www.csiro.au

Challenges analysing modern genomics data

Bill Wilson
CSIRO Mathematical and Information Sciences
Keio University :: Workshop on Data Science
24-27 March 2009

Australian Government

豪日交流基金
Australia-Japan FOUNDATION

CSIRO

Today

- Who are we?
- What are our skills?
- What do we work on?
- Current challenges…

Our Research Focus

- Multivariate statistics :: large biologically relevant datasets :: manageable & meaningful results
  - Gene expression microarray data
    - Affymetrix gene expression arrays
  - Whole genome (ie: large) genotyping data
    - Affymetrix Single Nucleotide Polymorphism arrays
- Statistical modelling of biological data
  - Gene expression outlier detection
  - Raw SNP probe data modelling
  - Genome wide Copy Number Variant (CNV) detection
- New methods (computational and statistical) for Gigabase sequencing data
  - QC
  - Getting more from the data

Our collaborations

Preventative Health flagship

- A focus on array technologies (transcription, SNPs, tiling)
- Statistical modelling, machine learning / Bayesian approaches
- P >> n analyses (multivariate statistics, classification methods)
- Experimental design

Others

- High throughput sequence analysis
  - No assembly
  - metagenomics
- Sequence Space Arrays

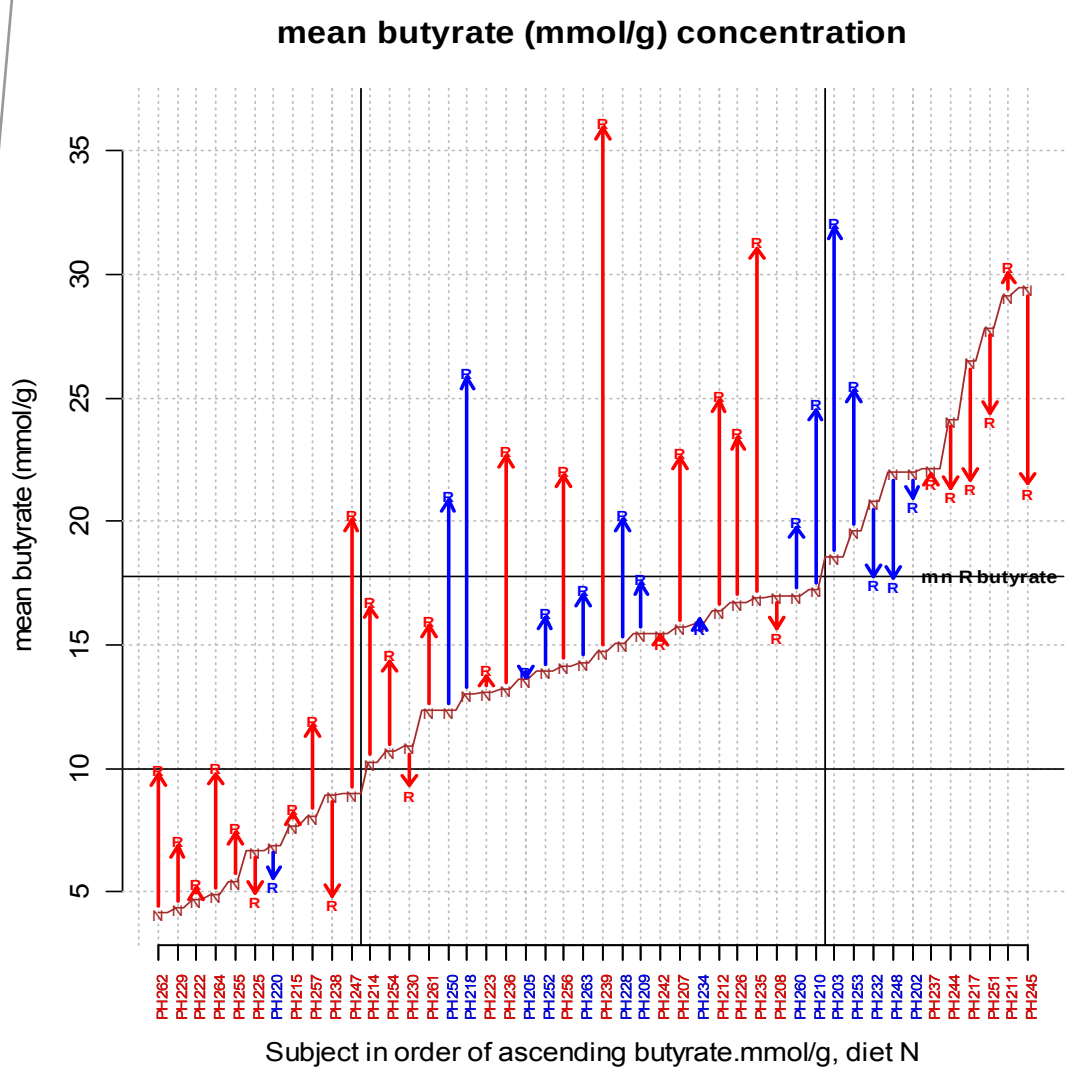
In Future

- Closer work with bench scientists
- Mutually beneficial collaborations

Cutting edge solutions for modern molecular data

Australian Government | 豪日交流基金 Australia-Japan FOUNDATION | CSIRO

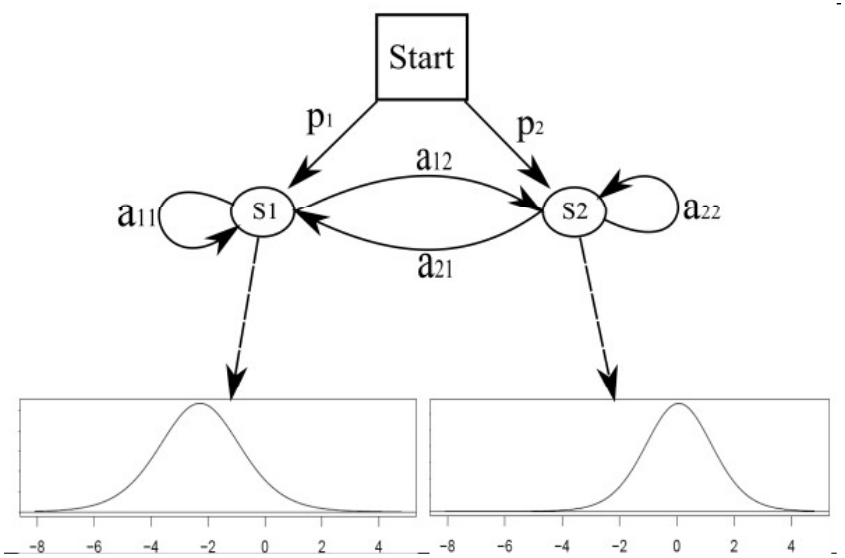
Butyrate variation in Humans



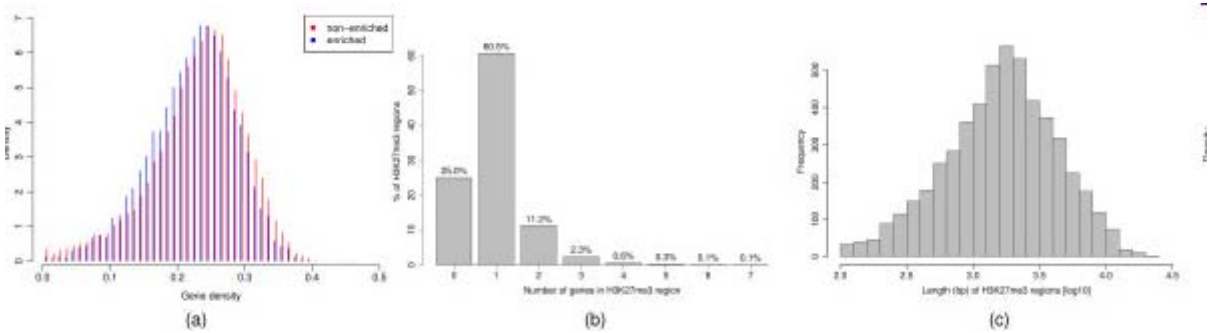
- Link between apoptosis (cell death) and [butyrate]
- Produced by bacteria in the gut fermenting starch
- Relationship to diet
- Relationship to weight
- Changes with resistant starch diets

Australian Government | 豪日交流基金 Australia-Japan FOUNDATION | CSIRO

Nucleosome positioning



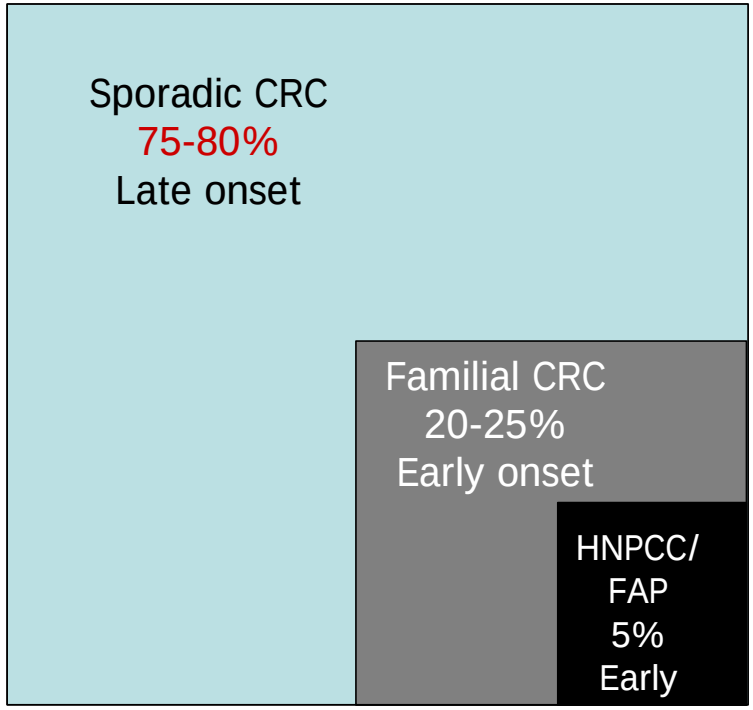
- Can sequence data be used to identify nucleosome positioning on DNA?
- Use positional information to correlate gene expression during biological processes



Australian Government
豪日交流基金 Australia-Japan FOUNDATION
CSIRO

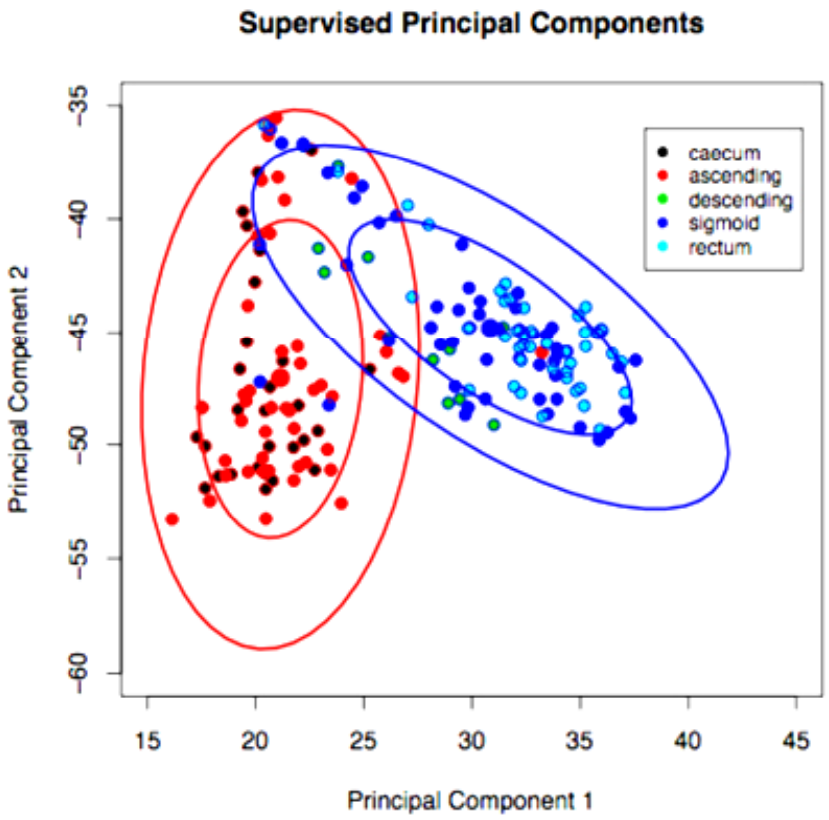
Colorectal Cancer

- About 25% of CRCs are in younger (<55) individuals or with a family history of CRC, suggesting a heritable susceptibility.
- If CRC is caught early enough it is almost completely curable



