Microbial Genomics and Bioinformatics
BM405
1.Introduction

The James Hutton Institute

Leighton Pritchard[1,2,3]
[1]Information and Computational Sciences,
[2]Centre for Human and Animal Pathogens in the Environment,
[3]Dundee Effector Consortium,
The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

Acceptable Use Policy

The James Hutton Institute

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available on SlideShare.

These slides, and supporting material including exercises, are available at https://github.com/widdowquinn/Teaching-Strathclyde-BM405

Table of Contents

The impact[a]

[a] Loman and Pallen (2015) *Nat. Rev. Micro.* doi:10.1038/nrmicro3565

The James Hutton Institute

Genome sequencing and bioinformatics have transformed our understanding of prokaryotic biology:

- function
- evolution
- interactions
- community structure
- real-time monitoring and diagnostics
- as a platform for synthetic biology

It now takes much longer to analyse than generate data

The endpoints

- **2003**: *Erwinia carotovora* subsp. *atroseptica*
- **2015**: *Dickeya* spp., *Campylobacter* spp., and *Escherichia coli*

Eca_Sanger_genome.dna

```
>Erwinia_carotovora
ttatccacattcaactgccgaatatttaatcttatttgagtaaattaacccacgatcctta
gccattcttctgctggatcttctggaatctgatgctgcgtgatgtcgatctcgaagaattt
caccgatccttttagcccccaattgggctagtagacgatccgctgtttggatcgcgccac
agaacgtgtcgtactctttactgccgatccctatcgcaccaaaattcacctgagaaagat
cgggcttttgctgttccaactgttcaaacagcgcttttaggttgtcagggaactccccctg
cgccgtgtgtcgacgttaccaccagccatataccgctctcgggaagtgcatcaagcgtag
ggccatgcagaatttcagtggagaaattatccttttctagcagttcagccaaatgctctg
ccacatattcggcgctgccaagggtactgccactgattaaggtaatgtctgccatagtga
tcccatccgataaaagacgggacattgtacgctgtgatctggctgggatctacctgtgga
taatatgggtataggaaaaaagcattacggggcgatcgtccgcataatggggttttgcag
ggagatcagtgtttcggtggactgaatttcatcaattgtttggatcttgttgataagtac
gtgttgcagcgccctcgatcgaacgacacatgaccttaataaagatactgtaatgcccagt
ggtgtagtaggcctcgaccacttcttccagattgtttagcttttccagcgccgagggata
atctttggcgcttttcagaatgatgccgataaagcagcacacgtcatagcctaactgttt
tggattcacgtccagacgcgtgccgataatgatgcccgcctgtttcatttttccacccg
aacgtggatcgttcctggactgacggaaaactgttttgccagttcggcataaggggtgcg
cgcattttccattaaggcggaaagaatgccgcgatcgagattatcgatctgataaatttc
tgccactttgagcctcctatttttatcgaataatgattttataccgctataaaatgacgga
ttaaaagcacaagtgcaatttatccataagattattgaatgactagcggtttttattgct
taatgaacaactaaatacgatgaataagagacgcggcaatgaaaaagcagtacatcgaaa
agcaacagcaaatcagcttcgttaaatctttcttctccagtcaactggagcagcttctgg
gattgattgaagtccaggcaccgattctcagccgtatcggcgatggcacacaggataact
tgtctggtacagaaaaaagcggtgcaggttaaggtaaaagccttaccggatgccacatttg
aagtggtgcactcactggcaaaatggaaacgcaaaacgctgggcgcgtatgattttagct
tcggtgaaggcatttatactcacatgaaggcgctgcgtccggacgaagatcgcctgagcc
cgatccactcggtttatgtcgatcagtgggattgggagcgcgtgatggaggatggcgagc
gcaatgctgaatatctgaaatcgacggtcacgcgtatttatcaaggcattaaagcgactg
aggctgcggtacatcaggcgtttggcattcagcctttcctgccagagcagattcatttttg
tgcataccgaaaaccttgctgaagcgttatcccgatctggacgccaaagggcgtgagcgag
```

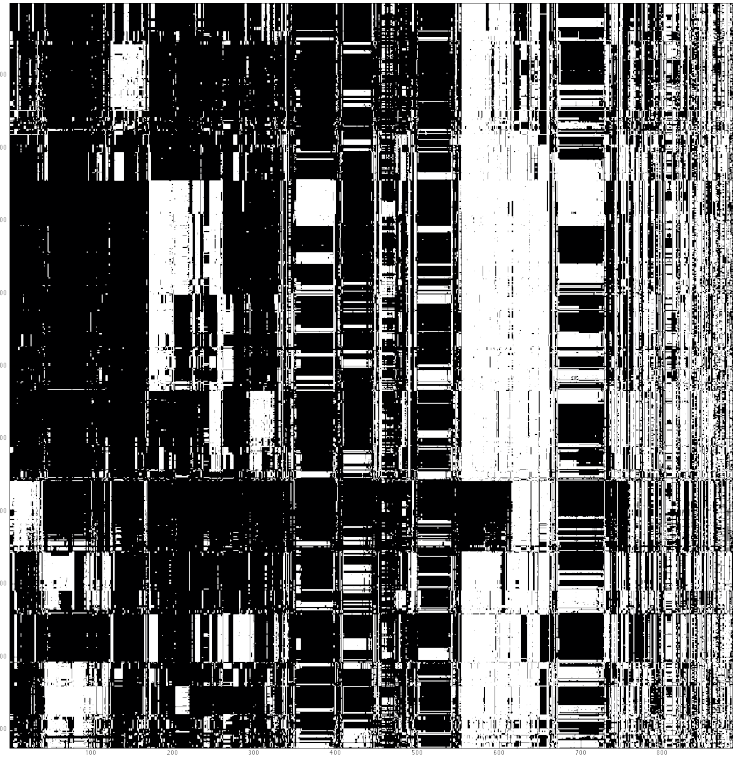
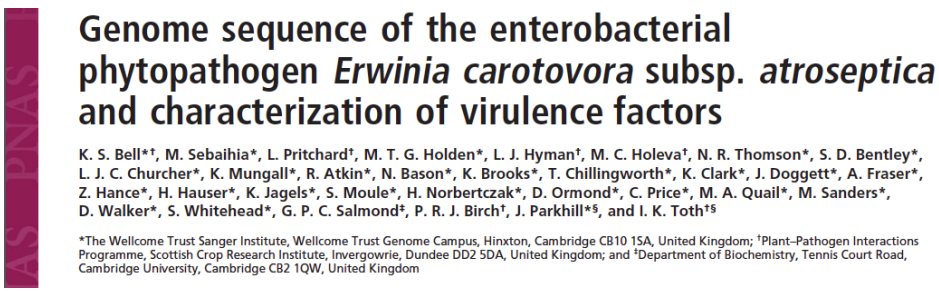


Table of Contents

2003: E. carotovora subsp. atroseptica

The James Hutton Institute

- **£250k collaboration** between SCRI, University of Cambridge, WT Sanger Institute
- Single isolate: *E. carotovora* subsp. *atroseptica* SCRI1043
- The first sequenced enterobacterial plant pathogen (32 authors!) [1]



PNAS

Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors

K. S. Bell*†, M. Sebaihia*, L. Pritchard†, M. T. G. Holden*, L. J. Hyman†, M. C. Holeva†, N. R. Thomson*, S. D. Bentley*, L. J. C. Churcher*, K. Mungall*, R. Atkin*, N. Bason*, K. Brooks*, T. Chillingworth*, K. Clark*, J. Doggett*, A. Fraser*, Z. Hance*, H. Hauser*, K. Jagels*, S. Moule*, H. Norbertczak*, D. Ormond*, C. Price*, M. A. Quail*, M. Sanders*, D. Walker*, S. Whitehead*, G. P. C. Salmond‡, P. R. J. Birch†, J. Parkhill*§, and I. K. Toth†§

*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom; †Plant–Pathogen Interactions Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, United Kingdom; and ‡Department of Biochemistry, Tennis Court Road, Cambridge University, Cambridge CB2 1QW, United Kingdom

- All repeats and gaps bridged and sequenced directly
- **Result**: a single, complete, high-quality 5Mbp circular chromosome at 10.2X coverage: 106,500 reads

[1] Bell *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101**: **30**:11105-11110. doi:10.1073/pnas.0402424101

2003: E. carotovora subsp. atroseptica

The James Hutton Institute

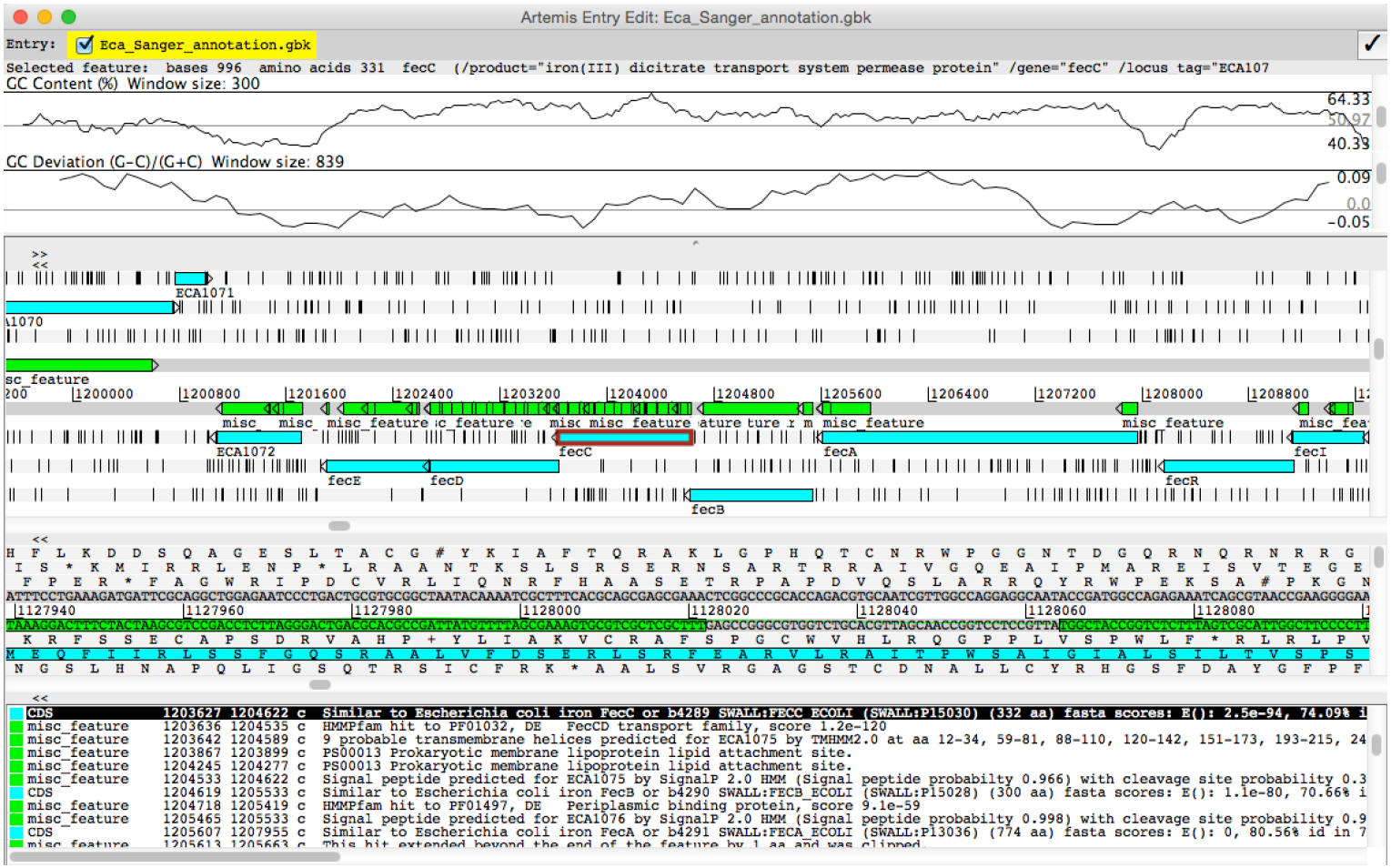
A genome sequence is a starting point...

- Manual annotation by the Sanger Pathogen Sequencing Unit
- Literature searches and comparisons
- **Six people, for six months $\approx$ three person-years**
- Genes: `BLAST`, `GLIMMER`, `ORPHEUS`
- Functional domains: `PFAM`, `SIGNALP`, `TMHMM`
- Metabolism: `KEGG`
- ncRNA: `RFAM`

2003: E. carotovora subsp. atroseptica

The James Hutton Institute

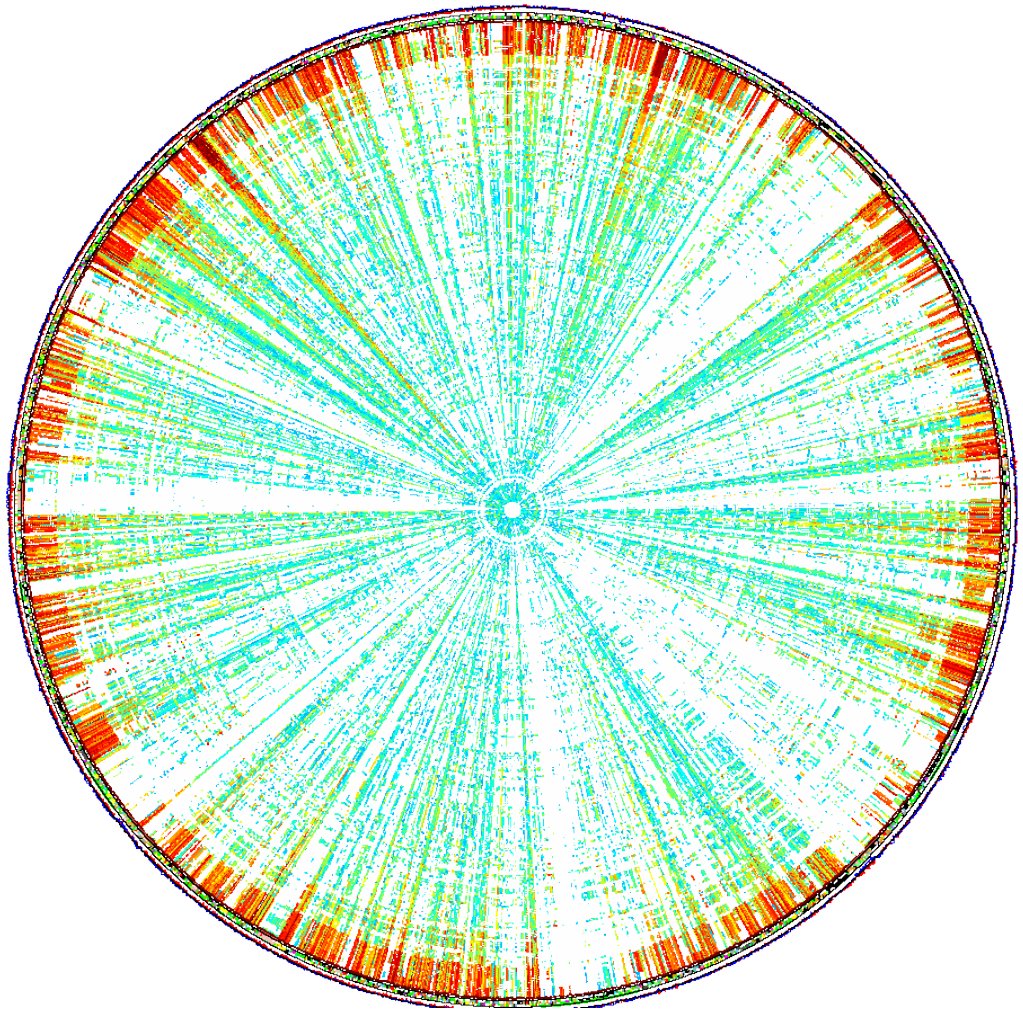
Working (`Eca_Sanger_annotation.gbk`) and published (`NC_004547.gbk`) annotation files are in the `data` directory



2003: E. carotovora subsp. atroseptica

The James Hutton Institute

Compared against all 142 available bacterial genomes[2]



[2] `data/Pba` directory in the accompanying GitHub repository

Table of Contents

2013: Dickeya spp.

Sequenced and annotated 25 new isolates of *Dickeya*

- 25 *Dickeya* isolates, at least six species
- Multiple sequencing methods: 454, Illumina (SE, PE)
- Minor publications (6, 8 authors)[3,4]


genomeA
Journals.ASM.org

Draft Genome Sequences of 17 Isolates of the Plant Pathogenic Bacterium *Dickeya*

Leighton Pritchard,[a] Sonia Humphris,[b] Gerry S. Saddler,[c] John G. Elphinstone,[d] Minna Pirhonen,[e] Ian K. Toth[b]

Information and Computational Sciences (ICS), James Hutton Institute, Invergowrie, Dundee, Scotland, United Kingdom[a]; Cellular and Molecular Sciences (CMS), James Hutton Institute, Invergowrie, Dundee, Scotland, United Kingdom[b]; Science and Advice for Scottish Agriculture (SASA), Edinburgh, United Kingdom[c]; Food and Environment Research Agency, Sand Hutton, York, United Kingdom[d]; Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland[e]


genomeA
Journals.ASM.org

Draft Genome Sequences of Four *Dickeya dianthicola* and Four *Dickeya solani* Strains

Leighton Pritchard,[a] Sonia Humphris,[b] Steve Baeyen,[c] Martine Maes,[c] Johan Van Vaerenbergh,[c] John Elphinstone,[e] Gerry Saddler,[d] Ian Toth[b]

Information and Computational Sciences (ICS), James Hutton Institute, Invergowrie, Dundee, Scotland, United Kingdom[a]; Cellular and Molecular Sciences (CMS), James Hutton Institute, Invergowrie, Dundee, Scotland, United Kingdom[b]; ILVO, Merelbeke, Belgium[c]; Food and Environment Research Agency (FERA), Sand Hutton, York, United Kingdom[d]; Science and Advice for Scottish Agriculture (SASA), Roddinglaw Road, Edinburgh, United Kingdom[e]

- **Results**: 12-237 fragments containing 4.2-5.1Mbp, at 6-84X coverage, 170k-4m reads
- **Automated annotation: RAST** with manual corrections

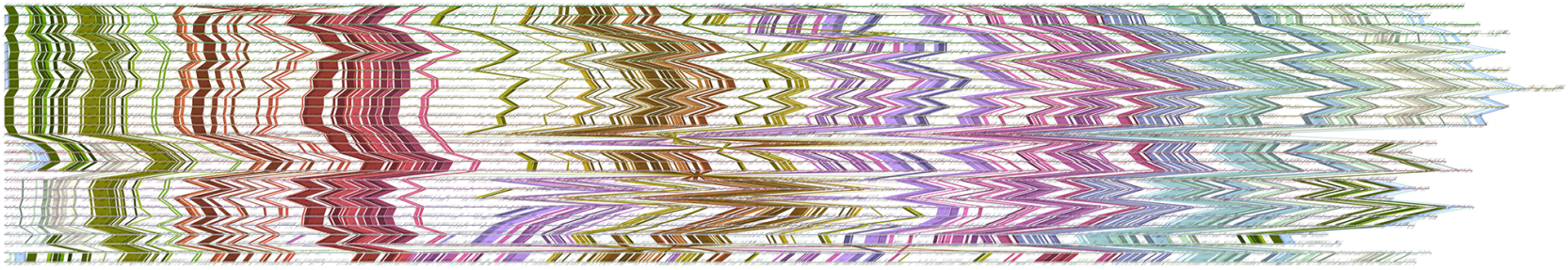
[3] Pritchard *et al.* (2013) *Genome Ann.* **1** (4) doi:10.1128/genomeA.00087-12

[4] Pritchard *et al.* (2013) *Genome Ann.* **1** (6) doi:10.1128/genomeA.00978-13

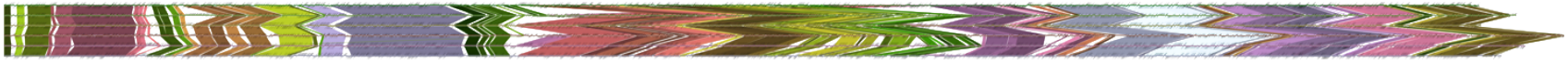
2013: Dickeya spp.

The James Hutton Institute

Within-genus comparisons: large-scale synteny and rearrangement

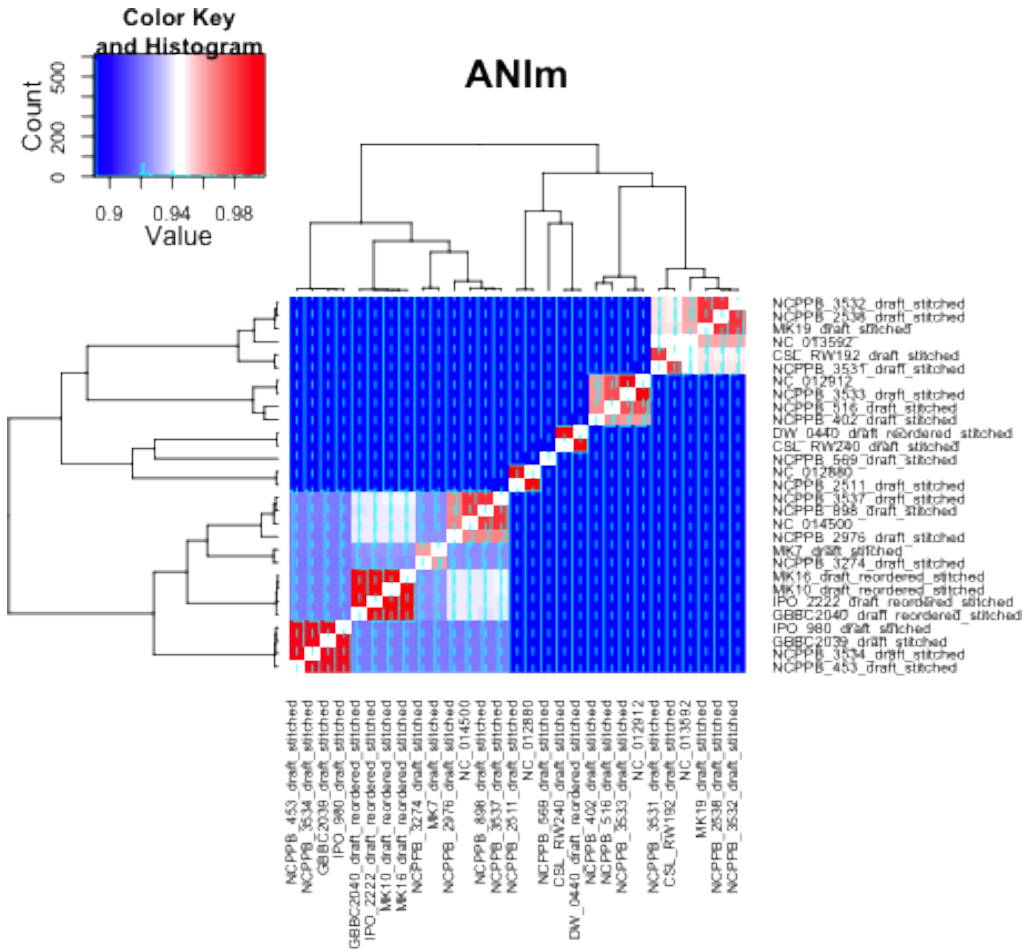


Within-species comparisons: e.g. indels, HGT



2013: Dickeya spp.

Within-genus comparisons: whole genome-based species delineation[5]

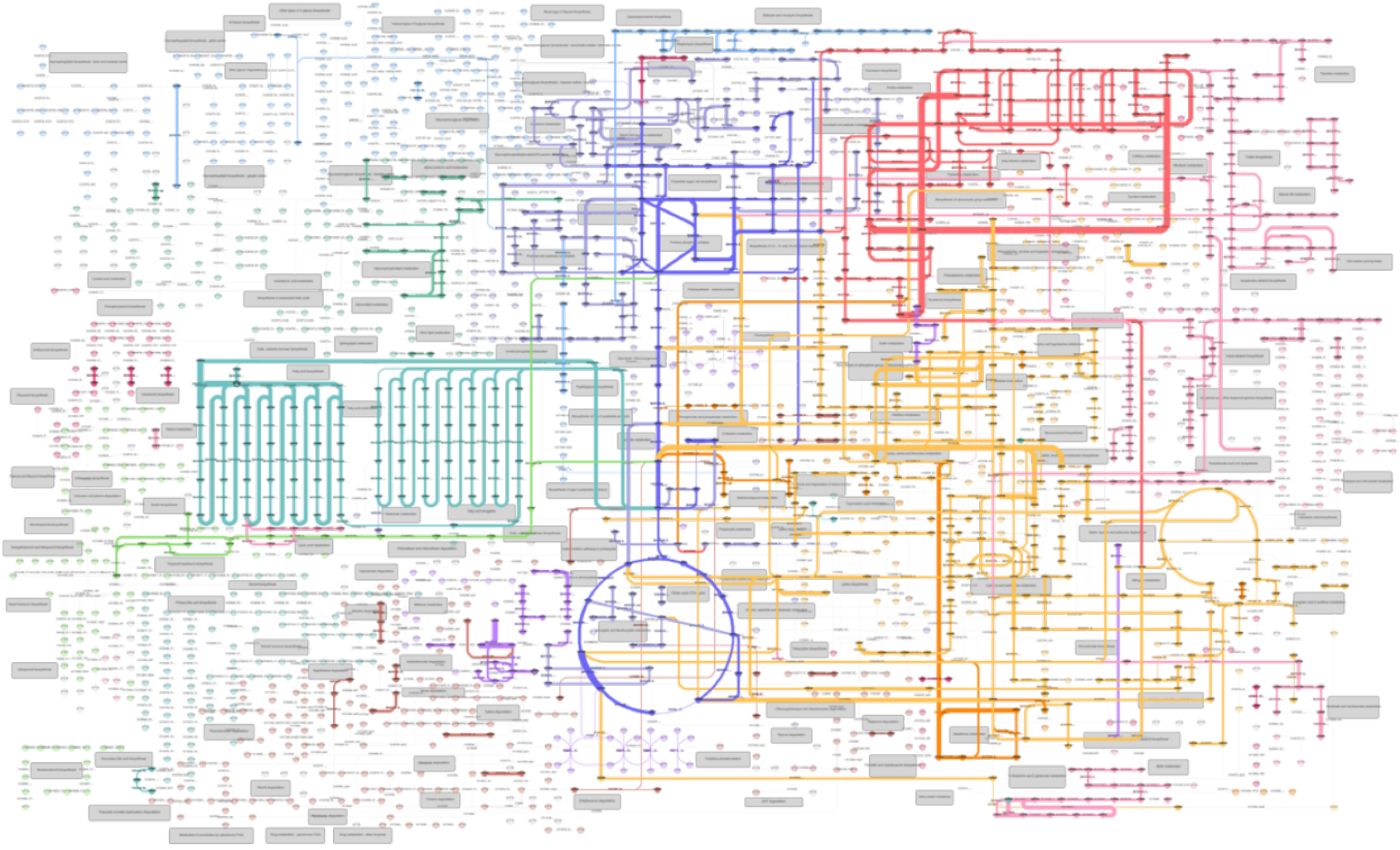


[5] van der Wolf *et al.* (2014) *Int. J. Syst. Evol. Micr.* **64**:768-774 doi:10.1099/ijs.0.052944-0

2013: Dickeya spp.

The James Hutton Institute

Within-genus comparisons: differences in metabolism



2014: E. coli

The James Hutton Institute

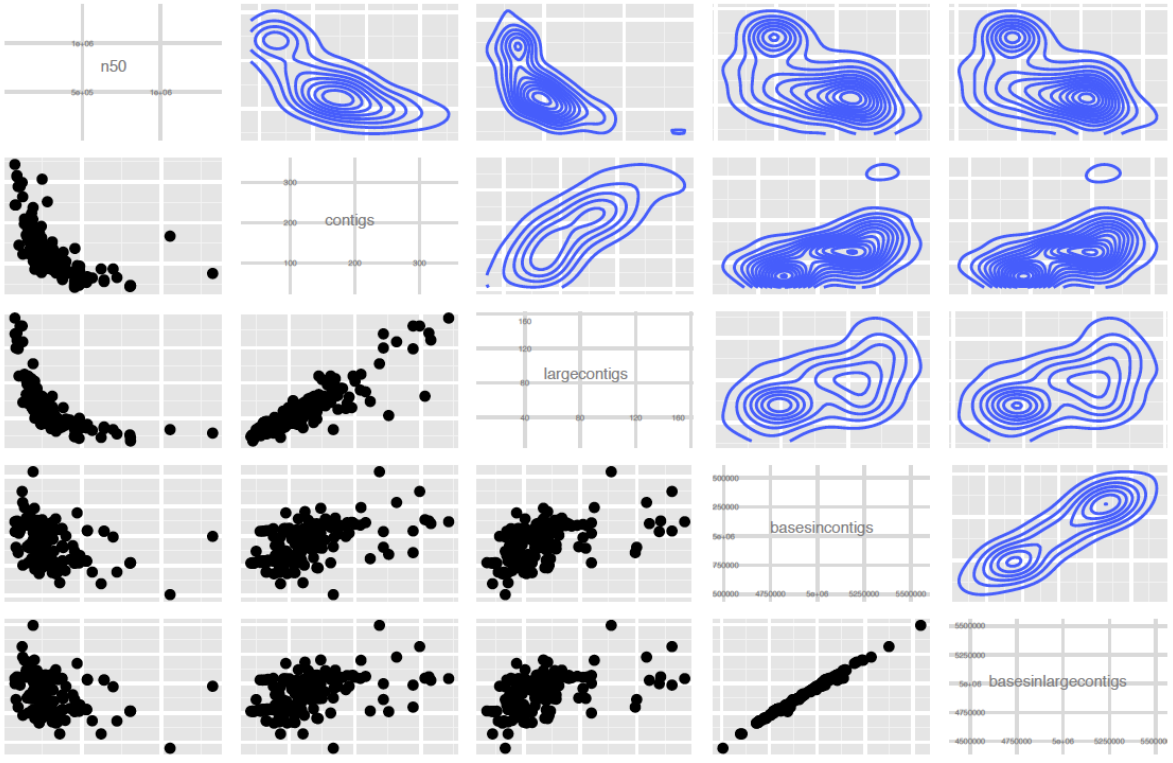
Sequenced and annotated $\approx$ 190 isolates of *E. coli*
All bacteria environmental, sampled from lysimeters

- Illumina paired-end sequencing. **Total cost of sequencing 190 bacteria: $\approx$£11k**
- **Automated annotation: PROKKA**

2014: E. coli

Sequencing output variable - even though same preps, "same" bacteria, similar sources.

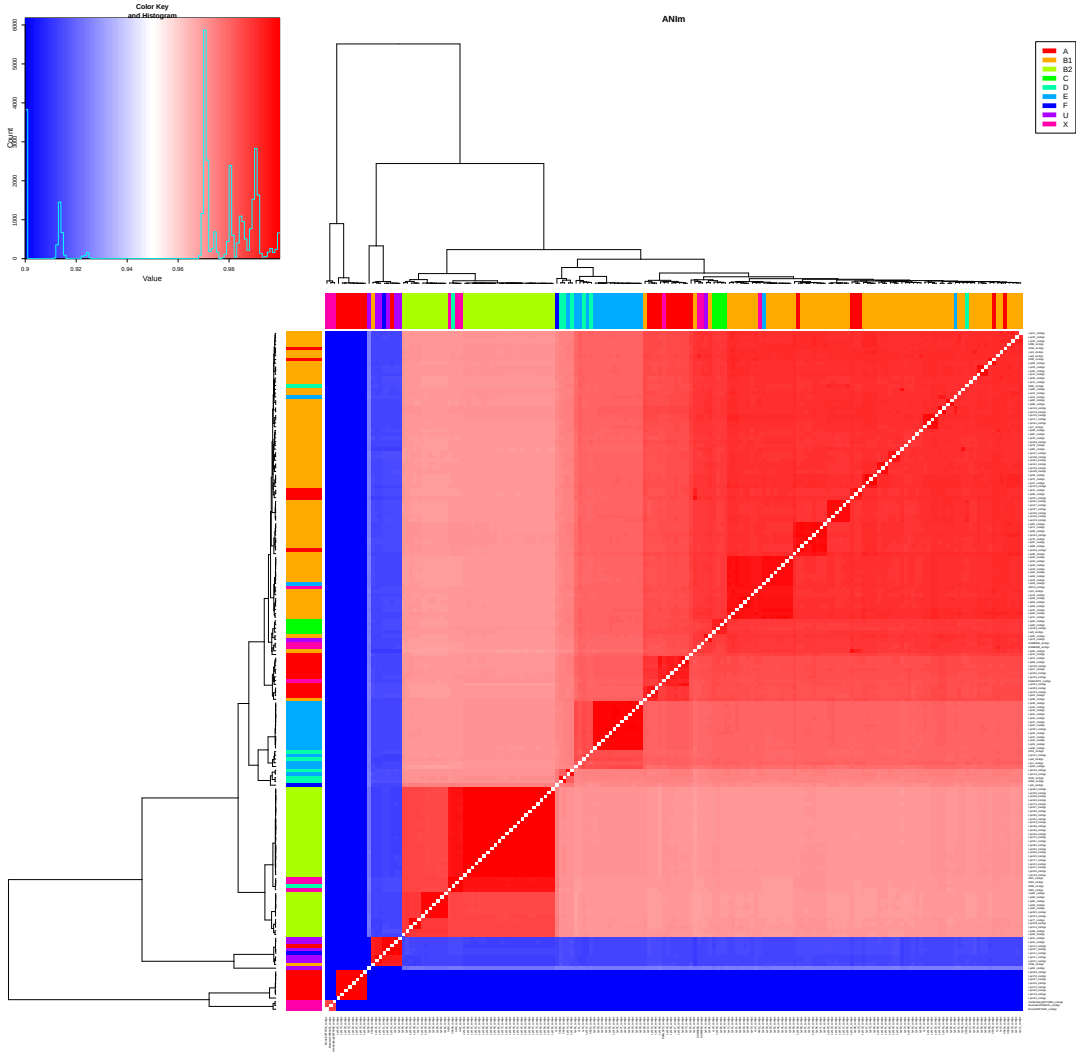
- **Results**: 5-3000 contigs (median ≈ 125); 9kbp-7.1Mbp (median ≈ 5Mbp); 170k-4m reads



2014: E. coli

The James Hutton Institute

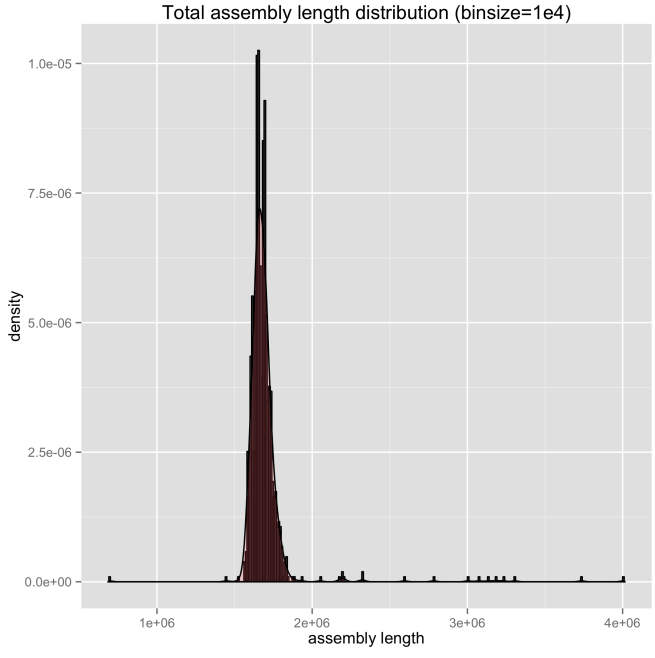
Genome sequencing enables within-species classification



2014: Campylobacter spp.

Sequenced $\approx$ 1034 isolates of *Campylobacter*
Clinical, animal, food-associated isolates

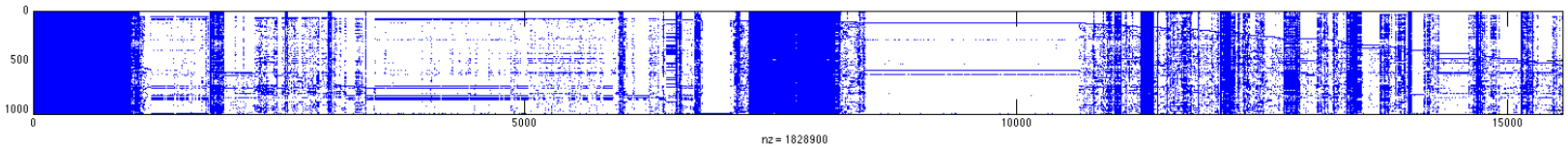
- Illumina paired-end sequencing. **Total cost of sequencing >1000 bacteria: $\approx$£60k**
- **Automated annotation: PRODIGAL**



2014: Campylobacter spp.

The James Hutton Institute

- Identified 15554 gene families from genecalls.
- To calculate, took 23 days on institute cluster (4e12 pairwise protein comparisons!).



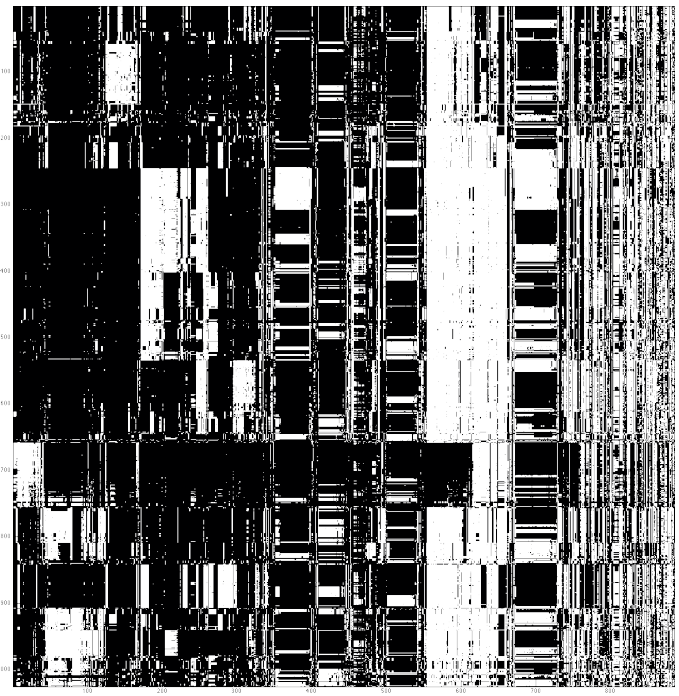


Table of Contents

So what's changed?

The James Hutton Institute

- **Cost**: £250k per genome, to £60 per genome.
  Now cheaper to sequence a genome than to analyse it!
- **Location**: sequencing centre, to benchtop
- **Data**: volume has increased massively - what you get back from machines, and what's out there to work with
  More data is better, but also more challenging.
- **Speed**: typical sequencing run time can be less than a day
- **Software**: more software to do more things (but not always better...)
- New kinds of **experiment**: genomes, exomes, variant calling, methylated sequences, ...
- New kinds of **application**: diagnostics, epidemic tracking, metagenomics, ...

So what's changed?

The James Hutton Institute

Having a single genome is useful, but having thousands really helps **comparative genomics**:
combining genomic data, evolutionary and comparative biology

- Transfer functional understanding of model systems (e.g. *E. coli*) to non-model organisms
- Genomic differences may underpin phenotypic (host range, virulence, physiological) differences
- Genome comparisons aid identification of functional elements on the genome
- Studying genomics changes reveals evolutionary processes and constraints

Table of Contents

Revolutions One and Two[a]

[a] Loman and Pallen (2015) *Nat. Rev. Micro.* doi:10.1038/nrmicro3565

The James Hutton Institute

Revolution One: whole-genome shotgun

- First bacterial genomes: *Haemophilus influenzae* (1995); *E. coli*, *Bacillus subtilis* (1997)
- (Oh, and the human genome)

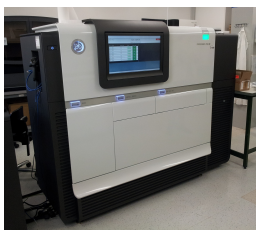
Revolution Two: high-throughput sequencing

- "Next-generation" sequencing (now "last-generation").
- 454 GS20 (2005), Illumina GAII (2007).
- metagenomics; surveillance sequencing; SNP-based comparisons; transposon-sequencing for functional genomics; ChIP-seq; ...

Not all HT sequencing is the same

The James Hutton Institute

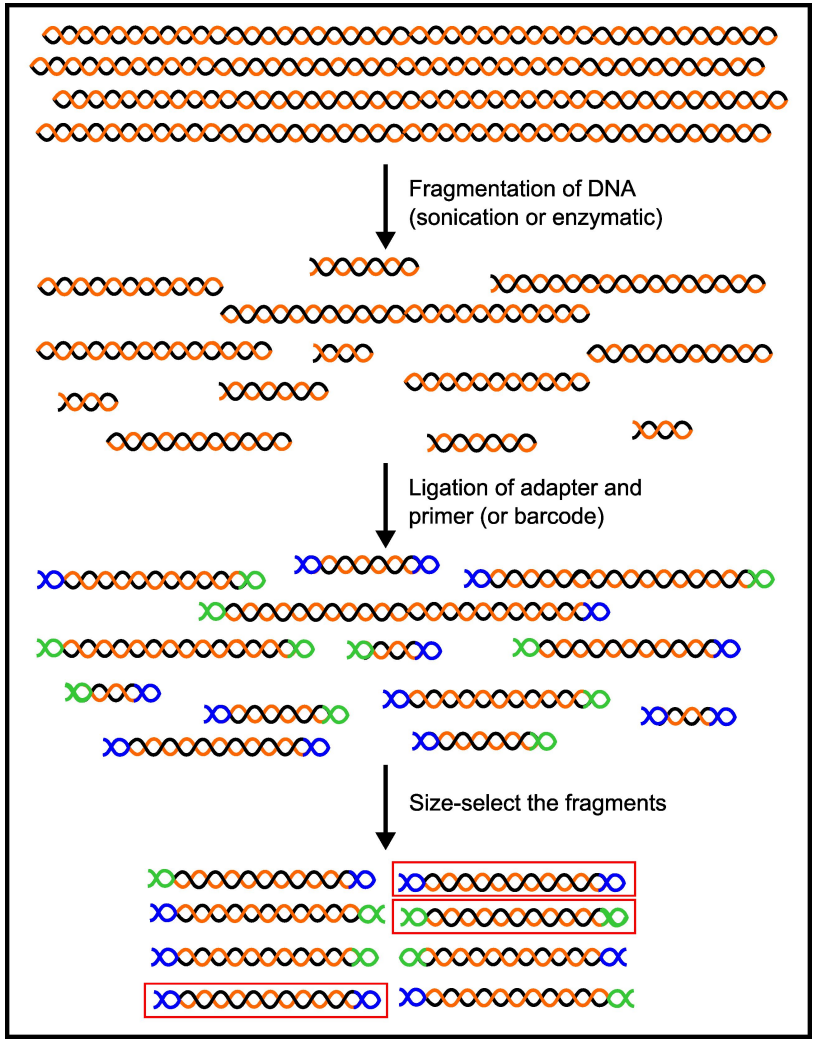
It's all about the biology, but it all starts with the data. Sequencing technology (including library prep.) affects your sequence data.

