edinburgh genomics.

Existing and emerging sequencing technologies

Mick Watson

Head of Bioinformatics

Edinburgh Genomics

Edinburgh Genomics

- Edinburgh Genomics
  - Academic sequencing facility @ University of Edinburgh
  - Current technologies

6 x Illumina HiSeq 2500

3 x Illumina MiSeq

2 x ABI 3730

  - Previous technologies
    - Roche 454
    - ABI SOLiD

http://genomics.ed.ac.uk/

the gene pool

ARK-Genomics
Centre for Comparative & Functional Genomics

SEQUENCING BY SYNTHESIS

Sequencing by synthesis

Simple model

- Sample is fragmented, usually amplified
- Denatured to single stranded
- The technology then measure incorporation of bases into the second strand

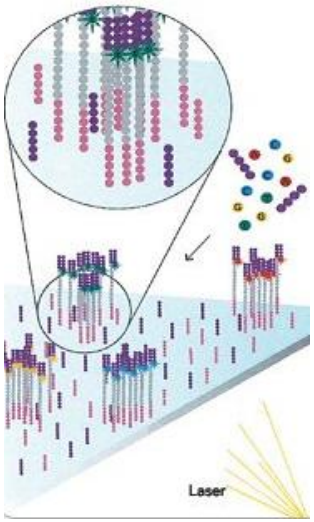
- The vast majority of sequencing data produced to date is from SBS technologies

edinburgh genomics.

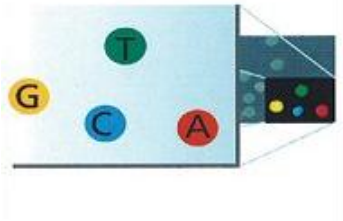
Illumina

- **The market leader**
- HiSeq 2500
  - 150 million reads per lane
  - Max read length 150bp x 2
- HiSeq 2000
  - 180 million reads per lane
  - Max read length 100bp x 2
- MiSeq
  - 15 million reads per lane
  - Max read length 250bp x 2

- Max output 45Gb per lane



Fluorescently labelled nucleotides are added



Laser captures image to determine first base

edinburgh genomics.

Ion Proton/Torrent

- **The challenger**
- Ion Proton/Torrent
  - Records production of H+ during SBS
  - Complex emulsion PCR library prep
- New technology
- Not much data publicly available
- Best ever throughput (Proton):
  - 86.6M reads ~ 200bp in length
  - 12.9Gb per run/chip
- Torrent: longer reads (~400bp), fewer reads (3 - 4 million)

life technologies™

edinburgh genomics.

Roche 454

- **The old workhorse**
- Roche 454
  - Records production pyrophosphate during SBS
  - Complex emulsion PCR library prep
- Earliest NGS
- Hasn't really kept up!
- FLX: 1 million ~400bp reads
- FLX+: 1 million ~800bp reads
- Junior: 80000 ~400bp reads

- Max output: ~ 1Gb

edinburgh genomics.

ABI SOLiD

- **The abject failure**
- ABI SOLiD
  - Records SBS in “colorspace”
  - Incorporation of dinucleotides recorded as colours
- Most labs have abandoned the technology in favour of Illumina citing:
  - Unreliability
  - Too short read length
  - Too little throughput

“We have seen little evidence that SOLiD data work well for genome assembly”

- Broad Institute

life technologies™

SINGLE MOLECULE SEQUENCING

Sequencing by synthesis

Even More Simple model

- The DNA molecules in the sample are detected directly
- Read using e.g. engineered polymerase or protein nanopore
- PCR-free so no bias is introduced
- Avoids “consensus” sequencing of SBS (sort of!)

Pacific Biosciences

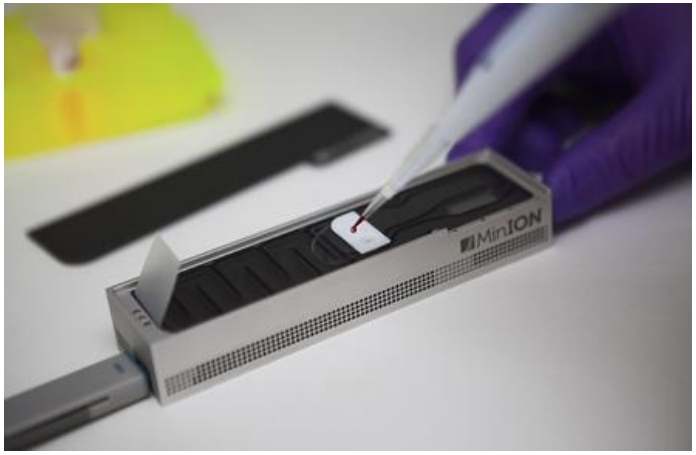
- **The Phoenix**
- Pacific Biosciences RS
  - Uses an engineered polymerase to read single molecules of DNA
- High raw error rate:
  - 15% errors, mostly indels
- Correction can improve this
  - 99% consensus accuracy
- Mean read lengths: 5 KB
- Maximum read lengths: ~30Kb
- No. Reads: 30-80000
- Throughput: 150Mb

THE FUTURE

edinburgh genomics.

