TGAC
The Genome Analysis Centre™
BBSRC Greater Norwich Development Partnership

Building Excellence in Genomics and Computational Bioscience

www.tgac.ac.uk @GenomeAnalysis

Next generation sequencing data analysis for metagenomics

Richard Leggett
Richard.Leggett@tgac.ac.uk
@richardmleggett

BBSRC bioscience for the future
Greater Norwich Development Partnership
South Norfolk COUNCIL
NORWICH City Council
Norfolk County Council

Typical pipelines

QC
FASTQC
PRINSEQ
PCR duplicate removal

Subsampling
Scripts
Digital normalisation

Merge overlapping
FLASH
PEAR

Read Trimming
trimmomatic
Sickle
FASTX-Toolkit

Contamination screening
kontaminant
BLAST

LOTS more!

pplacer
Phymm

Sequence placement

Guppy
Krona

Visualisation

BLAST
RapSearch 2
Diamond
Kraken

Alignment or assignment

MEGAN
MetaPhlAn

Taxonomic analysis

MEGAN
MetaPhlAn
Krona

Visualisation & comparison

Assembly
MEGAHIT
Ray Meta
MetaVelvet

Gene prediction
Glimmer
GeneMark
Prodigal

Variant analysis
MaryGold

Typical pipelines

QC
FASTQC
PRINSEQ
PCR duplicate removal

Subsampling
Scripts
Digital normalisation

Merge overlapping
FLASH
PEAR

Read Trimming
trimmomatic
Sickle
FASTX-Toolkit

Contamination screening
kontaminant
BLAST

LOTS more!

pplacer
Phymm
Sequence placement
Guppy
Krona
Visualisation

BLAST
RapSearch 2
Diamond
Kraken
Alignment or assignment
MEGAN
MetaPhlAn
Taxonomic analysis
MEGAN
MetaPhlAn
Krona
Visualisation & comparison

Assembly
MEGAHIT
Ray Meta
MetaVelvet
Gene prediction
Glimmer
GeneMark
Prodigal
Variant analysis
MaryGold

Typical pipelines

QC
FASTQC
PRINSEQ
PCR duplicate removal

Subsampling
Scripts
Digital normalisation

Merge overlapping
FLASH
PEAR

Read Trimming
trimmomatic
Sickle
FASTX-Toolkit

Contamination screening
kontaminant
BLAST

LOTS more!

pplacer
Phymm
Sequence placement
Guppy
Krona
Visualisation

BLAST
RapSearch 2
Diamond
Kraken
Alignment or assignment
MEGAN
MetaPhlAn
Taxonomic analysis
MEGAN
MetaPhlAn
Krona
Visualisation & comparison

Assembly
MEGAHIT
Ray Meta
MetaVelvet
Gene prediction
Glimmer
GeneMark
Prodigal
Variant analysis
MaryGold

Typical pipelines

QC
FASTQC
PRINSEQ
PCR duplicate removal

Subsampling
Scripts
Digital normalisation

Merge overlapping
FLASH
PEAR

Read Trimming
trimmomatic
Sickle
FASTX-Toolkit

Contamination screening
kontaminant
BLAST

LOTS more!

pplacer
Phymm

Sequence placement

Guppy
Krona

Visualisation

BLAST
RapSearch 2
Diamond
Kraken

Alignment or assignment

MEGAN
MetaPhlAn

Taxonomic analysis

MEGAN
MetaPhlAn
Krona

Visualisation & comparison

Assembly
MEGAHIT
Ray Meta
MetaVelvet

Gene prediction
Glimmer
GeneMark
Prodigal

Variant analysis
MaryGold

Typical pipelines

QC
FASTQC
PRINSEQ
PCR duplicate removal

Subsampling
Scripts
Digital normalisation

Merge overlapping
FLASH
PEAR

Read Trimming
trimmomatic
Sickle
FASTX-Toolkit

Contamination screening
kontaminant
BLAST

LOTS more!

pplacer
Phymm

Sequence placement

Guppy
Krona

Visualisation

BLAST
RapSearch 2
Diamond
Kraken

Alignment or assignment

MEGAN
MetaPhlAn

Taxonomic analysis

MEGAN
MetaPhlAn
Krona

Visualisation & comparison

Assembly
MEGAHIT
Ray Meta
MetaVelvet

Gene prediction
Glimmer
GeneMark
Prodigal

Variant analysis
MaryGold

In this session…

QC
FASTQC
PRINSEQ
PCR duplicate removal

Subsampling
Scripts
Digital normalisation

Merge overlapping
FLASH
PEAR

Read Trimming
trimmomatic
Sickle
FASTX-Toolkit

Contamination screening
kontaminant
BLAST

LOTS more!

pplacer
Phymm
Sequence placement
Guppy
Krona
Visualisation

BLAST
RapSearch 2
Diamond
Kraken
Alignment or assignment
MEGAN
MetaPhlAn
Taxonomic analysis
MEGAN
MetaPhlAn
Krona
Visualisation & comparison

Assembly
MEGAHIT
Ray Meta
MetaVelvet
Gene prediction
Glimmer
GeneMark
Prodigal
Variant analysis
MaryGold

Firstly… sequence data…

```
@HWI-ST790:234:D0W8BACXX:1:1101:1792:2000 1:N:0:GCCAA
ACNATTAACAACCTTGGTGTTCAGCATGAGAACTTATCTGCAGCTGAGTCTCGTATCCGTGACG
+
CC#4ADDFHHHHHIIIIEGHIIIIIIIIIIGIIIIIIIIIIIIIIIIIDGHHIDHHIII6@FGI
@HWI-ST790:234:D0W8BACXX:1:1101:2592:1999 1:N:0:GCCAA
CTNGAATGCAGGTAGAATACATCTCCCGGATAAGCCTCGCGGCCCCCGGGGCGGGGGGGGAGAG
+
:=#44AA?:<DFFE>FED?3A<EHH>FIF?ADGCGBA?D##########################
@HWI-ST790:234:D0W8BACXX:1:1101:4221:1999 1:N:0:GCCAA
GGNAAATACGAAAGATAAGCTACGCAAGAAACGAAGGATTACTGCGAAAGGCTGCGATGCGGCA
+
@@#4=BDDFDFHDIIBGIHHHIGGIIIBHHIF=ABB@?B<DE@BF<FHH@@EHACD<B3=8@:B
```

FASTQ files

- e.g. Illumina read files
- 4 lines per read
- Stores sequence and quality information

FASTQ files

- Sanger format quality scores 0-93
- Encoded with ASCII characters 33-126
- Older versions of Illumina software slightly different

Quality value representations

Char	ASCII	Q	Char	ASCII	Q	Char	ASCII	Q	Char	ASCII	Q	Char	ASCII	Q
!	33	0	4	52	19	G	71	38	Z	90	57	m	109	76
"	34	1	5	53	20	H	72	39	[	91	58	n	110	77
#	35	2	6	54	21	I	73	40	\	92	59	o	111	78
$	36	3	7	55	22	J	74	41	]	93	60	p	112	79
%	37	4	8	56	23	K	75	42	^	94	61	q	113	80
&	38	5	9	57	24	L	76	43	_	95	62	r	114	81
'	39	6	:	58	25	M	77	44	`	96	63	s	115	82
(	40	7	;	59	26	N	78	45	a	97	64	t	116	83
)	41	8	<	60	27	O	79	46	b	98	65	u	117	84
*	42	9	=	61	28	P	80	47	c	99	66	v	118	85
+	43	10	>	62	29	Q	81	48	d	100	67	w	119	86
,	44	11	?	63	30	R	82	49	e	101	68	x	120	87
-	45	12	@	64	31	S	83	50	f	102	69	y	121	88
.	46	13	A	65	32	T	84	51	g	103	70	z	122	89
/	47	14	B	66	33	U	85	52	h	104	71	{	123	90
0	48	15	C	67	34	V	86	53	i	105	72	|	124	91
1	49	16	D	68	35	W	87	54	j	106	73	}	125	92
2	50	17	E	69	36	X	88	55	k	107	74	~	126	93
3	51	18	F	70	37	Y	89	56	l	108	75			

FASTQ files

- Q score relates to probability, p, that base is incorrect:

$$Q_{\text{sanger}} = -10 \log_{10} p$$

- What this means…

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

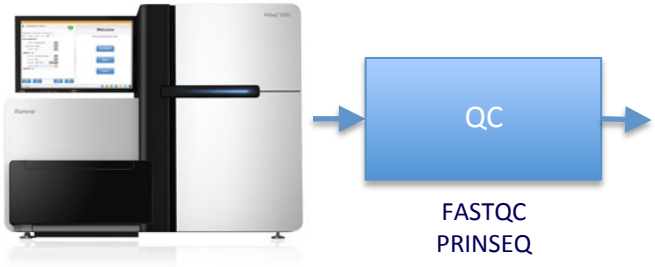
FASTA files

- e.g. assembler contigs
- Stores ID and sequence data only
- Sequence data can cover multiple lines

Sequence ID → >contig1
Sequence → ACNATTAACAACCTTGGTGTTCAGCATGAGAACTTATCTGCAGCTGAGTCTCGTATCCGTGACG
CTGAGTCTCGTATCCGTGACGGTTAGGGCGATTAGCATAGA

```
>contig2
TGACTAGCGGATTCGATTCGGAGGCTTATGGGCATTCCAGATGCAGCTAGCAGATGACATAGAT
GGGCATT
>contig3
CCCCCCTGACTAGCGGATTCGGTTCAGCATGAGTACGAATTCGGAGGCTTATGGGCATTCCAGA
AGCGTGCAGCTAGCAGATGAAGCGCATAGATGGGCTATTGTTCAGCATGAGCTGATCAACTACG
TACGGGACTGAGATGCCATGCAGTTGG
>contig4
TGACTAGCTAGTGGATTGACGAC
…
```

Quality Control

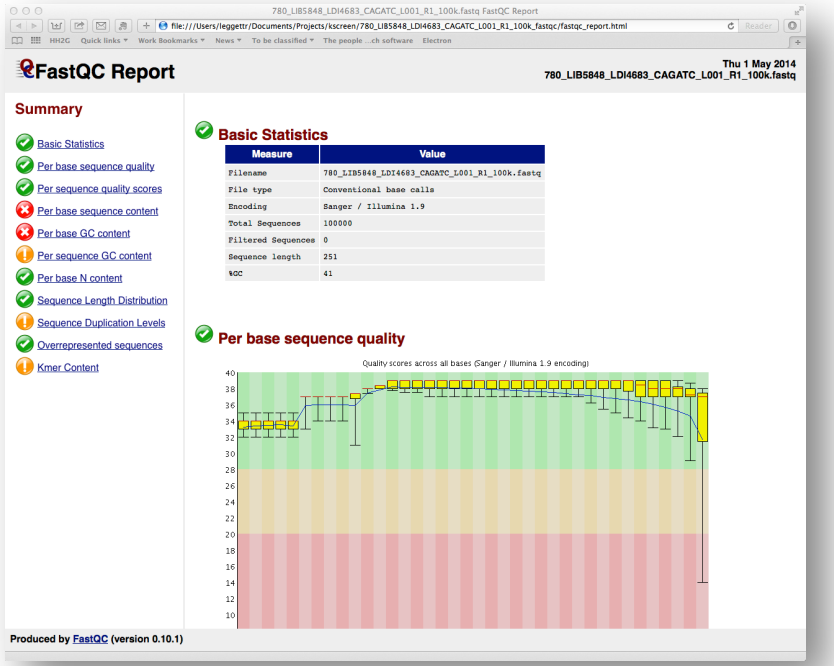


Why?

