

# CCR Collaborative Bioinformatics Resource (CCBR)

---

INTRODUCTION TO NGS DATA ANALYSIS

BTEP SERIES 2017

# bioinformatics.cancer.gov

**Bioinformatics @ CCR**

Home **Ask for Help** Training & Education CCBR Publications

## Bioinformatics Support at CCR

The CCR Collaborative Bioinformatics Resource (CCBR) is an organizational umbrella which provides a mechanism for CCR researchers to obtain many different types of bioinformatics assistance to further their research...

[Read More](#)

### About CCBR

[View](#) [Edit](#) [Revisions](#)

- › Who we are
- › How CCBR Works
- › Pipeliner: for analysis of Exome-Seq, Genome-Seq and RNA-Seq data
- › NGS Experimental Design: Best Practices

## CCR Collaborative Bioinformatics Resource (CCBR)

The CCR Collaborative Bioinformatics Resource (CCBR) is a resource group which provides a mechanism for CCR researchers to obtain many different types of bioinformatics assistance to further their research goals. The group has expertise in a broad range of bioinformatics topics, and as such, its goal is to provide a simplified central access point for CCR researchers.

The CCBR group includes members of the CCR Office of Science and Technology Resources (OSTR), Frederick National Laboratory for Cancer Research (FNLCR) and the Center for Biomedical Informatics and Information Technology (CBIIIT). The CCBR may also direct projects to other available CCR bioinformaticians as needs demand. Requests for any type of Bioinformatics support should be through the [CCBR Project Submission Form](#).

# CCBR support includes:

---

Consulting on experimental design, help with analysis and interpretation of biological data produced by large-scale genomics technologies including Next-generation sequencing (RNA-Seq, Exome-Seq, ChIP-Seq, Whole genome Sequencing), and microarrays

Support for the development of methods for new technologies provided by the Office of Science and Technology Resources (OSTR)

Provide training classes to CCR scientists focusing on software used in the analysis of their own data

# CCBR Members

---

## Office of Science and Technology Resources (OSTR)

*Maggie Cam (Head)*

## Center for Biomedical Informatics and Information Technology (CBIIT)

*Chunhua Yan*

*Ying Hu*

*Richard Finney*

## Frederick National Laboratory of Cancer Research (Leidos)

*Parthav Jailwala (Manager)*

*Fathi Elloumi*

*Justin Lack*

*Bong-Hyun Kim*

*George Nelson*

*Alexei Lobanov*

*Jack Chen*

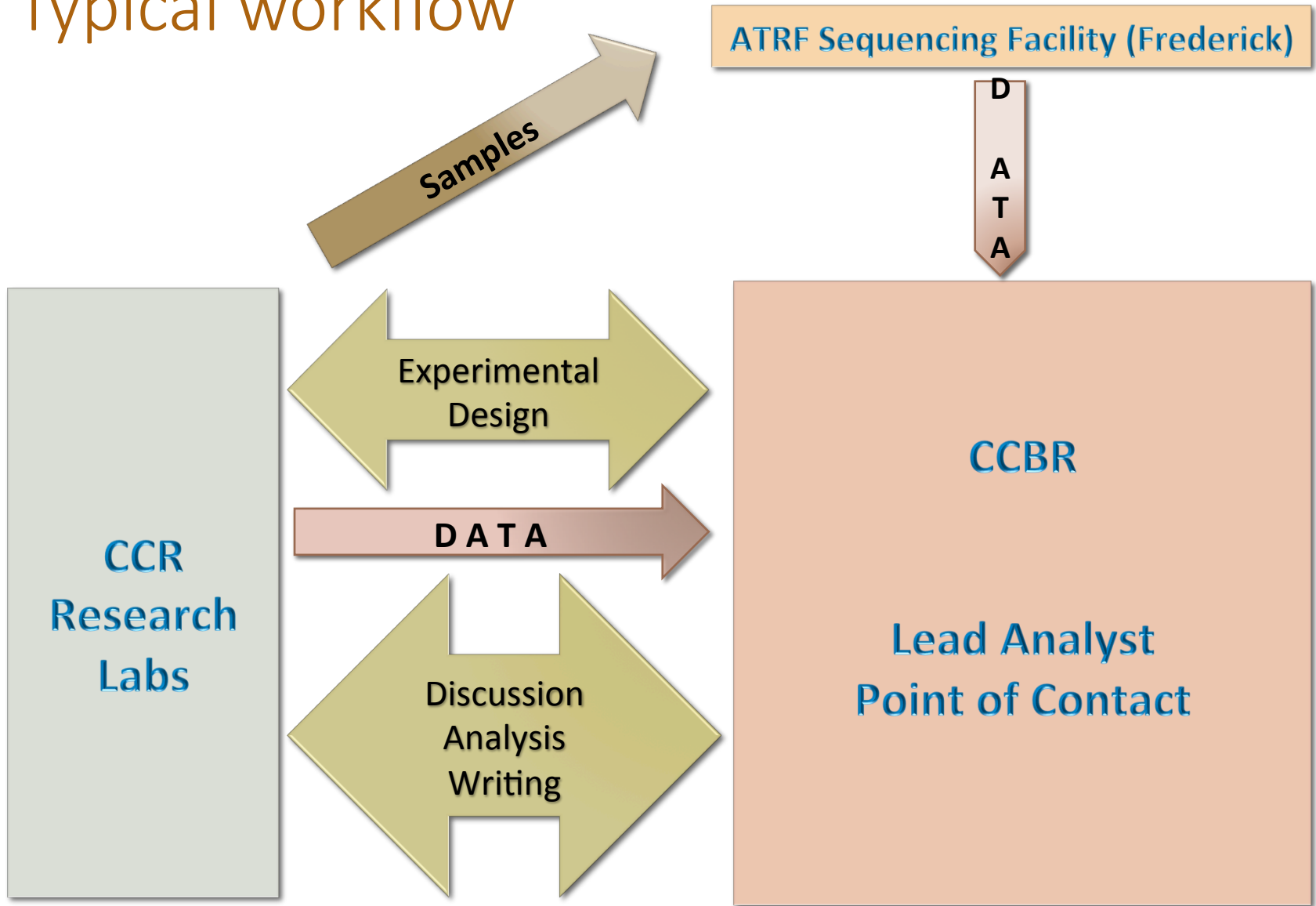
*Ashley Walton*

*Vishal Koparde*

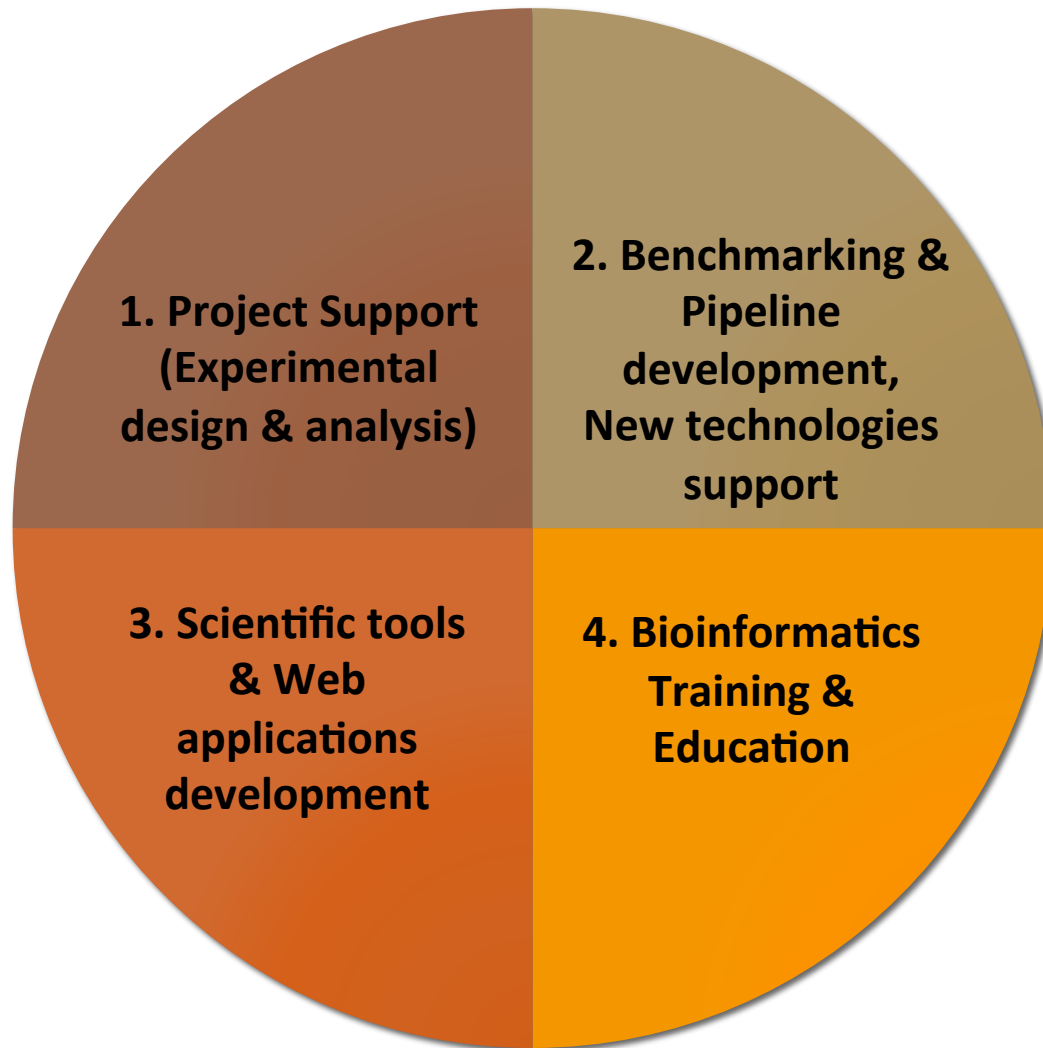
***Soon to be part of CDSL (CCR Cancer Data Science Lab)***



# Typical workflow



# CCBR Functional Areas



# 1. Project Support

---

# Current CCBR Projects

Total Number of	Total Number
Principal Investigators	95
Projects	189

## **Data Mining:**

*Analyzing public microarray data for Merkel cell carcinoma*

## **Basic/Bench Research:**

*Study of alternative splicing function of Rbfox1 in knock out and transgenic mice*  
*Disruption of Pol II Elongation with O-GlcNAcylation Inhibitors*

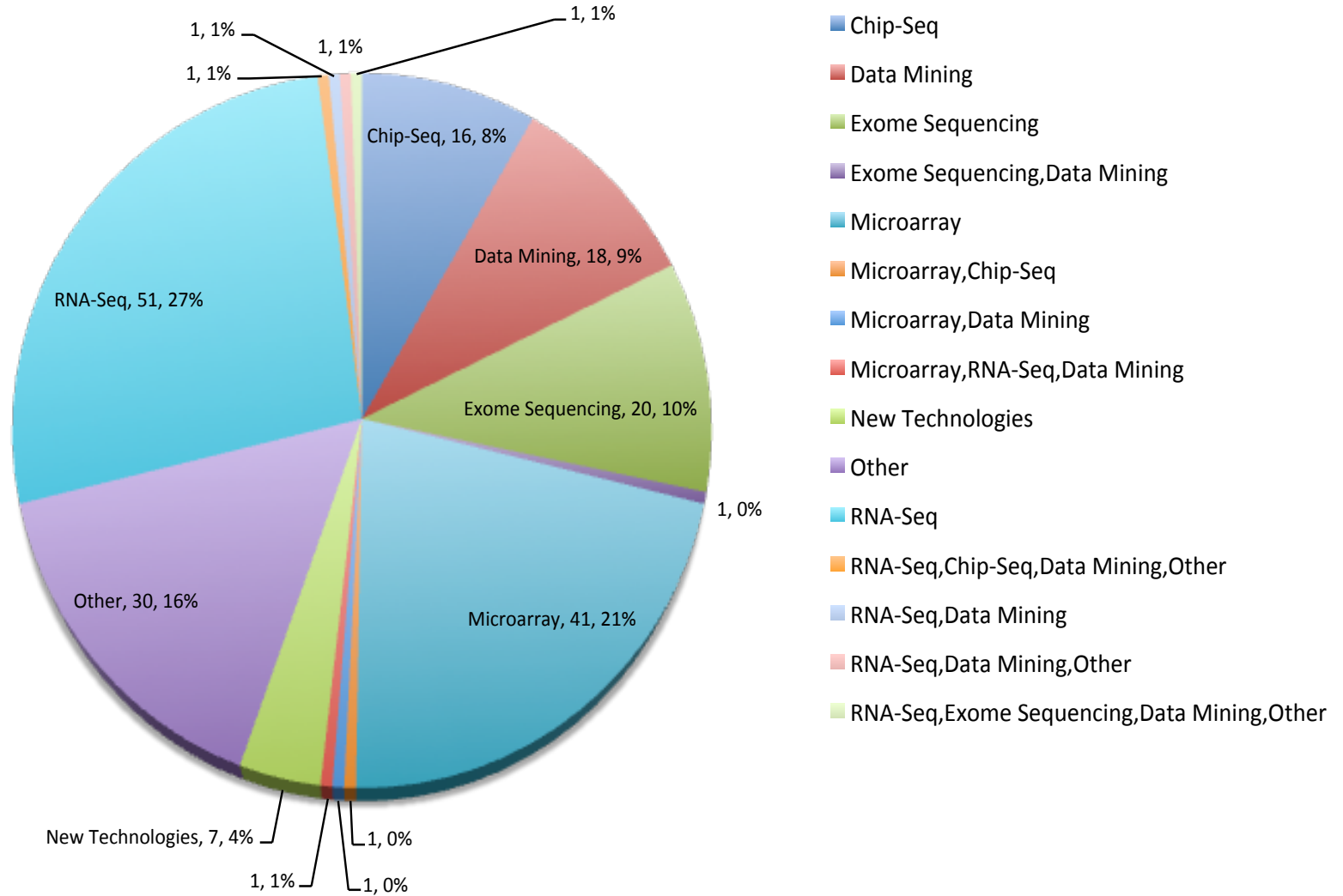
## **Translational Research:**

*Genomic characterization of mouse model for GBM*  
*Gene expression analysis of Kras-induced lung cancer mouse model*