- Roche/454 
- Illumina 
- Ion Torrent 
- Pacific Bioscience (PacBio) 

The basic principle

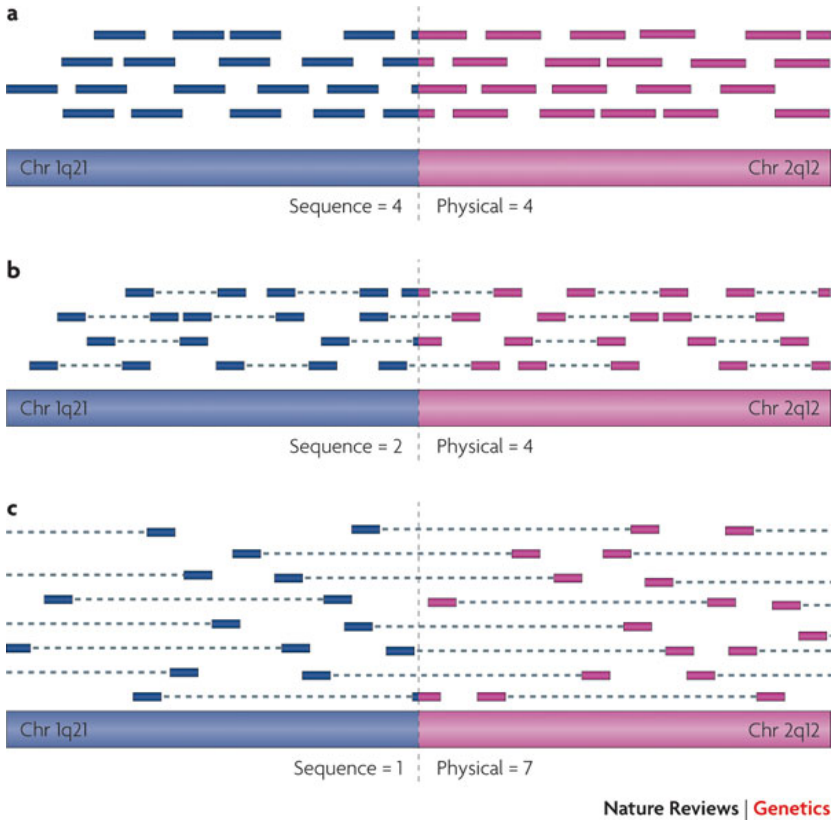
The James Hutton Institute

DNA source is fragmented, and the fragments are sequenced.



HTS: PE vs SE

High-throughput sequencing (e.g. Illumina), reads may be single-end, or paired-end.



Putting the jigsaw back together is sequence assembly.

Four different chemistries[a]

[a] Loman *et al.* (2012) *Nat. Rev. Micro.* **31**:294-296 doi:10.1038/nbt.2522

The James Hutton Institute

Reads differ by technology, and may require different bioinformatic treatment...

- **Roche/454**: Pyrosequencing (long reads, but expensive, and high homopolymer errors) (700-800bp, 0.7Gbp, 23h)
- **Illumina**: Reversible terminator (cost-effective, massive throughput, but short read lengths) (2x150bp, 1.5Gbp, 27h)
- **Ion Torrent**: Proton detection (short run times, good throughput, high homopolymers errors) (200bp, 1Gbp, 3h)
- **PacBio**: Real-time sequencing (very long reads, high error rate, expensive) (3-15kbp, 3Gbp/day, 20min)

... different error profiles, varying capability to assemble/determine variation

Costs of sequencing[a]

[a] Miyamoto *et al.* (2014) *BMC Genomics* **15**:699 doi:10.1186/1471-2164-15-699

The James Hutton Institute

Cost and required DNA comparison

	GS Jr	Ion PGM	MiSeq	PacBio
Instrument cost	$108K	$50K	$99K	$900K
Sequence yield per run	35Mb	2Gb (400bp)	8 Gb	1 Gb/8 SMRT cells
Running cost	$1000/1run(35Mb)	$437/Gb	$93/Gb	$1800/Gb
Sequence Run time	10 hr	7.3 hr	39 hr	16 hr/8 SMRT cells
Other time consuming steps	Library prep: 3hr emPCR: 6hr	Library prep: 3.5 hr emPCR : 8 hr	Library prep: 7 hr	Library prep: 5 hr
DNA requirements	500 ng with 1.8 OD	250 ng	250 ng	100ng (250bp library) - 5µg (20kb library)

Table of Contents

Benchmarked performance

The James Hutton Institute

Apply several sequencing technologies to the same sample(s).
Benchmark comparisons inform appropriate choice of sequencing technology[6,7,8,9,10,11,12]

Progress in technologies is driving research very rapidly.
Always look for most recent/relevant benchmarks.

Bioinformatic methods also need to be benchmarked.

---

[6] Miyamoto *et al.* (2014) *BMC Genomics* **15**:699 doi:10.1186/1471-2164-15-699

[7] Salipante *et al.* (2014) *Appl. Environ. Micro.* **80**:7583-7591 doi:10.1128/AEM.02206-14

[8] Frey *et al.* (2014) *BMC Genomics* **15**:96 doi:10.1186/1471-2164-15-96

[9] Koshimizu *et al.* (2013) *PLoS One* **8**:e74167 doi:10.1371/journal.pone.0074167

[10] Quail *et al.* (2012) *BMC Genomics* **13**:341 doi:10.1186/1471-2164-13-341

[11] Loman *et al.* (2012) *Nat. Biotech.* **30**:434-439 doi:10.1038/nbt.2198

[12] Lam *et al.* (2011) *Nat. Biotech.* **1** (6) doi:10.1038/nbt.2065

Benchmarking on Vibrio[a]

[a] Miyamoto *et al.* (2014) *BMC Genomics* **15**:699 doi:10.1186/1471-2164-15-699

- Sequenced *Vibrio parahaemolyticus* (2x chromosomes, closed reference genome) with four technologies
- Chose an assembler for each tech, and assembled reads
- Excess reads with Ion/MiSeq: used random subsets of reads to determine required coverage
- Aligned assemblies (MUMmer) to known high-quality chromosome sequence, to measure error

Benchmarking on Vibrio[a]

[a] Miyamoto *et al.* (2014) *BMC Genomics* **15**:699 doi:10.1186/1471-2164-15-699

The James Hutton Institute

Table 1 Data statistics for sequence run and assemblies

Sequencer	**GS Jr**	**Ion PGM**	**MiSeq**	**PacBio**
Number of reads	115611	4982888	39656630	120230*
Total bp	48285593	1443005019	9953814130	374942687
Coverage	9	279	1927	73
Mean length	418	290	251	3119
Assembler	**Newbler**	**Newbler**	**CLC**	**Sprai**
Number of bp used for assembly	48285593	400000107	299809460	374942687
Number of reads used	115611	1380757	1194460	120230*
Coverage	9	77	58	73
Number of contigs	309	61	34	31
Total bases	5053921	5075085	5103771	5298335
Max length	164926	895358	732626	3288561
N50 contig length	30451	392606	431440	3288561

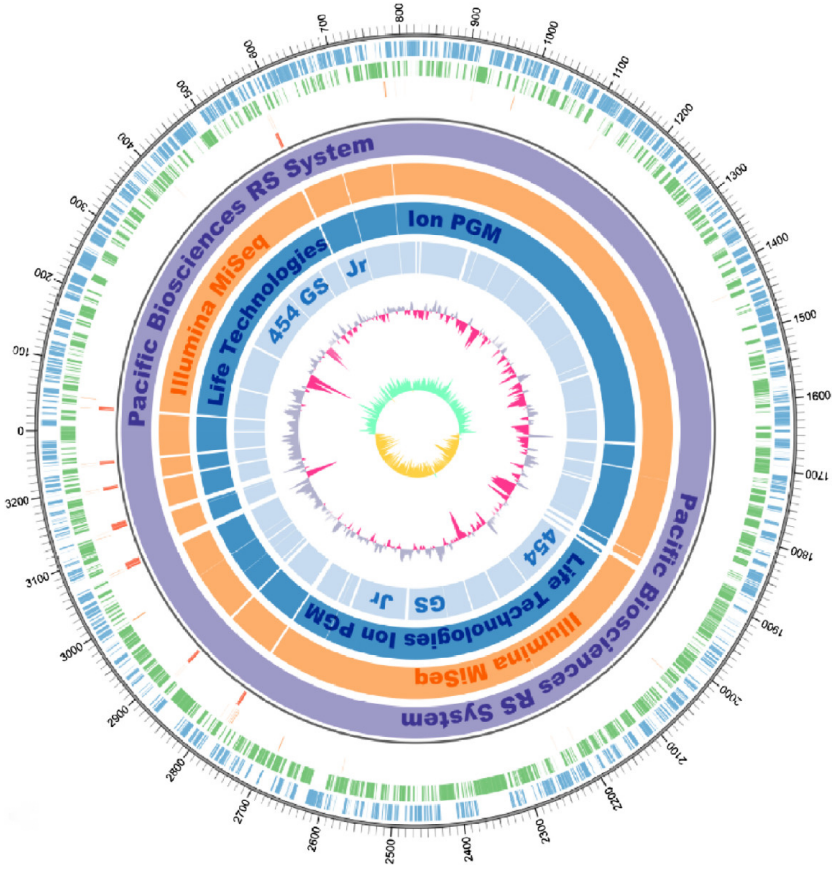
GS Jr, Ion PGM, and MiSeq data are based on a single run. PacBio data are from three cells. The upper part of the table shows read statistics and the lower part shows the statistics of the best assembly. *Number of reads of PacBio is the number of subreads longer than 500 bp.

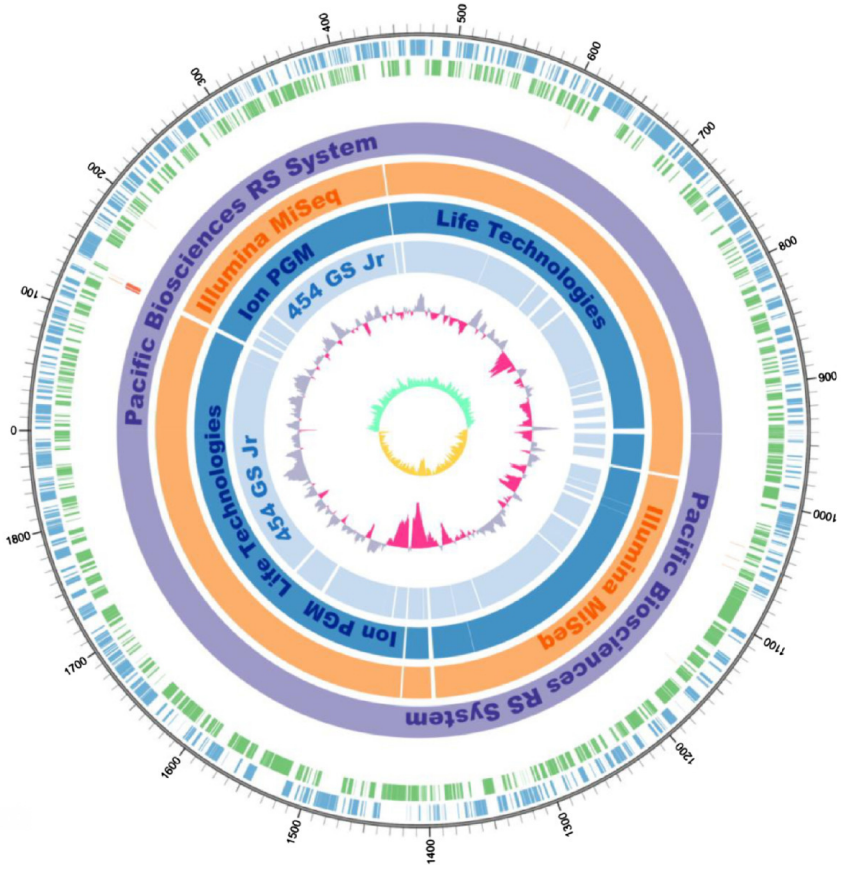
Benchmarking on Vibrio[a]

[a] Miyamoto *et al.* (2014) *BMC Genomics* **15**:699 doi:10.1186/1471-2164-15-699

The James Hutton Institute

De novo assembly and alignment against *Vibrio parahaemolyticus* (2x chromosomes)





Benchmarking on Vibrio[a]

[a] Miyamoto *et al.* (2014) *BMC Genomics* **15**:699 doi:10.1186/1471-2164-15-699

The James Hutton Institute

- More and longer reads do not always give the best assemblies: read depth, read distribution, error rate also matters
- Optimal assemblies were obtained at around 60x-80x coverage, for Illumina and Ion.
- Multiple rRNA regions are fragmented in short-read assemblies
- PacBio generated single chromosome contigs
- Assembly of multiple-chromosome bacteria is currently feasible

Variability in published genomes as methods are not standard (e.g. sequencing technology, assembler, parameter settings and pre-processing)...

Table of Contents

Revolution Three[a]

[a] Loman and Pallen (2015) *Nat. Rev. Micro.* doi:10.1038/nrmicro3565

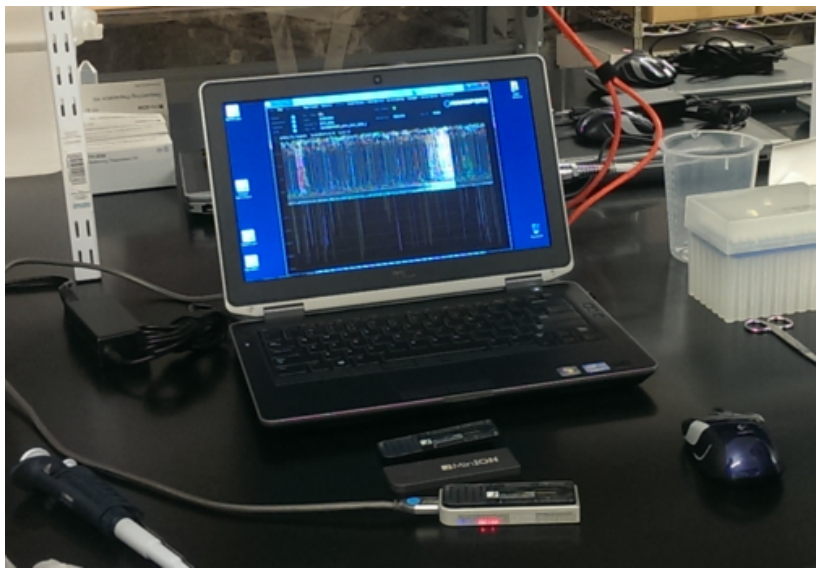
The James Hutton Institute

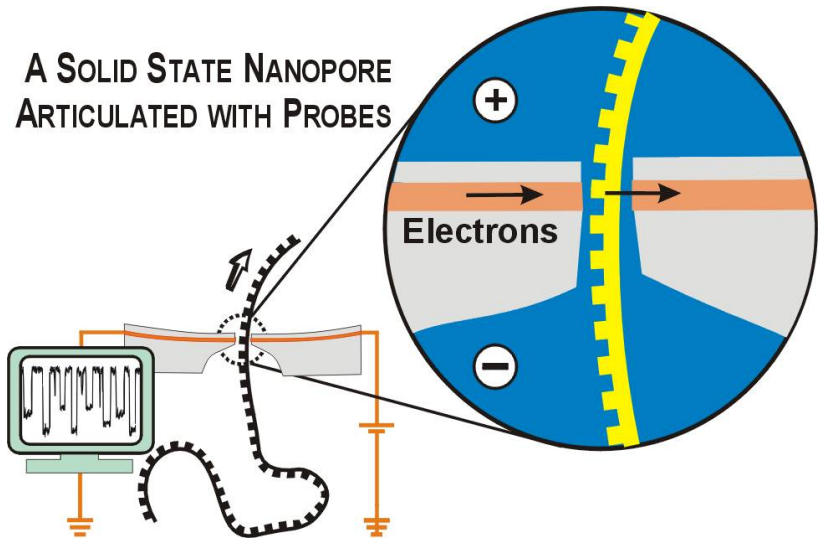
Revolution Three: single-molecule long-read sequencing

- Living through the revolution, now
- PacBio (SMRT): large machine, expensive
- Nanopore: portable device, inexpensive
- Less mature, less accurate, improving rapidly

The future dominant sequencer?

Oxford Nanopore. A sequencer the size of your hand.





- Microfluidics, single-molecule sequencing; 11-70kbp reads
- Reports current across pore (tiny electron microscope) as molecule moves through
- \$10/Mbp, 110Mbp per flowcell[13]

[13] Yaniv Erlich (2013) Future Continuous blog

Early data[a]

[a] Quick *et al.* (2014) *GigaScience* **3**:22 doi:10.1111/1755-0998.12324

The James Hutton Institute

It's a fast-moving area, and results are improving.

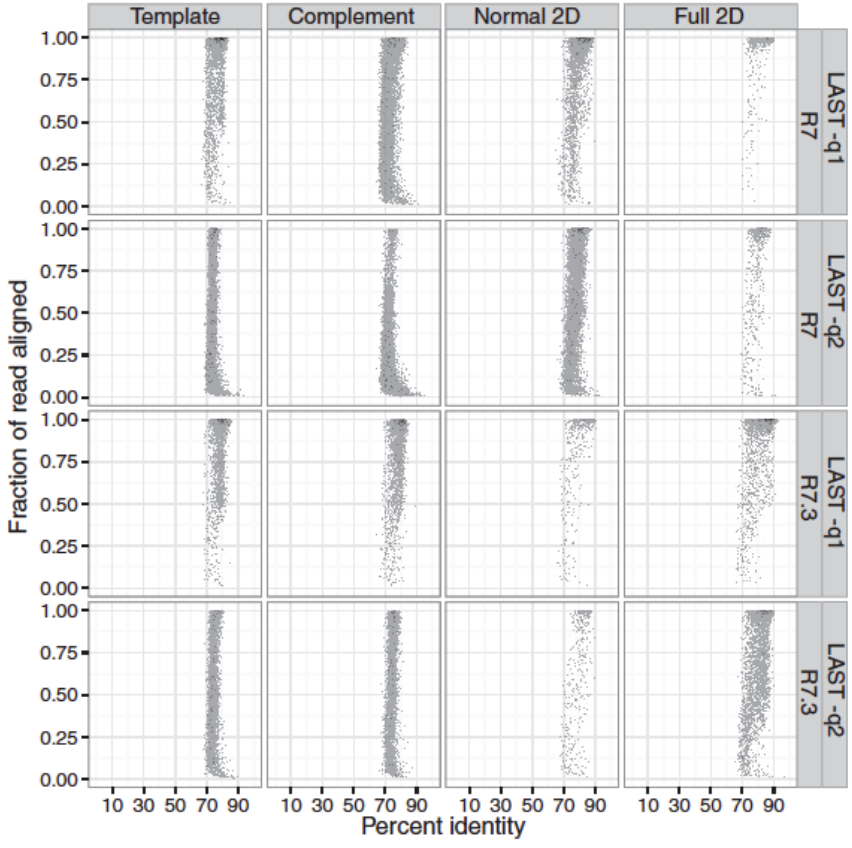


Figure 4 Alignment identity and completeness. Each plot reflects the alignment identity and the proportion of the read aligned for all 2D reads, as well as the underlying template and complement sequences. The top two panels reflect the alignment results for normal and full 2D reads from the R7 flowcell, and the bottom two panels reflect the R7.3 flowcell. Left panels employ a mismatch penalty of 1 and right panels reflect a mismatch penalty of 2. Overall, the lower mismatch penalty increases the identity and fraction of the read that aligned and this effect is greatest for full 2D reads.

Developing tools

Oxford Nanopore's open beta went out without analysis tools. Tools (Poretools, poRe, etc.) were written/tested/validated by the user community[14],[15]

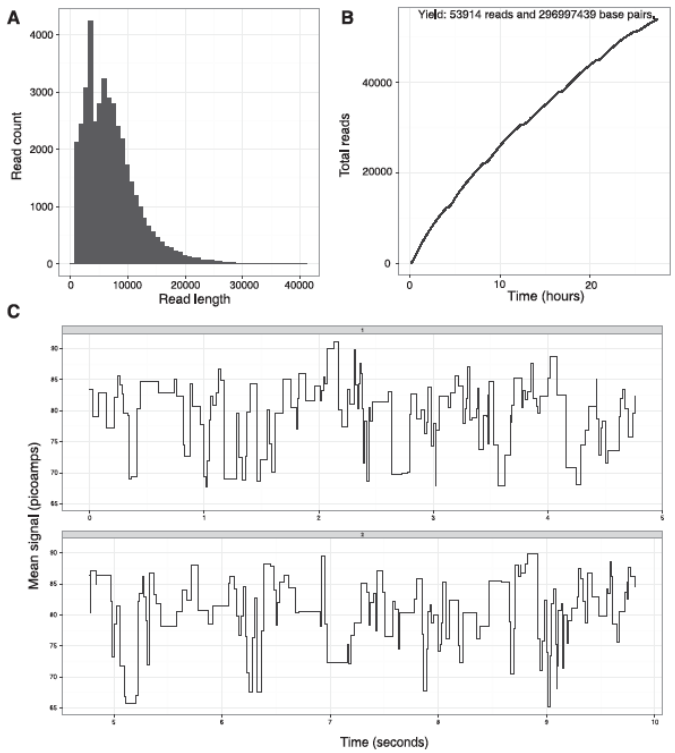


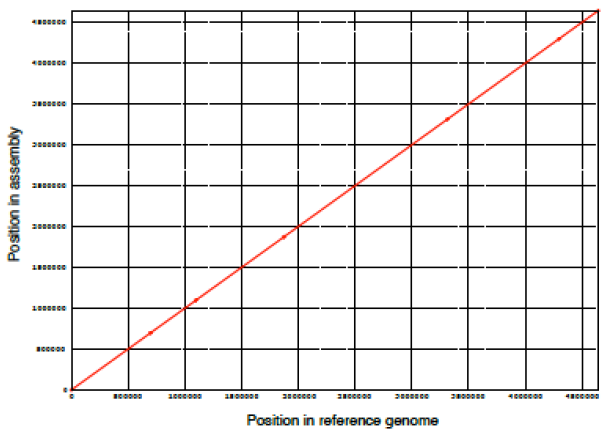
Fig. 1. Example `poretools` visualizations from a set of FAST5 files generated by a single MinION™ run. Panel **A** shows a histogram of read lengths. Panel **B** shows a collector's curve of reads over time. Panel **C** shows an example squiggle plot of detected event transitions originating from MinION™

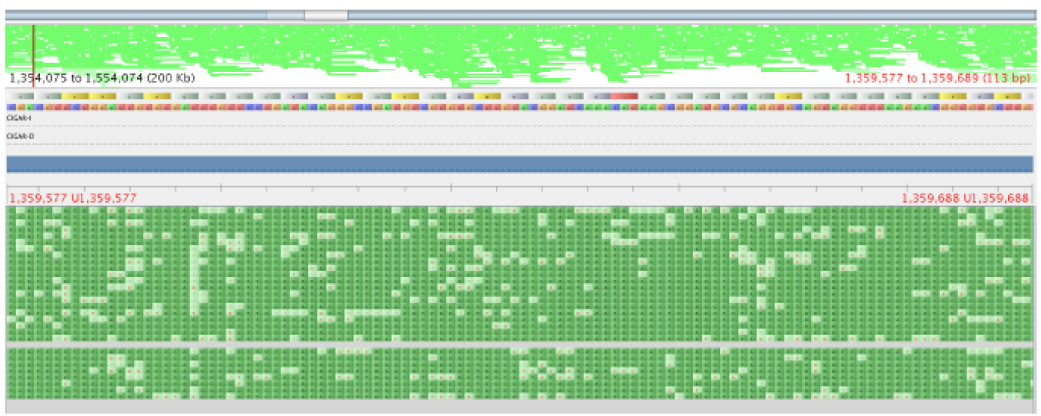
[14] Loman and Quinlan (2014) *Bioinformatics* doi:10.1093/bioinformatics/btu555

[15] Watson *et al.* (2014) *Bioinformatics* doi:10.1093/bioinformatics/btu590

Recent applications

- Amplicon sequencing (16S metagenomics) of bacteria and viruses [16]
- Real-time viral diagnostics [17]
- Scaffolding of a bacterial genome [18]
- Complete *de novo* assembly of a bacterial genome [19]





[16] Kilianski *et al.* (2015) *GigaScience* doi:10.1186/s13742-015-0051-z

[17] Greninger *et al.* (2015) *Genome Med.* doi:10.1186/s13073-015-0220-9

[18] Karlsson *et al.* (2015) *Sci. Reports* doi:10.1038/srep11996

[19] Loman *et al.* (2015) *Nat. Meth.* doi:10.1038/nmeth.3444

The three revolutions[a]

[a] Loman and Pallen (2015) *Nat. Rev. Micro.* doi:10.1038/nrmicro3565

The James Hutton Institute

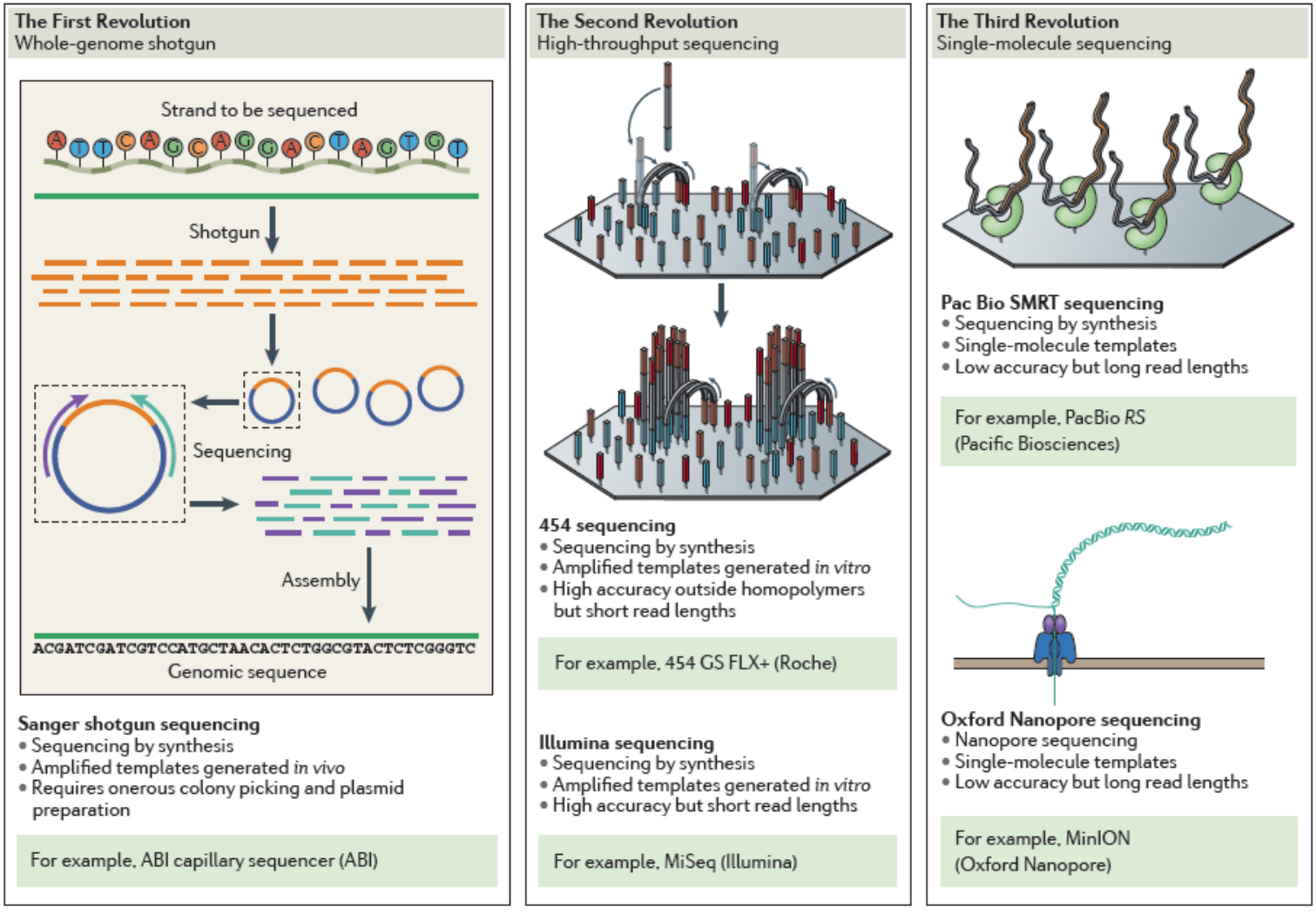
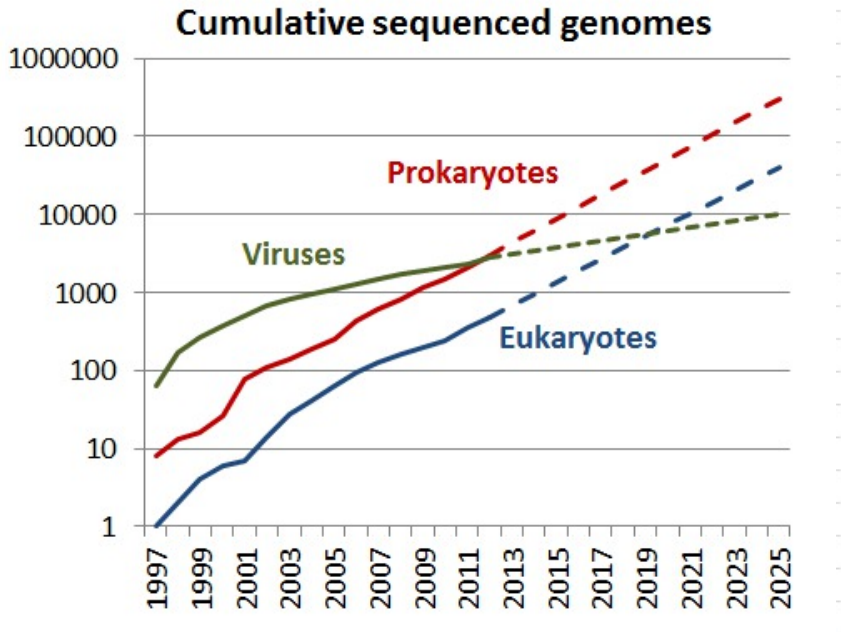
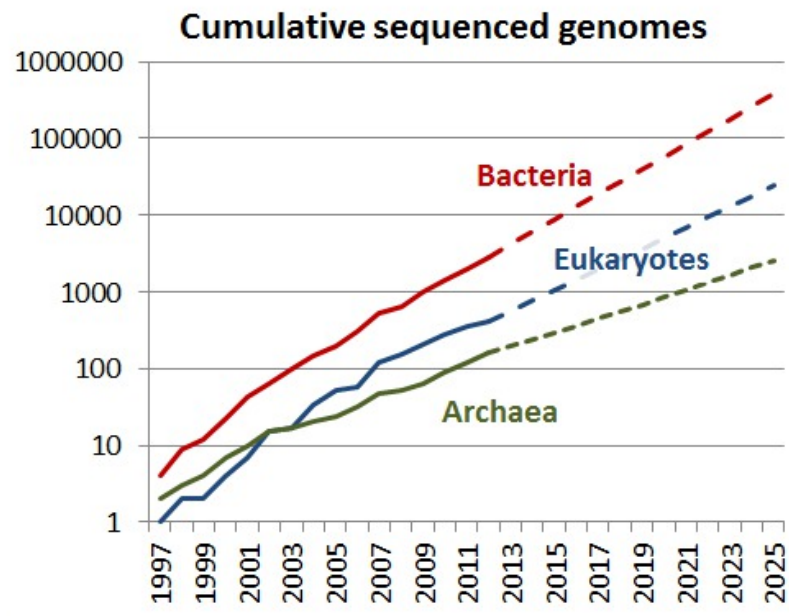


Table of Contents

Predicting the future is hard...

"How many genomes will we have, and when?"
Su *et al.* attempted to answer this[20]:





[20] http://sulab.org/2013/06/sequenced-genomes-per-year/

After that, the flood...

The James Hutton Institute

High-throughput sequencing methods have completely changed the landscape of microbiology
(Nearly) complete, (mainly) accurate sequence data is now inexpensive (and cheaper than analysis)

- GOLD (19/2/2014): 3,011 "finished" ; 9,891 "permanent draft" genomes
- GOLD (10/11/2015): 7,657 "finished" ; 27,438 "permanent draft" genomes; 50,673 prokaryotes
- NCBI WGS (19/2/2014): 17,023 microbial genomes
- NCBI Genome (10/11/2015): 55,033 prokaryotic genomes

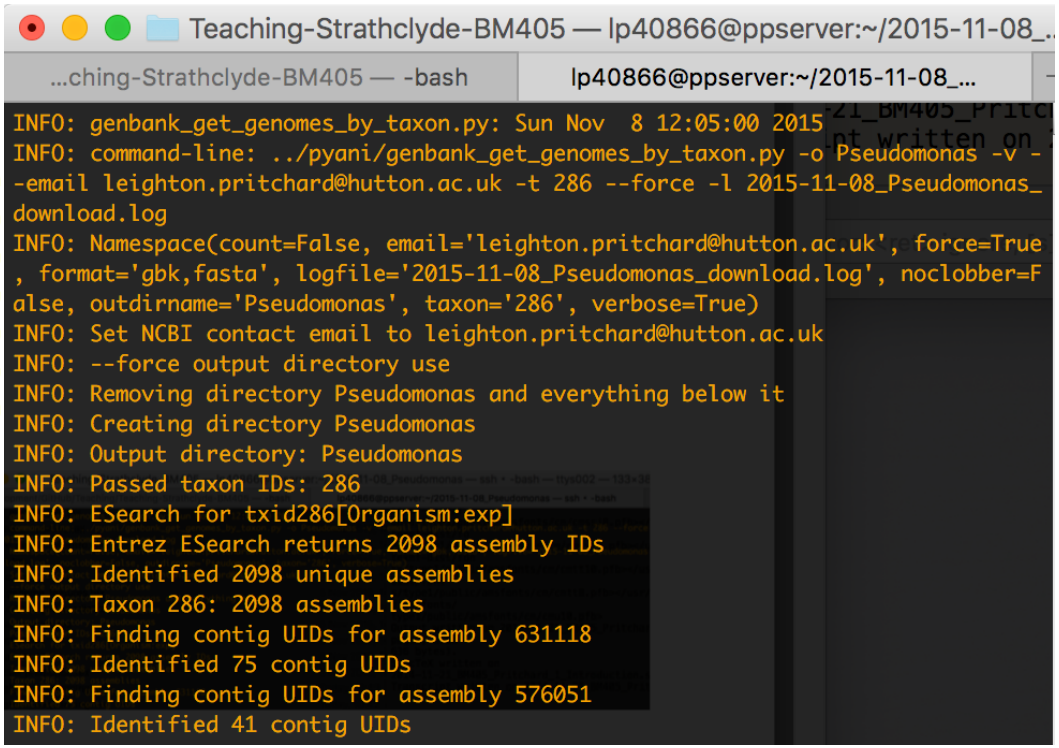
Pseudomonas

The James Hutton Institute

In 2011, 25 isolate sequences[21]; in 2015, 2098 genomes:

SUMMARY

One reason for the success of *Pseudomonas syringae* as a model pathogen has been the availability of three complete genome sequences since 2005. Now, at the beginning of 2011, more than 25 strains of *P. syringae* have been sequenced and many more will soon be released. To date, published analyses of *P. syringae* have been largely descriptive, focusing on catalogues of genetic differences among strains and between species. Numerous powerful statistical tools are now available that have yet to be applied to *P. syringae* genomic data for robust and quantitative reconstruction of evolutionary events. The aim of this review is to provide a snapshot of the current status of *P. syringae* genome sequence data resources, including very recent and unpublished studies, and thereby demonstrate the richness of resources available for this species. Furthermore, certain specific opportunities and challenges in making the best use of these data resources are highlighted.



We're going to need bigger bioinformatics...

[21] Studholme (2011) *Mol. Plant Pathol.* doi:10.1111/j.1364-3703.2011.00713.x

Microbial Genomics and Bioinformatics
BM405
2.Assembly

The James Hutton Institute

Leighton Pritchard[1,2,3]
[1]Information and Computational Sciences,
[2]Centre for Human and Animal Pathogens in the Environment,
[3]Dundee Effector Consortium,
The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

Acceptable Use Policy

The James Hutton Institute

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available on SlideShare.

These slides, and supporting material including exercises, are available at https://github.com/widdowquinn/Teaching-Strathclyde-BM405

What do you get from sequencing

The James Hutton Institute

Sequence reads. Usually lots of them.
Size/number/errors depend on technology used.

Table 1 Data statistics for sequence run and assemblies

Sequencer	**GS Jr**	**Ion PGM**	**MiSeq**	**PacBio**
Number of reads	115611	4982888	39656630	120230*
Total bp	48285593	1443005019	9953814130	374942687
Coverage	9	279	1927	73
Mean length	418	290	251	3119
Assembler	**Newbler**	**Newbler**	**CLC**	**Sprai**
Number of bp used for assembly	48285593	400000107	299809460	374942687
Number of reads used	115611	1380757	1194460	120230*
Coverage	9	77	58	73
Number of contigs	309	61	34	31
Total bases	5053921	5075085	5103771	5298335
Max length	164926	895358	732626	3288561
N50 contig length	30451	392606	431440	3288561

GS Jr, Ion PGM, and MiSeq data are based on a single run. PacBio data are from three cells. The upper part of the table shows read statistics and the lower part shows the statistics of the best assembly. *Number of reads of PacBio is the number of subreads longer than 500 bp.

[22] Miyamoto *et al.* (2014) *BMC Genomics* **15**:699 doi:10.1186/1471-2164-15-699

Sequence Read Data Formats

The James Hutton Institute

Two common read data sequence formats:

- **FASTQ**: Related to FASTA, a *de facto* standard for sequence reads
- **SAM/BAM**: Sequence alignment/mapping format, two flavours - uncompressed and compressed

New formats are required to handle very large numbers of genomes

- **CRAM**: Reference-based sequence compression

You might also receive assembled genomes directly from a sequencing partner

Table of Contents

FASTQ[a]

[a] Cock *et al.* (2009) *Bioinformatics* **38**:1767-1771 doi:10.1093/nar/gkp1137

The James Hutton Institute

```
@HISEQ2500-09:168:HA424ADXX:2:1101:1404:2061 1:N:0:ATCTCTCTCACCAACT
CGGTCTTGGGATAGATGGGTTGCAGGTTGCGGTAAAGCTCGGACTCCAGAGCGTCCAGGGTAGACTGGCTAATCTTCTGCTCTTTATCGATCATTATTTC
+
@@CBDDFFHHDFDHEGHIICGIFHHIIIIFHGGHIEHHIIIIGHGHIIIIIGGHHFFFFC@CBCCCDDBDCDDDDDDDDCCDDDD3@ABDDDDDEEEDE@
```

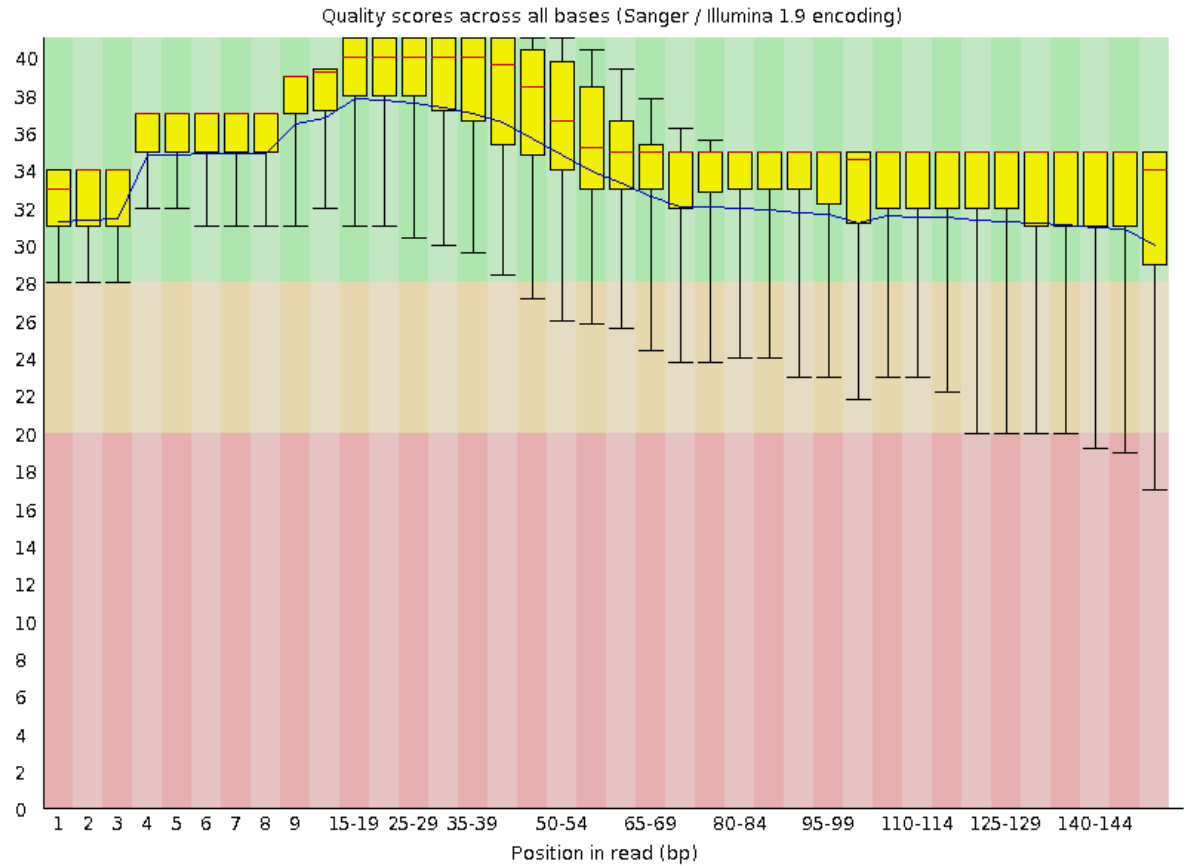
Files typically have `.fq`, `.fastq` extension.
Four lines per sequence

1. Header: sequence identifier and optional description, starts with "@"
2. Raw sequence (`[ACGTN]`)
3. *Optional* header, repeats line 1, starts with "+"
4. Quality scores, numbers encoded as ASCII
   $Q_{phred} = -10 \log_{10} e$, where e is the estimated probability that a base call is incorrect (like a pH).

Quality Control

The James Hutton Institute

The quality of basecalls (error rate) varies between and along reads.

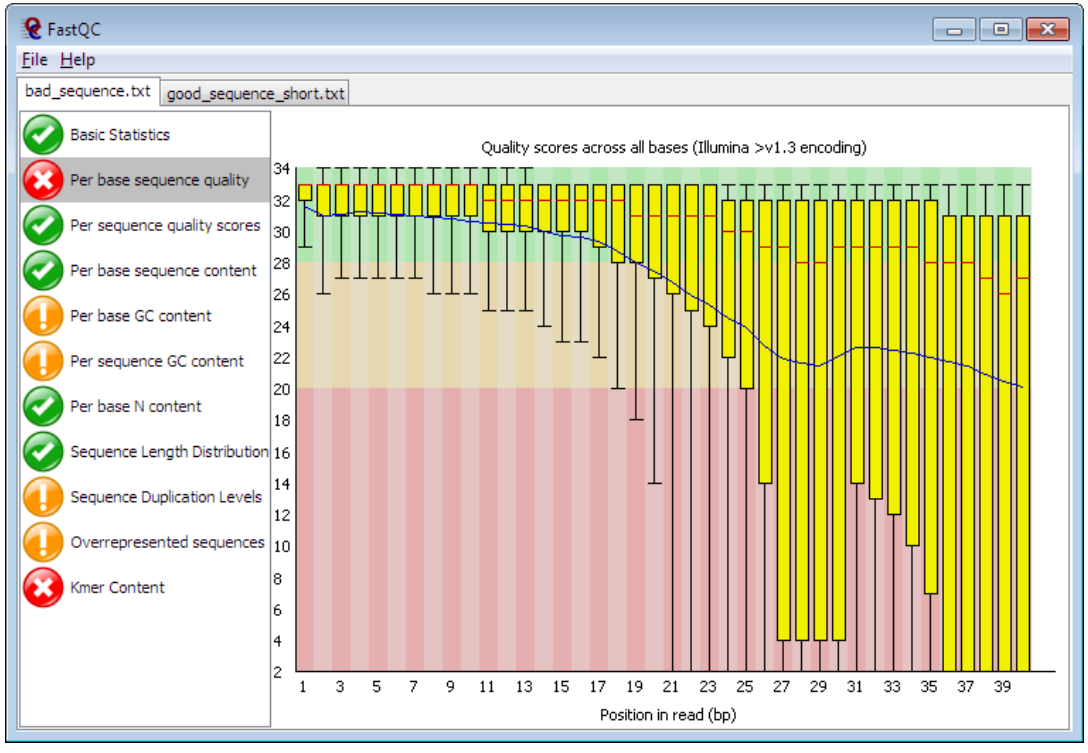


(real data from our *E.coli* sequencing: good quality)

Quality Control

The James Hutton Institute

Some datasets are better than others.



