- Faculty of Science, Engineering & Technology

UNIVERSITY OF TASMANIA

Microsatellite evolution in ancient and modern penguins

Bennet McComish

School of Mathematics and Physics

Microsatellites

Tandem repeats of motifs up to 6bp, e.g. $(AC)_6$ = ACACACACACAC

Length is highly polymorphic.

Ubiquitous in eukaryote genomes.

Most evolve neutrally, and are widely used as genetic markers in population genetics, ecology.

Some are also involved in disease in humans and other mammals.

Thought to mutate by replication slippage.

Repeats can be imperfect, e.g. one locus has three alleles:

1. $(AAAG)_{12}$
2. $(AAAG)_{22}A(AAAG)_{12}$
3. $(AAAGAGAG)_6(A)_4(AG)_3$ $(AAAG)_3(AG)_9AA(AG)_3(AAAG)_2$ $(AG)_2(AAAG)_2(AGAGAAAG)_{15}$ $(AAAG)_{24}$

or compound, e.g. $(AGG)_8(CTC)_6$

Point mutation may be important in these cases.

Microsatellite evolution in ancient and modern penguins

Microsatellite models

Symmetric models:

- Constant rate of mutation in both directions
- Rate proportional to current length
- Can change by multiple repeat units

Very simplistic, and don't have a stationary distribution.

Asymmetric models

- Different rates up and down – biased upwards if the current number of repeats is small, downwards if large.
- Can be generalised to allow multi-step changes, point mutations.

These models have stationary distributions.

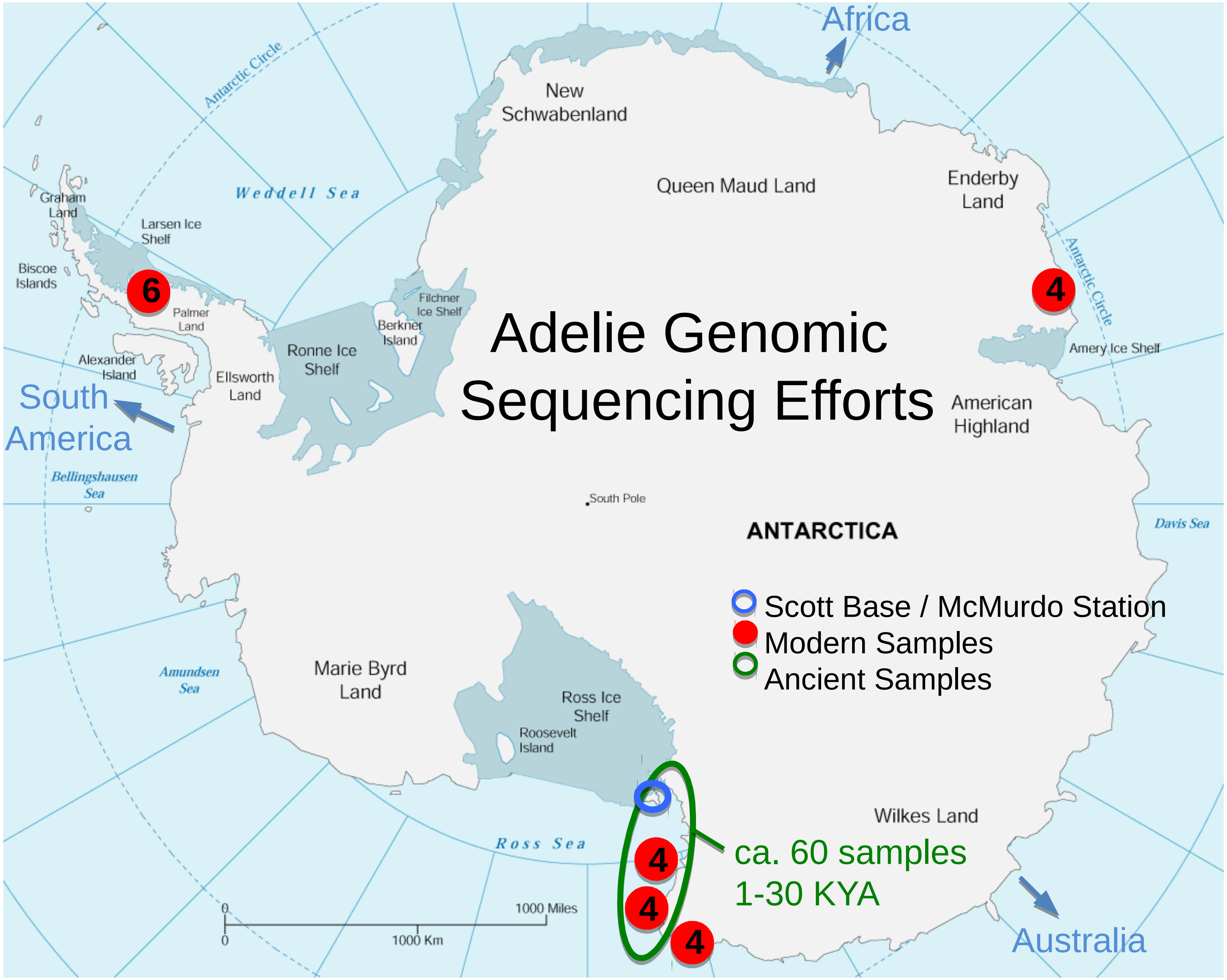
Adélie penguin data

Adélie penguins breed in multiple locations around the coast of Antarctica.

Same nesting sites used for thousands of years – dead chicks preserved.

We have high-coverage (~30x) genome sequence reads for 22 modern samples from five sites.

32 ancient genomes up to 25,000 years old from several sites currently being sequenced at lower coverage (up to 10x).



