

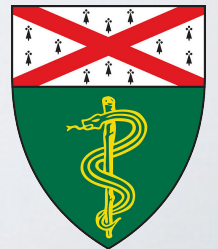
exceRpt

a computational exRNA-seq analysis pipeline

Rob Kitchen

Yale

2015 - 04 - 23



Extracellular RNA Communication

NIH Common Fund

- > data management
- > reference profiles
- > exRNA biogenesis
- > exRNA biomarker
- > exRNA therapy


currently 30 funded projects

exRNA.org

exRNA
RESEARCH PORTAL


ABOUT PROJECTS PUBLICATIONS **RESOURCES** EVENTS JOBS BLOG

Resources




Protocols

The consortium is working internally to compare Standard Operating Procedures (SOPs), i.e. protocols used by different labs for extracellular vesicle preparation and RNA analysis. Vetted results are here.




Data

As consortium members begin to publish, datasets associated with their publications will be posted here. Please contact us if you have an extracellular RNA or EV dataset you would like to make available to the community.



Standards

The consortium is developing data and metadata standards to enable collection and comparison of extracellular RNA datasets from multiple sources.



Software


Here you will find bioinformatics tools and pipelines developed by the consortium for the analysis of extracellular RNA.

If you want to contribute a resource or to be informed when new resources become available here, please contact us at info@exrna.org.

exRNA
About
Projects
Publications
Resources

Quick Links
ExRNA Atlas @ BioGPS
ExRNA @ WikiPathways
ExoCarta
Vesiclepedia

Contact Us
info@exrna.org
twitter.com/exRNA
Help
Search



exRNA-seq analysis software

small RNA-seq
analysis toolkit

exceRpt

quantification (RPM) of
miRNA, tRNA, piRNA,
snoRNA, circularRNA,
& long transcript fragments



long RNA-seq
analysis toolkit

RSEQTools

quantification (RPKM) of
[multi-exon] transcripts



agenda

1. **exceRpt** smallRNA-seq analysis suite
 - filtering, QC, and alignment
 - quantification & normalisation
 - visualisations
 - differential expression (soon)
2. **use-case** involving 345 samples
3. **downstream analysis**
 - miRNA-mRNA targets from miRTarBase
 - network analysis
 - enrichment analysis
 - pathway analysis

exRNA-seq analysis software

small RNA-seq
analysis toolkit

exceRpt

quantification (RPM) of
miRNA, tRNA, piRNA,
snoRNA, circularRNA,
& long transcript fragments



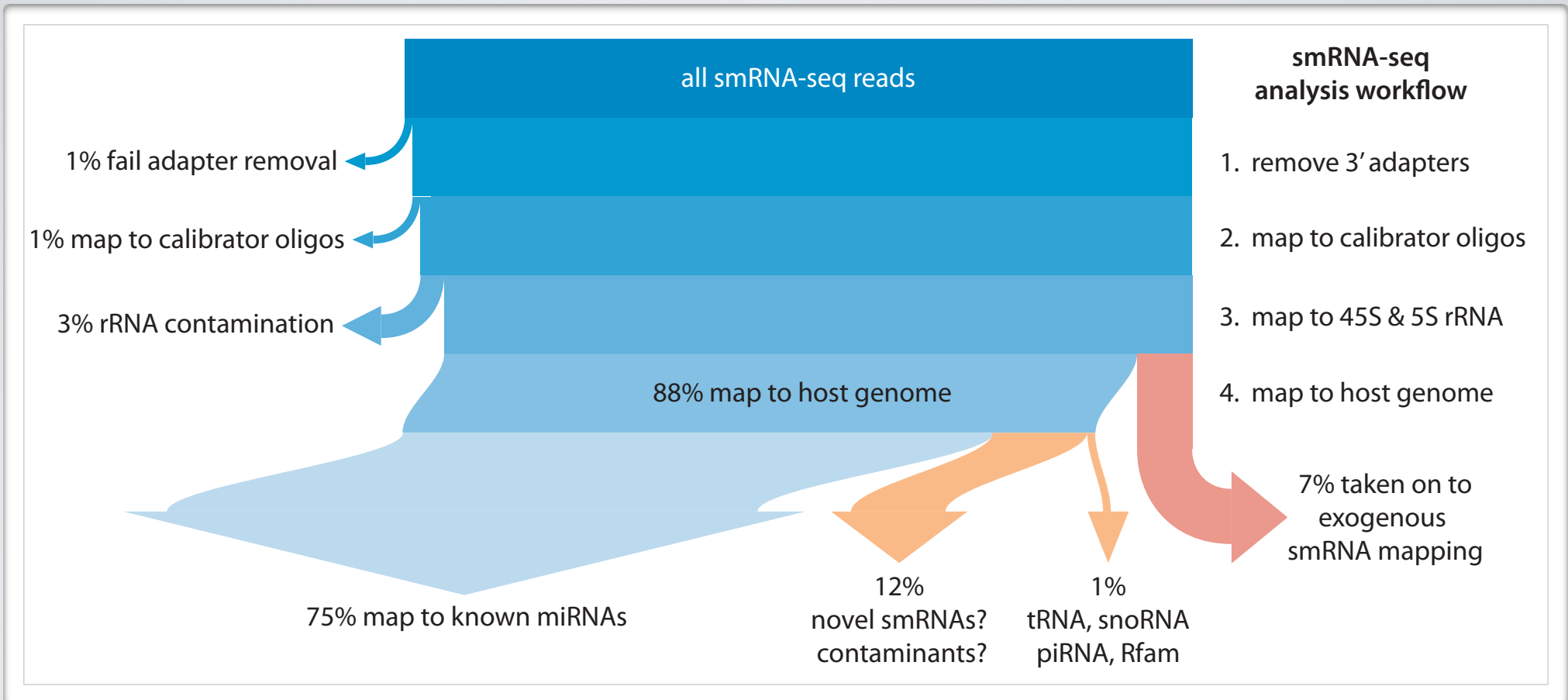
long RNA-seq
analysis toolkit

RSEQTools

quantification (RPKM) of
[multi-exon] transcripts



for a typical cellular sample...

