Next Generation Sequencing – The Role of New Sequence Technologies in Shaping the Future of Veterinary Science

Hosted by the RCVS Charitable Trust

RCVS

charitable trust

ROSLIN

ARK-Genomics
Centre for Comparative & Functional Genomics

Sequencing the rumen microbial population (the microbiome)

Opportunities for biotech and the environment

Mick Watson
Director of ARK-Genomics

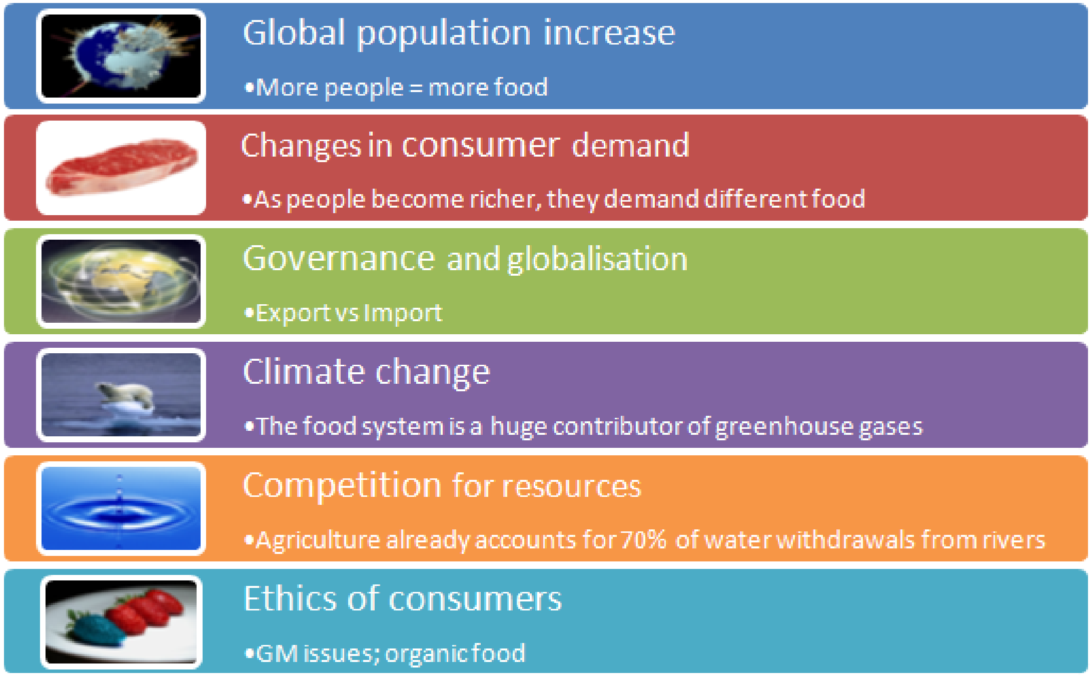
THE UNIVERSITY *of* EDINBURGH

BBSRC
bioscience for the future

FOOD SECURITY

Challenges in food security

- The World's food system doesn't work:
  - 1.5bn overweight, 500m obese[1]
  - 925m experience hunger, +1bn "hidden hunger"[2]
- Moving forward, there are a number of key pressures:



Global population increase
- More people = more food

Changes in consumer demand
- As people become richer, they demand different food

Governance and globalisation
- Export vs Import

Climate change
- The food system is a huge contributor of greenhouse gases

Competition for resources
- Agriculture already accounts for 70% of water withdrawals from rivers

Ethics of consumers
- GM issues; organic food

[1] WHO [http://www.who.int/mediacentre/factsheets/fs311/en/]

[2] Foresight report "The Future of Food and Farming: Challenges and choices for global sustainability"

Over the next 50 years, the world's farmers and ranchers will be called upon to produce more food than has been produced in the past 10,000 years combined, and to do so in environmentally sustainable ways.