## **Clinical Research:**

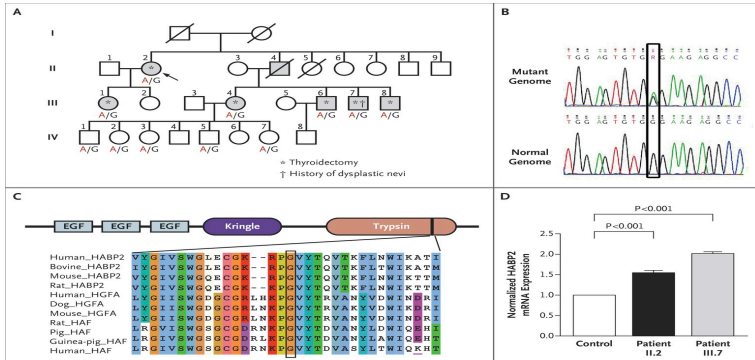
*Germline Exome-Seq analysis of Familial Non-Medullary Thyroid Cancer*  
*Exome-Seq analysis of adrenocortical cancer germ line and tumor DNA*

## Breakdown of Project Types

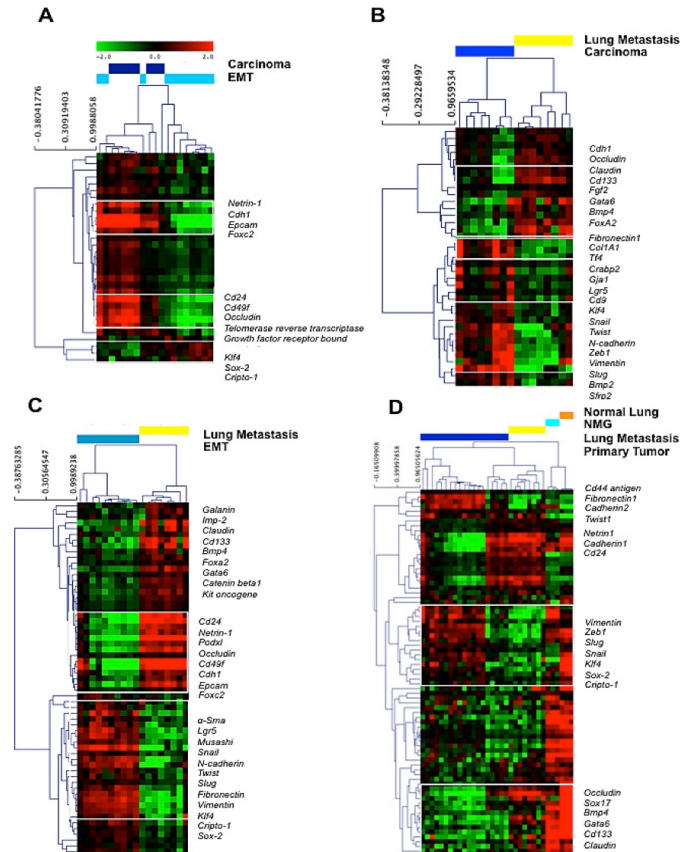


# Some Projects

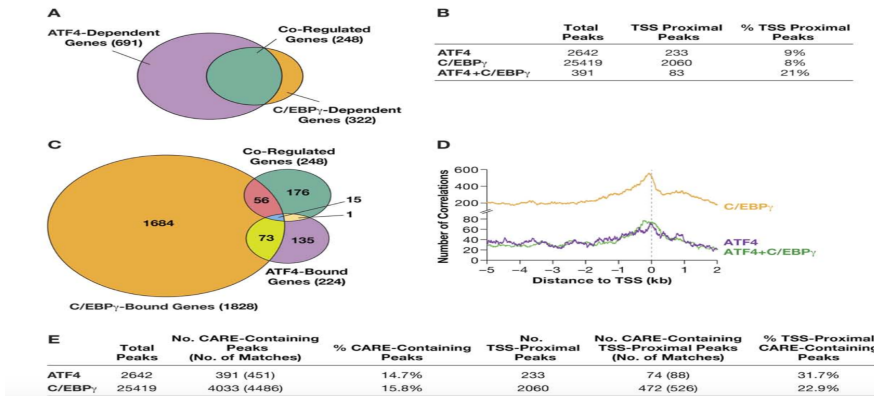
## Germline mutations – Exome Seq Analysis



## Gene Expression Analysis



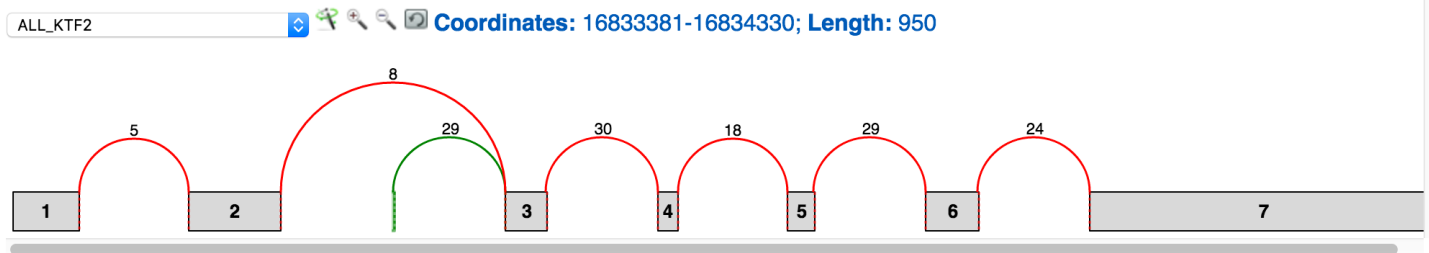
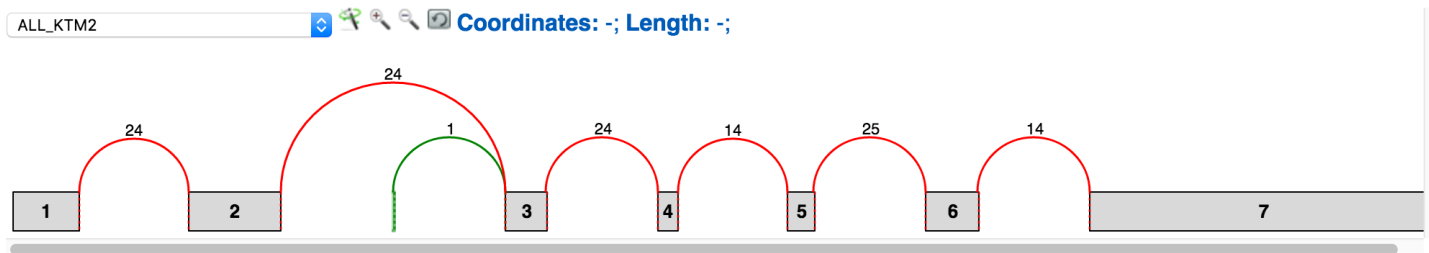
## Integrated CHIP-Seq & Microarray Analysis

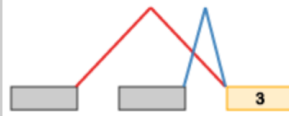

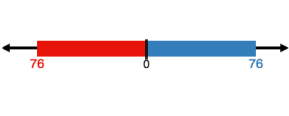






# Splice Variant Analysis

Gene name: **Tfec**; chr6:-:16833381-16898441;   
 Gene ID: **ENSMUSG00000029553**;

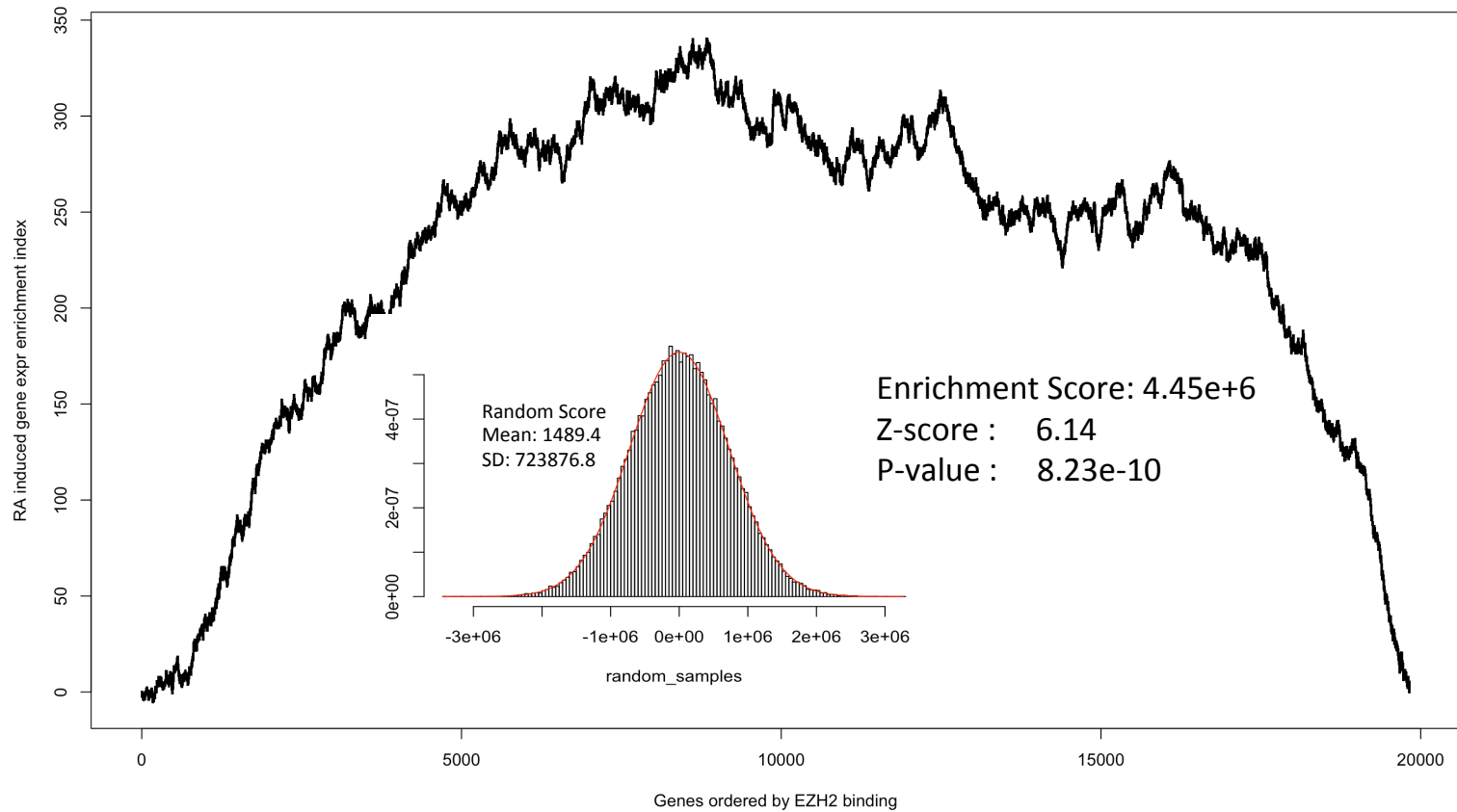


LSV ID	LSV Type	PSI KTM2	← More in KTM2   More in KTF2 →	PSI KTF2	LSV links
ENSMUSG00000029553:16845364-16845478:source					
LSV ID	LSV Type	PSI KTM2	← More in KTM2   More in KTF2 →	PSI KTF2	LSV links

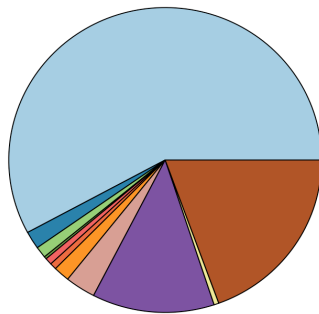
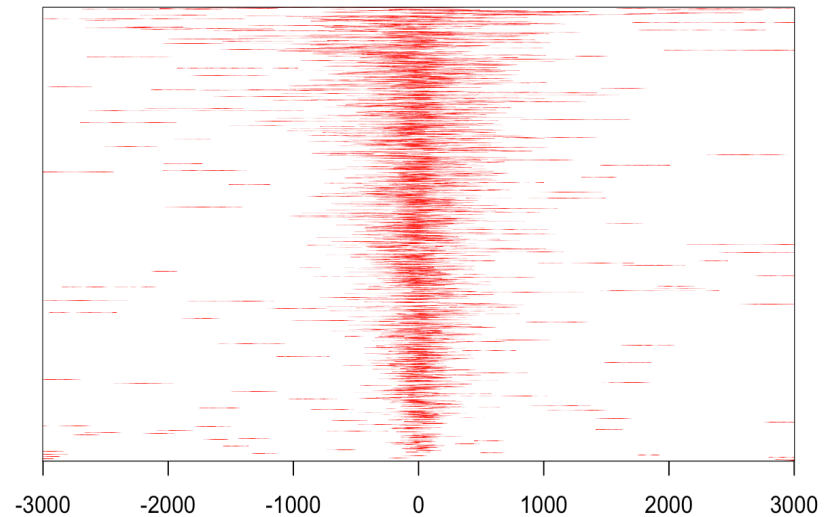
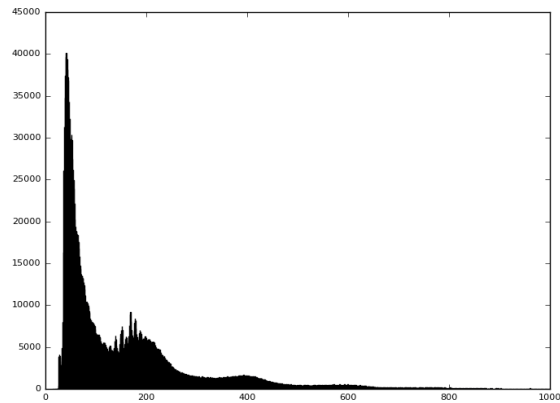
  1/1   10 



