UNIVERSITY OF TASMANIA

Microsatellite evolution in Adélie 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 models

Infinite allele model (IAM)

- Mutation involves any number of repeats and always results in a new allele.

Microsatellite models

Infinite allele model (IAM)

- Mutation involves any number of repeats and always results in a new allele.

K-allele model (KAM)

- K possible allelic states, any allele has constant probability of mutation to any other state.

Microsatellite models

Infinite allele model (IAM)

- Mutation involves any number of repeats and always results in a new allele.

K-allele model (KAM)

- K possible allelic states, any allele has constant probability of mutation to any other state.

Stepwise mutation model (SMM)

- Loss or gain (with equal probability) of a single repeat.

Microsatellite models

Infinite allele model (IAM)

- Mutation involves any number of repeats and always results in a new allele.

K-allele model (KAM)

- K possible allelic states, any allele has constant probability of mutation to any other state.

Stepwise mutation model (SMM)

- Loss or gain (with equal probability) of a single repeat.

Generalised stepwise model (GSM)

- Each mutation adds or removes X repeats, where X follows a geometric distribution.

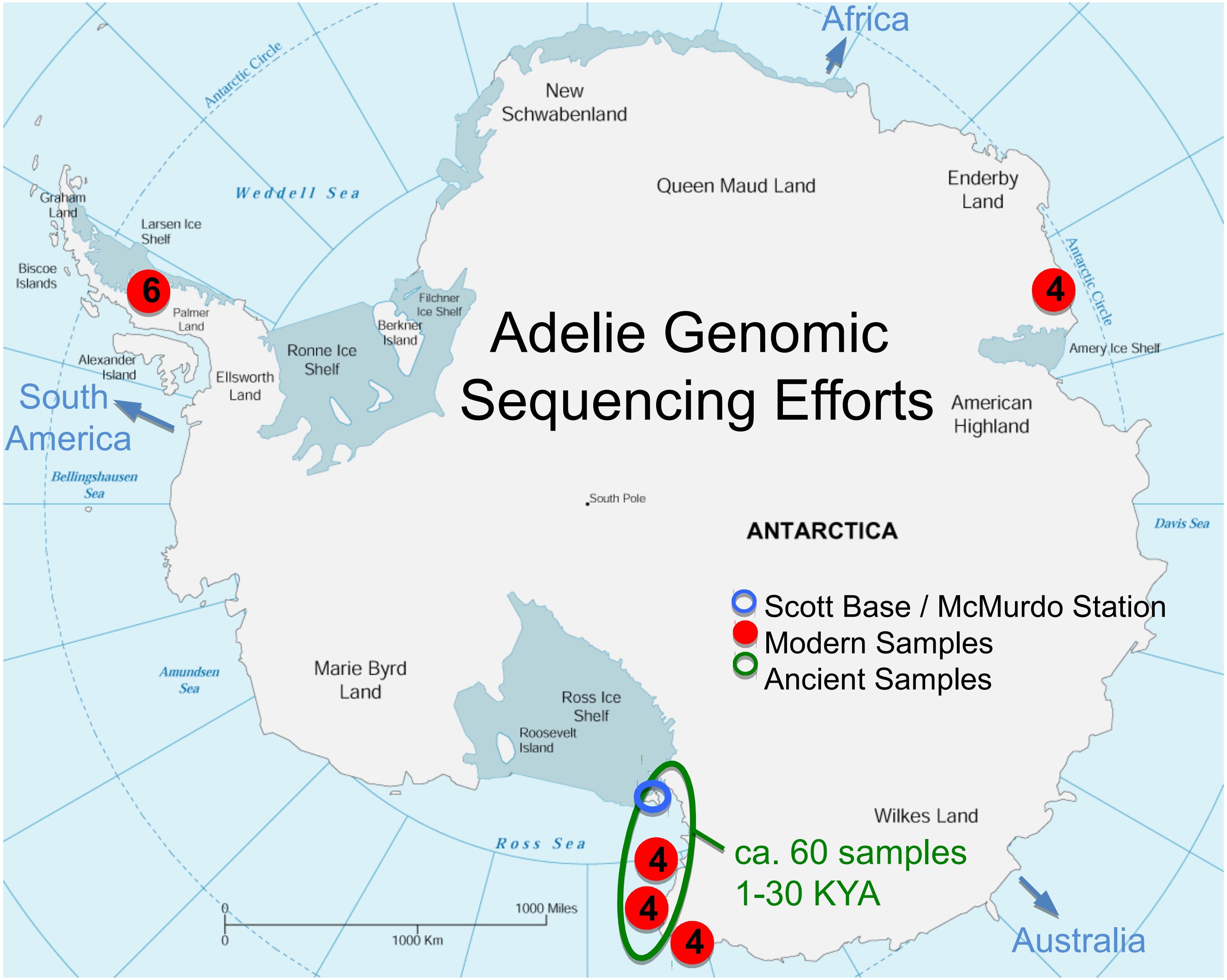
Adélie penguin data

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

Have been nesting on exposed areas of coastline for thousands of years – dead chicks and guano preserved.

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

Also lower-coverage reads for 22 ancient genomes up to 30,000 years old from several sites.



