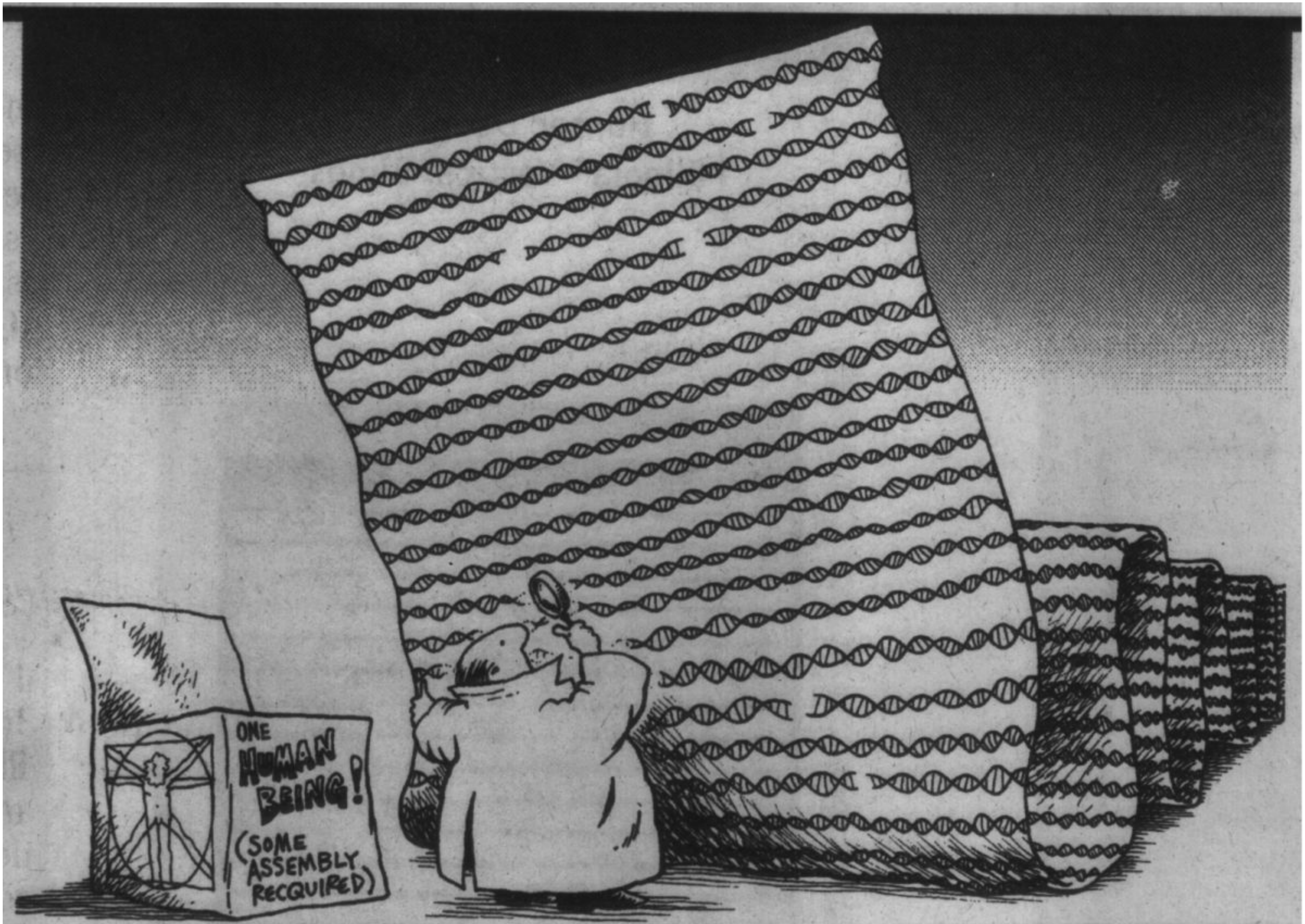
UNIVERSITY OF EXETER

Assembly and annotation using Galaxy

Konrad Paszkiewicz

Sequencing Service, University of Exeter, UK.

25th May 2011



BY AUTH FOR THE PHILADELPHIA INQUIRER

Overview

- Why de-novo assembly?
- What is de-novo assembly?
- Types of assemblers
- Annotation
- A toy example in Galaxy
- Future developments

Sequencing - 2007

PRODUCTION

Rooms of equipment
Subcloning > picking > prepping
35 FTEs
3-4 weeks

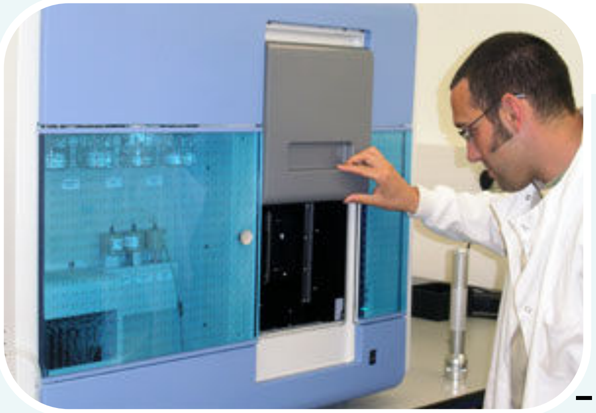
SEQUENCING

74x Capillary Sequencers
10 FTEs
15-40 runs per day
1-2Mb per instrument per day
120Mb total capacity per day

2nd generation sequencing - Today

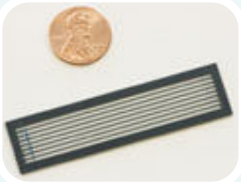
PRODUCTION

1x Cluster Station
1 FTE
1 day



SEQUENCING

1x Genome Analyzer
Same FTE as above
1 run per 3-10 days
- 90Gb per instrument per run



UNIVERSITY OF EXETER

Why de-novo assembly?

Why is de-novo sequence assembly useful?

- No reference genome available
- What is the most suitable reference genome?
  (e.g. species definition problem in bacteria)

– What's new in a genome?

  - Remapping will not tell you what is new in a genome (e.g. plasmids, novel genes, novel chromosomes)

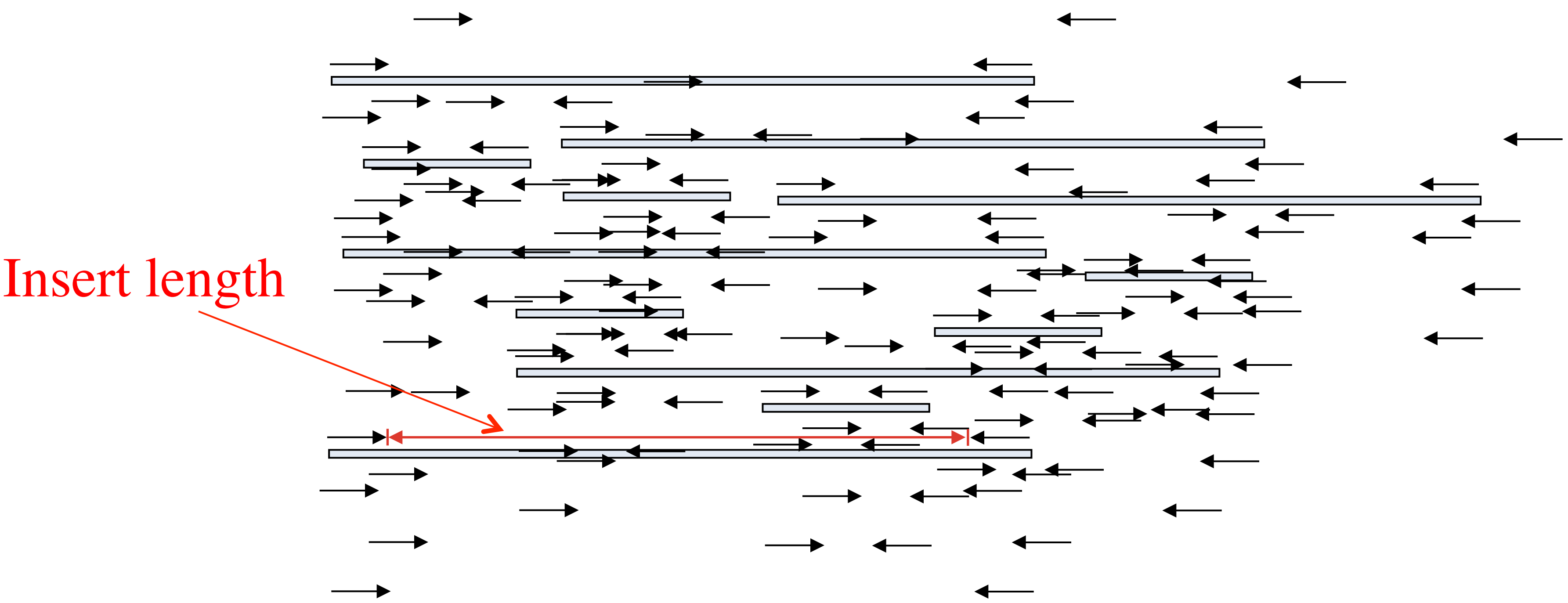
– What's really missing from a genome?

  - Remapping may fail to detect homologous regions

What is de-novo assembly?

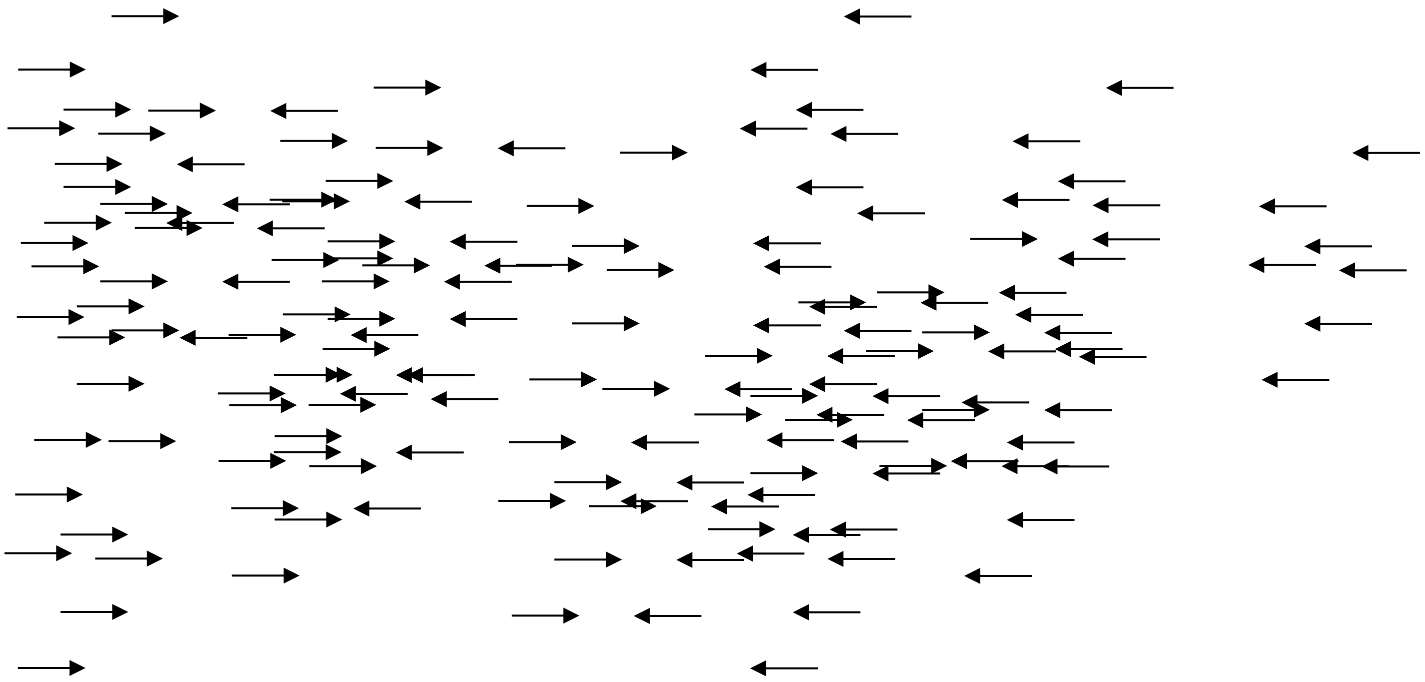
De-novo sequence assembly

1. Sequence DNA fragments from each end



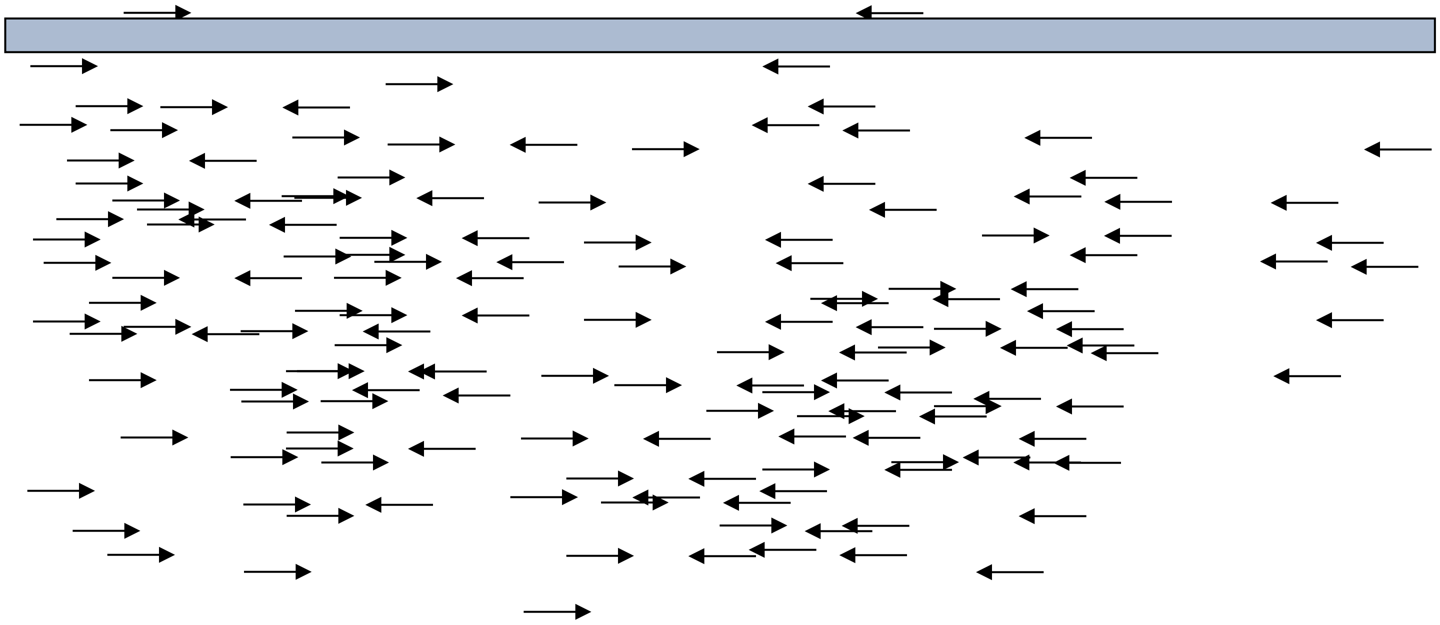
De-novo Sequence Assembly

1. Sequence DNA fragments from each end
2. Reads aligned to generate contigs



De-novo Sequence Assembly

1. Sequence DNA fragment from each end
2. Reads aligned to generate contigs

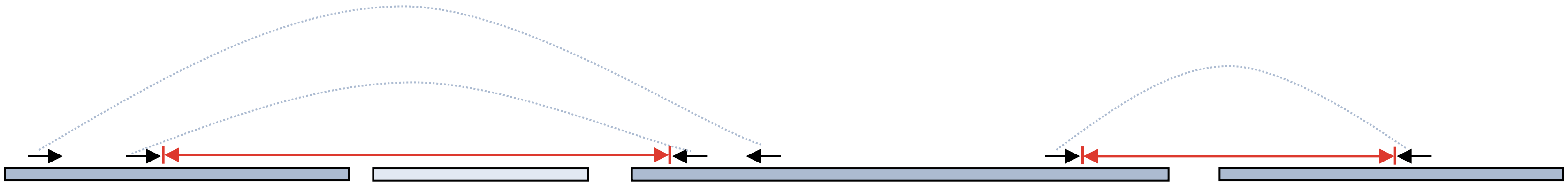


De-novo Sequence Assembly

1. Sequence clones from each end
2. Reads aligned to generate contigs
3. Supercontigs derived from paired reads on different contigs

De-novo Sequence Assembly

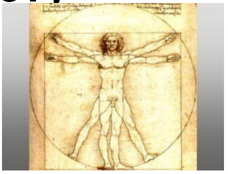
1. Sequence reads from each end
2. Reads aligned to generate contigs
3. Supercontigs derived from paired reads on different contigs



4. Ordering of contigs is determined
5. Different insert lengths and read lengths can resolve ambiguities

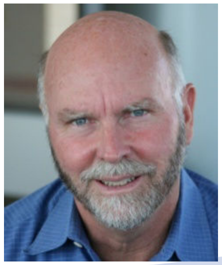
De-novo assembly:
It’s not just for genomes.

