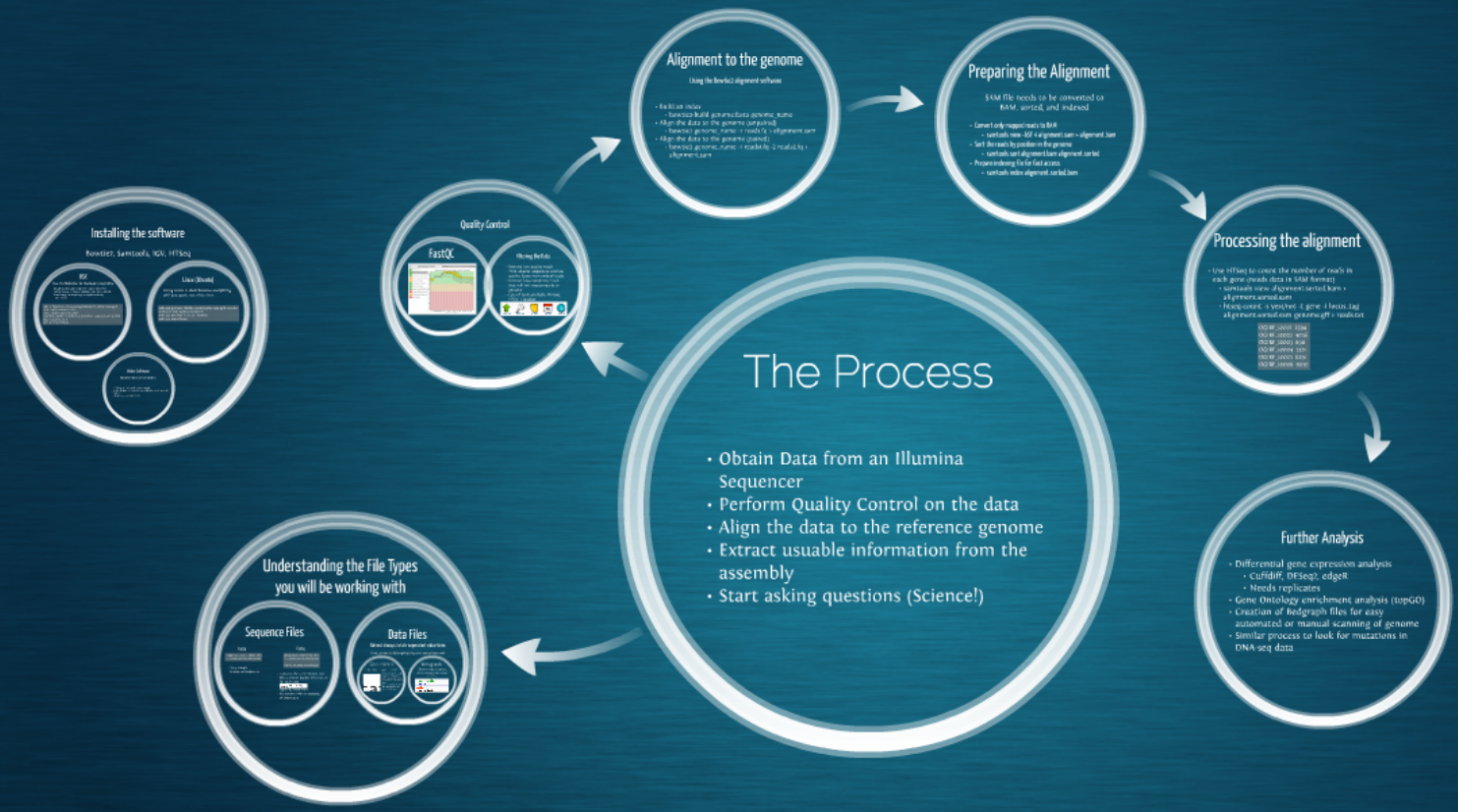