- Look for problems at an early stage.
- Understand if sequencing worked.
- Know your data = better results!

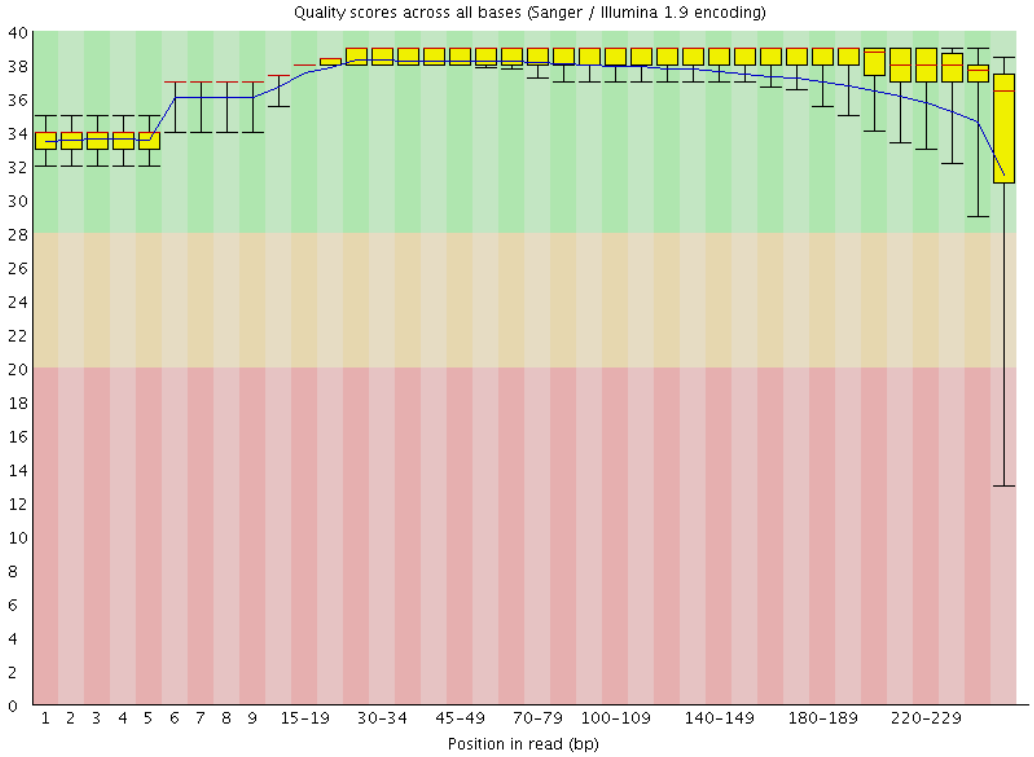
FASTQC

- Developed by Simon Andrews, Babraham Institute
- Quality control checks on sequence data
  - Quality scores
  - GC content
  - Duplicated sequence
  - Adaptors
  - Over-represented sequence
  - and more…
  - Run interactive or CLI



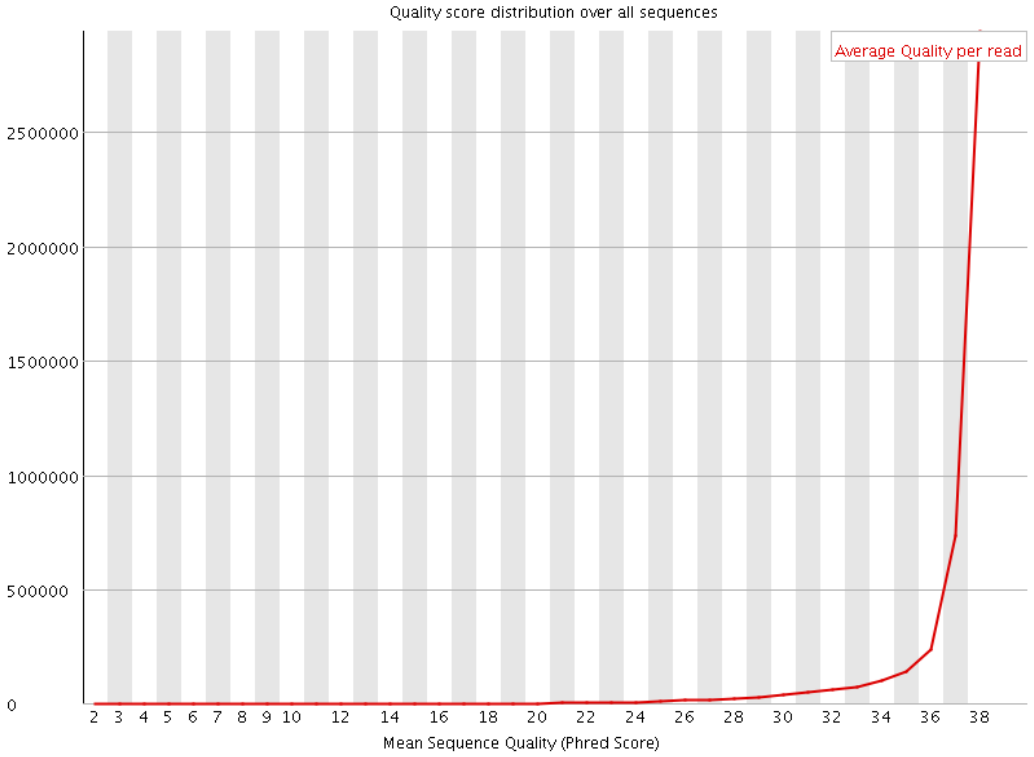
FASTQC

- Per-base quality
- “Average quality at position x”



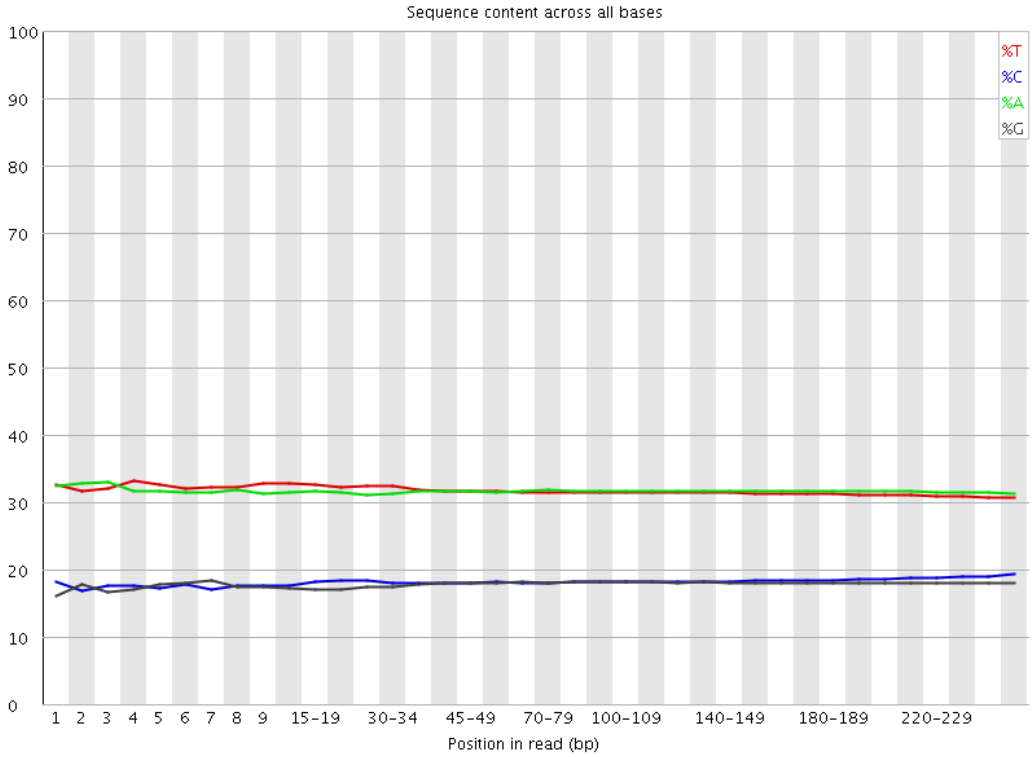
FASTQC

- Per-sequence quality
- “Number of reads with average quality x”