Adelie Genomic Sequencing Efforts

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

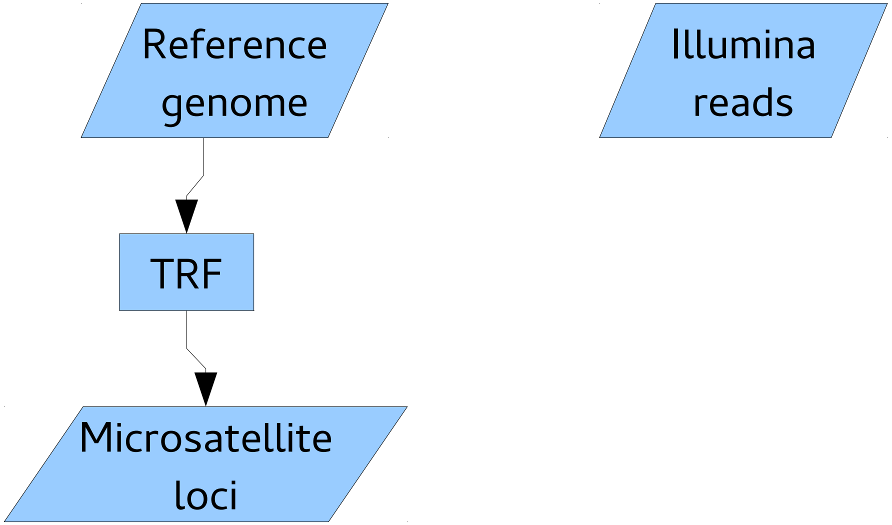
ca. 60 samples
1-30 KYA

Microsatellite detection

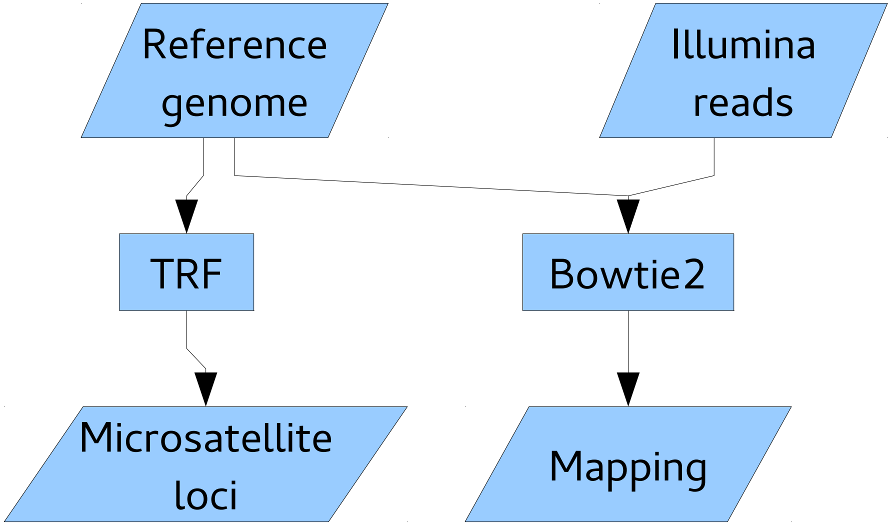
Reference genome

Illumina reads

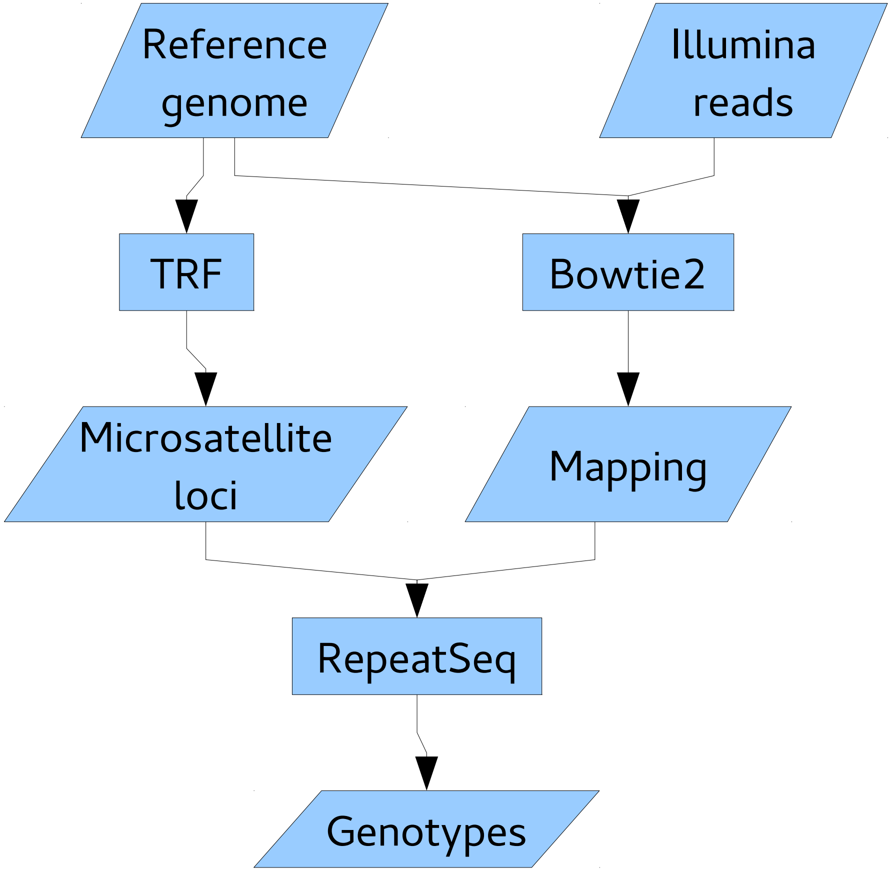