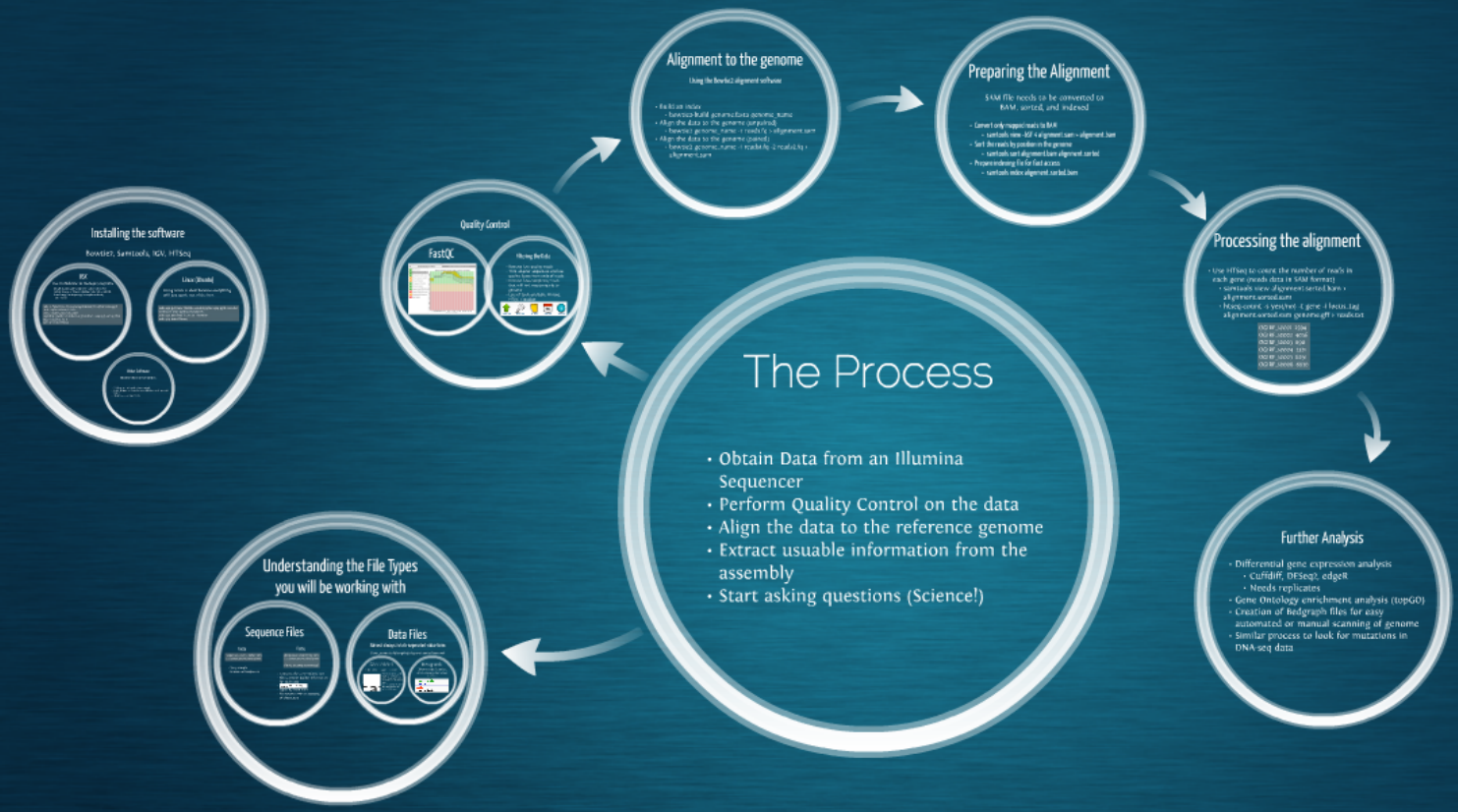