1. Traditional single homogenous genome assembly

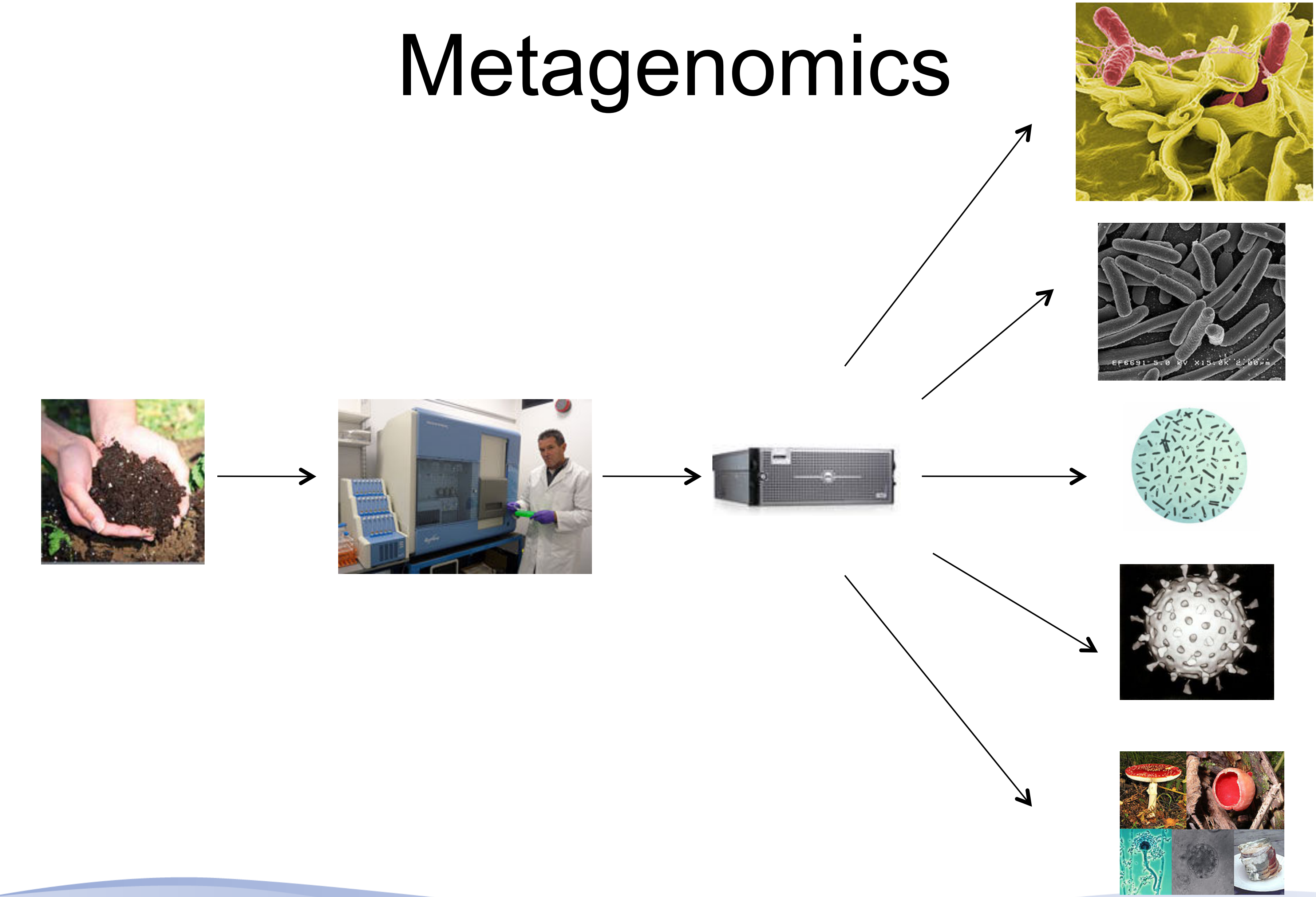


2. Single organism transcriptomes without a reference
   - Estimates of expression

3. Genomic/transcriptomic assembly of symbionts and metagenomes



Metagenomics



Denovo Sequence Assembly

- **Caveats**
  - No assembly is perfect
  - Assemblies from 2nd generation tend to be worse in a number of ways than Sanger based-assemblies

    + Easier to generate data

    + Easier to generate lots of assemblies

    - Shorter reads/higher error rates

    - Man/brainpower is more thinly spread

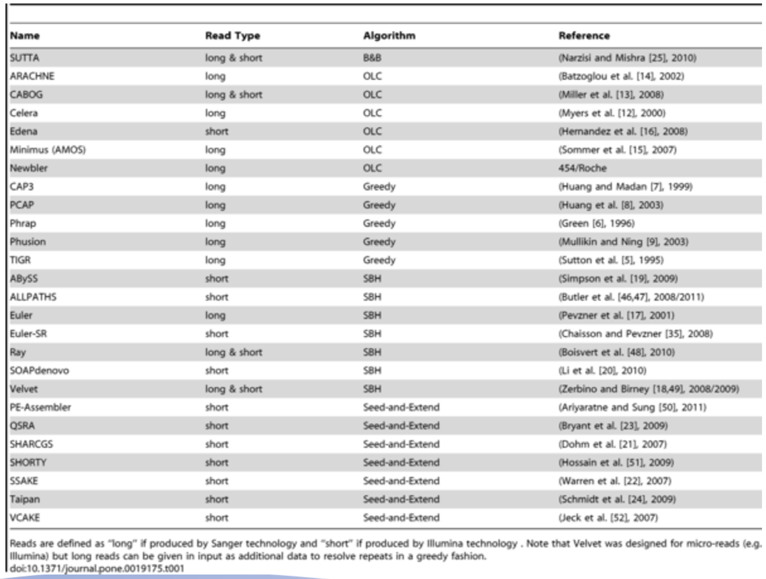
    - Harder to evaluate assemblies

    - Harder to annotate and compare between samples

Types of assemblers

Types of assemblers

- 4 categories, many variations
- Each tends to have its own niche
- Memory and hardware requirements can differ substantially
- Galaxy has support either in-built or via Galaxy Tool-shed for Velvet, MIRA, AbySS, Phrap Newbler
- **Typically a parameter scan is need to get the 'best' assembly**



Name	Read Type	Algorithm	Reference
SUTTA	long & short	B&B	(Narzisi and Mishra [25], 2010)
ARACHNE	long	OLC	(Batzoglou et al. [14], 2002)
CABOG	long & short	OLC	(Miller et al. [13], 2008)
Celera	long	OLC	(Myers et al. [12], 2000)
Edena	short	OLC	(Hernandez et al. [16], 2008)
Minimus (AMOS)	long	OLC	(Sommer et al. [15], 2007)
Newbler	long	OLC	454/Roche
CAP3	long	Greedy	(Huang and Madan [7], 1999)
PCAP	long	Greedy	(Huang et al. [8], 2003)
Phrap	long	Greedy	(Green [6], 1996)
Phusion	long	Greedy	(Mullikin and Ning [9], 2003)
TIGR	long	Greedy	(Sutton et al. [5], 1995)
ABySS	short	SBH	(Simpson et al. [19], 2009)
ALLPATHS	short	SBH	(Butler et al. [46,47], 2008/2011)
Euler	long	SBH	(Pevzner et al. [17], 2001)
Euler-SR	short	SBH	(Chaisson and Pevzner [35], 2008)
Ray	long & short	SBH	(Boisvert et al. [48], 2010)
SOAPdenovo	short	SBH	(Li et al. [20], 2010)
Velvet	long & short	SBH	(Zerbino and Birney [18,49], 2008/2009)
PE-Assembler	short	Seed-and-Extend	(Ariyaratne and Sung [50], 2011)
QSRA	short	Seed-and-Extend	(Bryant et al. [23], 2009)
SHARCGS	short	Seed-and-Extend	(Dohm et al. [21], 2007)
SHORTY	short	Seed-and-Extend	(Hossain et al. [51], 2009)
SSAKE	short	Seed-and-Extend	(Warren et al. [22], 2007)
Taipan	short	Seed-and-Extend	(Schmidt et al. [24], 2009)
VCAKE	short	Seed-and-Extend	(Jeck et al. [52], 2007)

Reads are defined as "long" if produced by Sanger technology and "short" if produced by Illumina technology . Note that Velvet was designed for micro-reads (e.g. Illumina) but long reads can be given in input as additional data to resolve repeats in a greedy fashion.
doi:10.1371/journal.pone.0019175.t001

Narzisi G, Mishra B, Comparing De Novo Genome Assembly:
The Long and Short of It. 2011 PLoS ONE 6(4):

De novo assembly of short sequence reads
Paszkiewicz, K. Studholme, D.
Briefings in Bioinformatics
August 2010 11(5): 457-472

Annotation

Annotation

Identification of

genes
exons
promoters
signal peptides
regulatory regions
alleles
non-coding RNAs
repeats...

2 broad categories of annotation methodology:

Sequence homology-based (e.g. Blast)
Profile/HMM-based (e.g. PFAM, TMHMM, SignalP)

Annotation

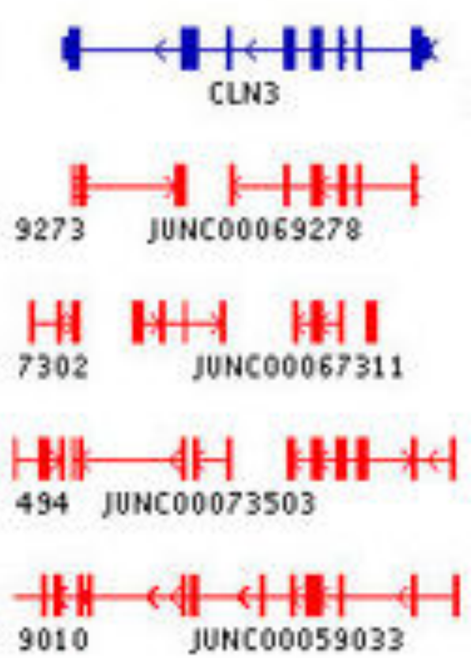
To do this effectively it is often necessary to gather additional data:

e.g.

ChIP-Seq
RNA-seq

Annotation

Exon structure
Transcription start sites



Annotated gene structure

A toy example in Galaxy

Denovo sequencing project

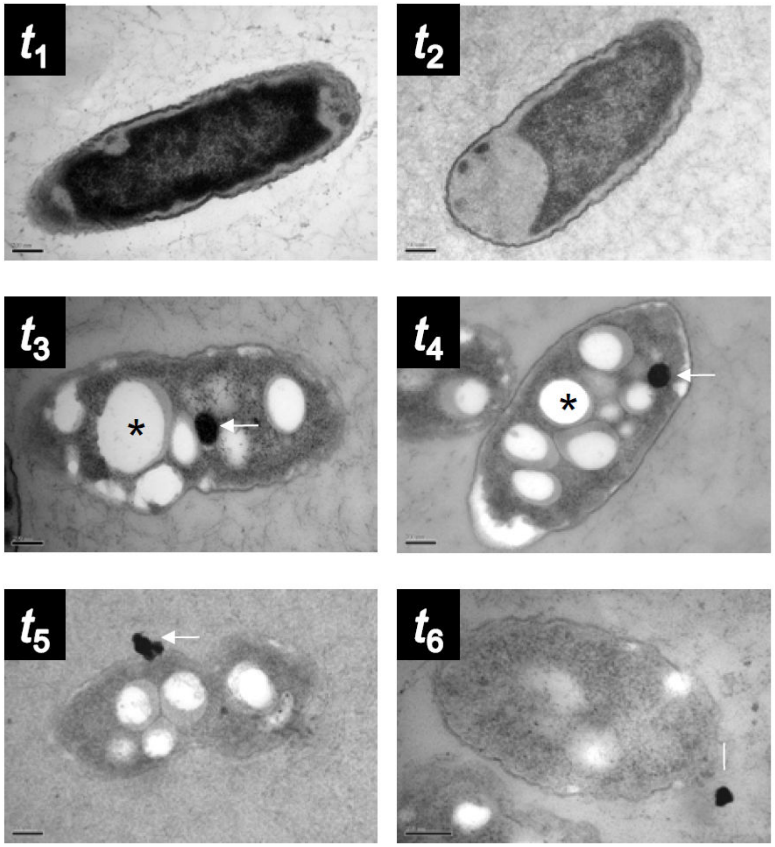
A new beta-proteobacterium which secretes elemental metal

60% GC content

Approximately 8 Mb genome

Aim: Which genes(s) are responsible for translocation?

Method: 1 lane Illumina sequencing
Mass spectrometry



Process

1. Uploading files from Illumina sequencing
2. Filtering reads
3. De-novo assembly
4. Annotation
5. Locating secretion protein using mass-spectrometry information

1. Uploading files

2nd generation sequencing output formats

Illumina

↓

FASTQ (various flavours)

SoLID/ABI-Life

↓

Colourspace FASTA

Roche 454

↓

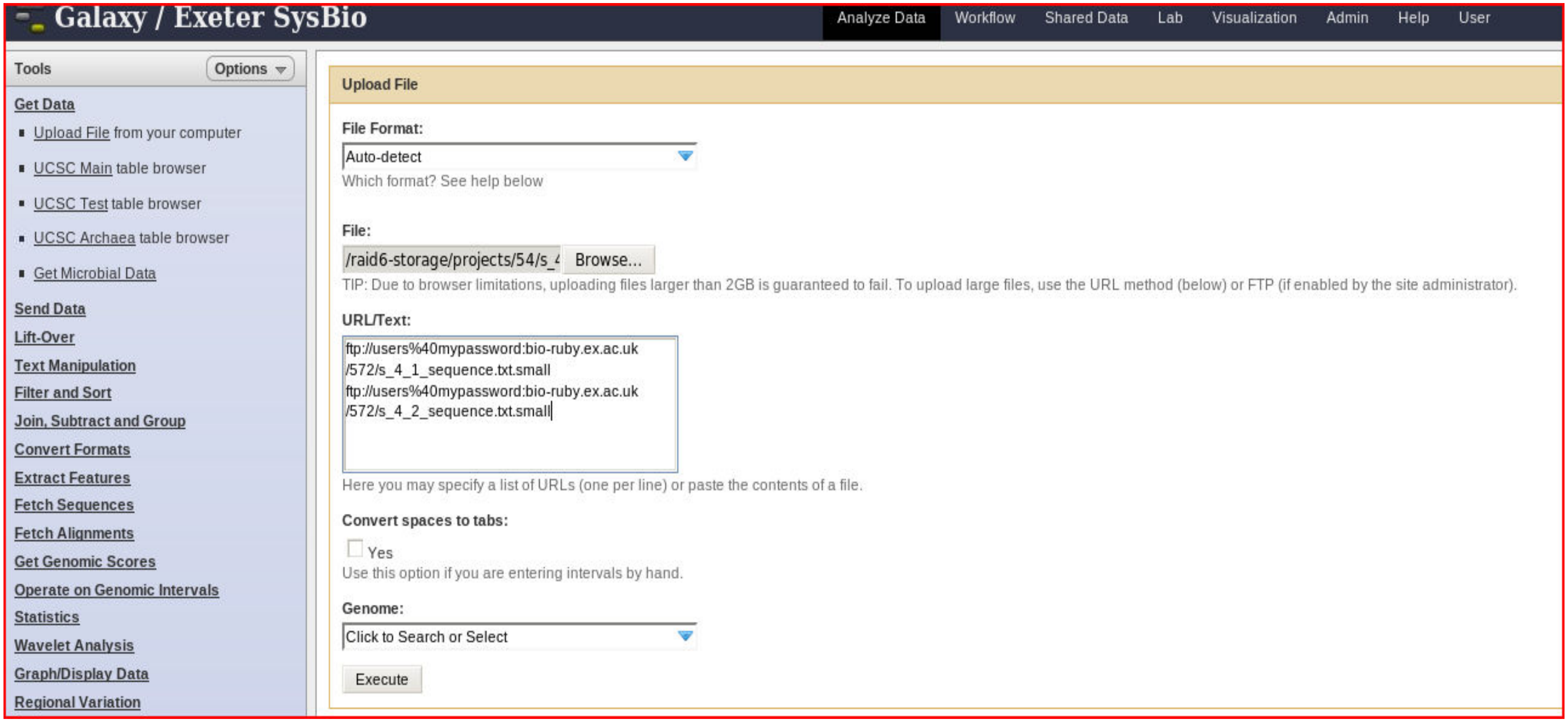
SFF

Ion Torrent

↓

SFF or FASTQ

Uploading FASTQ files



Upload File

File Format: Auto-detect

Which format? See help below

File: /raid6-storage/projects/54/s_4

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

```
ftp://users%40mypassword:bio-ruby.ex.ac.uk
/572/s_4_1_sequence.txt.small
ftp://users%40mypassword:bio-ruby.ex.ac.uk
/572/s_4_2_sequence.txt.small
```