Microsatellite detection



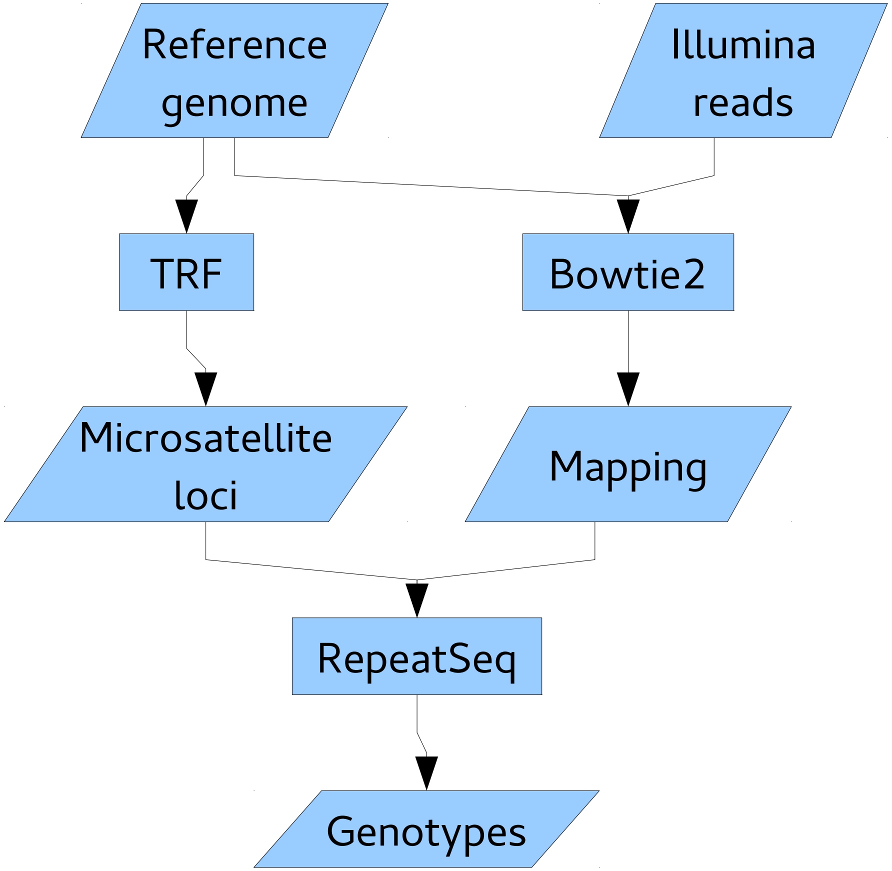
Microsatellite detection



Microsatellite detection



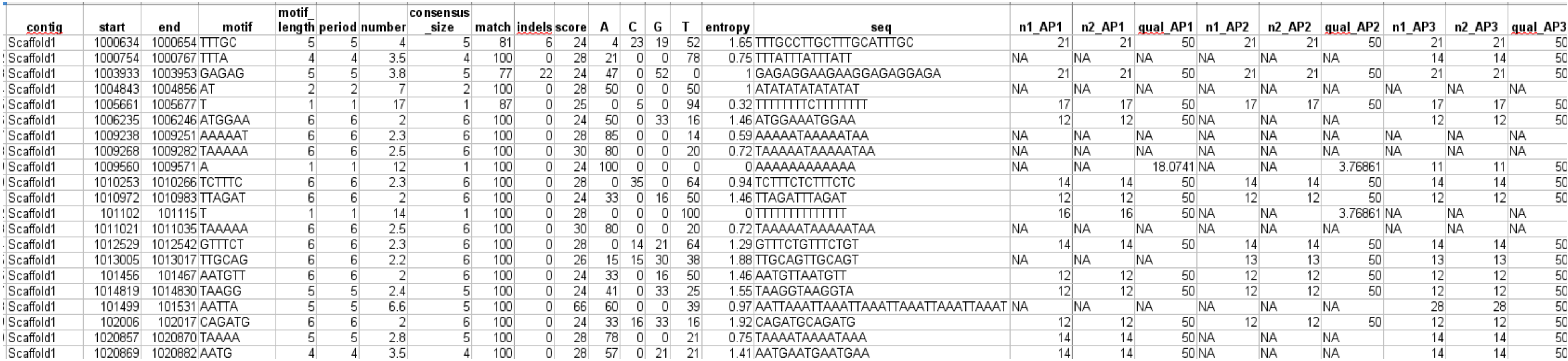
Microsatellite detection



Numbers of loci detected in Adélie reference genome using Tandem Repeat Finder:

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

Genotype data

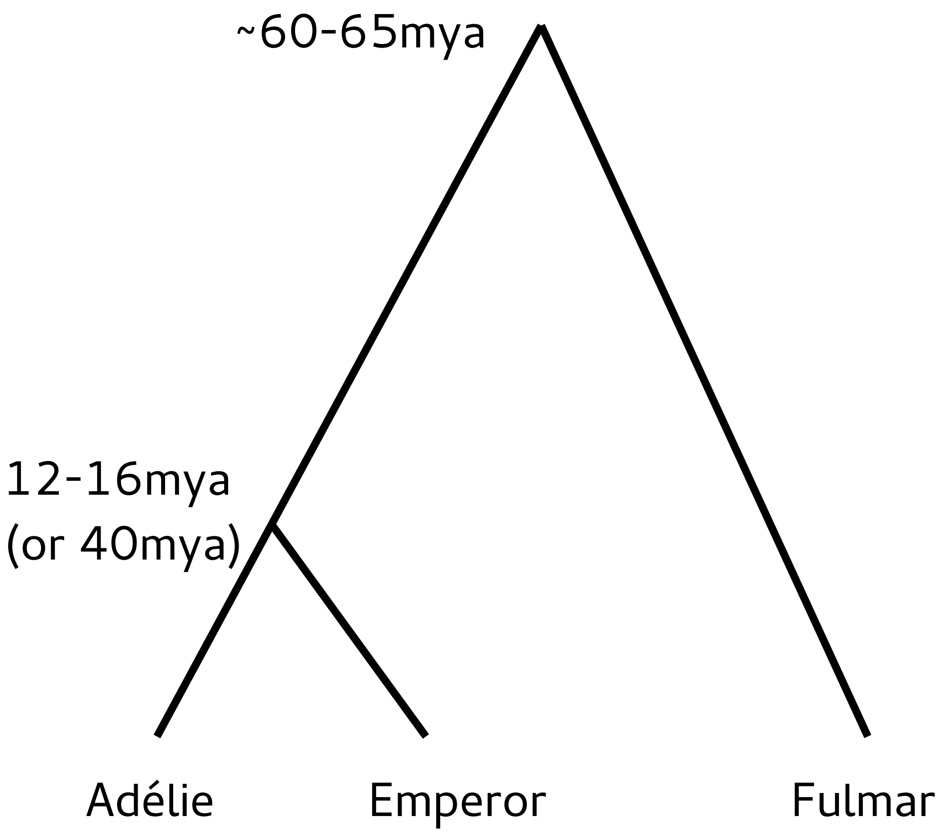


contig	start	end	motif	motif_ length	period	number	consensus _size	match	indels	score	A	C	G	T	entropy	seq	n1_AP1	n2_AP1	qual_AP1	n1_AP2	n2_AP2	qual_AP2	n1_AP3	n2_AP3	qual_AP3
Scaffold1	1000634	1000654	TTTGC	5	5	4	5	81	6	24	4	23	19	52	1.65	TTTGCCTTGCTTTGCATTTGC	21	21	50	21	21	50	21	21	50
Scaffold1	1000754	1000767	TTTA	4	4	3.5	4	100	0	28	21	0	0	78	0.75	TTTATTTATTTATT	NA	NA	NA	NA	NA	NA	14	14	50
Scaffold1	1003933	1003953	GAGAG	5	5	3.8	5	77	22	24	47	0	52	0	1	GAGAGGAAGAAGGAGAGGAGA	21	21	50	21	21	50	21	21	50
Scaffold1	1004843	1004856	AT	2	2	7	2	100	0	28	50	0	0	50	1	ATATATATATATAT	NA	NA	NA	NA	NA	NA	NA	NA	NA
Scaffold1	1005661	1005677	T	1	1	17	1	87	0	25	0	5	0	94	0.32	TTTTTTTTCTTTTTTTT	17	17	50	17	17	50	17	17	50
Scaffold1	1006235	1006246	ATGGAA	6	6	2	6	100	0	24	50	0	33	16	1.46	ATGGAAATGGAA	12	12	50	NA	NA	NA	12	12	50
Scaffold1	1009238	1009251	AAAAAT	6	6	2.3	6	100	0	28	85	0	0	14	0.59	AAAAATAAAAATAA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Scaffold1	1009268	1009282	TAAAAA	6	6	2.5	6	100	0	30	80	0	0	20	0.72	TAAAAATAAAAATAA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Scaffold1	1009560	1009571	A	1	1	12	1	100	0	24	100	0	0	0	0	AAAAAAAAAAAA	NA	NA	18.0741	NA	NA	3.76861	11	11	50
Scaffold1	1010253	1010266	TCTTTC	6	6	2.3	6	100	0	28	0	35	0	64	0.94	TCTTTCTCTTTCTC	14	14	50	14	14	50	14	14	50
Scaffold1	1010972	1010983	TTAGAT	6	6	2	6	100	0	24	33	0	16	50	1.46	TTAGATTTAGAT	12	12	50	12	12	50	12	12	50
Scaffold1	101102	101115	T	1	1	14	1	100	0	28	0	0	0	100	0	TTTTTTTTTTTTTT	16	16	50	NA	NA	3.76861	NA	NA	NA
Scaffold1	1011021	1011035	TAAAAA	6	6	2.5	6	100	0	30	80	0	0	20	0.72	TAAAAATAAAAATAA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Scaffold1	1012529	1012542	GTTTCT	6	6	2.3	6	100	0	28	0	14	21	64	1.29	GTTTCTGTTTCTGT	14	14	50	14	14	50	14	14	50
Scaffold1	1013005	1013017	TTGCAG	6	6	2.2	6	100	0	26	15	15	30	38	1.88	TTGCAGTTGCAGT	NA	NA	NA	13	13	50	13	13	50
Scaffold1	101456	101467	AATGTT	6	6	2	6	100	0	24	33	0	16	50	1.46	AATGTTAATGTT	12	12	50	12	12	50	12	12	50
Scaffold1	1014819	1014830	TAAGG	5	5	2.4	5	100	0	24	41	0	33	25	1.55	TAAGGTAAGGTA	12	12	50	12	12	50	12	12	50
Scaffold1	101499	101531	AATTA	5	5	6.6	5	100	0	66	60	0	0	39	0.97	AATTAAATTAAATTAAATTAAATTAAATTAAAT	NA	NA	NA	NA	NA	NA	28	28	50
Scaffold1	102006	102017	CAGATG	6	6	2	6	100	0	24	33	16	33	16	1.92	CAGATGCAGATG	12	12	50	12	12	50	12	12	50
Scaffold1	1020857	1020870	TAAAA	5	5	2.8	5	100	0	28	78	0	0	21	0.75	TAAAATAAAATAAA	14	14	50	NA	NA	NA	14	14	50
Scaffold1	1020869	1020882	AATG	4	4	3.5	4	100	0	28	57	0	21	21	1.41	AATGAATGAATGAA	14	14	50	NA	NA	NA	14	14	50