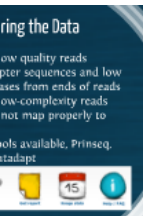
Analysis of Bacterial Transcriptome Data

Jonathan Goodson



Analysis of Bacterial Transcriptome Data

Jonathan Goodson



The Process

- Obtain Data from an Illumina Sequencer
- Perform Quality Control on the data
- Align the data to the reference genome
- Extract usuable information from the assembly
- Start asking questions (Science!)

Understanding the File Types you will be working with

Sequence Files

Fasta

```
>Sequence Name - Other info
ATCGTAGCGGATCGTAGCGAT
```

- Very simple
- Name and sequence

Fastq

```
@Sequence_identifying_info
ATCGTAGCGGATCGTAGCGAT
+
I?F0=CCHCDH@FGGHHBE@?
```

- Contains the same information
- Also contains quality information for each base
- $Q_{\text{sanger}} = -10 \log_{10} p$
- Typically from 0-40
- Represents 0-40 as sequence of characters

Data Files

Almost always in tab-seperated value form

Gene_name(tab)length(tab)position(tab)strand

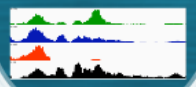
SAM/BAM

Contains data for each read mapping (or not) to a genome

- Very large files (>500MB)
- Center of all analysis
- BAM files are computer-readable formatted SAM files
- Can be sorted/indexed for high speed access and reduced disk space

Bedgraph

Tab-seperated value file, each line contains a span of the genome (bases 1-24) and a number of reads that map to each base



Sequence Files

Fasta

```
>Sequence Name - Other info
ATCGTAGCGGATCGTAGCGAT
```

- Very simple
- Name and sequence

Fastq

```
@Sequence_identifying_info
ATCGTAGCGGATCGTAGCGAT
+
I?F0=CCHCDH@FGGHHBE@?
```

- Contains the same information
- Also contains quality information for each base
- $Q_{\text{sanger}} = -10 \log_{10} p$
- Typically from 0-40
- Represents 0-40 as sequence of characters

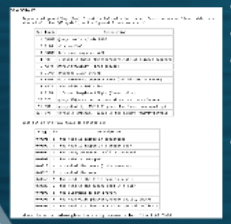
Data Files

Almost always in tab-seperated value form

Gene_name(tab)length(tab)position(tab)strand

SAM/BAM

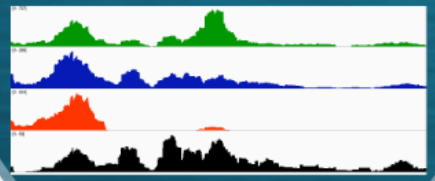
Contains data for each read mapping (or not) to a genome



- Very large files (>500Mb)
- Center of all analysis
- BAM files are computer-readable formatted SAM files
- Can be sorted/indexed for high speed access and reduced disk space

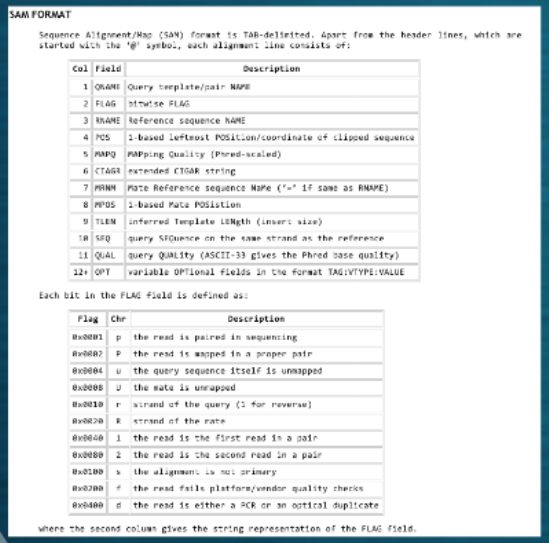
Bedgraph

Tab-seperated value file, each line contains a span of the genome (bases 1-24) and a number of reads that map to each base.



SAM/BAM

Contains data for each read mapping (or not) to a genome



SAM FORMAT

Sequence Alignment/Map (SAM) format is TAB-delimited. Apart from the header lines, which are started with the '@' symbol, each alignment line consists of:

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSition
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Each bit in the FLAG field is defined as:

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

where the second column gives the string representation of the FLAG field.