Reads can be trimmed, or discarded.

Including poor reads compromises assembly.

FASTQ encoding[a]

[a] Cock *et al.* (2009) *Bioinformatics* **38**:1767-1771 doi:10.1093/nar/gkp1137

The James Hutton Institute

More than one version of FASTQ, differ by quality encoding
Numbers converted to ASCII start at different values

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	¡	225	·
2	<STX>	34	"	66	B	98	b	130	Ç	162	¢	194	¬	226	‚
3	<ETX>	35	#	67	C	99	c	131	É	163	£	195	√	227	„
4	<EOT>	36	$	68	D	100	d	132	Ñ	164	§	196	ƒ	228	‰
5	<ENQ>	37	%	69	E	101	e	133	Ö	165	•	197	≈	229	Â
6	<ACK>	38	&	70	F	102	f	134	Ü	166	¶	198	Δ	230	Ê
7	<BEL>	39	'	71	G	103	g	135	á	167	ß	199	«	231	Á
8	<BS>	40	(	72	H	104	h	136	à	168	®	200	»	232	Ë
9	<TAB>	41	)	73	I	105	i	137	â	169	©	201	…	233	È
10	<LF>	42	*	74	J	106	j	138	ä	170	™	202		234	Í
11	<VT>	43	+	75	K	107	k	139	ã	171	´	203	À	235	Î
12	<FF>	44	,	76	L	108	l	140	å	172	¨	204	Ã	236	Ï
13	<CR>	45	-	77	M	109	m	141	ç	173	≠	205	Õ	237	Ì
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	Œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	œ	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	–	240	
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	Ò
18	<DC2>	50	2	82	R	114	r	146	í	178	≤	210	“	242	Ú
19	<DC3>	51	3	83	S	115	s	147	ì	179	≥	211	”	243	Û
20	<DC4>	52	4	84	T	116	t	148	î	180	¥	212	‘	244	Ù
21	<NAK>	53	5	85	U	117	u	149	ï	181	µ	213	’	245	ı
22	<SYN	54	6	86	V	118	v	150	ñ	182	∂	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	∑	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	∏	216	ÿ	248	¯
25	<EM>	57	9	89	Y	121	y	153	ô	185	π	217	Ÿ	249	˘
26	<SUB>	58	:	90	Z	122	z	154	ö	186	∫	218	/	250	˙
27	<ESC>	59	;	91	[	123	{	155	õ	187	ª	219	€	251	°
28	<FS>	60	<	92	\	124	|	156	ú	188	º	220	‹	252	¸
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	›	253	˝
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	fi	254	˛
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	fl	255	ˇ

FASTQ encoding[a]

[a] Cock *et al.* (2009) *Bioinformatics* **38**:1767-1771 doi:10.1093/nar/gkp1137

The James Hutton Institute

Versions vary by sequencer and period.
Most now settled on Sanger format (occasionally see historical data).
Quality scores (Q_{phred}) offset to lie in the given range:

1. **Sanger**: 33-126, used in SAM/BAM, and Illumina 1.8+
2. **Illumina 1.0-1.2**: 59-126
3. **Illumina 1.3-1.8**: 64-126

Knowing where your data comes from, and the data format and version, is always important.

Table of Contents

SAM[a]

[a] https://github.com/samtools/hts-specs

The James Hutton Institute

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Intended to represent read alignments, also used for raw reads. Tab-delimited plain text. Headers (*optional*) start with "@"

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

BAM[a]/CRAM[b]

[a] https://github.com/samtools/hts-specs
[b] http://www.ebi.ac.uk/ena/software/cram-toolkit

The James Hutton Institute

BAM is a compressed version of SAM.

- BGZF compression.
- Random access within compressed file, through indexing.

CRAM format may come to dominate, especially in archives, as datasets get larger:

- Reference-based compression.[23]
- Highly suited to compression and archiving of *very* large amounts of sequence data.[24]

[23] Fritz *et al.* (2011) *Genome Res.* **21**:734-740 doi:10.1101/gr.114819.110
[24] Cochrane *et al.* (2012) *GigaScience* **1**:2 doi:10.1186/2047-217X-1-2

Table of Contents

Read repositories

The James Hutton Institute

Repositories are centrally-maintained locations that keep sequence read data from multiple projects
Submission to a repository is a requirement for publication. And the right thing to do!

- **ENA**: The European Nucleotide Archive (http://www.ebi.ac.uk/ena), maintained by EBI/EMBL
- **SRA**: The Short Read Archive (http://www.ncbi.nlm.nih.gov/sra), maintained in the US by NCBI

ENA
European Nucleotide Archive

Sequence Assembly

The James Hutton Institute

Once you have reads, you can assemble a genome.

Two main approaches to read assembly:

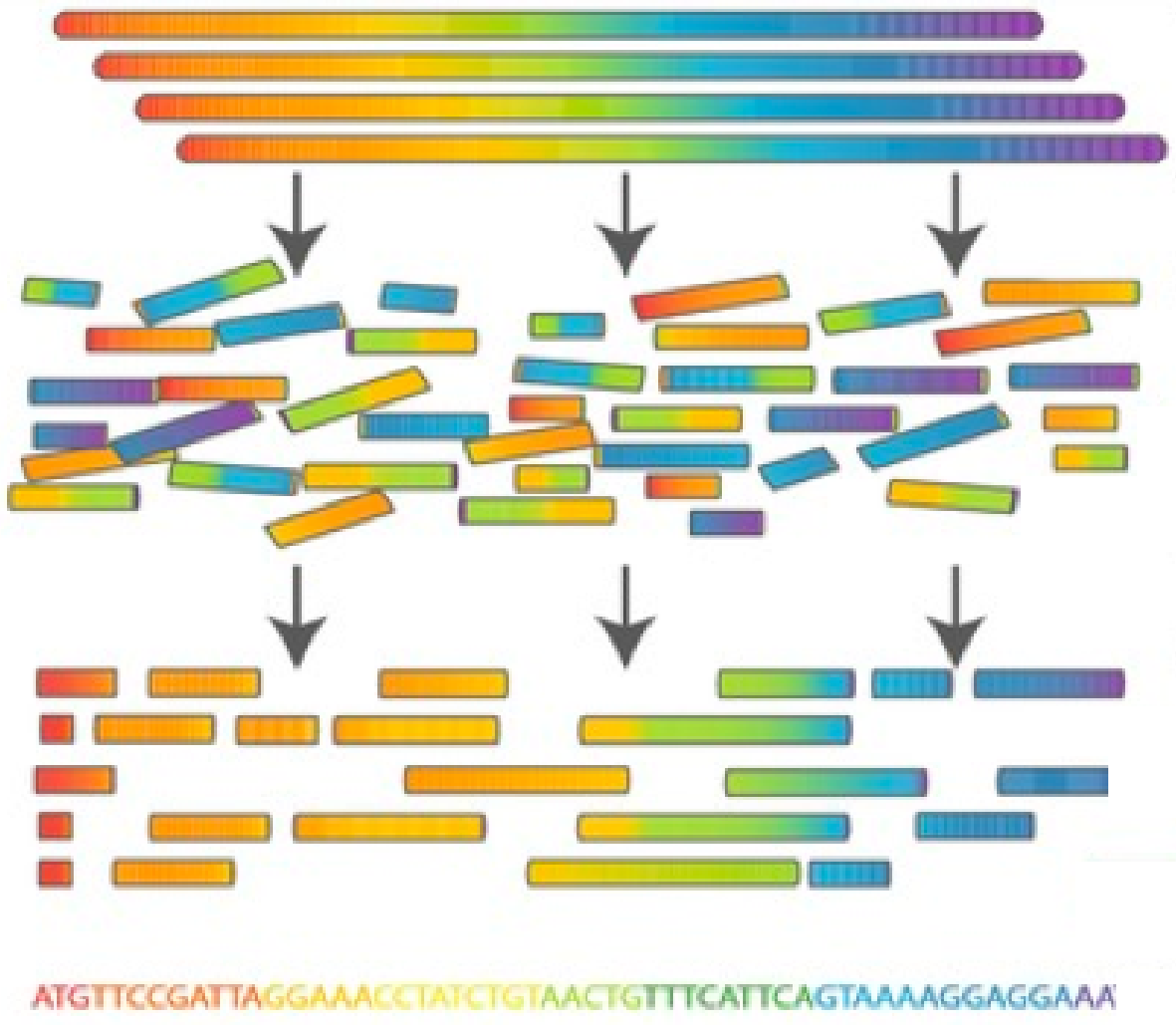
- **Overlap-Layout-Consensus**: Typically used with smaller sets of longer reads (e.g. 454, PacBio, Ion, Nanopore)
- **de Bruijn assembly**: Typically used with many, shorter reads (e.g. Illumina), but also useful for longer reads

See e.g. Leland Taylor's thesis (http://gcat.davidson.edu/phast/docs/Thesis_PHAST_LelandTaylor.pdf), and PHAST (http://gcat.davidson.edu/phast/index.html).

Table of Contents

Overlap-Layout-Consensus

The James Hutton Institute



Overlap-Layout-Consensus

The oldest approach, originally used with smaller sets of fewer reads.

Can be time consuming (all-vs-all comparisons), but offset with graph-based OLC algorithms (e.g. SGA).

Now more important again, with long-read data.

- Celera Assembler[25]
- Newbler (the Roche/454 GS assembler)[26]
- String Graph Assembler[27]

---

[25] http://wgs-assembler.sourceforge.net/

[26] http://www.454.com/products/analysis-software/

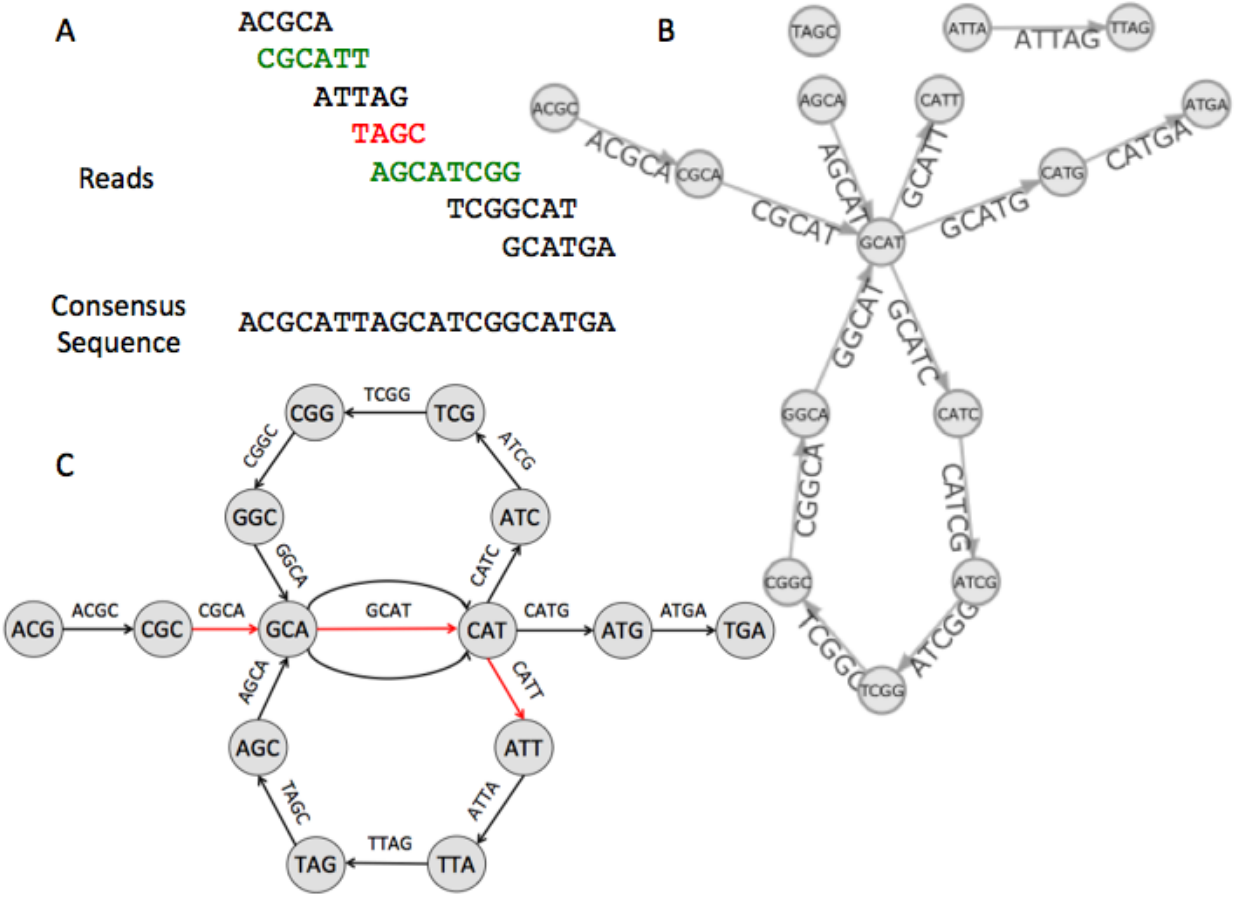
[27] Simpson and Durbin (2012) *Genome Res.* **22**:549-556 doi:10.1101/gr.126953.111

Table of Contents

de Bruijn graph assembly

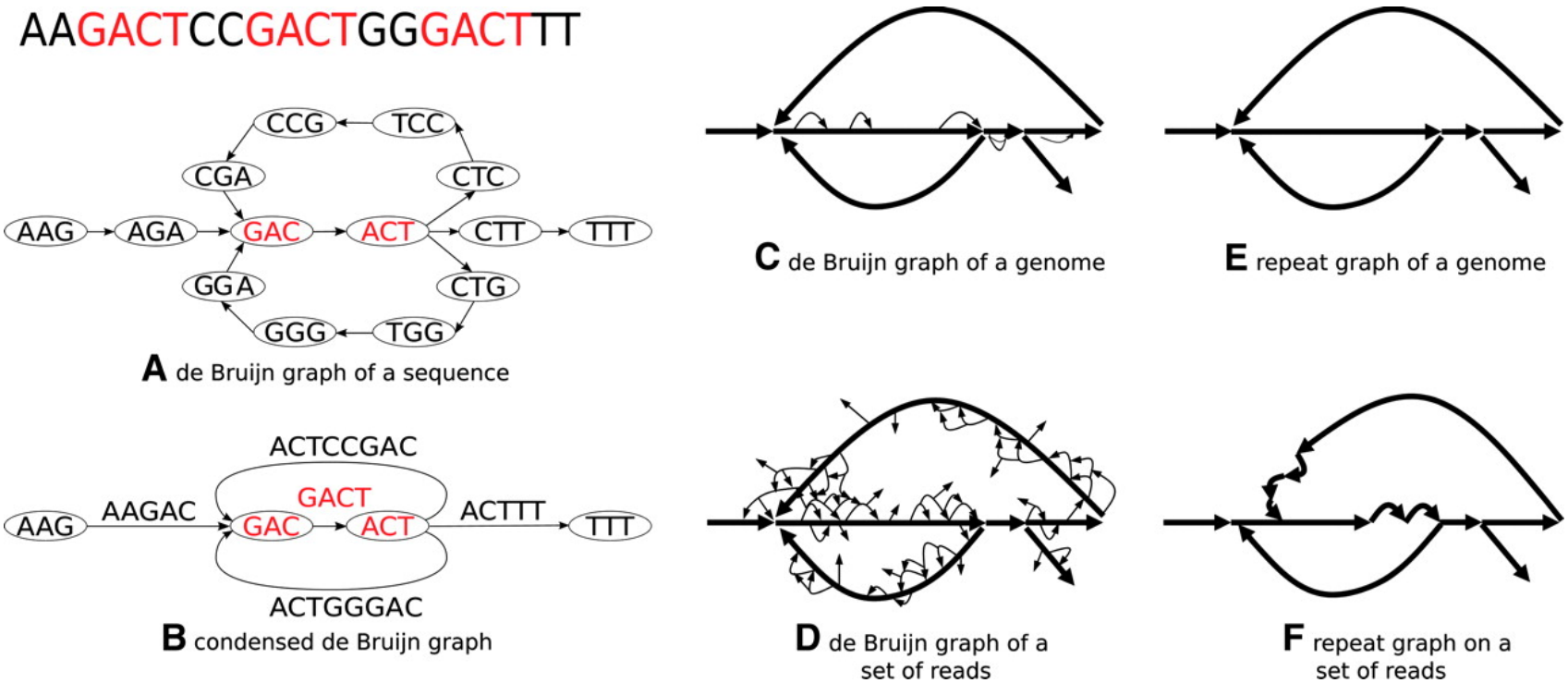
Used for short reads (e.g. Illumina):

k-mer based graph (choice of *k* important):



de Bruijn graph assembly

k-mer based genome and read graphs[28]
"True" edges = genome; "Error" edges = wrong assembly



[28] Chaisson *et al.* (2009) *Genome Res.* **19**:336-346 doi:10.1101/gr.079053.108

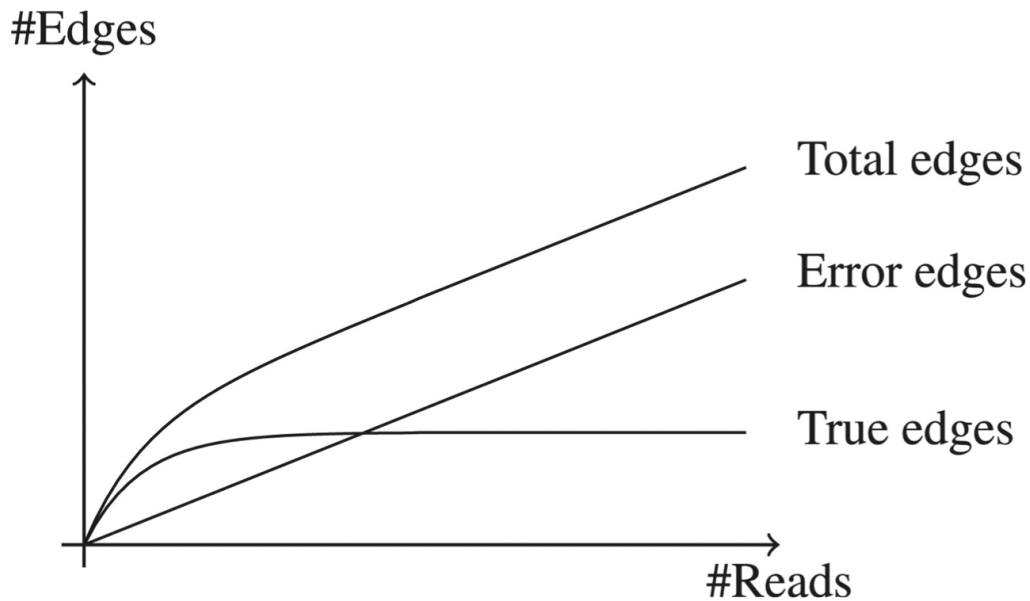
de Bruijn graph assembly

The James Hutton Institute

All sequencing technologies have basecall errors.

- The proportion of errors is approximately constant per read
- Basecall errors lead to edge errors
- The more reads you have, the more errors there are

Increased coverage does not ensure increased accuracy[29]



[29] Conway and Bromage (2011) *Bioinformatics* **27**:479-486 doi:10.1093/bioinformatics/btq697

de Bruijn graph assembly

The James Hutton Institute

Fast, and scales well to large datasets, as it never computes all-against-all overlaps.

Sensitive to sequencing errors, but resolves short repeats (graph bulges and whirls).

Notable tools:

- Velvet[30]
- CLC Assembly Cell[31]
- Cortex[32]

---

[30] Zerbino and Birney (2008) *Genome Res.* **18**:821-829 doi:10.1101/gr.074492.107

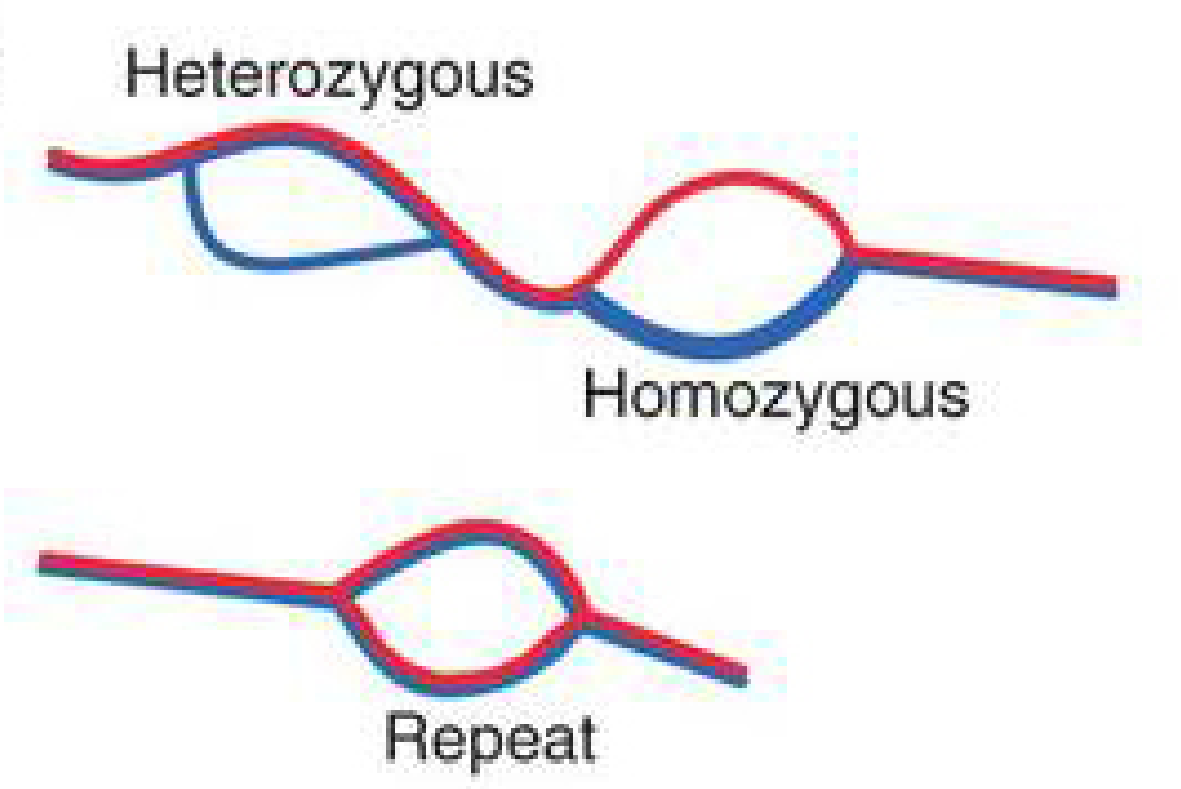
[31] http://www.clcbio.com/products/clc-assembly-cell/

[32] Iqbal *et al.* (2012) *Nat. Genet.* **44**:226-232 doi:10.1038/ng.1028

"Coloured" de Bruijn graph assemblies

The James Hutton Institute

Cortex[33] allows for on-the-fly identification of complex variation, and genotyping, by tracking "coloured" edges in the graph. Colours $\approx$ different isolates/organisms (e.g. a reference)



[33] Iqbal *et al.* (2012) *Nat. Genet.* **44**:226-232 doi:10.1038/ng.1028

Table of Contents

Why map reads?[a]

[a] Trapnell *et al.* (2009) *Nat. Biotech.* **27**:455-457 doi:10.1038/nbt0509-455

The James Hutton Institute

"Resequencing" an organism (sequencing a close relative, looking for SNPs/indels)

RNA-seq, ChIP-seq, etc. - coverage $\approx$ expression/binding

To see where reads map on an assembled genome

- Is coverage even? (can indicate repeats)
- Are there SNPs/indels? (heterogeneous population)
- Assembly problems?

Short-Read Sequence Alignment[a]

[a] Trapnell *et al.* (2009) *Nat. Biotech.* **27**:455-457 doi:10.1038/nbt0509-455

The James Hutton Institute

An embarrassment of tools (over 60 listed on Wikipedia)
Main approaches:

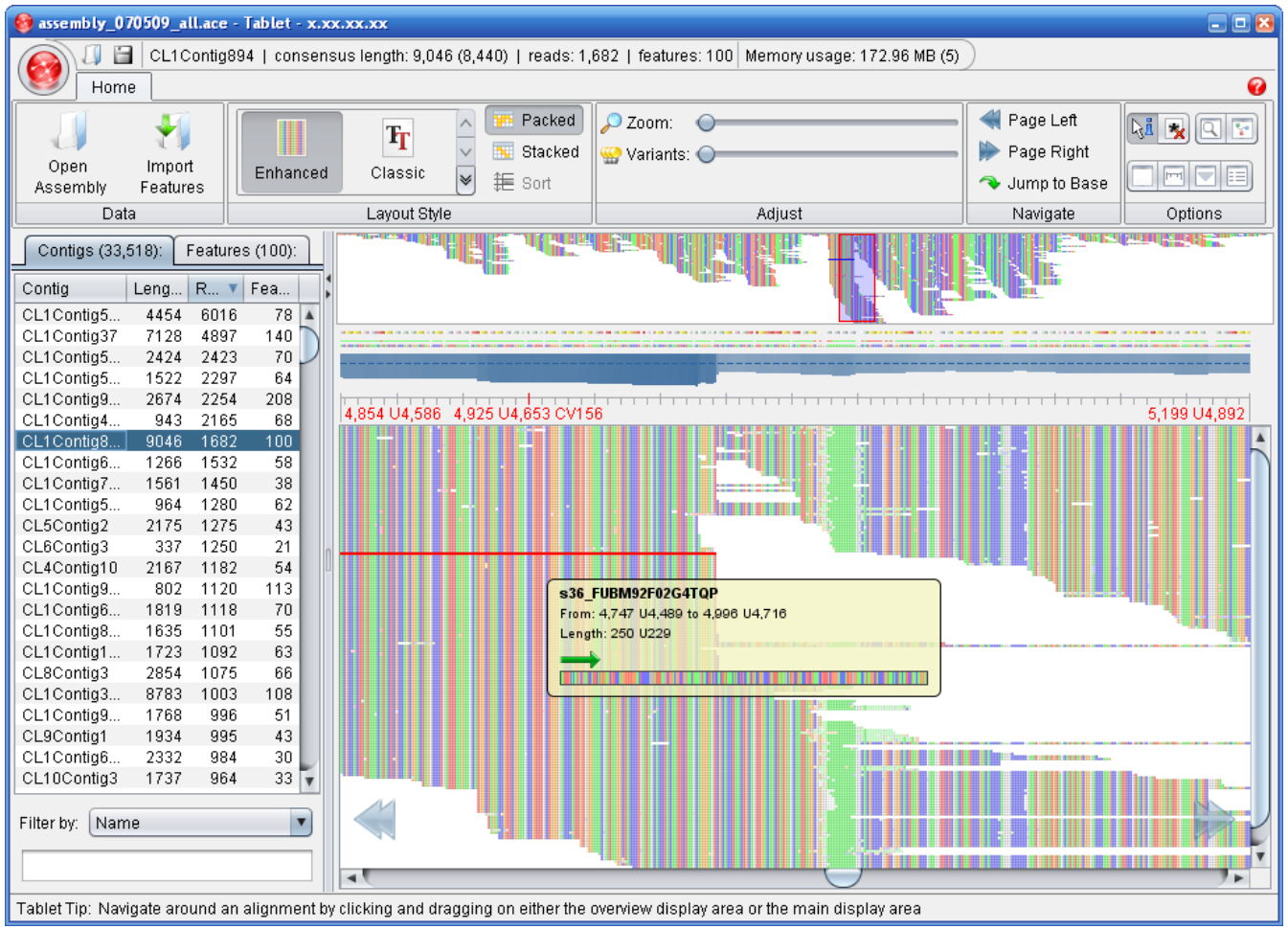
- **Alignment**: Smith-Waterman mathematically guaranteed to be the best alignment available (e.g. BFAST, MOSAIK); approximation to S-W (e.g. BLAST); ungapped or gapped alignment (e.g. MAQ, FAST, mrFAST, SOAP). Can be slow.
- **Burrows-Wheeler Transform**: Makes reusable index of the genome (e.g. Bowtie, BWA), can be extended to consider sequence probability (e.g. BWA-PSSM). Can be very fast.

Other tools may employ different algorithms, some designed to be parallelised on GPUs/FPGAs (e.g. NextGenMap, XpressAlign)

Visualising Read Mapping

The James Hutton Institute

Several tools available, e.g. Tablet (the best...)[34]



[34] Milne *et al.* (2013) *Brief. Bioinf.* **14**:193-202 doi:10.1093/bib/bbs012

Table of Contents

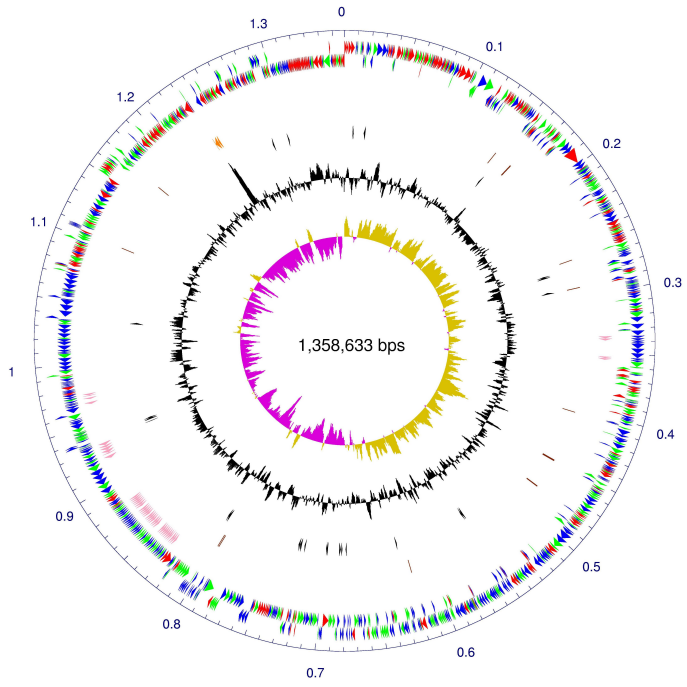
In an ideal world

The James Hutton Institute

Ideally, you would have one sequence per chromosome/plasmid. (and no errors): a **closed/complete** genome.

Eca_Sanger_genome.dna

```
>Erwinia_carotovora
ttatccacattcaactgccgaatatttaatcttatttgagtaaattaacccacgatccta
gccattcttctgctggatcttctggaatctgatgctgcgtgatgtcgatctcgagaatttt
caccgatccttttagccccccaattgggctagtagacgatccgctgtttggatcgcgccac
agaacgtgtcgtactctttactgccgatccctatcgcaccaaaattcacctgagaaagat
cgggcttttgctgttccaactgttcaaacagcgcttttaggttgtcagggaactccccctg
cgccgtgtgtcgacgttaccaccagccatataccgctctcgggaagtgcatcaagcgtag
ggccatgcagaatttcagtggagaaattatccttttctagcagttcagccaaatgctctg
ccacatattcggcgctgccaagggtactgccactgattaaggtaatgtctgccatagtga
tcccatccgataaaagacgggacattgtacgctgtgatctggctgggatctacctgtgga
taatatgggtataggaaaaaagcattacggggcgatcgtccgcataatggggttttgcag
ggagatcagtgtttcggtggactgaatttcatcaattgtttggatcttgttgataagtac
gtgttgcagcgcctcgatcgaacgacacatgaccttaataaagatactgtaatgcccagt
ggtgtagtaggcctcgaccacttcttccagattgtttagcttttccagcgccgagggata
atctttggcgcttttcagaatgatgccgataaagcagcacacgtcatagcctaactgttt
tggattcacgtccagacgcgtgccgataatgatgcccgcctgtttcatttttccacccg
aacgtggatcgttcctggactgacggaaaactgttttgccagttcggcataaggggtgcg
cgcattttccattaaggcggaaagaatgccgcgatcgagattatcgatctgataaatttc
tgccactttgagcctcctatttttatcgaataatgattttataccgctataaaatgacgga
ttaaaagcacaagtgcaatttatccataagattattgaatgactagcggttttattgct
taatgaacaactaaatacgatgaataagagacgcggcaatgaaaaagcagtacatcgaaa
agcaacagcaaatcagcttcgttaaatctttcttctccagtcaactggagcagcttctgg
gattgattgaagtccaggcaccgattctcagccgtatcggcgatggcacacaggataact
tgtctggtacagaaaaaagcggtgcaggttaaggtaaaagccttaccggatgccacatttg
aagtggtgcactcactggcaaaatggaaacgcaaaacgctgggcgcgtatgattttagct
tcggtgaaggcatttatactcacatgaaggcgctgcgtccggacgaagatcgcctgagcc
cgatccactcggtttatgtcgatcagtgggattgggagcgcgtgatggaggatggcgagc
gcaatgctgaatatctgaaatcgacggtcacgcgtatttatcaaggcattaaagcgactg
aggctgcggtacatcaggcgtttggcattcagcctttcctgccagagcagattcatttg
tgcataccgaaaaccttgctgaaagcgttatcccgatctggacgccaaagggcgtgagcgag
```

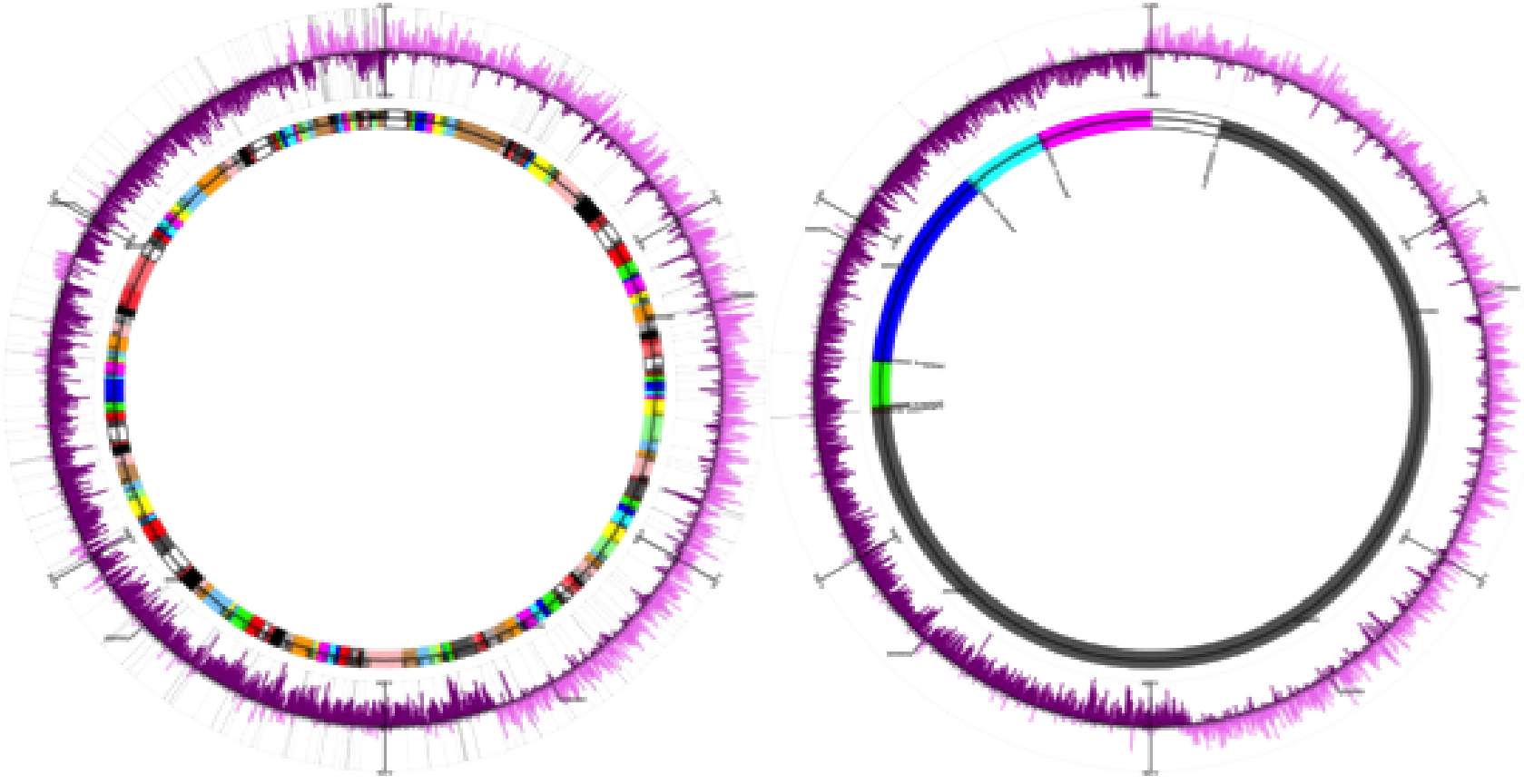


PacBio, Sanger, manual closing, Nanopore(?)

More realistically...

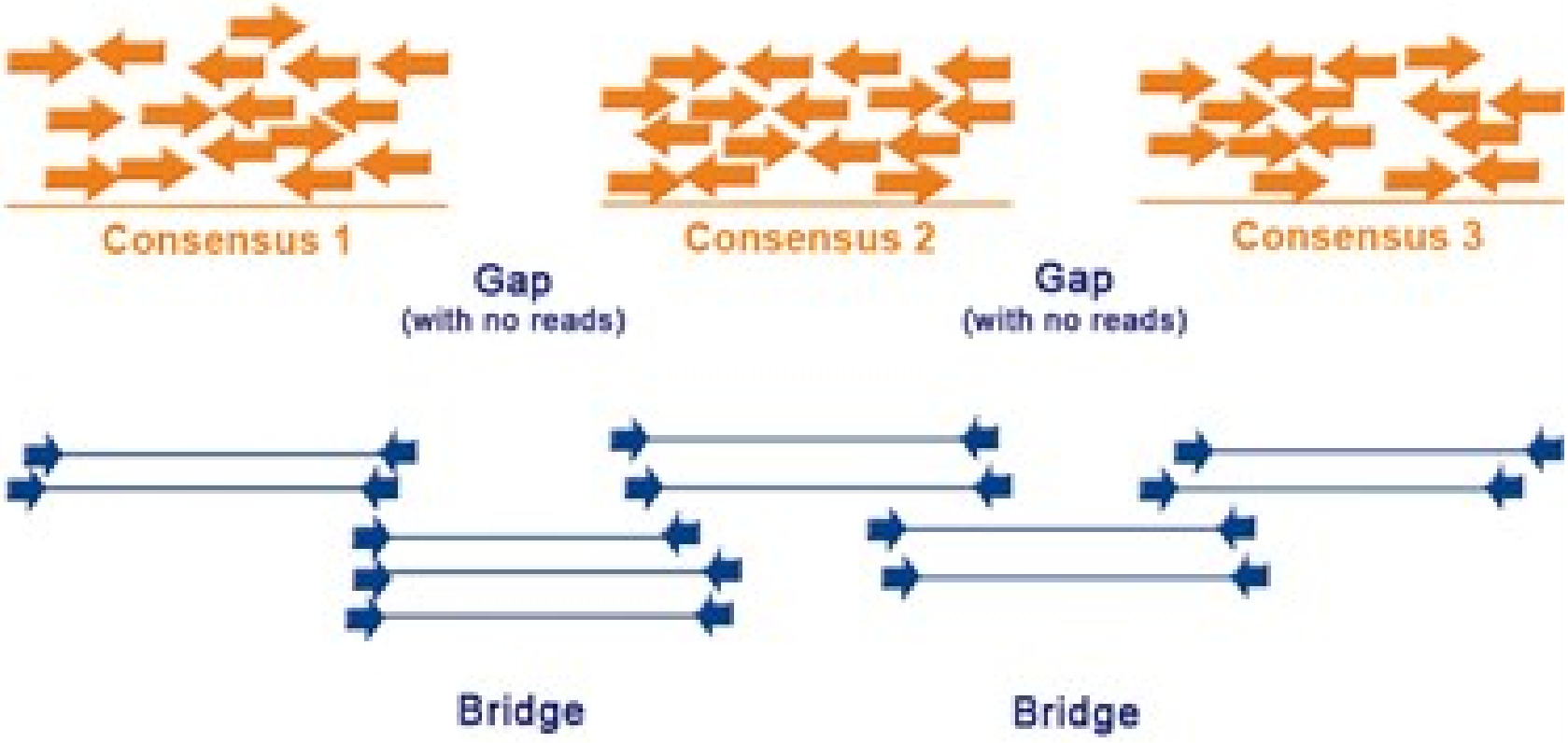
The James Hutton Institute

Typically, a number of assembled fragments (contigs or scaffolds) are returned in FASTA format: a **draft**, *disordered* genome. Around 250 contigs for a 5Mbp genome is usual with Illumina



Ordering contigs

Contigs can be ordered correctly into *scaffolds* if paired-end reads span gaps, or long reads are available (typically done during assembly).
Gaps are usually filled with `Ns` (length estimated)

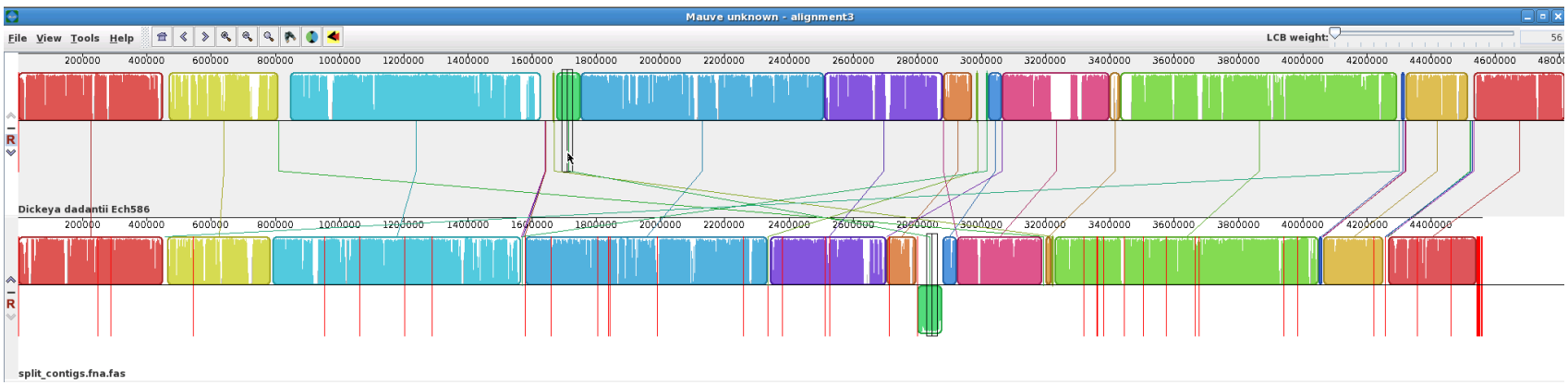


Ordering contigs

The James Hutton Institute

Contigs and scaffolds can also be reordered by alignment to a reference genome.