Adelie Genomic Sequencing Efforts

○ Scott Base / McMurdo Station
● Modern Samples
○ Ancient Samples

ca. 60 samples
1-30 KYA

Microsatellite detection

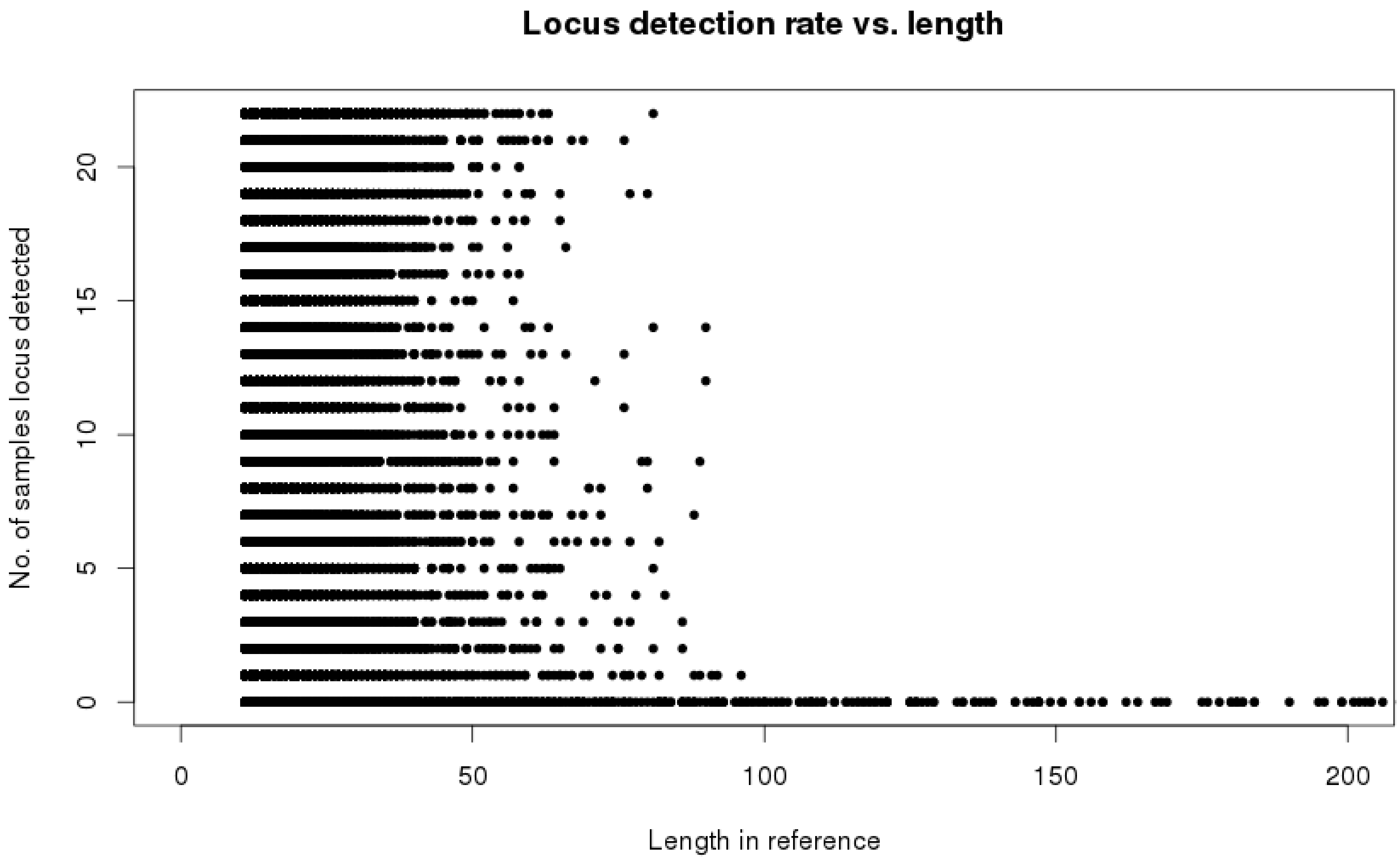
Run Tandem Repeat Finder (Benson, 1999) on best available reference genome – numbers of loci detected:

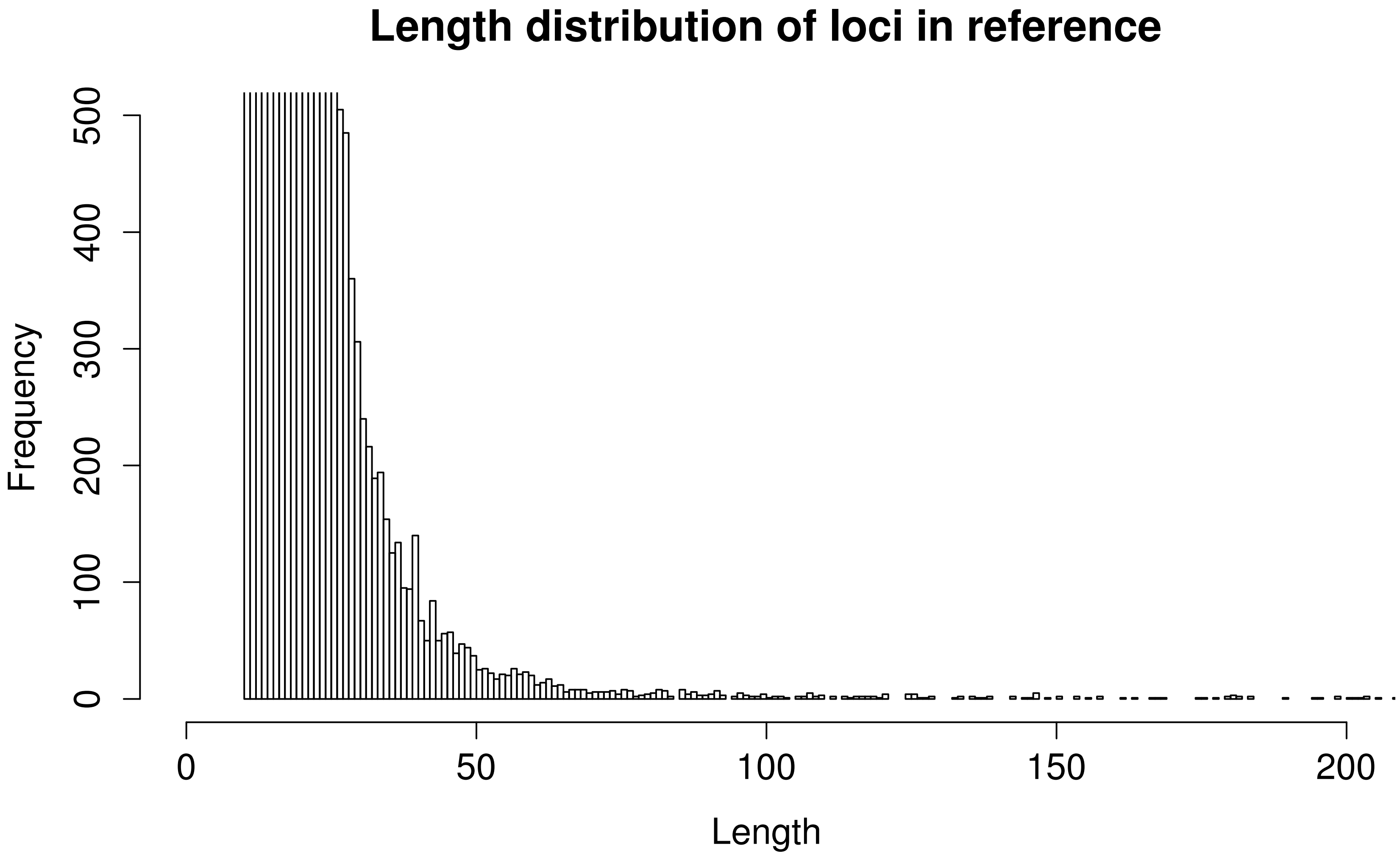
Motif length	Number of loci
1	175,604
2	41,411
3	61,014
4	105,862
5	232,325
6	529,492

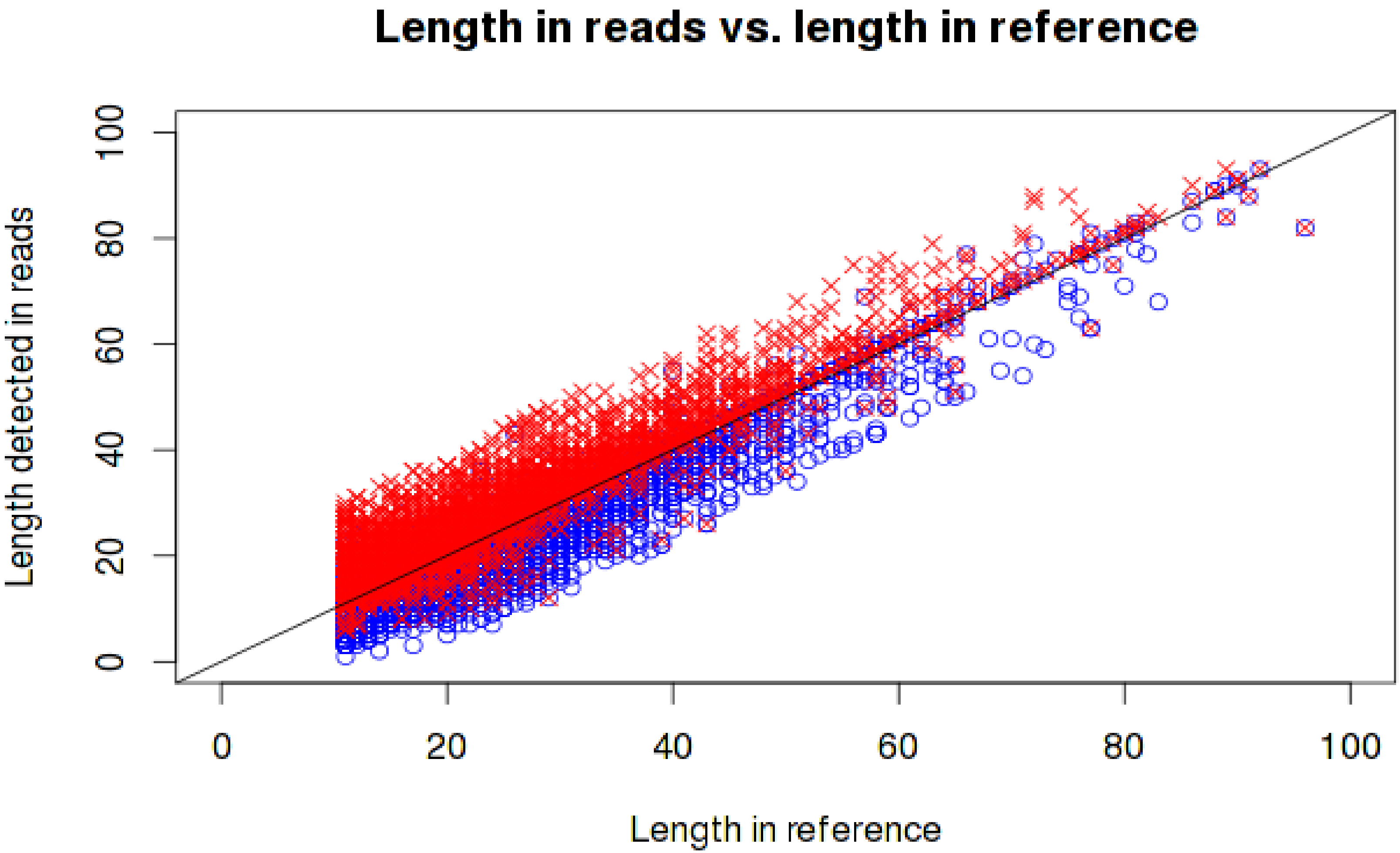
Map reads to reference using Bowtie2 (Langmead, 2012).