Oxford Nanopore

- **The Fantasy**
- ONT Gridion
  - Uses a protein nanopore to read bases as they pass through
- Lower raw error rate:
  - claims 94% accurate
- Read lengths up to 100Kb!
- Gridion will have 2000 nanopores
- Each pore will read a molecule every 4 minutes
- Minion: tiny USB sequencer that sequences directly from blood



Moleculo

- **The Start-up**
- Moleculo
  - Single molecule fragments (~10Kb) are isolated in droplets
  - These single molecules are fragmented and tagged
  - All fragments sequenced by Illumina technology, data separated by tag and assembled
- Produces ~10Kb pseudo-reads, assemblies of 10Kb single molecules
- No data around

moleculo

illumina®

KNOWN ISSUES

Known Issues

- Known Issues
  - All SBS technologies introduce bias due to PCR step
  - Illumina traditionally has trouble with high GC
  - Ion Proton/Torrent has trouble with high AT
  - **Both** 454 and Ion Proton/Torrent suffer from mis-reading homopolymers
  - Pac Bio has high raw error rate, correction not always possible

OPINION

Disclaimer

Disclaimer

Everything after this slide is my opinion only, apart from the improvements that I know are coming to Illumina technology

16s sequencing

- 16s amplicon sequencing
  - Historically 454 has been the sequencer of choice. This will change. 454 produces 1M ~400bp reads
  - The MiSeq can achieve 15 million ~450bp reads (overlapping 2x250) for a fraction of the cost
  - 2x250bp is coming to the HiSeq 2500 – that's 150 million ~450bp reads for around ½ of 454 costs
  - 2x400bp reads are coming to the MiSeq
  - The PacBio read lengths raise the possibility of full length (2Kb) 16s sequencing

Whole genome metagenomics

- Whole genome metagenomics
  - There is only one party in town: Illumina
    - 45Gb of data for around £2k
    - 2x150bp produce decent assemblies
    - 2x250bp will produce even better
  - Lots of software to support Illumina metagenomics
    - MetaVelvet, Meta-IDBA/UD, MetAMOS, Ray-Meta

RUMEN METAGENOMICS

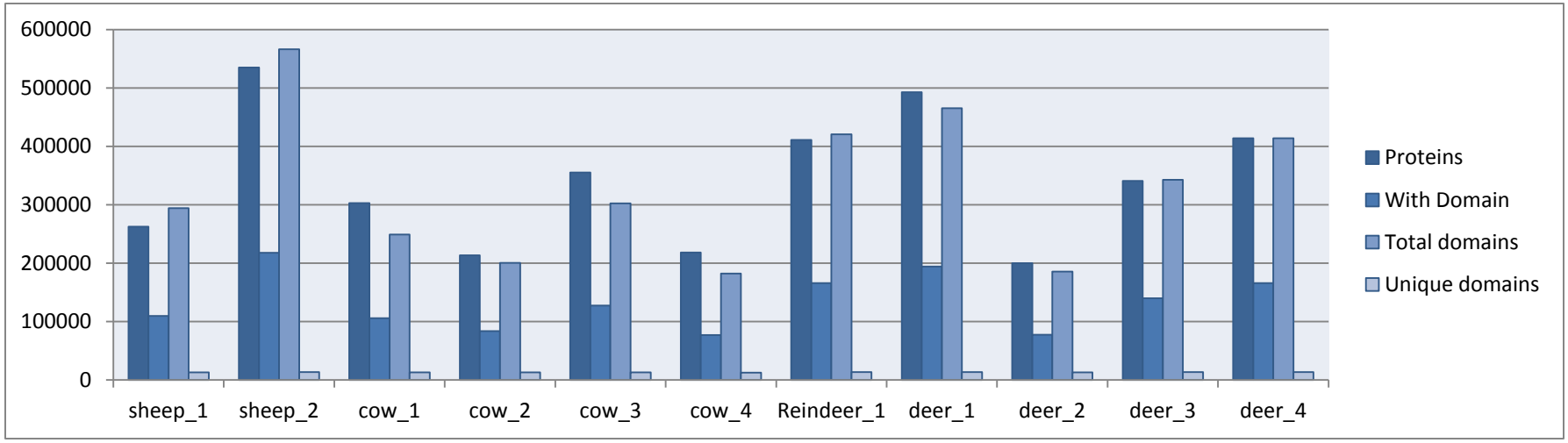
edinburgh genomics.