- Mauve/progressiveMauve[35]
- MUMmer[36]



[35] Darling *et al.* (2004) *Genome Res.* **14**:1394-1403 doi:10.1101/gr.2289704

[36] Kurtz *et al.* (2004) *Genome Biol.* **5**:R12 doi:10.1186/gb-2004-5-2-r12

Where next?[a]

[a] Lefebure *et al.* (2010) *Genome Biol. Evol.* **2**:646-655 doi:10.1093/gbe/evq048

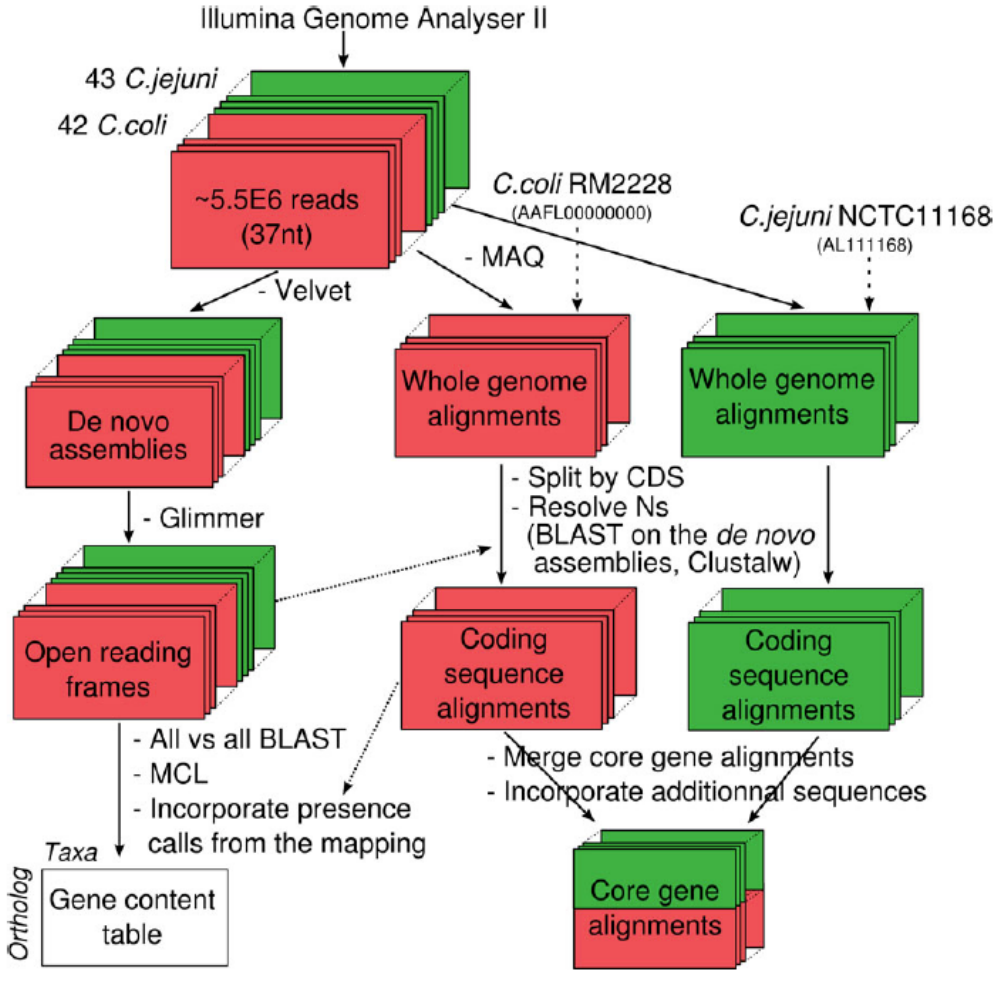


FIG. 1.—Pipeline combining de novo assemblies and read mapping, yielding a gene content table and core gene alignments.

Microbial Genomics and Bioinformatics
BM405
3.Whole Genome Comparisons

The James Hutton Institute

Leighton Pritchard[1,2,3]
[1]Information and Computational Sciences,
[2]Centre for Human and Animal Pathogens in the Environment,
[3]Dundee Effector Consortium,
The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

Acceptable Use Policy

The James Hutton Institute

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available on SlideShare.

These slides, and supporting material including exercises, are available at https://github.com/widdowquinn/Teaching-Strathclyde-BM405

Table of Contents

The Power of Comparative Genomics

The James Hutton Institute

Massively enabled by high-throughput sequencing, and the availability of thousands of sequenced isolates.

Computational comparisons more powerful and precise than experimental comparative genomics: **the ultimate microbial typing solution**

Three broad areas/scales:

- Comparison of bulk genome properties
- Whole genome sequence comparisons
- *Comparison of features/functional components*

Table of Contents

Nucleotide frequency/genome size

The James Hutton Institute

- Very easy to calculate from complete/draft genome
- Can calculate for individual contigs/scaffolds/regions
- Usually reported in GUI genome browsers

Trivial to determine using, e.g. Python

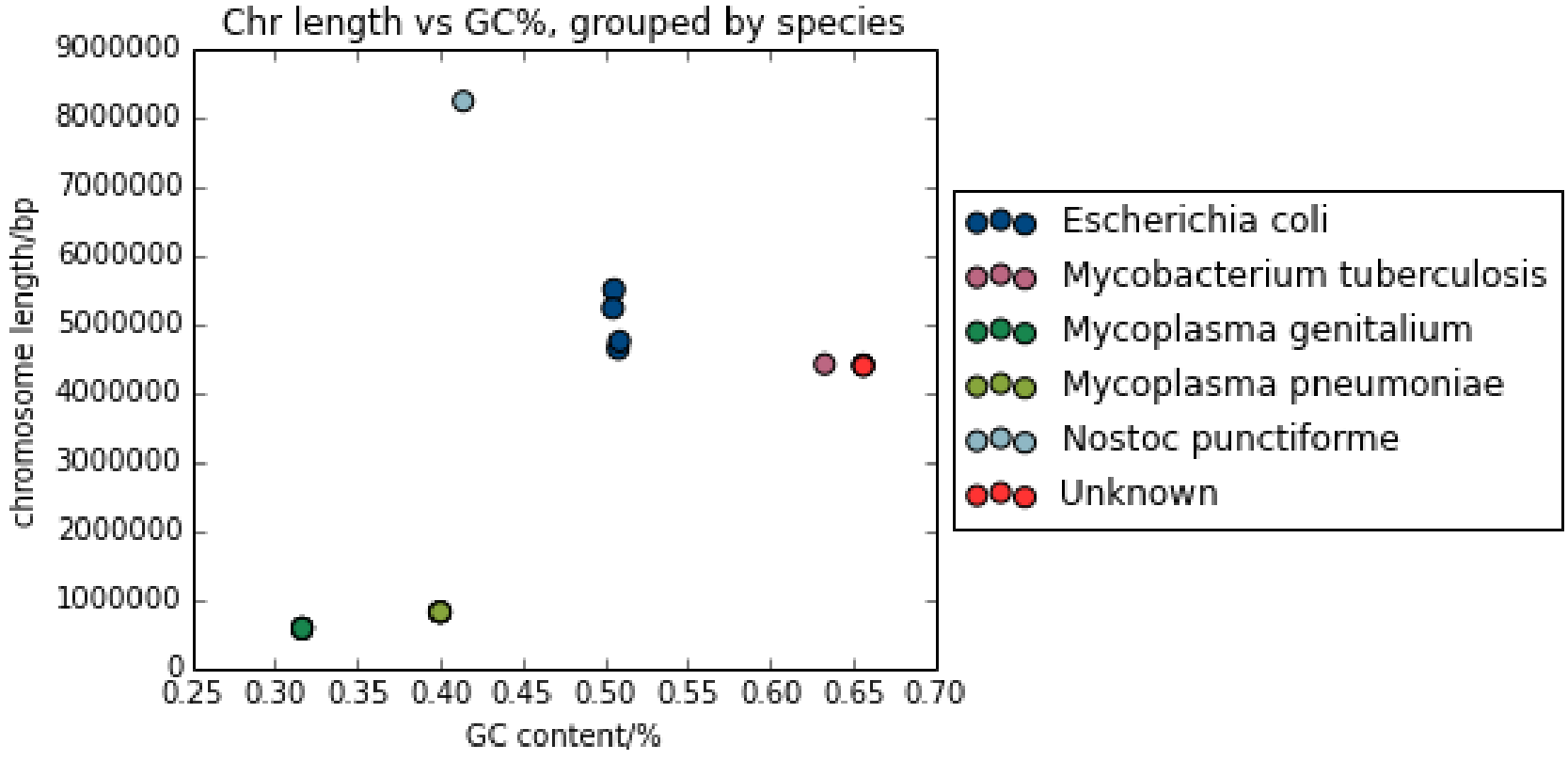
```
In [1]: from Bio import SeqIO
In [2]: s = SeqIO.read("data/NC_000912.fna", "fasta")
In [3]: a, c, g, t = s.seq.count("A"), s.seq.count("C"), s.seq.count("G"), s.seq.count("T")
In [4]: float(g + c)/len(s)
Out[4]: 0.40008010837904245
In [5]: float(g - c)/(g+c)
Out[5]: 0.002397259225467894
```

Nucleotide frequency/genome size

The James Hutton Institute

GC content and chromosome size can be characteristic

See `data/bacteria_size` for example iPython notebook exercise



Blobology[a]

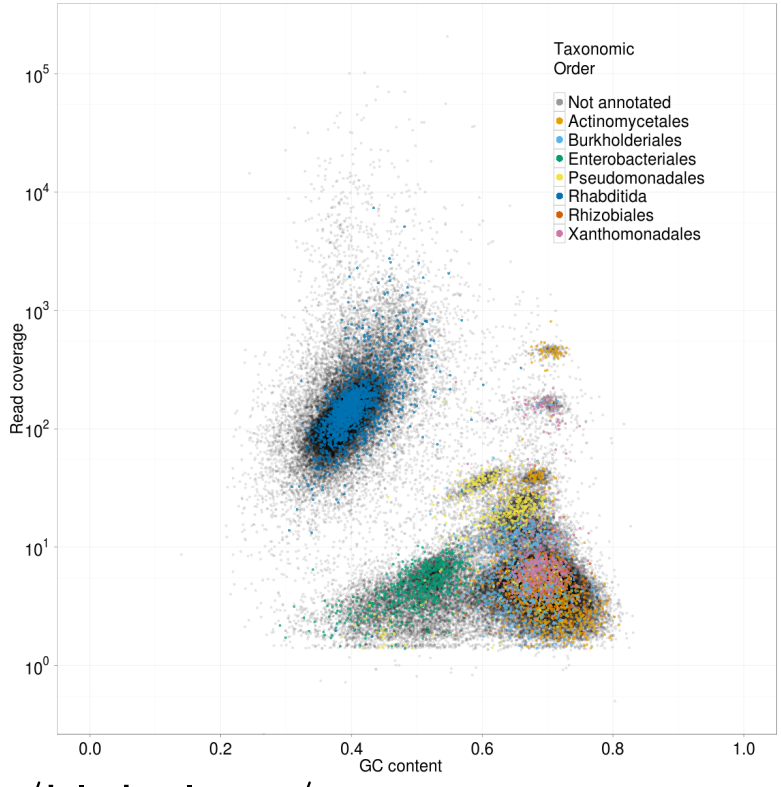
[a] Kumar and Blaxter *et al.* (2011) *Symbiosis* **3**:119-126 doi:10.1007/s13199-012-0154-6

The James Hutton Institute

Sequencing samples may be contaminated or contain microbial symbionts.

Expect more host than symbiont/contaminant DNA

GC content and read coverage can be used to separate contigs, following assembly and mapping



http://nematodes.org/bioinformatics/blobology/

k-mers

The James Hutton Institute

- Nucleotides: `[ACGT]`
- Dinucleotides: `[AA|AC|AG|AT|CA|CC|...]` (16 dimers)
- Trinucleotides: `[AAA|AAC|AAG|AAT|ACA|...]` (64 trimers)
- *k*-mers: 4^k *k*-mers

(see example in `data/shiny`)

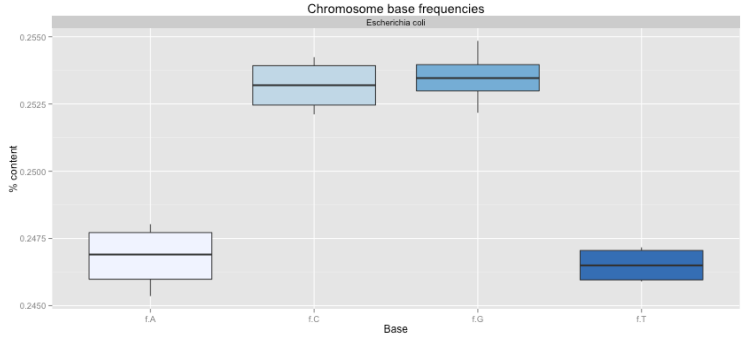
k-mers

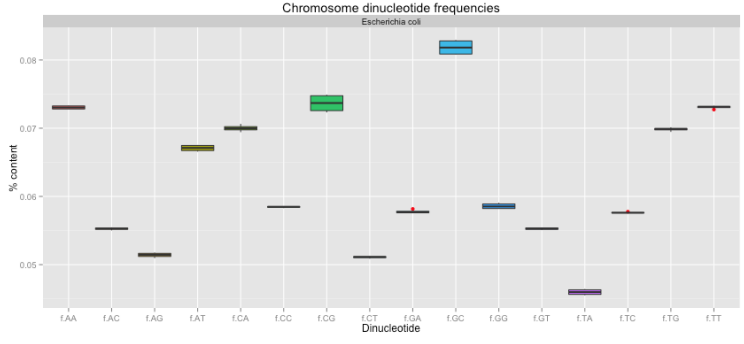
GC content = point value; *k*-mer frequencies = vector (list)

Diagnostic differences in *k*-mer frequency, and variability.

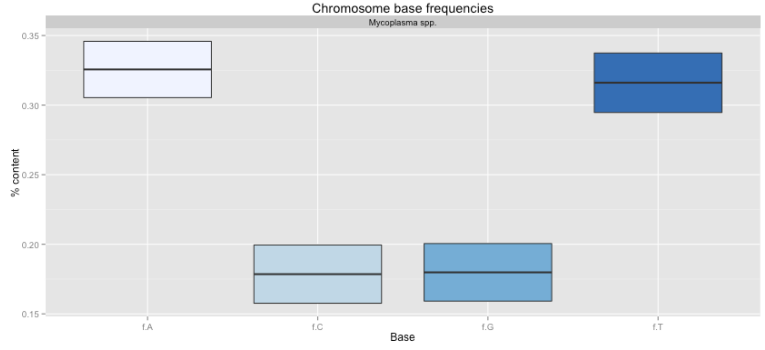
The basis of several comparison tools

E.coli





Mycoplasma spp.



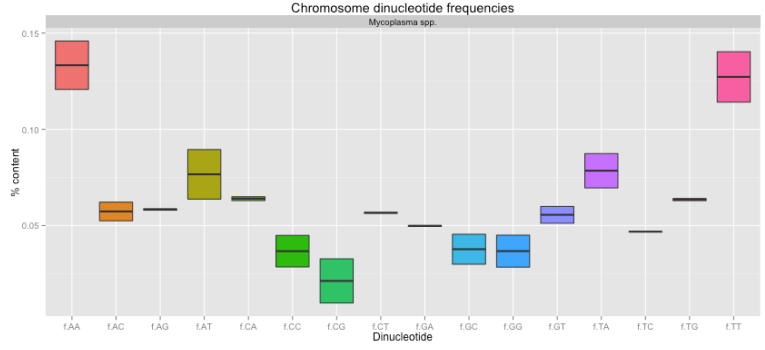
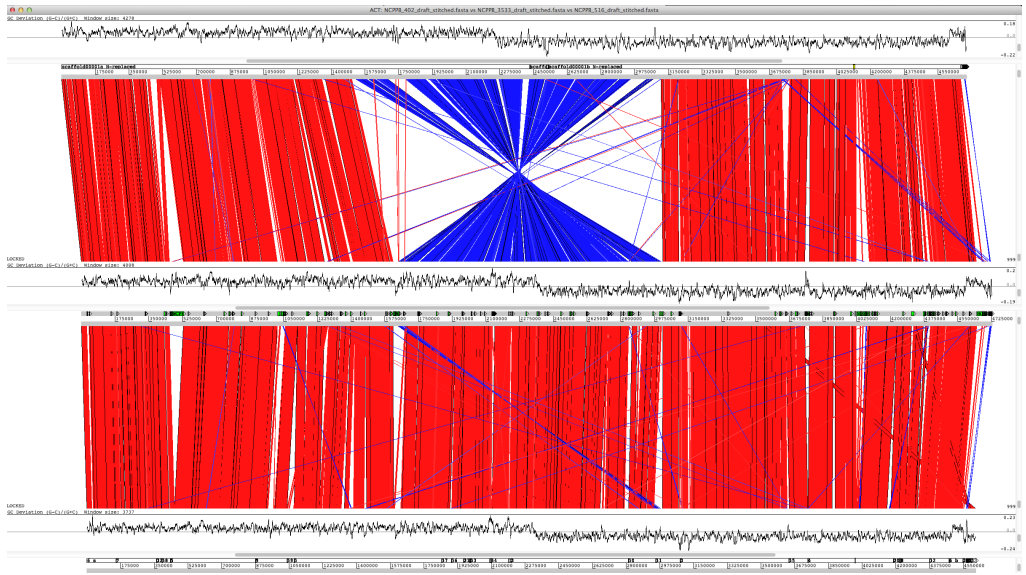


Table of Contents

What to align, and why?



To be useful, aligned genomes should:

- derive from a sufficiently recent common ancestor, so homologous regions can be identified
- derive from a sufficiently distant common ancestor, so that there are "interesting" differences to be identified
- **help to answer your biological question**

How to align, and why?

The James Hutton Institute

Naive sequence aligners (Needleman-Wunsch, Smith-Waterman) are not appropriate for genome alignment

- Computationally expensive on large sequences
- Cannot handle rearrangements

Very many alternative alignment algorithms proposed

- **megaBLAST** http://www.ncbi.nlm.nih.gov/blast/html/megablast.html
- **MUMmer** http://mummer.sourceforge.net/
- BLAT http://genome.ucsc.edu/goldenPath/help/blatSpec.html
- LASTZ http://www.bx.psu.edu/~rsharris/lastz/
- LAGAN http://lagan.stanford.edu/lagan_web/index.shtml
- and many, many more...

Example exercises in `data/whole_genome_alignment`.

megaBLAST

The James Hutton Institute

Optimised for speed, over BLASTN[37]

- Genome-level searches
- Queries on large sequence sets
- Long alignments of very similar sequence

Uses the greedy algorithm by Zhang *et al.*[38], **not** BLAST algorithm.

- Concatenates queries ("query packing") to improve performance
- Two modes: **megaBLAST** and discontinuous (**dc-megablast**) for divergent sequences

BLASTN now uses the megaBLAST algorithm by default

---

[37] http://www.ncbi.nlm.nih.gov/blast/Why.shtml

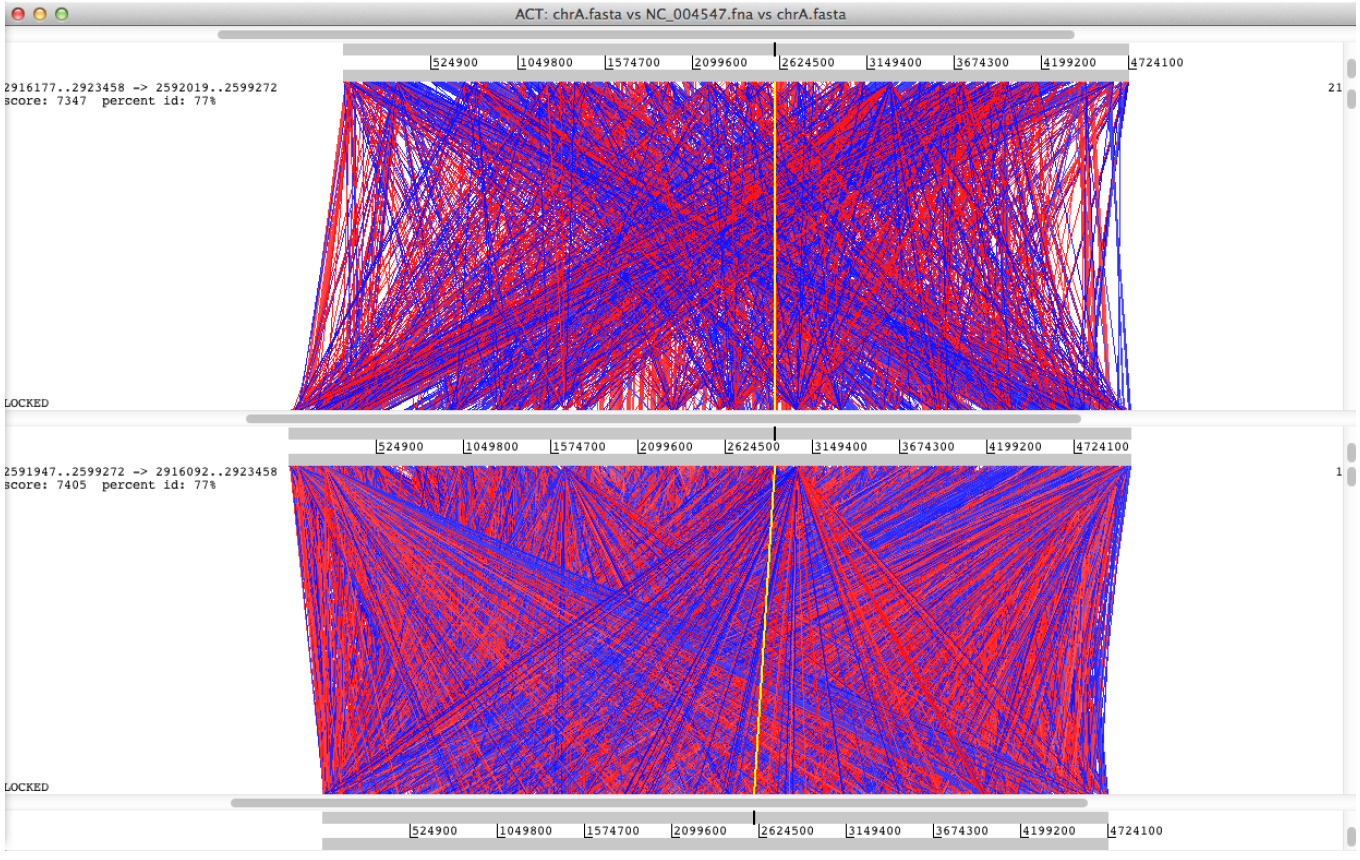
[38] Zhang *et al.* (2000) *J. Comp. Biol.* **7**:203-214 doi:10.1089/10665270050081478

BLAST vs megaBLAST

The James Hutton Institute

megaBLAST is faster, but does it give the same biological results?

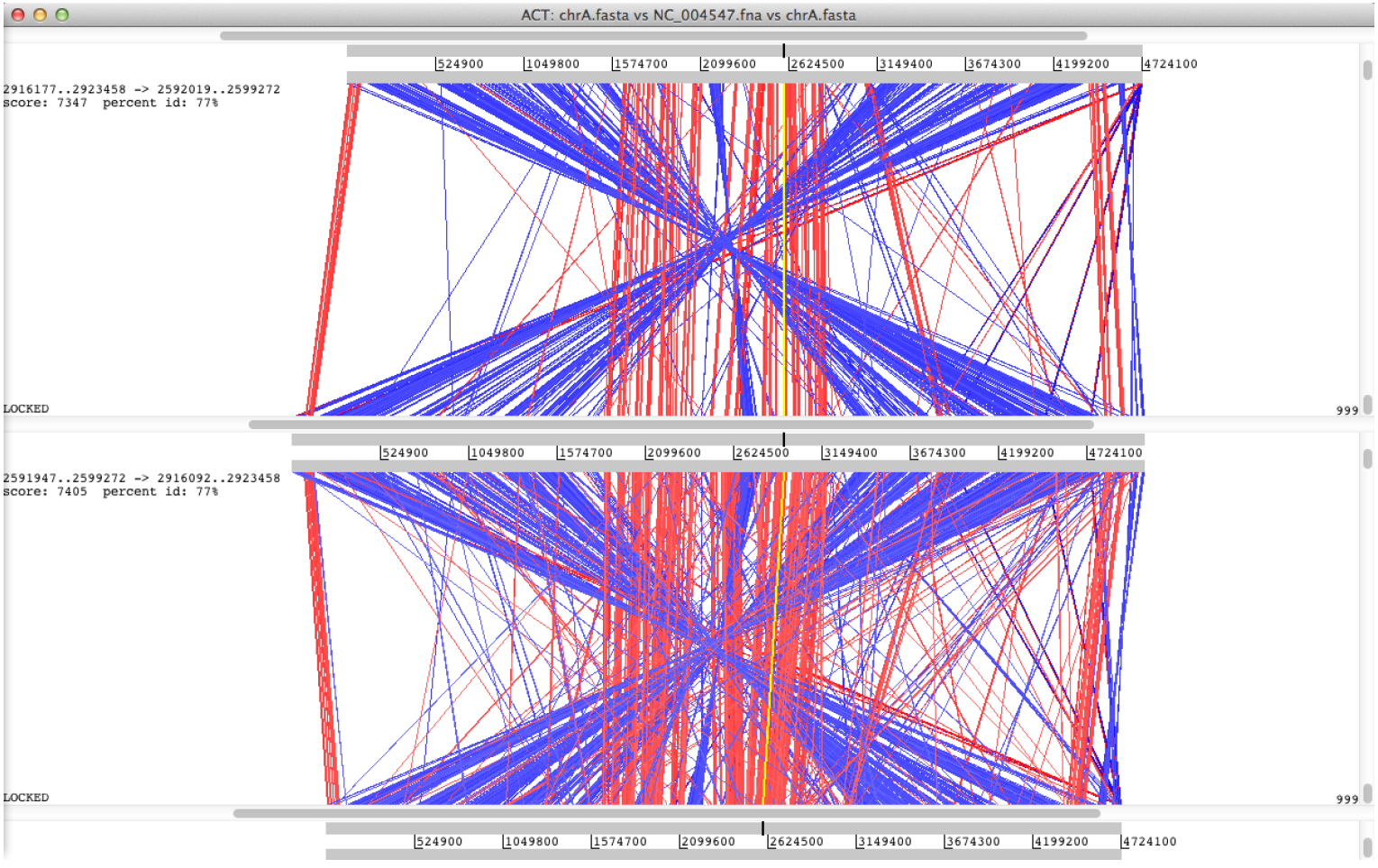
megaBLAST (top) and BLAST (bottom) pairwise comparisons:



BLAST vs megaBLAST

The James Hutton Institute

Filter out weak matches - not quite identical:



MUMmer[a]

[a] Kurtz *et al.* (2004) *Genome. Biol.* **5**:R12 doi:10.1186/gb-2004-5-2-r12

The James Hutton Institute

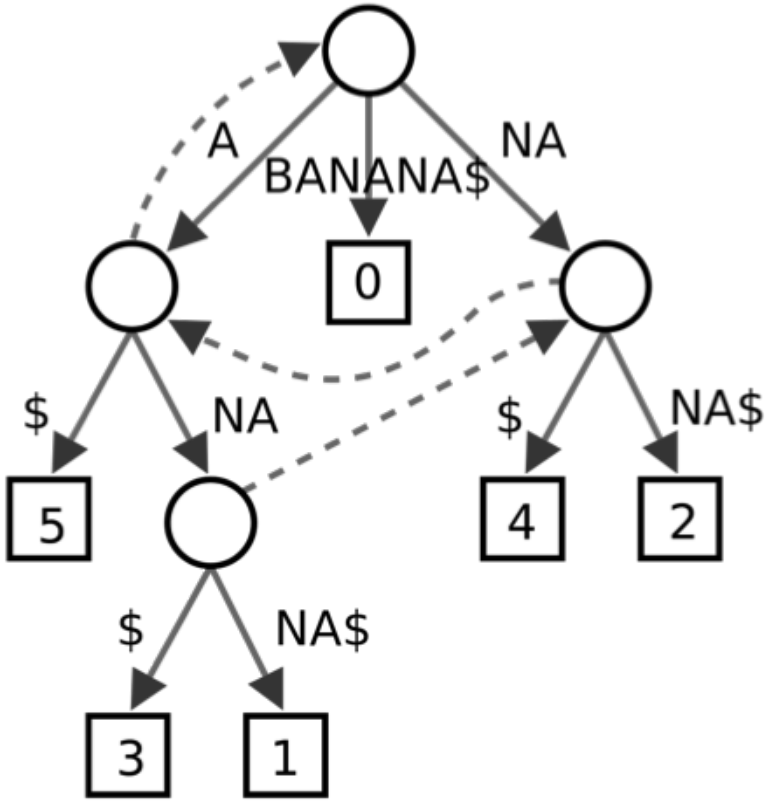
Uses *suffix trees* for pattern matching: very fast even for large sequences

- Finds *maximal exact matches*
- Memory use depends only on the reference sequence size

Suffix trees:

(http://en.wikipedia.org/wiki/Suffix_tree)

- Can be built and searched in $O(n)$ time
- But useful algorithms are nontrivial



The MUMmer algorithm[a]

[a] Kurtz *et al.* (2004) *Genome. Biol.* **5**:R12 doi:10.1186/gb-2004-5-2-r12

The James Hutton Institute

1. Identify a non-overlapping subset of maximal exact matches: often *Maximal Unique Matches (MUMs)*
2. Cluster into *alignment anchors*
3. Extend between anchors to produce the final alignment

This is the basis of a very flexible suite of programs that align different kinds of sequence: `mummer`, `nucmer`, `promer`

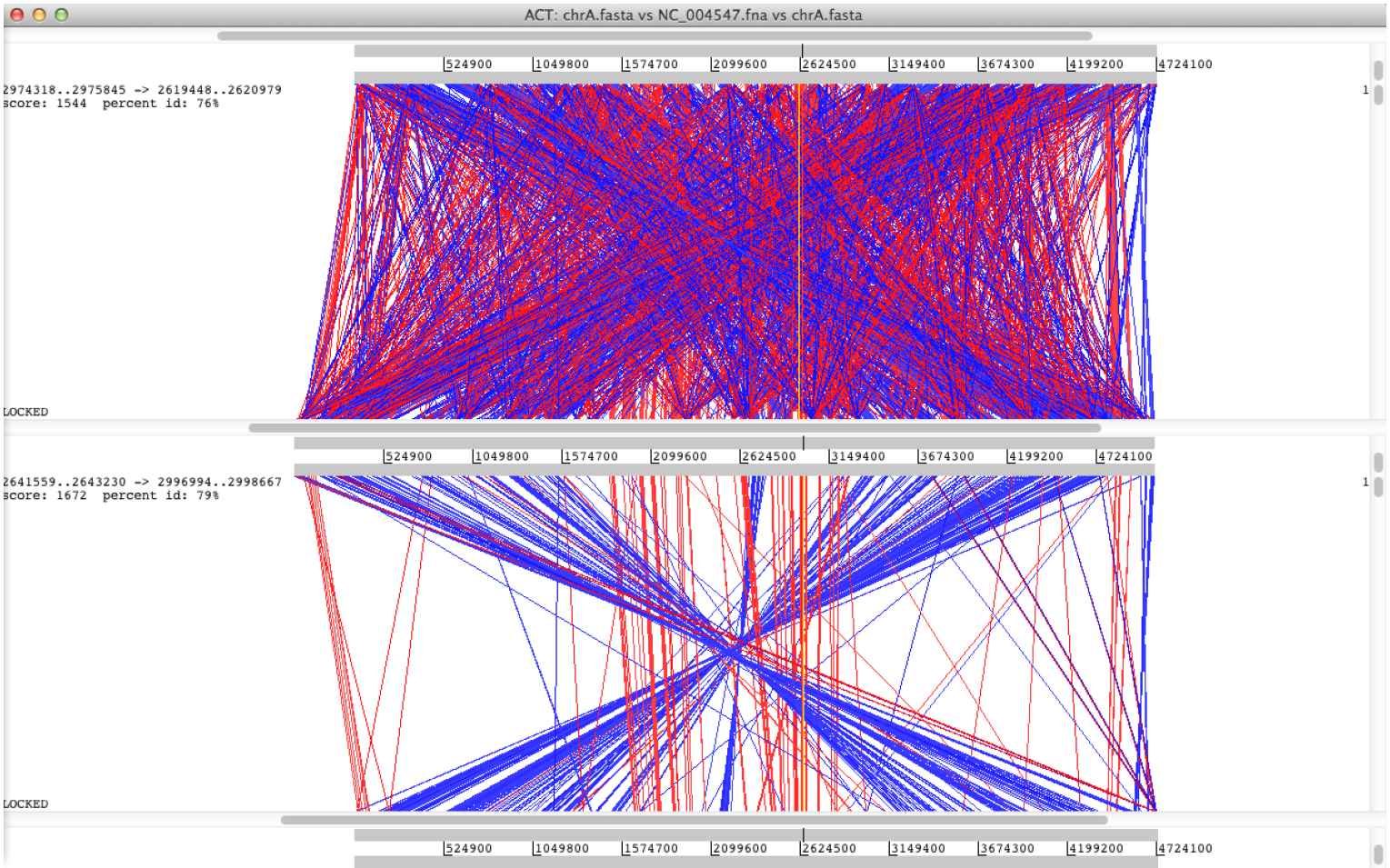
- nucleotide and (more sensitive) "conceptual protein" alignments
- used for genome comparisons, assembly scaffolding, repeat detection, ...
- the basis of other aligners/assemblers (e.g. Mugsy, AMOS)

MUMmer vs megaBLAST

The James Hutton Institute

MUMmer identifies fewer weak matches

megaBLAST (top) and MUMmer (bottom) pairwise comparisons:



MUMmer vs megaBLAST

The James Hutton Institute

Filter out weak BLAST matches - not quite identical:

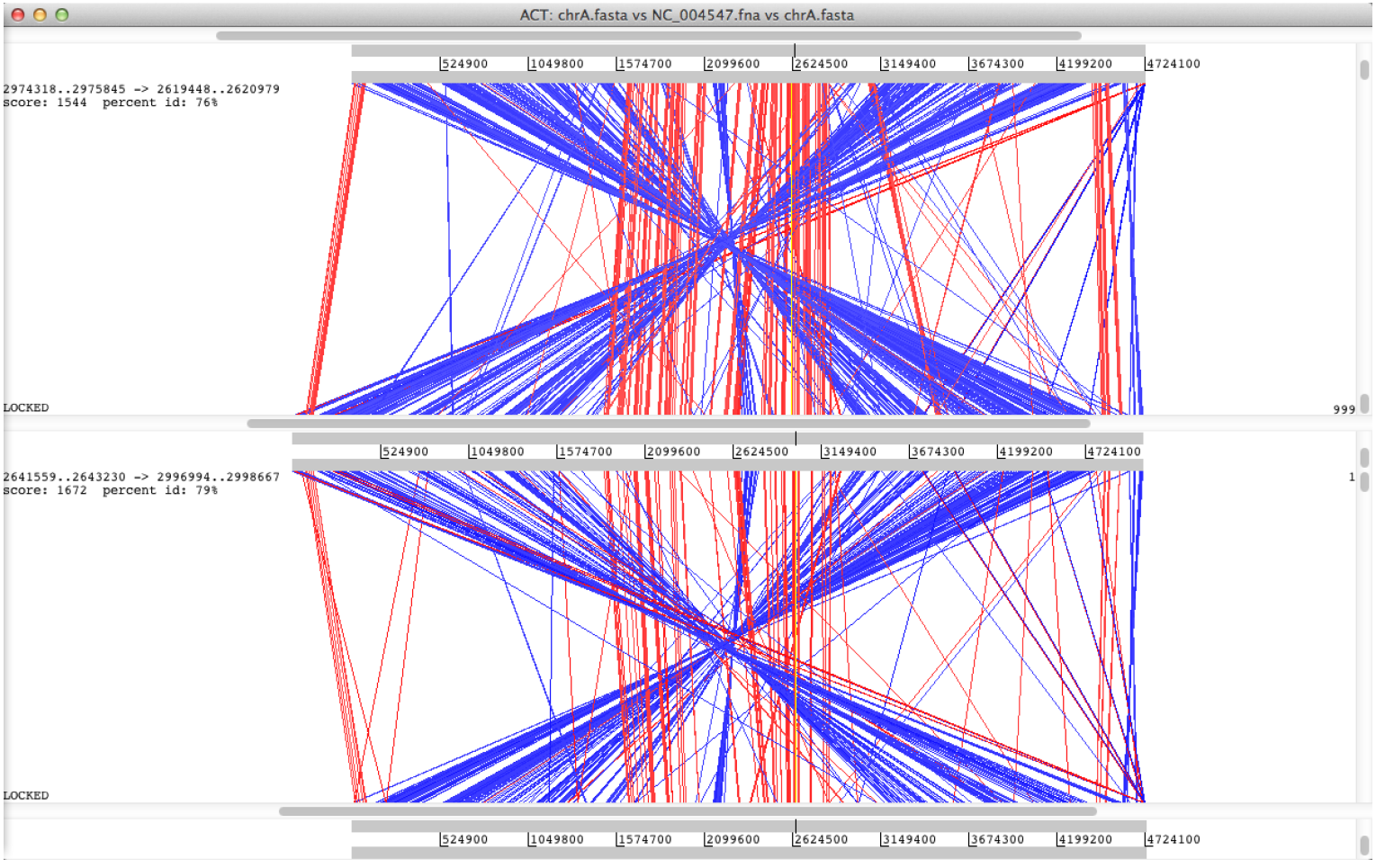


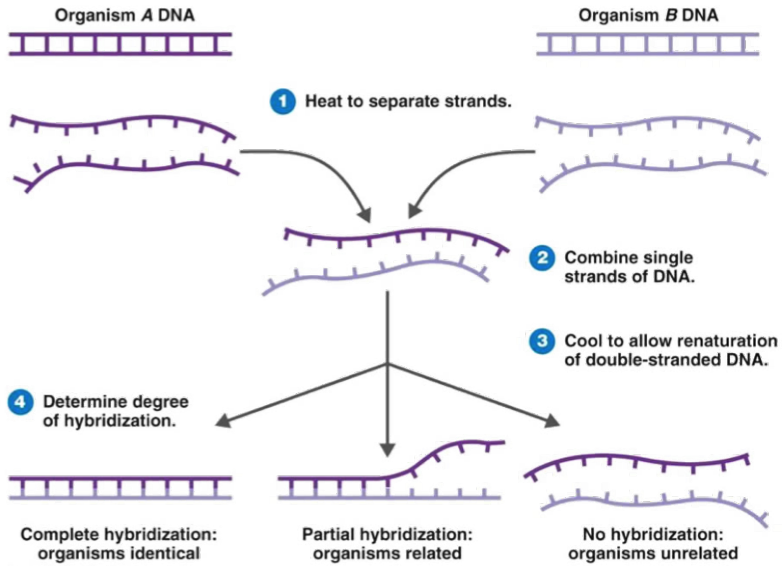
Table of Contents

DNA-DNA hybridisation[a]

[a] Morello-Mora and Amann (2001) *FEMS Micro. Rev.* **25**:39-67 doi:10.1016/S0168-6445(00)00040-1

The James Hutton Institute

- "Gold Standard" for prokaryotic taxonomy, since 1960s. "70% identity ≈ same species."
- Denature DNA from two organisms.
- Allow to anneal. Reassociation ≈ similarity, measured as ΔT of denaturation curves.



Proxy for sequence similarity - replace with genome analysis[39]?

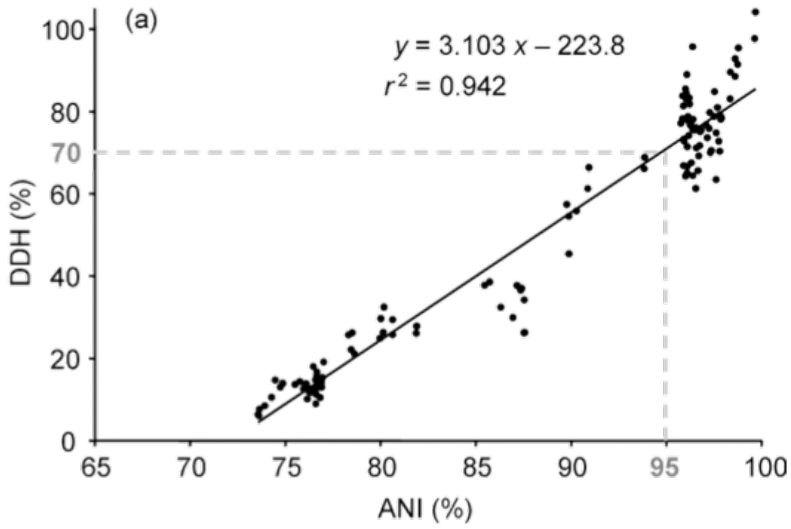
[39] Chan *et al* (2012) *BMC Microbiol.* **12**:302 doi:10.1186/1471-2180-12-302

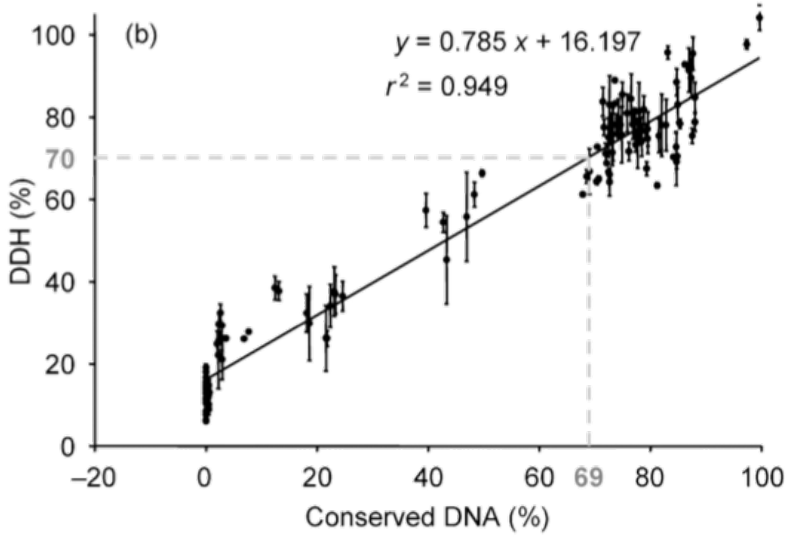
Average Nucleotide Identity (ANIb)[a]

[a] Goris *et al.* (2007) *Int. J. Syst. Biol.* **57**:81-91 doi:10.1099/ijs.0.64483-0

1. Break genomes into 1020t fragments
2. **ANIb**: Mean % identity of all BLASTN matches with $> 30\%$ identity and $> 70\%$ fragment coverage.