# Integrated ChIP-Seq/RNA-Seq Analysis

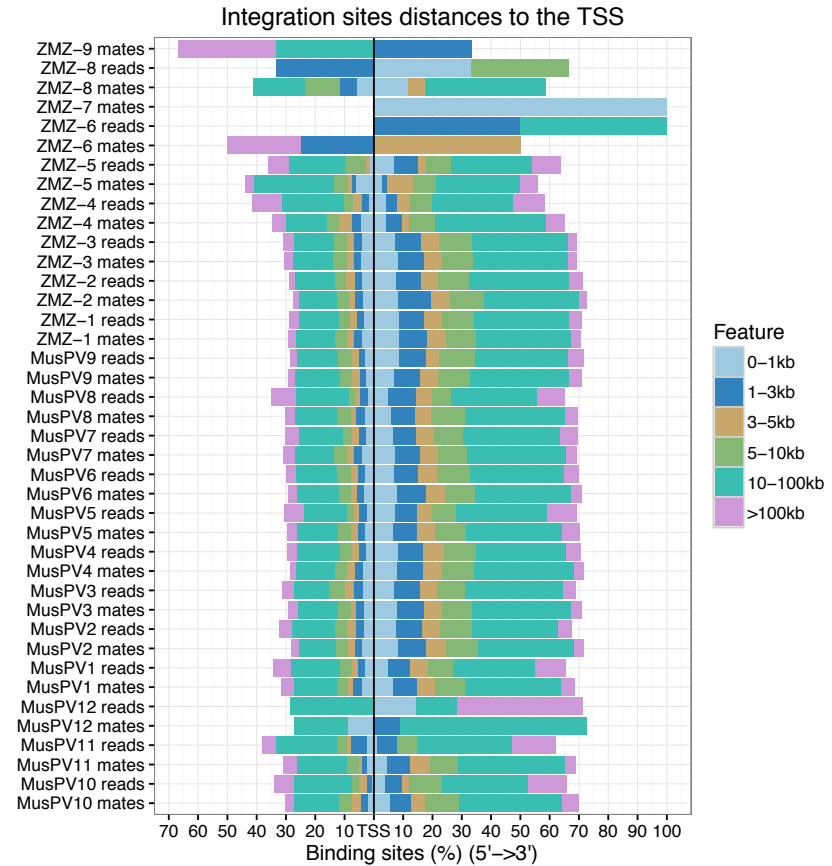
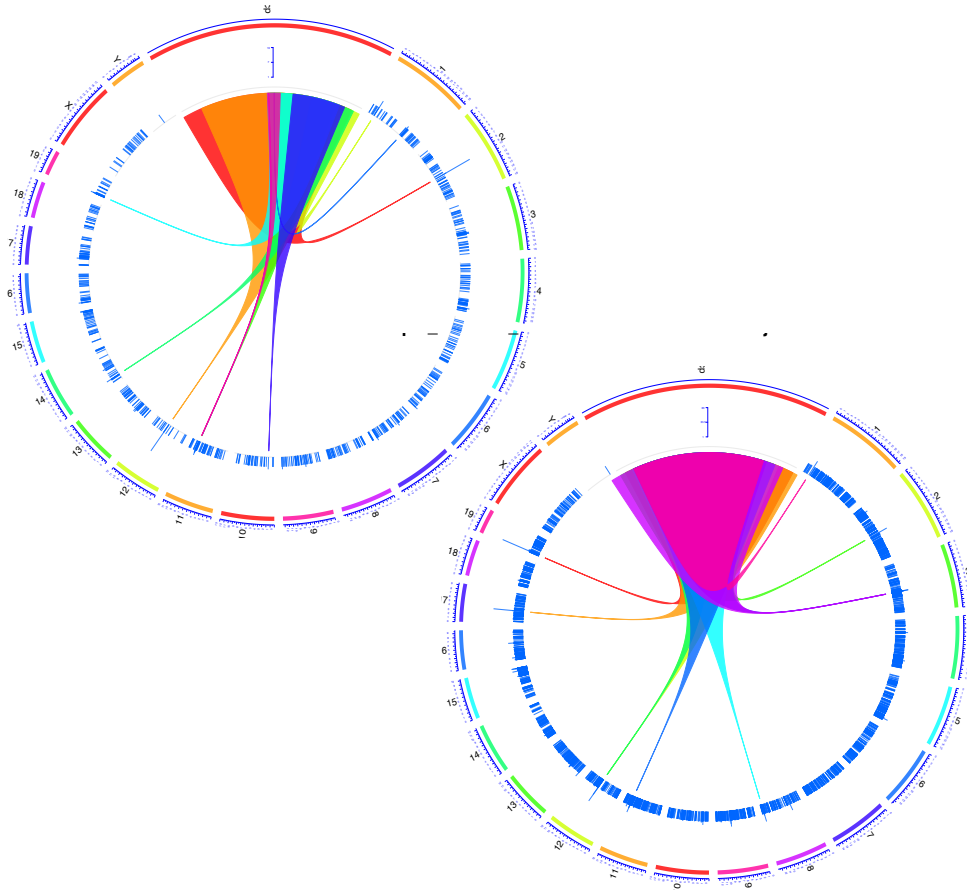


# Chip-Seq/ATAC-Seq Project



- Promoter (<=1kb) (57.88%)
- Promoter (1-2kb) (1.88%)
- Promoter (2-3kb) (1.2%)
- 5' UTR (0.27%)
- 3' UTR (0.71%)
- 1st Exon (0.69%)
- Other Exon (1.66%)
- 1st Intron (3.14%)
- Other Intron (12.64%)
- Downstream (<=3kb) (0.5%)
- Distal Intergenic (19.43%)

# Virus Integration Analysis



# Collaboration Website: iMeetCentral

The screenshot shows the iMeetCentral interface for project CCB-769. At the top, there is a blue navigation bar with the CCB-769 logo, a search bar for workspaces and files, and user options like 'Create', 'Custom Menu', and 'Maggie Cam'. Below this is a breadcrumb trail: Dashboard > Workspaces > CCB-769. The main navigation menu includes Home, Project Details, Communications, Results, Upload and View Files, Project Management, Settings, People, Properties, and an Add Tab button. The page title is 'CCBR-769 Home' with an 'Options' dropdown. The main heading is 'RNASeq Analysis of Inherited Kidney Cancer'. A 'CCBR-769 People' section shows a row of user avatars and a 'Manage' button. A 'Recent activity in "CCBR-769"' section features a comment by Parthav Jaiwala about a meeting. A 'Communications' section lists three recent posts: 'Next meeting: Discuss RNASeq results' (Nov 28, 2016), 'Meeting to discuss Initial QC results' (Oct 12, 2016), and 'Contrasts' (Sep 8, 2016). A detailed view of the 'Next meeting' post is shown on the right, including the text of the post and a comment by Christopher Ricketts.

CCBR-769 Home Options ▾

## RNASeq Analysis of Inherited Kidney Cancer

CCBR-769 People Manage

Recent activity in "CCBR-769" ↻

Parthav Jaiwala commented on [Next meeting: Discuss RNASeq results in CCB-769](#)  
7 weeks ago · [Comment](#) · [Like](#) · [Subscribe](#)

### Next meeting: Discuss RNASeq results

Hi Chris, Martin, As per our last meeting, we decided to review the clustering of samples by groups on a PCA plot, to decide which of the desired contrasts are meaningful to carry out. We have resumed analysis of your RNA-Seq samples and we should be ready to review the PCA results by the end of this week. If you are available...

**Christopher Ricketts**  
more  
1 month ago

**Parthav Jaiwala**  
more  
1 month ago

**Parthav Jaiwala**  
Hi Chris,  
Attached is the sample metadata spreadsheet. Please add information on the type of cell-line/sample (primary or metastatic status), as well as any other information about specific groups, that is relevant to the analyses.  
Early next week, I will send across the revised PCA plots, sample-to-sampl... [more](#)  
1 month ago

To send email messages to members, start **new** or select topic in **Communications**:

### Communications New

**Next meeting: Discuss RNASeq results** Nov 28, 2016  
Partha Jaiwala  
Hi Chris, Martin, As per our last meeting, we decided to review the clustering of samples by groups on a PCA plot, to decide which of the desired contrasts are meaningful to carry out. We have resumed analysis of your RNA-Seq samples and we should be ready to review the PCA results by the end of this week. If you are available this Friday (12/02), how about we have a meeting to discuss the project further? I am available either 10am or 11am, so please let me know whatever works at your end. Th...  
[3 comments](#)

**Meeting to discuss Initial QC results** Oct 12, 2016  
Partha Jaiwala  
Hi Chris, Martin: We have completed Initial QC of your RNA-Seq samples for this project. How about we have a meeting to review these results and plan the next steps. Please let us know if you are available tomorrow (Thursday) at 3pm to discuss these results. If tomorrow does not work at your end, we are also available to meet Friday afternoon at 2pm or 3pm. Thanks, Parthav  
[2 comments](#)

**Contrasts** Sep 8, 2016  
Partha  
Hi Chris, Thanks. Yes, a total of 24 comparisons (contrasts) may be overwhelming for downstream analyses and interpretation of results. The best way we can decide if some of these phenotypes should be combined, is by looking at how close/distant these

## 2016

- McCullen MV, Li H, **Cam M**, Sen S, McVicar DW, Anderson SK. [Ly49 Pro1 element activity is associated with gene activation, not gene expression: Pro 1 does not function as an enhancer in the Ly49 genes and Pro1 transcripts are not present in mature Ly49-expressing NK cells.](#) *Genes and Immunity advance online publication 28 July 2016; doi: 10.1038/gene.2016.31*
- Weyemi U, Redon CE, Sethi TK, Burrell AS, **Jailwala P, Kasoji M, Abrams N, Merchant A**, Bonner WM. [Twist1 and Slug mediate H2AX-regulated Epithelial-Mesenchymal Transition in breast cells.](#) *Cell Cycle. 2016 Jun 17:0. [Epub ahead of print]*
- Matter MS, Marquardt JU, Andersen JB, Quintavalle C, Korokhov N, Stauffer JK, Kaji K, Decaens T, Quagliata L, **Eloumi F**, Hoang T, Molinolo A, Conner EA, Weber A, Heikenwalder M, Factor VM, Thorgeirsson SS. [Oncogenic driver genes and the inflammatory microenvironment dictate liver tumor phenotype.](#) *Hepatology. 2016 Jun;63(6):1888-99. doi: 10.1002/hep.28487. Epub 2016 Mar 15.*
- Mendoza-Villanueva D, Balamurugan K, Ali HR, Kim SR, Sharan S, Johnson RC, **Merchant AS**, Caldas C, Landberg G, Sterneck E. [The C/EBP \$\delta\$  protein is stabilized by estrogen receptor  \$\alpha\$  activity, inhibits SNAI2 expression and associates with good prognosis in breast cancer.](#) *Oncogene. 2016 May 16. doi: 10.1038/onc.2016.156. [Epub ahead of print].*
- Bae HR, Leung PS, Tsuneyama K, Valencia JC, Hodge DL, Kim S, Back T, Karwan M, **Merchant AS**, Baba N, Feng D, Park O, Gao B, Yang GX, Eric Gershwin M, Young HA. [Chronic Expression of Interferon Gamma Leads to Murine Autoimmune Cholangitis with a Female Predominance.](#) *Hepatology. 2016 May 14. doi: 10.1002/hep.28641. [Epub ahead of print].*
- Rothermel LD, Sabesan AC, Stephens DJ, Chandran SS, Paria BC, Srivastava AK, Somerville R, Wunderlich JR, Lee CC, Xi L, Pham TH, Raffeld M, **Jailwala P, Kasoji M**, Kammula US. [Identification of an Immunogenic Subset of Metastatic Uveal Melanoma.](#) *Clin Cancer Res. 22:2237-49, 2016.*
- Kennedy MW, Chalamalasetty RB, Thomas S, Garriock RJ, **Jailwala P**, Yamaguchi TP. [Sp5 and Sp8 recruit  \$\beta\$ -catenin and Tcf1-Lef1 to select enhancers to activate Wnt target gene transcription.](#) *Proc Natl Acad Sci U S A. 113:3545-50, 2016.*
- Weyemi U, Redon CE, Choudhuri R, Aziz T, Maeda D, Boufraqueh M, Parekh PR, Sethi TK, **Kasoji M, Abrams N, Merchant A**, Rajapakse VN, Bonner WM. [The histone variant H2A.X is a regulator of the epithelial-mesenchymal transition.](#) *Nat Commun 7:10711, 2016.*
- Dine JL, O'Sullivan CC, Voeller D, Greer YE, Chavez KJ, Conway CM, Sinclair S, Stone B, Amiri-Kordestani L, **Merchant AS**, Hewitt SM, Steinberg SM, Swain SM, Lipkowitz S. [The TRAIL receptor agonist drozitumab targets basal B triple-negative breast cancer cells that express vimentin and Axl.](#) *Breast Cancer Res Treat. 155:235-51, 2016.*
- Castro NP, **Merchant AS**, Saylor KL, Anver MR, Salomon DS, Golubeva YG. [Adaptation of Laser Microdissection Technique for the Study of a Spontaneous Metastatic Mammary Carcinoma Mouse Model by NanoString Technologies.](#) *PLoS One. 11:e0153270, 2016.*