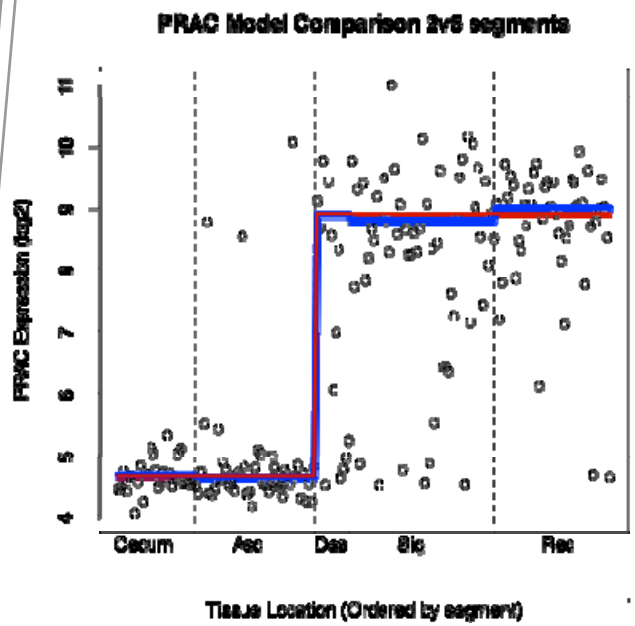
- J.P. Terdiman *et al.* (1999) AJG 94, 2344-2356.

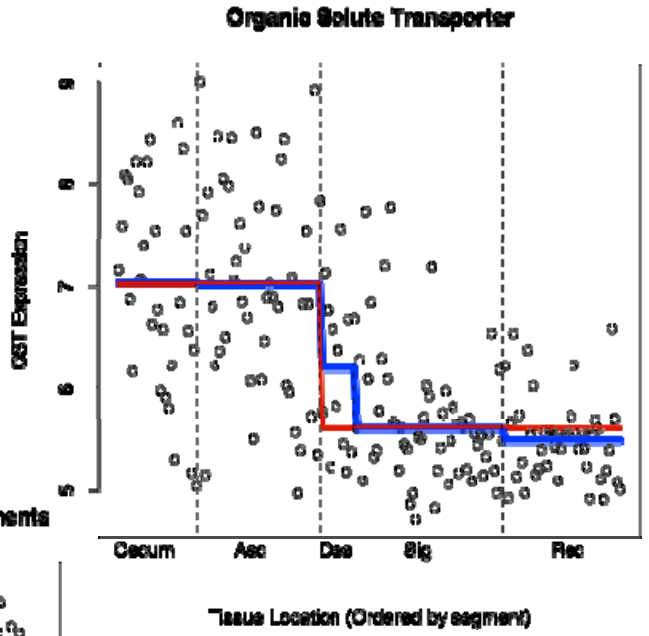
Our analysis - Data Structure

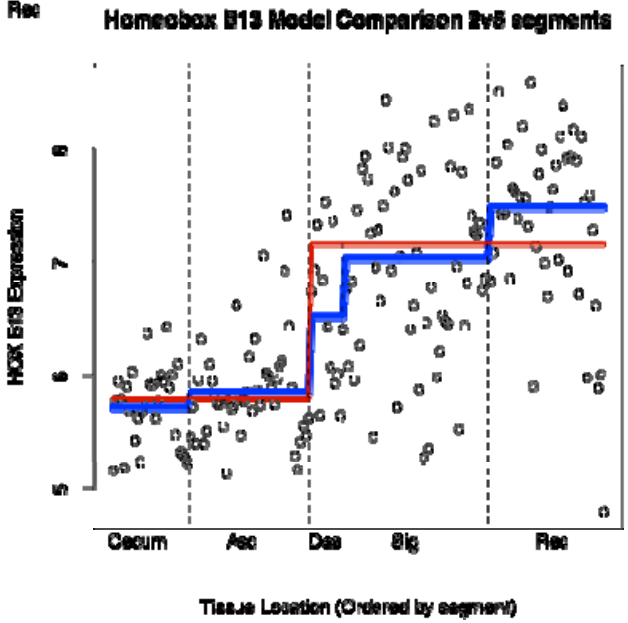


- Analysis of gene expression in the colon
  - Based on gene expression from 23000 probesets, the left and right colon appear to be quite different

Gene expression in the colon





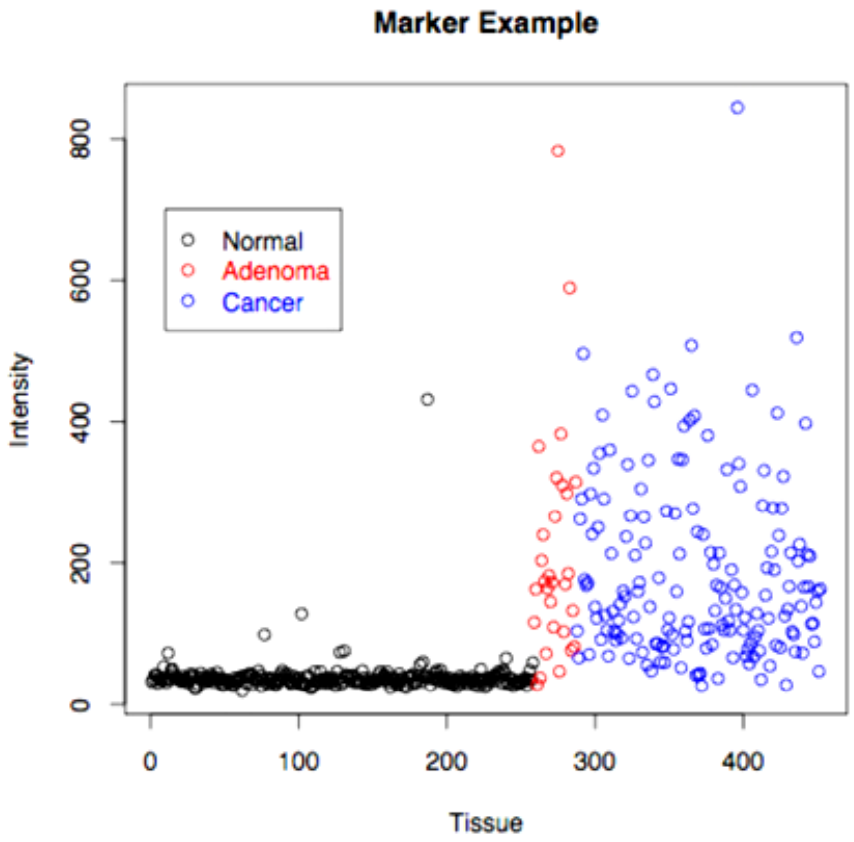


Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

Our analysis - Marker Discovery



- Analysis of gene expression in the colon
  - A molecular marker for disease?
  - This gene separates adenomas from normal tissue … but not adenomas from cancer.

SNP array data can be used for copy number

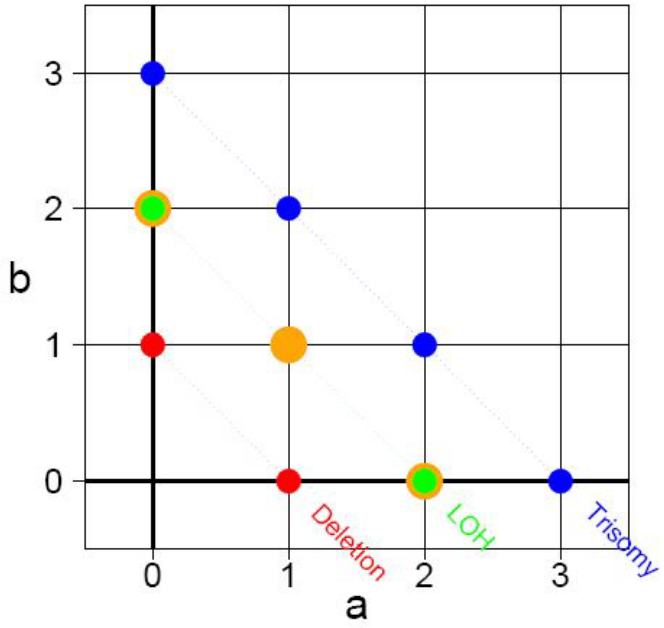
- **Genotyping**
  - AA
  - AB/BA
  - BB

  - Identification of alleles, but not number of copies

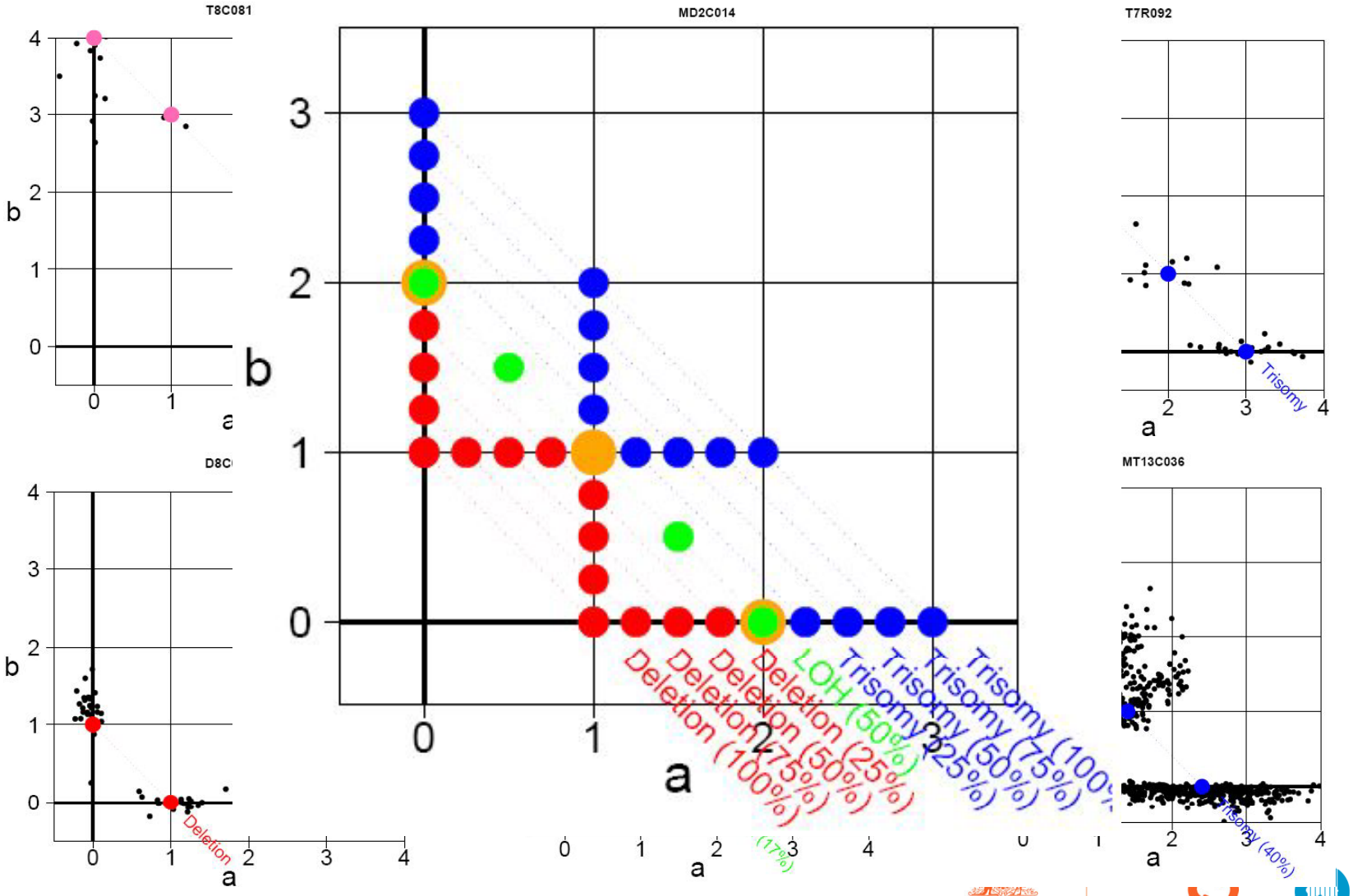
- Copy Number
  - 1
  - 2 (normal)
  - 3

