sequence is present for downstream analysis.

Sample name	Total reads	Read length	Avg. read quality	% N	% GC
SRR924146 (SRR924146_1.fastq.gz)	54,255,978	90.00	34.79	0%	51.11%
SRR924146 (SRR924146_2.fastq.gz)	54,255,978	90.00	33.46	0%	51.19%
SRR924528 (SRR924528_1.fastq.gz)	47,255,516	90.00	35.18	0%	51.67%
SRR924528 (SRR924528_2.fastq.gz)	47,255,516	90.00	33.24	.01%	51.80%
SRR924529 (SRR924529_1.fastq.gz)	51,972,705	90.00	34.82	0%	50.54%
SRR924529 (SRR924529_2.fastq.gz)	51,972,705	90.00	32.99	.01%	50.67%





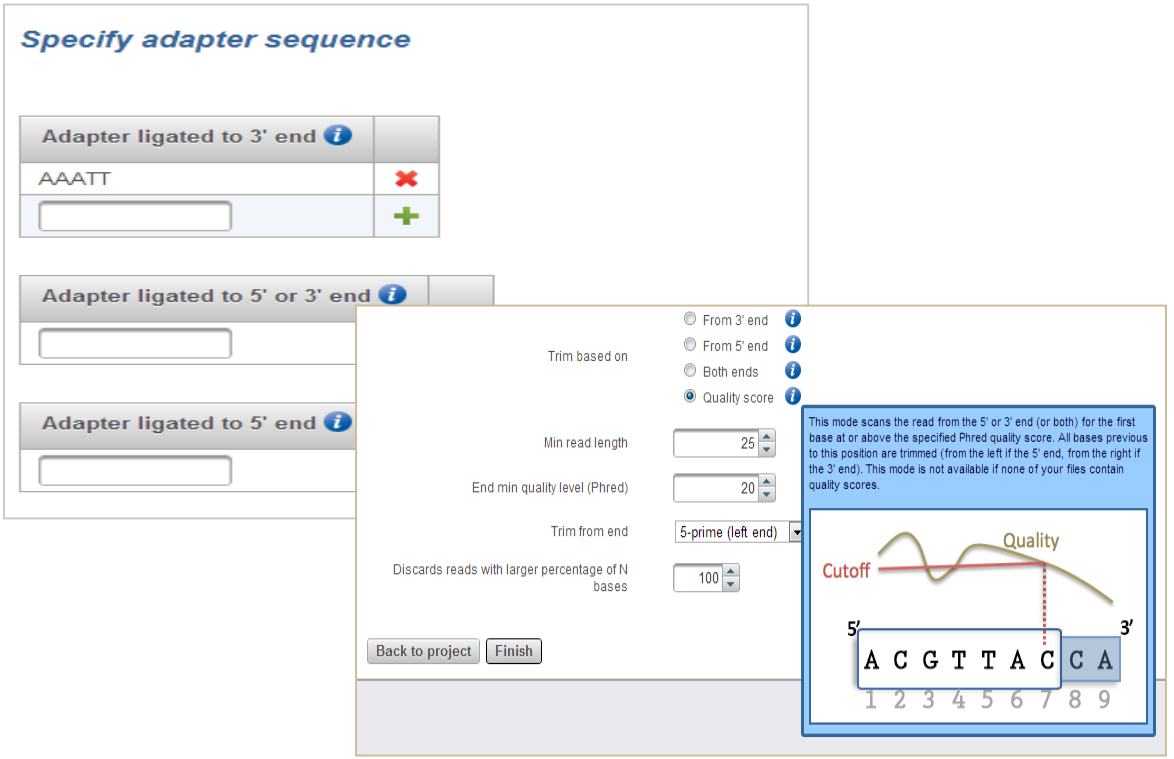




Pre-alignment Processing

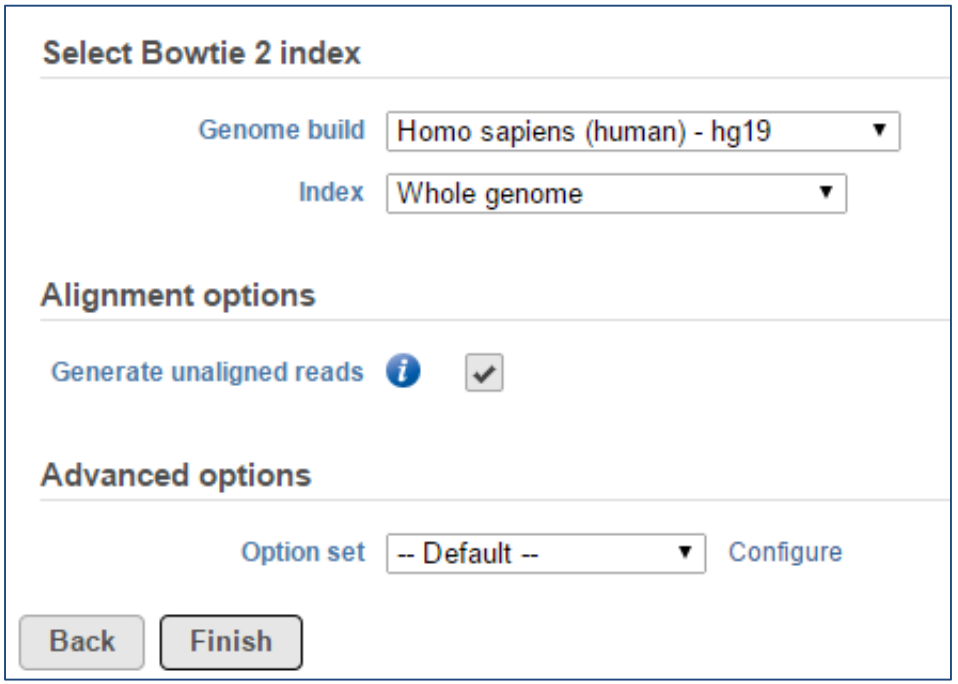
- Processing of the fastq files allows for the removal of non-biological sequence and low quality sequence.

  - Trim based on 3’ or 5’ end
  - Trim both end
  - Trim based on quality score
  - Trim adapter--cutadapt



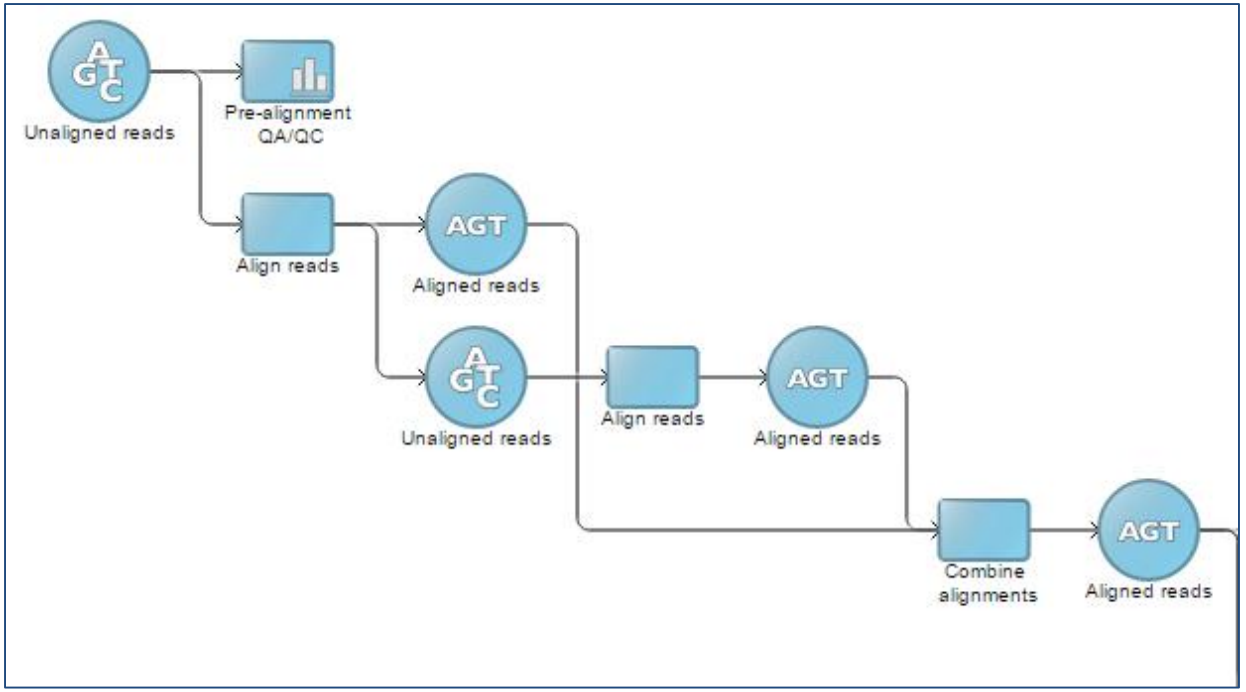
Alignment

- The process of alignment is used to map all of these reads to a reference sequence, providing information with regards to the start and stop positions of each read within the reference sequence as well as metrics for the quality of the mapping.



