What's in a Mutt? An Intro to Dog DNA Analysis

Lecture 4

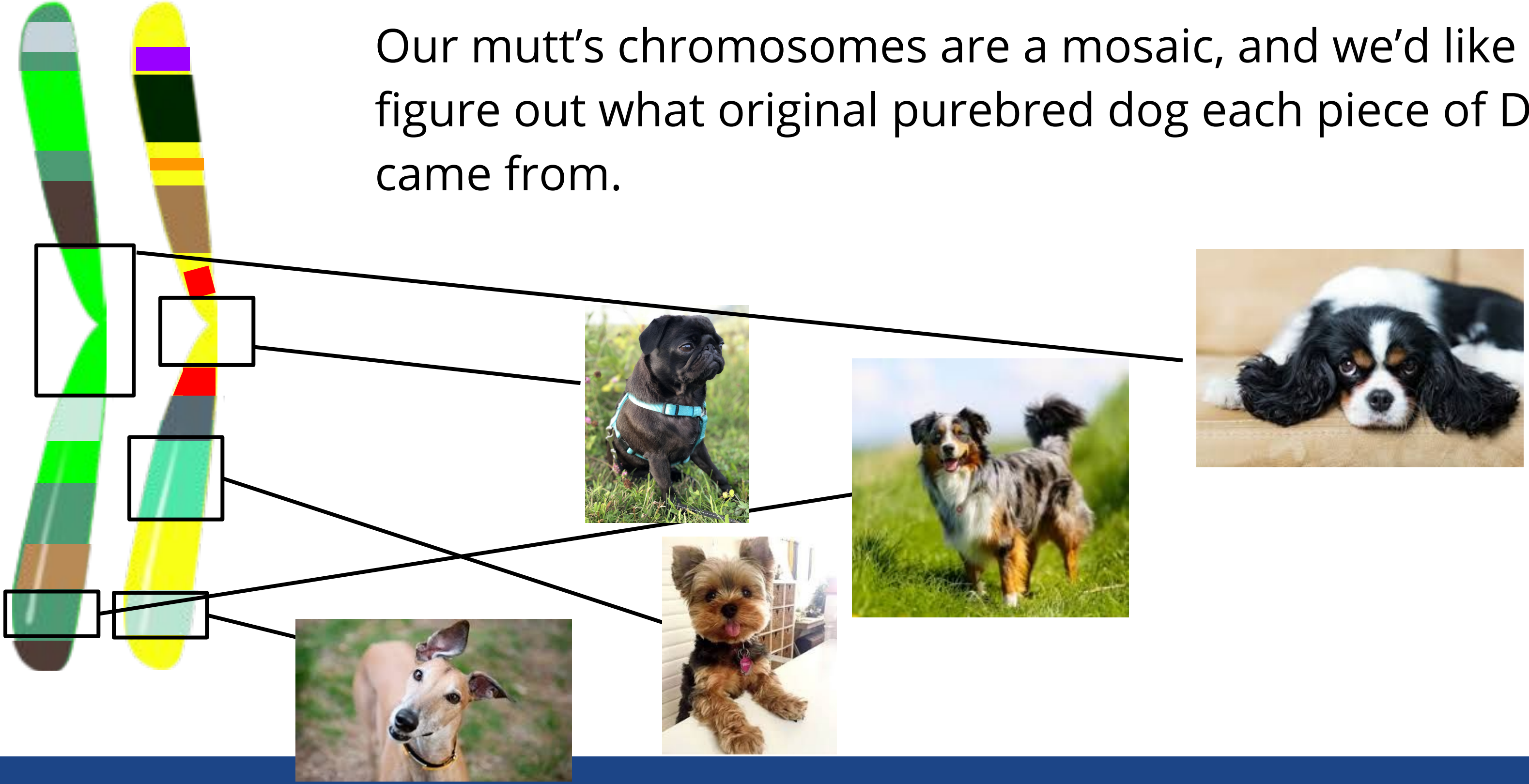
Jan 14th, 2019

Recap

Our mutt’s chromosomes are a mosaic, and we’d like to figure out what original purebred dog each piece of DNA came from.