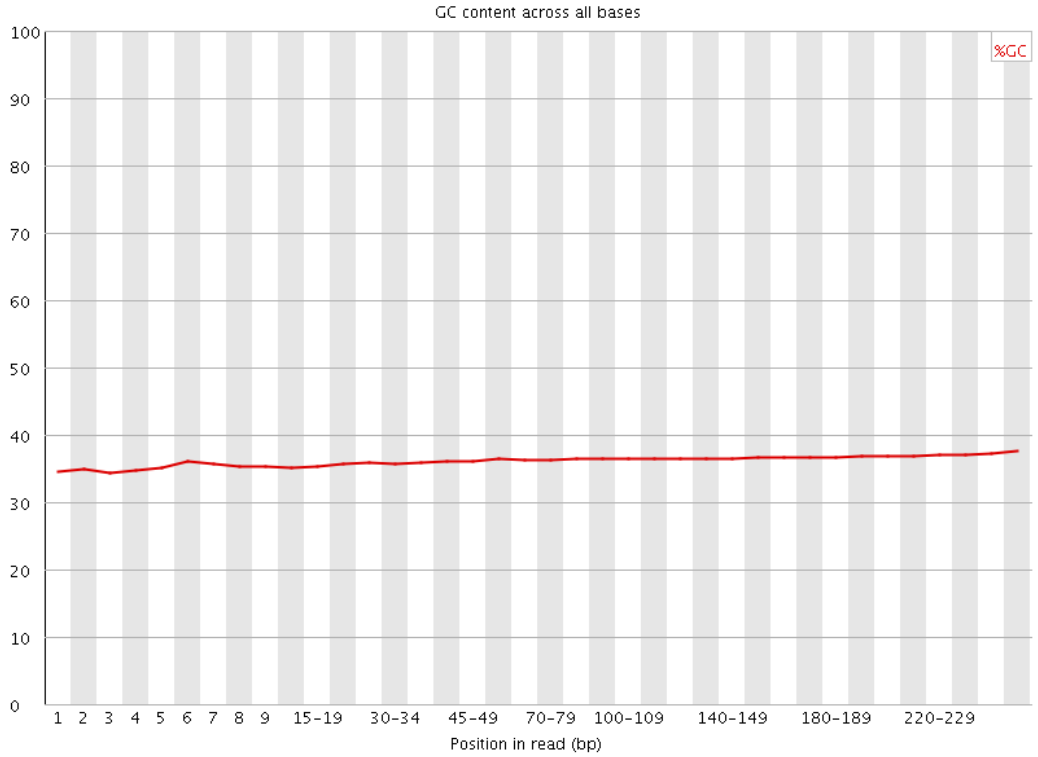
FASTQC

- Per-base sequence content
- “% of reads with A, C, G, T at position x”



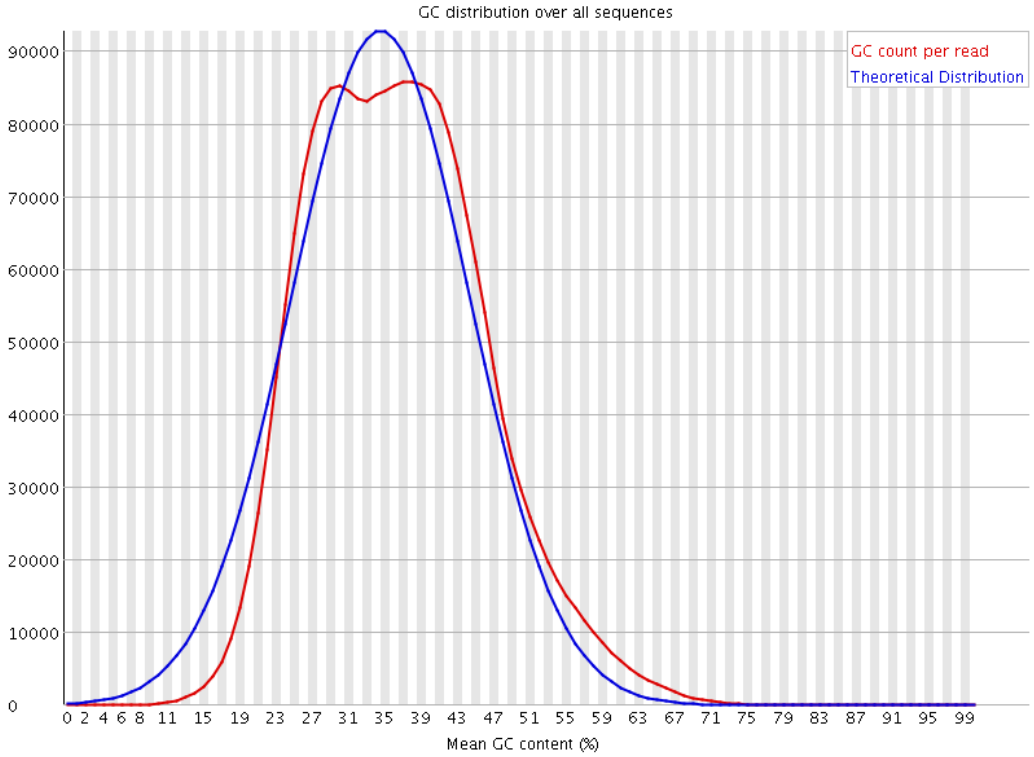
FASTQC

- Per-base GC content
- “Average GC % at position x across all reads”



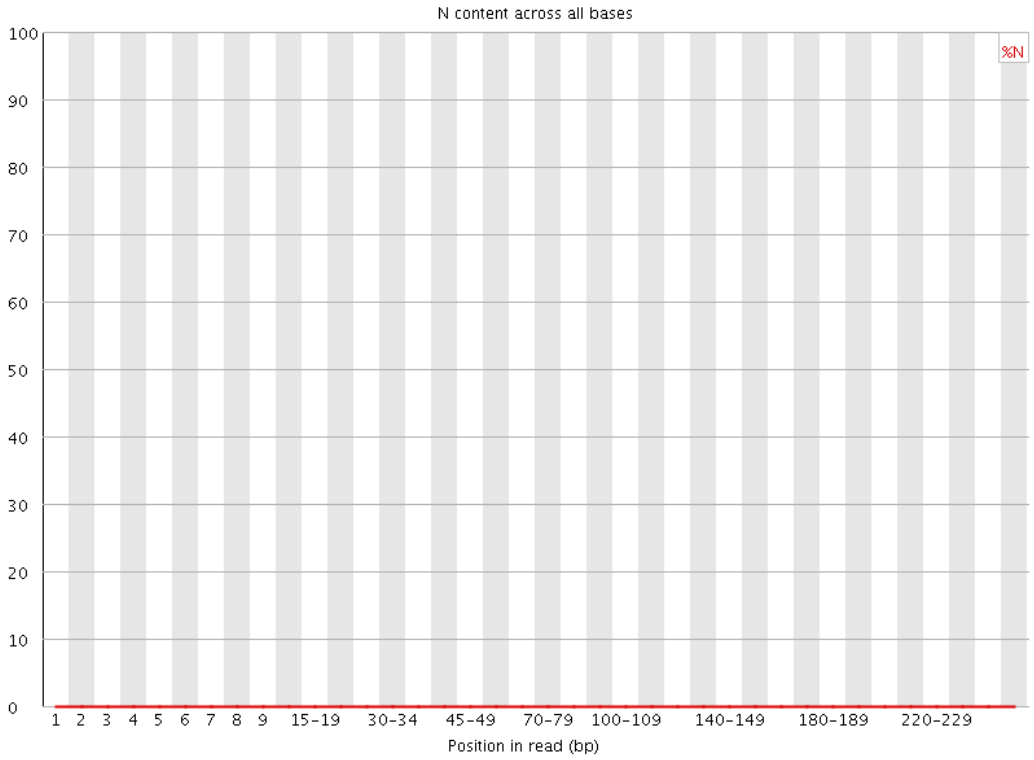
FASTQC

- Per-sequence GC content
- “Number of reads with GC % x”



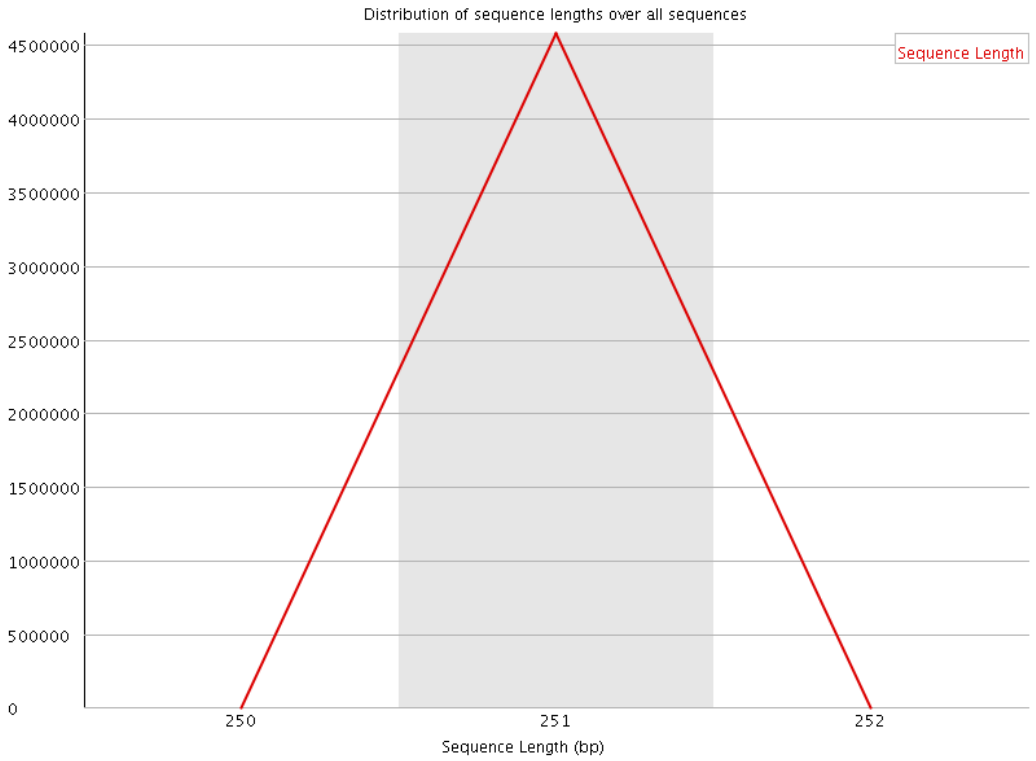
FASTQC

- Per-base N content
- “Percentage of Ns at position x”



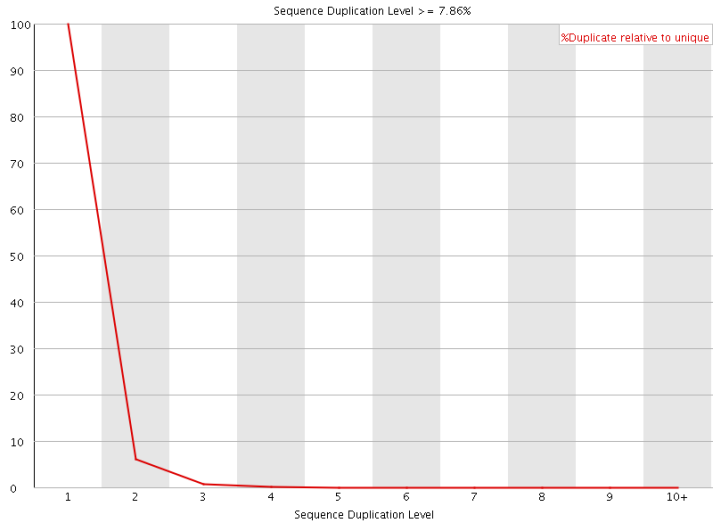
FASTQC

- Sequence length distribution
- “Number of reads of length x”



FASTQC

- Duplication levels
- “Relative number of sequences with duplication x”
- Shown relative to sequences that appear once.
- First 50 bases used.



