Comparative Genomics and Visualisation
BS32010
2.Bulk Genome Properties

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-2015-03-17-UoD_compgenvis

Table of Contents

The James Hutton Institute

Bulk genome comparisons

The James Hutton Institute

You don't have to sequence genomes to compare them
(but it helps)
PART 1: Experimental

Genome comparisons predate NGS

The James Hutton Institute

- Sequence data wasn't always cheap and abundant
- Practical, experimental genome comparisons were needed

Bulk genome comparisons

The James Hutton Institute

Calculate values for individual genomes, then compare them.

Bulk genome properties

The James Hutton Institute

- Large-scale summary measurements
- Measure genomes independently - compare values later

Bulk genome properties

The James Hutton Institute

- Large-scale summary measurements
- Measure genomes independently - compare values later
    - What kinds of measurements/properties?

Bulk genome properties

The James Hutton Institute

- Large-scale summary measurements
- Measure genomes independently - compare values later
    - Number of chromosomes
    - Ploidy
    - Chromosome size
    - Nucleotide (A,C,G,T) frequency

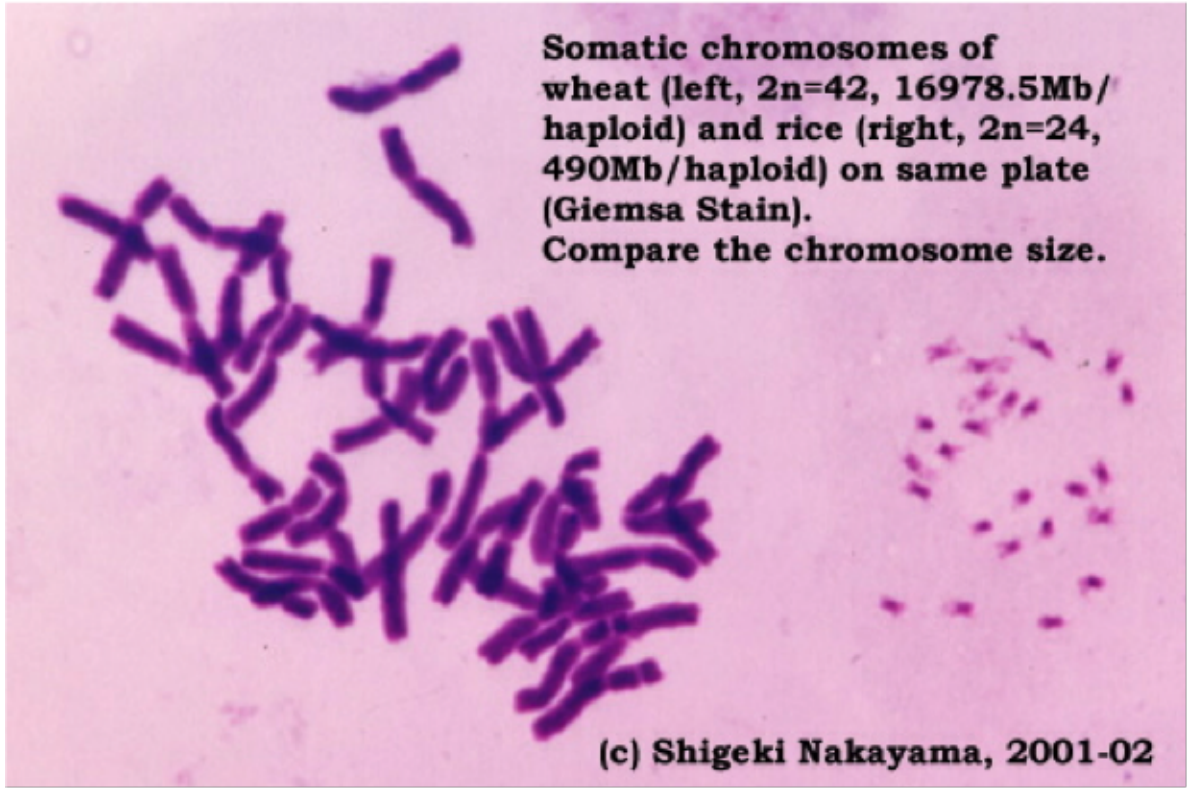
Chromosome stains [a] [b]