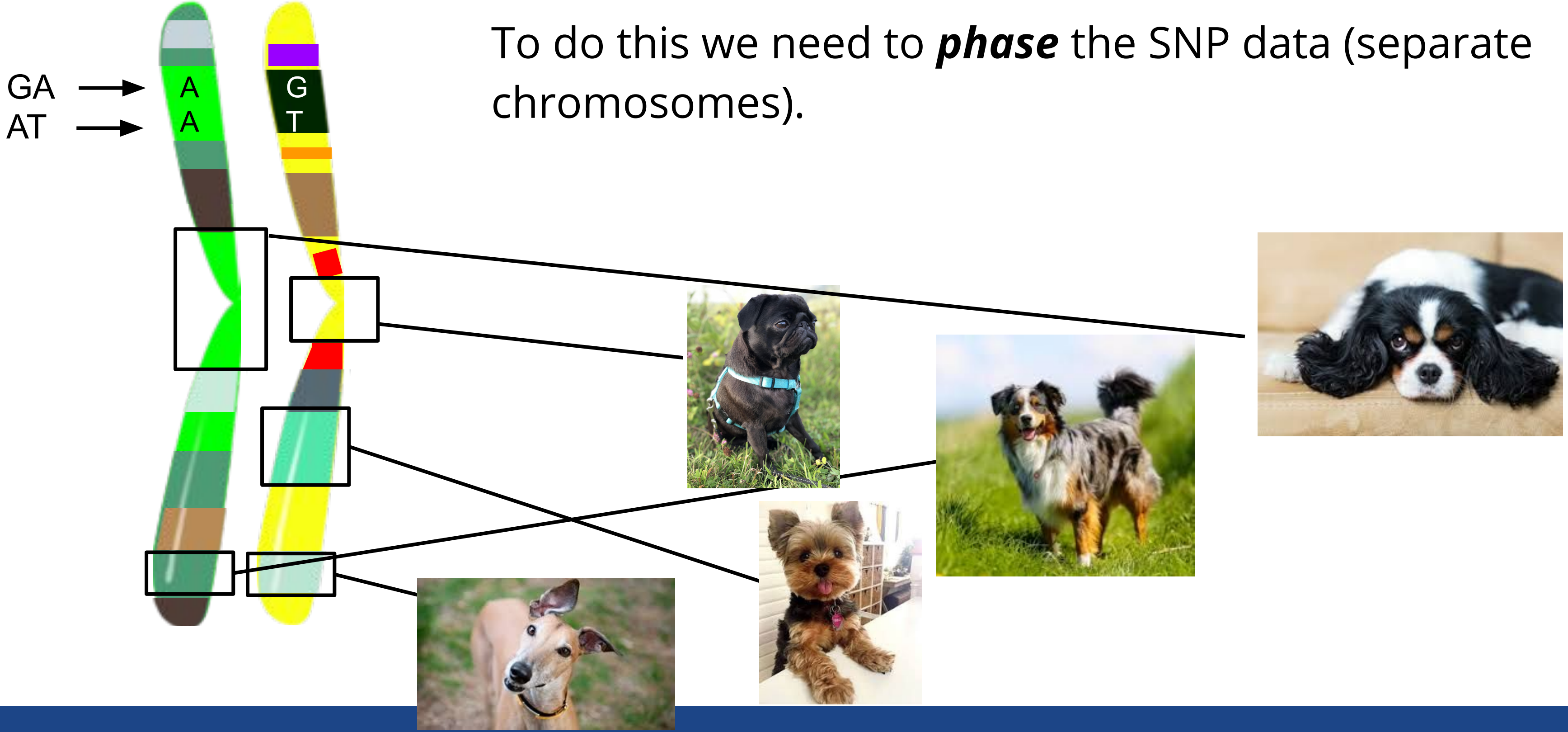
Recap

Our mutt's chromosomes are a mosaic, and we'd like to figure out what original purebred dog each piece of DNA came from.



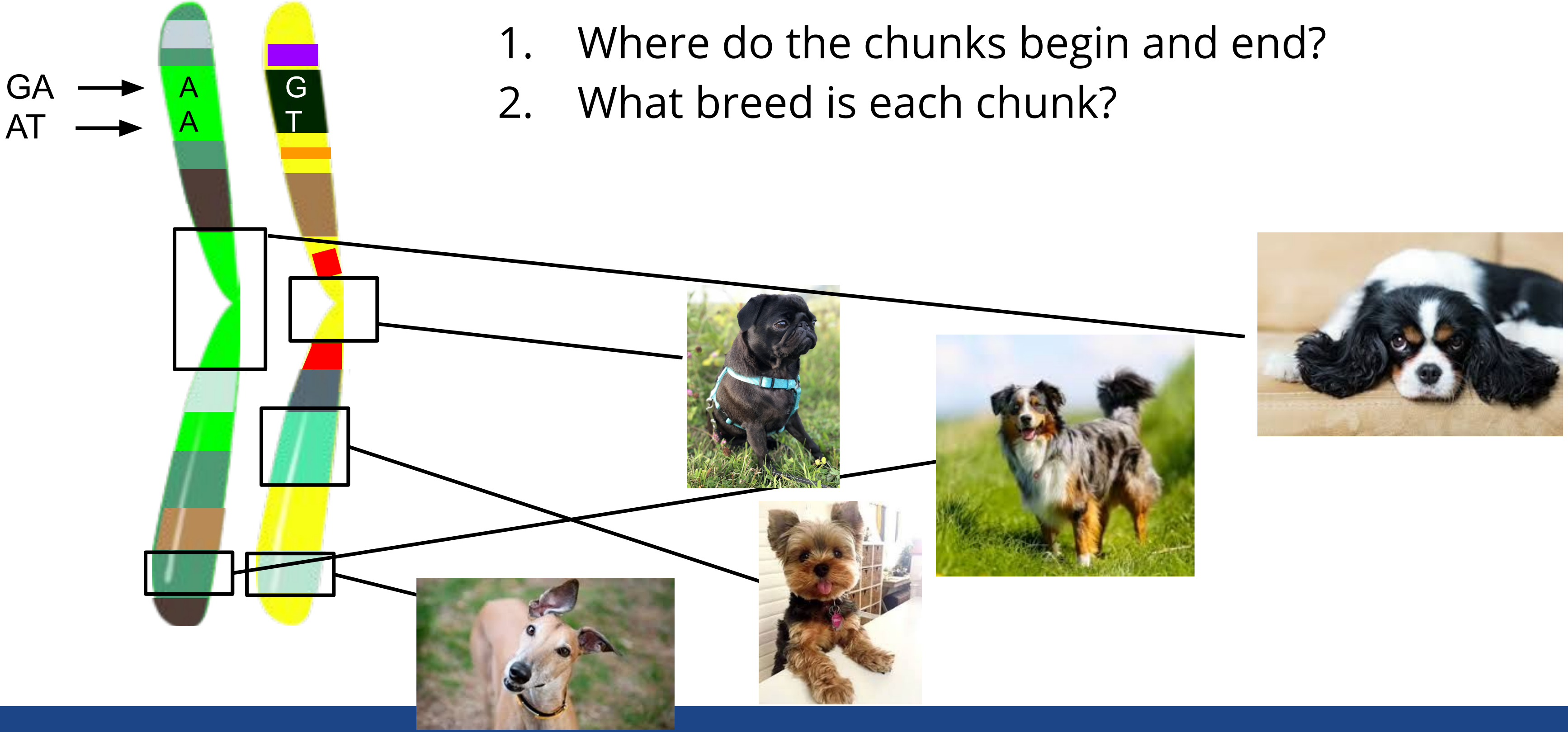
Recap

To do this we need to ***phase*** the SNP data (separate chromosomes).



Recap

1. Where do the chunks begin and end?
2. What breed is each chunk?



Recap: Comparing to purebreds

For now, let’s assume we know what breed each chunk is.

How might we go about determining the breed of each?

Comparing to purebreds

For now, let’s assume we know what breed each chunk is.