…but will not remove duplicates for you…

PCR duplicate removal

- Can reduce analysis overhead and remove bias.
- Options:
  - FASTX Toolkit e.g.

    ```
    fastx_collapser -i input.fastq -j output.fastq
    ```

    Can be slow, single end only.
  - SamTools rmdup e.g.

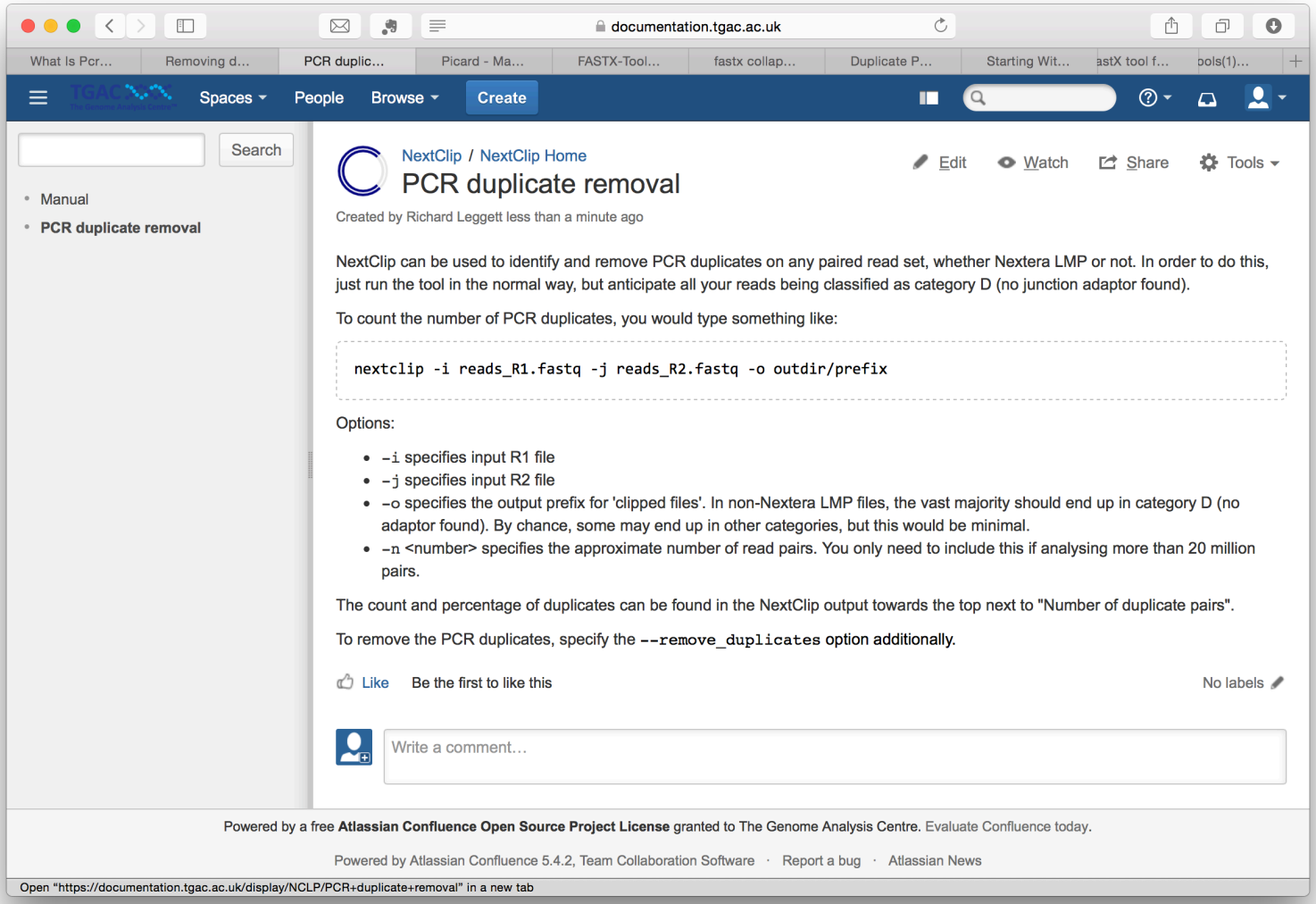
    ```
    samtools rmdup input.srt.bam output.bam
    ```

    Requires alignment, e.g. with BWA first.
  - NextClip – can use just for PCR analysis/removal
    - Pair-end only
    - To run…

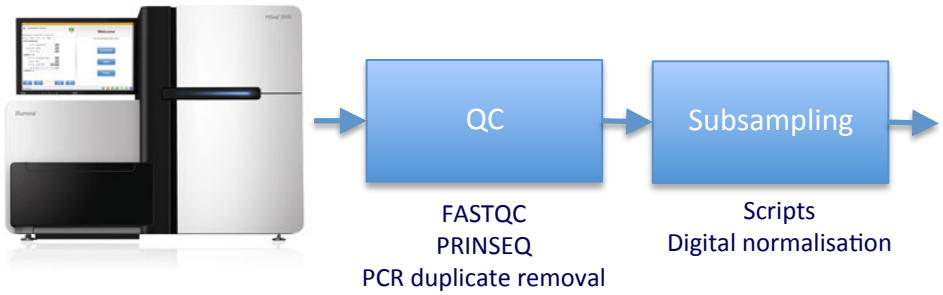
    ```
    nextclip -i reads_R1.fastq -j reads_R2.fastq
             -o outdir/prefix
    ```

NextClip

- https://documentation.tgac.ac.uk/



Subsampling



Why?

- Quick look at data.
- Protein BLAST can be SLOW.
- You can have too much coverage in assembly.

Subsampling

- How?
  - People tend to do their own thing!
  - Mainly “quick” scripts – Google it.
  - Could take top X reads of a file…
  - …but much better to random sample.
  - Also, consider *digital normalisation* for assembly.

Random read sampling

- https://www.biostars.org/p/6544/

Question: Selecting Random Pairs From Fastq?

9

I'd like to select a random subset of Illumina PE reads from a set of files. I realize this is pretty trivial to do, but I thought I'd ask if there's a simple program available to do so? It's important that it retains both ends of each pair, and in the right order, and that it can work by streaming over the input sequences (obviously they don't fit in RAM).

Or do I just have to write it myself?

illumina fastq sequence random code • 15k views

4.6 years ago by Ketil ♦ 3.6k Germany

ADD COMMENT • link • Not following ▾ modified 14 months ago by Alex Reynolds ♦ 12k • written 4.6 years ago by Ketil ♦ 3.6k

2

Not sure I follow - are you saying a random sampling might introduce bias that wasn't there already? This seems to go against the definition of 'random'.

ADD REPLY • link written 4.6 years ago by Ketil ♦ 3.6k

Do you expect your coverage to be even, otherwise you might get in hot water by introducing biases when randomly sampling over the whole genome/region: http://biostar.stackexchange.com/questions/4340/downsampling-bam-files

ADD REPLY • link written 4.6 years ago by Allpowerde • 1.1k

14

In one command line:

```
paste f1.fastq f2.fastq |\ #merge the two fastqs
awk '{ printf("%s",$0); n++; if(n%4==0) { printf("\n");} else { printf("\t\t");}
}' |\ #merge by group of 4 lines
shuf  |\ #shuffle
head |\ #only 10 records
sed 's/\t\t/\n/g' |\ #restore the delimiters
awk '{print $1 > "file1.fastq"; print $2 > "file2.fatsq"}' #split in two files.
```

4.6 years ago by Pierre Lindenbaum ♦ 75k France

Random read sampling

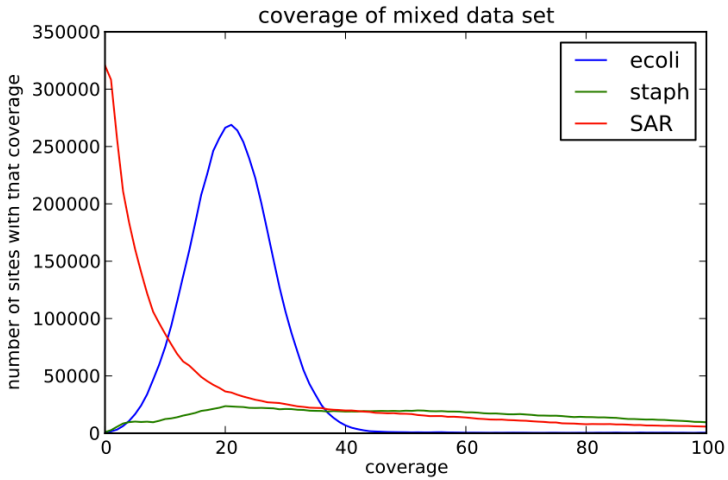
- My own Perl script available: https://github.com/richardmleggett/scripts
- To run:

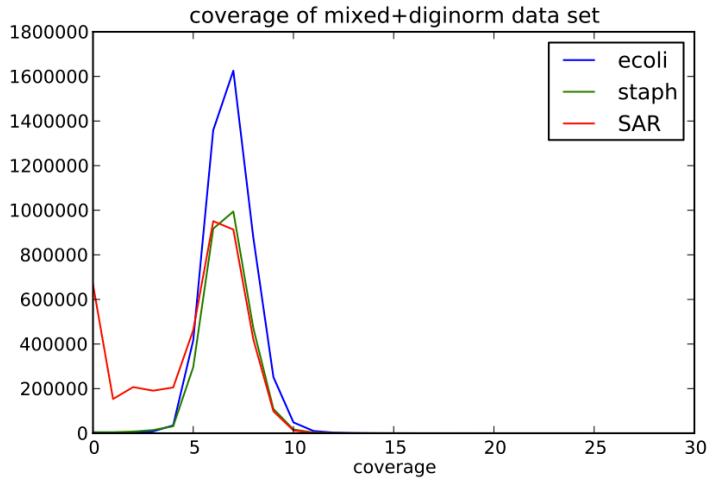
```
subsample.pl -fastq -a R1in.fastq -b R2in.fastq
             -c R1out.fastq -d R2out.fastq
```

- Optionally, also get the “remainder” of reads using -e and -f options.