Jacques Diouf, FAO Director General, 2007

ROSLIN AND ARK-GENOMICS

The Roslin Institute

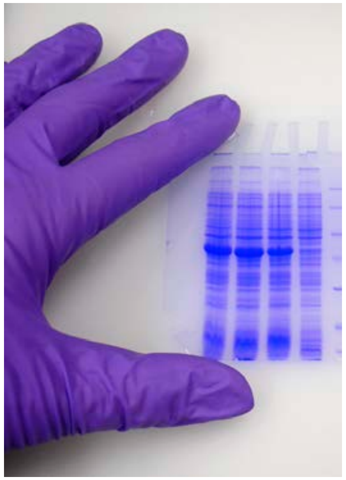
BBSRC
bioscience for the future

LIVESTOCK GENETICS

Bioscience underpinning health

ANIMAL HEALTH

BIOTECH



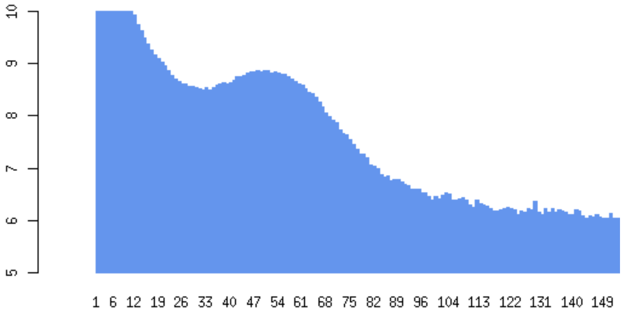
HUMAN HEALTH

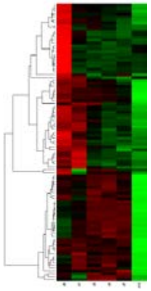
Food security

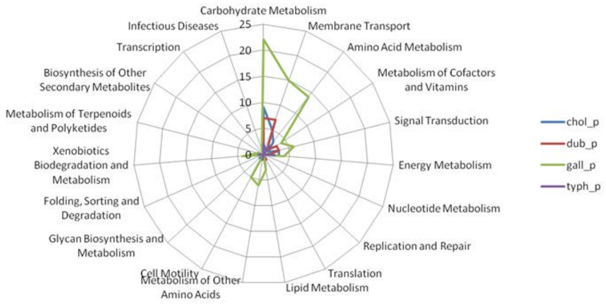
- ARK-Genomics
  - High-throughput facility focusing on the genetics and genomics of animals
  - Based at the Roslin Institute, University of Edinburgh
  - Offering research, collaborations and service provision
  - Investing in the latest genomics technologies
    - Sequencing
    - Genotyping
    - Transcriptomics
    - Comparative Genomics
    - Bioinformatics

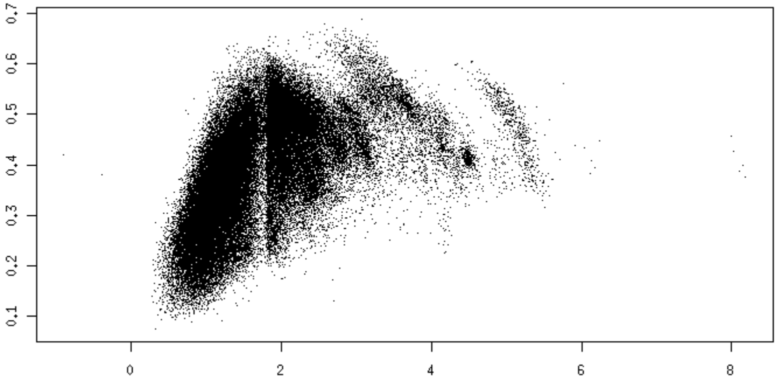
Current Research

- Virus discovery
- Pathogen genomics
- Host genomics
  - Re-sequence: Chicken
  - New: e.g. Falcon, Elephant
- Host-pathogen interactions
- Metagenomics
- Industrial Biotechnology









THE RUMINANT GUT MICROBIOME

Prevailing theory of the individual

- An individual consists of at least 10x as many bacterial cells as "host" cells
- Each individual is a "supra-organism"
  - a composite of host and microbial cells contribute the functions necessary for the individual to survive
- The genetic landscape of any individual is a composite of the host genome and the genomes of the millions of microbial symbionts that live on and within that individual
- It is clearly important to take a holistic view when examining any animal phenotype.