- Very large files (>500Mb)
- Center of all analysis
- BAM files are computer-readable formatted SAM files
- Can be sorted/indexed for high speed access and reduced disk space

SAM FORMAT

Sequence Alignment/Map (SAM) format is TAB-delimited. Apart from the header lines, which are started with the '@' symbol, each alignment line consists of:

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Each bit in the FLAG field is defined as:

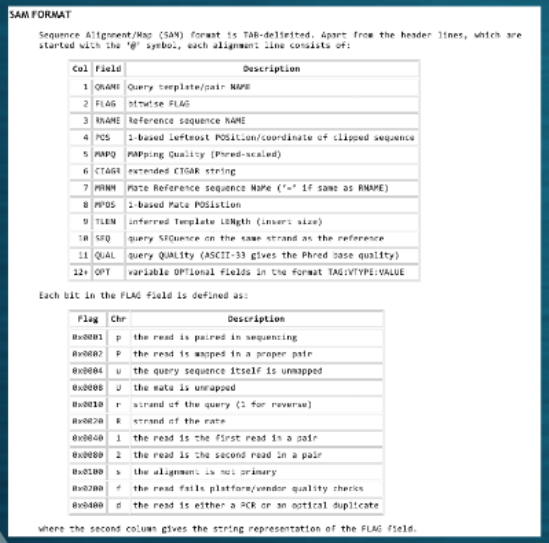
Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

where the second column gives the string representation of the FLAG field.

- Ve
- Ce
- BA
  re
  fil
- Ca
  hig
  re

SAM/BAM

Contains data for each read mapping (or not) to a genome



SAM FORMAT

Sequence Alignment/Map (SAM) format is TAB-delimited. Apart from the header lines, which are started with the '@' symbol, each alignment line consists of:

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSition
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Each bit in the FLAG field is defined as:

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

where the second column gives the string representation of the FLAG field.

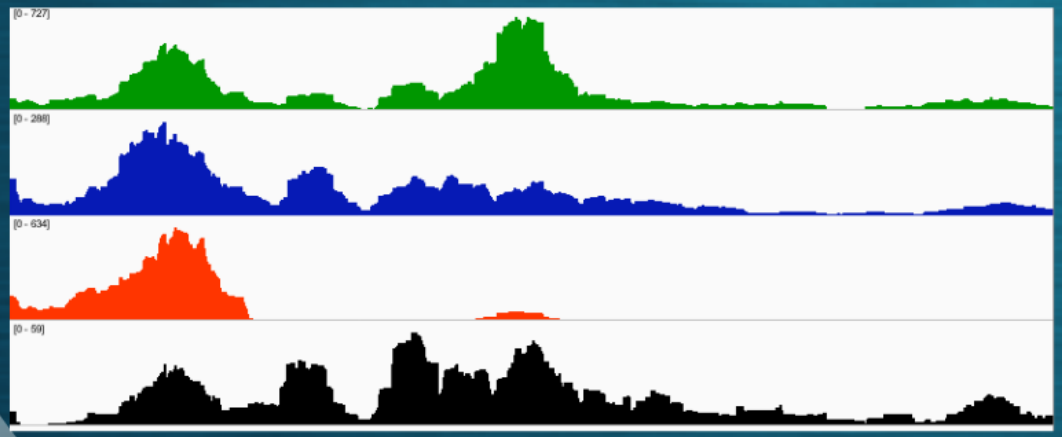
- Very large files (>500Mb)
- Center of all analysis
- BAM files are computer-readable formatted SAM files
- Can be sorted/indexed for high speed access and reduced disk space

oMb)
is
uter-
SAM

ed for
nd

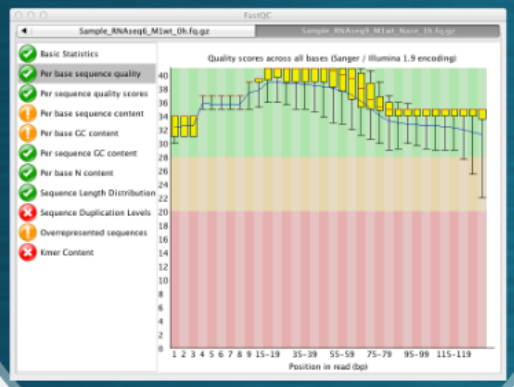
Bedgraph

Tab-seperated value file, each line contains a span of the genome (bases 1-24) and a number of reads that map to each base.