What did we sequence?

Sample	Desc	#Reads (millions)	Read type	Gbp
Ag2	Sheep, highland pasture	61.84	100x2	12.37
Bg2	Sheep, highland pasture	87.12	100x2	17.42
1099_C1	Cattle, maize sileage	56.60	100x2	11.32
1043_C2	Cattle, maize sileage	55.89	100x2	11.18
1033_C1	Cattle, maize sileage	63.60	100x2	12.72
983	Cattle, maize sileage	217.79	100x2	43.56
D1a	Red Deer, rough grazing	149.51	150x2	29.90
D2a	Red Deer, rough grazing	125.77	150x2	25.15
D3b	Red Deer, rough grazing	171.13	150x2	34.23
D4b	Red Deer, rough grazing	160.55	150x2	32.11
R1b	Reindeer, Summer Pasture	149.40	150x2	29.88
R2b	Reindeer, Summer Pasture	209.29	150x2	41.86
				301.70

edinburgh genomics.

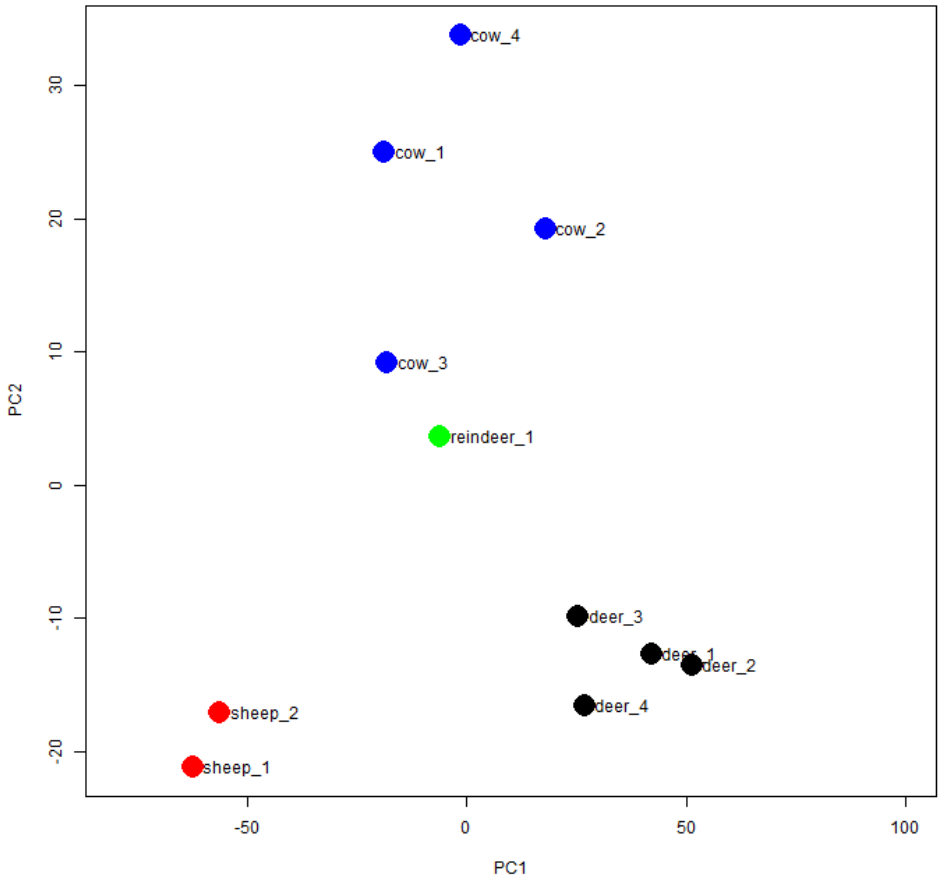
Gene predictions and domains

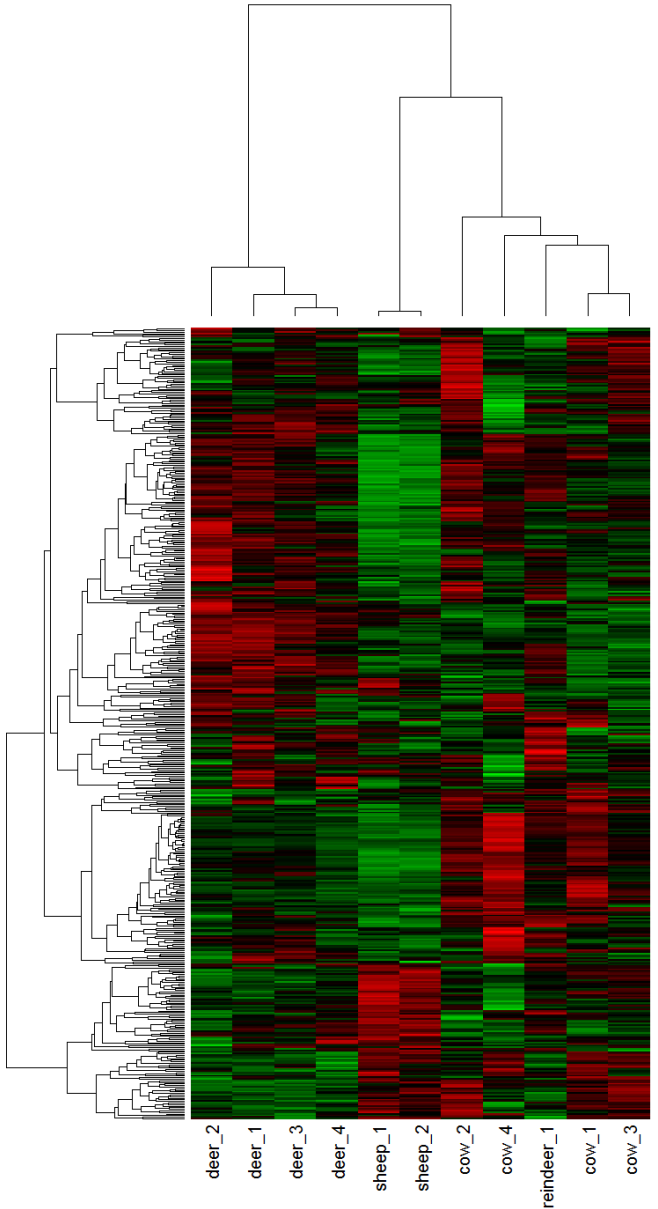


Sample	Proteins	With Domain	Total domains	Unique domains
sheep_1	262578	109432	294242	12972
sheep_2	534761	217719	566000	13517
cow_1	302624	105701	248925	13072
cow_2	213662	83562	200267	13031
cow_3	355222	127535	302140	13298
cow_4	218302	76966	182265	12723
Reindeer_1	411158	165709	420309	13566
deer_1	492563	194275	465101	13572
deer_2	199967	77375	185724	13017
deer_3	340798	139906	342889	13477
deer_4	414010	165756	413926	13540
	3745645	**1463936**	**3621788**	**145785**

edinburgh genomics.

