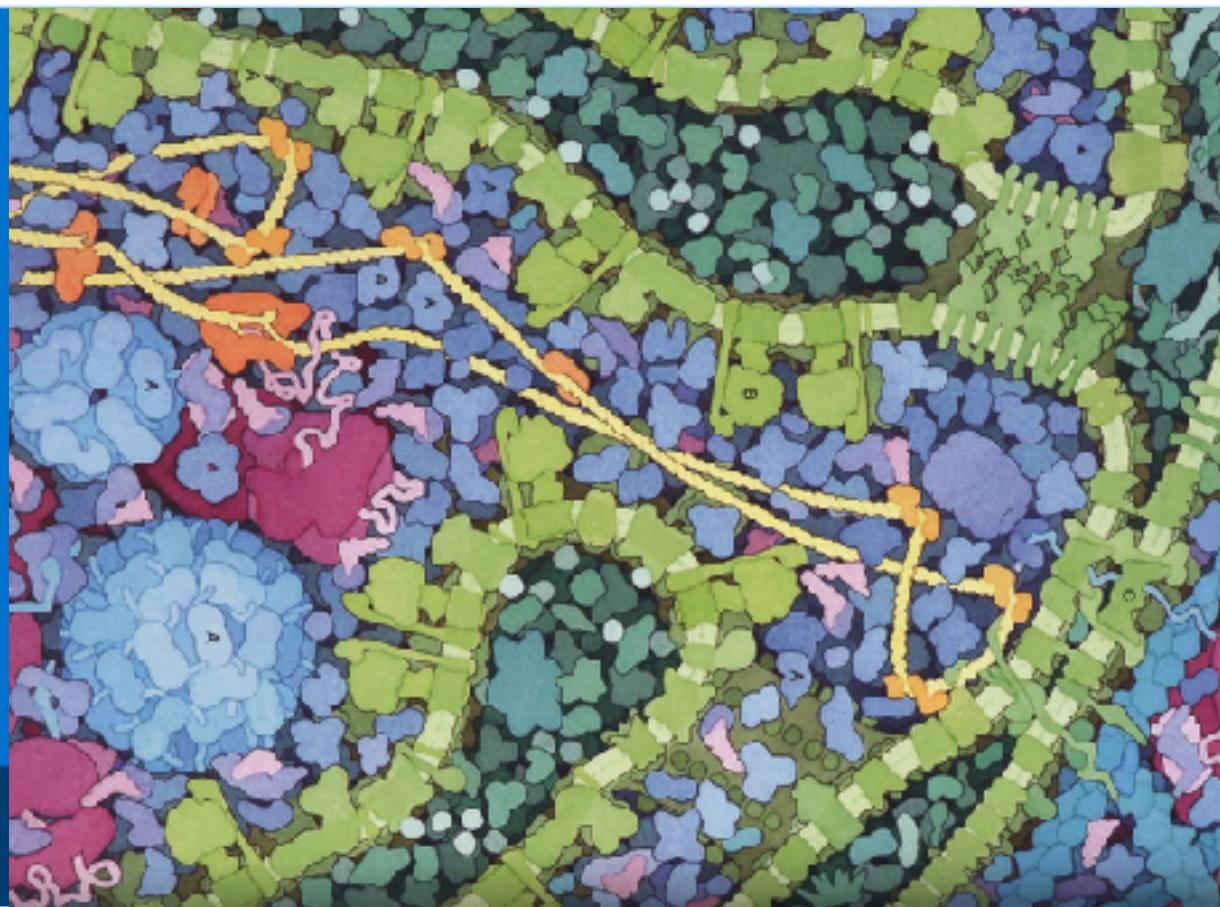


Wellcome Trust Advanced Course - RNA Transcriptomics 2018

Dr. Anton Enright
Group Leader,
Head of Genomics,
Dept of Pathology,
University of Cambridge
aje39@cam.ac.uk

Department of Pathology

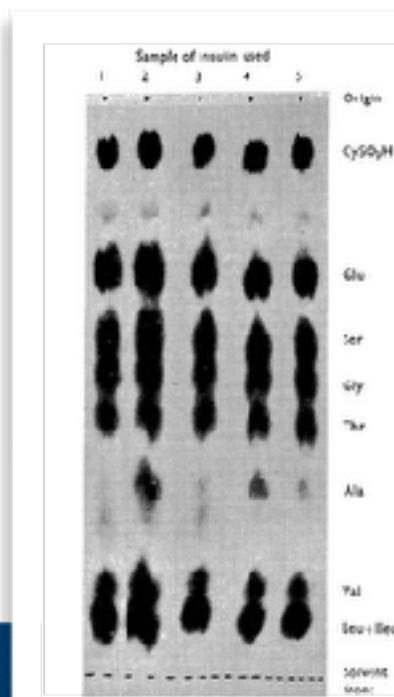
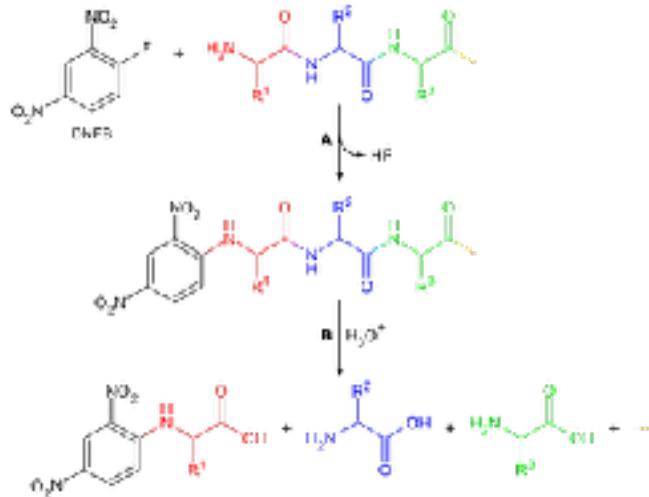


Overview - Computational Analysis

- **Course Modules and Analysis**
 - Background - Progress in Sequencing
 - Main Biological Focus
 - Different RNA protocols used
 - Analysis Framework
 - Goals

Sequencing in molecular biology

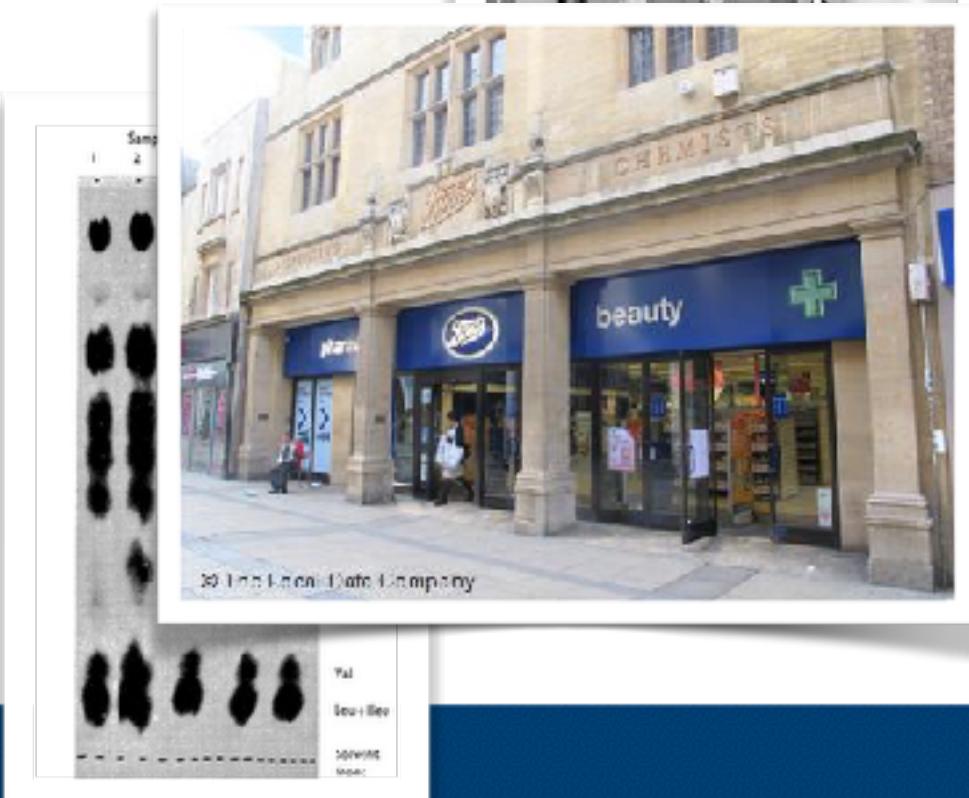
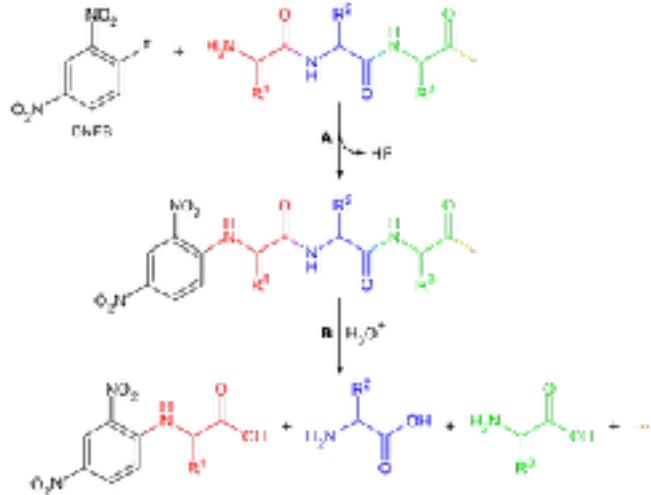
- Birth of sequencing - Fred Sanger 1951
- Proteins thought to be quite amorphous
- Looking for composition of amino acids in Bovine Insulin
- N-terminal Labelling then hydrolysis
- Chromatographic separation



Fred Sanger
Biochemistry Dept circa 1952

Sequencing in molecular biology

- Birth of sequencing - Fred Sanger 1951
 - Proteins thought to be quite amorphous
 - Looking for composition of amino acids in Bovine Insulin
 - N-terminal Labelling then hydrolysis
 - Chromatographic separation

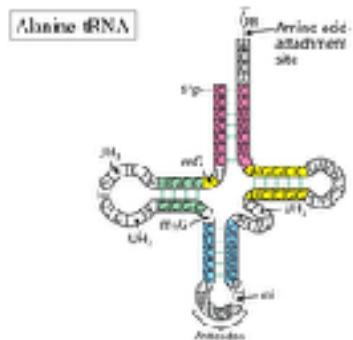


Sequencing Nucleotides - RNA was first

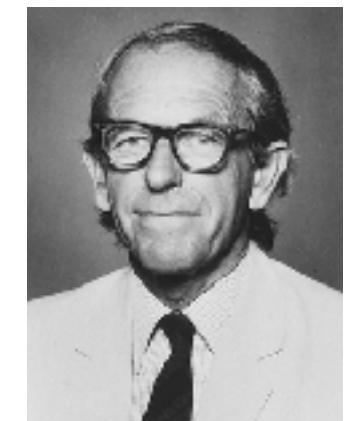
- Double Ribonuclease digestion of RNA molecules - 1965, 1967



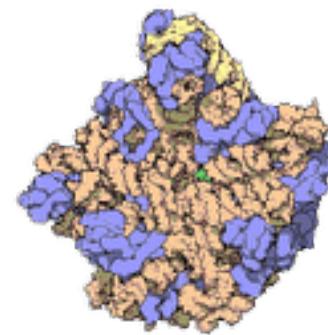
Robert Holley
Cornell



Alanine tRNA



Fred Sanger
LMB



5s Ribosomal RNA

DNA Sequencing was first pioneered in Cambridge

By Fred Sanger, original methods were slow and laborious

DNA Sequencing was first pioneered in Cambridge

By Fred Sanger, original methods were slow and laborious

C*

CG*

CGT*

CGTA*

CGTAC*

CGTACG*

CGTACGT*

CGTACGTA*

CGTACGTAG*

DNA Sequencing was first pioneered in Cambridge

By Fred Sanger, original methods were slow and laborious

C*

CG*

CGT*

CGTA*

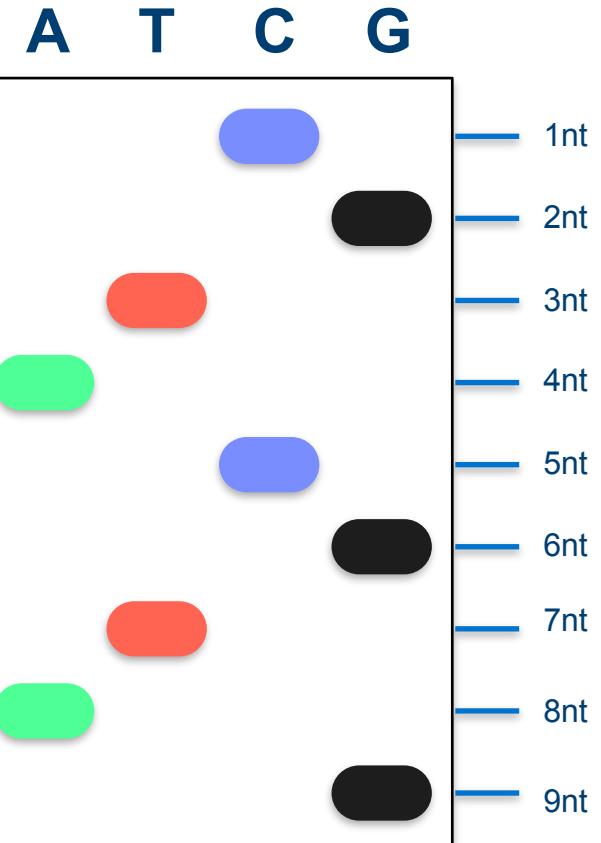
CGTAC*

CGTACG*

CGTACGT*

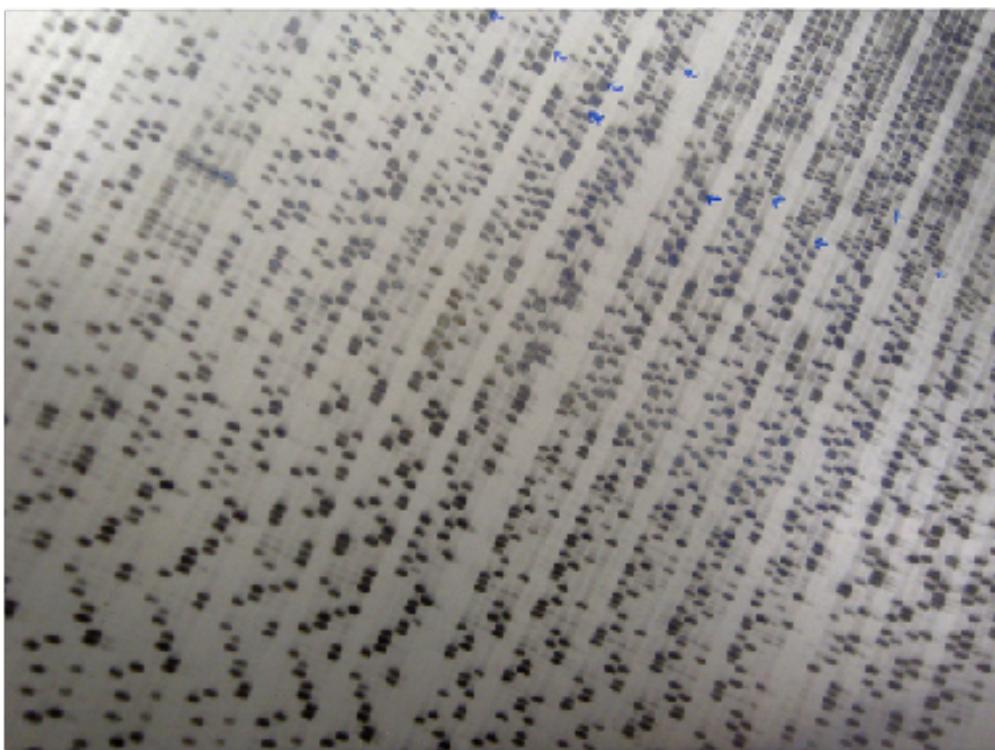
CGTACGTA*

CGTACGTAG*

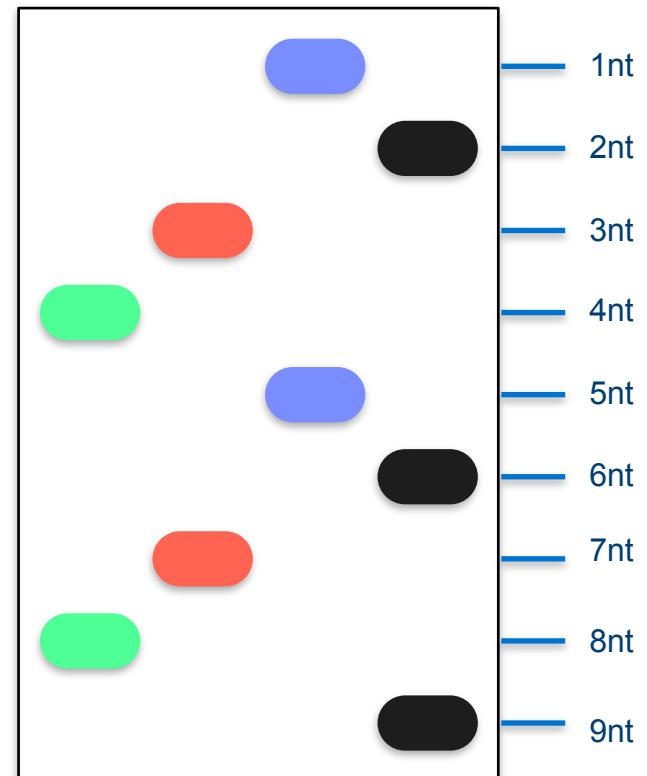


DNA Sequencing was first pioneered in Cambridge

By Fred Sanger, original methods were slow and laborious

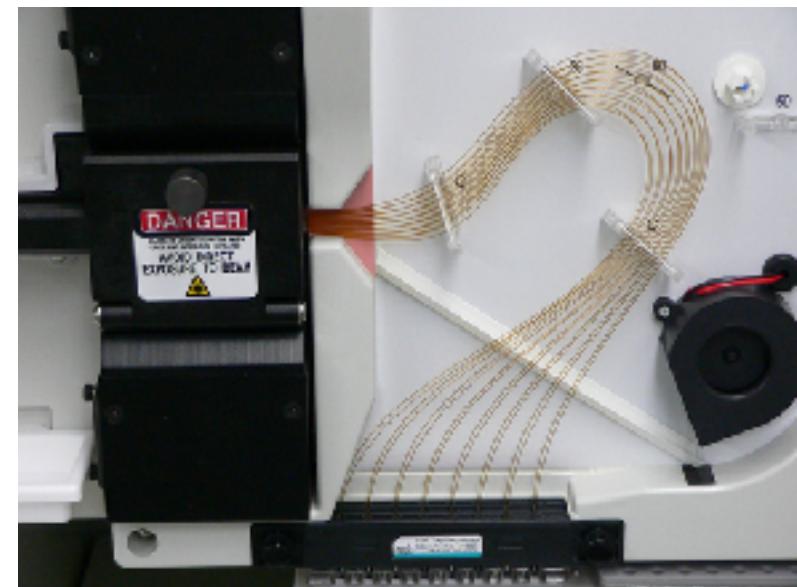
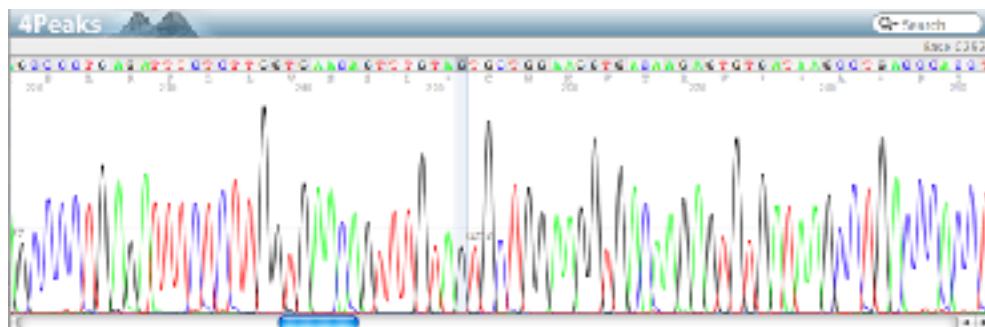


A T C G



The Human Genome Project

- Cambridge was part of a massive project to sequence the first human genome
- There are 3.2 Billion nucleotides (A,T,G and C) in the sequence
- The first genome took 10 years and cost \$ 3 Bn



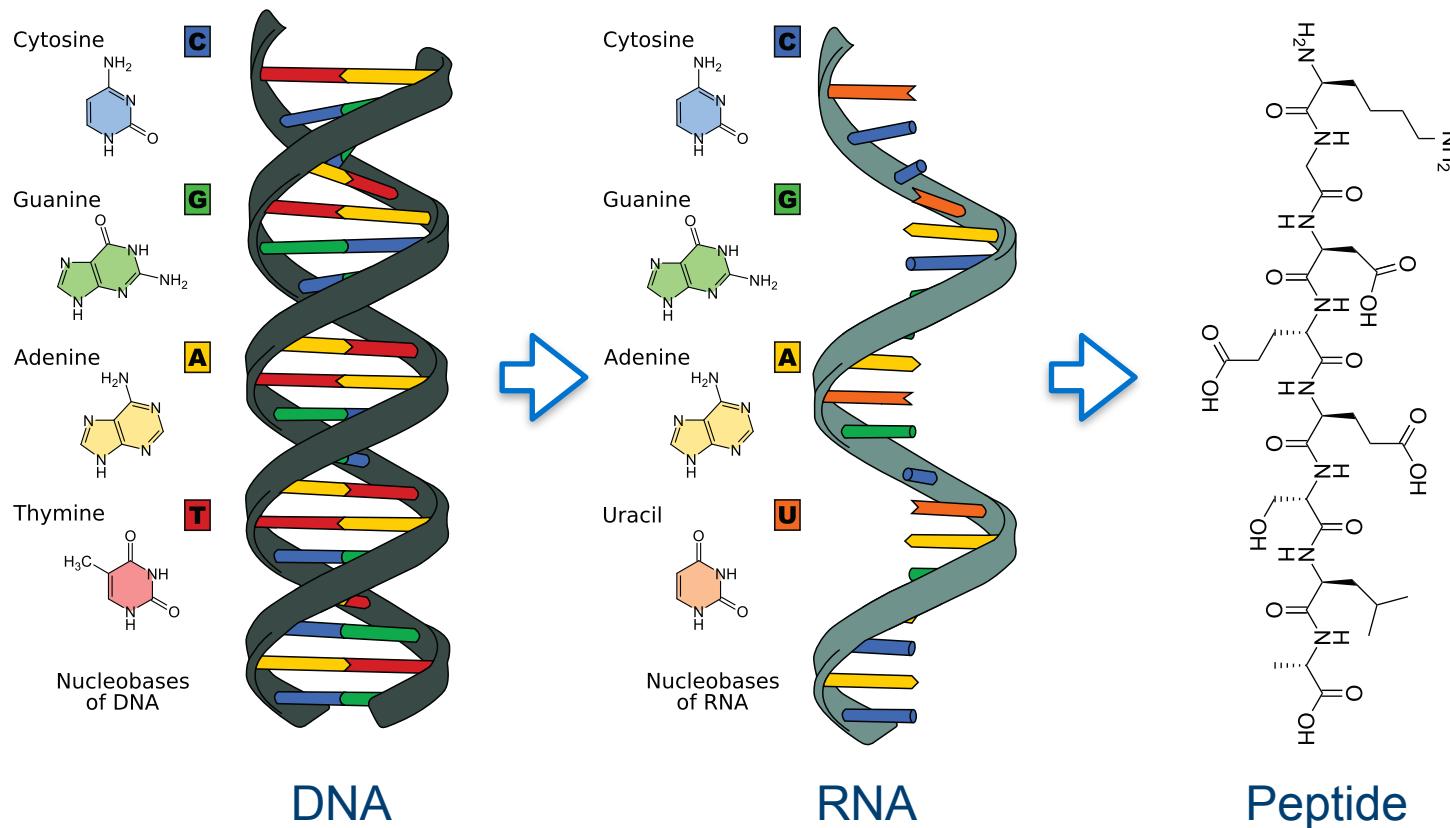
A Paradigm Shift

- The original technology to produce the first genome (early 2000s) obsolete
- Since 2010, significantly faster and cheaper approaches available
- The first human genome cost \$3Bn and took 10 years
- **A full human genome now costs about £2,500 and takes 3-5 days**

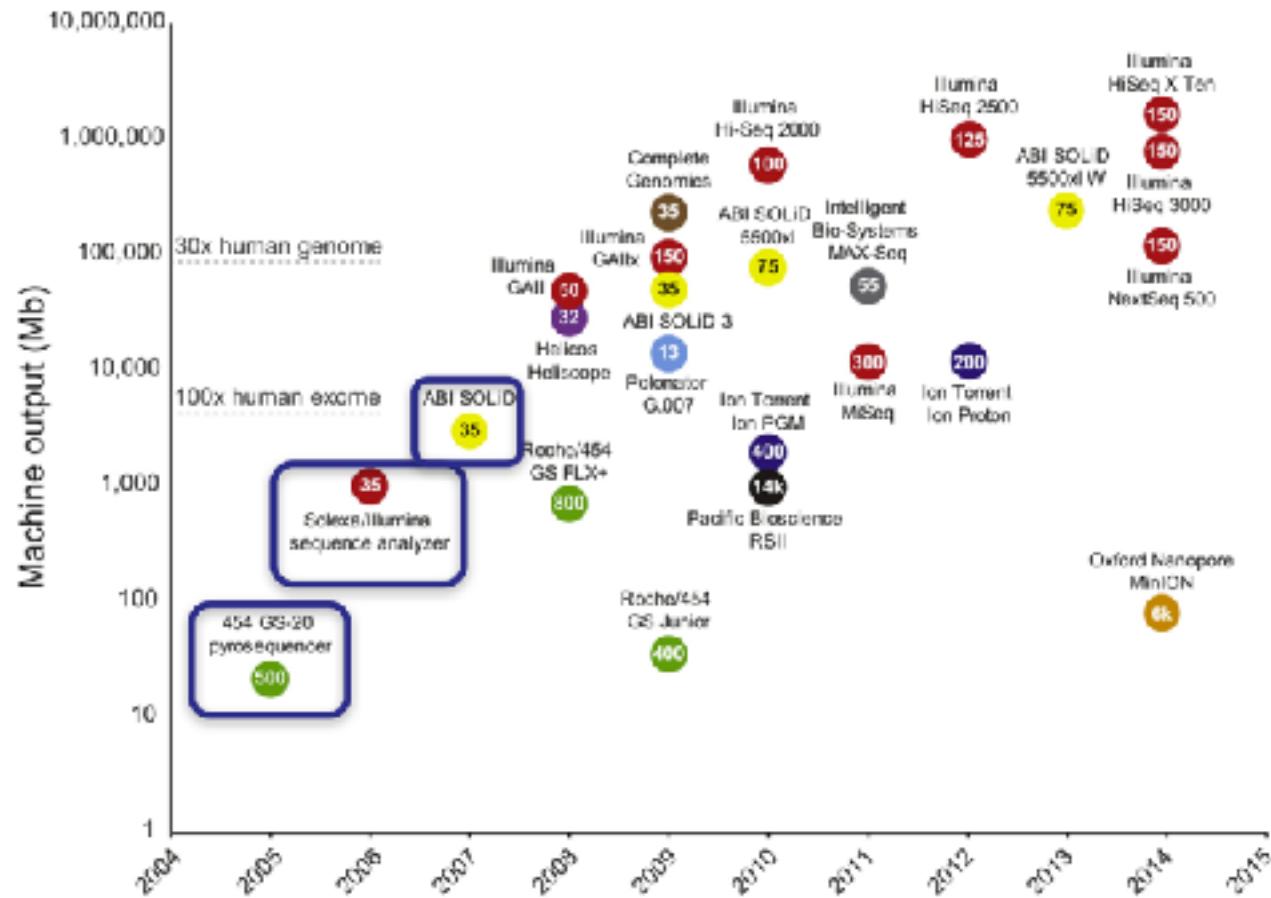


RNA Transcriptomics

RNA is the most dynamic and pervasive molecule at the interface between information storage and function



The Next Generation - High Output, Low Cost



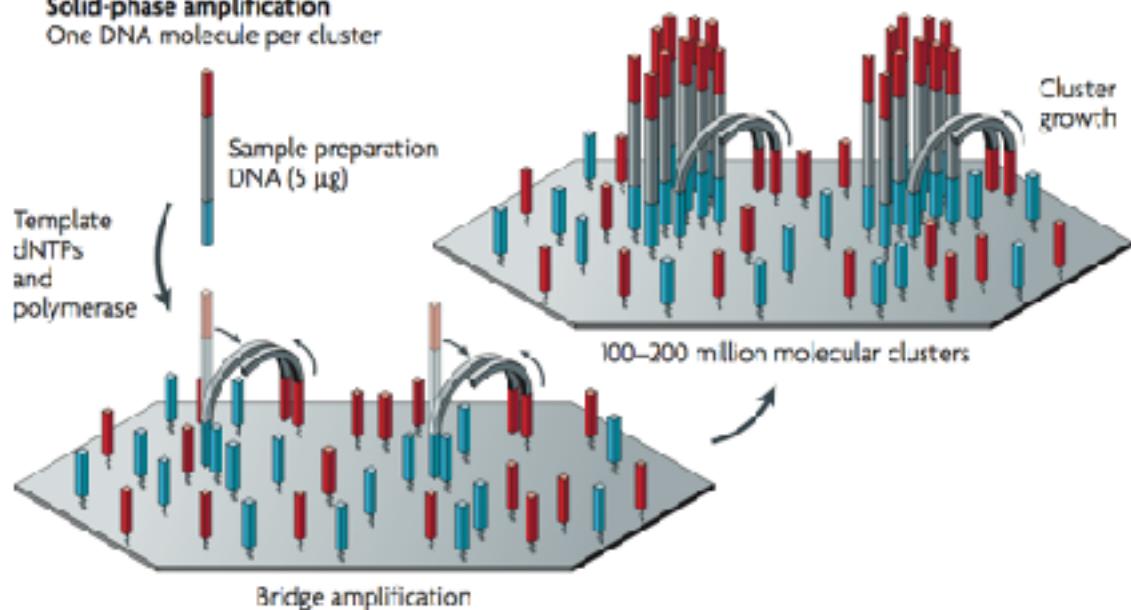
Reuter JA, Spacek DV & Snyder MP, Mol Cell. 2015 May 21;58(4):586-97.

Solid Phase Immobilisation Sequencing (e.g. Illumina)

- Solid Phase methods immobilisation

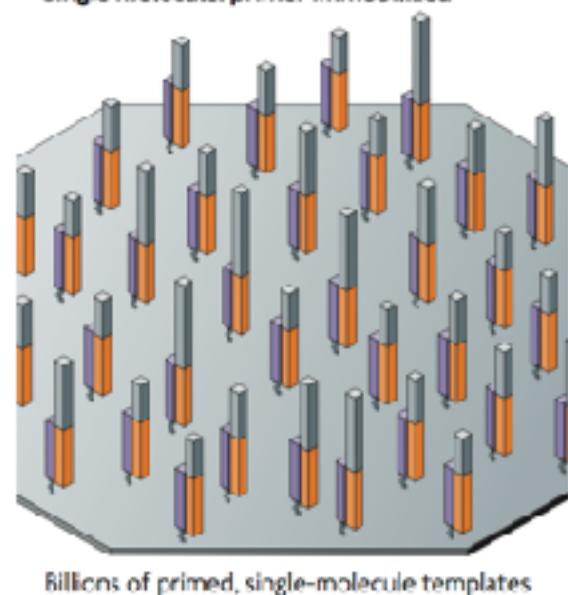
Images from: Metzker ML, Nat Rev Genet. 2010 Jan; 11(1):31-46.

b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster



Illumina Sequencing (PCR)

c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized

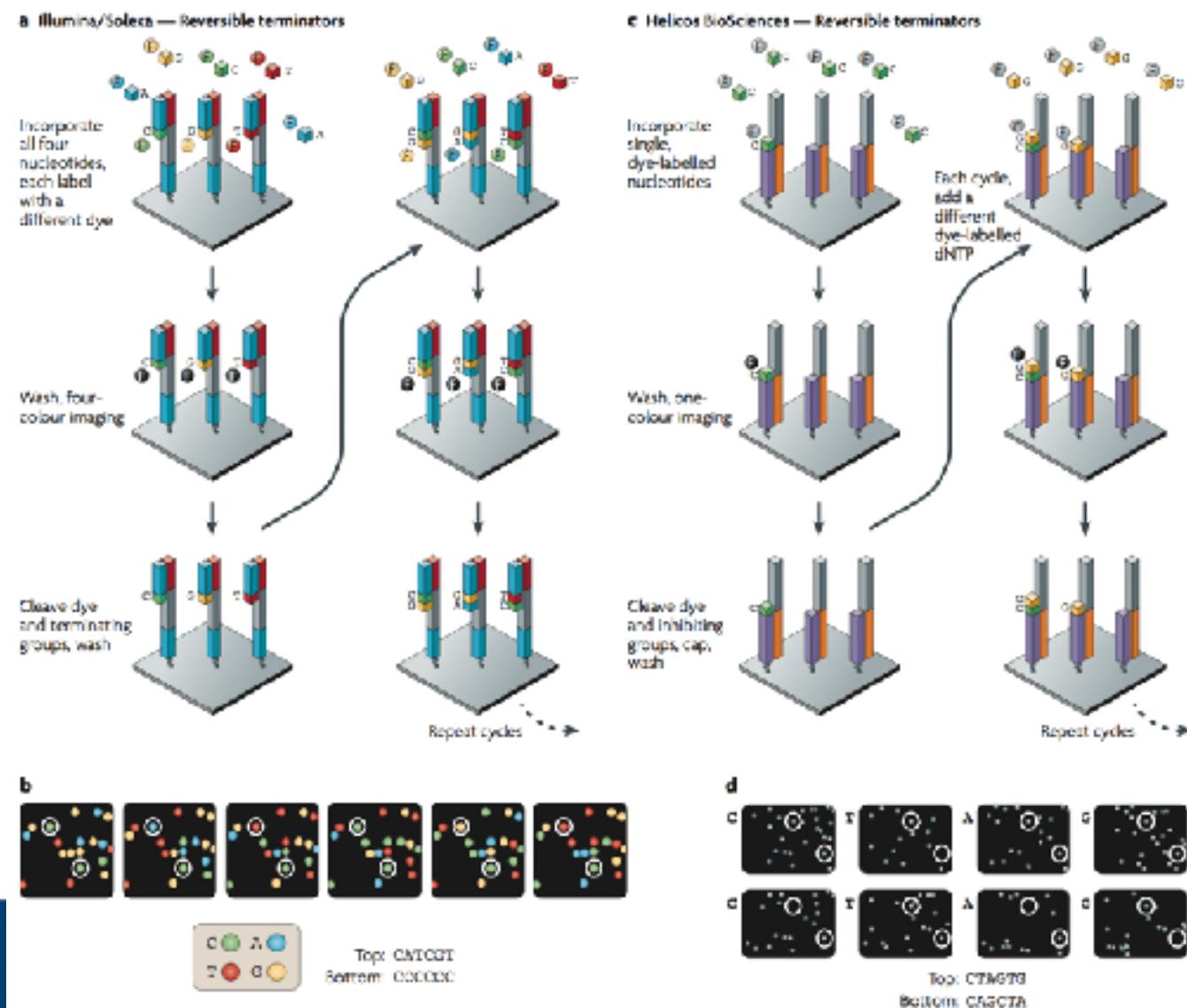


Helicos (Single Molecule)

Illumina - Reversible terminator sequencing

Images from: Metzker ML, Nat Rev Genet. 2010 Jan;11(1):31-46.

- Nucleotides are fluorescently labelled
- They contain terminators attached to the fluorophore
- After all four bases are incorporated by a polymerase the clusters are laser imaged
- These terminators are removed after a cycle and the next base is added

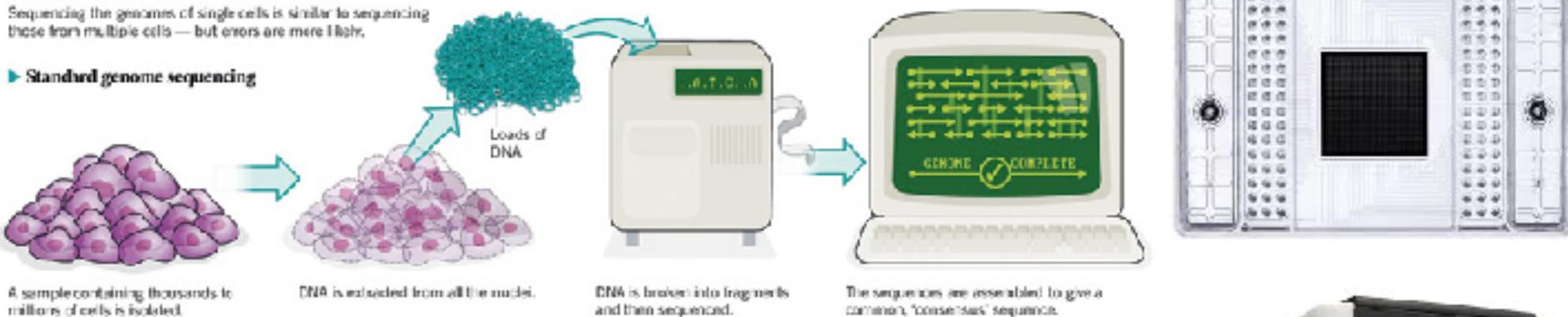


Single Cell Genomics

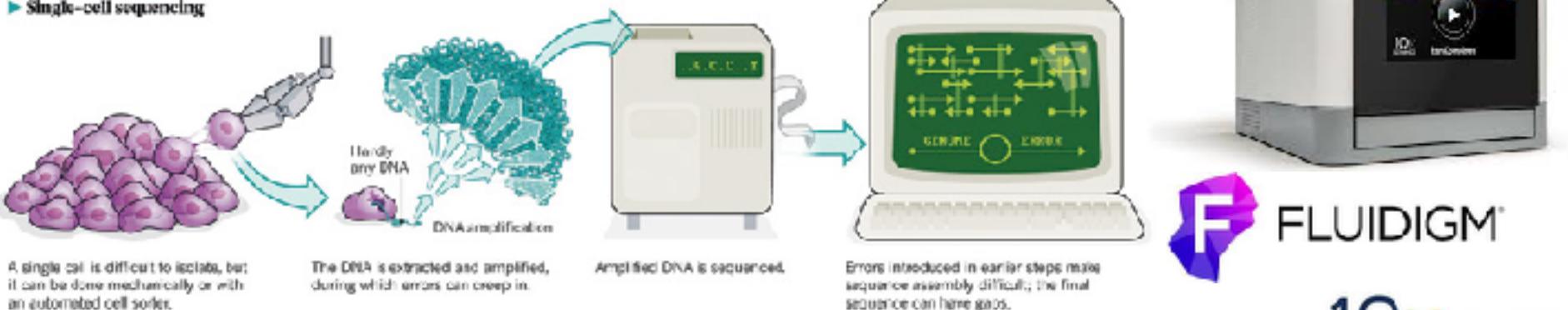
ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing these from multiple cells — but errors are more likely.

► Standard genome sequencing



► Single-cell sequencing

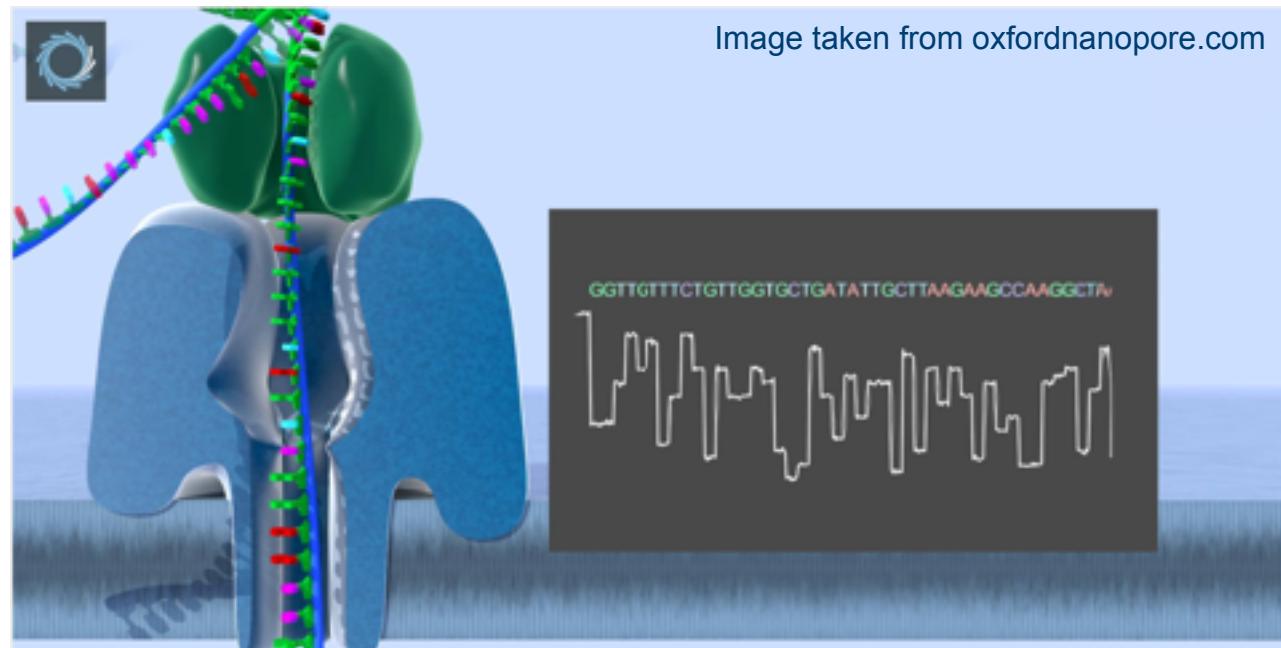


10X GENOMICS

4th Gen Approaches to sequencing

Cheap single molecule sequencing with Nanopores

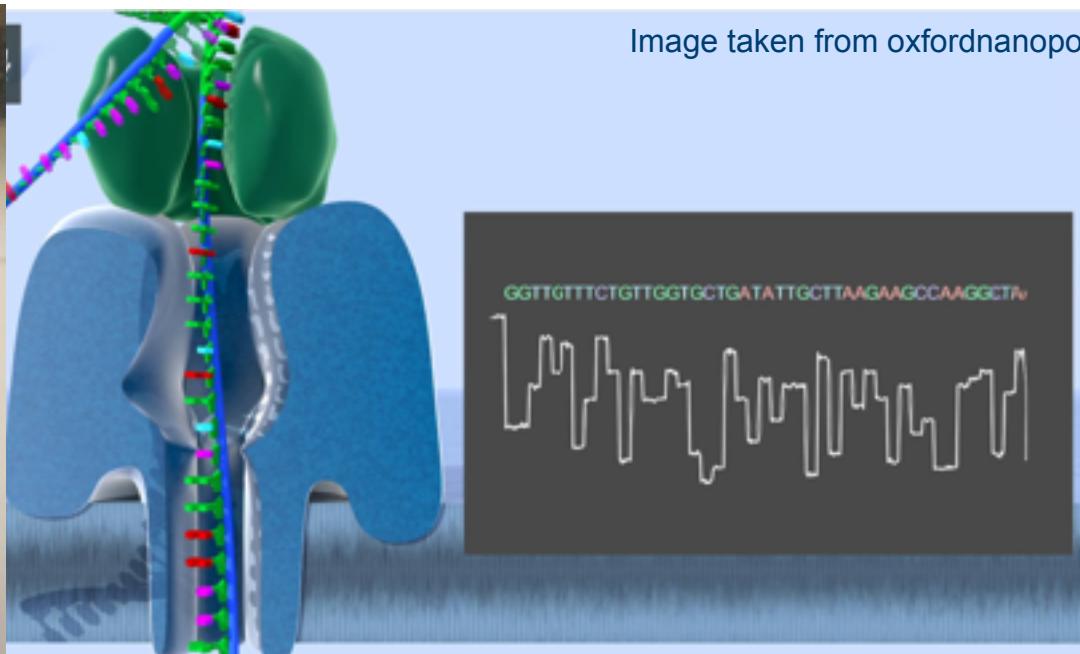
- gDNA fragmented and adapters attached
- gDNA ligated to a molecular motor
- Molecular motor attaches to a protein pore across an artificial membrane
- Voltage across each pore is measured in real-time
- Nucleotides can be called based on current changes



- Very long reads, 50-100kb possible
- Currently relatively high error rates
- Low cost, tiny machine

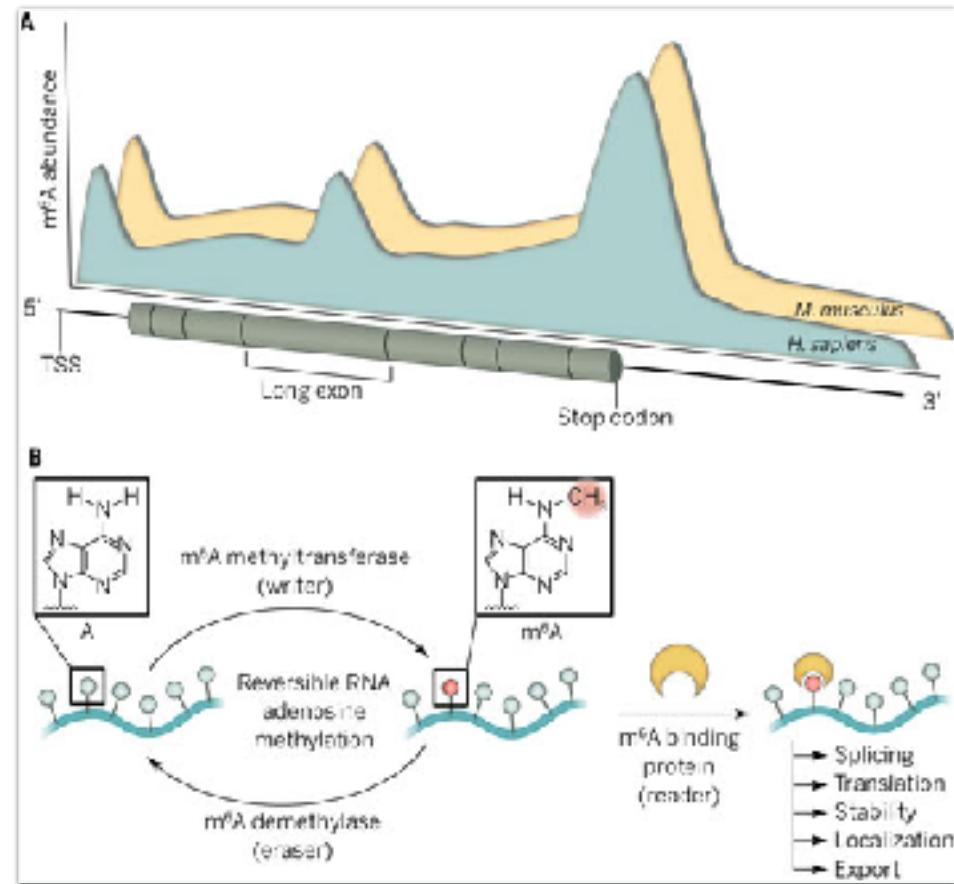
4th Gen Approaches to sequencing

Cheap single molecule sequencing with Nanopores



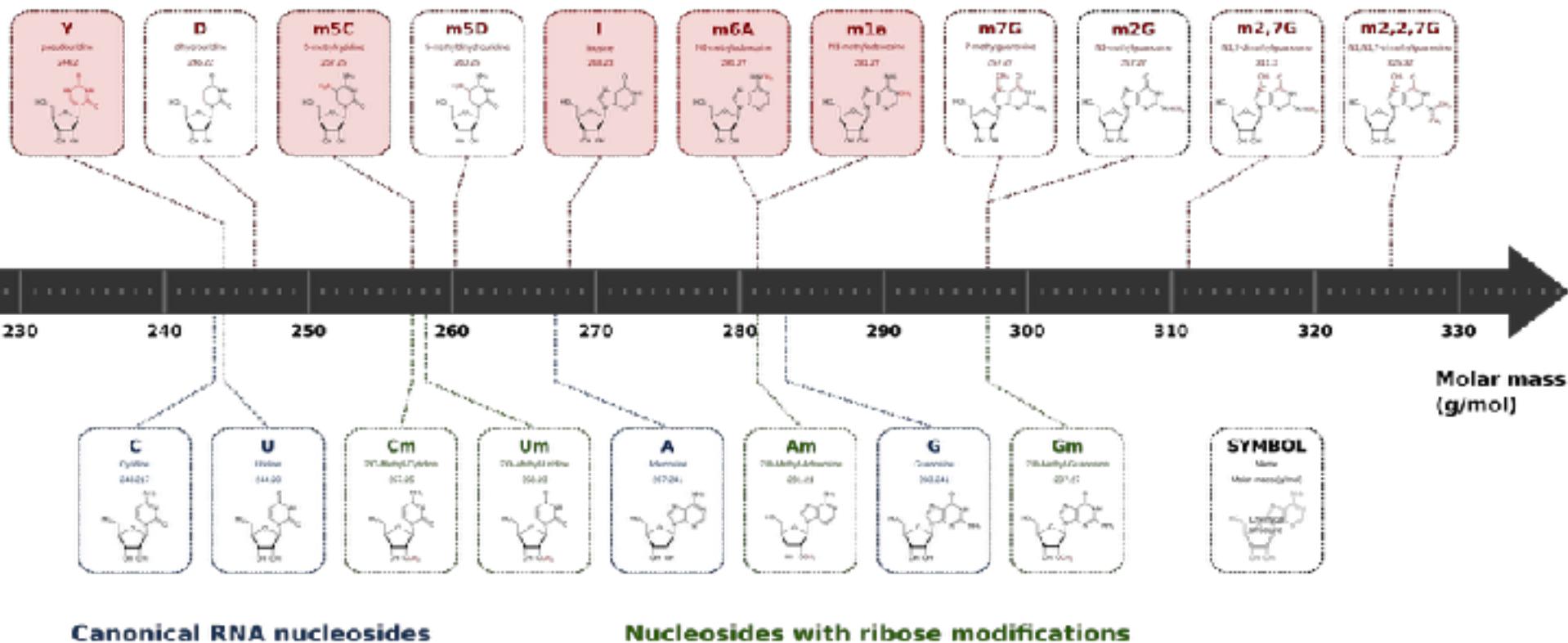
- Very long reads, 50-100kb possible
- Currently relatively high error rates
- Low cost, tiny machine

From Epi-Genetics to Epi-Transcriptomics

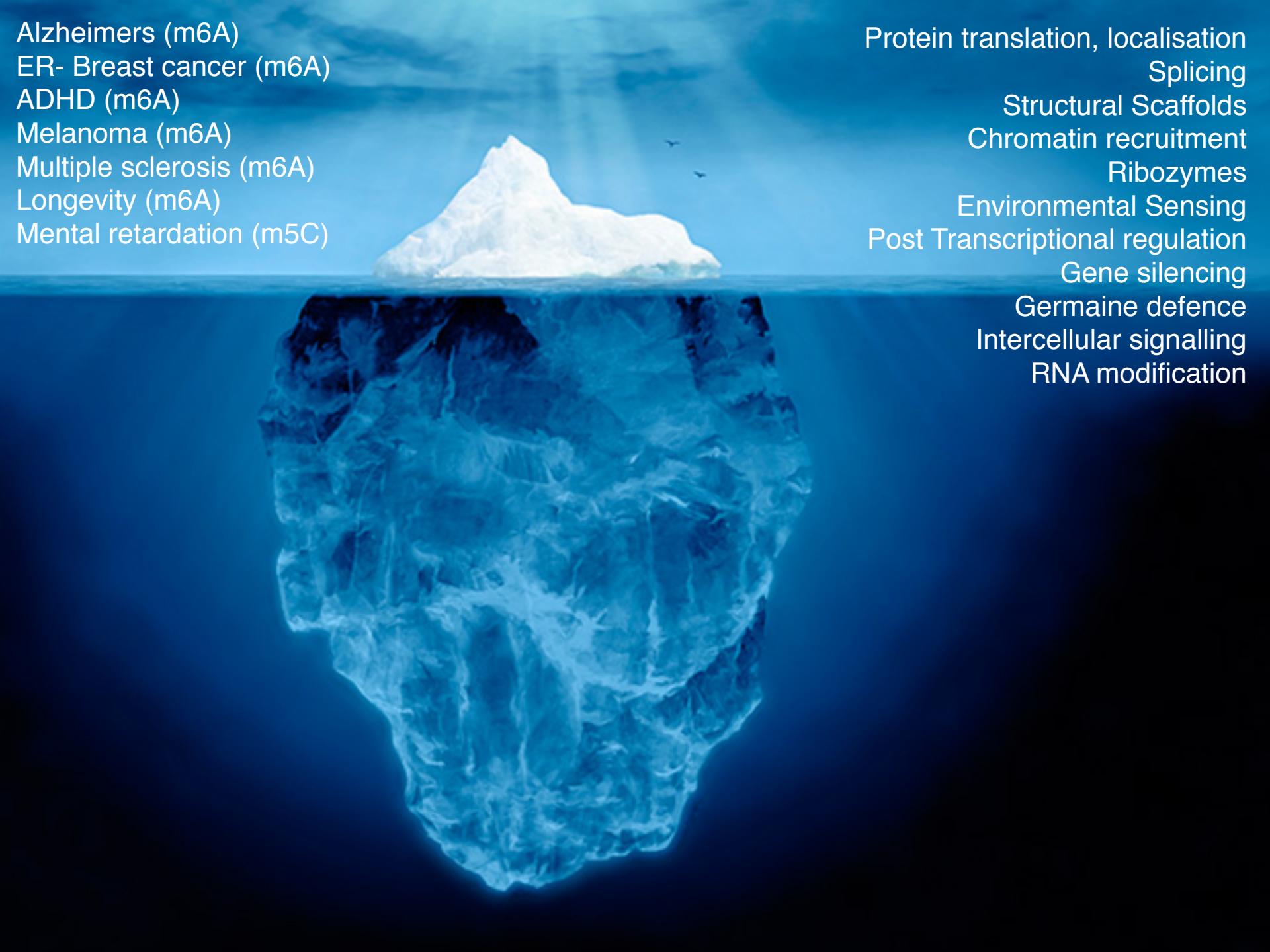


RNA modifications

Nucleosides with nitrogenous bases modifications



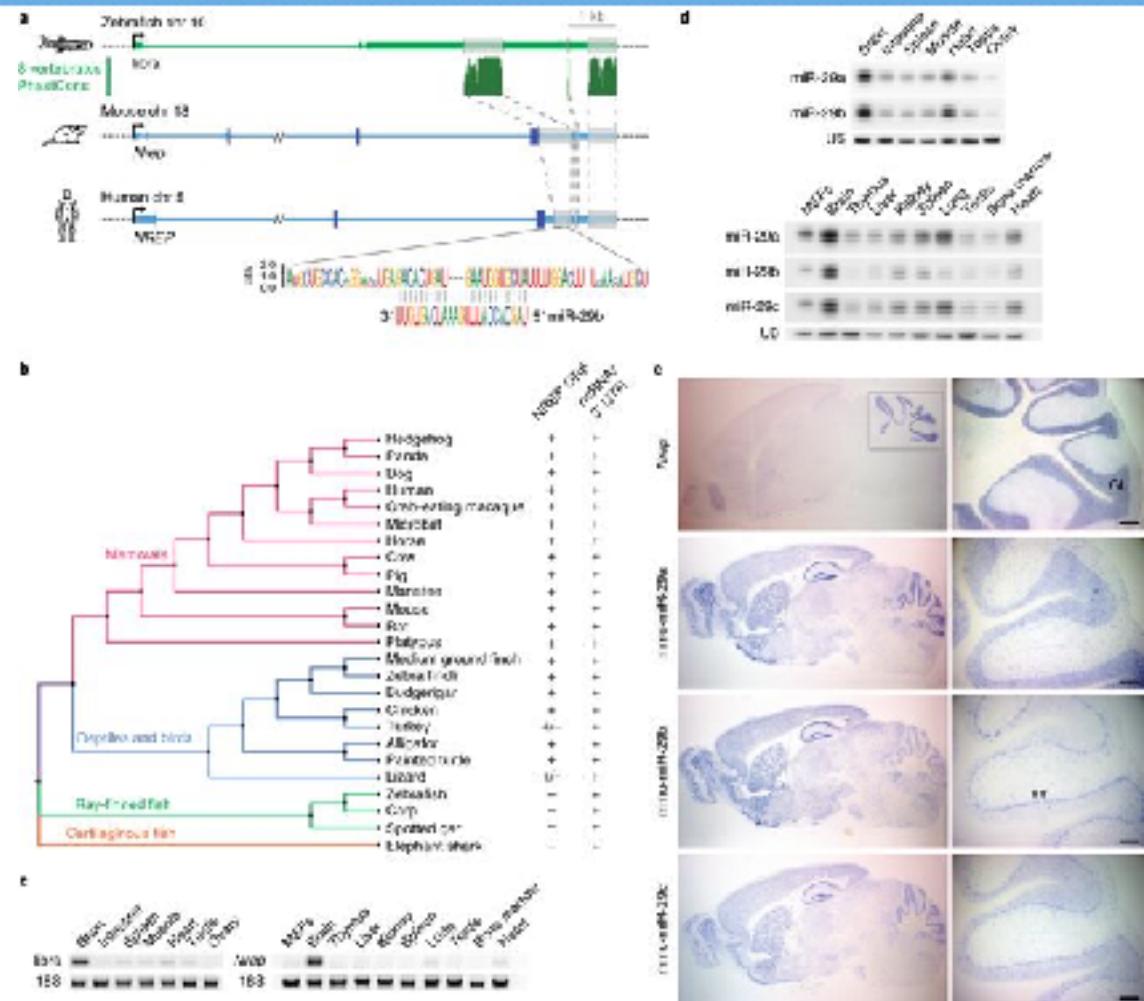
Alzheimers (m6A)
ER- Breast cancer (m6A)
ADHD (m6A)
Melanoma (m6A)
Multiple sclerosis (m6A)
Longevity (m6A)
Mental retardation (m5C)



Protein translation, localisation
Splicing
Structural Scaffolds
Chromatin recruitment
Ribozymes
Environmental Sensing
Post Transcriptional regulation
Gene silencing
Germaine defence
Intercellular signalling
RNA modification

WTAC RNA Epi-transcriptomics 2018

- Experimental data from **Alena Shkumatava and Allison Mallory**
- Conserved lncRNAs (cyrano, libra) Nrep
- Contain a non-canonical microRNA binding site



MicroRNA degradation by a conserved target RNA regulates animal behavior Angelo Bitetti, Allison C Mallory, Elisabetta Golini, Claudia Carrieri, Héctor Carreño Gutiérrez, Emerald Perlas, Yuvia A Pérez-Rico, Glauco P Tocchini-Valentini, Anton J Enright, William HJ Norton, Silvia Mandillo, Dónal O'Carroll, Alena Shkumatava - **NSMB 2018**

This years experiment

- **Functional elucidation of the role and effects of the NREP lncRNA**
 - **mouse NPCs with wt lncRNAs**
 - **mouse NPCs with mutant microRNA binding sites in the lncRNAs**
 - **mouse NPCs with reprogrammed microRNA binding sites**
 - **mouse NPCs with a 500bp deletion in the lncRNA**
 - Look for differential microRNA expression
 - Look for concomitant mRNA dysregulation
 - **Small RNA Seq, Single Cell RNA Seq, Direct Nanopore Seq (cDNA, Direct RNA)**

Computational Modules

- **Introduction to R and BioConductor**
- **Small RNA**
 - Sequence analysis
 - Differential expression analysis
 - Target detection
- **mRNA**
 - RNASeq data analysis and differential expression
 - Single Cell Analysis and differential expression
- **Analysis of Nanopore sequencing data**
 - cDNA
 - direct RNA
- **Bringing it all together**

Computational Materials

- **Computational Protocols available indefinitely (Open Source)**
 - Github
 - WWW
- **Computational Practicals performed on virtual machines (Virtual Box)**
- **Materials and Virtual Machines (VMs) available after the course**
- **You can repeat and reanalyse any of this data again at home, at your leisure.**

Practical Webpages - GitHub

The screenshot shows a GitHub repository interface for 'Course-and-Practicals'. The repository contains several files, including 'ctrl_vs_mc210.r' which is highlighted. The code in this file is used for 'Analysis of Sample Median Data'.

```
ctrl_mc210=apply(p[,c(2,3)],1,median)
ctrl_mc210=as.numeric(ctrl_mc210)

plot(ctrl_mc210,mc210_mc210,pch=20,main="ctrl vs mc210",col="darkblue",cex=0.8)
text(ctrl_mc210,mc210_mc210,mtc210_mc210,col="darkblue",cex=0.8)
text(ctrl_mc210,mc210_mc210,mtc210_mc210,col="red",cex=0.8)
text(ctrl_mc210,mc210_mc210,mtc210_mc210,col="green",cex=0.8)
```

The repository also includes a 'ctrl_vs_mc210.html' file, which displays a scatter plot titled 'ctrl vs mc210'. The x-axis is labeled 'ctrl_mc210' and the y-axis is labeled 'mc210_mc210'. The plot shows a strong positive linear correlation between the two variables, with data points colored by group (ctrl_mc210) and a red regression line.

- <http://tinyurl.com/wtac2018>

VirtualBox Installation

The image shows two side-by-side screenshots. On the left is a screenshot of the Oracle VM VirtualBox Manager application window. It has a toolbar at the top with icons for New, Settings, Start, and Shutdown. Below the toolbar is a section titled "Welcome to VirtualBox!" containing text and a small logo. On the right is a screenshot of the VirtualBox website at www.virtualbox.org. The page features a large "VirtualBox" logo and a "Download VirtualBox" section with links to binaries and source code.

Oracle VM VirtualBox Manager

Welcome to VirtualBox!

The left part of this window is a list of virtual machines. The list is empty now because you haven't created any yet.

In order to create a new virtual machine, click the **NEW** button in the main toolbar located above the window.

You can press the **Alt** key to get information about a specific item in the list.

About

Screenshots

Downloads

Documentation

User docs

Technical docs

Contribute

Community

virtualbox.org

VirtualBox

Download VirtualBox

Here, you will find links to VirtualBox binaries and its source code.

VirtualBox binaries

By downloading, you agree to the terms and conditions of the respective license.

- VirtualBox platform packages. The binaries are released under the terms of the GPL version 2.
 - VirtualBox 4.3.28 for Windows hosts [x86/amd64](#)
 - VirtualBox 4.3.28 for OS X hosts [x86/amd64](#)
 - VirtualBox 4.3.28 for Linux hosts [amd64](#)
 - VirtualBox 4.3.28 for Solaris hosts [amd64](#)
- VirtualBox 4.3.28 Oracle VM VirtualBox Extension Pack [All supported platforms](#)

Support for USB 2.0 devices, VirtualBox RDP and PXE boot for Intel cards. See [this chapter](#) from the [User Manual](#) for an introduction to this Extension Pack. The Extension Pack binaries are released under the [VirtualBox Personal Use and Evaluation License \(PUEL\)](#).
Please install the extension pack with the same version as your installed version of VirtualBox!
If you are using [VirtualBox 4.2.28](#), please download the extension pack [here](#).
If you are using [VirtualBox 4.1.36](#), please download the extension pack [here](#).
If you are using [VirtualBox 4.0.28](#), please download the extension pack [here](#).
- VirtualBox 4.3.28 Software Developer Kit (SDK) [All platforms](#)

See the [changelog](#) for what has changed.
You might want to compare the:
 - SHA256 checksums or the
 - MDS checksumsto verify the integrity of downloaded packages.

<https://www.virtualbox.org>

VirtualBox Installation

The image shows two side-by-side screenshots. On the left is a screenshot of the Oracle VM VirtualBox Manager application window. It displays a dialog box for creating a new virtual machine. The dialog has fields for 'Name' (WTAC image), 'Type' (Linux), and 'Version' (Ubuntu (64-bit)). A large blue starburst graphic is visible in the background of the manager window. On the right is a screenshot of the official VirtualBox website at [virtualbox.org](https://www.virtualbox.org). The page features a large 'VirtualBox' logo and a 'Download VirtualBox' section. This section contains links for downloading binaries for various hosts (Windows, OS X, Linux, Solaris) and an 'Extension Pack'. It also includes a link to the 'Software Developer Kit (SDK)'.

<https://www.virtualbox.org>

VirtualBox Installation

The image shows two side-by-side screenshots. On the left is a screenshot of the Oracle VM VirtualBox Manager interface, specifically the 'Create New Virtual Machine' wizard. It displays a 'Memory size' slider with a current value of 4 MB and a maximum of 4096 MB. On the right is a screenshot of the official VirtualBox website at [virtualbox.org](https://www.virtualbox.org). The page features a large 'VirtualBox' logo and a 'Download VirtualBox' section. This section contains links for various binary packages, including 'VirtualBox 4.3.28 for Windows hosts', 'VirtualBox 4.3.28 for OS X hosts', 'VirtualBox 4.3.28 for Linux hosts', and 'VirtualBox 4.3.28 for Solaris hosts'. It also mentions the 'VirtualBox 4.3.28 Oracle VM VirtualBox Extension Pack' and the 'VirtualBox 4.3.28 Software Developer Kit (SDK)'. A sidebar on the left of the website page lists links for 'About', 'Screenshots', 'Downloads', 'Documentation', 'End-user docs', 'Technical docs', 'Contribute', and 'Community'.

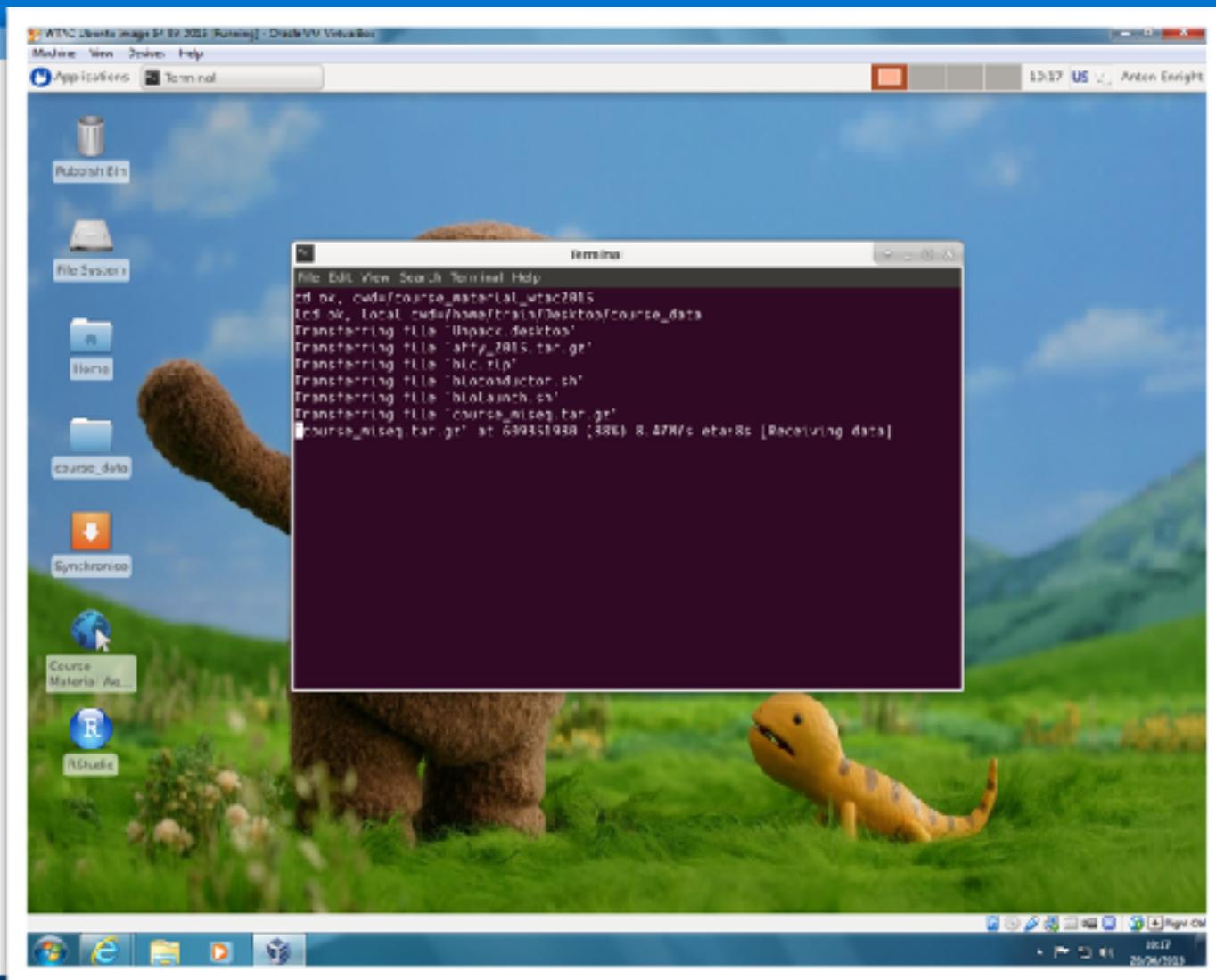
<https://www.virtualbox.org>

VirtualBox Installation

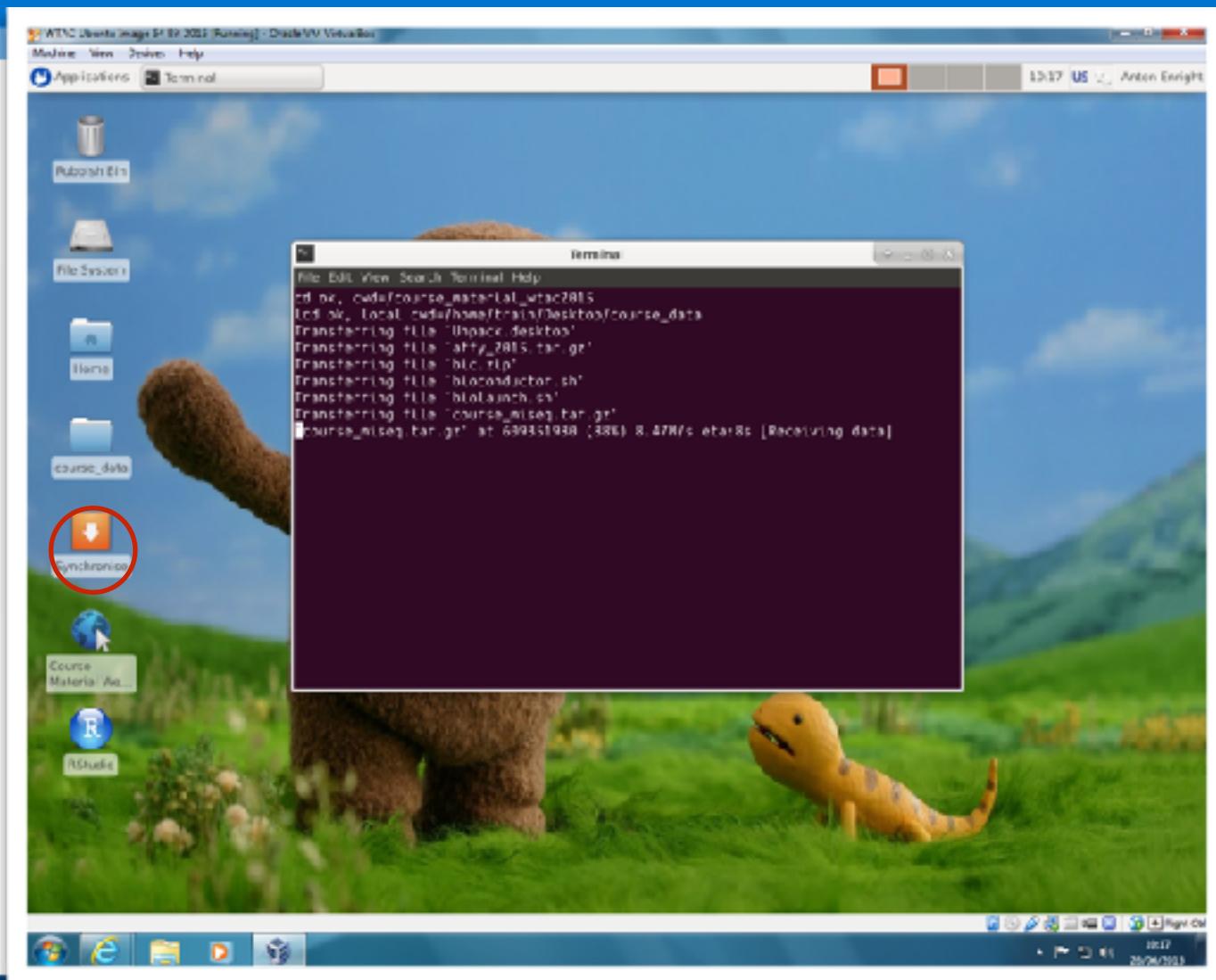
The image shows two side-by-side screenshots. On the left is a screenshot of the Oracle VM VirtualBox Manager application. It has a toolbar with 'New', 'Settings', 'Start', and 'Shutdown' buttons. Below that is a section titled 'Hard drive' with a large blue starburst graphic. It contains text about adding a virtual machine, creating a new hard drive, or selecting an existing one. There are three radio button options: 'Do not add a virtual hard drive' (selected), 'Create a virtual hard drive now', and 'Use an existing virtual hard drive'. A 'Empty' button is at the bottom. On the right is a screenshot of the VirtualBox download page at [virtualbox.org](https://www.virtualbox.org). The page features a large 'VirtualBox' logo and a 'Download VirtualBox' section. It includes links for 'About', 'Screenshots', 'Downloads', 'Documentation', 'End-user docs', 'Technical docs', 'Contribute', and 'Community'. The 'Downloads' section is expanded, showing links for various platform packages and the Oracle VM VirtualBox Extension Pack. It also includes a note about SHA256 and MD5 checksums for package integrity verification.