[a] IRGSP (2005) *Nature* doi:10.1038/nature03895

[b] Brenchley *et al.* (2012) *Nature* doi:10.1038/nature11650

Count chromosomes, estimate size

Wheat: 17Gbp; Rice: 390Mbp



Chromosome count/size [a] [b]

[a] Kamisugi *et al.* (1993) *Chromosome Res.* **1**(3): 189-196

[b] Wang *et al.* (2013) *Nat. Rev. Genet.* doi:10.1038/nrg3375

- chromosome count and ploidy can vary widely

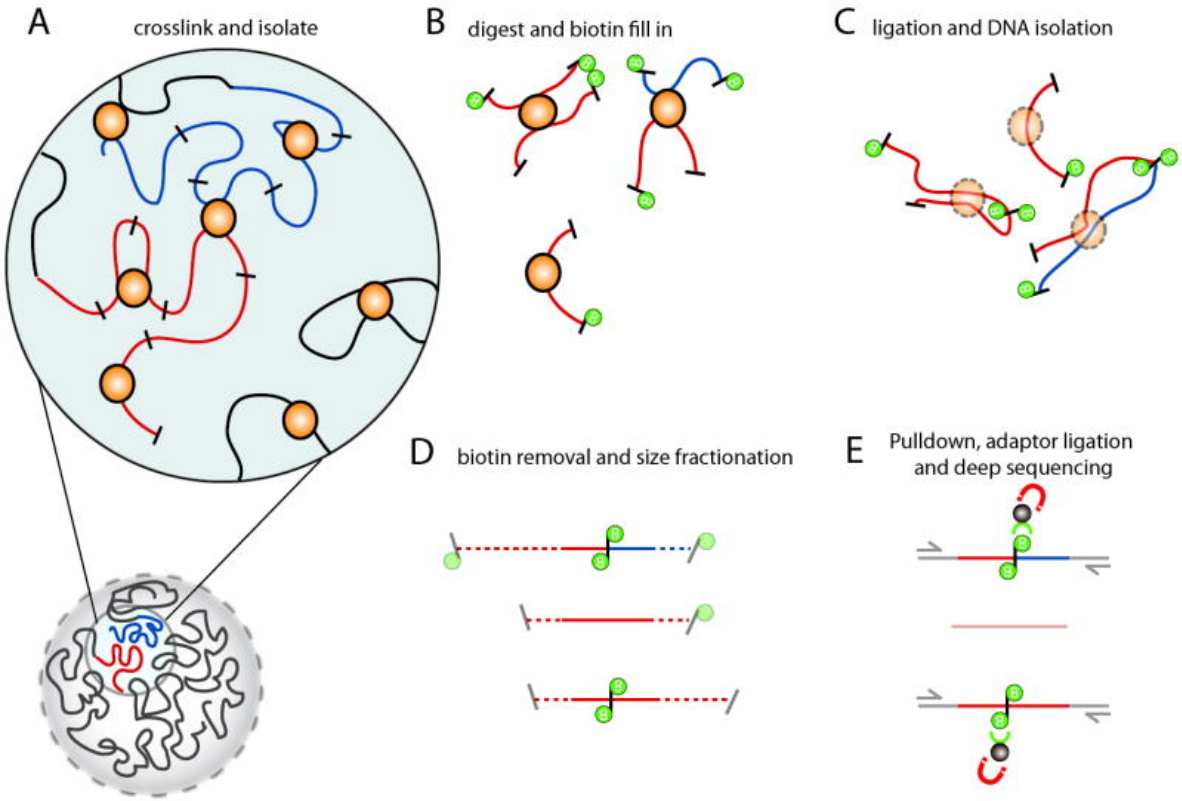
Organism	Chromosomes	Ploidy
E. coli	1	1
Human (*H. sapiens*)	46	2
Rice (*O. sativa*)	24	1
Adders-tongue (*Ophioglossum reticulatum*)	1260	84
Domestic (not wild) wheat somatic	42	6
Domestic (not wild) wheat gametic	14	2

Physical study of chromosomes [a]