Objective:

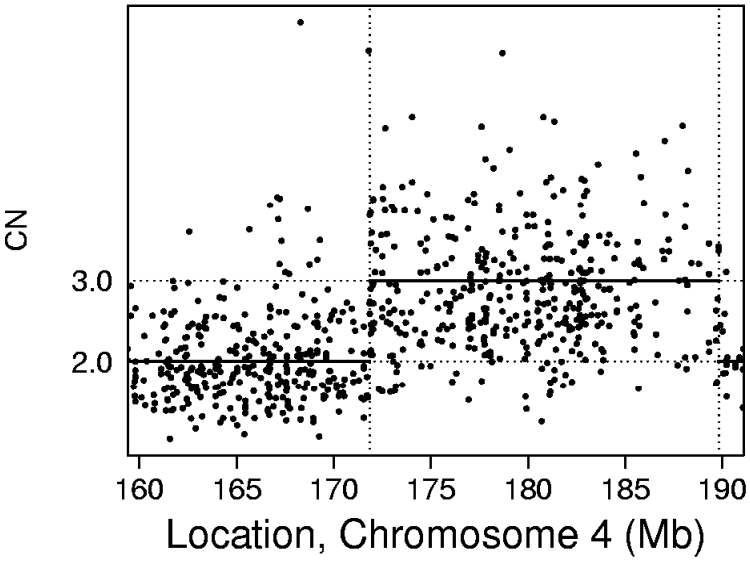
Automatically detect copy number changes for small stretches of DNA

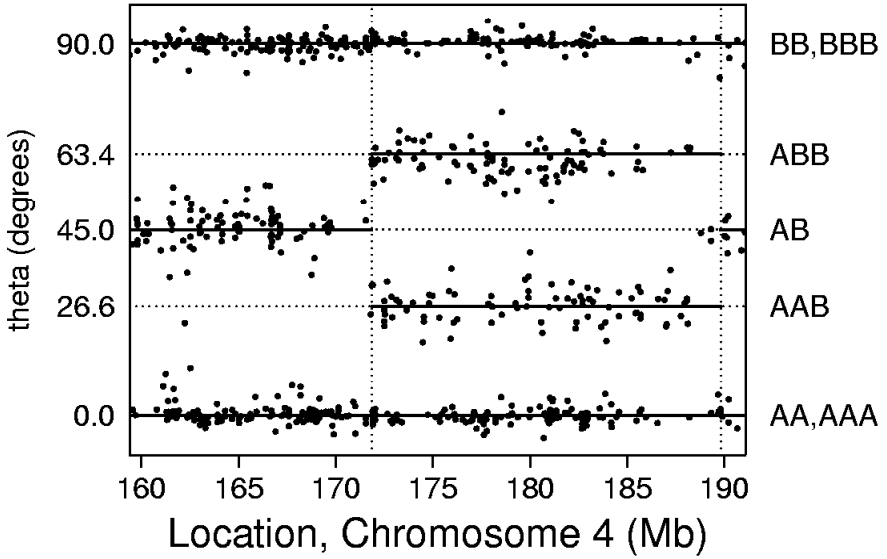


Copy Number Variants in real data



Copy number – genome scanning





Understanding high throughput sequencing

Next generation sequencing

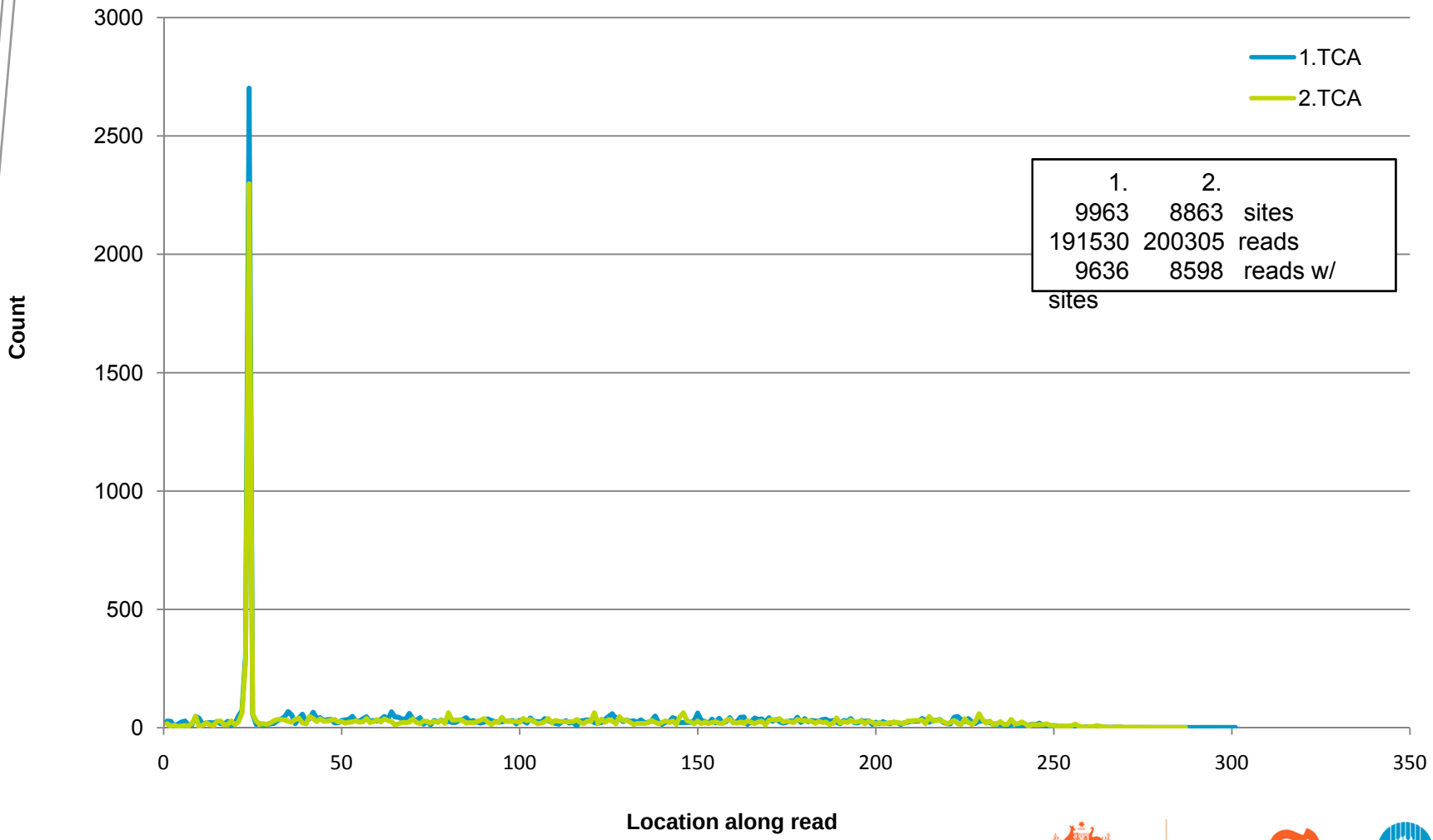
- New sequencing technologies produce Gb of sequencing data.
- Short reads, 35-70bp.
- Few to many errors.
  - Some error models
- Problems with storage.
  - Raw image files, extracted sequences.
- The ‘problem’ is just getting worse.
  - Better data, longer reads, more capacity
- Human genome resequencing efforts.
  - Genetic loci associated with disease
- Bacterial adaptation to environment
  - Cheese production
- Understanding metagenomic communities
  - Gut health
  - Biodiversity
  - Energy production - coal, oil

What does real data look like?

- 454 sequence data
  - Emulsion method.
  - Long reads 100bp+
  - Error prone, especially homopolymeric runs
  - Look at mRNA in tissue.

- Sequencing strategy
  - Use a primer with a restriction site.
  - Isolate mRNA
  - Add linker to end of mRNA
  - PCR of primer
  - Restriction digest
  - Sequence
  - Publish!
  - Two samples sequenced
  - Low coverage, low number of contigs
  - What does it look like at 25mer scale?