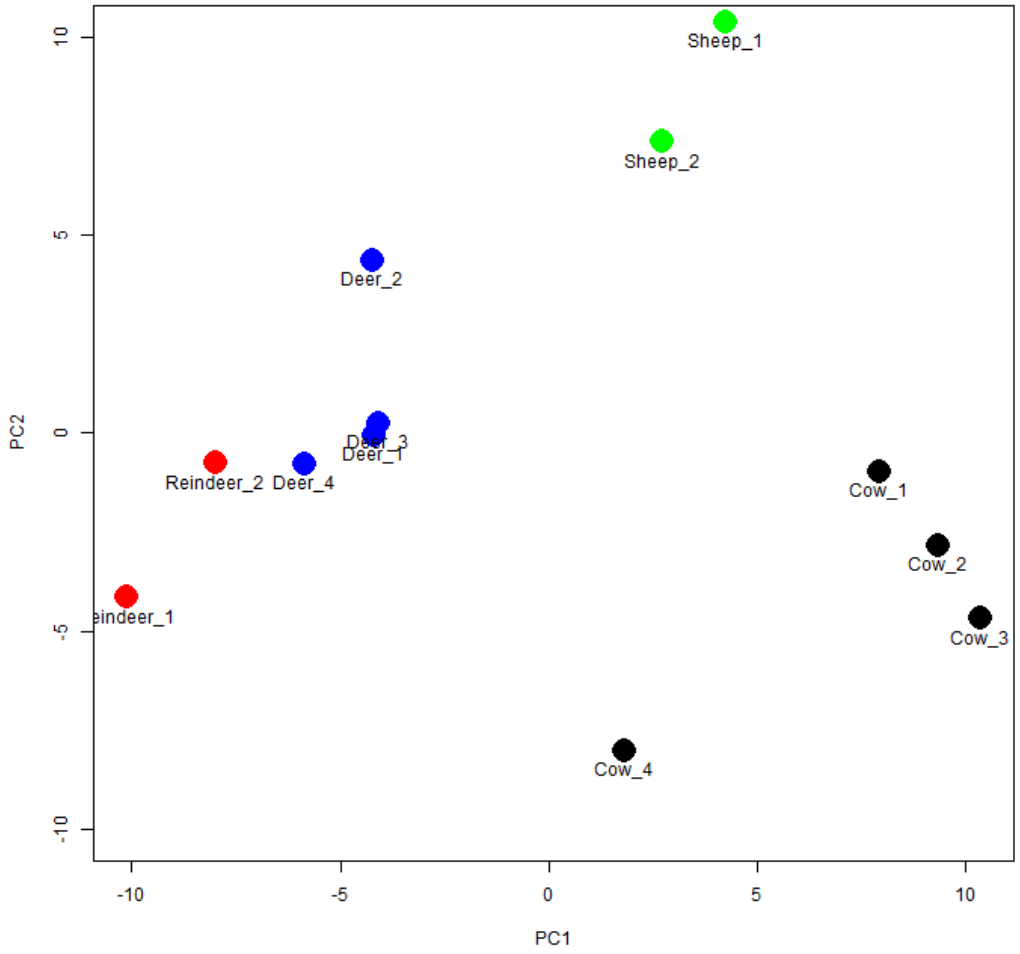
Clustering of Pfam clans

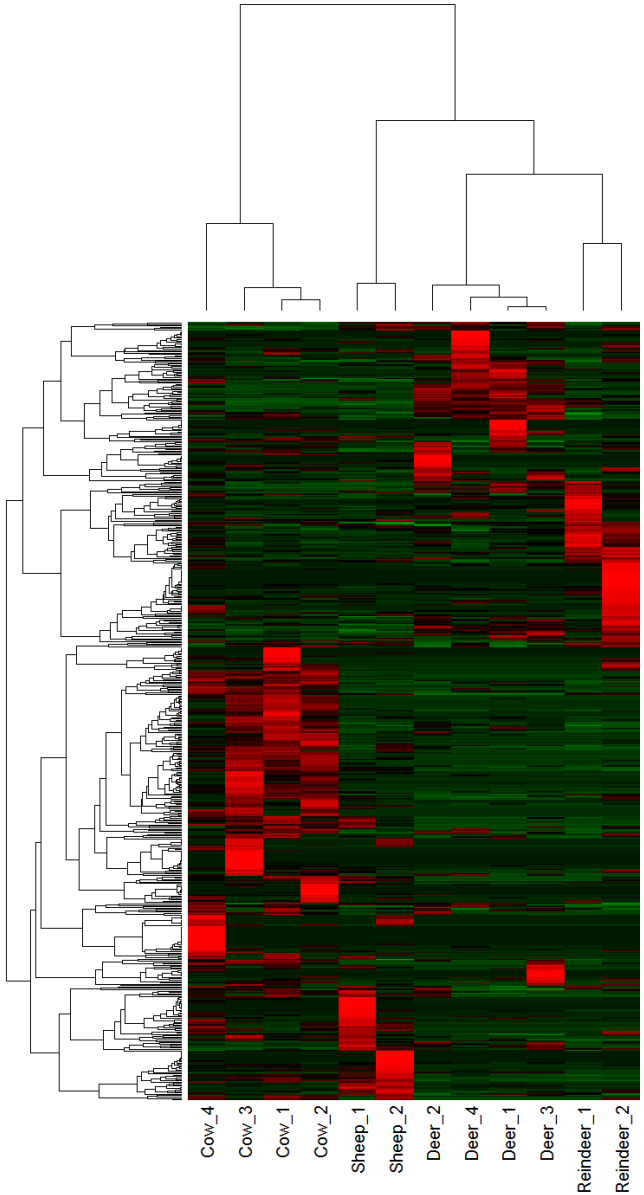




edinburgh genomics.