[a] Belton *et al.* (2012) *Methods* doi:10.1016/j.ymeth.2012.05.001

The James Hutton Institute

- Genome size and chromosome count do **not** indicate organism "complexity"
- Modern physical study of chromosomes still produces surprises (e.g. Hi-C: chromatin interaction)



Nucleotide content [a]

[a] Krane *et al.* (1991) *Nucl. Acids Res.* **19**(19): 5181-5185

The James Hutton Institute

- Radiolabel monophosphates from genomic DNA
- Separate by thin-layer chromatography
- Compare label ratios using scanner

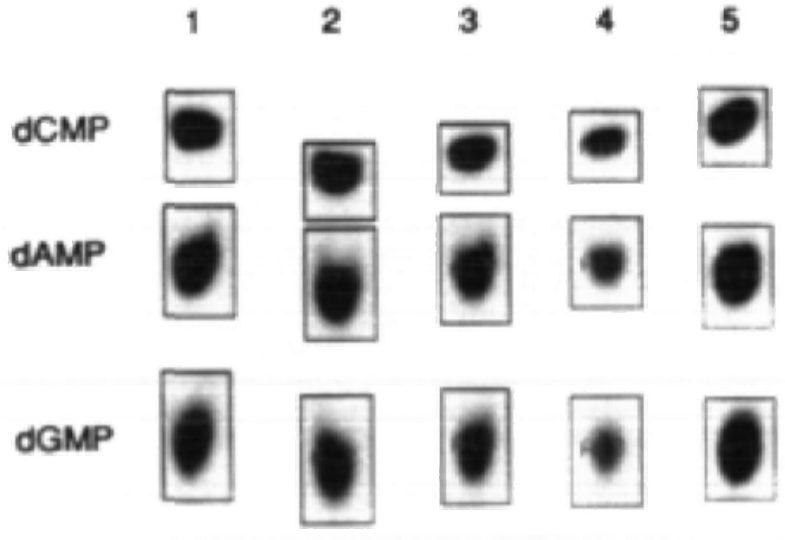


Table 4. Base composititon of yeast artificial chromosomes containing of human DNA inserts.

YAC	Replicate	Corrected G:A	Corrected C:A	%G+C from G:A	%G+C from C:A	Average GC content	Std dev.
yKM19-3	1	0 737	0.799	42.4	44.4	43.4	
(180 kb)	2	0.703	0.771	41.3	43.5	42.4	0 58
	3	0.707	0.766	41.4	43.4	42.4	
yCF-4	1	0.724	0.766	42.0	43.4	42 7	
(330 kb)	2	0.708	0.731	41.4	42.2	41 8	0.45
	3	0.701	0 759	41.2	43.2	42.2	
yW30-5	1	0.722	0 747	41.9	42.7	42.3	
(300 kb)	2	0 744	0.765	42.7	43.3	43.0	0.47
	3	0 713	0.739	41.6	42 5	42.1	
yJ311-2	1	0.725	0.747	42.0	42.8	42 4	
(230 kb)	2	0.728	0.757	42.1	42.4	42.6	0 10
	3	0.728	0.736	42.1	42.4	42.5	
yJ311-5	1	0.695	0 726	41.0	42.1	41.5	
(200 kb)	2	0 746	0.756	42.7	43.0	42.9	0.70
	3	0 731	0 733	42.2	42.3	42.3	
yHPRT	1	0 718	0.744	41.8	42 7	42.2	
(680 kb)	2	0 718	0.725	41.8	42.0	41 9	0.46
	3	0.694	0.713	41.0	41.6	41.3	

Table of Contents

The James Hutton Institute

This happened...

The James Hutton Institute

- Cheap, accurate, high-throughput sequencing

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 can 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)

After that, the flood...