Digital normalisation

- “Abundance normalization” for assembly.





- http://ivory.idyll.org/blog/what-is-diginorm.html
- Implemented in khmer tool. Paper:
  - http://f1000research.com/articles/4-900/

Merging overlapping reads

QC

FASTQC
PRINSEQ
PCR duplicate removal

Subsampling

Scripts
Digital normalisation

Merge overlapping

FLASH
PEAR

Why?

- Overlapping Paired Ends.
- Longer matches.
- Better assemblies.

FLASH

- Tunable mismatch ratio and minimum overlap.
- Outputs merged reads, plus files of unmerged.
- Simple to run:

```
flash R1.fastq R2.fastq -o output_prefix
```

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 21 2011, pages 2957–2963
doi:10.1093/bioinformatics/btr507

Genome analysis

Advance Access publication September 7, 2011

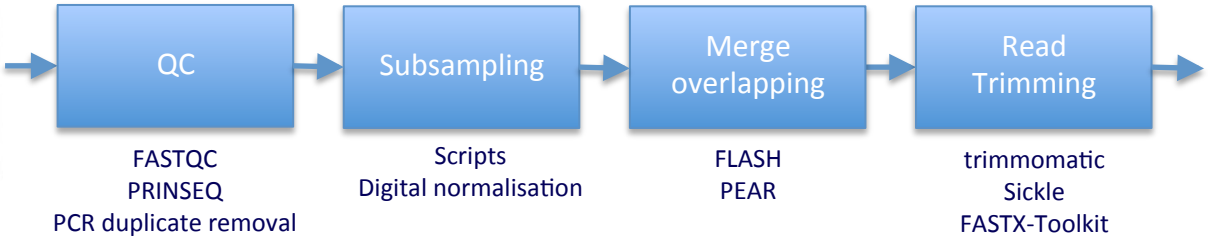
FLASH: fast length adjustment of short reads to improve genome assemblies

Tanja Magoč* and Steven L. Salzberg

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Associate Editor: Martin Bishop

Read trimming



Why?

- Remove adaptor contamination.
- Remove poor quality bases.
- Get better matching/assembly.

Read trimming

- May include:
  - Quality clipping – removing low quality bases from end
  - Ns – remove reads containing Ns
  - Adaptor trimming – removing primer/adaptor contamination
- Lots of tools:
  - Sickle – windowed quality trimming
  - Trimmomatic – adaptor and quality trimming
  - FASTX-Toolkit – adaptor and quality trimming
  - Cutadapt – adaptor trimming
  - and many more…

Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20

```
Read:       TGACGGCTATCGATCGATTAGCGGATAGCCATGACGCAAT
Qualities:  12345779:;>ACDFGHIIIIIIGFFDCA?<:97533210
```

Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20

Read: TGACGGCTATCGATCGATTAGCGGATAGCCATGACGCAAT

Qualities: 12345779:;>ACDFGHIIIIIIGFFDCA?<:97533210

Mean quality 17.5

Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20

Read: TGACGGCTATCGATCGATTAGCGGATAGCCATGACGCAAT

Qualities: 12345779:;>ACDFGHIIIIIIGFFDCA?<:97533210

Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20

Read: TGACGGCTATCGATCGATTAGCGGATAGCCATGACGCAAT

Qualities: 12345779:;>ACDFGHIIIIIIGFFDCA?<:97533210

Read trimming

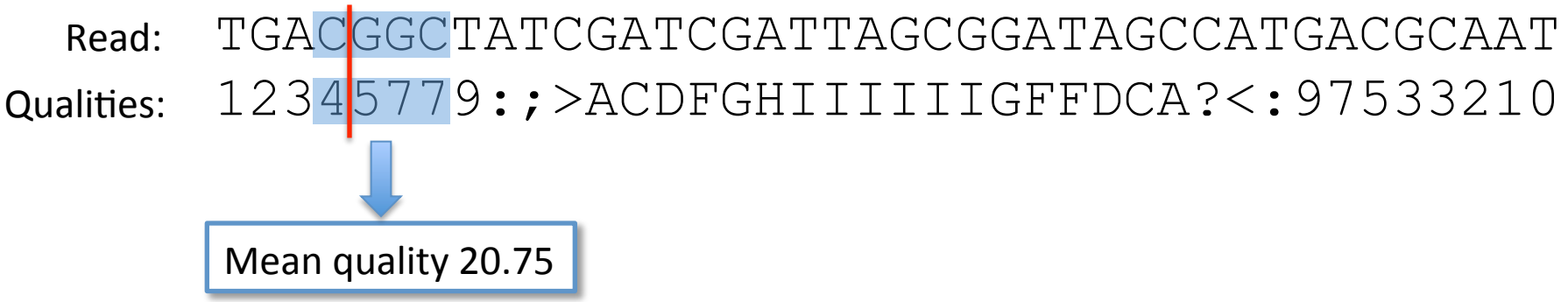
- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20

Read: TGACGGCTATCGATCGATTAGCGGATAGCCATGACGCAAT

Qualities: 12345779:;>ACDFGHIIIIIIGFFDCA?<:97533210

Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20



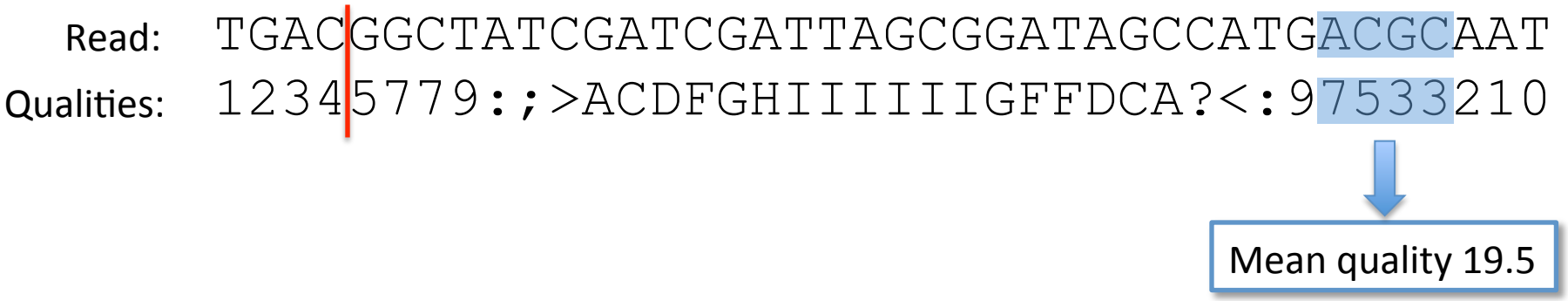
Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20

```
Read:       TGACGGCTATCGATCGATTAGCGGATAGCCATGACGCAAT
Qualities:  12345779:;>ACDFGHIIIIIIGFFDCA?<:97533210
```

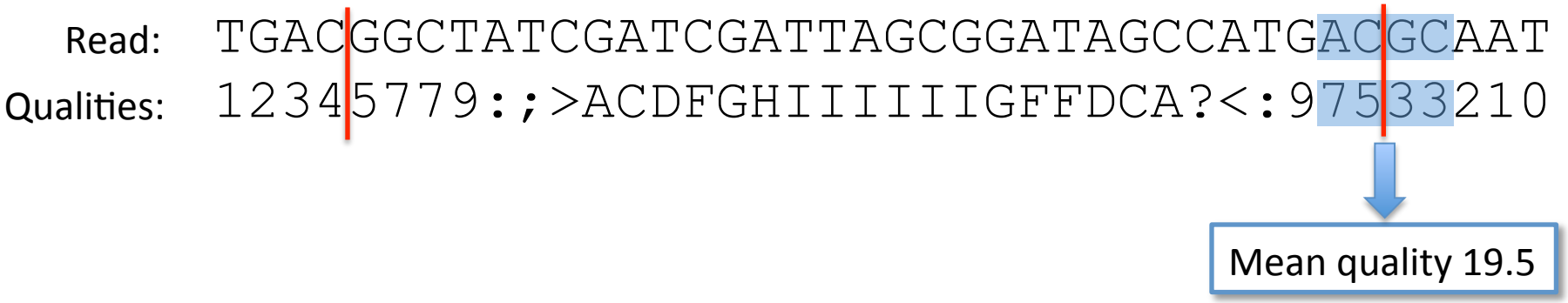
Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20



Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20



Read trimming

- Sliding windows
  - Sickle uses window size 0.1 x read length
  - Example: Read length 40, window size 4, quality threshold 20