Use RepeatSeq (Highnam, 2013) to genotype samples based on read mapping.

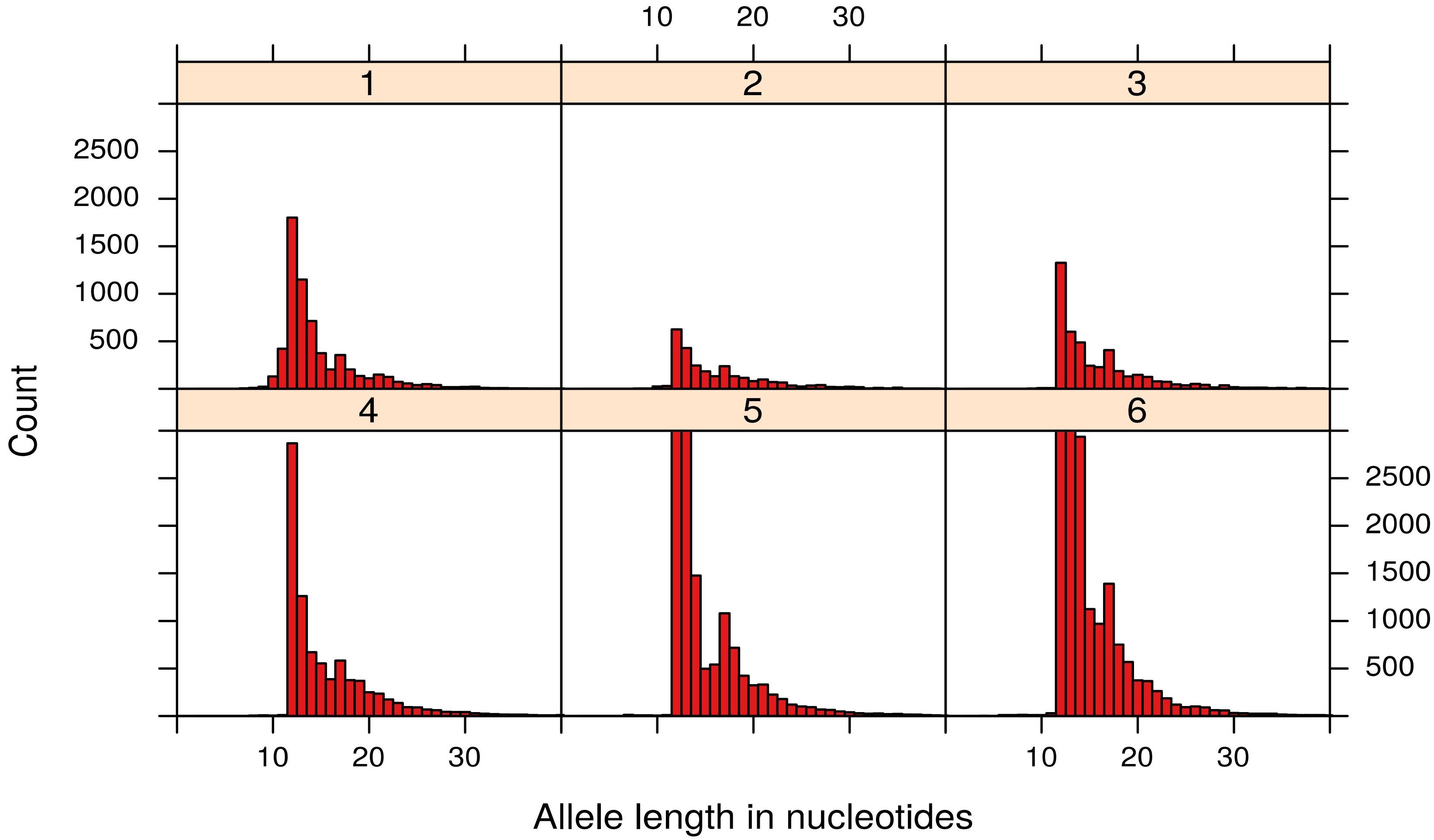
Convert ouput for easy processing in R – we have: motif, position and length in reference, lengths observed in samples, quality scores.


Locus detection rate vs. length
No. of samples locus detected
0
5
10
15
20
0
50
100
150
200
Length in reference