Multiple Stage Alignment

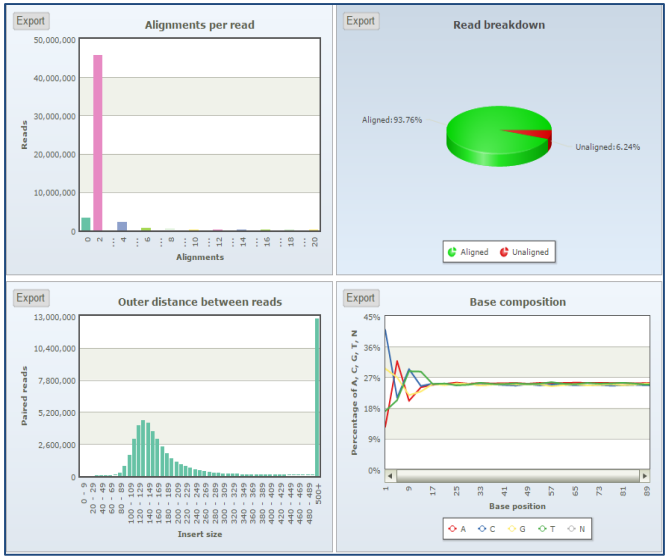
- Combine multiple alignments:
  - increase the alignment rate
- Align to different references
  - remove contamination



Post-alignment QA/QC

- Quality control of aligned sequencing data is necessary to ensure that mapping to a reference sequence was successful.

Sample name	Total reads	Total alignments	Aligned	Unique	Paired	Coverage	Avg. coverage depth	Avg. length	Avg. quality	Avg. mapping quality	%GC
SRR924146	54,255,978	109,016,432	92.00%	0%	85.75%	6.09%	51.23 (SD 1,126.83)	90.00	34.51	2.22 (SD 1.06)	50.05%
SRR924528	47,255,516	94,445,222	91.51%	0%	85.27%	5.22%	51.73 (SD 1,043.86)	90.00	34.65	2.23 (SD 1.06)	50.57%
SRR924529	51,972,705	105,400,250	93.76%	0%	88.14%	5.76%	52.40 (SD 944.29)	90.00	34.23	2.20 (SD 1.07)	49.83%
SRR924530	45,812,679	91,571,018	92.36%	0%	86.81%	5.49%	47.65 (SD 834.12)	90.00	33.85	2.20 (SD 1.07)	50.26%
SRR924531	50,046,772	100,212,304	92.78%	0%	87.32%	5.23%	54.84 (SD 966.15)	90.00	34.26	2.23 (SD 1.05)	50.22%
SRR924532	49,222,642	98,345,686	92.71%	0%	87.40%	5.76%	48.90 (SD 826.42)	90.00	34.25	2.22 (SD 1.06)	49.44%

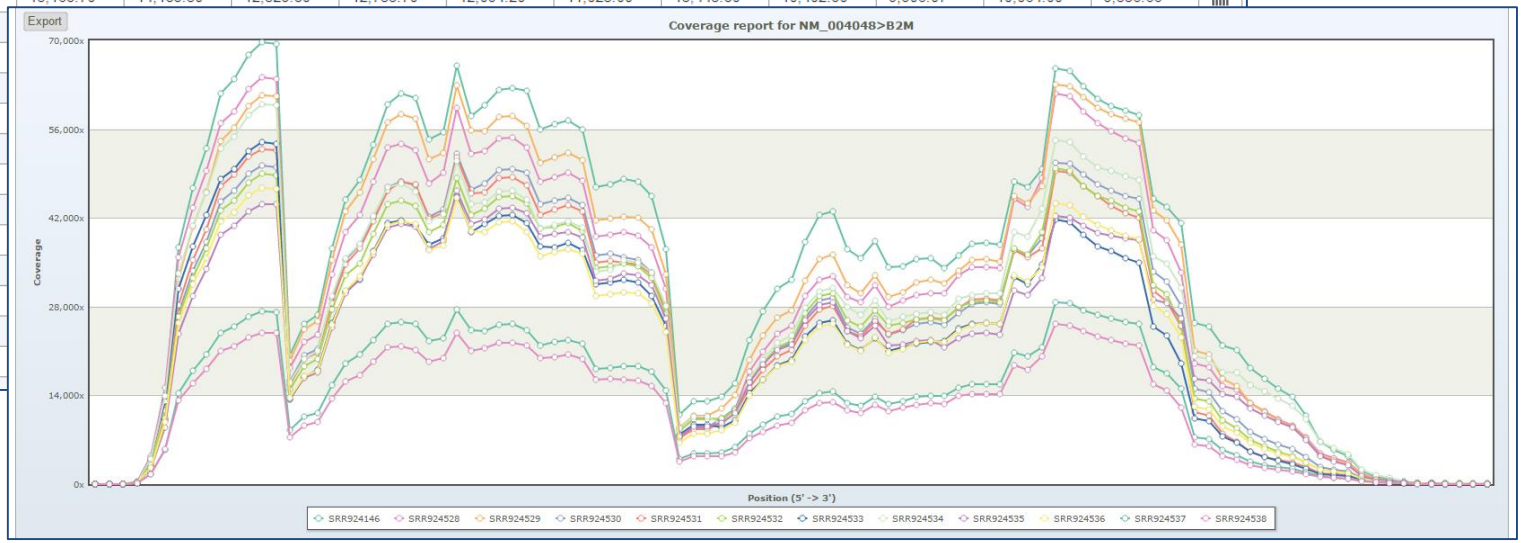


Coverage Report

- The coverage reports allows for assessment of how much of your regions of interest are represented in the sequencing data.

Sample name	1x	20x	30x	50x	100x	Average coverage	Average quality
SRR924528	59.82	29.00	22.70	15.06	7.14	37.00	34.67
SRR924146	60.96	31.42	25.17	17.21	8.46	42.38	34.56
SRR924529							
SRR924530							
SRR924532							
SRR924531							
SRR924533							

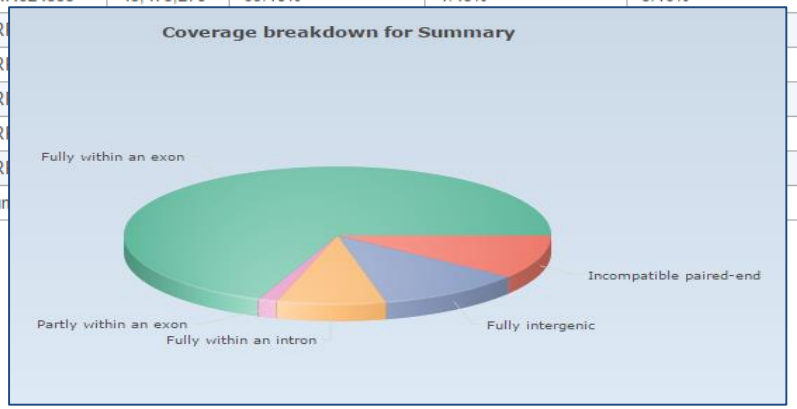
Region name	SRR924146	SRR924528	SRR924529	SRR924530	SRR924531	SRR924532	SRR924533	SRR924534	SRR924535	SRR924536	SRR924537	SRR924538	
NM_004048>B2M	32,868.30	28,542.80	29,790.70	24,287.40	23,614.70	23,325.90	21,309.00	25,891.00	21,839.70	21,005.80	12,841.80	11,300.30	
NM_001402>EEF1A1	20,014.00	17,384.10	16,437.90	14,624.50	14,865.70	14,665.60	12,944.70	15,391.90	13,667.60	12,902.60	11,775.20	10,736.80	
NM_001010>RPS6	16,852.00	15,168.70	14,189.50	12,529.80	12,753.70	12,094.20	11,025.00	13,445.80	10,402.90	9,906.07	10,064.00	9,536.66	
NM_213725>RPLP1	16,031.20												
NM_021104>RPL41	15,523.30												
NM_001003>RPLP1	15,121.50												
NM_005514>HLA-B	13,137.80												
NM_021109>TMSB4X	12,584.40												
NM_001035267>RPL41	12,479.30												
NM_022551.5>RPS18	10,392.00												
NM_001199802>RPL11	10,203.50												
NM_000975>RPL11	10,181.10												
NM_000967>RPL3	9,965.70												
NM_001033853>RPL3	9,843.11												
NM_005617>RPS14	9,792.04												

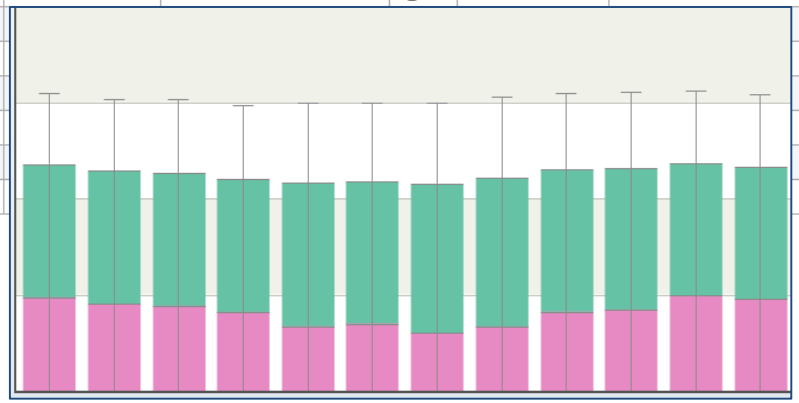


Quantification

- Quantification is the process of estimating gene abundance based upon mapped reads in relation to position of genes/transcripts.
- Flow provides options for an E/M algorithm or Cufflinks.