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs: □ Yes

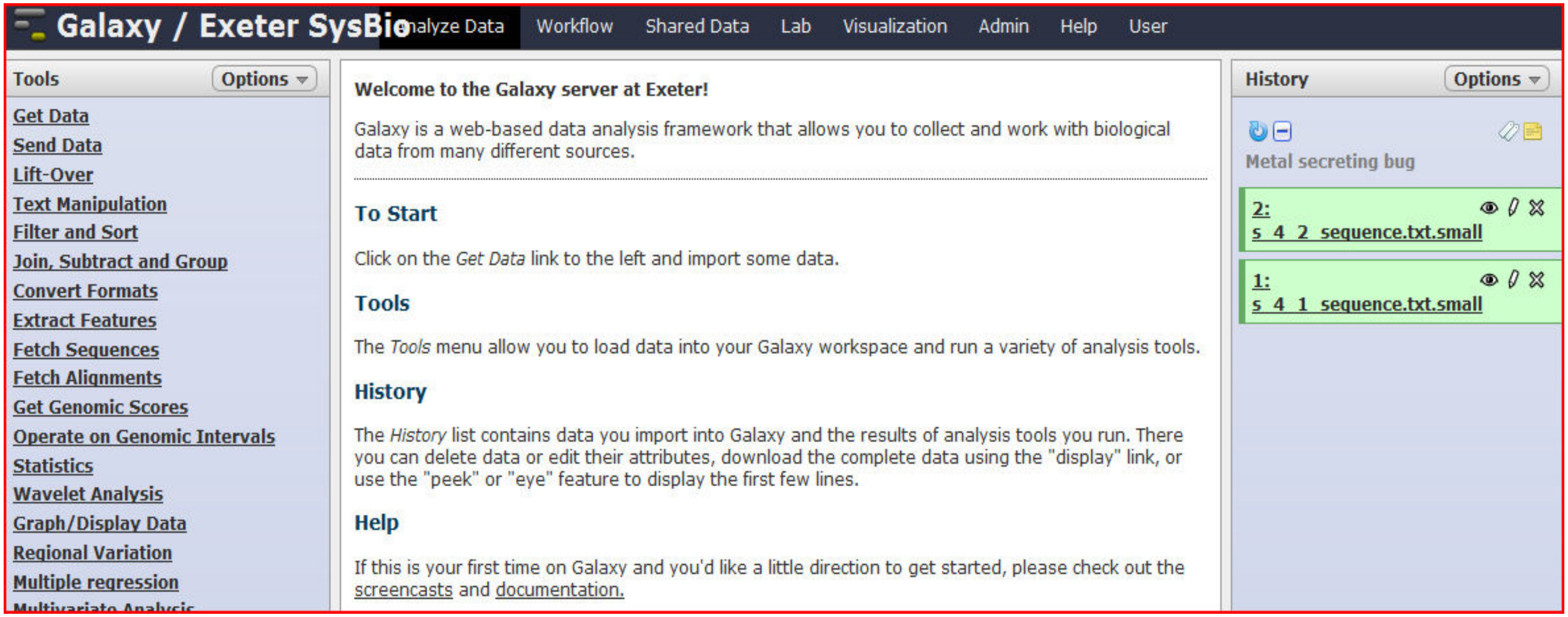
Use this option if you are entering intervals by hand.

Genome: Click to Search or Select

Or (maybe) directly from SRA/ENA

Or via Galaxy LIMS

Uploading FASTQ files



Welcome to the Galaxy server at Exeter!

Galaxy is a web-based data analysis framework that allows you to collect and work with biological data from many different sources.

To Start

Click on the *Get Data* link to the left and import some data.

Tools

The *Tools* menu allow you to load data into your Galaxy workspace and run a variety of analysis tools.

History

The *History* list contains data you import into Galaxy and the results of analysis tools you run. There you can delete data or edit their attributes, download the complete data using the "display" link, or use the "peek" or "eye" feature to display the first few lines.

Help

If this is your first time on Galaxy and you'd like a little direction to get started, please check out the screencasts and documentation.

2. Filtering reads

All platforms have errors and artefacts

Illumina

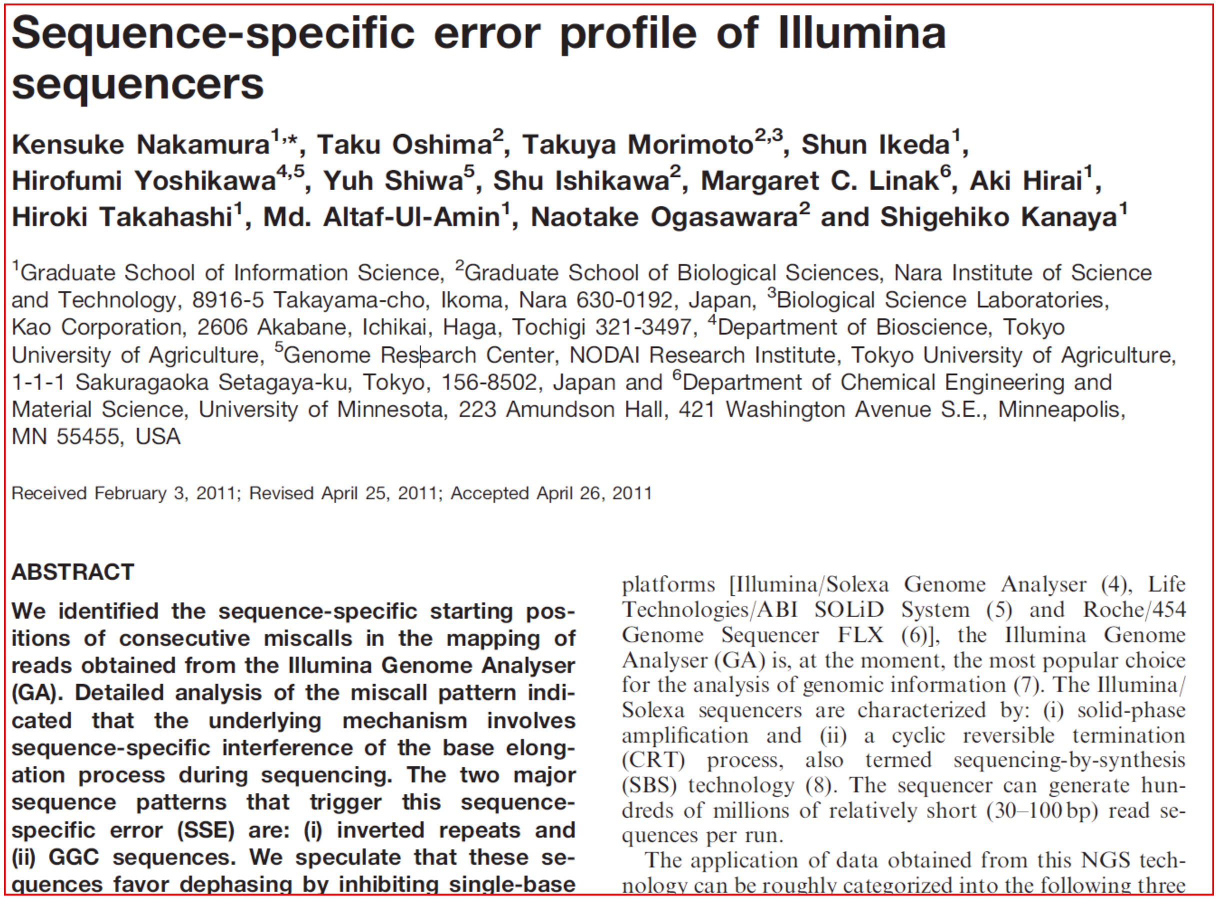
SoLID/ABI-Life

Roche 454

Ion Torrent

1. Removal of low quality bases
2. Removal of adaptor sequences
3. Platform specific artefacts (e.g homopolymers)

Illumina artefacts



Sequence-specific error profile of Illumina sequencers

Kensuke Nakamura[1,*], Taku Oshima[2], Takuya Morimoto[2,3], Shun Ikeda[1], Hirofumi Yoshikawa[4,5], Yuh Shiwa[5], Shu Ishikawa[2], Margaret C. Linak[6], Aki Hirai[1], Hiroki Takahashi[1], Md. Altaf-Ul-Amin[1], Naotake Ogasawara[2] and Shigehiko Kanaya[1]

[1]Graduate School of Information Science, [2]Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan, [3]Biological Science Laboratories, Kao Corporation, 2606 Akabane, Ichikai, Haga, Tochigi 321-3497, [4]Department of Bioscience, Tokyo University of Agriculture, [5]Genome Research Center, NODAI Research Institute, Tokyo University of Agriculture, 1-1-1 Sakuragaoka Setagaya-ku, Tokyo, 156-8502, Japan and [6]Department of Chemical Engineering and Material Science, University of Minnesota, 223 Amundson Hall, 421 Washington Avenue S.E., Minneapolis, MN 55455, USA

Received February 3, 2011; Revised April 25, 2011; Accepted April 26, 2011

ABSTRACT

We identified the sequence-specific starting positions of consecutive miscalls in the mapping of reads obtained from the Illumina Genome Analyser (GA). Detailed analysis of the miscall pattern indicated that the underlying mechanism involves sequence-specific interference of the base elongation process during sequencing. The two major sequence patterns that trigger this sequence-specific error (SSE) are: (i) inverted repeats and (ii) GGC sequences. We speculate that these sequences favor dephasing by inhibiting single-base

platforms [Illumina/Solexa Genome Analyser (4), Life Technologies/ABI SOLiD System (5) and Roche/454 Genome Sequencer FLX (6)], the Illumina Genome Analyser (GA) is, at the moment, the most popular choice for the analysis of genomic information (7). The Illumina/Solexa sequencers are characterized by: (i) solid-phase amplification and (ii) a cyclic reversible termination (CRT) process, also termed sequencing-by-synthesis (SBS) technology (8). The sequencer can generate hundreds of millions of relatively short (30–100 bp) read sequences per run.

The application of data obtained from this NGS technology can be roughly categorized into the following three

Nakamura, K. et al Sequence-specific error profile of Illumina sequencers

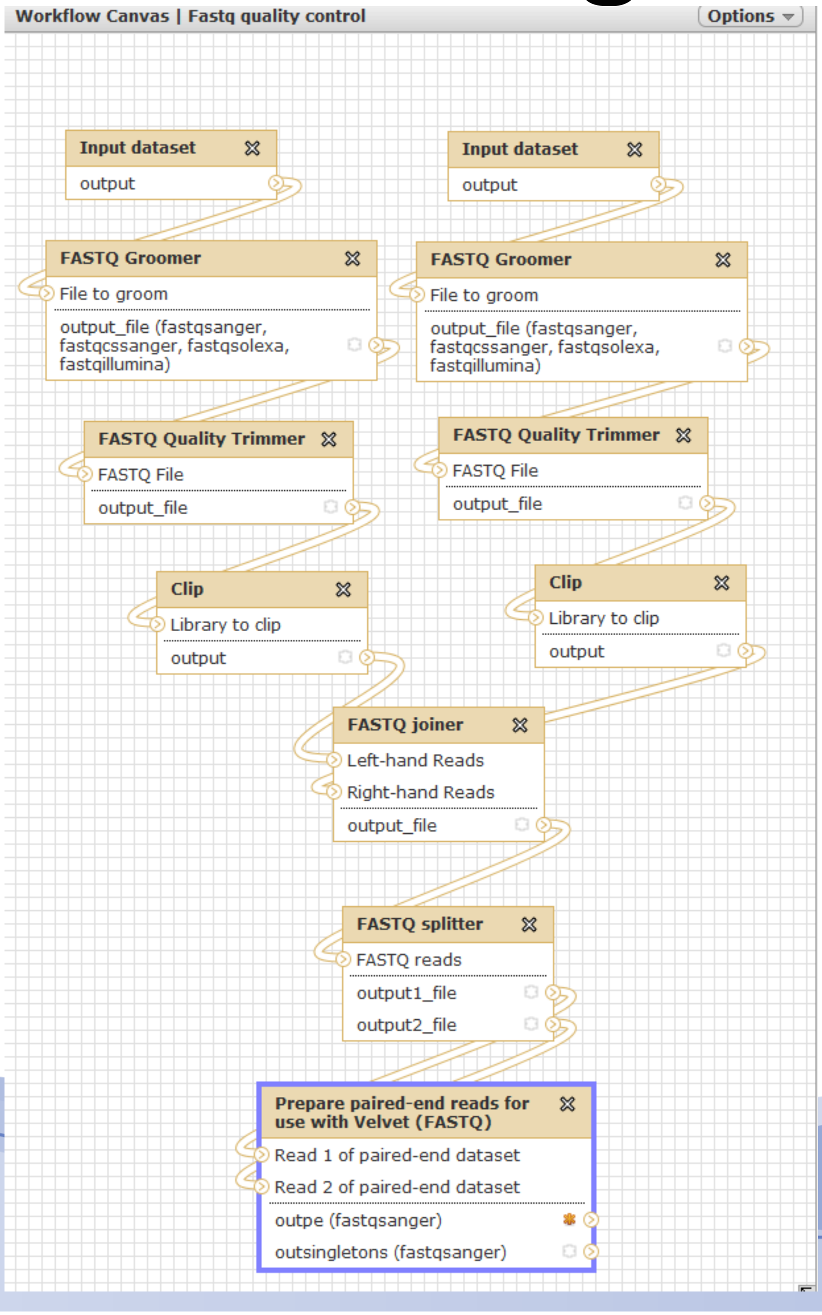
Nucl. Acids Res. (2011) May 16, 2011

Illumina artefacts

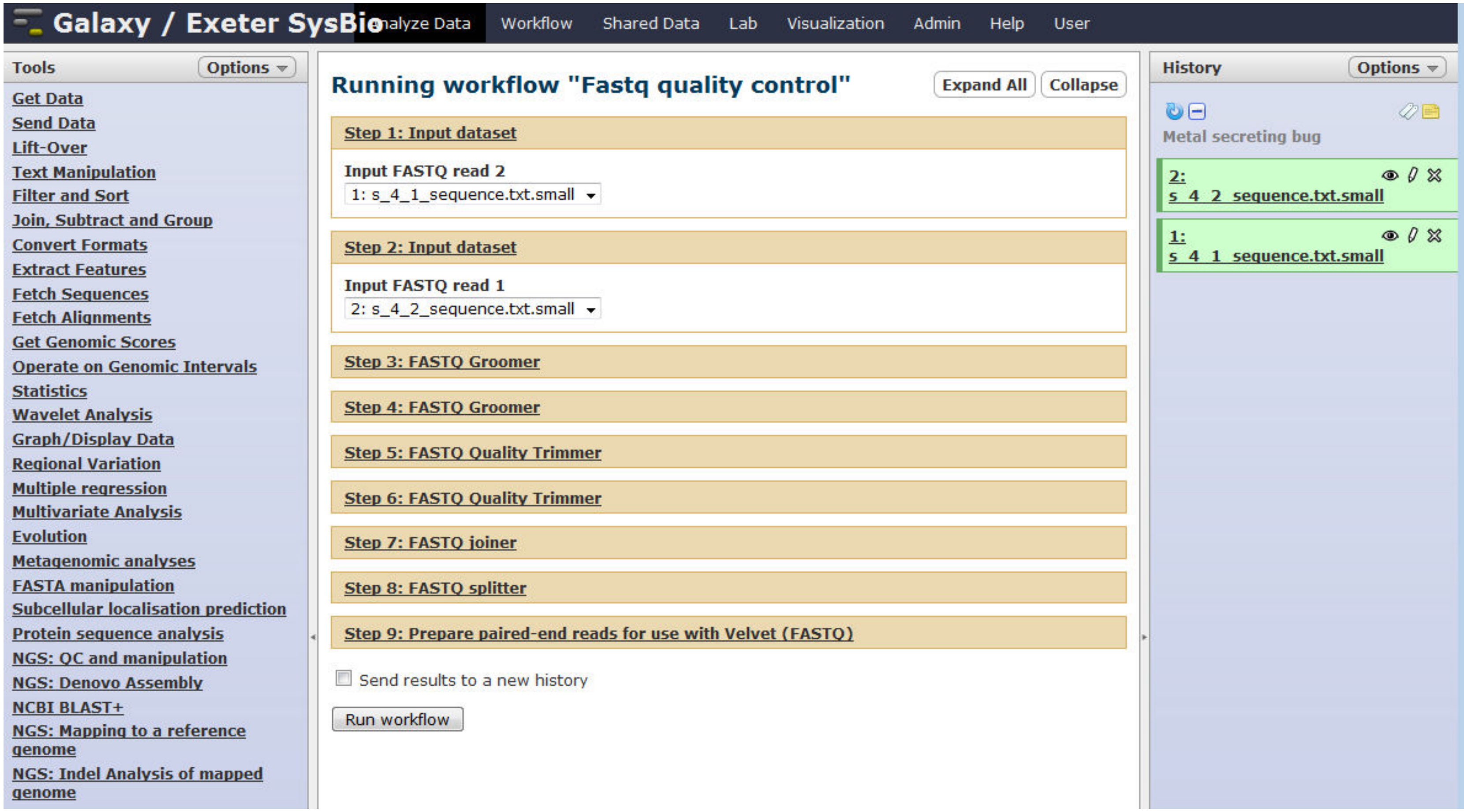
1. GC rich regions are under represented
   a. PCR
   b. Sequencing
2. Substitutions more common than insertions
3. GGC/GCC motif is associated with low quality and mismatches
4. Filtering low quality reads exacerbates low coverage of GC regions

Assembly and/or filtering software should account for this technology specific bias but doesn't yet

Quality controlling workflow



Quality controlling workflow



Running workflow "Fastq quality control"