<https://www.virtualbox.org>

Virtual Machine

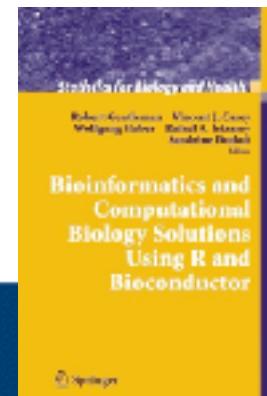


Virtual Machine



R & BioConductor

- Freely available
- Updated constantly
- Available here:
 - <http://www.r-project.org/>
 - <http://www.bioconductor.org/>
- LIMMA user guide and the Springer book series are well worth investigating
- DESeq2 Vignette and Manual Pages excellent also
- Full R courses available



small RNA Seq Workflow

- **Read cleanup and QC in REAPER (Enrightlab R Module)**
 - Adapter Removal
 - Contaminant Removal
 - Size Selection 15-28nt
- **Read mapping using ChimiRa (Enrightlab)**
 - BLAST Based mapping against miRBase precursors
- **Differential Expression Analysis (Francesca Buffa)**
 - microRNA calls in multiple samples

Chimira Web Server for small RNA Seq

<http://wwwdev.ebi.ac.uk/enright-srv/chimira>

Bioinformatics Advance Access published June 20, 2015

Application Note

Chimira: Analysis of small RNA Sequencing data and microRNA modifications

Dimitrios M. Vitsios and Anton J. Enright*

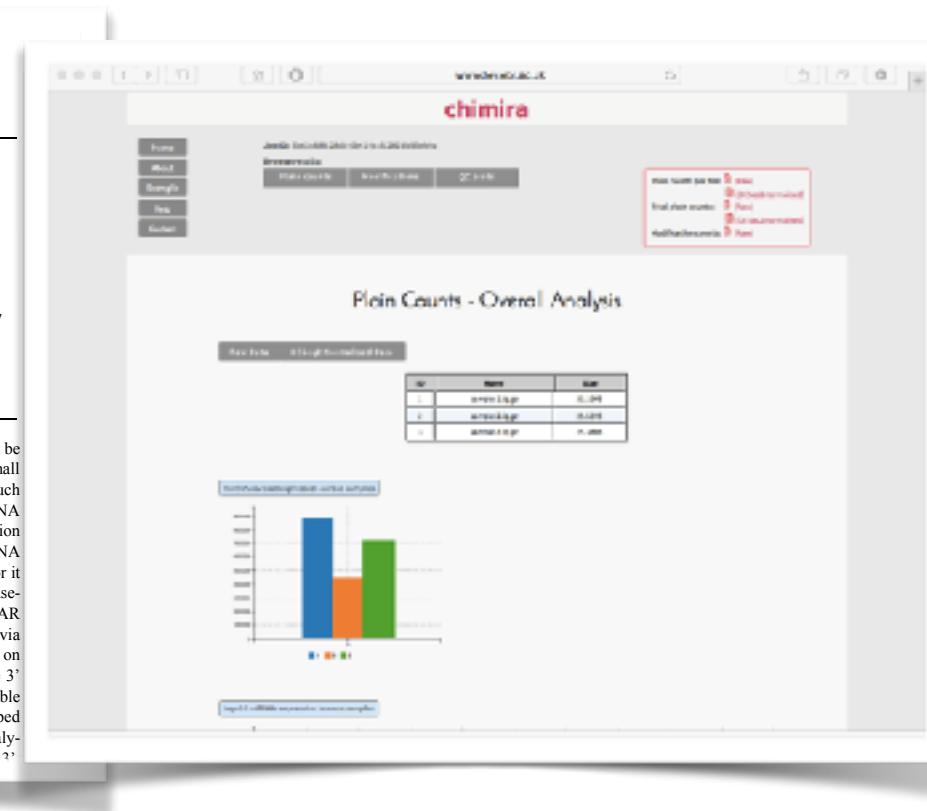
EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Associate Editor: Prof. Ivo Hofacker

ABSTRACT

Summary: Chimira is a web-based system for microRNA (miRNA) analysis from small RNA-Seq data. Sequences are automatically cleaned, trimmed, size selected and mapped directly to miRNA hairpin sequences. This generates count-based miRNA expression data for subsequent statistical analysis. Moreover, it is capable of identifying epi-transcriptomic modifications in the input sequences. Supported modification types include multiple types of 3'-modifications (e.g. uridylation, adenylation), 5'-modifications and also internal modifications or variation (ADAR editing or SNPs). Besides cleaning and mapping of input sequences to miRNAs (Griffiths-Jones et al., 2008), Chimira provides a simple and intuitive set of tools for the analysis and interpretation of the results (see also Supplementary material). These allow the visual study of the differential expression between two specific samples or sets of samples, the identification of the most highly expressed miRNAs within sample pairs (or sets of

plore the full profile of all modifications and/or edits that can be identified in small RNA-Seq data. The functional roles of small RNAs in different conditions may be greatly influenced by such modifications. This can be accomplished by aligning small RNA sequences against their hairpin precursors. The alignment region spanning each miRNA is analysed to detect bases in the miRNA sequence that could not possibly have derived from the precursor it aligns to. These unalignable nucleotides are likely either: i) base-calling errors, ii) single nucleotide polymorphisms, iii) ADAR edits or iv) post-transcriptional miRNA modifications (e.g. via TUTases). Base-calling errors are pseudo-random depending on the platform used and usually more likely to occur towards the 3' end of sequences. In order to study this diverse pool of possible miRNA post-transcriptional modifications, we have developed Chimira. This is a cohesive platform for the processing and analysis of small RNA NGS data allowing simultaneous detection of 3'



mRNA Seq: Analysis Types

Reference Transcriptome

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2

Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

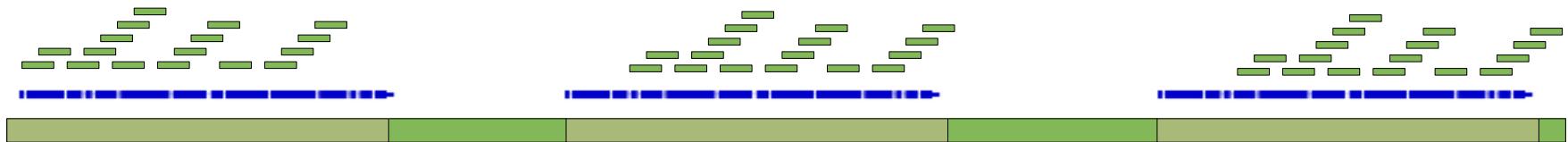
mRNA Seq: Analysis Types

Reference Transcriptome

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

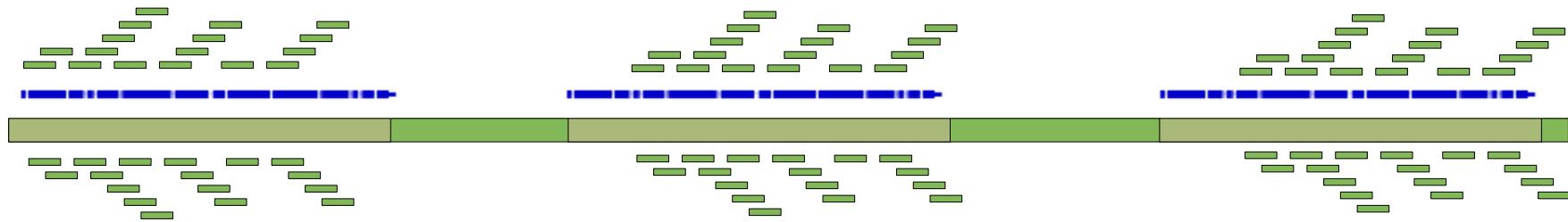
mRNA Seq: Analysis Types

Reference Transcriptome

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

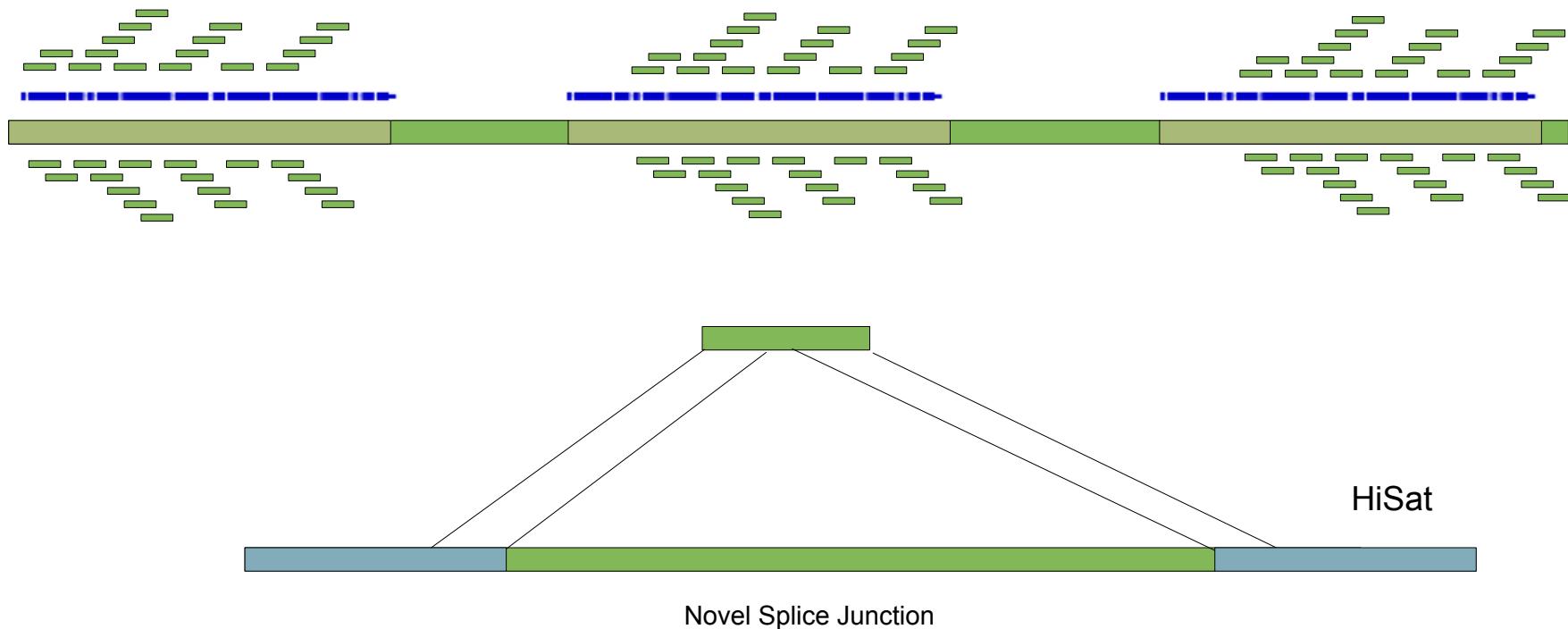
mRNA Seq: Analysis Types

Reference Transcriptome

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

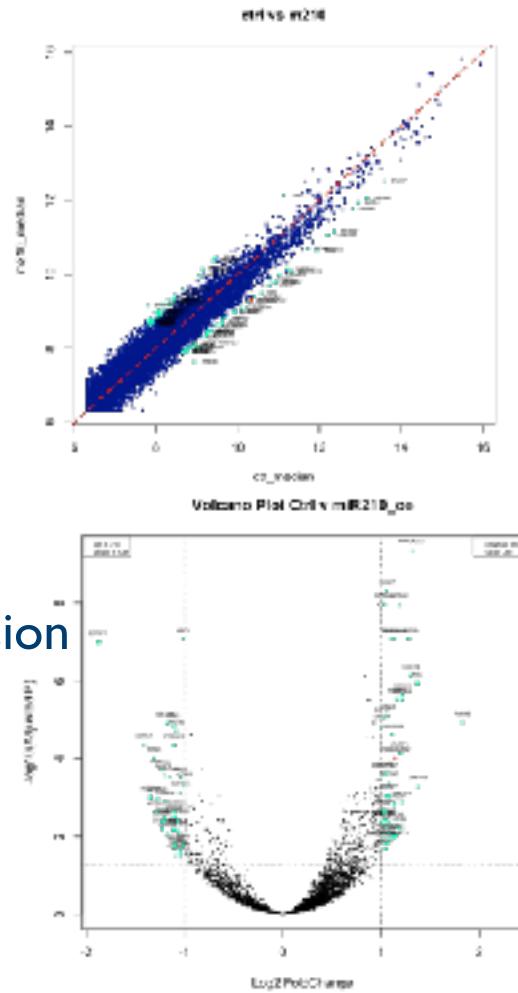
R/BioConductor Downstream analysis with DESeq2

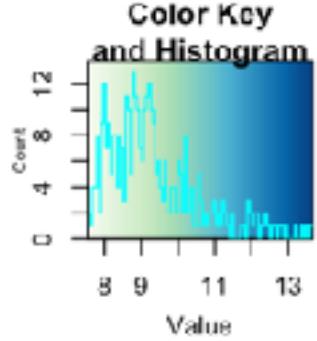


Results in count data and FPKMs

mRNA Seq Analysis

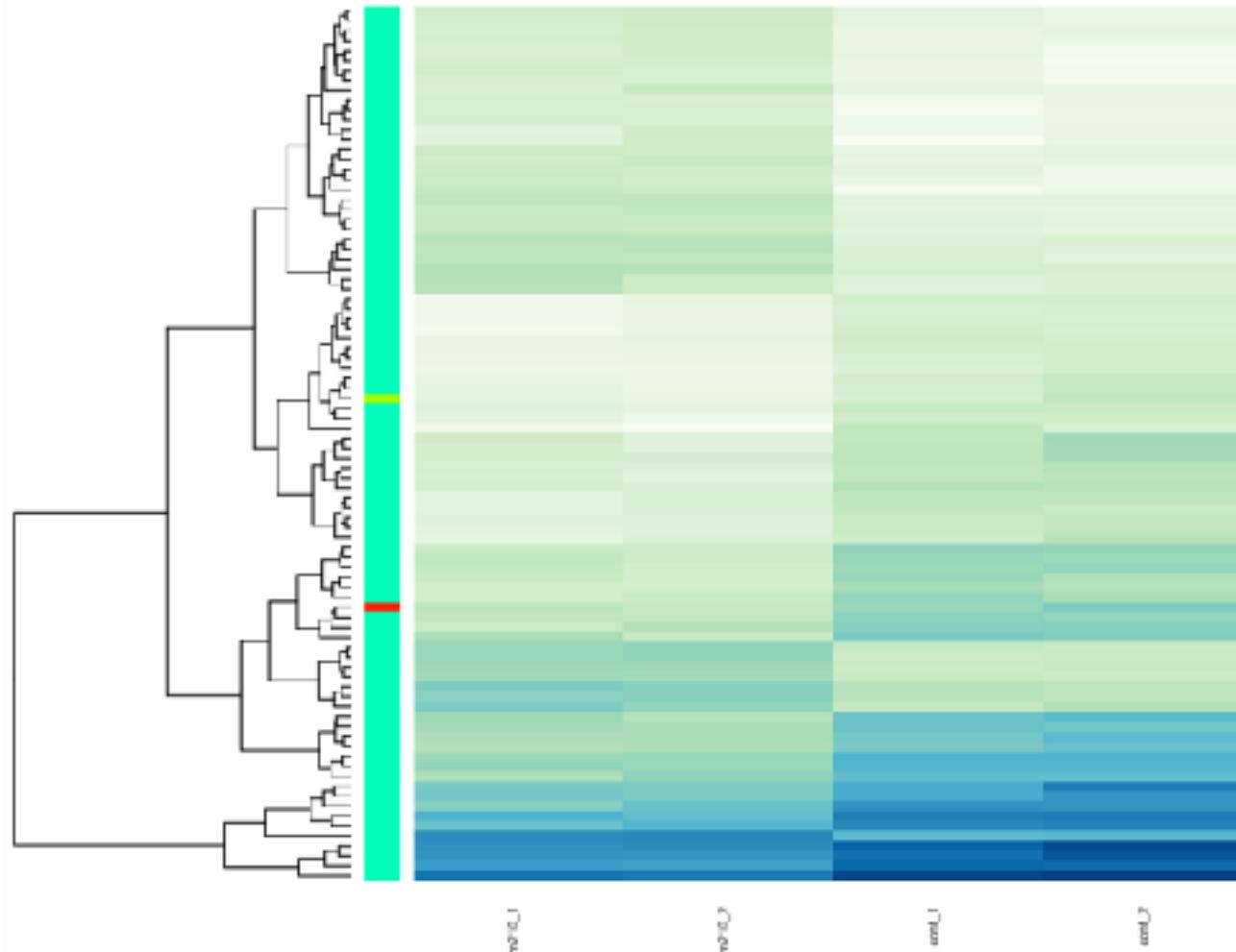
- HiSat2 vs Human Reference Genome from Ensembl
 - Does NOT run easily in windows, better in UNIX or Mac
- Genome Annotations from Ensembl GTF
- HTSeq-count and DESeq2 to perform differential expression
- Analysis of Results in R/BioConductor



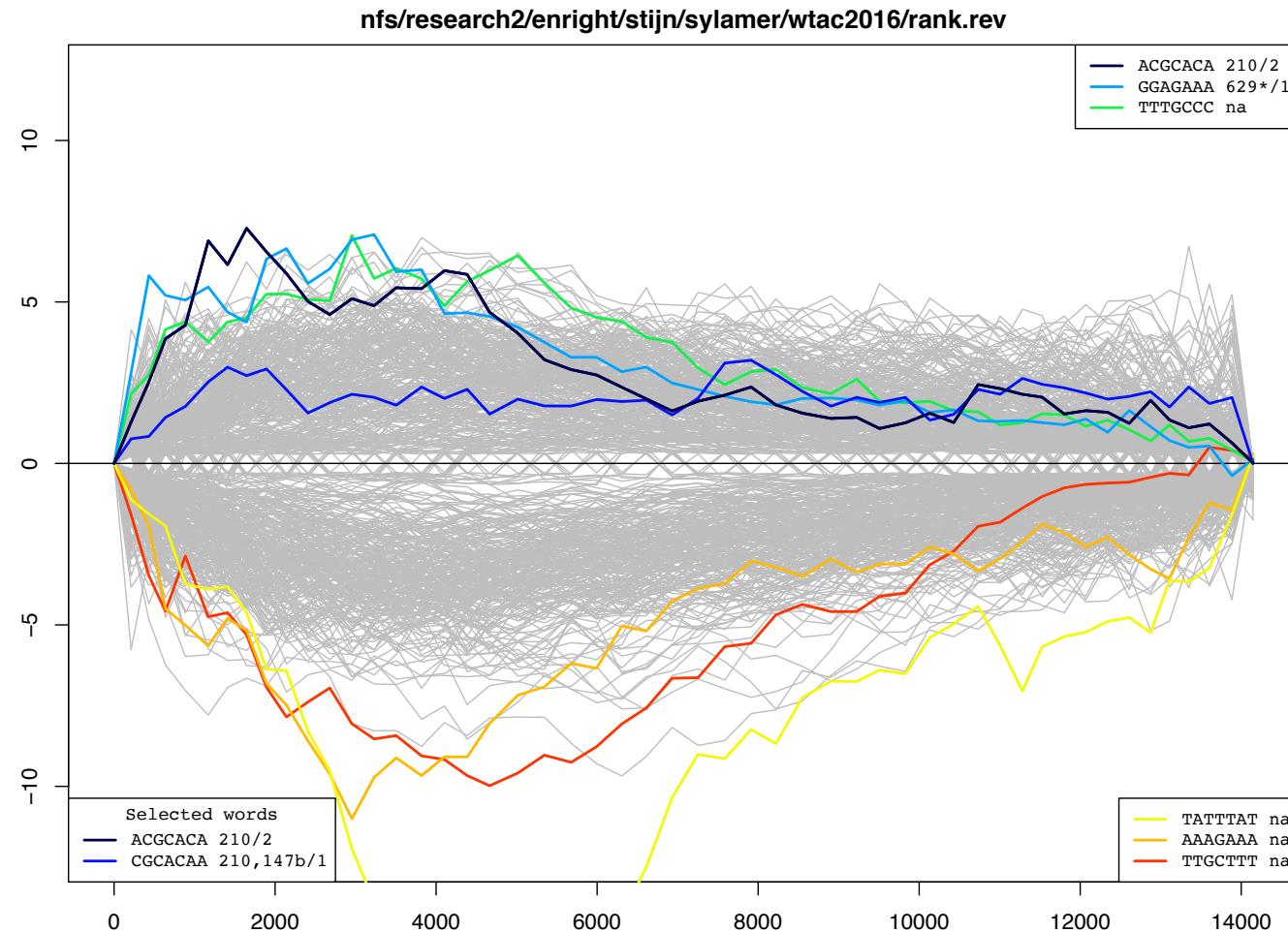


Heatmap Sig. Hits

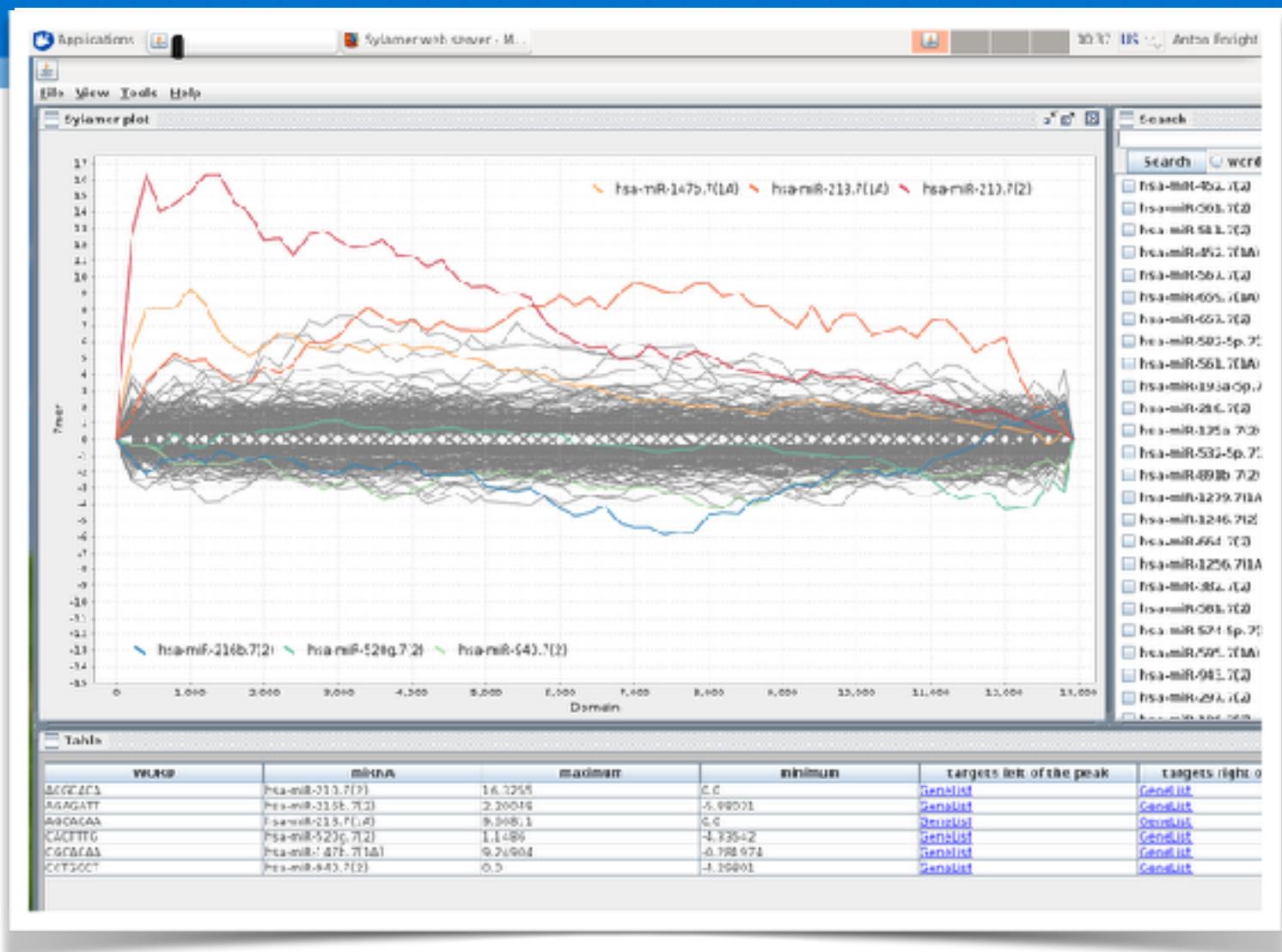
- HiSeq
 - DoE
 - Gene
 - HTS
 - Ana



Functional MicroRNA Analysis Sylamer

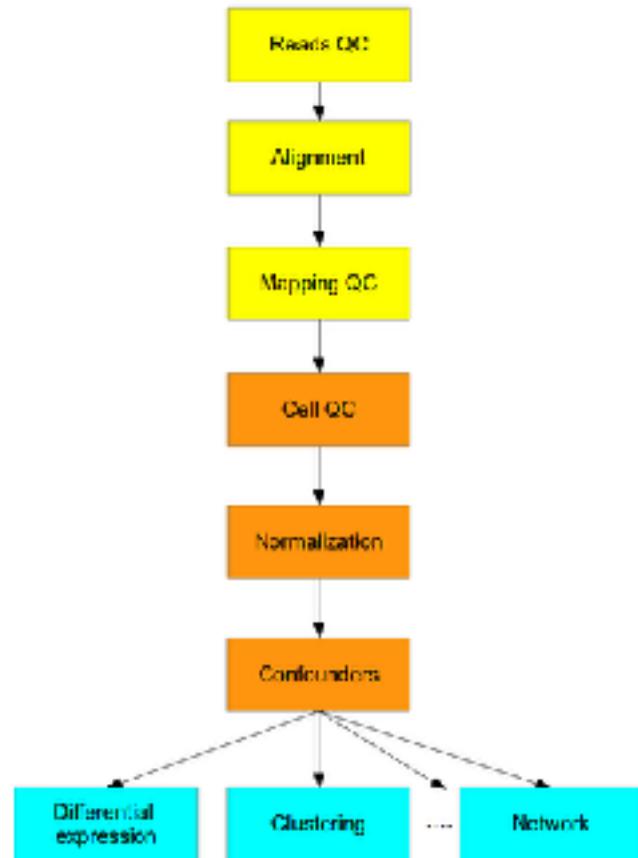
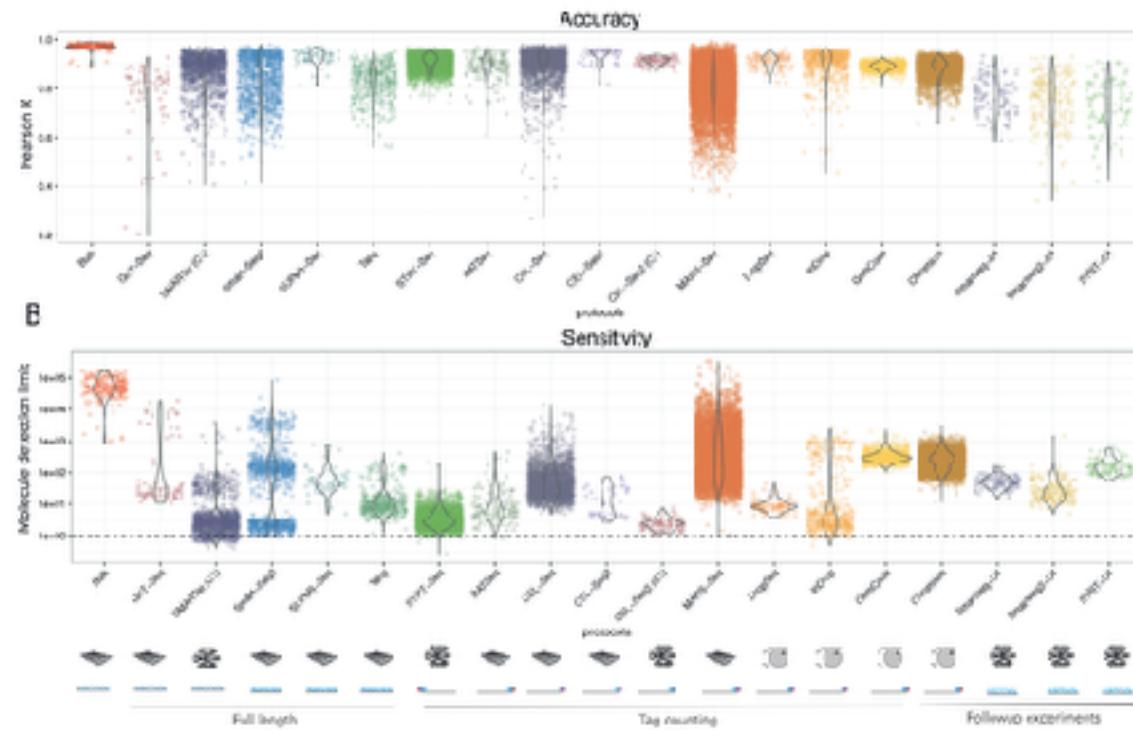


Functional MicroRNA Analysis Sylamer



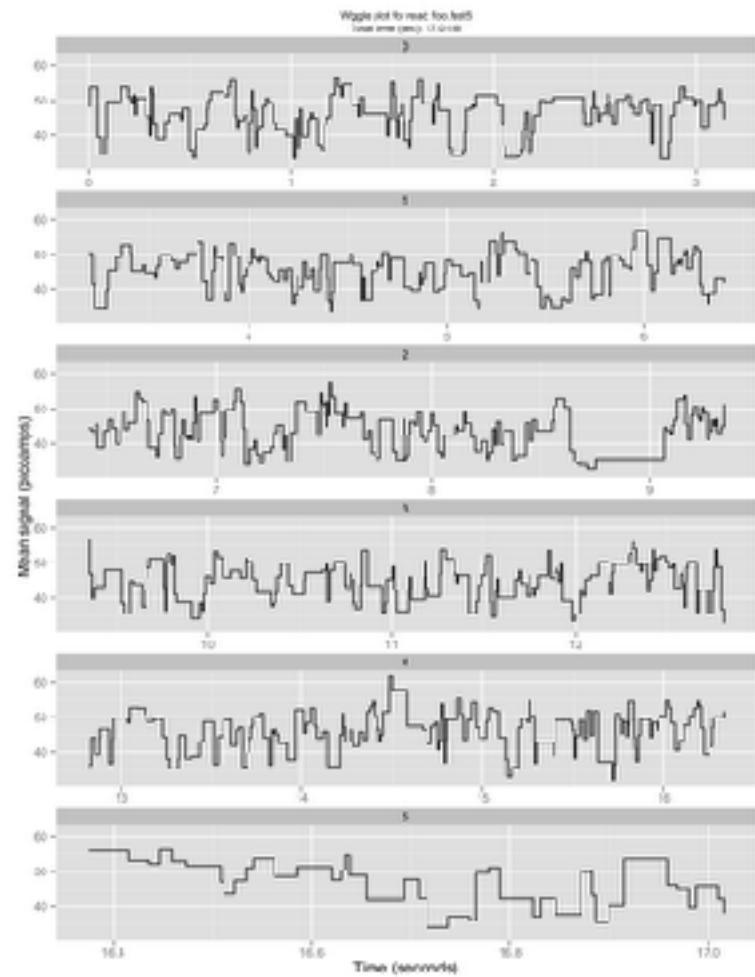
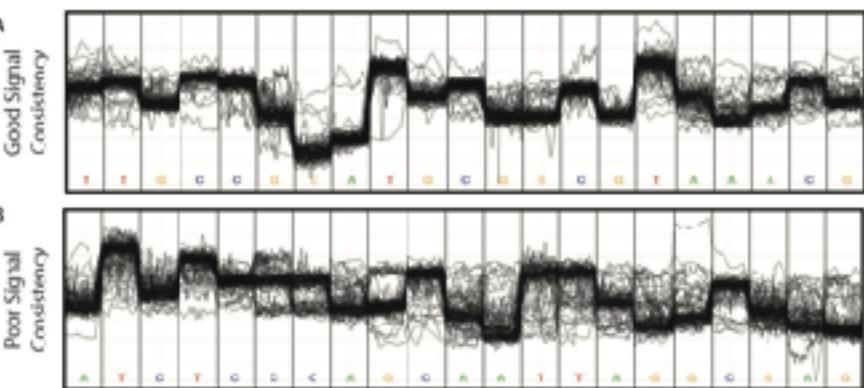
Single Cell Analysis

- Martin Hemberg and Tallulah Andrews



Nanopore Analysis

- Run Setup
- Live Monitoring of Sequencing Run
- Basecalling
- QC of pores, sequences and samples
- Analysis of detected sequences
- Comparison between cDNA and RNA



Computational Section - Instructors & Assistants

- **Anton Enright (small RNA Seq, mRNA Seq, Nanopore Analysis)**
 - Jack Monahan & Adrien Leger &
- **Jiannis Ragoussis (cDNA Nanopore Seq)**
 - Spyridon Oikonomopoulos
- **Francesca Buffa (Small RNA Seq, Small RNA Functional Analysis, Networks)**
 - Alessandro Barberis
- **Martin Hemberg (Single Cell RNA Seq)**
 - Tallulah Andrews

Final Points

Final Points

- This is a new course

Final Points

- This is a new course
- We are using new technologies

Final Points

- This is a new course
- We are using new technologies
- We are generating a tremendous amount of data in a short time

Final Points

- This is a new course
- We are using new technologies
- We are generating a tremendous amount of data in a short time
- We aim to teach the whole experiment from design to analysis

Final Points

- This is a new course
- We are using new technologies
- We are generating a tremendous amount of data in a short time
- We aim to teach the whole experiment from design to analysis

Final Points

- This is a new course
- We are using new technologies
- We are generating a tremendous amount of data in a short time
- We aim to teach the whole experiment from design to analysis
- Things can and will go wrong.....

R & BioConductor

- Freely available



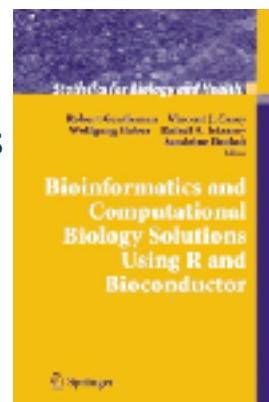
- Updated constantly

- Available here:

- <http://www.r-project.org/>
- <http://www.bioconductor.org/>

- DeSeq2 & LIMMA user guide and the Springer book series

- Full R courses available



R Commands

- Parentheses used to give options and parameters
- Syntax very important
- Commas used to separate options
- **commandname(option1,option2,option3)**

- When using text you should always put in quotes
- **print("My Name is Anton")**

R Objects

- You create objects using the `=` sign in R
- Sometimes people use `<-` instead of `=`
 - `mynewobject <- 3`
 - `mynewobject = 3`
 - `mynewobject = "I stole Kates Chocolate"`
 - `mynewobject = median(c(43,534,643,432))`

R Data Types

Vector (Lists)

character	the cat ran over the dog
numeric	43 546 34 123 43 54
logical	T F F T F T

Factor

list	AB AC AB AD AD DE
levels(list)	AB AC AD DE

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	I	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types - Vector

Vector (Lists)

character `the cat ran over the dog`

numeric `43 546 34 123 43 54`

logical `T F F T F T`

```
mycharacterlist = c("the","cat","ran","over","the","dog")
```

```
mynumericlist = c(43,546,34,123,43,54)
```

```
mylogicalist = c(TRUE, FALSE, FALSE, TRUE, FALSE, TRUE)
```

R Data Types - Factor

Factor

```
list    AB AC AB AD AD DE
```

```
levels(list) AB AC AD DE
```

```
mylist = c("AB","AC","AB","AD","AD","DE")
```

```
myfactor = as.factor(mylist)
```

```
levels(myfactor)
```

R Data Types - Matrix

```
y = matrix(1:25, nrow=5,ncol=5)
```

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types - Matrix

```
y = matrix(1:25, nrow=5,ncol=5)
```

```
y[1,2]
```

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types - Matrix

```
y = matrix(1:25, nrow=5,ncol=5)
```

```
y[1,2]
```

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types - Matrix

```
y = matrix(1:25, nrow=5,ncol=5)
```

```
y[1,2]
```

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types - Matrix

```
y = matrix(1:25, nrow=5,ncol=5)
```

```
y[1,2]
```

```
y[4,]
```

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types - Matrix

```
y = matrix(1:25, nrow=5,ncol=5)
```

```
y[1,2]
```

```
y[4,]
```

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types - Matrix

```
y = matrix(1:25, nrow=5,ncol=5)
```

```
y[1,2]
```

```
y[4,]
```

```
y[,3]
```

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

R Data Types

Vector (Lists)

character `the cat ran over the dog`

numeric `43 546 34 123 43 54`

logical `T F F T F T`

Factor

list `AB AC AB AD AD DE`

levels(list) `AB AC AD DE`

Matrix

	col1	col2	col3	col4	col5	col6
Row1	12	43	453	43	34	233
Row2	43	546	34	123	43	54
Row3	33	43	234	43	34	234
Row4	34	342	34	34	55	362
Row5	43	546	34	123	43	54
Row6	33	343	76	96	45	766

Data.Frame

	col1	col2	col3	col4	col5	col6
Row1	The	43	453	T	34	233
Row2	cat	546	34	F	43	54
Row3	ran	43	234	T	34	234
Row4	over	342	34	F	55	362
Row5	the	546	34	T	43	54
Row6	dog	343	76	T	45	766

Practicals for Today

- Introduction to R/BioConductor
- Raw analysis of small RNA Illumina FASTQ files
 - Adapter Identification
 - Adapter Stripping
 - Raw Read QC and Analysis
 - Mapping to miRBase with the Chimira webserver
 - Generation of counts tables
 - Full DESeq2 Statistical Analysis of count data
- Analysis of smallRNA tissue seq
- Analysis of experiment miR210 over-expression
 - smallRNA Seq
 - mRNA Seq (matched)
 - Sylamer analysis for miR-210 target prediction

First Example Dataset

A conserved RNA regulates miRNA turnover and animal behavior through a near-perfect miRNA site

Angelo Bitetti^{1,2,3,*}, Allison C Mallory^{1,2,3,*}, Claudia Carrieri⁴, Elisabetta Golini⁵, Hector Carreño Gutierrez⁶, Emerald Perlas⁷, Yuvia A. Pérez-Rico^{1,2,3,8}, Glauco P. Tocchini-Valentini⁵, Anton J. Enright⁹, William H. J. Norton⁶, Silvia Mandillo⁵, Dónal O'Carroll⁴ and Alena Shkumatava^{1,2,3}

Mouse ES Cells
Sequenced on Illumina

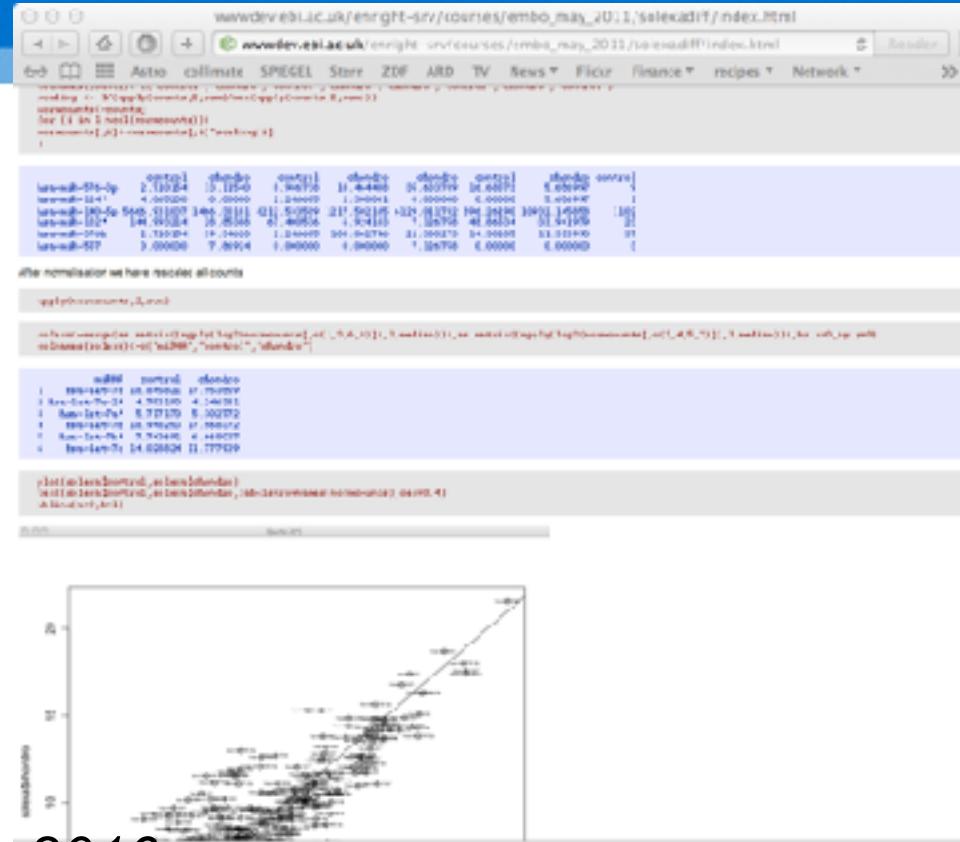
3 Wildtype
3 Scrambled controls (Nrep lncRNA)

Second Example Dataset

Breast Cancer Samples
(Ioannis Raggousis,
McGill University, Montreal)

MCF7 Line
3 Scrambled Controls
3 miR-210 knockdowns

Sequenced on Illumina
MiSeq
Wellcome Trust Advanced Course 2016



Considerations for smallRNA seq analysis

- Single-end sequencing 50nt
- Typically more samples fail in a smallRNA seq analysis
- Replicates are essential
- Aim for 4-5 biological replicates per sample if possible
- Huge sequencing depth is not usually required per sample
- 4M reads per sample can usually be sufficient
- 20M reads per sample is ideal
- Spike-ins are sometimes used
 - Can be misleading due to sequencing biases
 - Can be hard to control quantity and accuracy

Downstream Analysis

microRNA Normalisation
and Differential expression

Small RNA Seq Normalisation

- **Lane Depth Correction**
 - Correct total miRNA mapping counts to be equal across all samples
- **Statistical approaches better for normalisation and differential expression**
 - DEseq
 - BaySeq
 - EdgeR

Lane1	Lane2	Lane3	Lane1	Lane2	Lane3
10	34	33	23.1081081	34	30.1875
11	32	19	25.4180189	37	27.5625
24	55	44	55.4504505	65	59.75
11	28	21	25.4180189	35	24.9375
11	22	21	25.4180189	27	24.9375
7	5	6	16.1756257	5	7.125
Total	171	141	Total	171	171
Scaling	0.000710534	1.000230528			

Mitigating biases ?

- Try and ensure that each replicate of a sample has a different adapter sequence ligated.
 - i.e. Don't multiplex replicates identically.
- 5' Ligation of random nucleotides



- Spike-Ins ?



Normalisation of microRNA Count Data

- Correction via total depth normalisation not a good idea
- Count based data should always be normalised with a negative binomial approach
 - **DESeq2**
 - **EdgeR**
 - **Voom**
- Spike-ins difficult to get working correctly for smallRNA runs
- snoRNA can sometimes work to assist normalisation of microRNA samples with extreme differences (e.g. Dicer Knockdown).

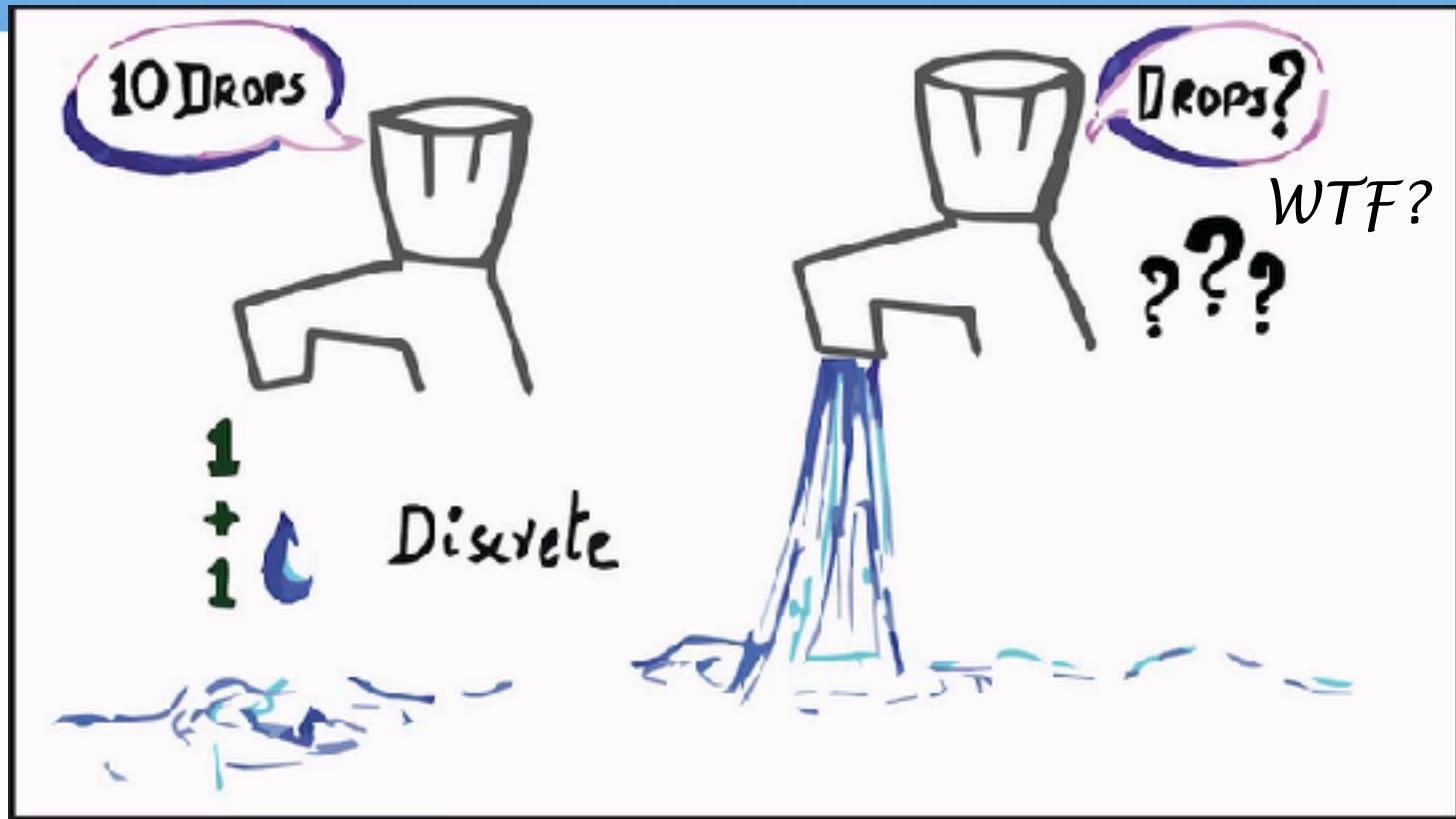
Count Data is Different! - NGS Analysis



Count Data is Different! - NGS Analysis

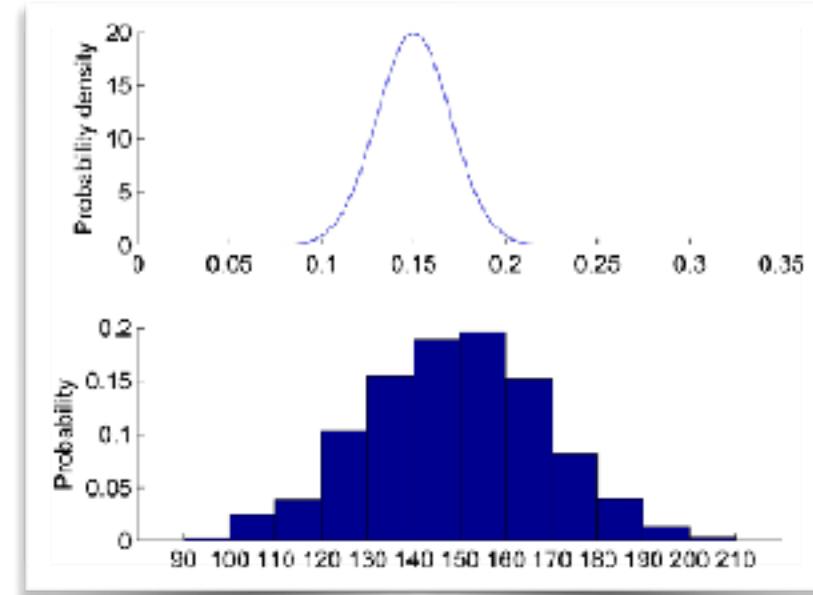


Count Data is Different! - NGS Analysis



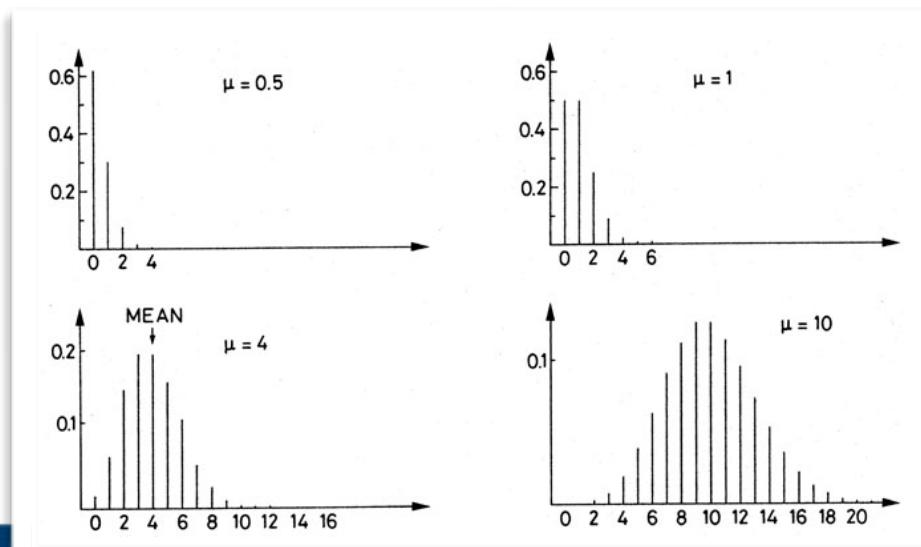
Sequencing Data Statistics

- Sequencing data produces counts not measurements
 - Forms a discrete distribution
 - Very different from continuous distributions (e.g. intensities from a microarray probe)
 - Positive values
 - Skewed
 - Heteroscedastic
 - Massive dynamic range
 - Large differences sample to sample



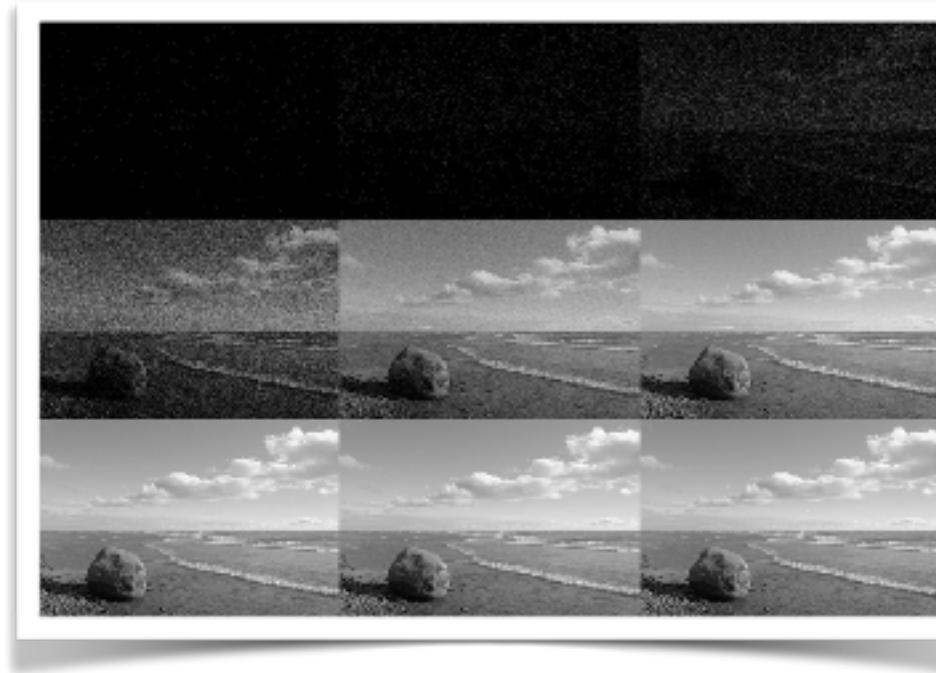
Discrete Count Data - How to model it

- **The Poisson Distribution**
- Example: A short, light rain shower with r drops per meter².
- What is the probability to find k drops on a paving stone of size 1m²?
- For poisson, the variance should equal the mean



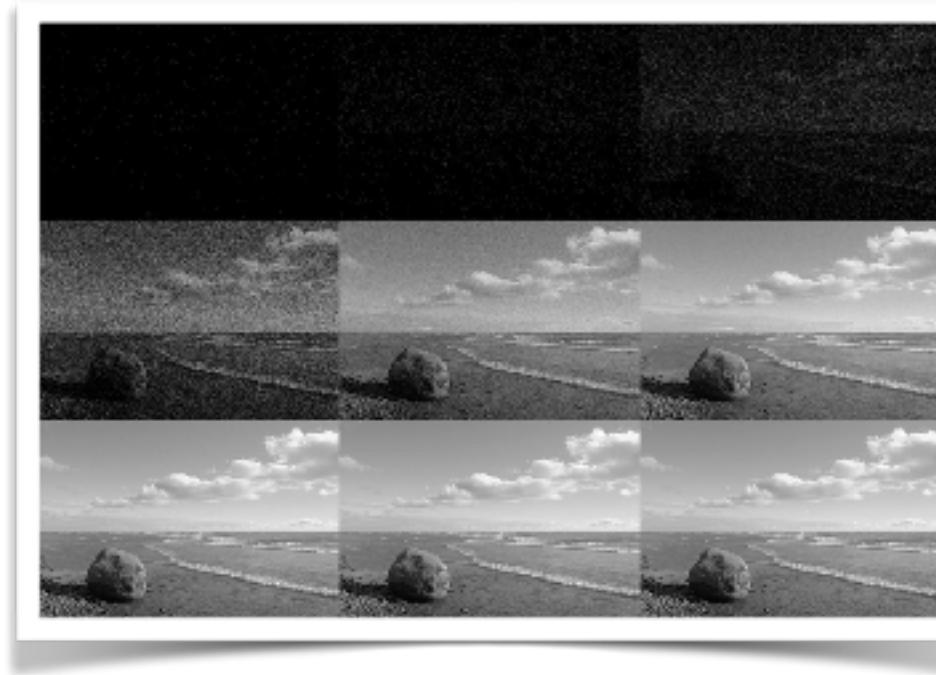
Discrete Count Data

- Shot Noise



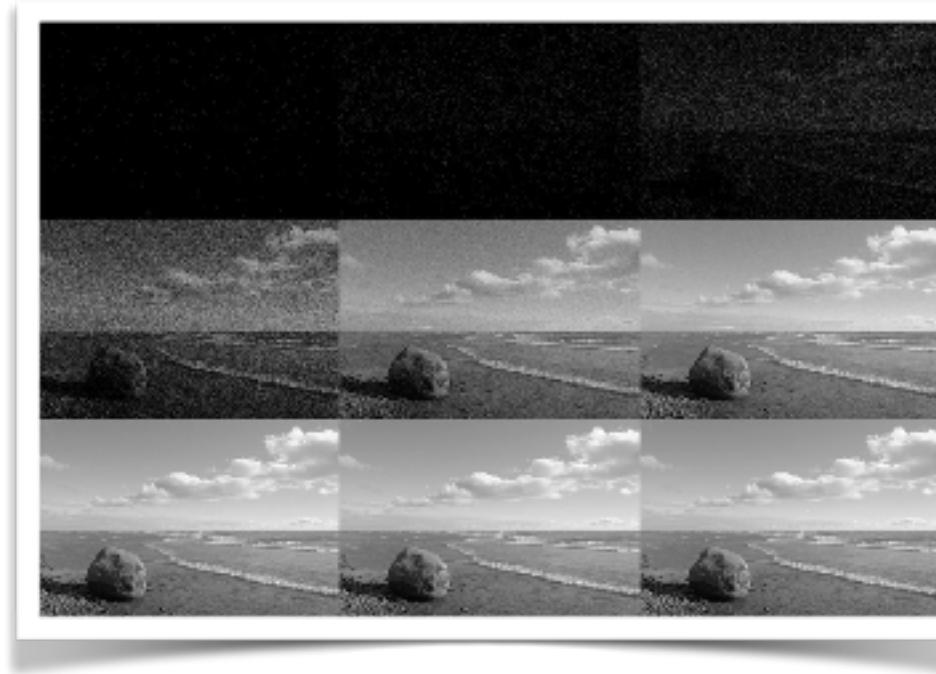
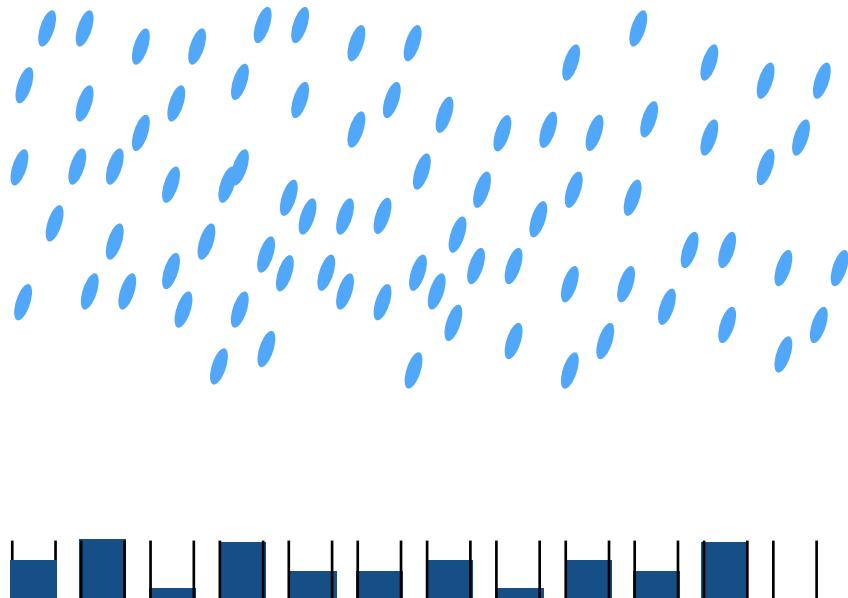
Discrete Count Data

- Shot Noise



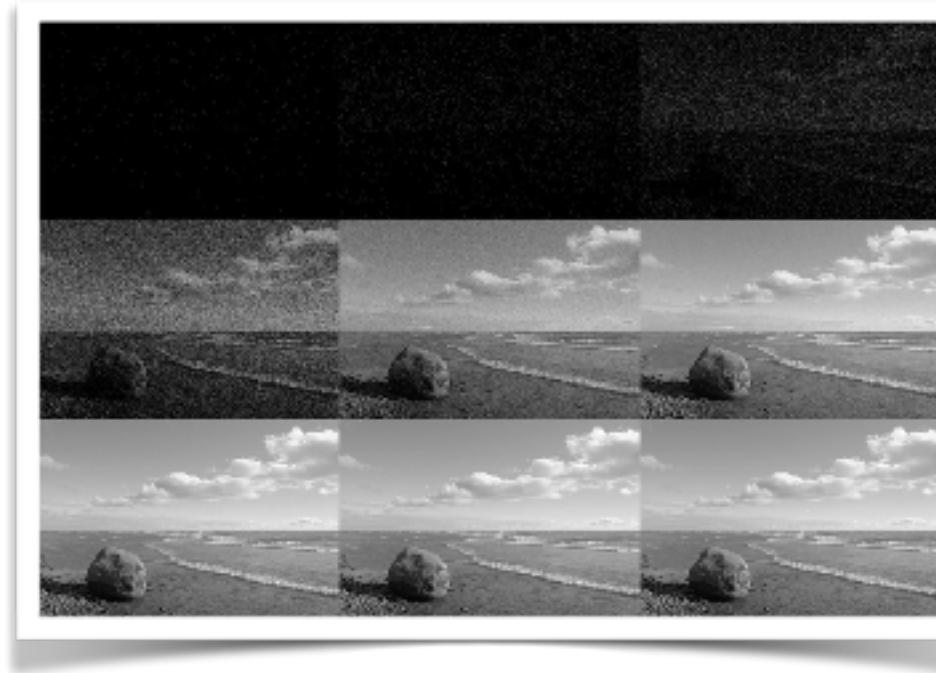
Discrete Count Data

- Shot Noise



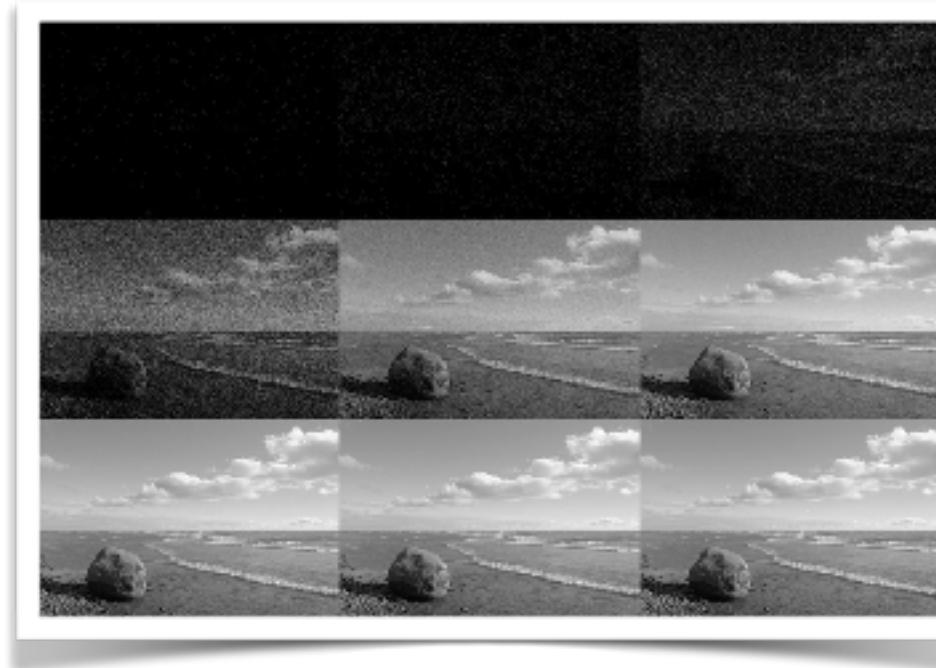
Discrete Count Data

- Shot Noise



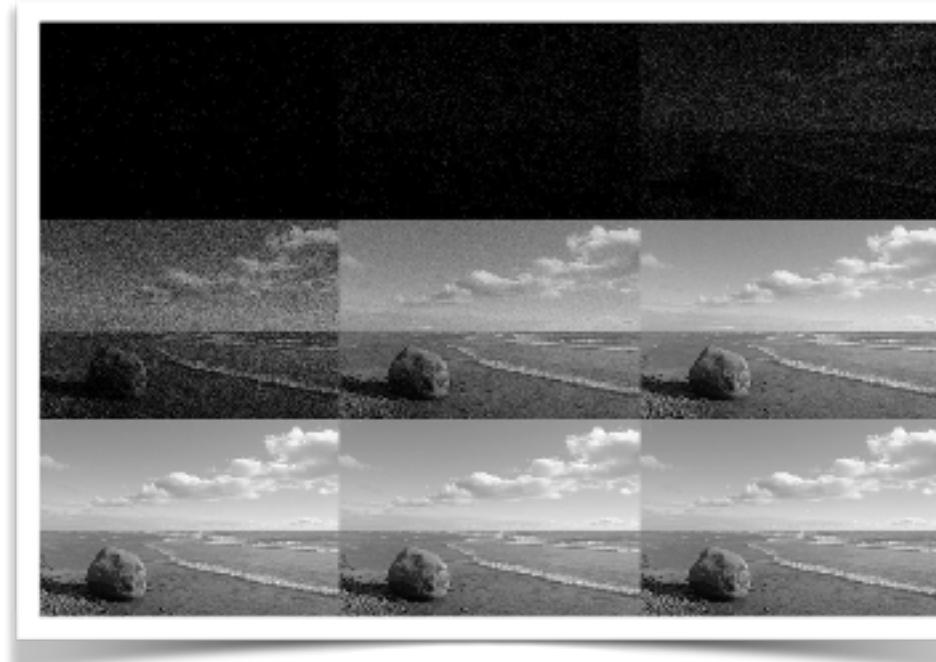
Discrete Count Data

- Shot Noise



Discrete Count Data

- Shot Noise

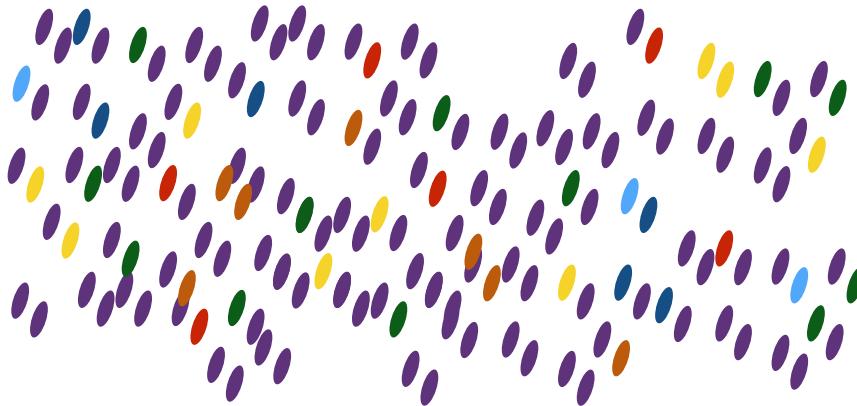




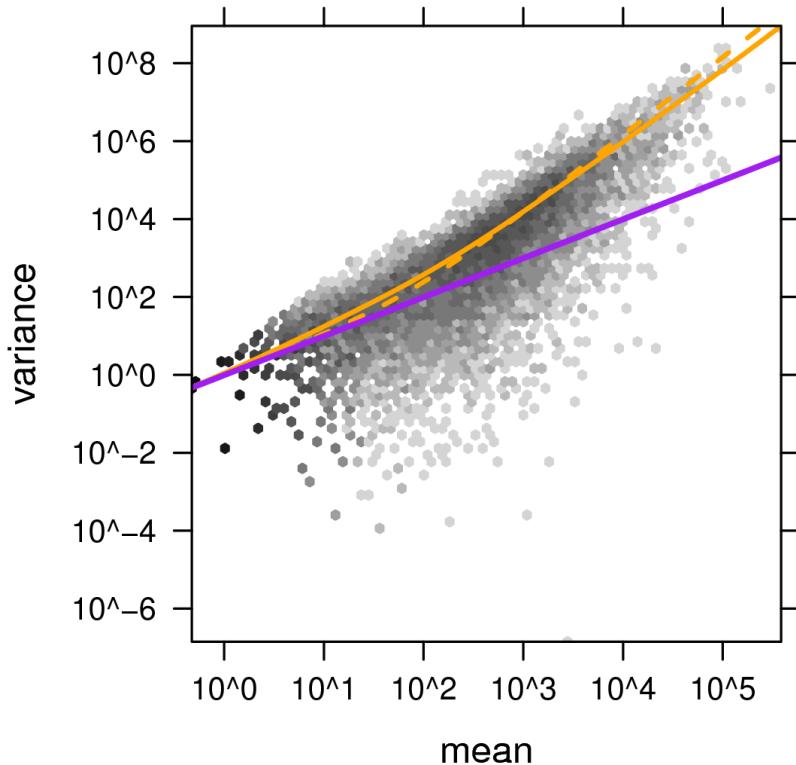
Cygnus Wall - North America Nebula - Cambridge 2014

It's Raining Genes... Hallelujah!