Identifying older microsatellite loci

Run Tandem Repeat Finder on more distantly related reference genomes (emperor penguin and northern fulmar).

Map modern Adélie reads to these genomes, genotype samples and process output as before.



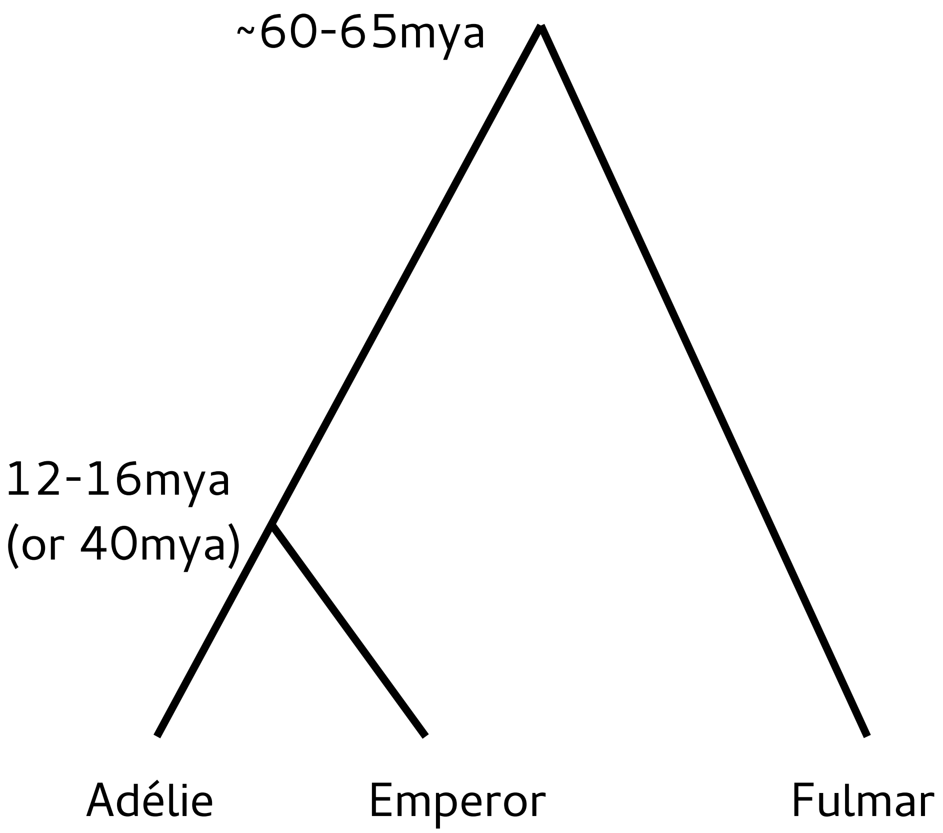
Identifying older microsatellite loci

Run Tandem Repeat Finder on more distantly related reference genomes (emperor penguin and northern fulmar).

Map modern Adélie reads to these genomes, genotype samples and process output as before.

Assumption: Loci that can be genotyped must be older than divergence between species.

We expect older loci to have longer alleles on average.



Identifying older microsatellite loci

808,828 loci genotyped using emperor reference, 327,668 using fulmar, but many of these may not be microsatellites in Adélie.

Look only at polymorphic loci – these should be "active" microsatellites.