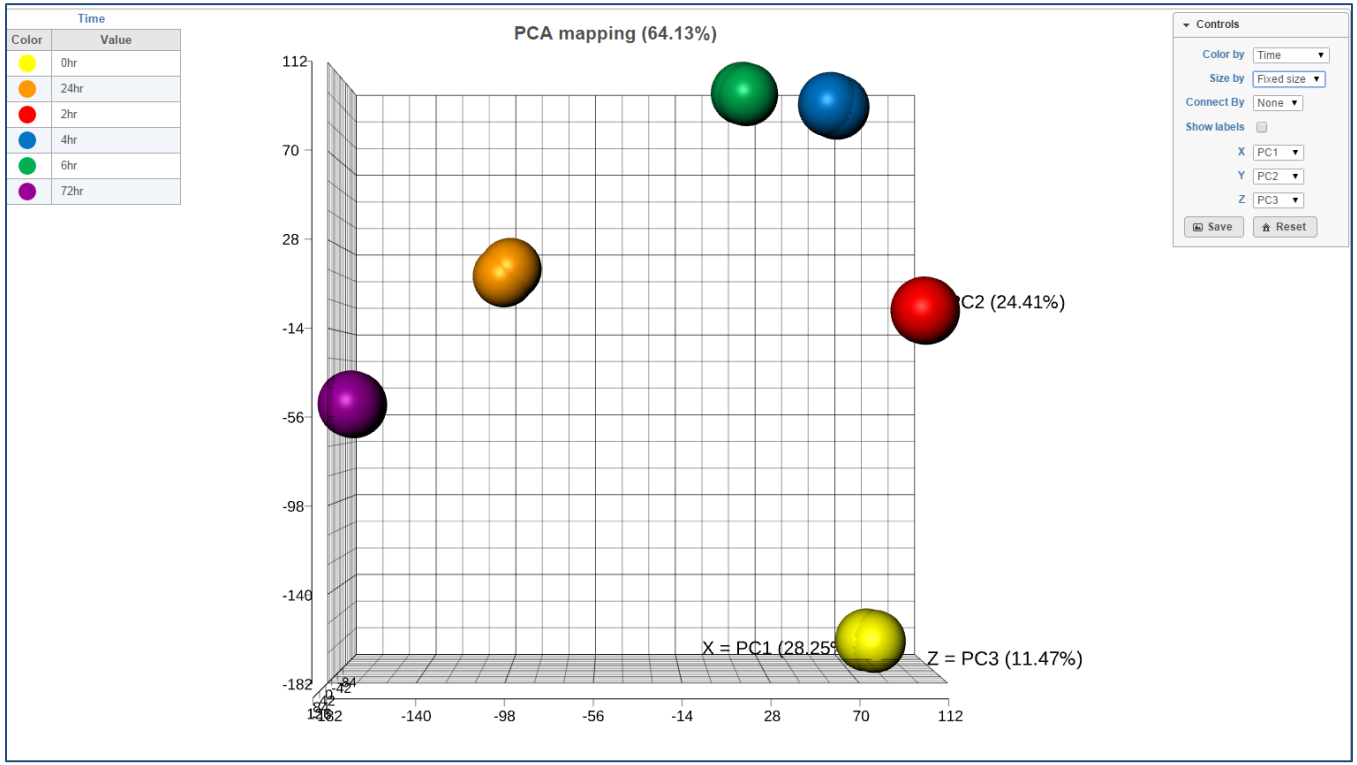
Sample name	Total reads	Fully within an exon	Partly within an exon	Fully within an intron	Fully intergenic	Incompatible paired-end	View	Total junctions	Compatible junctions
SRR924146	49,916,169	64.62%	1.44%	10.28%	14.09%	9.57%		15,289,673	12,059,280
SRR924528	43,245,716	65.17%	1.44%	9.85%	13.74%	9.80%		13,567,216	10,726,269
SRR924529	48,727,334	68.14%	1.49%	9.17%	11.67%	9.53%		14,514,990	11,419,235
SRR924530	42,312,362	67.76%	1.55%	9.16%	11.74%	9.79%		12,507,952	9,799,789
SRR924531	46,433,309	67.66%	1.42%	9.25%	11.77%	9.90%		14,469,011	11,437,153
SRR924532	45,634,196	68.10%	1.51%	9.39%	11.21%	9.79%		14,038,887	11,122,241
SRR924533	43,475,270	69.10%	1.48%	8.16%	10.66%	10.60%		14,393,536	11,514,878





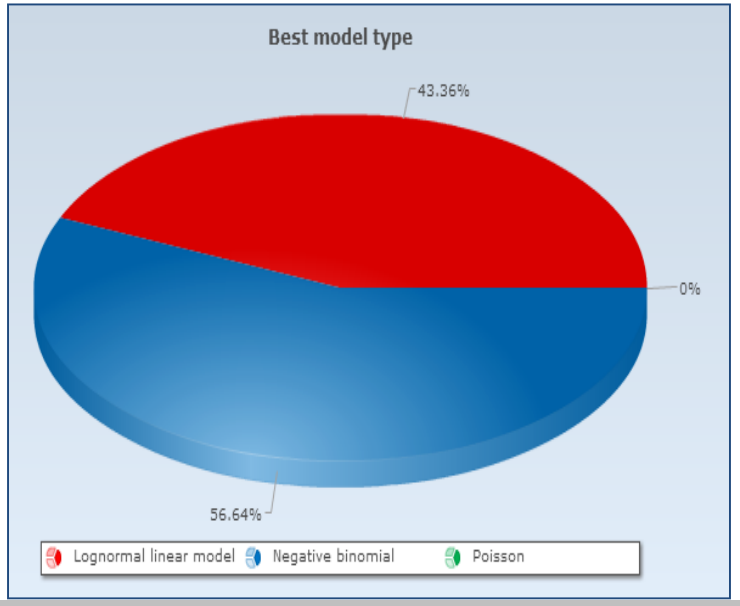
PCA

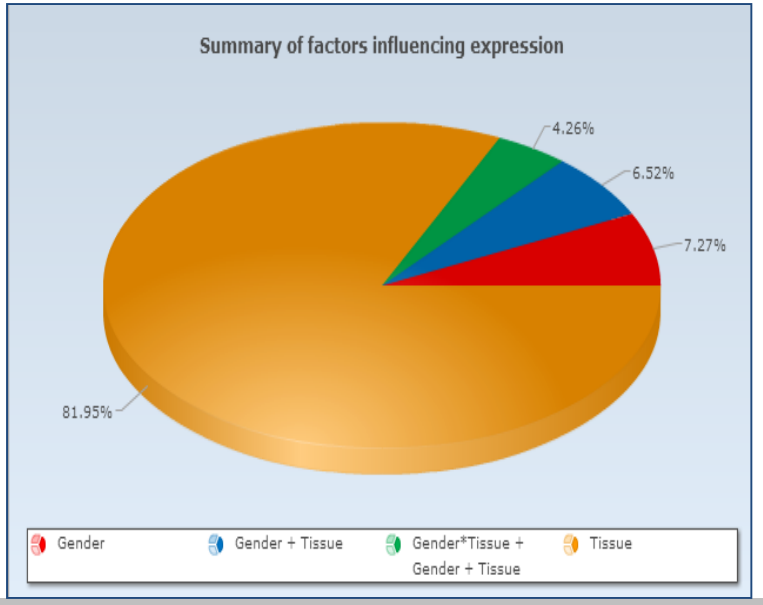
- Principal component analysis is an exploratory technique that uses dimensional reduction to capture the primary sources of variability in the data.