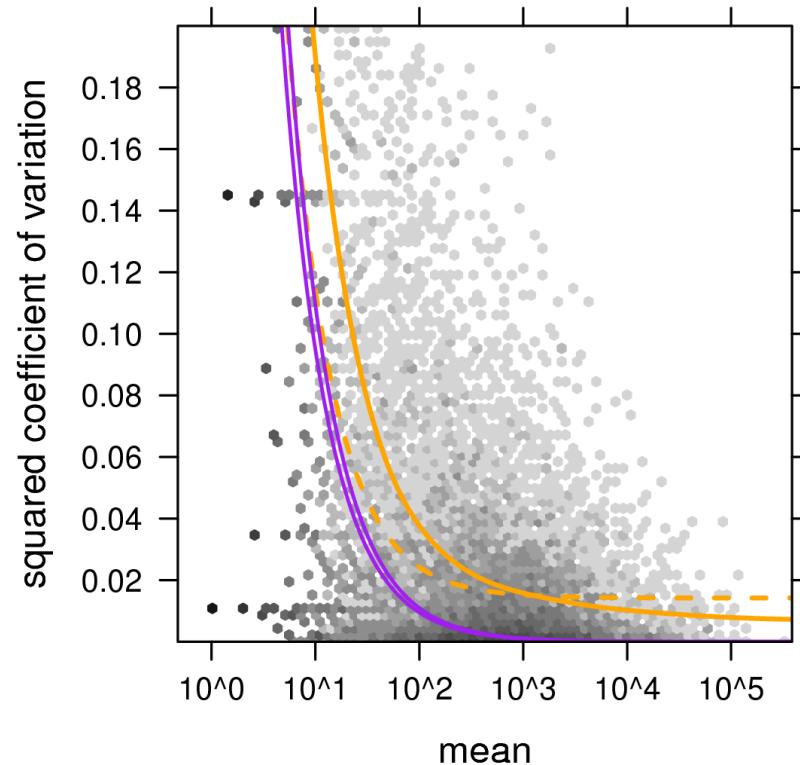
- Shot Noise is very different for individual genes



Sequencing Data - Variance - 2 replicates

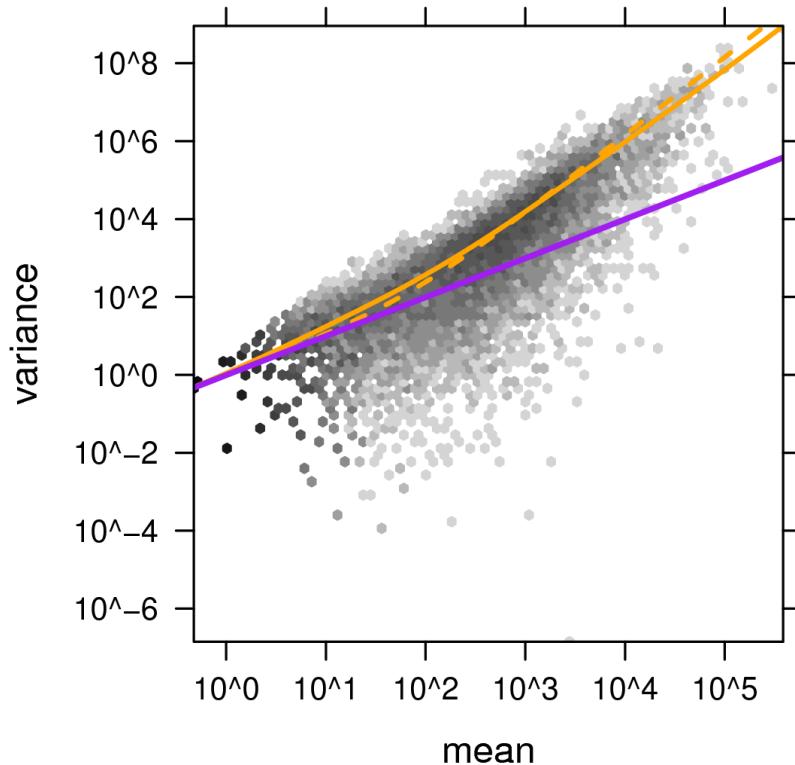


Poisson
Poisson + constant CV
Poisson + local regression

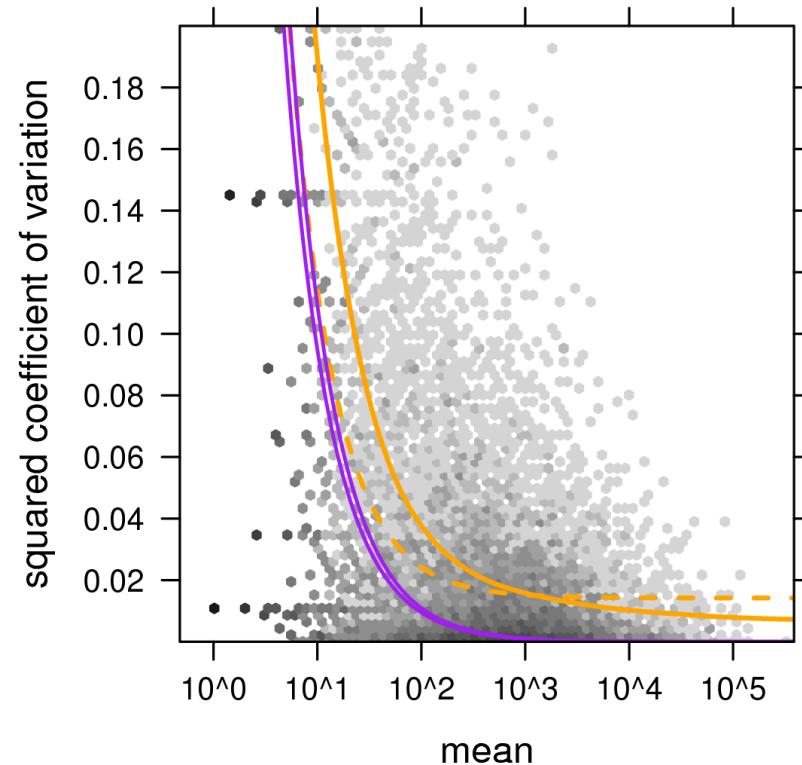


$v = \mu$ —————
 $v = \mu + \alpha \mu^2$ - - -
 $v = \mu + f(\mu^2)$ ————

Sequencing Data - Variance - 2 replicates



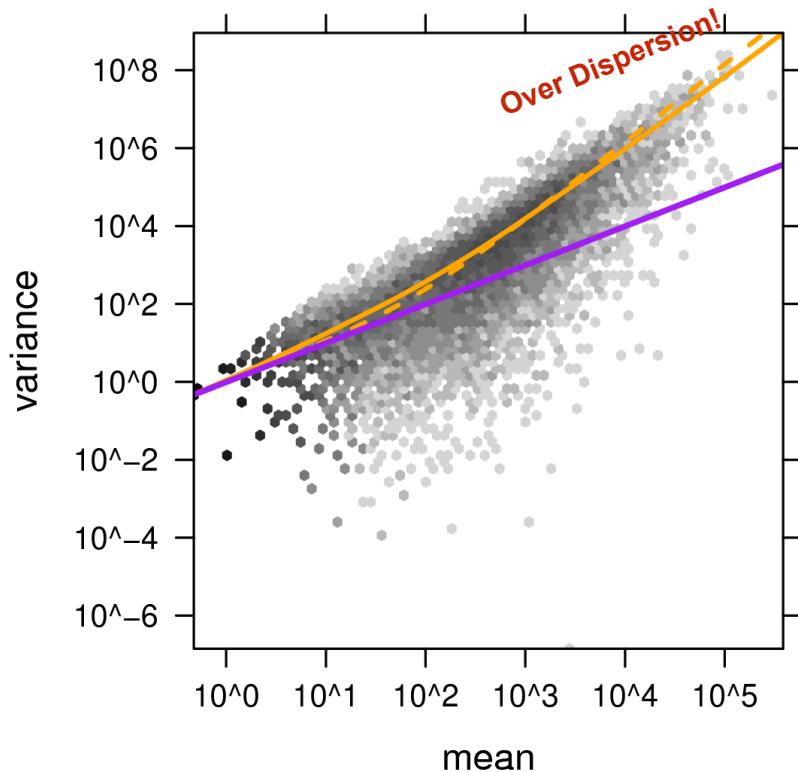
Poisson
Poisson + constant CV
Poisson + local regression



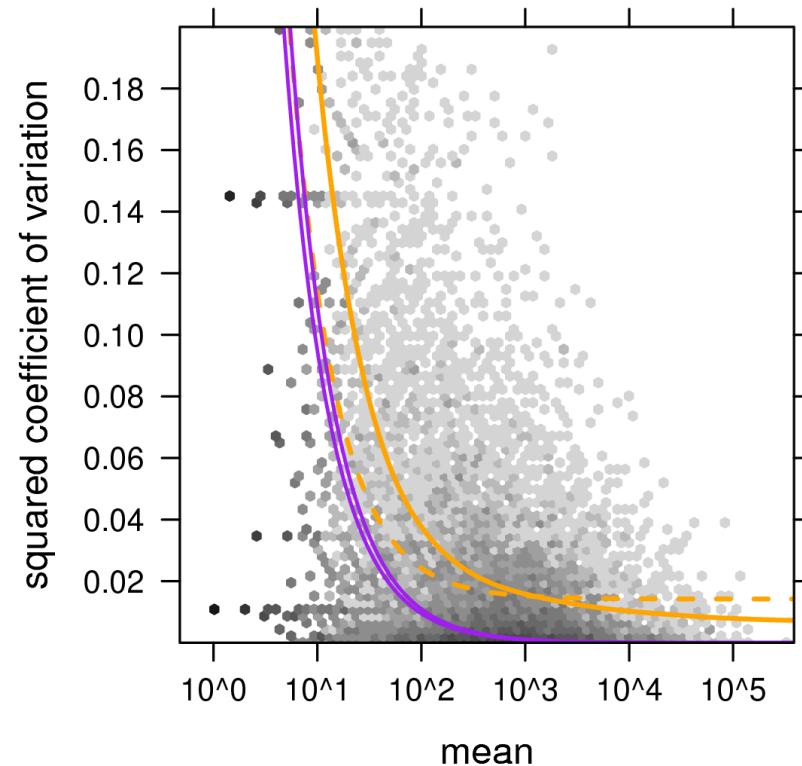
$v = \mu$ —————
 $v = \mu + \alpha \mu^2$ - - -
 $v = \mu + f(\mu^2)$ ————

Poisson Invalidated for genes with high expression as their variance exceeds the model
The variance should be approximately equal to the mean

Sequencing Data - Variance - 2 replicates



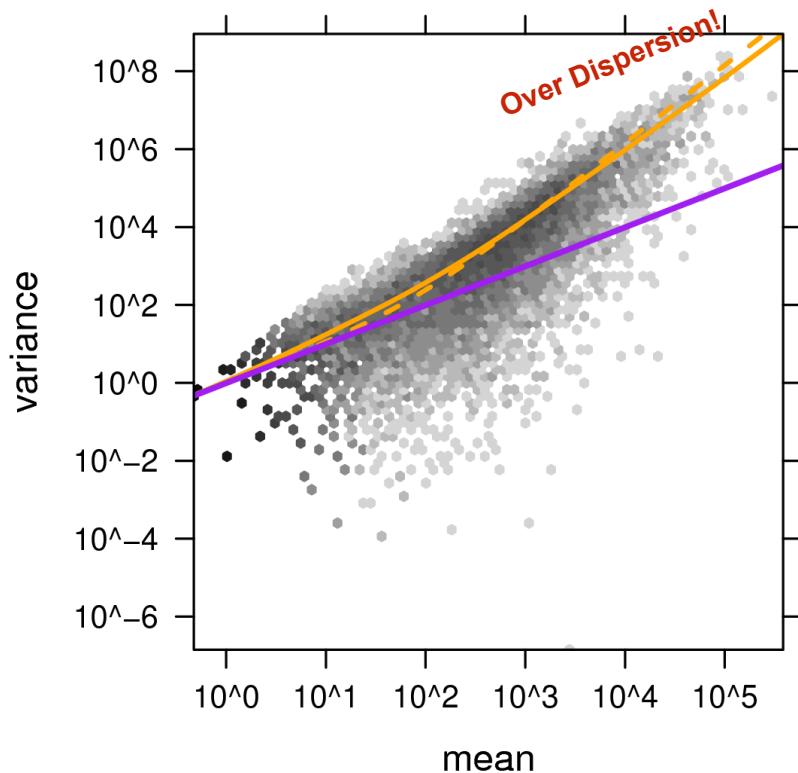
Poisson
Poisson + constant CV
Poisson + local regression



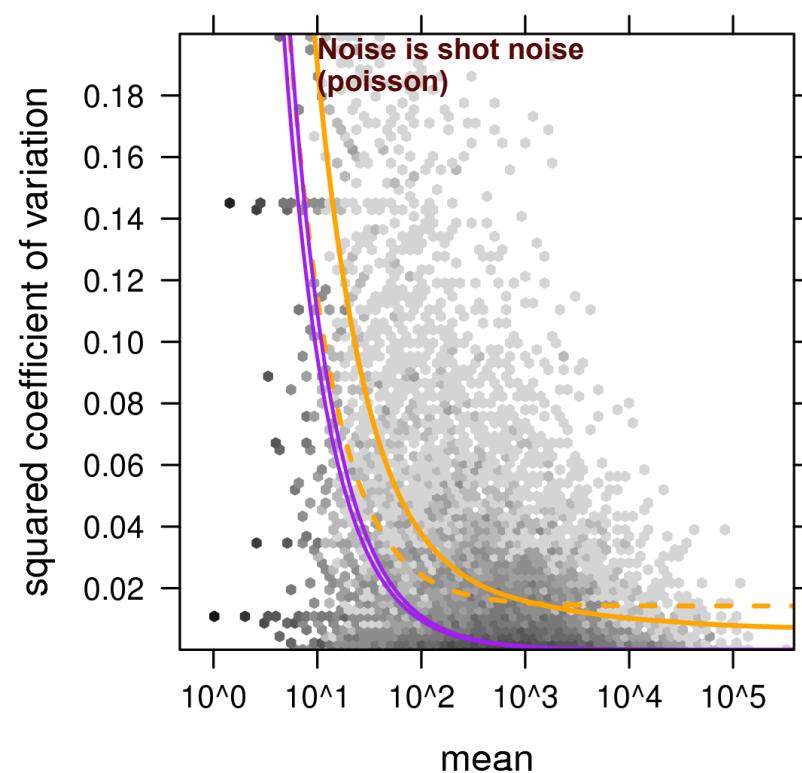
$\nu = \mu$
 $\nu = \mu + \alpha \mu^2$
 $\nu = \mu + f(\mu^2)$

Poisson Invalidated for genes with high expression as their variance exceeds the model
The variance should be approximately equal to the mean

Sequencing Data - Variance - 2 replicates



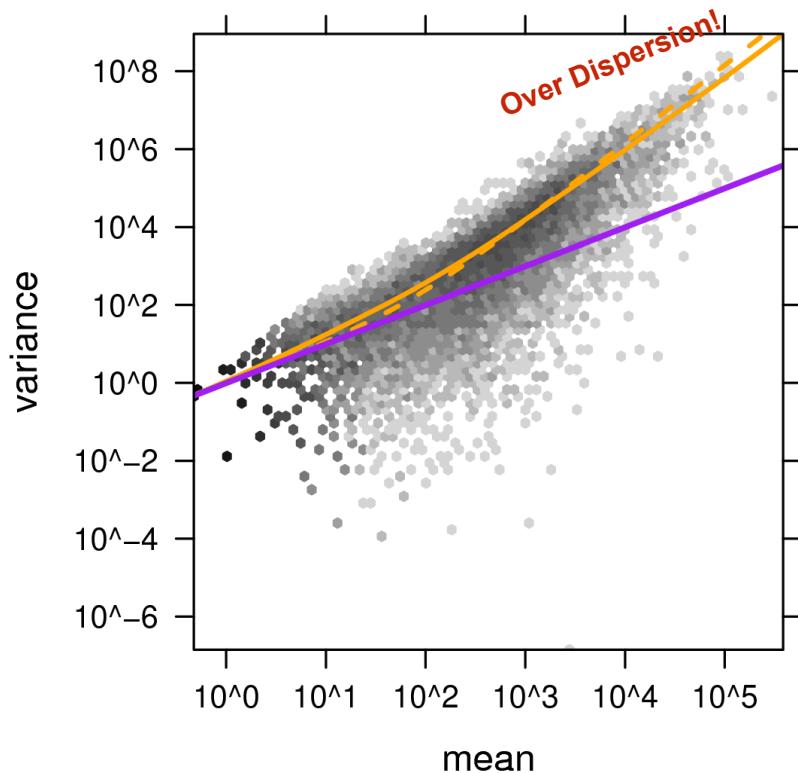
Poisson
Poisson + constant CV
Poisson + local regression



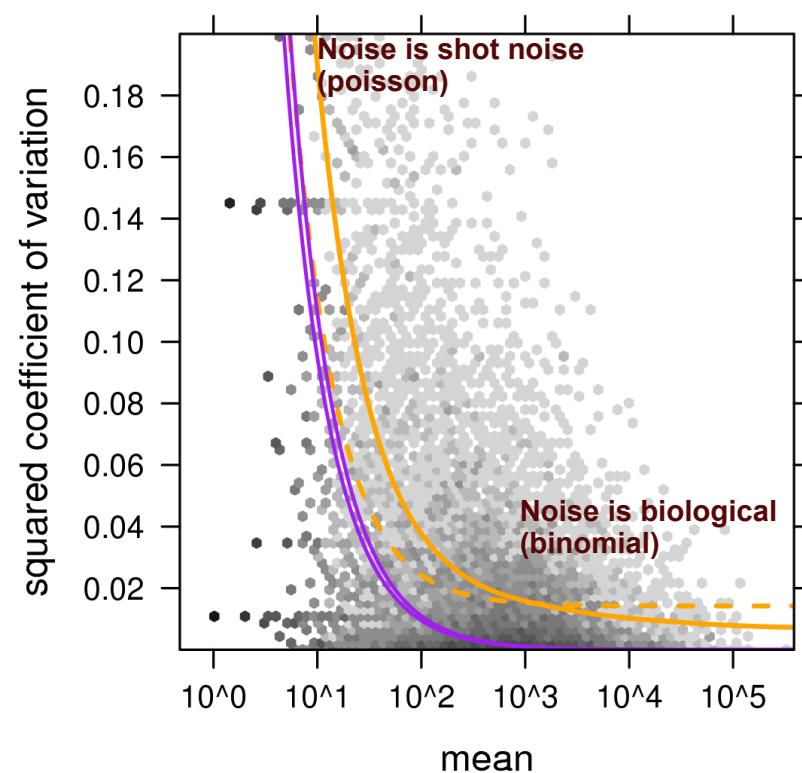
$v = \mu$ —————
 $v = \mu + \alpha \mu^2$ - - -
 $v = \mu + f(\mu^2)$ ————

Poisson Invalidated for genes with high expression as their variance exceeds the model
The variance should be approximately equal to the mean

Sequencing Data - Variance - 2 replicates



Poisson
Poisson + constant CV
Poisson + local regression



$v = \mu$ —————
 $v = \mu + \alpha \mu^2$ - - -
 $v = \mu + f(\mu^2)$ ————

Poisson Invalidated for genes with high expression as their variance exceeds the model
The variance should be approximately equal to the mean

Sequencing Data - Summary

- **Shot Noise**

- unavoidable, appears even with perfect replication
- dominant noise for weakly expressed genes

- **Technical Noise**

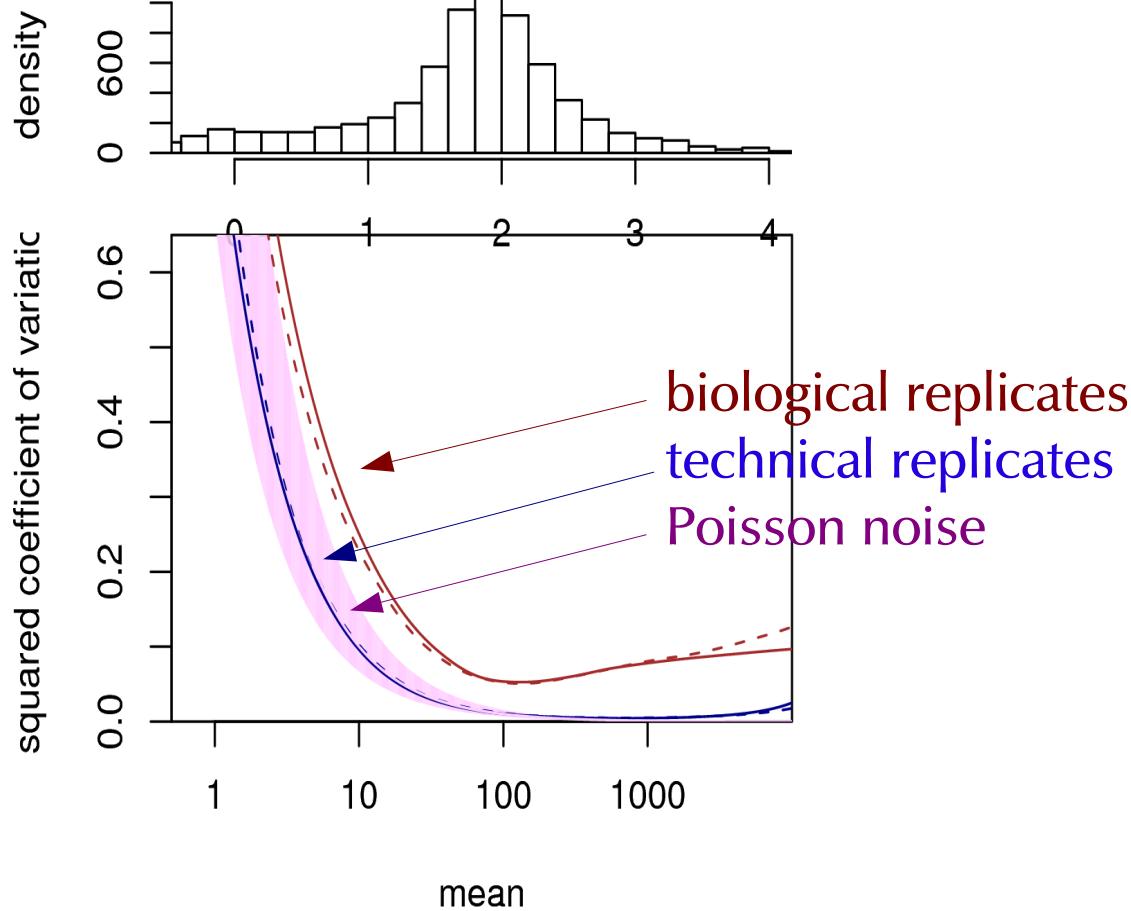
- from sample preparation and sequencing
- negligible (if all goes well)

- **Biological Noise**

- unaccounted-for differences between samples
- Dominant noise for strongly expressed genes

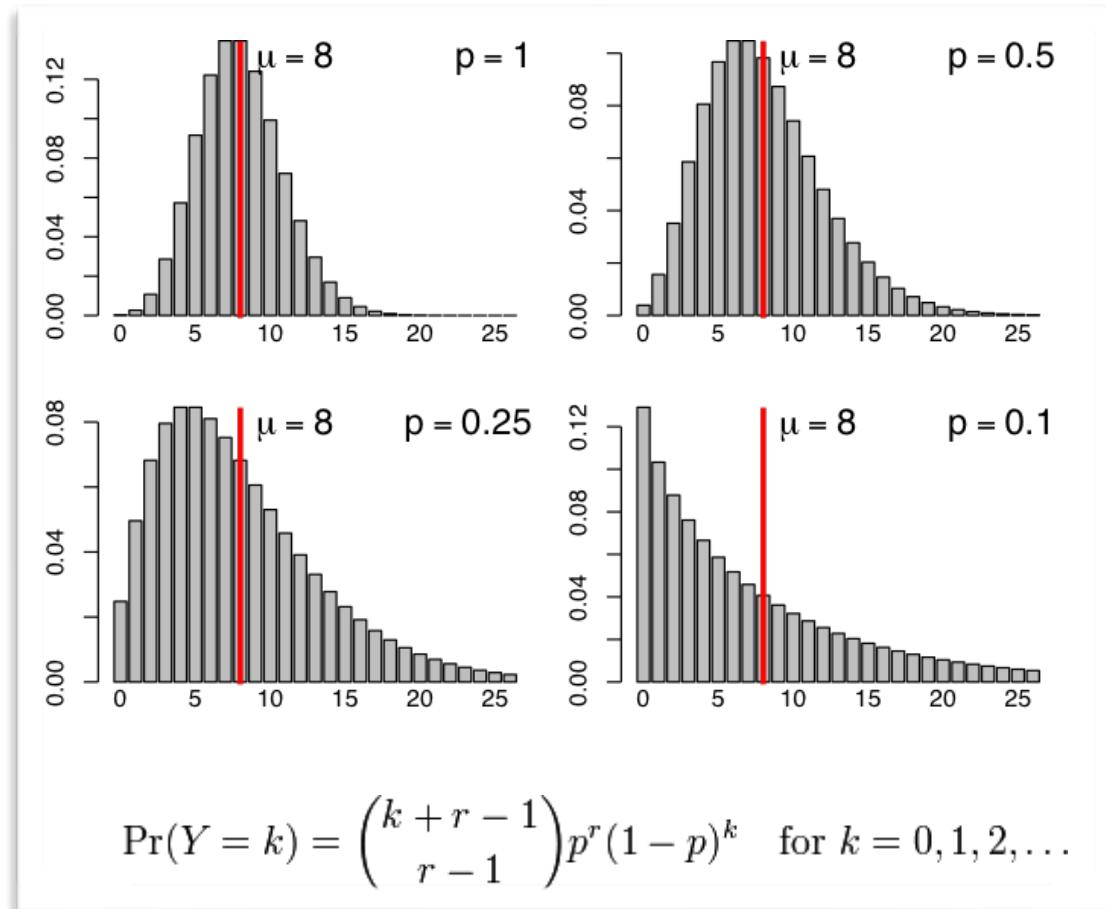
Sequencing Data - Summary

- **Shot Noise**
 - unavoidable
 - dominant at low mean
- **Technical noise**
 - from sample preparation
 - negligible at high mean
- **Biological noise**
 - unaccounted for
 - Dominant at high mean



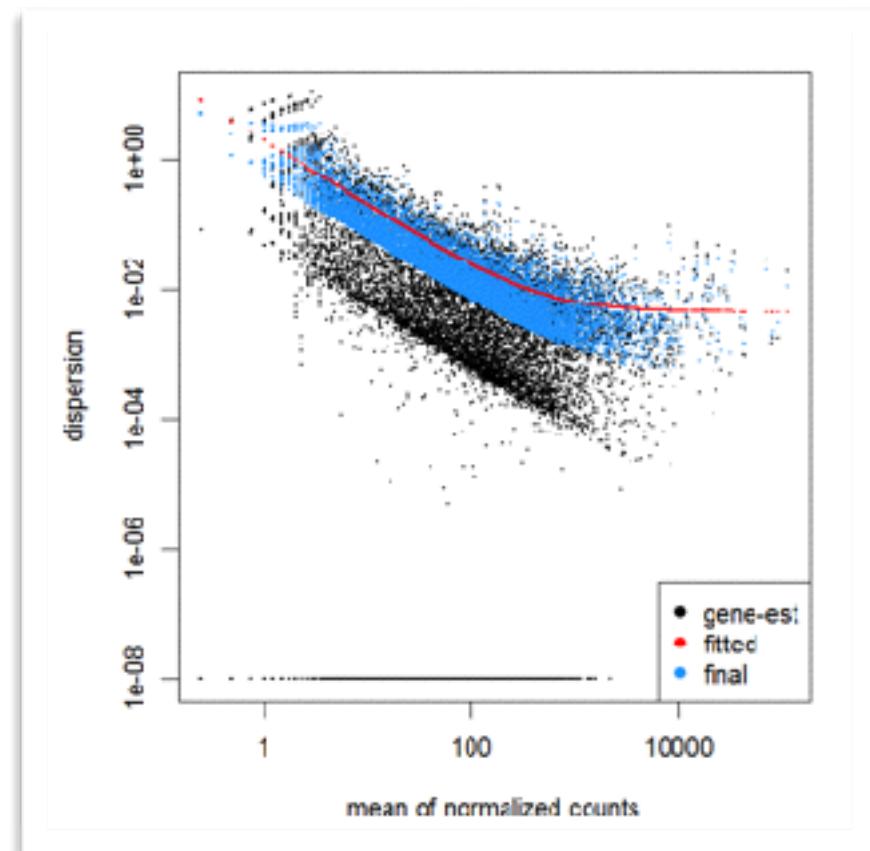
How to resolve this issue ?

- The negative binomial distribution

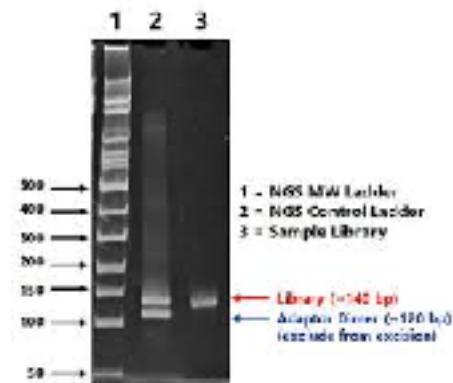
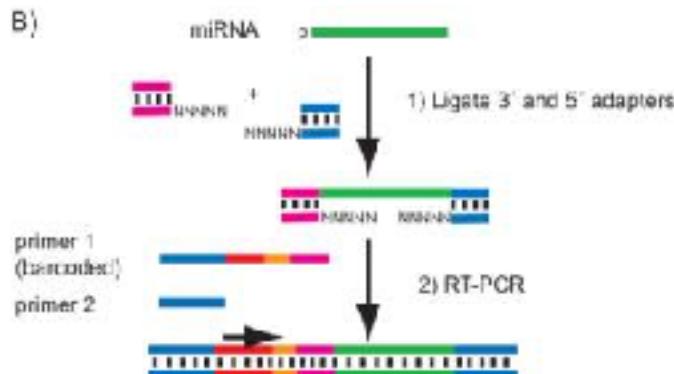
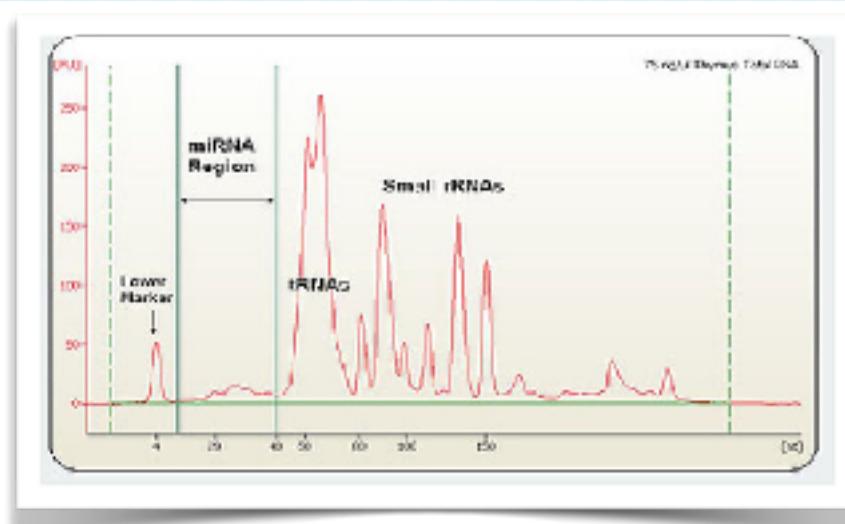
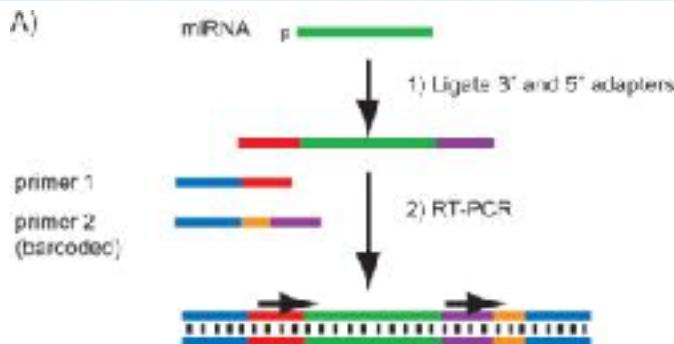


Negative Binomial Approaches

- DESeq2 - Simon Anders & Wolfgang Huber
- EdgeR - Robinson & Smyth

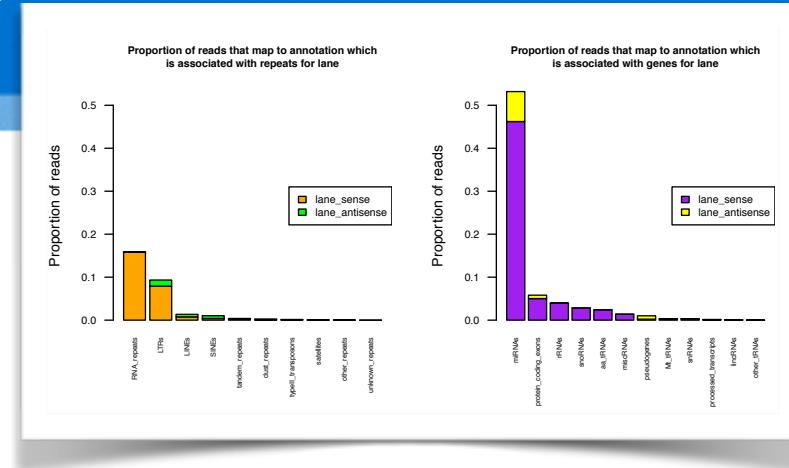
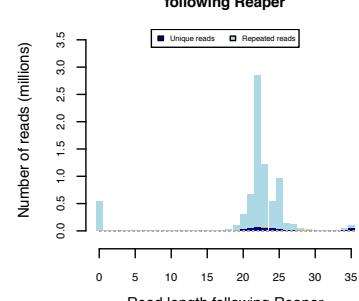
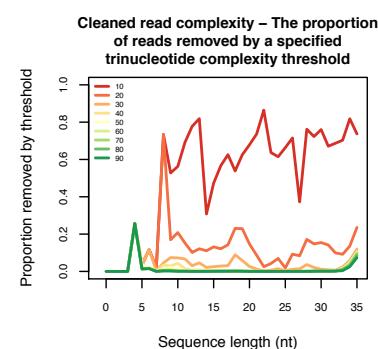
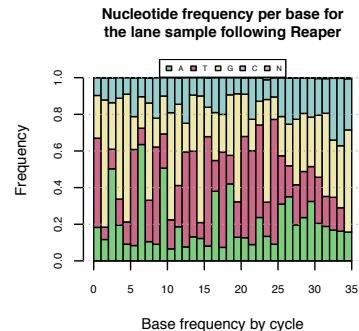
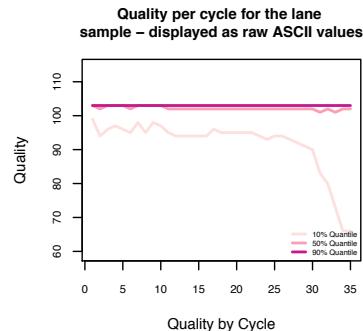
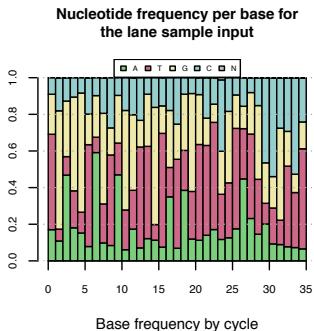


smallRNA Seq



Tools for small RNAseq Analysis

- Galaxy
- FASTX toolkit
- UEA Toolkit
- R & BioConductor
- Reaper, Tally and Kraken



Downstream Analysis

Workflow: Raw miRNA NGS to Results

```
@HS24_10147:1:1101:1067:1989#0
GCCCGGCTAGCTCAGTCGGTAGAGCATGGACTGGAATTCTCGGGTGCN
+
BCCFFFFFHHHFHHJGHIIJIJJJIJH?FHHHEHJIGIJJJJJHJJ!
@HS24_10147:1:1101:2218:1975#0
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
BBCFFFFDHHHHHJJJJJJJJFJIIJJJJJJJJ@HIJIIJIIHII!
@HS24_10147:1:1101:2722:1968#0
CGACCTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCGATGTATCTCN
+
?@@FFFFFAHDAEHFHGIIGGIAE9DFHIGFFFGEIGFD=CCCCF!
@HS24_10147:1:1101:2977:1942#0
NCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
!1BDFFDDHHHHIJJIIJJJJJJJJJJJJJJJJGHIJJJJJJJJ!
@HS24_10147:1:1101:5876:1978#0
TACAGTCCGACGATCTGGAATTCTCGGGTGCCAAGGCTCCAGTCACCGAN
+
=:==DDDDDD<DFHIHGGHIIHGEFHGBFBDECHI<?@FGDFGCGDE!
@HS24_10147:1:1101:8742:1944#0
NCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
```

Workflow: Raw miRNA NGS to Results

```
@HS24_10147:1:1101:1067:1989#0
GCCCGGCTAGCTCAGTCGGTAGAGCATGGACTGGAATTCTCGGGTGCN
+
BCCFFFFFHHHFHHJGHIIJIJJJIJH?FHHHEHJIGIJJJJJHJJ!
@HS24_10147:1:1101:2218:1975#0
TCGCTTGGTGCAGATCGGGACTTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
BBCFFFFDHHHHHJJJJJJJJFJIIJJJJJJJJ@HIJIIHII!
@HS24_10147:1:1101:2722:1968#0
CGACCTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCGATGTATCTCN
+
?@@FFFFFAHDAEHFHGIIGGIAE9DFHIGFFHGEGIGFD=CCCCF!
@HS24_10147:1:1101:2977:1942#0
NCGCTTGGTGCAGATCGGGACTTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
!1BDFFDDHHHHIJJIIJJJJJJJJJJJJJJGHIJJJJJJJJ!
@HS24_10147:1:1101:5876:1978#0
TACAGTCCGACGATCTGGAATTCTCGGGTGCCAAGGCTCCAGTCACCGAN
+
=:D-DDDDDD<DFHIHGGHIIHGEFHGBFBDECHI<?@FGDFGCGDE!
@HS24_10147:1:1101:8742:1944#0
NCGCTTGGTGCAGATCGGGACTTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
```

Workflow: Raw miRNA NGS to Results

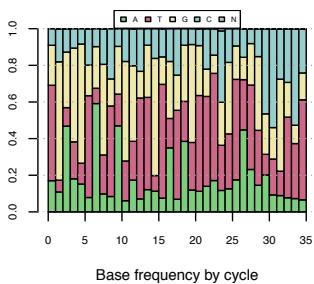
```
@HS24_10147:1:1101:1067:1989#0
```

```
GCCCCGGCTA  
+>1_t11_w33_x157998  
GTTTCGGTAGTGTAGTGGTTATCACGTTGCCT  
@HS24_101  
TCGCTTGGI  
+>2_t11_w32_x80579  
GCATTGGTGGTCAGTGGTAGAATTCTCGCCT  
BBCFFFFDHE  
@HS24_101  
CGACCTGGA  
+>3_t10_w36_x27074  
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATT  
@HS24_101  
?@@FFFFFE  
@HS24_101  
NCGCTTGGI  
+>4_t10_w35_x24313  
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAAT  
+>5_t11_w34_x22696  
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAA  
@HS24_101  
TTCACAGTGGCTAACGTTCTG  
+>6_t8_w20_x20049  
!1BDFFDDH  
@HS24_101  
TACAGTCCG  
+>7_t11_w33_x17558  
GCATTGGTGGTTCAAGTGGTAGAATTCTCGCCTG  
@HS24_101  
=:=DDDDDD  
@HS24_101  
NCGCTTGGI  
+>8_t8_w32_x16093  
GCATTGGTGGTTCAAGTGGTAGAATTCTCGCCT  
+>9_t12_w31_x15490  
GCATTGGTGGTTCAAGTGGTAGAATTCTCGGCC  
@HS24_101  
>10_t11_w33_x14783  
TCCCTGGTGGTCTAGTGGTTAGGATTGGCGCT  
+>11_t8_w31_x14593  
GCATTGGTGGTTCAAGTGGTAGAATTCTCGCCT
```

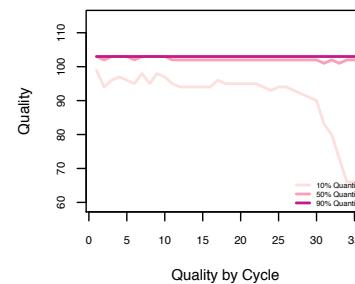
Workflow: Raw miRNA NGS to Results

HS24_101_1_101_1067_100040

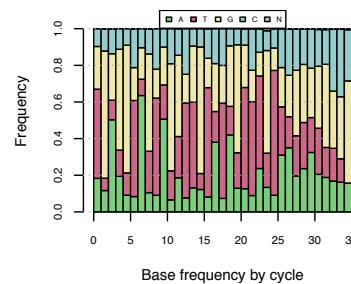
Nucleotide frequency per base for the lane sample input



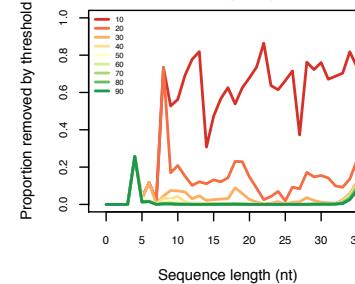
Quality per cycle for the lane sample – displayed as raw ASCII values



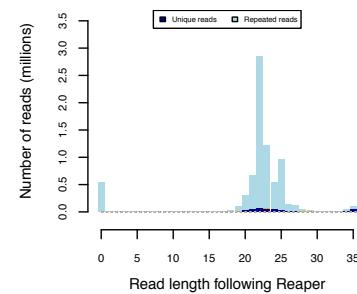
Nucleotide frequency per base for the lane sample following Reaper



Cleaned read complexity – The proportion of reads removed by a specified trinucleotide complexity threshold



Read length plot for lane sample following Reaper



```

NCGCTTGGT    TTACAGTGCTAAGTTCTG
+             >7_t11_w33_x17558
!1BDFFDDH    GCATTGGTGGTTCACTGGTAGAATTCTGCCTG
@HS24_101     >8_t8_w32_x16093
TACAGTCCG    GCATTTGTGGTTCACTGGTAGAATTCTGCCT
+             >9_t12_w31_x15490
=:DDDDDDD    GCATTGGTGGTTCACTGGTAGAATTCTGCC
@HS24_101     >10_t11_w33_x14783
NCGCTTGGT    TCCCTGGTGGTCTAGTGGTAGGATTGGCGCT
+             >11_t8_w31_x14593
                GCATTGTGGTTCACTGGTAGAATTCTGCCT
  
```

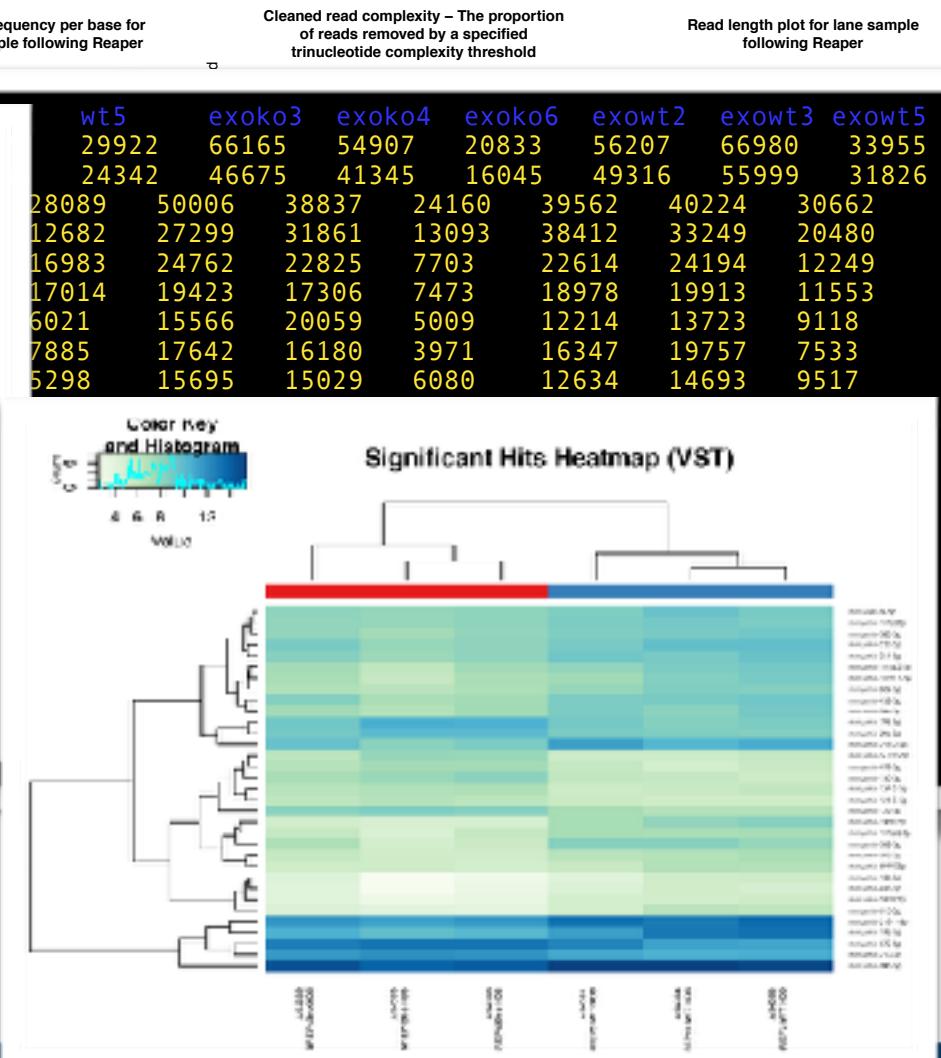
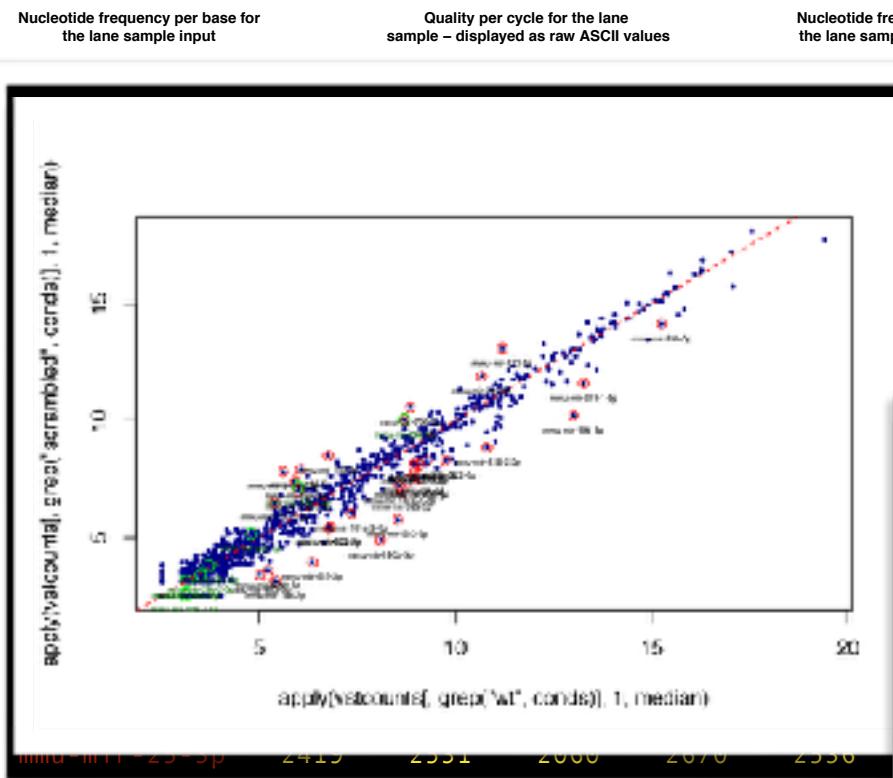
Workflow: Raw miRNA NGS to Results

04/24_10147_1_1101_1067_1000#0

Nucleotide frequency per base for the lane sample input	Quality per cycle for the lane sample – displayed as raw ASCII values						Nucleotide frequency per base for the lane sample following Reaper			Cleaned read complexity – The proportion of reads removed by a specified trinucleotide complexity threshold			Read length plot for lane sample following Reaper		
	ko3	ko4	ko6	wt1	wt2	wt5	exoko3	exoko4	exoko6	exowt2	exowt3	exowt5			
mmu-mir-21a-5p	33561	26530	38601	34692	32023	29922	66165	54907	20833	56207	66980	33955			
mmu-mir-92a-1-3p	28479	29132	26069	37165	36005	24342	46675	41345	16045	49316	55999	31826			
mmu-mir-27b-3p	29042	23071	47952	23737	25318	28089	50006	38837	24160	39562	40224	30662			
mmu-mir-191-5p	12427	11341	17954	13162	14266	12682	27299	31861	13093	38412	33249	20480			
mmu-mir-142-5p	25394	25189	23562	28321	25120	16983	24762	22825	7703	22614	24194	12249			
mmu-mir-182-5p	11649	10632	22673	10975	13652	17014	19423	17306	7473	18978	19913	11553			
mmu-mir-16-2-5p	9108	10131	7848	8892	8200	6021	15566	20059	5009	12214	13723	9118			
mmu-let-7f-2-5p	8067	10065	9573	9193	9740	7885	17642	16180	3971	16347	19757	7533			
mmu-mir-92a-2-3p	7579	8116	7360	8248	7350	5298	15695	15029	6080	12634	14693	9517			
mmu-let-7i-5p	6779	6875	6660	6448	6025	5110	13050	12705	4014	9939	11520	5813			
mmu-mir-22-3p	4009	3224	10072	3418	3199	6876	12495	10807	5810	8814	10434	9580			
mmu-mir-342-3p	627	607	767	670	591	581	6404	4218	2507	11565	9414	5303			
mmu-mir-26a-2-5p	5656	6673	4293	5680	5627	3160	9024	10408	1889	8836	9367	3577			
mmu-mir-5109-3p	509	533	989	641	647	750	3964	3432	1583	3604	3459	8811			
mmu-let-7a-2-5p	2374	3183	2259	2434	2849	1831	7530	7570	1765	7058	8507	3296			
mmu-mir-30e-5p	4385	4553	3438	4604	4269	2656	7653	7930	2277	6561	6912	4121			
mmu-mir-150-5p	4437	6858	1210	5429	4831	986	6483	6694	768	5442	6969	1413			
mmu-mir-5099-3p	4162	5294	6956	6017	4928	5860	5766	4972	2184	4623	5696	2644			
mmu-mir-186-5p	2670	2818	2888	2601	2505	1950	6329	6524	1965	5874	5847	3680			
mmu-mir-148a-3p	3214	3768	5707	3413	4515	4453	5320	6501	3769	5812	5436	5619			
mmu-mir-181a-2-5p	1178	1709	2342	1401	1971	1785	5358	6478	1676	5687	5467	2561			
mmu-mir-25-3p	2419	2531	2060	2670	2536	1630	6147	6279	2676	5828	6328	4631			

Workflow: Raw miRNA NGS to Results

0HS24_10147.1;1101;1067;1989#0

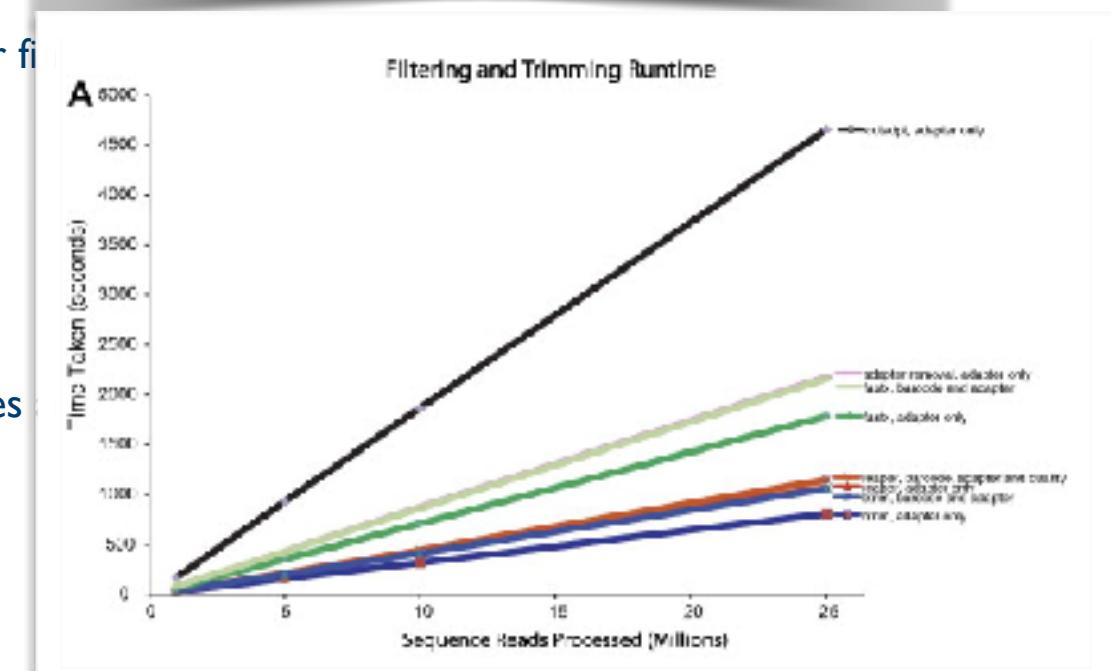
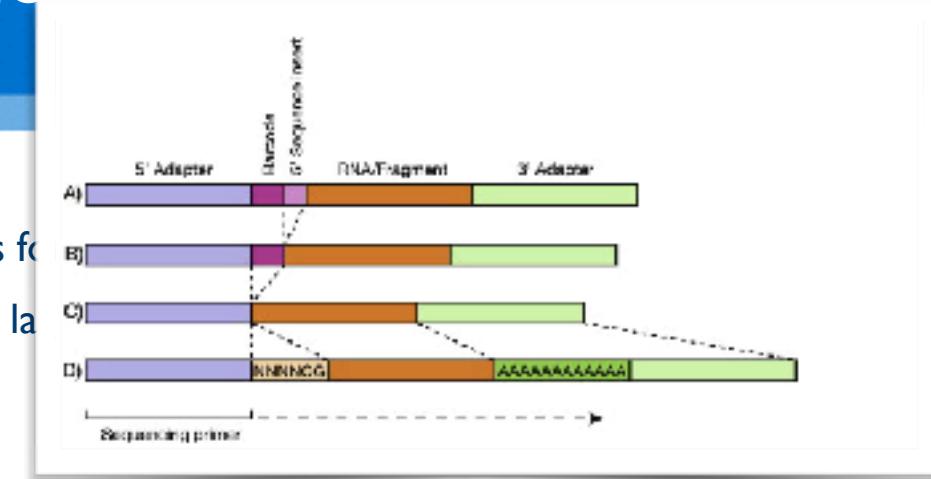


The Kraken Suite of Tools for smallRNA Analysis

- Minion
 - De bruijn graph 3' assembly analysis for small RNA reads
 - Can identify adapter sequences and large-scale contaminants
- Reaper
 - Extremely fast and accurate adapter finder and trimmer
 - Also cleans
 - Low-complexity tracts
 - PolyA, PolyN
 - Low-scoring tracts
 - Deals with complex read geometries and random seqs
- Tally

The Kraken Suite of Tools for smallRNA Analysis

- Minion
 - De bruijn graph 3' assembly analysis for smallRNAs
 - Can identify adapter sequences and label them
- Reaper
 - Extremely fast and accurate adapter filtering
 - Also cleans
 - Low-complexity tracts
 - PolyA, PolyN
 - Low-scoring tracts
 - Deals with complex read geometries
- Tally



The Kraken Suite of Tools for smallRNA Analysis

- Minion
 - De bruijn graph 3' assembly analysis for small RNA reads
 - Can identify adapter sequences and large-scale contaminants
- Reaper
 - Extremely fast and accurate adapter finder and trimmer
 - Also cleans
 - Low-complexity tracts
 - PolyA, PolyN
 - Low-scoring tracts
 - Deals with complex read geometries and random seqs
- Tally

Turning Processed Reads into Counts on microRNAs

```
@HS24_10147:1:1101:1067:1989#0
GCCCGGCTAACCTGACTGGCTACACCAATCCGACTCCATTCTGGCTCCCN
+
BCCFFI >1_t11_w33_x157998
@HS24 GTTTCCGTAGTGTAGTGTTATCACGTTGCCT
TCGCTI >2_t11_w32_x80579
+ GCATTGGTGGTTCAGTGGTAGAATTCTGCCT
BBCFFI >3_t10_w36_x27074
```

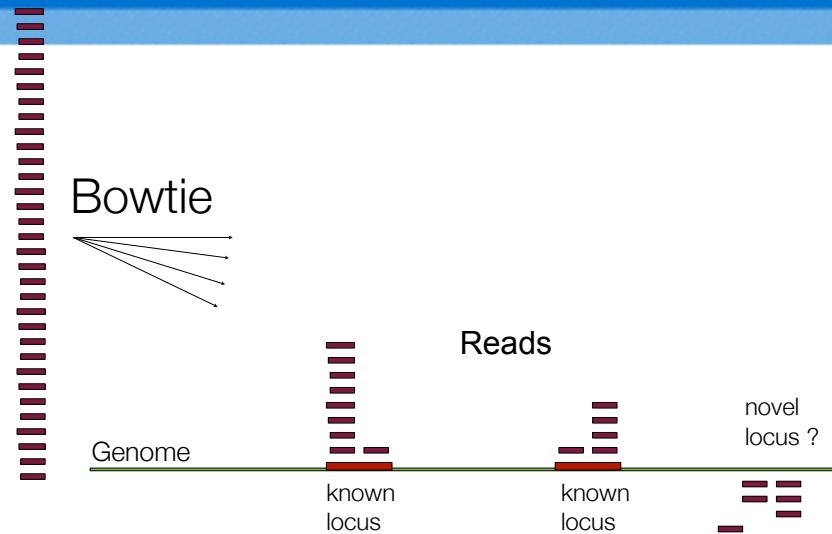
	ko3	ko4	ko6	wt1	wt2	wt5	exoko3	exoko4	exoko6	exowt2	exowt3	exowt5
mmu-mir-21a-5p	33561	26530	38601	34692	32023	29922	66165	54907	20833	56207	66980	33955
mmu-mir-92a-1-3p	28479	29132	26069	37165	36005	24342	46675	41345	16045	49316	55999	31826
mmu-mir-27b-3p	29042	23071	47952	23737	25318	28089	50006	38837	24160	39562	40224	30662
mmu-mir-191-5p	12427	11341	17954	13162	14266	12682	27299	31861	13093	38412	33249	20480
mmu-mir-142-5p	25394	25189	23562	28321	25120	16983	24762	22825	7703	22614	24194	12249
mmu-mir-182-5p	11649	10632	22673	10975	13652	17014	19423	17306	7473	18978	19913	11553
mmu-mir-16-2-5p	9108	10131	7848	8892	8200	6021	15566	20059	5009	12214	13723	9118
mmu-let-7f-2-5p	8067	10065	9573	9193	9740	7885	17642	16180	3971	16347	19757	7533
mmu-mir-92a-2-3p	7579	8116	7360	8248	7350	5298	15695	15029	6080	12634	14693	9517
mmu-let-7i-5p	6779	6875	6660	6448	6025	5110	13050	12705	4014	9939	11520	5813
mmu-mir-22-3p	4009	3224	10072	3418	3199	6876	12495	10807	5810	8814	10434	9580
mmu-mir-342-3p	627	607	767	670	591	581	6404	4218	2507	11565	9414	5303
mmu-mir-26a-2-5p	5656	6673	4293	5680	5627	3160	9024	10408	1889	8836	9367	3577
mmu-mir-5109-3p	509	533	989	641	647	750	3964	3432	1583	3604	3459	8811
mmu-let-7a-2-5p	2374	3183	2259	2434	2849	1831	7530	7570	1765	7058	8507	3296
mmu-mir-30e-5p	4385	4553	3438	4604	4269	2656	7653	7930	2277	6561	6912	4121
mmu-mir-150-5p	4437	6858	1210	5429	4831	986	6483	6694	768	5442	6969	1413
mmu-mir-5099-3p	4162	5294	6956	6017	4928	5860	5766	4972	2184	4623	5696	2644
mmu-mir-186-5p	2670	2818	2888	2601	2505	1950	6329	6524	1965	5874	5847	3680
mmu-mir-148a-3p	3214	3768	5707	3413	4515	4453	5320	6501	3769	5812	5436	5619
mmu-mir-181a-2-5p	1178	1709	2342	1401	1971	1785	5358	6478	1676	5687	5467	2561
mmu-mir-25-3p	2419	2531	2060	2670	2536	1630	6147	6279	2676	5828	6328	4631

Mapping small RNA Reads

- Genome Based Approach
- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours

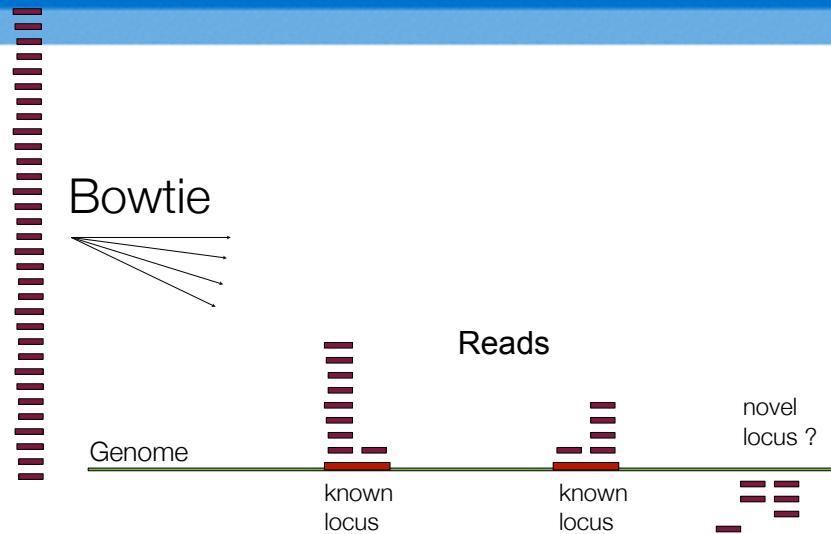
Mapping small RNA Reads

- **Genome Based Approach**
- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours



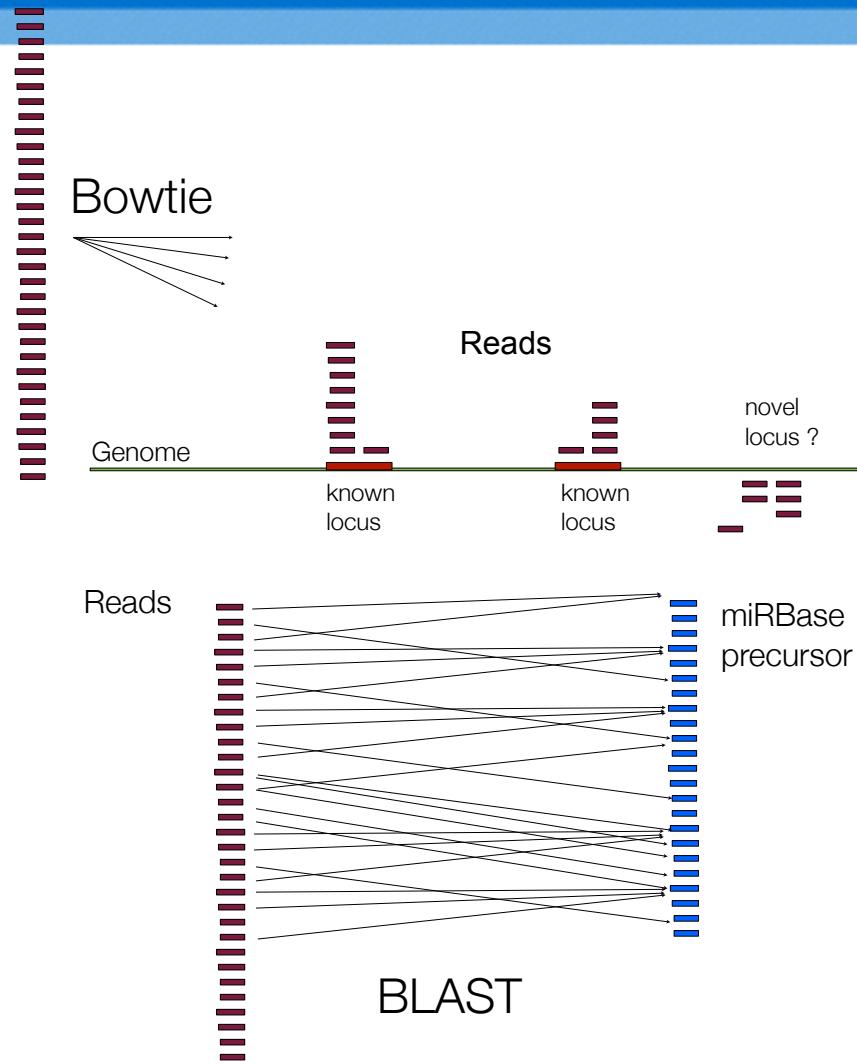
Mapping small RNA Reads

- **Genome Based Approach**
- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours
- **Read versus Precursor Approach**
 - Compare reads directly against miRBase sequences
 - Fast and accurate
 - look for >95% identity and no more than 1-2 mismatches
 - Runtime: 10-15 minutes



Mapping small RNA Reads

- **Genome Based Approach**
- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours
- **Read versus Precursor Approach**
- Compare reads directly against miRBase sequences
 - Fast and accurate
 - look for >95% identity and no more than 1-2 mismatches
 - Runtime: 10-15 minutes



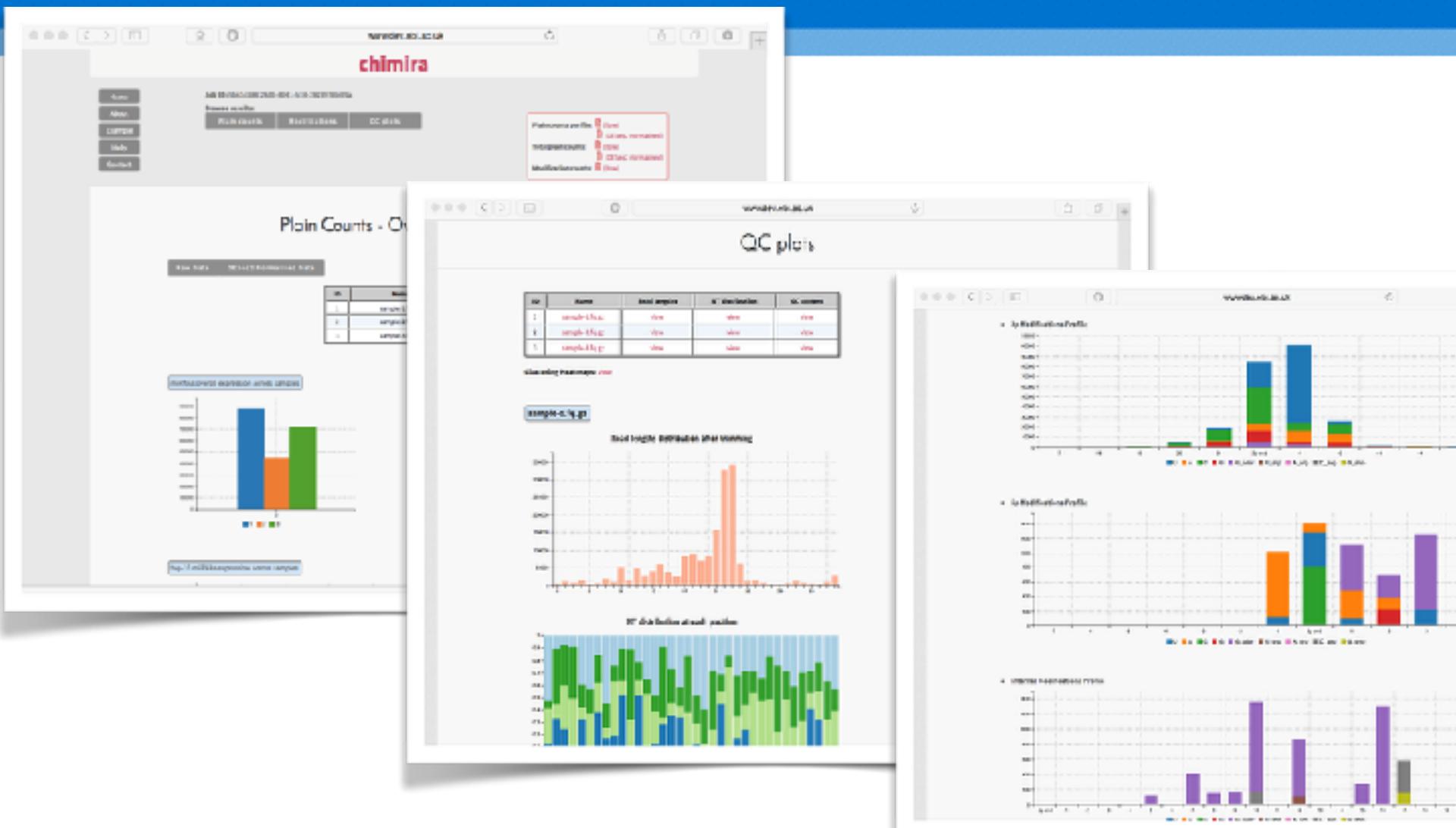
Bringing this together: ChimiRa

A simple web-based *microRNA NGS analysis toolkit*

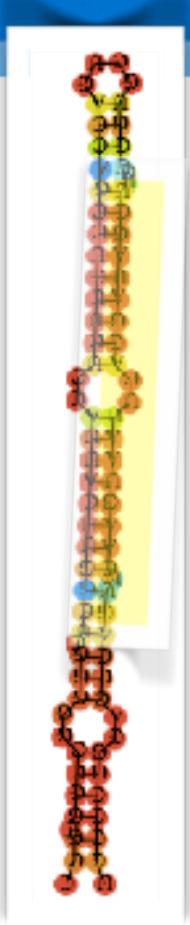
- Automatically preprocesses data - adapter removal, deduplication
- Reads compared directly against known microRNA precursors
 - Multiple Species available
 - Up to 2 mismatches allowed (configurable)
 - Reads assigned to best match
 - Ambiguous reads assigned randomly
- 5p or 3p side of hairpin called automatically
- Possibility to detect 5p / 3p modifications and RNA editing
 - tutase
 - ADAR

ChimiRa Interface

wwwdev.ebi.ac.uk/enright-srv/chimira



MicroRNA Read Level Analysis



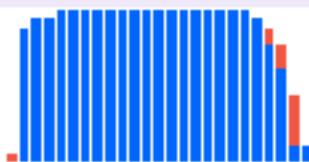
mmu-mir-26b-5p

Depth: 5741 reads

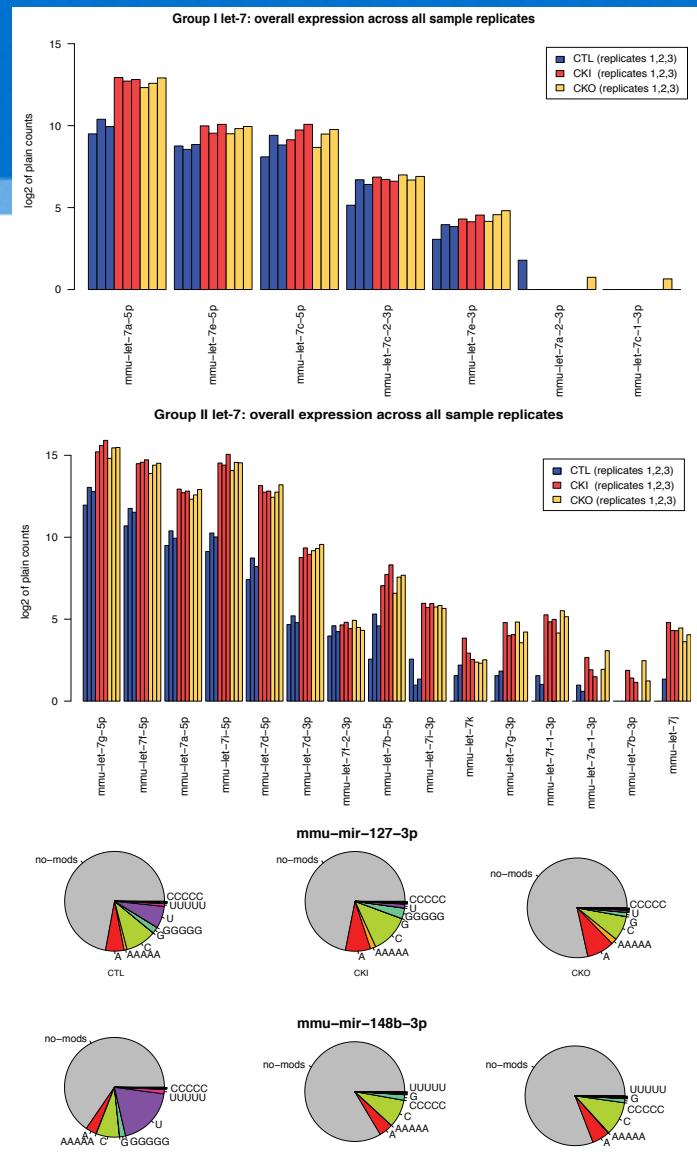
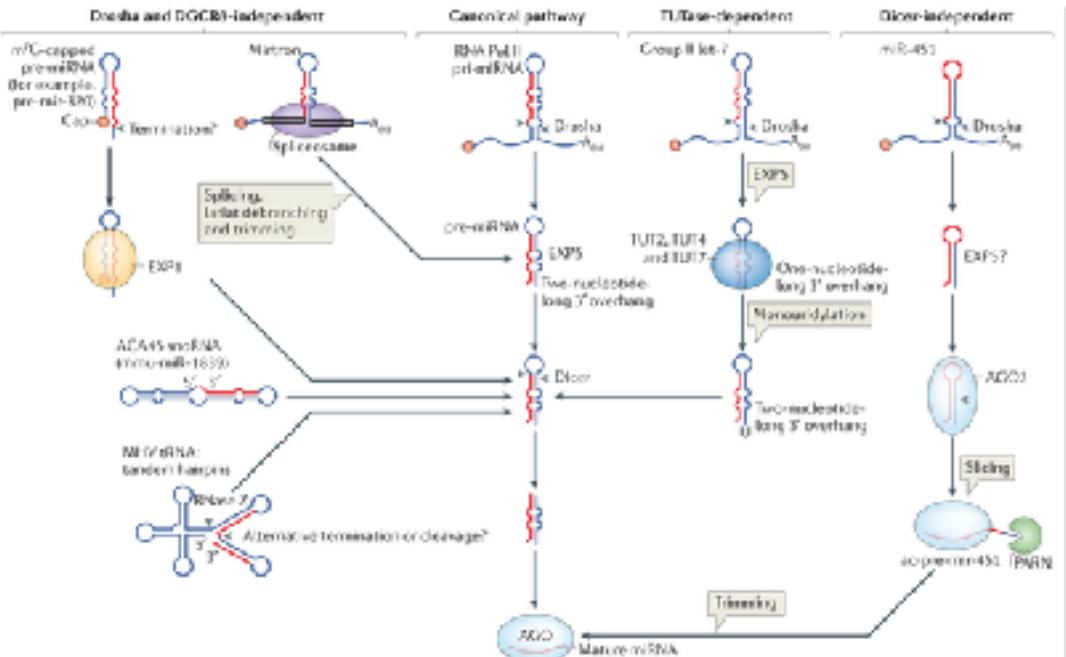
Aligned Reads on Precursor

```
UUCAGGUAAUCGUUUAUAGUUU 103_t9_w22 Depth:393 Modifications: no_modification
UUCAGGUAAUCGUUUAUAGUUU 481_t10_w21 Depth:698 Modifications: no_modification
UUCAGGUAAUCGUUUAUAGUUU 771_t9_w23 Depth:360 Modifications: nont_3p_U
UUCAGGUAAUCGUUUAUAGUUU 896_t11_w22 Depth:291 Modifications: nont_3p_A
UUCAGGUAAUCGUUUAUAGUUU 1300_t9_w23 Depth:175 Modifications: nont_3p_A
UUCAGGUAAUCGUUUAUAGUUU 1665_t10_w20 Depth:149 Modifications: no_modification
UUCAGGUAAUCGUUUAUAGUUU 3129_t11_w23 Depth:58 Modifications: nont_3p_AA
UUCAGGUAAUCGUUUAUAGUUU 3388_t10_w23 Depth:52 Modifications: nont_3p_AU
UUCAGGUAAUCGUUUAUAGUUU 5769_t9_w23 Depth:26 Modifications: no_modification
UUCAGGUAAUCGUUUAUAGUUU 8212_t10_w24 Depth:17 Modifications: nont_3p_AA
UUCAGGUAAUCGUUUAUAGUUU 8687_t9_w23 Depth:16 Modifications: nont_3p_CG
UUCAGGUAAUCGUUUAUAGUUU 9287_t11_w21 Depth:14 Modifications: nont_3p_A
UUCAGGUAAUCGUUUAUAGUUU 10438_t9_w23 Depth:13 Modifications: nont_3p_C
UUCAGGUAAUCGUUUAUAGUUU 11348_t9_w22 Depth:11 Modifications: nont_3p_A
UUCAGGUAAUCGUUUAUAGUUU 13435_t6_w19 Depth:9 Modifications: no_modification
UUCAGGUAAUCGUUUAUAGUUU 13682_t8_w24 Depth:9 Modifications: no_modification
UUCAGGUAAUCGUUUAUAGUUU 15177_t9_w19 Depth:8 Modifications: no_modification
UUCAGGUAAUCGUUUAUAGUUU 15767_t12_w23 Depth:7 Modifications: nont_3p_C
```

Distribution across Precursor



MicroRNA 3p modification Uridylation by tutase enzymes

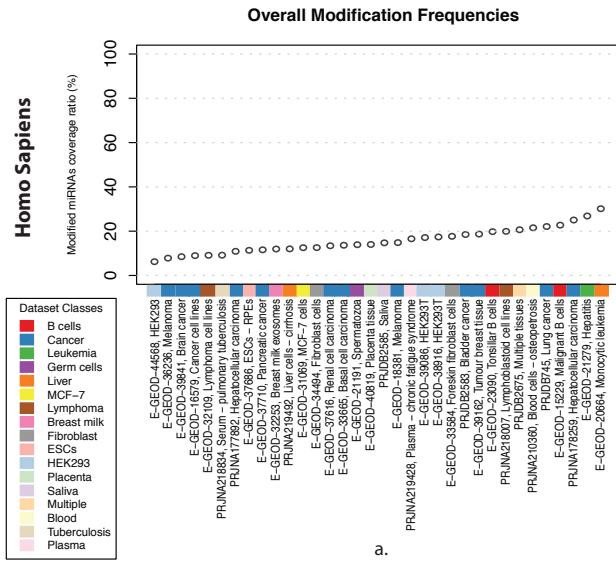


TUTase mutant data (mouse)

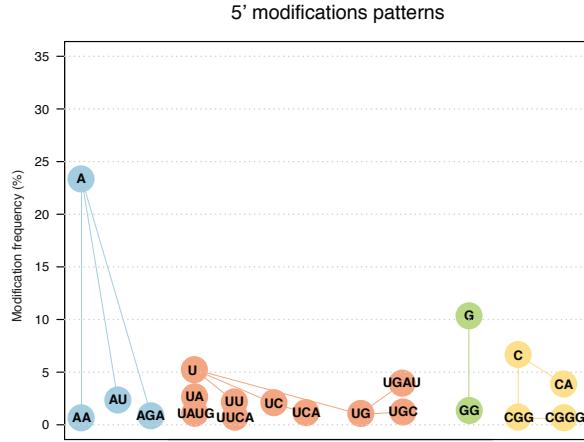
Francisco Sánchez Madrid Lab (T-cells)
Dónal O'Carroll Lab (Germline)

Global analysis of microRNA modification

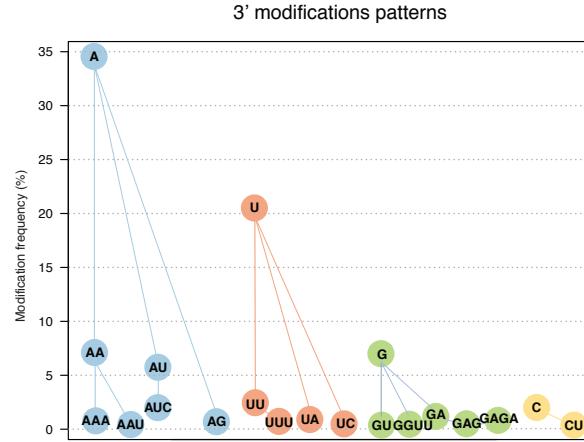
Homo Sapiens



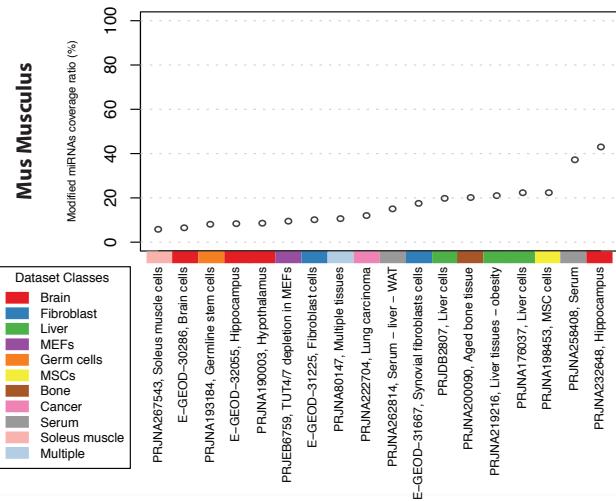
a.



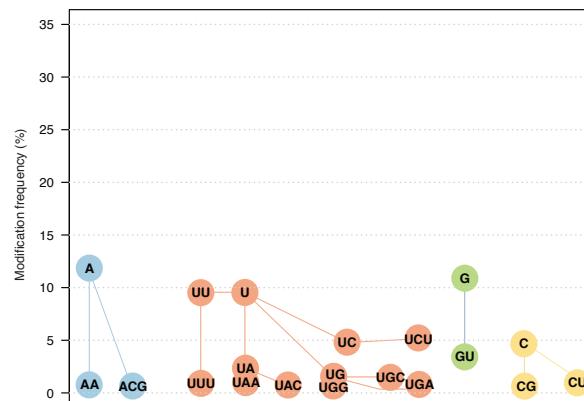
b.



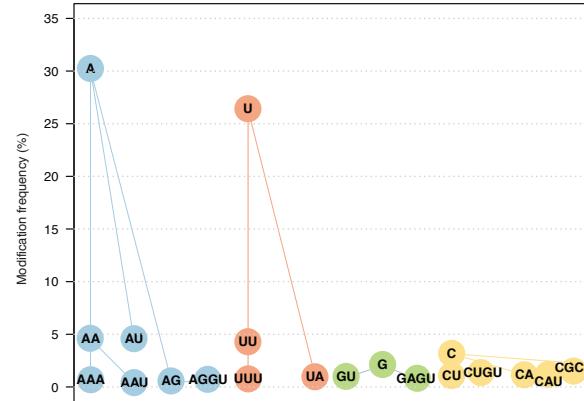
Mus Musculus



d.



e.