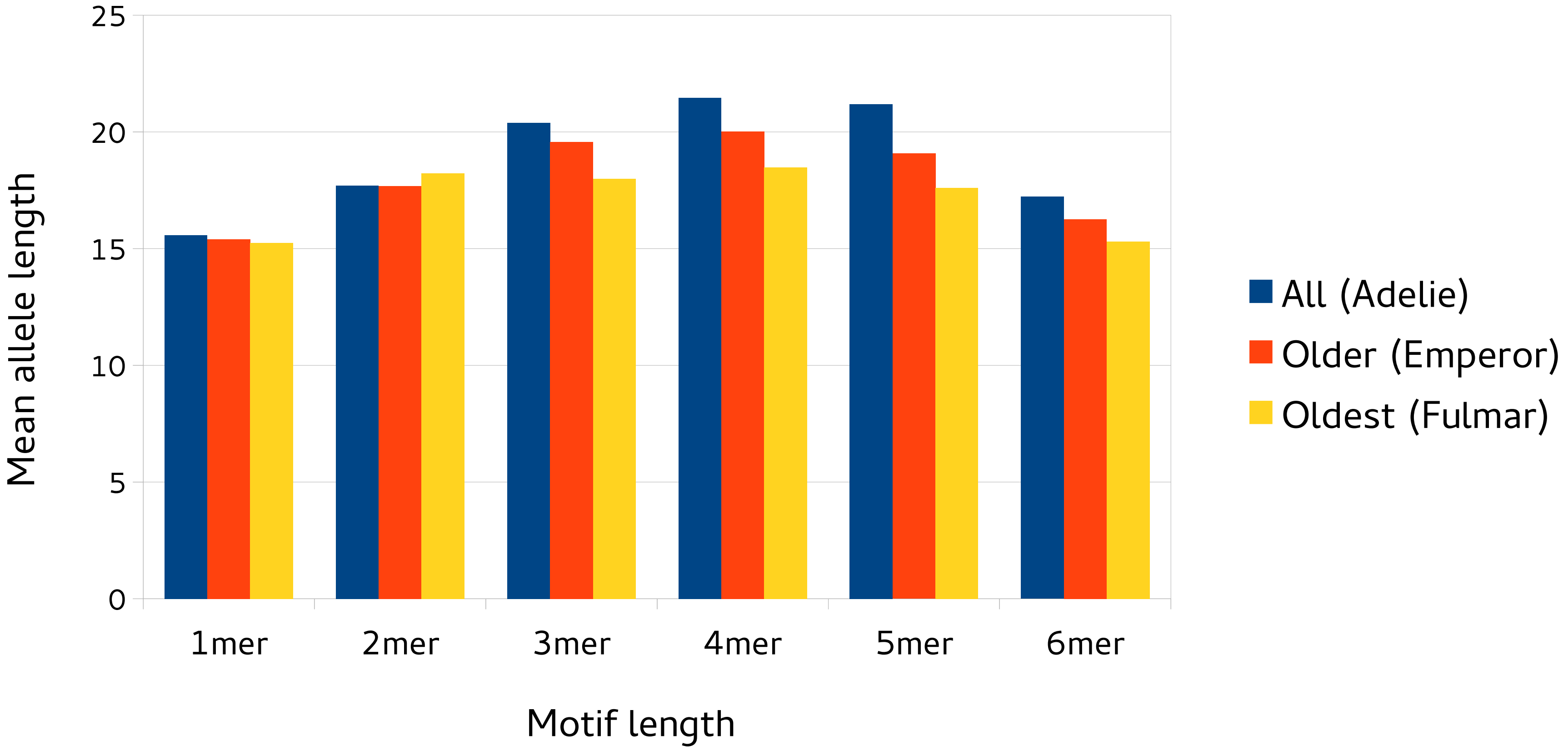
Identifying older microsatellite loci

808,828 loci genotyped using emperor reference, 327,668 using fulmar, but many of these may not be microsatellites in Adélie.

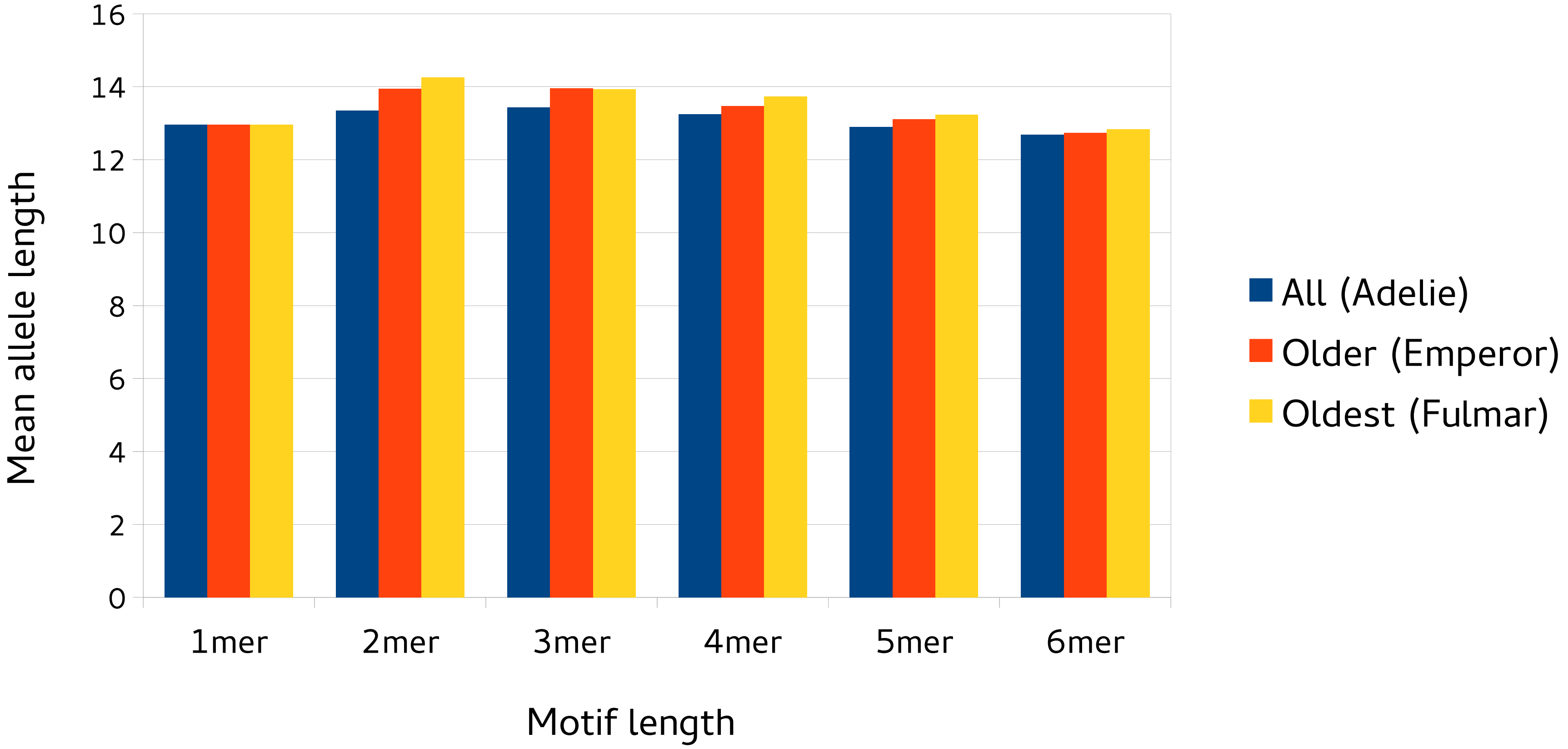
Look only at polymorphic loci – these should be "active" microsatellites.