Why study it?

- Energy from food

"Our results indicate that the obese microbiome has an increased capacity to harvest energy from the diet. Furthermore, this trait is transmissible: colonization of germ-free mice with an 'obese microbiota' results in a significantly greater increase in total body fat than colonization with a 'lean microbiota'"

Turnbaugh et al (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444(7122):1027-31

- Novel enzyme discovery

"An initial assembly of the metagenomic sequence resulted in 179,092 scaffolds... Only 47 (0.03%) of the assembled scaffolds showed high levels of similarity to previously sequenced genomes available in GenBank. These results suggest that the vast majority of the assembled scaffolds represent segments of hitherto uncharacterized microbial genomes."

Hess M et al (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 331(6016):463-7.

Methane emissions

- Globally, ruminant livestock produce about 80 million metric tons of methane annually, accounting for about 28% of global methane emissions from human-related activities.
- With about 100 million cattle in the U.S. and 1.2 billion large ruminants in the world, ruminants are one of the largest methane sources.
- In the U.S., cattle emit about 5.5 million metric tons of methane per year into the atmosphere, accounting for 20% of U.S. methane emissions

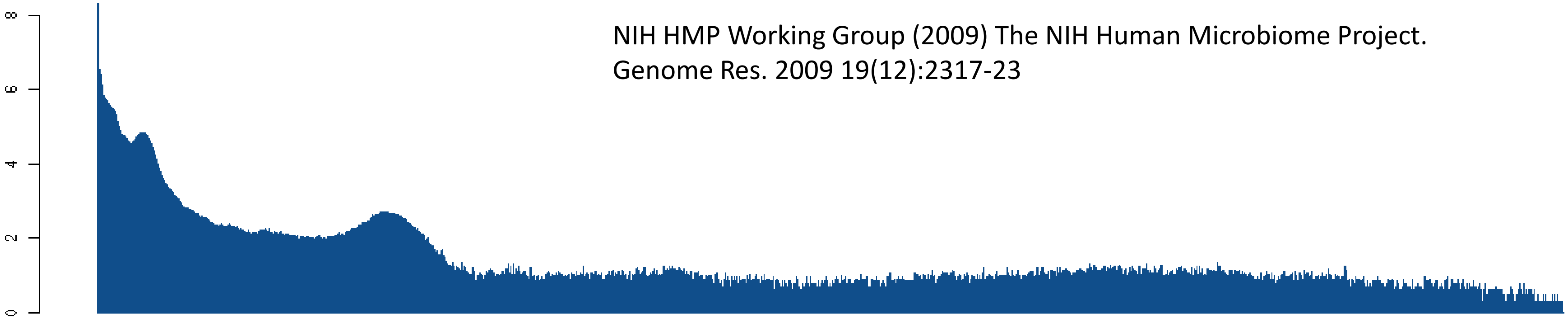
- It’s not the animal – it’s the rumen methanogenic bacteria

METAGENOMIC ASSEMBLY

What does a metagenomic sample look like?