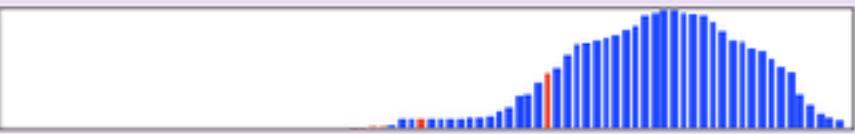
UNIVERSITY OF

mmu-mir-5109-3p

Depth: 1496 reads

Aligned Reads on Precursor

Distribution across Precursor

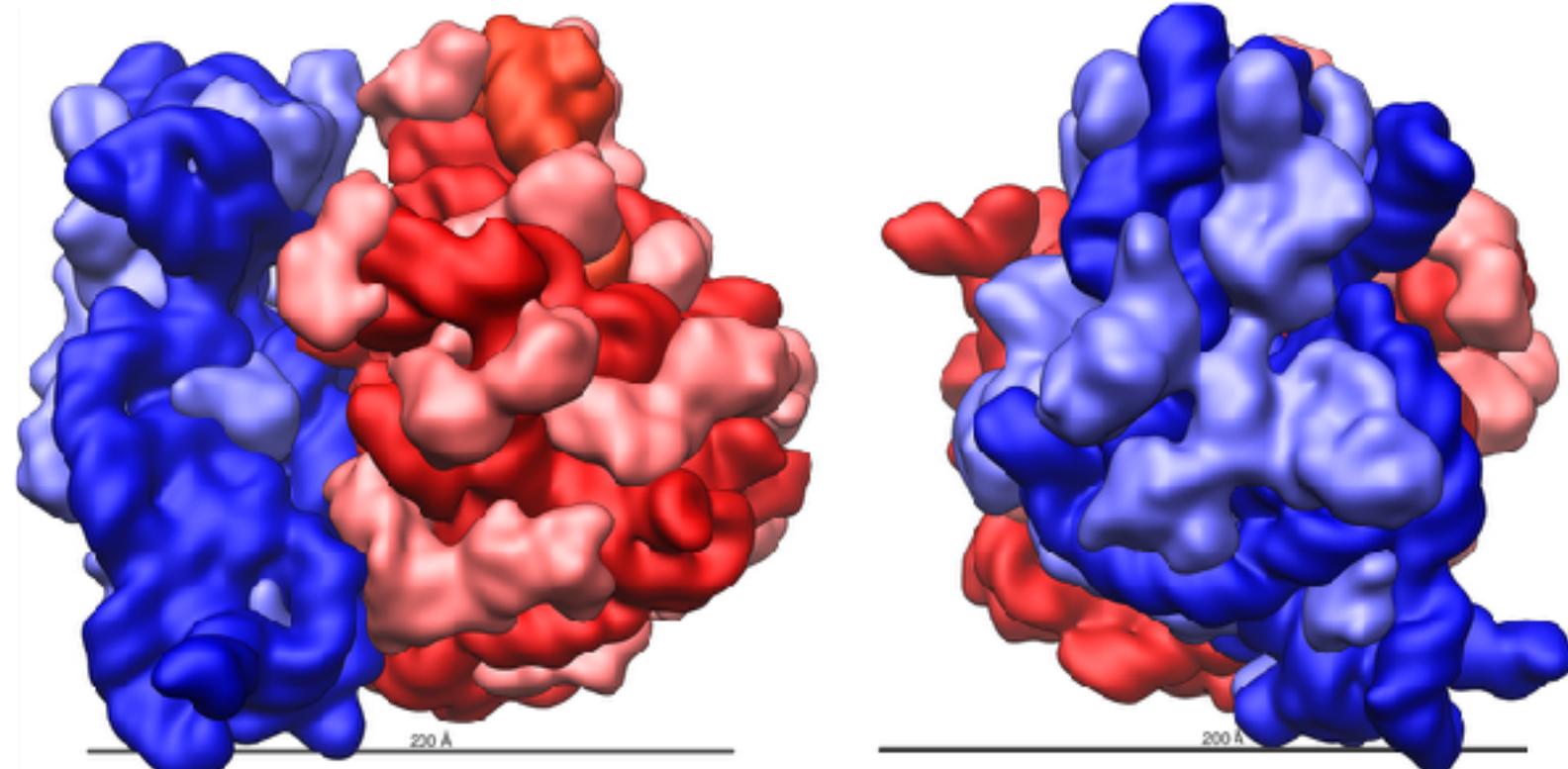




UNIVERSITY OF

mmu-mir-5109-3p

Depth: 1496 reads



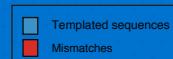
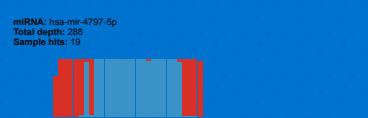
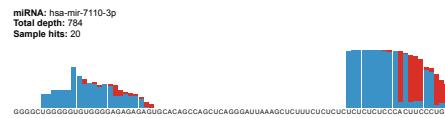
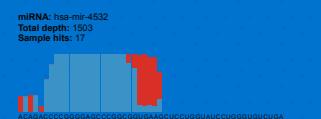
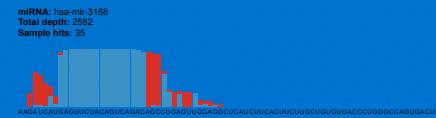
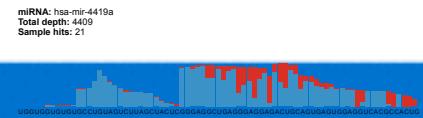
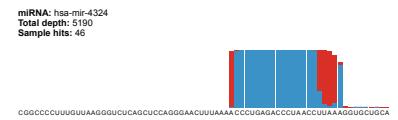
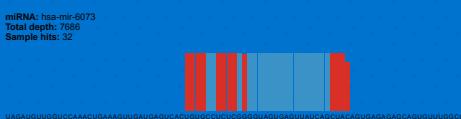
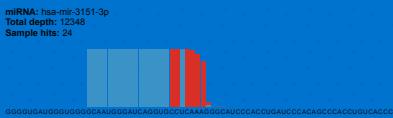
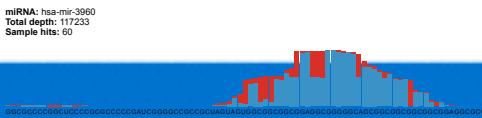
..... 7671_t12_w11 Depth:13 Modifications: no_modification
..... 7619_t8_w13 Depth:13 Modifications: no_modification
..... 7627_t8_w27 Depth:13 Modifications: no_modification

Distribution across Precursor





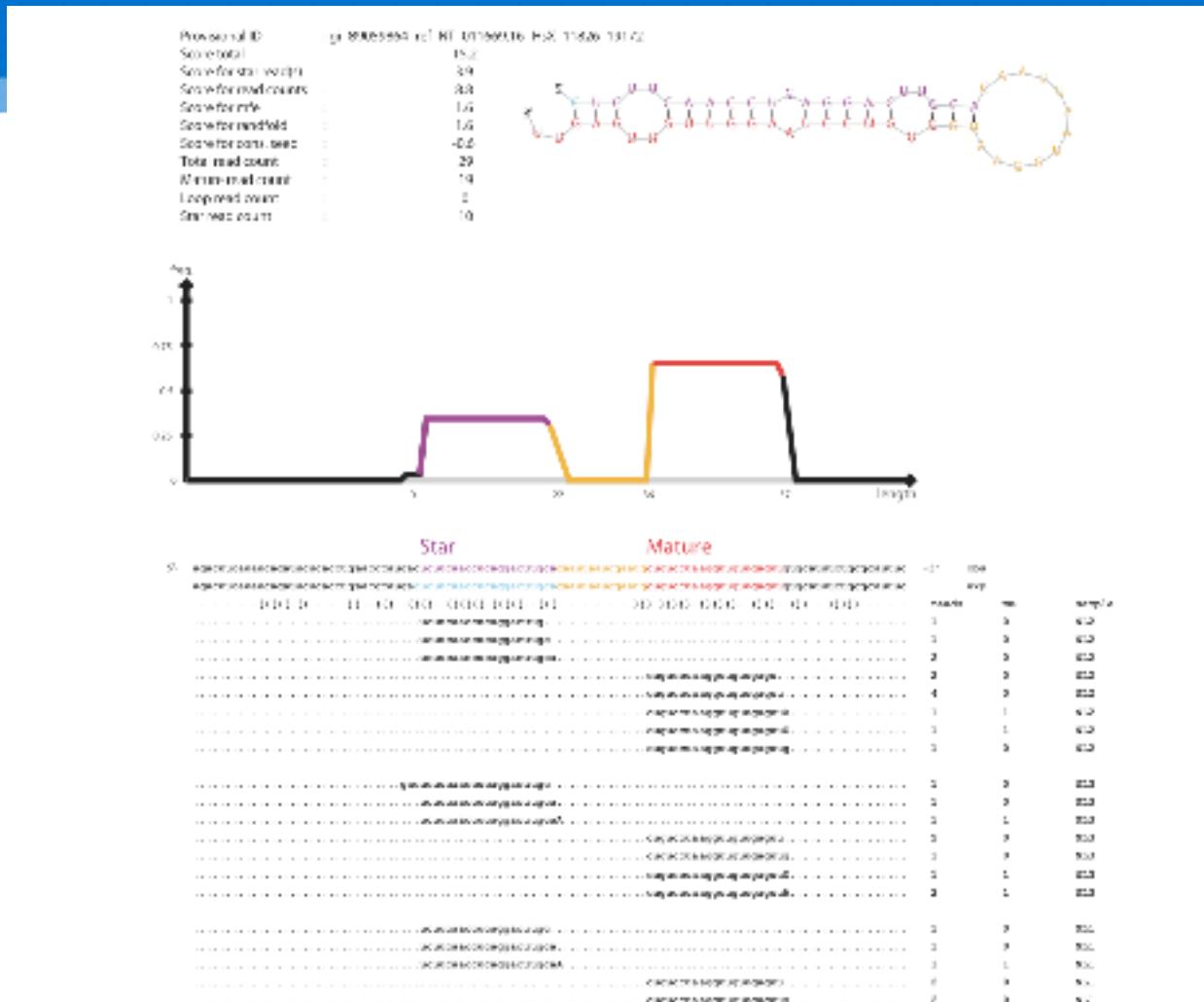
**mRNA: hsa-mir-1303
Total depth: 170430
Sample hits: 79**



Downstream Analysis

de novo microRNA Discovery

Identification of novel miRNAs



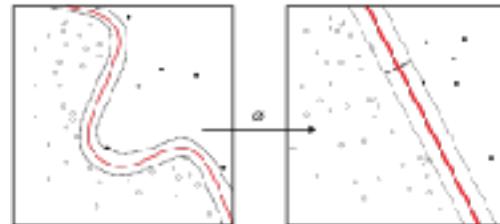
miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. NAR 2012

de novo microRNA discovery

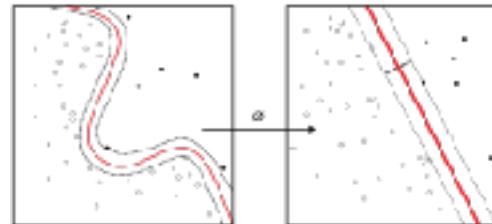
Can we use these features of microRNA alignment coverage to better predict novel microRNAs ?

Previous approaches rely on microRNA hairpin structure and conservation analysis (e.g. miRDeep)



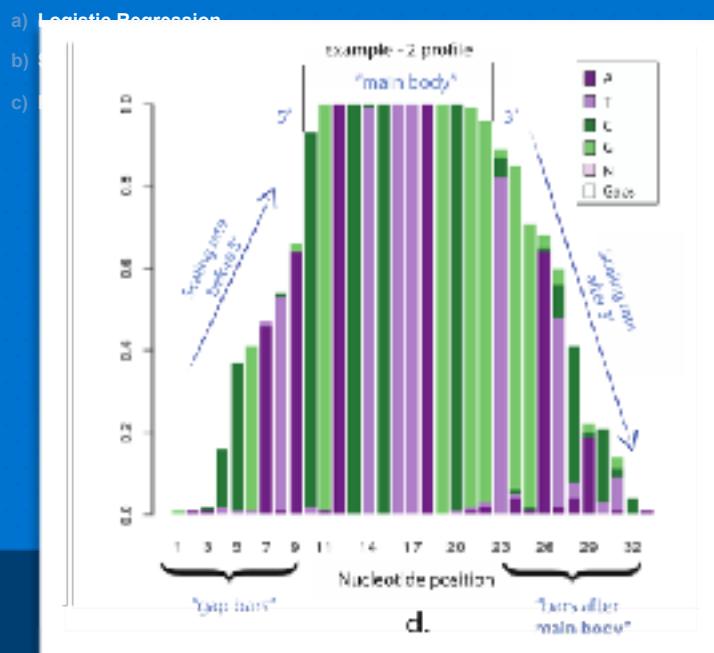
Machine Learning approach

- a) Logistic Regression
- b) Support Vector Machines
- c) Random Forest



Machine Learning approach

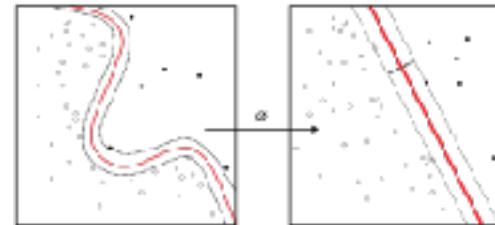
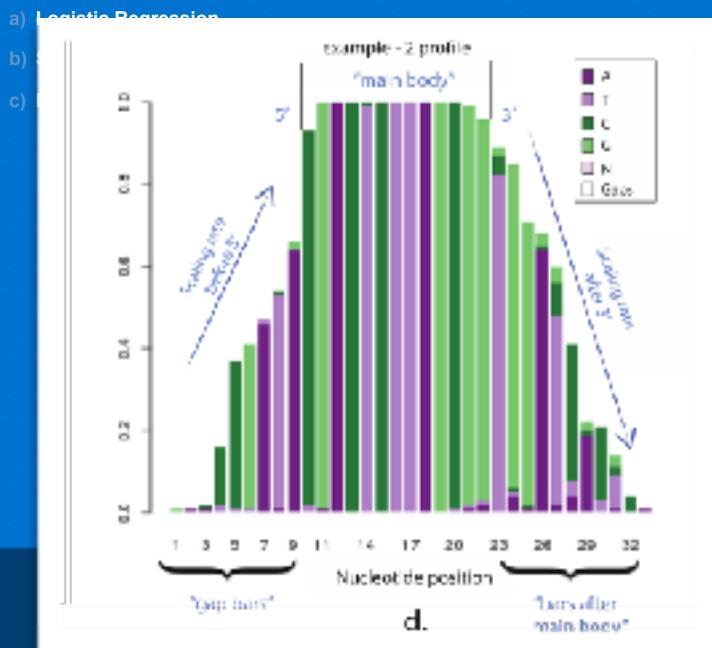
1. Core miRNA features





Machine Learning approach

1. Core miRNA features

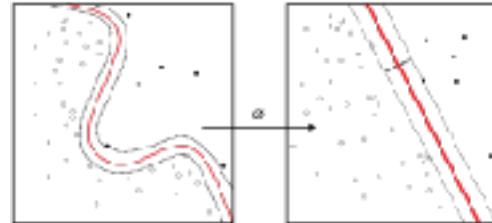
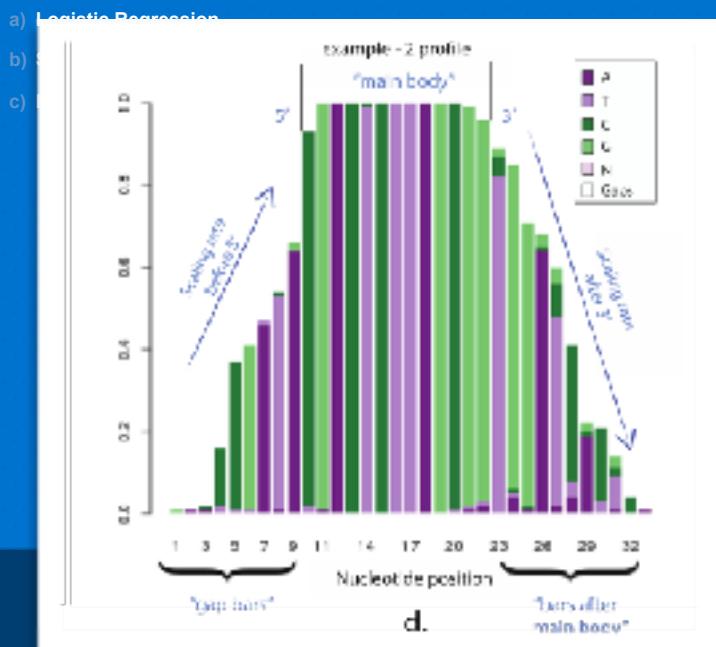


2. Sequence complexity features

- **gcs**: C+G skew
- **cpg**: CpG skew
- **cwf**: Complexity by Wootton & Federhen
- **ce**: Entropy
- **cz**: Complexity as compression ratio (using Gzip)
- **cmN**: Complexity as Markov model size of N
- **ctN**: Trifnov's complexity with order N
- **cIN**: Linguistic complexity with order N

Machine Learning approach

1. Core miRNA features



2. Sequence complexity features

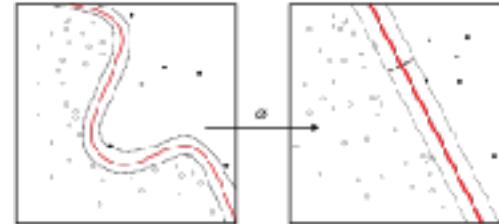
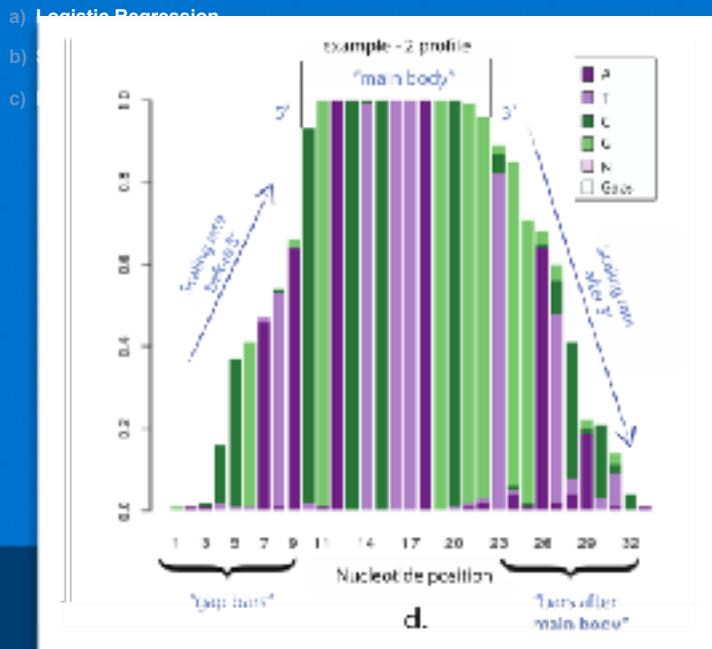
- **gcs:** C+G skew
- **cpg:** CpG skew
- **cwf:** Complexity by Wootton & Federhen
- **ce:** Entropy
- **cz:** Complexity as compression ratio (using Gzip)
- **cmN:** Complexity as Markov model size of N
- **ctN:** Trifnov's complexity with order N
- **cIN:** Linguistic complexity with order N

3. Genomic features

- **hairpin size**
- **number of unpaired bp**
- **min. free energy**
- **loop distance from hairpin stem**
- etc...

Machine Learning approach

1. Core miRNA features

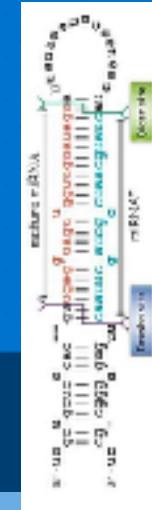


2. Sequence complexity features

- **gcs**: C+G skew
- **cpg**: CpG skew
- **cwf**: Complexity by Wootton & Federhen
- **ce**: Entropy
- **cz**: Complexity as compression ratio (using Gzip)
- **cmN**: Complexity as Markov model size of N
- **ctN**: Trifnov's complexity with order N
- **clN**: Linguistic complexity with order N

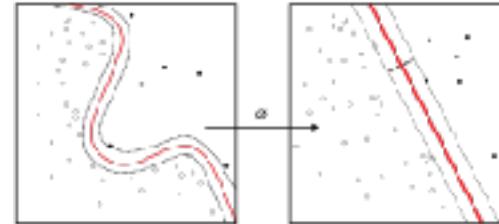
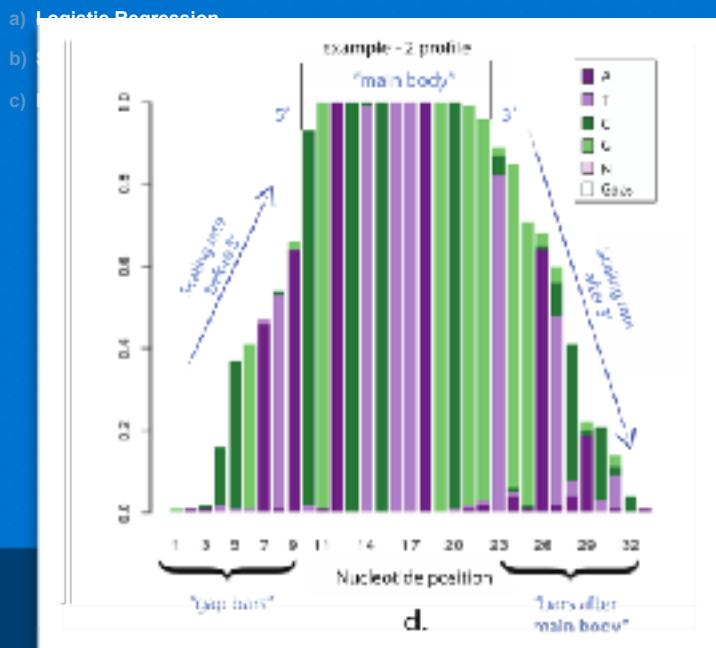
3. Genomic features

- hairpin size
- number of unpaired bp
- min. free energy
- loop distance from hairpin stem
- etc...



Machine Learning approach

1. Core miRNA features



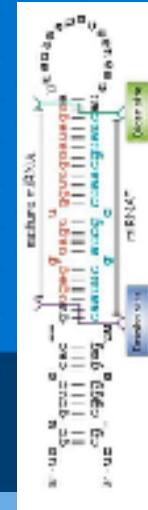
2. Sequence complexity features

- **gcs**: C+G skew
- **cpg**: CpG skew
- **cwf**: Complexity by Wootton & Federhen
- **ce**: Entropy
- **cz**: Complexity as compression ratio (using Gzip)
- **cmN**: Complexity as Markov model size of N
- **ctN**: Trifnov's complexity with order N
- **cIN**: Linguistic complexity with order N

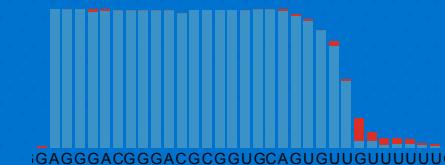
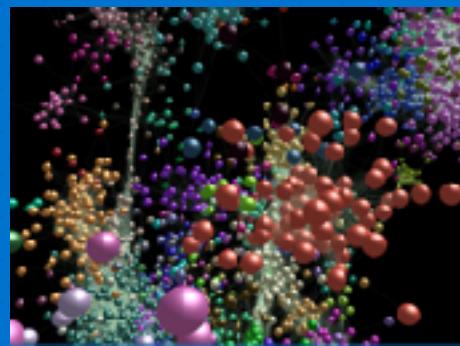
3. Genomic features

- hairpin size
- number of unpaired bp
- min. free energy
- loop distance from hairpin stem
- etc...

34 features overall



mirnovo: workflow



3' adapter trimming [**reaper**] & de-duplication [**tally**]

Sequence clustering with **vsearch**

Clusters filtering:
Retain clusters with a min. number of depth and variants

Assess temp. consensus sequence for each cluster

Align against **rfam** and retain tRNA and rRNA hits.

Clusters refinement:

- Pairwise alignment of all consensus sequences [**swan**]
- Clustering of temp. consensus sequences [**mcl**]
- Merge similar clusters

Multiple-Sequence Alignment within each cluster [**muscle**]

Map against mirbase to retrieve info for know hits (if species is)

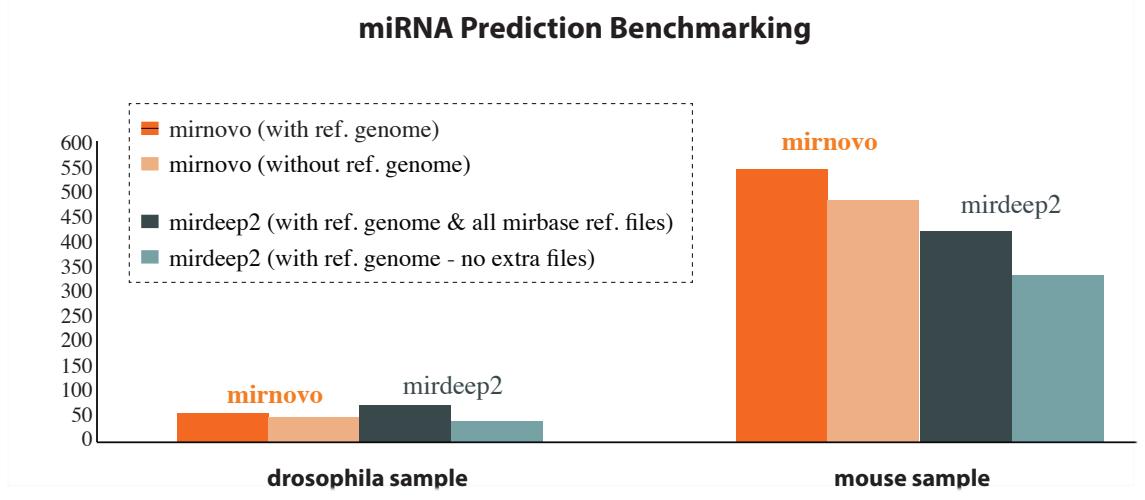
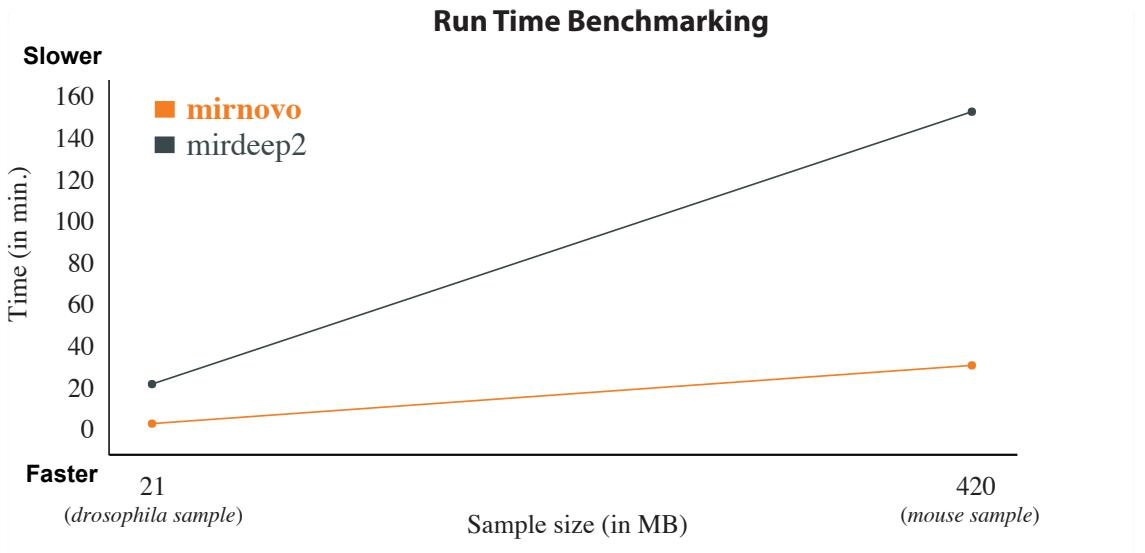
Calculate core and sequence complexity features for each cluster

Align against genome (if it is available)

- Extract genomic features for each cluster

Create feature table

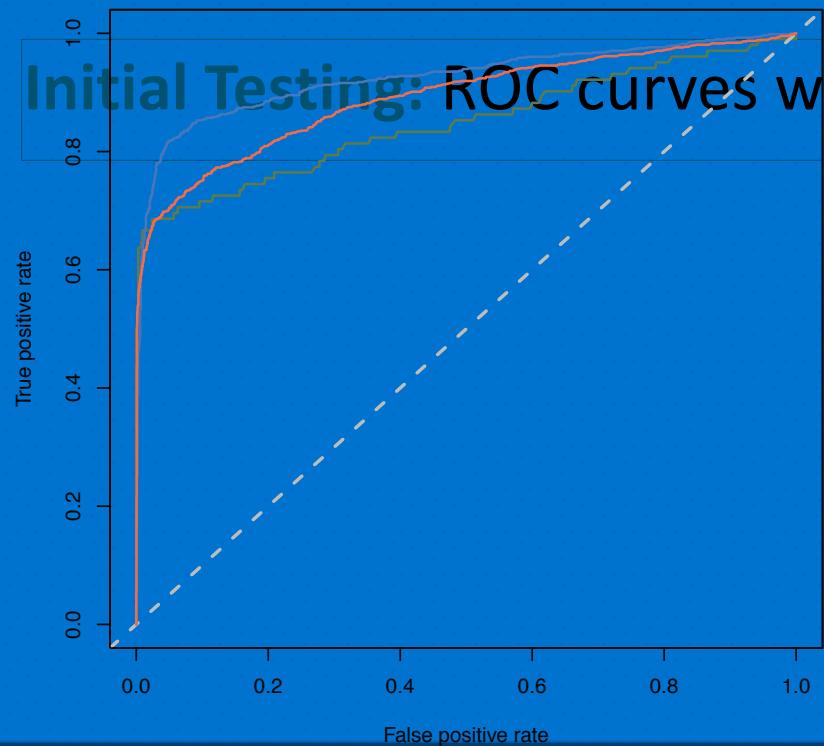
Predict with a pre-trained classifier:
random forest, SVM, Deep Learning, Logistic Regression



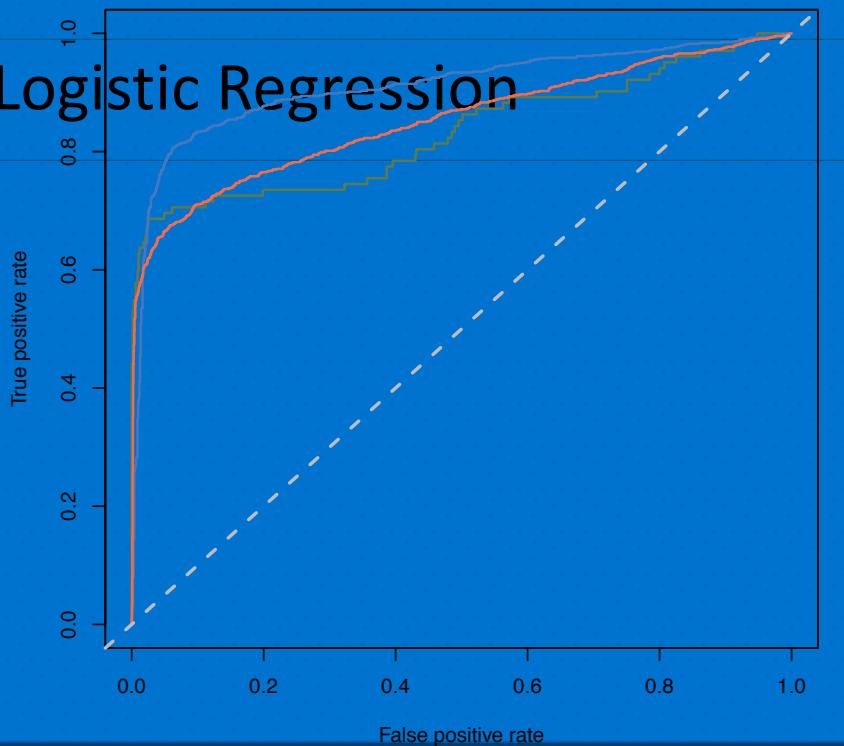
ROC Curves from test datasets

- human placenta sample
- mouse liver sample
- drosophila sample

Using Random Forest Classifier



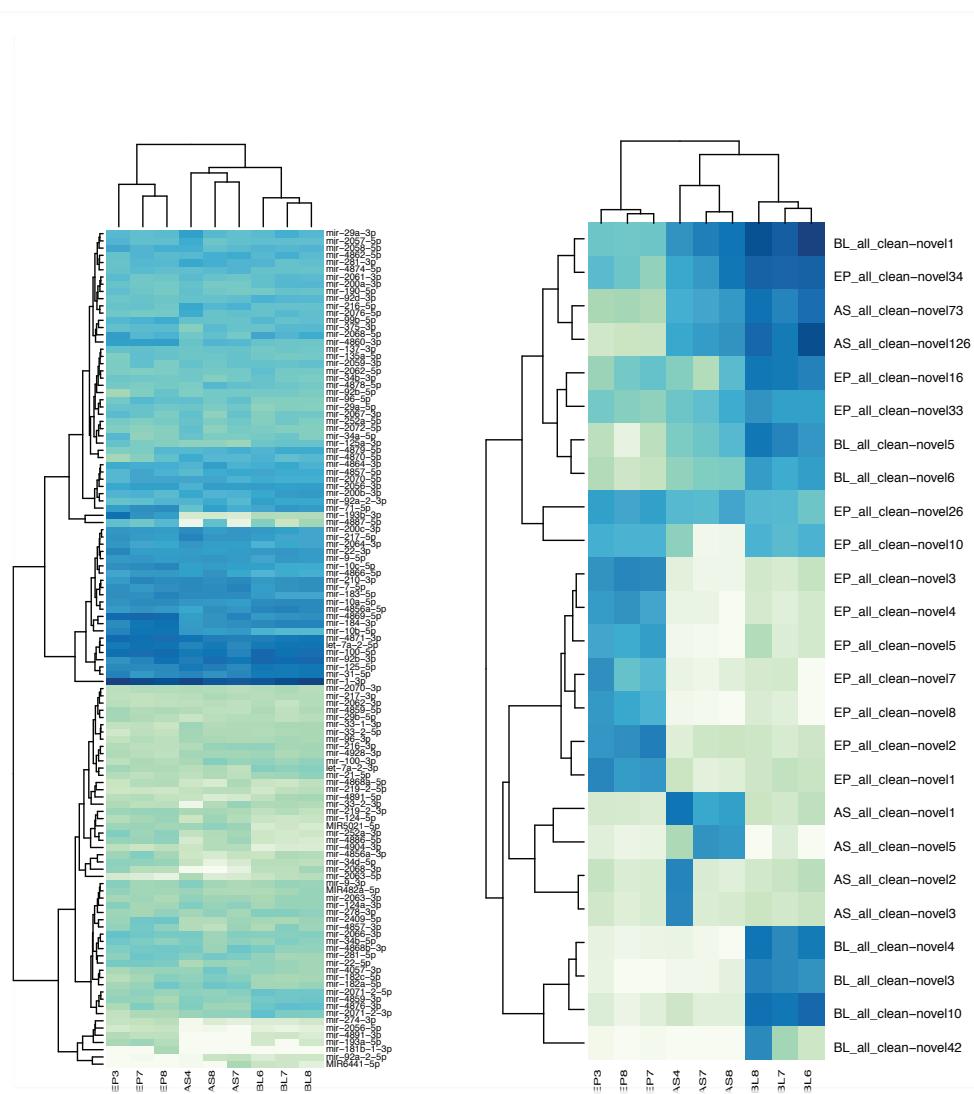
Using SVM Classifier



Initial Testing: ROC curves with Logistic Regression



Running in genome

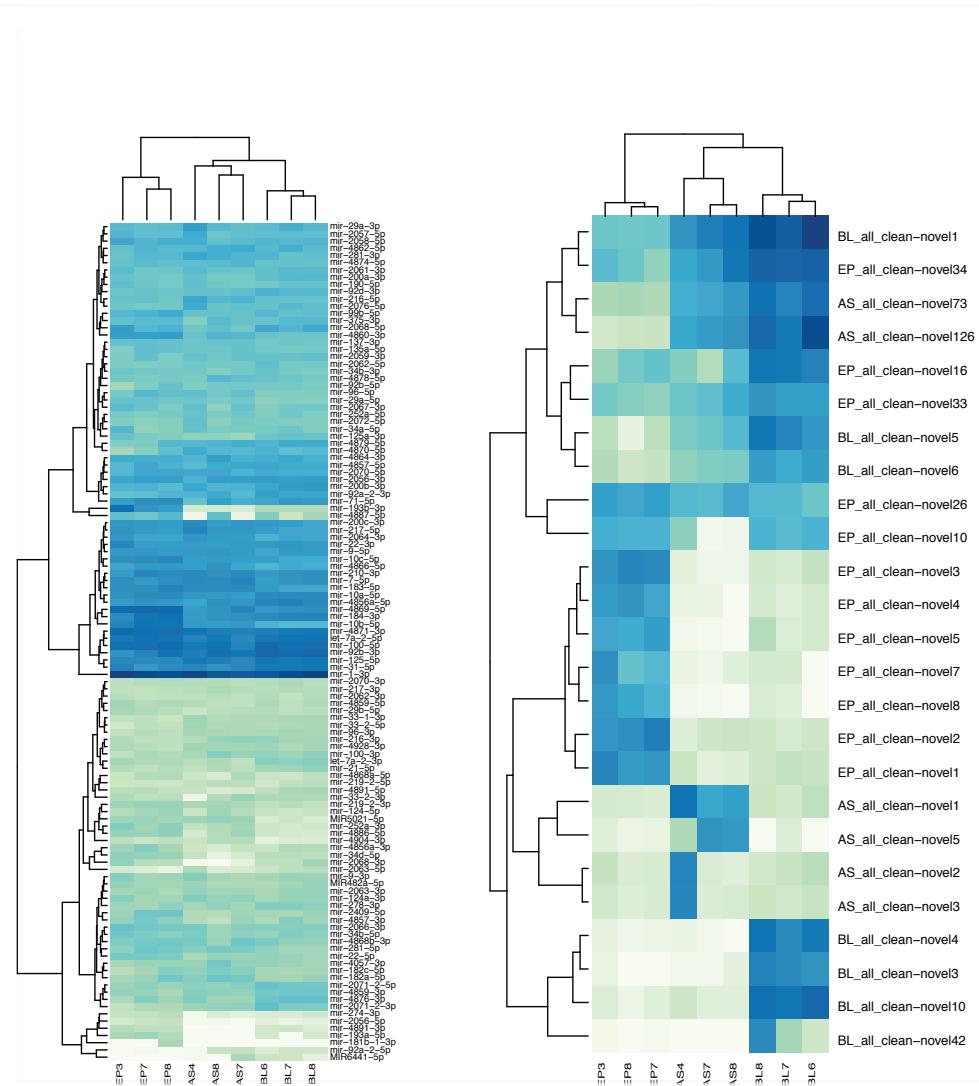




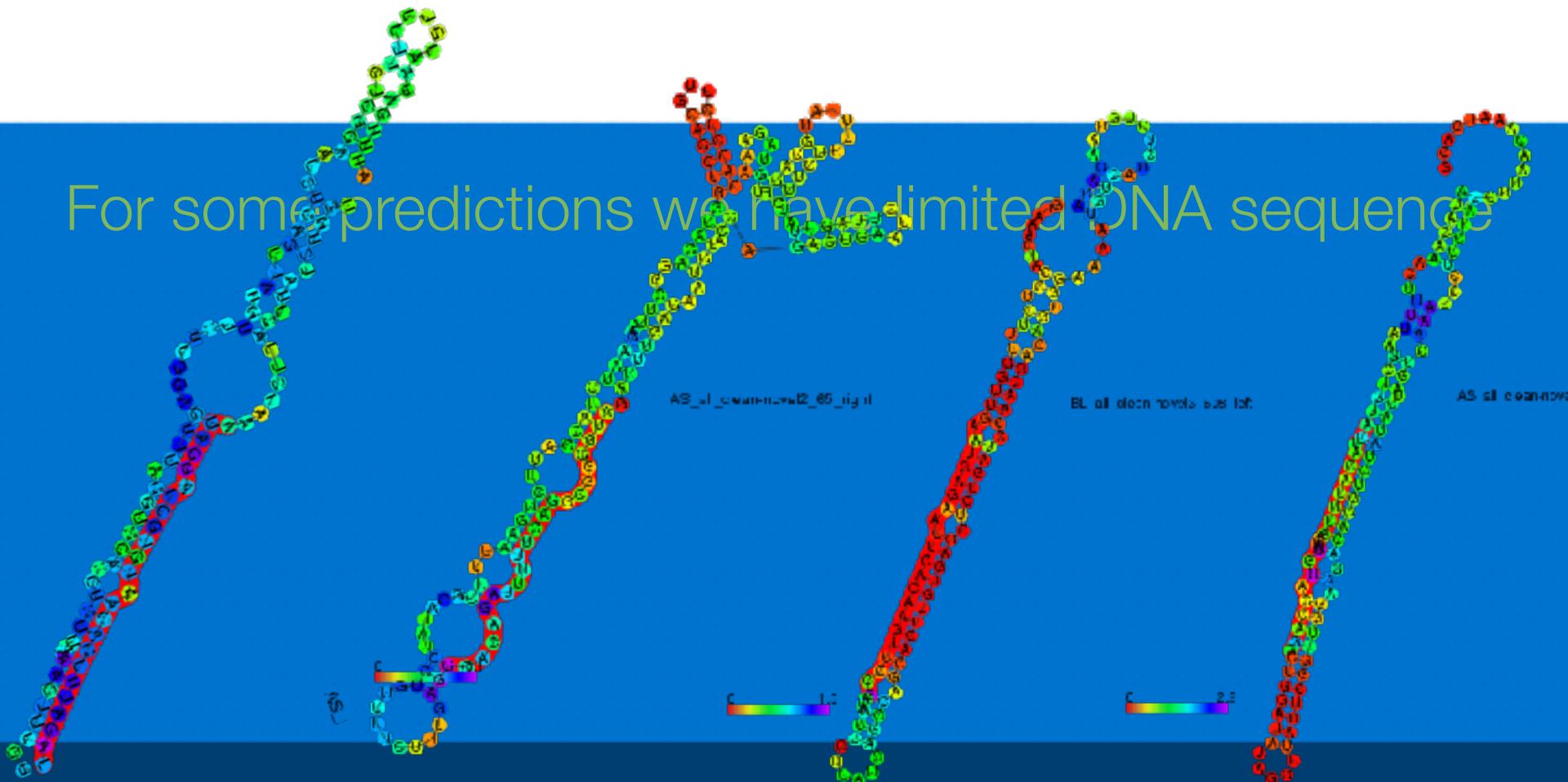
Running in genome

Amphioxus (lancelet fish)

With Elia Benito-Gutierrez (Zoology)

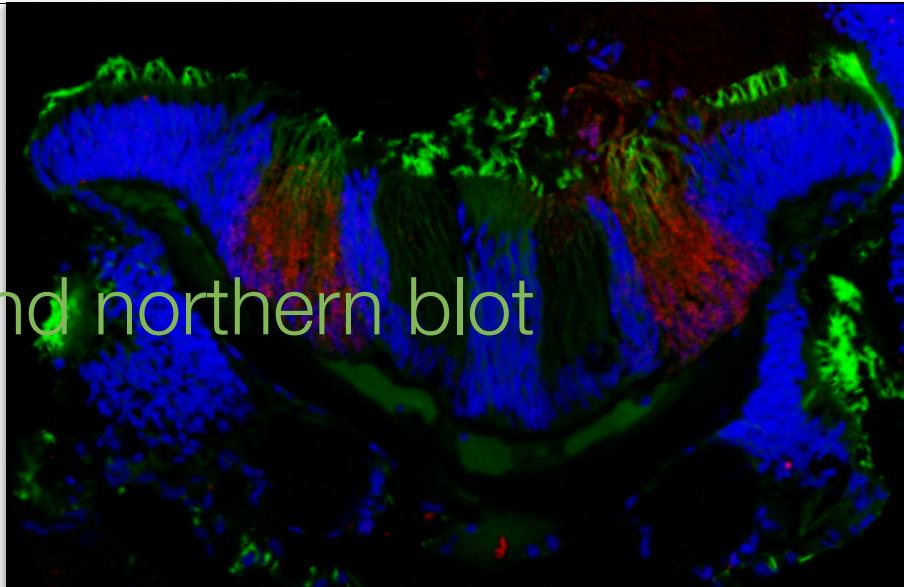
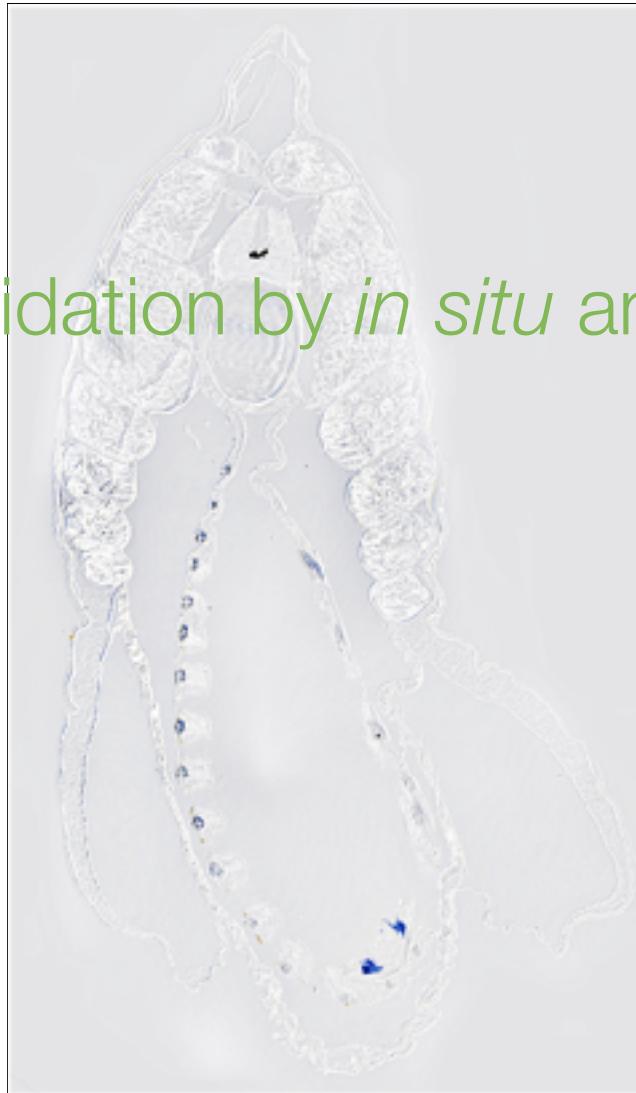


For some predictions we have limited DNA sequence

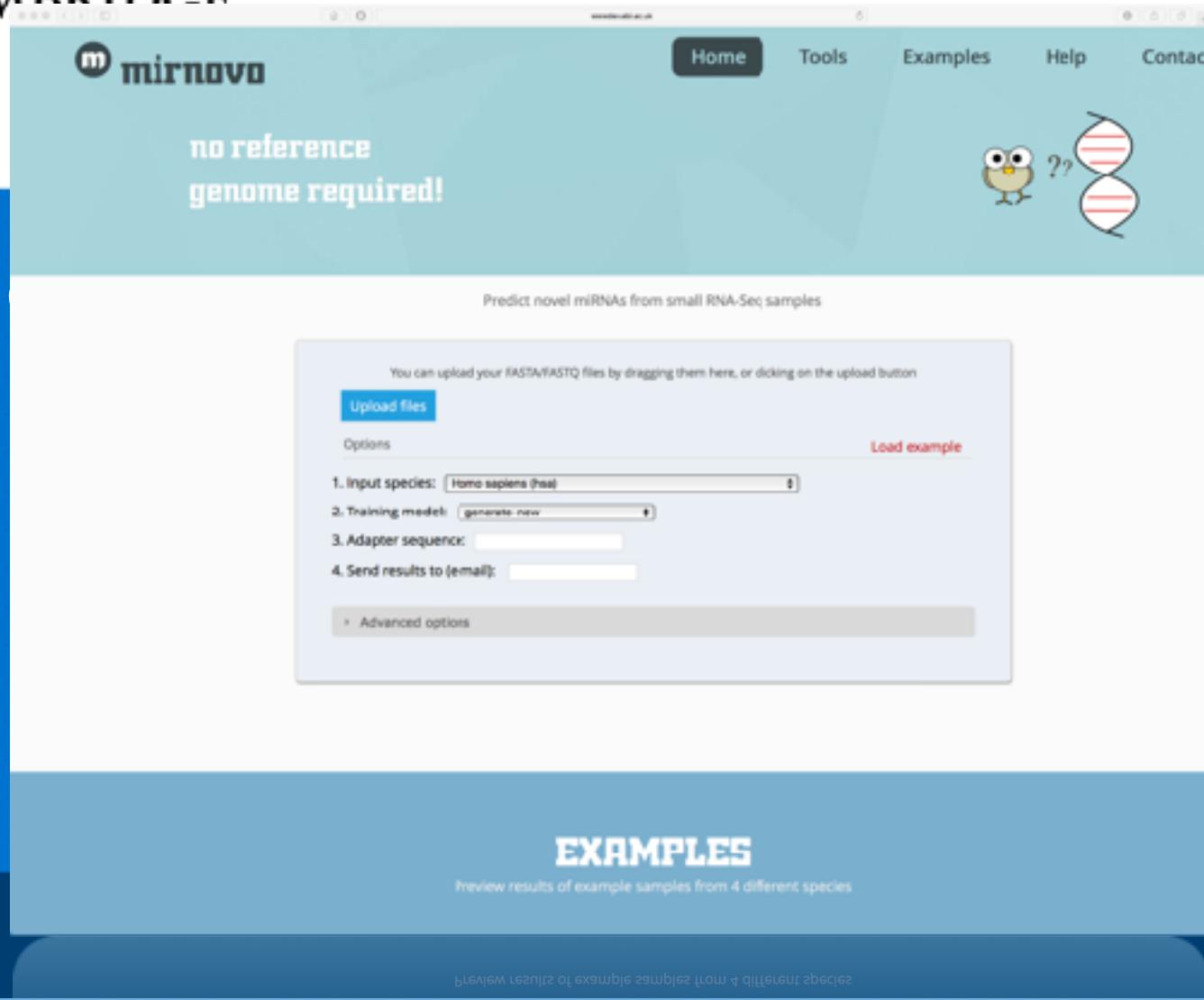




Validation by *in situ* and northern blot



mirn



The screenshot shows the mirnovo web application interface. At the top, there is a navigation bar with links for Home, Tools, Examples, Help, and Contact. To the left of the main content area, there is a large, stylized 'm' logo.

mirnovo

no reference genome required!

Predict novel miRNAs from small RNA-Seq samples

You can upload your FASTA/FASTQ files by dragging them here, or clicking on the upload button

Upload files

Options [Load example](#)

1. Input species:

2. Training model:

3. Adapter sequence:

4. Send results to (email):

[Advanced options](#)

EXAMPLES
Preview results of example samples from 4 different species

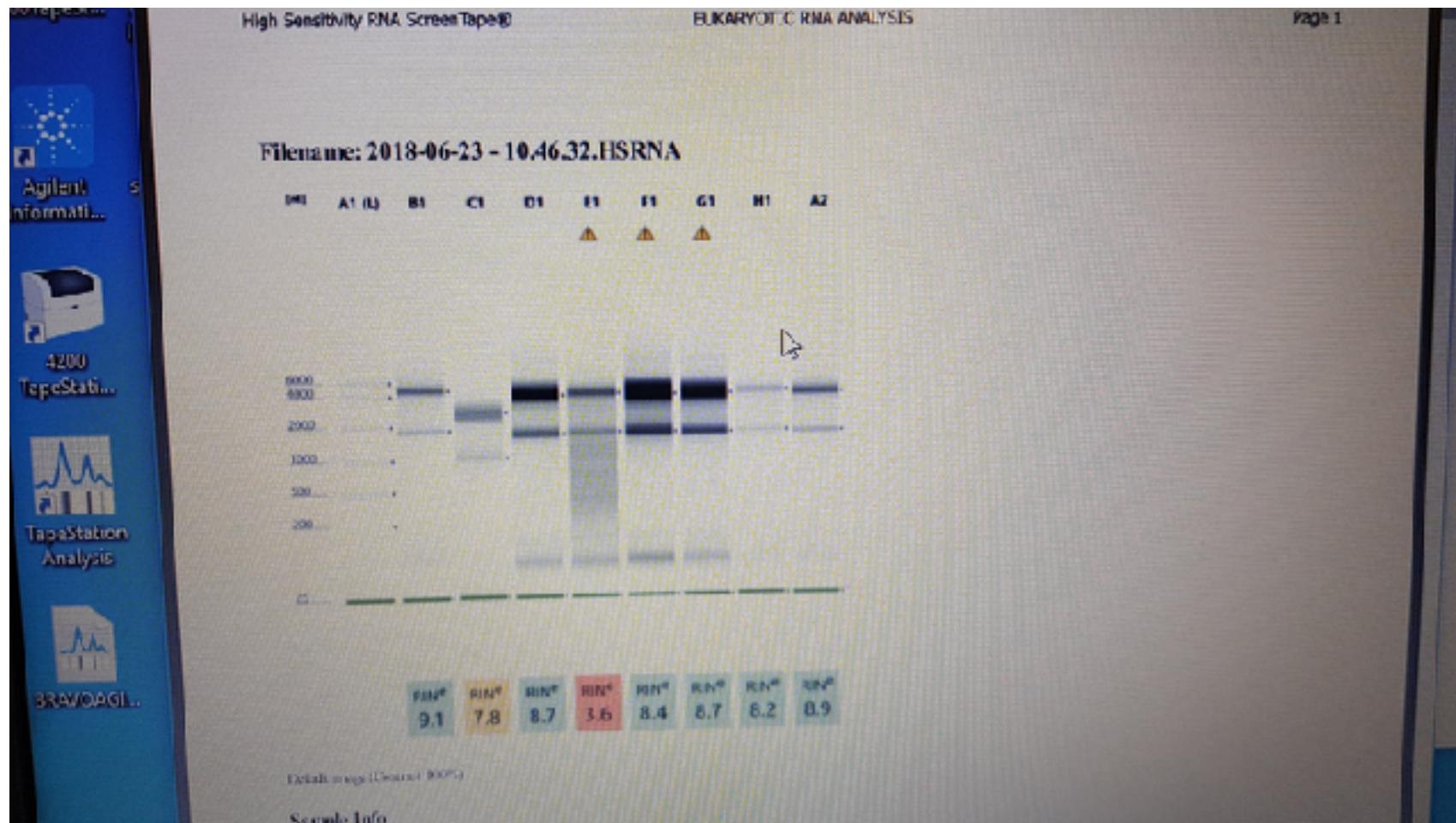
samples from different 4 most samples a different 4 samples to preview

EXAMPLES

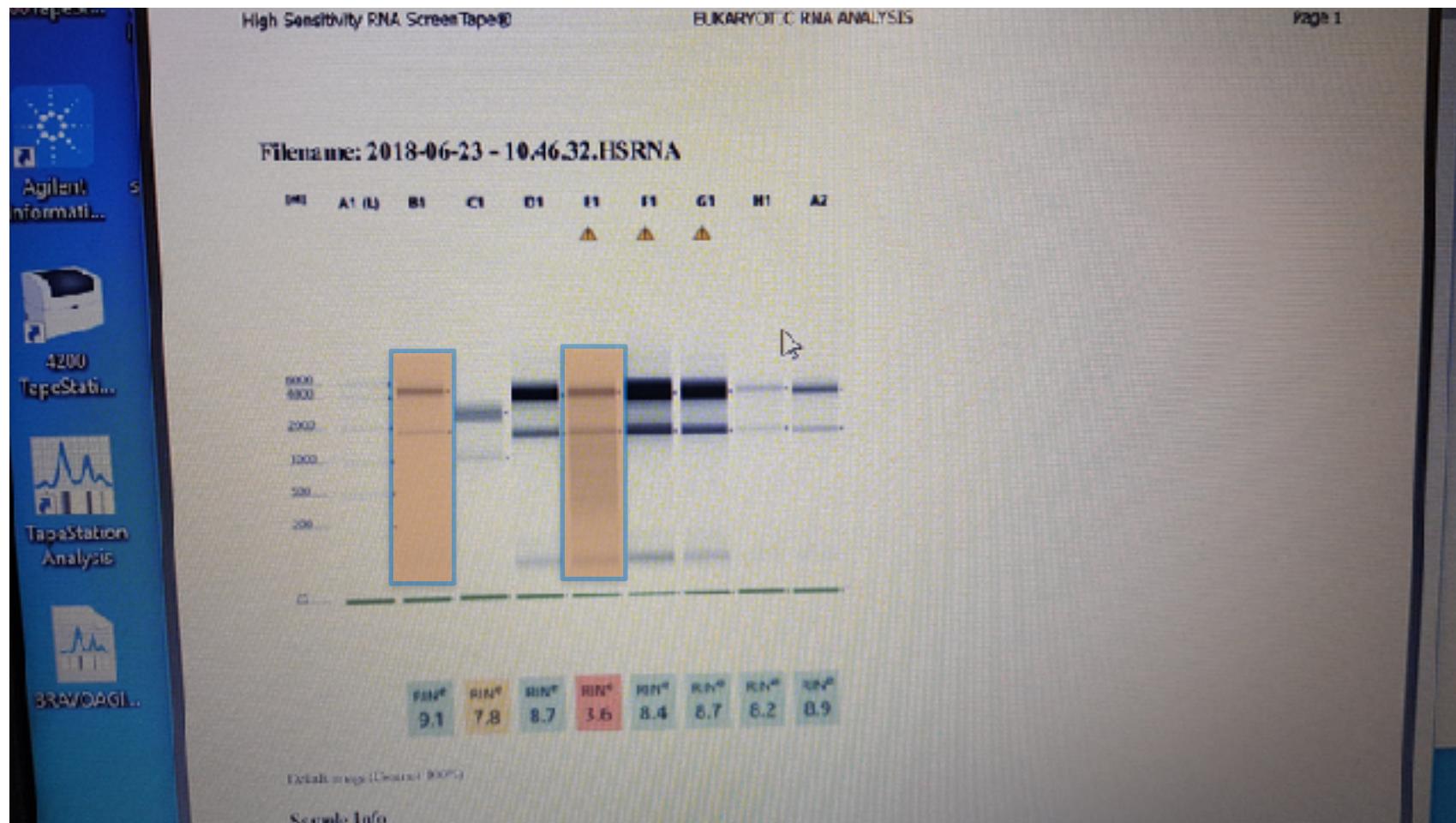
Nanopore Direct RNA

- polyA+ Yield from Alena and Allison's cells was very low (1% mRNA)
- 5-6mg Total RNA per sample
- 30-60ng polyA+ RNA
- **Will prepare new samples**
 - MCF7 Breast Cancer
 - Yeast Nanopore Control
 - RNA Methylation Sample (Gurdon Institute)

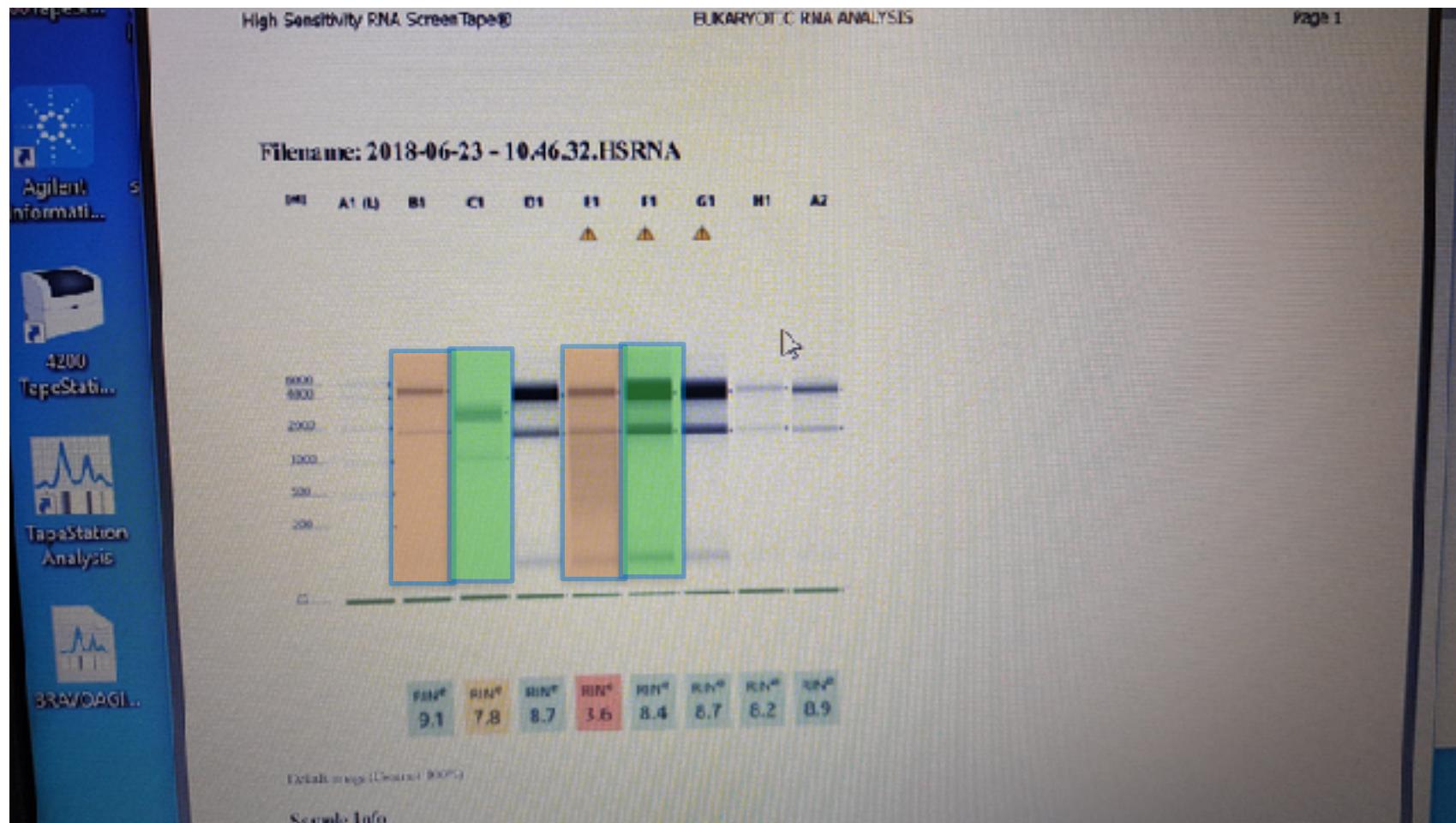
Nanopore Direct RNA



Nanopore Direct RNA



Nanopore Direct RNA



mRNA-Seq Data

mRNA Seq - Example Dataset

Breast Cancer Samples (MCF7)

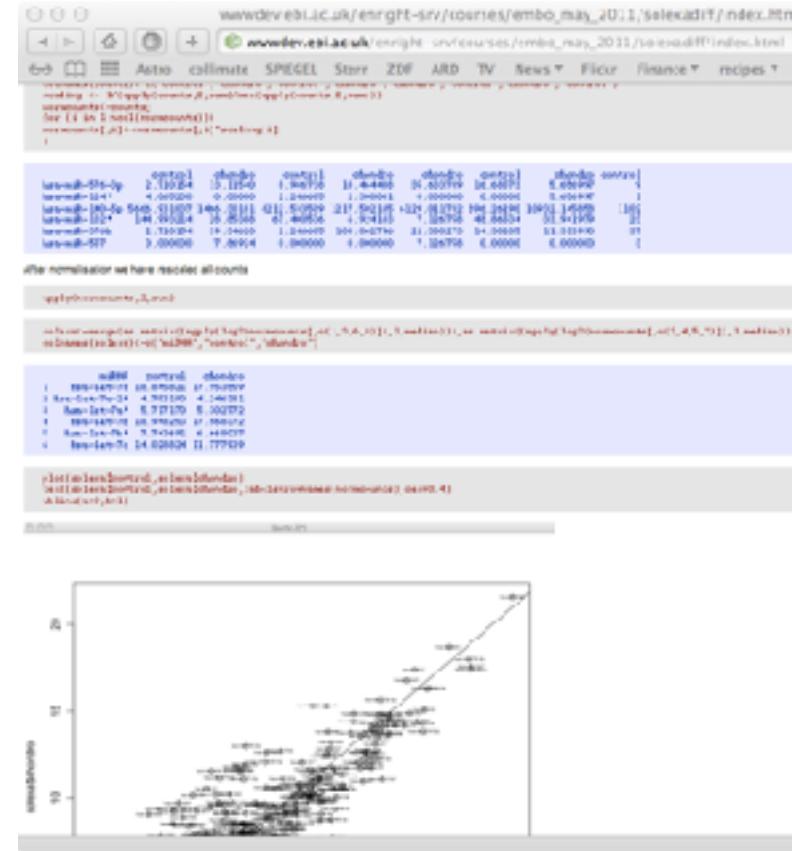
Jiannis Raggousis, McGill University, Montreal

MCF7 Line

3 Scrambled Controls

3 miR-210 knockdowns

Sequenced on Illumina
MiSeq
Wellcome Trust Advanced Course 2016



mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, RSEM,

Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, RSEM,

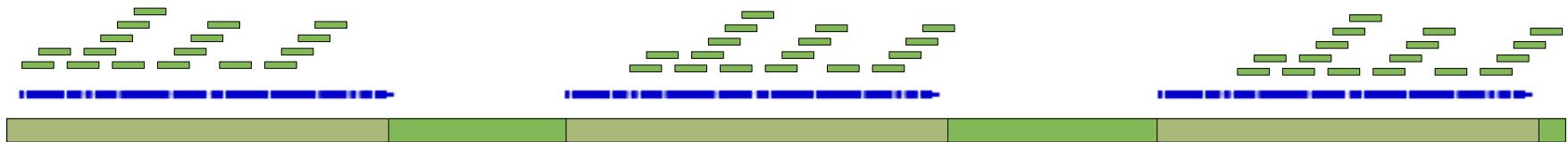


Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, RSEM,

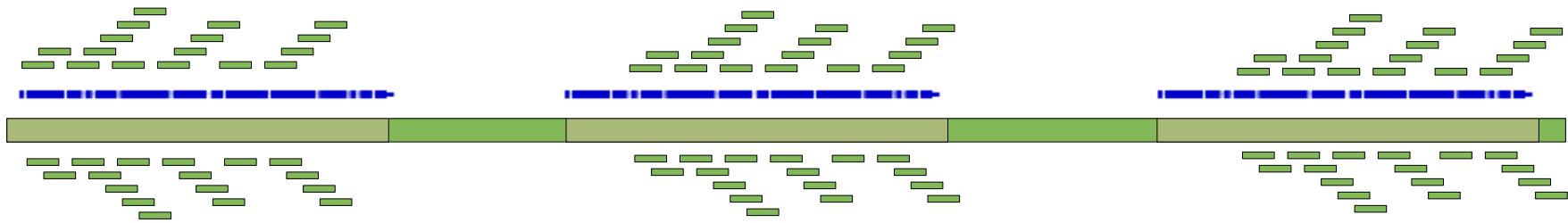


Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, RSEM,

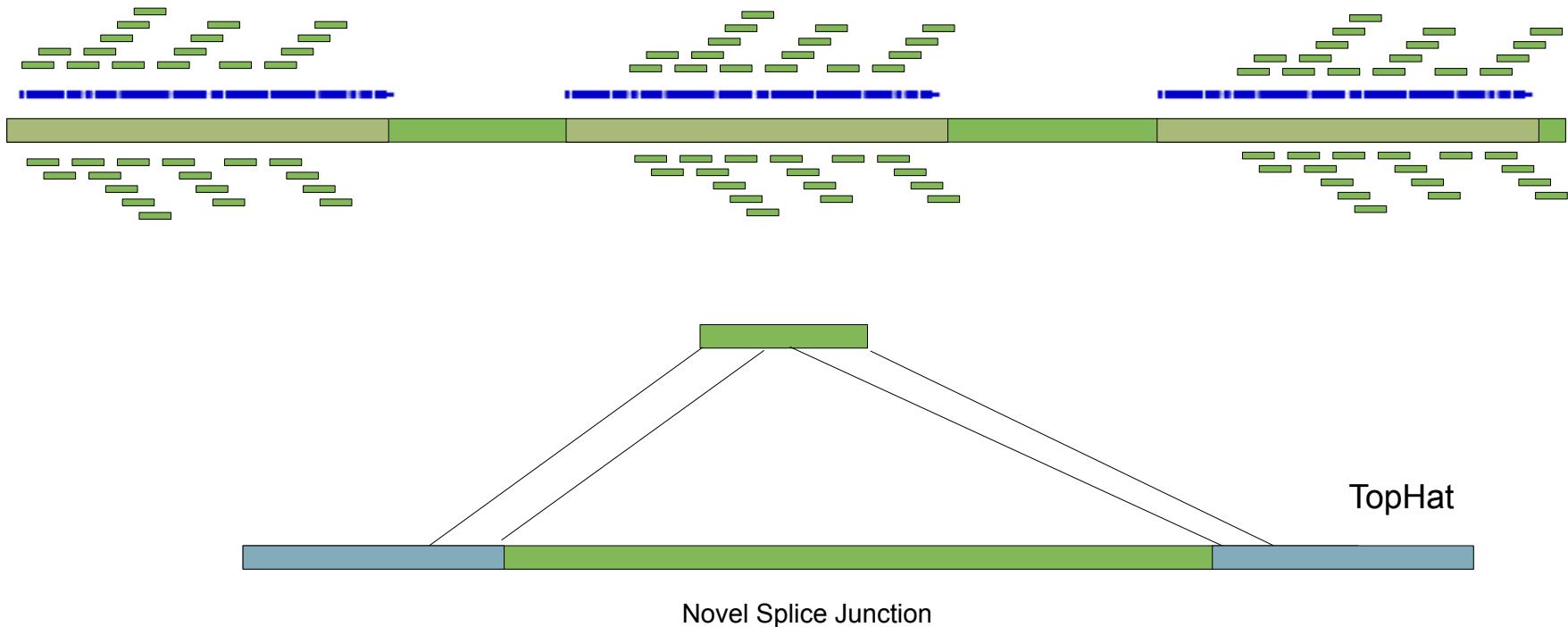


Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, RSEM,



Results in count data and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information

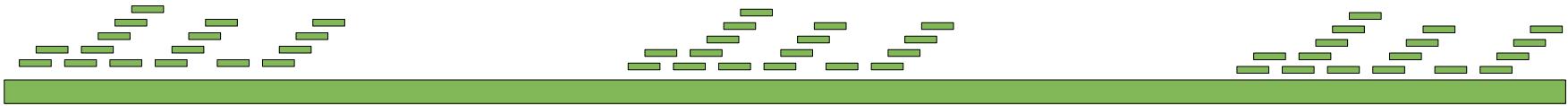
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



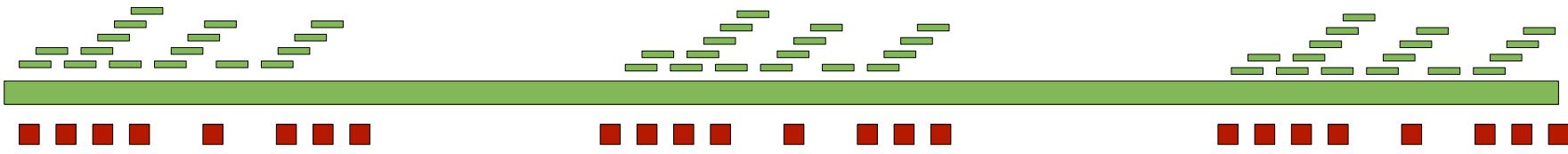
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



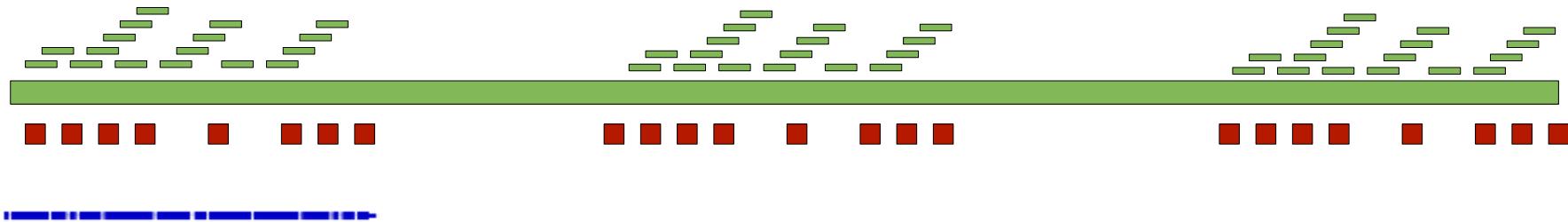
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



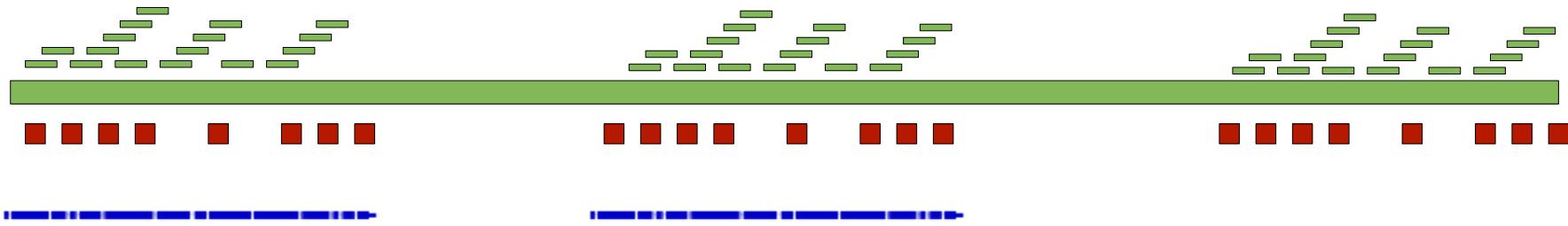
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



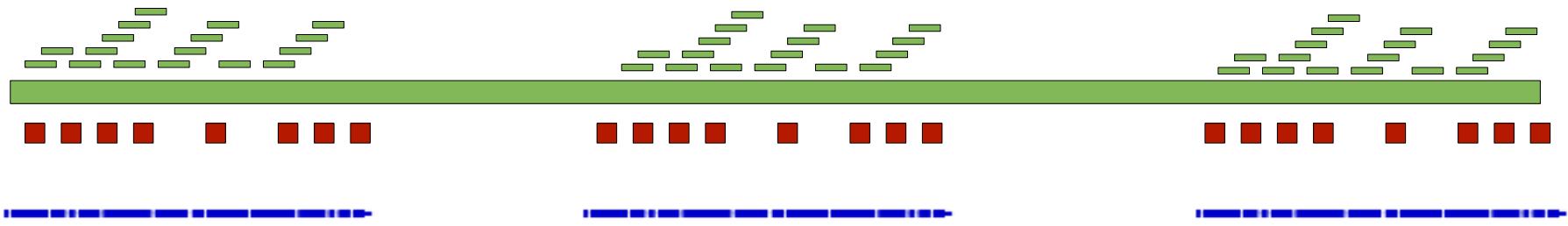
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome