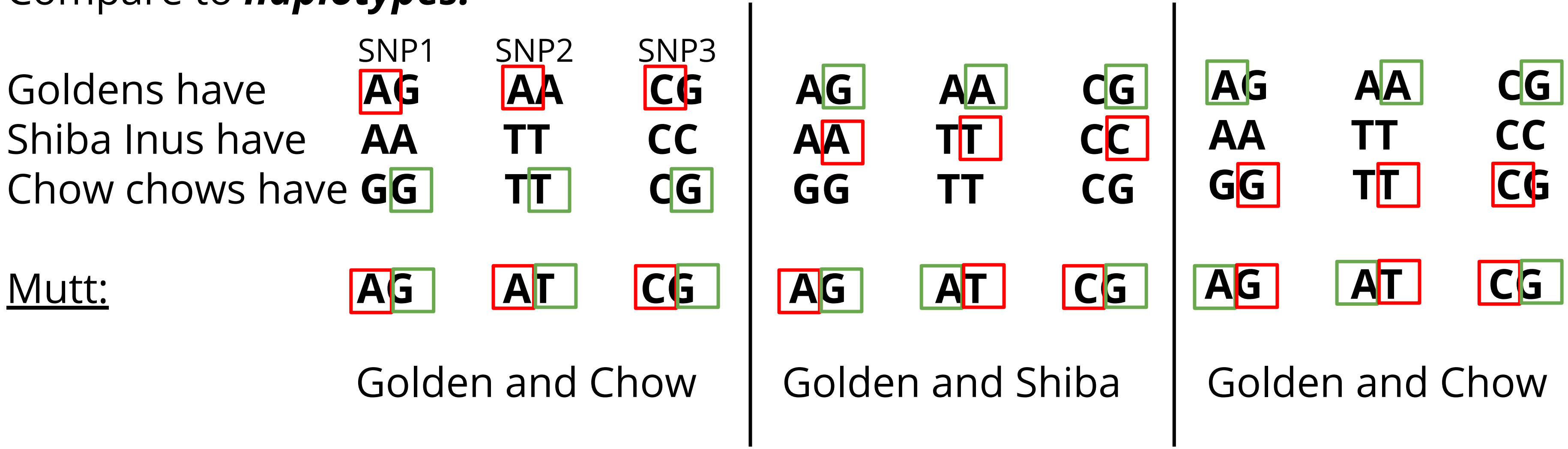
Compare to ***haplotypes:***

	SNP1	SNP2	SNP3
Goldens have	**AG**	**AA**	**CG**
Shiba Inus have	**AA**	**TT**	**CC**
Chow chows have	**GG**	**TT**	**CG**
Mutt:	**AG**	**AT**	**CG**

Comparing to purebreds

For now, let's assume we know what breed each chunk is.

Compare to ***haplotypes:***

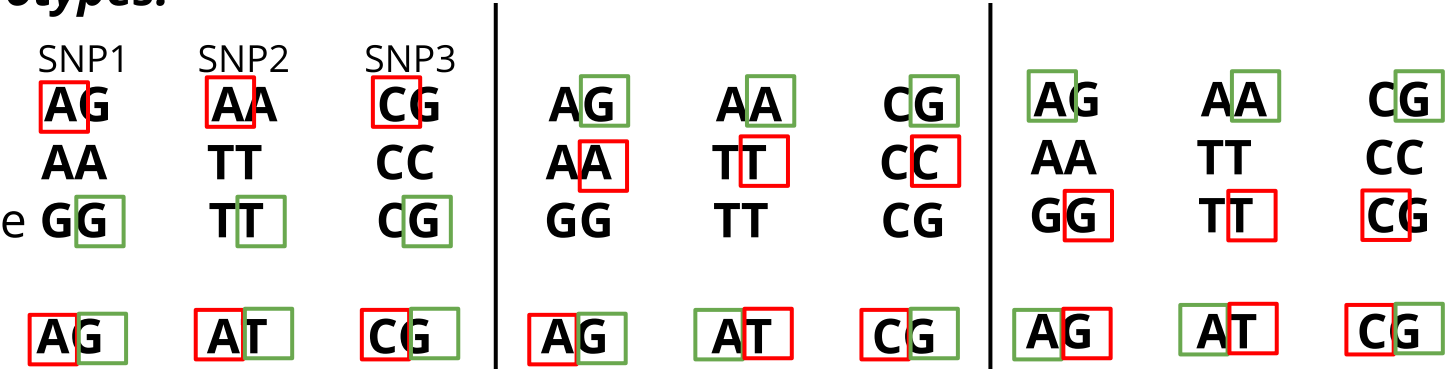


	SNP1	SNP2	SNP3						
Goldens have	AG	AA	CG	AG	AA	CG	AG	AA	CG
Shiba Inus have	AA	TT	CC	AA	TT	CC	AA	TT	CC
Chow chows have	GG	TT	CG	GG	TT	CG	GG	TT	CG
<u>Mutt:</u>	AG	AT	CG	AG	AT	CG	AG	AT	CG
	Golden and Chow			Golden and Shiba			Golden and Chow		

Comparing to purebreds

Fourth combo [ATG] and [GAC] not possible; could be Golden and Unknown

Compare to ***haplotypes:***



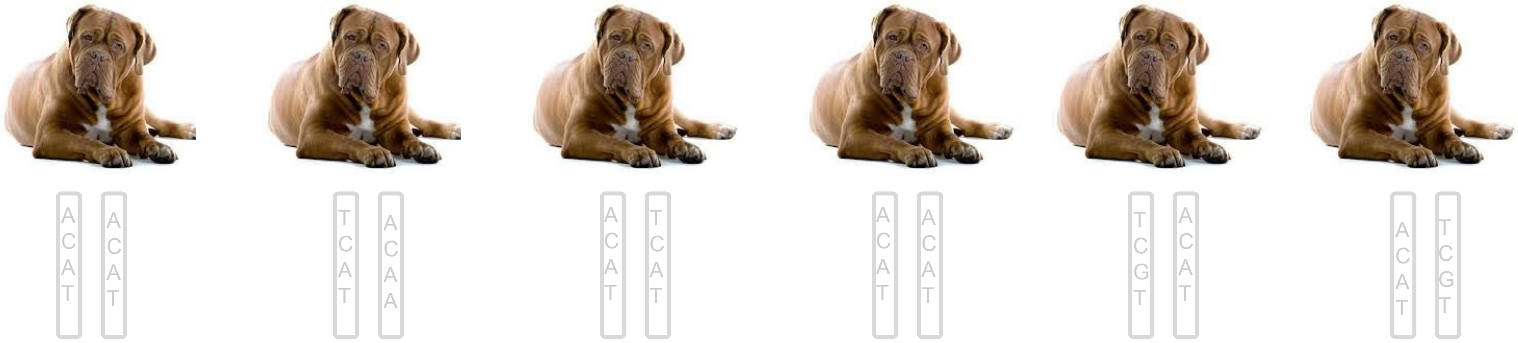
	SNP1	SNP2	SNP3						
Goldens have	AG	AA	CG	AG	AA	CG	AG	AA	CG
Shiba Inus have	AA	TT	CC	AA	TT	CC	AA	TT	CC
Chow chows have	GG	TT	CG	GG	TT	CG	GG	TT	CG
Mutt:	AG	AT	CG	AG	AT	CG	AG	AT	CG
	Golden and Chow			Golden and Shiba			Golden and Chow		

Comparing to purebreds

How is this picture different from what our purebred data actually look like?