- DDH:ANIb linear
- DDH:%ID linear
- 70%ID $\approx$ 95%ANIb



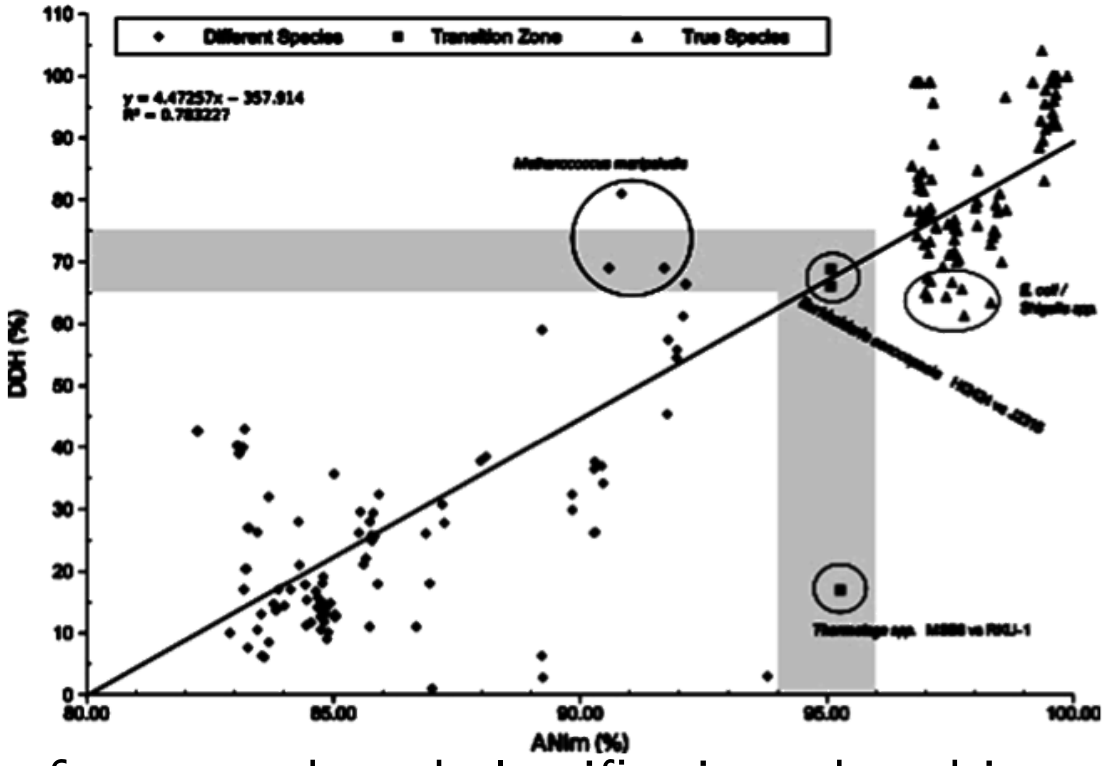


Average Nucleotide Identity (ANIm)[a]

[a] Richter and Rossello-Mora (2009) *Proc. Natl. Acad. Sci. USA* **106**:19126-19131 doi:10.1073/pnas.0906412106

1. Align genomes (MUMmer)
2. **ANIm**: Mean % identity of all matches

- DDH:ANIm linear
- 70%ID ≈ 95%ANIb



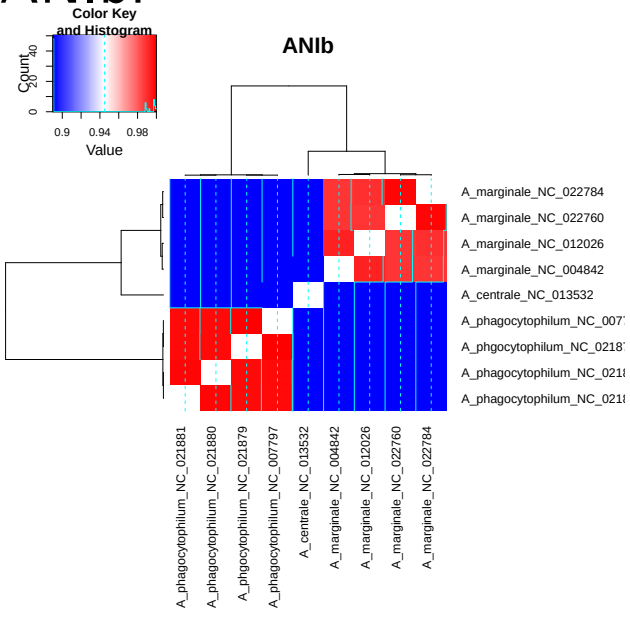
TETRA: tetranucleotide frequency-based classifier introduced in same paper.

ANI/TETRA comparison

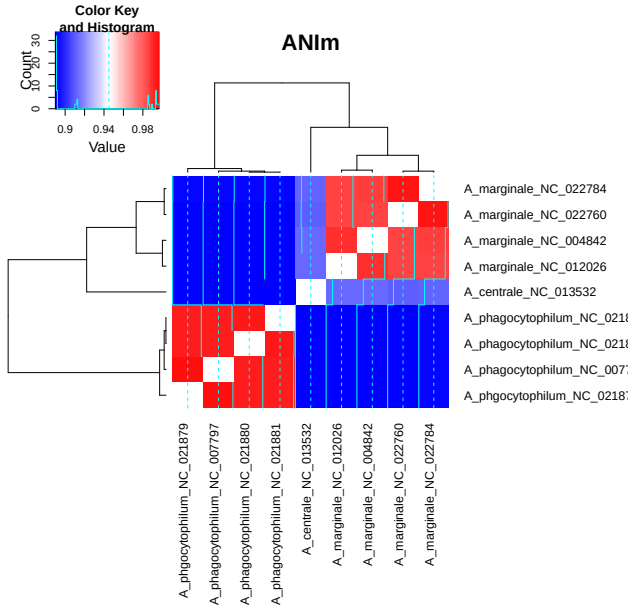
The James Hutton Institute

All three methods applied to *Anaplasma* spp.

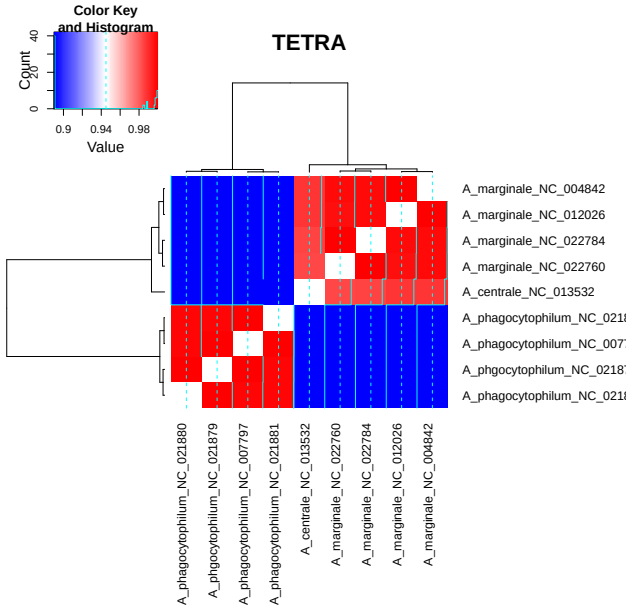
ANIb:



ANIm:



TETRA:



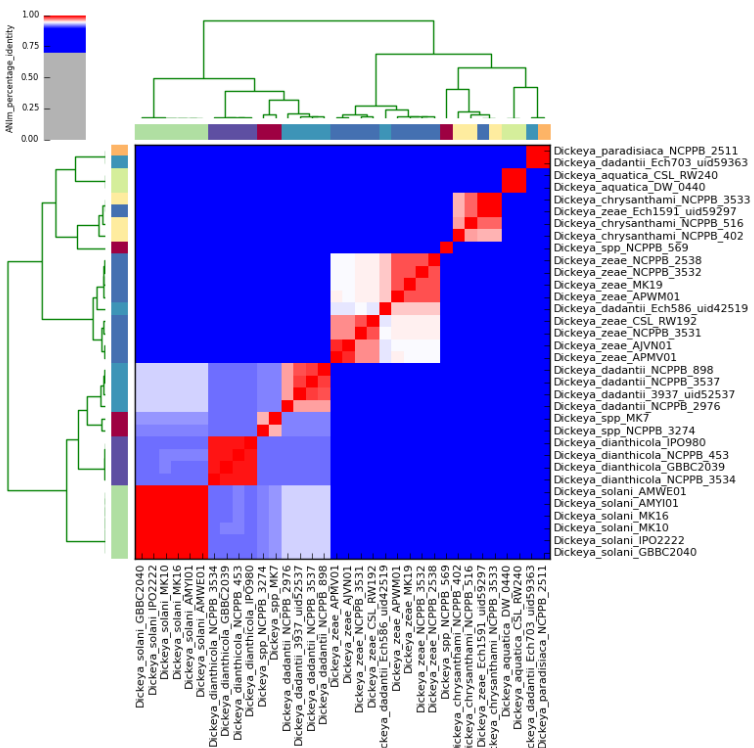
ANIb discards information, relative to ANIm: less sensitive

ANIb/ANIm $\approx$ evolutionary history; TETRA $\approx$ bulk composition

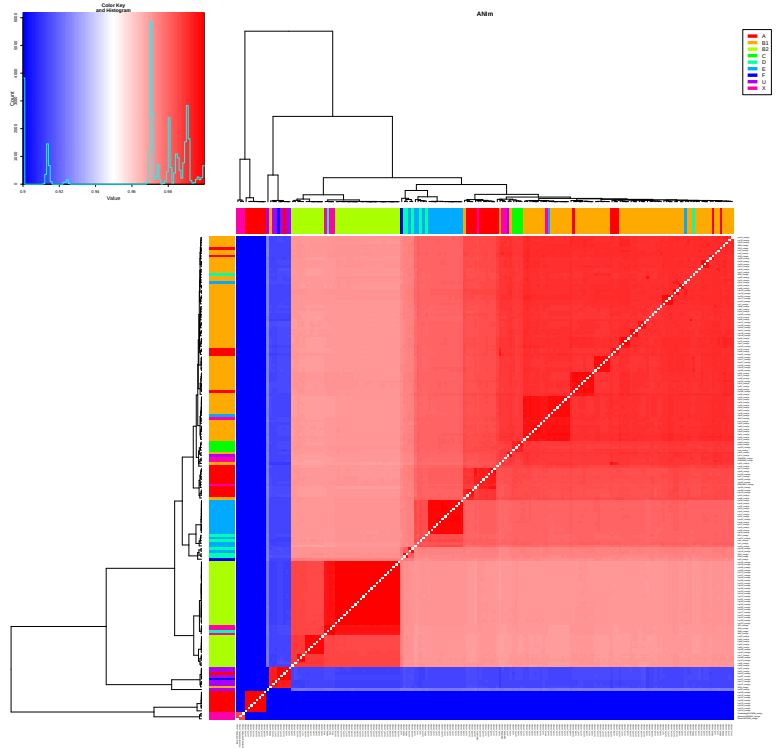
ANI in practice

Practical applications[40] (note: no gene content used)

34 *Dickeya* isolates: species structure



180 *E.coli* isolates: subtyping



[40] van der Wolf *et al.* (2014) *Int. J. Syst. Evol. Micr.* **64**:768-774 doi:10.1099/ijs.0.052944-0

Table of Contents

The James Hutton Institute

Collinearity and Synteny

The James Hutton Institute

Genome rearrangements occur, but there can still be conservation of sequence similarity and ordering.

- Two elements are **collinear** if they lie in the same linear sequence
- Two elements are **syntenous** (or *syntenic*) if:
  - (*orig.*) they lie on the same chromosome
  - (*mod.*) there is conservation of blocks of order within the same chromosome

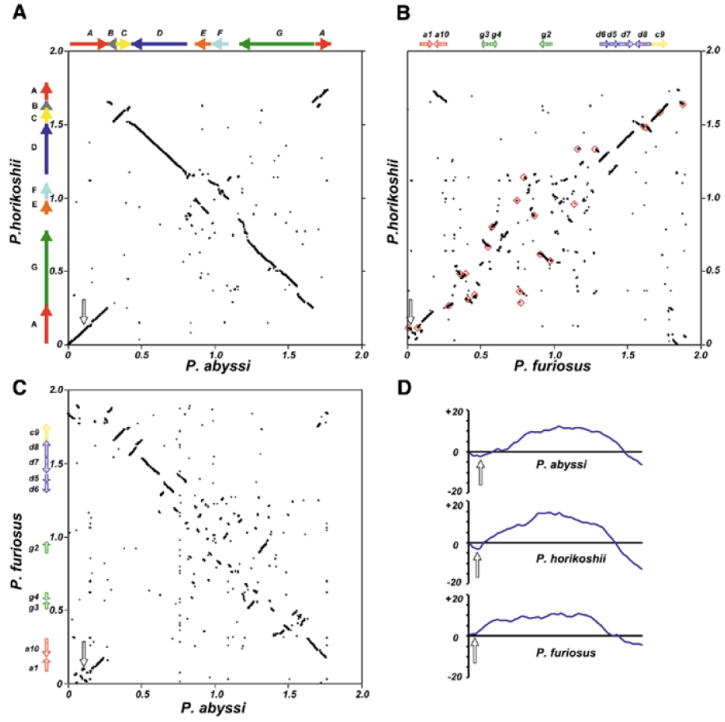
Signs of evolutionary constraints, like sequence conservation or synteny, may indicate functional genome regions.

Pyrococcus spp.[a]

[a] Zivanovic *et al.* (2002) *Nuc. Acids Res.* **30**:1902-1910 doi:10.1093/nar/30.9.1902

The James Hutton Institute

Comparison of *Pyrococcus* genomes (*P. horikoshii*, *P. abyssi*, *P. furiosus*) shows chromosome-shuffling.



Transposition a major cause of genomic disruption

Vibrio mimicus [a]

[a] Hasan *et al.* (2010) *Proc. Natl. Acad. Sci. USA* **107**:21134-21139 doi:10.1073/pnas.1013825107

The James Hutton Institute

Chromosome C-II carries genes associated with environmental adaptation; C-I carries virulence genes.
C-II has undergone extensive rearrangement; C-I has not.

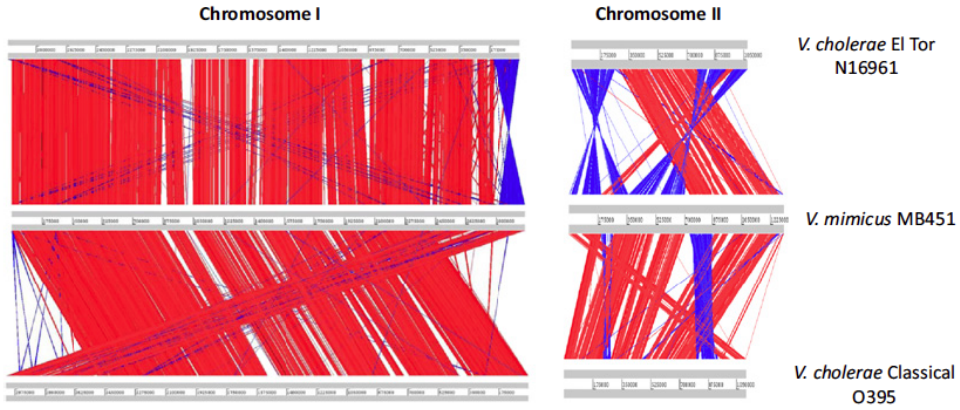


Fig. 2. Linear pairwise comparison of the *Vibrio mimicus* genome by Artemis Comparison Toll. Regions with similarity are highlighted by connecting red or blue lines between the genomes; red lines indicate homologous blocks of sequence, and blue lines indicate inversions. Gaps indicate unique DNA. The gray bars represent forward and reverse strands.

Suggests modularity of genome organisation, as a mechanism for adaptation (HGT, two-speed genome).

Serratia symbiotica [a]

[a] Burke and Moran (2011) *Genome Biol. Evol.* **3**:195-208 doi:10.1093/gbe/evr002

S. symbiotica is a recently evolved symbiont of aphids

Massive genomic decay is an adaptation to the new environment.

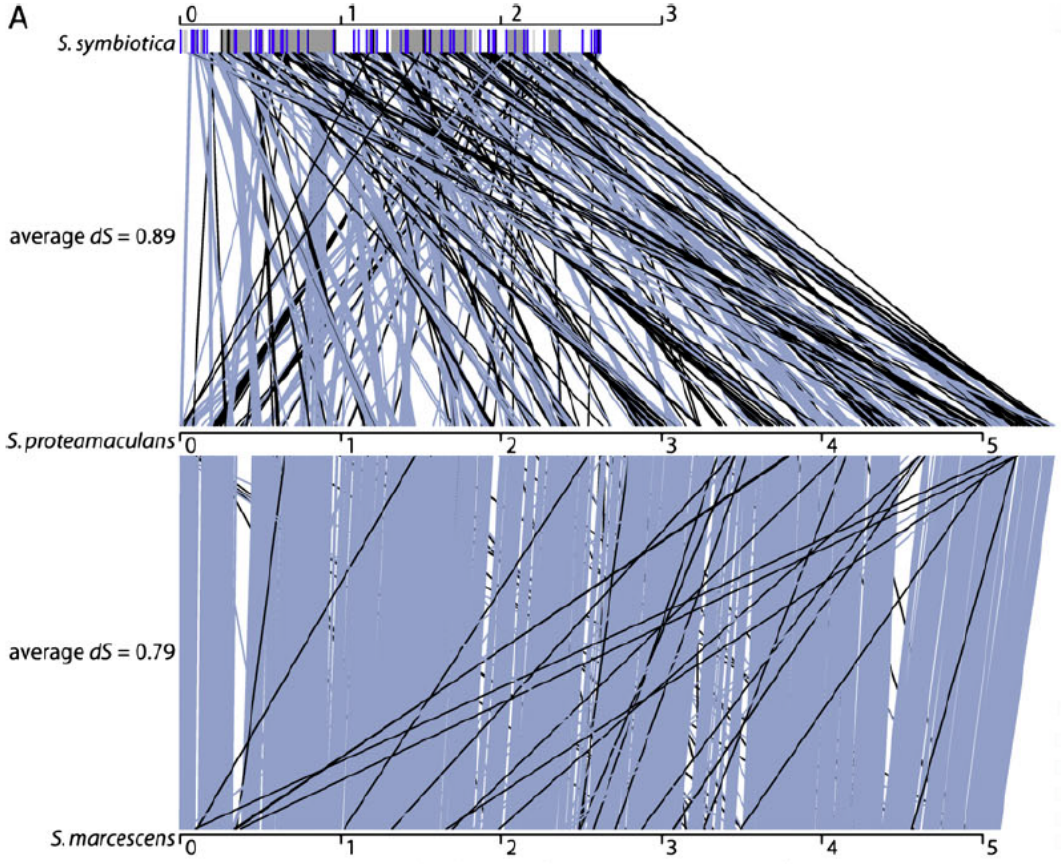


Table of Contents

The James Hutton Institute

Multiple genome alignment is hard

The James Hutton Institute

Can we not just align all our genomes, together?

No. Because it's really, really hard.

Analogous to problems with multiple sequence alignment (three or more sequences).

- Computationally extremely expensive ($O(L^n)$, L=length of sequence, n=number of sequences)
- NP-complete problem: no known efficient way to find a solution

Heuristic (approximate) methods are used, most commonly:

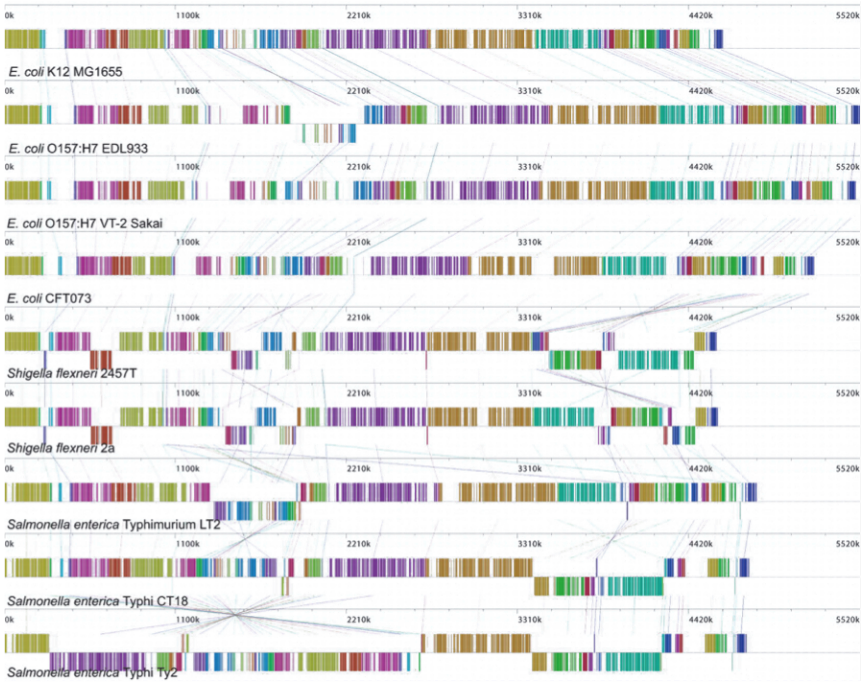
- Progressive alignment
- Iterative alignment

Mauve[a]

[a] Darling *et al.* (2004) *Genome Res.* **14**:1394-1403 doi:10.1101/gr.2289704

The James Hutton Institute

Progressive alignment tool, with a GUI. Application to nine enterobacteria: rearrangement of homologous backbone.



Alternatives include MLAGAN[41] and MUMmer[42]

[41] Brudno *et al.* (2003) *Genome Res.* **13**:721-731 doi:10.1101/gr.926603

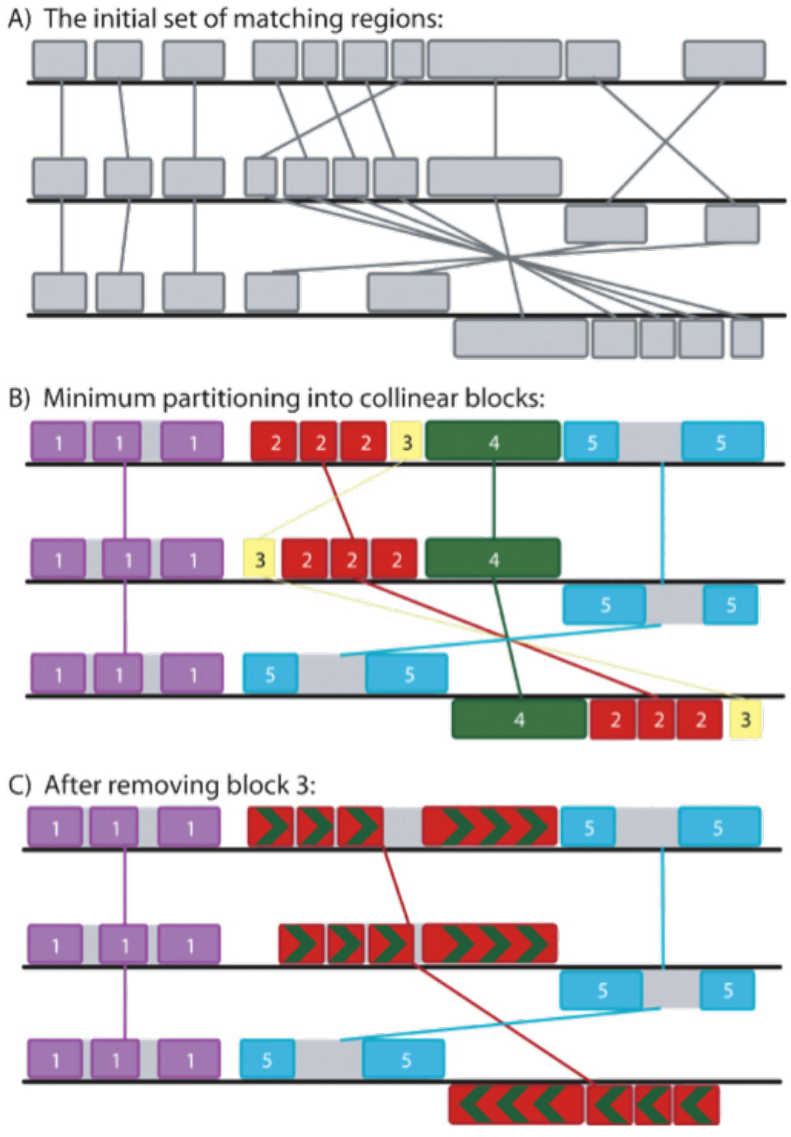
[42] Kurtz *et al.* (2004) *Genome Biol.* **5**:R12 doi:10.1186/gb-2004-5-2-r12

Mauve algorithm[a]

[a] Darling *et al.* (2004) *Genome Res.* **14**:1394-1403 doi:10.1101/gr.2289704

The James Hutton Institute

1. Find local alignments (*multi-MUMs*)
2. Build guide tree from multi-MUMs
3. Select subset of multi-MUMs as anchors, and partition into Local Collinear Blocks (LCBs): consistently ordered subsets
4. Progressive alignment against guide tree

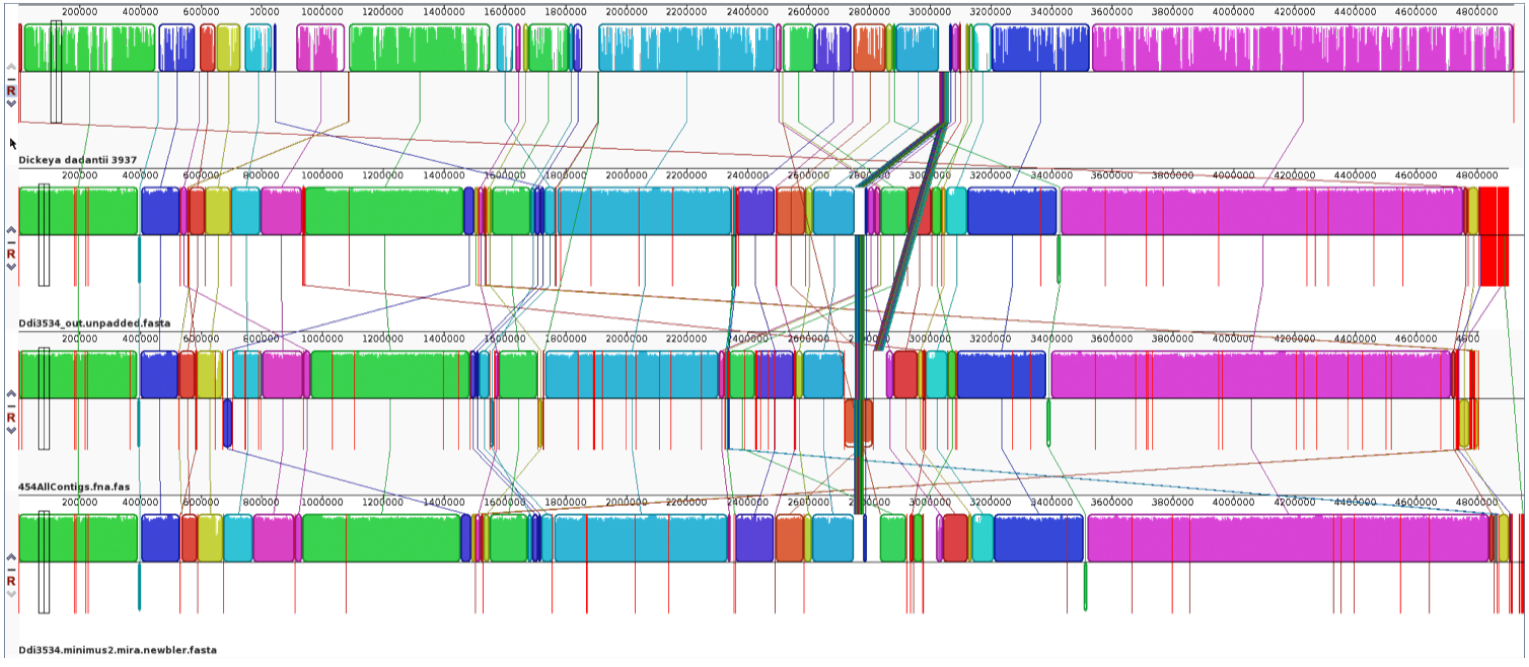


Reordering contigs[a]

[a] Darling *et al.* (2004) *Genome Res.* **14**:1394-1403 doi:10.1101/gr.2289704

The James Hutton Institute

Mauve also enables draft genome reordering.
Once LCBs are identified, can apply Mauve Contig Mover to reorder contigs



Example exercise in `data/whole_genome_alignment`

Table of Contents

Chromosome painting[a]

[a]Yahara *et al.* (2013) *Mol. Biol. Evol.* **30**:1454-1464 doi:10.1093/molbev/mst055

The James Hutton Institute

"Chromosome painting" infers recombination-derived 'chunks'
Genome's haplotype constructed in terms of recombination events from a 'donor' to a 'recipient' genome

FIG. 1. Chromosome painting in silico. Each lane indicates the chromosome of a strain shown on the right. The strains are classified by fineSTRUCTURE into subgroups labeled by colors (table 1 and fig. 2) on the left. A color along the chromosome indicates the subgroup that donated a chunk of SNPs through homologous recombination. All genomic positions are transformed to those of a reference strain (26695).

Chromosome painting[a]

[a] Yahara *et al.* (2013) *Mol. Biol. Evol.* **30**:1454-1464 doi:10.1093/molbev/mst055

Recombination events summarised in a *coancestry matrix*.
H. pylori: most within geographical bounds, but asymmetrical donation from Amerind/East Asian to European isolates.

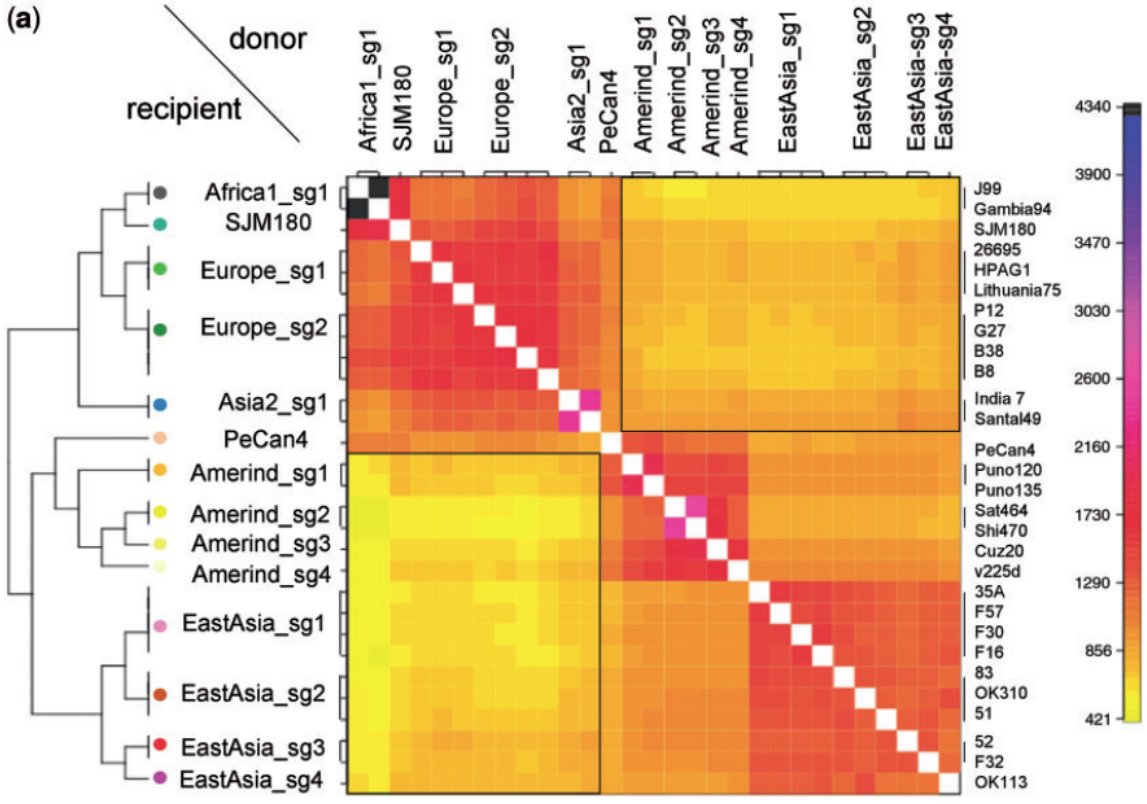


Table of Contents

P.aeruginosa nosocomial acquisition[a]

[a] Quick *et al.* (2014) *BMJ Open* **4**: e006278. doi:10.1136/bmjopen-2014-006278

The James Hutton Institute

Motivation

Nosocomial water transmission of *P.aeruginosa* an urgent concern

Setup

Burns patients (30) screened for *P.aeruginosa* on admission

Samples taken from patients and environment

All *P.aeruginosa* isolates (141) WGS sequenced

Outcome

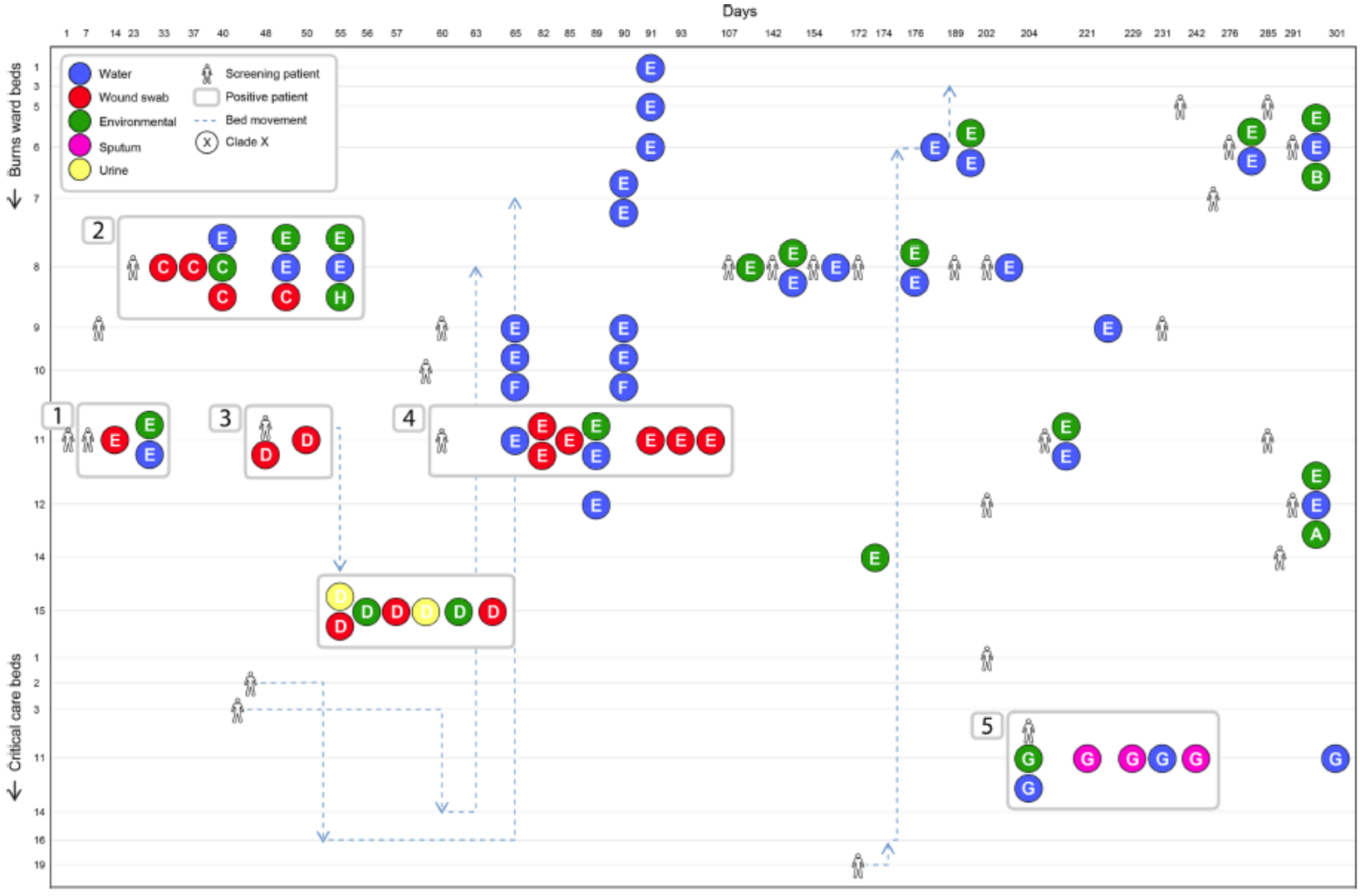
Clustering of isolates by room and outlet

Three patient isolates identical to water isolates from same room

Biofilm from thermostatic mixer valve a possible source

P.aeruginosa nosocomial acquisition[a]

[a]Quick *et al.* (2014) *BMJ Open* **4**: e006278. doi:10.1136/bmjopen-2014-006278



P.aeruginosa nosocomial acquisition[a]

[a] Quick *et al.* (2014) *BMJ Open* **4**: e006278. doi:10.1136/bmjopen-2014-006278

The James Hutton Institute

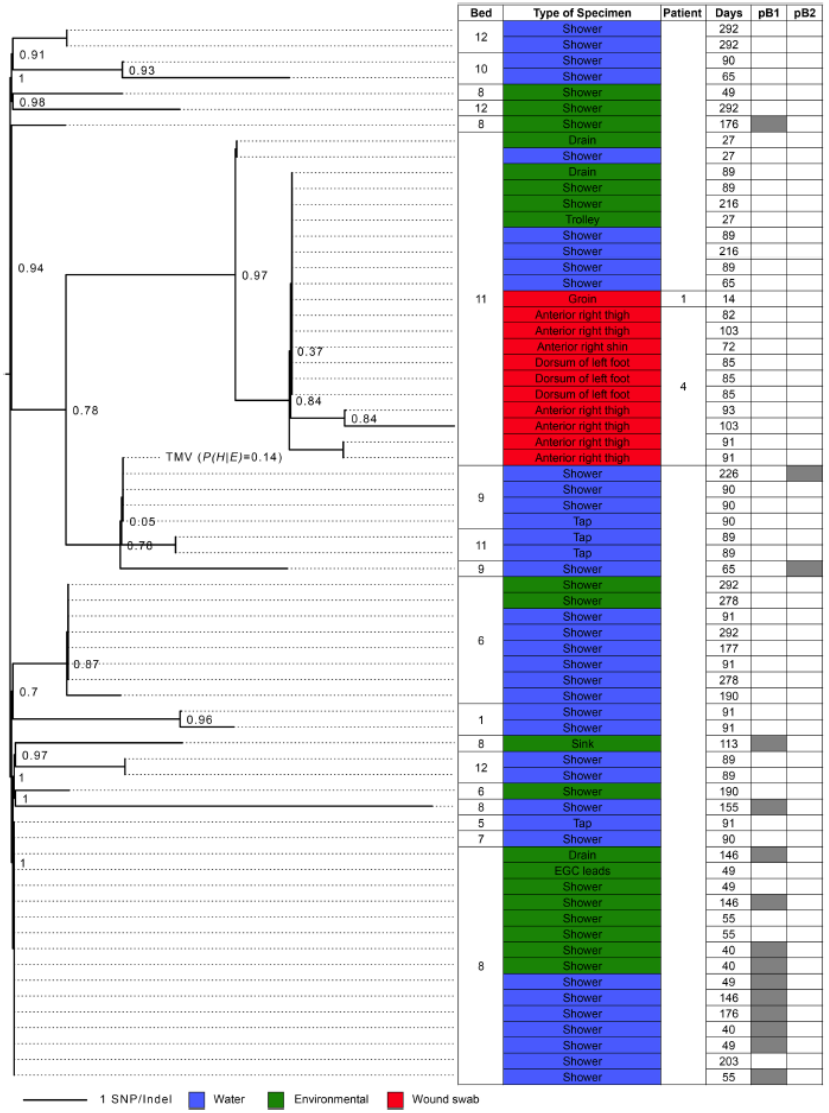
Methods

- **Illumina MiSeq** WGS of 141 isolates
- Metagenomic sequencing of biofilm
- **Simulated sequencing** of 55 published *P. aeruginosa*
- **BWA** mapping against **PAO1 reference genome**
- SNPs called with **SAMtools** & **VarScan**
- **ML reconstruction** with **FastTree**
- *De novo* assembly with **Velvet** for MLST prediction

Sequences and bioinformatic methods shared online:
http://www.github.com/joshquick/snp_calling_scripts

P.aeruginosa nosocomial acquisition[a]

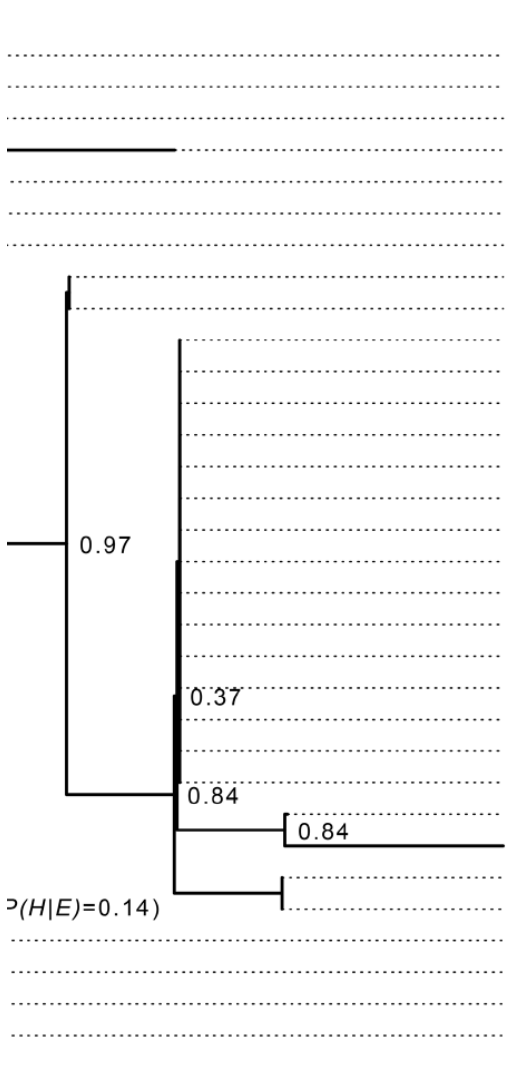
[a]Quick *et al.* (2014) *BMJ Open* **4**: e006278. doi:10.1136/bmjopen-2014-006278



Bed	Type of Specimen	Patient	Days	pB1	pB2
12	Shower		292		
	Shower		292		
10	Shower		90		
	Shower		65		
8	Shower		49		
12	Shower		292		
8	Shower		176		
	Drain		27		
	Shower		27		
	Drain		89		
	Shower		89		
	Shower		216		
	Trolley		27		
	Shower		89		
	Shower		216		
	Shower		89		
	Shower		65		
11	Groin	1	14		
	Anterior right thigh		82		
	Anterior right thigh		103		
	Anterior right shin		72		
	Dorsum of left foot		85		
	Dorsum of left foot	4	85		
	Dorsum of left foot		85		
	Anterior right thigh		93		
	Anterior right thigh		103		
	Anterior right thigh		91		
	Anterior right thigh		91		
	Shower		226		
9	Shower		90		
	Shower		90		
	Tap		90		
11	Tap		89		
	Tap		89		
9	Shower		65		
	Shower		292		
	Shower		278		
	Shower		91		
6	Shower		292		
	Shower		177		
	Shower		91		
	Shower		278		
	Shower		190		
1	Shower		91		
	Shower		91		
8	Sink		113		
12	Shower		89		
	Shower		89		
6	Shower		190		
8	Shower		155		
5	Tap		91		
7	Shower		90		
	Drain		146		
	EGC leads		49		
	Shower		49		
	Shower		146		
	Shower		55		
	Shower		55		
	Shower		40		
8	Shower		40		
	Shower		49		
	Shower		146		
	Shower		176		
	Shower		40		
	Shower		49		
	Shower		203		
	Shower		55		

P.aeruginosa nosocomial acquisition[a]

[a]Quick *et al.* (2014) *BMJ Open* **4**: e006278. doi:10.1136/bmjopen-2014-006278



Bed	Type of Specimen	Patient	Days
12	Shower		292
	Shower		292
10	Shower		90
	Shower		65
8	Shower		49
12	Shower		292
8	Shower		176
11	Drain		27
	Shower		27
	Drain		89
	Shower		89
	Shower		216
	Trolley		27
	Shower		89
	Shower		216
	Shower		89
	Shower		65
	Groin	1	14
	Anterior right thigh	4	82
	Anterior right thigh		103
	Anterior right shin		72
	Dorsum of left foot		85
	Dorsum of left foot		85
	Dorsum of left foot		85
	Anterior right thigh		93
	Anterior right thigh		103
	Anterior right thigh		91
	Anterior right thigh		91
9	Shower		226
	Shower		90
	Shower		90
	Tap		90

P.aeruginosa nosocomial acquisition[a]

[a] Quick *et al.* (2014) *BMJ Open* **4**: e006278. doi:10.1136/bmjopen-2014-006278

The James Hutton Institute

Strengths

A *P. aeruginosa* source could be tracked by WGS

Insights into transmission: water to patient a likely route

Sensitivity - identifies microevolution

Limitations

Small sample size: 5/30 patients infected, gave 55/141 isolates

Not clear that causal inferences are general

300-day sampling, not real-time crisis analysis

Good existing reference genome set for this bacterium

Sequencing cost: $\approx$£8k; Staff cost: $\approx$£15k

Microbial Genomics and Bioinformatics
BM405
4.Genome Features

The James Hutton Institute

Leighton Pritchard[1,2,3]
[1]Information and Computational Sciences,
[2]Centre for Human and Animal Pathogens in the Environment,
[3]Dundee Effector Consortium,
The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

Acceptable Use Policy

The James Hutton Institute

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available on SlideShare.