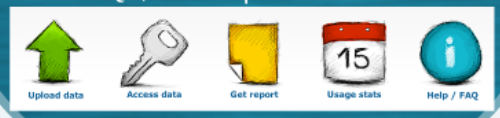
Quality Control

FastQC

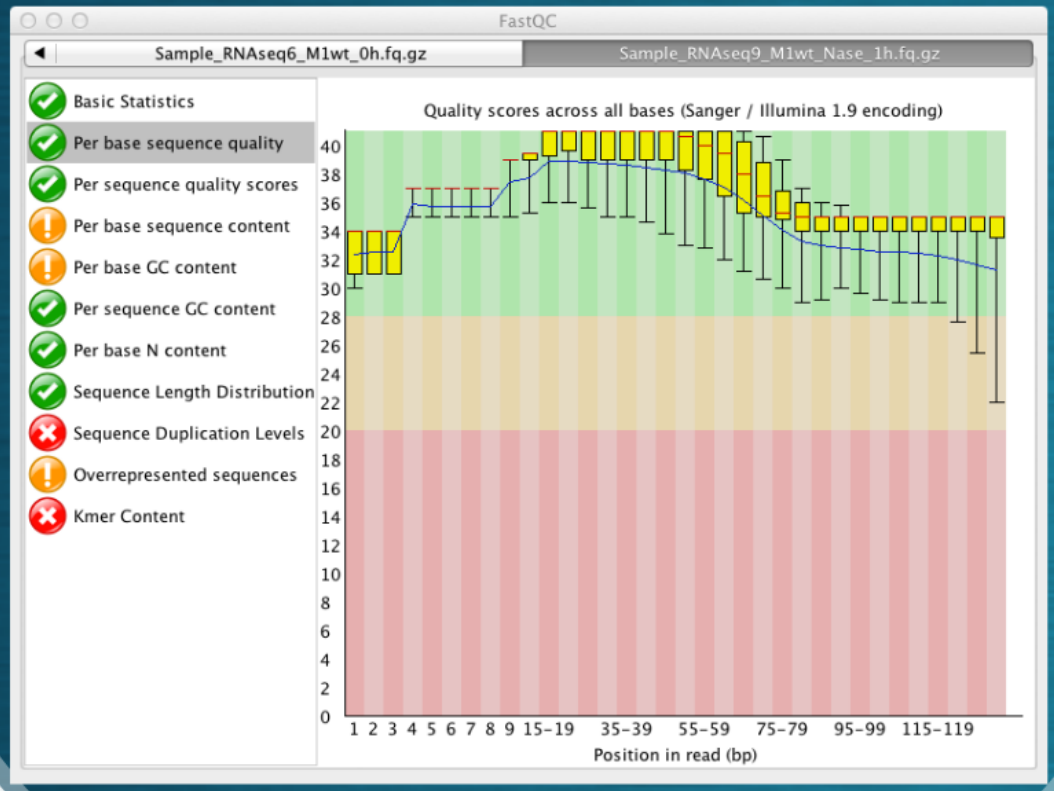


Filtering the Data

- Remove low quality reads
- Trim adapter sequences and low quality bases from ends of reads
- Remove low-complexity reads that will not map properly to genome
- Lots of tools available, Prinseq, HTQC, Cutadapt