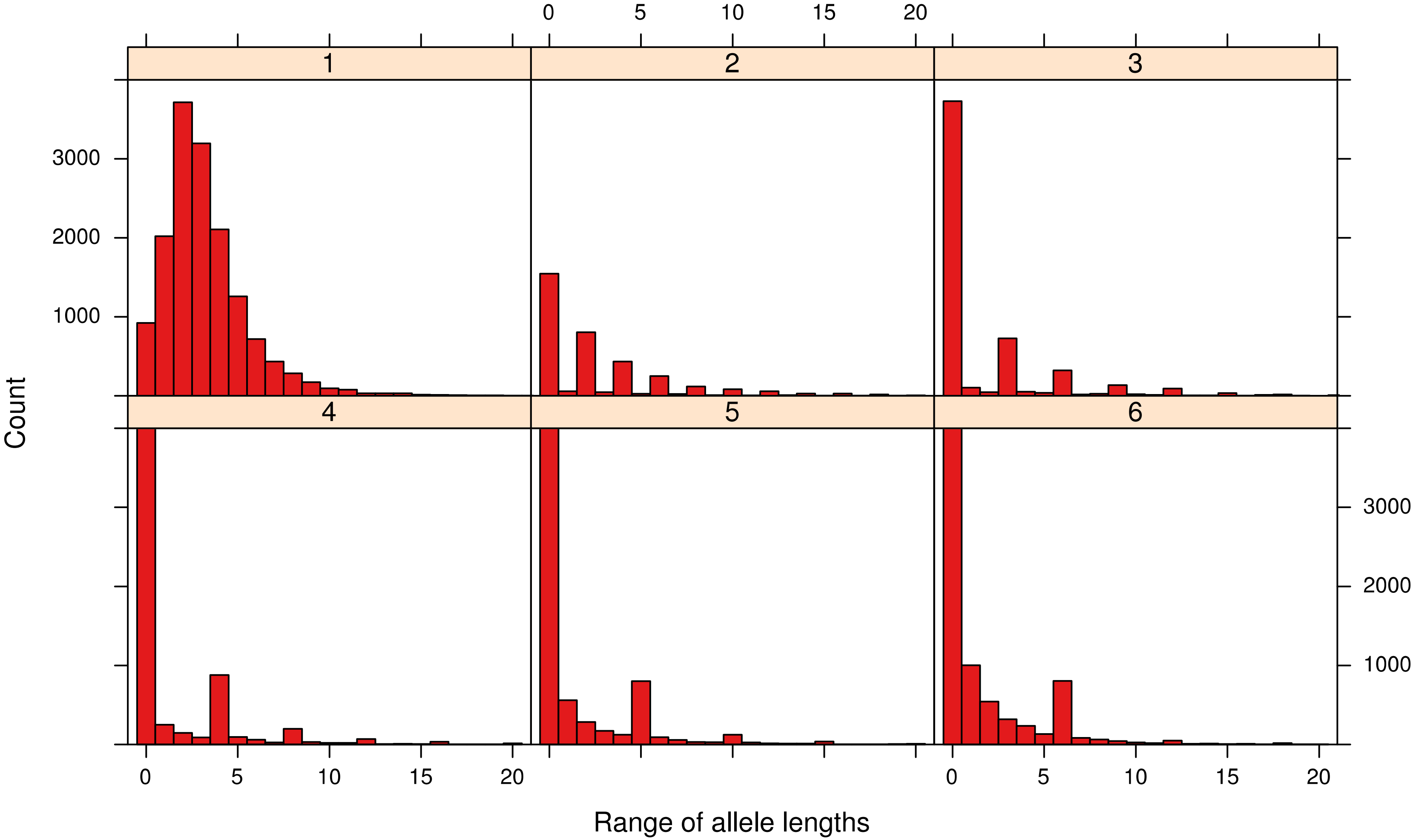

Length distribution of loci in reference
Frequency
0
100
200
300
400
500
0
50
100
150
200
Length


Length in reads vs. length in reference
Length detected in reads
0
20
40
60
80
100
Length in reference
0
20
40
60
80
100

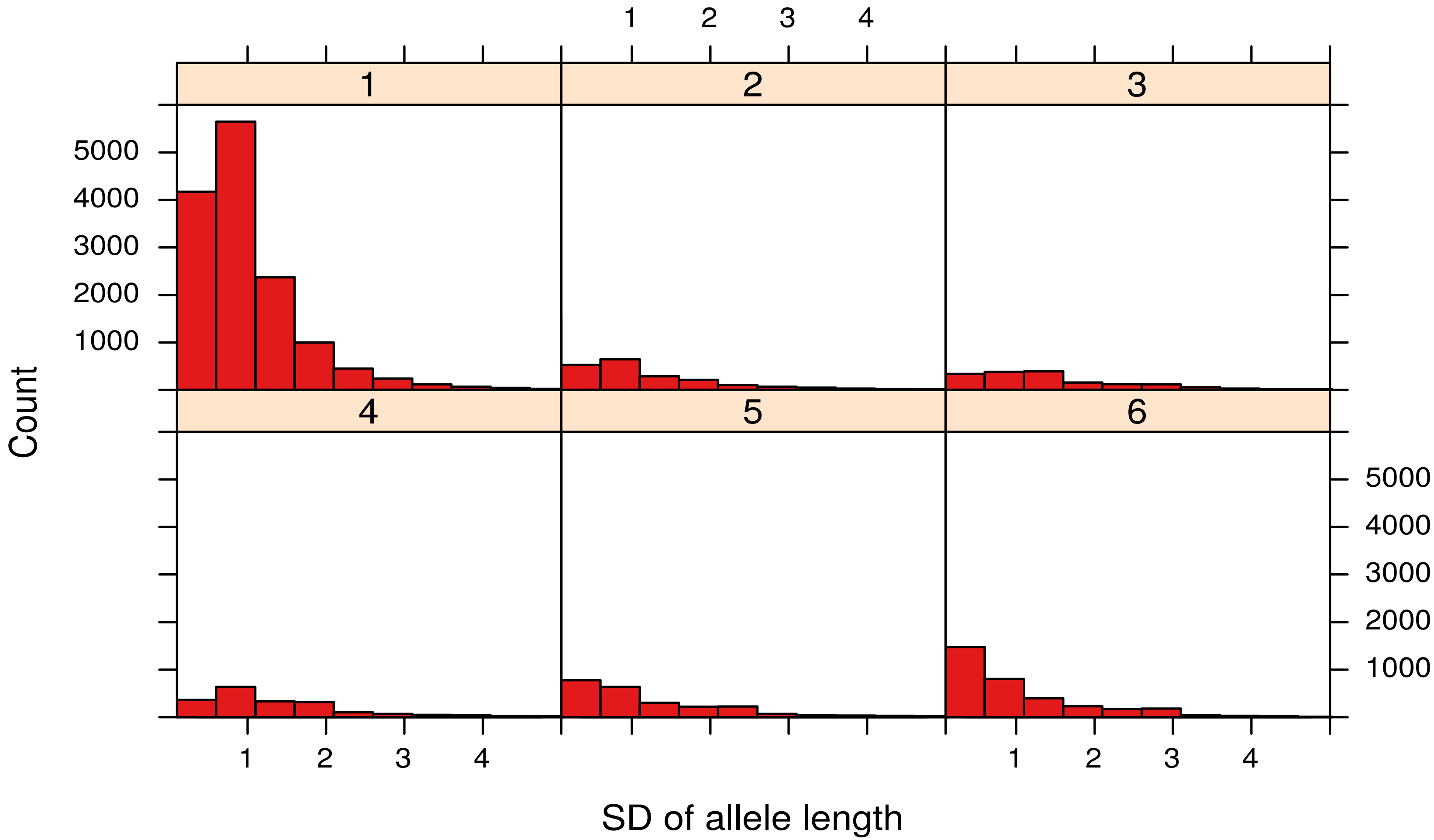
Length of most commonly observed allele, by motif length