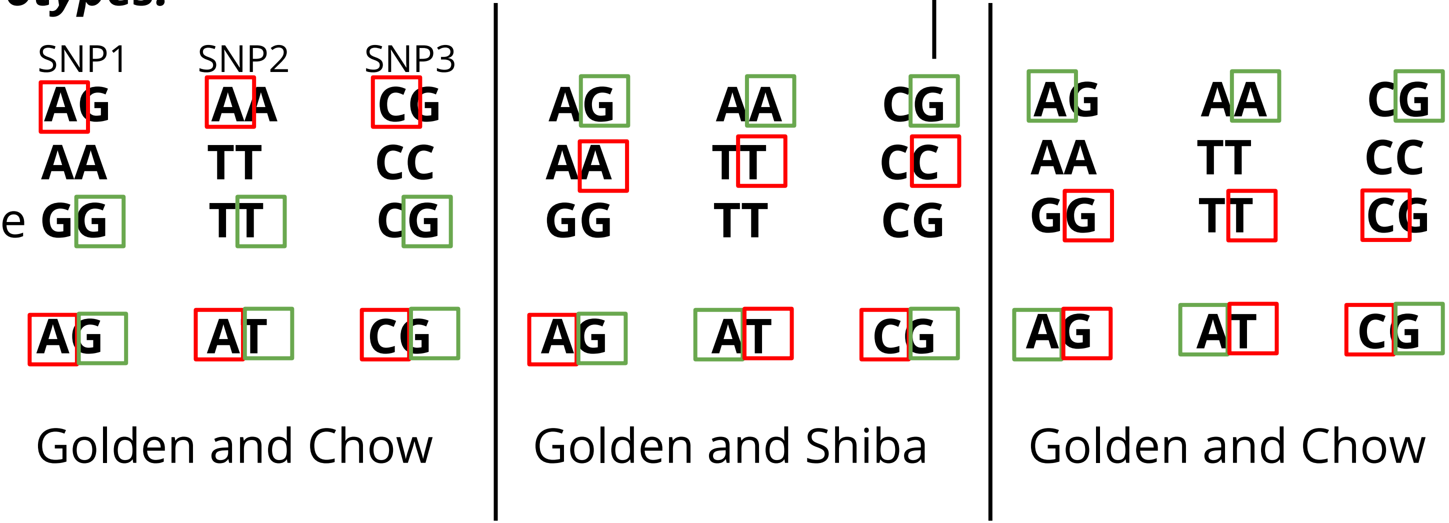
Comparing to purebreds

- Six dogs per breed
  - So we see multiple genotypes per purebred
- Phased purebred data
  - So we might only see certain allele combinations for adjacent SNPs



Comparing to purebreds

Compare to ***haplotypes:***



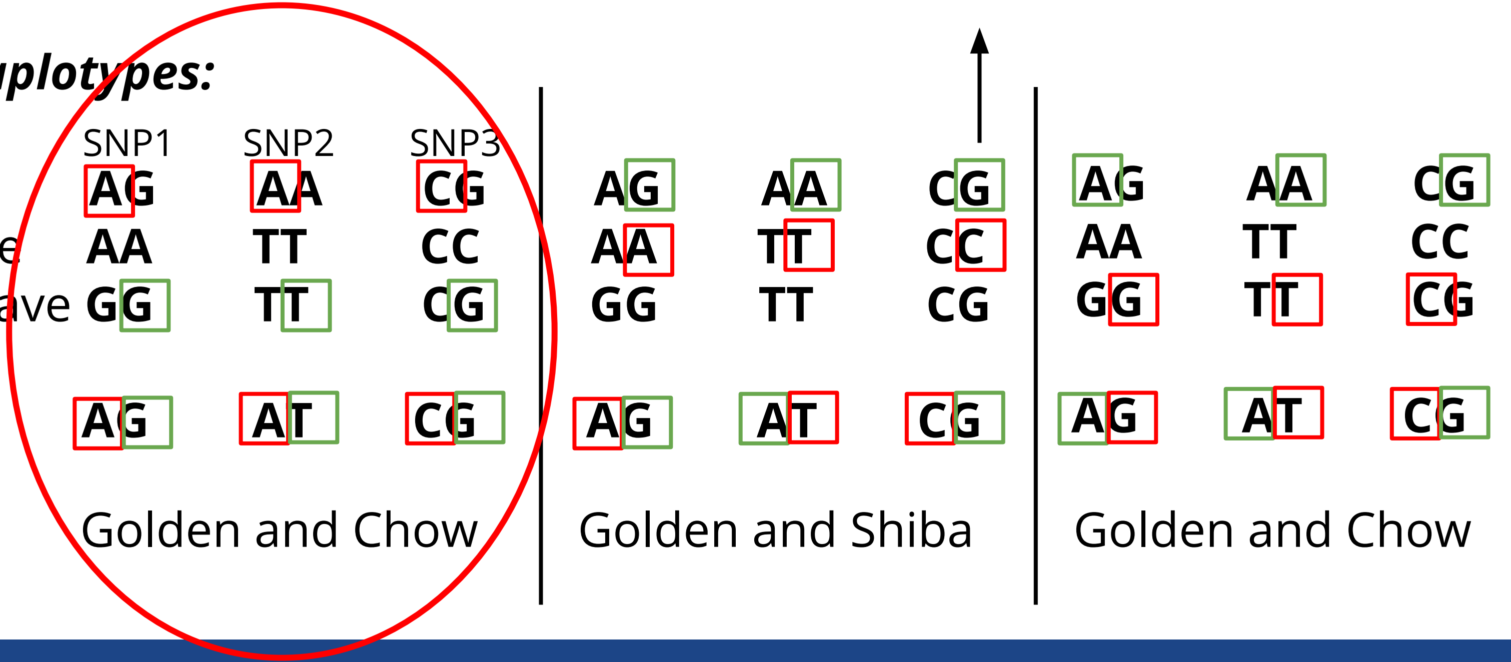
	SNP1	SNP2	SNP3						
Goldens have	AG	AA	CG	AG	AA	CG	AG	AA	CG
Shiba Inus have	AA	TT	CC	AA	TT	CC	AA	TT	CC
Chow chows have	GG	TT	CG	GG	TT	CG	GG	TT	CG
<u>Mutt:</u>	AG	AT	CG	AG	AT	CG	AG	AT	CG
	Golden and Chow			Golden and Shiba			Golden and Chow		