Step 1: Input dataset

Input FASTQ read 2: 1: s_4_1_sequence.txt.small

Step 2: Input dataset

Input FASTQ read 1: 2: s_4_2_sequence.txt.small

Step 3: FASTQ Groomer

Step 4: FASTQ Groomer

Step 5: FASTQ Quality Trimmer

Step 6: FASTQ Quality Trimmer

Step 7: FASTQ joiner

Step 8: FASTQ splitter

Step 9: Prepare paired-end reads for use with Velvet (FASTQ)

☐ Send results to a new history

Run workflow

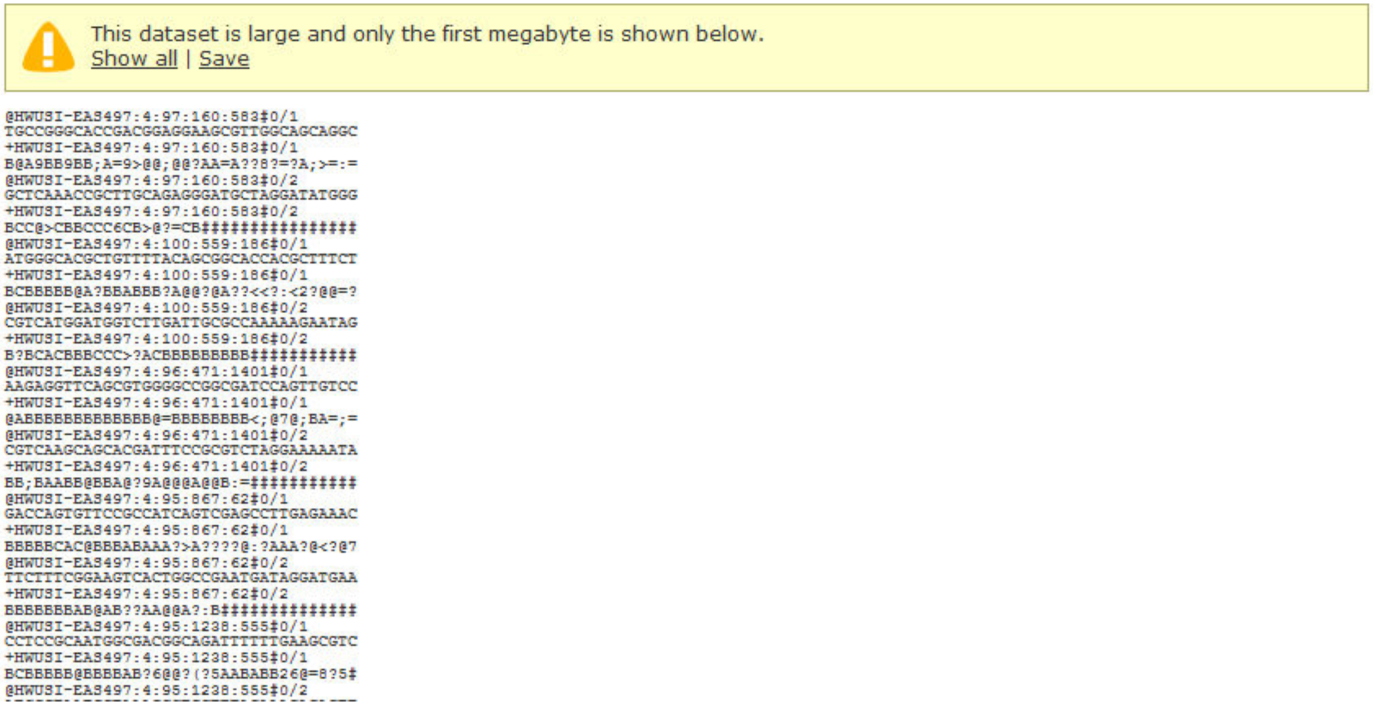
History: Metal secreting bug

2: s_4_2_sequence.txt.small

1: s_4_1_sequence.txt.small

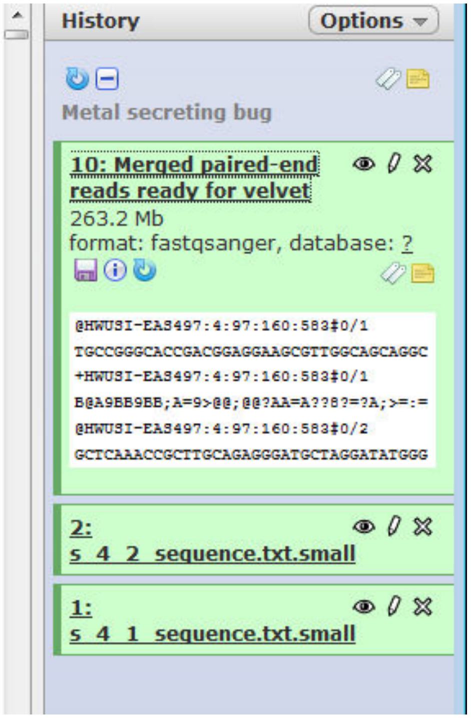
UNIVERSITY OF EXETER

Quality controlling workflow



This dataset is large and only the first megabyte is shown below.
Show all | Save

```
@HWUSI-EAS497:4:97:160:583#0/1
TGCCGGGCACCGACGGAGGAAGCGTTGGCAGCAGGC
+HWUSI-EAS497:4:97:160:583#0/1
B@A9BB9BB;A=9>@@;@@?AA=A??8?=?A;>=:=
@HWUSI-EAS497:4:97:160:583#0/2
GCTCAAACCGCTTGCAGAGGGATGCTAGGATATGGG
+HWUSI-EAS497:4:97:160:583#0/2
BCC@>CBBCCC6CB>@?=CB################
@HWUSI-EAS497:4:100:559:186#0/1
ATGGGCACGCTGTTTTACAGCGGCACCACGCTTTCT
+HWUSI-EAS497:4:100:559:186#0/1
BCBBBBB@A?BBABBB?A@@?@A??<<?:<2?@@=?
@HWUSI-EAS497:4:100:559:186#0/2
CGTCATGGATGGTCTTGATTGCGCCAAAAAGAATAG
+HWUSI-EAS497:4:100:559:186#0/2
B?BCACBBBCCC>?ACBBBBBBBBB###########
@HWUSI-EAS497:4:96:471:1401#0/1
AAGAGGTTCAGCGTGGGGCCGGCGATCCAGTTGTCC
+HWUSI-EAS497:4:96:471:1401#0/1
@ABBBBBBBBBBBBB@=BBBBBBBB<;@7@;BA=;=
@HWUSI-EAS497:4:96:471:1401#0/2
CGTCAAGCAGCACGATTTCCGCGTCTAGGAAAAATA
+HWUSI-EAS497:4:96:471:1401#0/2
BB;BAABB@BBA@?9A@@@A@@B:=###########
@HWUSI-EAS497:4:95:867:62#0/1
GACCAGTGTTCCGCCATCAGTCGAGCCTTGAGAAAC
+HWUSI-EAS497:4:95:867:62#0/1
BBBBBCAC@BBBABAAA?>A????@:?AAA?@<?@7
@HWUSI-EAS497:4:95:867:62#0/2
TTCTTTCGGAAGTCACTGGCCGAATGATAGGATGAA
+HWUSI-EAS497:4:95:867:62#0/2
BBBBBBBAB@AB??AA@@A?:B##############
@HWUSI-EAS497:4:95:1238:555#0/1
CCTCCGCAATGGCGACGGCAGATTTTTTGAAGCGTC
+HWUSI-EAS497:4:95:1238:555#0/1
BCBBBBB@BBBBAB?6@@?(?5AABABB26@=8?5#
@HWUSI-EAS497:4:95:1238:555#0/2
```



Metal secreting bug

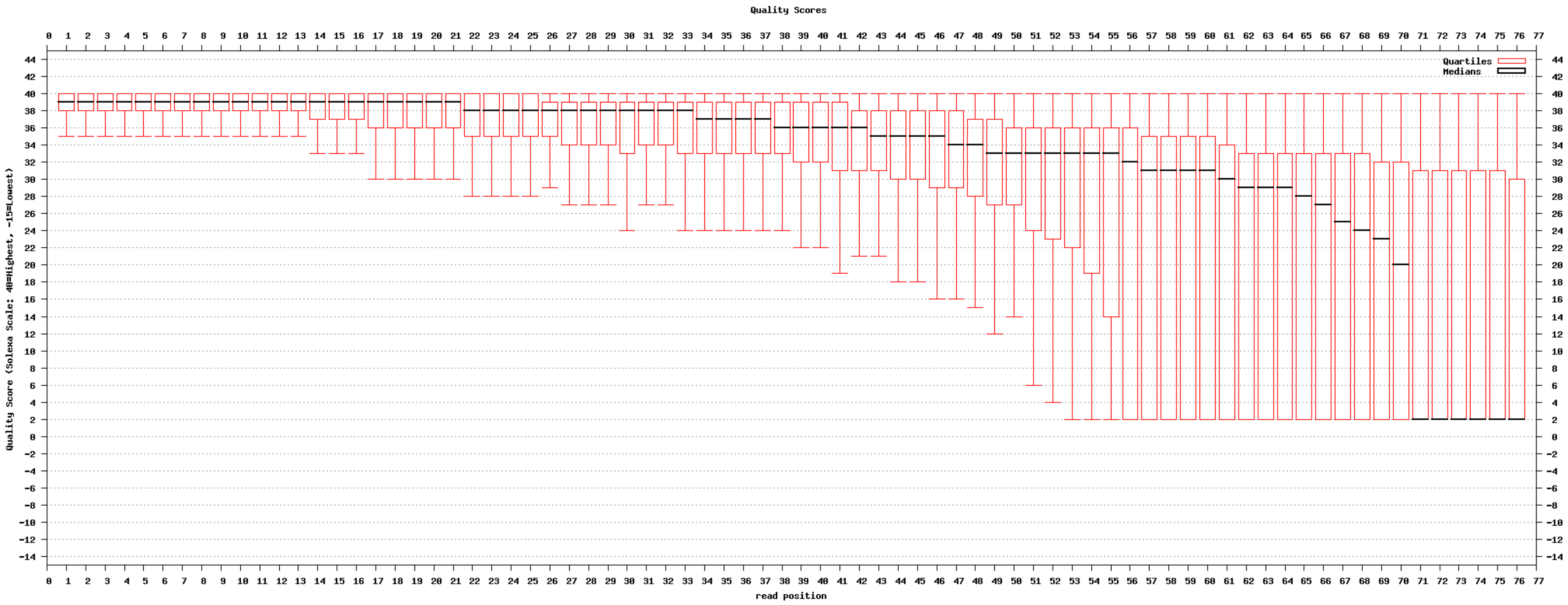
10: Merged paired-end reads ready for velvet
263.2 Mb
format: fastqsanger, database: ?

```
@HWUSI-EAS497:4:97:160:583#0/1
TGCCGGGCACCGACGGAGGAAGCGTTGGCAGCAGGC
+HWUSI-EAS497:4:97:160:583#0/1
B@A9BB9BB;A=9>@@;@@?AA=A??8?=?A;>=:=
@HWUSI-EAS497:4:97:160:583#0/2
GCTCAAACCGCTTGCAGAGGGATGCTAGGATATGGG
```

2: s_4_2_sequence.txt.small

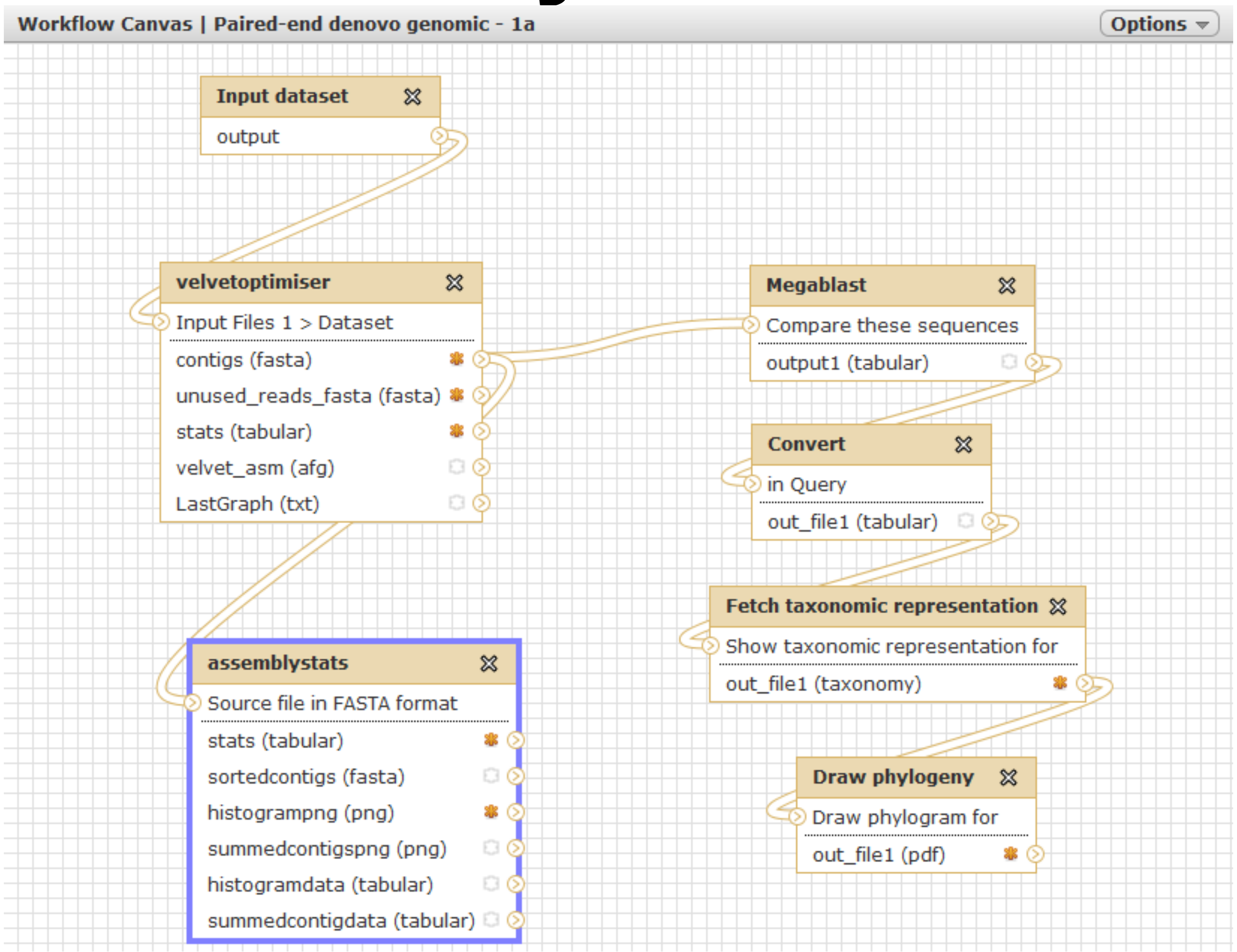
1: s_4_1_sequence.txt.small

Quality visual summaries



3. De-novo Assembly

Assembly workflow



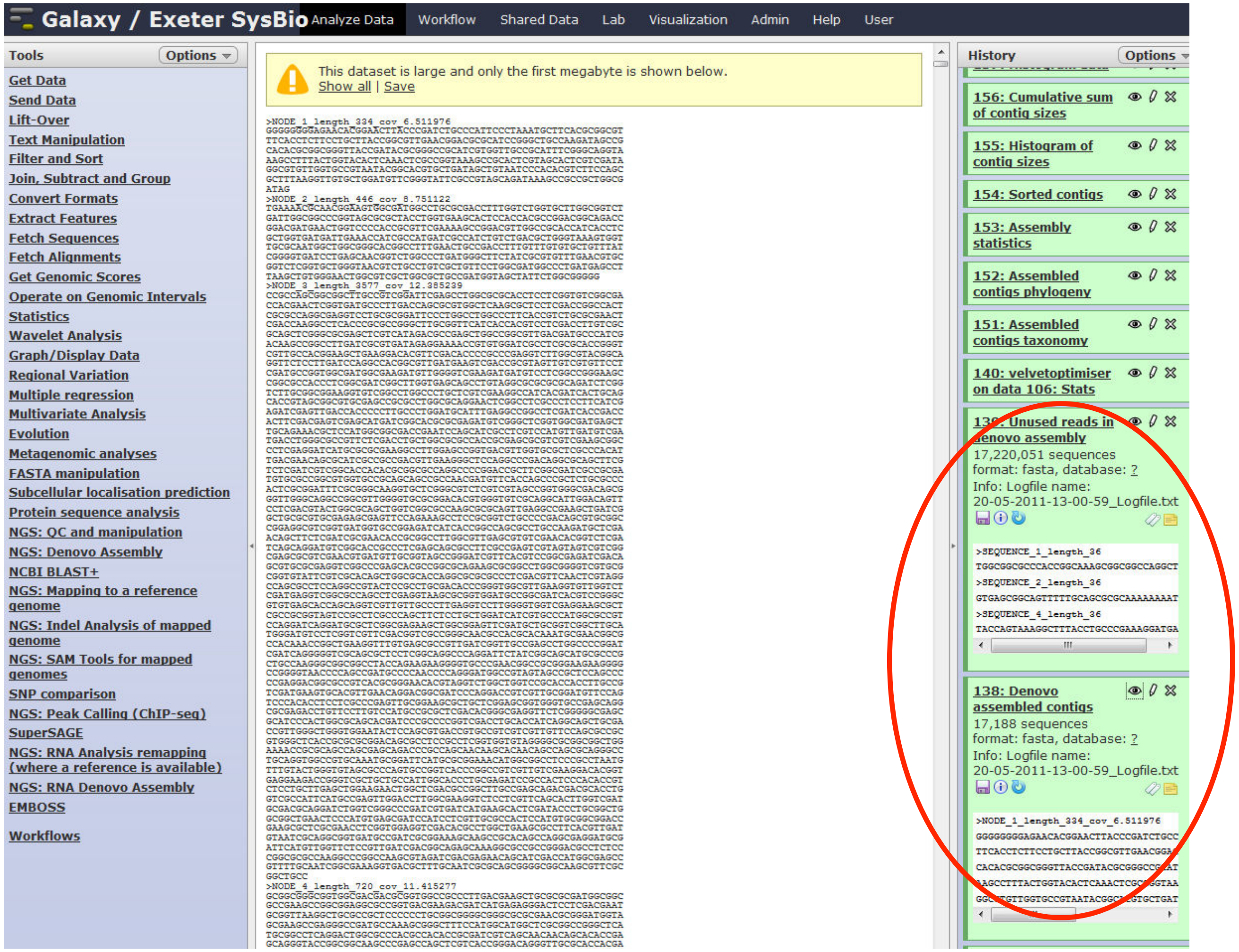
Velvet optimiser for genomic de-novo assembly

- De-bruijn graph assembler
- Runs a selection of k-mer lengths and parameters
- Selects optimum assembly based on contig length and N50 size (adjustable)
- Originally written by Simon Gladman, CSIRO
- Available at the Galaxy Tool Shed

However...