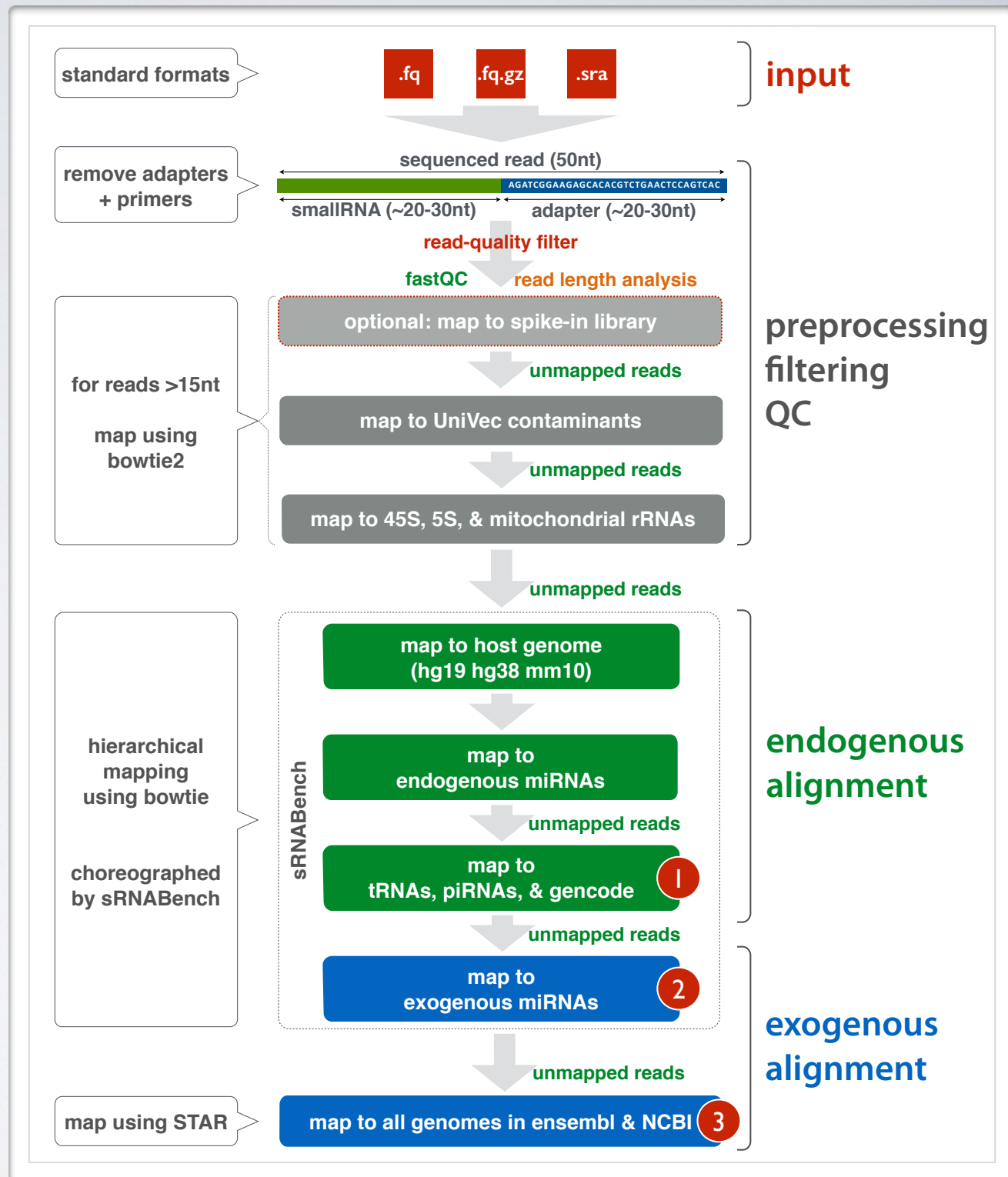


- exRNA samples typically much noisier
- cascade of read-alignment steps mitigates contamination

exceRpt

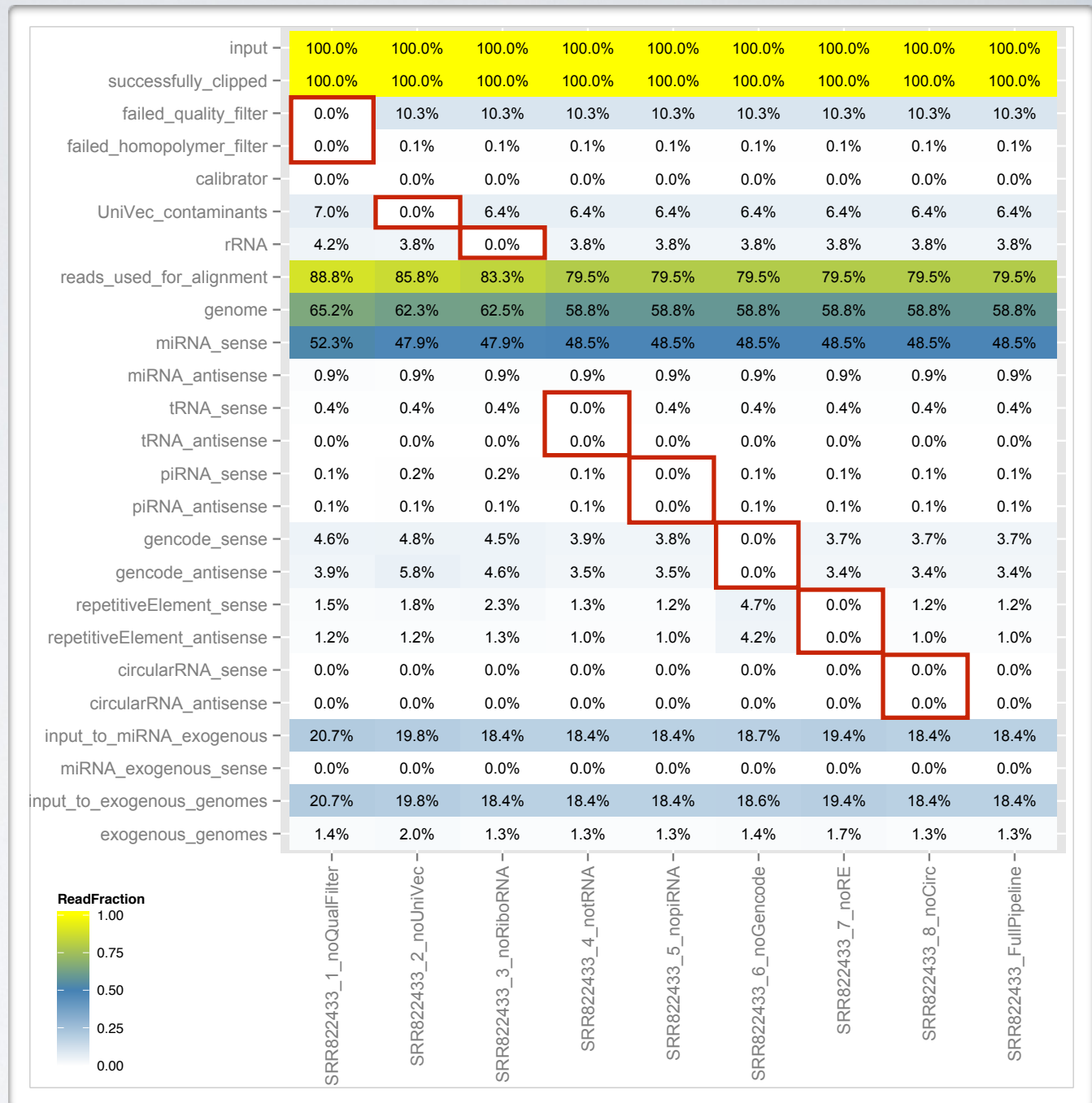
- automatic pre-processing and QC of sequence reads
- absolute quantitation by quantification of exogenous spike-in sequences
- explicit rRNA filtering & QC
- quantify many different smallRNA types
- choice of 3 end-points