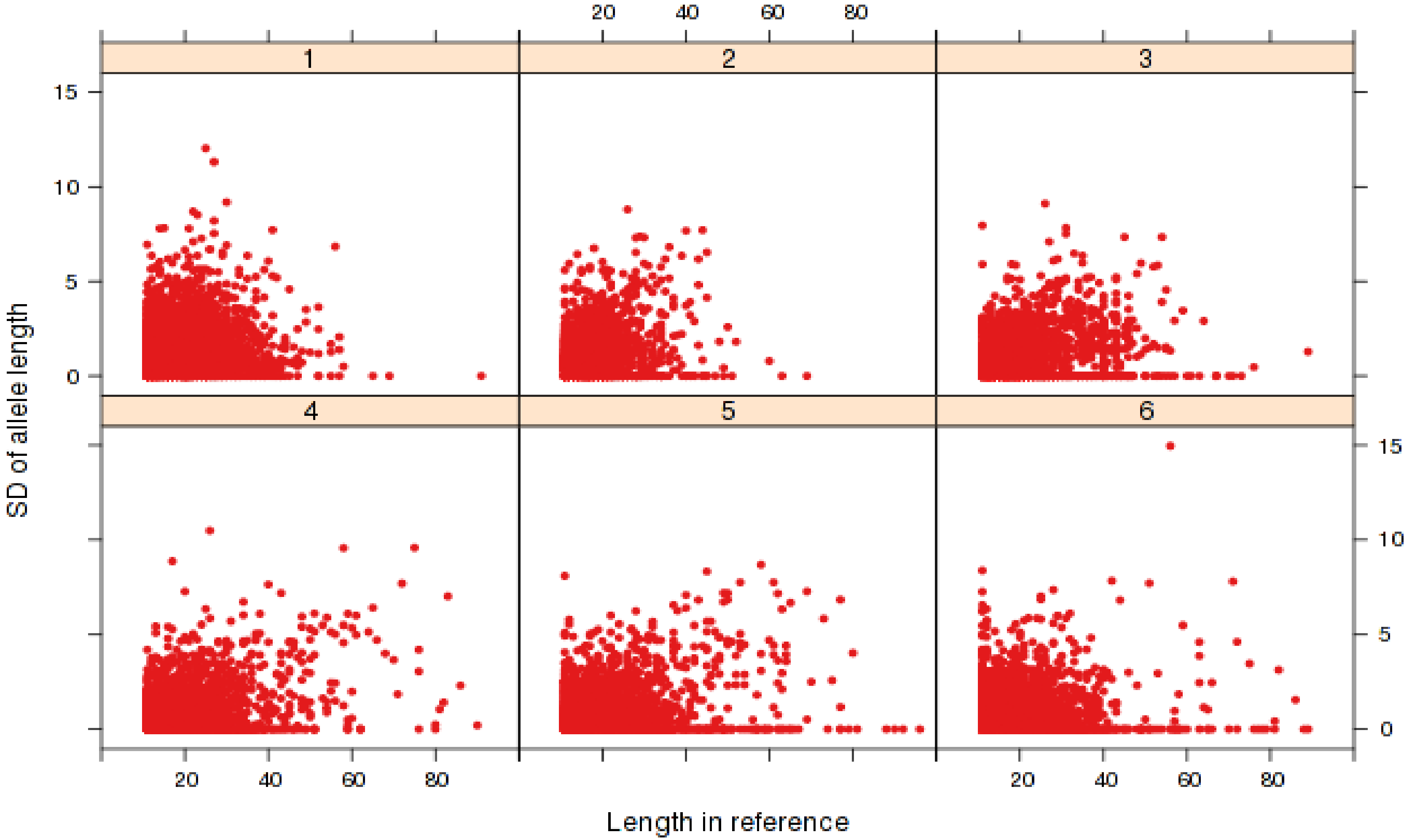
Range of allele lengths, by motif length