These slides, and supporting material including exercises, are available at https://github.com/widdowquinn/Teaching-Strathclyde-BM405

Table of Contents

Genome Features

The James Hutton Institute

- Genome features are annotated regions of the genome.
- Typically represent functional elements.
- May be simple (single region), or complex (subfeatures)

Why annotate genome features?

The James Hutton Institute

- Almost all use of genomics depends on annotation: **annotation quality is critical to downstream use of genomics in biology**
- Annotation is *curation* (a live, active process), not cataloguing
- Automated annotation from curated data (public databases) is the **only** game in town, given the data quantities we generate
- But you can't propagate something that doesn't exist: **up to 30% of metabolic activity has no known gene associated with it**[43]
- Biocurators can spend as much time "de-annotating" literature-based annotations as entering new data[44]

---

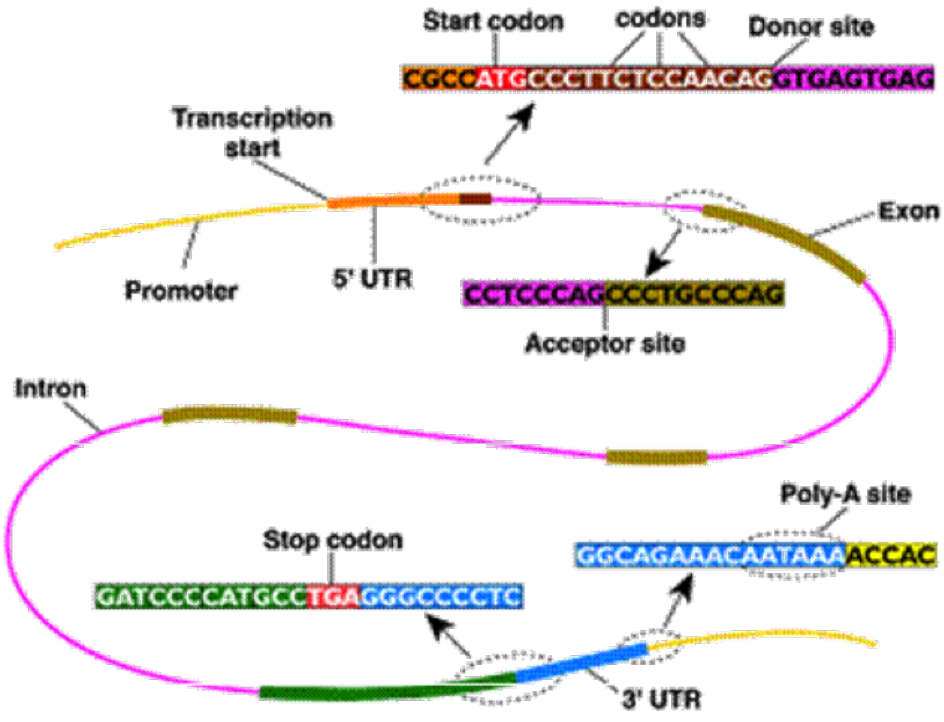
[43] Chen and Vitkup (2007) *Trends Biotech.* doi:10.1016/j.tibtech.2007.06.001

[44] Bairoch (2009) *Nat. Preced.* doi:10.1038/npre.2009.3092.1

Gene Features

Gene features have significant substructure, especially in eukaryotes.

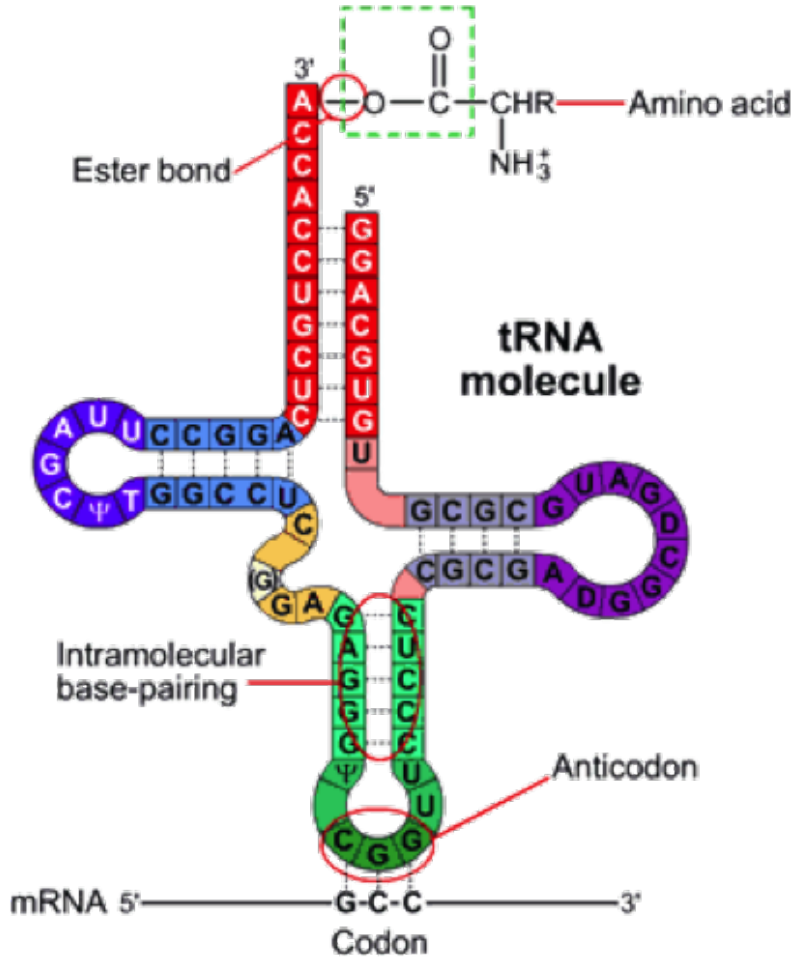
- 5‘ UTR
- translation start
- intron start/stop
- exon start/stop
- translation stop
- translation terminator
- 3‘ UTR



ncRNA Features

The James Hutton Institute

- tRNA - transfer RNA
- rRNA - ribosomal RNA
- CRISPRs - bacterial/archaeal defence (used for genome editing)
- many other classes

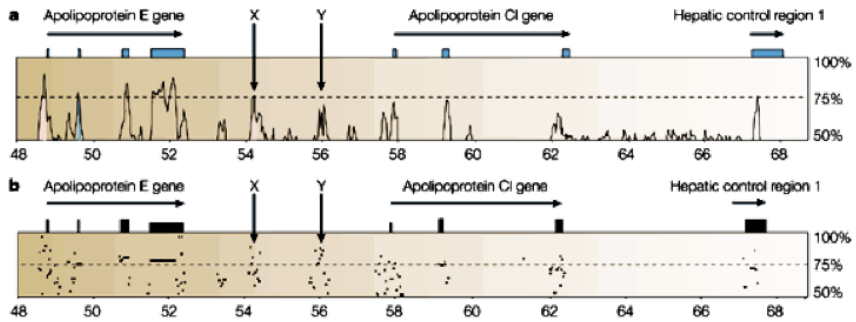


Regulatory/Repeat Features

The James Hutton Institute

Regulatory sites

- transcription start sites
- RNA polymerase binding sites
- Transcription Factor Binding Sites (TFBS)



Repetitive regions and mobile elements

- tandem repeats
- (retro-)transposable elements
- phage inclusions

Principles of feature prediction

The James Hutton Institute

Two main approaches to feature prediction:

- *ab initio* prediction - start from first principles, using only the genome sequence:
  - Unsupervised methods - not trained on a dataset
  - Supervised methods - trained on a dataset
- homology matches
  - alignment to features from related organisms (comparative genomics, annotation transfer)
  - from known gene products (e.g. proteins, ncRNA)
  - from transcripts/other intermediates (e.g. ESTs, cDNA, RNAseq)

Dedicated tools available for many different classes of feature.

Table of Contents

Prokaryotic CDS Prediction Methods

The James Hutton Institute

Using CDS prediction as an illustrative example for all feature prediction.

Sequence conservation (evolutionary constraint; an unsupervised, *a priori* method) can be useful

- Prokaryotes "easier" than eukaryotes for gene/CDS prediction
- Less uncertainty in predictions (isoforms, gene structure)
  - Very gene-dense (over 90% of chromosome is coding sequence)
  - No intron-exon structure

Prokaryotic CDS Prediction Methods

The James Hutton Institute

ORFs are plentiful:

- Problem is: "which possible ORF contains the true gene, and which start site is correct?"
- Still not a solved problem

Finding Open Reading Frames

The James Hutton Institute

The simplest approach: find ORFs (sequence between two consecutive in-frame stop codons)

- ORF finding is naive, does not consider:
  - Start codon
  - Promoter/RBS motifs
  - Wider context (e.g. overlapping genes)

Dedicated tools, e.g. `Glimmer`, `Prodigal`, `RAST`, `GeneMarkS` usually better.

Two ab initio CDS Prediction Tools

The James Hutton Institute

- `Glimmer`[45]
  - Interpolated Markov models
  - Can be trained on "gold standard" datasets
- `Prodigal`[46]
  - Log-likelihood model based on GC frame plots, followed by dynamic programming
  - Can be trained on "gold standard" datasets

Applying these to an example bacterial chromosome...

---

[45] Delcher *et al.* (2007) *Bioinformatics* **23**:673-679 doi:10.1093/bioinformatics/btm009

[46] Hyatt *et al.* (2010) *BMC Bioinf.* **11**:119 doi:10.1186/1471-2105-11-119

Comparing predictions in Artemis[a]

[a] Carver *et al.* (2012) *Bioinformatics* **28**:464-469 doi:10.1093/bioinformatics/btr703

The James Hutton Institute

Not every ORF (green) is predicted to encode for a coding sequence (CDS; blue/orange).
Self-contradictory CDS calls (orange); even automated annotation needs manual curation.

Comparing predictions in Artemis

The James Hutton Institute

`Glimmer`(green)/`Prodigal`(blue) CDS prediction methods do not always agree (presence/absence, start position).

How do we know which (if either) is best?

Table of Contents

Using a "Gold Standard": validation[a]

[a] Pritchard and Broadhurst (2014) *Methods Mol. Biol.* **1127**:53-64 doi:10.1007/978-1-62703-986-4_4

The James Hutton Institute

A general approach for *all* predictive methods

- Define a known, "correct" set of true/false, positive/negative etc. examples - the "gold standard"
- Evaluate your predictive method against that set for
  - sensitivity, specificity, accuracy, precision, etc.

This ought to be done by the method developers, but often wise to evaluate in your own system.

Many methods available, coverage beyond the scope of this introduction

Contingency Tables

		Condition (Gold standard)	
		True	False
Test outcome	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Performance Metrics

Sensitivity = TPR = $TP/(TP + FN)$

Specificity = TNR = $TN/(FP + TN)$

FPR = 1 − Specificity = $FP/(FP + TN)$

If you don't have this information, you can't interpret predictive results properly.

"Gold Standard" results

The James Hutton Institute

- Tested `glimmer`[47] and `prodigal`[48] on two enterobacterial close relatives as "gold standards" (still not perfect...)
  1. Manually annotated (>3 expert person years)
  2. Community-annotated (many research groups, interested in their own subset of genes)
- **Both methods trained directly on the annotated genes in each organism!**

[47] Delcher *et al.* (2007) *Bioinformatics* **23**:673-679 doi:10.1093/bioinformatics/btm009

[48] Hyatt *et al.* (2010) *BMC Bioinf.* **11**:119 doi:10.1186/1471-2105-11-119

"Gold Standard" results

The James Hutton Institute

Manually annotated: 4550 CDS

genecaller	`glimmer`	`prodigal`
predicted	4752	4287
missed	**284** (6%)	407 (9%)
Exact Prediction		
sensitivity	62%	**71**%
FDR	41%	**25**%
PPV	59%	**75**%
Correct ORF		
sensitivity	**94**%	91%
FDR	10%	**3**%
PPV	90%	**97**%

"Gold Standard" results

The James Hutton Institute

Community annotated: 4475 CDS

genecaller	`glimmer`	`prodigal`
predicted	4679	4467
missed	**112** (3%)	156 (3%)
Exact Prediction		
sensitivity	62%	**86**%
FDR	31%	**14**%
PPV	69%	**86**%
Correct ORF		
sensitivity	**97**%	**97**%
FDR	7%	**3**%
PPV	93%	**97**%

Gene/CDS Prediction

The James Hutton Institute

- Alternative CDS (and all other) prediction methods are unlikely to give identical results, or perform equally well
- **There is No Free Lunch** (this is a theorem: http://en.wikipedia.org/wiki/No_free_lunch_theorem)
- To assess/choose between methods, performance metrics are required
- Even on prokaryotes (a relatively simple case), current best methods for CDS prediction are imperfect
- Manual correction is often required (usually the most demanding and time-consuming part of the process).

Table of Contents

Prokaryotic Annotation Pipelines[a]

[a] Richardson and Watson (2012) *Brief. Bioinf.* **14**:1-12 doi:10.1093/bib/bbs007

The James Hutton Institute

Many choices, including `RAST`[49], `PROKKA`[50], `BaSYS`[51], etc.
Often perform both CDS/feature calling and functional prediction.
Two broad approaches:

1. Heavyweight: maintain database and resource, often annotating by homology, e.g. `RAST`
2. Lightweight: chain together multiple third-party packages, e.g. `PROKKA`

Pipelines take a lot of tedium (and control) out of annotating bacterial genomes, but have the same issues as every other prediction tool.

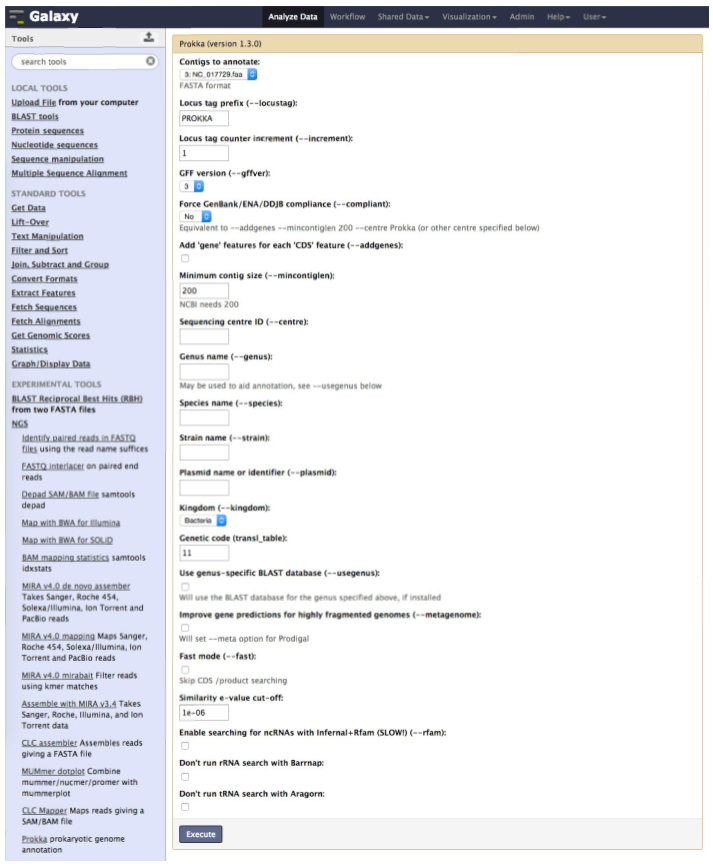
[49] Aziz *et al.* (2008) *BMC Genomics* **9**:75 doi:10.1186/1471-2164-9-75
[50] Seemann (2014) *Bioinformatics* **30**:2068-2069 doi:10.1093/bioinformatics/btu153
[51] Van Domselaar *et al.* (2005) *Nuc. Acids Res.* **33**:W455-W459 doi:10.1093/nar/gki593

PROKKA[a]

[a] Seemann (2014) *Bioinformatics* **30**:2068-2069 doi:10.1093/bioinformatics/btu153

The James Hutton Institute

- Lightweight, and fast.
- Runs locally. (5Mbp genome takes ≈10min on my desktop; more detailed ncRNA prediction takes ≈20min)
- Flexible: built-in databases can be replaced by user databases.
- Uses freely-accessible third-party tools for prediction



Simple to run (at the command-line, or in `Galaxy`[52]).

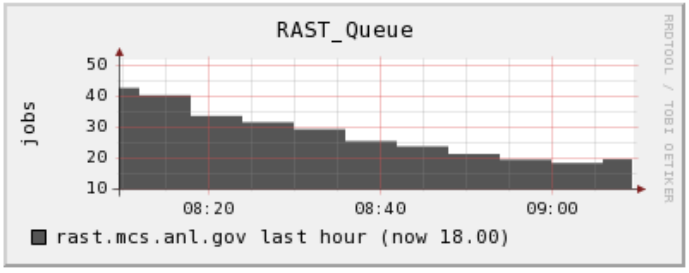
[52] Goecks *et al.* (2010) *Genome Biol.* **11**:R86 doi:10.1186/gb-2010-11-8-r86

RAST[a]

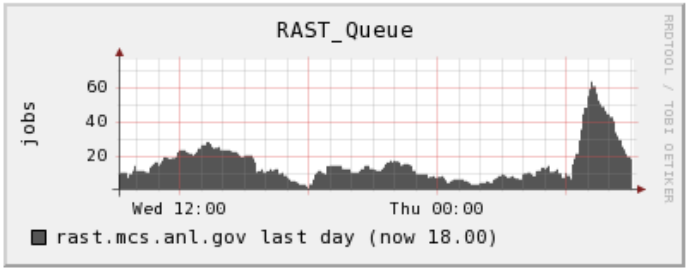
[a] Aziz *et al.* (2008) *BMC Genomics* **9**:75 doi:10.1186/1471-2164-9-75

- Server-based (http://rast.nmpdr.org/). Queues likely.
- Relies on SEED and FIGFam databases, held at NMPDR
- **FIGFam**: isofunctional homologue families
- **Produces metabolic reconstruction**

RAST Hourly Queue



RAST Daily Queue



RAST Weekly Queue

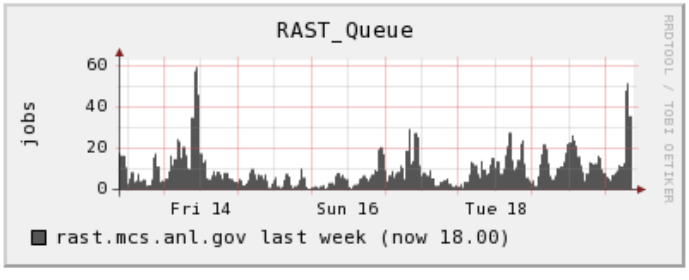


Table of Contents

Principles of function prediction

The James Hutton Institute

At genome scale, we realistically have to automate function prediction.
Function prediction is just like any other prediction method.
"Does this sequence imply that function?"

Two main approaches to function prediction:

- *ab initio* prediction (on basis of feature sequence/context only)
  - Unsupervised methods - not trained on an exemplar dataset
  - Supervised methods - trained on an exemplar dataset
- homology matches (sequence similarity)
  - alignment to features with known/predicted functions

Homology-based function prediction

The James Hutton Institute

Two proteins with similar sequence may have similar function.
But...

- How similar do they have to be (and where) to share the same function?
- What do we mean by 'same function', anyway? Interaction/substrate specificity? Participation in a pathway? Contribution to a structure? Biochemical interconversion? ...
- How confident can we be in the comparator (annotated) sequence: was *that* function determined experimentally?

Gene Ontology (GO)[a]

[a] Ashburner *et al.* (2000) *Nat. Genet.* **25**:25-29 doi:10.1038/75556

The James Hutton Institute

The Gene Ontology provides a common vocabulary for describing biological function, and unifying functional descriptions.

Ontologies (controlled vocabularies) are central to information-sharing.

Gene Ontology Consortium: http://geneontology.org/

Many annotation tools and databases produce GO output, or compatible controlled vocabulary terms, e.g.

- Blast2GO[53]: BLAST-based annotation
- PHI-Base[54]: microbial pathogen-host interaction specific functions
- GOPred[55]: combines several protein function classifiers

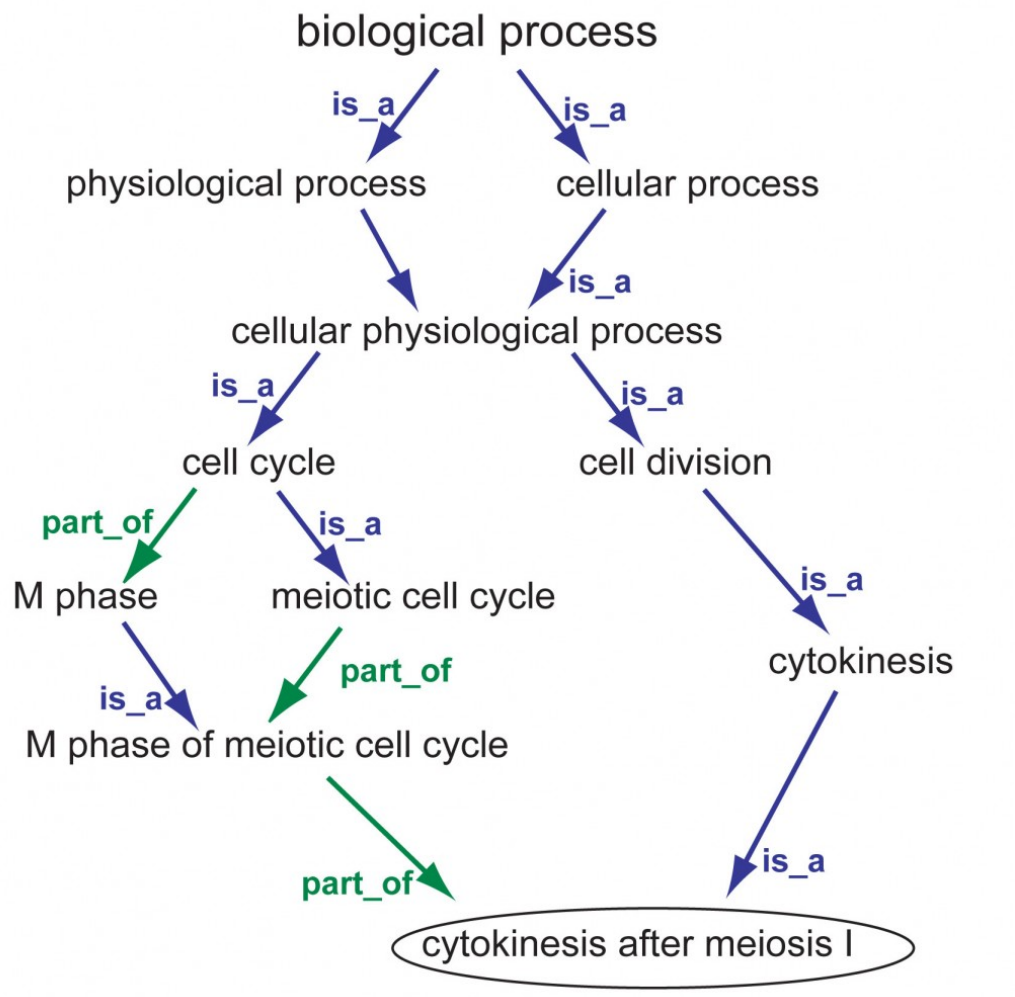
[53] Conesa *et al.* (2005) *Bioinformatics* **21**:3674-3676 doi:10.1093/bioinformatics/bti610

[54] Winnenburg *et al.* (2006) *Nuc. Acids Res.* **34**:D459-D464 doi:10.1093/nar/gkj047

[55] Sarac *et al.* (2010) *PLoS One* **5**:e12382 doi:10.1371/journal.pone.0012382

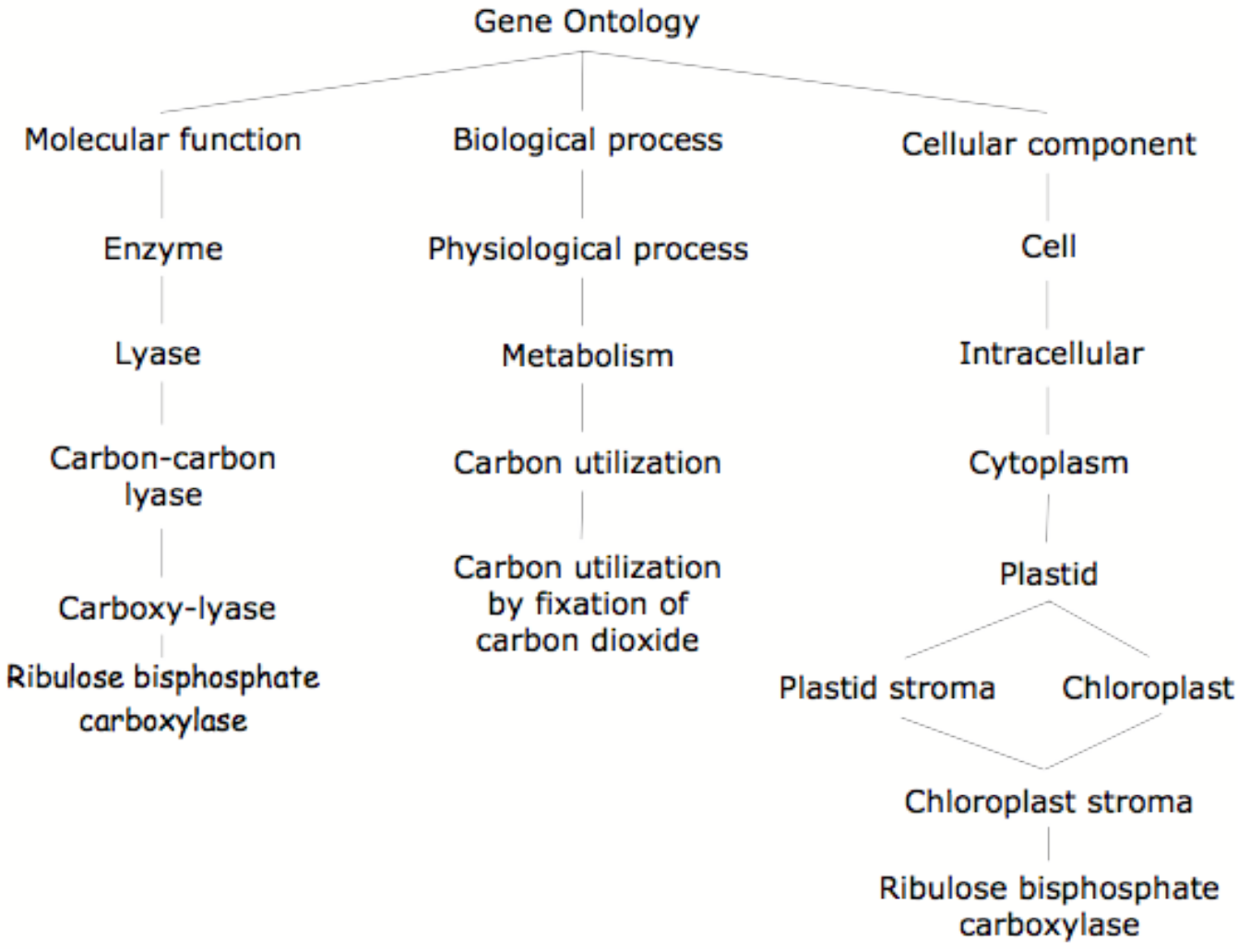
Gene Ontology (GO)[a]

[a] Ashburner *et al.* (2000) *Nat. Genet.* **25**:25-29 doi:10.1038/75556



Gene Ontology (GO)[a]

[a] Ashburner *et al.* (2000) *Nat. Genet.* **25**:25-29 doi:10.1038/75556



Are database annotations reliable?[a]

[a] Schnoes *et al.* (2013) *PLoS Comp. Biol.* **9**:e1003063 doi:10.1371/journal.pcbi.1003063

The James Hutton Institute

Are protein function annotations in databases determined experimentally, or by annotation transfer?

High throughput experiments and genome annotations are conducted without validation of function, and placed in databases.

- GO databases record annotation origin by publication
- GO databases record evidence codes, e.g.: **EXP**=Inferred from Experiment; **ISS**=Inferred from Sequence Similarity
- 0.14% of contributing publications provide 25% of all experimentally validated annotations in the Uniprot-GOA compilation.
- There are biases in functional annotation.

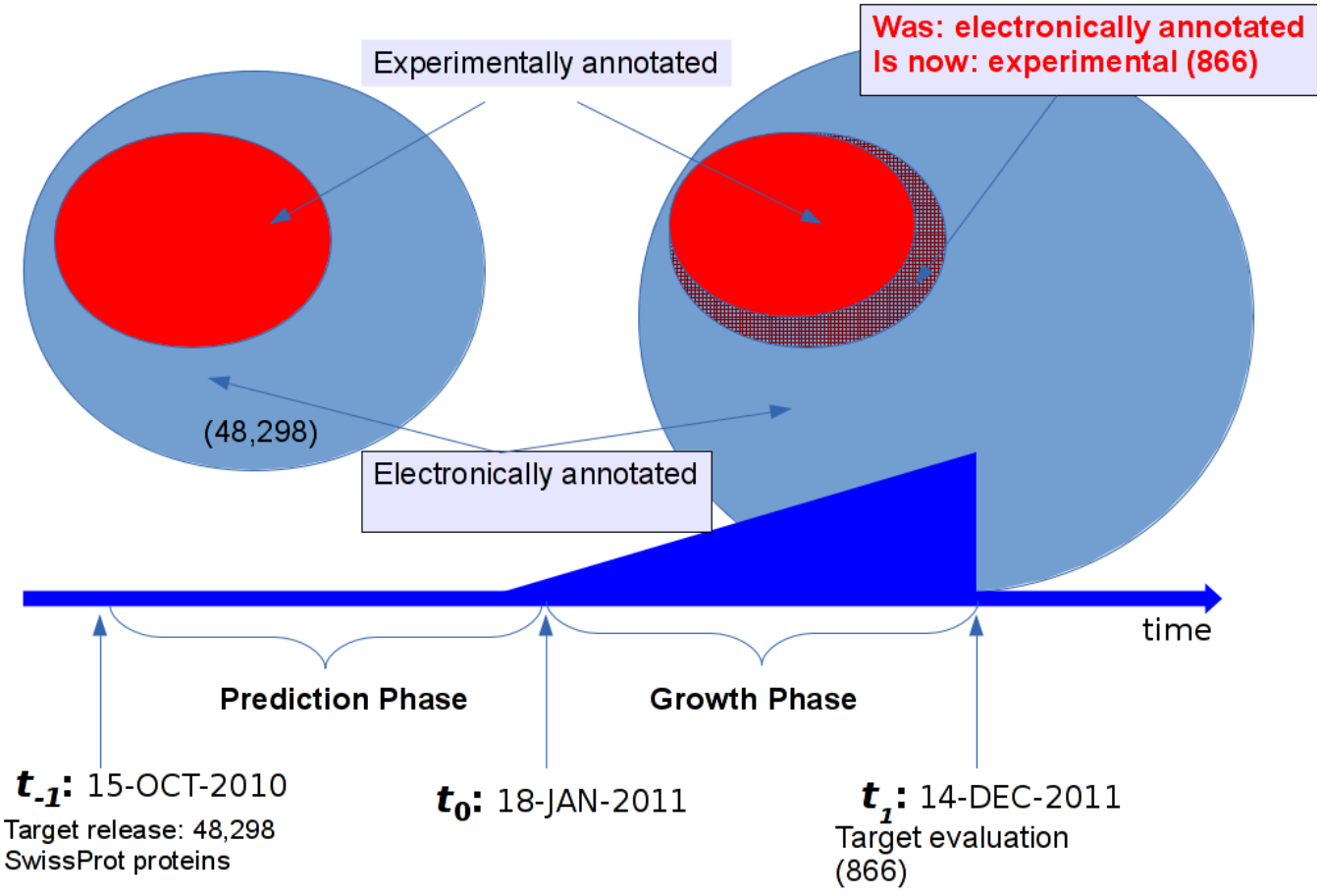
No clear solution to this kind of bias - **but we have to recognise and account for it**.

Are database annotations reliable?[a]

[a] Radivojac *et al.* (2013) *Nat. Meth.* **10**:221-227 doi:10.1038/nmeth.2340

The James Hutton Institute

The Critical Assessment of Function Annotation (CAFA) project.



Do biased database annotations matter?

The James Hutton Institute

Experimental annotations of proteins are incomplete. But is that important?
Tested by simulation, and following databases for three years.[56]

1. Yes. It matters.
2. Current large scale annotations are meaningful and almost surprisingly reliable.
3. The nature and level of data incompleteness, and type of classification model have an effect.
4. "Low precision, high recall" (i.e. less discriminating) tools most significantly affected.

Molecular function prediction is usually more reliable than biological process prediction[57]

---

[56] Jiang *et al.* (2014) *Bioinformatics* **30**:i609-i616 doi:10.1093/bioinformatics/btu472

[57] Cozzetto *et al.* (2013) *BMC Bioinf.* **14**:S3-S1 doi:10.1186/1471-2105-14-S3-S1

CAFA results[a]

[a] Radivojac *et al.* (2013) *Nat. Meth.* **10**:221-227 doi:10.1038/nmeth.2340

The James Hutton Institute

The Critical Assessment of Function Annotation (CAFA) 2013 results. (F-measure combines precision and recall)

- You **can** do better than BLAST.
- Best-performing methods do comparably well.
- Best methods used evolutionary relationships, structure, and expression data.
- Machine Learning works best.

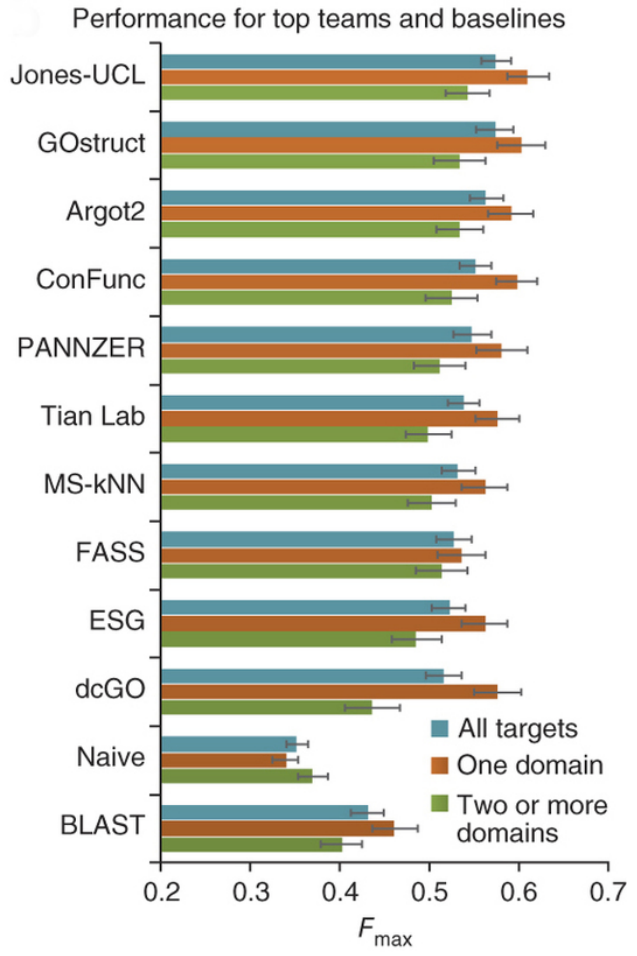


Table of Contents

A wee trip to the doctor

The James Hutton Institute

- You go for a checkup, and are tested for disease X
- The test has **sensitivity** $= 0.95$ (predicts disease where there is disease)
- The test has **FPR** $= 0.01$ (predicts disease where there is no disease)

A wee trip to the doctor

The James Hutton Institute

- You go for a checkup, and are tested for disease X
- The test has **sensitivity** $= 0.95$ (predicts disease where there is disease)
- The test has **FPR** $= 0.01$ (predicts disease where there is no disease)
- Your test is *positive*
- **What is the probability that you have disease X?**
  - **0.01, 0.05, 0.50, 0.95, 0.99?**
- (Audience Participation!)

A wee trip to the doctor

The James Hutton Institute

- What is the probability that you have disease X?
- **Unless you know the baseline occurrence of disease X, you cannot determine this.**

A wee trip to the doctor

The James Hutton Institute

- What is the probability that you have disease X?
- **Unless you know the baseline occurrence of disease X, you cannot determine this.**
- Baseline occurrence: f_X
  - $f_X = 0.01 \implies P(\text{disease}|+\text{ve}) = 0.490 \approx 0.5$
  - $f_X = 0.8 \implies P(\text{disease}|+\text{ve}) = 0.997 \approx 1.0$

Table of Contents

Why Performance Metrics Matter[a]

[a] Pritchard and Broadhurst (2014) *Methods Mol. Biol.* **1127**:53-64 doi:10.1007/978-1-62703-986-4_4

The James Hutton Institute

- Imagine a paper describing a predictor for protein functional class (e.g. Type III effector)
- The paper reports **sensitivity** $= 0.95$, **FPR** $= 0.01$
- You run the predictor on 4,500 proteins in a new genome
- It predicts 50 members of the class. How many of them are likely to be true positives?

Why Performance Metrics Matter[a]

[a] Pritchard and Broadhurst (2014) *Methods Mol. Biol.* **1127**:53-64 doi:10.1007/978-1-62703-986-4_4

The James Hutton Institute

- Imagine a paper describing a predictor for protein functional class (e.g. Type III effector)
- The paper reports **sensitivity** $= 0.95$, **FPR** $= 0.01$
- You run the predictor on 4,500 proteins in a new genome
- It predicts 50 members of the class. How many of them are likely to be true positives?
- *We need a baseline level of that class (f_X) in the genome to determine this.*
- We estimate ≈ 45 members in protein complement, so $f_X = 0.01$
  - $f_X = 0.01 \implies P(\text{class}|+\text{ve}) = 0.490 \approx 0.5$

Bayes' Theorem

The James Hutton Institute

- May seem counter-intuitive: 95% sensitivity, 99% specificity $\Longrightarrow$ **50% chance** of any prediction being incorrect
- Probability given by Bayes' Theorem
  - $P(X|+) = \frac{P(+|X)P(X)}{P(+|X)P(X)+P(+|\bar{X})P(\bar{X})}$

Let's play a game...

The James Hutton Institute

2, 4, 6, ...

Bayes' Theorem

- May seem counter-intuitive: 95% sensitivity, 99% specificity $\implies$ **50% chance** of any prediction being incorrect
- Probability given by Bayes' Theorem
  - $P(X|+) = \frac{P(+|X)P(X)}{P(+|X)P(X)+P(+|\bar{X})P(\bar{X})}$
- This step commonly overlooked in the literature
  - confirmation bias
  - people want to see positive examples/tell a story
  - people want to think their predictor works

A cautionary tale[a]

[a] Arnold *et al.* (2009) *PLoS Pathog.* **5**:e1000376 doi:10.1371/journal.ppat.1000376

The James Hutton Institute

- Paper describes `EffectiveT3`, a type III effector prediction tool
- Reported **sensitivity** $\approx$ **0.71, FPR** $\approx$ **0.15**
- Applied tool to 739 complete bacterial and archaeal genomes

A cautionary tale[a]

[a] Arnold *et al.* (2009) *PLoS Pathog.* **5**:e1000376 doi:10.1371/journal.ppat.1000376

The James Hutton Institute

- Paper describes `EffectiveT3`, a type III effector prediction tool
- Reported **sensitivity $\approx$ 0.71, FPR $\approx$ 0.15**
- Applied tool to 739 complete bacterial and archaeal genomes
- Organisms with an identifiable T3SS: 2-7% of genome predicted to be secreted
- **Organisms without an identifiable T3SS (or known not to have one): 1-10% of genome predicted to be secreted**
- "*The surprisingly high number of (false) positives in genomes without T3SS exceeds the expected false positive rate*"
- This is not a surprise, statistically.

A cautionary tale[a]

[a] Arnold *et al.* (2009) *PLoS Pathog.* **5**:e1000376 doi:10.1371/journal.ppat.1000376

The James Hutton Institute

Probability that an `EffectiveT3` positive prediction corresponds to a secreted protein is given by Bayes' Theorem

- $P(X|+) = \frac{P(+|X)P(X)}{P(+|X)P(X)+P(+|\bar{X})P(\bar{X})}$
  - $P(+|X)$ = sensitivity = 0.71
  - $P(+|\bar{X})$ = FPR = 0.15
  - $P(X)$ = base rate $\approx$ 0.03 [58]
  - $\implies P(X|+) \approx 0.13$

Only 13% of predictions likely to be positive!

How many predicted type III secreted proteins were there...