Differential Expression Detection

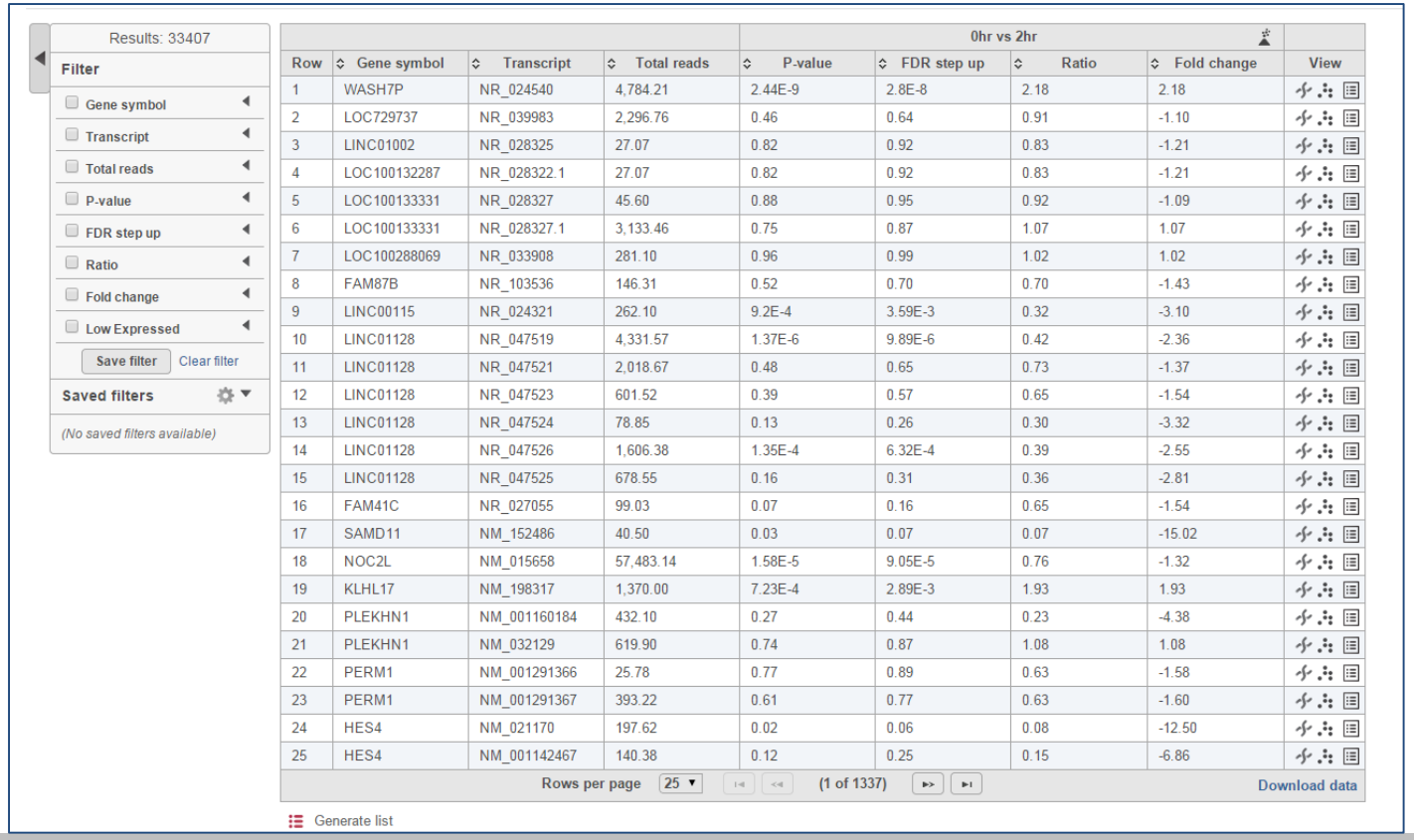
- Differential expression analysis provides a means to statistically identify what genes/transcripts differ between groups.
- Flow provide multiple options for this analysis: gene specific analysis (GSA), a mixed model ANOVA, and Cuffdiff.
  - The GSA identifies a statistical model that is the best for a specific gene and use the best model to test for differential expression for each gene independently.





Generation of the Feature List

- The feature list provides the associated stats for each gene/transcript in the analysis and allows for interactive filtering to find what is significant.



Results: 33407

Filter

- Gene symbol
- Transcript
- Total reads
- P-value
- FDR step up
- Ratio
- Fold change
- Low Expressed

Save filter | Clear filter

Saved filters

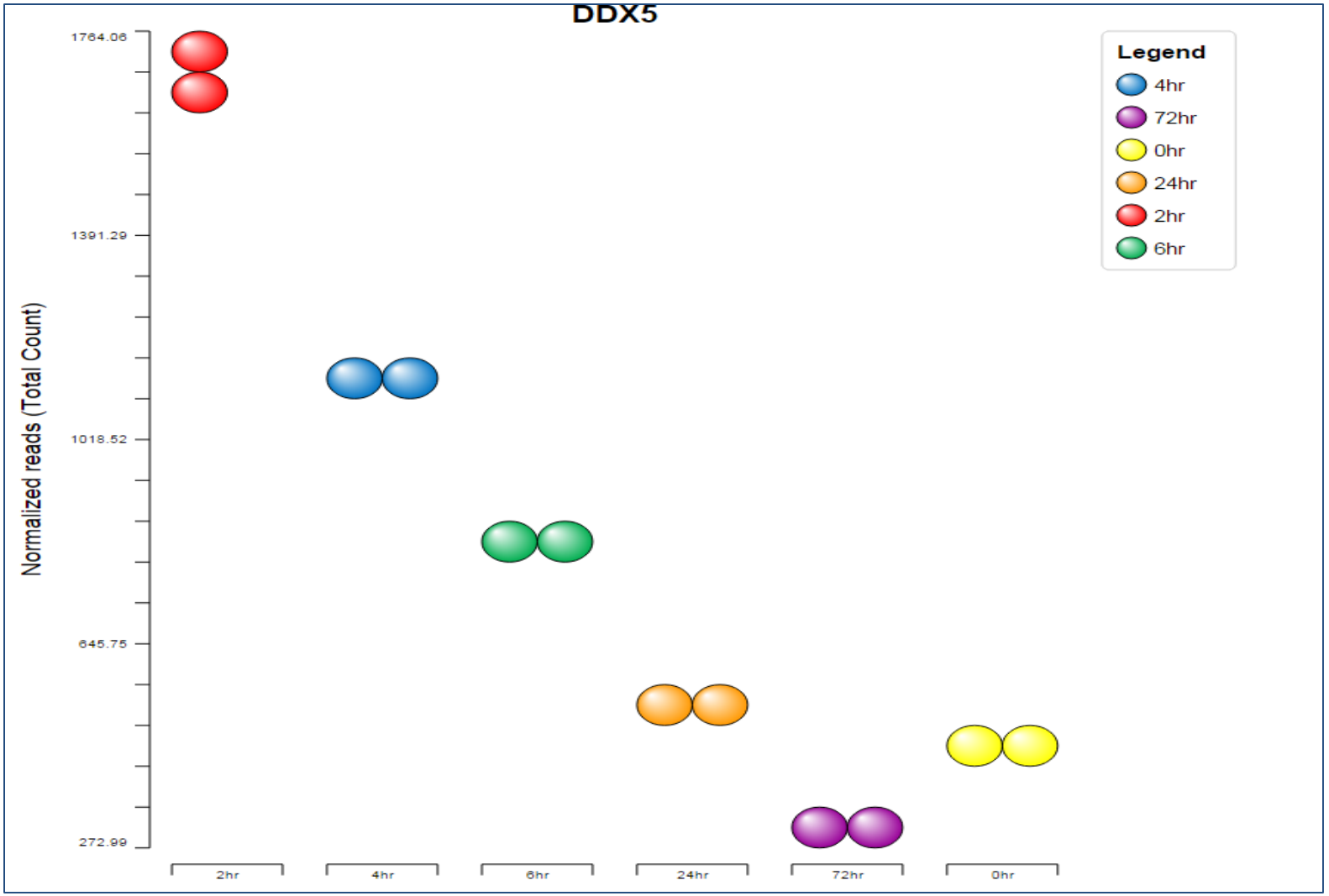
(No saved filters available)

				0hr vs 2hr				
Row	Gene symbol	Transcript	Total reads	P-value	FDR step up	Ratio	Fold change	View
1	WASH7P	NR_024540	4,784.21	2.44E-9	2.8E-8	2.18	2.18	
2	LOC729737	NR_039983	2,296.76	0.46	0.64	0.91	-1.10	
3	LINC01002	NR_028325	27.07	0.82	0.92	0.83	-1.21	
4	LOC100132287	NR_028322.1	27.07	0.82	0.92	0.83	-1.21	
5	LOC100133331	NR_028327	45.60	0.88	0.95	0.92	-1.09	
6	LOC100133331	NR_028327.1	3,133.46	0.75	0.87	1.07	1.07	
7	LOC100288069	NR_033908	281.10	0.96	0.99	1.02	1.02	
8	FAM87B	NR_103536	146.31	0.52	0.70	0.70	-1.43	
9	LINC00115	NR_024321	262.10	9.2E-4	3.59E-3	0.32	-3.10	
10	LINC01128	NR_047519	4,331.57	1.37E-6	9.89E-6	0.42	-2.36	
11	LINC01128	NR_047521	2,018.67	0.48	0.65	0.73	-1.37	
12	LINC01128	NR_047523	601.52	0.39	0.57	0.65	-1.54	
13	LINC01128	NR_047524	78.85	0.13	0.26	0.30	-3.32	
14	LINC01128	NR_047526	1,606.38	1.35E-4	6.32E-4	0.39	-2.55	
15	LINC01128	NR_047525	678.55	0.16	0.31	0.36	-2.81	
16	FAM41C	NR_027055	99.03	0.07	0.16	0.65	-1.54	
17	SAMD11	NM_152486	40.50	0.03	0.07	0.07	-15.02	
18	NOC2L	NM_015658	57,483.14	1.58E-5	9.05E-5	0.76	-1.32	
19	KLHL17	NM_198317	1,370.00	7.23E-4	2.89E-3	1.93	1.93	
20	PLEKHN1	NM_001160184	432.10	0.27	0.44	0.23	-4.38	
21	PLEKHN1	NM_032129	619.90	0.74	0.87	1.08	1.08	
22	PERM1	NM_001291366	25.78	0.77	0.89	0.63	-1.58	
23	PERM1	NM_001291367	393.22	0.61	0.77	0.63	-1.60	
24	HES4	NM_021170	197.62	0.02	0.06	0.08	-12.50	
25	HES4	NM_001142467	140.38	0.12	0.25	0.15	-6.86	