- We need a method of benchmarking the assembly using biological knowledge
- GC value
- Genome size ~ Total number of bp in contigs?
- Fraction of genes fully assembled
  - Measured against closely related genome
- Manual finishing, gap closure only if really necessary
- Most assemblies only need to be ‘good-enough’... whatever that means...

Assembly results



This dataset is large and only the first megabyte is shown below.
Show all | Save

```
>NODE_1_length_334_cov_6.511976
GGGGGGGGAGAACACGGAACTTACCCGATCTGCCCATTCCCTAAATGCTTCACGCGGCGT
TTCACCTCTTCCTGCTTACCGGCGTTGAACGGACGCGCATCCGGGCTGCCAAGATAGCCG
CACACGCGGCGGGTTACCGATACGCGGGCCGCATCGTGGTTGCCGCATTTCGGGCAGGTA
AAGCCTTTACTGGTACACTCAAACTCGCCGGTAAAGCCGCACTCGTAGCACTCGTCGATA
GGCGTGTTGGTGCCGTAATACGGCACGTGCTGATAGCTGTAATCCCACACGTCTTCCAGC
GCTTTAAGGTTGTGCTGGATGTTCGGGTATTCGCCGTAGCAGATAAAGCCGCCGCTGGCG
ATAG
>NODE_2_length_446_cov_8.751122
TGAAAACGCAACGGAAGTGGCGATGGCCTGCGCGACCTTTGGTCTGGTGCTTGGCGGTCT
GATTGGCGGCCCGGTAGCGCGCTACCTGGTGAAGCACTCCACCACGCCGGACGGCAGACC
GGACGATGAACTGGTCCCCACCGCGTTCGAAAAGCCGGACGTTGGCCGCACCATCACCTC
GCTGGTGATGATTGAAACCATCGCCATGATCGCCATCTGTCTGACGCTGGGTAAAGTGGT
TGCGCAATGGCTGGCGGGCACGGCCTTTGAACTGCCGACCTTTGTTTGTGTGCTGTTTAT
CGGGGTGATCCTGAGCAACGGTCTGGCCCTGATGGGCTTCTATCGCGTGTTTGAACGTGC
GGTCTCGGTGCTGGGTAACGTCTGCCTGTCGCTGTTCCTGGCGATGGCCCTGATGAGCCT
TAAGCTGTGGGAACTGGCGTCGCTGGCGCTGCCGATGGTAGCTATTCTGGCGGGGG
>NODE_3_length_3577_cov_12.385239
CCGCCAGCGGCGGCTTGCCGTCGGATTCGAGCCTGGCGCGCACCTCCTCGGTGTCGGCGA
CCACGAACTCGGTGATGCCCTTGACCAGCGCGTGGCTCAAGCGCTCCTCGACCGGCCACT
CGCGCCAGGCGAGGTCCTGCGCGGATTCCCTGGCCTGGCCCTTCACCGTCTGCGCGAACT
CGACCAAGGCCTCACCCGCGCCGGGCTTGCGGTTCATCACCACGTCCTCGACCTTGTCGC
GCAGCTCGGGCGCGAGCTCGTCATAGACGCCGAGCTGGCCGGCGTTGACGATGCCCATCG
ACAAGCCGGCCTTGATCGCGTGATAGAGGAAAACCGTGTGGATCGCCTCGCGCACCGGGT
CGTTGCCACGGAAGCTGAAGGACACGTTCGACACCCCGCCCGAGGTCTTGGCGTACGGCA
GGTTCTCCTTGATCCAGGCCACGGCGTTGATGAAGTCGACCGCGTAGTTGTCGTGTTCCT
CGATGCCGGTGGCGATGGCGAAGATGTTGGGGTCGAAGATGATGTCCTCGGCCGGGAAGC
CGGCGCCACCCTCGGCGATCGGCTTGGTGAGCAGCCTGTAGGCGCGCGCGCAGATCTCGG
TCTTGCGGCGGAAGGTGTCGGCCTGGCCCTGCTCGTCGAAGGCCATCACGATCACTGCAG
CACCGTAGCGGCGTGCGAGCCGCGCCTGGCGCAGGAACTCGGCCTCGCCCTCCTTCATCG
AGATCGAGTTGACCACCCCCTTGCCCTGGATGCATTTGAGGCCGGCCTCGATCACCGACC
ACTTCGACGAGTCGAGCATGATCGGCACGCGCGAGATGTCGGGCTCGGTGGCGATGAGCT
TGCAGAAACGCTCCATGGCGGCGACCGAATCCAGCATCGCCTCGTCCATGTTGATGTCGA
TGACCTGGGCGCCGTTCTCGACCTGCTGGCGCGCCACCGCGAGCGCGTCGTCGAAGCGGC
CCTCGAGGATCATGCGCGCGAAGGCCTTGGAGCCGGTGACGTTGGTGCGCTCGCCCACAT
TGACGAACAGCGCATCGCCGCCGACGTTGAAGGGCTCCAGGCCCGACAGGCGCAGCTTCG
TCTCGATCGTCGGCACCACACGCGGCGCCAGGCCCCGGACCGCTTCGGCGATCGCCGCGA
TGTGCGCCGGCGTGGTGCCGCAGCAGCCGCCAACGATGTTCACCAGCCCGCTCTGCGCCC
ACTCGCGGATTTCGCGGGCAAGGTGCTCGGGCGTCTCGTCGTAGCCGGTGGGCGACAGCG
GGTTGGGCAGGCCGGCGTTGGGGTGCGCGGACACGTGGGTGTCGCAGGCATTGGACAGTT
CCTCGACGTACTGGCGCAGCTGGTCGGCGCCAAGCGCGCAGTTGAGGCCGAAGCTGATCG
GCTGCGCGTGCGAGAGCGAGTTCCAGAAAGCCTCCGCGGTCTGCCCCGACAGCGTGCGGC
CGGAGGCGTCGGTGATGGTGCCGGAGATCATCACCGGCCAGCGCCTGCCAAGATGCTCGA
ACAGCTTCTCGATCGCGAACACCGCGGCCTTGGCGTTGAGCGTGTCGAACACGGTCTCGA
TCAGCAGGATGTCGGCACCGCCCTCGAGCAGCGCCTTCGCCGAGTCGTAGTAGTCGTCGG
CGAGCGCGTCGAACGTGATGTTGCGGTAGCCGGGATCGTTCACGTCCGGCGAGATCGACA
GCGTGCGCGAGGTCGGCCCGAGCACGCCGGCGCAGAAGCGCGGCCTGGCGGGGTCGTGCG
CGGTGTATTCGTCGCACAGCTGGCGCACCAGGCGCGCGCCCTCGACGTTCAACTCGTAGG
CCAGCGCCTCCAGGCCGTACTCCGCCTGCGACACCCGGGTGGCGTTGAAGGTGTTGGTCT
CGATGAGGTCGGCGCCAGCCTCGAGGTAAGCGCGGTGGATGCCGGCGATCACGTCCGGGC
GTGTGAGCACCAGCAGGTCGTTGTTGCCCTTGAGGTCCTTGGGGTGGTCGAGGAAGCGCT
CGCCGCGGTAGTCCGCCTCGCCCAGCTTCTCCTGCTGGATCATCGTGCCCATGGCGCCGT
CCAGGATCAGGATGCGCTCGGCGAGAAGCTGGCGGAGTTCGATGCTGCGGTCGGCTTGCA
TGGGATGTCCTCGGTCGTTCGACGGTCGCCGGGCAACGCCACGCACAAATGCGAACGGCG
CCACAAACCGGCTGAAGGTTTGTGAGCGCCGTTGATCGGTTGCCGAGCCTGGCCCCGGAT
CGATCAGGGGGTCGCAGCGCTCCTCGGCAGGCCCAGGATTCTATCGGCAGCATGCGCCCG
CTGCCAAGGGCGGCGGCCTACCAGAAGAAGGGGTGCCCGAACGGCCGCGGGAAGAAGGGG
CCGGGGTAACCCCAGCCGATGCCCCAACCCCAGGGATGGCCGTAGTAGCCGCTCCAGCCC
CCGAGGACGGCGCCGTCACGCGGGAACACGTAGGTCTGGCTGGTCCGCACCACCTTGCCG
TCGATGAAGTGCACGTTGAACAGGACGGCGATCCCAGGACCGTCGTTGCGGATGTTCCAG
TCCCACACCTCCTCGCCCGAGTTGCGGAAGCGCTGCTCGGAGCGGTGGGTGCCGAGCAGG
CGCGAGACCTGTTCCTTGTCCATGCCGCGCTCGACACGGGCGAGGTTCTCGGGGGCGAGC
GCATCCCACTGGCGCAGCACGATCCCGCCCCGGTCGACCTGCACCATCAGGCAGCTGCGA
CCGTTGGGCTGGGTGGAATACTCCAGCGTGACCGTGCCGTCGTCGTTGTTCCAGCGCCGC
GTGGGCTCACCGCGCGCGGACAGCGCCTCCGCCTCGGTGGTGTAGGGGCGCGGCGGCTGG
AAAACCGCGCAGCCAGCGAGCAGACCCGCCAGCAACAAGCACAACAGCCAGCGCAGGGCC
TGCAGGTGGCCGTGCAAATGCGGATTCATGCGCGGAAACATGGCGGCCTCCCGCCTAATG
TTTGTACTGGGTGTAGCGCCCAGTGCCGGTCACCCGGCCGTCGTTGTCGAAGGACACGGT
GAGGAAGACCGGGTCGCTGCTGCCATTGGCACCCTGCGAGATCCGCCACTCCCACACCGT
CTCCTGCTTGAGCTGGAAGAACTGGCTCGACGCCGGCTTGCCGAGCAGACGACGCACCTG
GTCGCCATTCATGCCGAGTTGGACCTTGGCGAAGGTCTCCTCGTTCAGCACTTGGTCGAT
GCGACGCAGGATCTGGTCGGGCCCGATCGTGATCATGAAGCACTCGATACCCTGCGGCTG
GCGGCTGAACTCCCATGTGAGCGATCCATCCTCGTTGCGCCACTCCATGTGCGGCGGACC
GAAGCGCTCGCGAACCTCGGTGGAGGTCGACACGCCTGGCTGAAGCGCCTTCACGTTGAT
GTAATCGCAGGCGGTGATGCCGATCGCGGAAAGCAAGCCGCACAGCCAGGCGAGGATGCG
ATTCATGTTGGTTCTCCGTTGATCGACGGCAGAGCAAAGGCGCCGCCGGGACGCCTCTCC
CGGCGCGCCAAGGCCCGGCCAAGCGTAGATCGACGAGAACAGCATCGACCATGGCGAGCC
GTTTTGCAATCGGCGAAAGGTGACGCTTTGCAATCGCGCAGCGGGGCGGCAAGCGTTCGC
GGCTGCC
>NODE_4_length_720_cov_11.415277
GCGGCGGGCGGTGGCGACGACGCGGTGGCCGCCCTTGACGAAGCTGCGCGCGATGGCGGC
GCCGAAGCCGGCGGAGGCGCCGGTGACGAAGACGATCATGAGAGGGACTCCTCGACGAAT
GCGGTTAAGGCTGCGCCGCTCCCCCCTGCGGCGGGGCGGGCGCGCGAACGCGGGATGGTA
GCGAAGCCGAGGGCCGATGCCAAAGCGGGCTTTCCATGGCATGGCTCGCGGCCGGGCTCA
TGCGGCCTCAGGACTGGCGCCCACGCCACACCGCGATCGTCAGCAACAACAGCACACCGA
GCAGGGTACCGGCGGCAAGCCCGAGCCAGCTCGTCACCGGGACAGGGTTGCGCACCACGA
```

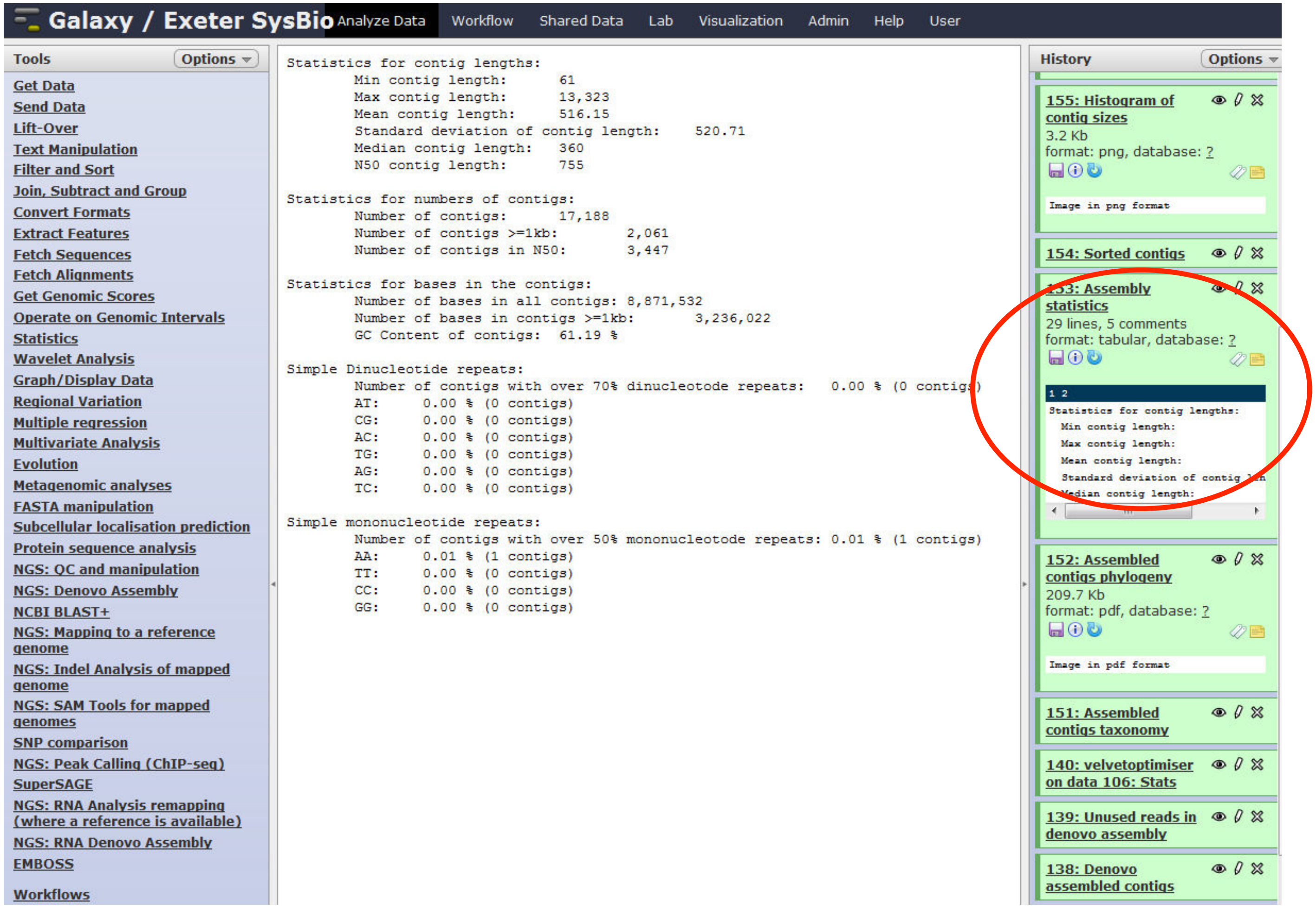
139: Unused reads in denovo assembly
17,220,051 sequences
format: fasta, database: ?
Info: Logfile name:
20-05-2011-13-00-59_Logfile.txt

```
>SEQUENCE_1_length_36
TGGCGGCGCCCACCGGCAAAGCGGCGGCCAGGCT
>SEQUENCE_2_length_36
GTGAGCGGCAGTTTTTGCAGCGCGCAAAAAAAAT
>SEQUENCE_4_length_36
TACCAGTAAAGGCTTTACCTGCCCGAAAGGATGA
```