[58] Boch and Bonas (2010) *Annu. Rev. Phytopathol.* **48**:419-436 doi:10.1146/annurev-phyto-080508-081936

A cautionary tale[a]

[a] Arnold *et al.* (2009) *PLoS Pathog.* **5**:e1000376 doi:10.1371/journal.ppat.1000376

The James Hutton Institute

TTSS status	G+C content	Number of proteins	Positives	Z-Score
-	56.3%	1841	3.7%	0.6
-	43.1%	1963	1.6%	-2.3
-	[illegible]	[illegible]	[illegible]	8.3
-	[illegible]	[illegible]	[illegible]	6.8
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
+	57.9%	5169	3.8%	7.2
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
+	58.3%	5607	3.7%	[illegible]
-	42.8%	2120	10.8%	1[illegible]
-	44.8%	2385	12.2%	2[illegible]
-	40.1%	3541	3.3%	[illegible]
-	45.8%	4008	3.5%	[illegible]
+	51.4%	4510	2.8%	[illegible]
+	52.0%	4614	2.7%	[illegible]
+	52.1%	4659	2.8%	[illegible]
+	52.1%	4617	2.8%	[illegible]
+	52.2%	4205	2.8%	[illegible]
+	52.2%	3964	2.6%	[illegible]
+	52.1%	4779	2.5%	[illegible]
+	52.2%	4805	2.6%	[illegible]
+	52.2%	4091	2.7%	[illegible]
+	52.1%	5601	3.1%	[illegible]
+	52.2%	4627	2.8%	[illegible]
+	51.9%	4753	2.9%	1.9
+	52.1%	4312	2.7%	0.8
+	52.2%	4523	2.8%	1.8
-	53.6%	3645	3.0%	5.0
+	46.2%	4489	4.7%	7.6
-	46.3%	4394	4.7%	7.2
+	46.2%	4687	4.8%	7.7
-	46.3%	4440	4.8%	7.3
-	45.1%	3754	4.9%	5.0

196 Type III effectors?

TTSS status	G+C content	Number of proteins	Positives	Z-Score
-	58.4%	4777	3.3%	-1.0
-	66.9%	2157	3.6%	1.1
-	62.4%	4587	7.7%	18.9
-	65.4%	4506	7.2%	20.7
-	59.2%	1631	6.1%	6.9
-	60.2%	1998	6.2%	8.8
-	60.1%	1728	6.4%	6.2
-	59.9%	2416	5.3%	7.3
-	72.5%	3079	5.0%	12.6
-	72.4%	2940	4.0%	8.0
-	53.5%	2272	5.8%	8.5
-	63.1%	2950	9.3%	19.4
-	53.8%	3056	5.9%	9.5
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
-	54.1%	3076	6.4%	11.9
-	61.4%	2119	10.3%	21.1

218 Type III effectors, no T3SS?

0.038 x 5169 x 0.13 ≈ 26

[No. +ve x *P*(T3E|+ve)]

1-2% (Collmer *et al.* 2002); 1% (Boch and Bonas, 2010)

Interpreting genome-scale predictions[a]

[a] Pritchard and Broadhurst (2014) *Methods Mol. Biol.* **1127**:53-64 doi:10.1007/978-1-62703-986-4_4

The James Hutton Institute

- Statistics at genome-scale can be counterintuitive.
- **Use Bayes' Theorem!**
- Predictions identify groups, not individual members of the group. e.g.
  - Test for airport smugglers has $P(\text{smuggler}|+) = 0.9$
  - Test gives 100 positives
- Which specific individuals are truly smugglers?

Interpreting genome-scale predictions[a]

[a] Pritchard and Broadhurst (2014) *Methods Mol. Biol.* **1127**:53-64 doi:10.1007/978-1-62703-986-4_4

The James Hutton Institute

- Statistics at genome-scale can be counterintuitive.
- **Use Bayes' Theorem!**
- Predictions identify groups, not individual members of the group. e.g.
  - Test for airport smugglers has $P(\text{smuggler}|+) = 0.9$
  - Test gives 100 positives
- Which specific individuals are truly smugglers?
- The test *does not* allow you to determine this - you need more evidence for each individual
- Same principle applies to other classifiers, (including protein functional class prediction) - watch for 'cherry-picking' in publications

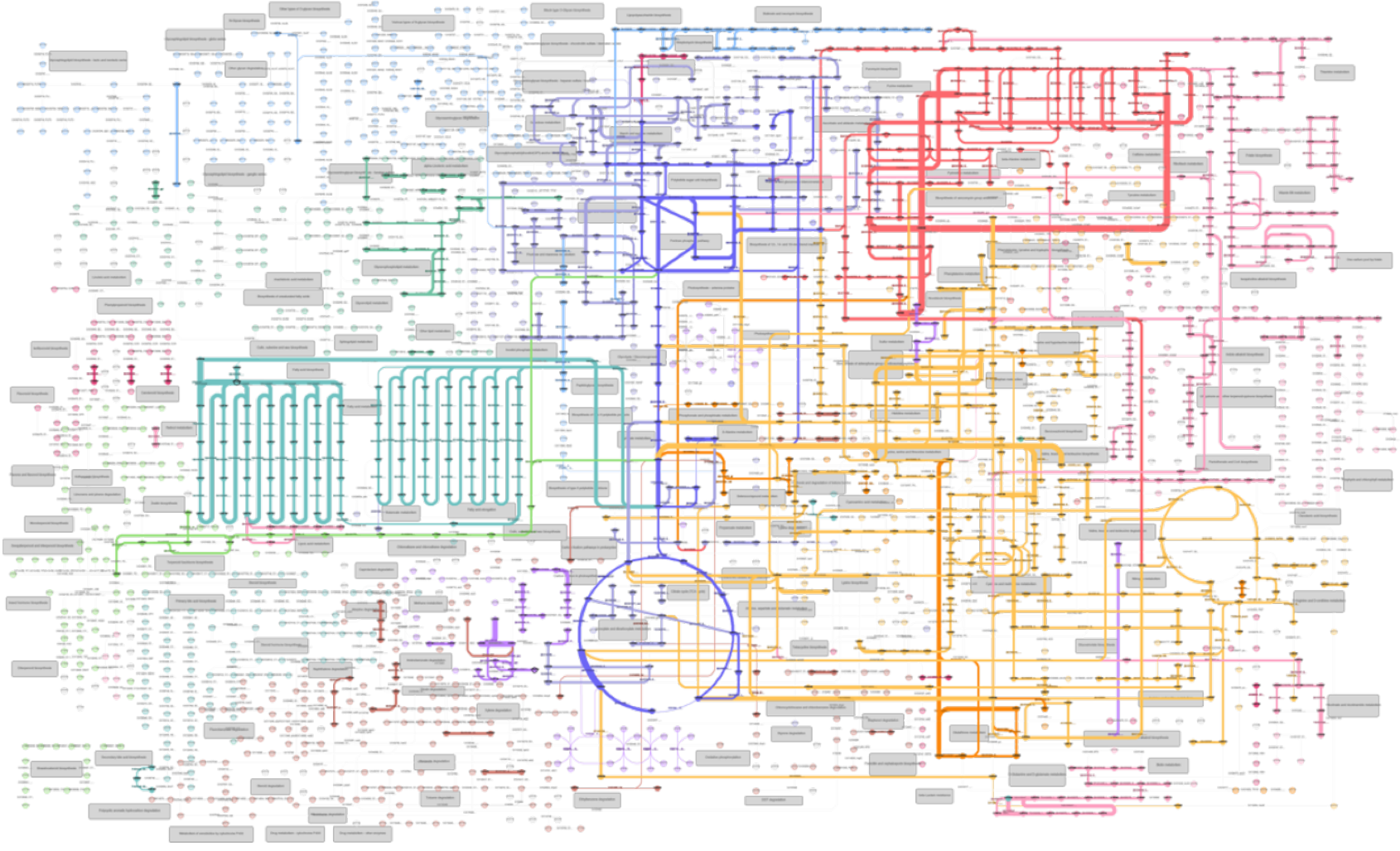
Table of Contents

Reconstructing metabolism[a]

[a] Thiele and Palsson (2010) *Nat. Protoc.* **5**:93-121 doi:10.1038/nprot.2009.203

The James Hutton Institute

Once metabolic functional annotation has been assigned to features, we can do comparative analysis of metabolism.



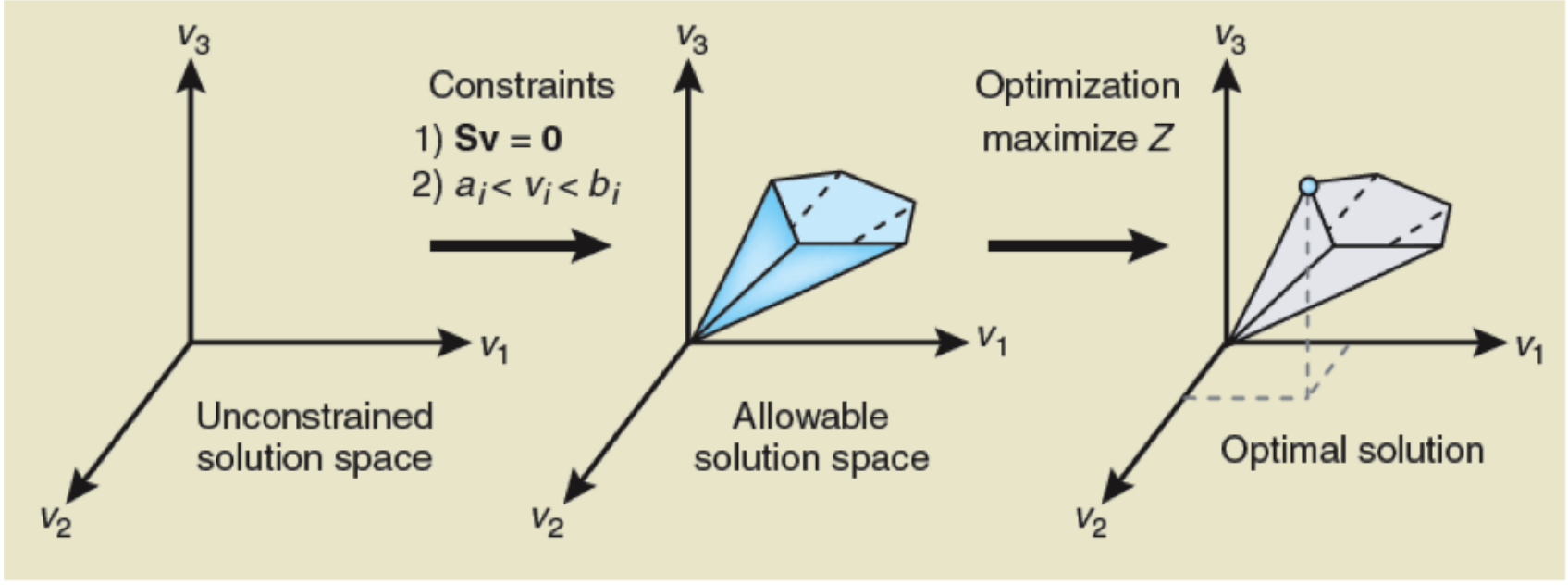
Dynamic models of metabolism[a]

[a] Orth *et al.* (2010) *Nat. Biotech.* **28**:245-248 doi:10.1038/nbt.1614

The James Hutton Institute

By using constraint-based models (e.g. Flux Balance Analysis), we can make these into dynamic representations of bacterial metabolism.

- Upper, lower bounds to reaction rates
- Define objective phenotype
- Calculate conditions resulting in flux
- *in silico* knockouts



E. coli metabolism[a]

[a] Monk *et al.* (2013) *Proc. Natl. Acad. Sci. USA* **110**:20338-20343 doi:10.1073/pnas.1307797110

The James Hutton Institute

E. coli has a very long history of metabolic reconstruction[59]
Recent modelling work predicts which nutrients support growth

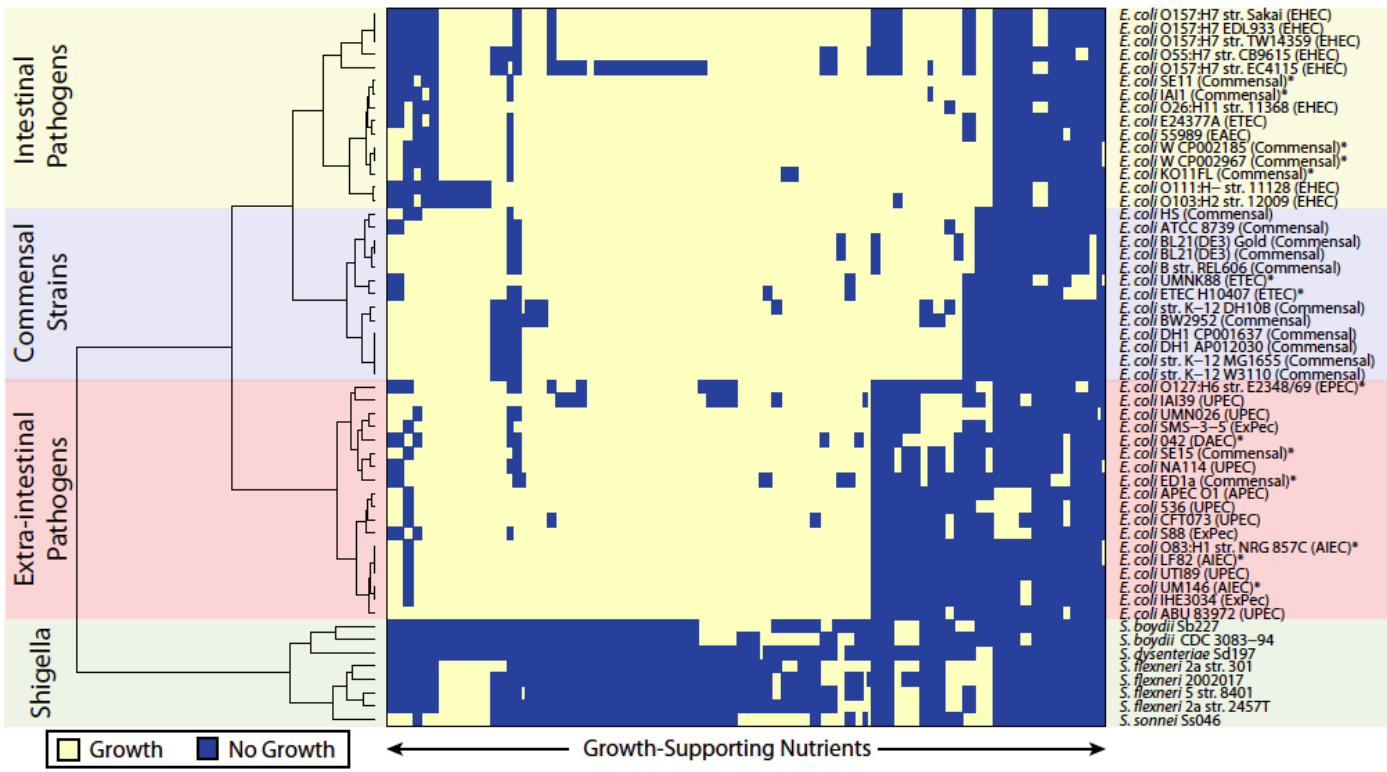


Fig. 2. Clustering of species by unique growth-supporting conditions. Predicted metabolic phenotypes on the variable growth-supporting nutrient conditions composed of different carbon, nitrogen, phosphorous, and sulfur nutrient sources in aerobic and anaerobic conditions. Strains are clustered based on their ability to sustain growth in each different environment. Rows represent individual strains, and columns represent different nutrient conditions. In general, strains clustered into their respective pathotypes of commensal *E. coli* strains, intestinal pathogenic *E. coli* strains, extra-intestinal pathogenic *E. coli* strains, and *Shigella* strains. An asterix symbol indicates those strains that clustered outside of their respective pathotype. All growth conditions are listed in Dataset S1.

[59] Reed and Palsson (2000) *J. Bact.* **185**:2692-2699 doi:10.1128/JB.185.9.2692-2699.2003

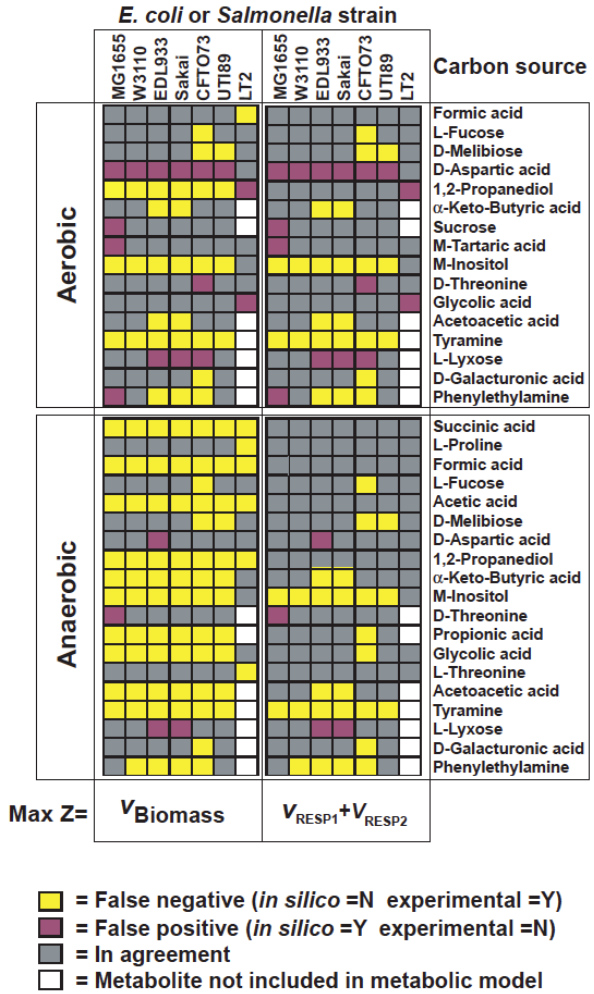
E. coli metabolism[a]

[a] Baumler *et al.* (2011) *BMC Syst. Biol.* **5**:182 doi:10.1186/1752-0509-5-182

The James Hutton Institute

Models are complex, and experimental validation is essential

There's more we don't know...



Microbial Genomics and Bioinformatics
BM405
5.Finding Equivalent Features

The James Hutton Institute

Leighton Pritchard[1,2,3]
[1]Information and Computational Sciences,
[2]Centre for Human and Animal Pathogens in the Environment,
[3]Dundee Effector Consortium,
The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

Acceptable Use Policy

The James Hutton Institute

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available on SlideShare.

These slides, and supporting material including exercises, are available at https://github.com/widdowquinn/Teaching-Strathclyde-BM405

Table of Contents

What makes genome features equivalent?

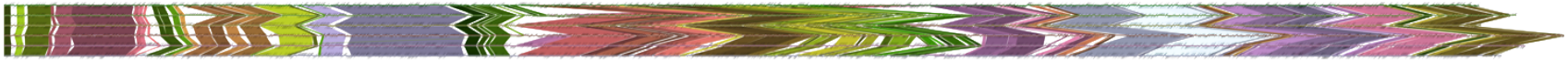
When we compare two features (e.g. genes) between two or more genomes, there must be some basis for making the comparison That is, they have to be *equivalent* in some way, such as:

- common evolutionary origin
- functional similarity
- a family-based relationship

It's common to define equivalence of genome features in terms of evolutionary relationship.

Why look at equivalent features?

The James Hutton Institute



The real power of genomics is comparative genomics!

- Makes catalogues of genome components comparable between organisms
- Differences, e.g. presence/absence of equivalents may support hypotheses for functional or phenotypic difference
- Can identify characteristic signals for diagnosis/epidemiology
- Can build parts lists and wiring diagrams for systems and synthetic biology

Evolutionary relationships[a]

[a] Fitch (1970) *Syst. Zool.* **19**:99-113 doi:10.2307/2412448

The James Hutton Institute

Equivalencies and relationships can be quite complex.
We need precise terms to describe relationships between genome features.

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. Physiological Chem., U. Wisconsin, Madison 53706) 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113.—This work provides a means by which it is possible to determine whether two groups of related proteins have a common ancestor or are of independent origin. A set of 16 random

- **analogy**: functional similarity
- **homology**: evolutionary common ancestor

Table of Contents

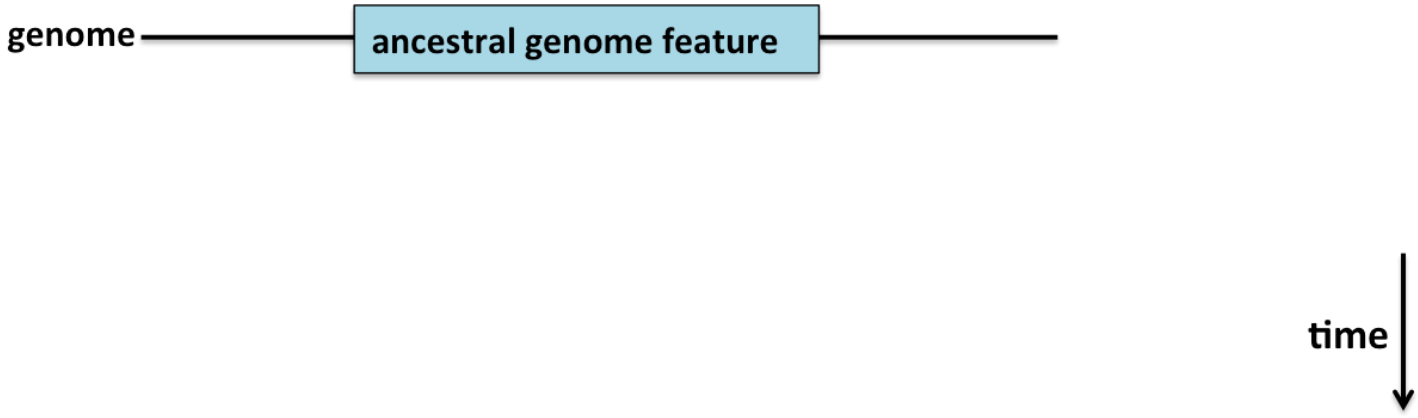
Who let the -logues out?[a]

[a] Fitch (2000) *Trends Genet.* **16**:227-231 doi:10.1016/S0168-9525(00)02005-9

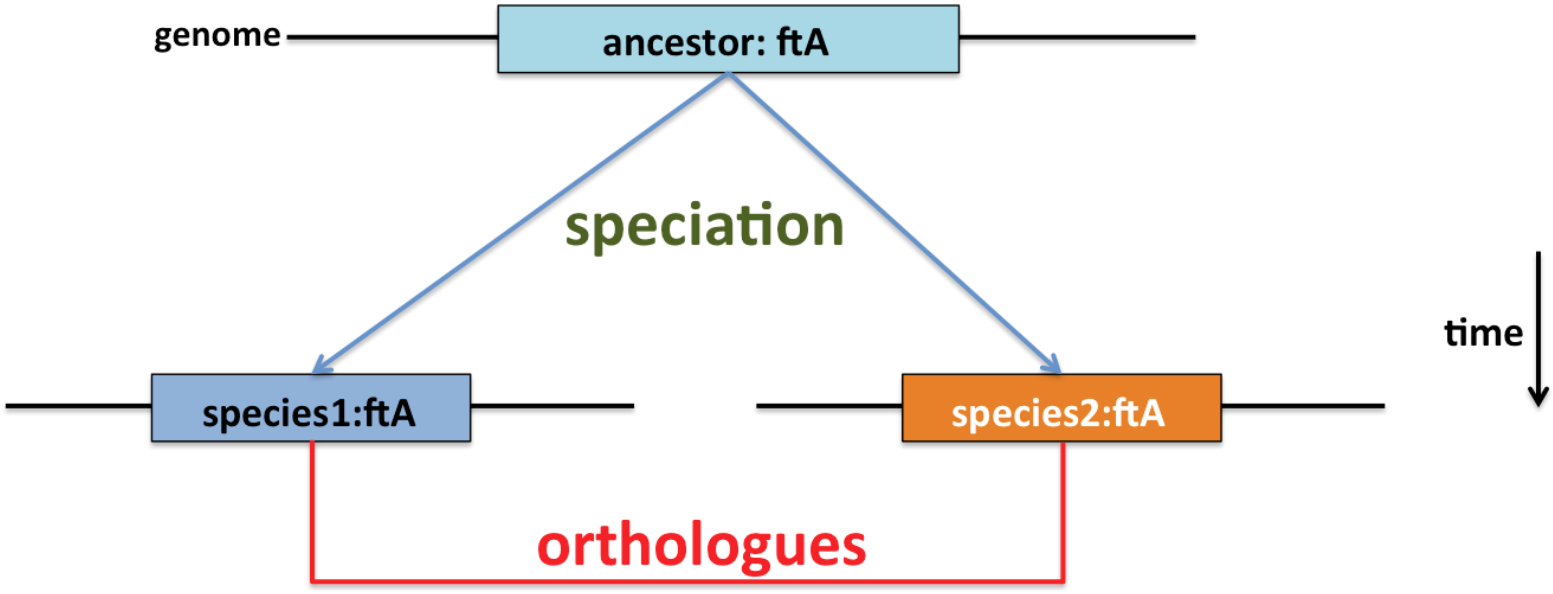
The James Hutton Institute

- **homologues**: elements that are similar because they share a common ancestor. **There are NOT degrees of homology**
- **analogues**: elements that are (functionally?) similar, and this may be through common ancestry *or* some other means, e.g. convergent evolution
- **orthologues**: homologues that diverged through speciation
- **paralogues**: homologues that diverged through duplication within the same genome

Who let the -logues out?

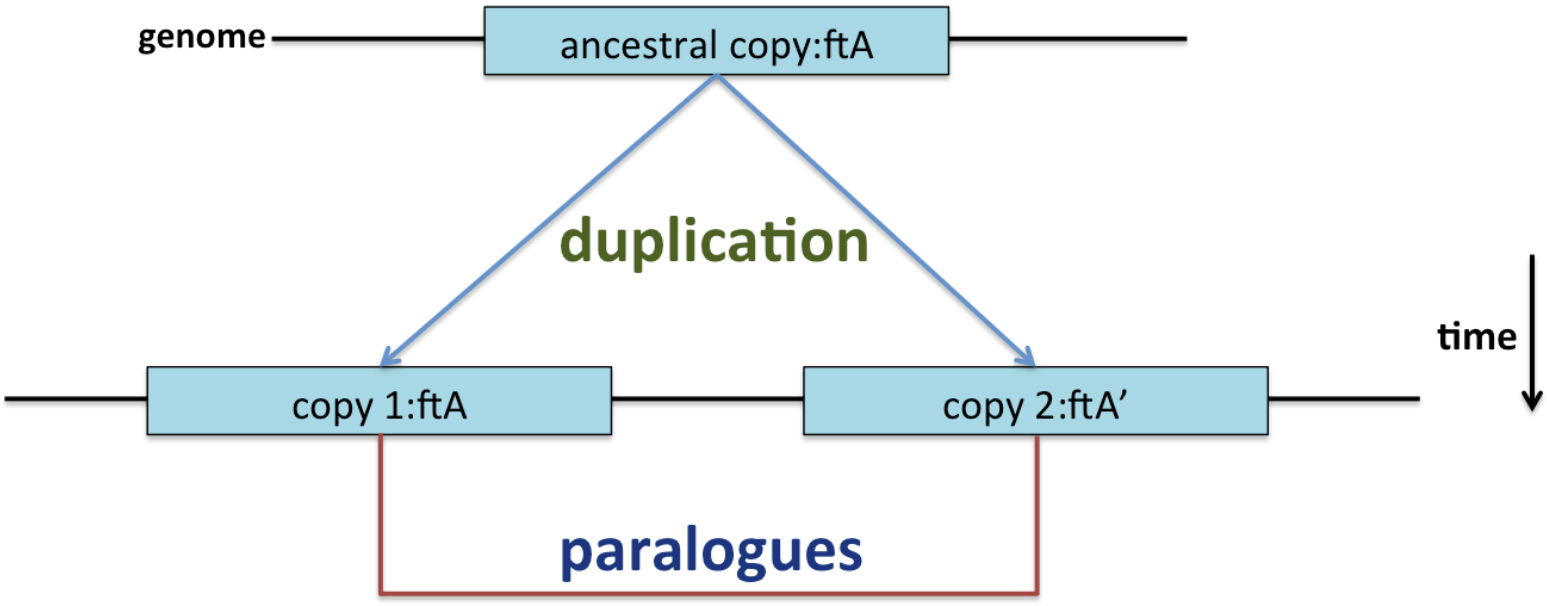


Who let the -logues out?



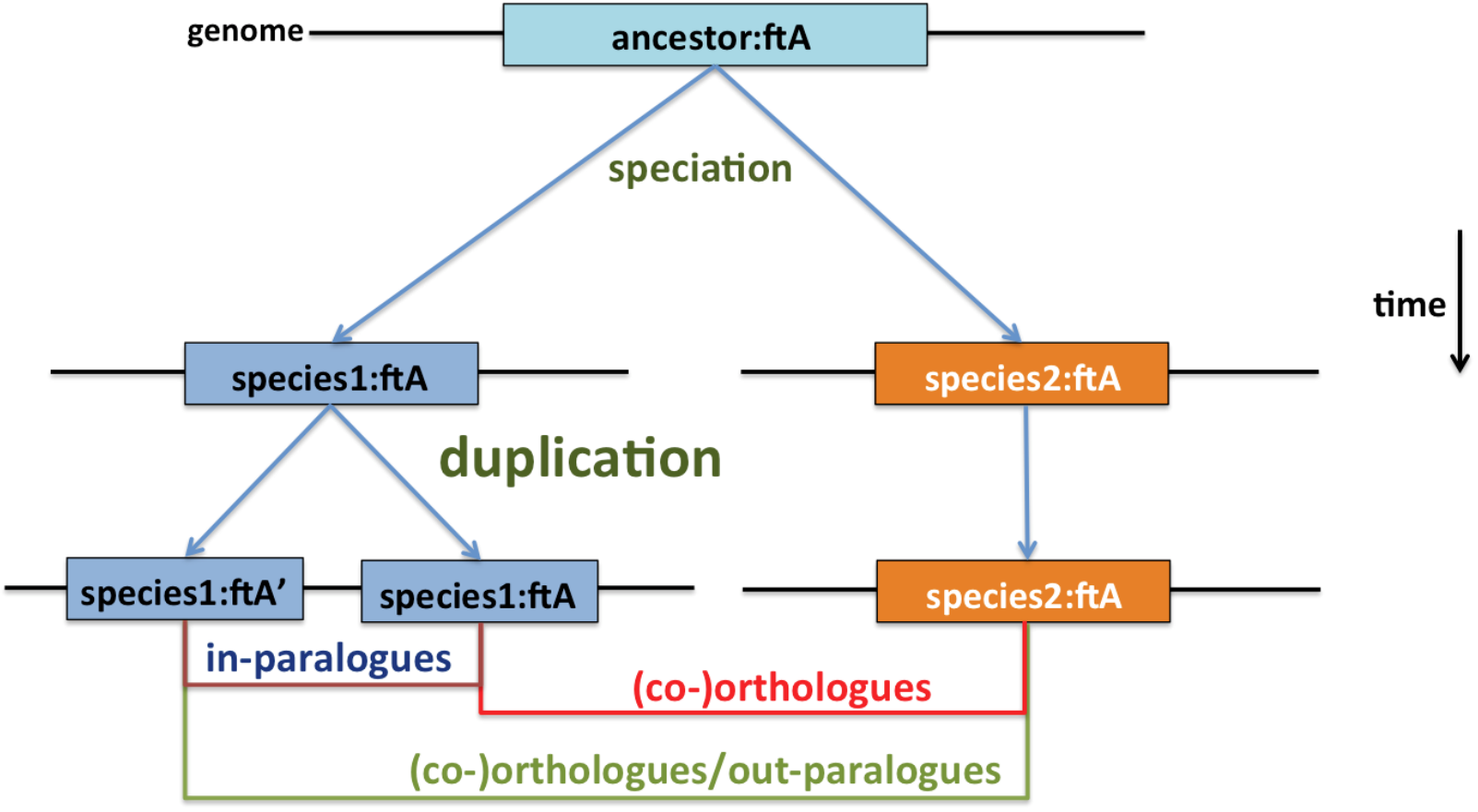
- **Orthologues:** homologues that diverged through speciation

Who let the -logues out?

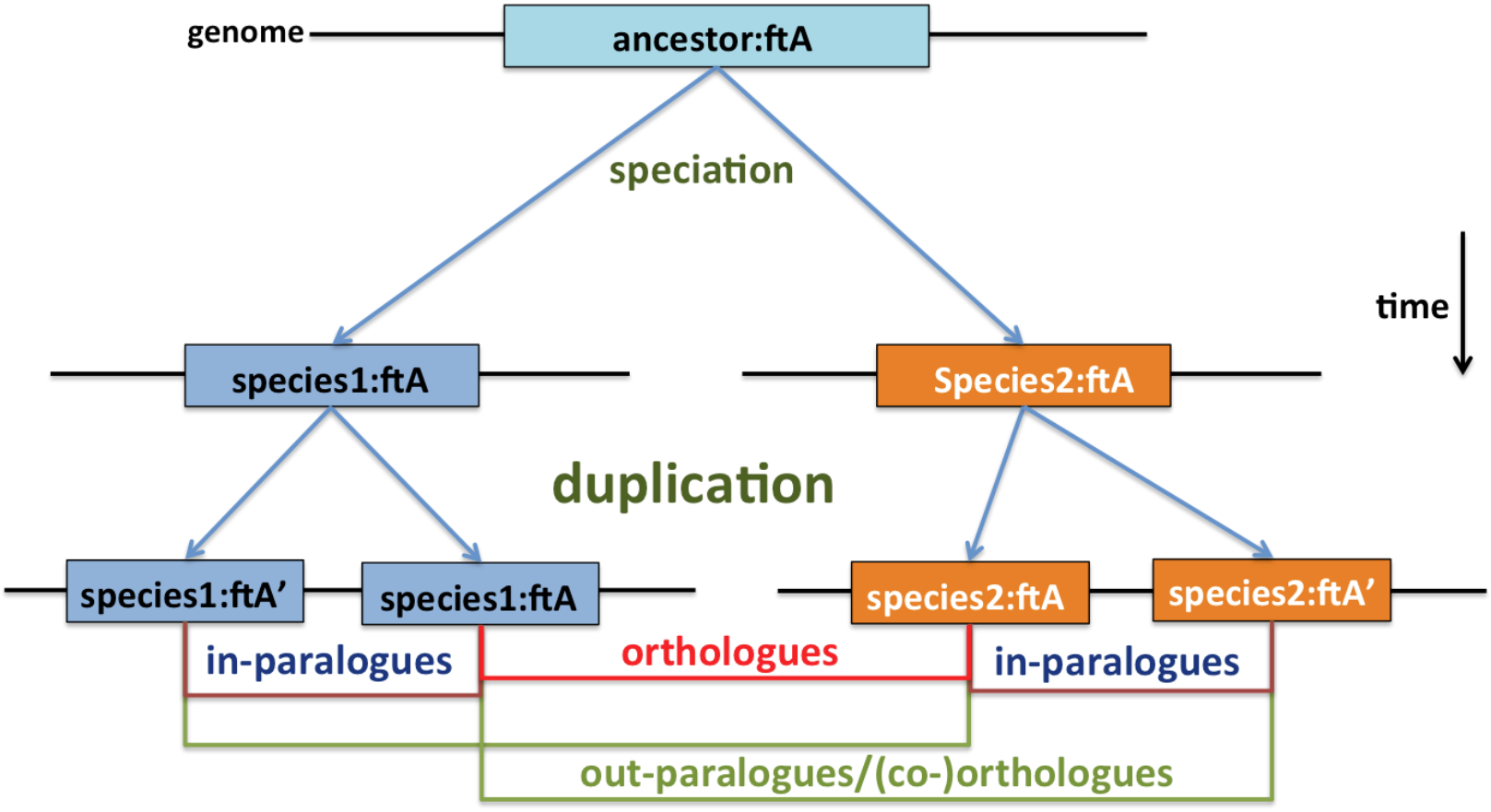


Paralogues: homologues that diverged through duplication within the same genome

Who let the -logues out?



Who let the -logues out?



ITYFIALMCTT[a]

[a] Kristensen *et al.* (2011) *Brief. Bioinf.* **12**:379-391 doi:10.1093/bib/bbr030

The James Hutton Institute

But it's a little more complicated than that.
Biology is not well-behaved.

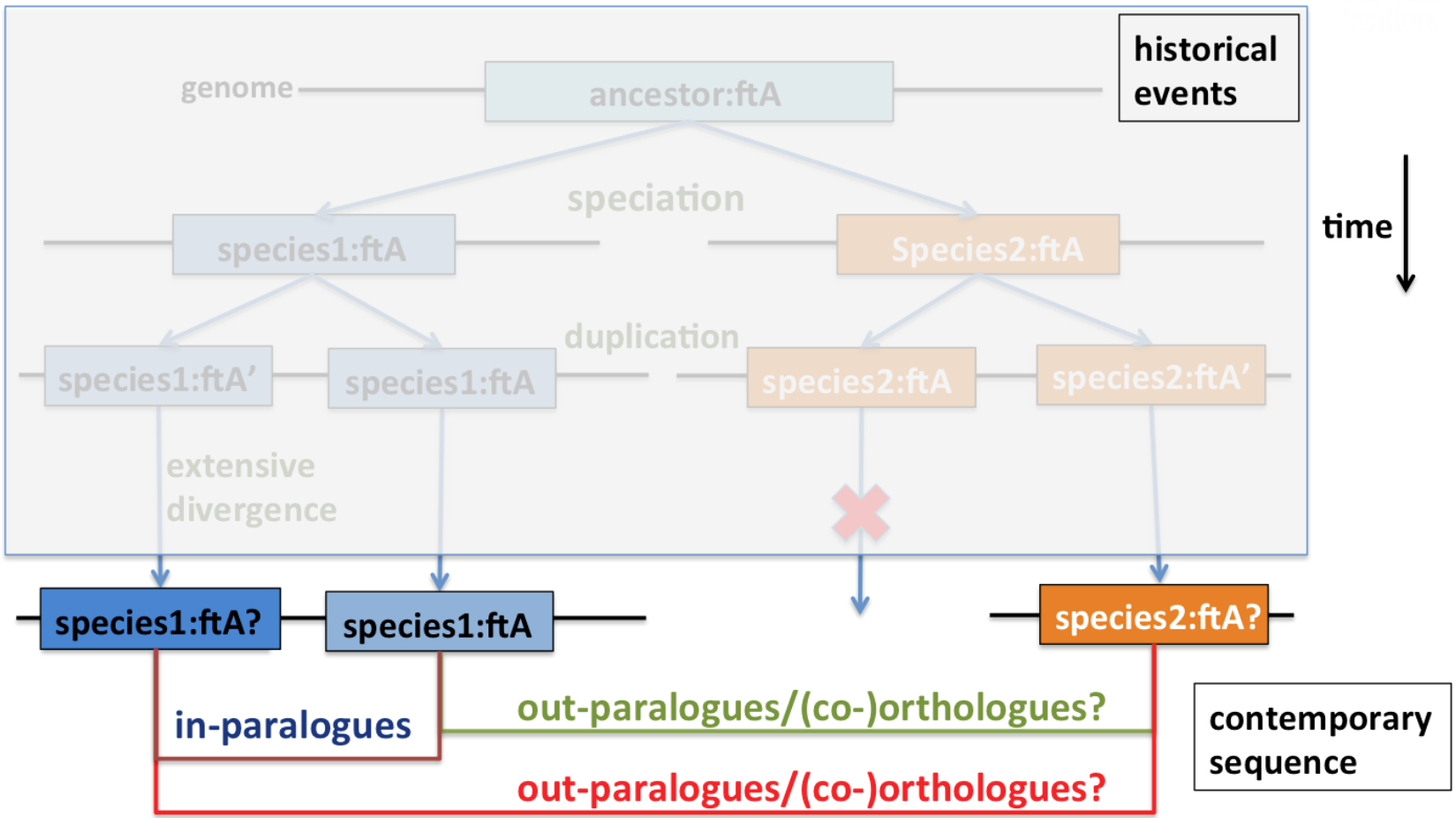
- Gene loss
- Homologues may diverge so widely that they can be hard to recognise
- Reconstructed evolutionary trees may not be robust inferences of speciation (or relevant to it, in prokaryotes)
- There is no record of history - we can only make inferences

All classifications of orthology/paralogy are inferences!

ITYFIALMCTT[a]

[a] Kristensen *et al.* (2011) *Brief. Bioinf.* **12**:379-391 doi:10.1093/bib/bbr030

The James Hutton Institute



All classifications of orthology/paralogy are inferences!

Ensembl Compara[a]

[a] Vilella *et al.* (2009) *Genome Res.* **19**:327-335 doi:10.1101/gr.073585.107

The James Hutton Institute

Some tools/databases, e.g. Ensembl Compara, use slightly different definitions (almost everything's an "orthologue")

- `within_species_paralog`:
  same-species paralogue (in-paralogue)
- `ortholog_one2one`:
  orthologue
- `ortholog_one2many`:
  orthologue/paralogue relationship
- `orthology_many2many`:
  orthologue/paralogue relationship

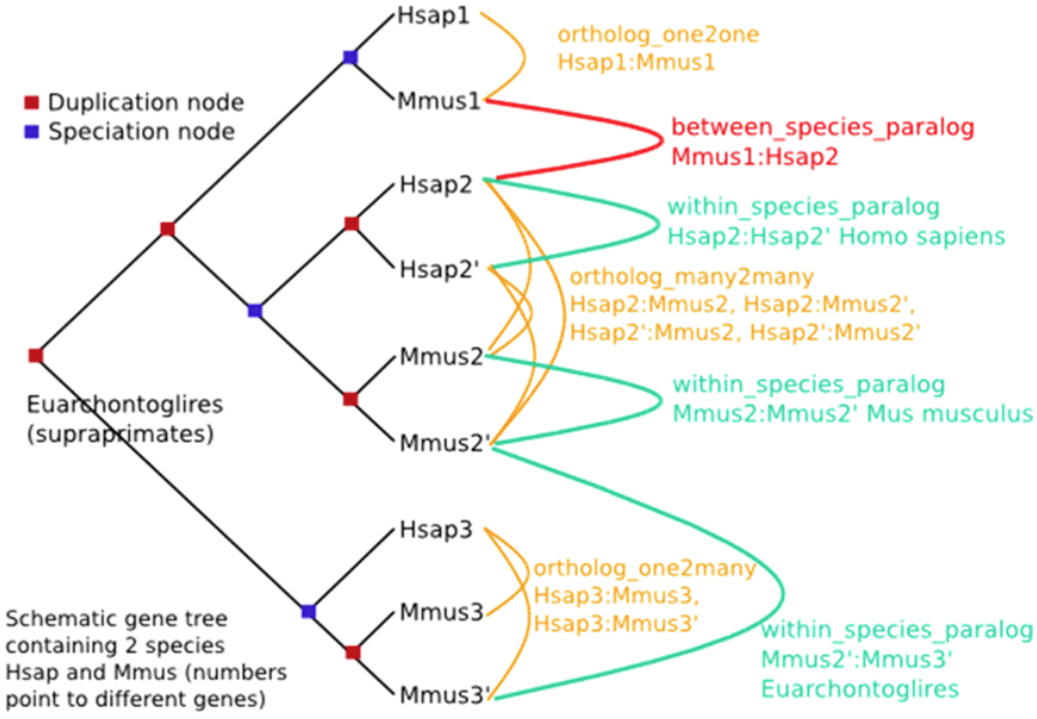


Table of Contents

Why focus on orthologues?