# How to get started

---

Contact [maggie.cam@nih.gov](mailto:maggie.cam@nih.gov) or [CCBR@mail.nih.gov](mailto:CCBR@mail.nih.gov)

Drop by: Bldg 37, Rm 3041 (office hours 10-12am)

For significant help, trigger project request:  
[Bioinformatics.cancer.gov](http://Bioinformatics.cancer.gov) (“Ask for Help”)

Appointment: Discuss experimental design, analysis,  
goals and timelines

After data arrives: analyst is assigned to project and  
additional meeting(s) for further discussion

## 2. Research & Development

---

# R&D

*Joint effort of ATRF Sequencing Facility Bioinformatics Team and CCBR*

---

Benchmarking of current and new algorithms for routine use

- Exome-seq, RNA-Seq, CHIP-Seq, miR-Seq

Pipeline development

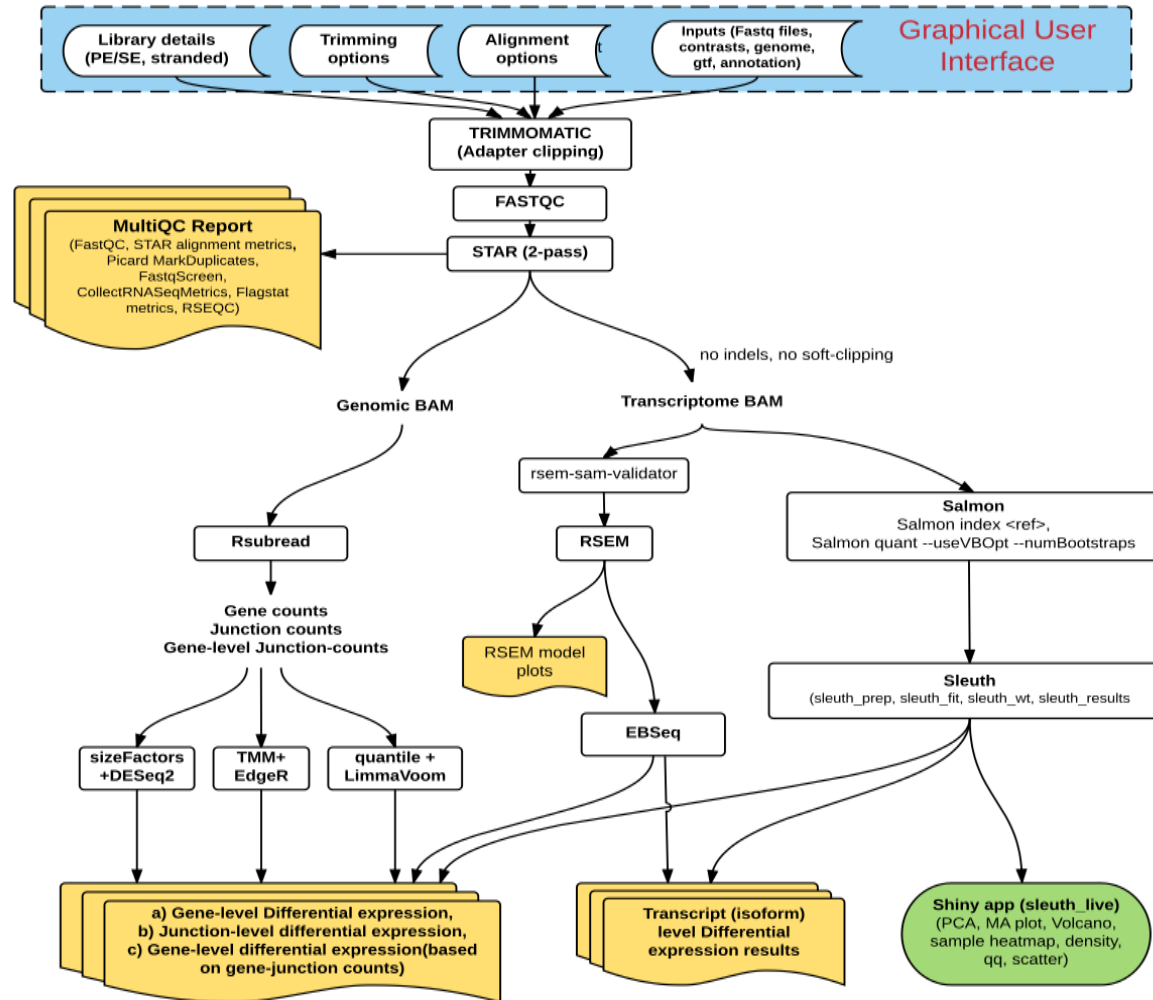
- Methods for standard workflows selected based on benchmarking
- Streamline upstream QC methods from ATRF and downstream analysis
- Standardize methods for reproducibility, updated versions as needed
- Publish in GitHub for CCR/NIH, also useful for others

Additional tool development

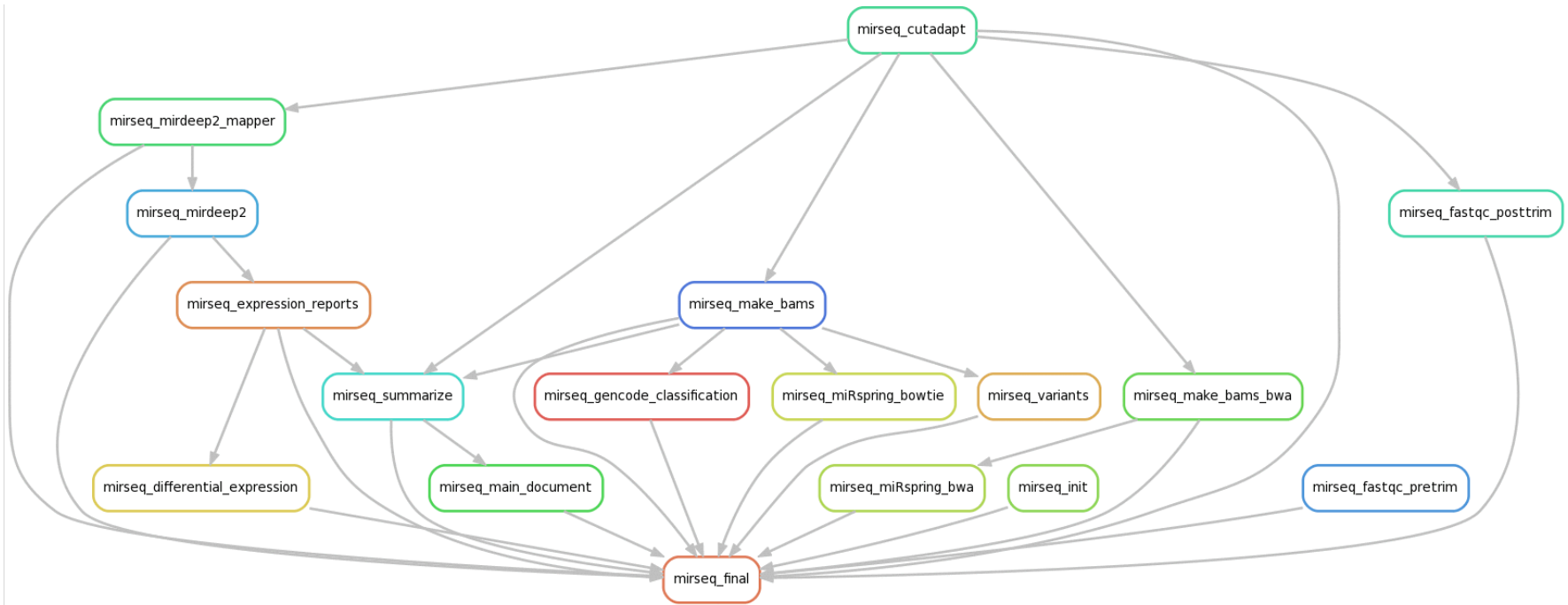
- Novel tools and algorithms arising from collaborative projects
- Published in GitHub and/or made available through web application