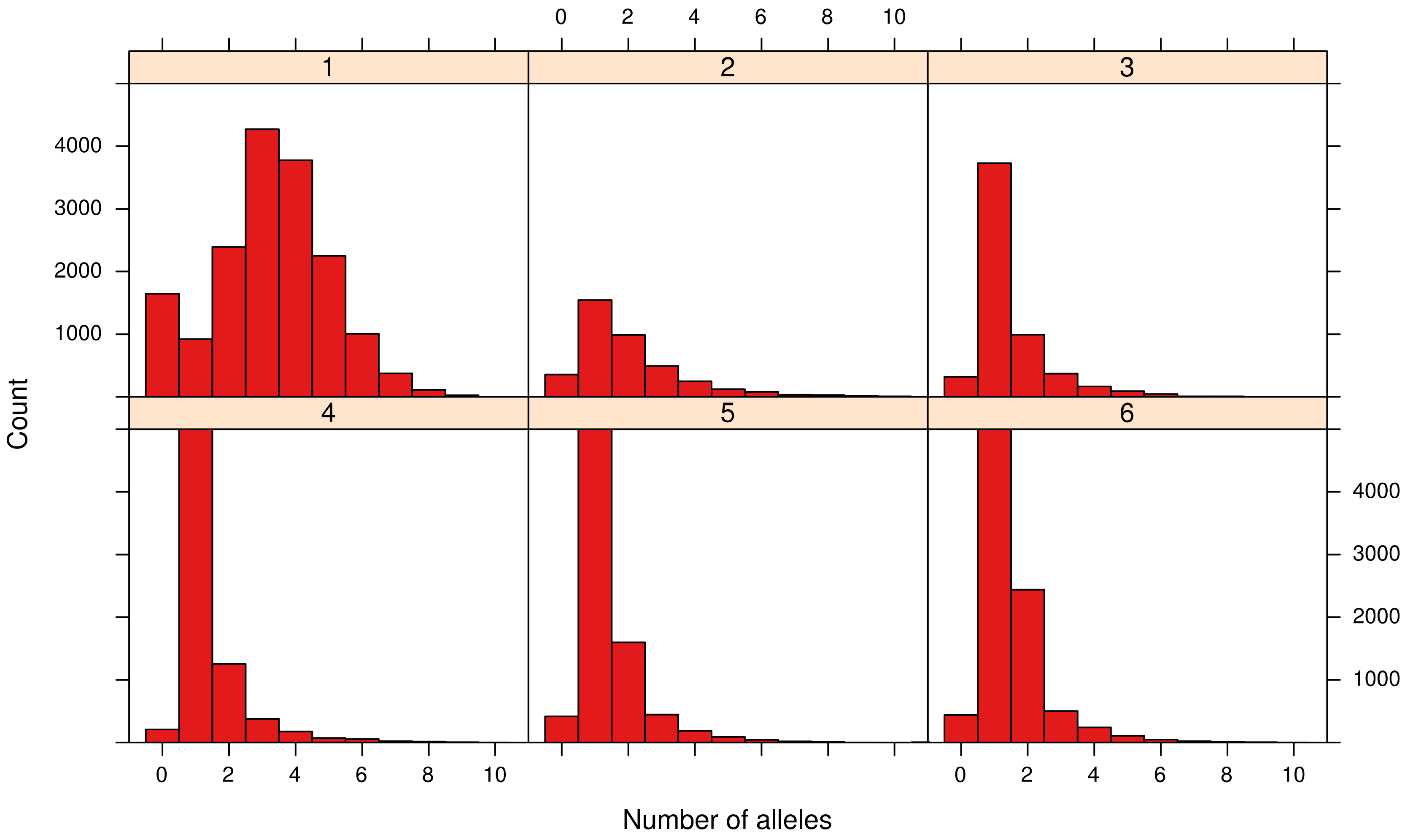
Distribution of allele length SD, by motif length