Let’s say for SNP3, for goldens, we see *G* 10% of the time, shiba: 2%, and chows: 30%.

Comparing to purebreds

Let’s say for SNP3, for goldens,
we see *G* 10% of the time,
shiba: 2%, and chows: 30%.

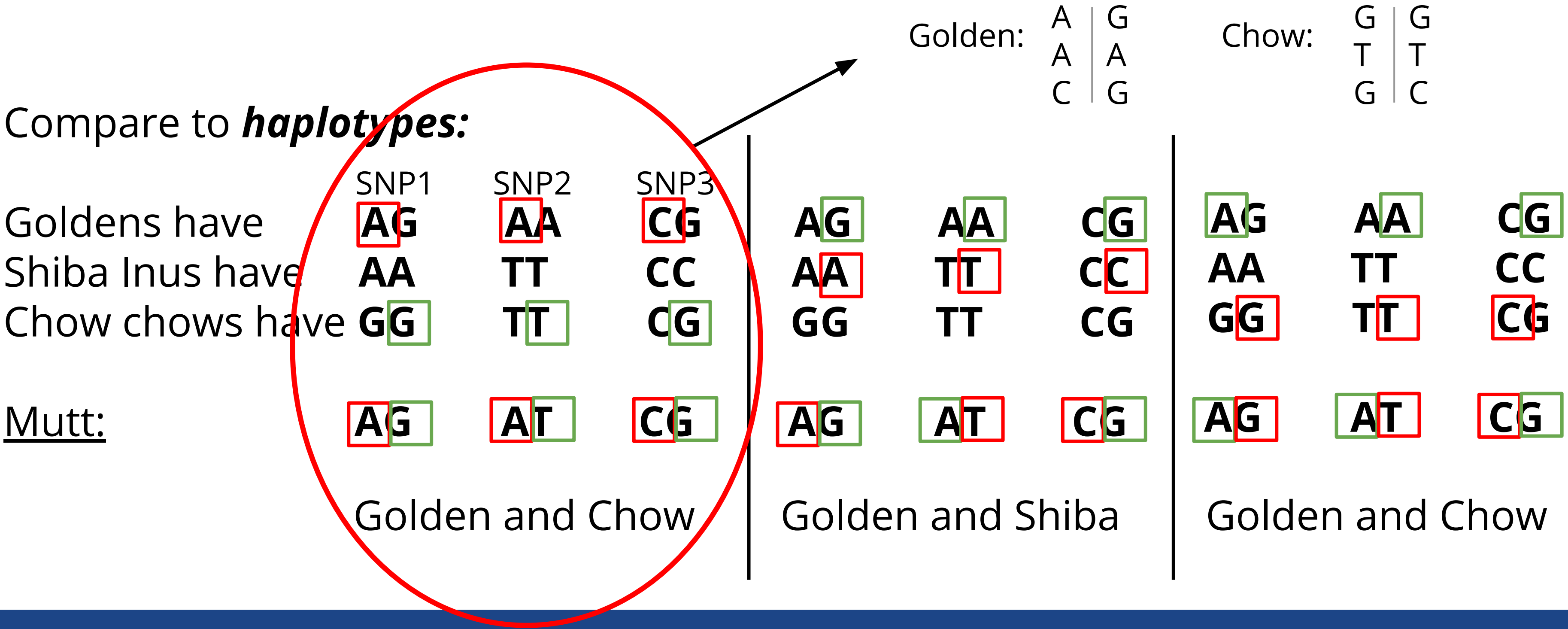
Compare to ***haplotypes:***



	SNP1	SNP2	SNP3						
Goldens have	AG	AA	CG	AG	AA	CG	AG	AA	CG
Shiba Inus have	AA	TT	CC	AA	TT	CC	AA	TT	CC
Chow chows have	GG	TT	CG	GG	TT	CG	GG	TT	CG
Mutt:	AG	AT	CG	AG	AT	CG	AG	AT	CG
	Golden and Chow			Golden and Shiba			Golden and Chow		

Comparing to purebreds

So based on our mutt, the most likely phasing for a golden and a chow with these genotypes is:



Golden:			Chow:		
	A	G		G	G
	A	A		T	T
	C	G		G	C

Compare to ***haplotypes:***

	SNP1	SNP2	SNP3
Goldens have	AG	AA	CG
Shiba Inus have	AA	TT	CC
Chow chows have	GG	TT	CG
Mutt:	AG	AT	CG