138: Denovo assembled contigs
17,188 sequences
format: fasta, database: ?
Info: Logfile name:
20-05-2011-13-00-59_Logfile.txt

```
>NODE_1_length_334_cov_6.511976
GGGGGGGGAGAACACGGAACTTACCCGATCTGCC
TTCACCTCTTCCTGCTTACCGGCGTTGAACGGA
CACACGCGGCGGGTTACCGATACGCGGGCCGCAT
AAGCCTTTACTGGTACACTCAAACTCGCCGGTAA
GGCGTGTTGGTGCCGTAATACGGCACGTGCTGAT
```

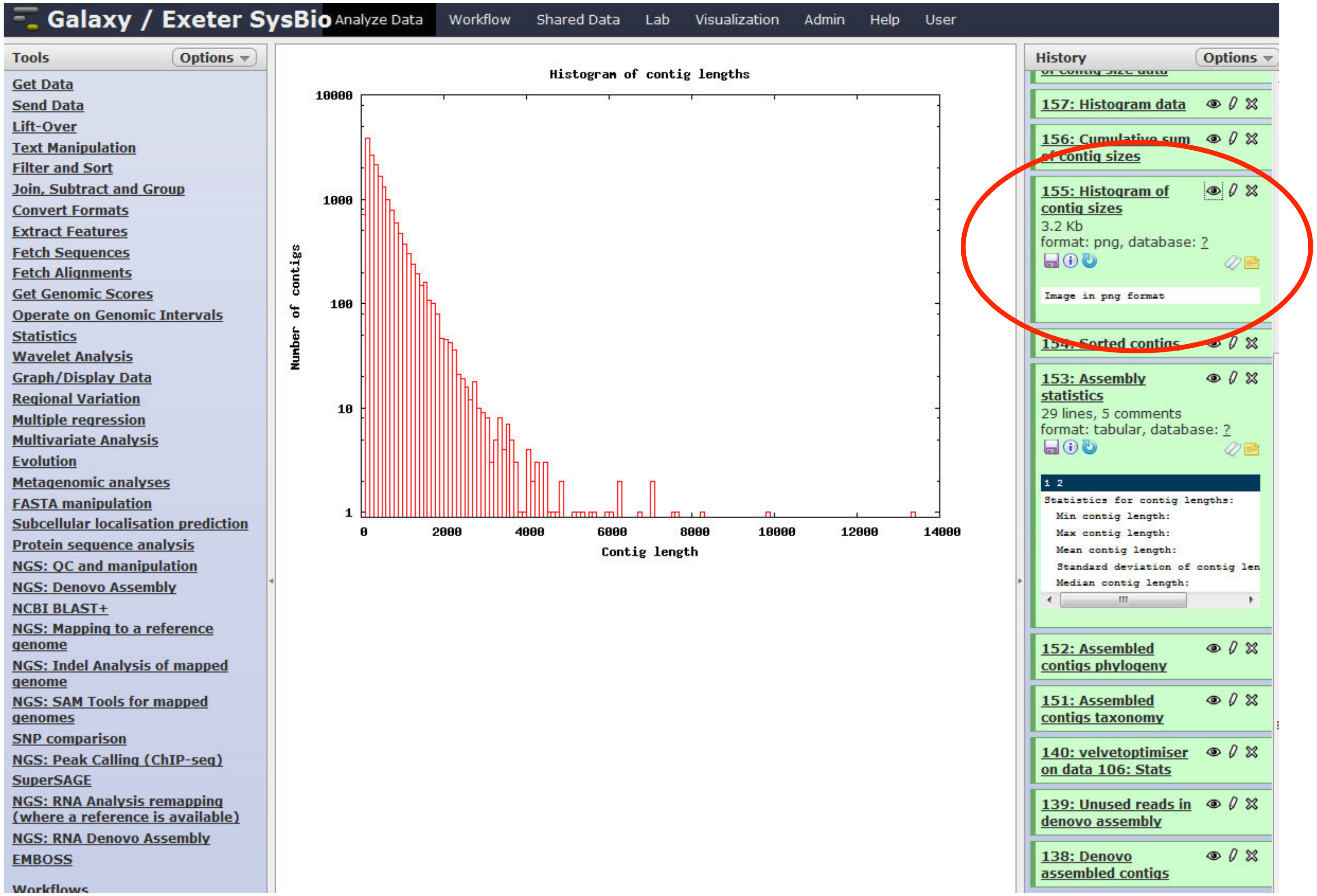
Assembly statistics



```
Statistics for contig lengths:
        Min contig length:      61
        Max contig length:      13,323
        Mean contig length:     516.15
        Standard deviation of contig length:    520.71
        Median contig length:   360
        N50 contig length:      755

Statistics for numbers of contigs:
        Number of contigs:      17,188
        Number of contigs >=1kb:        2,061
        Number of contigs in N50:       3,447

Statistics for bases in the contigs:
        Number of bases in all contigs: 8,871,532
        Number of bases in contigs >=1kb:       3,236,022
        GC Content of contigs:  61.19 %

Simple Dinucleotide repeats:
        Number of contigs with over 70% dinucleotode repeats:   0.00 % (0 contigs)
        AT:     0.00 % (0 contigs)
        CG:     0.00 % (0 contigs)
        AC:     0.00 % (0 contigs)
        TG:     0.00 % (0 contigs)
        AG:     0.00 % (0 contigs)
        TC:     0.00 % (0 contigs)

Simple mononucleotide repeats:
        Number of contigs with over 50% mononucleotode repeats: 0.01 % (1 contigs)
        AA:     0.01 % (1 contigs)
        TT:     0.00 % (0 contigs)
        CC:     0.00 % (0 contigs)
        GG:     0.00 % (0 contigs)
```

```
1 2
Statistics for contig lengths:
  Min contig length:
  Max contig length:
  Mean contig length:
  Standard deviation of contig len
  Median contig length:
```

Assembly statistics



157: Histogram data

156: Cumulative sum of contig sizes

155: Histogram of contig sizes
3.2 Kb
format: png, database: ?

Image in png format

154: Sorted contigs

153: Assembly statistics
29 lines, 5 comments
format: tabular, database: ?

```
1 2
Statistics for contig lengths:
  Min contig length:
  Max contig length:
  Mean contig length:
  Standard deviation of contig len
  Median contig length:
```

152: Assembled contigs phylogeny

151: Assembled contigs taxonomy

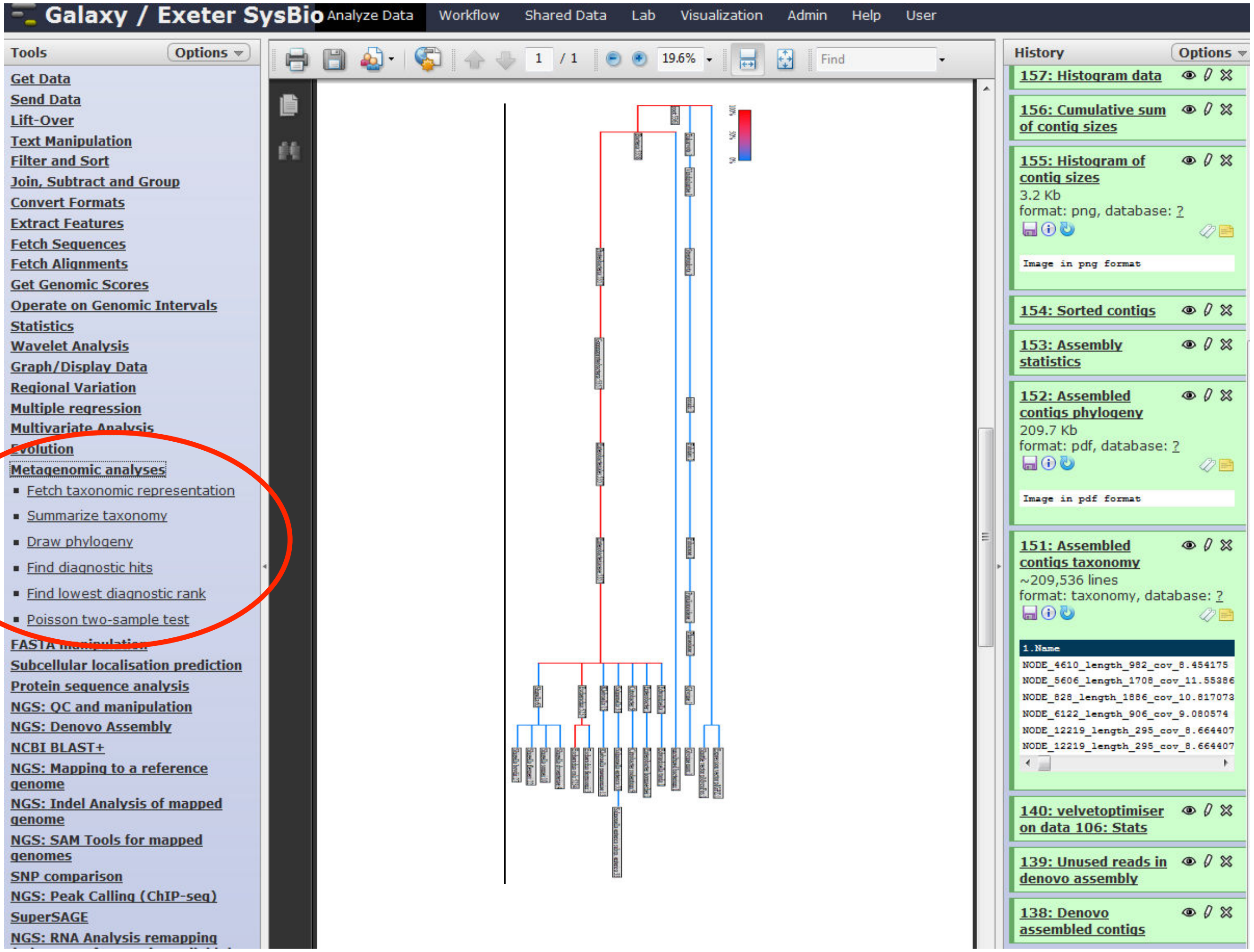
140: velvetoptimiser on data 106: Stats

139: Unused reads in denovo assembly

138: Denovo assembled contigs

UNIVERSITY OF EXETER

Taxonomy of contigs



157: Histogram data

156: Cumulative sum of contig sizes

155: Histogram of contig sizes
3.2 Kb
format: png, database: ?

Image in png format

154: Sorted contigs

153: Assembly statistics

152: Assembled contigs phylogeny
209.7 Kb
format: pdf, database: ?

Image in pdf format

151: Assembled contigs taxonomy
~209,536 lines
format: taxonomy, database: ?

```
1.Name
NODE_4610_length_982_cov_8.454175
NODE_5606_length_1708_cov_11.55386
NODE_828_length_1886_cov_10.817073
NODE_6122_length_906_cov_9.080574
NODE_12219_length_295_cov_8.664407
NODE_12219_length_295_cov_8.664407
```

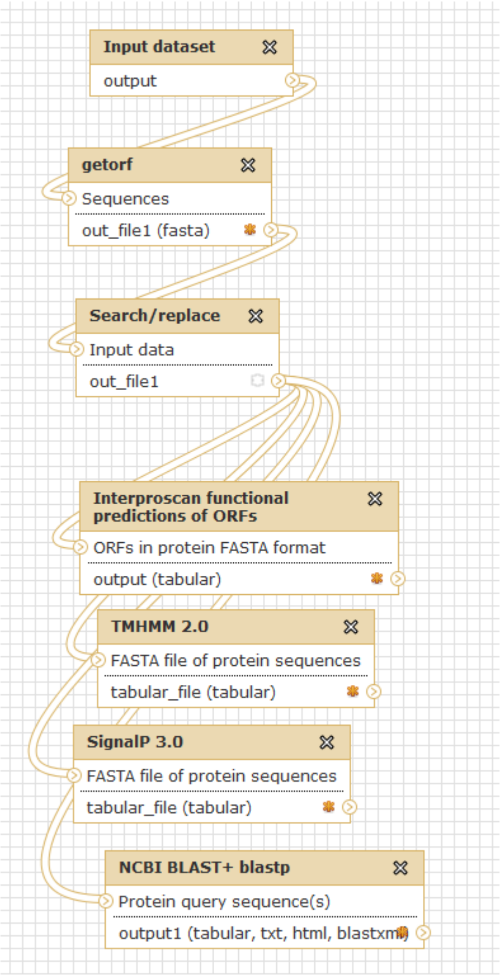
140: velvetoptimiser on data 106: Stats

139: Unused reads in denovo assembly

138: Denovo assembled contigs

4. Annotation

Annotation workflow



Input dataset

output

getorf

Sequences

out_file1 (fasta)

Search/replace

Input data

out_file1

Interproscan functional predictions of ORFs

ORFs in protein FASTA format

output (tabular)

TMHMM 2.0

FASTA file of protein sequences

tabular_file (tabular)

SignalP 3.0

FASTA file of protein sequences

tabular_file (tabular)

NCBI BLAST+ blastp

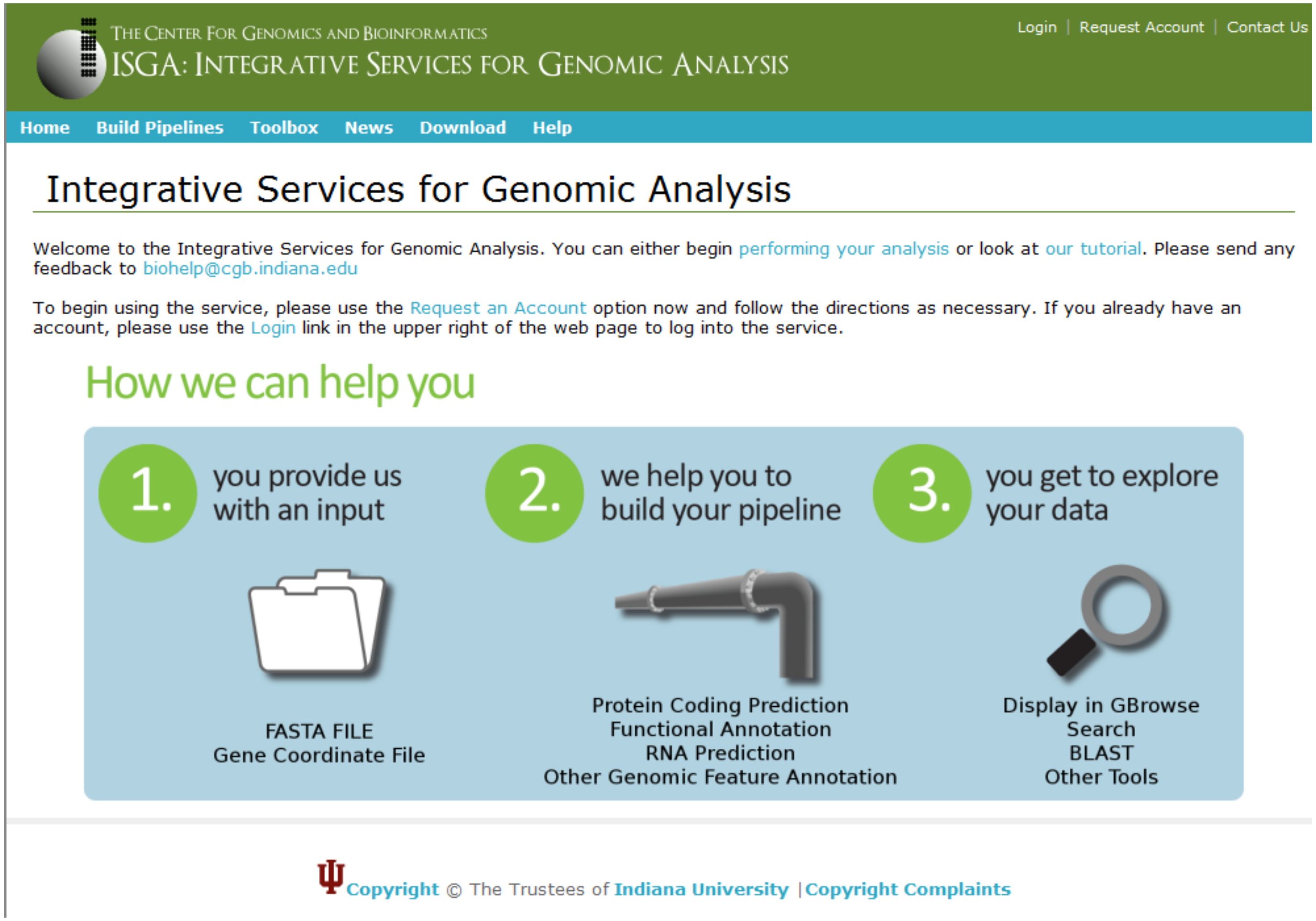
Protein query sequence(s)

output1 (tabular, txt, html, blastxml)

Still to be included

- De-novo gene prediction
- EST and other evidence needs to be included
- tRNAs
- RepeatMasker
- Non-coding features
- Other annotation software pipelines

Can we incorporate these?