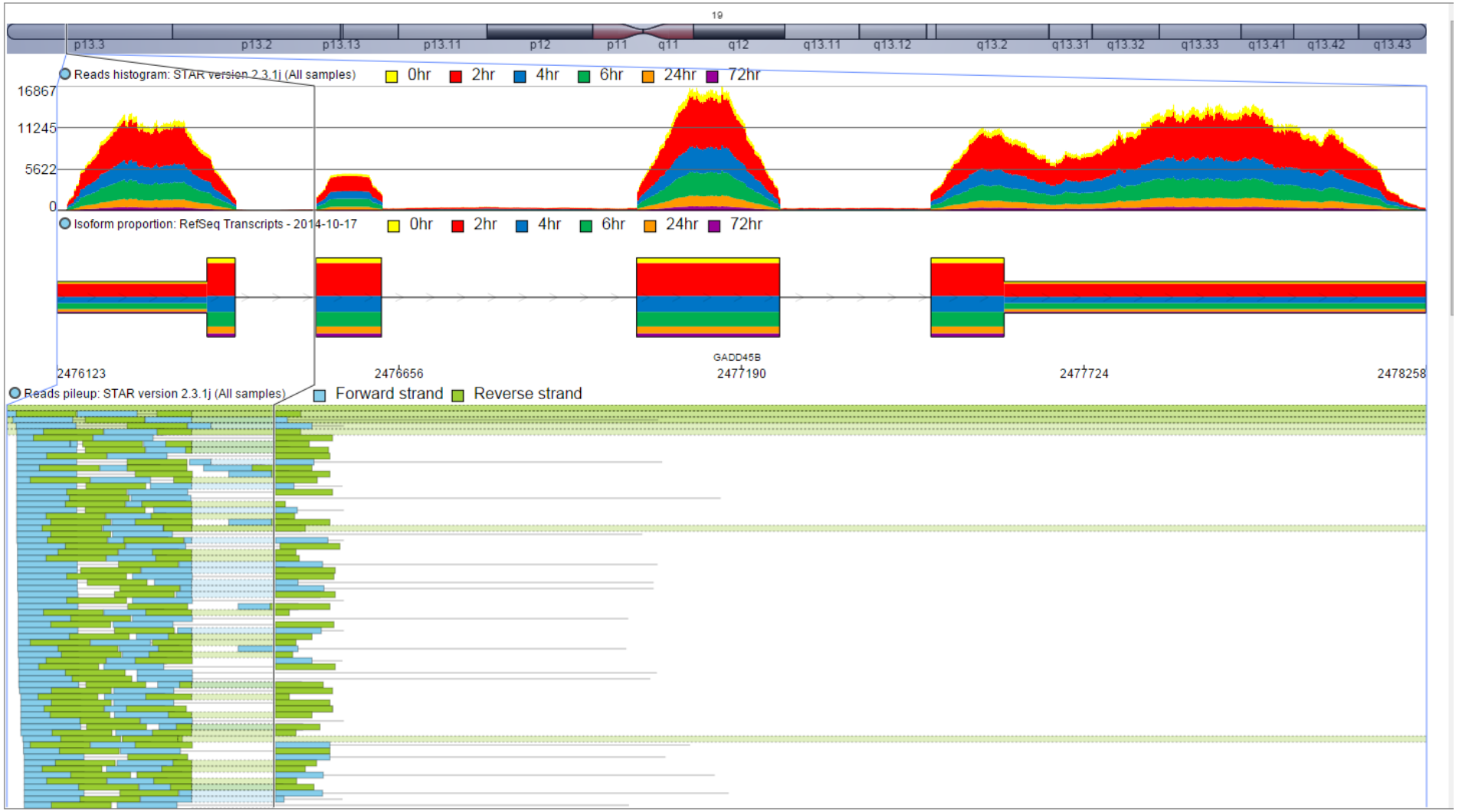
Rows per page 25 (1 of 1337) Download data

Generate list

Dot Plot

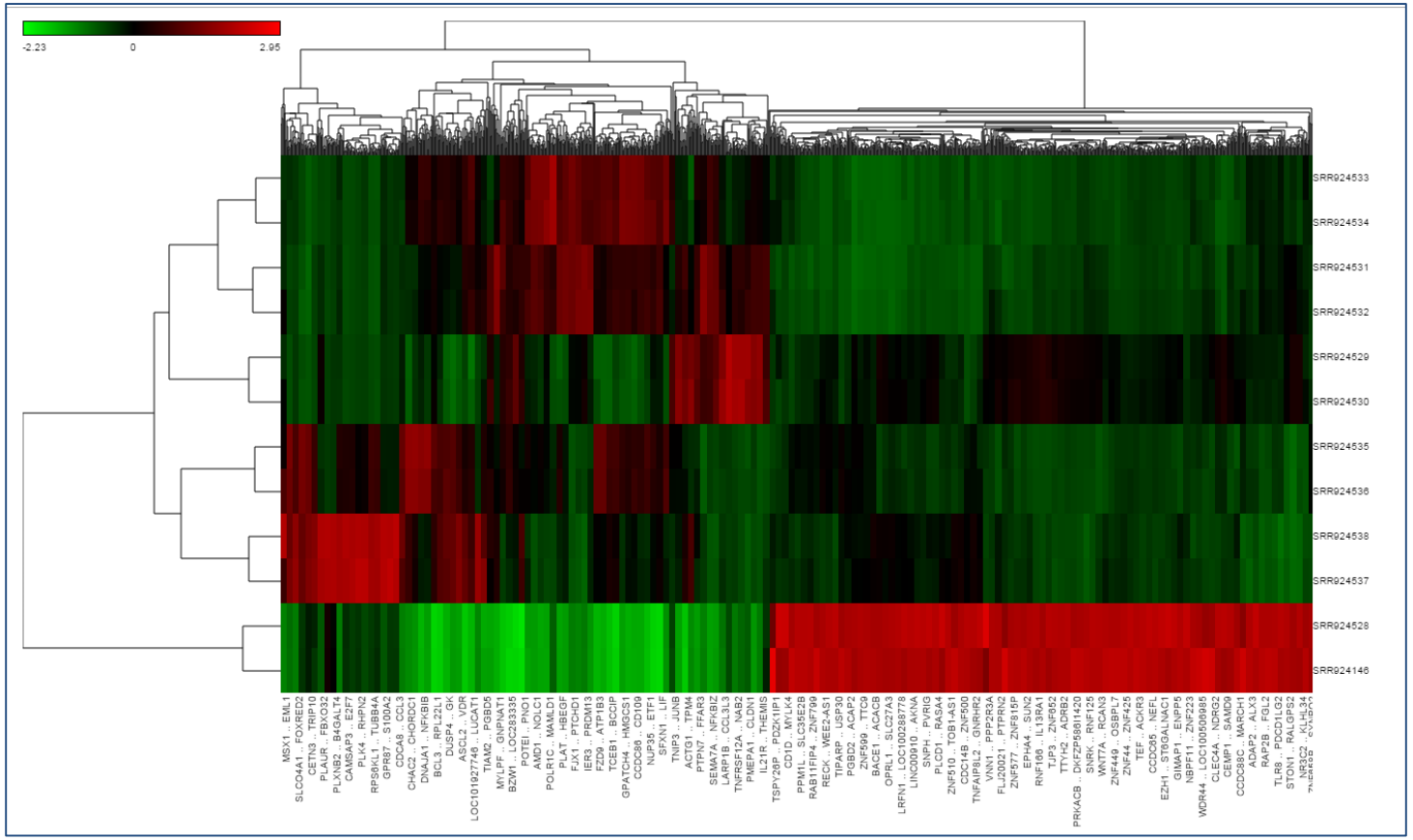


Genome View



Hierarchical Clustering

Hierarchical Clustering allows for the visualization of genes and samples of interest in a single plot, grouping samples and genes based upon similarity.



Biological Interpretation

- Enrichment analysis allows for testing a list of significant genes to determine if they are over represented in any gene set/pathway.

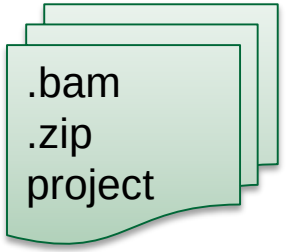
Gene set	Description	Enrichment score	P-value	Genes in list	Genes not in list	
GO:0002376	immune system process	46.58	5.87E-21			
GO:0065007	biological regulation	44.73	3.74E-20			
GO:0080090	regulation of primary metabolic process	41.75	7.39E-19			
GO:0031323	regulation of cellular metabolic process	41.38	1.07E-18	492	5,105	
GO:0006955	immune response	40.74	2.02E-18	139	877	
GO:0060255	regulation of macromolecule metabolic process	40.57	2.41E-18	453	4,603	
GO:0019222	regulation of metabolic process	39.88	4.77E-18	519	5,513	
GO:0002682	regulation of immune system process	36.95	9E-17	145	983	
GO:0050789	regulation of biological process	36.25	1.8E-16	756	9,069	
GO:0050794	regulation of cellular process	34.70	8.49E-16	729	8,706	
GO:0048583	regulation of response to stimulus	34.61	9.32E-16	286	2,616	
GO:0002684	positive regulation of immune system process	33.79	2.12E-15	103	610	
GO:0080134	regulation of response to stress	33.56	2.65E-15	125	825	
GO:0048518	positive regulation of biological process	32.24	9.99E-15	409	4,241	
GO:0051239	regulation of multicellular organismal process	32.18	1.05E-14	221	1,892	
GO:0019219	regulation of nucleobase-containing compound metabolic process	29.78	1.16E-13	370	3,797	
GO:0070887	cellular response to chemical stimulus	29.66	1.31E-13	200	1,698	
GO:0010556	regulation of macromolecule biosynthetic process	28.58	3.87E-13	333	3,351	
GO:0048522	positive regulation of cellular process	28.40	4.65E-13	369	3,823	
GO:0031326	regulation of cellular biosynthetic process	28.24	5.41E-13	343	3,489	
GO:0051171	regulation of nitrogen compound metabolic process	27.93	7.43E-13	373	3,889	
GO:0006952	defense response	27.60	1.03E-12	138	1,043	
GO:0009889	regulation of biosynthetic process	27.10	1.7E-12	344	3,534	
GO:0006950	response to stress	26.17	4.3E-12	250	2,370	
GO:2001141	regulation of RNA biosynthetic process	25.97	5.27E-12	301	3,013	