Golden and Chow

AG	AA	CG
AA	TT	CC
GG	TT	CG
AG	AT	CG

Golden and Shiba

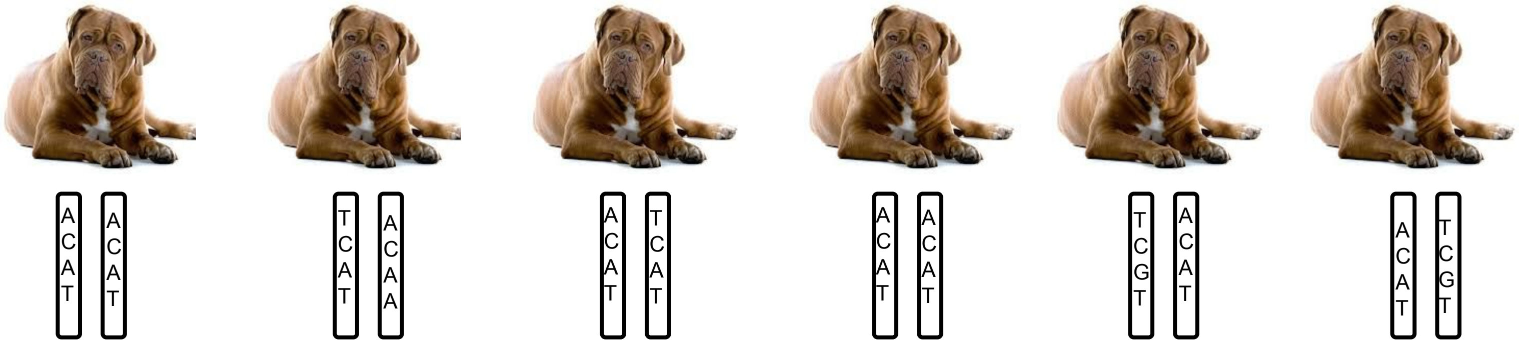
AG	AA	CG
AA	TT	CC
GG	TT	CG
AG	AT	CG

Golden and Chow

Comparing to purebreds

- Six dogs per breed
  - So we see multiple genotypes per purebred

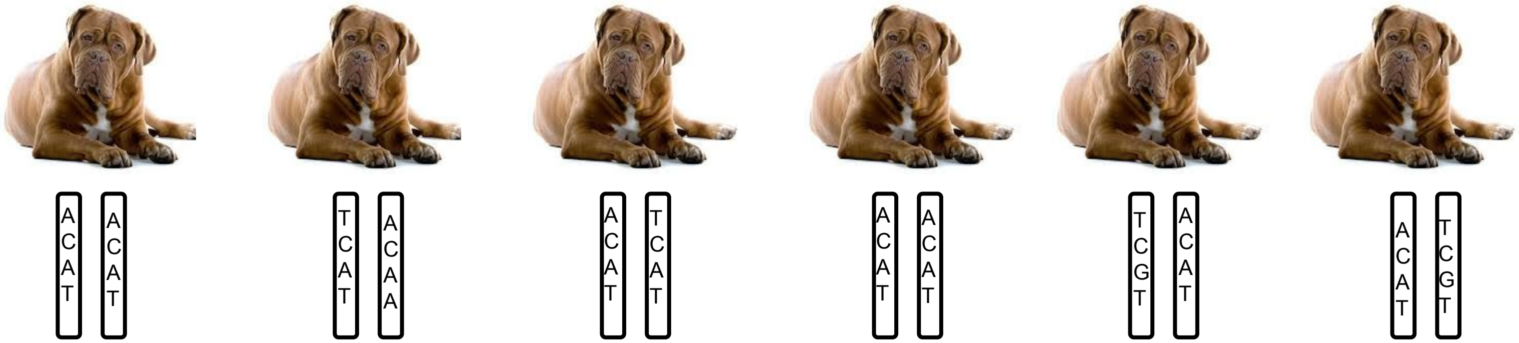
- Phased purebred data
  - So we might only see certain allele combinations for adjacent SNPs



Comparing to purebreds

- Six dogs per breed
- Phased purebred data

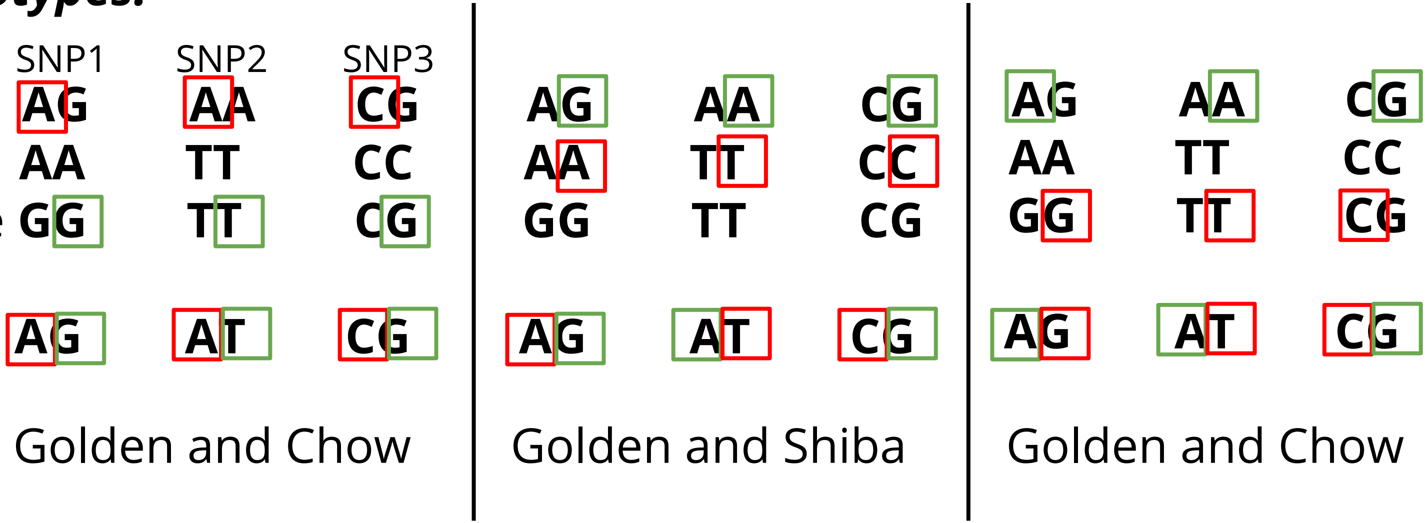
Now we have phased purebreds, so we can use this info too!