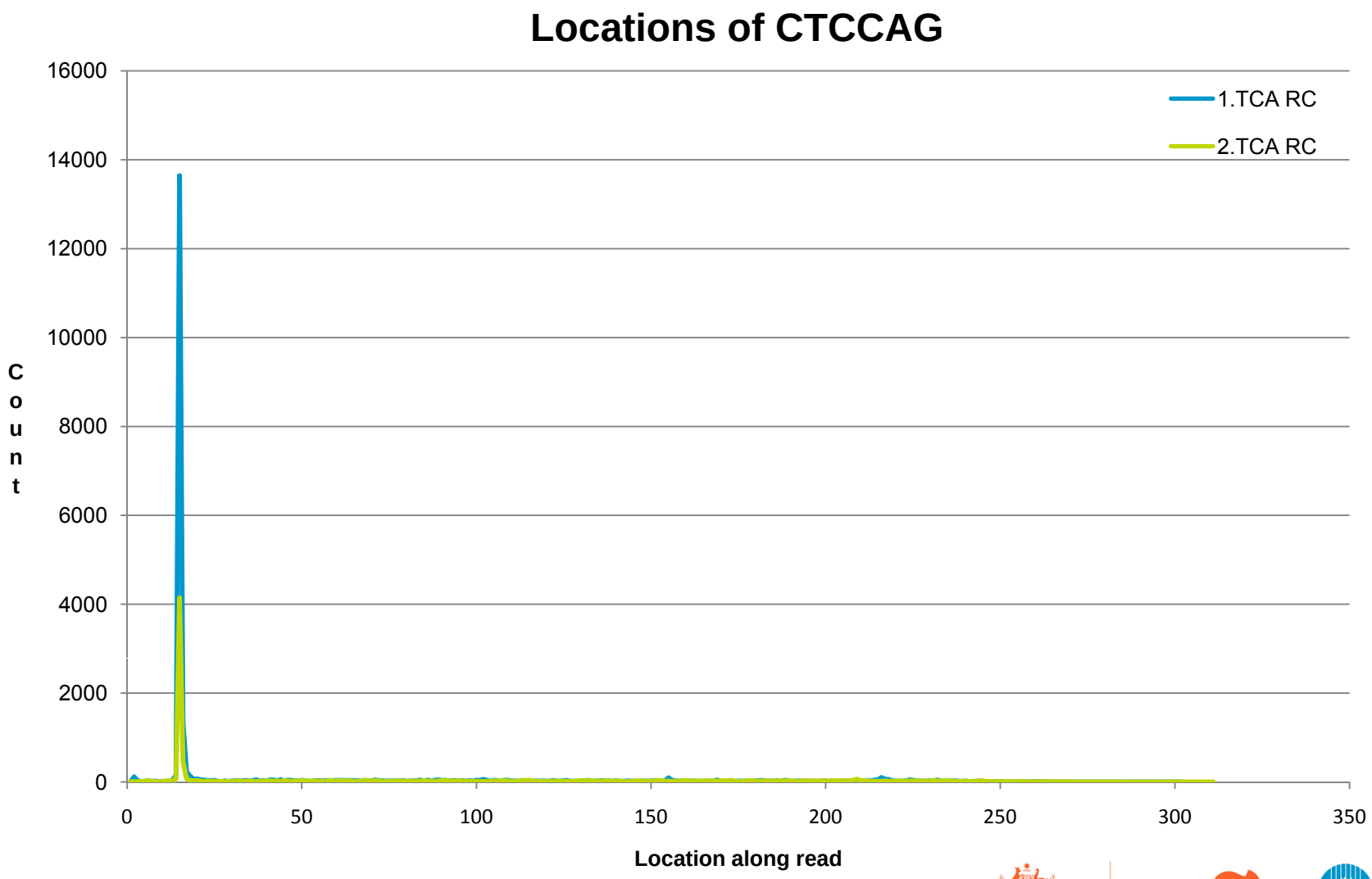
Locations of CTGGAG



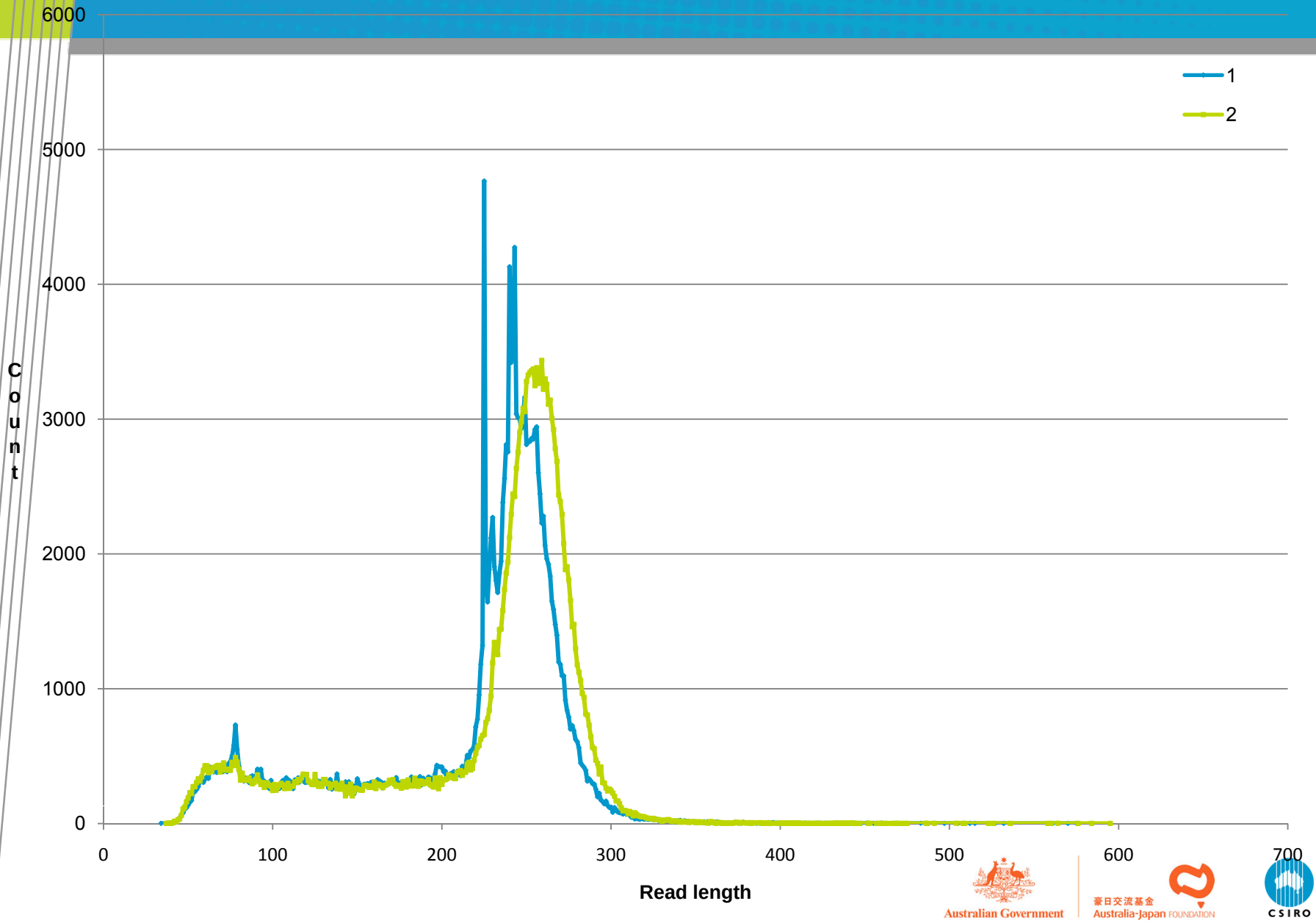
Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO



454 Read Lengths

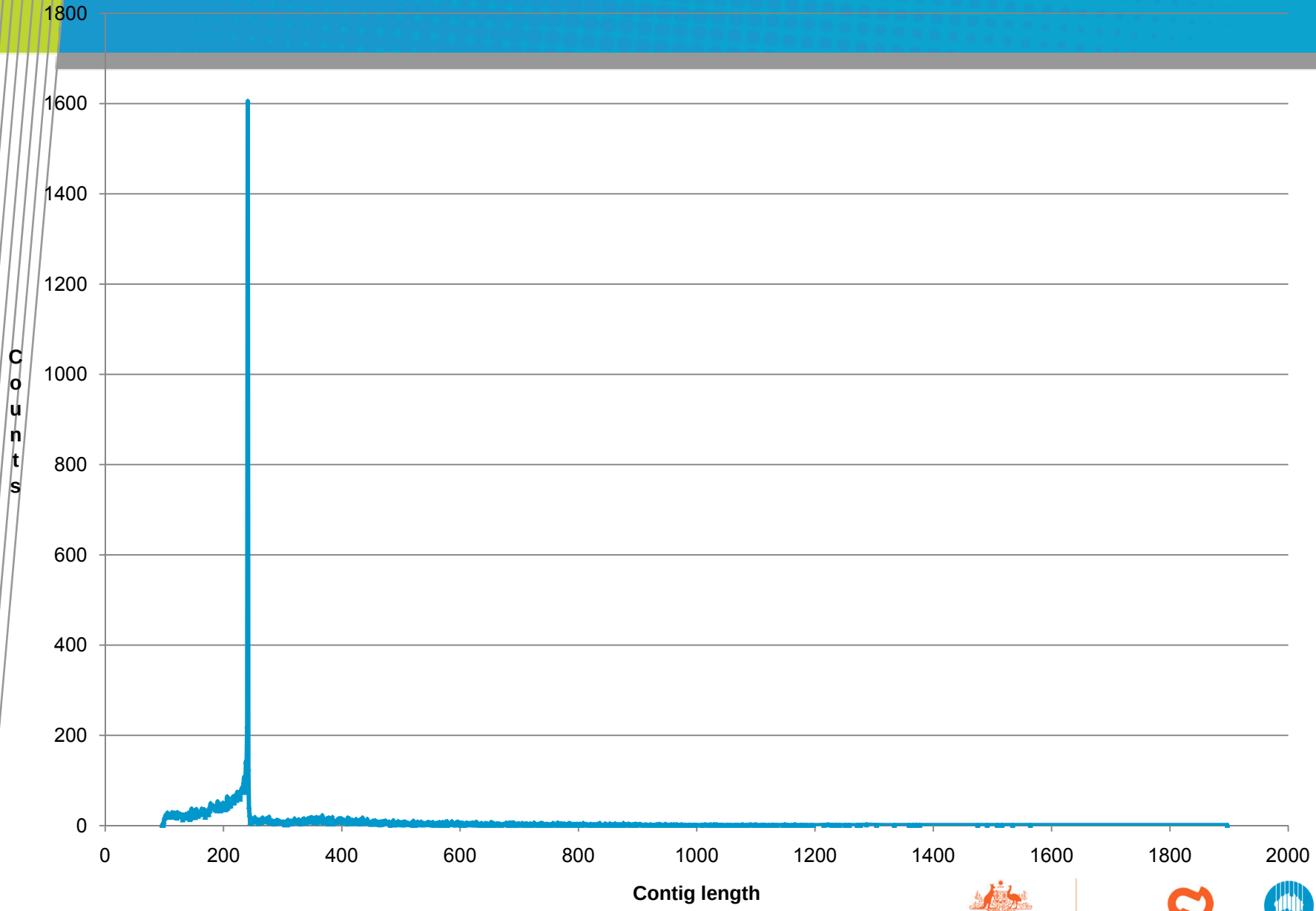


Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

Contigs by length



Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

A new way of analysing the data

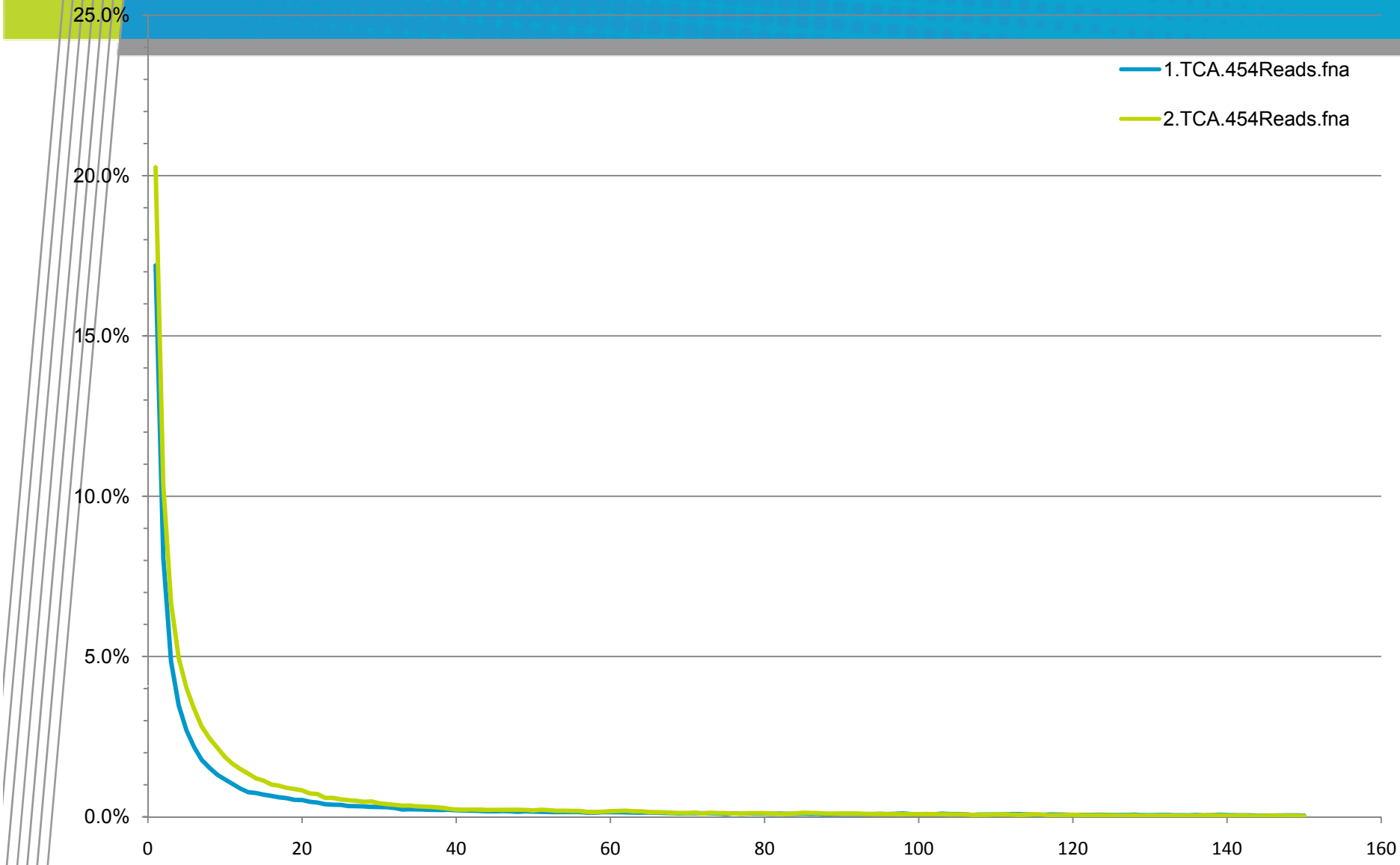
- Cut up sequence data to 25mers
  - ± unique in genome sequences, even Humans
  - This spreads the error to many sequences
  - All possible 25mers $4^{25} \approx 10^{15}$
  - Would take 158 thousand years years to read.
- Use pattern matching at 100%
  - No mismatching problems
- Look at frequency of 25mers
  - QC, annotations
- Cleaver techniques
  - Assembly
  - Annotation
  - SNP detection
  - Sample comparisons

```
GGAAATGGATGGTGGATGACTCCAGAAATCCGT

GGAAATGGATGGTGGATGACTCCAG
 GAAATGGATGGTGGATGACTCCAGA
  AAATGGATGGTGGATGACTCCAGAA
   AATGGATGGTGGATGACTCCAGAAA
    ATGGATGGTGGATGACTCCAGAAAT
     TGGATGGTGGATGACTCCAGAAATC
      GGATGGTGGATGACTCCAGAAATCC
       GATGGTGGATGACTCCAGAAATCCG
        ATGGTGGATGACTCCAGAAATCCGT
```

Australian Government | 豪日交流基金 Australia-Japan FOUNDATION | CSIRO

Tiled 25-mers from 454 data



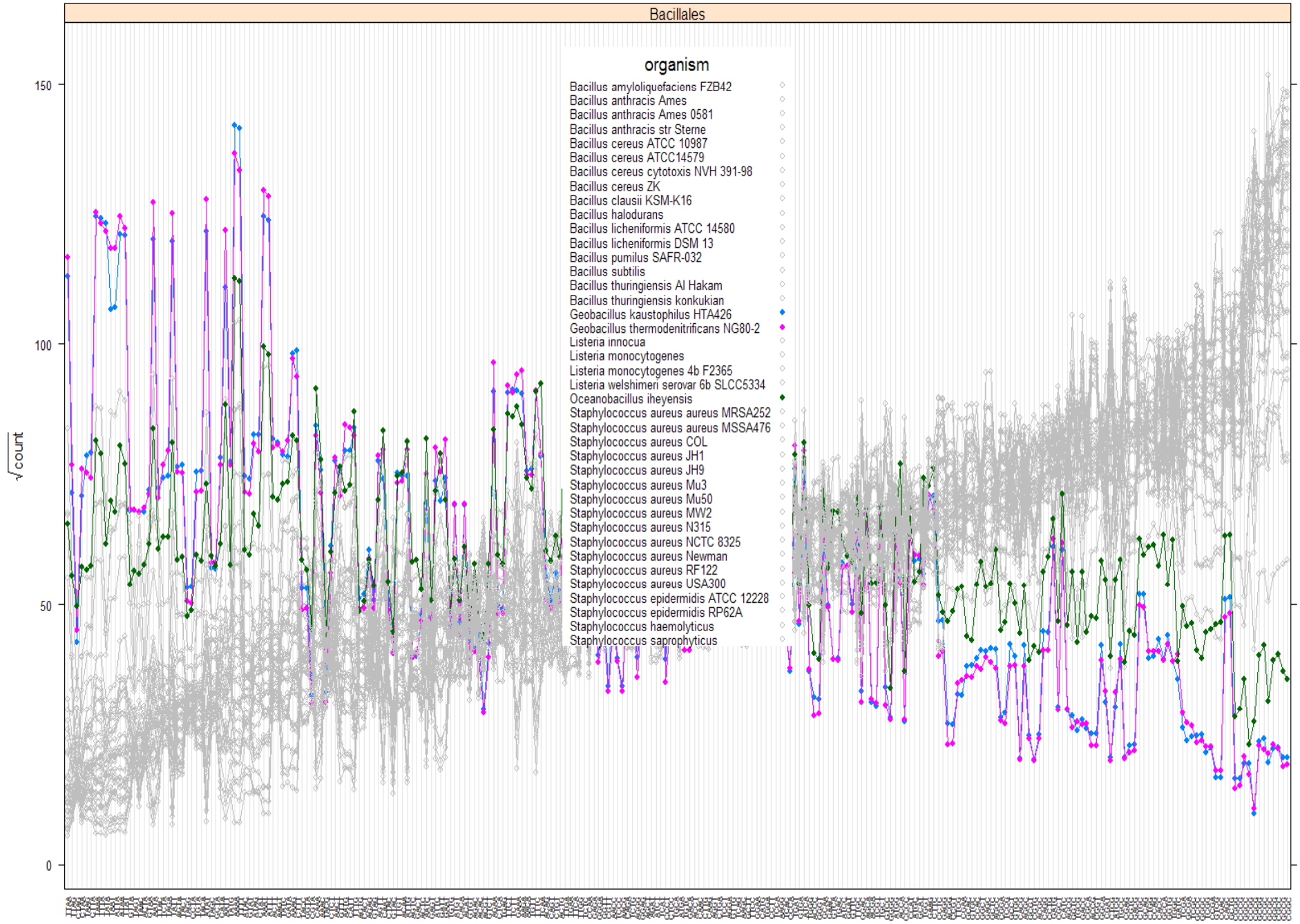
Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

HTS! What are the issues?
What do bacterial genomes look like?

- Take 700 bacterial genomes
- Cut up into 25mers
  - Very sparse data
- Make 100% comparisons, graph
  - Data visualisation
  - Bin data for visualisation
  - Creates a ‘spectrum’ for each ‘species’
- Bacterial clades have most similarity
- Highly conserved sequences are identifiable
- New relationships discovered
- Novel insights into identifying bacterial in mixtures


Bacillales
organism
Bacillus amyloliquefaciens FZB42
Bacillus anthracis Ames
Bacillus anthracis Ames 0581
Bacillus anthracis str Sterne
Bacillus cereus ATCC 10987
Bacillus cereus ATCC14579
Bacillus cereus cytotoxis NVH 391-98
Bacillus cereus ZK
Bacillus clausii KSM-K16
Bacillus halodurans
Bacillus licheniformis ATCC 14580
Bacillus licheniformis DSM 13
Bacillus pumilus SAFR-032
Bacillus subtilis
Bacillus thuringiensis Al Hakam
Bacillus thuringiensis konkukian
Geobacillus kaustophilus HTA426
Geobacillus thermodenitrificans NG80-2
Listeria innocua
Listeria monocytogenes
Listeria monocytogenes 4b F2365
Listeria welshimeri serovar 6b SLCC5334
Oceanobacillus iheyensis
Staphylococcus aureus aureus MRSA252
Staphylococcus aureus aureus MSSA476
Staphylococcus aureus COL
Staphylococcus aureus JH1
Staphylococcus aureus JH9
Staphylococcus aureus Mu3
Staphylococcus aureus Mu50
Staphylococcus aureus MW2
Staphylococcus aureus N315
Staphylococcus aureus NCTC 8325
Staphylococcus aureus Newman
Staphylococcus aureus RF122
Staphylococcus aureus USA300
Staphylococcus epidermidis ATCC 12228
Staphylococcus epidermidis RP62A
Staphylococcus haemolyticus
Staphylococcus saprophyticus
√count
150
100
50
0

Comparing all bacteria

- Take all bacterial genomes and compare to each other
- Look at overlap
- Very sparse matrix
- Bacteria are more different than Vertebrates
- Use this information for analysis!!