The James Hutton Institute

Formalise the idea of *corresponding genes* in different organisms.
Orthologues serve two purposes:

- **Evolutionary equivalence**
- **Functional equivalence** ("The Ortholog Conjecture"[60])

Applications in comparative genomics, functional genomics and phylogenetics.[61]
Over 30 databases attempt to describe orthologous relationships (http://questfororthologs.org/orthology_databases[62])

---

[60] Chen and Zhang (2012) *PLoS Comp. Biol.* **8**:e1002784 doi:10.1371/journal.pcbi.1002784

[61] Dessimoz (2011) *Brief. Bioinf.* **12**:375-376 doi:10.1093/bib/bbr057

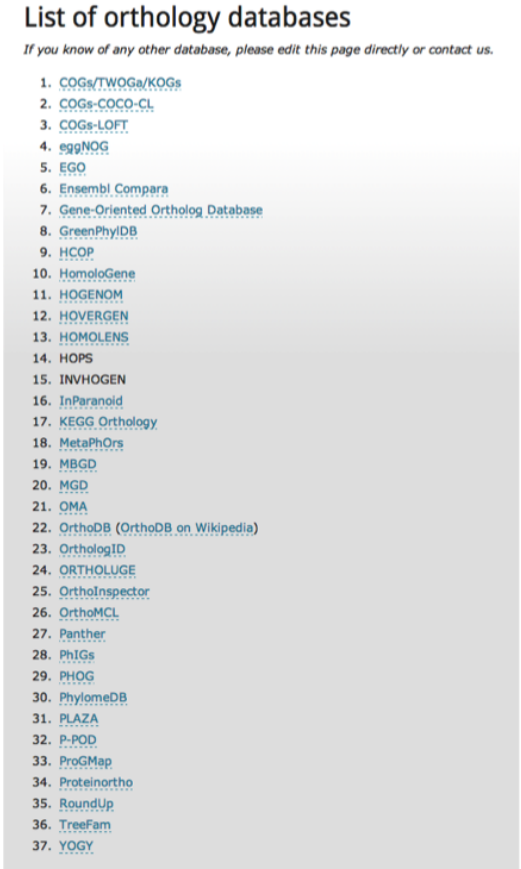
[62] Altenhoff and Dessimoz (2009) *PLoS Comp. Biol.* **5**:e1000262 doi:10.1371/journal.pcbi.1000262

Finding orthologues

The James Hutton Institute

Multiple methods and databases[63,64,65]

- **Pairwise genome**
  - RBBH (aka BBH, RBH), RSD, InParanoid, RoundUp
- **Multi-genome**
  - Graph-based: COG, eggNOG, OrthoDB, OrthoMCL, OMA, MultiParanoid
  - Tree-based: TreeFam, Ensembl Compara, PhylomeDB, LOFT



List of orthology databases

If you know of any other database, please edit this page directly or contact us.

1. COGs/TWOGs/KOGs
2. COGs-COCO-CL
3. COGs-LOFT
4. eggNOG
5. EGO
6. Ensembl Compara
7. Gene-Oriented Ortholog Database
8. GreenPhylDB
9. HCOP
10. HomoloGene
11. HOGENOM
12. HOVERGEN
13. HOMOLENS
14. HOPS
15. INVHOGEN
16. InParanoid
17. KEGG Orthology
18. MetaPhOrs
19. MBGD
20. MGD
21. OMA
22. OrthoDB (OrthoDB on Wikipedia)
23. OrthologID
24. ORTHOLUGE
25. OrthoInspector
26. OrthoMCL
27. Panther
28. PhIGs
29. PHOG
30. PhylomeDB
31. PLAZA
32. P-POD
33. ProGMap
34. Proteinortho
35. RoundUp
36. TreeFam
37. YOGY

---

[63] Kristensen *et al.* (2011) *Brief. Bioinf.* **12**:379-391 doi:10.1093/bib/bbr030

[64] Trachana *et al.* (2011) *Bioessays* **33**:769-780 doi:10.1002/bies.201100062

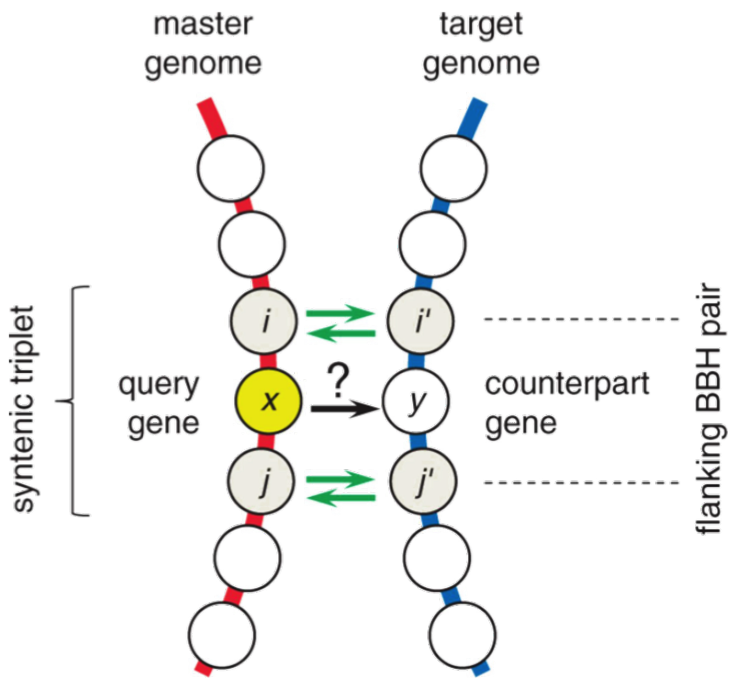
[65] Salichos and Rokas (2011) *PLoS One* **6**:e18755 doi:10.1371/journal.pone.0018755.g006

Table of Contents

Which prediction methods work best?

The James Hutton Institute

Taking advantage of prokaryotic operon structure: **if the outer pair of a syntenic triplet of genes are orthologous, the middle gene is also likely to be orthologous.**[66]
Specifically testing reciprocal best hits (RBH).

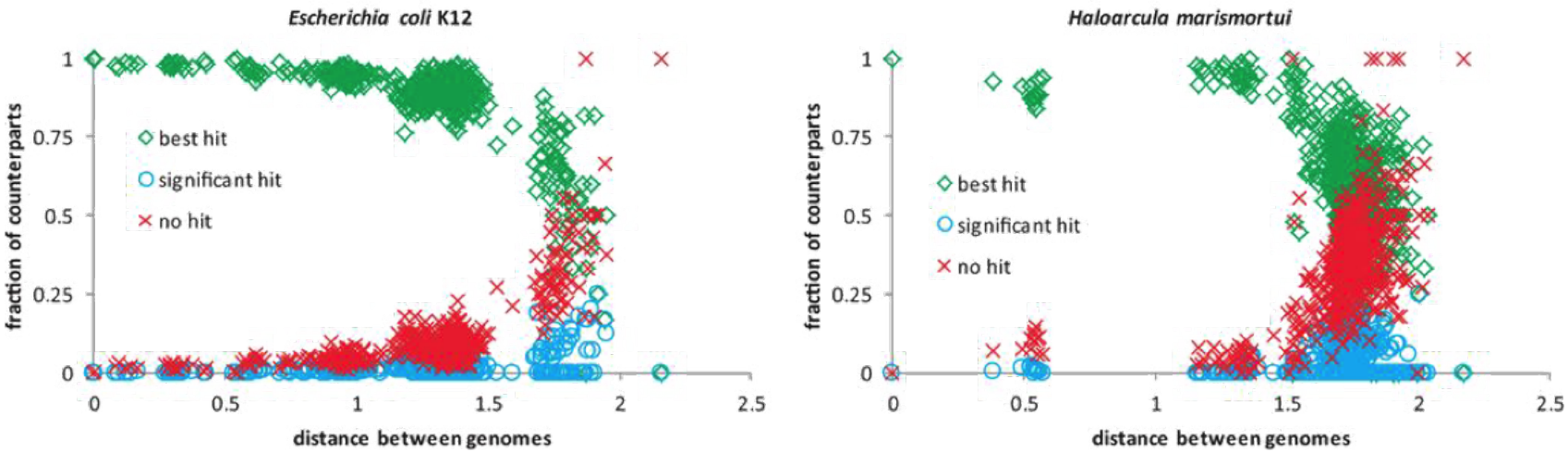


[66] Wolf and Koonin (2012) *Genome Biol. Evil.* **4**:1286-1294 doi:10.1093/gbe/evs100

Which prediction methods work best?

- Tested on 573 prokaryotic genomes
- 88-99% of RBH found in syntenic triplets
- Overwhelming majority of middle genes are RBH

RBH reliably finds orthologues.[67]



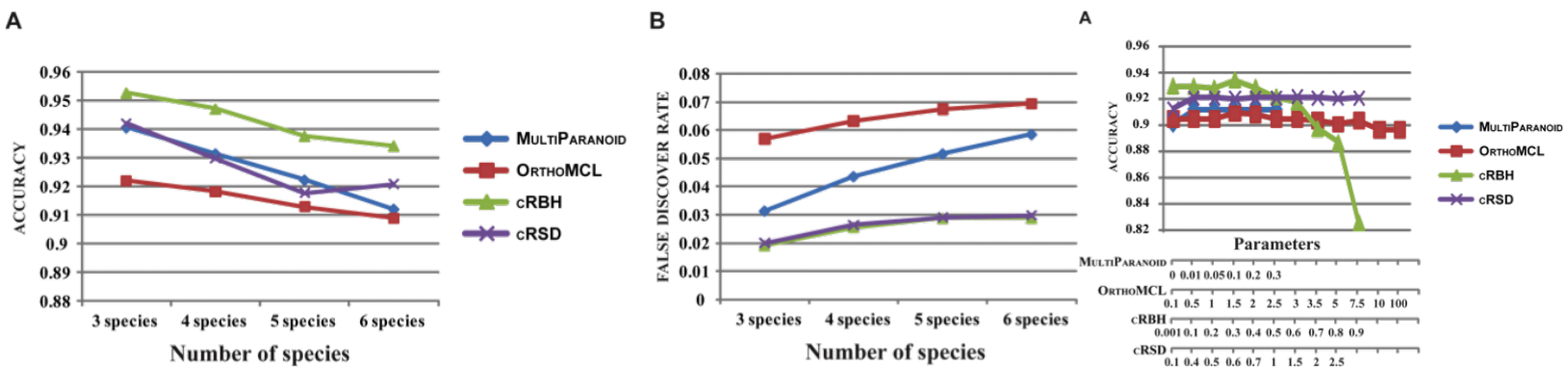
[67] Wolf and Koonin (2012) *Genome Biol. Evil.* **4**:1286-1294 doi:10.1093/gbe/evs100

Which prediction methods work best?

Four methods tested against 2,723 curated orthologues from six *Saccharomycetes*

- RBBH (and cRBH); RSD (and cRSD); MultiParanoid; OrthoMCL
- Rated by statistical performance metrics: sensitivity, specificity, accuracy, FDR

cRBH most accurate and specific, with lowest FDR.[68]



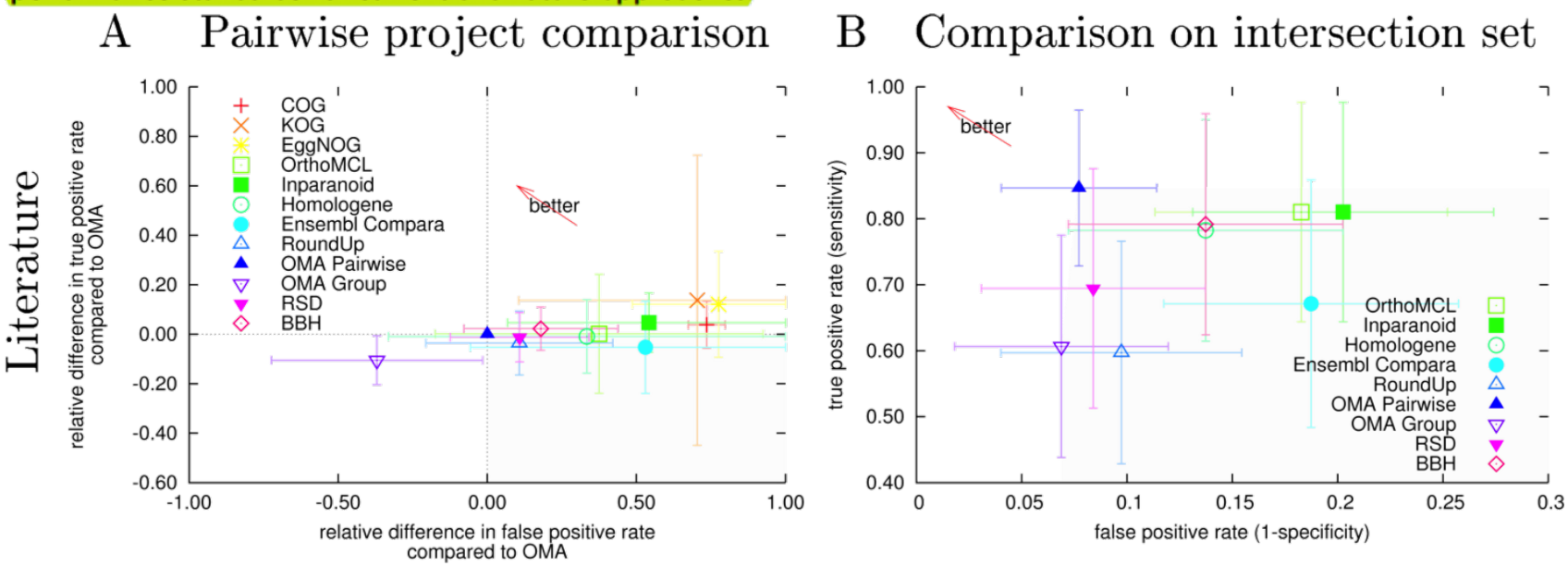
[68] Salichos and Rokas (2011) *PLoS One* **6**:e18755 doi:10.1371/journal.pone.0018755.g006

Which prediction methods work best?

The James Hutton Institute

Testing on literature-based benchmarks for grouping by function and correct branching of phylogeny.[69]

phylogenetic tests. Furthermore, we show that standard bidirectional best-hit often outperforms projects with more complex algorithms. First, the present study provides guidance for the broad community of orthology data users as to which database best suits their needs. Second, it introduces new methodology to verify orthology. And third, it sets performance standards for current and future approaches.



[69] Altenhoff and Dessimoz (2009) *PLoS Comp. Biol.* **5**:e1000262 doi:10.1371/journal.pcbi.1000262

Which prediction methods work best?

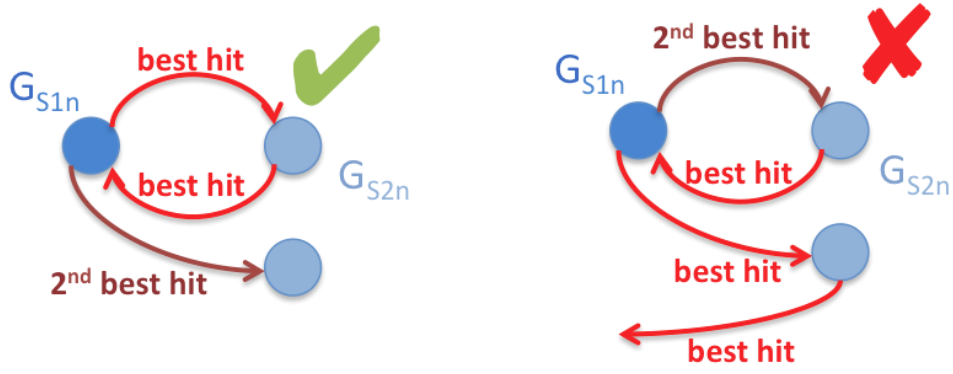
The James Hutton Institute

- Performance varies by choice of method, and interpretation of "orthology"
- Biggest influence is genome annotation quality
- Relative performance varies with choice of benchmark
- **(clustering) RBH outperforms more complex algorithms under many circumstances**

What is this magic RBH method?

The James Hutton Institute

- S1, S2 are the gene sequence sets from two organisms
- Use sequence search tool (BLAST/FASTA):
  - Query=S1, Subject=S2
  - Query=S2, Subject=S1



- Optionally filter hits (e.g. on %identity and %coverage)
- Find all pairs of sequences {G_{S1n}, G_{S2n}} in S1, S2 where G_{S1n} is the best BLAST match to G_{S2n} and G_{S2n} is the best BLAST match to G_{S1n}.

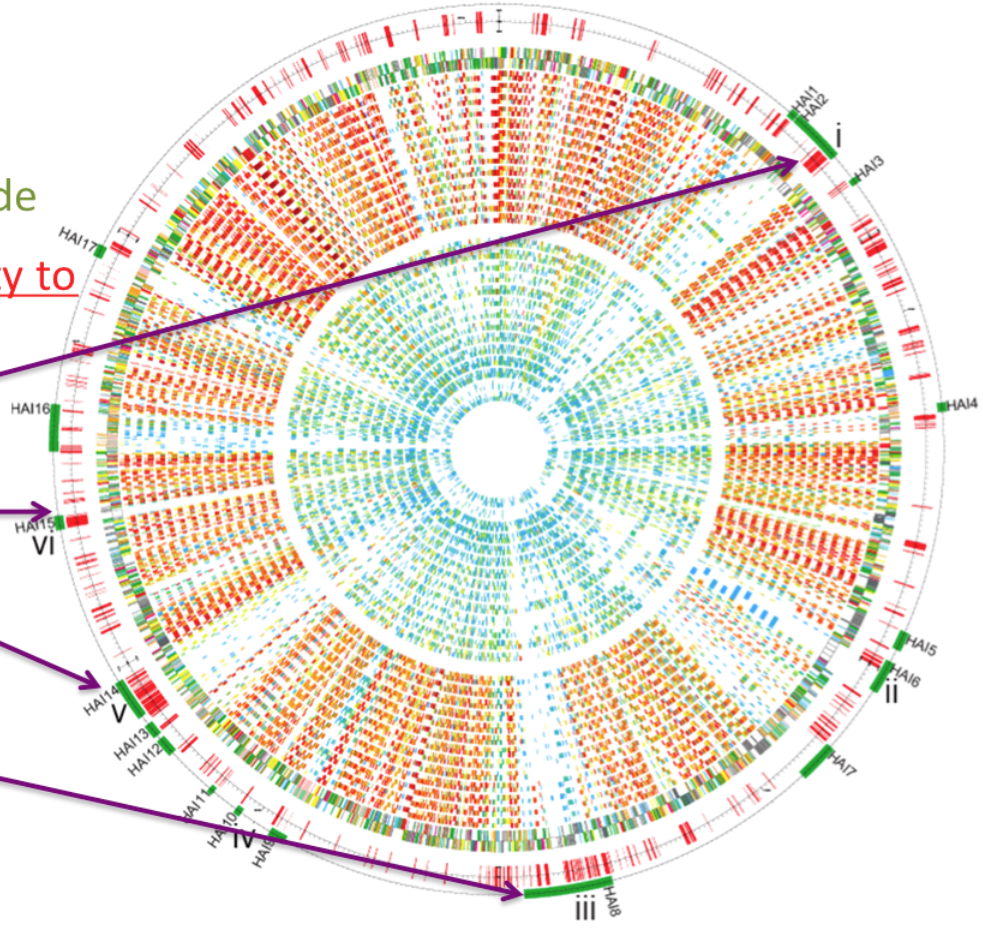
Table of Contents

Functional adaptation in Pba[a]

[a] Toth *et al.* (2006) *Ann. Rev. Phytopath.* **44**:305-336 doi:10.1146/annurev.phyto.44.070505.143444

The James Hutton Institute

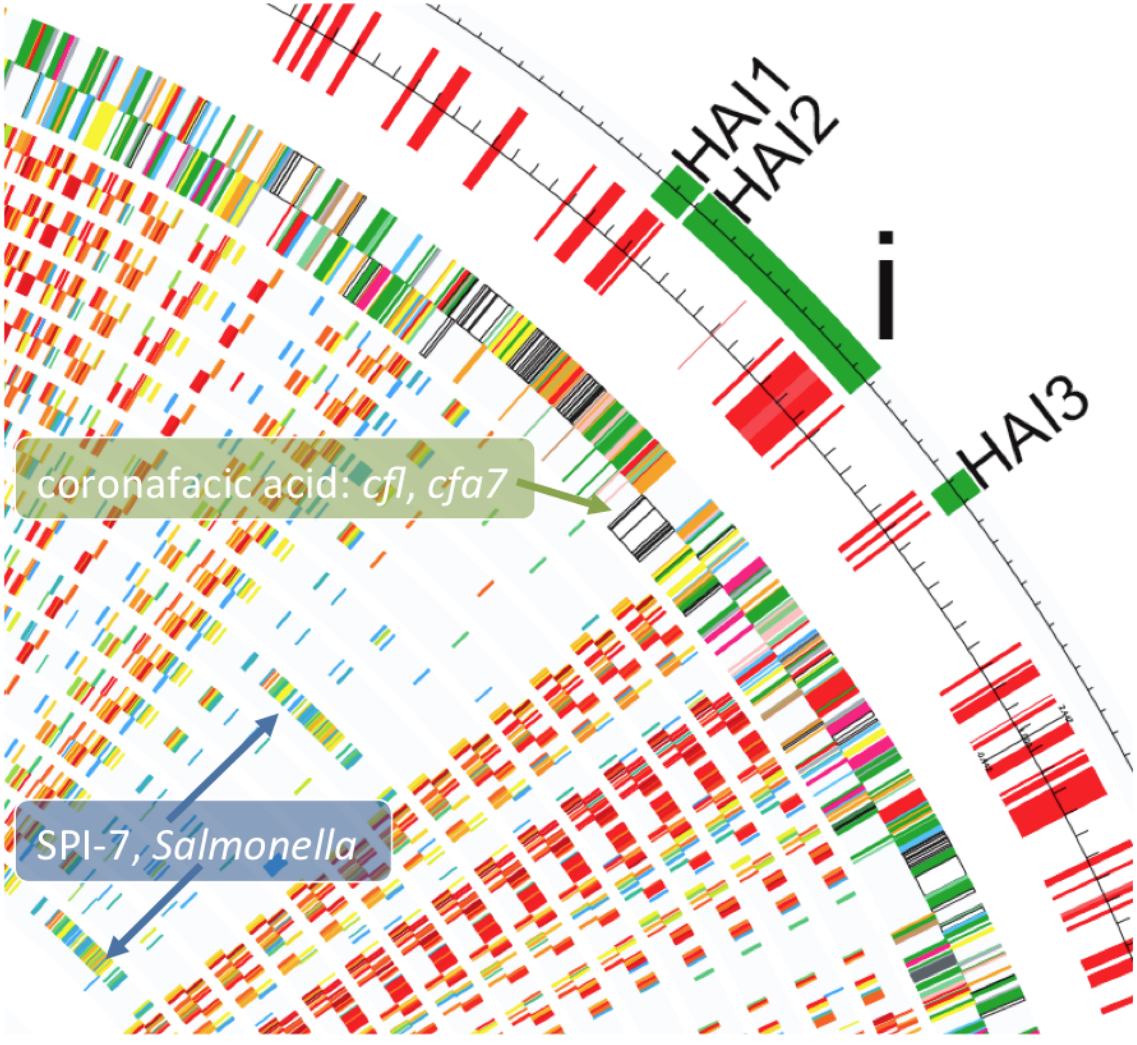
- Comparison against plant- (13) and animal-associated (14) bacteria
- Plant-associated in centre
- Animal-associated on outside
- Red marks: greater similarity to plant-associated bacteria
- **HAI2**: Phytotoxin
- **HAI15**: Adherence
- **HAI14**: Nitrogen fixation
- **HAI8**: T3SS



Functional adaptation in Pba[a]

[a] Toth *et al.* (2006) *Ann. Rev. Phytopath.* **44**:305-336 doi:10.1146/annurev.phyto.44.070505.143444

The James Hutton Institute



Coronatine (*P. syringae*) interferes with jasmonate responses in host, as a jasmonate mimic

Coronafacic acid – *Pseudomonas syringae* phytotoxin precursor (coronatine)
- payload

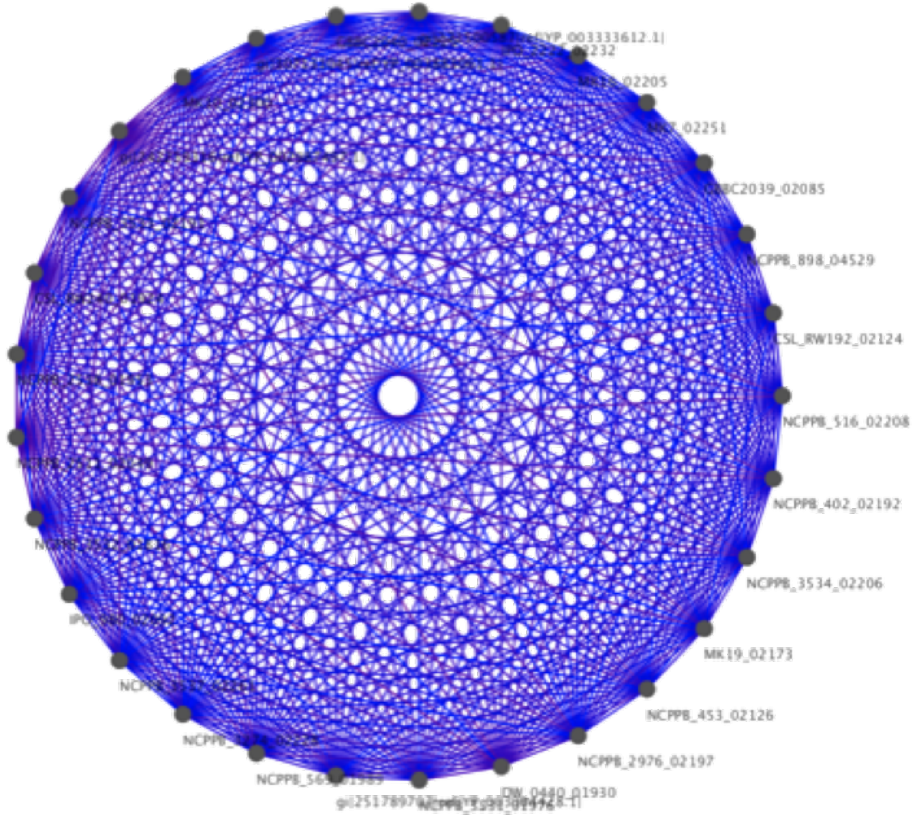
SPI-7 - *Salmonella* Typhi Pathogenicity island
- delivery system

Table of Contents

Core genome

The James Hutton Institute

Once equivalent genes have been identified, those present in all related isolates can be identified: **the core genome**.
The *core* genome is expected to underpin common function.
A core RBH cluster (*clique*) for 29 genomes:

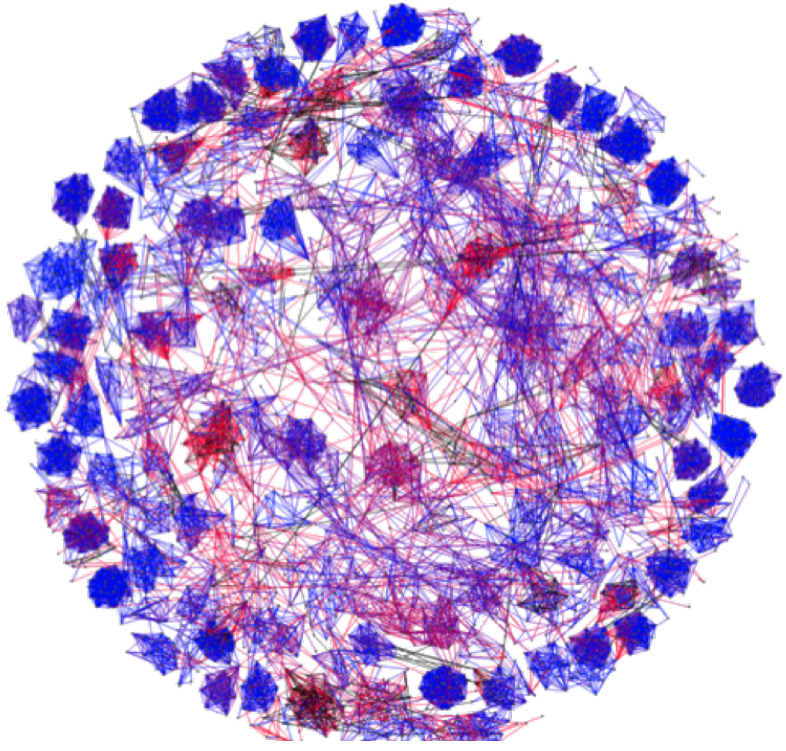


Accessory genome

The James Hutton Institute

The remaining genes are **the accessory genome**, and are expected to mediate function that distinguishes between isolates.

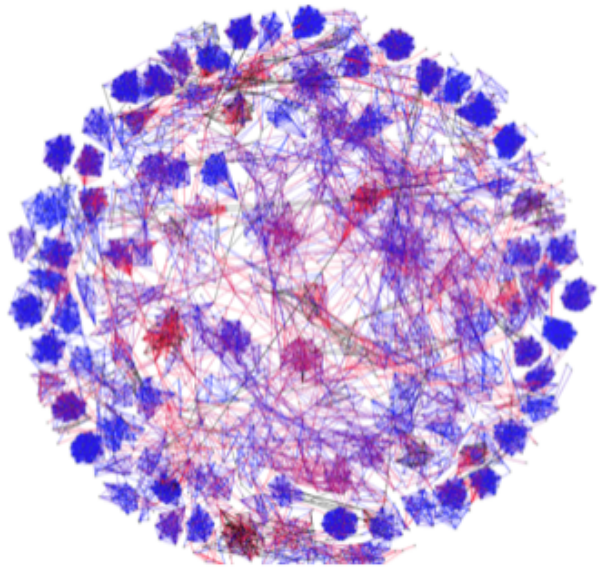
An accessory RBH cluster for 29 genomes:



Accessory clusters

Accessory RBH clusters can be pruned, to identify the accessory genome specific to subgroups of isolates:

Species	Weak Pruning	Full Pruning
Core Genome	2201	2201
D. chrysanthemi	32	36
D. dadantii	11	14
D. dianthicola	102	127
D. paradisiaca	404	441
D. solani	120	157
D. zeae	33	40



- Accessory: RBBH with all other members of same species, but no other *Dickeya*
- Weak pruning: remove all RBBH <80% identity, <40% coverage
- Full pruning: trim graph (by Mahalanobis distance) until minimal cliques found

These genes may be responsible for subgroup-specific phenotypes

Accessory genome

Accessory genomes act as a cradle for adaptive evolution[70]

This is particularly so for pathogens, such as *Pseudomonas* spp.[71]

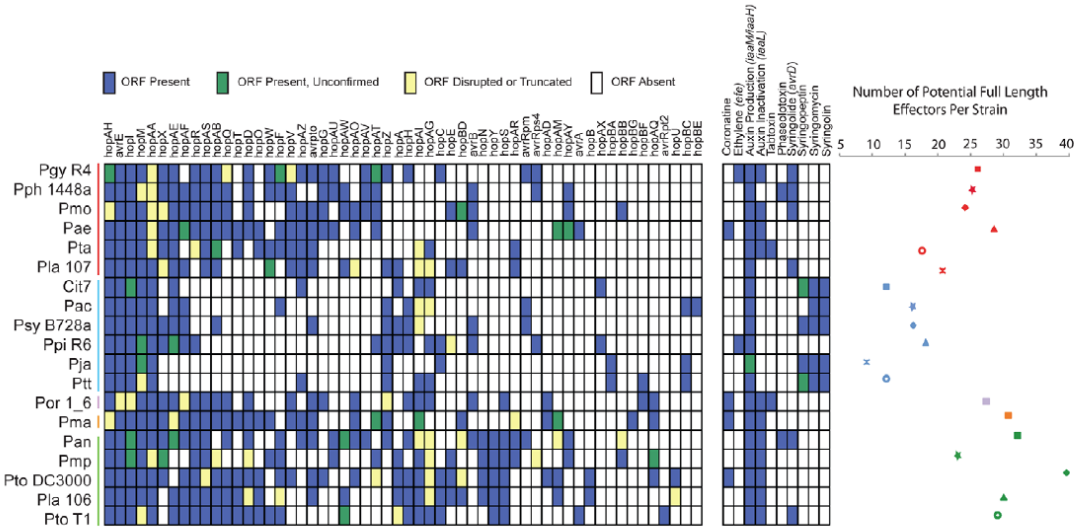


Figure 3. ***P. syringae*** **isolates harbor extensive diversity in virulence gene repertoires.** TTE, toxin, and plant hormone biosynthesis genes are listed across the top, *P. syringae* genomes, color-coded by phylogenetic group as in Figure 1. At the left, a blue box indicates presence of full-length ORFs or complete pathways within each genome. Green boxes indicate that genes or pathways are present by similarity searches, but the presence of full-length genes could not be verified by PCR, or the pathways are potentially incomplete. Yellow boxes indicate that genes are either significantly truncated or are disrupted by insertion sequence elements. White boxes indicate absence of genes or pathways from the strains based on homology searches. At the far right, the total number of potentially functional TTE proteins is shown for each genome and displayed according to the color-coded strain and group symbols shown in Figure 1.
doi:10.1371/journal.ppat.1002132.g003

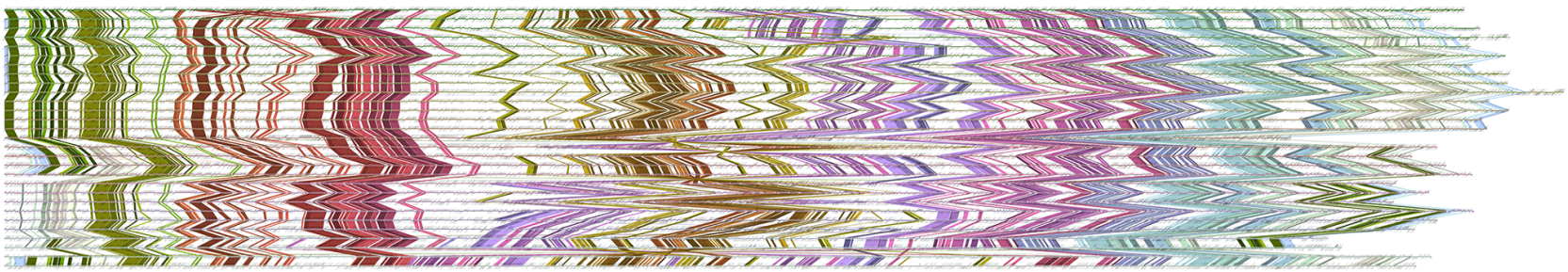
[70] Croll and Mcdonald (2012) *PLoS Path.* **8**:e1002608 doi:10.1371/journal.ppat.1002608

[71] Baltrus *et al.* (2011) *PLoS Path.* **7**:e1002132 doi:10.1371/journal.ppat.1002132.t002

Core genome synteny

The James Hutton Institute

Using tools like i-ADHoRe[72] that identify synteny and collinearity, the structural organisation of the core genome can be determined:



For *Dickeya*, the core genome appears to be structurally well-conserved across all isolates.

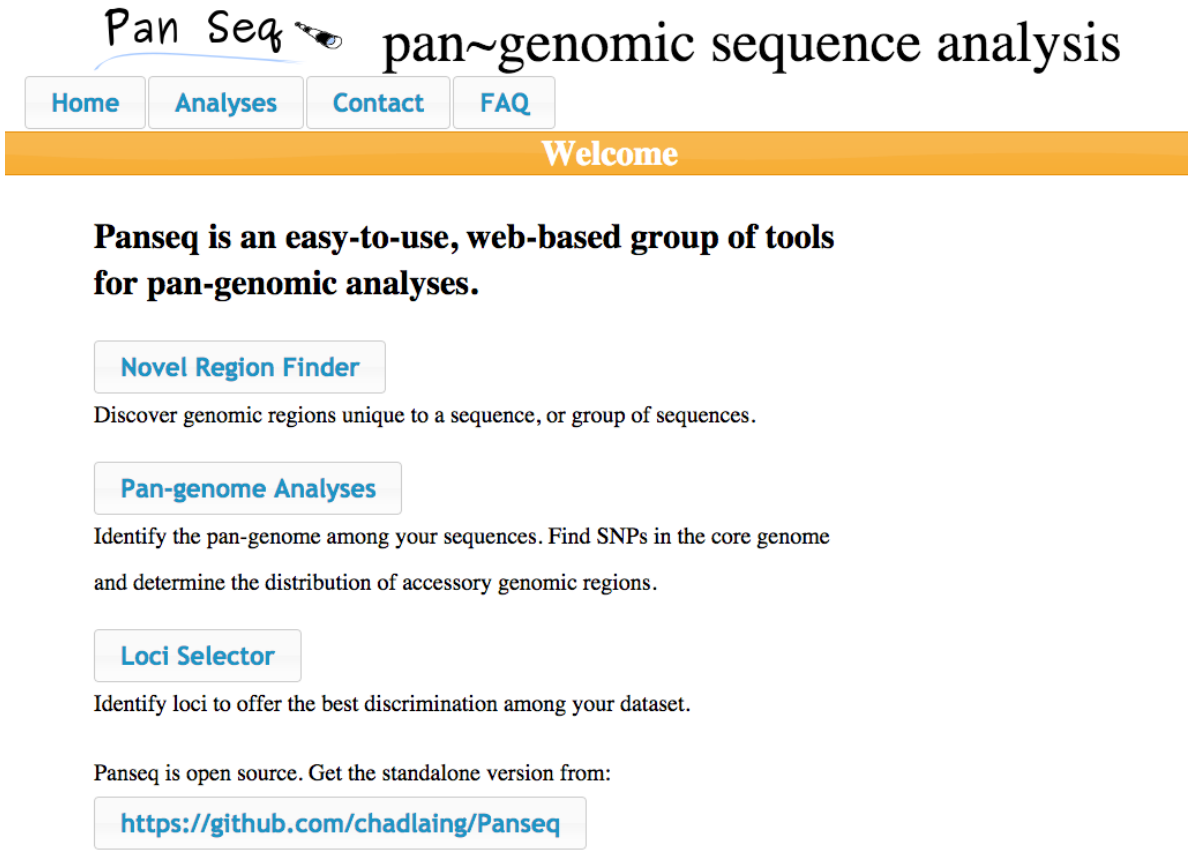
[72] Proost *et al.* (2012) *Nuc. Acids Res.* **40**:e11 doi:10.1093/nar/gkr955

Panseq[a]

[a] Laing *et al.* (2010) *BMC Bioinf.* **11**:461 doi:10.1186/1471-2105-11-461

The James Hutton Institute

Panseq is an online tool for identification of core and accessory genomes, available at https://lfz.corefacility.ca/panseq/, and https://github.com/chadlaing/Panseq for standalone use



Harvest[a]

[a] Treangen *et al.* (2014) *Genome Biol.* **15**:524 doi:10.1186/s13059-014-0524-x

The James Hutton Institute

Visualising and organising comparison/pangenome data across thousands of bacteria is difficult.

The `Harvest` suite of tools enables alignment and visualisation of thousands of genomes:

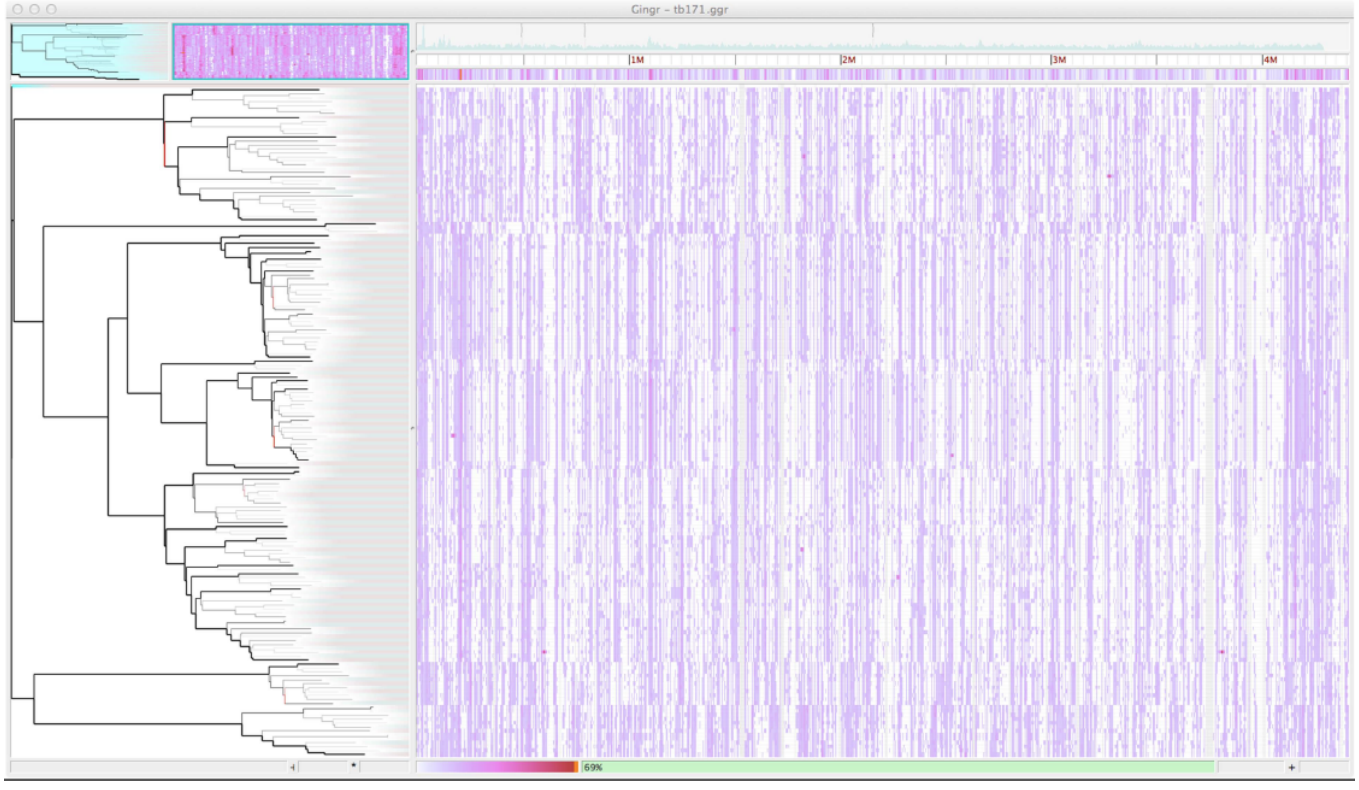


Table of Contents

Things I didn't get to

The James Hutton Institute

- **Genome-Wide Association Studies (GWAS):**
  - Try http://genenetwork.org/ to play with some data
- **Prediction of regulatory elements, e.g.**
  - Kellis *et al.* (2003) *Nature* doi:10.1038/nature01644
  - King *et al.* (2007) *Genome Res.* doi:10.1101/gr.5592107
  - Chaivorapol *et al.* (2008) *BMC Bioinf.* doi:10.1186/1471-2105-9-455
  - CompMOBY: http://genome.ucsf.edu/compmoby/
- **Detection of Horizontal/Lateral Gene Transfer (HGT/LGT), e.g.**
  - Tsirigos & Rigoutsos (2005) *Nucl. Acids. Res.* doi:10.1093/nar/gki187
- **Phylogenomics, e.g.**
  - Delsuc *et al.* (2005) *Nat. Rev. Genet.* doi:10.1038/nrg1603

Table of Contents

Conclusions

The James Hutton Institute

- **Comparative genomics is a powerful set of techniques for:**
  - Understanding and identifying evolutionary processes and mechanisms
  - Reconstructing detailed evolutionary history of a set of organisms
  - Identifying and understanding common genomic features of organisms
  - Providing hypotheses about gene function for experimental investigation
- **A huge amount of data is available to work with**
  - And it's only going to get much, much larger
- **Results feed into many areas of study:**
  - Medicine and health
  - Agriculture and food security
  - Basic biology in all fields
  - Systems and synthetic biology

Conclusions

The James Hutton Institute

- **Comparative genomics is essentially based around comparisons**
  - What is similar between two genomes? What is different?
- **Comparative genomics is evolutionary genomics**
- **Large datasets benefit from visualisation for effective interpretation**
  - Much scope for improvement in visualisation
- **Tools with the same purpose give different output**
  - BLAST vs MUMmer
  - RBBH vs MCL
  - **Choice of application matters for correctness and interpretation! – understand what the application does, and its limits.**

Conclusions

The James Hutton Institute

- Comparative genomics is
  - Fun
  - Indoor work, in the warm and dry
  - Not a job that involves heavy lifting

Licence: CC-BY-SA

The James Hutton Institute

By: Leighton Pritchard