Read: TGAC GCAAT

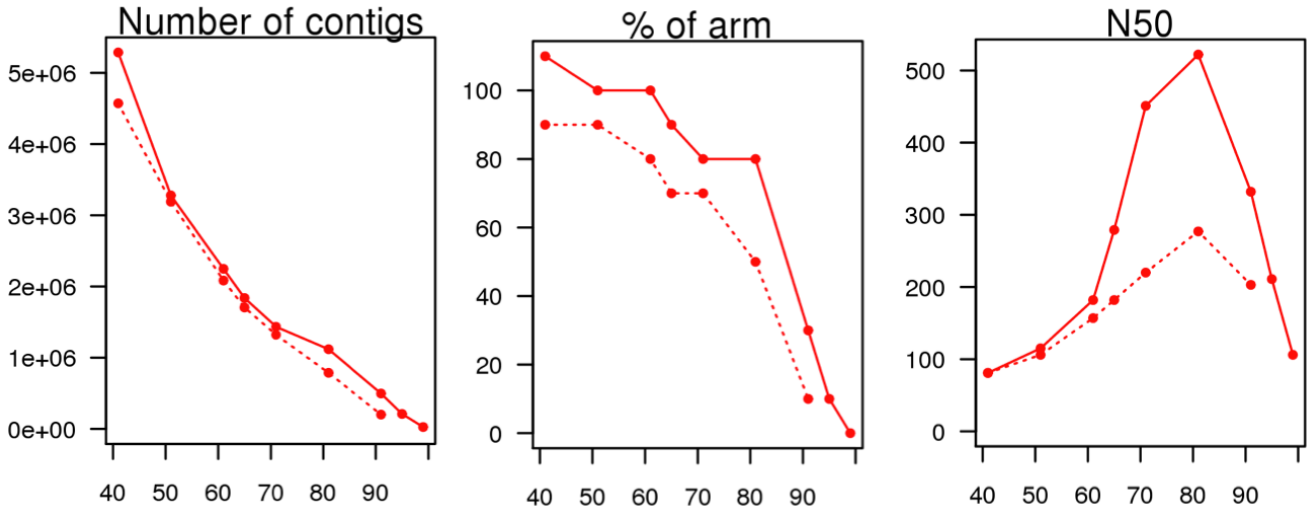
Qualities: 1234 33210

Trimmed read:

```
GGCTATCGATCGATTAGCGGATAGCCATGAC
5779:;>ACDFGHIIIIIIGFFDCA?<:975
```

Read trimming

- Quality trimming can have a cost – eg. hexaploid wheat



untrimmed data
trimmed reads (to Q30)

- Gene content of trimmed assembly was lower
- Indicates cost to quality trimming
- Good quality sequence lost as well as low quality

(Data from Paul Bailey, Bernardo Clavijo at TGAC)

Read trimming

- Adaptor trimming approaches (eg. Trimmomatic):
  - Database of adaptors and primers
  - Matching bases increase score
  - Mismatches reduce score
  - Good enough match triggers clipping
- Most tools also allow
  - Hard clipping at start or end
  - Minimum read size specification

Contamination screening

Why?

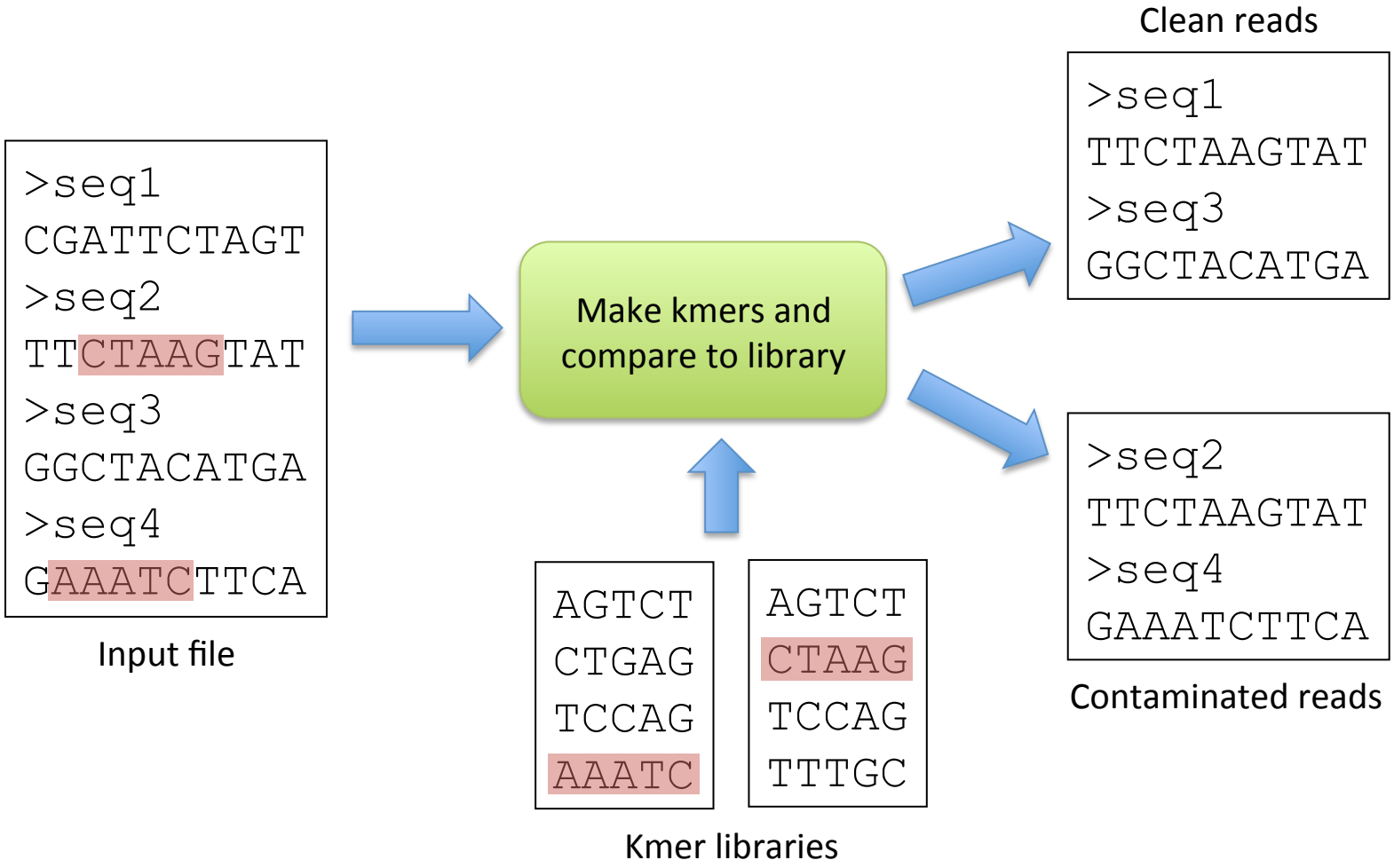
- Remove contaminants.
- Remove host.
- Simplify analysis.

Contamination screening

1. Alignment against potential contaminants or host references
   - e.g. BWA, BLAST
   - Remove reads with good alignment
2. Kontaminant, a TGAC-developed tool for kmer-based screening and filtering…

Kontaminant

- Matches kmers in reads to pre-computed reference:

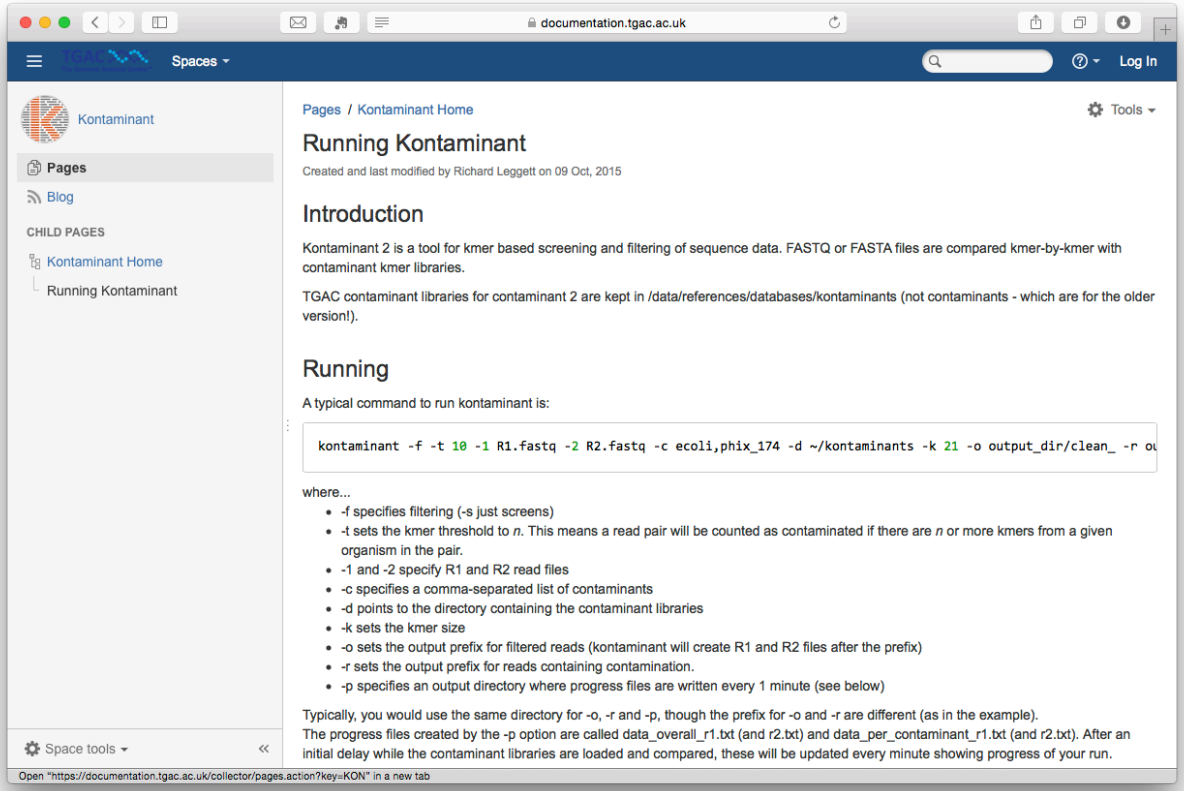


Kontaminant

- Example:

```
kontaminant -f -1 R1.fastq -2 R2.fastq -c ecoli -k 21
            -o outdir/kept_ -r outdir/removed_
```