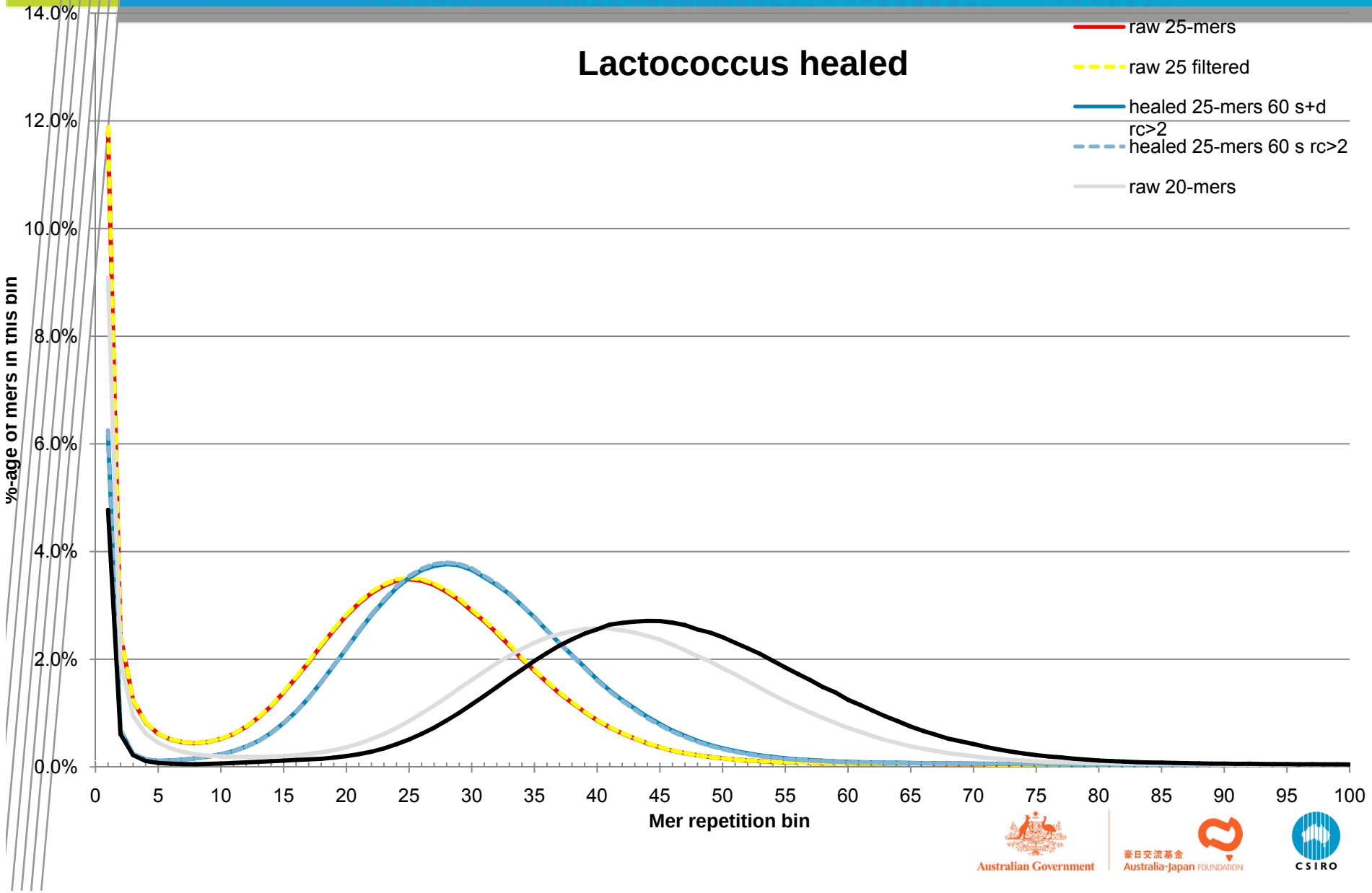
	Yersinia_enterocolitica_8081	Yersinia_enterocolitica_8081	Yersinia_pestis_Angola	Yersinia_pestis_Angola	Yersinia_pestis_Angola	Yersinia_pestis_Antiqua	Yersinia_pestis_Antiqua	Yersinia_pestis_Antiqua	Yersinia_pestis_Antiqua	Yersinia_pestis_biovar_Mediaevails	Yersinia_pestis_biovar_Mediaevails
	NC_008791	NC_008800	NC_010157	NC_010158	NC_010159	NC_008120	NC_008121	NC_008122	NC_008150	NC_005810	NC_005813
Yersinia_enterocolitica_8081											
66725	101.88%	8.21%	43.89%	0.23%	0.69%	0.26%	0.00%	44.07%	0.68%	0.67%	44.06%
4526838	0.12%	100.99%	0.01%	0.01%	3.59%	0.01%	0.00%	0.01%	3.59%	3.57%	0.01%
Yersinia_pestis_Angola											
65906	44.44%	0.45%	102.00%	0.02%	1.70%	0.02%	0.00%	100.89%	1.69%	1.69%	101.11%
102309	0.15%	0.36%	0.01%	103.11%	10.94%	88.20%	9.60%	3.43%	10.72%	10.72%	3.43%
4260358	0.01%	3.81%	0.03%	0.26%	101.04%	0.27%	0.09%	0.12%	98.99%	98.54%	0.12%
Yersinia_pestis_Antiqua											
8442	0.00%	0.00%	0.00%	116.39%	45.82%	45.91%	100.00%	22.89%	45.82%	45.83%	22.87%
68294	43.06%	0.43%	97.36%	5.14%	7.50%	5.68%	2.83%	102.07%	7.51%	7.50%	101.47%
94525	0.18%	0.40%	0.01%	95.46%	12.33%	104.09%	4.10%	4.10%	12.37%	12.36%	4.10%
4431554	0.01%	3.67%	0.03%	0.25%	95.16%	0.26%	0.09%	0.12%	100.66%	99.28%	0.12%
Yersinia_pestis_biovar_Mediaevails											
8336	0.00%	0.00%	0.00%	117.06%	46.37%	46.46%	98.93%	23.16%	46.35%	46.38%	23.16%
21576	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
68309	43.04%	0.43%	97.56%	5.14%	7.50%	5.67%	2.83%	101.45%	7.51%	7.49%	102.07%
103048	0.17%	0.38%	0.01%	92.53%	11.24%	93.61%	3.75%	3.76%	11.25%	11.24%	3.76%
4412864	0.01%	3.67%	0.03%	0.25%	95.13%	0.26%	0.09%	0.12%	99.70%	100.71%	0.12%
Yersinia_pestis_CO92											
8412	0.00%	0.00%	0.00%	116.74%	45.95%	46.04%	100.00%	22.96%	45.93%	45.96%	22.96%
68428	42.94%	0.43%	97.40%	5.13%	7.49%	5.66%	2.82%	101.62%	7.50%	7.49%	101.47%
92941	0.17%	0.40%	0.01%	95.74%	12.58%	101.83%	4.17%	4.17%	12.61%	12.60%	4.17%
4449691	0.01%	3.67%	0.00%	0.25%	95.51%	0.26%	0.09%	0.09%	99.69%	99.02%	0.09%
Yersinia_pestis_KIM											
100730	0.16%	0.37%	0.01%	91.14%	11.56%	96.69%	3.82%	3.82%	11.62%	11.59%	3.82%
4407747	0.01%	3.71%	0.03%	0.25%	95.74%	0.26%	0.09%	0.12%	99.76%	99.26%	0.12%
Yersinia_pestis_Nepal516											
8431	0.00%	0.00%	0.00%	116.33%	45.70%	45.80%	99.80%	22.83%	45.69%	45.71%	22.83%
100833	0.16%	0.37%	0.01%	91.31%	11.60%	96.85%	3.84%	3.84%	11.63%	11.62%	3.84%
4337710	0.01%	3.29%	0.03%	0.25%	95.07%	0.27%	0.09%	0.12%	99.76%	99.26%	0.12%
Yersinia_pestis_Pestoides_F											
69640	42.09%	0.47%	95.22%	6.86%	11.05%	7.38%	2.78%	98.97%	11.09%	11.08%	99.20%
136925	0.24%	0.45%	0.04%	69.14%	8.53%	69.68%	1.39%	1.42%	8.56%	8.56%	1.42%
4356622	0.01%	3.75%	0.03%	0.25%	97.33%	0.27%	0.09%	0.12%	99.23%	98.87%	0.12%
Yersinia_pseudotuberculosis_IP_31758											
58131	0.00%	0.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
152872	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
4628469	0.01%	3.23%	0.00%	0.08%	77.82%	0.08%	0.00%	0.00%	81.04%	81.13%	0.00%
Yersinia_pseudotuberculosis_IP32953											
26956	0.00%	0.00%	0.09%	0.00%	0.00%	0.00%	0.00%	0.09%	0.00%	0.00%	0.09%
65890	44.66%	0.50%	94.08%	0.03%	1.95%	0.03%	0.00%	95.00%	1.97%	1.95%	95.37%
4655829	0.01%	3.53%	0.03%	0.23%	81.11%	0.25%	0.08%	0.11%	84.47%	84.55%	0.11%
Yersinia_pseudotuberculosis_PB1_											
66563	43.54%	0.39%	91.80%	1.87%	5.71%	1.88%	0.00%	92.76%	5.80%	5.78%	93.08%
4609939	0.00%	3.56%	0.03%	0.22%	81.27%	0.24%	0.08%	0.11%	84.60%	84.69%	0.11%
Yersinia_pseudotuberculosis_YPIII											
4604382	0.01%	3.08%	0.05%	0.06%	80.17%	0.06%	0.00%	0.05%	83.51%	83.53%	0.05%

	Yersinia_enterocolitica_8081	Yersinia_enterocolitica_8081	Yersinia_pestis_Angola	Yersinia_pestis_Angola	Yersinia_pestis_Angola	Yersinia_pestis_Antiqua	Yersinia_pestis_Antiqua	Yersinia_pestis_Antiqua	Yersinia_pestis_Antiqua	Yersinia_pestis_biovar_Mediaevails	Yersinia_pestis_biovar_Mediaevails
	NC_008791	NC_008800	NC_010157	NC_010158	NC_010159	NC_008120	NC_008121	NC_008122	NC_008150	NC_005810	NC_005813
Yersinia_enterocolitica_8081											
66725	101.88%	8.21%	43.89%	0.23%	0.69%	0.26%	0.00%	44.07%	0.68%	0.67%	44.06%
4526838	0.12%	100.99%	0.01%	0.01%	3.59%	0.01%	0.00%	0.01%	3.59%	3.57%	0.01%
Yersinia_pestis_Angola											
65906	44.44%	0.45%	102.00%	0.02%	1.70%	0.02%	0.00%	100.89%	1.69%	1.69%	101.11%
102309	0.15%	0.36%	0.01%	103.11%	10.94%	88.20%	9.60%	3.43%	10.72%	10.72%	3.43%
4260358	0.01%	3.81%	0.03%	0.26%	101.04%	0.27%	0.09%	0.12%	98.99%	98.54%	0.12%
Yersinia_pestis_Antiqua											
8442	0.00%	0.00%	0.00%	116.39%	45.82%	45.91%	100.00%	22.89%	45.82%	45.83%	22.87%
68294	43.06%	0.43%	97.36%	5.14%	7.50%	5.68%	2.83%	102.07%	7.51%	7.50%	101.47%
94525	0.18%	0.40%	0.01%	95.46%	12.33%	104.09%	4.10%	4.10%	12.37%	12.36%	4.10%
4431554	0.01%	3.67%	0.03%	0.25%	95.16%	0.26%	0.09%	0.12%	100.66%	99.28%	0.12%
Yersinia_pestis_biovar_Mediaevails											
8336	0.00%	0.00%	0.00%	117.06%	46.37%	46.46%	98.93%	23.16%	46.35%	46.38%	23.16%
21576	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
68309	43.04%	0.43%	97.56%	5.14%	7.50%	5.67%	2.83%	101.45%	7.51%	7.49%	102.07%
103048	0.17%	0.38%	0.01%	92.53%	11.24%	93.61%	3.75%	3.76%	11.25%	11.24%	3.76%
4412864	0.01%	3.67%	0.03%	0.25%	95.13%	0.26%	0.09%	0.12%	99.70%	100.71%	0.12%
Yersinia_pestis_CO92											
8412	0.00%	0.00%	0.00%	116.74%	45.95%	46.04%	100.00%	22.96%	45.93%	45.96%	22.96%
68428	42.94%	0.43%	97.40%	5.13%	7.49%	5.66%	2.82%	101.62%	7.50%	7.49%	101.47%
92941	0.17%	0.40%	0.01%	95.74%	12.58%	101.83%	4.17%	4.17%	12.61%	12.60%	4.17%
4449691	0.01%	3.67%	0.00%	0.25%	95.51%	0.26%	0.09%	0.09%	99.69%	99.02%	0.09%
Yersinia_pestis_KIM											
100730	0.16%	0.37%	0.01%	91.14%	11.56%	96.69%	3.82%	3.82%	11.62%	11.59%	3.82%
4407747	0.01%	3.71%	0.03%	0.25%	95.74%	0.26%	0.09%	0.12%	99.76%	99.26%	0.12%
Yersinia_pestis_Nepal516											
8431	0.00%	0.00%	0.00%	116.33%	45.70%	45.80%	99.80%	22.83%	45.69%	45.71%	22.83%
100833	0.16%	0.37%	0.01%	91.31%	11.60%	96.85%	3.84%	3.84%	11.63%	11.62%	3.84%
4337710	0.01%	3.29%	0.03%	0.25%	95.07%	0.27%	0.09%	0.12%	99.76%	99.26%	0.12%
Yersinia_pestis_Pestoides_F											
69640	42.09%	0.47%	95.22%	6.86%	11.05%	7.38%	2.78%	98.97%	11.09%	11.08%	99.20%
136925	0.24%	0.45%	0.04%	69.14%	8.53%	69.68%	1.39%	1.42%	8.56%	8.56%	1.42%
4356622	0.01%	3.75%	0.03%	0.25%	97.33%	0.27%	0.09%	0.12%	99.23%	98.87%	0.12%
Yersinia_pseudotuberculosis_IP_31758											
58131	0.00%	0.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
152872	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
4628469	0.01%	3.23%	0.00%	0.08%	77.82%	0.08%	0.00%	0.00%	81.04%	81.13%	0.00%
Yersinia_pseudotuberculosis_IP32953											
26956	0.00%	0.00%	0.09%	0.00%	0.00%	0.00%	0.00%	0.09%	0.00%	0.00%	0.09%
65890	44.66%	0.50%	94.08%	0.03%	1.95%	0.03%	0.00%	95.00%	1.97%	1.95%	95.37%
4655829	0.01%	3.53%	0.03%	0.23%	81.11%	0.25%	0.08%	0.11%	84.47%	84.55%	0.11%
Yersinia_pseudotuberculosis_PB1_											
66563	43.54%	0.39%	91.80%	1.87%	5.71%	1.88%	0.00%	92.76%	5.80%	5.78%	93.08%
4609939	0.00%	3.56%	0.03%	0.22%	81.27%	0.24%	0.08%	0.11%	84.60%	84.69%	0.11%
Yersinia_pseudotuberculosis_YPIII											
4604382	0.01%	3.08%	0.05%	0.06%	80.17%	0.06%	0.00%	0.05%	83.51%	83.53%	0.05%