	Number of loci			Mean length		
	Adélie	Emperor	Fulmar	Adélie	Emperor	Fulmar
1mer	79,971	39,829	7,367	15.57	15.40	15.24
2mer	12,008	6,536	900	17.70	17.68	18.22
3mer	8,786	6,109	1,048	20.38	19.57	17.99
4mer	9,837	7,014	1,240	21.45	20.02	18.48
5mer	12,767	9,647	2,012	21.18	19.07	17.60
6mer	16,027	13,527	2,924	17.22	16.26	15.30
Total	139,396	82,662	15,491	17.18	16.85	16.18

Mean allele lengths in loci of different ages

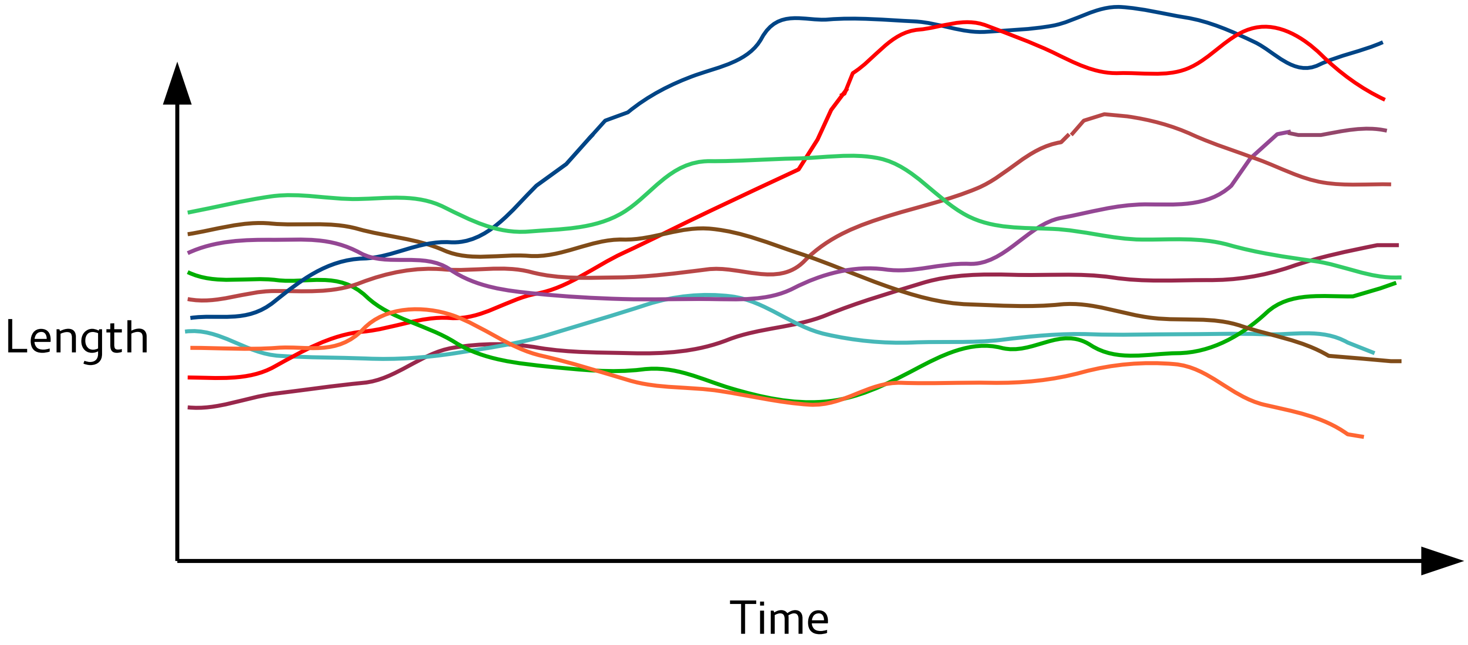


Mean allele lengths in short loci



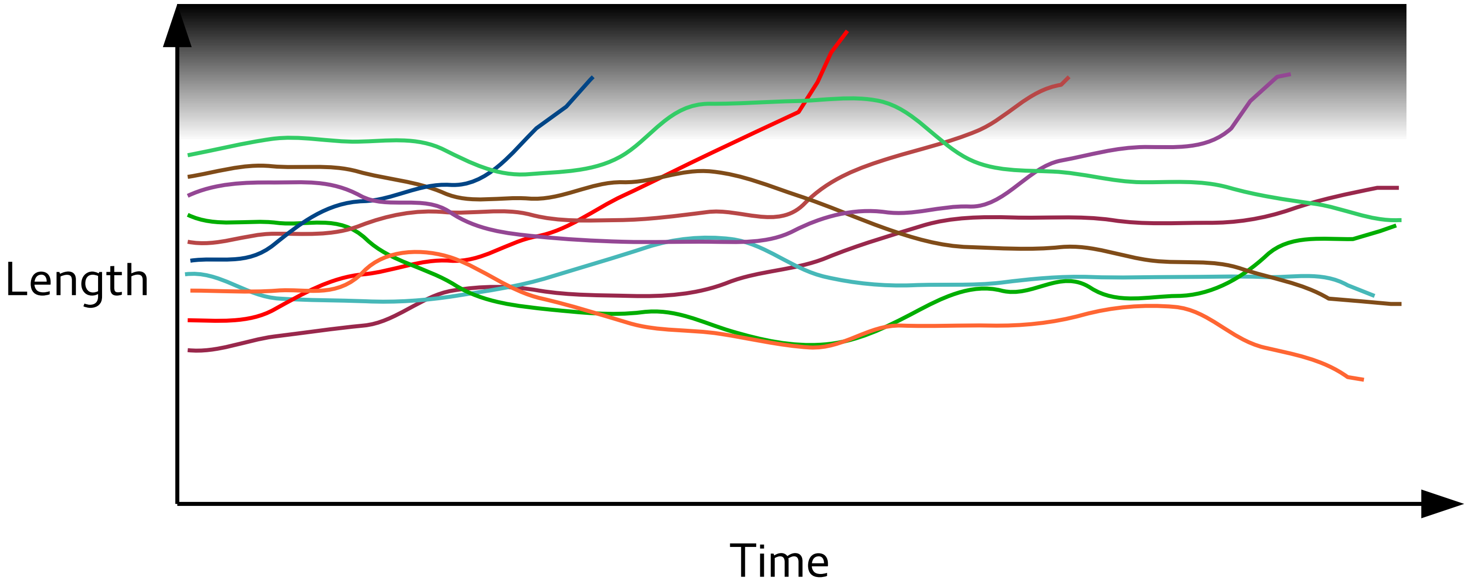
New model

Could be explained by a censoring effect - longer alleles are more likely to accumulate mutations, and cease to function as microsatellites as a result.



New model