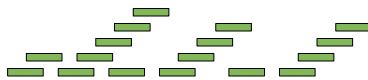
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



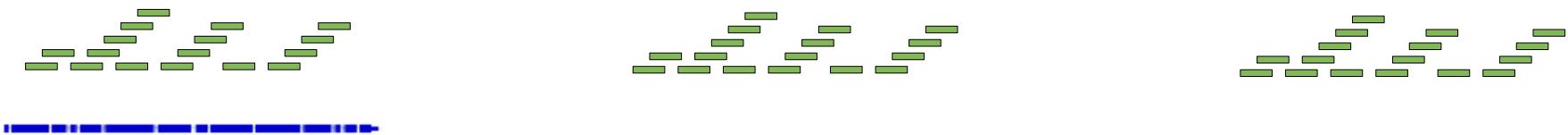
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



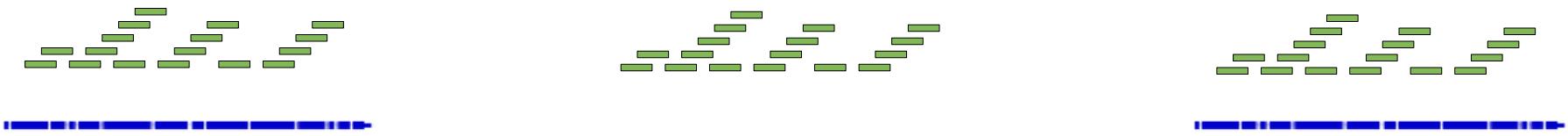
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



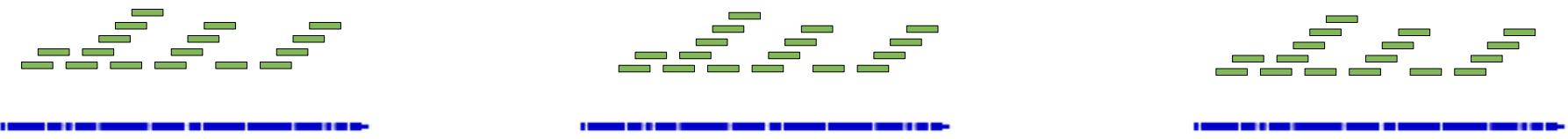
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



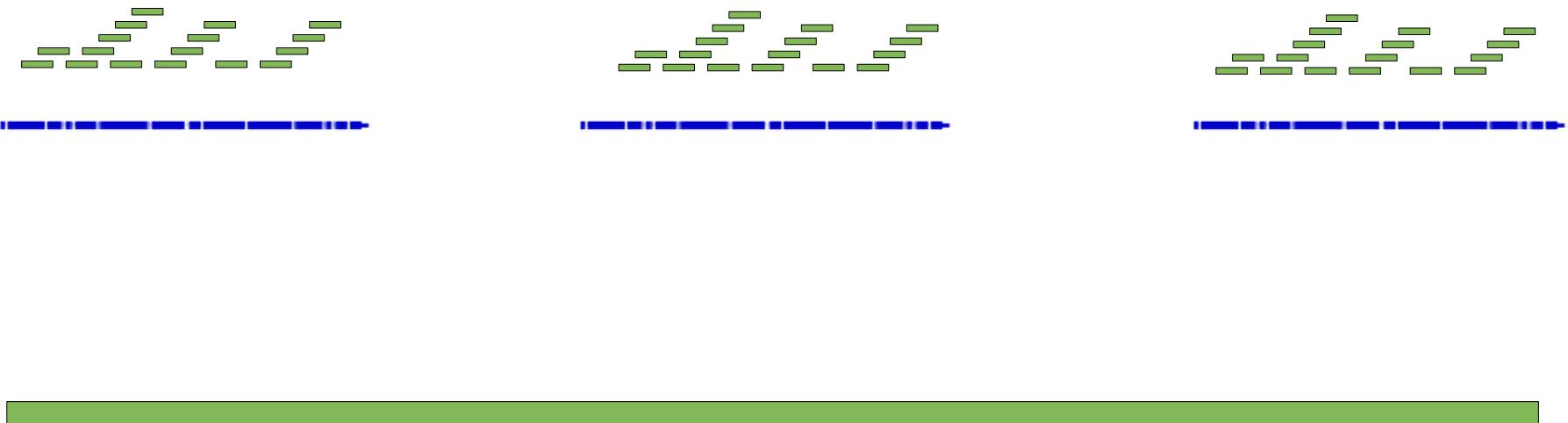
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



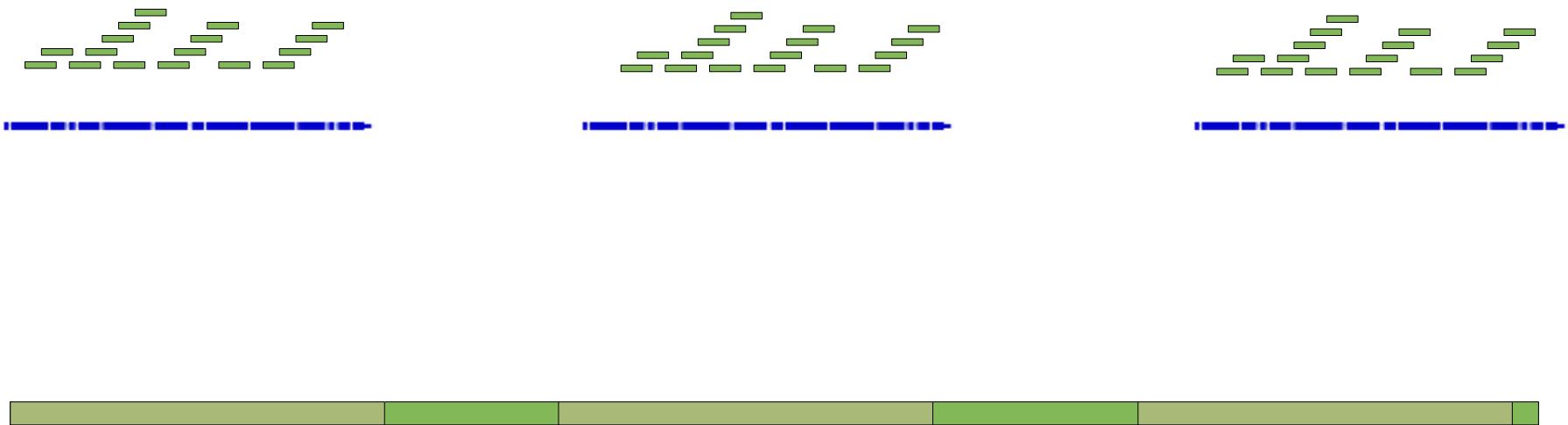
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

Quantitating Expression Levels

FPKM = fragments per kilobase

Normalised counts according to transcript length

Data Types encountered

- **FASTQ** - Raw Read Data
- **BAM** - Binary Alignment Format
- **SAM** - Text Alignment Format
- **ebwt** - Bowtie Genome Index File
- **GFF** - Gene Models on a reference
- **GTF** - Gene Models on a reference
- **BED** - Gene Models on a reference
- **TDF** - Coverage File (IGV)
- **SAMTOOLS** is a package for converting between different formats

SAM/BAM Format: e.g. SAMTOOLS

```
Coor ref
+r001/1
+r002
+r003
+r004
-r003
-r001/2
12345678901234 5678901234567890123456789012345
AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
    TTAGATAAAGGATA*CTG
    aaaAGATAA*GGATA
gcctaAGCTAA
ATAGCT.....TCAGC
    ttagctTAGGC
CAGCGCCAT
```

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002
r003
r004
r003
r001 83 ref 37 30 9M
0 AAAAGATAAGGATA *
0 AGCTAA * NM:i:1
0 ATAGCTTCAGC *
0 TAGGC * NM:i:0
*
```

mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg19

HiSat2 Single-End Mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2

Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg19

HiSat2 Single-End Mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

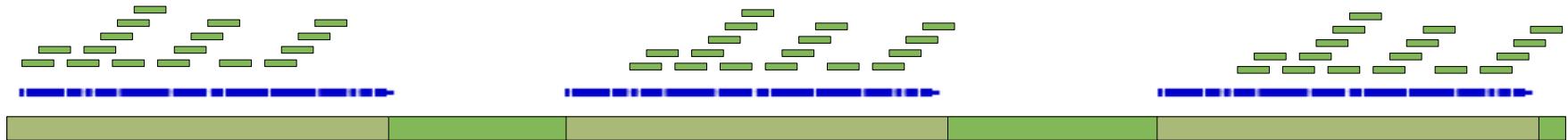
mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg19

HiSat2 Single-End Mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

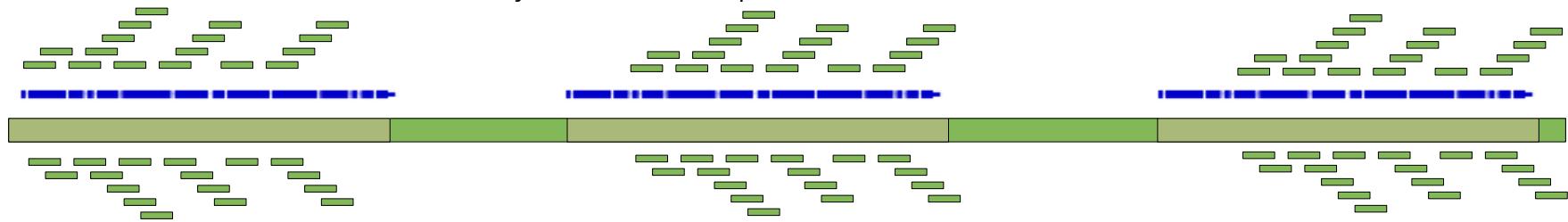
mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg19

HiSat2 Single-End Mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

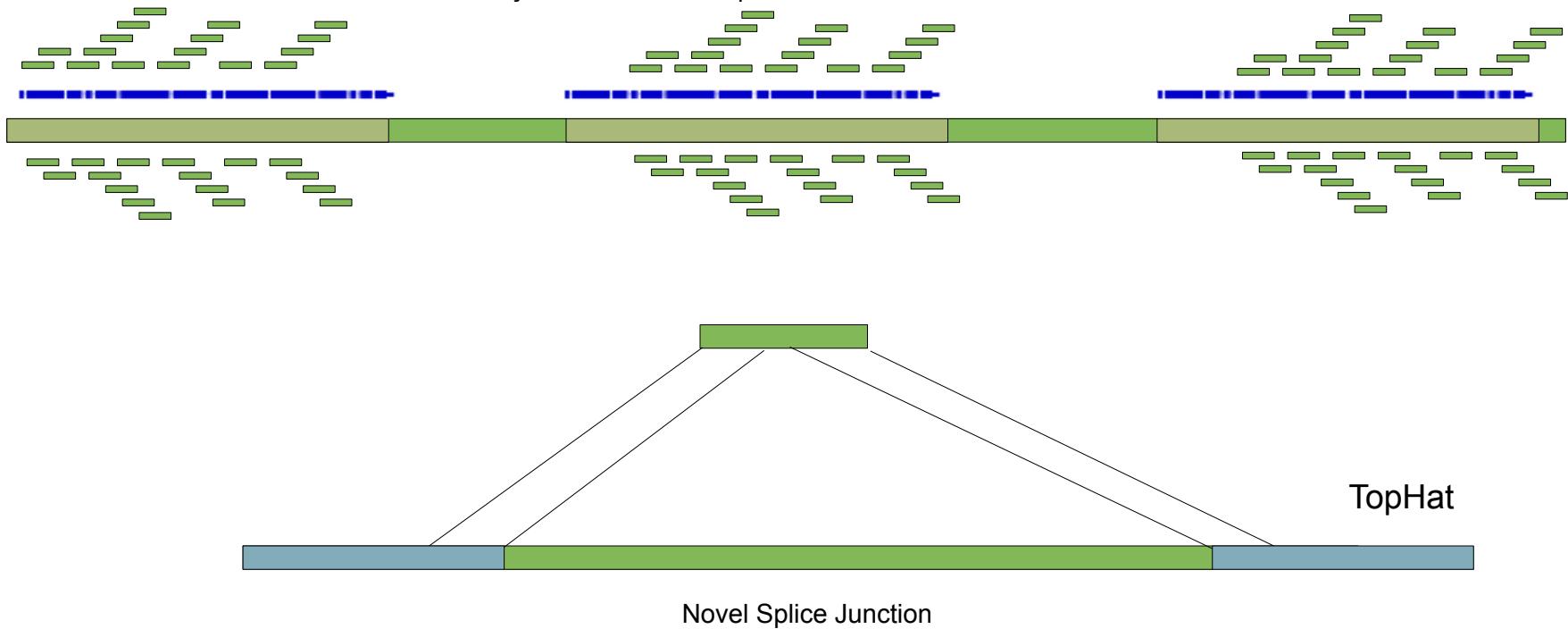
mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg19

HiSat2 Single-End Mode

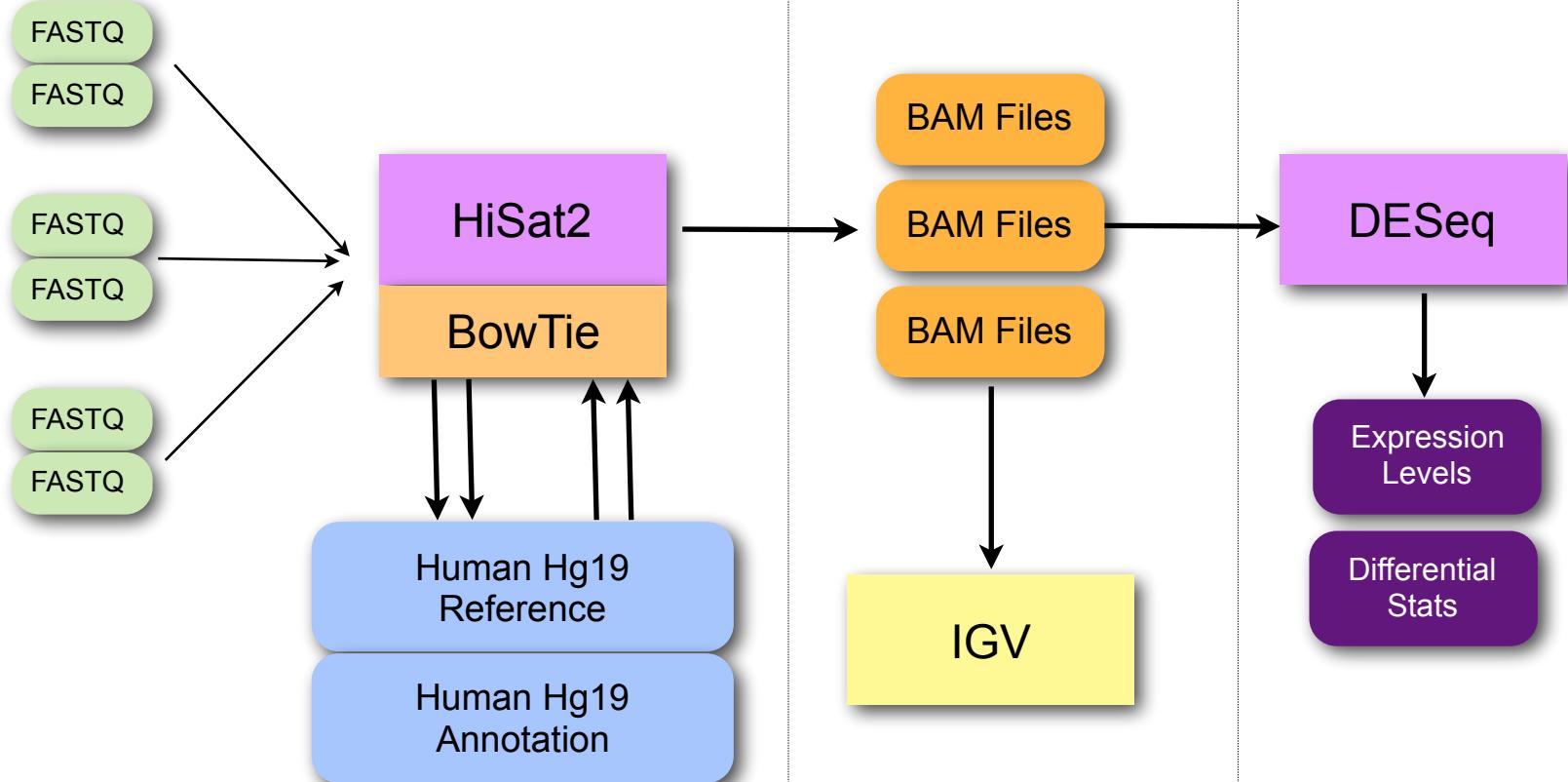
HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2

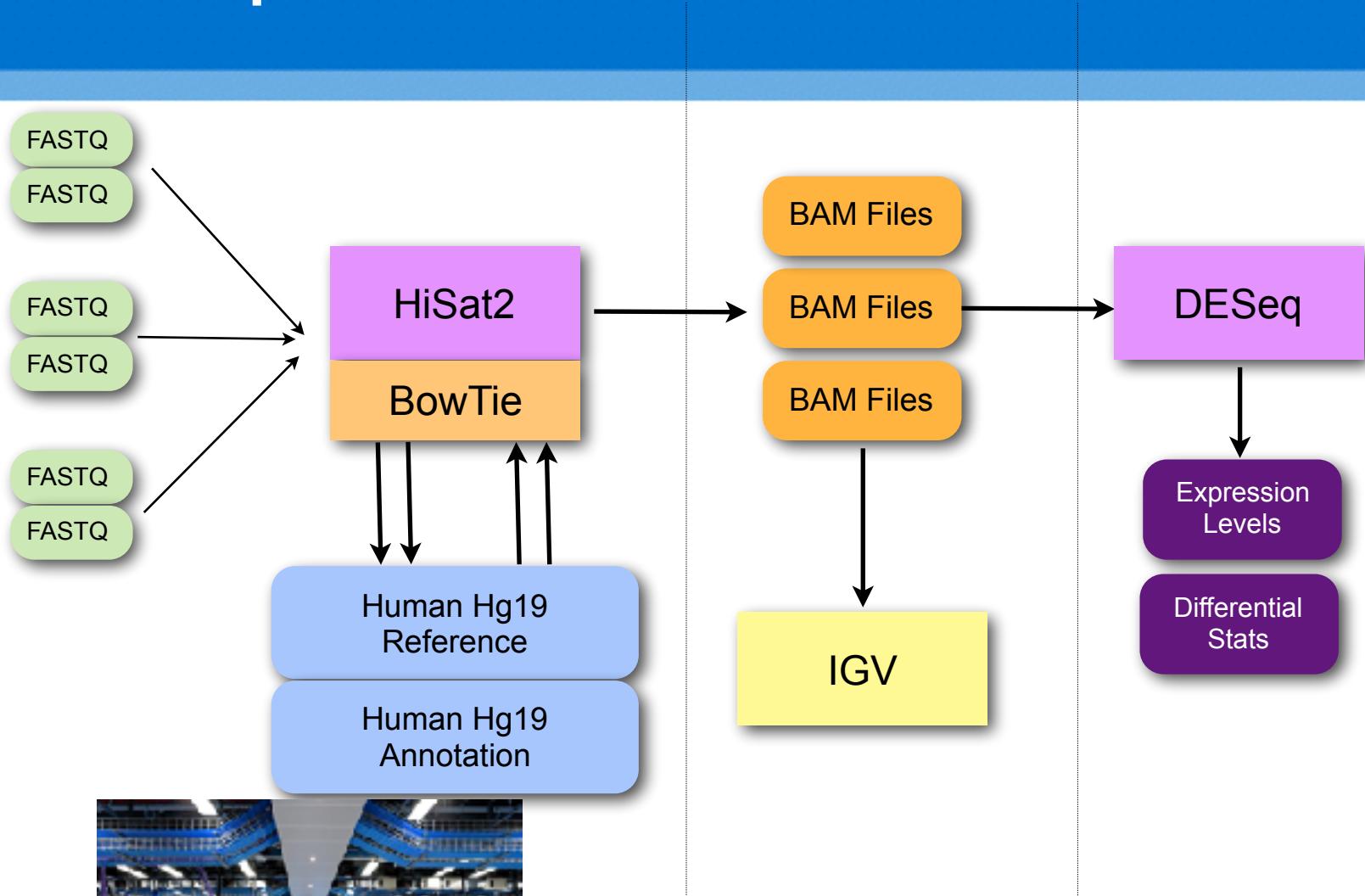


Results in count data and FPKMs

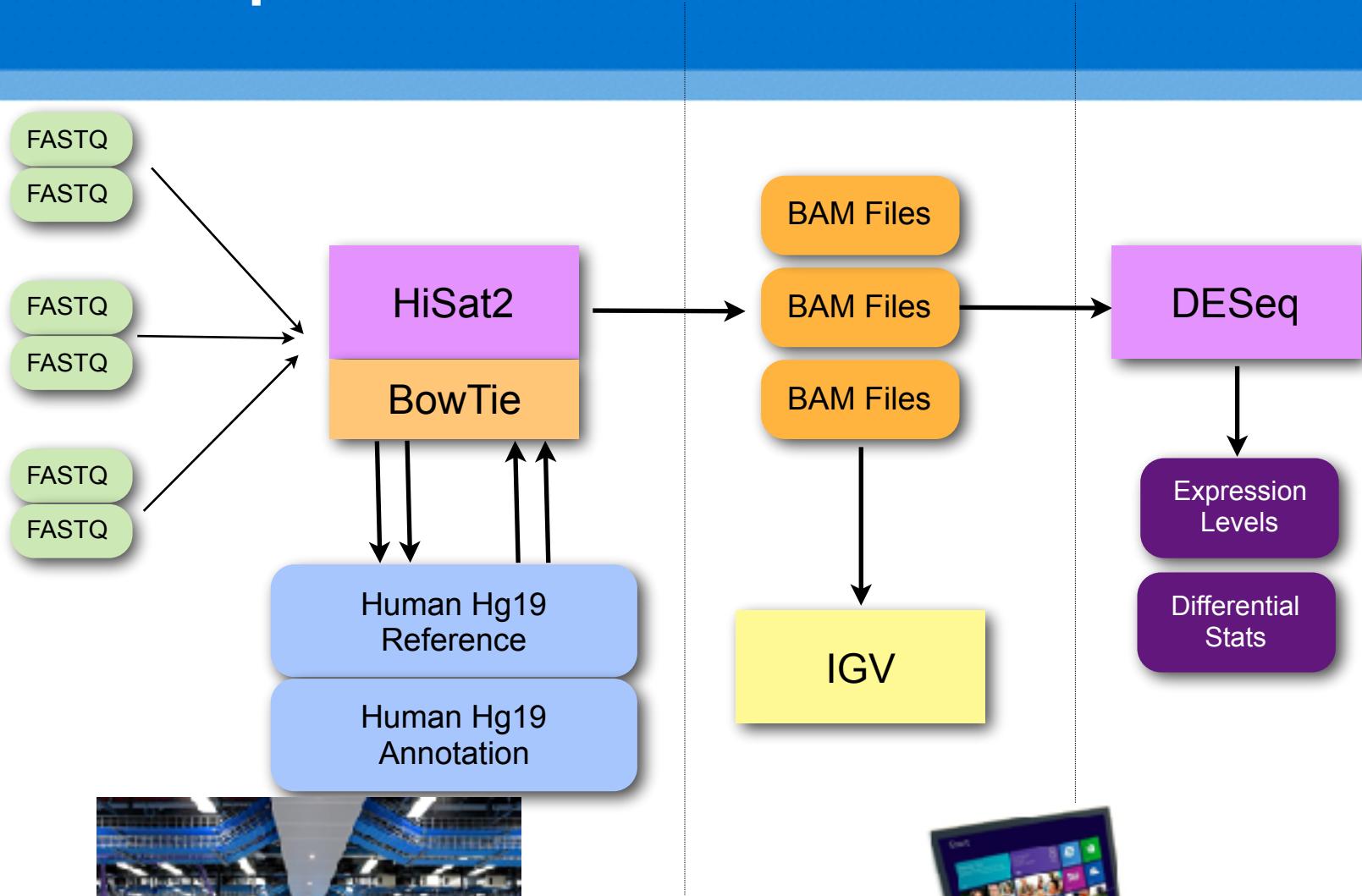
mRNA Seq Workflow



mRNA Seq Workflow

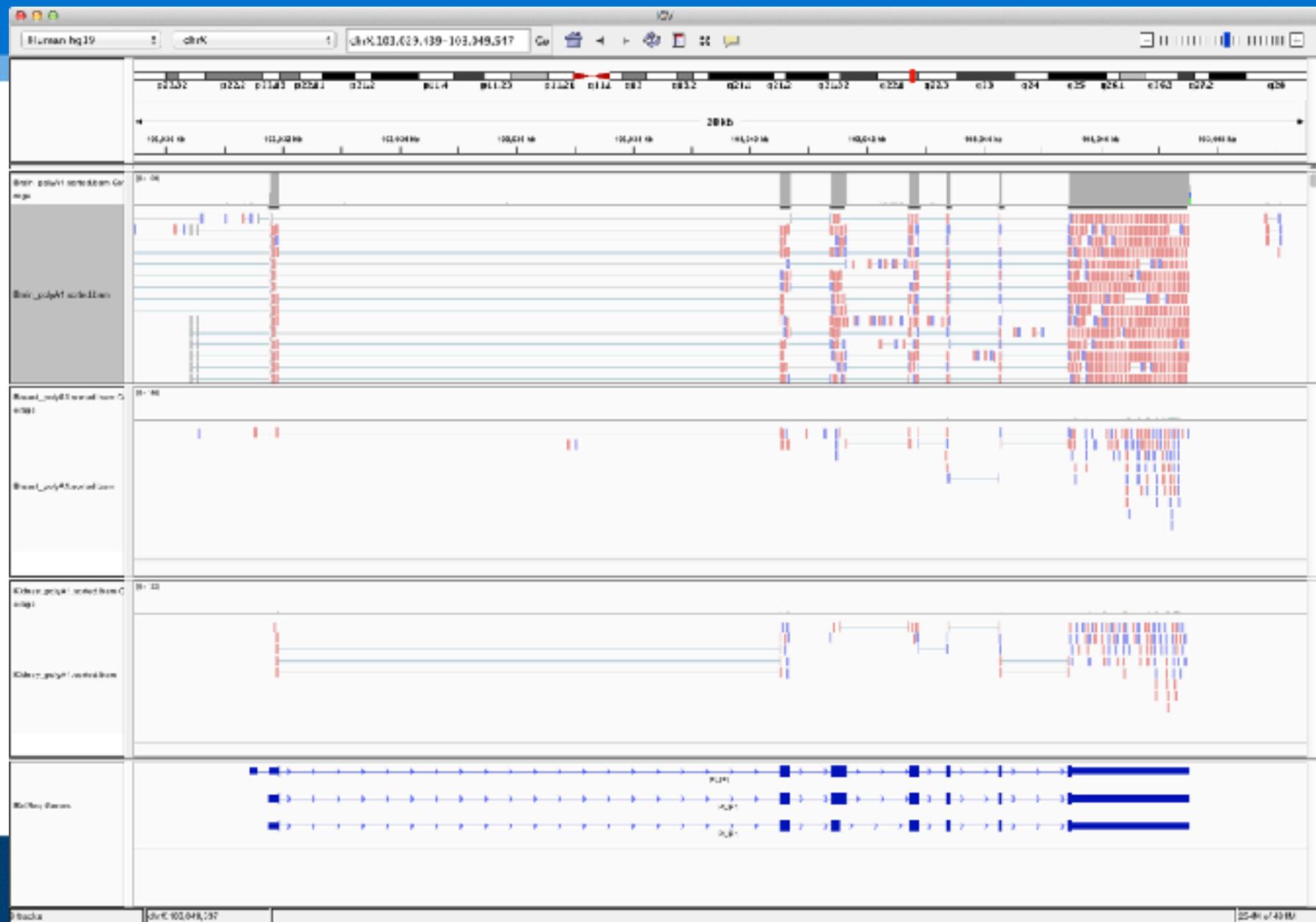


mRNA Seq Workflow



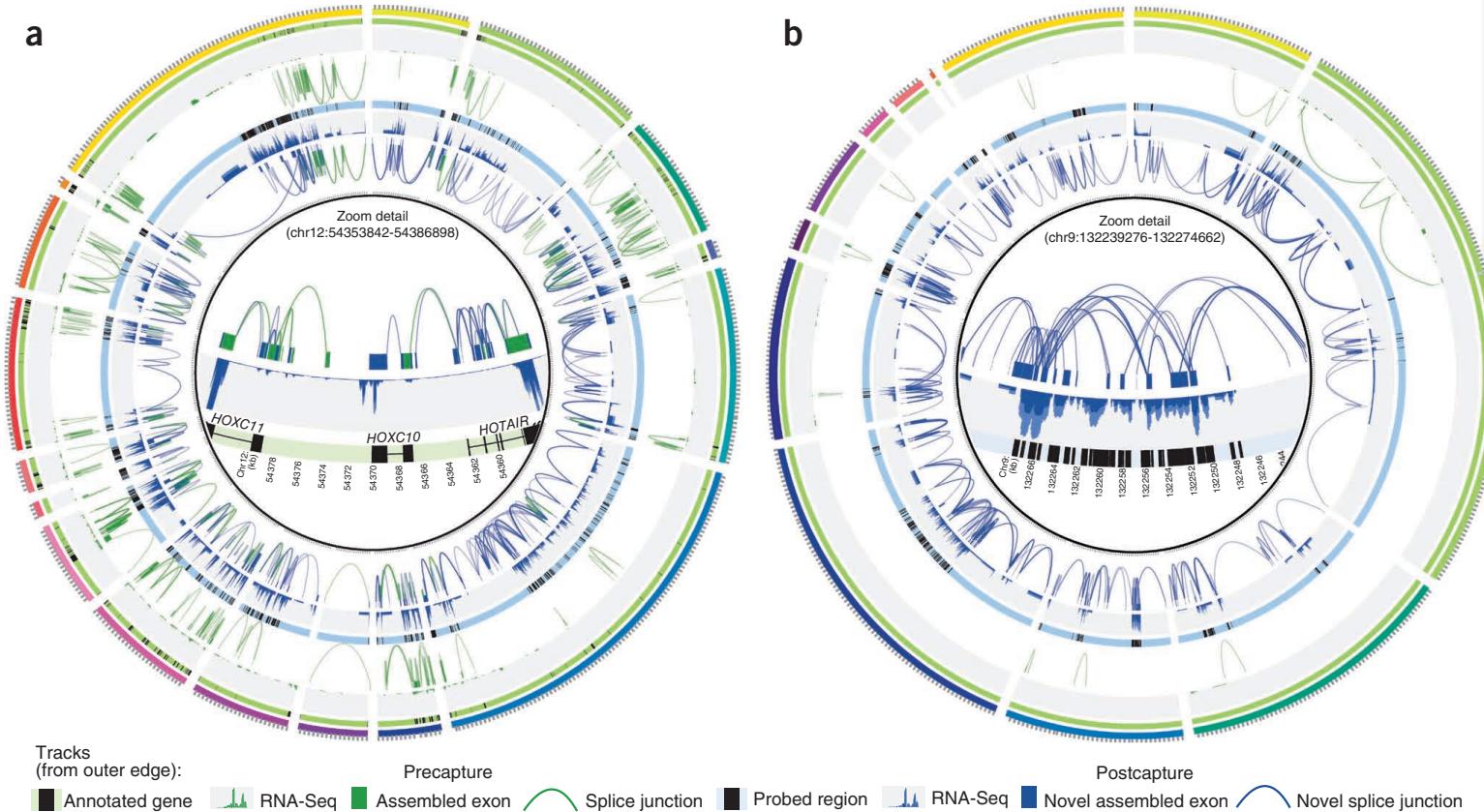
IGV

<http://www.broadinstitute.org/igv/>



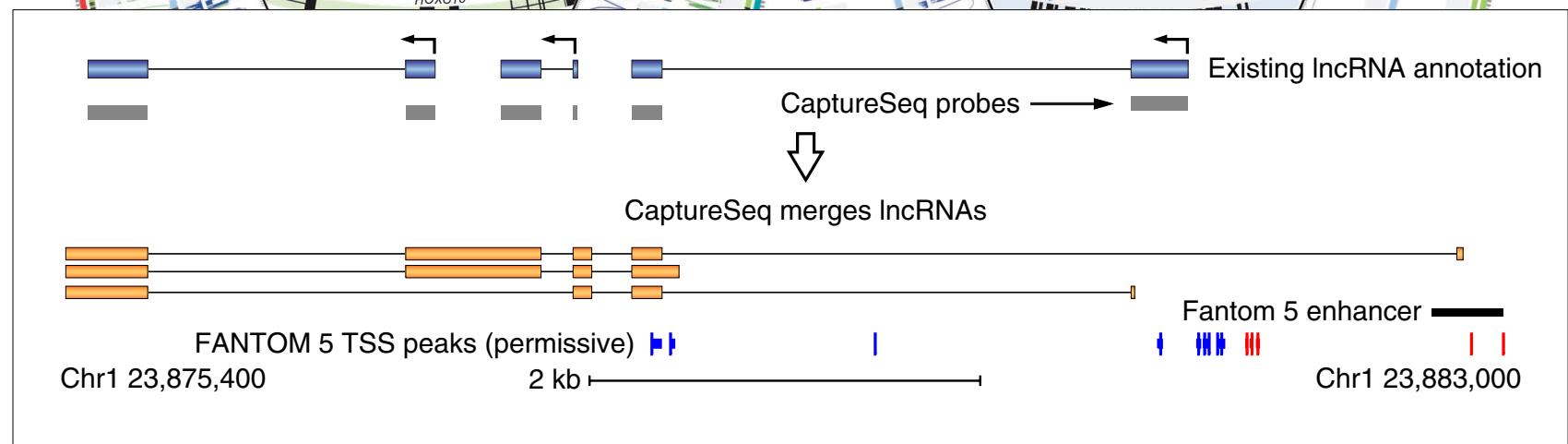
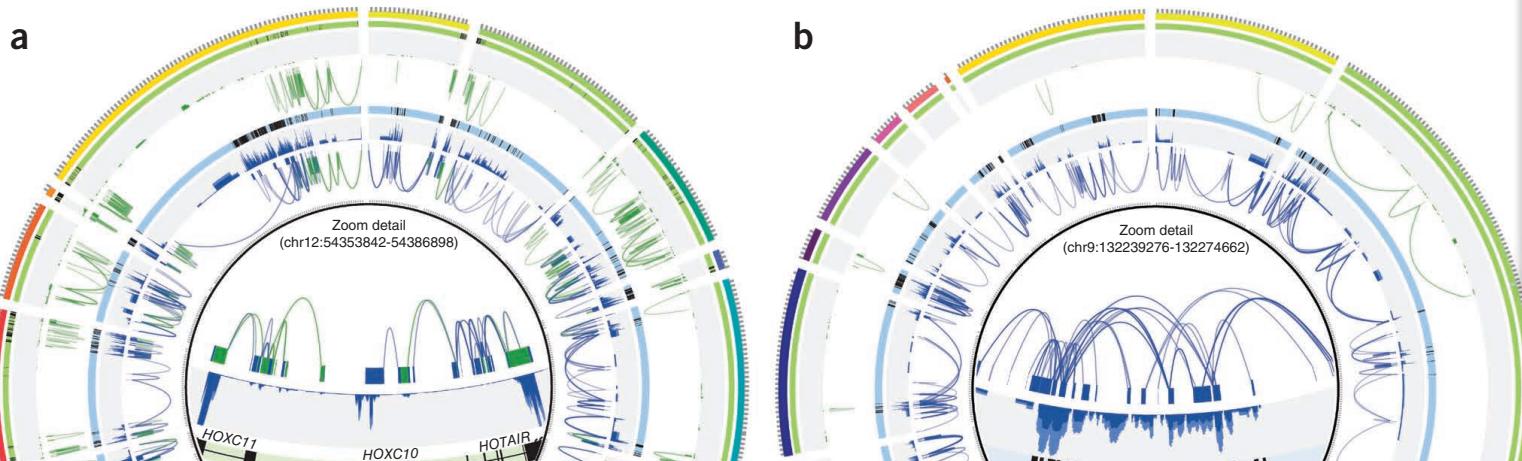
CaptureSeq

Improved definition of the mouse transcriptome via targeted RNA sequencing
G Bussotti, T Leonardi, MB Clark, TR Mercer, J Crawford, L Malquori, et al
Genome research 26 (5), 705-716 (2016)

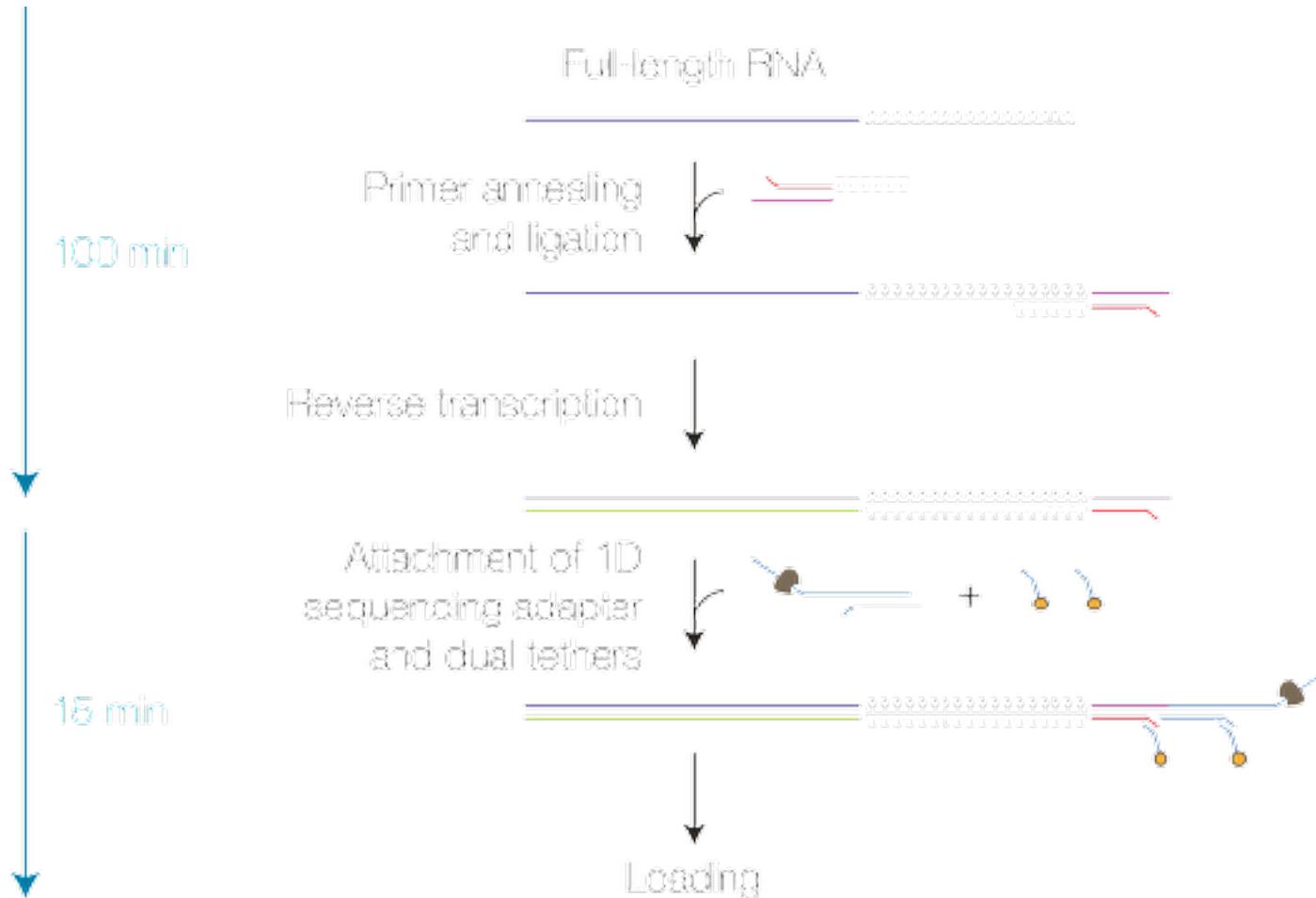


CaptureSeq

Improved definition of the mouse transcriptome via targeted RNA sequencing
G Bussotti, T Leonardi, MB Clark, TR Mercer, J Crawford, L Malquori, et al
Genome research 26 (5), 705-716 (2016)



Nanopore Direct RNA Sequencing Protocol



Nanopore Direct RNA Sequencing Protocol

- Input requirements
 - 200-600ng polyA+ RNA

Nanopore Direct RNA Sequencing Protocol

- **Input Samples**

- MCF7 PolyA+ (WTAC 2016)
- MCF7 PolyA+ (WTAC 2016)
- "Lucky Dip Mystery RNA" PolyA+
- "Lucky Dip Mystery RNA" PolyA+
- Nanopore Yeast Control
- Nanopore Yeast Control

PolyA+ Isolation

- **Trizol purification**
 - Frozen cell pellets
 - 200ul Trizol - Thaw and homogenise cell pellet (30-60secs)
 - Transfer to fresh eppendorf (lobind)
 - Add 800ul Trizol, vortex and stand at RT (5mins)
 - Add 200ul Chloroform, invert, vortex then spin at 14k rpm at 4C for 15 mins
 - Remove aqueous layer to new tube

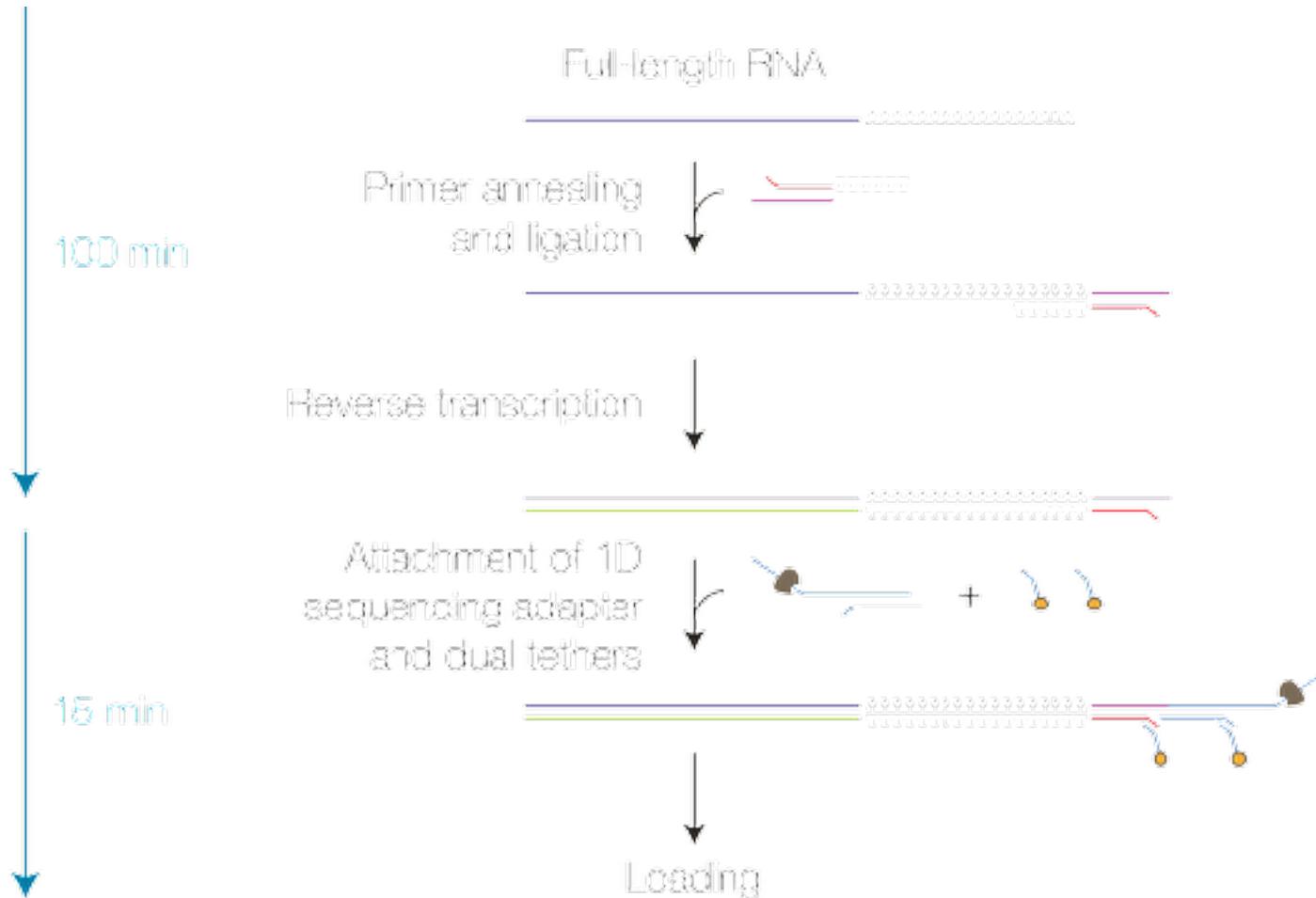
Nanopore Direct RNA Sequencing Protocol

- **DNAse treatment**
- **PolyA+ Enrichment (poly dT beads)**
 - NEB kit
 - Max input amount 5mg
- **Recovery of polyA+ RNA = 5-10% ?**
 - 1% in our hands this week

Nanopore Direct RNA Sequencing Protocol

- Huge amount of input polyA+ RNA required
- We needed 3x as many cells as we thought
- NEB polyA+ kit can deal with a maximum of 5mg of total RNA
 - Another kit may have been better

Nanopore Direct RNA Sequencing Protocol



Nanopore Direct RNA Sequencing Protocol

- **Library Preparation - 0.2ml PCR tube**
 - 3ul - Ligation Reaction Buffer (NEBNext)
 - 9ul - PolyA+ RNA sample
 - 0.5ul - Nanopore RNA standard
 - 1ul - Reverse Transcription Adapter (RTA)
 - 1.5ul - T4 Ligase (keep in freezer until needed)
- Incubate 10mins RT

Nanopore Direct RNA Sequencing Protocol

- **Make RT Master Mix**

- 9ul Nuclease free water
- 2.0ul dNTPs
- 8ul 5x First Strand buffer
- 4ul 0.1M DTT (Stabilises enzymes)
- Add the Master Mix to the Previous sample and mix (pipette)
 - Add 2ul Superscript III RT enzyme
 - Cycle for 50C for 50 mins -> Then 70C for 10mins -> Then to 4C
 - Transfer to a new fresh 1.5ml DNA lo-bind tube

Nanopore Direct RNA Sequencing Protocol

- **Library cleanup**

- Agencore RNAClean XP beads - 72ul
- Rotating Mixer - 5 mins Room Temp (RT)
- Spin Down (Brief spin on microcentrifuge)
- Magnetic Rack
 - Wash with fresh, nuclease free 70% ethanol - Do NOT disturb pellet Wash by rotation 180 degrees on the rack
 - Remove ethanol
 - Brief spin and remove residual ethanol (do not disturb pellet)
 - Resuspend on 20ul Nuclease free water - 5 minute incubation RT
 - Pellet on magnet until eluate is clean and colourless
 - Take 20ul into a new lo-bind tube



Nanopore Direct RNA Sequencing Protocol

- **Final Steps**

- 20ul RT RNA from last step
- 8ul Ligation buffer (NEB NEXT)
- 6ul RNA adapter (RMX)
- 3ul Nuclease free water
- 3ul T4 DNA ligase
- Mix (pipette) and incubate 10 mins RT
- Add 40ul RNAClean XP beads
- Rotator mixer - 5mins RT
- Final cleanup - **Performed Twice**
 - Pellet on magnetic rack and remove supernatant carefully
 - Wash with 150ul of Wash buffer - Flick tube to resuspend beads
 - Back on Rack to pellet

Nanopore Direct RNA Sequencing Protocol

- **Elution**
 - 21ul Elution buffer
 - Incubate 10 mins RT
 - Final Pellet on magnetic rack
 - Remove and retain 21ul of eluate into a new clean 1.5ml lo-bind tube

Nanopore Direct RNA Sequencing Protocol

- **Flow Cell**
 - **Draw Back Buffer**
 - Set pipette (P1000) to 200ul and insert tip into priming port
 - Turn wheel of pipette to 220-230 slowly until some buffer is extracted, do not remove too much or flow cell will be damaged.
 - **Priming Flow Cell**
 - 600 RRB
 - 600 nuclease free water
 - Load 800ul via priming port slowly to avoid bubbles

Nanopore Direct RNA Sequencing Protocol

- **Priming Flow Cell**
- **Load 200ul of the priming mix into the SpotOn port slowly and carefully (avoid bubbles)**
- **Sample Mix**
 - 37.5ul RBB
 - 37.5ul RNase free water
 - Load dropwise into the SpotOn port.

Informatics Wrap Up

- Course Image and course data is available soon
- The datasets are large: >20Gb
- All the practicals are available here:
 - www.tinyurl.com/wtac2016
- Many other courses and practicals available.

Using Sylamer

- Expert command-line version

- User supplied gene-list
- User supplied 3'UTRs

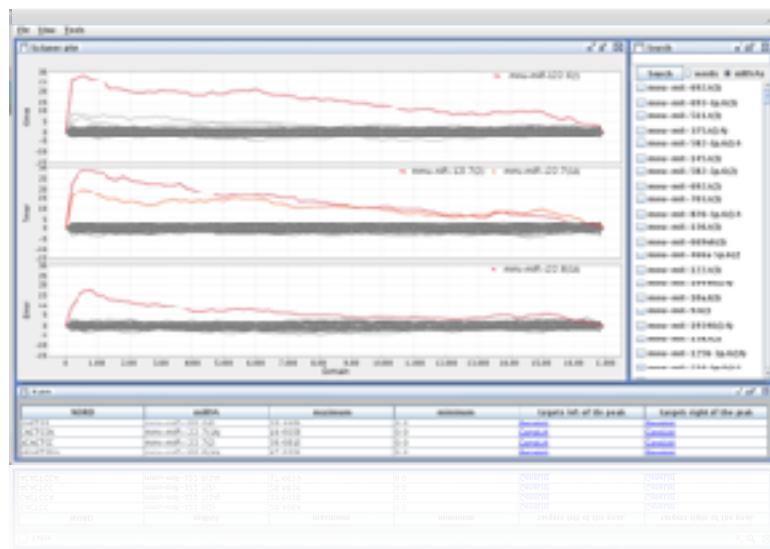
- Expert JAVA version

- User supplied genelist
- User supplied 3'UTRs

- Web-based version

- *SylArray*
- User supplied genelist
- www.ebi.ac.uk/enright-srv/sylarray

The screenshot shows the SylArray web server interface. At the top, there's a navigation bar with tabs for 'EMBL-EBI', 'SylArray', 'All Databases', 'Search Test Data', 'Tools', 'Help', and 'About Us'. Below the navigation is a sidebar with links for 'SylArray', 'Enright Group', 'Dr Enright', 'About SylArray', 'SylArray algorithm', 'Run New Analysis', 'miRNA-computational Targets', and 'miRNAs'. The main content area has a heading 'SYLARRAY' with a logo. It contains instructions: 'Web-server for automated detection of miRNA effects from expression data.', '1. Please paste your desired genelist into the box below.', and 'or upload a genelist.txt file [choose file] in this interface'. There's also a dropdown menu '2. Please select the type of identifier used in the inputfile' set to 'Chromosome identifier'. Below that is a 'Submit' button and a 'Clear all' button. To the right, there's a section for '3. Submit all data or clear submitted information.' with 'Submit' and 'Clear all' buttons. Further down are sections for '4. Advanced options.' and '5. ANALYSIS DETAILS.' with tabs for 'Seeds', 'ID List', and 'Patterns'.



Bartonicek, N. & Enright, A.J. SylArray: a web server for automated detection of miRNA effects from expression data. *Bioinformatics* 26, 2900–2901 (2010)

Computational Methods for smallRNA Analysis

Anton Enright

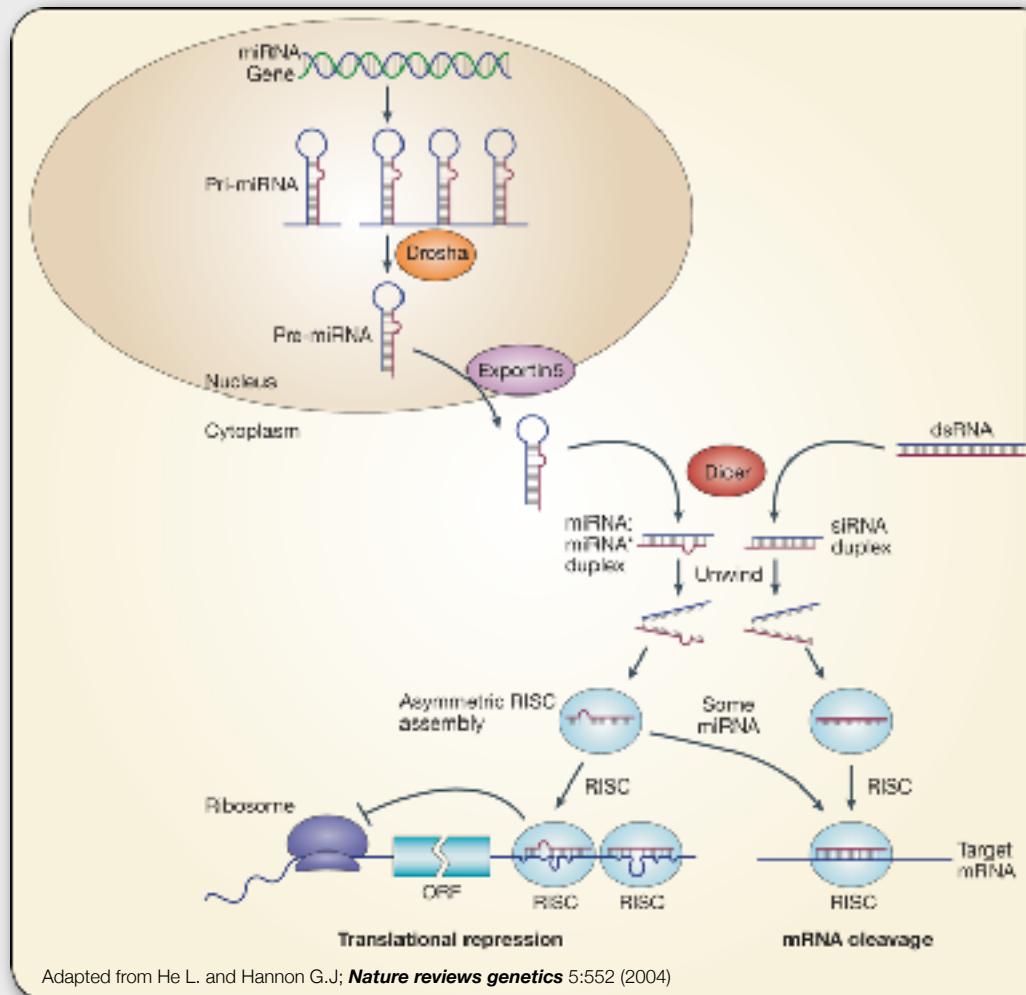
Group Leader

EMBL - European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridge, CB10 1SD, UK

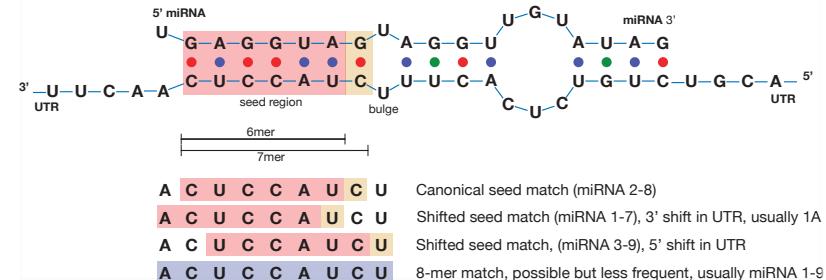
<http://www.ebi.ac.uk/enright/>
aje@ebi.ac.uk



microRNA methods development 2001-2016

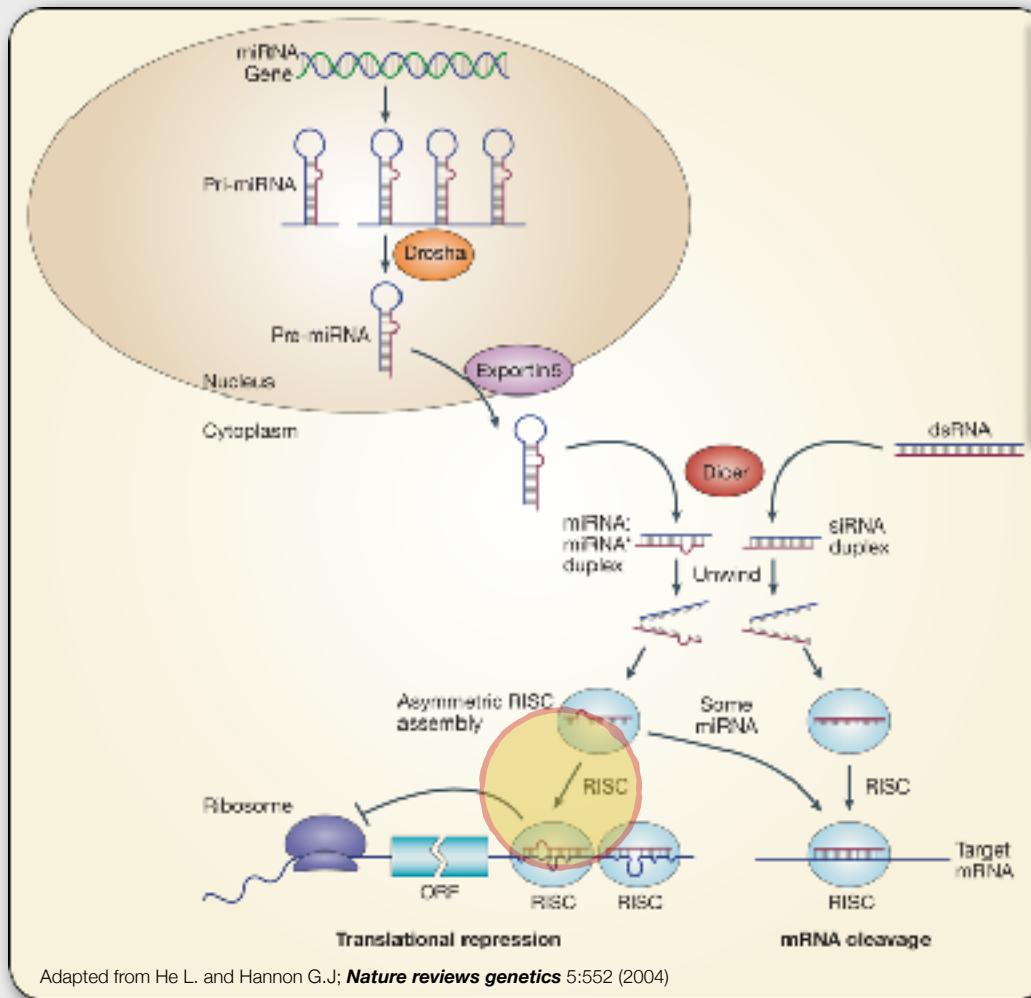


Anatomy of a microRNA target site

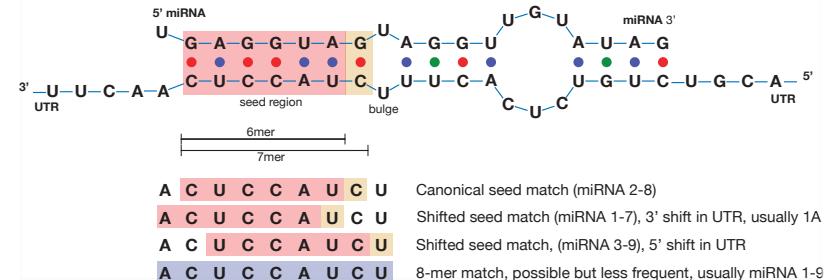


- **miRanda** - *de novo* target prediction
- **miRBase** - microRNA database
- **microcosm** - microRNA target database
- **Sylamer** - expression guided target prediction

microRNA methods development 2001-2016



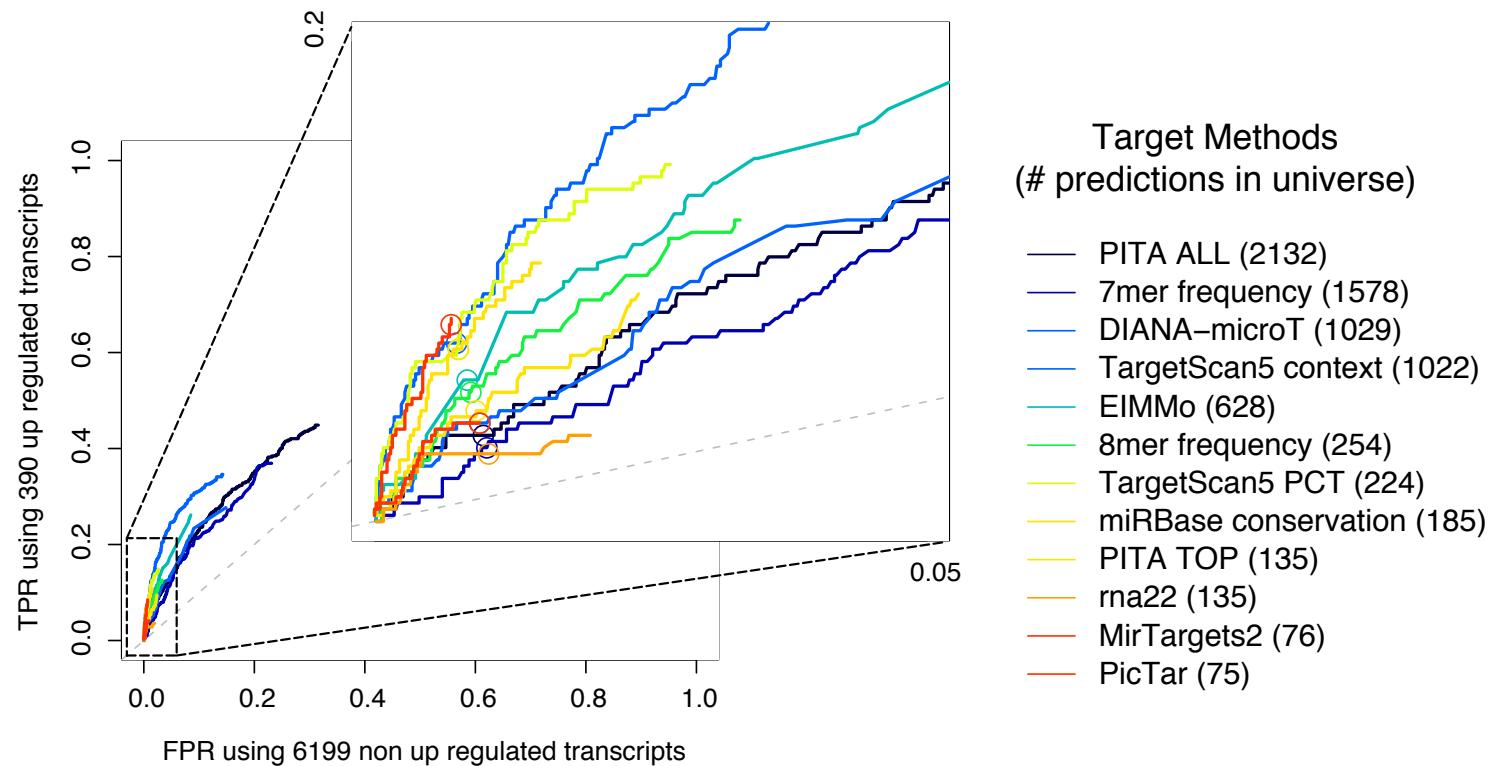
Anatomy of a microRNA target site



- **miRanda** - *de novo* target prediction
- **miRBase** - microRNA database
- **microcosm** - microRNA target database
- **Sylamer** - expression guided target prediction

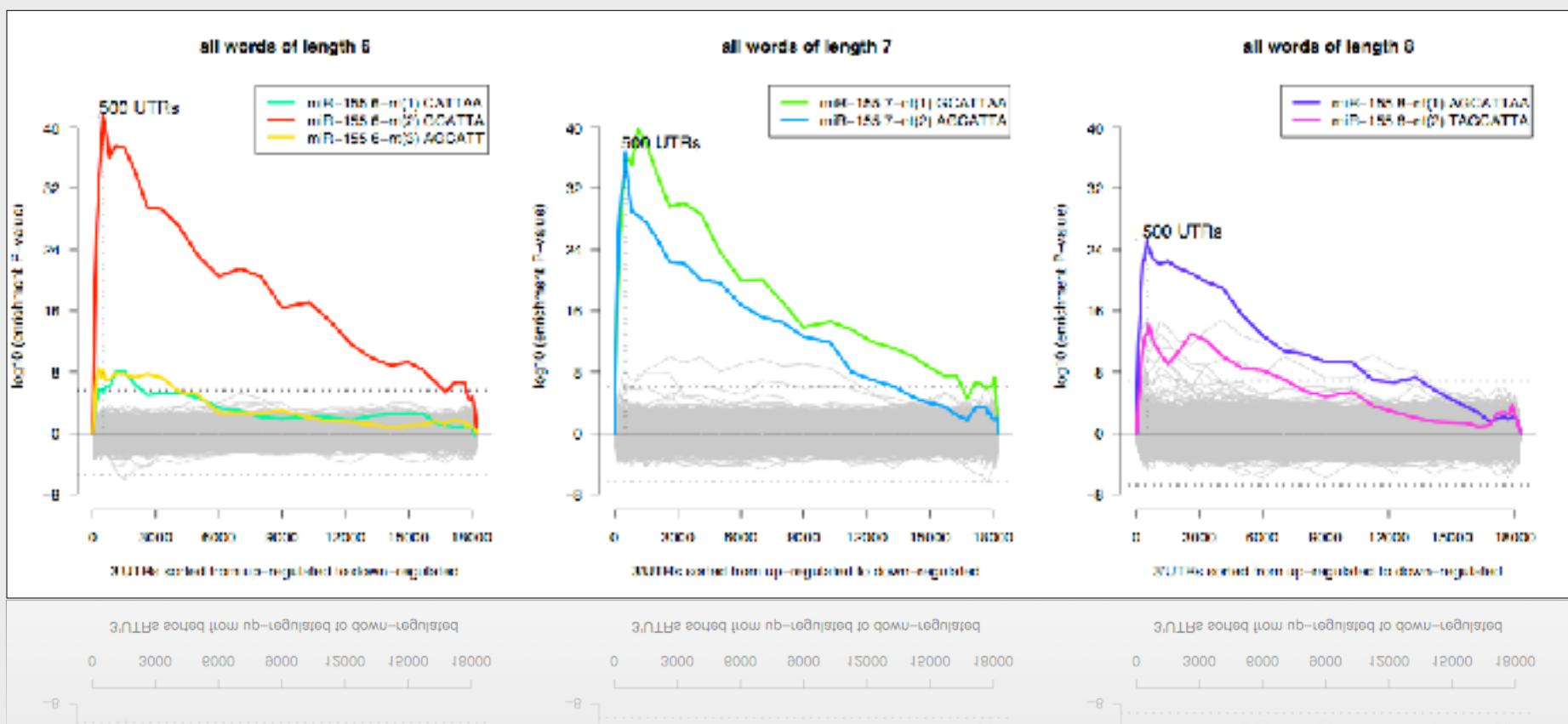
Comparison of Target Prediction Methods

10 mouse tissues large-scale validation



Systematic discovery of microRNA signatures from large-scale gene expression data
Abreu-Goodger C., van Dongen S, Enright A.J. (*In Preparation*)

Validation rate via Sylamer analysis > 75%



Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs.

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., van Dongen, S., Inoue, K., Enright, A. J., and Schier, A. F. (2006). *Science* 312, 75–79.

Fast assessment of miRNA binding and siRNA off-targets from Expression Data

van Dongen S., Abreu-Goodger C., Enright AJ. *Nature Methods* 2008

Sylamer - New Web Server - Sylarray2

<http://wwwdev.ebi.ac.uk/enright-dev/sylarray2/>

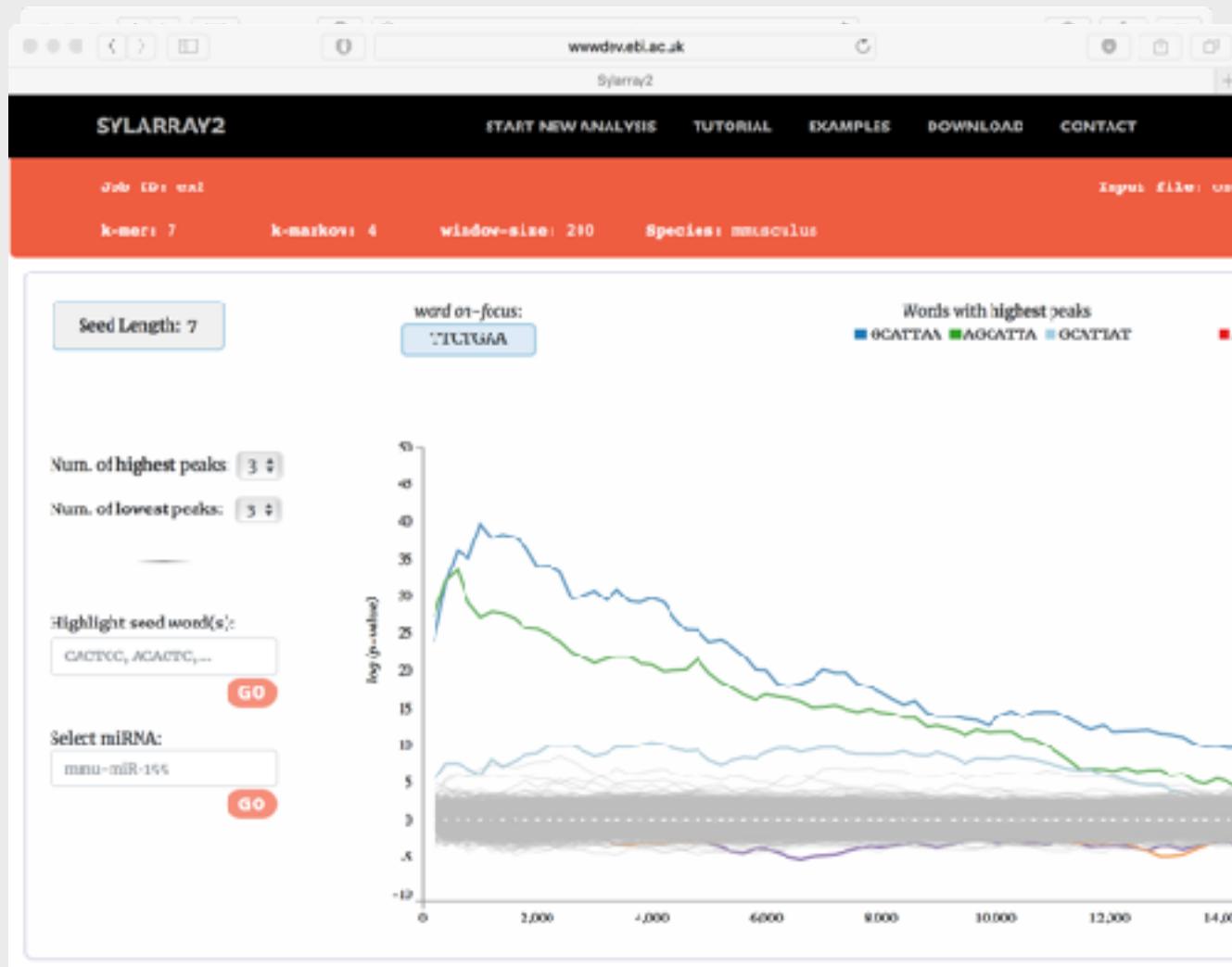
Sylamer - New Web Server - Sylarray2

The screenshot shows a web browser window with the URL wwwdev.ebi.ac.uk/enright-dev/sylarray2/ in the address bar. The page title is "SYLARRAY2". The main heading is "Start Analysis" with the subtitle "Detect miRNA targets from gene lists". A central input area contains the following text: "You can upload your input file by dragging it here or by clicking on the upload button." Below this is a blue "Upload gene list" button. To the left is an "Options" section and to the right is a "Load example" link. The "Options" section includes the following fields:

- Select reference UTRs from species:
- or upload a custom FASTA file with UTRs (.fa or .fa.gz - gzip compressed):
- Please make sure that the gene IDs in your gene list and custom UTR file are of the same type.
- Input Gene IDs:
- k-mer: k-markov: win-size:

<http://wwwdev.ebi.ac.uk/enright-dev/sylarray2/>

Sylamer - New Web Server - Sylarray2



<http://wwwdev.ebi.ac.uk/enright-dev/sylarray2/>

Sylamer - Single Sum Significance (SSS)

- Look at different types of motif that correspond to miRNA seed matches
 - Combine different flavours of non-canonical seeds

Sylamer - Single Sum Significance (SSS)

- Look at different types of motif that correspond to miRNA seed matches
 - Combine different flavours of non-canonical seeds

TGCATTA GCATTATGAGA
CGTAAT

Sylamer - Single Sum Significance (SSS)

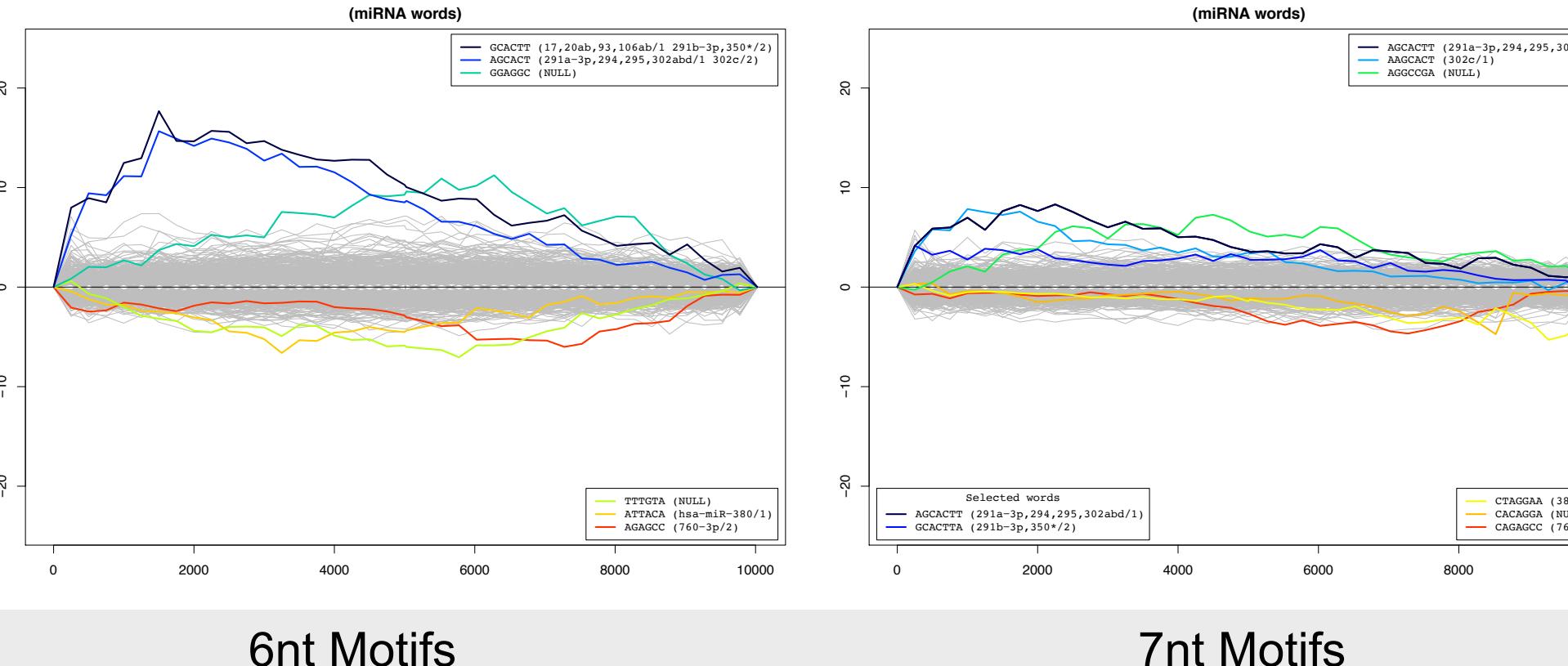
- Look at different types of motif that correspond to miRNA seed matches
 - Combine different flavours of non-canonical seeds

TGCATTA GCATTATGAGA
CGTAAT

ACGTAA
GTAATC
ACGTAAT
ACGTAATC

Single Sum Significance - miR302

Paediatric Germ Cell Tumours

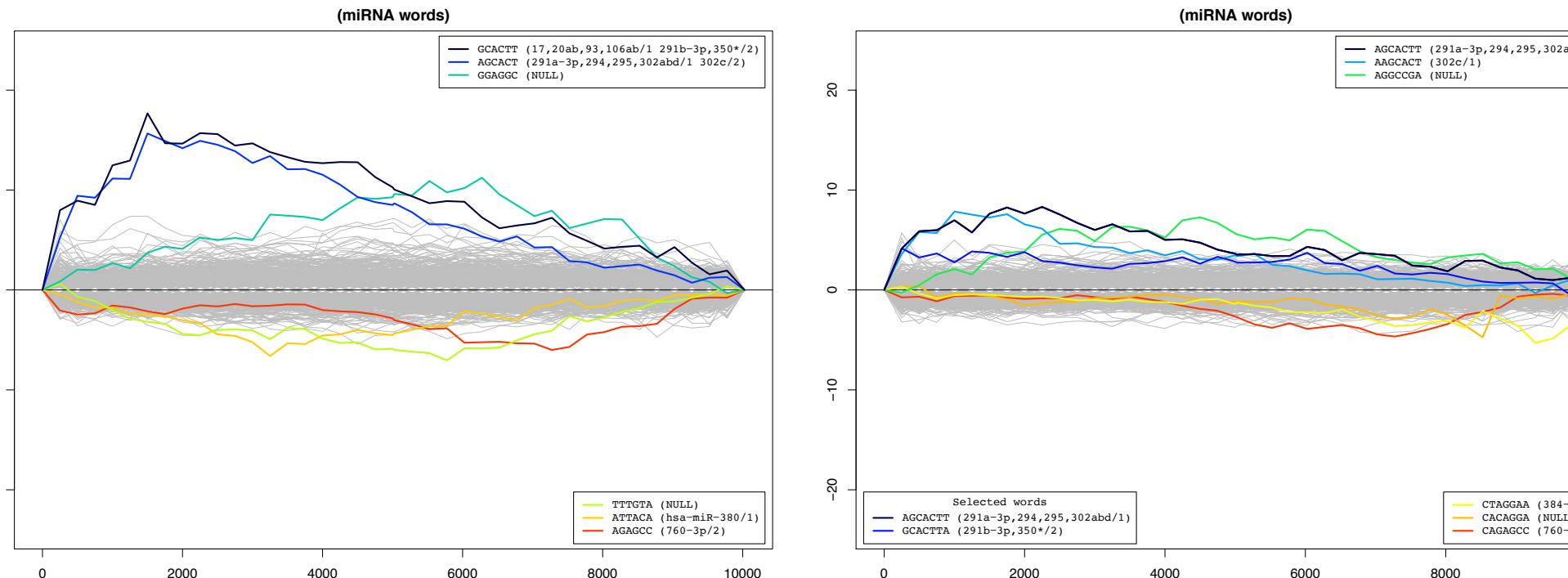


Single Sum Significance

- Good for cases with complicated microRNA responses or weaker signals

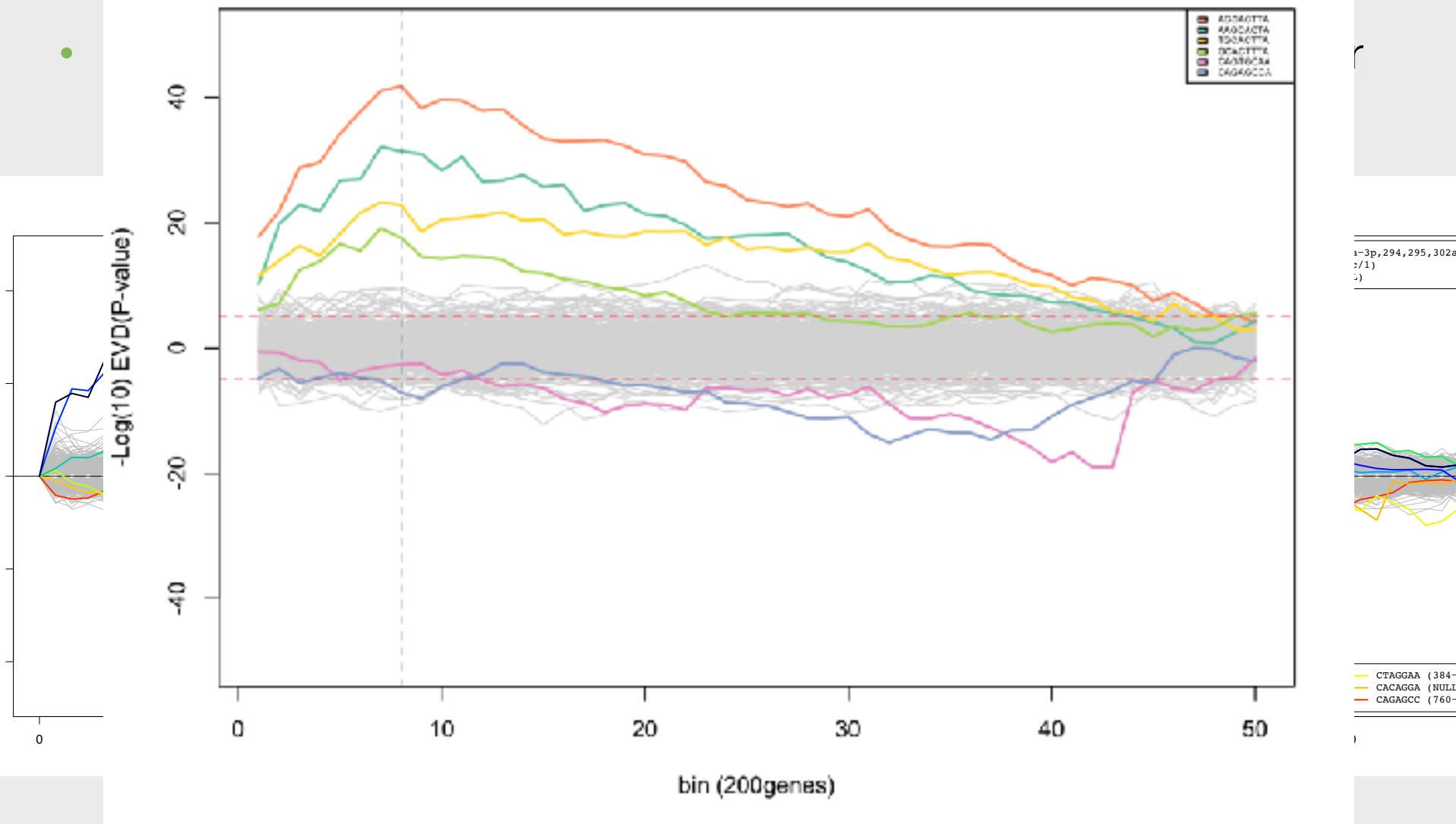
Single Sum Significance

- Good for cases with complicated microRNA responses or weaker signals



Si

Single-Sum Significance Sylamer



Analysis of microRNA Expression Data

Workflow: Raw miRNA NGS to Results