What do real bacterial genomes look like?

- Generate Illumina data, 35mers
- Create library of all 25mers
  - Spread the errors
- Bin these based on first 4-mers
- Graph and compare for different species.
- Produces a frequency curve for each dataset.
- Identify contaminations, and sequence errors
- Can errors be healed?
- Where is the real data??
- Modeling data
- Next steps …

Real sequence data



Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

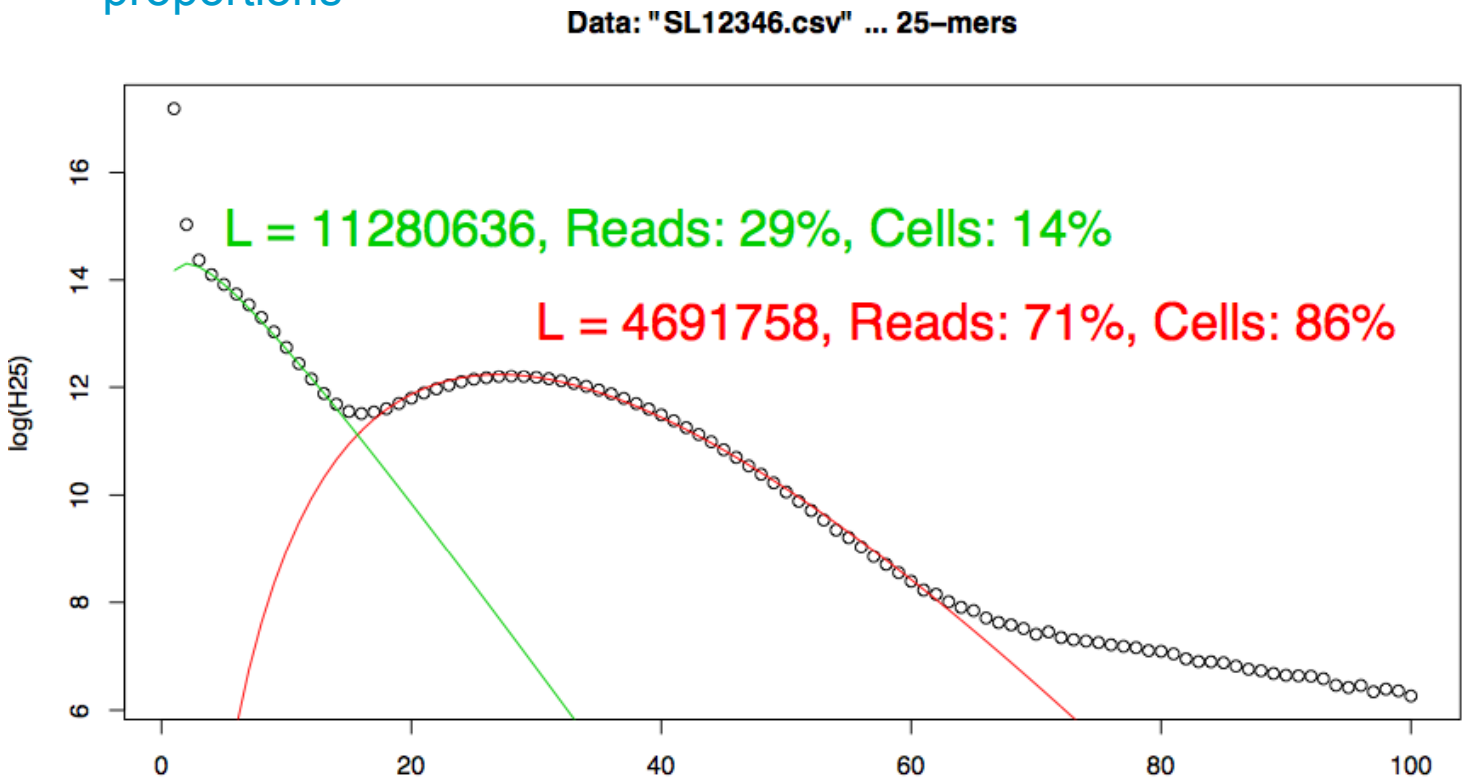
How to model this data? Real data!

- Look at frequency of all 25mers per sequencing run.
- Fast to compare, requires some disk space.
- Applicable to QC.
  - Quality of sequences
  - Shows biases in data collection
  - allows us to make better assemblies and gene annotations.
  - Applicable to SNP detection
- Take a random selection from some genome DNA
- Look at frequency distribution,
- Try to model
- What happened??
- A real example

Modeling contamination

- Green: 11Mbp genome
- Red: 4.8Mbp genome
- Present in sample at different proportions

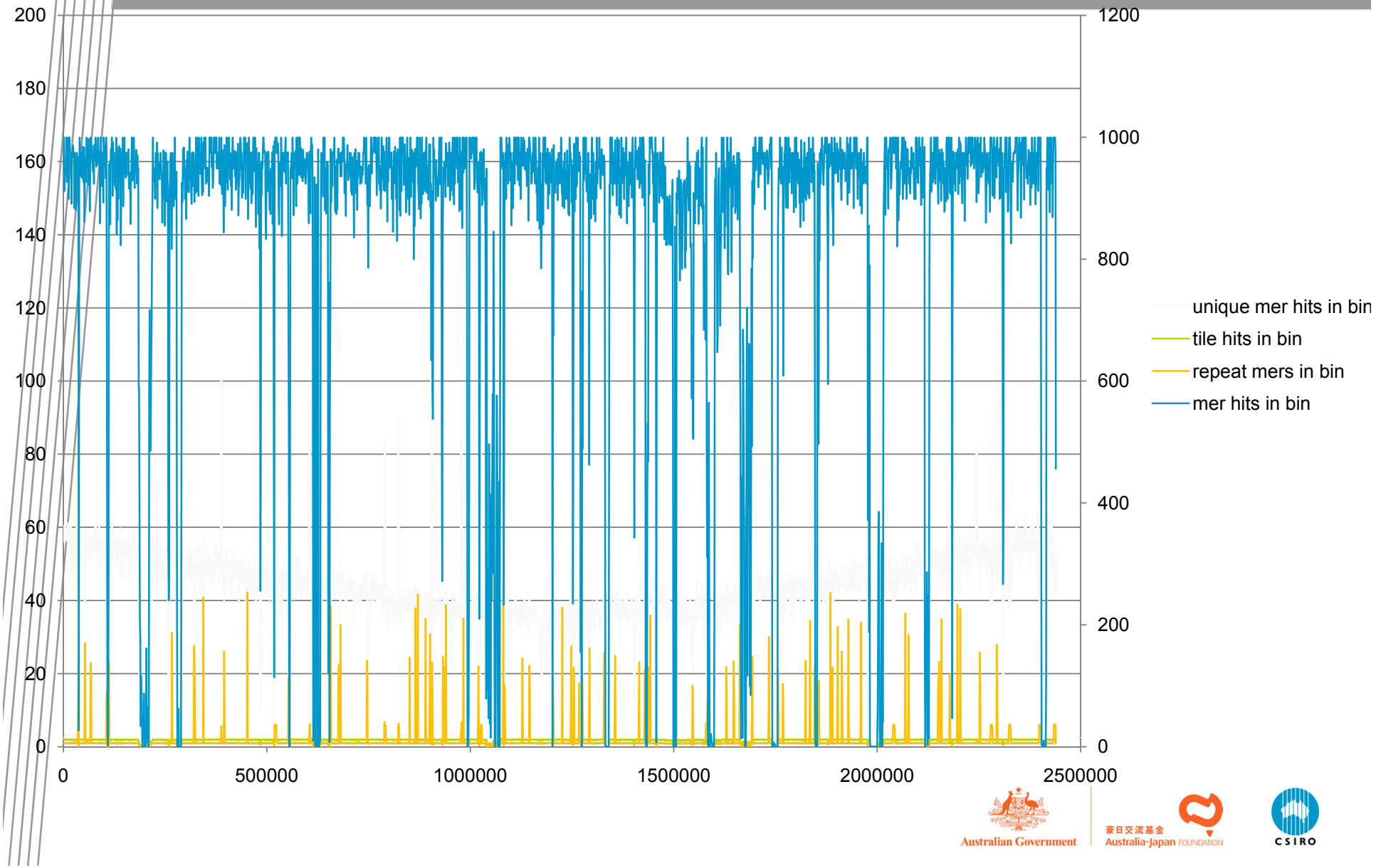
- Fit a gamma model
  - Works better than Poisson



What's happening at the genome level?

- We can see the frequencies of 'hits', where do they come from?
- Is there unequal sampling of the genome?
- Are some regions over or under represented?
- Map reads to a physical 'reference' genome
- Look at differences between raw and 'healed' data.