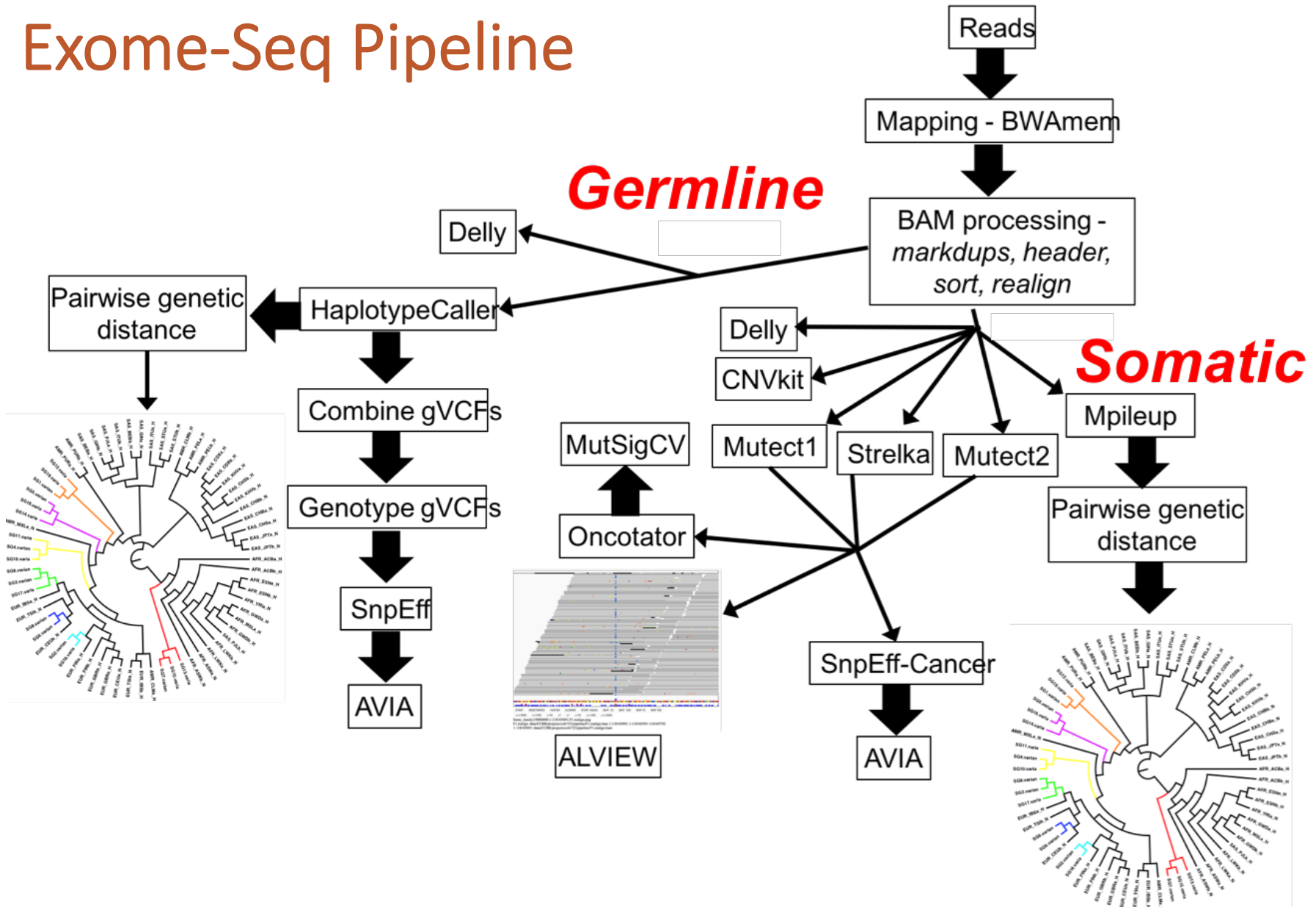
# RNA-Seq Pipeline



# miR-Seq Pipeline (coming soon)



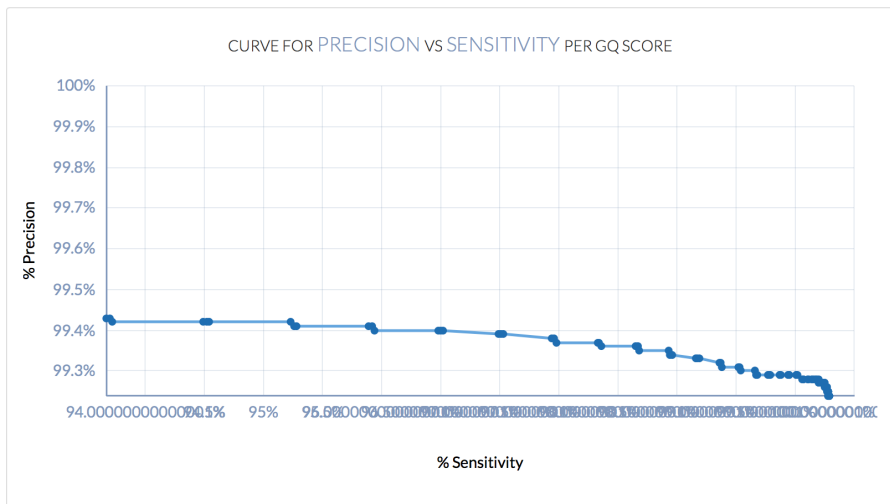
# Exome-Seq Pipeline



# Precision FDA Challenge

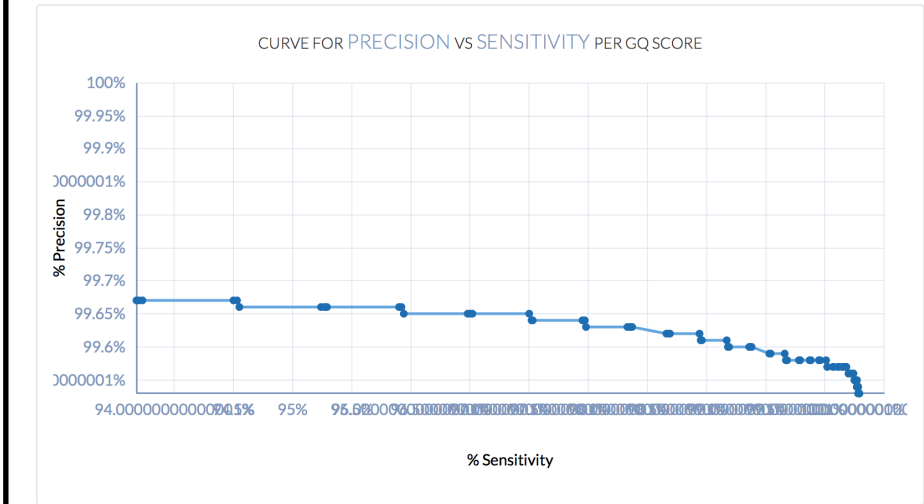
Old, GRCh37 reference genome

PRECISION	99.24%	TRUE-POSITIVES	3,144,270
RECALL	99.79%	FALSE-POSITIVES	24,156
F-MEASURE	99.51%	FALSE-NEGATIVES	6,749

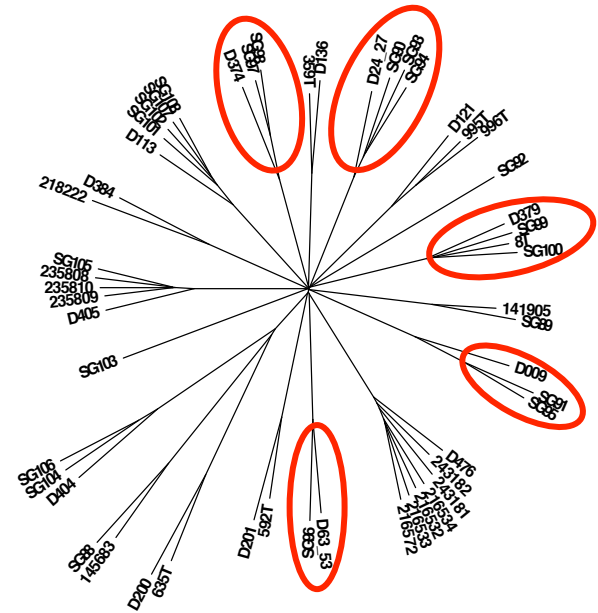
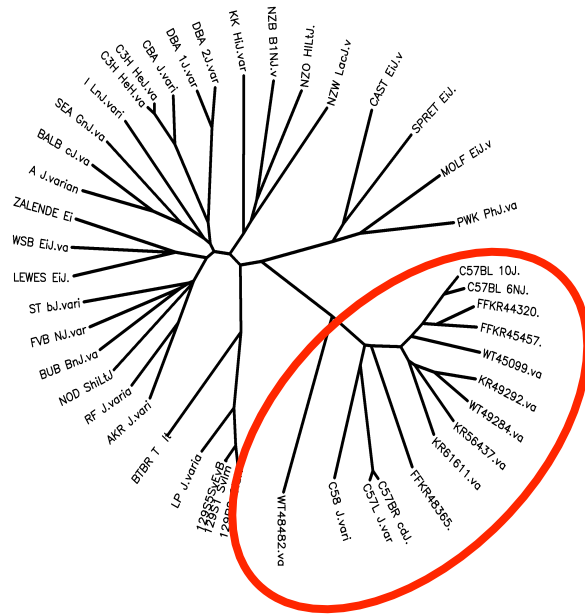


New, hs37d5 reference genome

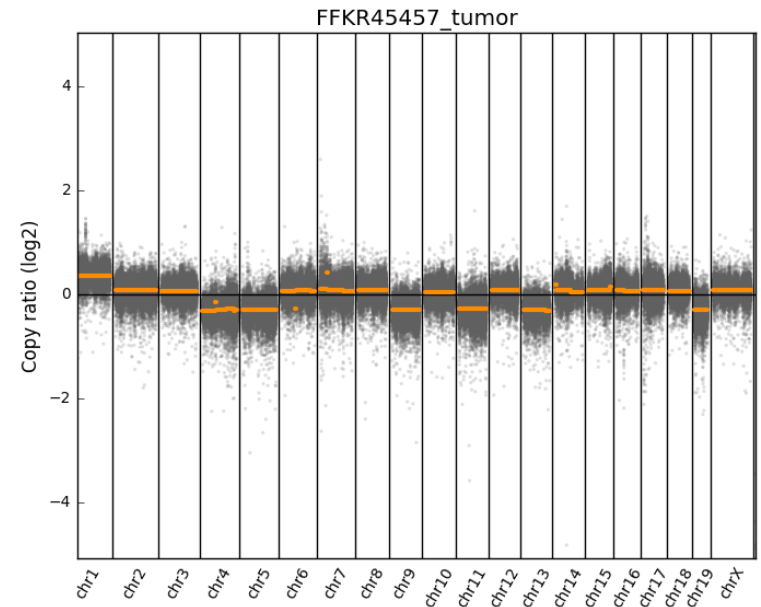
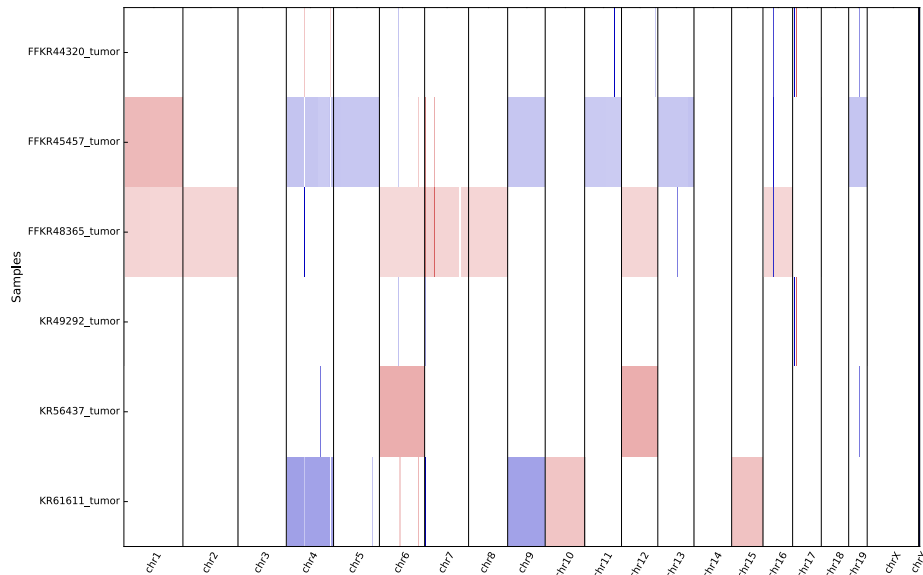
PRECISION	99.53%	TRUE-POSITIVES	3,144,362
RECALL	99.79%	FALSE-POSITIVES	14,933
F-MEASURE	99.66%	FALSE-NEGATIVES	6,657



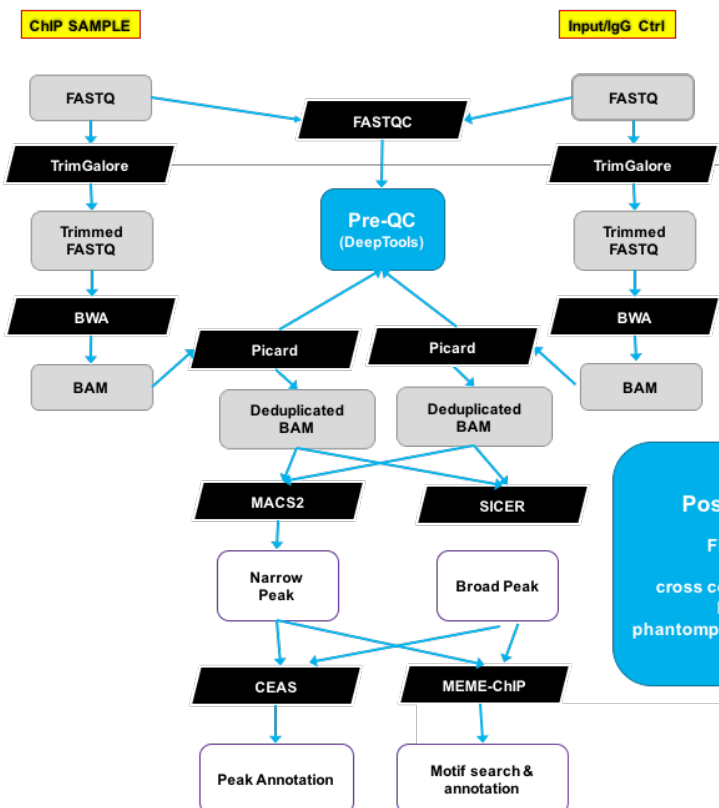
# QC – Sample Identification



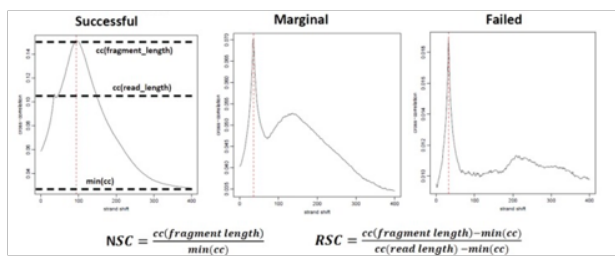
# CNVkit – amplification/deletion detection



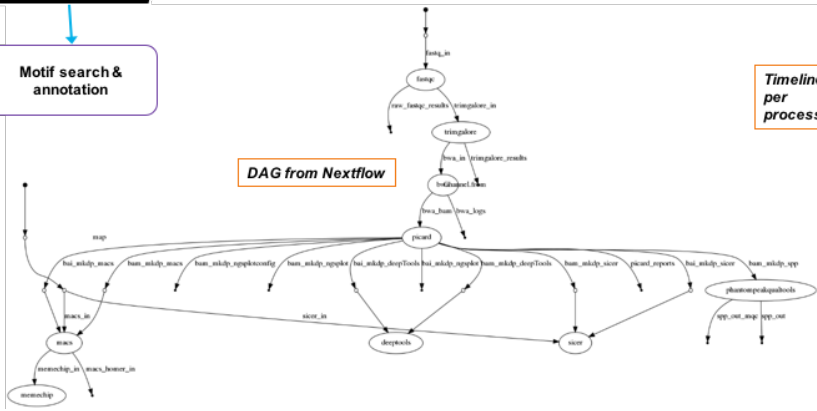
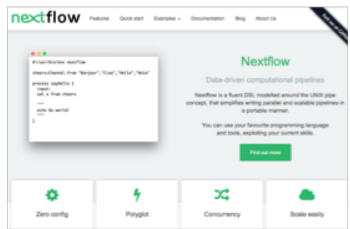
# ChIP-Seq Pipeline (coming soon)



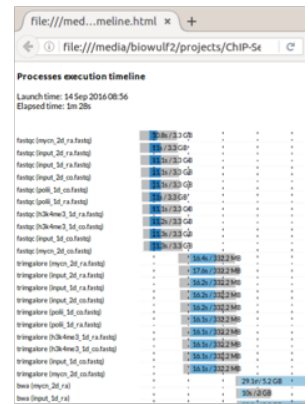
```
[kimb8@biowulf ChIP-Seq-Pipeline]$
[kimb8@biowulf ChIP-Seq-Pipeline]$ nextflow run main.nf --reads='example/*.fastq' --macsconfig='example/macsetsetup.config' -with-timeline timeline.html -with-dag dag.png
NEXTFLOW ~ version 0.21.3
Launching main.nf
[warm up] executor > slurm
[a0/b76b27] Submitted process > fastqc (polii_1d_co.fastq)
[10/5f375] Submitted process > fastqc (input_2d_ra.fastq)
[e5/551f45] Submitted process > fastqc (input_2d_co.fastq)
[9d/1eeba2] Submitted process > fastqc (polii_1d_ra.fastq)
[3a/94cc96] Submitted process > fastqc (input_1d_co.fastq)
[25/ddcf3] Submitted process > fastqc (h3k4me3_1d_ra.fastq)
[c7/201647] Submitted process > fastqc (input_1d_ra.fastq)
[30/04cd5f] Submitted process > fastqc (mycn_2d_ra.fastq)
[a1/576ac9] Submitted process > fastqc (input_2d_co.fastq)
[33/2b9801] Submitted process > fastqc (h3k4me3_1d_co.fastq)
[39/daa8e4] Submitted process > trimgalore (polii_1d_co.fastq)
```



Powered by  
Nextflow



Timeline  
per  
process



# Microarray Analysis App (coming soon)

## CCBR Microarray analysis workflow

(For Affymetrix human and mouse data)

Project ID:

Select CEL files:  no files selected

Choose phenotype file:  no file selected

Choose contrast file:  no file selected

KEGG/GO Enrichment Pvalue threshold:

Which contrast to show:

KEGG/GO Enrichment logFC threshold:

Microarray Results Pre-normalization QC plots ▾ Post-normalization plots ▾ DEG-Enrichments-tables ▾ Help ▾