NCBI SRA:SRR041654

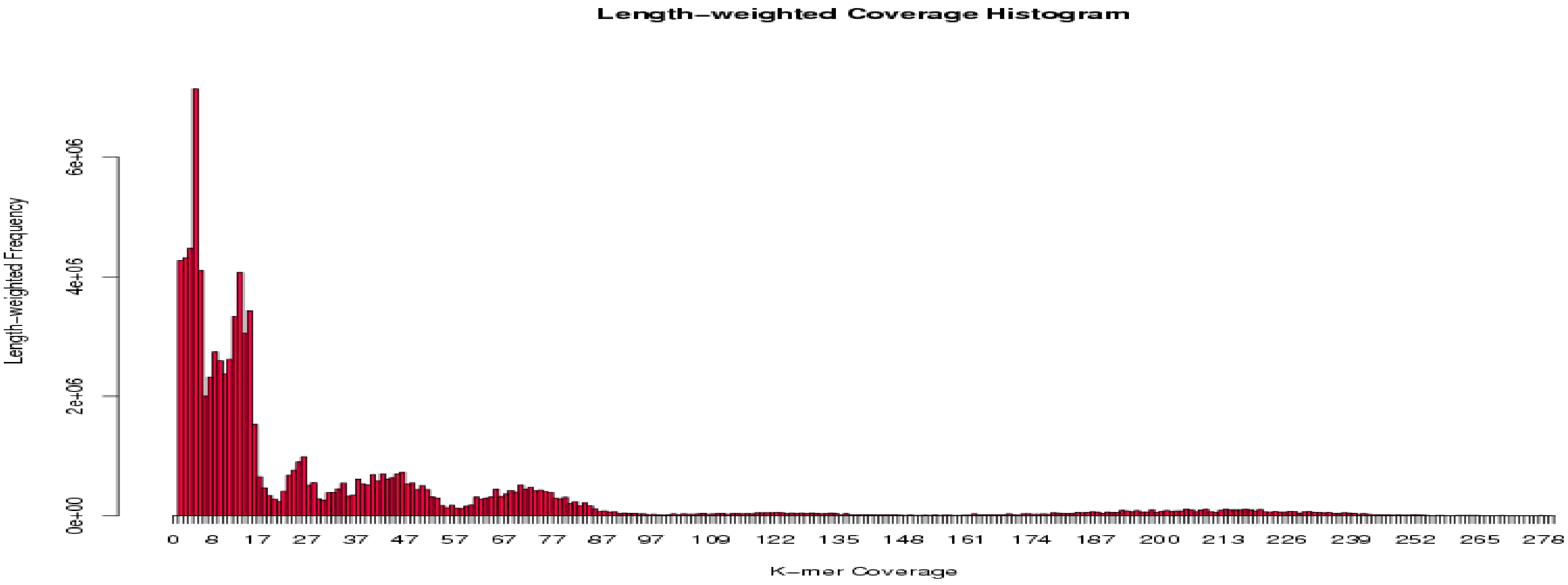
NIH HMP Working Group (2009) The NIH Human Microbiome Project.
Genome Res. 2009 19(12):2317-23



- Kmer coverage graph
- Y-axis is log10
- X-axis from 1 to 6000
- Several sub-populations of kmer can be seen
- Cannot differentiate low frequency kmers from errors

MetaVelvet

- Assembly of metagenomic samples:
  - Namiki T *et al* (2011) MetaVelvet : An extension of Velvet assembler to de novo metagenome assembly. Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine, New York, NY, USA