1 2 3



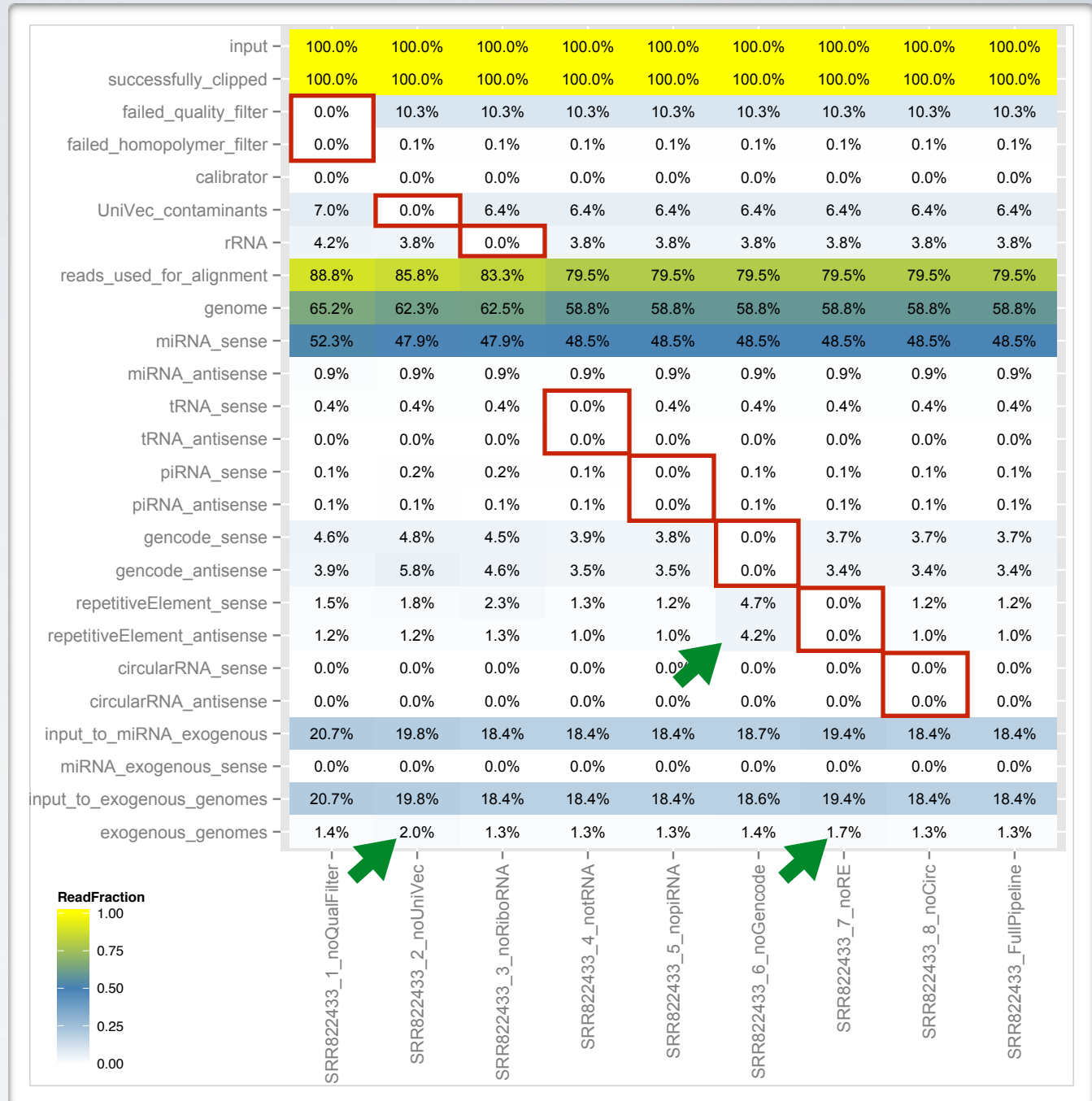
effect of filtering

- what happens to alignments when we individually remove upstream libraries?
- gencode appears to be a subset of the repetitive elements
- quality filter, UniVec, and repetitive elements have the largest impact



effect of filtering

- what happens to alignments when we individually remove upstream libraries?
- gencode appears to be a subset of the repetitive elements
- quality filter, UniVec, and repetitive elements have the largest impact



mappability

- what happens if we run the smallRNA libraries through the pipeline as if they were reads?
- miRNAs include all species, but most are highly conserved
- 1/20 piRNAs are rRNA

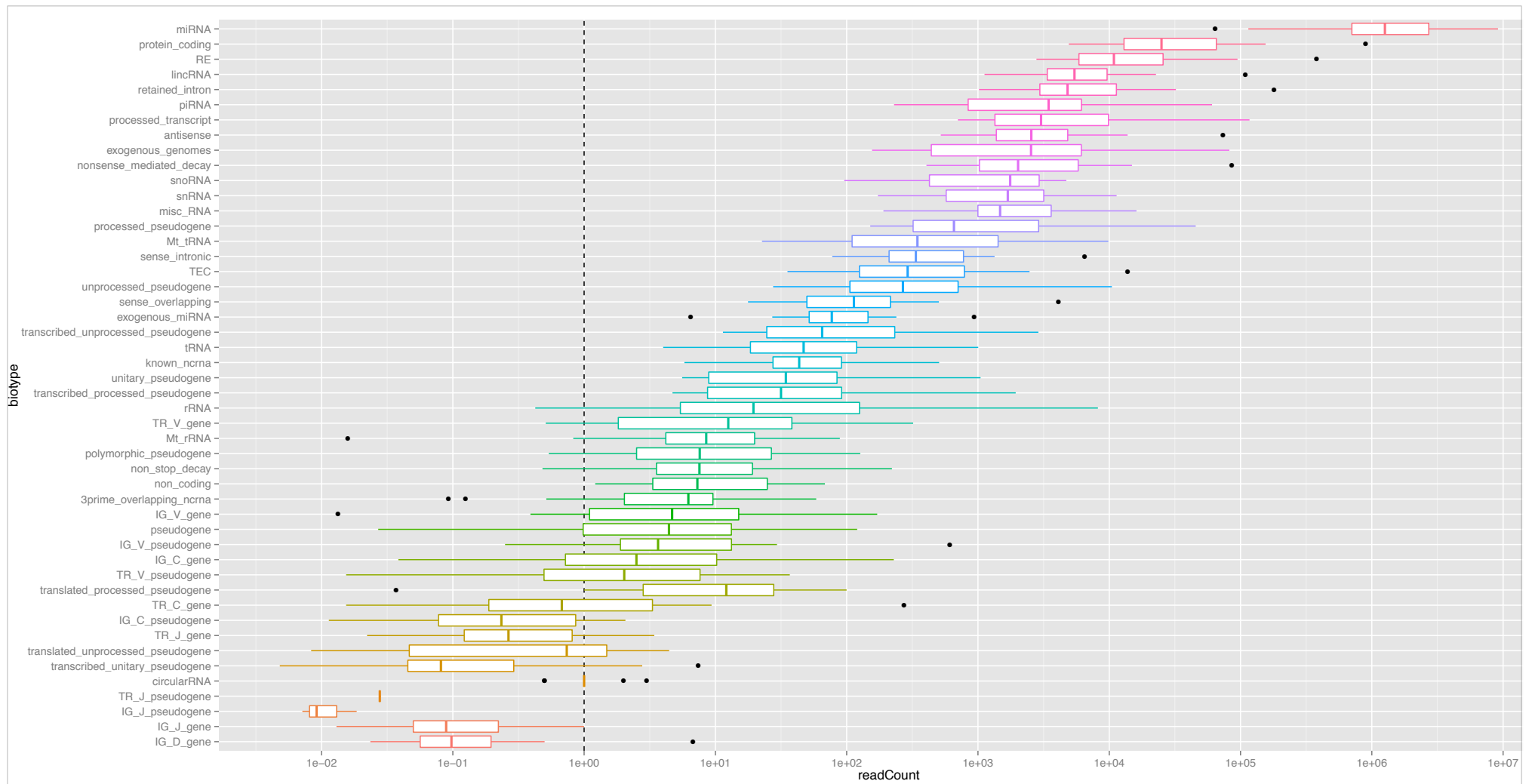
input	100.0%	100.0%	100.0%	100.0%
clipped	100.0%	100.0%	100.0%	100.0%
failed_quality_filter	0.0%	0.0%	0.0%	0.0%
failed_homopolymer_filter	0.3%	0.3%	0.3%	0.0%
calibrator	NA%	NA%	NA%	NA%
UniVec_contaminants	0.0%	0.0%	0.0%	0.0%
rRNA	0.1%	0.2%	0.1%	5.0%
reads_used_for_alignment	99.6%	99.5%	99.6%	95.0%
genome	75.5%	75.5%	73.9%	94.9%
miRNA_sense	32.5%	32.4%	27.7%	0.0%
miRNA_antisense	0.6%	0.6%	0.3%	0.0%
tRNA_sense	0.0%	0.0%	0.0%	0.2%
tRNA_antisense	0.0%	0.0%	0.0%	0.0%
piRNA_sense	0.0%	0.0%	0.0%	94.6%
piRNA_antisense	0.0%	0.0%	0.0%	0.1%
gencode_sense	7.8%	7.7%	8.3%	0.0%
gencode_antisense	4.9%	4.9%	5.8%	0.0%
repetitiveElement_sense	8.1%	7.9%	7.5%	0.0%
repetitiveElement_antisense	8.9%	9.1%	7.4%	0.0%
circularRNA_sense	0.0%	0.0%	0.0%	0.0%
circularRNA_antisense	0.0%	0.0%	0.0%	0.0%
input_to_miRNA_exogenous	18.7%	18.7%	20.8%	0.0%
miRNA_exogenous_sense	13.6%	13.6%	14.9%	0.0%
input_to_exogenous_genomes	1.4%	1.4%	1.4%	0.0%
exogenous_genomes	1.2%	1.2%	1.4%	0.0%
	miRBase21_hg19_FullPipeline	miRBase21_hg38_FullPipeline	miRBase21_mm10_FullPipeline	piRNAs_hg38_FullPipeline

mappability

- what happens if we run the smallRNA libraries through the pipeline as if they were reads?
- miRNAs include all species, but most are highly conserved
- 1/20 piRNAs are rRNA