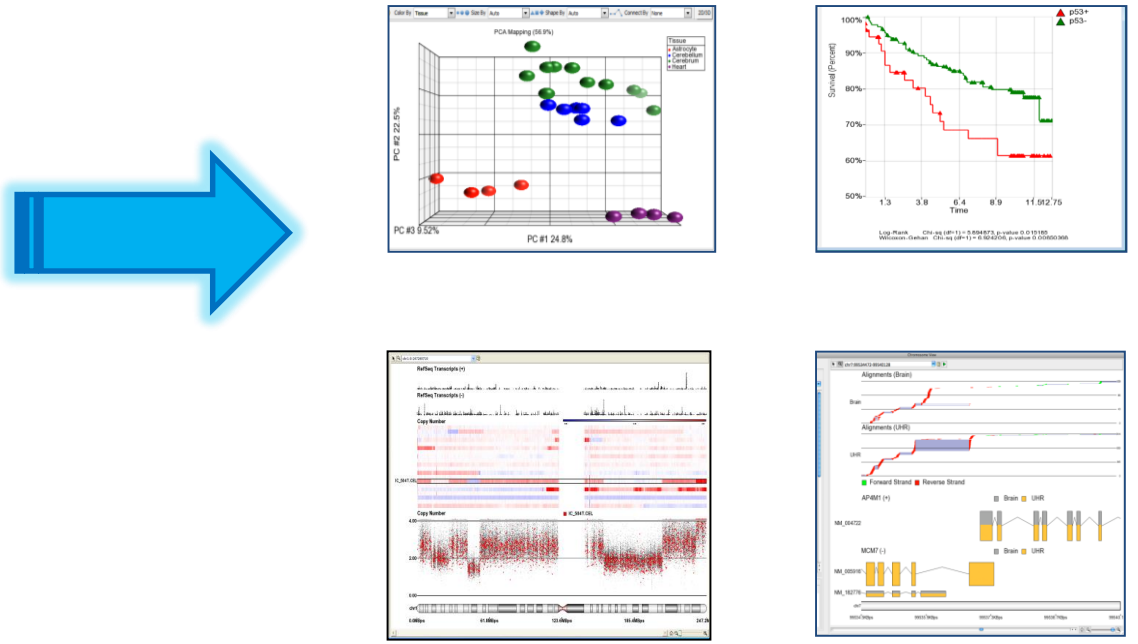
	In list	Not in list
In set	203	1,455
Not in set	942	15,428

Rows per page 25 (1 of 323) Download data

Input of NGS reads into Partek® Genomics Suite™

- Aligned data (.sam/.bam files)
- Project download from Partek® Flow ®



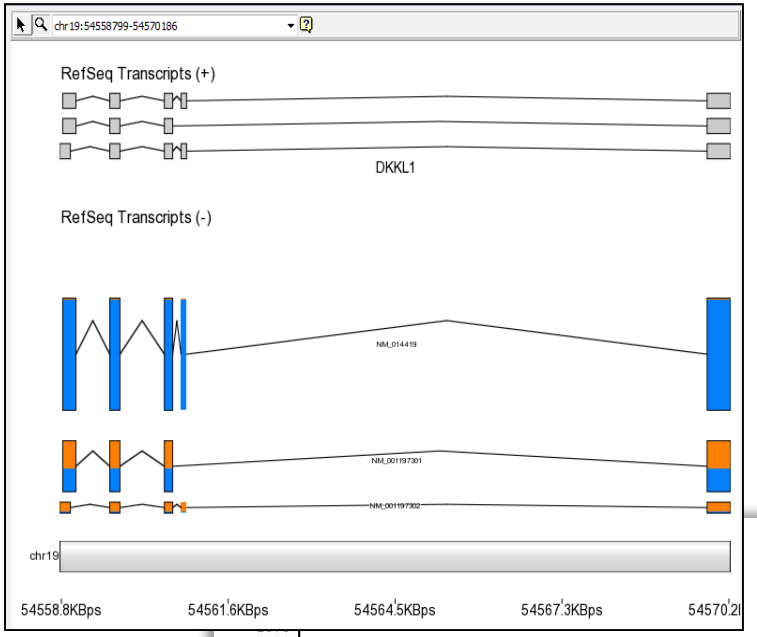


Data Analysis in Partek® Genomics Suite™

Partek®
turning data into discovery®

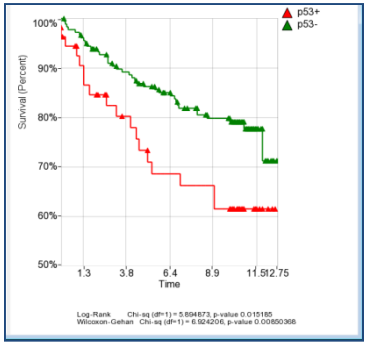
RNA-seq Workflows in PGS

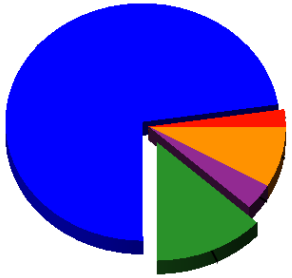
- RNA-Seq
  - Quantification
  - Differential expression detection
  - Allele specific expression
  - Integration with other genomic data



Inferential Statistics

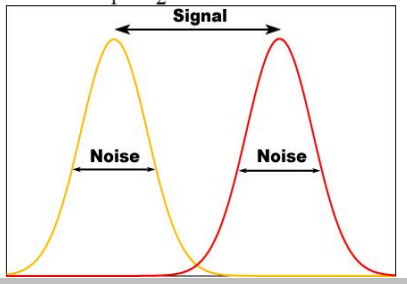
- Parametric
  - t-Test, z-Test, ANOVA, Welch's ANOVA, Pearson Correlation
- Non-parametric
  - Mann-Whitney, Kruskal-Wallis, Friedman, Chi-square, Rank correlation
- Others
  - Power Analysis
  - Survival Analysis
    - Cox regression
    - Kaplan-Meier Curve
  - Multiple test corrections
  - Descriptive statistics





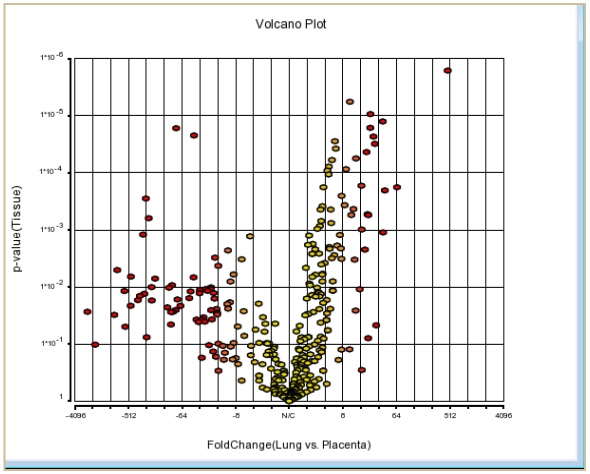
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} \sim \frac{signal}{noise}$$