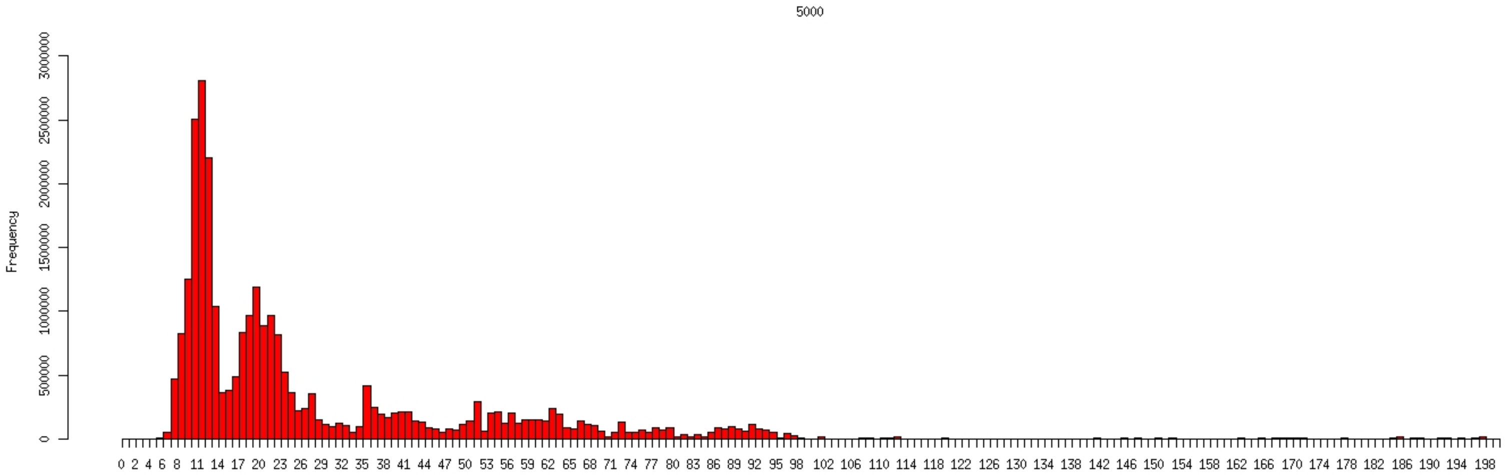
RUMEN METAGENOMIC ASSEMBLY

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

Assembly protocol

- Trim reads to Q30 (sickle)
- Assemble using Velvet
- Manual inspection of coverage peaks
- Re-assemble using MetaVelvet
- At this stage, no optimisation for K (used K:51)



Assembly stats

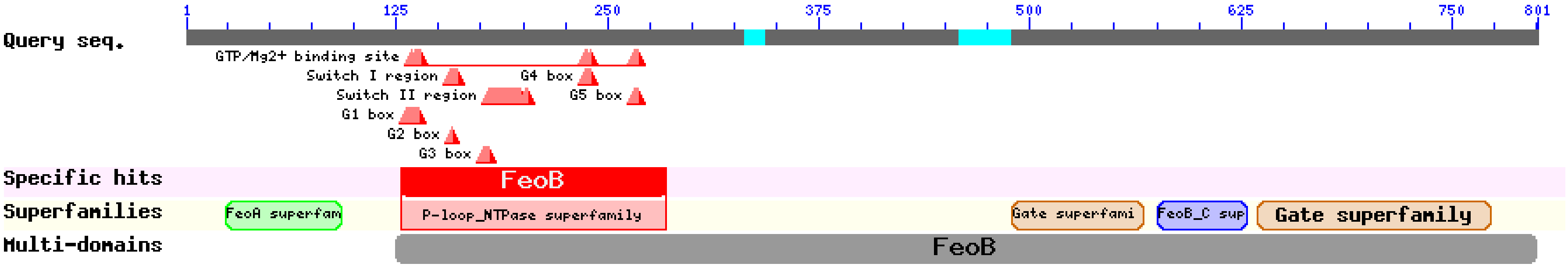
	Contigs > 1000bp				Contigs > 500bp			
Sample	**N50**	**Total**	**Number**	**Max**	**N50**	**Total**	**Number**	**Max**
Ag2	2502	171080118	73968	250047	1451	267044905	241461	250047
Bg2	2620	359972055	153624	152301	1525	553909015	499548	152301
1099_C1	1518	107617445	68547	53793	784	290103096	405130	53793
1043_C2	1623	50054937	29157	54895	530	238805983	441475	54895
1033_C1	1604	129661930	77631	89904	805	330320409	448607	89904
983	1432	54430150	35961	37263	656	222333169	364693	37263

- Fragmented assemblies, typical of metagenomics
- Likely lots of low-coverage genomes

“GENE PREDICTION”