- https://documentation.tgac.ac.uk/



Assembly

QC

FASTQC
PRINSEQ
PCR duplicate removal

Subsampling

Scripts
Digital normalisation

Merge overlapping

FLASH
PEAR

Read Trimming

trimmomatic
Sickle
FASTX-Toolkit

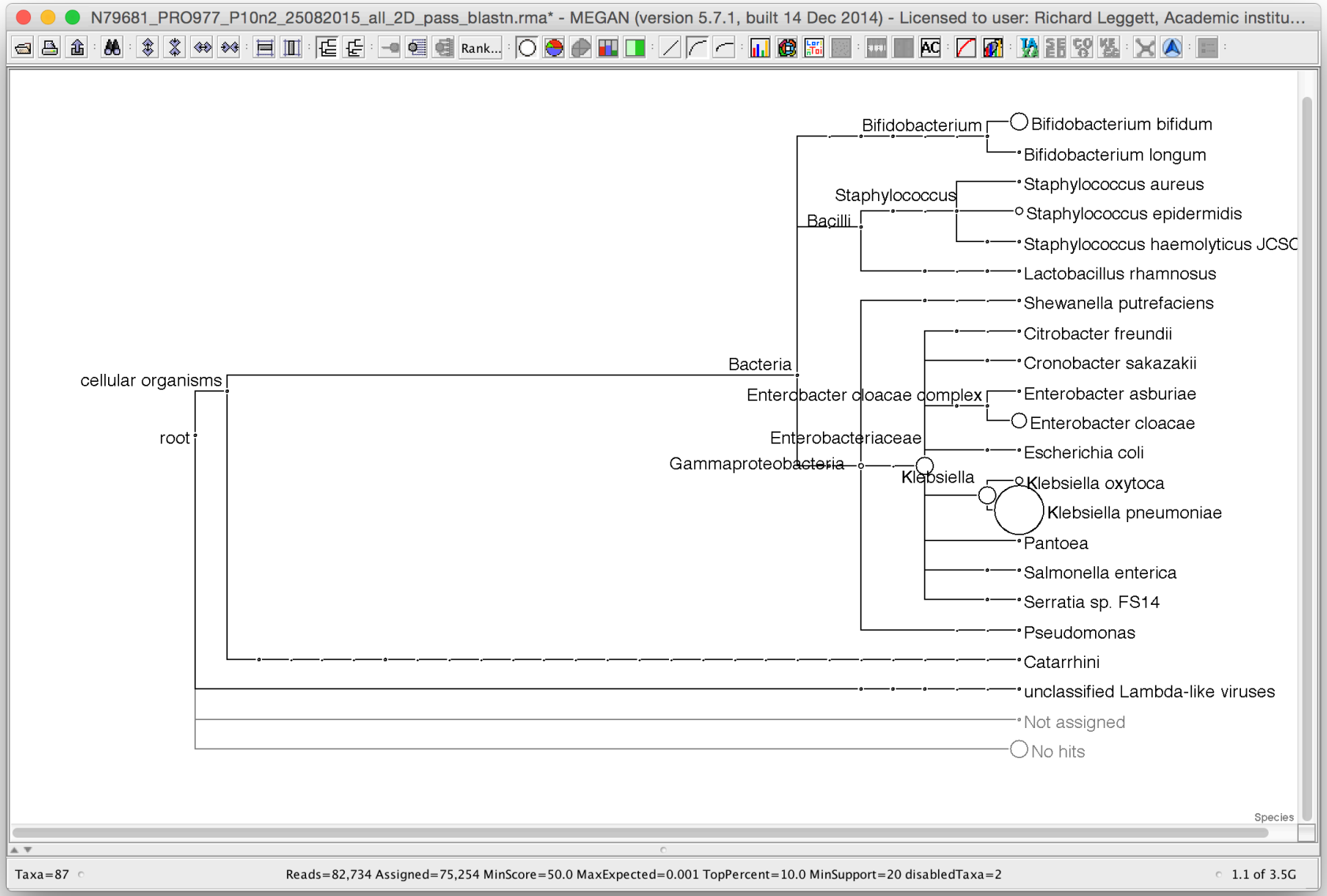
Contamination screening

kontaminant
BLAST

Assembly

MEGAHIT
Ray Meta
MetaVelvet

Coming up…

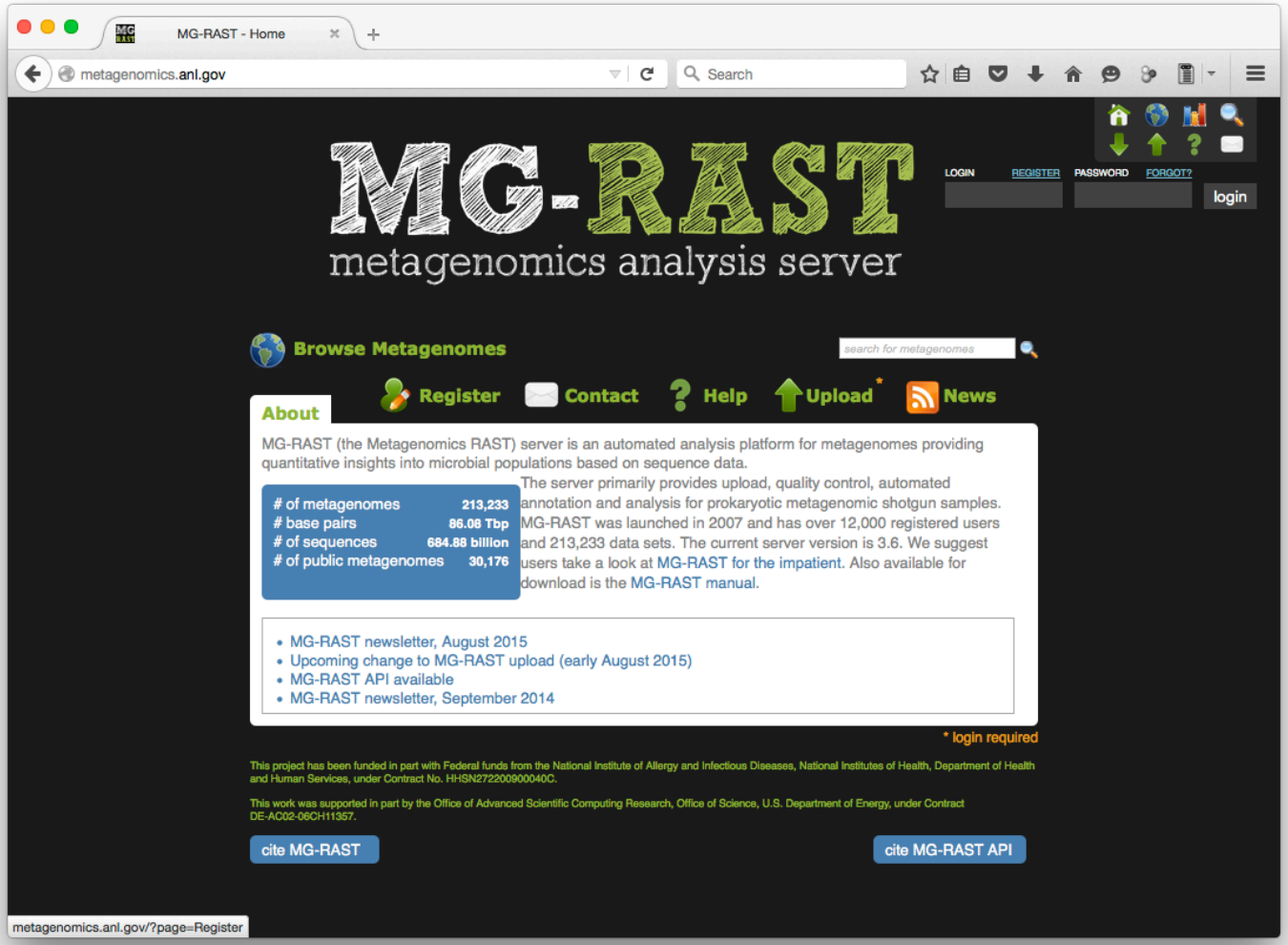

N79681_PRO977_P10n2_25082015_all_2D_pass_blastn.rma* - MEGAN (version 5.7.1, built 14 Dec 2014) - Licensed to user: Richard Leggett, Academic institu…
Rank...
Bifidobacterium
Bifidobacterium bifidum
Bifidobacterium longum
Staphylococcus
Staphylococcus aureus
Staphylococcus epidermidis
Bacilli
Staphylococcus haemolyticus JCSC
Lactobacillus rhamnosus
Shewanella putrefaciens
Citrobacter freundii
Bacteria
Cronobacter sakazakii
cellular organisms
Enterobacter cloacae complex
Enterobacter asburiae
Enterobacter cloacae
root
Enterobacteriaceae
Escherichia coli
Gammaproteobacteria
Klebsiella
Klebsiella oxytoca
Klebsiella pneumoniae
Pantoea
Salmonella enterica
Serratia sp. FS14
Pseudomonas
Catarrhini
unclassified Lambda-like viruses
Not assigned
No hits
Species
Taxa=87
Reads=82,734 Assigned=75,254 MinScore=50.0 MaxExpected=0.001 TopPercent=10.0 MinSupport=20 disabledTaxa=2
1.1 of 3.5G

Web portals

- A number of web-based portals exist where you can upload your data and perform analysis.
- Examples:
  - MG-RAST
  - EBI Metagenomics

MG-RAST

- http://metagenomics.anl.gov



MG-RAST

TGAC The Genome Analysis Centre™ BBSRC Greater Norwich Development Partnership

Taxonomic Hits Distribution HIDE

The pie charts below illustrate the distribution of taxonomic domains, phyla, and orders for the annotations. Each slice indicates the percentage of reads with predicted proteins and ribosomal RNA genes annotated to the indicated taxonomic level. This information is based on all the annotation source databases used by MG-RAST. An interactive Krona chart of the full taxonomy is also available
Click on a slice or legend to view all sequences annotated with the indicated taxonomic level in the analysis page.