- Histogram
- Maplots
- Boxplots
- RLE
- NUSE



# 3. Tool and web applications development

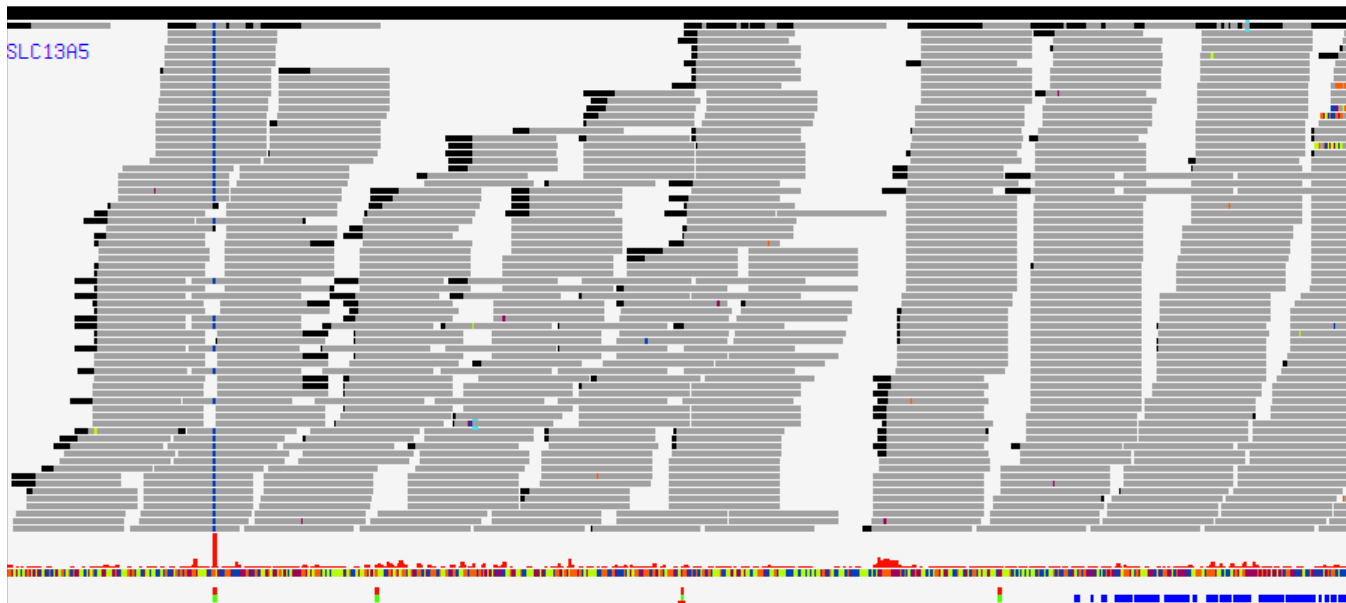
---

# Alview: bam file viewer

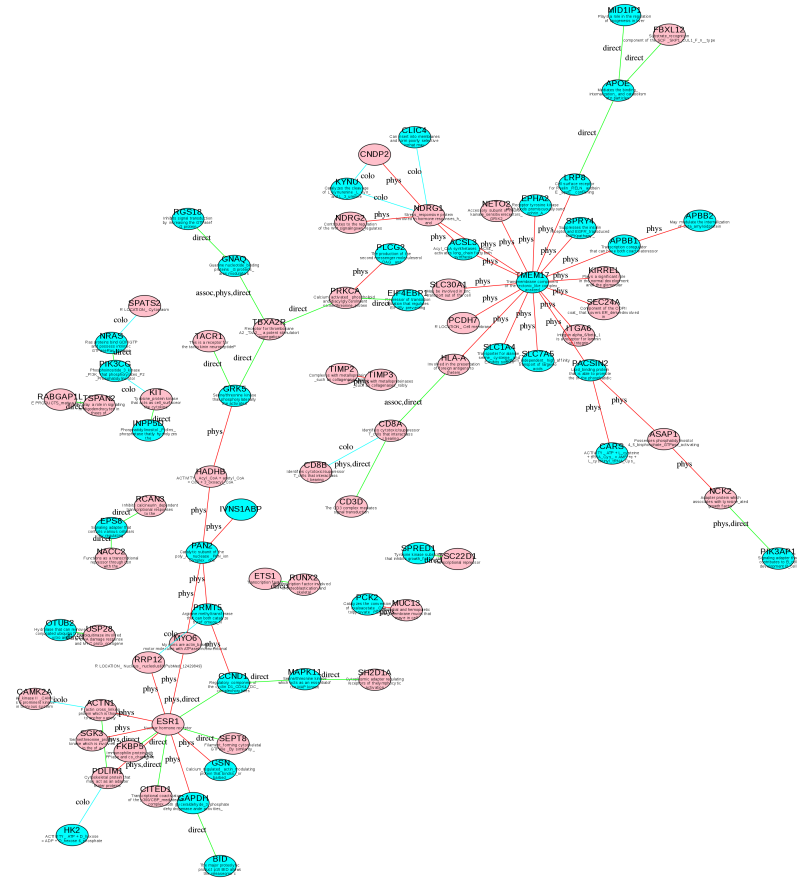
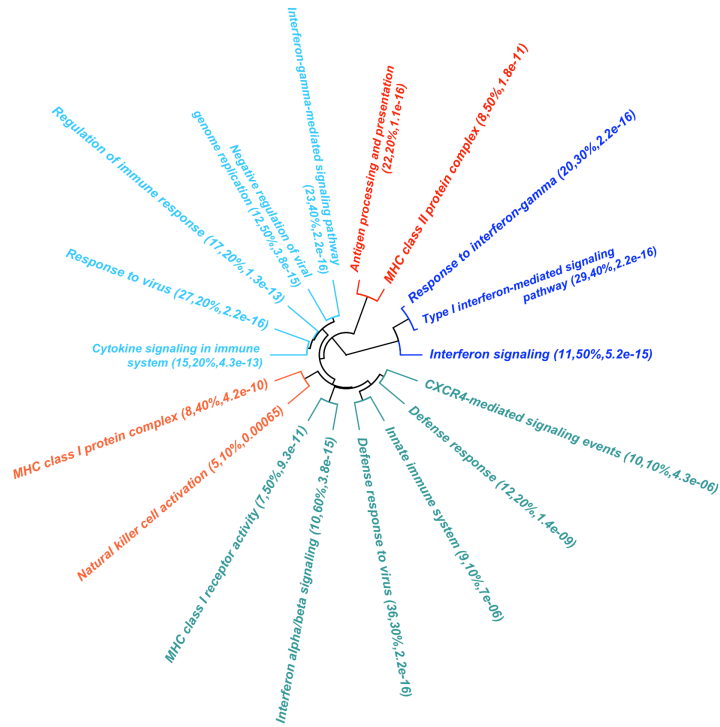
Alview - Use Menu to Select BAM File to View - enter genome position (or gene name)

Position:   Width:  Height:

base	<Page	Page>	lefthalf	righthalf	ZoomIN	ZoomOut	>10000	<100000	>100000	UCSC Link	PNG Save	Help
<10	10>	<100	100>	<1000	1000>	<10000	>10000	<100000	>100000			

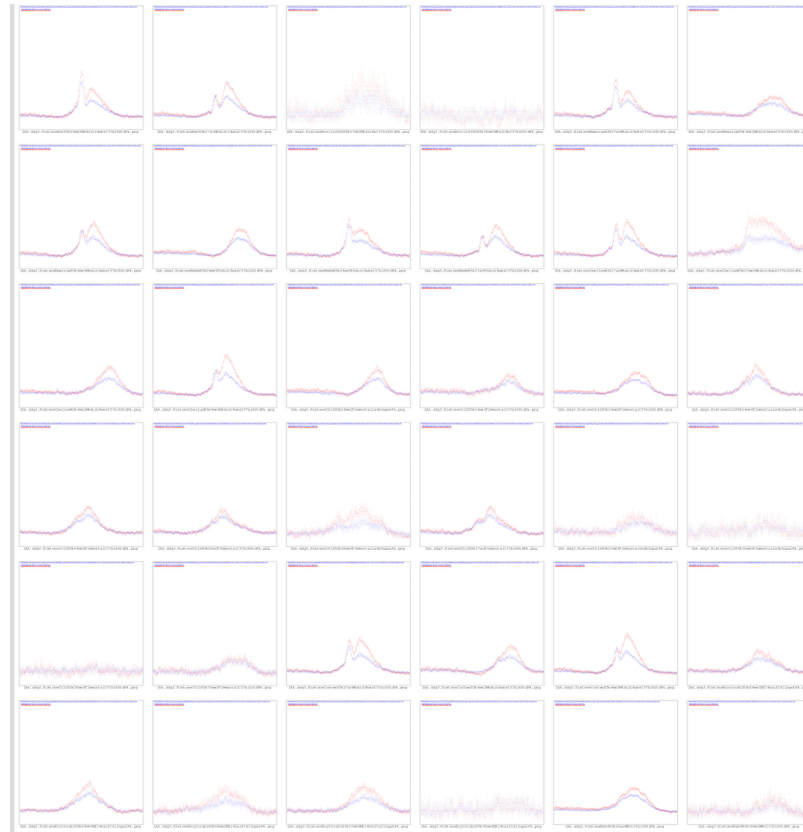


# Pathway Analysis Tool: Pathway Enrichment/Network



# Pathway Analysis Tool: Encode Miner

---



# cBioPortal for Mouse Cancer Models

**cBioPortal** for Cancer Genomics

**thehyve** You are logged in as nci\_user nci. Sign out.

HOME DATA SETS RESULTS TUTORIALS FAQ NEWS TOOLS ABOUT VISUALIZE YOUR DATA

Modify Query

Overview Mutations Expression Download Bookmark

TRP53

**TRP53:**

A135V/K136\_T137dup

# Mutations

0 9

0 200 400 600 800 1000 1200 1348 aa

3D Structure »

51 Missense 7 In-frame 12 Truncating 3 Other

Show / hide columns Showing 73 mutation(s) Search:

Sample ID	Cancer Study	AA change	Annotation	Type	COSMIC	#Mut in Sample
17740_tumor	Glioblastoma (NIH)	MUTATED		splice_region		53
1202810_tumor	Glioblastoma (NIH)	V214E		Missense		71
17243_tumor	Glioblastoma (NIH)	S258_G259insDS		splice_region		83

# 4. Bioinformatics Training

---

# BTEP Classes

---

Microarray

Best Practices/Experimental Design

Exome-Seq

RNA-Seq

Chip-Seq

# Training on CCBR Pipeliner

---

Small groups (max 6-7 people)

Biowulf account

Some command line knowledge

<https://ccbr.github.io/Pipeliner/>

mail: [ccbr@mail.nih.gov](mailto:ccbr@mail.nih.gov)



## CCBR Pipeliner

Welcome to the **CCBR Pipeliner**.

**Pipeliner** provides access to some of the NGS data analysis pipelines used by CCBR on the NIH *Biowulf* Linux Cluster. The program provides a graphical interface for configuring and executing NGS workflows.

**Pipeliner** is designed to work on NIH Biowulf Linux Cluster with its prerequisite software (i.e. Snakemake, slurm, component software and others).

At this time, support can only be provided for those who have access to NIH *Biowulf* Linux Cluster. More details can be obtained by going to CCBR's GitHub Webpage

[Goto Pipeliner GitHub Page »](#)

[Pipeliner Manual »](#)



# Some NGS Best Practices

---

RNA-Seq

Exome-Seq: Justin Lack

Chip-Seq: Bong-Hyun Kim

# Intro and Best Practices: RNA-Seq

---

# RNA-Seq Applications

---

## Differential Gene Expression

- Looks at genes that are at least at the detection limit of microarrays
- Most straightforward, requires less read depth (10-30 M reads)
- Can be more cost-effective than microarrays

## Differential Transcript Expression (Isoform switching)

- Still confined to known transcripts / isoforms
- Complexity is in the assignment of exons to particular isoforms
- Many algorithms can differ in results

## Transcript Discovery / Whole Transcriptome Profiling

- Interest is in looking for new isoforms or unannotated genes
- More complex in terms of bioinformatics analysis
- Can find false positives, depending on leniency of algorithm

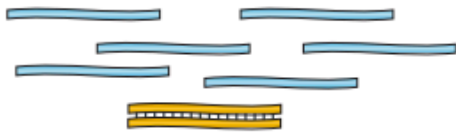
## Others

- SNP/Somatic Variant/Gene Fusion Detection

# Method – Preparation

## a Data generation

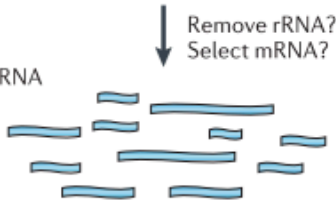
① mRNA or total RNA



② Remove contaminant DNA

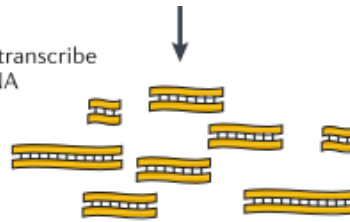


③ Fragment RNA

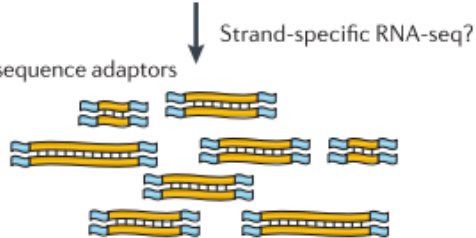


Remove rRNA?  
Select mRNA?