THE CENTER FOR GENOMICS AND BIOINFORMATICS
ISGA: INTEGRATIVE SERVICES FOR GENOMIC ANALYSIS
Login | Request Account | Contact Us
Home
Build Pipelines
Toolbox
News
Download
Help

Integrative Services for Genomic Analysis

Welcome to the Integrative Services for Genomic Analysis. You can either begin performing your analysis or look at our tutorial. Please send any feedback to biohelp@cgb.indiana.edu

To begin using the service, please use the Request an Account option now and follow the directions as necessary. If you already have an account, please use the Login link in the upper right of the web page to log into the service.

How we can help you

1. you provide us with an input

FASTA FILE
Gene Coordinate File

2. we help you to build your pipeline

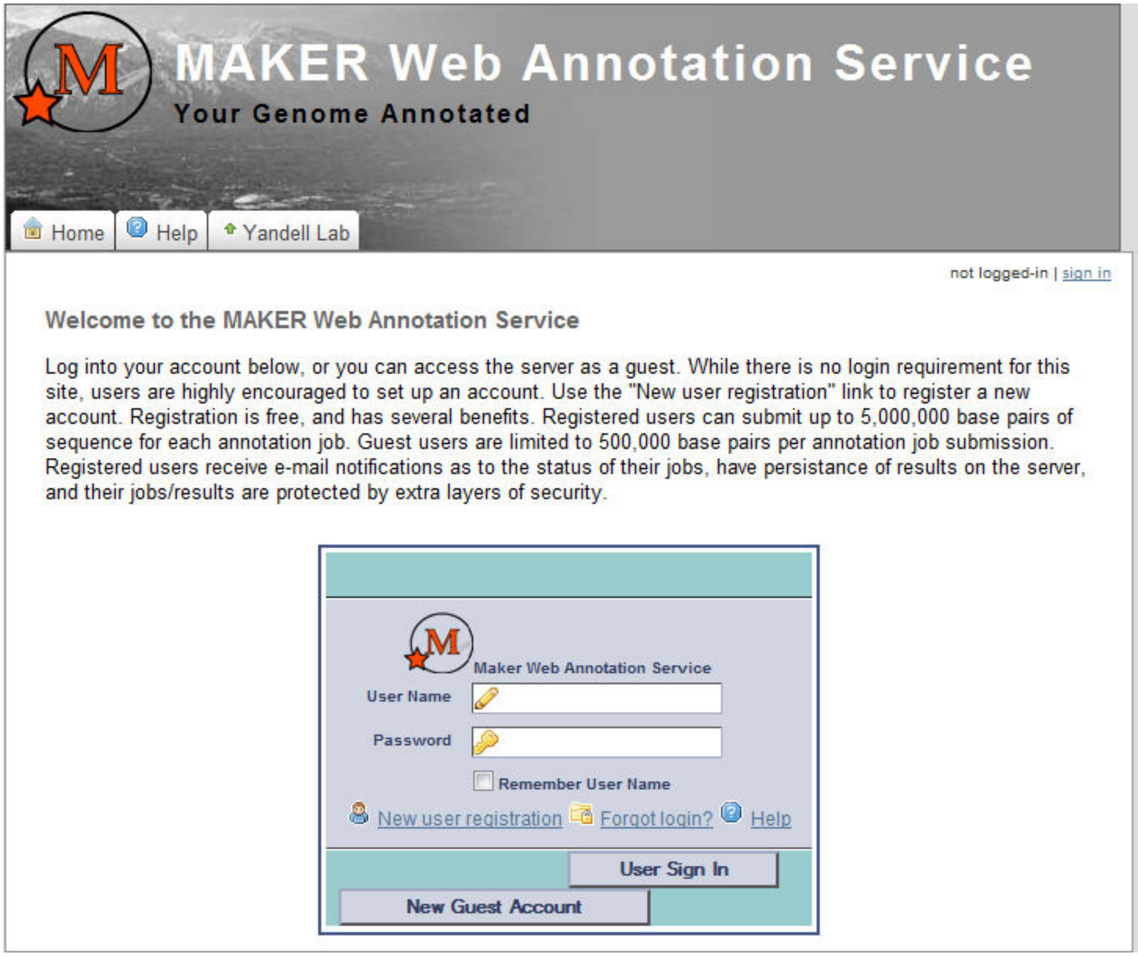
Protein Coding Prediction
Functional Annotation
RNA Prediction
Other Genomic Feature Annotation

3. you get to explore your data

Display in GBrowse
Search
BLAST
Other Tools

Copyright © The Trustees of Indiana University | Copyright Complaints

Can we incorporate these?



Welcome to the MAKER Web Annotation Service

Log into your account below, or you can access the server as a guest. While there is no login requirement for this site, users are highly encouraged to set up an account. Use the "New user registration" link to register a new account. Registration is free, and has several benefits. Registered users can submit up to 5,000,000 base pairs of sequence for each annotation job. Guest users are limited to 500,000 base pairs per annotation job submission. Registered users receive e-mail notifications as to the status of their jobs, have persistance of results on the server, and their jobs/results are protected by extra layers of security.

Do we want to incorporate these?

Is the service sustainable if it becomes really popular?

If so: Locally? Web services?

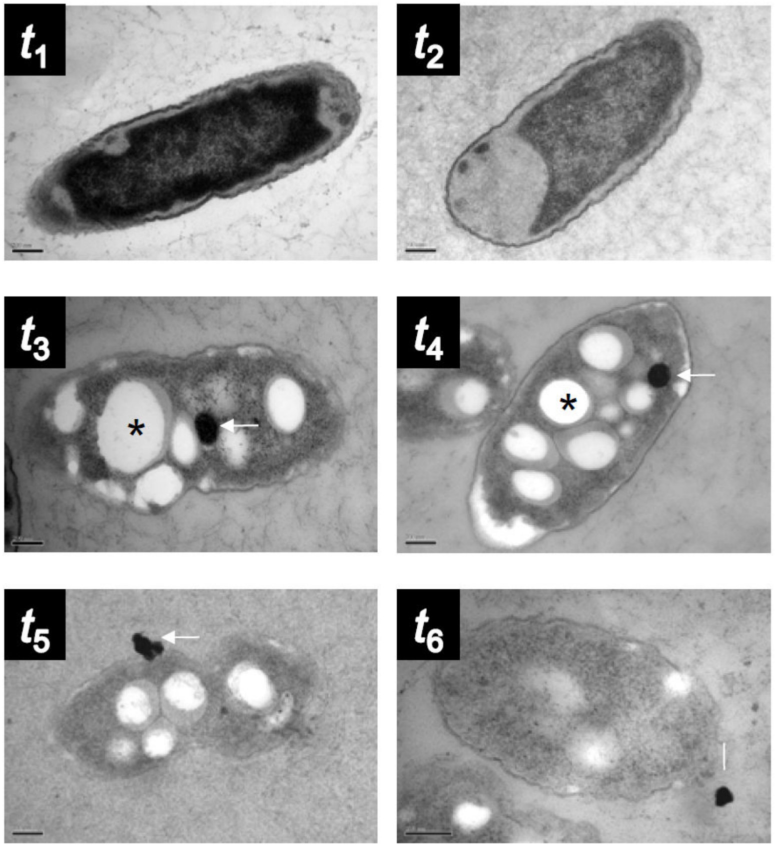
Denovo sequencing project

A new bacterium which secretes elemental metal

60% GC content

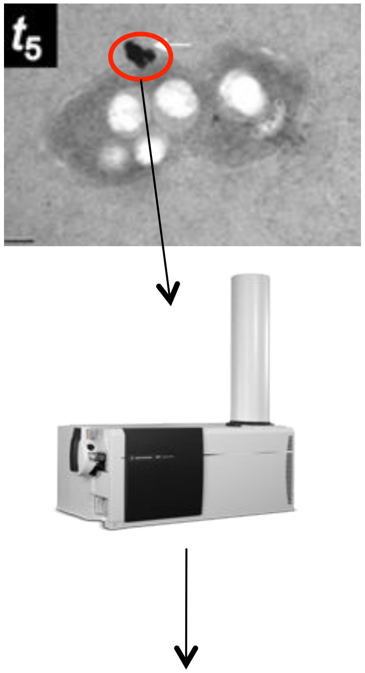
Approximately 8 Mb genome

Aim: Which genes(s) are responsible for translocation?



5. Where is the secretory protein?

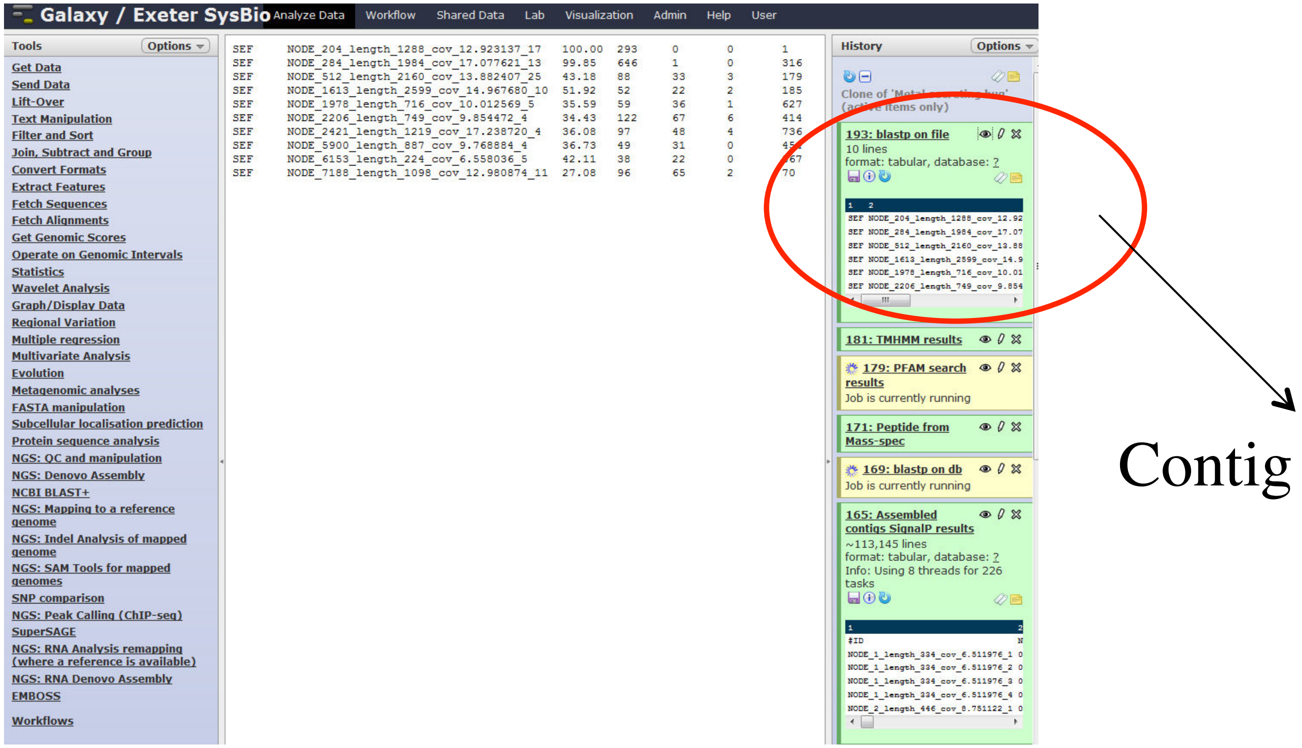
Mass spectrometry evidence



MTITASQSRTEVVVRSA..

Locate peptide within contigs ORFs using BlastP

MTITASQSRTEVVVRSA....



```
SEF     NODE_204_length_1288_cov_12.923137_17   100.00  293     0       0       1
SEF     NODE_284_length_1984_cov_17.077621_13   99.85   646     1       0       316
SEF     NODE_512_length_2160_cov_13.882407_25   43.18   88      33      3       179
SEF     NODE_1613_length_2599_cov_14.967680_10  51.92   52      22      2       185
SEF     NODE_1978_length_716_cov_10.012569_5    35.59   59      36      1       627
SEF     NODE_2206_length_749_cov_9.854472_4     34.43   122     67      6       414
SEF     NODE_2421_length_1219_cov_17.238720_4   36.08   97      48      4       736
SEF     NODE_5900_length_887_cov_9.768884_4     36.73   49      31      0       45
SEF     NODE_6153_length_224_cov_6.558036_5     42.11   38      22      0       67
SEF     NODE_7188_length_1098_cov_12.980874_11  27.08   96      65      2       70
```

Clone of 'Metal accreting bug'
(active items only)

193: blastp on file
10 lines
format: tabular, database: ?

```
1   2
SEF NODE_204_length_1288_cov_12.92
SEF NODE_284_length_1984_cov_17.07
SEF NODE_512_length_2160_cov_13.88
SEF NODE_1613_length_2599_cov_14.9
SEF NODE_1978_length_716_cov_10.01
SEF NODE_2206_length_749_cov_9.854
```

181: TMHMM results

179: PFAM search results
Job is currently running

171: Peptide from Mass-spec

169: blastp on db
Job is currently running

165: Assembled contigs SignalP results
~113,145 lines
format: tabular, database: ?
Info: Using 8 threads for 226 tasks

```
1                                 2
#ID                               N
NODE_1_length_334_cov_6.511976_1  0
NODE_1_length_334_cov_6.511976_2  0
NODE_1_length_334_cov_6.511976_3  0
NODE_1_length_334_cov_6.511976_4  0
NODE_2_length_446_cov_8.751122_1  0
```

Contig 204 ORF 17

Check with annotation tools

- SignalP predicts a signal peptide using both NN and HMM
- TMHMM also predicts that the peptide is external
- PFAM reports a DUF (Domain of Unknown Function)
- BlastP NR reports Hypothetical proteins

ORF located and characterised as coding for a novel metal export factor

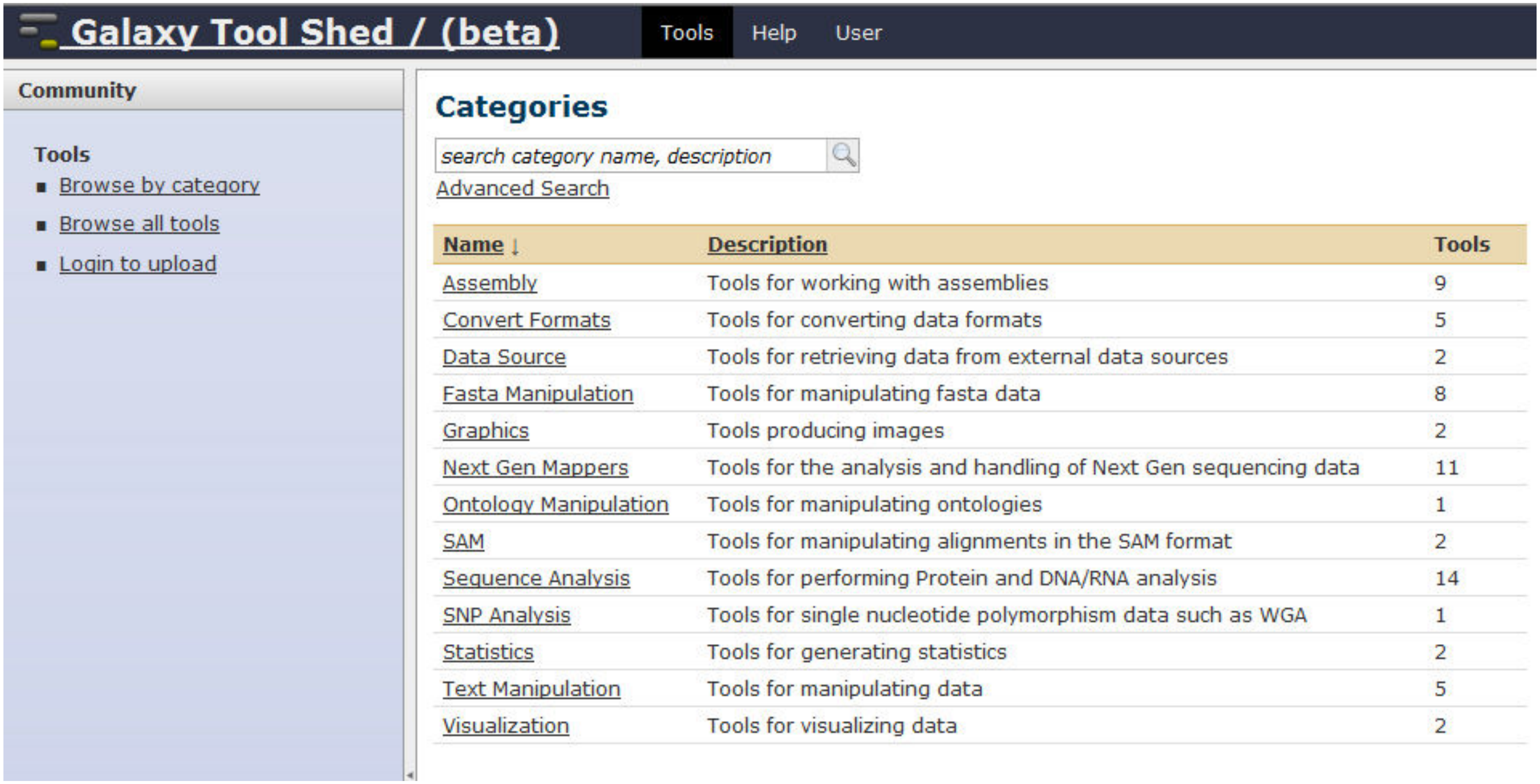
Summary

- Filtered and formatted raw data
- Assembled a draft 8 Mb genome – no finishing
- Evaluated metrics and taxonomy of contigs
- Called ORFs bacterial codon usage table
- Basic annotation with BlastP against NCBI NR
- PFAM, SignalP, TMHMM
- Identified peptide within contigs
- No hits in PFAM, NCBI NR. Signal peptide present
- Time frame < 1 day