View taxonomic interactive chart

domain Download chart data

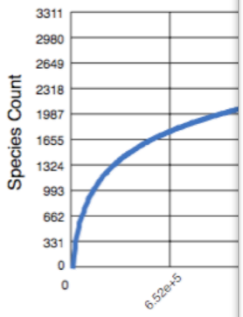
Rarefaction Curve HIDE

The plot below shows the rarefaction curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled. On the left, a steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individuals is sampled: more intensive sampling is likely to yield only few additional species.

Sampling curves generally rise very
rarefaction curves are calculated fr
the the complete dataset.

Download chart data

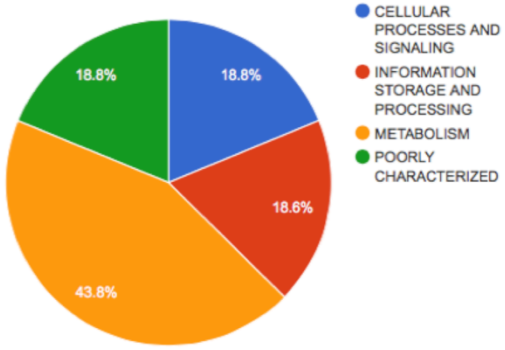
The image is currently dynamic. To



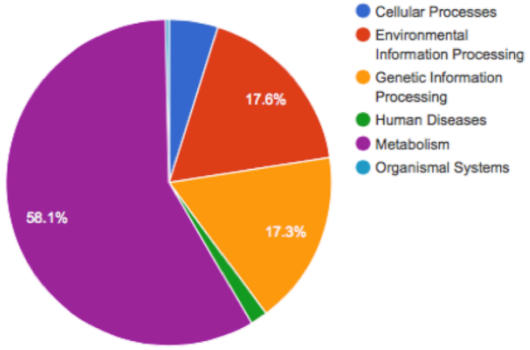
Functional Category Hits Distribution [?] HIDE

The pie charts below illustrate the distribution of functional categories for at the highest level supported by these functional hierarchies. Each slice indicates the percentage of reads with predicted protein functions annotated to the category for the given source. An interactive Krona chart of each functional hierarchy is also available
Click on a slice or legend to view all sequences annotated with the indicated category in the analysis page.

COG Download chart data
has 74,908 predicted functions
23.2% of predicted proteins
63.7% of annotated proteins
View COG interactive chart

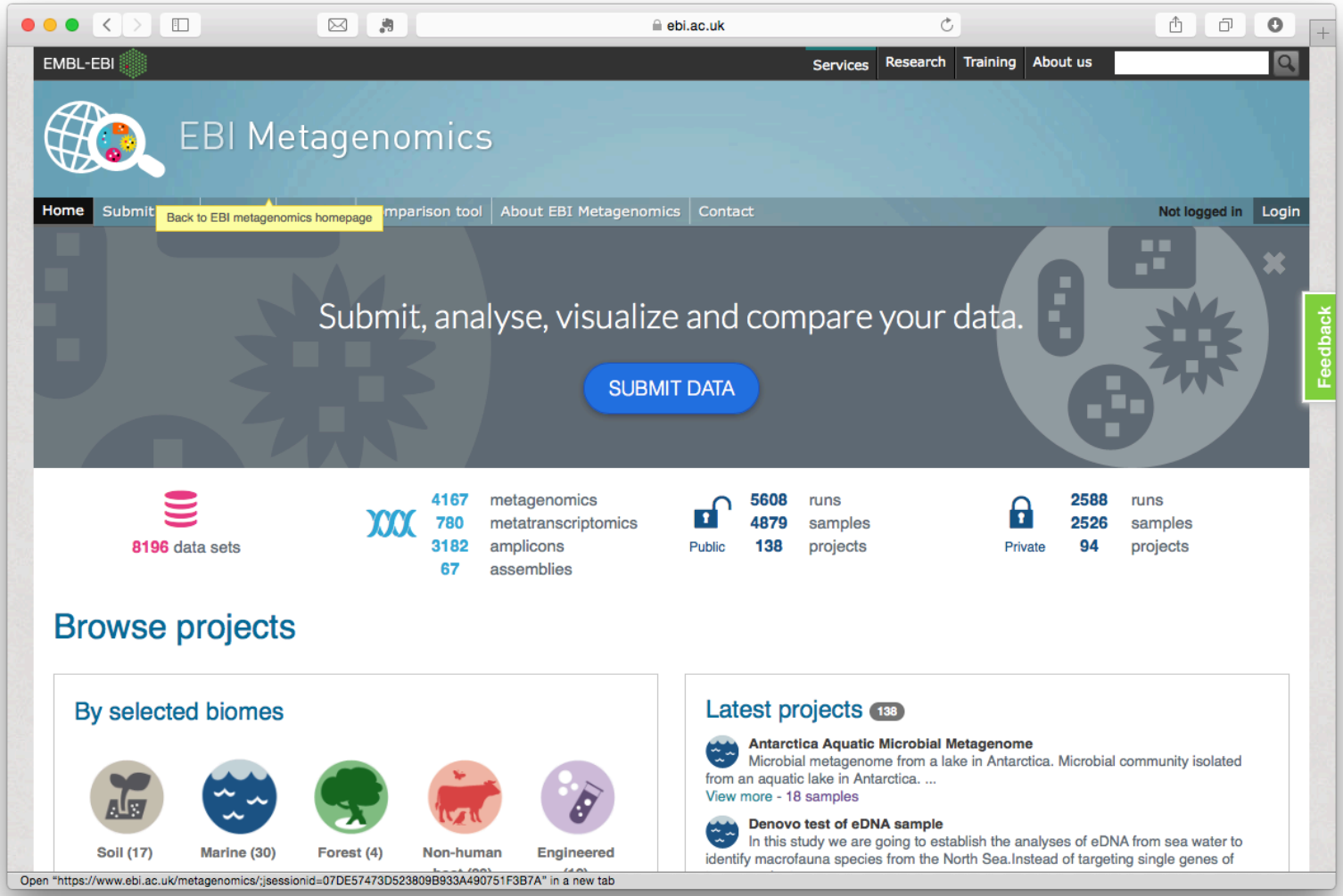


KO Download chart data
has 44,728 predicted functions
13.9% of predicted proteins
38.0% of annotated proteins
View KO interactive chart

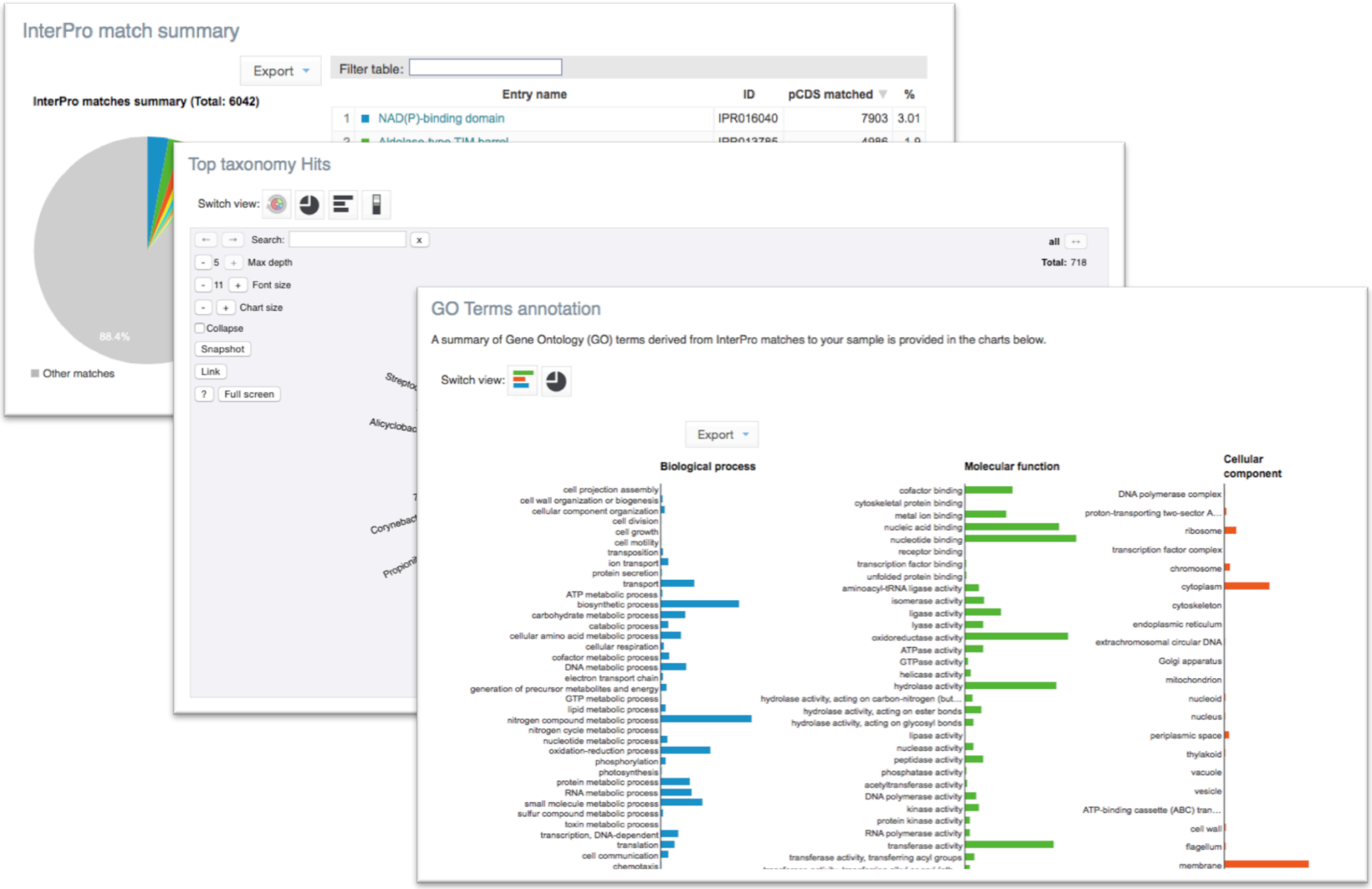


EBI Metagenomics

- https://www.ebi.ac.uk/metagenomics/



EBI Metagenomics



The End

Thank you for listening.