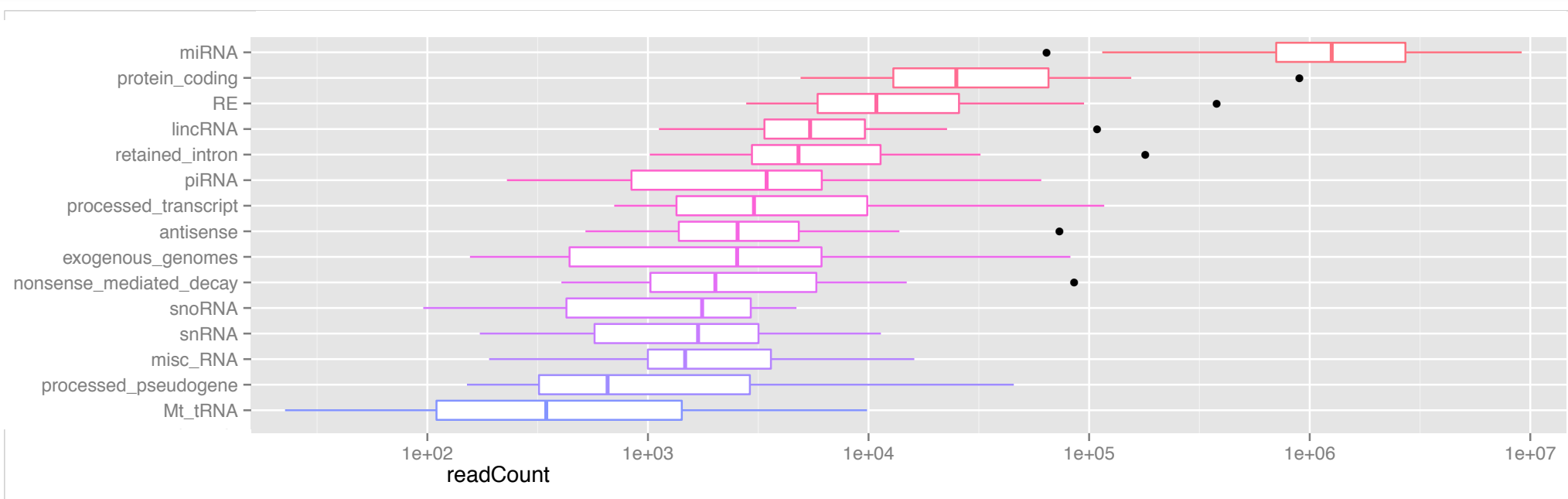
input	100.0%	100.0%	100.0%	100.0%
clipped	100.0%	100.0%	100.0%	100.0%
failed_quality_filter	0.0%	0.0%	0.0%	0.0%
failed_homopolymer_filter	0.3%	0.3%	0.3%	0.0%
calibrator	NA%	NA%	NA%	NA%
UniVec_contaminants	0.0%	0.0%	0.0%	0.0%
rRNA	0.1%	0.2%	0.1%	5.0%
reads_used_for_alignment	99.6%	99.5%	99.6%	95.0%
genome	75.5%	75.5%	73.9%	94.9%
miRNA_sense	32.5%	32.4%	27.7%	0.0%
miRNA_antisense	0.6%	0.6%	0.3%	0.0%
tRNA_sense	0.0%	0.0%	0.0%	0.2%
tRNA_antisense	0.0%	0.0%	0.0%	0.0%
piRNA_sense	0.0%	0.0%	0.0%	94.6%
piRNA_antisense	0.0%	0.0%	0.0%	0.1%
gencode_sense	7.8%	7.7%	8.3%	0.0%
gencode_antisense	4.9%	4.9%	5.8%	0.0%
repetitiveElement_sense	8.1%	7.9%	7.5%	0.0%
repetitiveElement_antisense	8.9%	9.1%	7.4%	0.0%
circularRNA_sense	0.0%	0.0%	0.0%	0.0%
circularRNA_antisense	0.0%	0.0%	0.0%	0.0%
input_to_miRNA_exogenous	18.7%	18.7%	20.8%	0.0%
miRNA_exogenous_sense	13.6%	13.6%	14.9%	0.0%
input_to_exogenous_genomes	1.4%	1.4%	1.4%	0.0%
exogenous_genomes	1.2%	1.2%	1.4%	0.0%
	miRBase21_hg19_FullPipeline	miRBase21_hg38_FullPipeline	miRBase21_mm10_FullPipeline	piRNAs_hg38_FullPipeline

total reads by biotype



total reads by biotype

- large contribution from miRNA and mRNA
- also some signal from exogenous sequences



exceRpt @ Genboree

The screenshot displays the exceRpt web interface. On the left is the 'Data Selector' panel with a tree view of data sources. On the top right is a 'Details' table. On the bottom right are 'Input Data' and 'Output Targets' sections.

Data Selector

Refresh Data Filter: Select a filter...

- Databases
 - Bowtie - Example Data
 - BWA hg19 - Example data
 - BWA mm9 - Example Data
 - FastQC - Example Data
 - Import Samples - Example Data
 - RSEQtools hg18 - Example Data
 - RSEQtools hg19 - Example Data
 - smallRNA-seq Pipeline - Example Data
 - Tracks
 - Lists & Selections
 - SampleSets
 - Samples
 - Files
 - smallRNaseqPipeline
 - SRR822433.fastq.gz
- Projects
 - exRNA
 - rob.kitchen_group
 - Databases
 - exRNA example
 - Projects

Details

Attribute	Value
Group	rob.kitchen_group
Role	administrator
Name	exRNA example
Description	Template for Human Genome, UCSC Build Hg19
Species	Homo sapiens

Input Data

SRR822433.fastq.gz

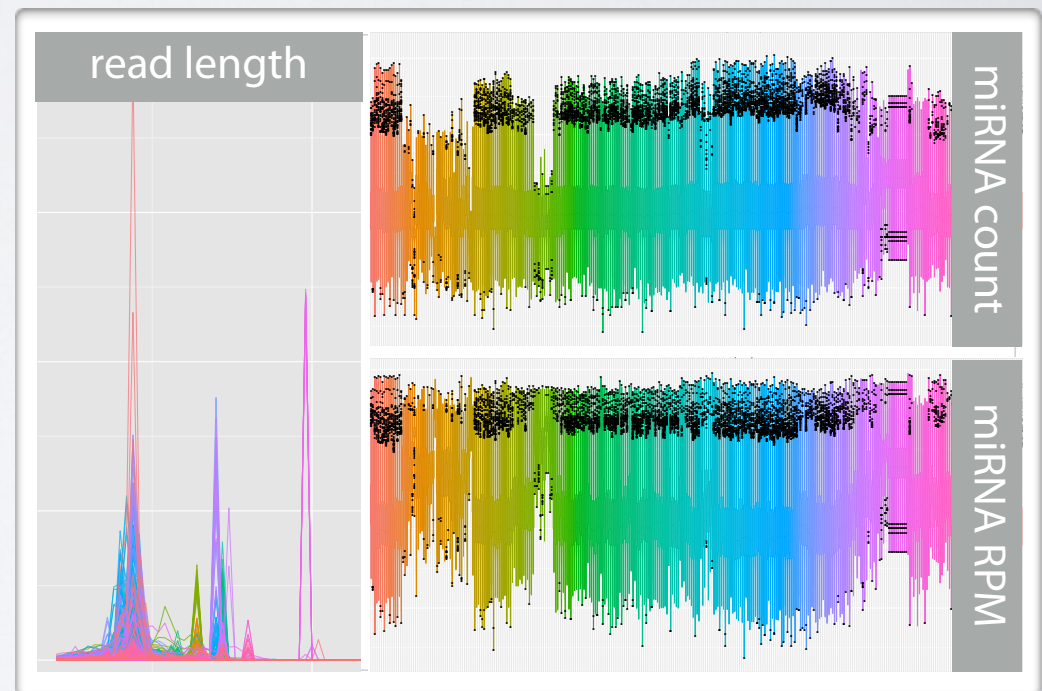
Output Targets

exRNA example

- extremely **simple to use** (1 input, 1 output)
- can process **multiple samples** in parallel
- very **customisable** (choice of smRNA libs, calibrators, etc)

downstream analysis

- exceRpt output from **multiple samples can be combined**, compared, and tested for differential abundance
- features:
 - intuitive **QC** visualisations
 - **filtering & normalisation**
 - differential expression
- processed expression data are **excel, R, matlab, & geneSpring** compatible



agenda

1. **exceRpt** smallRNA-seq analysis suite
 - filtering, QC, and alignment
 - quantification & normalisation
 - visualisations
 - differential expression (soon)
2. **use-case** involving 345 samples
3. **downstream analysis**
 - miRNA-mRNA targets from miRTarBase
 - network analysis
 - enrichment analysis
 - pathway analysis