```
@HS24_10147:1:1101:1067:1989#0
GCCCGGCTAGCTCAGTCGGTAGAGCATGGACTGGAATTCTCGGGTGCN
+
BCCFFFFFHHHFHHJGHIIJIJJJIJH?FHHHEHJIGIJJJJHJJ!
@HS24_10147:1:1101:2218:1975#0
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACCTCCAN
+
BBCFFFFDHHHHHJJJJJJJJFJIIJJJJJJJJ@HIJIIHII!
@HS24_10147:1:1101:2722:1968#0
CGACCTGGAATTCTCGGGTGCCAAGGAACCTCCAGTCACCGATGTATCTCN
+
?@@FFFFFAHDAEHFHGIIGGIAE9DFHIGFFFGEIGFD=CCCCF!
@HS24_10147:1:1101:2977:1942#0
NCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACCTCCAN
+
!1BDFFDDHHHHIJJIIJJJJJJJJJJJJJJGHIJJJJJJJJ!
@HS24_10147:1:1101:5876:1978#0
TACAGTCCGACGATCTGGAATTCTCGGGTGCCAAGGCTCCAGTCACCGAN
+
=:D-DDDDDD<DFHIHGGHIIHGEFHGBFBDECHI<?@FGDFGCGDE!
@HS24_10147:1:1101:8742:1944#0
NCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACCTCCAN
+
```

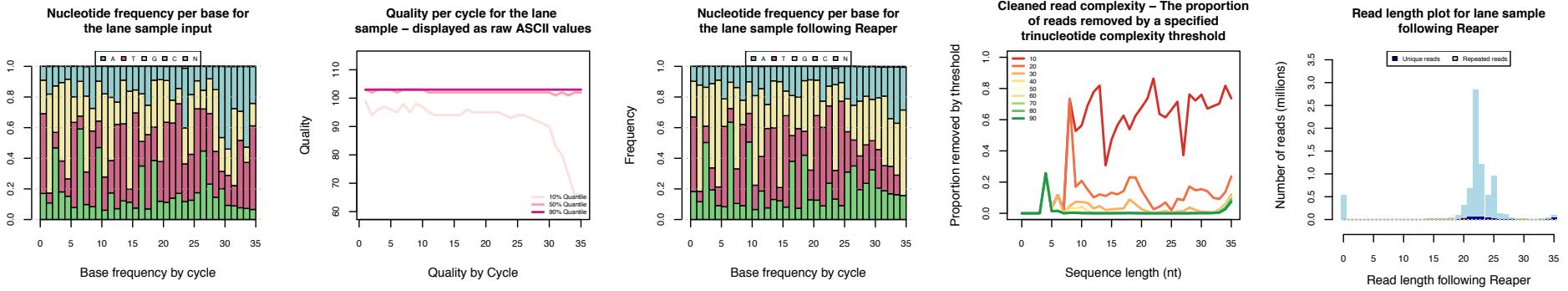
Workflow: Raw miRNA NGS to Results

```
@HS24_10147:1:1101:1067:1989#0
GCCCGGCTAGCTCAGTCGGTAGAGCATGGACTGGAATTCTCGGGTGCN
+
BCCFFFFFHHHFHHJGHIIJIJJJIJH?FHHHEHJIGIJJJJJHJJ!
@HS24_10147:1:1101:2218:1975#0
TCGCTTGGTGCAGATCGGGACTTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
BBCFFFFDHHHHHJJJJJJJJFJIIJJJJJJJJ@HIJIIHII!
@HS24_10147:1:1101:2722:1968#0
CGACCTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCGATGTATCTCN
+
?@@@FFFFFAHDAEHFHGIIGGIAE9DFHIGFFFGEGIGFD=CCCCF!
@HS24_10147:1:1101:2977:1942#0
NCGCTTGGTGCAGATCGGGACTTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
!1BDFFDDHHHHIJJIIJJJJJJJJJJJJJJGHIJJJJJJJJ!
@HS24_10147:1:1101:5876:1978#0
TACAGTCCGACGATCTGGAATTCTCGGGTGCCAAGGCTCCAGTCACCGAN
+
=:==DDDDDD<DFHIHGGHIIHGEFHGBFBDECHI<?@FGDFGCGDE!
@HS24_10147:1:1101:8742:1944#0
NCGCTTGGTGCAGATCGGGACTTGGAATTCTCGGGTGCCAAGGAACTCCAN
+
```

Workflow: Raw miRNA NGS to Results

```
@HS24_10147:1:1101:1067:1989#0
GCCCGGCTA
+BCCFFFFF
@HS24_101
TCGCTTGGI
+BBCFFFFD
@HS24_101
CGACCTGGA
+?@@FFFFF
@HS24_101
NCGCTTGGT
+!1BDFFDDH
@HS24_101
TACAGTCCG
+=:=DDDDDD
@HS24_101
NCGCTTGGT
+
>1_t11_w33_x157998
GTTCGTTAGTAGTGTTATCAGTCGCCT
>2_t11_w32_x80579
GCATTGGTGGTCAGTGGTAGAATTCTGCCT
>3_t10_w36_x27074
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATT
>4_t10_w35_x24313
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAAT
>5_t11_w34_x22696
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAA
>6_t8_w20_x20049
TTCACAGTGGCTAACGTTCTG
>7_t11_w33_x17558
GCATTGGTGGTTCAGTGGTAGAATTCTGCCTG
>8_t8_w32_x16093
GCATTGGTGGTTCAGTGGTAGAATTCTGCCT
>9_t12_w31_x15490
GCATTGGTGGTTCAGTGGTAGAATTCTGCC
>10_t11_w33_x14783
TCCCTGGTGGTCTAGTGGTTAGGATTGGCGCT
>11_t8_w31_x14593
GCATTGGTGGTTCAGTGGTAGAATTCTGCCT
```

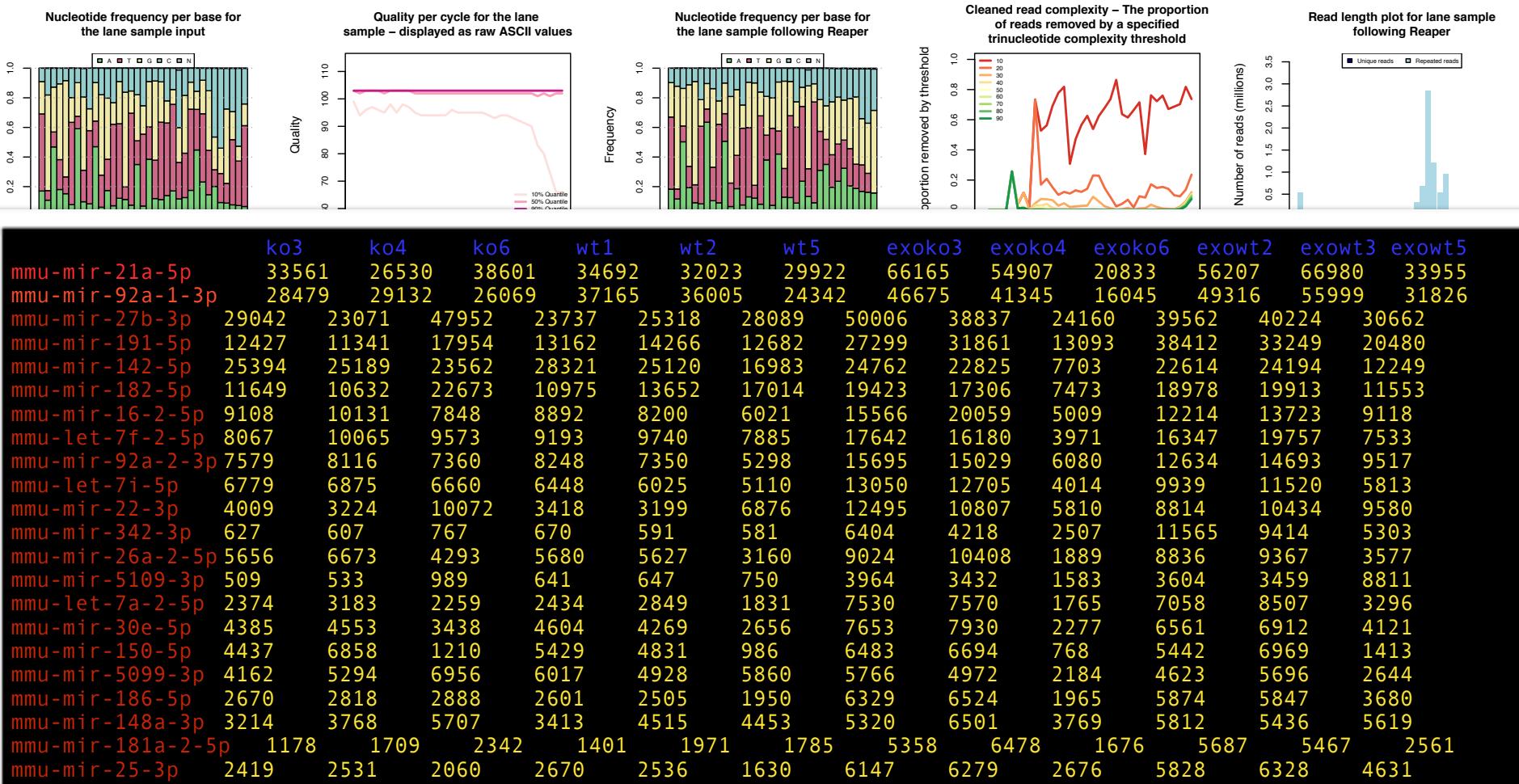
Workflow: Raw miRNA NGS to Results



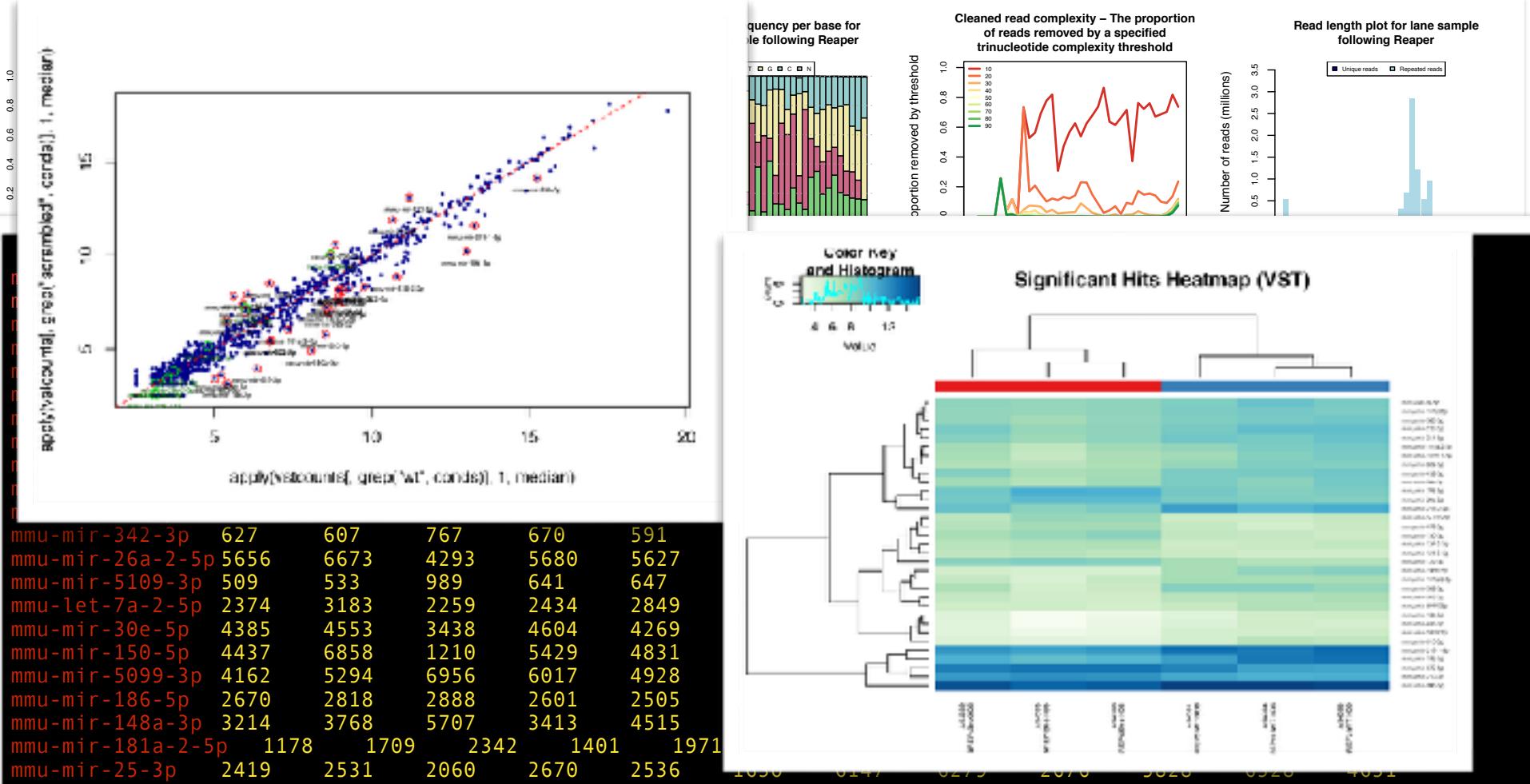
The visualization shows two side-by-side sequence files. The left file is raw data with various sequencing artifacts and low-quality bases represented by 'N' and 'D'. The right file is the same data after being processed by the Reaper quality filtering tool, where many of these artifacts have been removed, resulting in cleaner sequence data.

```
@HS24_101
NCGCTTGGT
+
!1BDFFDDH
@HS24_101
TACAGTCCG
+
=:DDDDDD
@HS24_101
NCGCTTGGT
+
>6_t8_w20_x20049
TTCACAGTGGCTAAGTTCTG
>7_t11_w33_x17558
GCATTGGTGGTTCAAGTGGTAGAATTCTGCCCTG
>8_t8_w32_x16093
GCATTGTGGTTCAAGTGGTAGAATTCTGCCCT
>9_t12_w31_x15490
GCATTGGTGGTTCAAGTGGTAGAATTCTGCC
>10_t11_w33_x14783
TCCCTGGTGGTCTAGTGGTTAGGATTGGCGCT
>11_t8_w31_x14593
GCATTGTGGTTCAAGTGGTAGAATTCTGCCCT
```

Workflow: Raw miRNA NGS to Results



Workflow: Raw miRNA NGS to Results



The Kraken Suite of Tools for smallRNA Analysis

- **Minion**

- De bruijn graph 3' assembly analysis for small RNA reads
- Can identify adapter sequences and large-scale contaminants

- **Reaper**

- Extremely fast and accurate adapter finder and trimmer
- Also cleans
 - Low-complexity tracts
 - PolyA, PolyN
 - Low-scoring tracts
- Deals with complex read geometries and random seqs

- **Tally**

- Collapses redundant reads together rapidly

The Kraken Suite of Tools for smallRNA Analysis

- **Minion**

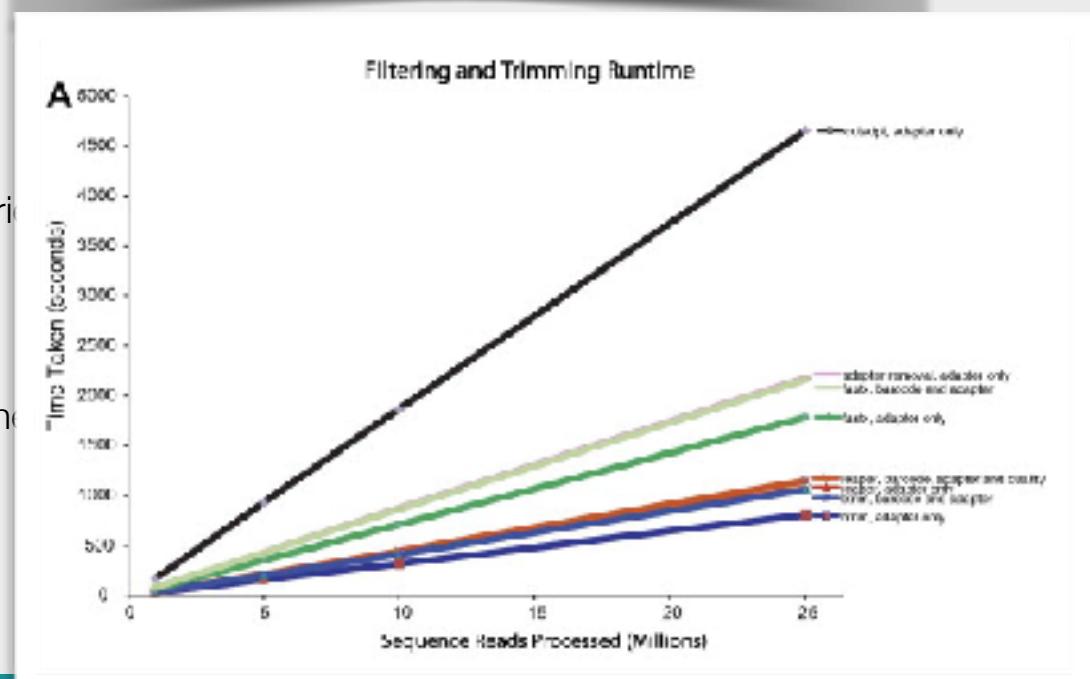
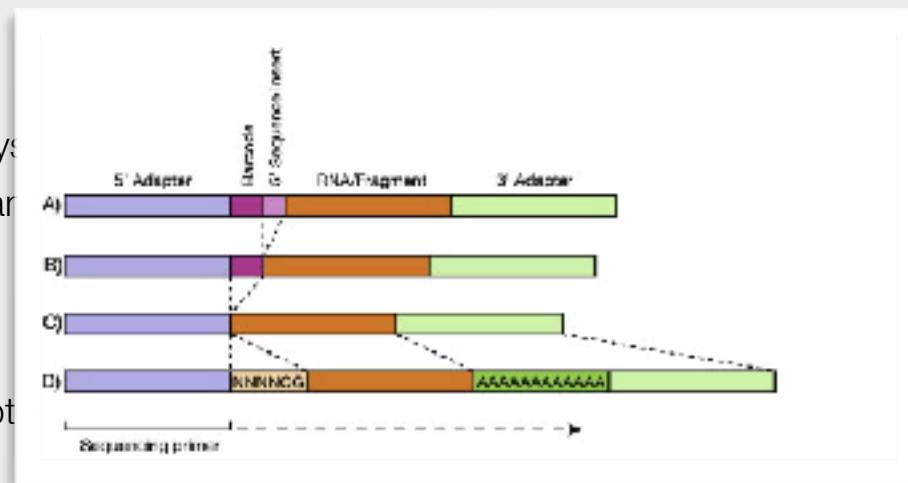
- De bruijn graph 3' assembly analysis
- Can identify adapter sequences and

- **Reaper**

- Extremely fast and accurate adapter removal
- Also cleans
 - Low-complexity tracts
 - PolyA, PolyN
 - Low-scoring tracts
- Deals with complex read geometries

- **Tally**

- Collapses redundant reads together



The Kraken Suite of Tools for smallRNA Analysis

- **Minion**

- De bruijn graph 3' assembly analysis for small RNA reads
- Can identify adapter sequences and large-scale contaminants

- **Reaper**

- Extremely fast and accurate adapter finder and trimmer
- Also cleans
 - Low-complexity tracts
 - PolyA, PolyN
 - Low-scoring tracts
- Deals with complex read geometries and random seqs

- **Tally**

- Collapses redundant reads together rapidly

Turning Processed Reads into Counts on microRNAs

```
@HS24_10147:1:1101:1067:1989#0  
GCCCGGCTAACCTCACTGGCTACACCAATCCCACCTCCAAATTCTGGGCTGGCN  
+  
BCCFFI >1_t11_w33_x157998  
@HS24 GTTTCGTAGTGTAGTGTTATCACGTTGCCT  
TCGCTI >2_t11_w32_x80579  
+ GCATTGGTGGTTCAGTGGTAGAATTCTGCCT  
BBCFFI >3_t10_w36_x27074
```

	ko3	ko4	ko6	wt1	wt2	wt5	exoko3	exoko4	exoko6	exowt2	exowt3	exowt5
mmu-mir-21a-5p	33561	26530	38601	34692	32023	29922	66165	54907	20833	56207	66980	33955
mmu-mir-92a-1-3p	28479	29132	26069	37165	36005	24342	46675	41345	16045	49316	55999	31826
mmu-mir-27b-3p	29042	23071	47952	23737	25318	28089	50006	38837	24160	39562	40224	30662
mmu-mir-191-5p	12427	11341	17954	13162	14266	12682	27299	31861	13093	38412	33249	20480
mmu-mir-142-5p	25394	25189	23562	28321	25120	16983	24762	22825	7703	22614	24194	12249
mmu-mir-182-5p	11649	10632	22673	10975	13652	17014	19423	17306	7473	18978	19913	11553
mmu-mir-16-2-5p	9108	10131	7848	8892	8200	6021	15566	20059	5009	12214	13723	9118
mmu-let-7f-2-5p	8067	10065	9573	9193	9740	7885	17642	16180	3971	16347	19757	7533
mmu-mir-92a-2-3p	7579	8116	7360	8248	7350	5298	15695	15029	6080	12634	14693	9517
mmu-let-7i-5p	6779	6875	6660	6448	6025	5110	13050	12705	4014	9939	11520	5813
mmu-mir-22-3p	4009	3224	10072	3418	3199	6876	12495	10807	5810	8814	10434	9580
mmu-mir-342-3p	627	607	767	670	591	581	6404	4218	2507	11565	9414	5303
mmu-mir-26a-2-5p	5656	6673	4293	5680	5627	3160	9024	10408	1889	8836	9367	3577
mmu-mir-5109-3p	509	533	989	641	647	750	3964	3432	1583	3604	3459	8811
mmu-let-7a-2-5p	2374	3183	2259	2434	2849	1831	7530	7570	1765	7058	8507	3296
mmu-mir-30e-5p	4385	4553	3438	4604	4269	2656	7653	7930	2277	6561	6912	4121
mmu-mir-150-5p	4437	6858	1210	5429	4831	986	6483	6694	768	5442	6969	1413
mmu-mir-5099-3p	4162	5294	6956	6017	4928	5860	5766	4972	2184	4623	5696	2644
mmu-mir-186-5p	2670	2818	2888	2601	2505	1950	6329	6524	1965	5874	5847	3680
mmu-mir-148a-3p	3214	3768	5707	3413	4515	4453	5320	6501	3769	5812	5436	5619
mmu-mir-181a-2-5p	1178	1709	2342	1401	1971	1785	5358	6478	1676	5687	5467	2561
mmu-mir-25-3p	2419	2531	2060	2670	2536	1630	6147	6279	2676	5828	6328	4631

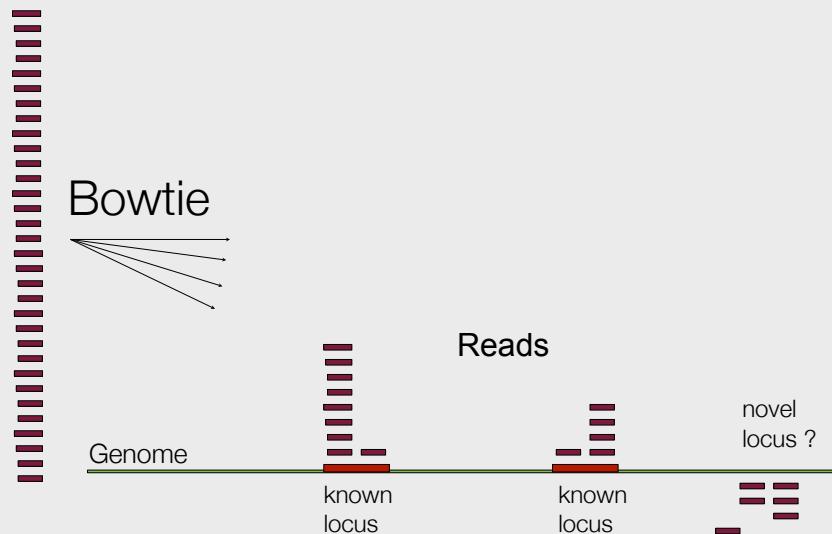
Mapping small RNA Reads

- **Genome Based Approach**
- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours

Mapping small RNA Reads

- **Genome Based Approach**

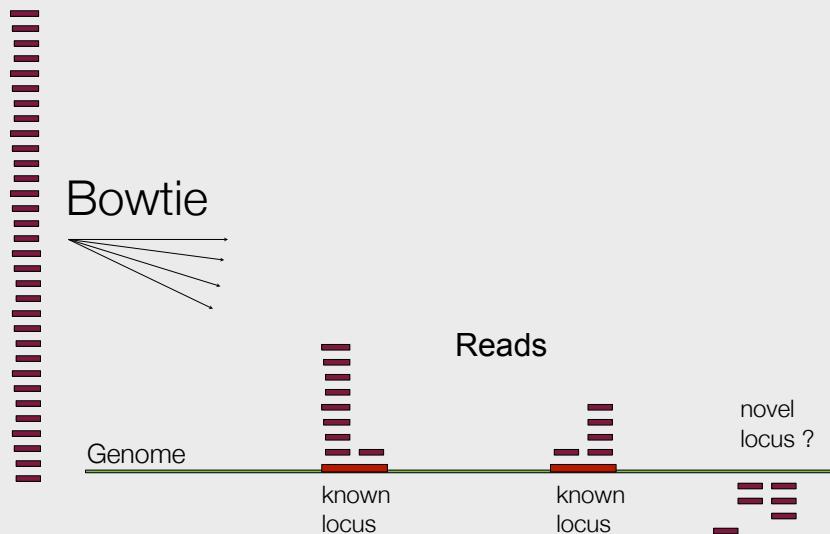
- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours



Mapping small RNA Reads

- **Genome Based Approach**

- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours



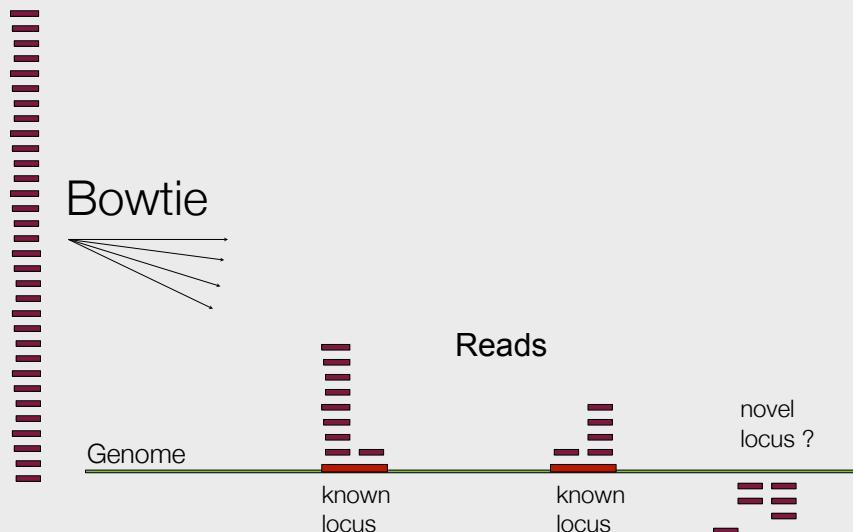
- **Read versus Precursor Approach**

- Compare reads directly against miRBase sequences
 - Fast and accurate
 - look for >95% identity and no more than 1-2 mismatches
 - Runtime: 10-15 minutes

Mapping small RNA Reads

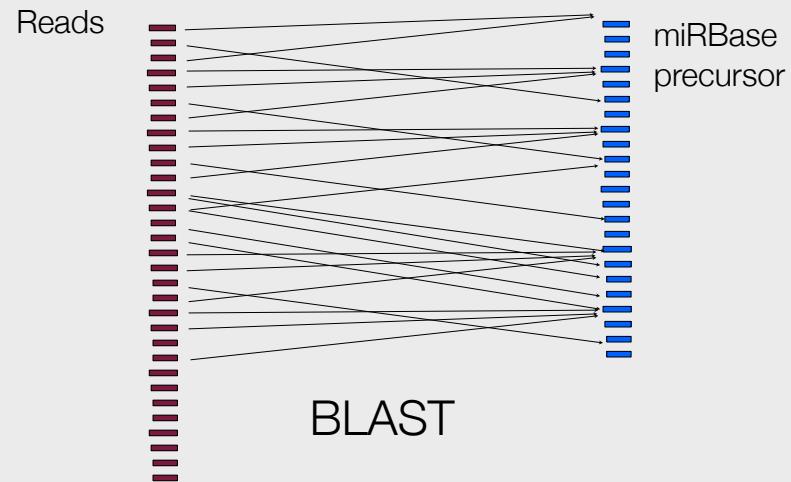
- **Genome Based Approach**

- Map all reads to the genome with an aligner tool
 - Bowtie: <http://bowtie-bio.sourceforge.net>
 - Select reads that overlap known miRNA locii
 - Problems resolving depth across loci
 - Runtime: 4-12 hours



- **Read versus Precursor Approach**

- Compare reads directly against miRBase sequences
 - Fast and accurate
 - look for >95% identity and no more than 1-2 mismatches
 - Runtime: 10-15 minutes



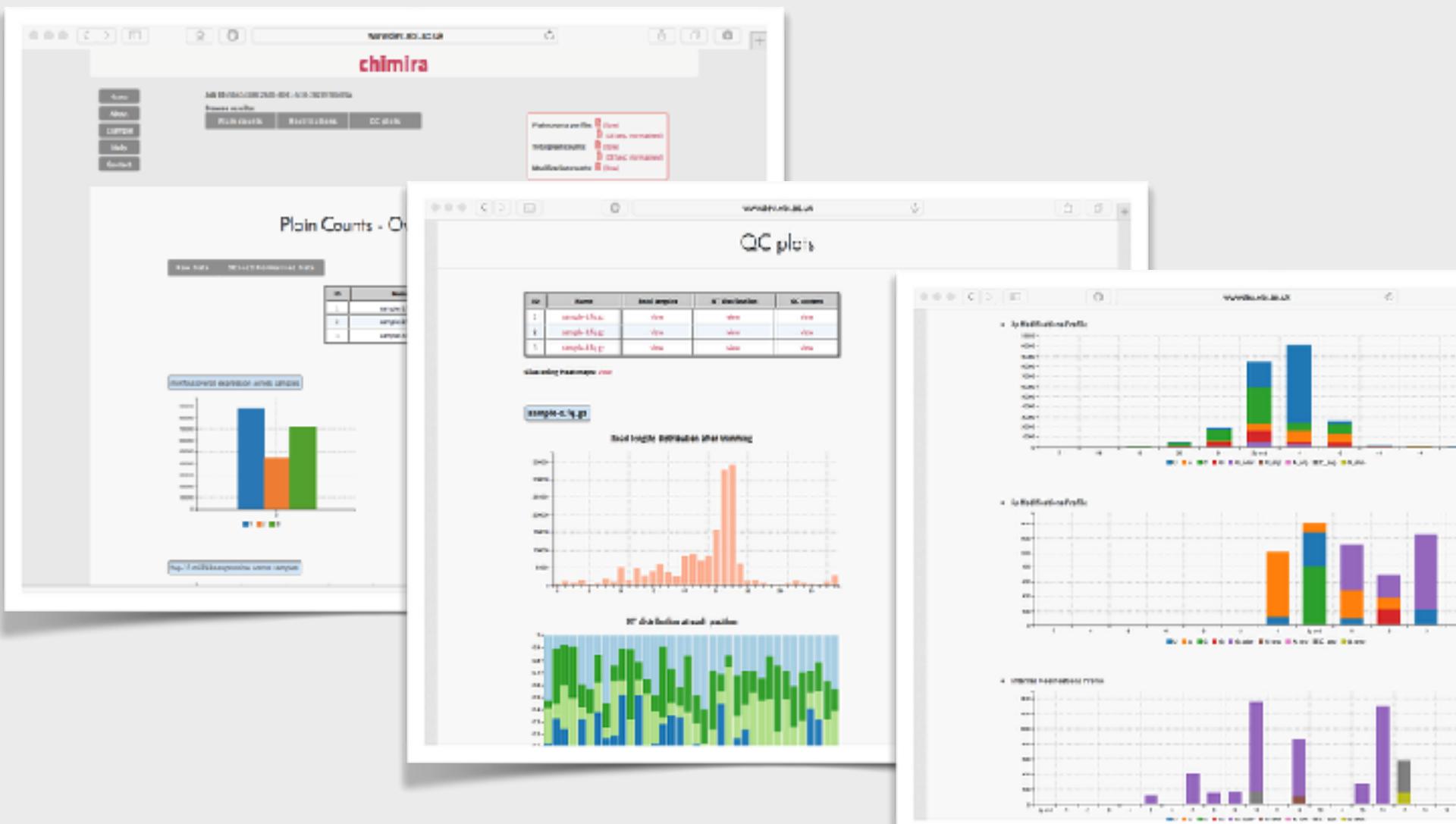
Bringing this together: ChimiRa

A simple web-based microRNA NGS analysis toolkit

- Automatically preprocesses data - adapter removal, deduplication
- Reads compared directly against known microRNA precursors
 - Multiple Species available
 - Up to 2 mismatches allowed (configurable)
 - Reads assigned to best match
 - Ambiguous reads assigned randomly
- 5p or 3p side of hairpin called automatically
- Possibility to detect 5p / 3p modifications and RNA editing
 - tutase
 - ADAR

ChimiRa Interface

wwwdev.ebi.ac.uk/enright-srv/chimira



MicroRNA Read Level Analysis

mmu-mir-26b-5p

Depth: 5741 reads

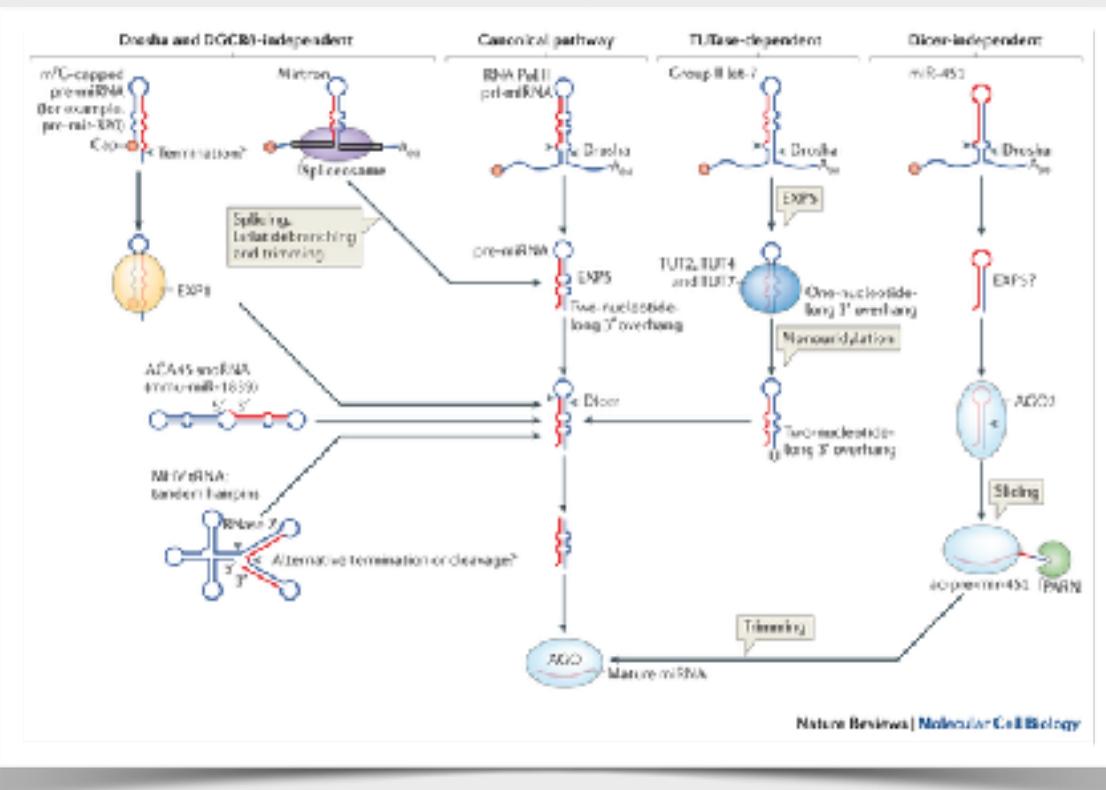
Aligned Reads on Precursor