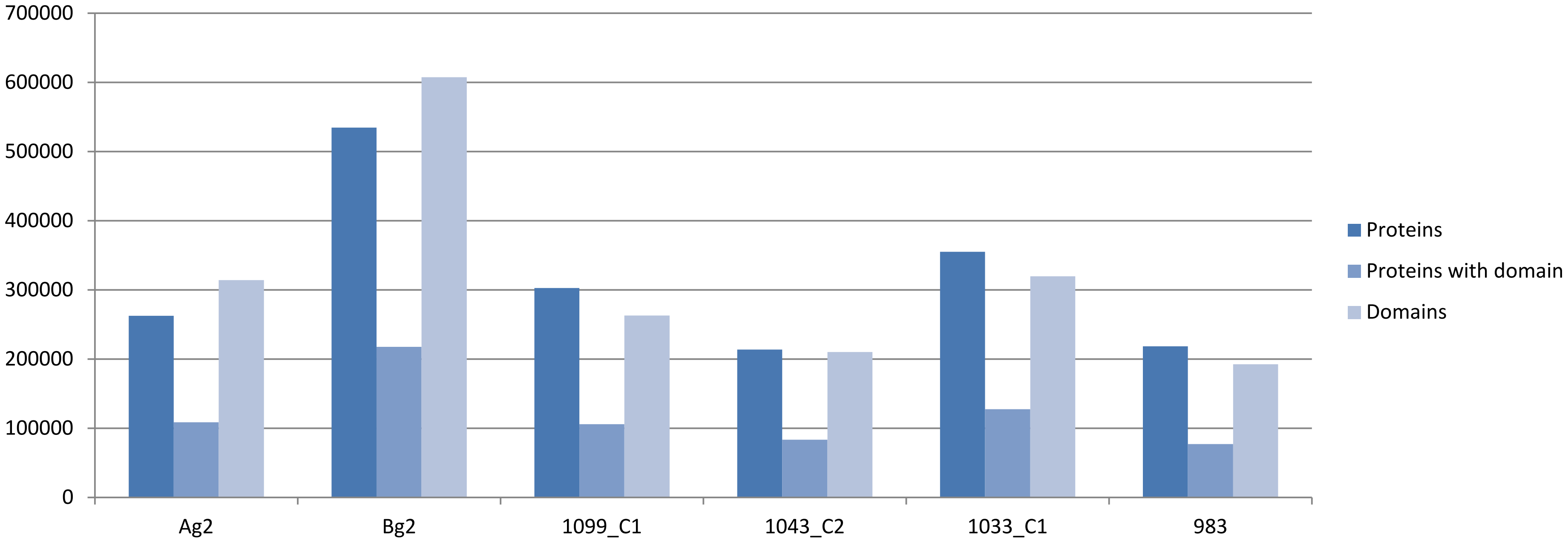
Gene prediction protocol

- Extracted long ORFs (> 200bp)
- Translate
- Compare to Pfam
  - Uses pfam_scan.pl -> hmmpfam (HMMER)
- Typical output: 801aa protein



- Involved in Fe transport
- 54% identical, 72% positive to previously sequenced protein
  - ferrous iron transporter B [*Odoribacter laneus*]