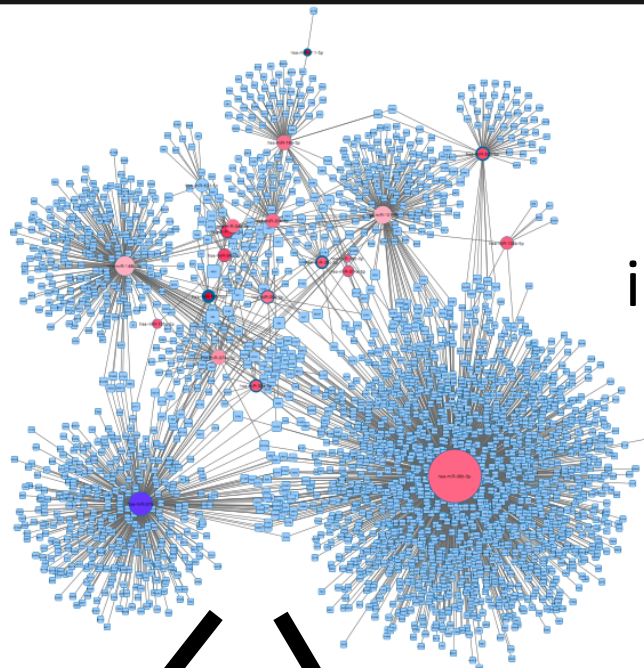


Downstream Analyses: Overview

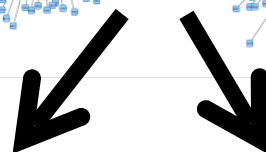
Genoree

miRNA	Accession	log2(Fold Change)	p-value
hsa-miR-325-3p	miR-325-3p	2.0670239	0.000136
hsa-miR-29b-3p	miR-29b-3p	1.954569	0.000136
hsa-miR-215-3p	miR-215-3p	1.933544	0.000136
hsa-miR-30c-3p	miR-30c-3p	1.877117	0.000136
hsa-miR-30b-3p	miR-30b-3p	1.7753465	0.000136
hsa-miR-654-3p	miR-654-3p	1.7056933	0.000136
hsa-miR-34c-3p	miR-34c-3p	1.6528728	0.000136
hsa-miR-29a-3p	miR-29a-3p	1.555888	0.000136
hsa-miR-138	miR-138	1.5294532	0.000136
hsa-miR-7-1-3p	miR-7-1-3p	1.4732764	0.000136
hsa-miR-1247-3p	miR-1247-3p	1.3954387	0.000136
hsa-miR-374b-3p	miR-374b-3p	1.3885012	0.000136
hsa-miR-122	miR-122	1.3745191	0.000136
hsa-miR-7-2-3p	miR-7-2-3p	1.3332179	0.000136
hsa-miR-135a-3p	miR-135a-3p	1.3189223	0.000136
hsa-miR-135b-3p	miR-135b-3p	1.3170213	0.000136
hsa-miR-135c-3p	miR-135c-3p	1.3169861	0.000136
hsa-miR-148b-3p	miR-148b-3p	1.3144326	0.000136
hsa-miR-148a-3p	miR-148a-3p	1.3137757	0.000136
hsa-miR-204-3p	miR-204-3p	1.3118857	0.000136
hsa-miR-987c-3p	miR-987c-3p	1.3118857	0.000136
hsa-miR-296	miR-296	1.3109332	0.000136
hsa-miR-987b-3p	miR-987b-3p	1.3109332	0.000136
hsa-miR-121-3p	miR-121-3p	1.3109332	0.000136
hsa-miR-30b-3p	miR-30b-3p	1.3109332	0.000136
hsa-miR-923	miR-923	1.3109332	0.000136

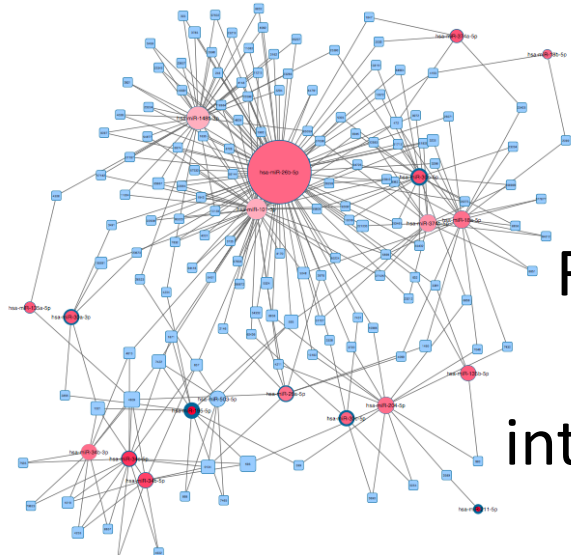
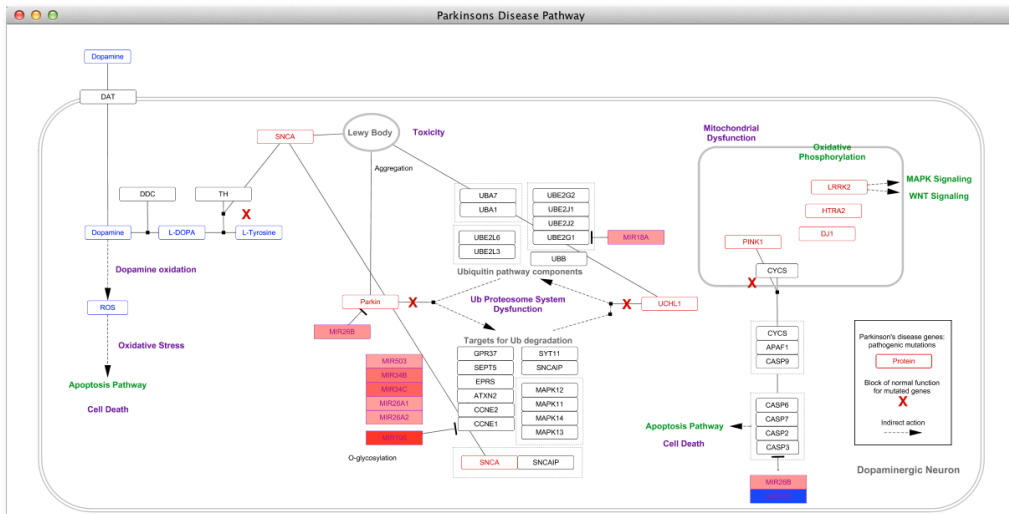
PD vs PDD diff exp
Burgos K., et al. (2014) PLoS ONE



All known interactions



Pathways



Focused set of interactions



1. Query Interactions: Genboree

Home Workbench Browser Profile Groups Projects Databases Tools Log Out Help [Genboree Home](#)

System/Network Data Genome Transcriptome Cistrome Epigenome Metagenome **Visualization** Help

Welcome to the Genboree Workbench! [Watch Intro Video](#)

Data Selector

Refresh Data Filter: Select a filter...

- Epigenome Informatics Demo Output Data
- Epigenome ToolSet Demo Input Data
- Epigenomics Roadmap Repository
- Examples and Test Data
- exRNA
 - exRNA KJENS1
 - Databases
 - ADPDb
 - Tracks
 - Lists & Selections
 - Lists of Files
 - AD CSF first40 results.zip
 - AD CSF results.zip
 - SampleSets
 - Samples
 - Files
 - OriginalDiffExpResults
 - KJ_diffExp_InputList.txt
 - PostProcessedRuns
 - ProcessPipelineRuns
 - smallRNAseqPipeline
 - Projects
 - exRNA Metadata Standards

Details

Attribute	Value
Group	exRNA KJENS1
Role	author
Name	ADPDb
Description	Database for the Alzheimer's and Parkinson's Disease Project from Kendall Jensen's lab at TGen - ERCC
Species	Homo sapiens

Input Data

↑ ↓ ✕ 🗑

- KJ_diffExp_InputList.txt

Output Targets

↑ ↓ ✕ 🗑

© 2001-2015
Bioinformatics Research Laboratory
 400D Jewish Wing, MS:BCM225
 1 Baylor Plaza
 Houston, TX 77030