④ Reverse transcribe  
into cDNA

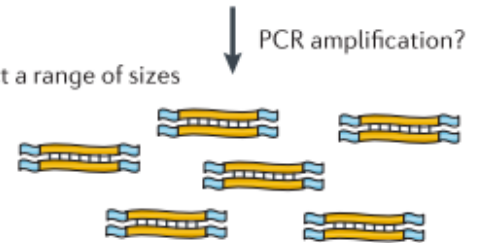


⑤ Ligate sequence adaptors



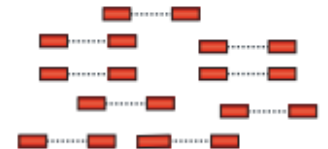
Strand-specific RNA-seq?

⑥ Select a range of sizes

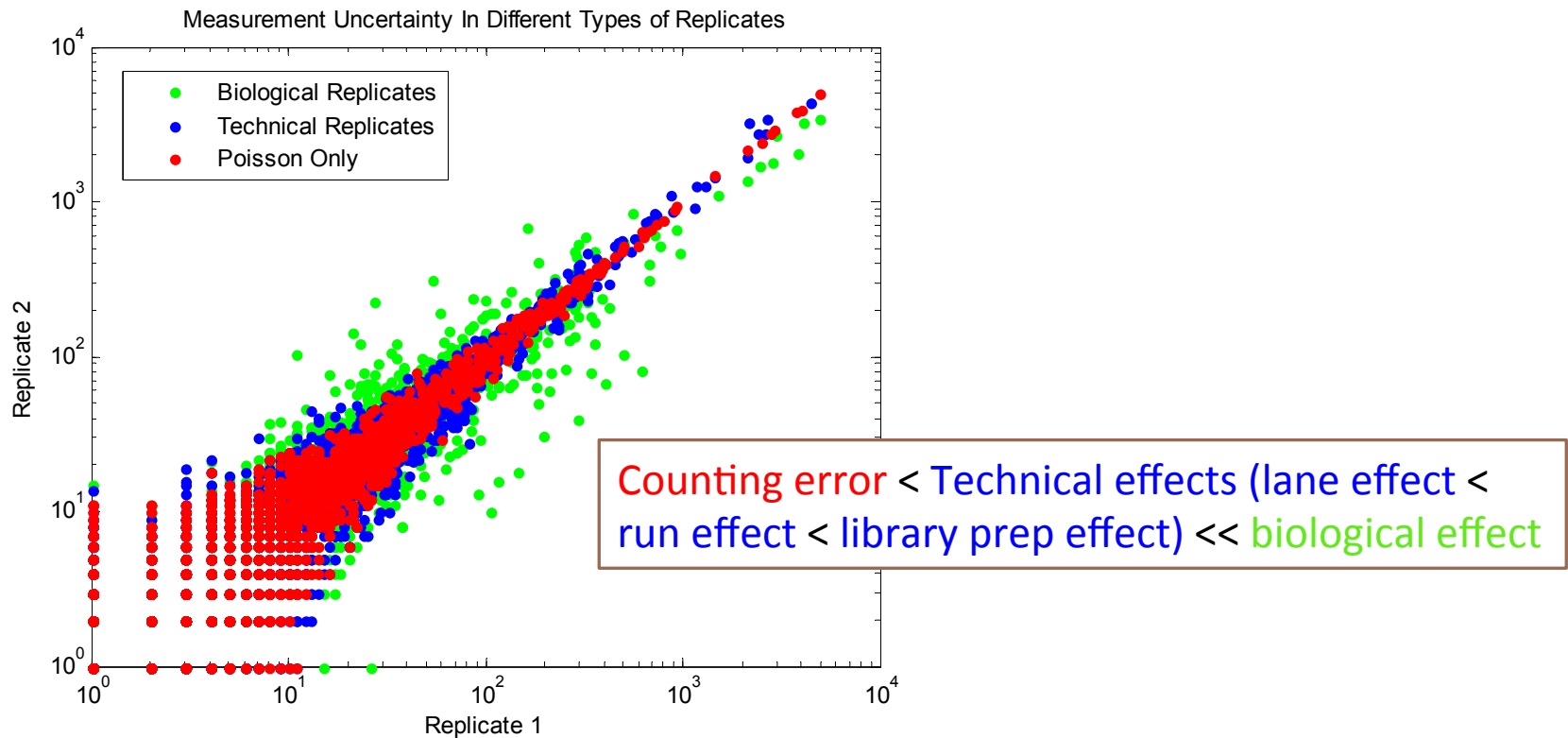


PCR amplification?

⑦ Sequence cDNA ends

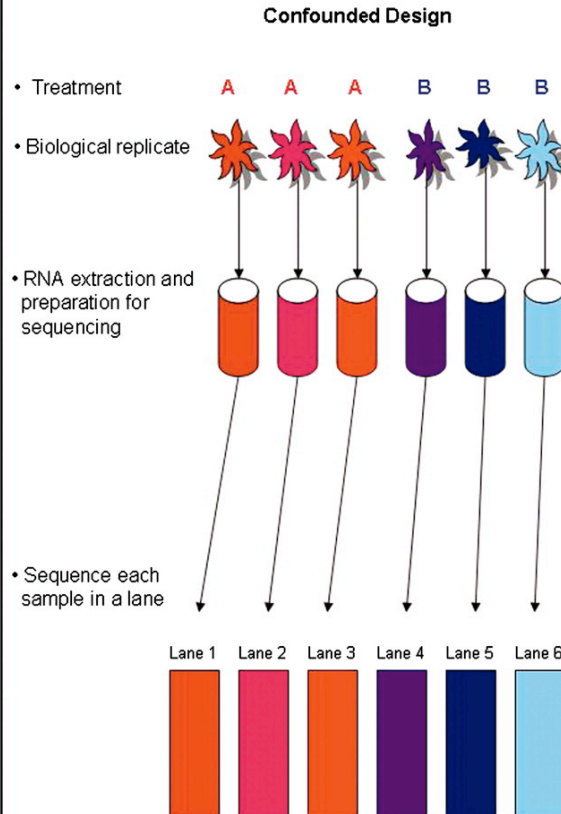
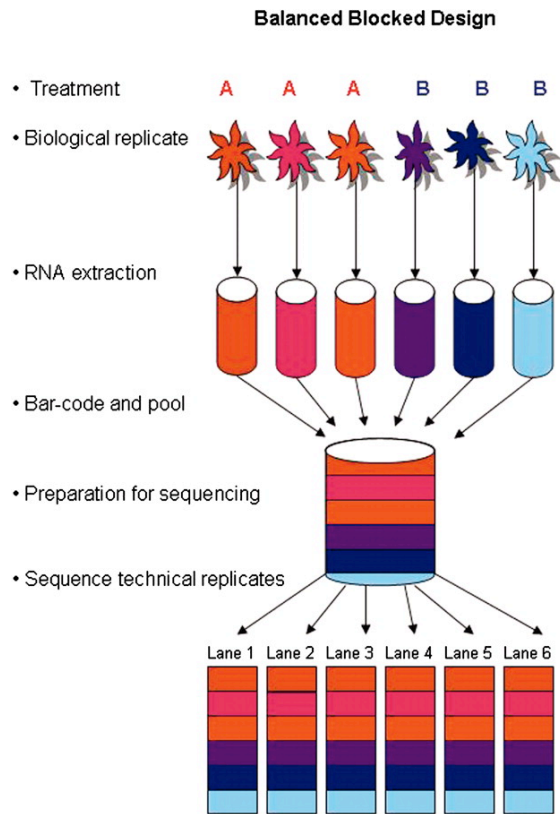


# Types of variance



*Busby et al, Bioinformatics 2013*  
*Marioni et al, Genome Res 2008*

# Experimental Design: avoiding lane effects



*- does not permit partitioning of batch and lane effects from the estimate of within-group biological variability*

*Auer and Doerge, Genetics 2010*

# Best Practices

---

1. Factor in at least 3 replicates (absolute minimum), but 4 if possible (optimum minimum). Biological replicates are recommended rather than technical replicates.

2. Always process your RNA extractions at the same time. Extractions done at different times lead to unwanted batch effects.

3. There are 2 major considerations for RNA-Seq libraries:

If you are interested in coding mRNA, you can select to use the mRNA library prep. The recommended sequencing depth is between **10-20M paired-end (PE)** reads. Your RNA has to be high quality (RIN > 8).

If you are interested in long noncoding RNA as well, you can select the total RNA method, with sequencing depth **~25-60M PE** reads. This is also an option if your RNA is degraded.

4. Ideally to avoid lane batch effects, all samples would need to be multiplexed together and run on the same lane. This may require an initial MiSeq run for library balancing. Additional lanes can be run if more sequencing depth is needed.

5. If you are unable to process all your RNA samples together and need to process them in batches, make sure that replicates for each condition are in each batch so that the batch effects can be measured and removed bioinformatically.

<https://bioinformatics.cancer.gov/content/rna-seq>

# A good review:

---

Conesa *et al. Genome Biology* (2016) 17:13  
DOI 10.1186/s13059-016-0881-8

Genome Biology

REVIEW

Open Access

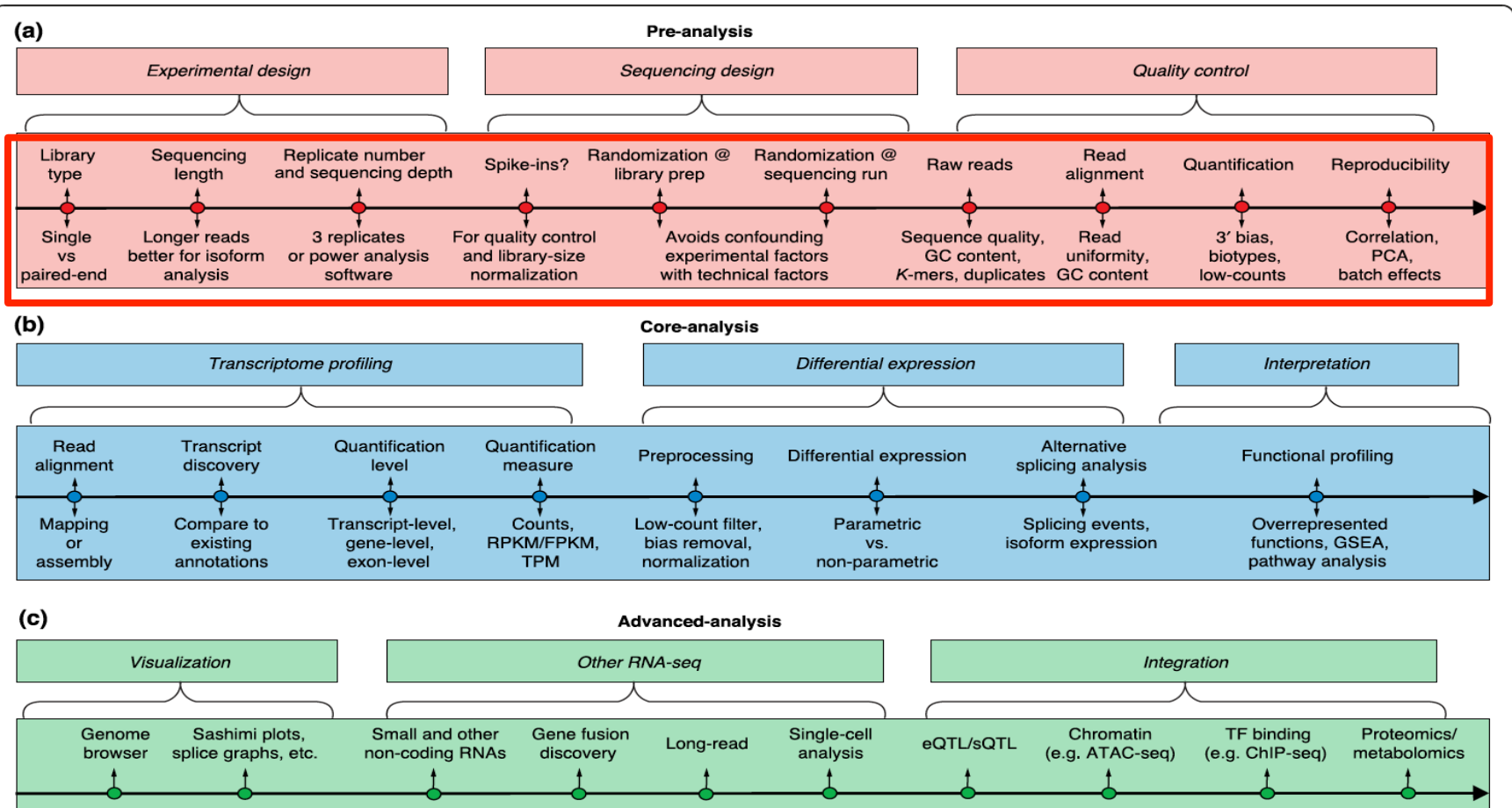
## A survey of best practices for RNA-seq data analysis



Ana Conesa<sup>1,2\*</sup>, Pedro Madrigal<sup>3,4\*</sup>, Sonia Tarazona<sup>2,5</sup>, David Gomez-Cabrero<sup>6,7,8,9</sup>, Alejandra Cervera<sup>10</sup>, Andrew McPherson<sup>11</sup>, Michał Wojciech Szczęśniak<sup>12</sup>, Daniel J. Gaffney<sup>3</sup>, Laura L. Elo<sup>13</sup>, Xuegong Zhang<sup>14,15</sup> and Ali Mortazavi<sup>16,17\*</sup>



# Generic roadmap for expt design & analysis



**Fig. 1** A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

# Samples vs Read depth

If on a tight budget, deciding between number of replicates vs sequencing depth, always higher replicates with lower sequencing depth leads to higher statistical power