```
UUCAGCAUUAUCGUUAGUUU 103_t9_w22 Depth:393 Modifications: no_modification
UUCAGCAUUAUCGUUAGUUU 481_t10_w21 Depth:698 Modifications: no_modification
UUCAGCAUUAUCGUUAGUUU 771_t9_w23 Depth:360 Modifications: nont_3p_U
UUCAGCAUUAUCGUUAGUUU 896_t11_w22 Depth:291 Modifications: nont_3p_A
UUCAGCAUUAUCGUUAGUUU 1300_t9_w23 Depth:175 Modifications: nont_3p_A
UUCAGCAUUAUCGUUAGUUU 1660_t10_w20 Depth:149 Modifications: no_modification
UUCAGCAUUAUCGUUAGUUU 3129_t11_w23 Depth:582 Modifications: nont_3p_AA
UUCAGCAUUAUCGUUAGUUU 3380_t10_w23 Depth:52 Modifications: nont_3p_AB
UUCAGCAUUAUCGUUAGUUU 5769_t9_w23 Depth:26 Modifications: no_modification
UUCAGCAUUAUCGUUAGUUU 8219_t10_w24 Depth:17 Modifications: nont_3p_BB
UUCAGCAUUAUCGUUAGUUU 8687_t9_w23 Depth:16 Modifications: nont_3p_CJ
UUCAGCAUUAUCGUUAGUUU 9287_t11_w21 Depth:14 Modifications: nont_3p_A
CUUCAGCAUUAUCGUUAGUUU 10438_t9_w23 Depth:13 Modifications: nont_3p_C
UUCAGCAUUAUCGUUAGUUU 11348_t9_w22 Depth:11 Modifications: nont_3p_X
UUCAGCAUUAUCGUUAGUUU 13435_t6_w19 Depth:9 Modifications: no_modification
UUCAGCAUUAUCGUUAGUUU 13682_t8_w24 Depth:9 Modifications: no_modification
UUCAGCAUUAUCGUUAGUUU 15177_t9_w19 Depth:8 Modifications: no_modification
UUCAGCAUUAUCGUUAGUUU 15767_t12_w23 Depth:7 Modifications: nont_3p_C
```

Distribution across Precursor

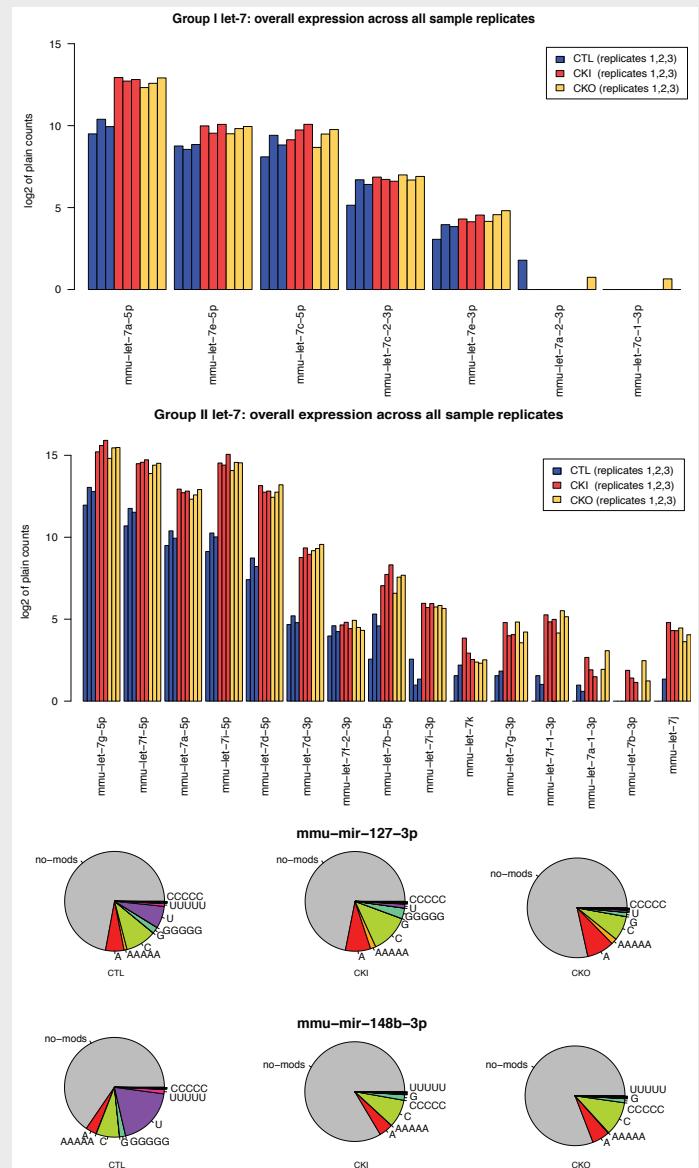


MicroRNA 3p modification Uridylation by tutase enzymes

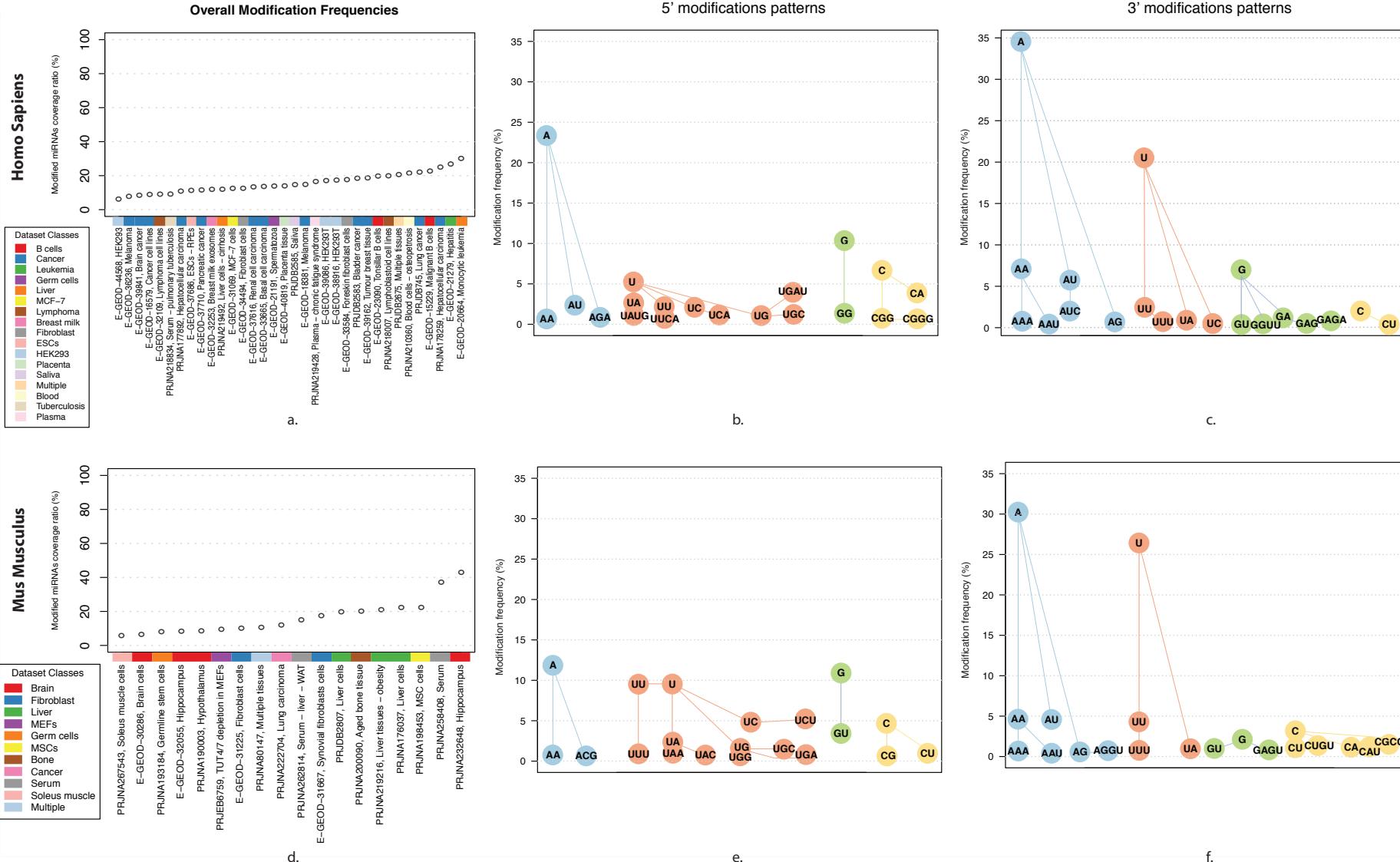


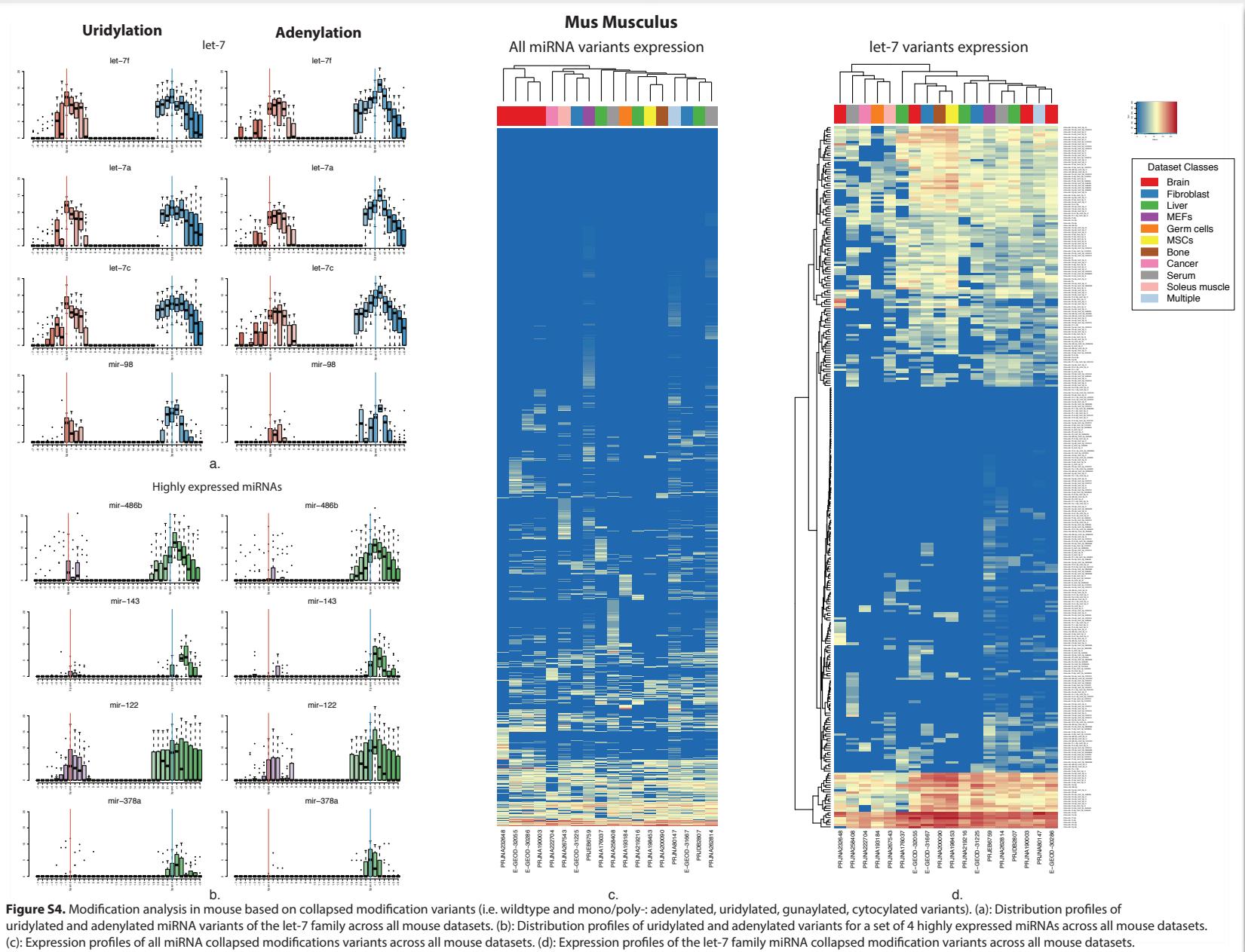
TUTase mutant data (mouse)

Francisco Sánchez Madrid Lab (T-cells) Dónal O'Carroll Lab (Germline)



Global analysis of microRNA modification





There is/was a lot of crap in miRBase

mmu-mir-5109-3p

Depth: 1496 reads

Aligned Reads on Precursor

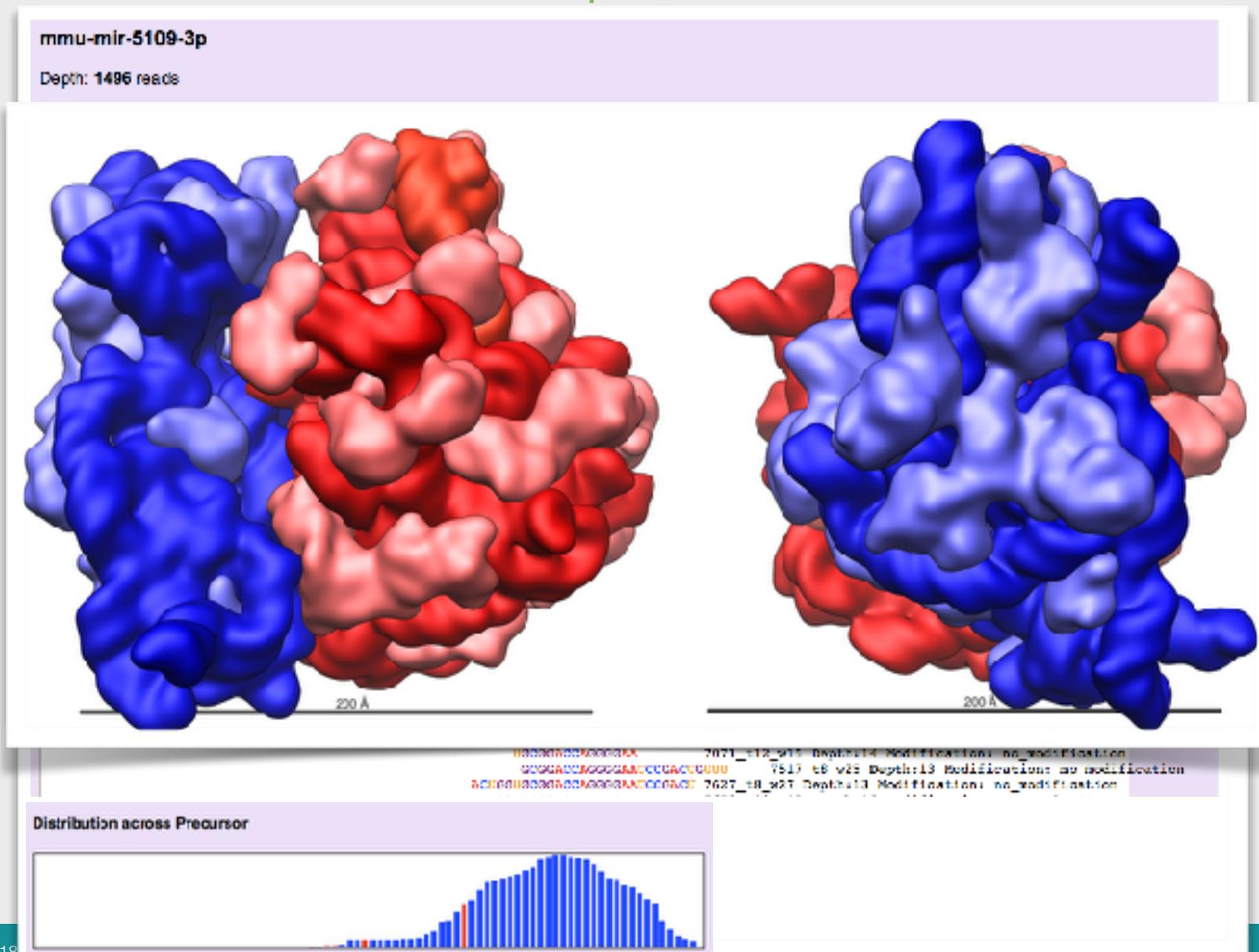
Sequence: ...
Depth: 1496 reads
Aligned Reads on Precursor

1333_t7_w17 Depth:104 Modifications: no_modification
1384_t8_w16 Depth:98 Modifications: no_modification
1705_t8_w20 Depth:75 Modifications: most_bp_CUGU
2352_t8_w15 Depth:51 Modifications: no_modification
2762_t7_w17 Depth:42 Modifications: no_modification
2913_t7_w27 Depth:40 Modifications: most_bp_AA
3433_t7_w18 Depth:33 Modifications: no_modification
3544_t7_w18 Depth:31 Modifications: no_modification
3601_t8_w14 Depth:31 Modifications: no_modification
3608_t8_w20 Depth:31 Modification: most_bp_UU
3609_t8_w15 Depth:30 Modifications: no_modification
3904_t8_w11 Depth:28 Modifications: no_modification
4272_t10_w17 Depth:25 Modification: no_modification
4514_t7_w14 Depth:24 Modifications: most_bp_AAAA
4523_t8_w24 Depth:23 Modifications: no_modification
4776_t8_w26 Depth:22 Modification: most_bp_CUG
4807_t7_w18 Depth:22 Modifications: no_modification
4863_t7_w24 Depth:22 Modification: most_bp_CUU
5147_t9_w19 Depth:20 Modification: most_bp_A
5167_t8_w16 Depth:20 Modifications: no_modification
5226_t9_w23 Depth:20 Modification: most_bp_UU
5423_t11_w15 Depth:19 Modifications: no_modification
5488_t8_w17 Depth:19 Modifications: most_bp_AAAA
5491_t9_w13 Depth:19 Modifications: no_modification
5661_t12_w15 Depth:18 Modifications: no_modification
5835_t8_w15 Depth:18 Modifications: no_modification
5926_t9_w19 Depth:17 Modification: most_bp_UU
6020_t11_w16 Depth:17 Modifications: no_modification
6231_t8_w27 Depth:16 Modifications: most_bp_AAAA
6602_t11_w16 Depth:15 Modifications: most_bp_A
6744_t7_w18 Depth:15 Modifications: no_modification
6980_t7_w24 Depth:14 Modifications: most_bp_AAAA
6984_t8_w18 Depth:14 Modifications: most_bp_CUGU
6989_t12_w15 Depth:14 Modifications: no_modification
7071_t12_w15 Depth:14 Modifications: no_modification
7519_t8_w15 Depth:13 Modifications: no_modification
7627_t8_w27 Depth:13 Modifications: no_modification

Distribution across Precursor



There is/was a lot of crap in miRBase



Analysis of NGS alignments shows many suspects



Figure S12. Coverage profiles of miRNAs that have been detected in human samples as potential artefacts. miRNAs with reads in less than 15 samples (5% of all samples) have been excluded from this analysis.

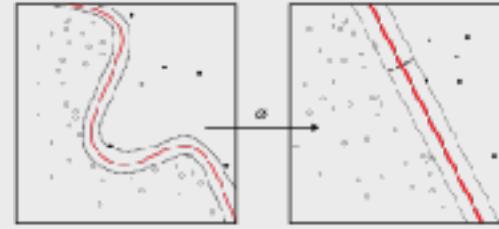
de novo microRNA Discovery

de novo microRNA discovery

- Can we use these features of microRNA alignment coverage to better predict novel microRNAs ?
- Previous approaches rely on microRNA hairpin structure and conservation analysis (e.g. miRDeep)
- Alignment features lend themselves well to machine learning approaches
- Possibility for a genome free approach ?
 - Reads can be aligned against each other into consensus sequences
 - Evidence of trimming, tailing and editing detectable

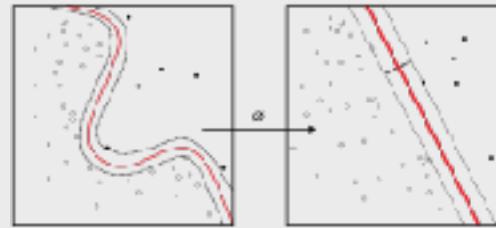
Machine Learning approach

- a) Logistic Regression
- b) Support Vector Machines
- c) **Random Forest**
- d) **Deep Learning**

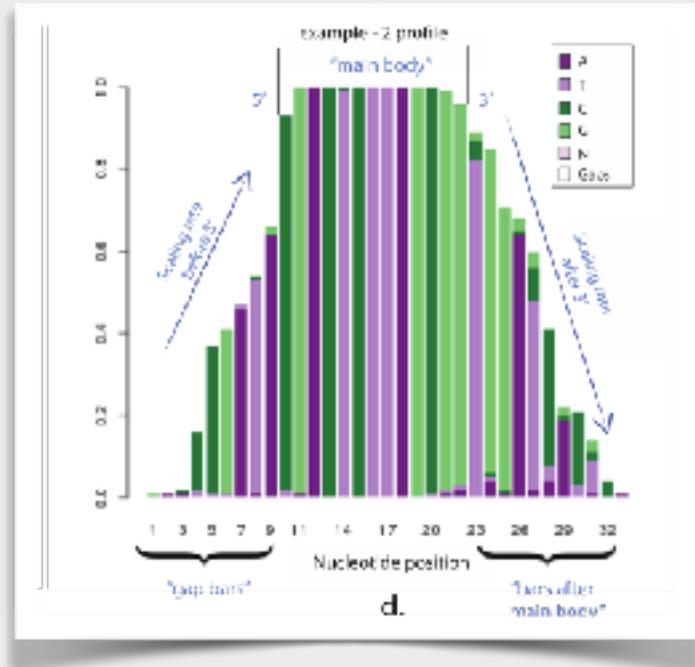


Machine Learning approach

- a) Logistic Regression
- b) Support Vector Machines
- c) **Random Forest**
- d) Deep Learning



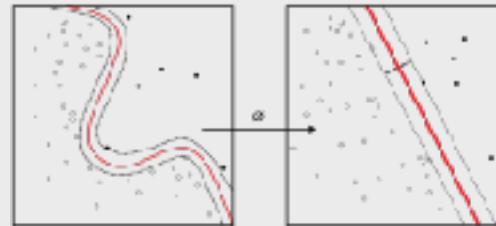
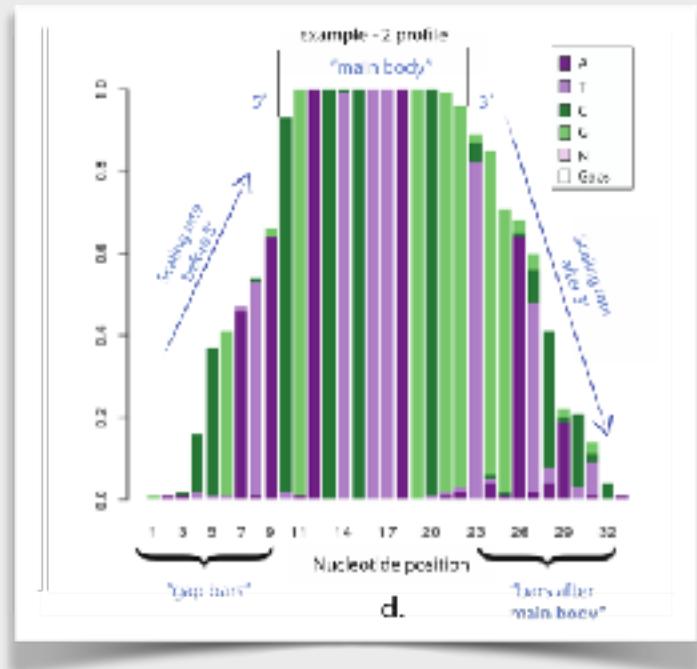
1. Core miRNA features



Machine Learning approach

- a) Logistic Regression
- b) Support Vector Machines
- c) **Random Forest**
- d) **Deep Learning**

1. Core miRNA features



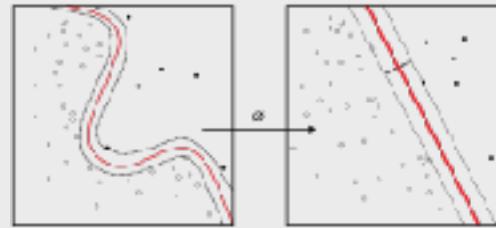
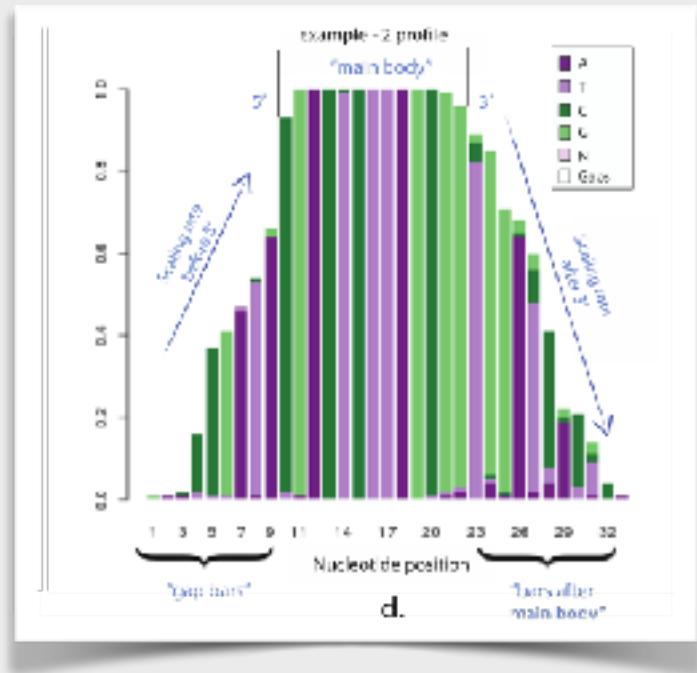
2. Sequence complexity features

- **gcs**: C+G skew
- **cpg**: CpG skew
- **cwf**: Complexity by Wootton & Federhen
- **ce**: Entropy
- **cz**: Complexity as compression ratio (using Gzip)
- **cmN**: Complexity as Markov model size of N
- **ctN**: Trifnov's complexity with order N
- **cIN**: Linguistic complexity with order N

Machine Learning approach

- a) Logistic Regression
- b) Support Vector Machines
- c) **Random Forest**
- d) **Deep Learning**

1. Core miRNA features



2. Sequence complexity features

- **gcs**: C+G skew
- **cpg**: CpG skew
- **cwf**: Complexity by Wootton & Federhen
- **ce**: Entropy
- **cz**: Complexity as compression ratio (using Gzip)
- **cmN**: Complexity as Markov model size of N
- **ctN**: Trifnov's complexity with order N
- **cIN**: Linguistic complexity with order N

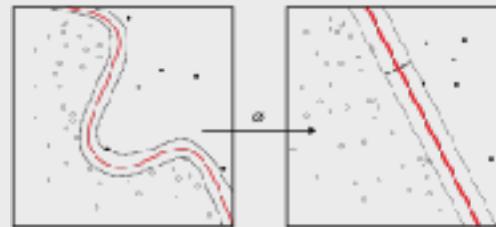
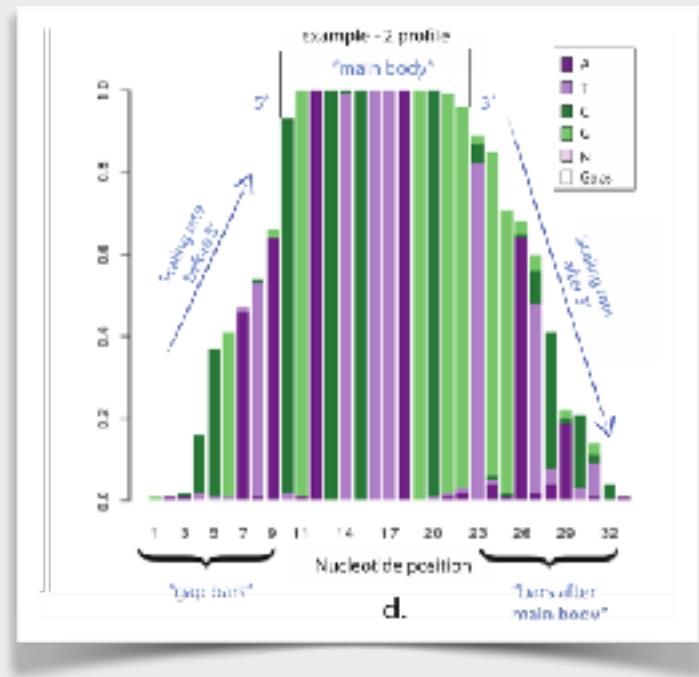
3. Genomic features

- **hairpin size**
- **number of unpaired bp**
- **min. free energy**
- **loop distance from hairpin stem**
- etc...

Machine Learning approach

- a) Logistic Regression
- b) Support Vector Machines
- c) **Random Forest**
- d) **Deep Learning**

1. Core miRNA features

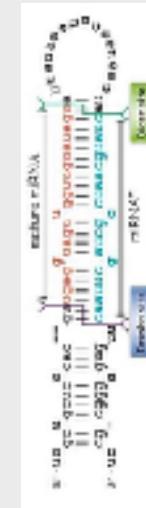


2. Sequence complexity features

- **gcs**: C+G skew
- **cpg**: CpG skew
- **cwf**: Complexity by Wootton & Federhen
- **ce**: Entropy
- **cz**: Complexity as compression ratio (using Gzip)
- **cmN**: Complexity as Markov model size of N
- **ctN**: Trifnov's complexity with order N
- **cIN**: Linguistic complexity with order N

3. Genomic features

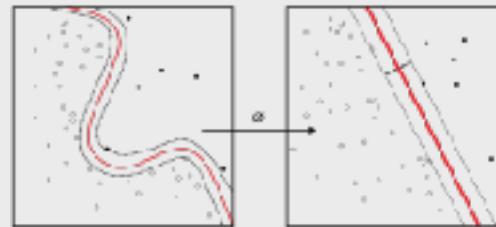
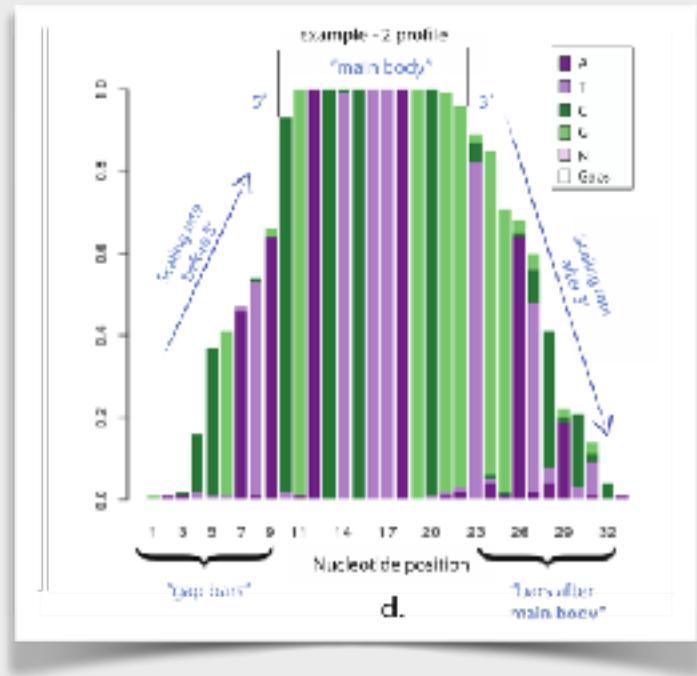
- **hairpin size**
- **number of unpaired bp**
- **min. free energy**
- **loop distance from hairpin stem**
- etc...



Machine Learning approach

- a) Logistic Regression
- b) Support Vector Machines
- c) **Random Forest**
- d) **Deep Learning**

1. Core miRNA features



2. Sequence complexity features

- **gcs**: C+G skew
- **cpg**: CpG skew
- **cwf**: Complexity by Wootton & Federhen
- **ce**: Entropy
- **cz**: Complexity as compression ratio (using Gzip)
- **cmN**: Complexity as Markov model size of N
- **ctN**: Trifnov's complexity with order N
- **cIN**: Linguistic complexity with order N

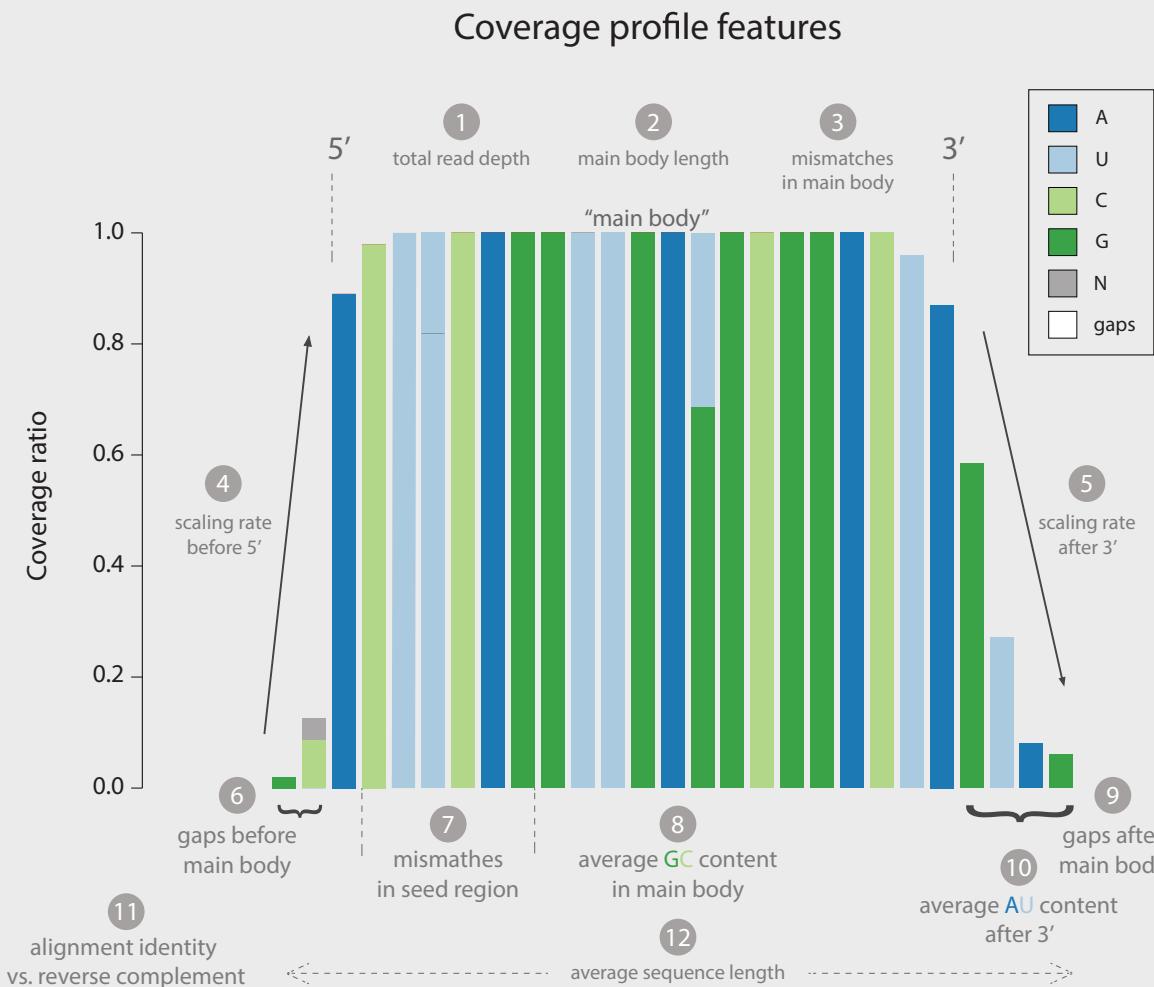
3. Genomic features

- **hairpin size**
- **number of unpaired bp**
- **min. free energy**
- **loop distance from hairpin stem**
- etc...

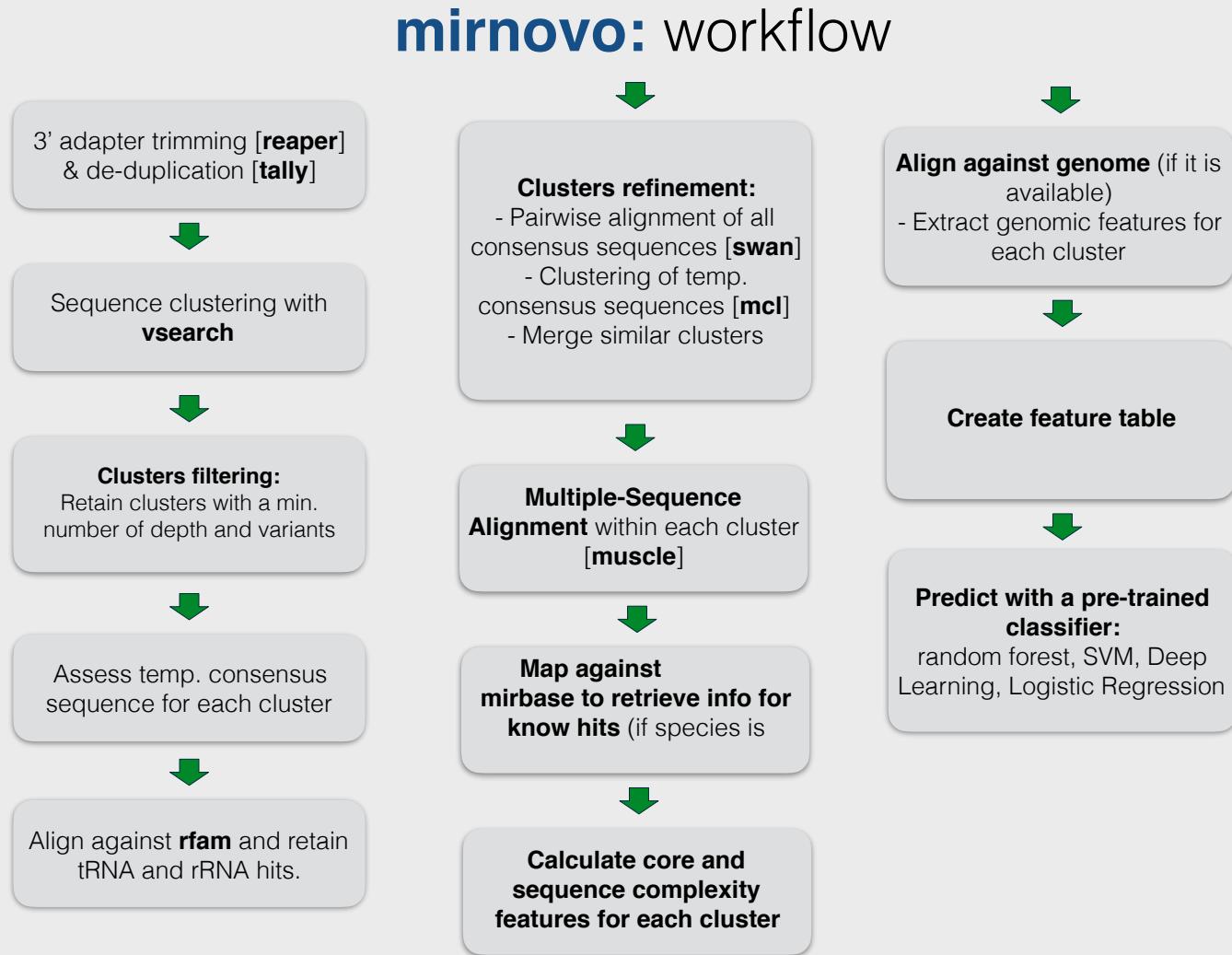
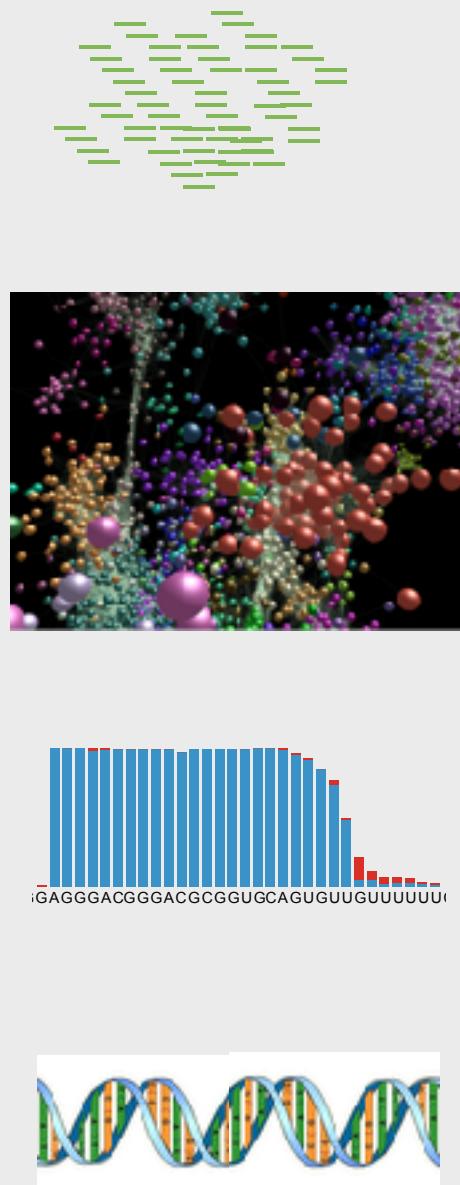
34 features overall



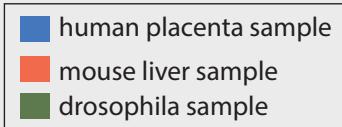
Coverage Feature Set



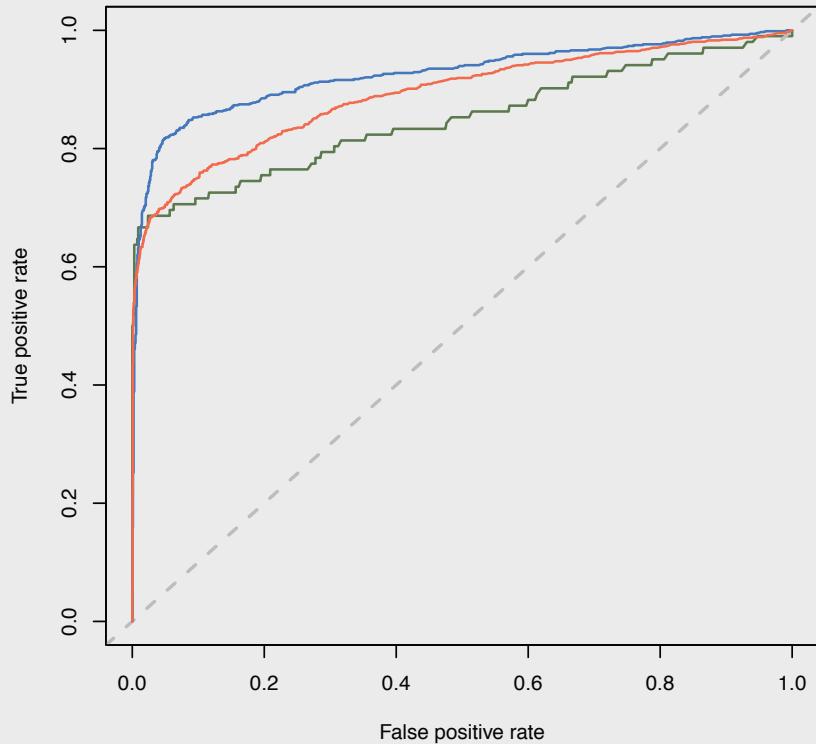
mirnovo: workflow



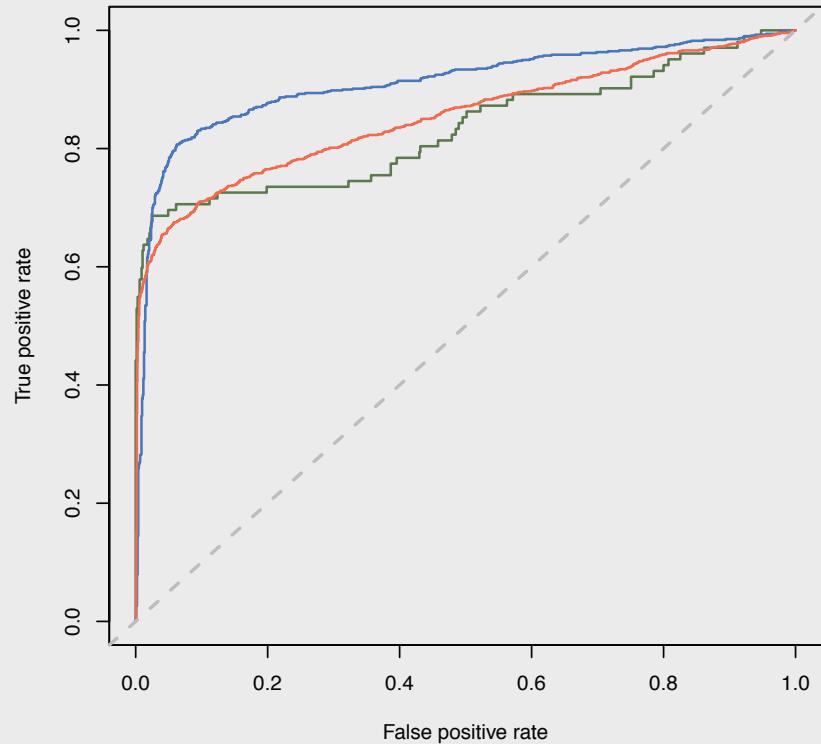
ROC Curves from test datasets



Using Random Forest Classifier



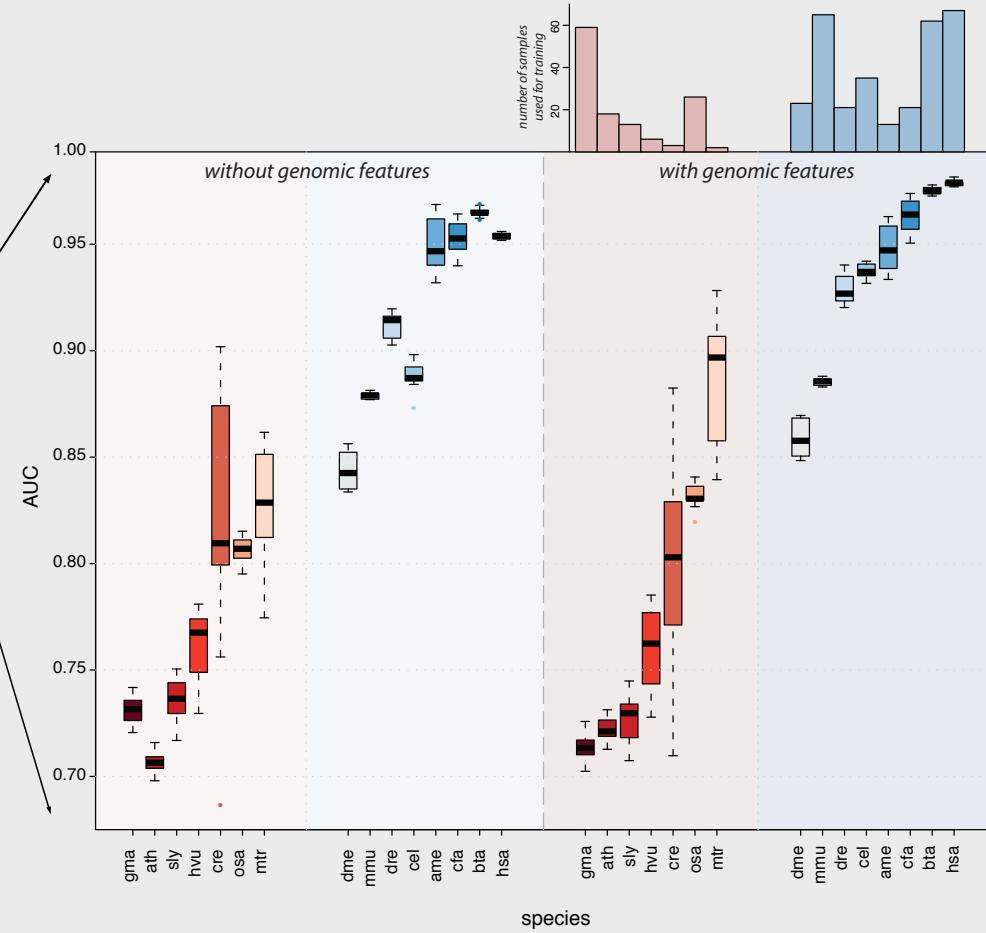
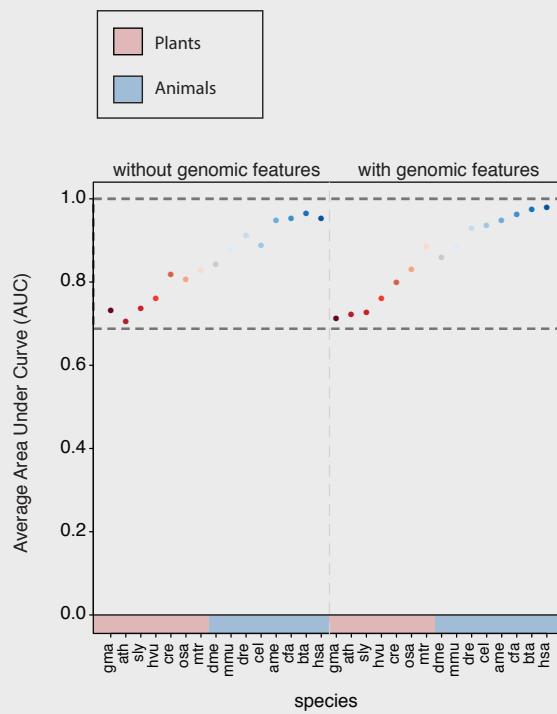
Using SVM Classifier



Performance

b

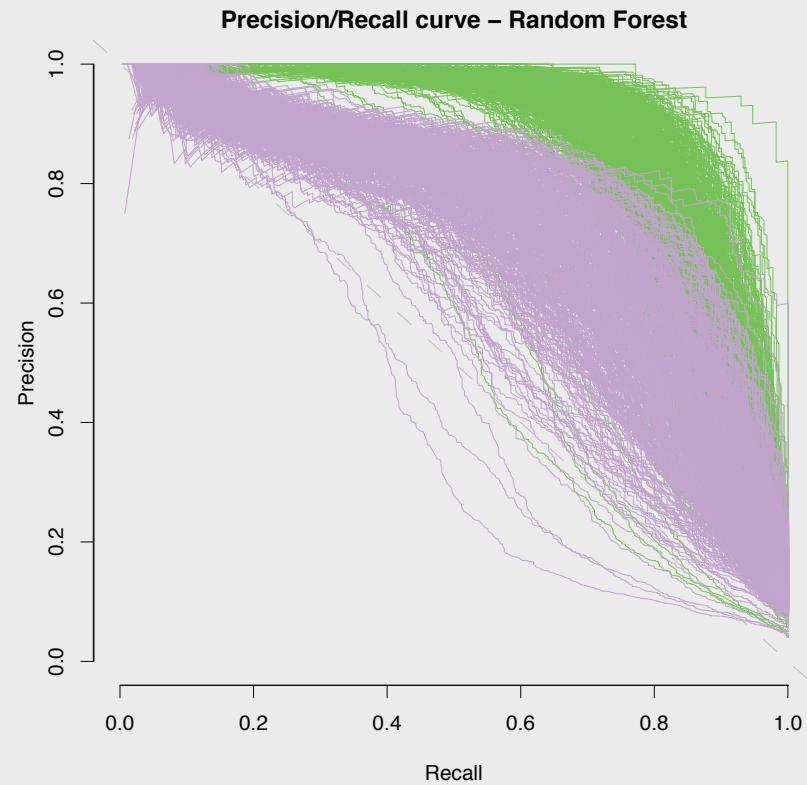
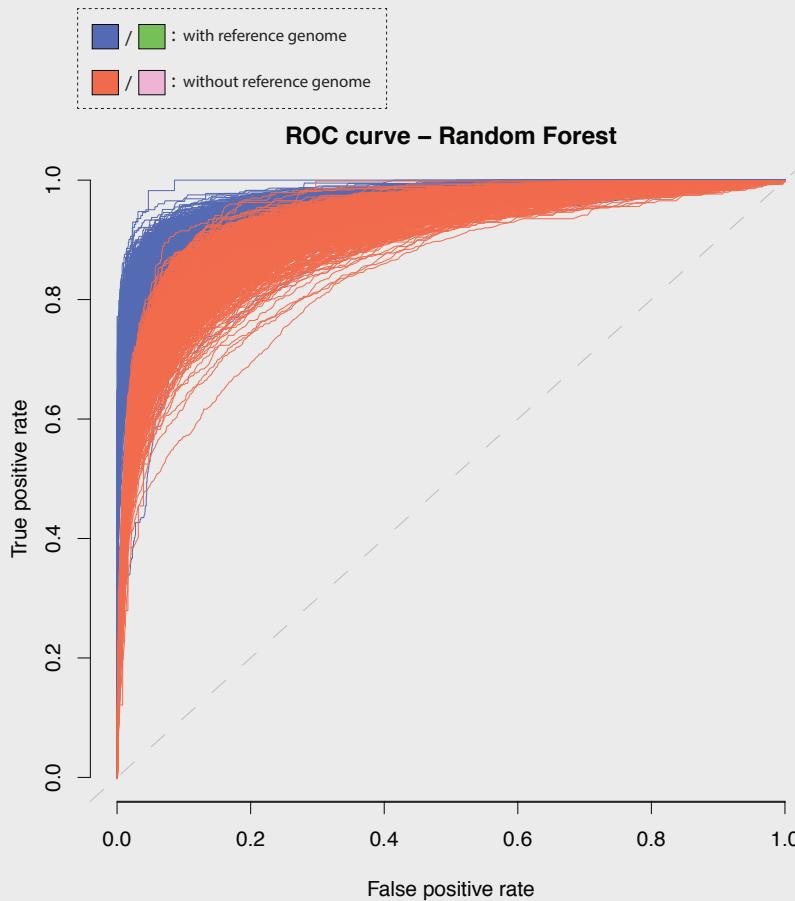
Training model performance with 10-fold cross-validation



GEUVADIS Performance

b

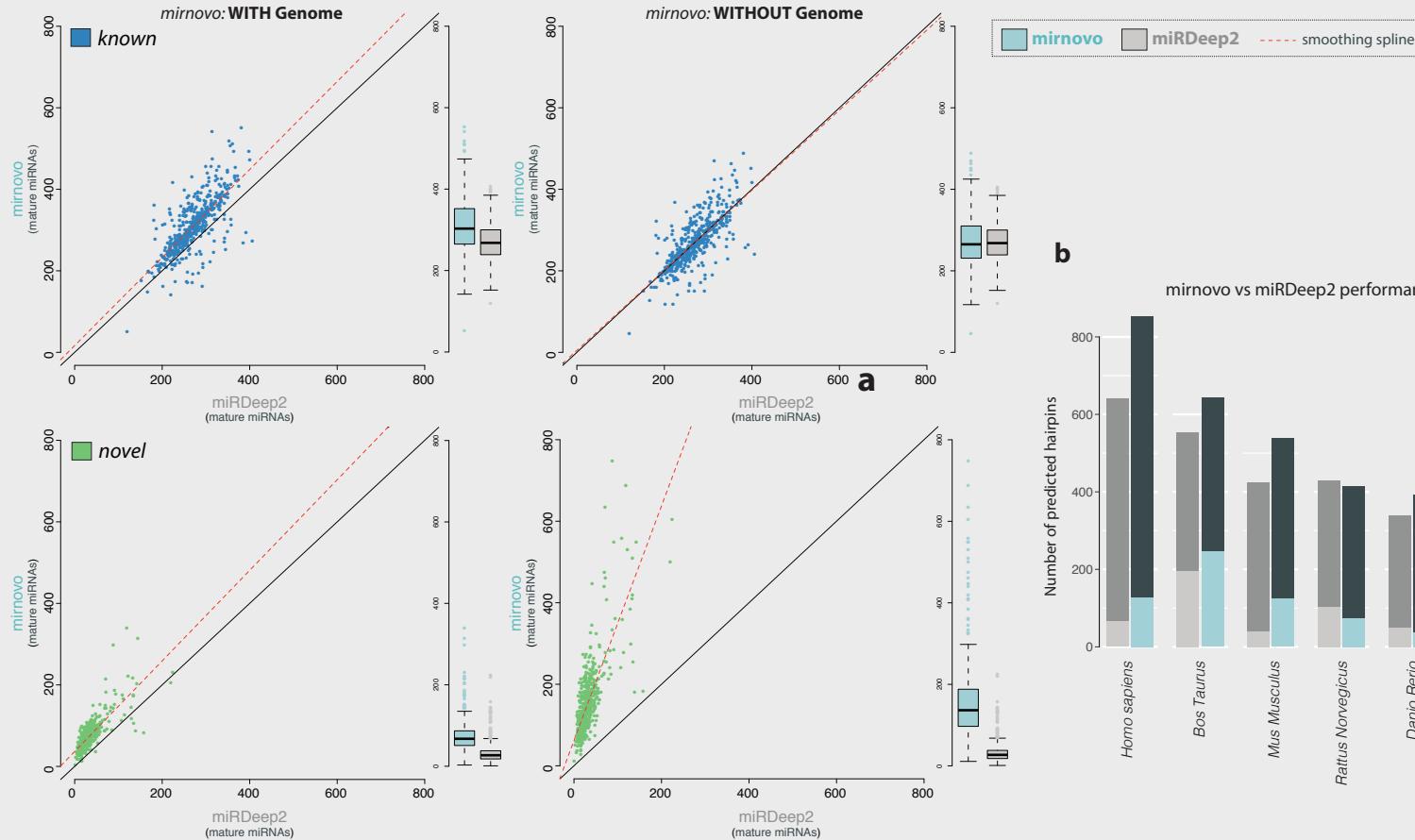
mirnovo overall performance on GEUVADIS dataset,
with and without a reference genome



Performance vs miRDeep2

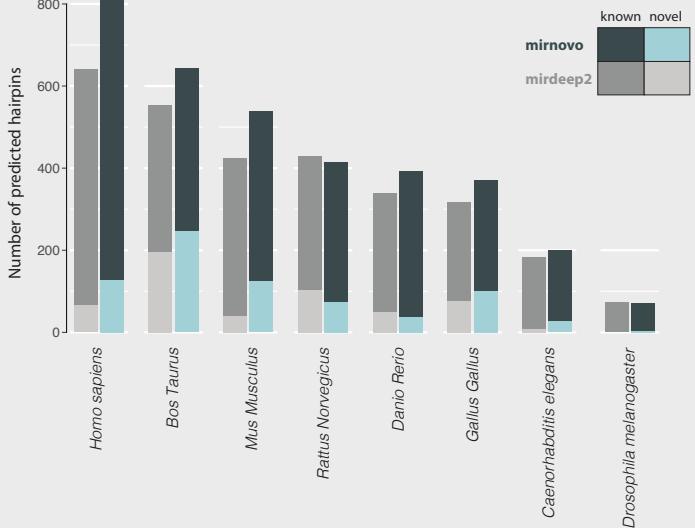
a

mirnovo vs miRDeep2 benchmarking (on GEUVADIS dataset)



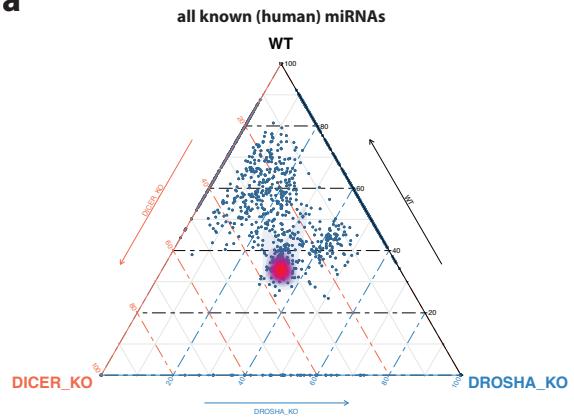
b

mirnovo vs miRDeep2 performance on 8 model organisms

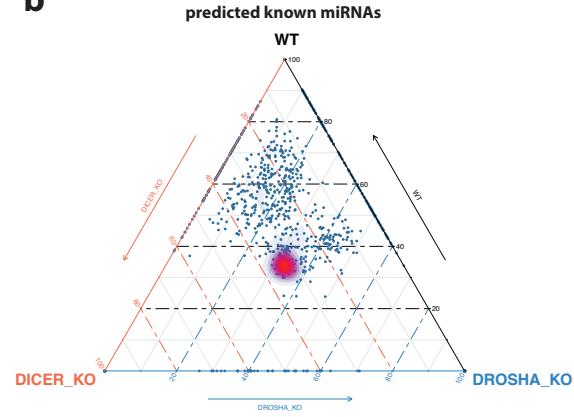


Expression of different sets of miRNAs across wild-type (WT), Drosha-Knockout and Dicer-Knockout human samples.

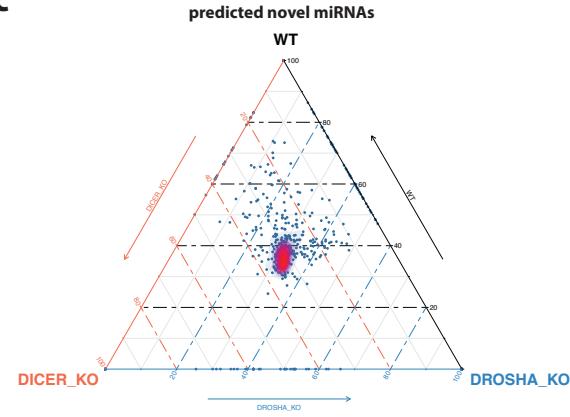
a



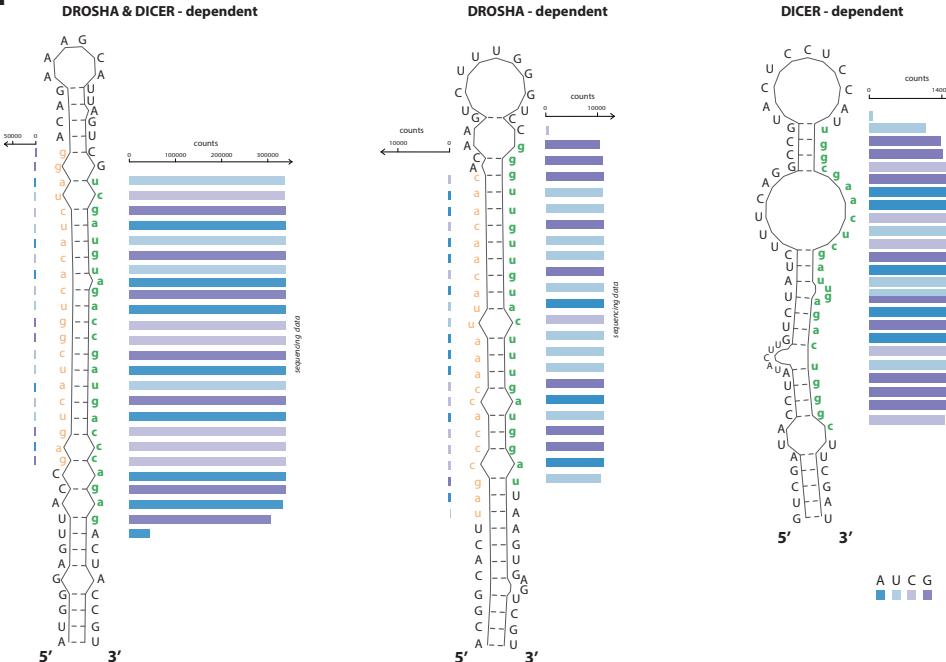
b



c



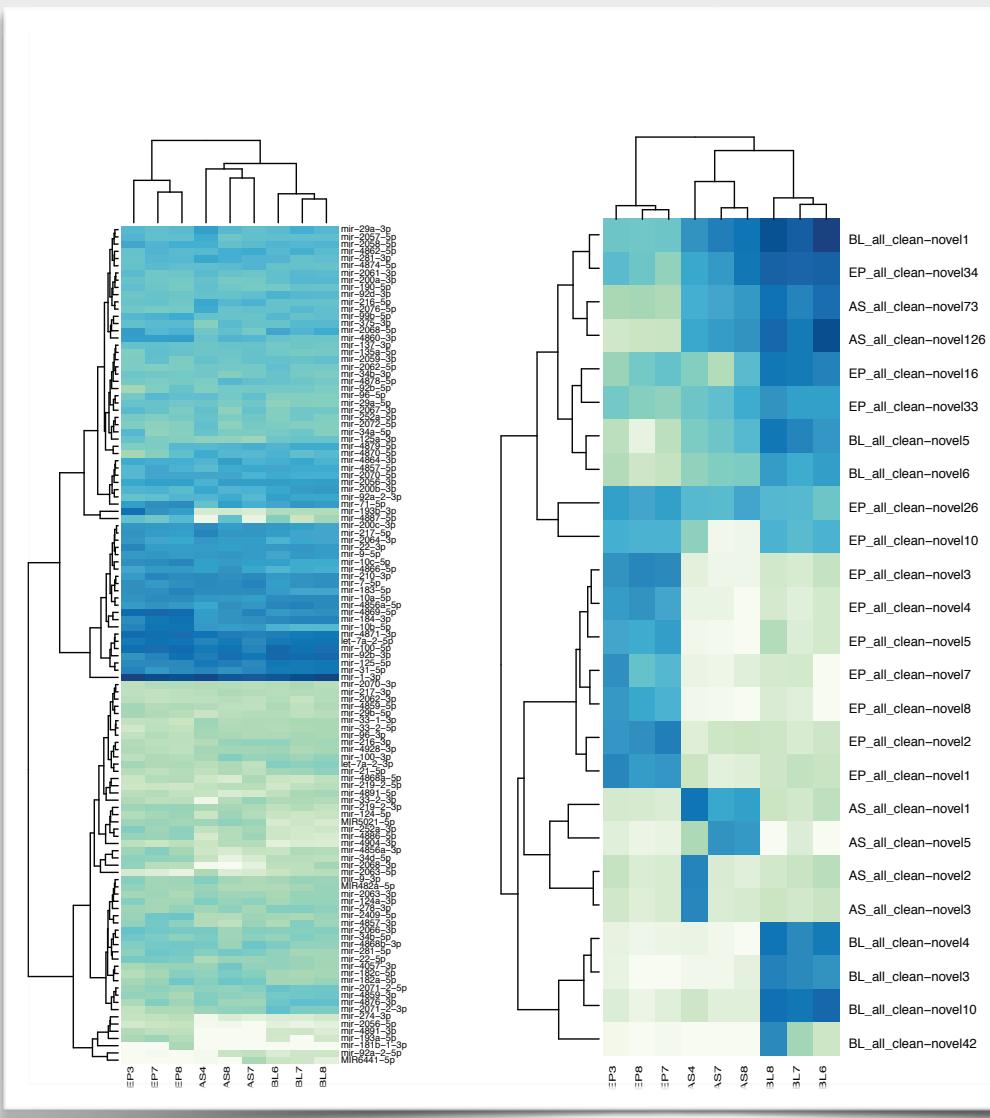
d



Running in genome-less mode



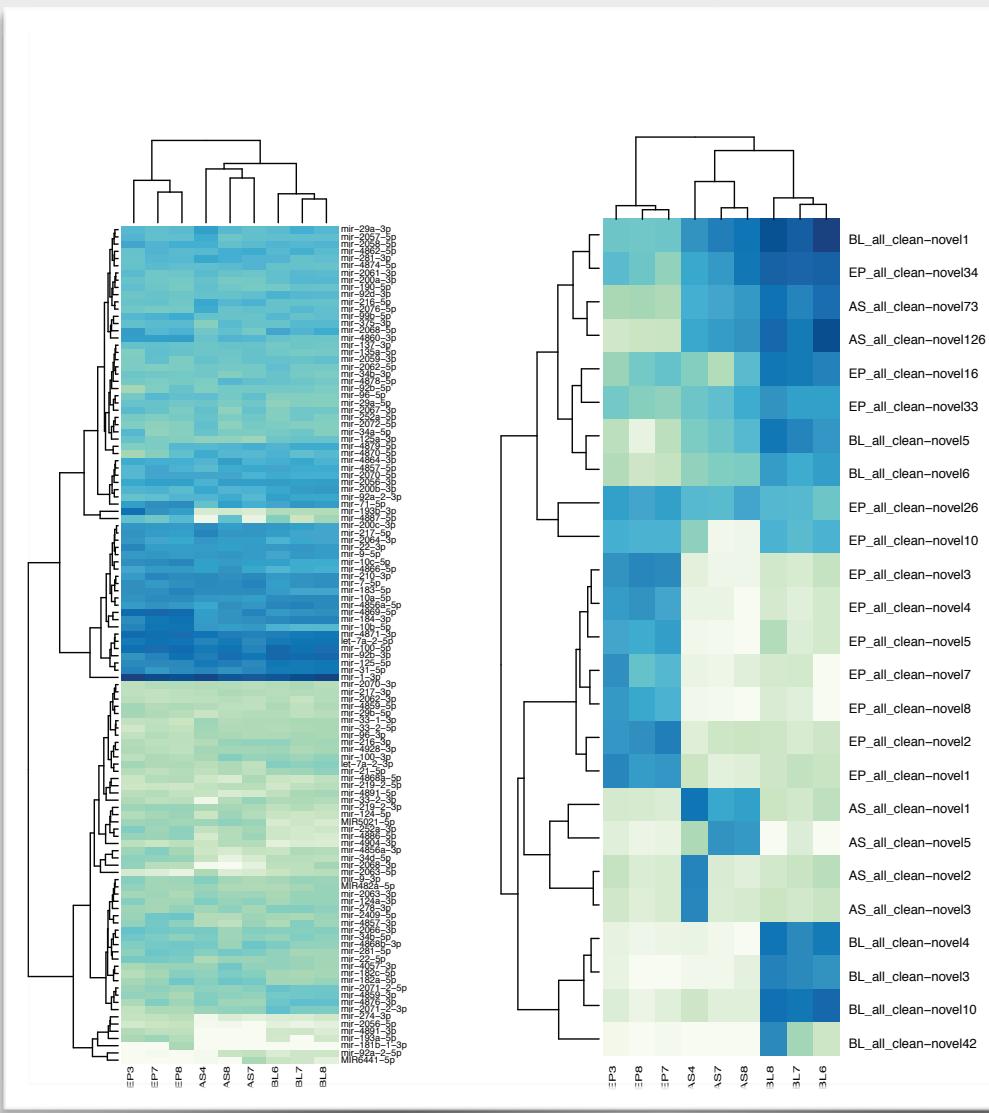
- Amphioxus (lancelet fish)
- With **Elia Benito-Gutierrez** (Zoology)



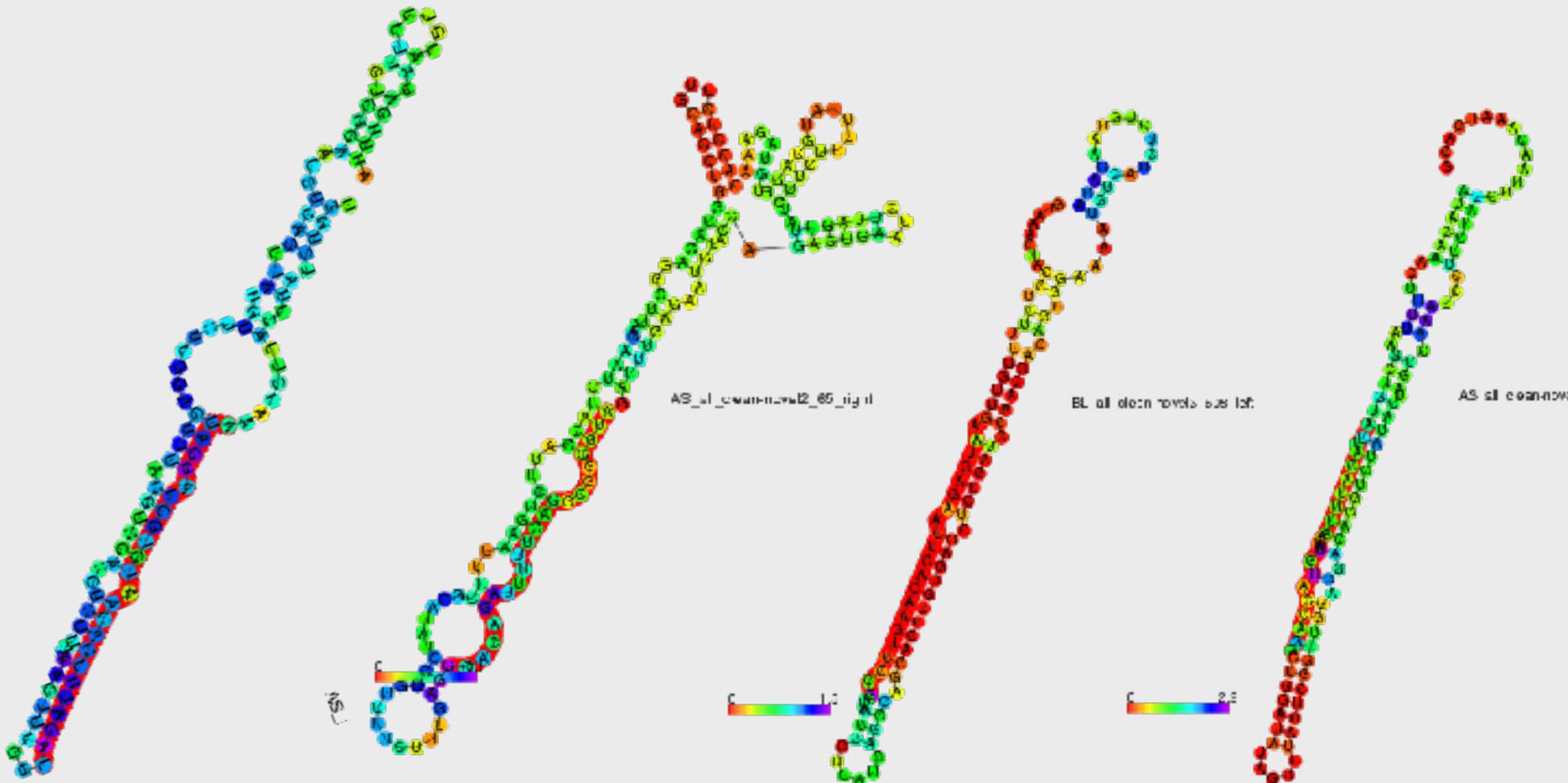
Running in genome-less mode



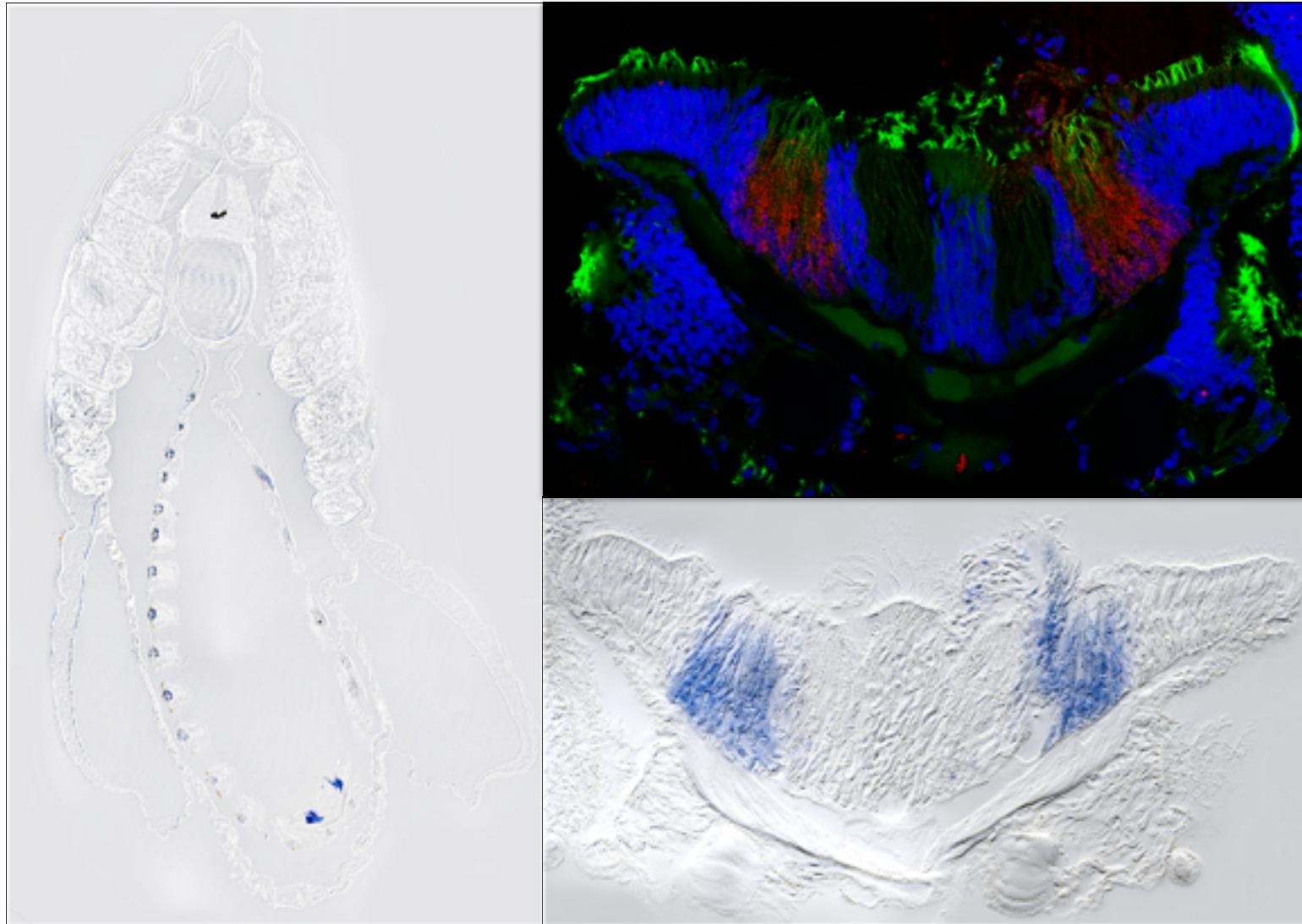
- Amphioxus (lancelet fish)
 - With **Elia Benito-Gutierrez** (Zoology)
 - 3 novel species, sig. diverged
 - Maldives
 - North sea
 - Mediterranean
 - Genomes not yet available
 - DNA Sequencing and assembly in progress



For some predictions we have limited DNA sequence

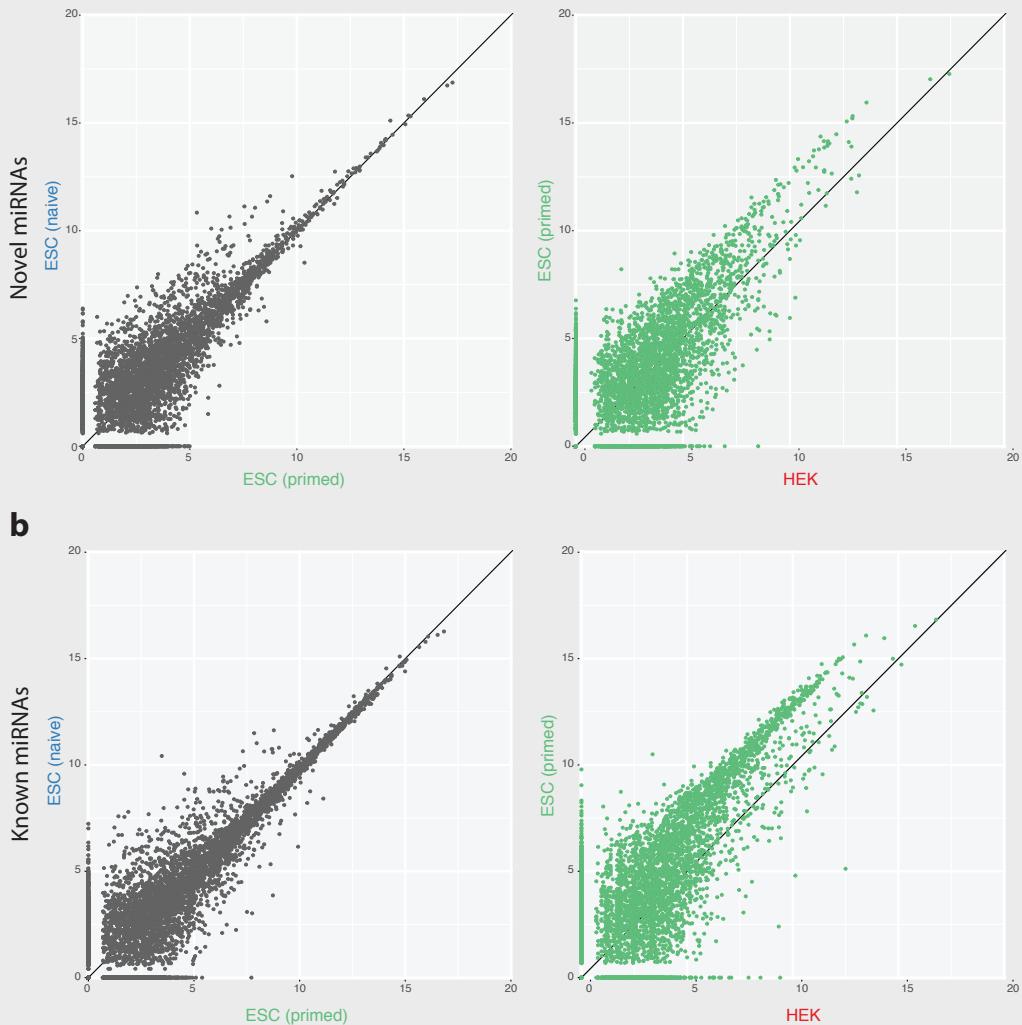


Validation by *in situ* and northern blot



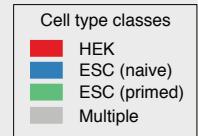
a

Differential expression of known and novel miRNAs in single-cells



C

Single-cell hierarchical clustering



156

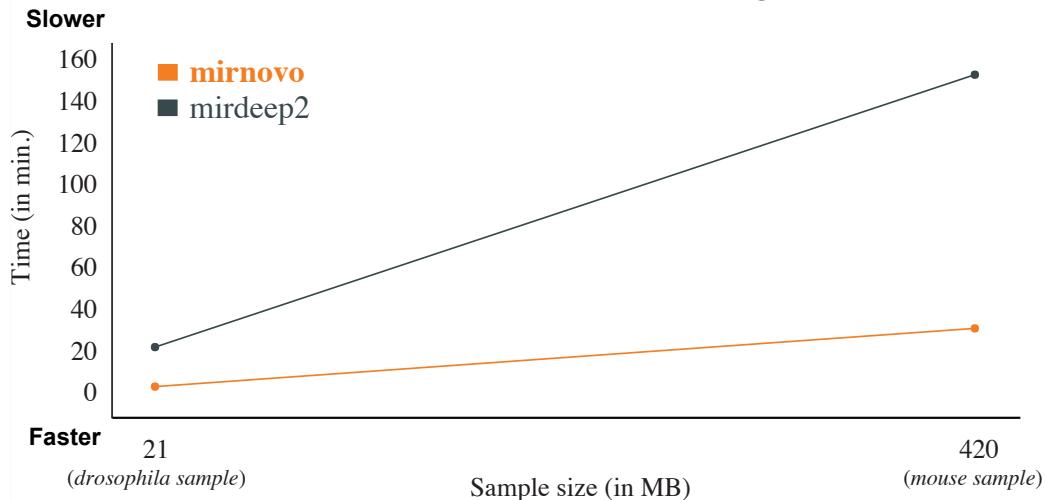
The screenshot shows the miRNova web application. At the top, there is a navigation bar with links for Home, Tools, Examples, Help, and Contact. The Home link is highlighted. To the right of the navigation bar is a cartoon owl and a microRNA hairpin icon. The main heading on the page is "no reference genome required!". Below this, a sub-headline says "Predict novel miRNAs from small RNA-Seq samples". A central input form allows users to upload FASTA/FASTQ files by dragging them here or clicking the "Upload files" button. It also includes fields for "Input species" (set to Homo sapiens (hsa)), "Training model" (set to generate new), "Adapter sequence", and "Send results to (email)". There is also an "Advanced options" link. Below this form is a large blue "EXAMPLES" section with the sub-headline "Preview results of example samples from 4 different species". At the bottom of the page, there is a footer with the text "Preview results of example samples from 4 different species" and a "REFERENCES" section.

Table S1. Comparison of novel miRNA prediction tools based on features availability. Cells in blue indicate a supported feature while white cells denote the lack of the feature. Web-server applications that are no longer available are denoted as ‘*offline*’.

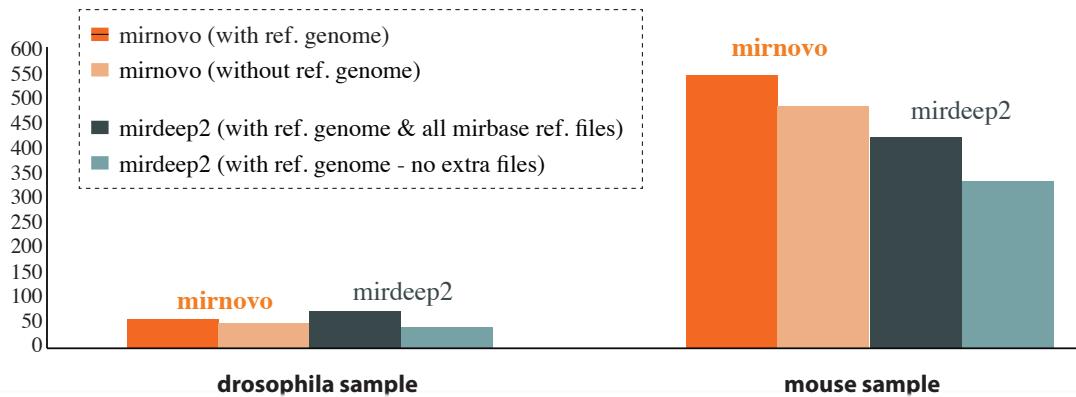
	Web-server	Stand-alone	Animals	Plants	Genomic features	Prediction without genome
miRDeep2						
miRAnalyzer	<i>offline*</i>					
miRTRAP	<i>offline*</i>					
mirTools 2.0						
miRDeep-P	<i>offline*</i>					
MiReNA	<i>offline*</i>					
miReader		<i>crashes during installation</i>				
MirPlex		<i>crashes during runtime</i>				
mirnovo						

* last checked on July 3, 2017

Run Time Benchmarking



miRNA Prediction Benchmarking



Acknowledgements

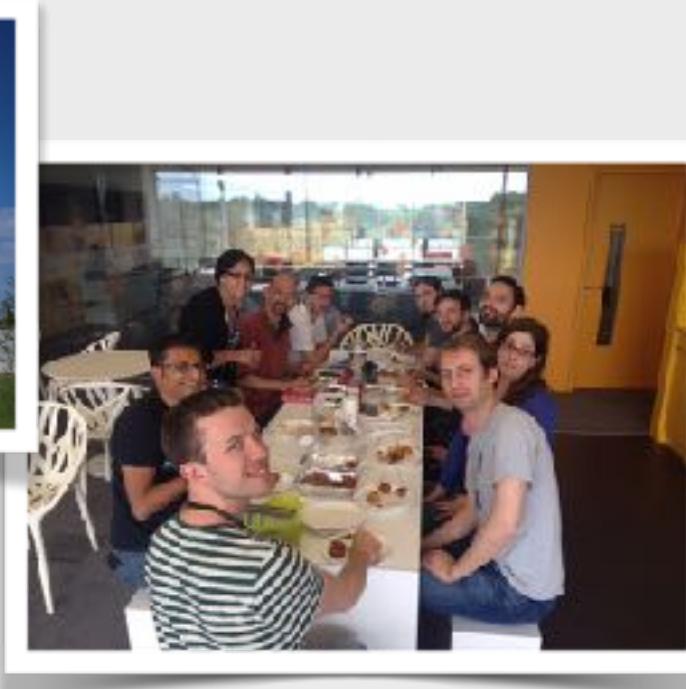
- EMBL-EBI

- Dimitrios Vitsios
- Elsa Kentepozidou
- Mat Davis
- Leonor Quintais
- Stijn van Dongen



- University of Edinburgh

- Dónal O'Carroll
- Marcos Morgan



- University of Cambridge (Zoology)

- Elia Benito-Gutierrez

- Wellcome Trust Sanger Institute

- Anna Protasio & Mathew Berriman



Acknowledgements

- EMBL-EBI

- Dimitrios Vitsios
- Elsa Kentepozidou
- Mat Davis
- Leonor Quintais
- Stijn van Dongen

- University of Edinburgh

- Dónal O'Carroll
- Marcos Morgan

- University of Cambridge

- Elia Benito-Gutierrez

- Wellcome Trust Sanger Institute

- Anna Protasio & Mathew Berriman



Wrap Up



Overview

- **Introduction to R and BioConductor**
- **Small RNA Sequencing Analysis**
 - Sample QC and cleaning - Reaper (R)
 - Sample Mapping - ChimiRa
 - Differential count analysis (R/BioConductor) - DESeq2
 - Nrep Experimental data
- **mRNA Sequencing Analysis (Illumina)**
 - MiR-210 Overexpression
 - Read Mapping - HiSat2
 - QC
 - Differential Count Analysis (R/BioConductor) - DESeq2
- **Direct RNA Sequencing (Nanopore)**
 - Library preparation
 - Flowcell loading and sequencing (MinKnow)
 - Basecalling (Albacore)
 - Mapping (Minimap2 vs Genome and Transcriptome)
 - Sequence analysis (R/BioConductor and IGV)

Practical Webpages

Wellcome Trust Advanced Courses - RNA Transcriptomics

WELLCOM GENOME CAMBRIDGE ADVANCED COURSES AND SCIENTIFIC CONFERENCES

UNIVERSITY OF CAMBRIDGE

2018 WTAC - RNA Transcriptomics, Hinxton, Cambridge, UK

Instructors: Anton Enright, Jack Monahan, Adrien Leger

Anton Enright (tae39@cam.ac.uk)	Jack Monahan (monahan@ebi.ac.uk)	Adrien Leger (aleg@ebi.ac.uk)

Department of Pathology,
Tennis Court Road,
Cambridge, UK.

WTAC logo

- <http://tinyurl.com/wtac2018>

Courses-and-Practicals/Practical_2.md at master · EnrightLab/Courses-and-Practicals

```
barplot(apply(wtaccounts.2,sum), las=2,col=cond_colours,main="Post-Normalised Counts",cex.lab=1.5,cex.main=1.5)
```

Post-Normalised Counts

Sample ID	Post-Normalised Counts (approx.)
1	2.8e+07
2	2.8e+07
3	3.2e+07
4	2.2e+07
5	2.5e+07
6	3.0e+07

```
# We will apply the Variance Stabilizing Transformation (VST) if it's better than log2 for counts.  
vst <- varStabilizingTransformation(wtaccounts)  
wtaccounts <- apply(vst)  
wtaccounts <- wtaccounts[order(apply(wtaccounts,1,sum),decreasing = TRUE),]
```

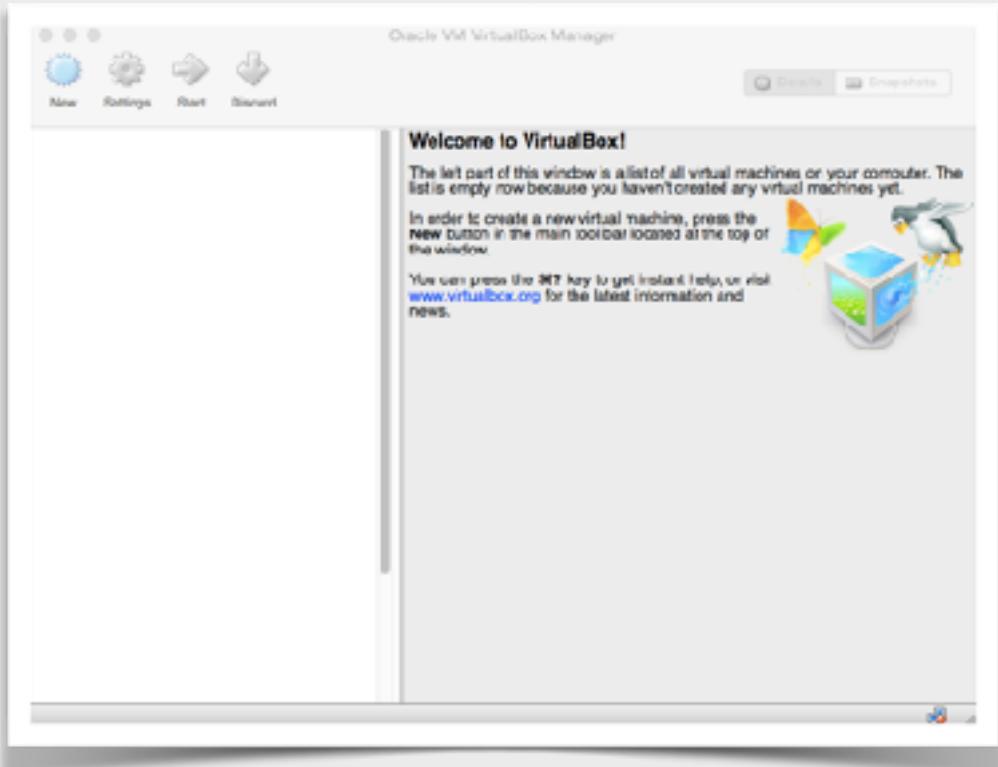
as an additional QC step we can calculate the sample-to-sample Pearson correlations and plot them in a heatmap.