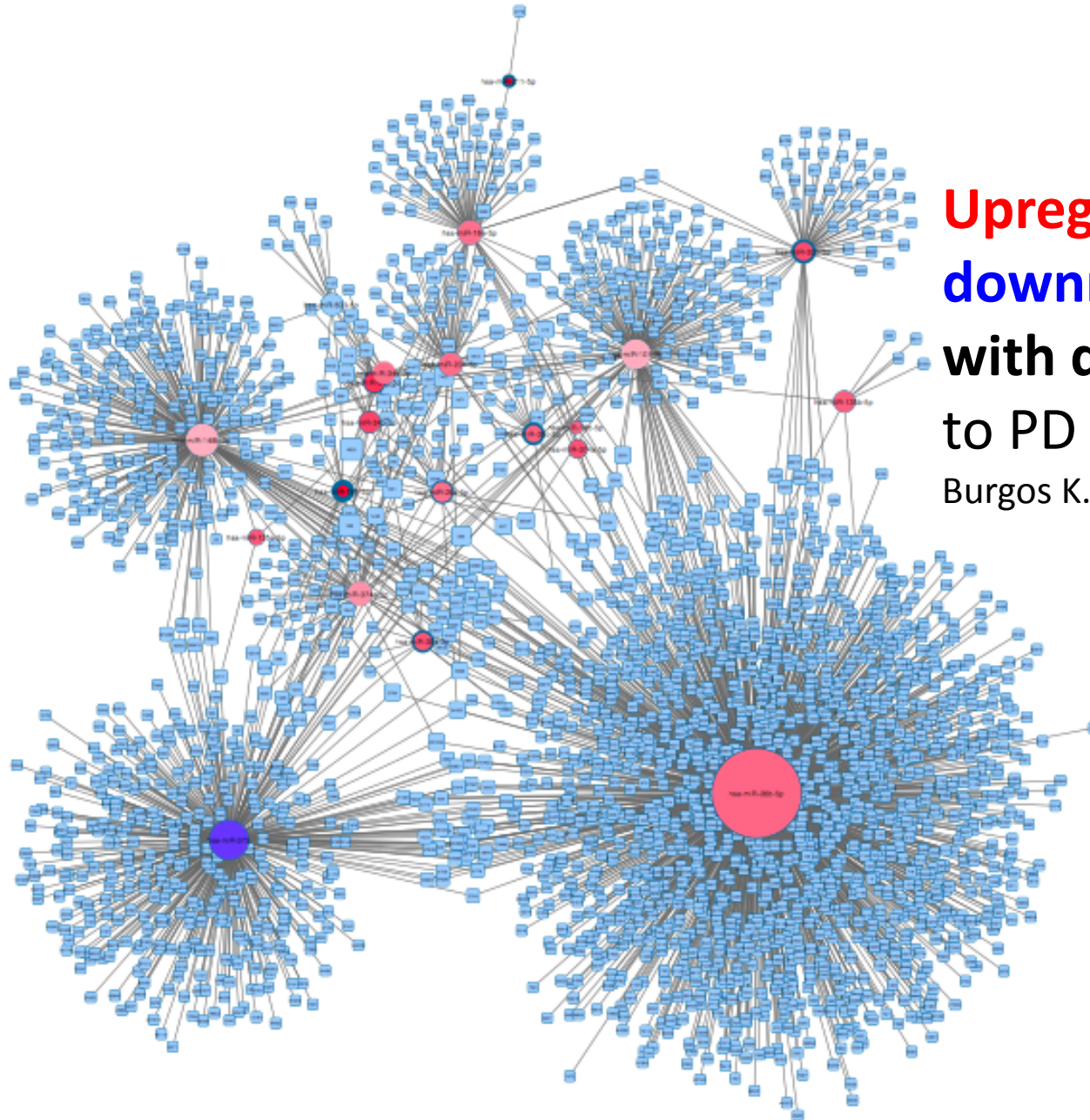
Genboree is built & maintained by the **Bioinformatics Research Laboratory** at **Baylor College of Medicine**.

Genboree is a hosted service. Code is available **free for academic use**.

[Questions or comments?](#)



2. Network Analysis: Cytoscape

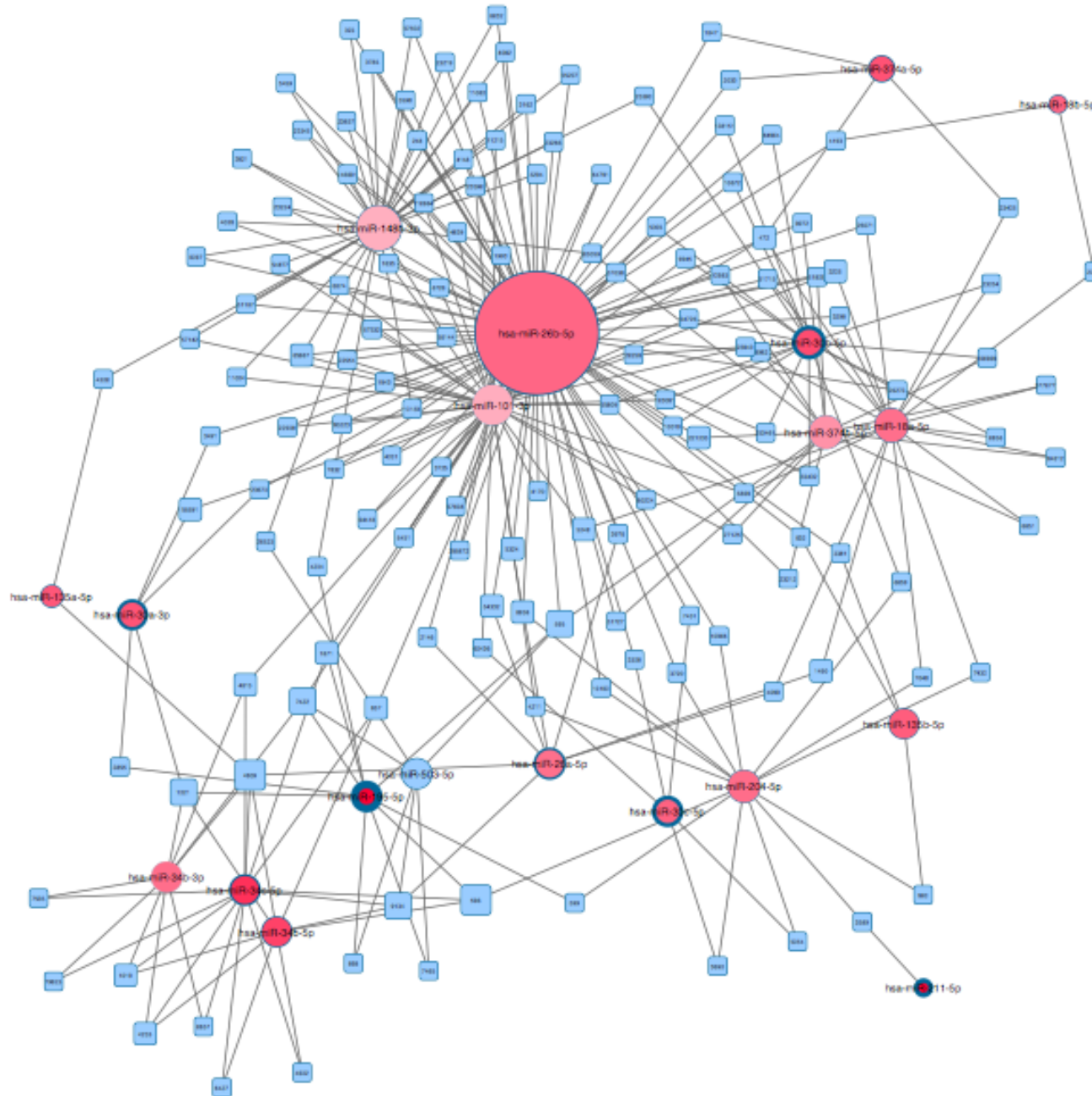


Upregulated and **downregulated** in **PD** with dementia relative to PD alone.