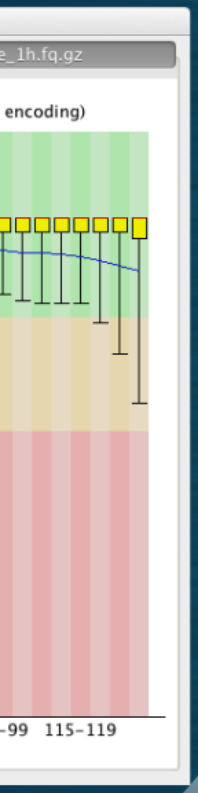


FastQC



- Re
- Tr
  qu
- Re
  th
  ge
- Lo
  HT

Filtering the Data



- Remove low quality reads
- Trim adapter sequences and low quality bases from ends of reads
- Remove low-complexity reads that will not map properly to genome
- Lots of tools available, Prinseq, HTQC, Cutadapt

Alignment to the genome

Using the Bowtie2 alignment software

- Build an index
  - bowtie2-build genome.fasta genome_name
- Align the data to the genome (unpaired)
  - bowtie2 genome_name -r reads.fq › alignment.sam
- Align the data to the genome (paired)
  - bowtie2 genome_name -1 reads1.fq -2 reads2.fq › alignment.sam

Preparing the Alignment

SAM file needs to be converted to BAM, sorted, and indexed

- Convert only mapped reads to BAM
  - samtools view -bSF 4 alignment.sam > alignment.bam
- Sort the reads by position in the genome
  - samtools sort alignment.bam alignment.sorted
- Prepare indexing file for fast access
  - samtools index alignment.sorted.bam