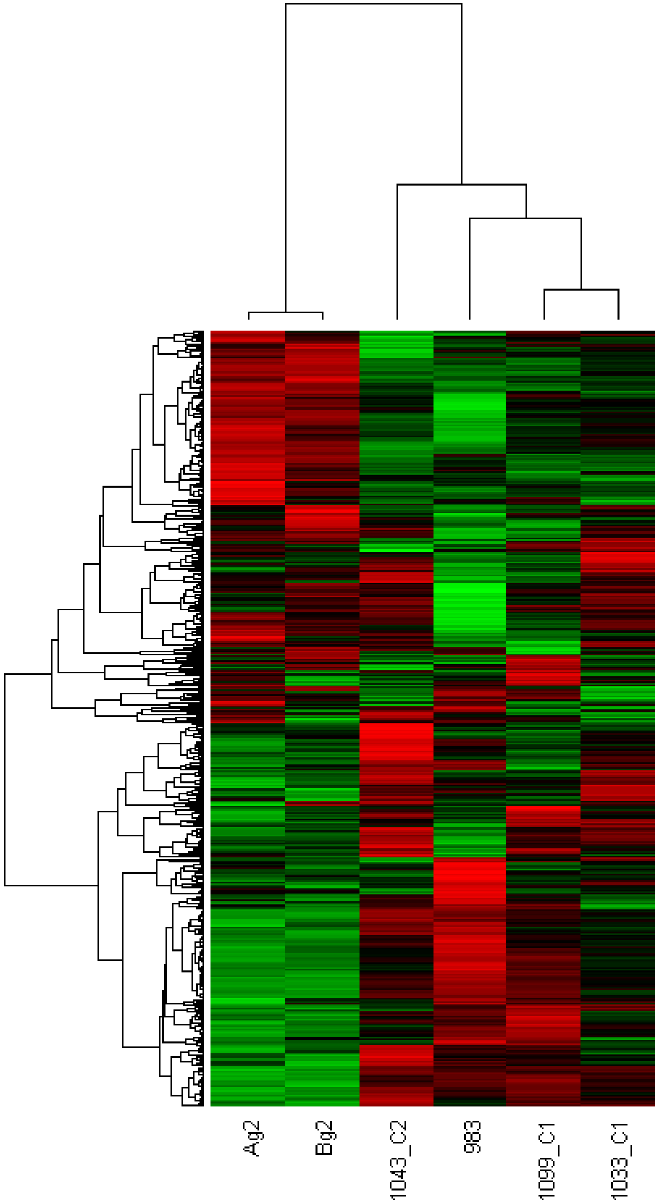
Gene predictions and domains



Sample	Proteins	Proteins with domain	Domains
Ag2	262578	108760	314117
Bg2	534761	217774	607496
1099_C1	302675	105834	262967
1043_C2	213664	83611	210409
1033_C1	355262	127638	319642
983	218392	77069	192464
	1887332	**720686**	**1907095**

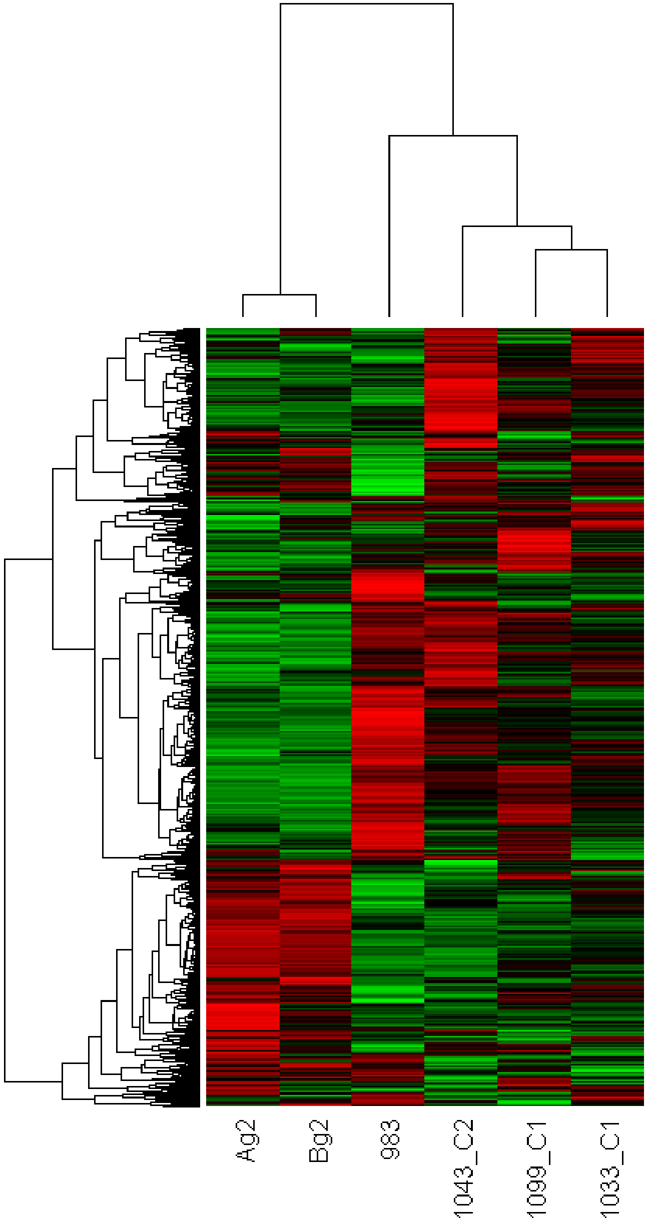
Clustering of Pfam clans:

- Clans are collections of Pfam families
- Method:
  - Count protein hits against each Pfam clan
  - Normalise to the total number of clans hit per sample
  - Cluster based on correlation matrices