Burgos K., et al. (2014) PLoS ONE

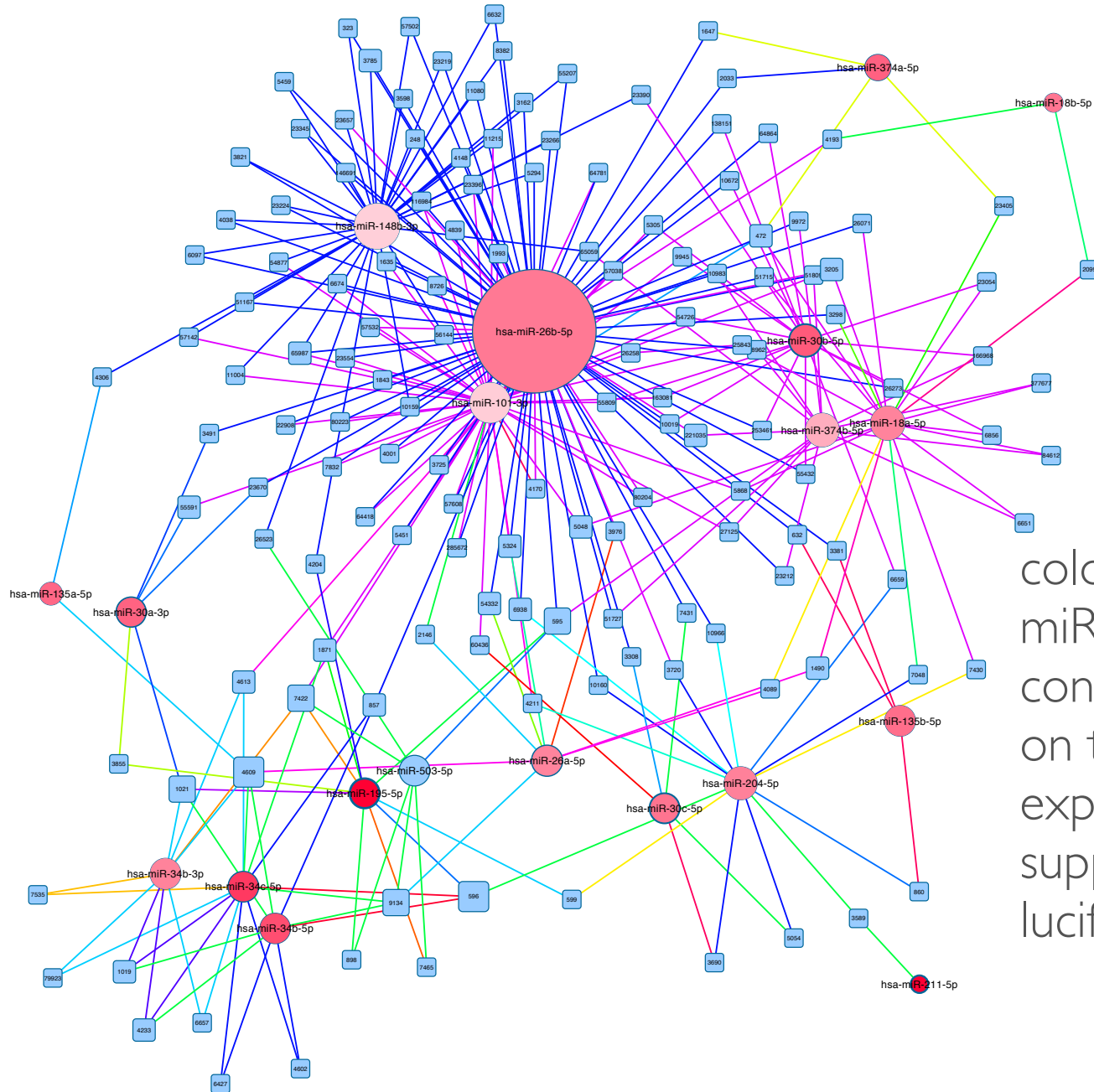


2. Network Analysis: Cytoscape





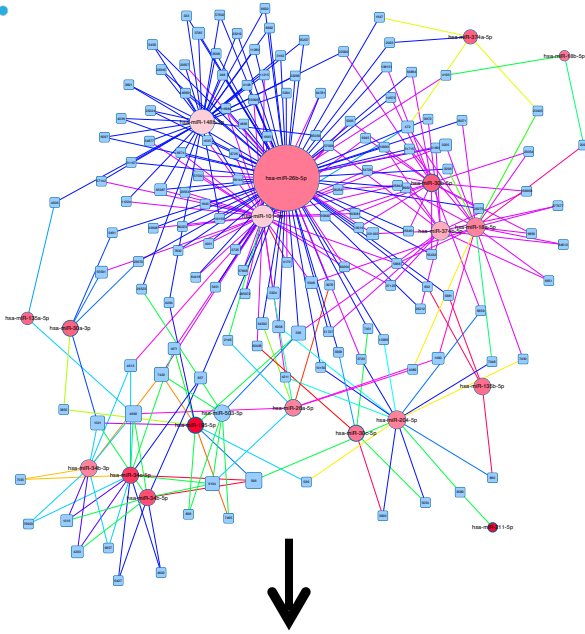
2. Network Analysis: Cytoscape



colour or filter
miRNA-mRNA
connections based
on type of
experimental
support e.g.
luciferase, CLIP, etc.



3. Enrichment Analysis: BiNGO



3.0+

BiNGO

BINGO

Calculates overrepresented GO terms in the network and display them as a network of significant GO terms.

★★★★☆ (54) 19530 downloads

GO-ID	Description	...	corr p-val	cluster freq
<input type="checkbox"/> 8284	positive regulation of cell proliferation	...	6.3743E-7	22/135 16.2%
<input type="checkbox"/> 30154	cell differentiation	...	2.4311E-6	41/135 30.3%
<input type="checkbox"/> 50793	regulation of developmental process	...	2.4311E-6	27/135 20.0%
<input type="checkbox"/> 48856	anatomical structure development	...	2.4311E-6	54/135 40.0%
<input type="checkbox"/> 48869	cellular developmental process	...	3.3453E-6	41/135 30.3%
<input type="checkbox"/> 43627	response to estrogen stimulus	...	3.3453E-6	11/135 8.1%
<input type="checkbox"/> 32502	developmental process	...	4.0164E-6	60/135 44.4%
<input type="checkbox"/> 42127	regulation of cell proliferation	...	5.0150E-6	27/135 20.0%
<input type="checkbox"/> 7275	multicellular organismal development	...	7.8071E-6	56/135 41.4%
<input type="checkbox"/> 48731	system development	...	9.0406E-6	49/135 36.2%
<input type="checkbox"/> 31960	response to corticosteroid stimulus	...	9.1021E-6	10/135 7.4%
<input type="checkbox"/> 48646	anatomical structure formation involved in morphogenesis	...	1.4801E-5	17/135 12.5%
<input type="checkbox"/> 48513	organ development	...	1.5414E-5	40/135 29.6%
<input type="checkbox"/> 48754	branching morphogenesis of a tube	...	2.4400E-5	9/135 6.6%