Comparing to purebreds

Let's say we only see the following phasing in goldens:
AAC / GAG

Compare to ***haplotypes:***

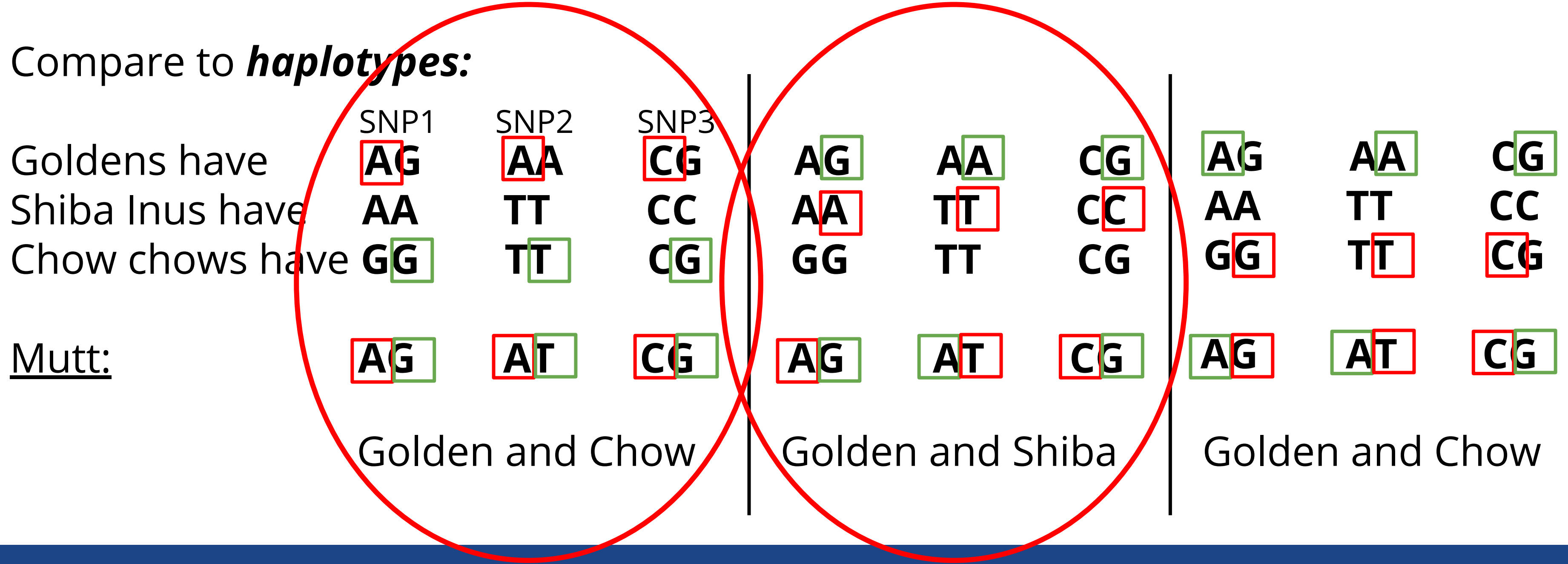


	SNP1	SNP2	SNP3						
Goldens have	AG	AA	CG	AG	AA	CG	AG	AA	CG
Shiba Inus have	AA	TT	CC	AA	TT	CC	AA	TT	CC
Chow chows have	GG	TT	CG	GG	TT	CG	GG	TT	CG
<u>Mutt:</u>	AG	AT	CG	AG	AT	CG	AG	AT	CG
	Golden and Chow			Golden and Shiba			Golden and Chow		

Comparing to purebreds

Let's say we only see the following phasing in goldens:
AAC / GAG

Compare to ***haplotypes:***



	SNP1	SNP2	SNP3						
Goldens have	AG	AA	CG	AG	AA	CG	AG	AA	CG
Shiba Inus have	AA	TT	CC	AA	TT	CC	AA	TT	CC
Chow chows have	GG	TT	CG	GG	TT	CG	GG	TT	CG
Mutt:	AG	AT	CG	AG	AT	CG	AG	AT	CG
	Golden and Chow			Golden and Shiba			Golden and Chow		

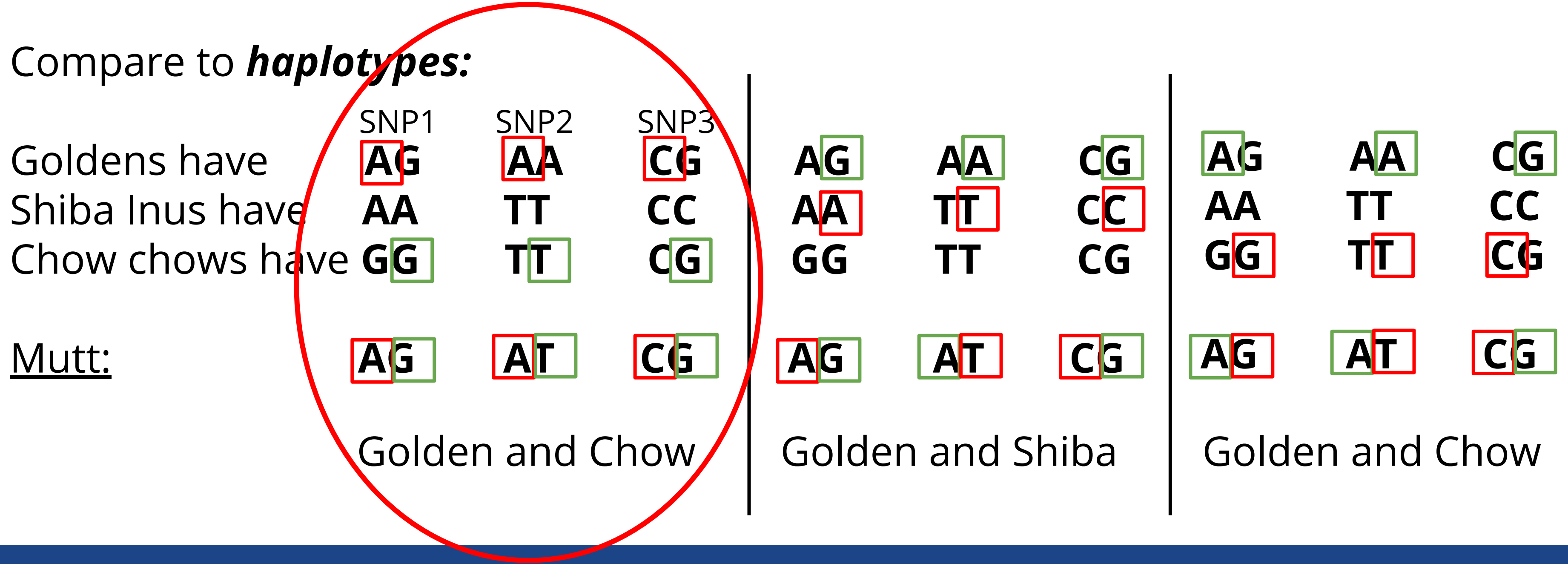
Comparing to purebreds

Let's say we only see the following phasing in goldens: AAC / GAG

✚

Let's say for SNP3, for goldens, we see *G* 10% of the time, shiba: 2%, and chows: 30%.