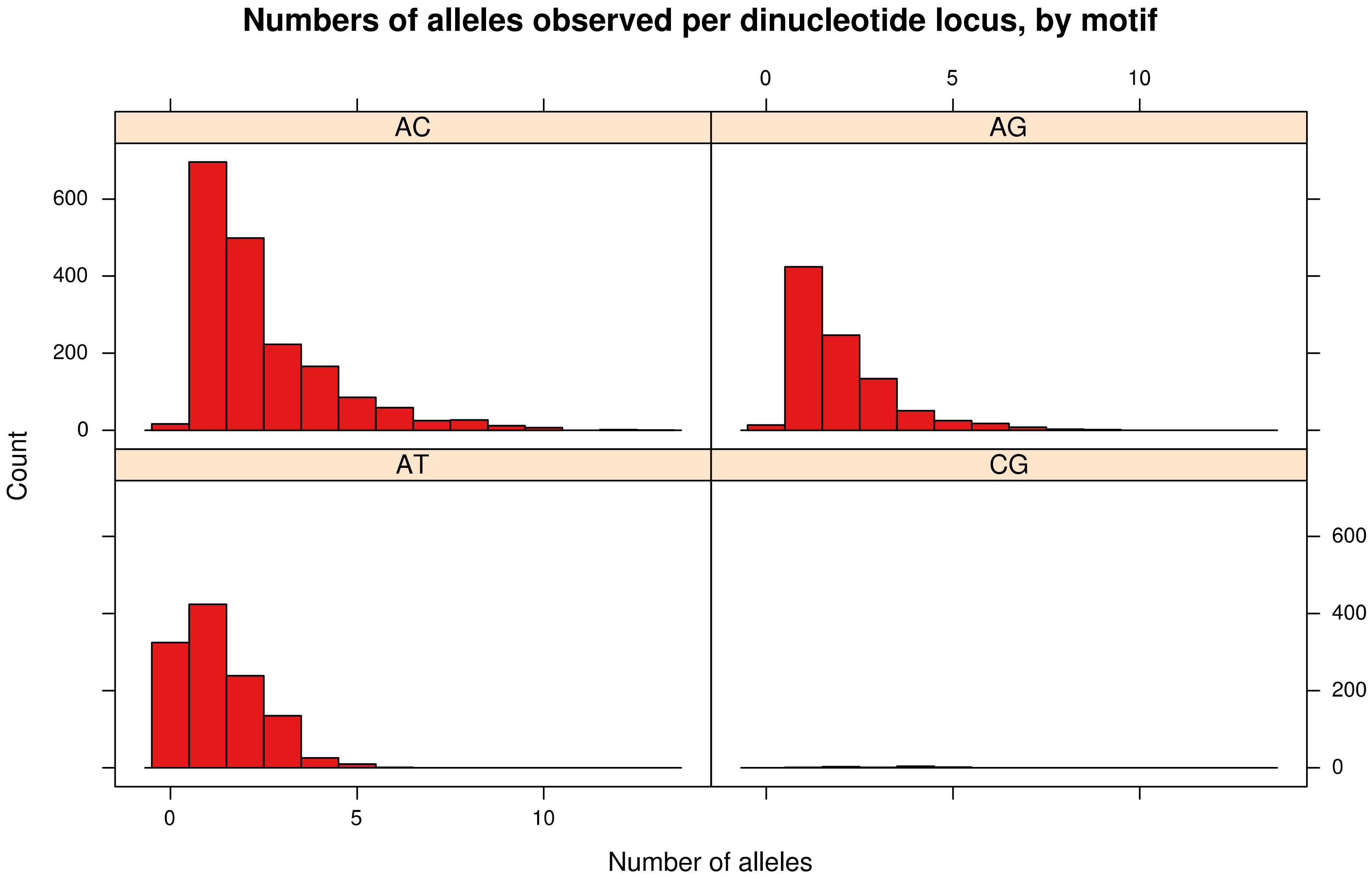


SD of allele length vs. length in reference, by motif length

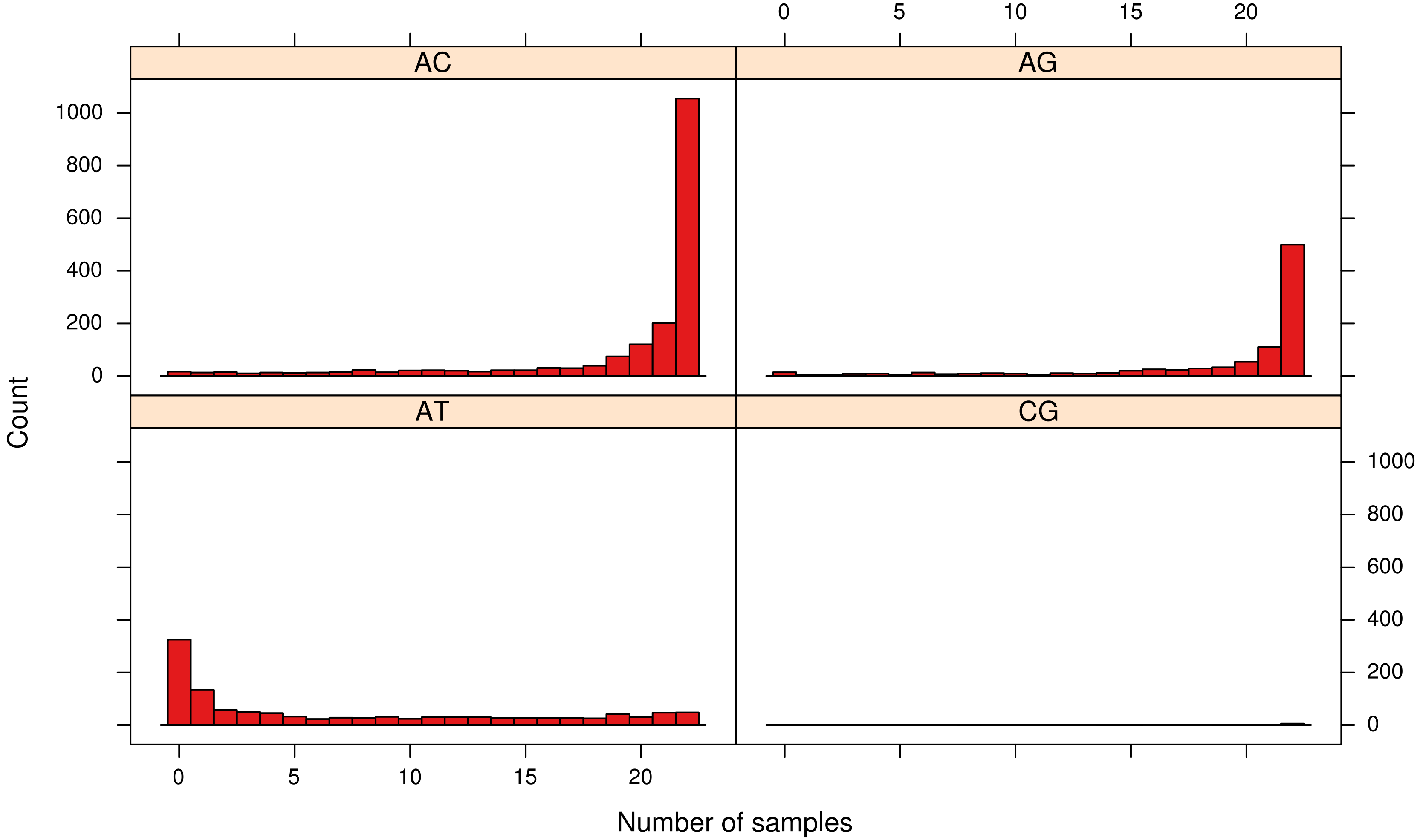


Numbers of alleles observed per locus, by motif length




Numbers of alleles observed per dinucleotide locus, by motif
AC
AG
AT
CG
Count
Number of alleles
0
5
10
0
200
400
600

Numbers of samples with data for each dinucleotide locus, by motif



What now?

Analyse ancient samples.

Need to choose summary statistics that can distinguish models. Any suggestions would be most welcome!

Look at purity of loci, for models that incorporate point mutation.

Ascertainment bias is a problem:

1. can't detect long loci (but these are rare).
2. AT-rich motifs less likely to be observed (lower coverage).

Different motifs will have to be analysed independently.

Thanks!

Barbara Holland

Human Frontier Science Program

Griffith University Ancient DNA Lab:

- Dave Lambert
- Matt Parks
- Sankar Subramanian

All of you for listening!