The James Hutton Institute

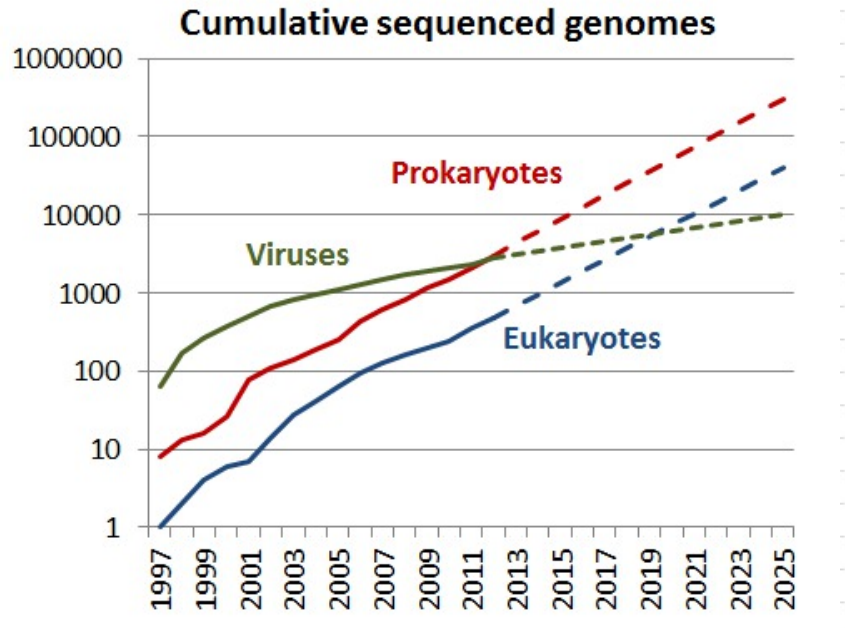
High-throughput sequencing methods have completely changed the landscape of biology
(Nearly) complete, (mainly) accurate sequence data is now inexpensive (and cheaper than analysis)

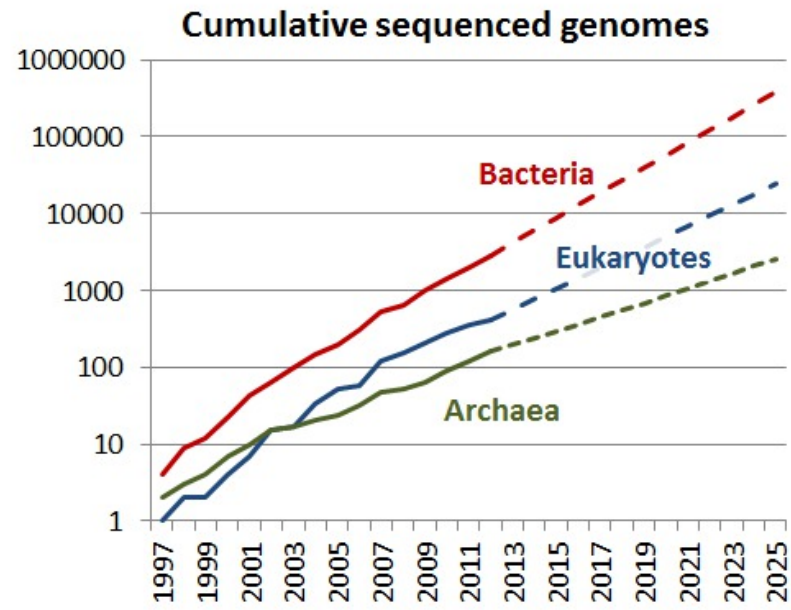
- GOLD (19/2/2014): 3,011 "finished" ; 9,891 "permanent draft" genomes
- GOLD (18/11/2014): 6,649 "finished" ; 23,552 "permanent draft" genomes
- NCBI WGS (19/2/2014): 17,023 microbial genomes
- NCBI WGS (18/11/2014): 26,026 microbial genomes

Predicting the future is hard...

The James Hutton Institute

Su *et al.* attempted to answer this[1]:

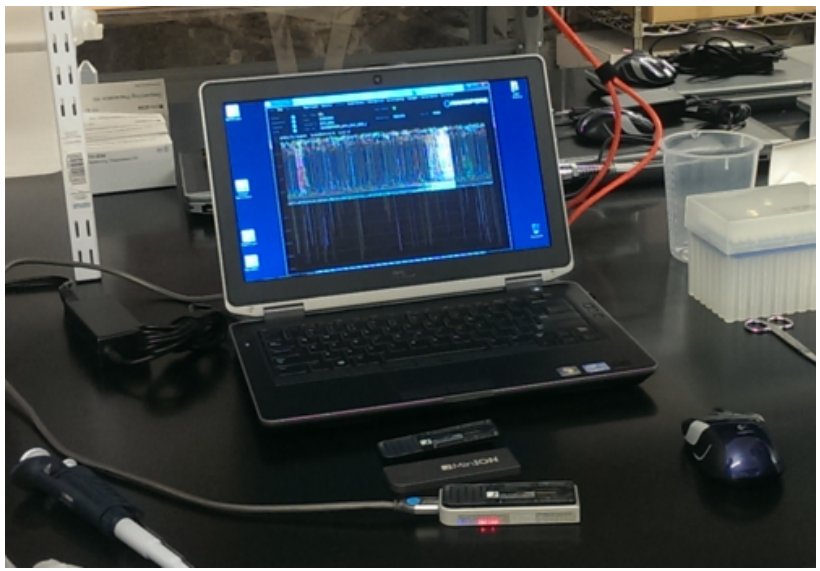


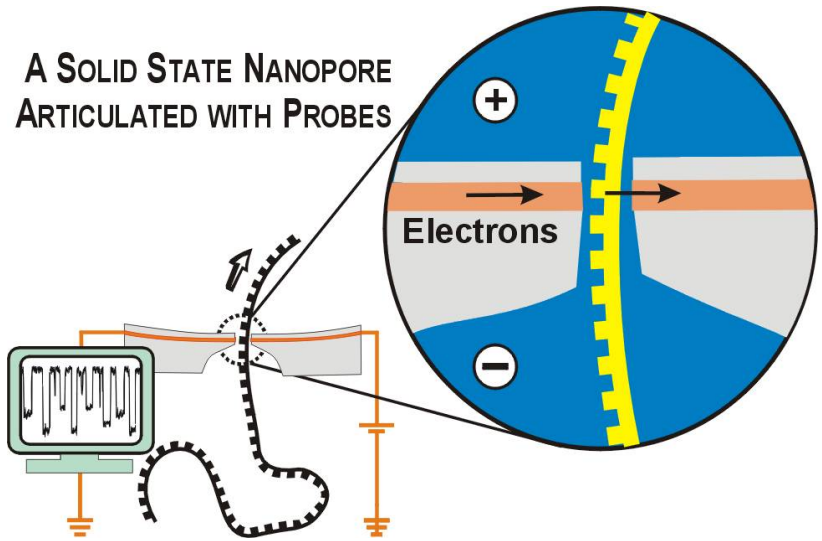


[1] http://sulab.org/2013/06/sequenced-genomes-per-year/

What's coming next?

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[2]

[2] Yaniv Erlich (2013) Future Continuous blog

Table of Contents

The James Hutton Institute

Bulk genome comparisons

The James Hutton Institute

You don't have to sequence genomes to compare them
(but it helps)
PART 2: in silico

Bulk genome comparisons

The James Hutton Institute

EXERCISE 1:

`ex01_gc_content.ipynb`

Nucleotide frequency/genome size

The James Hutton Institute

Very easy to calculate from complete or draft genome sequence

```
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
```

GC content, chromosome size can be characteristic of an organism.

Blobology [a] [b]

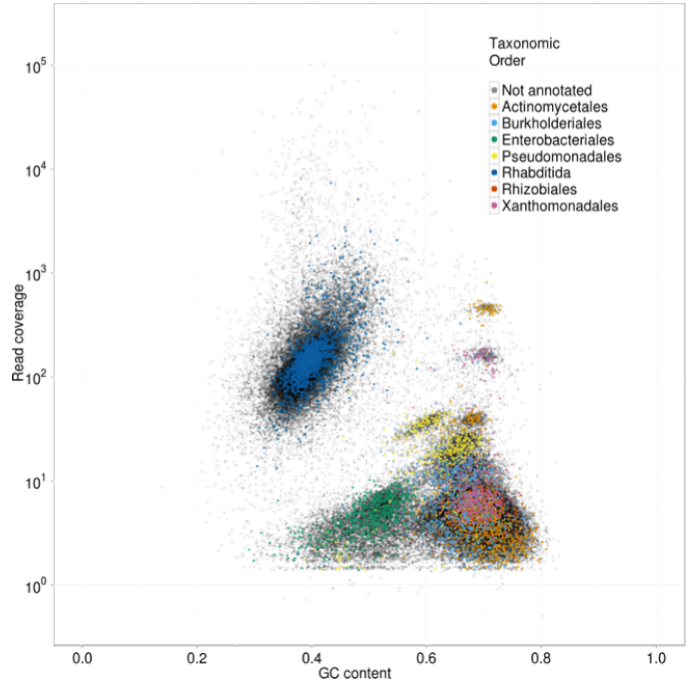
[a] Kumar & Blaxter (2011) *Symbiosis* doi:10.1007/s13199-012-0154-6
[b] http://nematodes.org/bioinformatics/blobology/