Other assemblers

- Minimus2 (Galaxy wrapper by Edward Kirton)
  - Merge contigs from different assemblies
- MIRA (Galaxy wrapper by Peter Cock, SCRI)
  - Recent upgrades for PacBio and Ion Torrent
- AbySS (Galaxy wrapper by Edward Kirton)
- Newbler (Galaxy wrapper by Edward Kirton)
  - Roche/454 proprietary assembler and remapper
- Phrap (Galaxy wrapper by Edward Kirton)
  - Sanger read assembly
- String Graph Assembler (Jared Simpson, Sanger)
  - Useful for large (> human) genomes with short reads

Available at Galaxy Toolshed



Tools

- Browse by category
- Browse all tools
- Login to upload

Categories

search category name, description

Advanced Search

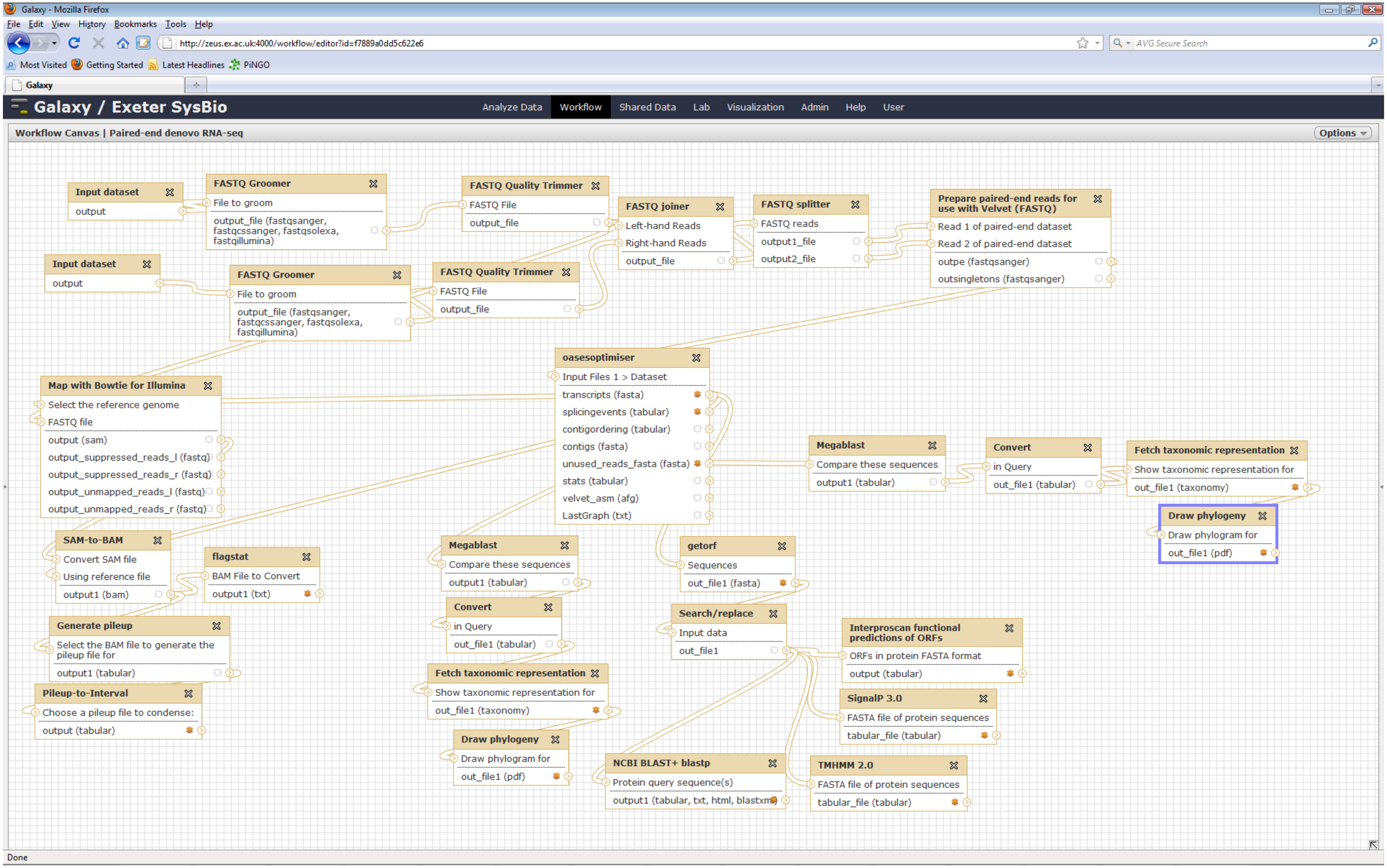
Name ↓	Description	Tools
Assembly	Tools for working with assemblies	9
Convert Formats	Tools for converting data formats	5
Data Source	Tools for retrieving data from external data sources	2
Fasta Manipulation	Tools for manipulating fasta data	8
Graphics	Tools producing images	2
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	11
Ontology Manipulation	Tools for manipulating ontologies	1
SAM	Tools for manipulating alignments in the SAM format	2
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	14
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	1
Statistics	Tools for generating statistics	2
Text Manipulation	Tools for manipulating data	5
Visualization	Tools for visualizing data	2

Other applications

Oases optimiser for de-novo RNA-seq

- Sister program of Velvet
- Runs a selection of kmer lengths
- Combines all results
- Uses these as a scaffold to assemble transcripts at shortest kmer length

Galaxy denovo RNA-seq Pipeline



Future developments

Community to-do/wish list

- Adding tools dedicated to evaluating assembly quality (e.g. Using EST sequences or related sequences)
- Tools to aid in finishing assemblies
- AFG or other assembly-format visualisation
- Collating and formatting annotation (e.g. GFF files)
- Metagenomics/transcriptomics (e.g. MetaVelvet)
- Gene prediction software
- Blast2Go
- Comparison of GO or PFAM terms between samples
- Enabling workflows of workflows
- AMOS tools (Amos validate etc), web-services

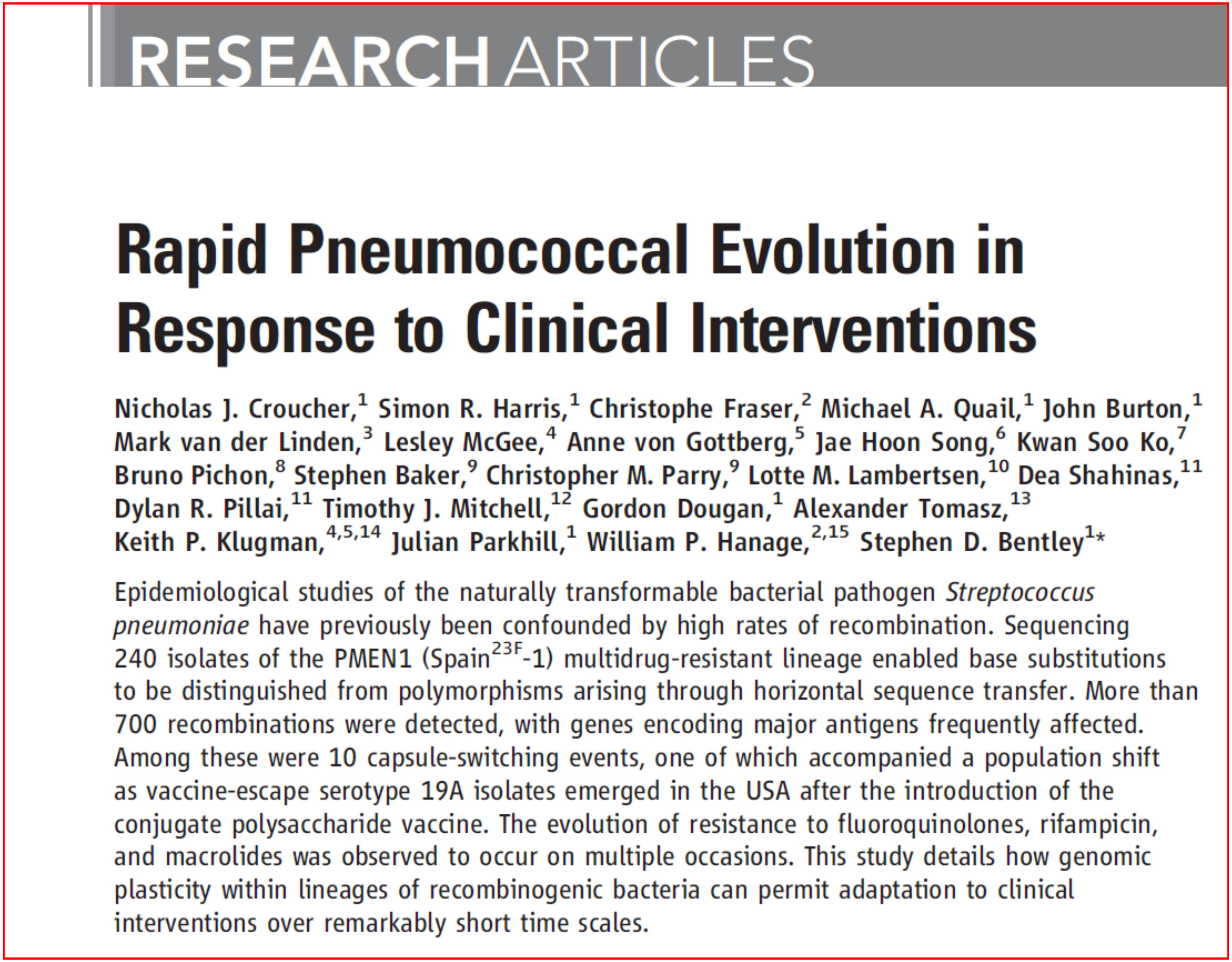
Future developments

A single Illumina GAIIx run can produce data for ~ 100 bacterial genomes in less than a week.

Cost: ~10,000 Euro

Question: How do we deal with 100s of comparisons between datasets in Galaxy?

Do we want to?
Do we have a choice?



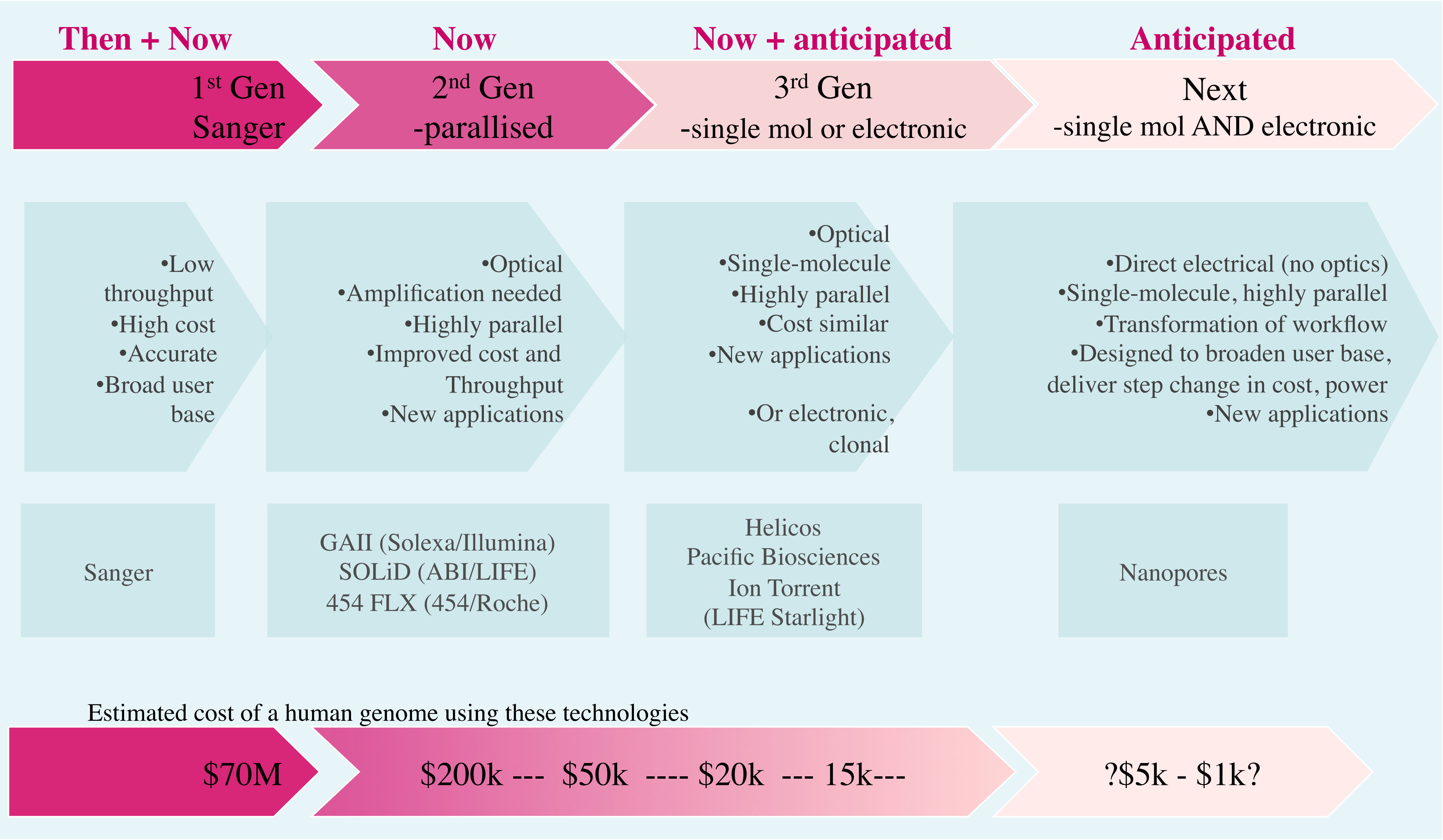
RESEARCH ARTICLES

Rapid Pneumococcal Evolution in Response to Clinical Interventions

Nicholas J. Croucher,[1] Simon R. Harris,[1] Christophe Fraser,[2] Michael A. Quail,[1] John Burton,[1] Mark van der Linden,[3] Lesley McGee,[4] Anne von Gottberg,[5] Jae Hoon Song,[6] Kwan Soo Ko,[7] Bruno Pichon,[8] Stephen Baker,[9] Christopher M. Parry,[9] Lotte M. Lambertsen,[10] Dea Shahinas,[11] Dylan R. Pillai,[11] Timothy J. Mitchell,[12] Gordon Dougan,[1] Alexander Tomasz,[13] Keith P. Klugman,[4,5,14] Julian Parkhill,[1] William P. Hanage,[2,15] Stephen D. Bentley[1]*

Epidemiological studies of the naturally transformable bacterial pathogen *Streptococcus pneumoniae* have previously been confounded by high rates of recombination. Sequencing 240 isolates of the PMEN1 (Spain^{23F}-1) multidrug-resistant lineage enabled base substitutions to be distinguished from polymorphisms arising through horizontal sequence transfer. More than 700 recombinations were detected, with genes encoding major antigens frequently affected. Among these were 10 capsule-switching events, one of which accompanied a population shift as vaccine-escape serotype 19A isolates emerged in the USA after the introduction of the conjugate polysaccharide vaccine. The evolution of resistance to fluoroquinolones, rifampicin, and macrolides was observed to occur on multiple occasions. This study details how genomic plasticity within lineages of recombinogenic bacteria can permit adaptation to clinical interventions over remarkably short time scales.

DNA sequencing generations



Then + Now	Now	Now + anticipated	Anticipated
1st Gen Sanger	2nd Gen -parallised	3rd Gen -single mol or electronic	Next -single mol AND electronic
•Low throughput •High cost •Accurate •Broad user base	•Optical •Amplification needed •Highly parallel •Improved cost and Throughput •New applications	•Optical •Single-molecule •Highly parallel •Cost similar •New applications •Or electronic, clonal	•Direct electrical (no optics) •Single-molecule, highly parallel •Transformation of workflow •Designed to broaden user base, deliver step change in cost, power •New applications
Sanger	GAII (Solexa/Illumina) SOLiD (ABI/LIFE) 454 FLX (454/Roche)	Helicos Pacific Biosciences Ion Torrent (LIFE Starlight)	Nanopores

Estimated cost of a human genome using these technologies

$70M	$200k --- $50k ---- $20k --- 15k---	?$5k - $1k?

Questions?

Konrad Paszkiewicz

k.h.paszkiewicz@exeter.ac.uk

“We need to start thinking about how to train people, both health-care professionals and scientists, to be facile in bioinformatics. We need to foster development of professionals who have expertise analyzing large data sets of the size that biologists haven't had to think about. We need to entice smart people into genomics.”

Eric Green,

Director National Human Genome Research Institute

Acknowledgements

University of Exeter

- Murray Grant
- Karen Moore
- Alex Moorhouse

Peter Cock – James Hutton Institute

Galaxy team and tool shed contributors