Unhealed **Lactococcus cremoris SK11 raw 1K**



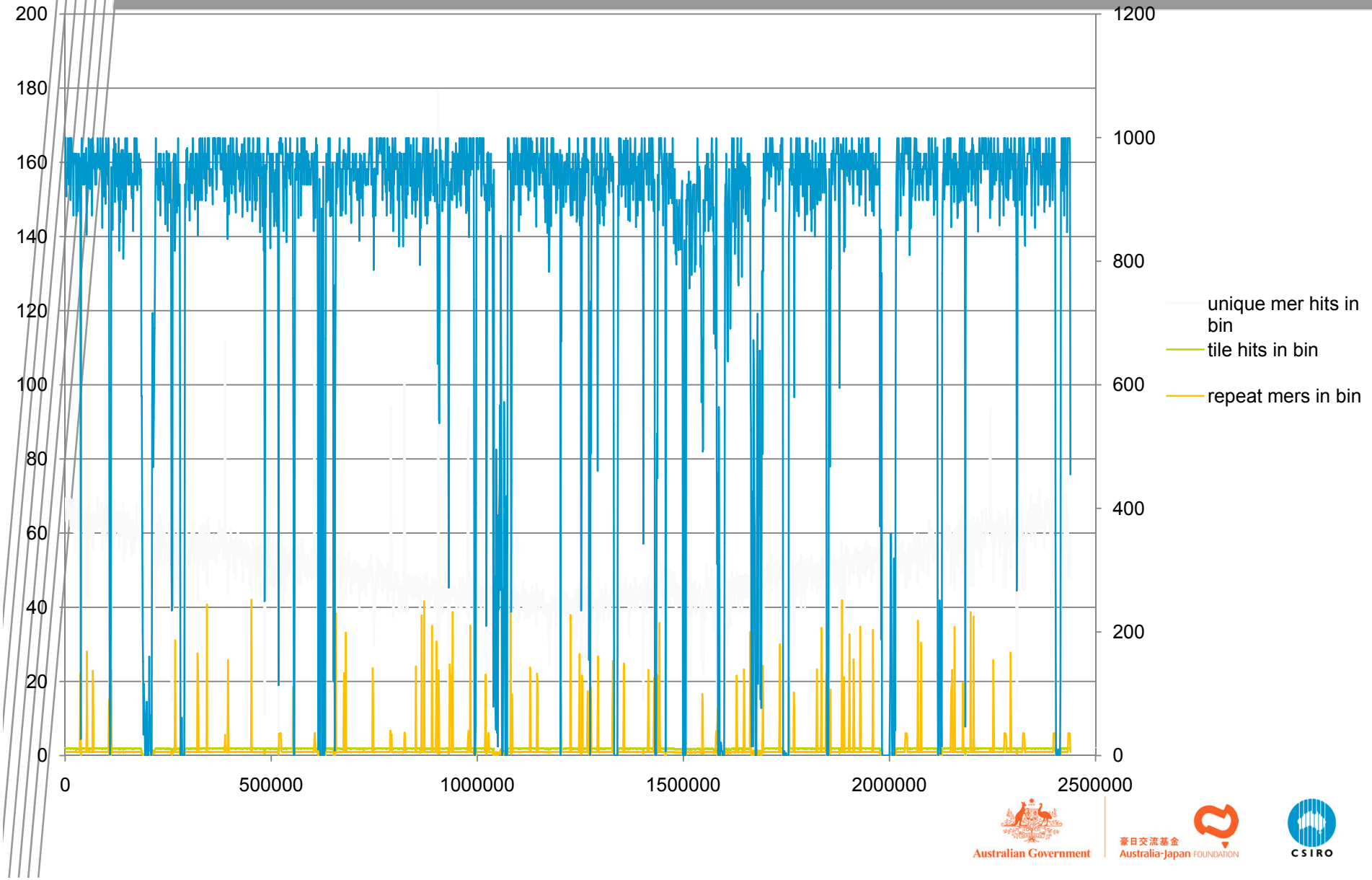
Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

Healed

Lactococcus cremoris SK11 healed 1K

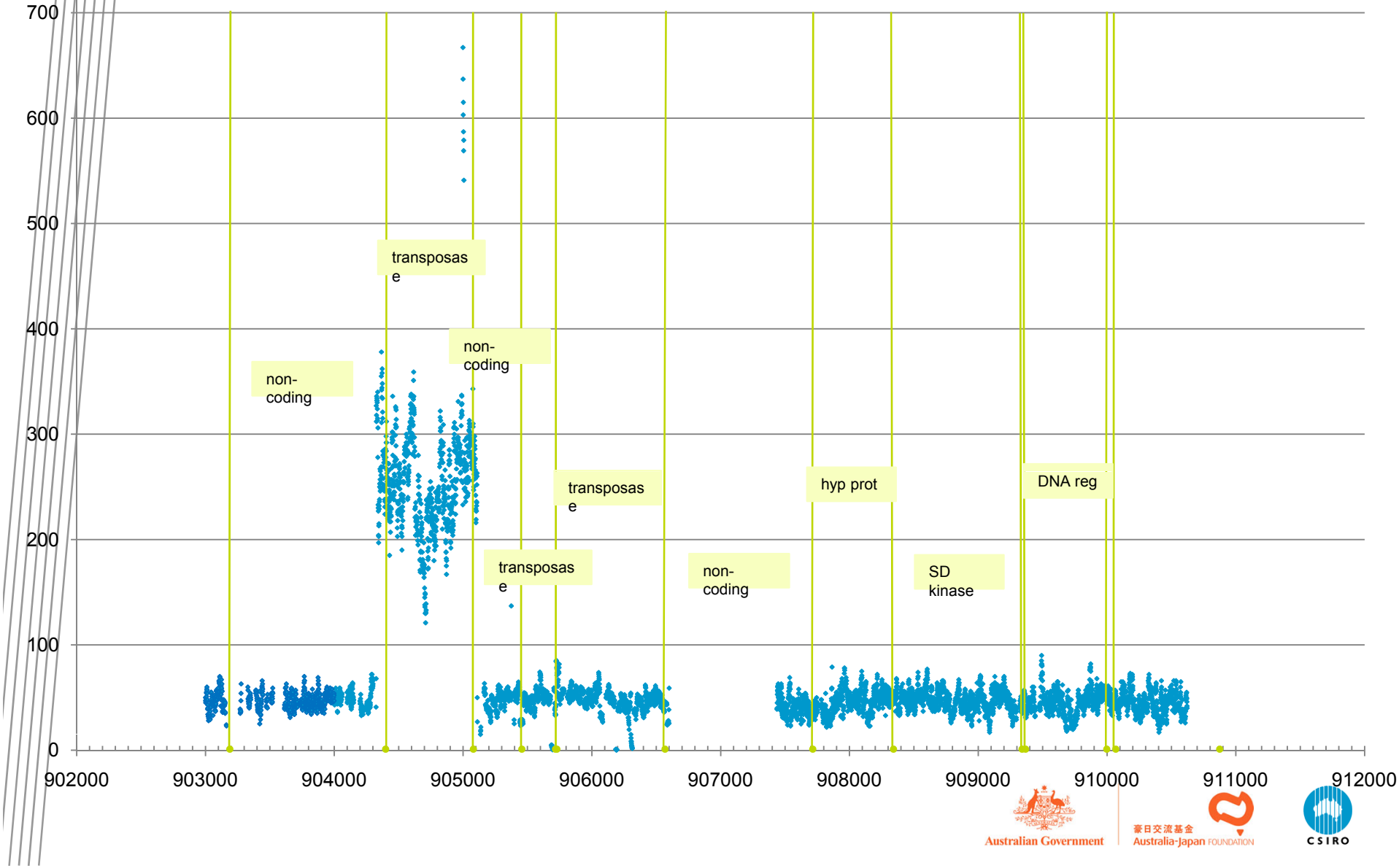


Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

Which sequences are over represented?

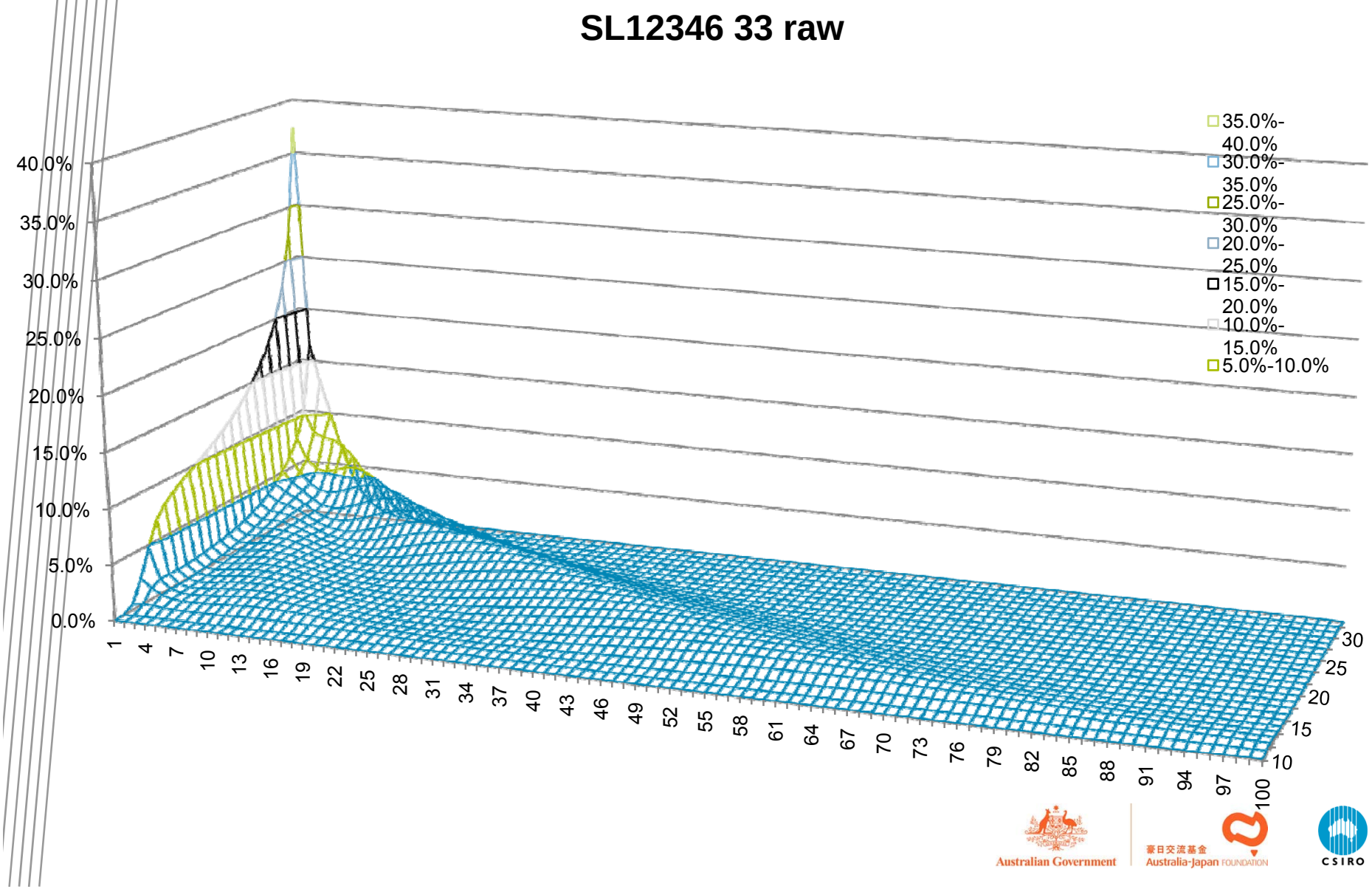


Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

What do mixtures look like?



Australian Government

豪日交流基金 Australia-Japan FOUNDATION

CSIRO

Future?

- Sequencing analysis is in a paradigm mud-pit
  - Pattern matching
  - Little statistics
- There are some interesting alternatives coming out in the literature.
  - De Bruijn graphs.
- How can statistics help?
- Turning the high throughput sequencing problem into a statistical problem

Thank you

CMIS Statistical Bioinformatics
Bill Wilson

Phone: +61 2 9325 3153
Email: bill.wilson@csiro.au
Web: www.csiro.au/CMIS

Australian Government
豪日交流基金 Australia-Japan FOUNDATION
CSIRO