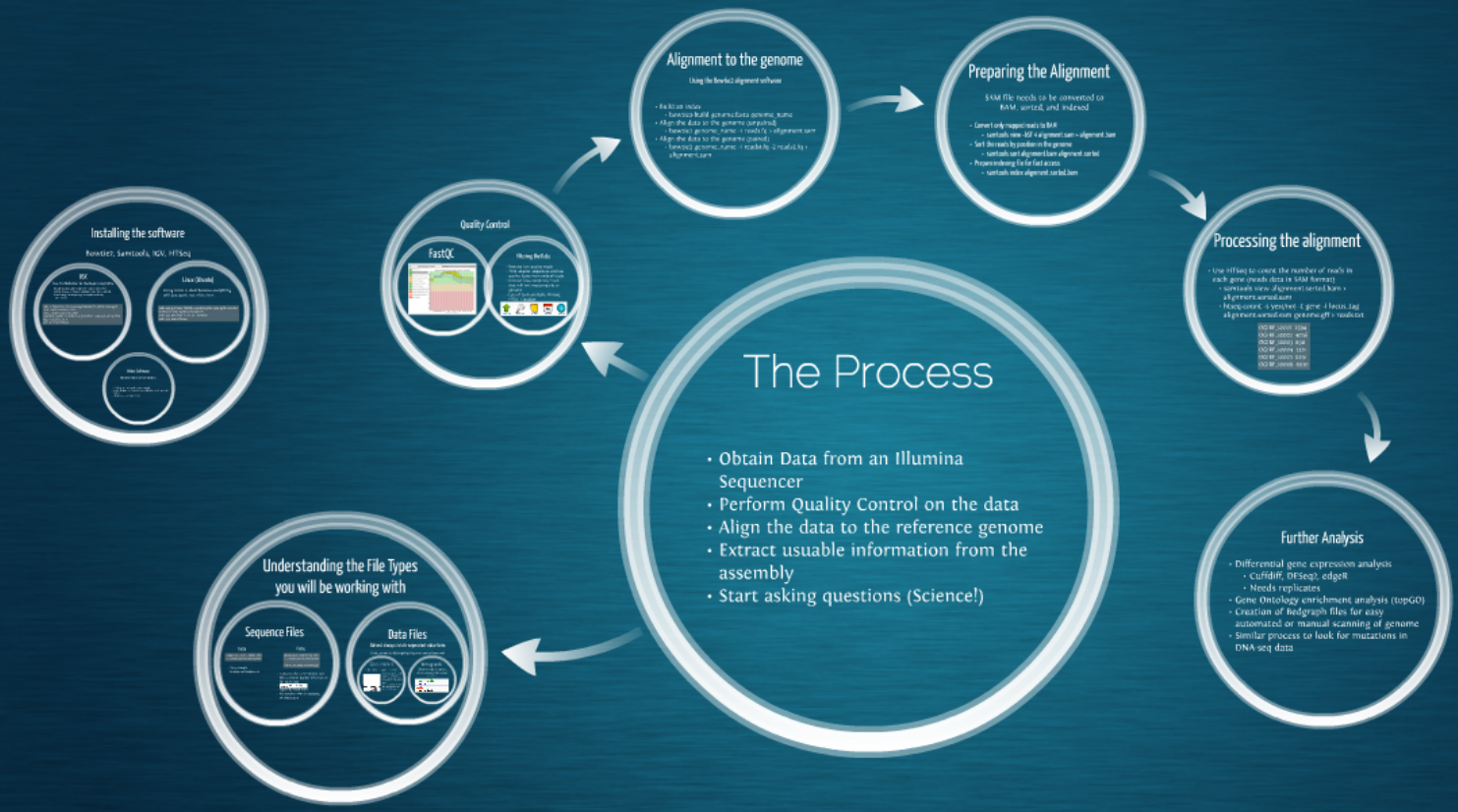
Processing the alignment

- Use HTSeq to count the number of reads in each gene (needs data in SAM format)
  - samtools view alignment.sorted.bam › alignment.sorted.sam
  - htseq-count -s yes(/no) -t gene -i locus_tag alignment.sorted.sam genome.gff › reads.txt

```
OG1RF_10001  2594
OG1RF_10002  4056
OG1RF_10003  890
OG1RF_10004  2321
OG1RF_10005  6291
OG1RF_10006  8930
```

Further Analysis

- Differential gene expression analysis
  - Cuffdiff, DESeq2, edgeR
  - Needs replicates
- Gene Ontology enrichment analysis (topGO)
- Creation of Bedgraph files for easy automated or manual scanning of genome
- Similar process to look for mutations in DNA-seq data



Analysis of Bacterial Transcriptome Data

Jonathan Goodson

Installing the software

Bowtie2, Samtools, IGV, HTSeq

OSX

Use Homebrew to manage programs

Install XCode Command Line Tools from either within XCode or the standalone installers available from http://developer.apple.com/downloads/index.action

```
ruby -e "$(curl -fsSL https://raw.github.com/mxcl/homebrew/go)"
brew tap homebrew/science
brew install samtools bowtie2
wget http://python-distribute.org/distribute_setup.py | sudo python
sudo easy_install pip
sudo pip install htseq
```

Linux (Ubuntu)

Using Linux is ideal because everything will just work out of the box.

```
sudo apt-get install build-essential python-pip python2.7-dev
python-numpy python-matplotlib
sudo apt-get install samtools bowtie2
sudo pip install htseq
```

Other Software

Most other software comes with installers

- IGV (http://www.broadinstitute.org/igv/)
- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- R (http://www.r-project.org/)

OSX

Use Homebrew to manage programs

Install XCode Command Line Tools from either within XCode or the standalone installers available from http://developer.apple.com/downloads/index.action

```
ruby -e "$(curl -fsSL https://raw.github.com/mxcl/homebrew/go)"
brew tap homebrew/science
brew install samtools bowtie2
wget http://python-distribute.org/distribute_setup.py | sudo python
sudo easy_install pip
sudo pip install htseq
```

Linux (Ubuntu)

Using Linux is ideal because everything will just work out of the box.

```
sudo apt-get install build-essential python-pip python2.7-dev
python-numpy python-matplotlib
sudo apt-get install samtools bowtie2
sudo pip install htseq
```

Other Software

Most other software comes with installers

- IGV (http://www.broadinstitute.org/igv/)
- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- R (http://www.r-project.org/)