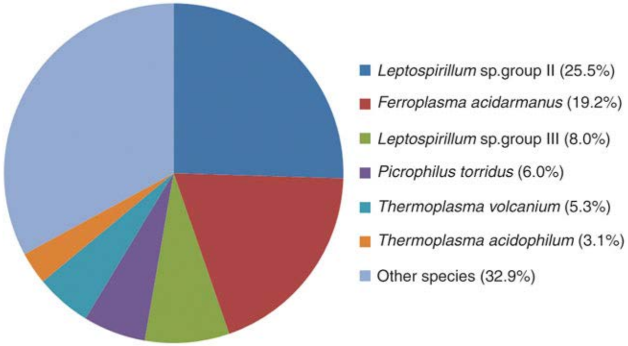


Clustering of Pfam families:

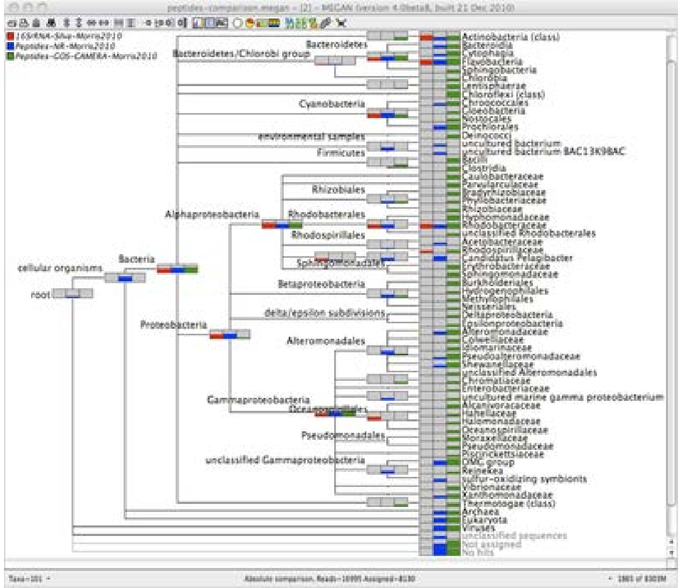
- Families are collections of Pfam domains
- Method:
  - Count protein hits against each Pfam families
  - Normalise to the total number of families hit per sample
  - Cluster based on correlation matrices



Taxon assignment



- Computationally difficult
  - What is the query?
  - What is the database?
- In the query we have 100M x 2 reads
- There are over 2000 genomes completed/draft
  - Under-representative of our dataset
- Many use “nr”
  - 17M sequences
- 200M x 17M sequence comparison
  - Not feasible



Basic approach

- Don’t assign the reads, assign the assembly!
- Searching ~100,000 sequences rather then millions!
- What is your cut-off? Using megablast, require
  - HSP of at least 100bp
  - % identity of 80%

	Sample	N50	Total	Number	Max	Hits	%
557_1	Ag2	2502	171080118	73968	250047	5867	7.93
557_2	Bg2	2620	359972055	153624	152301	12770	8.31
557_3	1099_C1	1518	107617445	68547	53793	4842	7.06
557_4	1043_C2	1623	50054937	29157	54895	2963	10.16
557_5	1033_C1	1604	129661930	77631	89904	6445	8.30
557_6	983	1432	54430150	35961	37263	1954	5.43

DISCUSSION, CONCLUSIONS

Rumen metagenomics

- It is possible to assemble contigs from deep sequencing of rumen microbiomes
- Even with deep-sequencing, there are many genomes at low coverage -> fragmented assembly
- It is possible to extract novel proteins/enzymes and predict domains/functions
- Sheep and cow microbiomes cluster separately according to their protein domain content
- The vast majority of the genomic landscape is novel – most contigs don’t hit anything known
- There is a huge potential for discovery using metagenomics approaches

Acknowledgements

- ARK-Genomics
  - Richard Talbot
  - Sarah Smith
  - Karen Troup
- Funders
  - BBSRC
  - TSB

BBSRC
bioscience for the future

Technology Strategy Board
Driving Innovation

- Rowett/Aberdeen
  - John Wallace

www.ark-genomics.org

Next Generation Sequencing – The Role of New Sequence Technologies in Shaping the Future of Veterinary Science

Hosted by the RCVS Charitable Trust

RCVS

charitable trust