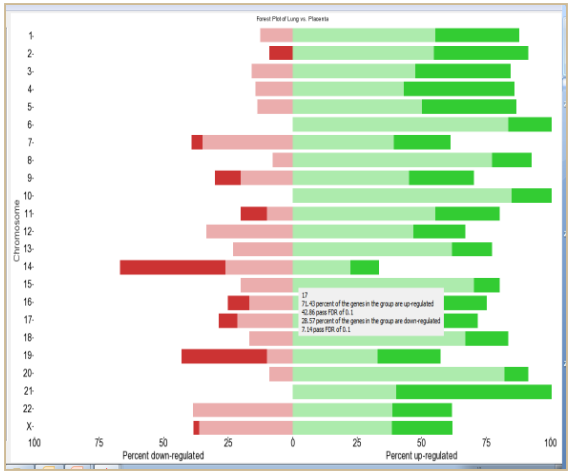
R^2

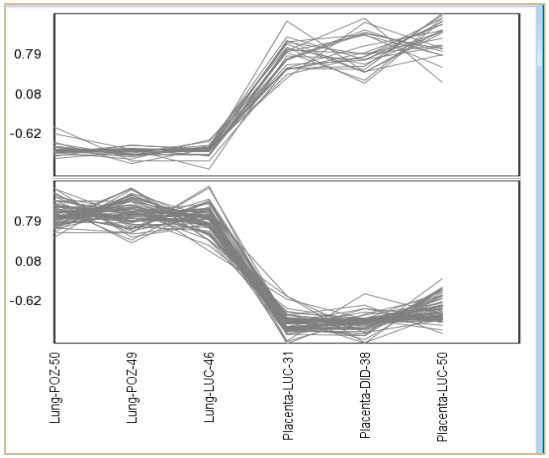


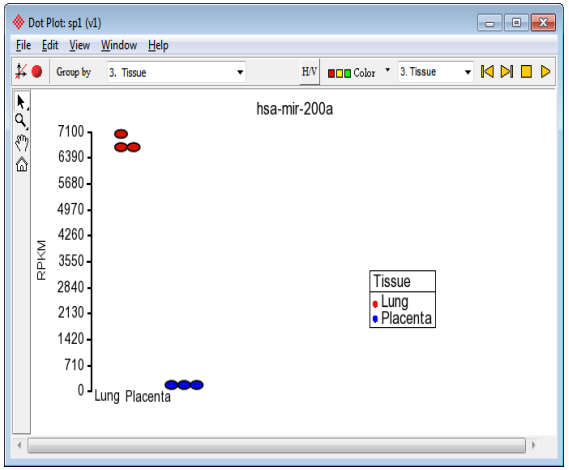
Visualization of Statistical Results

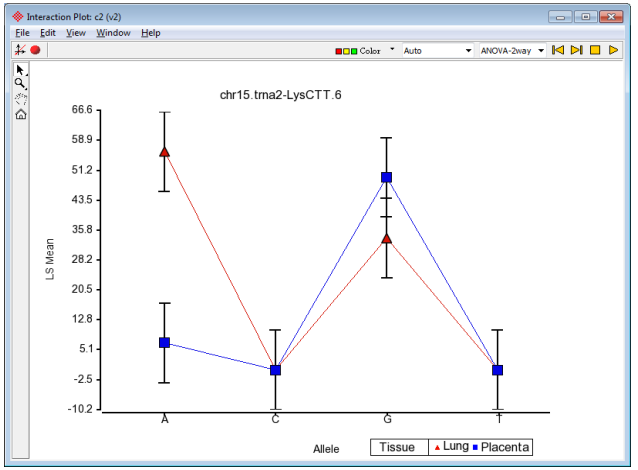
- Forest plot
- Volcano plot
- Dot plot
- Profile
- Interaction plot
- More…











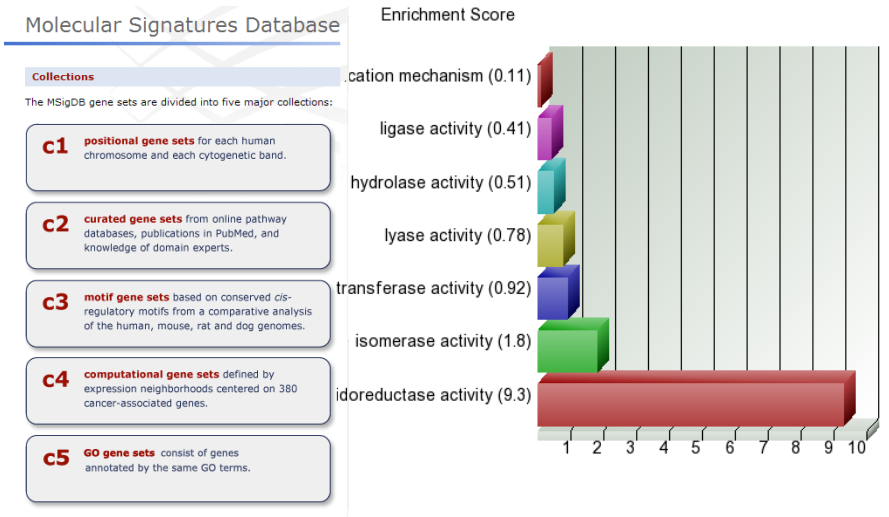
Save Images (.jpeg .svg .pdf and more…)

Biological Interpretation

- Biological relevance is not usually found in only a single gene

- Database:
  - Gene Ontology,
  - KEGG Pathways
  - Custom annotation
    - GMT, GAF, text file

- Method:
  - Enrichment: test if lead genes are over-represented in any pathway
  - Pathway ANOVA: detect differentially expressed pathway