The James Hutton Institute

Sequence data can be contaminated by other organisms

- Host and symbiont DNA have different %GC
- Host and symbiont DNA differ in coverage

- Assemble genome
- Map reads
- Plot coverage against %GC



Nucleotide k-mers

The James Hutton Institute

Sequence data is necessary to determine k-mers/frequencies
Not possible by experiment

- Nucleotides, $k = 1$, 4x1-mers
  `A,C,G,T`
- Dinucleotides, $k = 2$, 16x2-mers
  `AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT`
- Triucleotides, $k = 1$, 64x3-mers
- k-nucleotides, 4^kxk-mers

Finding k-mers

The James Hutton Institute

ACTIVITY:

https://widdowquinn.shinyapps.io/nucleotide_frequencies/

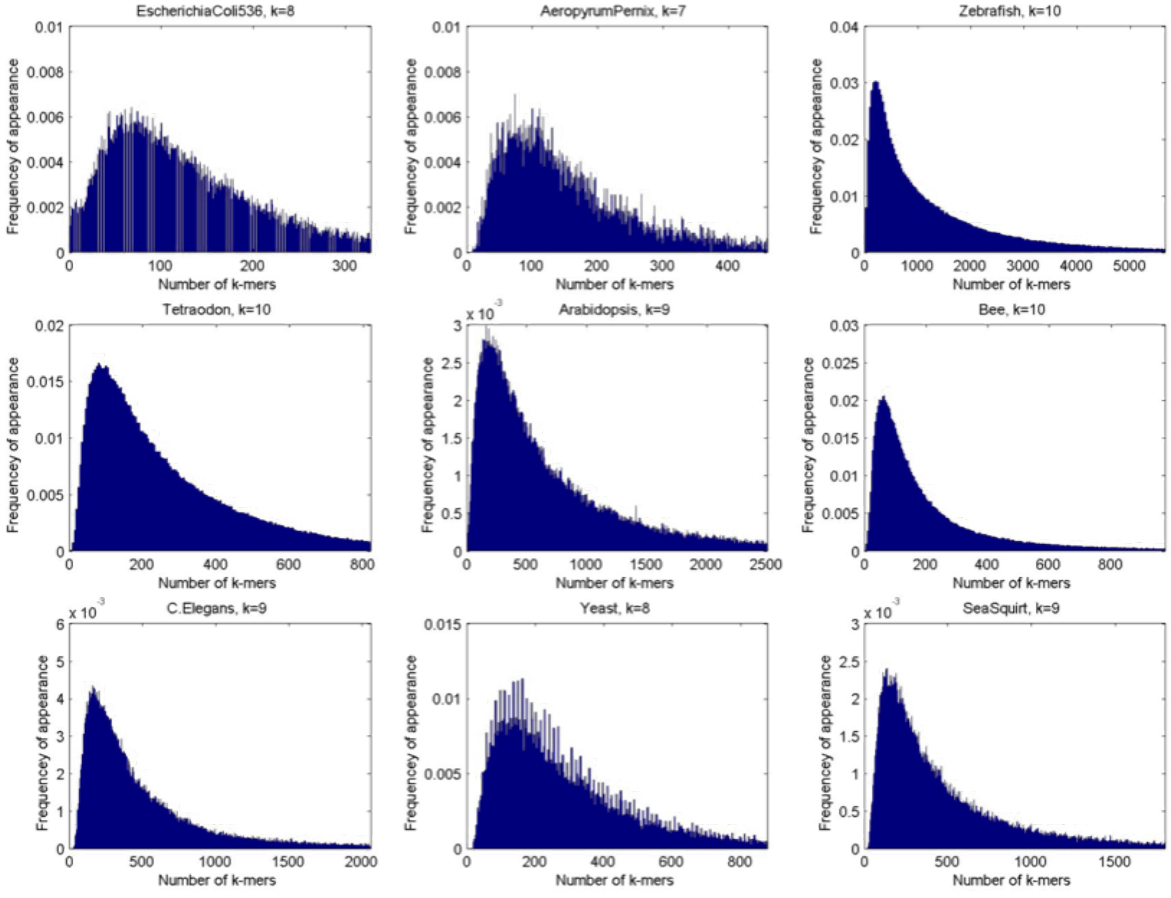
k-mer spectra [a]