4. Pathway Analysis: WikiPathways

pathway discussion edit history delete protect unwatch permissions log out

Alexander Pico, Kristina Hanspers, Egon Willighagen

Enter node name to highlight

Lewy Body Toxicity Aggregation

Mitochondrial Dysfunction Oxidative Phosphorylation

MAPK WNT

Ubiquitin pathway components

Ub Proteasome System Dysfunction

Targets for Ub degradation

Parkin UCHL1

Parkin's disease genes: pathogenic mutations

Block of normal function for mutated genes

Indirect action

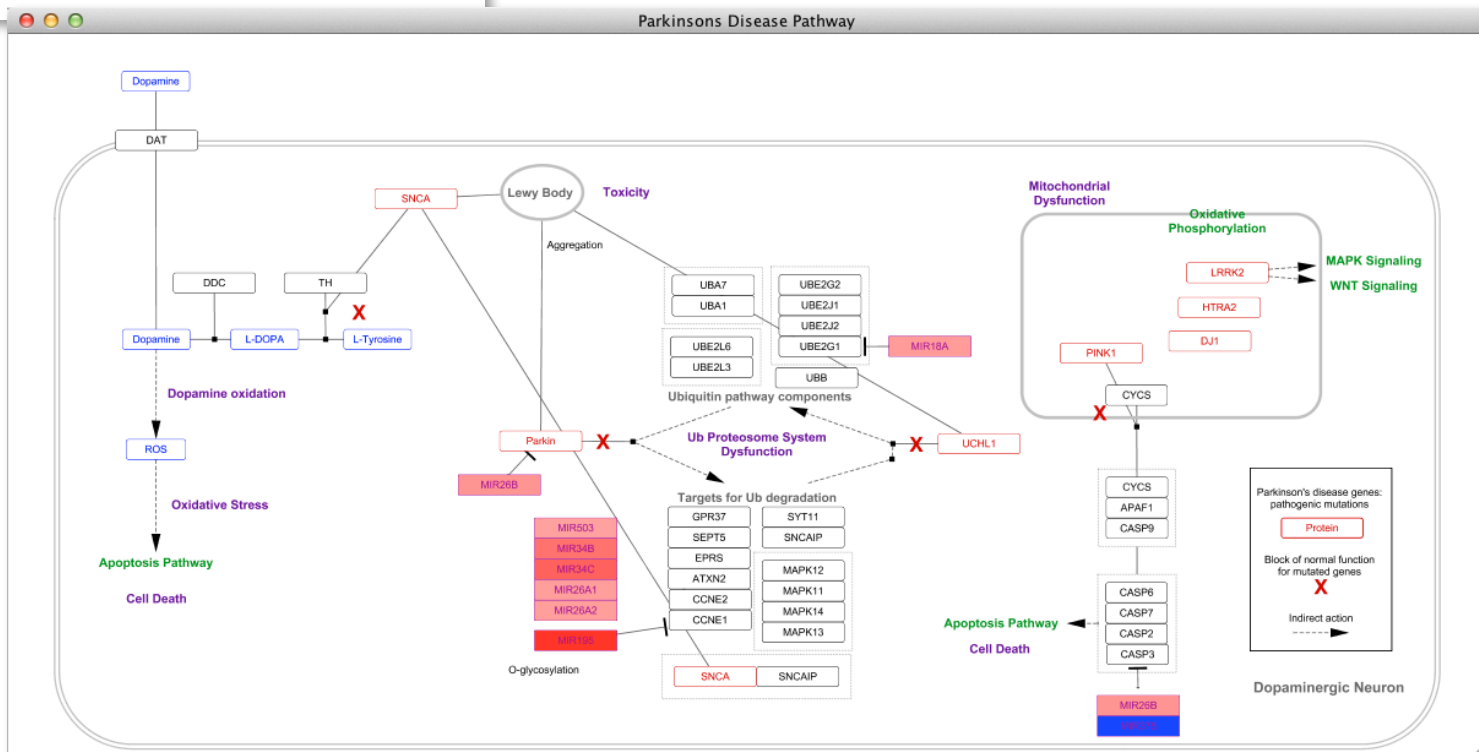
Download

wikipathways.org

WikiPathways 3.0+

WikiPathways web service client and GPML file format importer

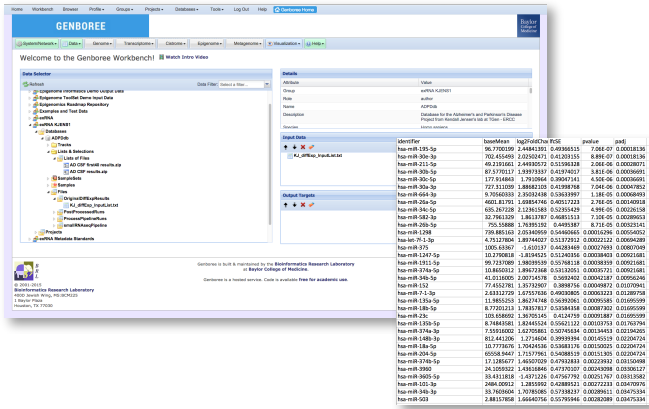
★★★★★ (8) 5548 downloads



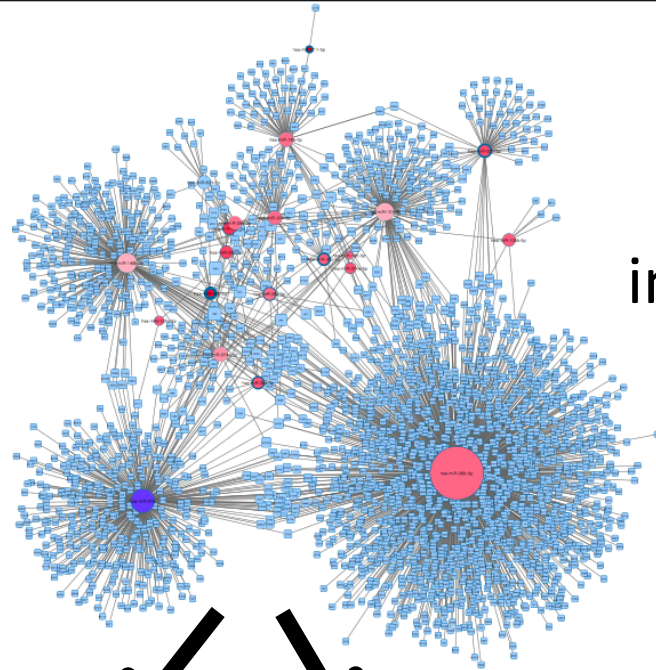


Downstream Analyses: Summary

Genoree



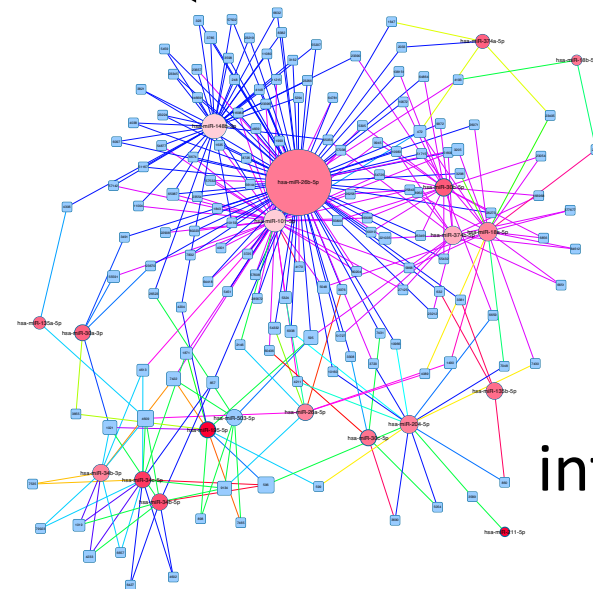
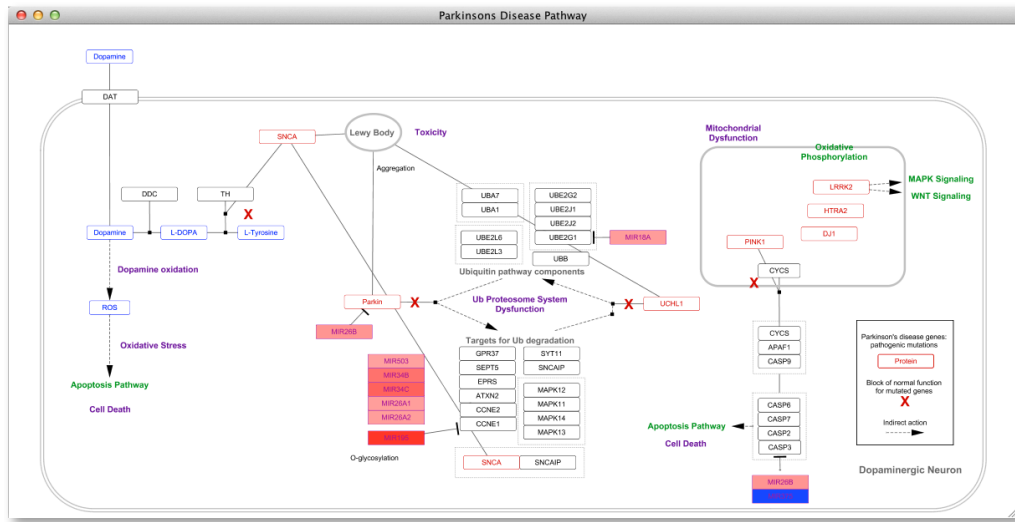
PD vs PDD diff exp
Burgos K., et al. (2014) PLoS ONE



All known interactions



Pathways



Focused set of interactions

summary

- trivial to use **exceRpt** thanks to:
 - automated, linear workflow
 - GUI, processing, & data/result storage on Genboree
- average runtime for high quality RNA-seq sample:
 - **1hr** for **endogenous**-only analysis
 - **2hr 30mins** for endogenous+**exogenous** alignments
- downstream **network analysis tools directly integrated** in genboree and automatically accept exceRpt results

genboree.org

exRNA.org

Acknowledgements



Mark Gerstein



Aleks Milosavljevic



Joel Rozowsky



**National Institutes
of Health**



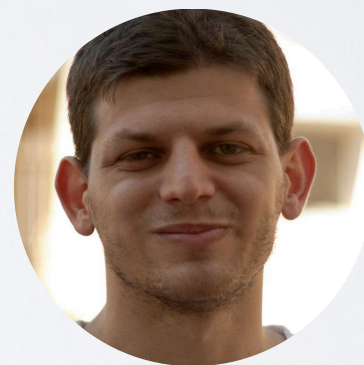
Matt Roth



**Sai Lakshmi
Subramanian**



**William
Thistlethwaite**



**Fabio
Navarro**



Alex Pico