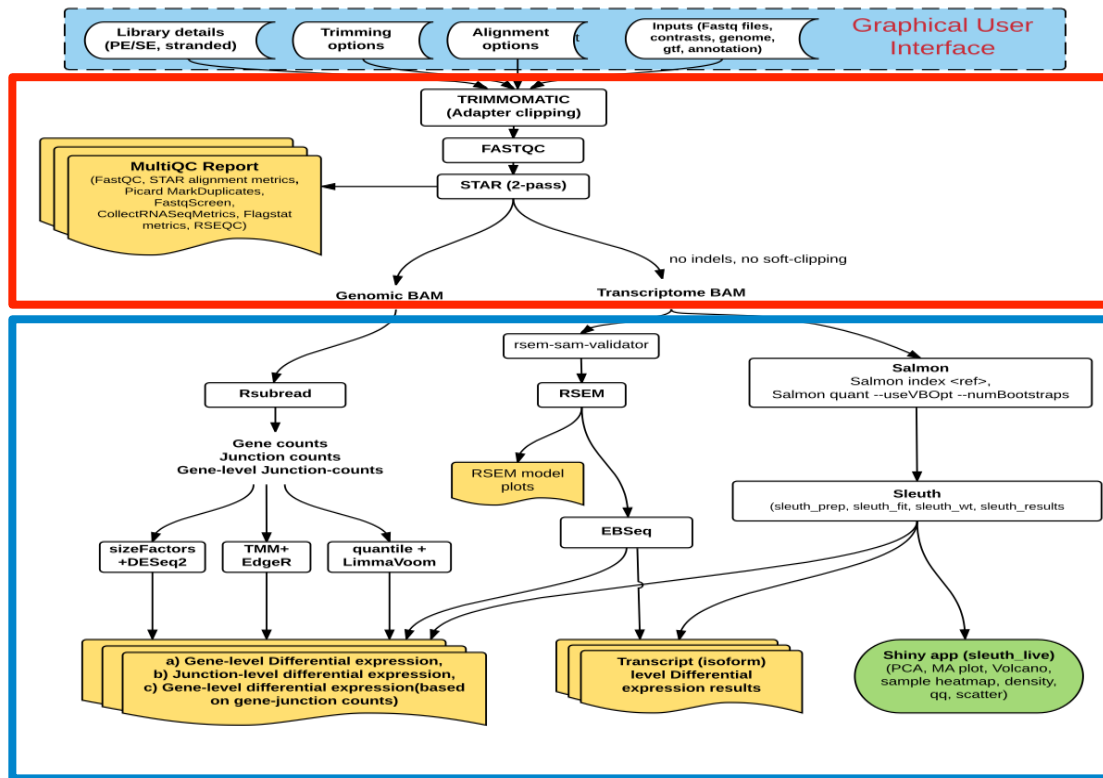
- 3M reads x 10 replicates = 30M reads yields 52% power
- 10mil reads x 3 replicates = 30M reads yields 33% power

**Table 1** Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default value), the statistical power increases with the difference between the two groups (effect size), the sequencing depth, and the number of replicates per group. This table shows the statistical power for a gene with 70 aligned reads, which was the median coverage for a protein-coding gene for one whole-blood RNA-seq sample with 30 million aligned reads from the GTEx Project [214]

# RNA-Seq Pipeline Workflow



STEP 1: INITIAL QC

STEP 2: COUNTING & DEG

# CCBR Pipeliner (QC Report, DEG Analysis)

**Project Information**

Project Id:  (Examples: CCBR-*nnn*,*Labname* or short project name)

Email address:  (Mandatory field: must use @nih.gov email address)

Flow Cell ID:  (Examples: FlowCellID, Labname, date or short project name)

**Global Settings**

Genome:  Pipeline Family:

**Project Description**  **RNAseq**

Data Directory:

FastQ files Found: 0

Working Directory:

**Options**

Pipeline:

Read Length is

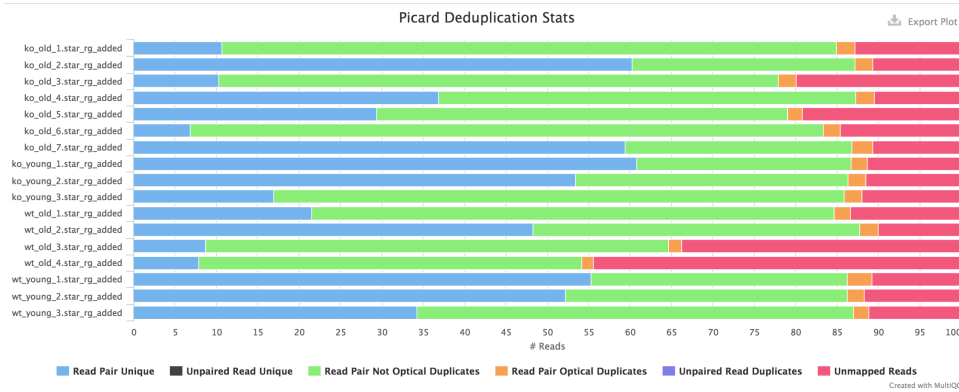
Reads are Unstranded

**Low Abundance Gene Thresholds**

Filter out genes <  read counts in <  samples

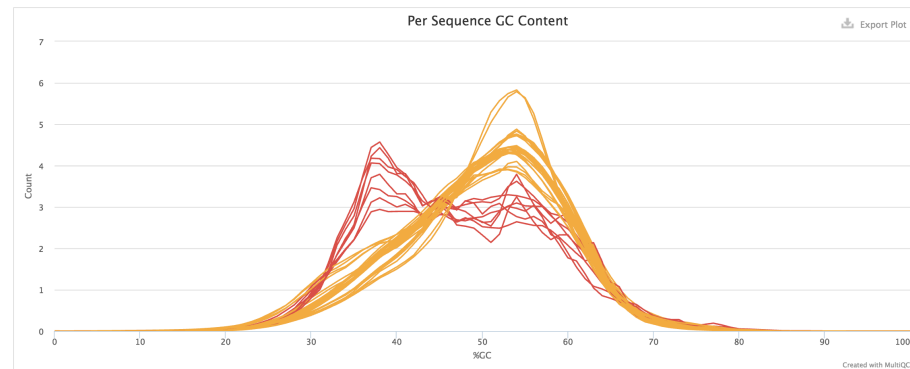
**Sample Information**

# QC: Low RNA input (0.1 – 1 ug total RNA or 10 - 100 ng isolated mRNA)



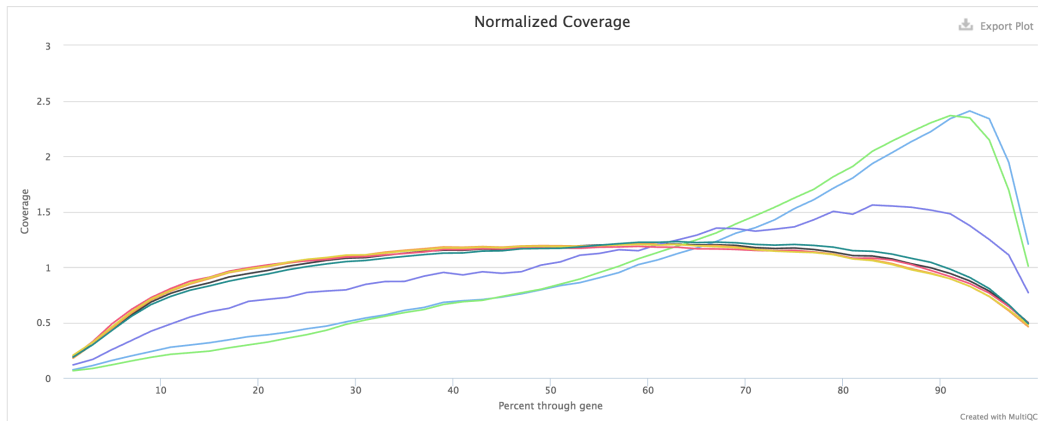
High duplication rates

GC Bias



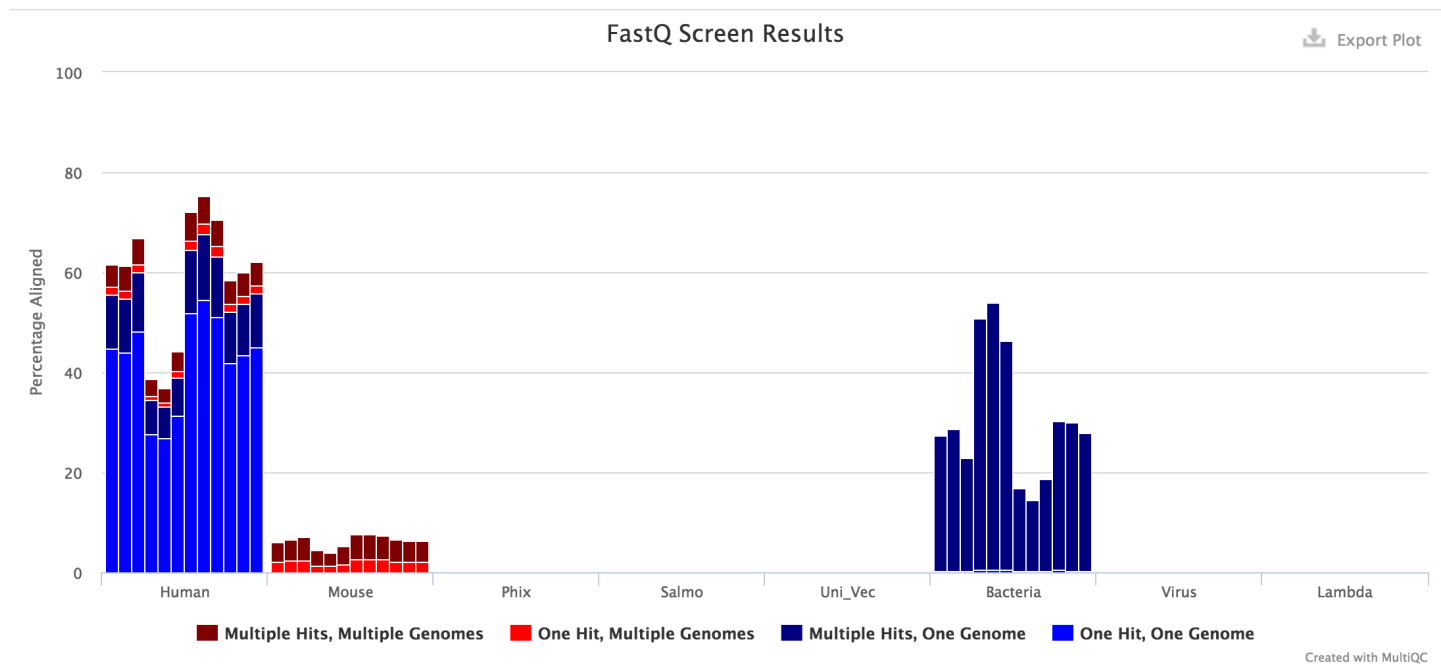
Clontech: 250pg – 10 ng RNA

# QC: Poor RNA Quality (RIN > 7, for FFPE or degraded, use ribominus)

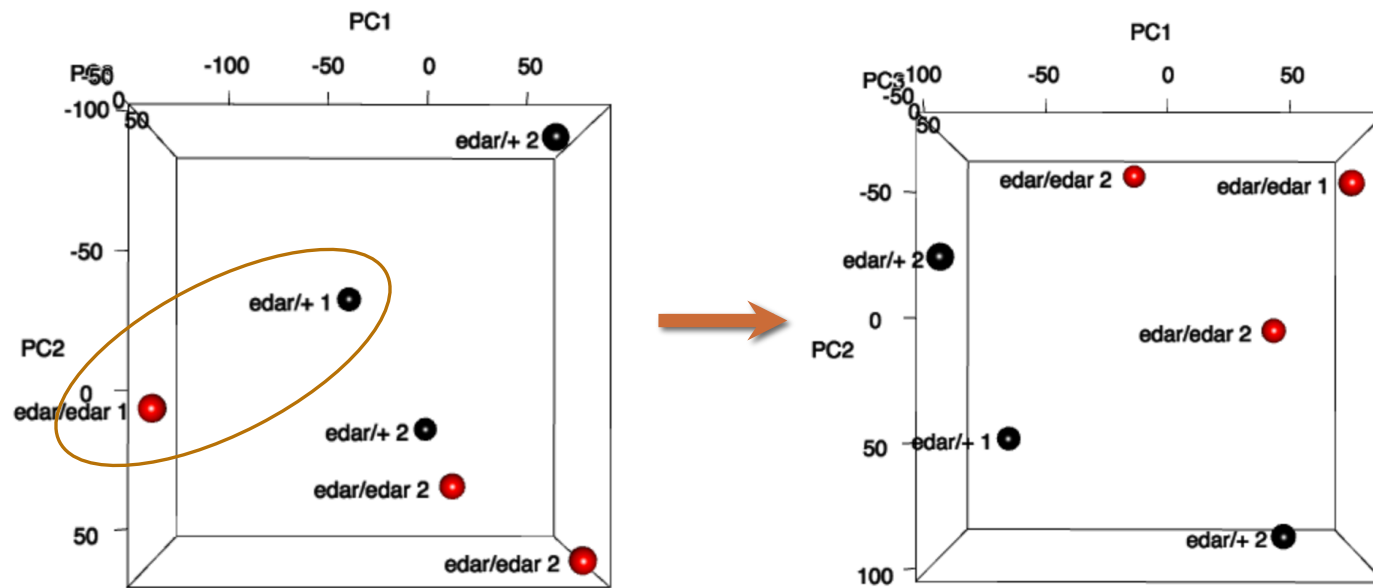


Degraded RNA showing  
3' bias in coverage

# QC: Contamination



# QC: Batch Effects



Litter effect: used batch removal



## Take Home Message:

---

While you are planning your RNA-Seq experiment (not after), please come talk to us.

[CCBR@mail.nih.gov](mailto:CCBR@mail.nih.gov)