[a] Chor *et al.* (2009) *Genome Biol.* doi:10.1186/gb-2009-10-10-r108

The James Hutton Institute

k-mer spectrum: frequency distribution of observed *k*-mer counts.

Most species have a unimodal *k*-mer spectrum



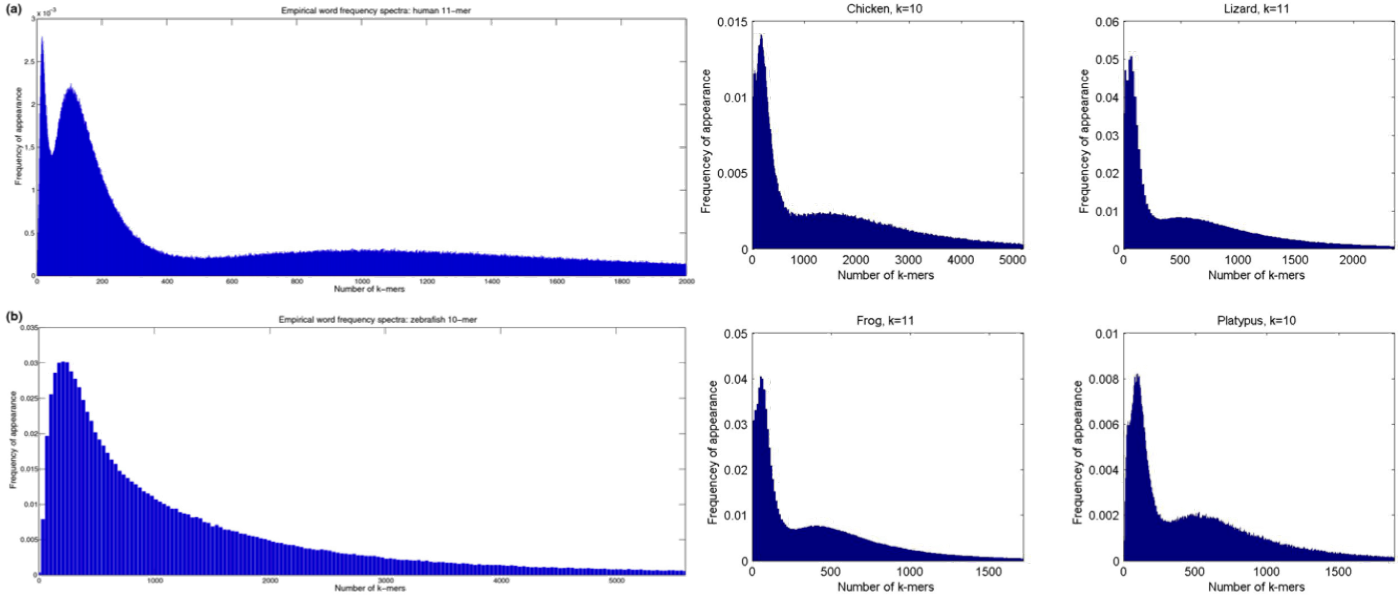
k-mer spectra [a]

[a] Chor *et al.* (2009) *Genome Biol.* doi:10.1186/gb-2009-10-10-r108

The James Hutton Institute

All mammals tested (and some other) species have *multimodal k-mer spectra*

Genomic regions differ in this property



Bulk genome comparisons

The James Hutton Institute

EXERCISE 2:

`ex02_kmer_spectra.ipynb`

Licence: CC-BY-SA

The James Hutton Institute

By: Leighton Pritchard