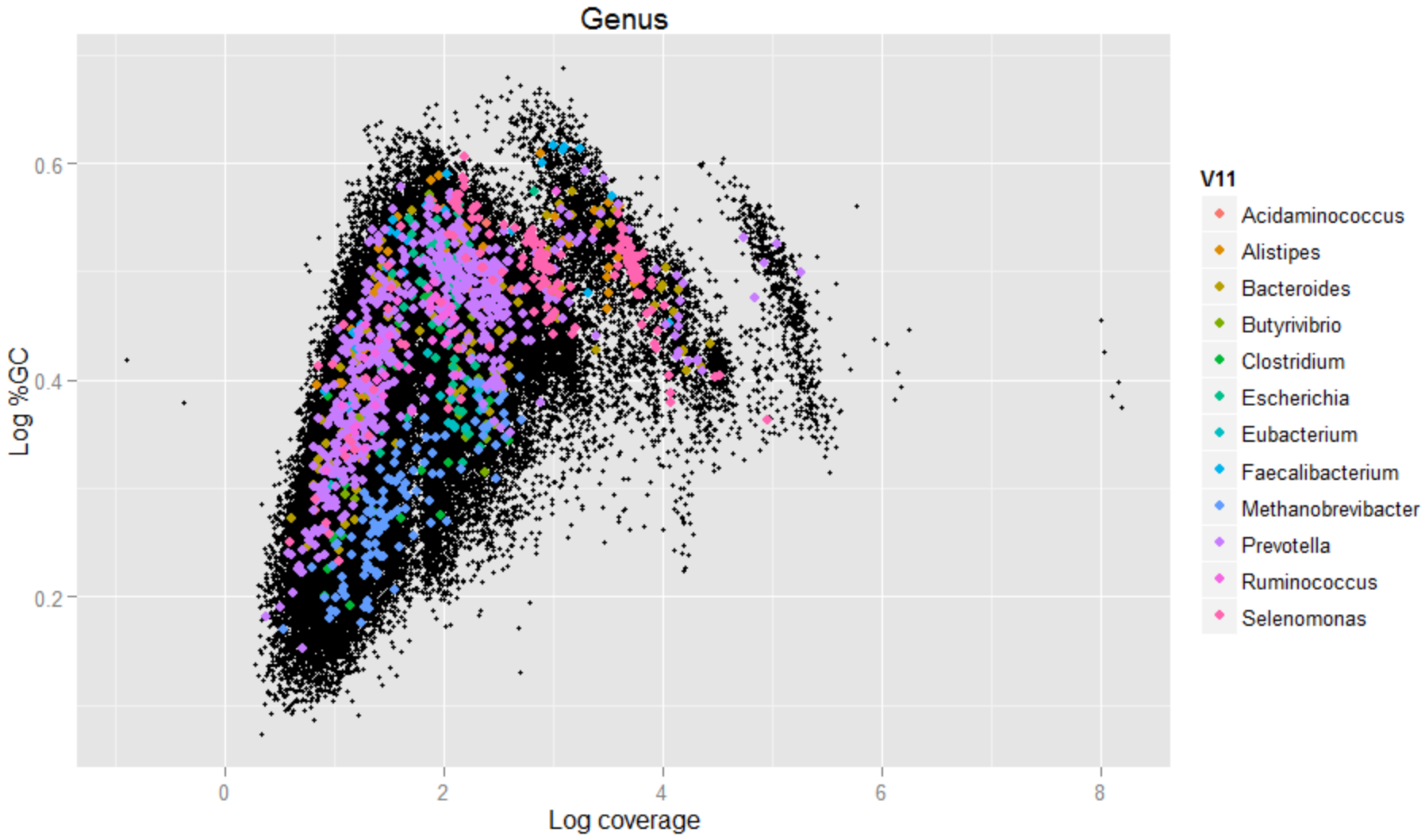
Clustering of known taxonomy





edinburgh genomics.

Known genera?



edinburgh genomics.

Acknowledgements

- My sequencing team
  - Richard Talbot
  - Karen Troup
  - Pablo Fuentes-Utrilla
- My bioinformatics team
  - Me
  - Frances Turner
  - Julia Loecherbach
- My collaborators
  - John Wallace
  - Ian Fotheringham
  - Franck Escallettes
- My funders
  - Technology strategy board
  - BBSRC
- My employers
  - The Roslin Institute
  - The University of Edinburgh

@BioMickWatson

biomickwatson.wordpress.com

genomics.ed.ac.uk

www.ark-genomics.org