```
heatmaps.2<-cor(wtaccounts[,1:ncol(wtaccounts)],col=hmcol,nab="Sample-to-Sample Correlation (Raw Counts)",
```

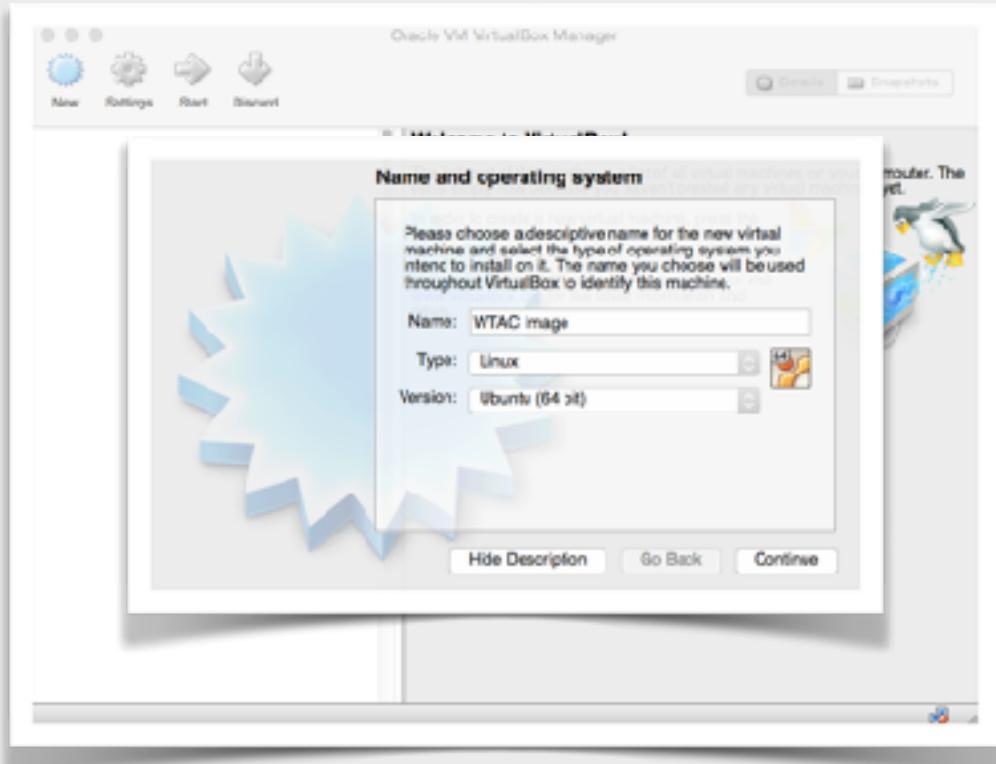
Color Key and Histogram

Sample-to-Sample Correlation (Raw Counts)

VirtualBox Installation

A screenshot of the VirtualBox download page from virtualbox.org. The page features a large "VirtualBox" logo at the top. On the left, there is a sidebar with links: About, Screenshots, Downloads, Documentation, End-user docs, Technical docs, Contribute, and Community. The main content area is titled "Download VirtualBox". It says "Here, you will find links to VirtualBox binaries and its source code." Below this is a section titled "VirtualBox binaries" with the subtext "By downloading, you agree to the terms and conditions of the respective license agreements." There are several bullet points under this section, each linking to a specific binary package. At the bottom, it says "See the [changelog](#) for what has changed. You might want to compare the" followed by two more bullet points about SHA256 checksums.

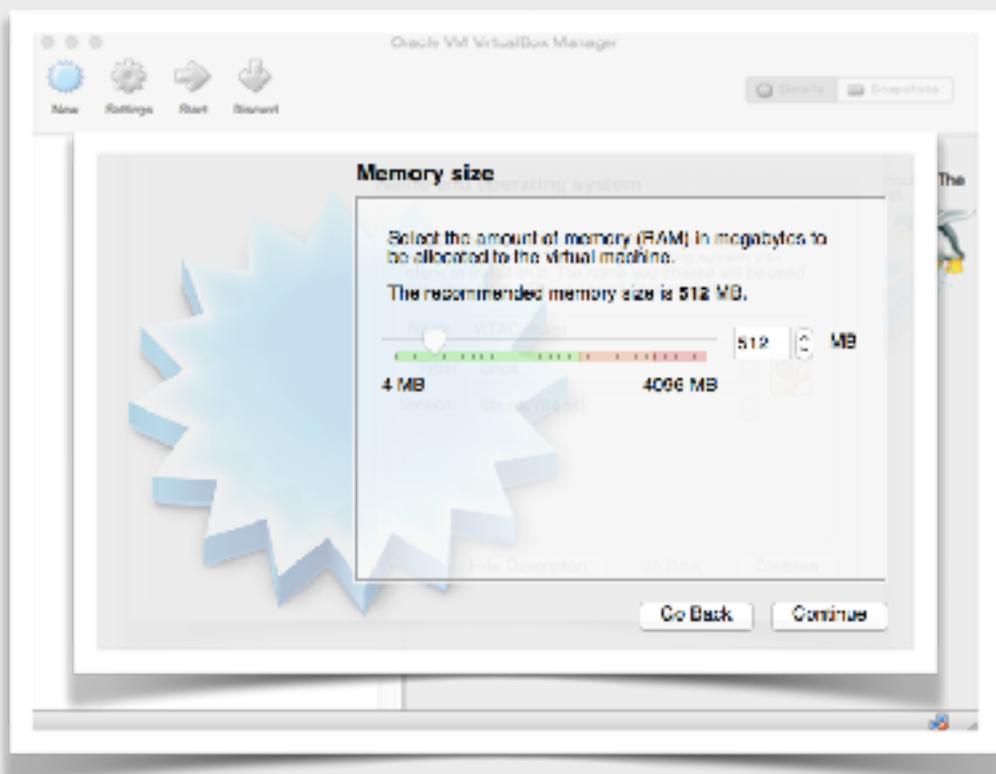
VirtualBox Installation



The screenshot shows the VirtualBox download page at virtualbox.org. The page features a large blue 'VirtualBox' logo at the top. Below it, a section titled 'Download VirtualBox' is shown. A sidebar on the left includes links for 'About', 'Screenshots', 'Downloads', 'Documentation', 'End-user docs', 'Technical docs', 'Contribute', and 'Community'. The main content area starts with a statement: 'By downloading, you agree to the terms and conditions of the respective license agreements.' followed by a list of download links:

- **VirtualBox platform packages.** The binaries are released under the GPL version 2.
 - [VirtualBox 4.3.28 for Windows hosts](#) (x86/amd64)
 - [VirtualBox 4.3.28 for OS X hosts](#) (x86/amd64)
 - [VirtualBox 4.3.28 for Linux hosts](#)
 - [VirtualBox 4.3.28 for Solaris hosts](#) (amd64)
- **VirtualBox 4.3.28 Oracle VM VirtualBox Extension Pack** (platforms)
Support for USB 2.0 devices, VirtualBox RDP and PXE boot and more. See the User Manual for an introduction to this Extension Pack. Binaries are released under the [VirtualBox Personal Evaluation License \(PUEL\)](#).
Please install the extension pack with the same version as your VirtualBox!
 - If you are using [VirtualBox 4.2.28](#), please download the extension pack.
 - If you are using [VirtualBox 4.1.36](#), please download the extension pack.
 - If you are using [VirtualBox 4.0.28](#), please download the extension pack.
- **VirtualBox 4.3.28 Software Developer Kit (SDK)** (All platforms)
See the [changelog](#) for what has changed.
You might want to compare the:
 - [SHA256 checksums](#) or the
 - [MD5 checksums](#)to verify the integrity of downloaded packages.

VirtualBox Installation



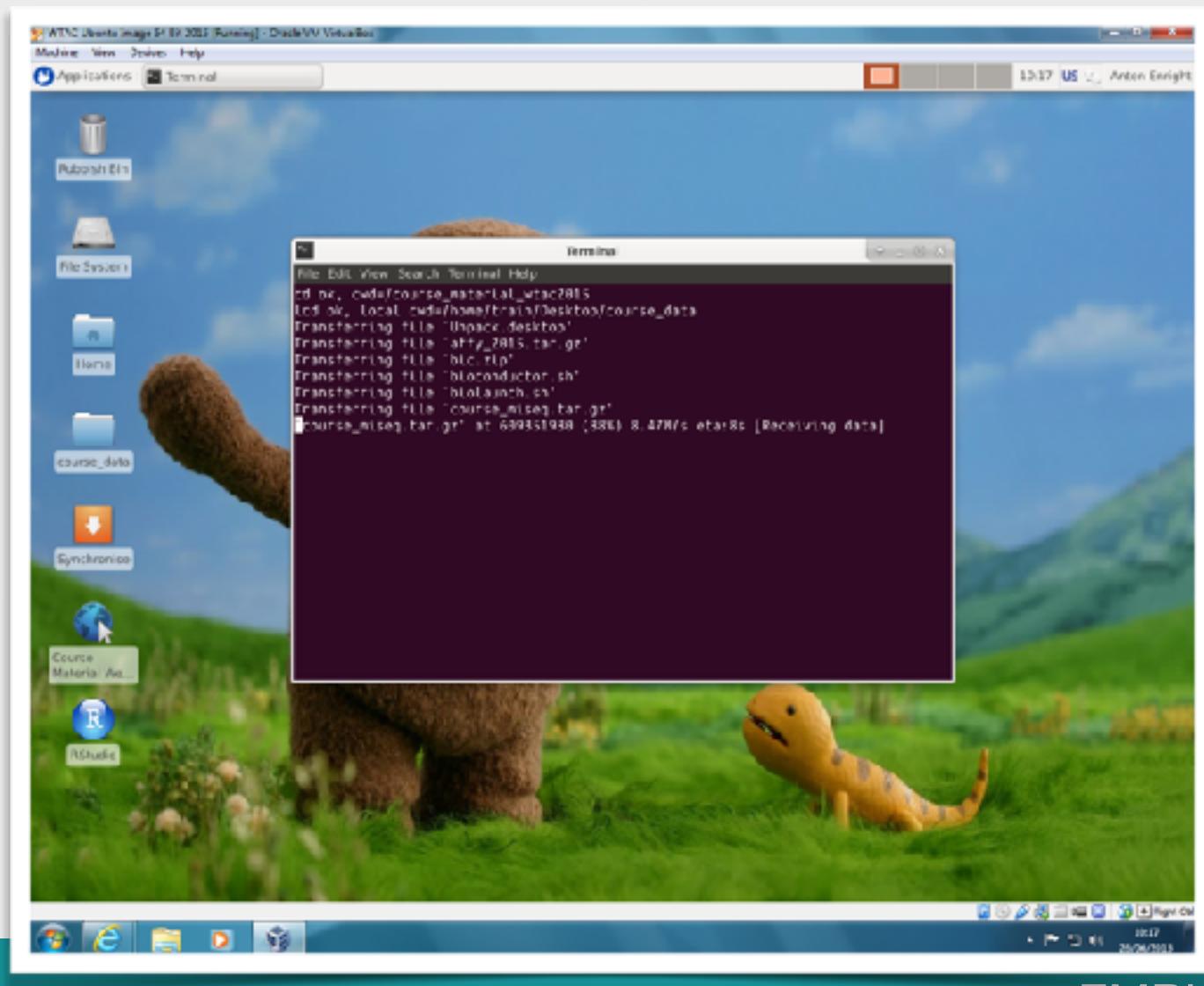
VirtualBox Installation



The screenshot shows the VirtualBox download page at virtualbox.org. The page features a large blue 'VirtualBox' logo at the top. Below it, a section titled 'Download VirtualBox' is shown. A message states: 'Here, you will find links to VirtualBox binaries and its source code.' Underneath, a section titled 'VirtualBox binaries' is present with the note: 'By downloading, you agree to the terms and conditions of the respective platforms.' There are two main bullet points: one for 'VirtualBox platform packages' (listing versions for Windows, OS X, Linux, and Solaris hosts) and one for 'VirtualBox 4.3.28 Oracle VM VirtualBox Extension Pack' (listing support for USB 2.0 devices, RDP, and PXE boot). Below these, a section for the 'VirtualBox 4.3.28 Software Developer Kit (SDK)' is shown, with a note about SHA256 and MD5 checksums for package integrity verification.

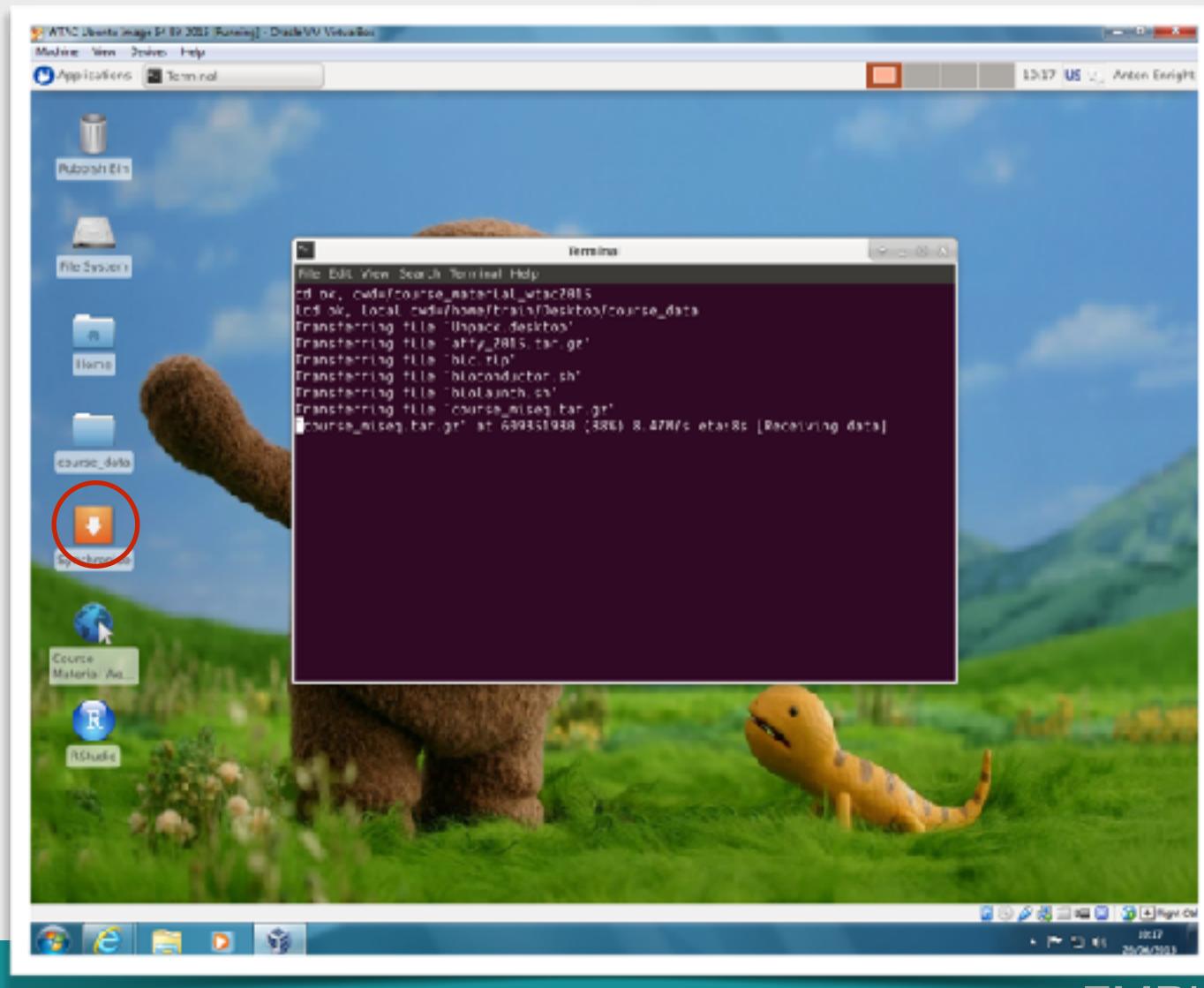
Course Image

- Virtual Box 8Gb Image + 15Gb Compressed Data



Course Image

- Virtual Box 8Gb Image + 15Gb Compressed Data



R & BioConductor

- Freely available
- Updated constantly
- Available here:
 - <http://www.r-project.org/>
 - <http://www.bioconductor.org/>
- DESeq2 Vignette and Manual Pages excellent also
- EdgeR
- Voom
- Full R courses available

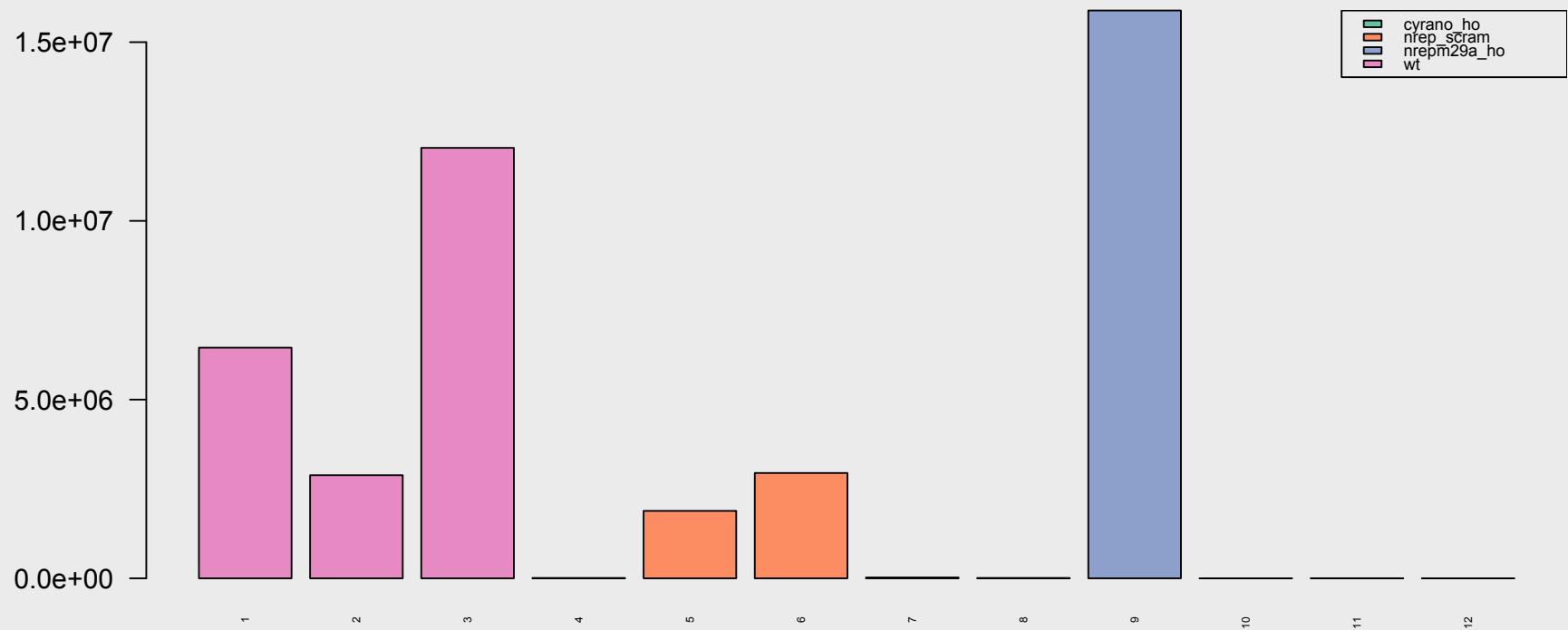


small RNA Seq Workflow

- **Read cleanup and QC in REAPER (Enrightlab R Module)**
 - Adapter Removal
 - Contaminant Removal
 - Size Selection 15-28nt
- **Read mapping using ChimiRa (Enrightlab)**
 - BLAST Based mapping against miRBase precursors
- **Read normalisation and differential expression analysis in R/BioConductor using DESeq2**
 - microRNA calls in multiple samples

Small RNA Seq

Pre Normalised Counts



Chimira Web Server for small RNA Seq

<http://wwwdev.ebi.ac.uk/enright-srv/chimira>

Bioinformatics Advance Access published June 20, 2015

Application Note

Chimira: Analysis of small RNA Sequencing data and microRNA modifications

Dimitrios M. Vitsios and Anton J. Enright*

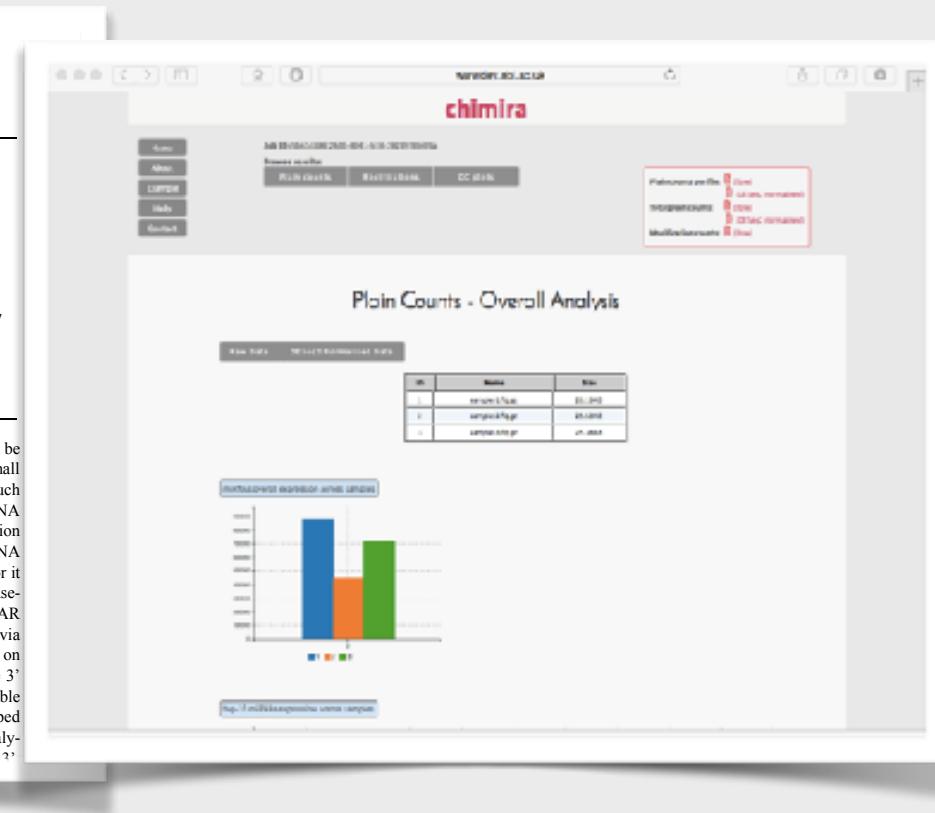
EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Associate Editor: Prof. Ivo Hofacker

ABSTRACT

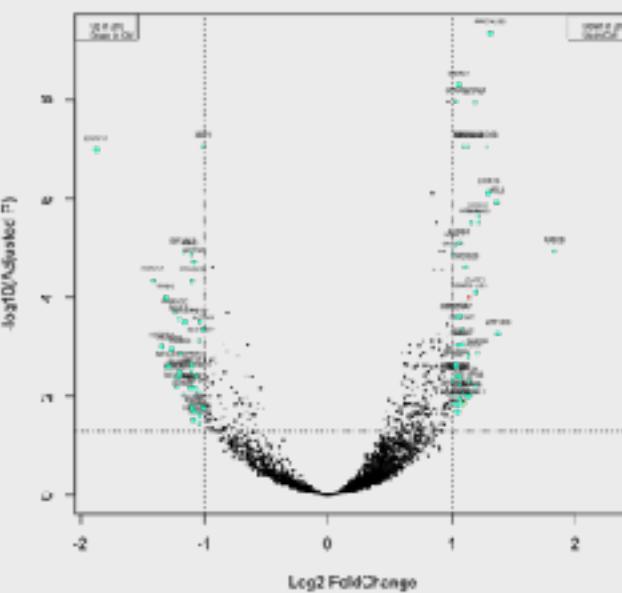
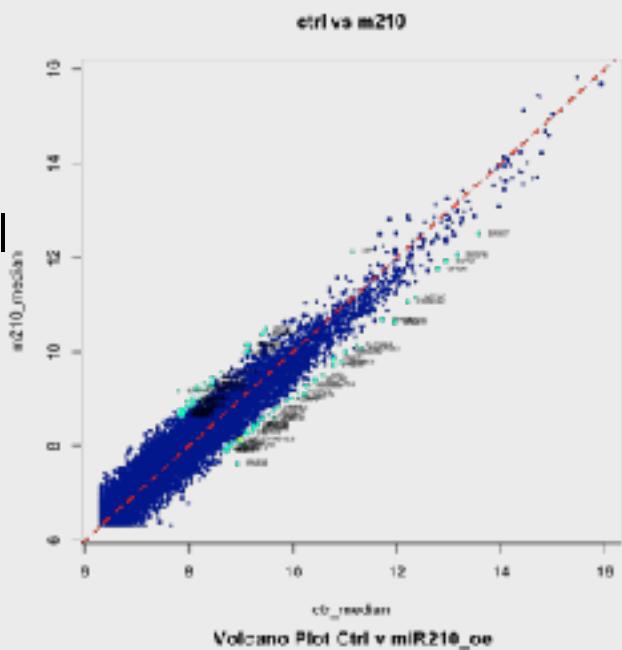
Summary: Chimira is a web-based system for microRNA (miRNA) analysis from small RNA-Seq data. Sequences are automatically cleaned, trimmed, size selected and mapped directly to miRNA hairpin sequences. This generates count-based miRNA expression data for subsequent statistical analysis. Moreover, it is capable of identifying epi-transcriptomic modifications in the input sequences. Supported modification types include multiple types of 3'-modifications (e.g. uridylation, adenylation), 5'-modifications and also internal modifications or variation (ADAR editing or SNPs). Besides cleaning and mapping of input sequences to miRNAs (Griffiths-Jones et al., 2008), Chimira provides a simple and intuitive set of tools for the analysis and interpretation of the results (see also Supplementary material). These allow the visual study of the differential expression between two specific samples or sets of samples, the identification of the most highly expressed miRNAs within sample pairs (or sets of

plore the full profile of all modifications and/or edits that can be identified in small RNA-Seq data. The functional roles of small RNAs in different conditions may be greatly influenced by such modifications. This can be accomplished by aligning small RNA sequences against their hairpin precursors. The alignment region spanning each miRNA is analysed to detect bases in the miRNA sequence that could not possibly have derived from the precursor it aligns to. These unalignable nucleotides are likely either: i) base-calling errors, ii) single nucleotide polymorphisms, iii) ADAR edits or iv) post-transcriptional miRNA modifications (e.g. via TUTases). Base-calling errors are pseudo-random depending on the platform used and usually more likely to occur towards the 3' end of sequences. In order to study this diverse pool of possible miRNA post-transcriptional modifications, we have developed Chimira. This is a cohesive platform for the processing and analysis of small RNA NGS data allowing simultaneous detection of 3'

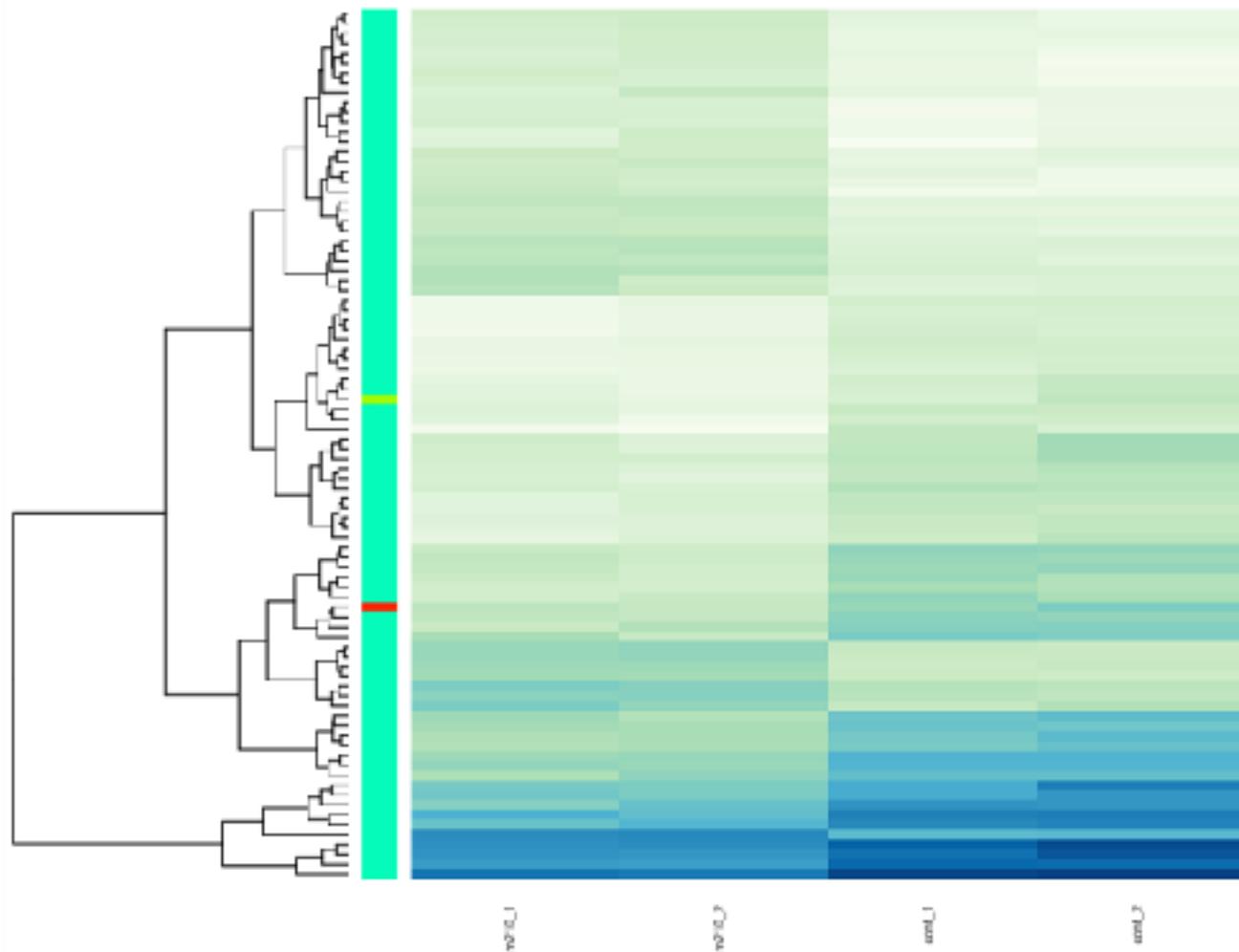


mRNA Seq Analysis

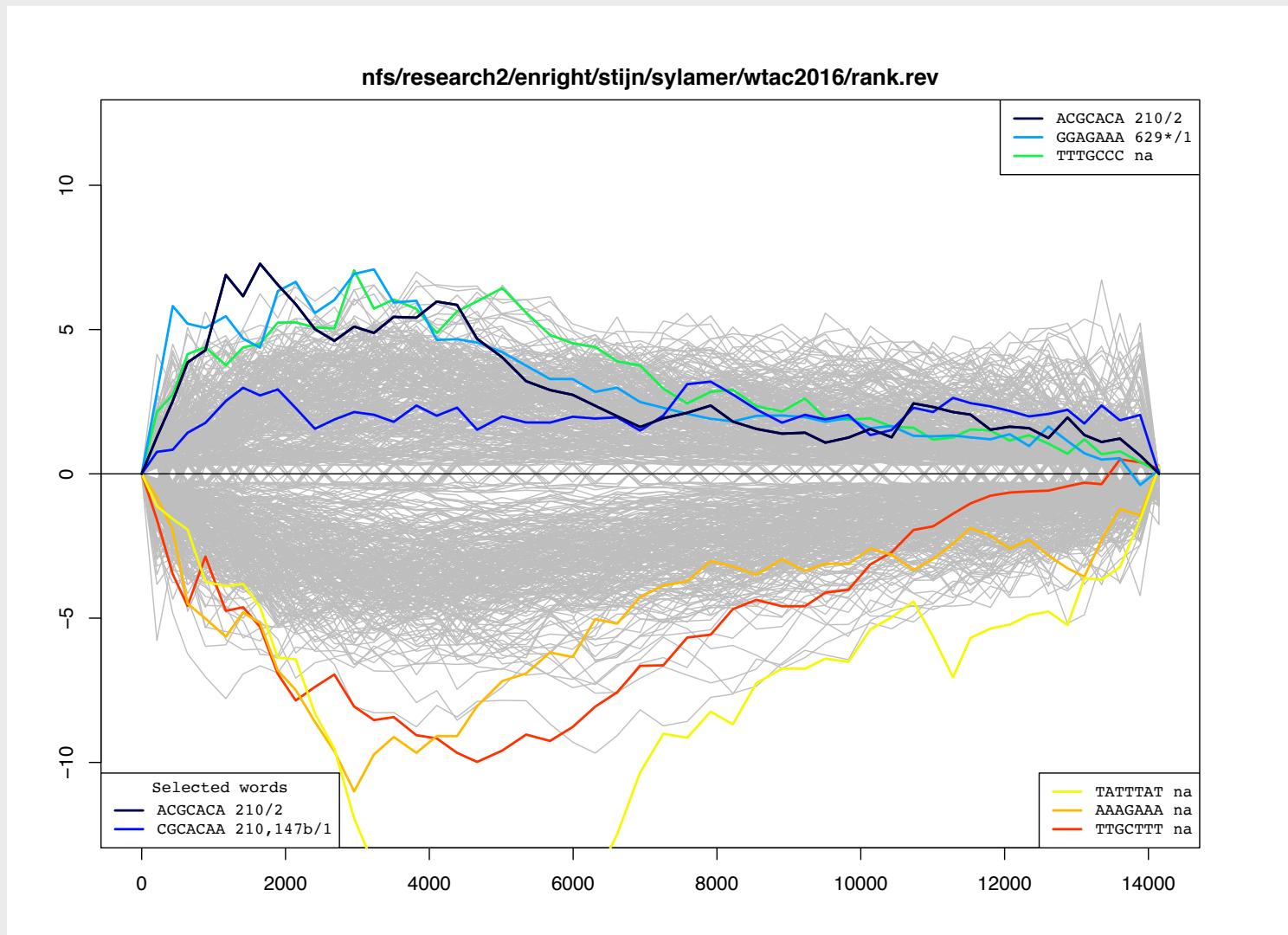
- HiSat2 vs Human Reference Genome from Ensembl
 - Does NOT run easily in windows, better in UNIX or Mac
- Genome Annotations from Ensembl GTF
- HTSeq-count and DESeq2 to perform differential expression
- Analysis of Results in R/BioConductor



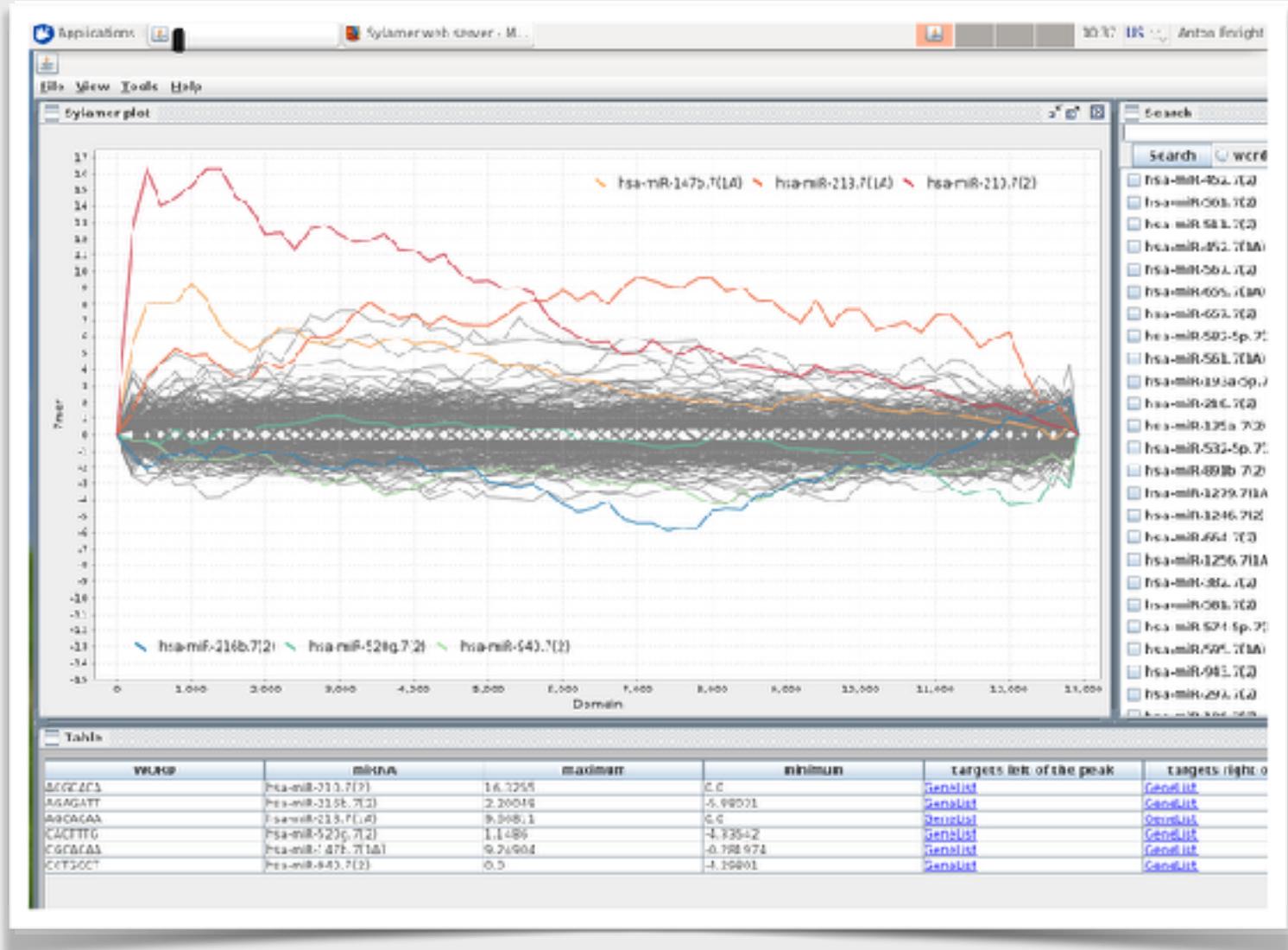
- HiSat2
- Do
- Genome
- HTSeq-express
- Analysis



Sylamer Analysis of miR-210 mRNA Seq



Sylamer Analysis of miR-210 mRNA Seq



mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg38

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2

Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg38

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

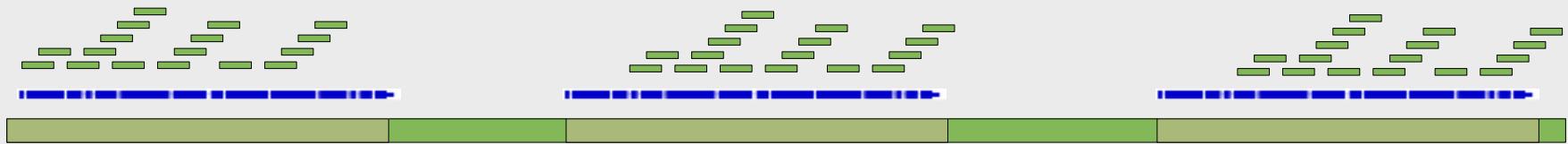
mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg38

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

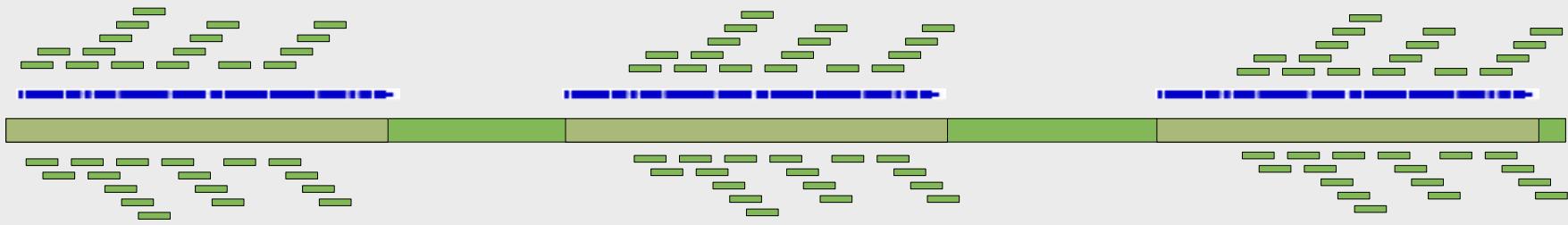
mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg38

HiSat2 Single end mode

HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2



Results in count data and FPKMs

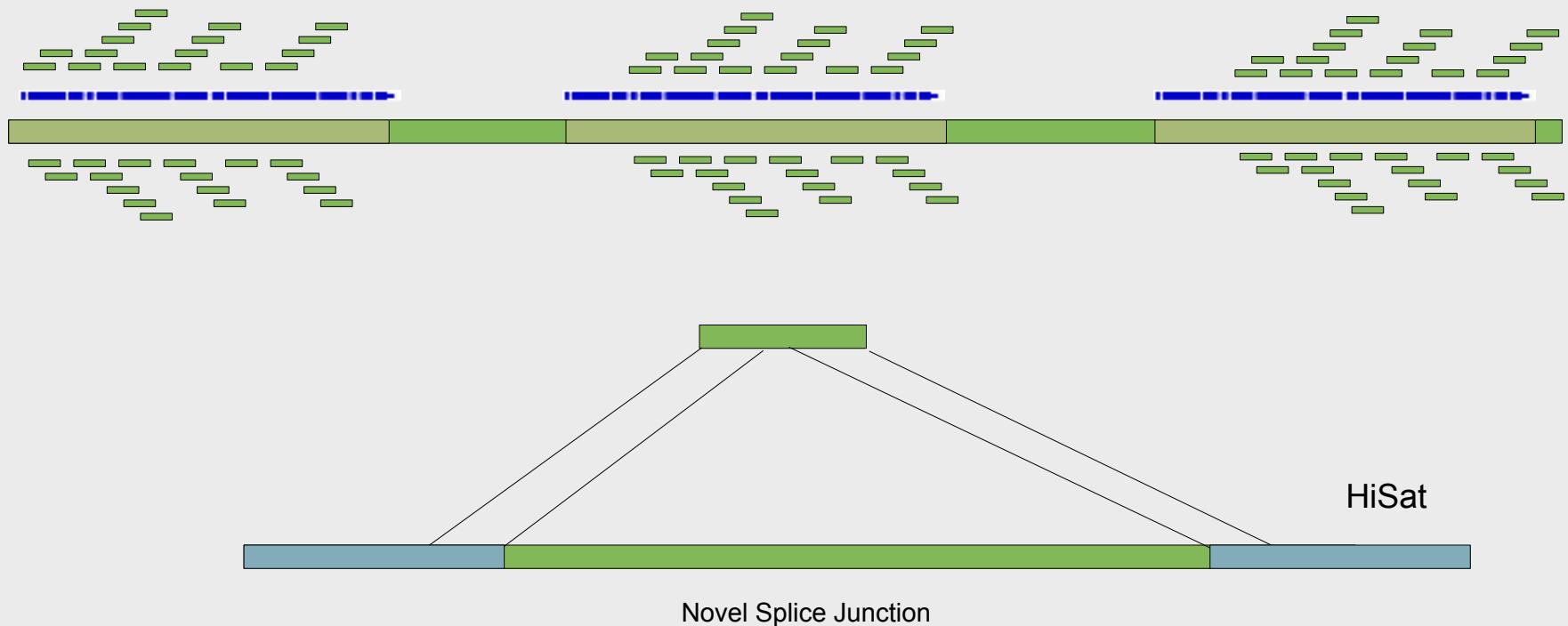
mRNA Seq: Analysis Types

Reference Transcriptome = Human Hg38

HiSat2 Single end mode

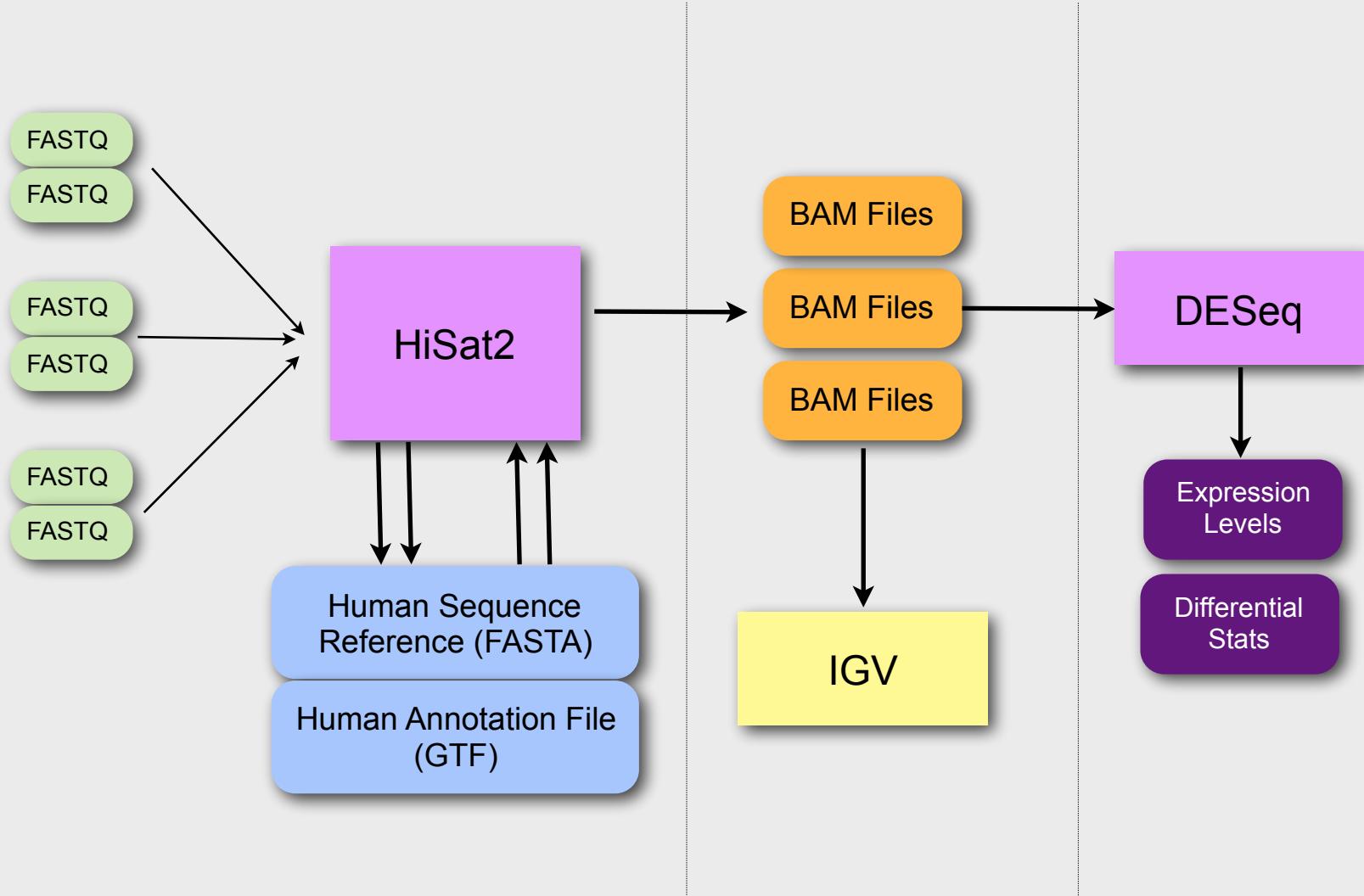
HTSeq-count Transcript Quantitation and Differential Statistics

R/BioConductor Downstream analysis with DESeq2

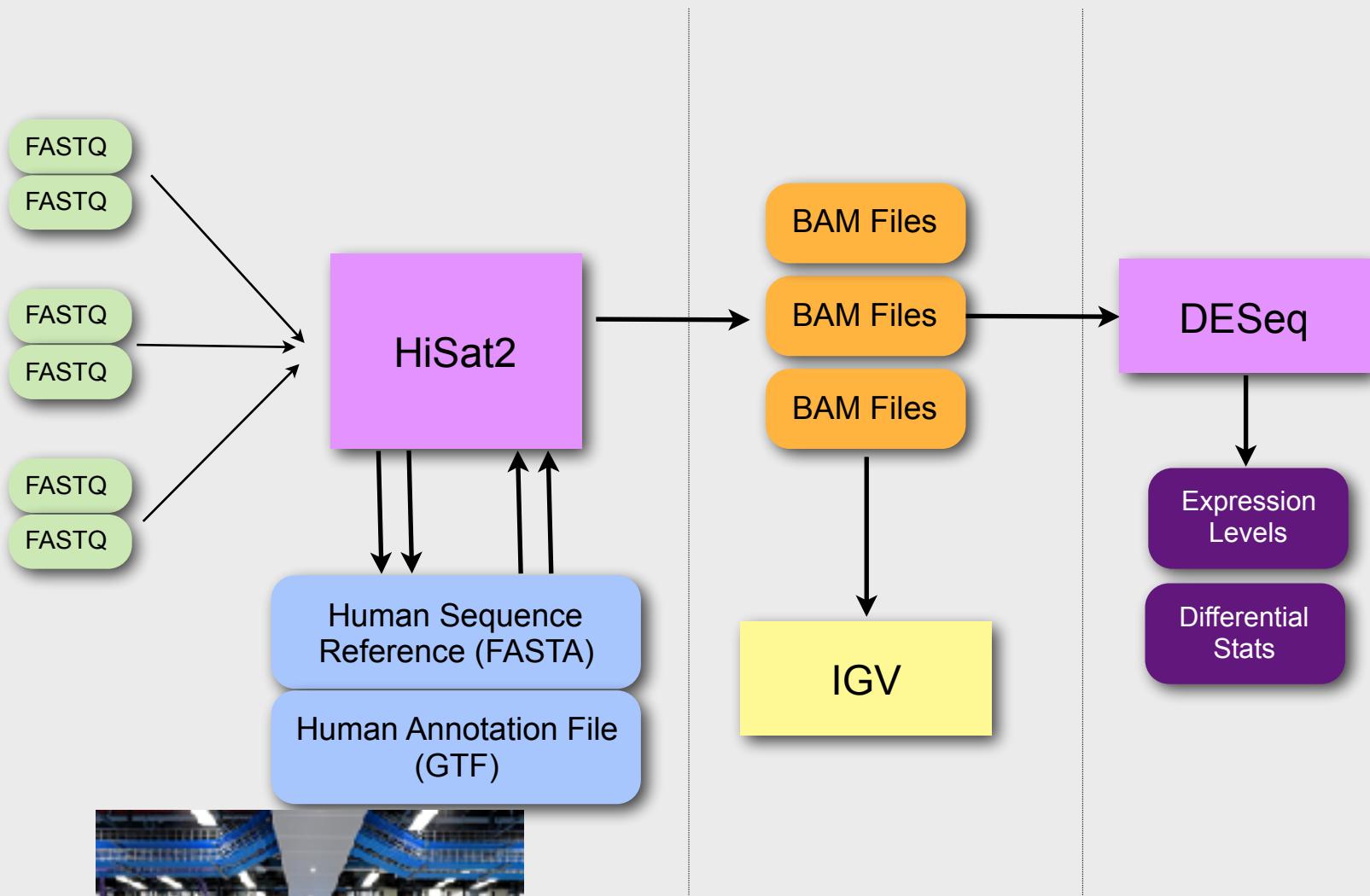


Results in count data and FPKMs

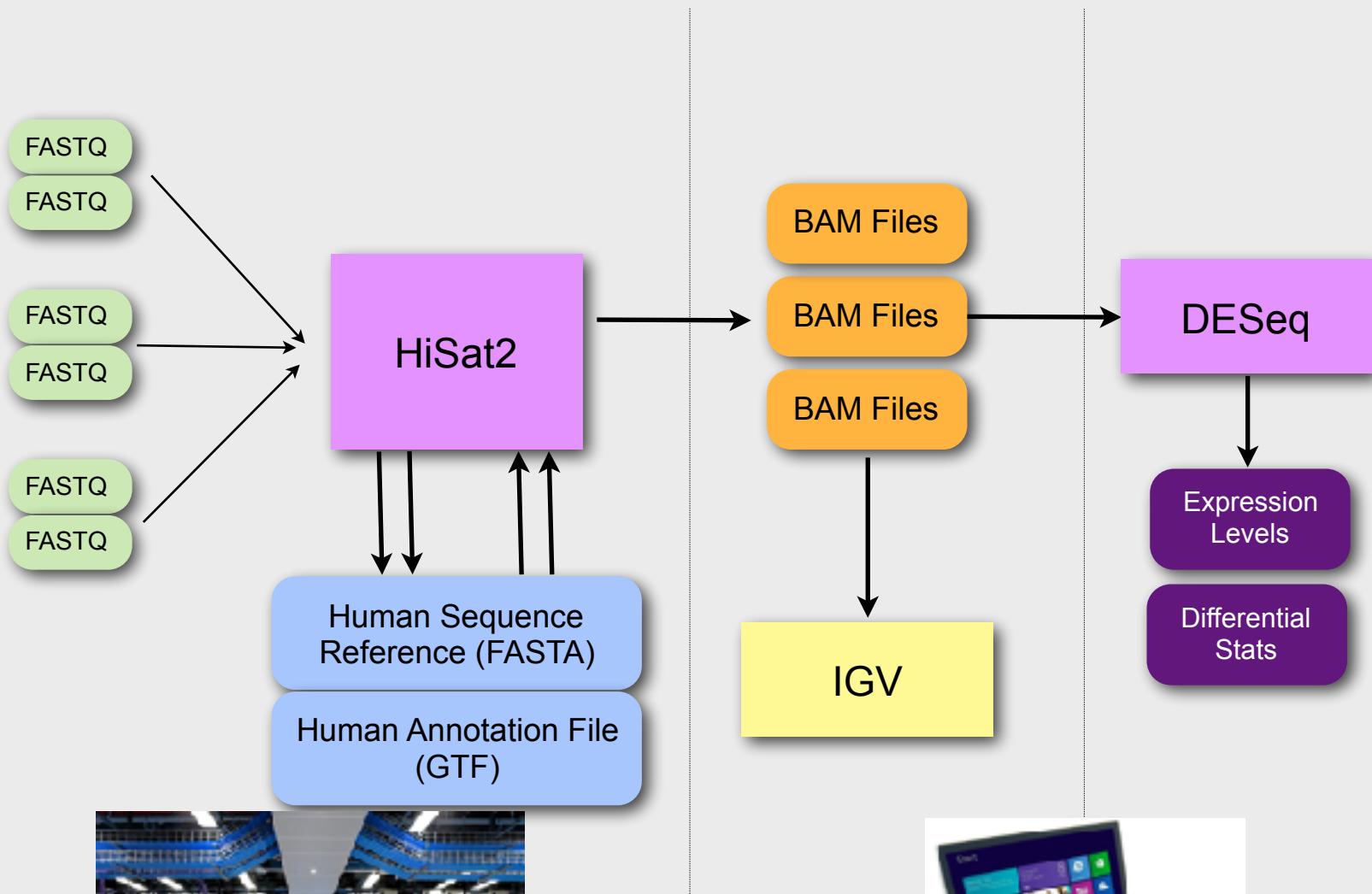
mRNA Seq Workflow



mRNA Seq Workflow



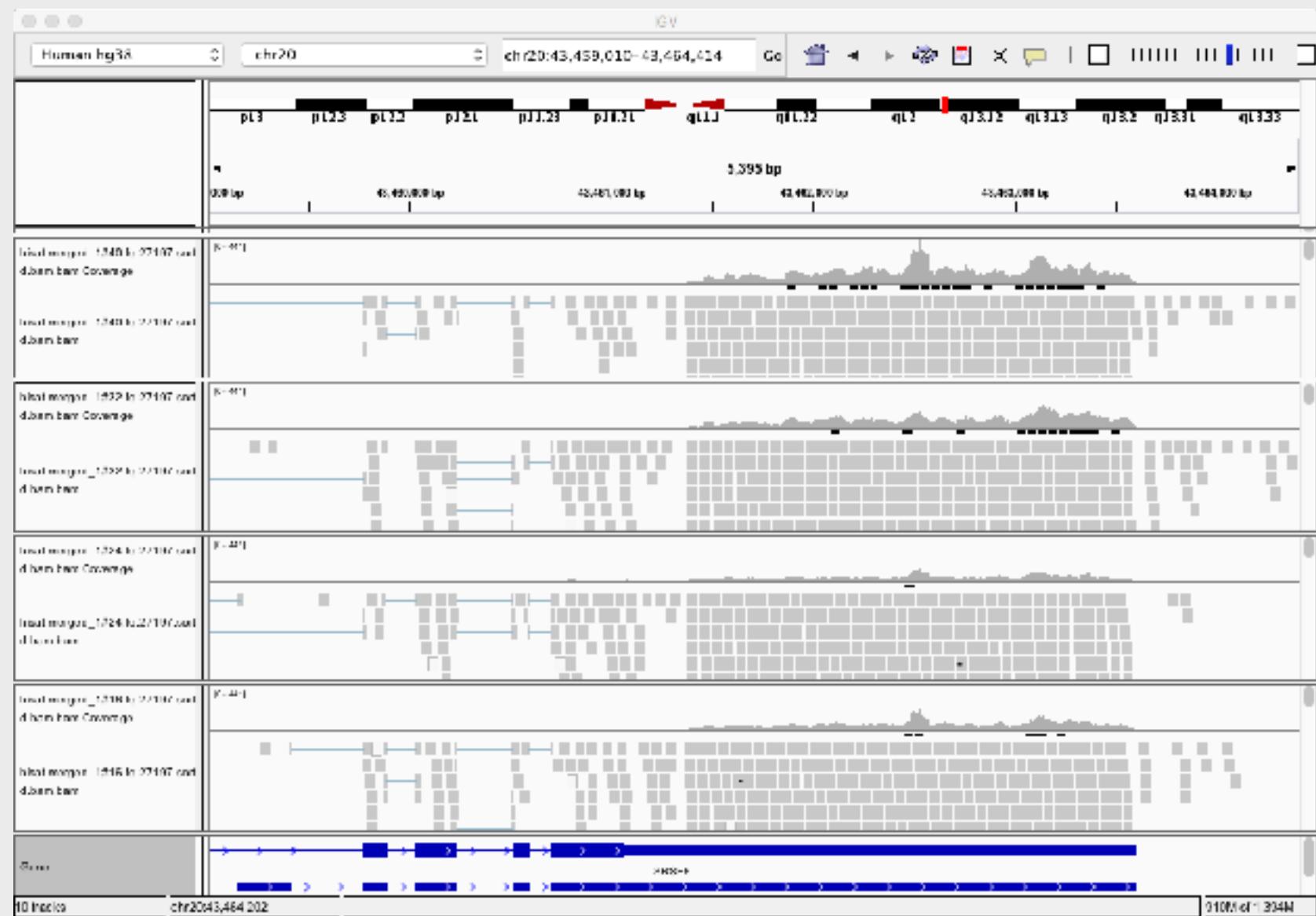
mRNA Seq Workflow



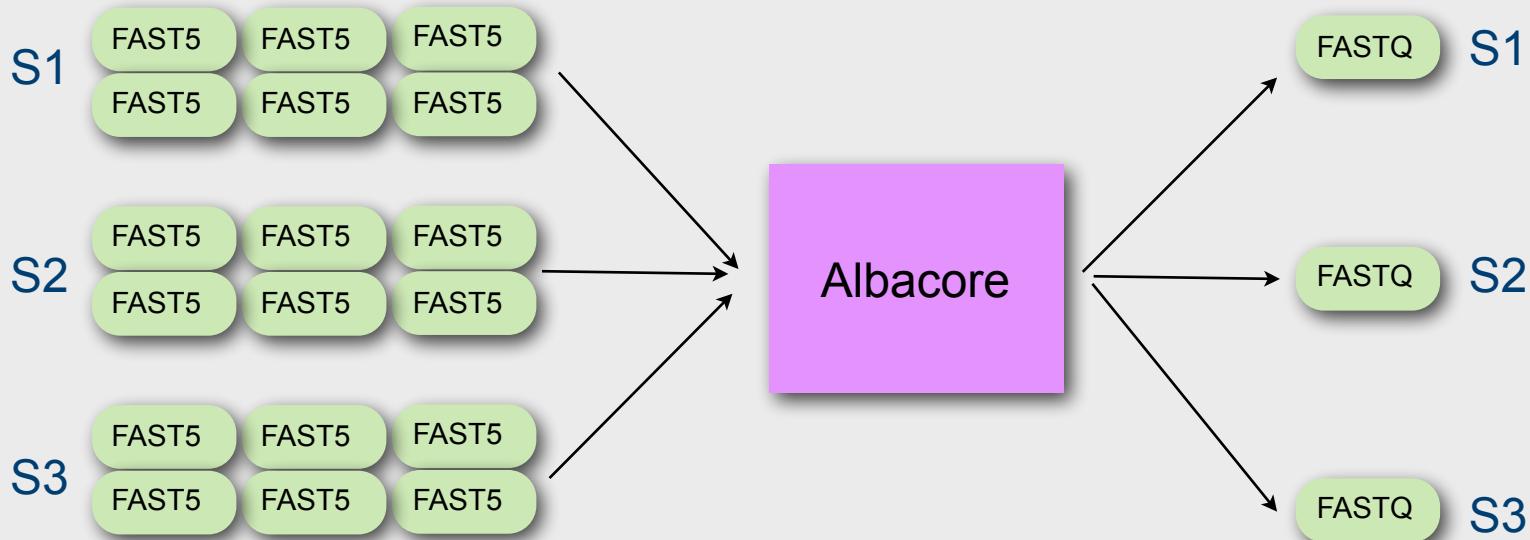
Mappers

- **Other Mappers Available**

- TopHat2 / Cufflinks
- STAR
- Kallisto
- Trinity - de novo assembly of transcriptomes
- Minimap2 - Nanopore mapping (long noisy reads)



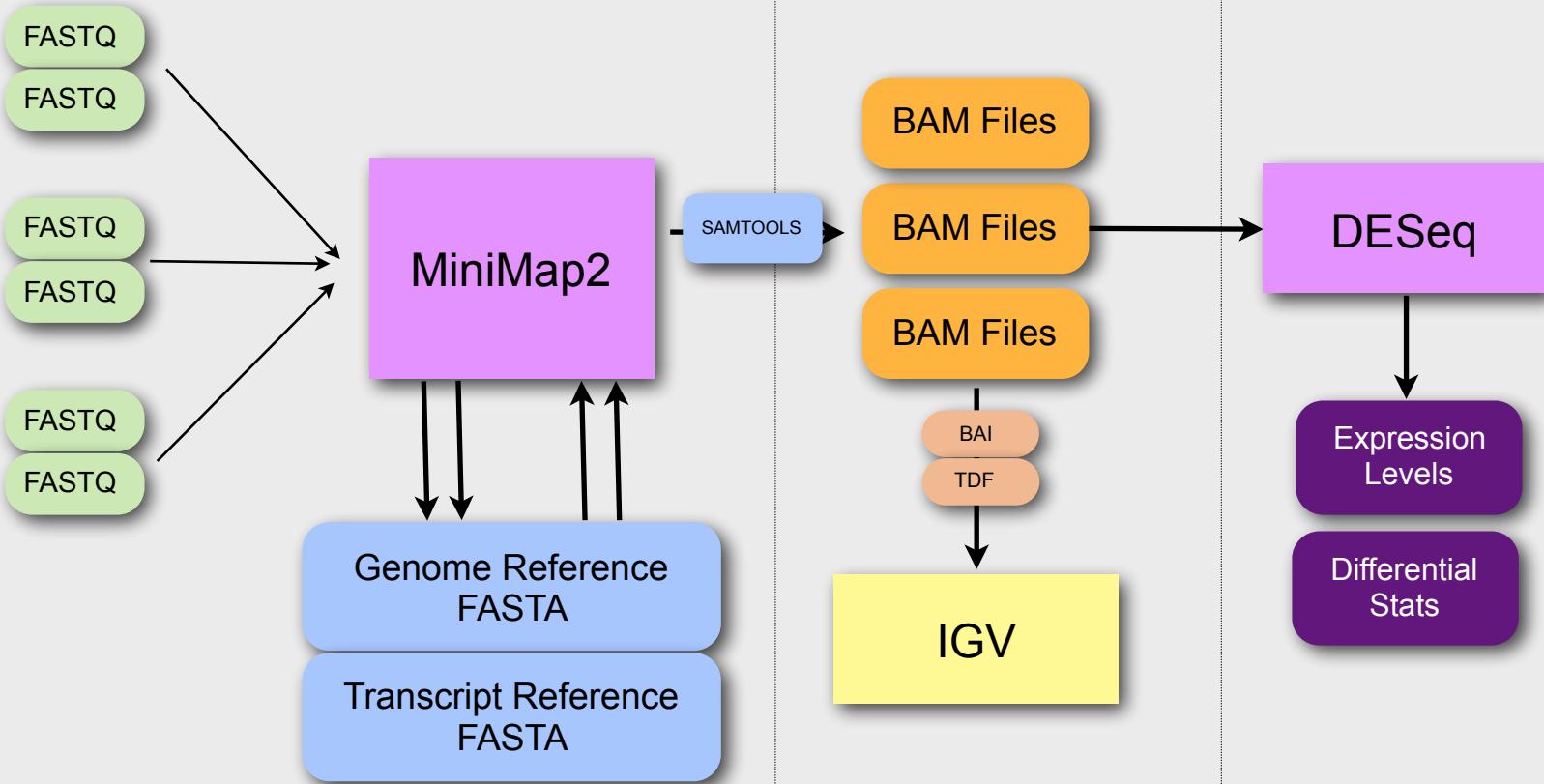
Nanopore Workflow - 1



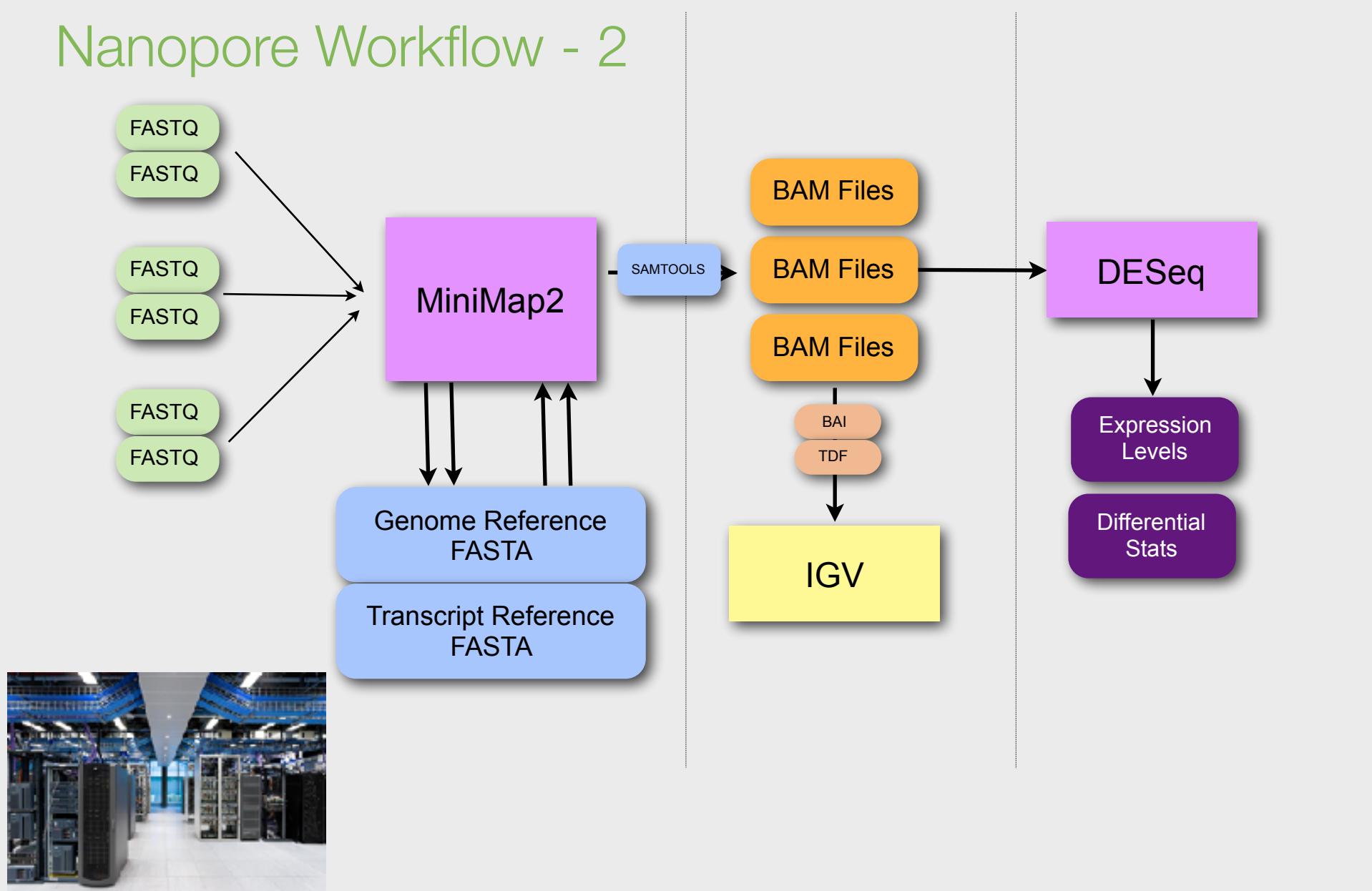
Generated 3.4Gigabases of sequence

Basecall folders contained between 200,000 and 2,000,000 FAST5 files
365 Gigabytes of raw FAST5 data
FASTQ files more manageable (3.5Gb)
Basecalling took >12 hours on 25 CPUs for the biggest sample
Basecalling while sequencing would be better

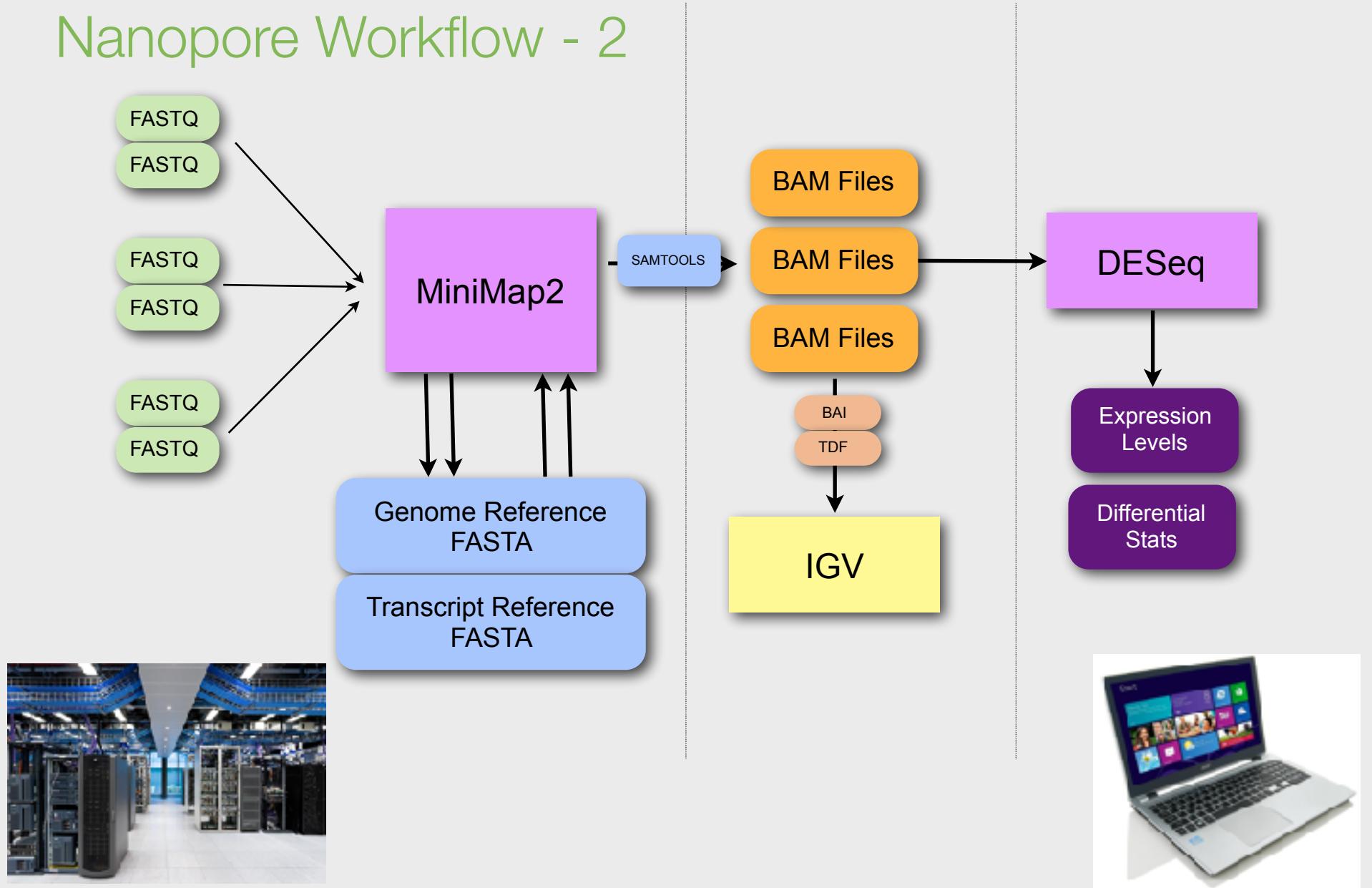
Nanopore Workflow - 2



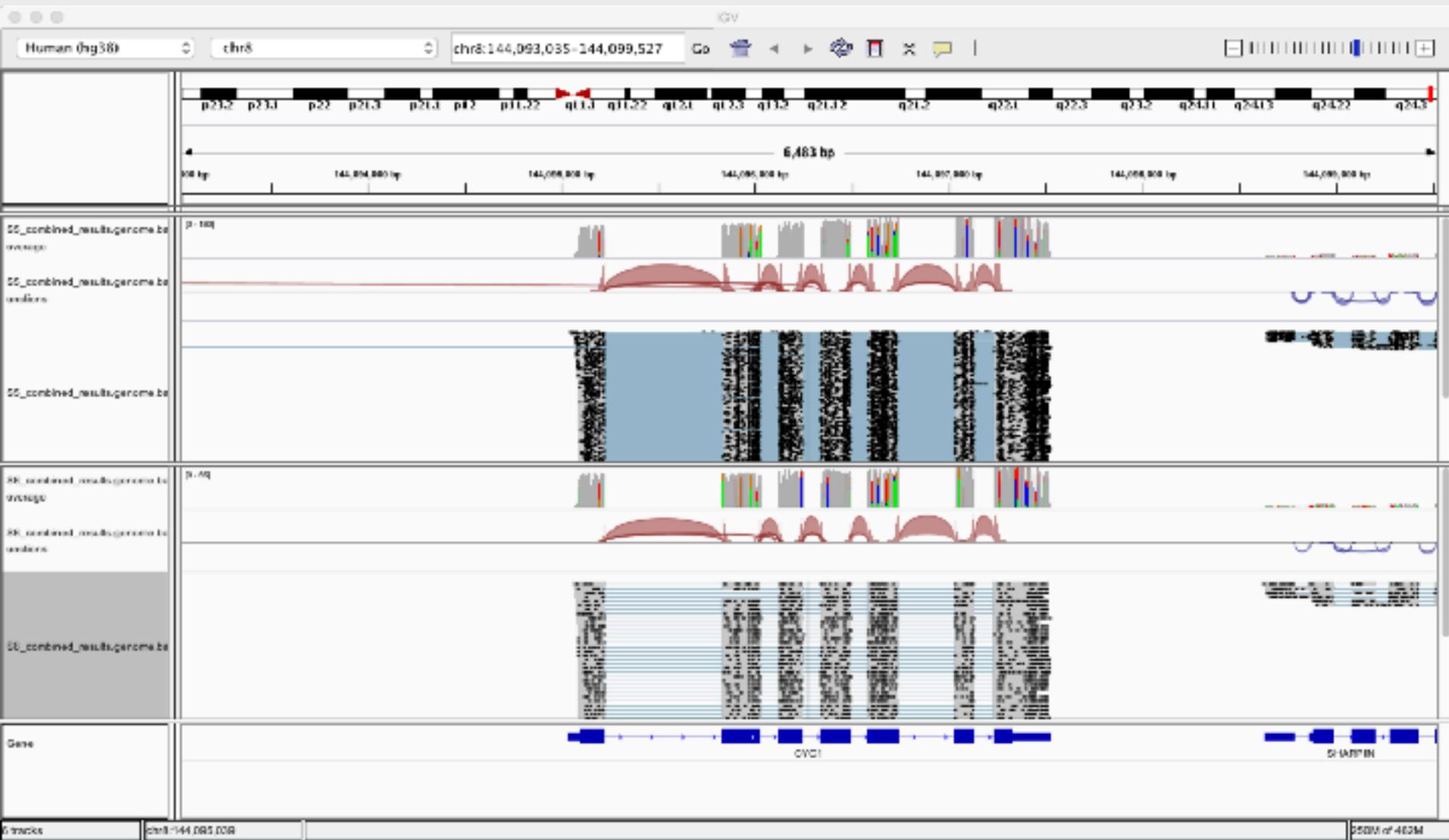
Nanopore Workflow - 2



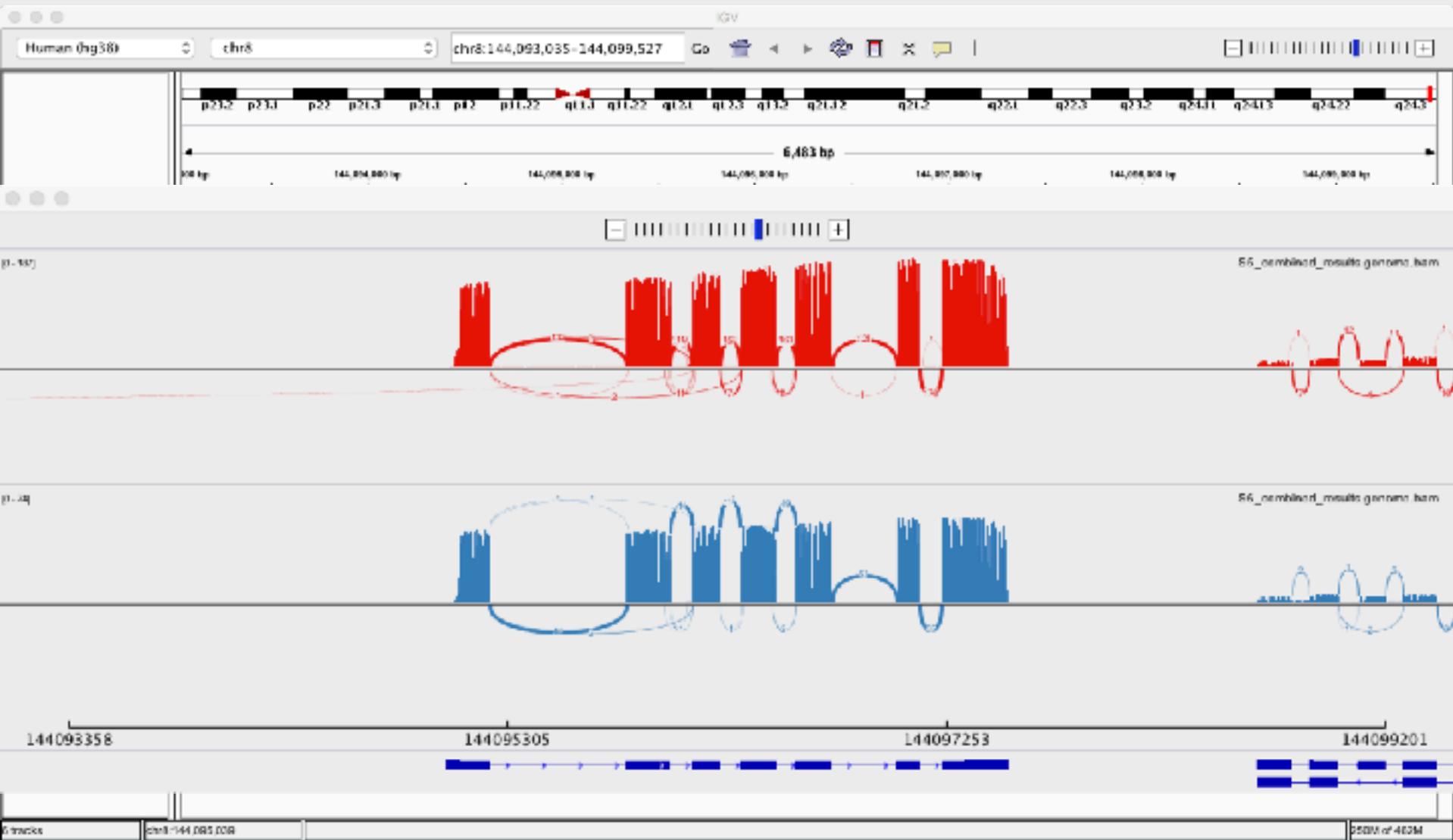
Nanopore Workflow - 2



DirectRNA Results



DirectRNA Results

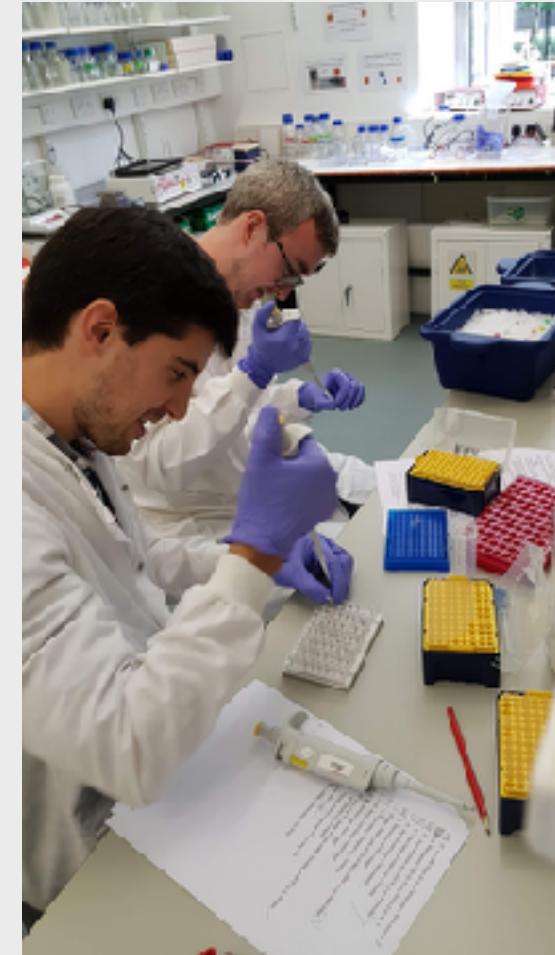


Experimental Findings - Direct RNA Seq

- Big differences in yields sample to sample
- Hard to normalise away the effect
- We will explore the MCF7 Data and look at it in more detail
- Yeast samples don't cluster well and had radically different yields
- We will explore RNA methylation changes in KO vs WT
- Alena and Allison producing a new set of mRNA-Seq (Illumina)

Direct RNA Sequencing - Conclusions

- Was a real struggle to prepare enough polyA+ RNA
 - Big thanks to Allison, Jack and Adrien
- PolyA+ Kit we used (NEB)
 - Maximum input of 5mg of Total RNA
 - We could have tried overloading the kit
- ThermoFisher Kit - Better Choice
 - Minimum input of 5mg
 - Maximum input of 100mg



Nanopore DirectRNA Seq

- polyA+ yields between 1-4%
- We did 2 polyA+ selections, probably one would be sufficient
- Nanopore recommends 600ng minimum 200ng....

Sample	Total RNA	polyA+	Yield Mb after 48hrs
S1	5mg	~ 100ng	260
S2	5mg	200ng	137
S3	5mg	160ng	120
S4	5mg	~ 100ng	251
S5	5mg	180ng	1222
S6	5mg	~ 100ng	607

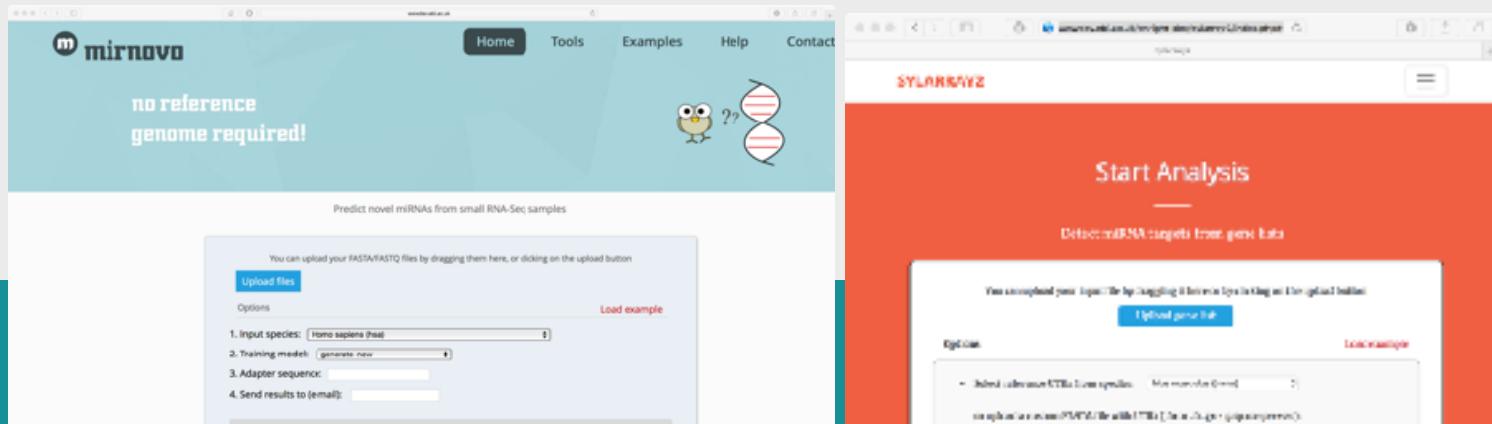
- Clearly the protocol works with less input material

Data and VMs

- Course Image available to download or to take away today
- The datasets are large: >20Gb
- All the practicals are available here:
 - www.tinyurl.com/wtac2018
- Many other courses and practicals available.

Links

- Our Practicals: tinyurl.com/wtac2018
- ChimiRa - MicroRNA Read Cleaning, Mapping and QC
 - <https://www.ebi.ac.uk/research/enright/software/chimira>
- MirNovo - Novel microRNA prediction
 - <http://wwwdev.ebi.ac.uk/enright-dev/mirnovo>
- Sylarray2 - Sylamer analysis on genelists (DEVELOPMENT)
 - <http://wwwdev.ebi.ac.uk/enright-dev/sylarray2/index.php>



Questions

Email:

aje39@cam.ac.uk

aleg@ebi.ac.uk

monahanj@ebi.ac.uk