Could be explained by a censoring effect - longer alleles are more likely to accumulate mutations, and cease to function as microsatellites as a result.



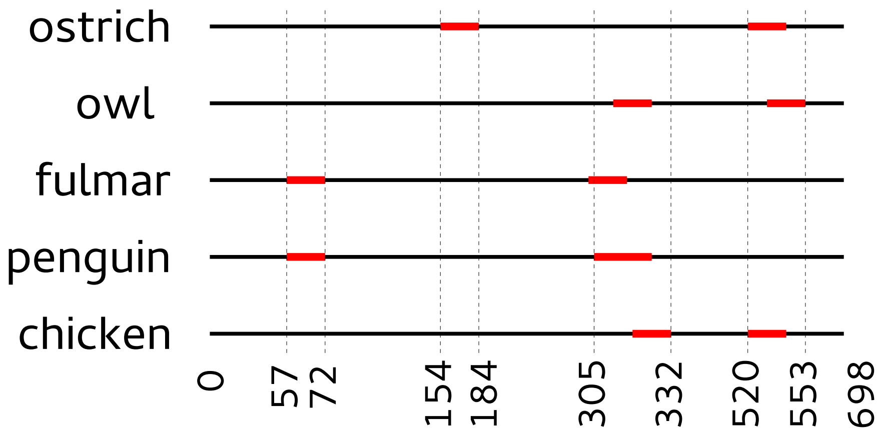
Better age estimates

We have an alignment of 48 complete bird genomes, and a phylogeny.

We have identified microsatellite loci in all 48 genomes.

We can map loci to standard coordinates using chicken genome as reference.

Use ancestral state reconstruction under a Dollo model to estimate when each microsatellite was 'born' on the tree.



Thanks!

Barbara Holland

Human Frontier Science Program

Griffith University Ancient DNA Lab:

- Dave Lambert
- Matt Parks
- Sankar Subramanian

All of you for listening!