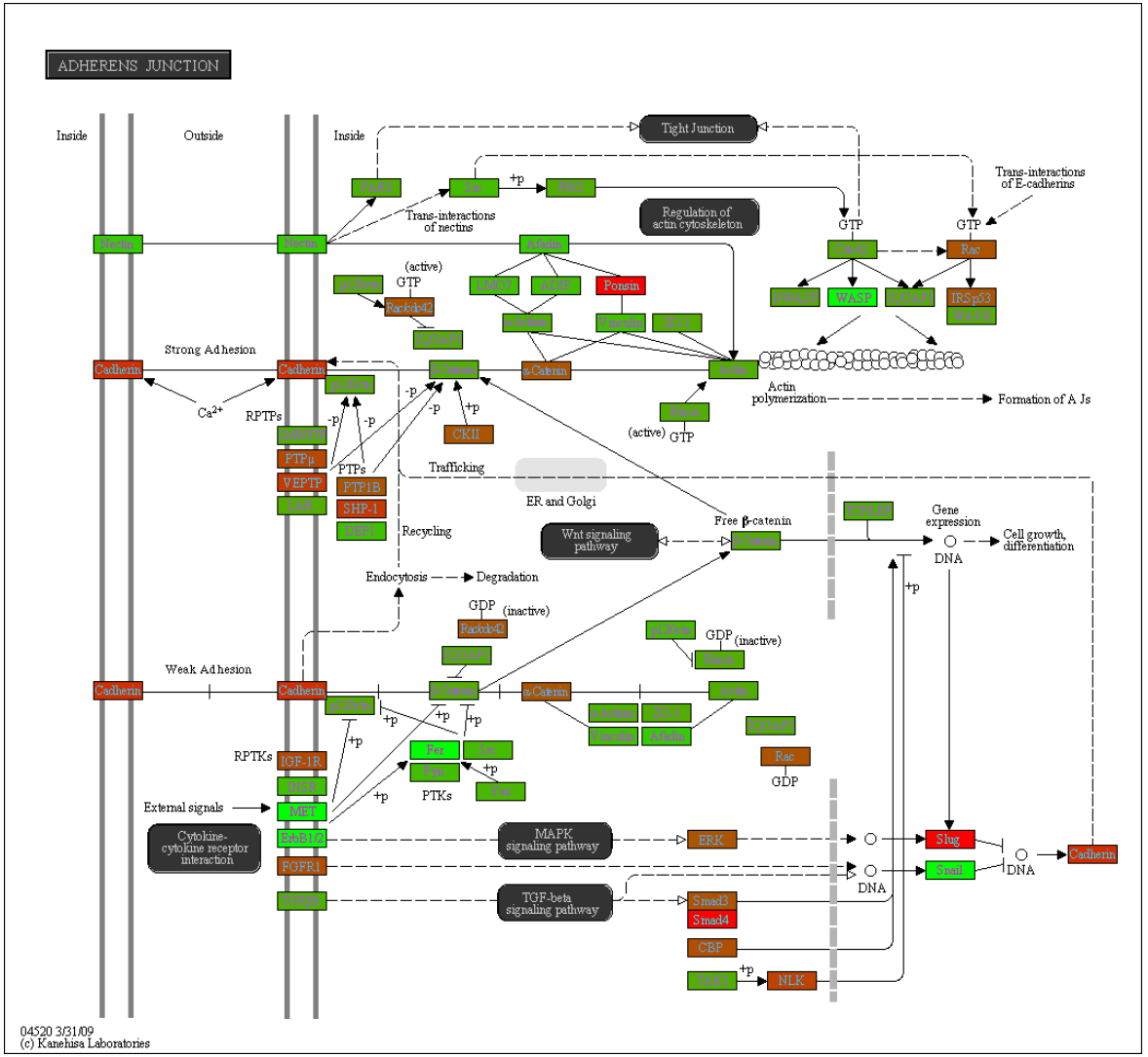


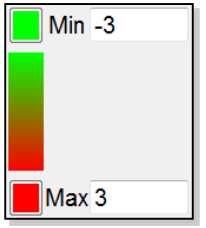
KEGG Organisms: Complete Genomes

Eukaryotes: 180 Bacteria: 2149 Archaea: 149

[Genomes | Draft

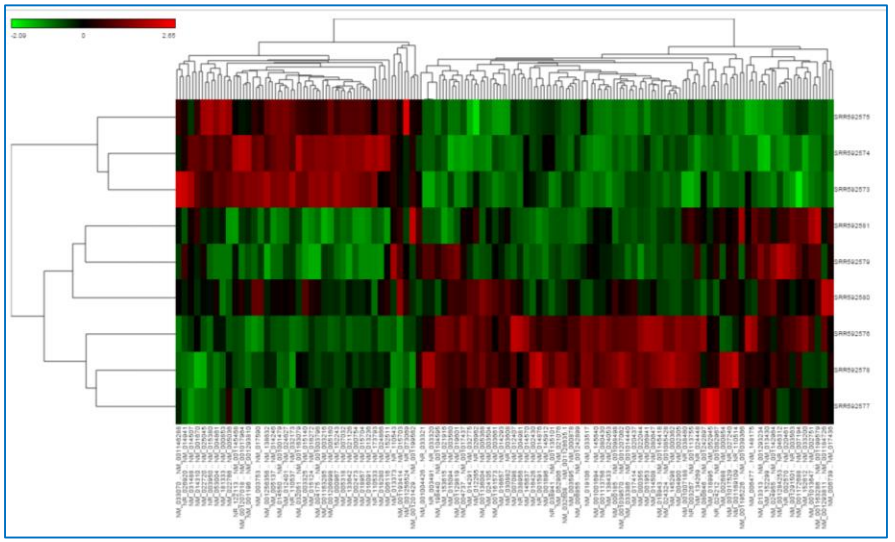
Pathway ANOVA

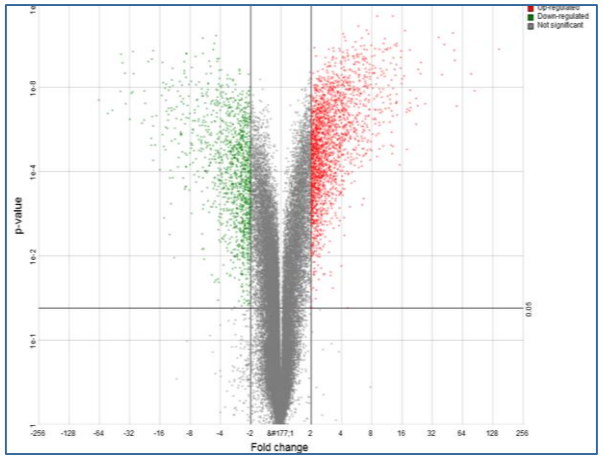


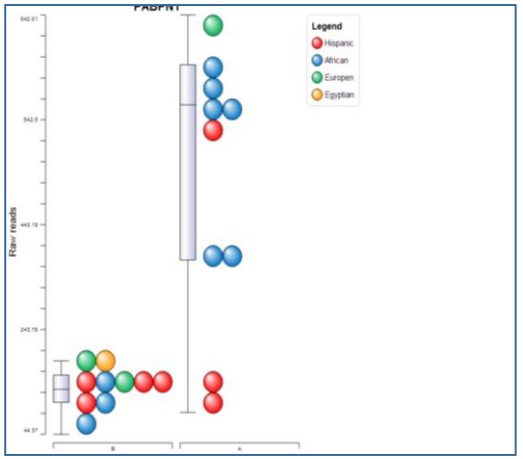


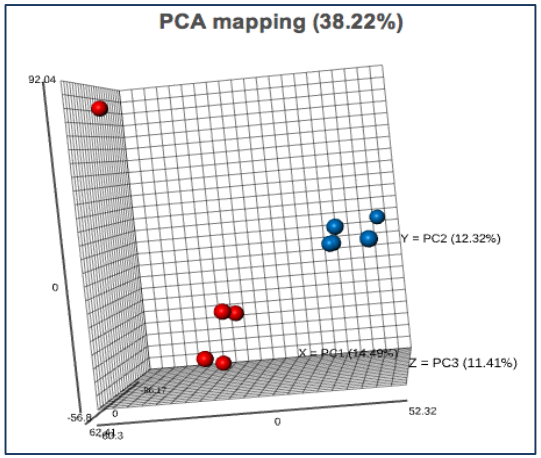
Highlight of Partek® Flow ®

- Flexible data storage
- Context sensitive interface
- Visual analysis pipeline
- Broad choice of analysis tools
- Comprehensive statistics and visualization



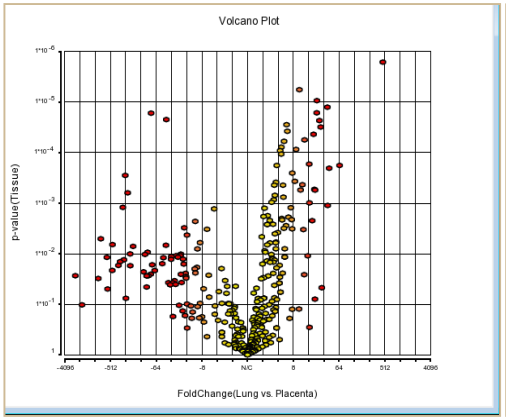


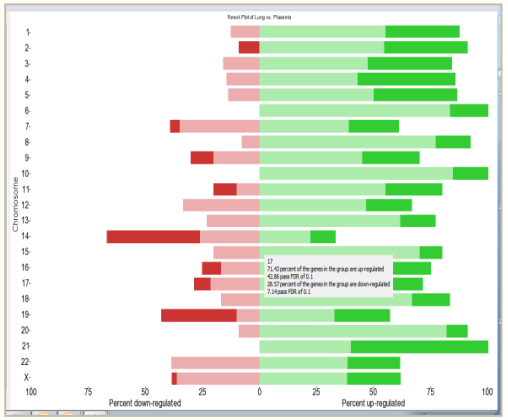


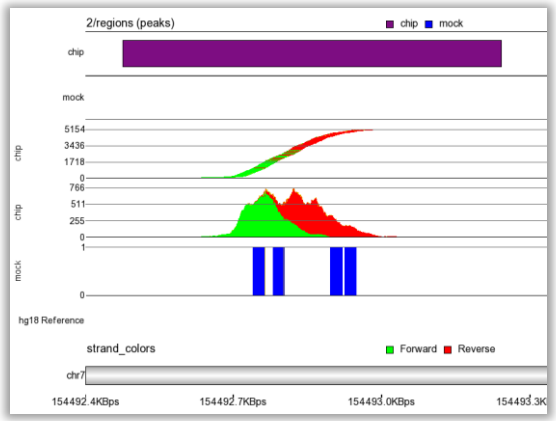


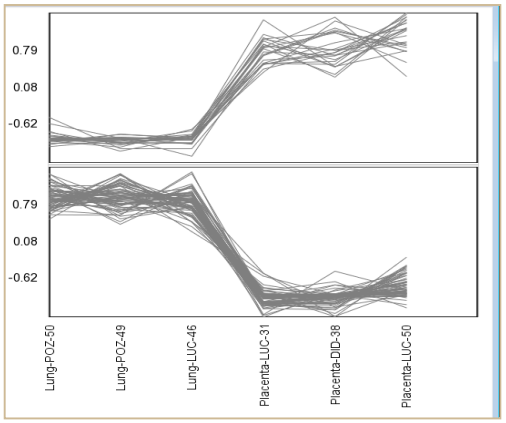
Highlight of Partek ® Genomics Suite™ 6.6

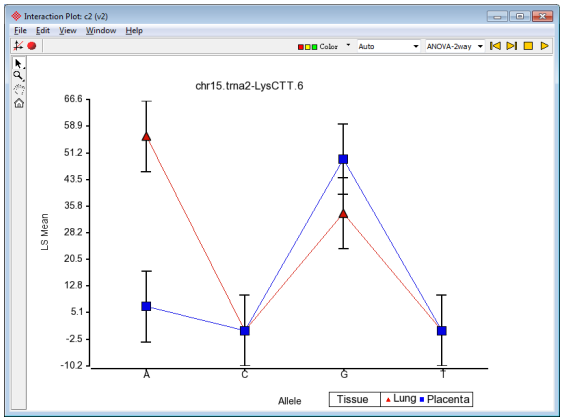
- Simple workflows for microarray and NGS assays
- Powerful statistics
- Interactive visualization
- Flexible data integration



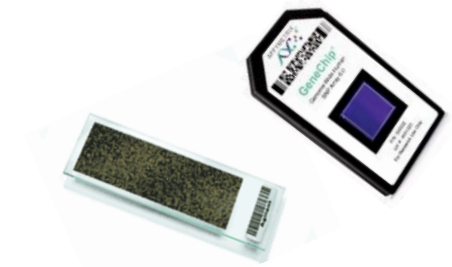






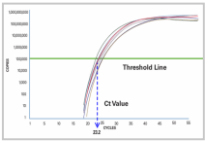


Partek Provide Solutions for Any Technology



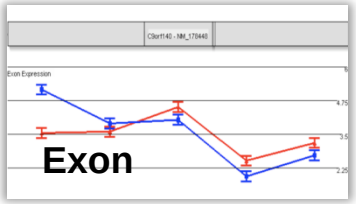
Partek® Flow-GS-Pathway

RT-PCR

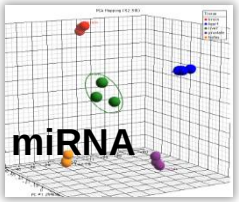


Partek Provide Solutions for Any Assay

Gene Expression

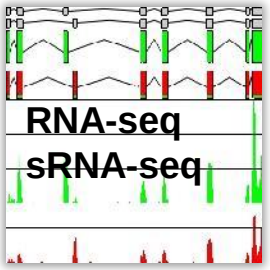


Exon



miRNA

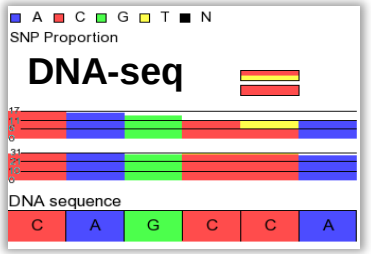
Partek® Flow-GS-Pathway



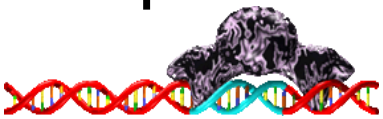
RNA-seq

sRNA-seq

ChIP-Seq

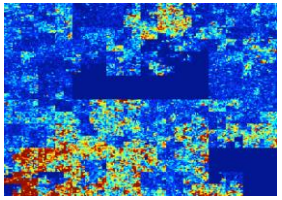


ChIP-chip



CN/LOH /ASCN

Methylation



Partek is Your Partner

Self-learning

- Help > On-line tutorials
- Recorded webinars

Regional Technical Support

- Email: support@partek.com
- Phone: +1-314-878-2329

Partek® Flow® @NIH

- Instructions for setting up a Flow account can be found here: partekflow.cit.nih.gov
- Account set up require 3 main steps:
  - Set up a Helix account
  - Obtain storage space
  - Request Flow account