Compare to ***haplotypes:***

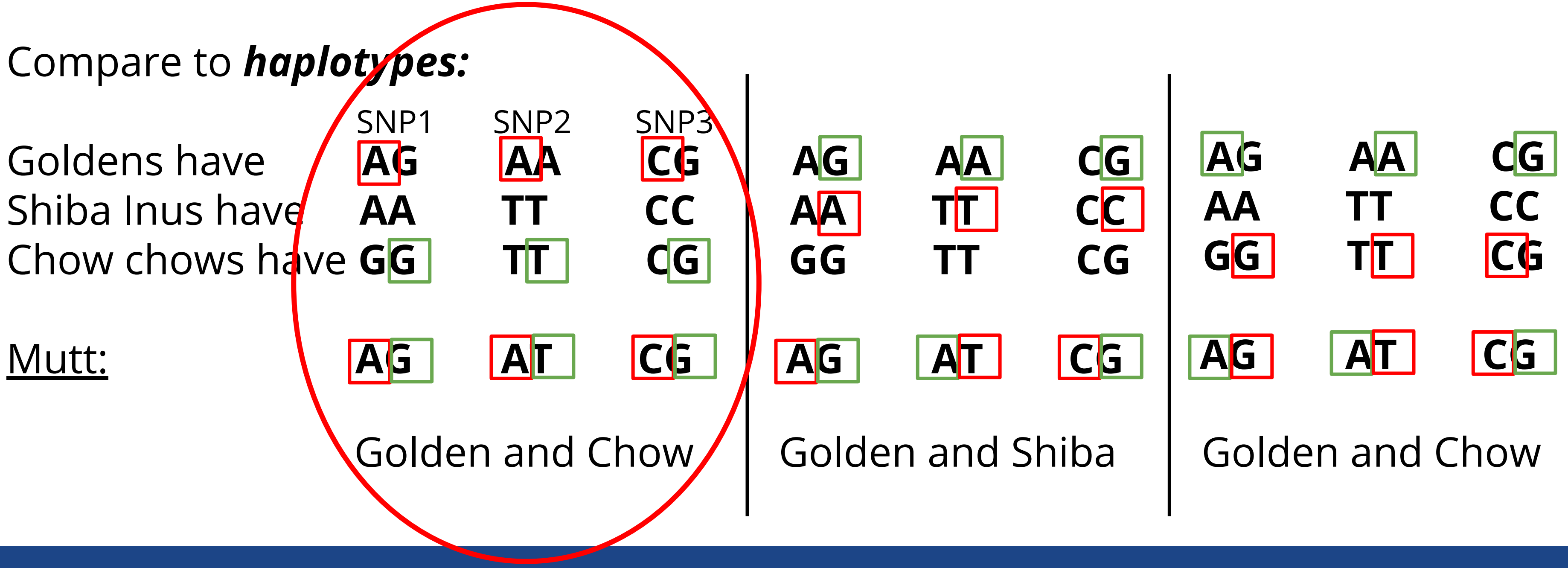


	SNP1	SNP2	SNP3						
Goldens have	AG	AA	CG	AG	AA	CG	AG	AA	CG
Shiba Inus have	AA	TT	CC	AA	TT	CC	AA	TT	CC
Chow chows have	GG	TT	CG	GG	TT	CG	GG	TT	CG
Mutt:	AG	AT	CG	AG	AT	CG	AG	AT	CG
	Golden and Chow			Golden and Shiba			Golden and Chow		

Comparing to purebreds

Phasing ✚ Allele frequencies

Compare to ***haplotypes:***



	SNP1	SNP2	SNP3
Goldens have	AG	AA	CG
Shiba Inus have	AA	TT	CC
Chow chows have	GG	TT	CG
Mutt:	AG	AT	CG

Golden and Chow

AG	AA	CG
AA	TT	CC
GG	TT	CG
AG	AT	CG

Golden and Shiba

AG	AA	CG
AA	TT	CC
GG	TT	CG
AG	AT	CG

Golden and Chow

Hidden Markov Models (HMMs) with SupportMix

We’ll use a program called SupportMix, which takes in:

1. Phased SNPs from purebred dogs
2. Phased SNPs from our mutts
3. A “genetic linkage map” of the centiMorgan distances between SNPs

Output: For each mutt, gives the best guess breed for each SNP, and the probability the given guess is correct

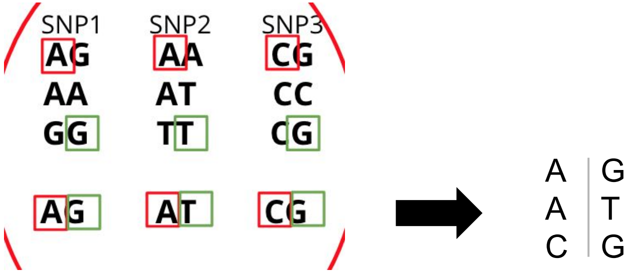
Method: Hidden Markov Model

Hidden Markov Models (HMMs) with SupportMix

We'll use a program called SupportMix, which takes in:

1. Phased SNPs from purebred dogs
2. Phased SNPs from our mutts

When we phased our purebred dogs, we also got out mutt phasings. So, we can phase mutts and purebreds together to get phased mutts!



Note: We use different sets of purebred dogs to phase the mutts than we use with SupportMix (6 from each breed to phase the mutts, and 6 *others* from each breed that we phase with each other and/or with *other* mutts) to compare to.

Hidden Markov Models (HMMs) with SupportMix

We’ll use a program called SupportMix, which takes in:

1. Phased SNPs from purebred dogs
2. Phased SNPs from our mutts
3. A “genetic linkage map” of the centiMorgan distances between SNPs

Output: For each mutt, gives the best guess breed for each SNP, and the probability the given guess is correct

Method: Hidden Markov Model

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T

Oversimplified again, let's consider these the most common haplotype for each breed

Fido	A	C	G	T	T	C	G	A	C	T

Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T
Fido	A	C	G	T	T	C	G	A	C	T

Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T
Fido	A	C	G	T	T	C	G	A	C	T

Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T

Fido	A	C	G	T	T	C	G	A	C	T

Slides adapted from the Jackson Laboratory

HMMs

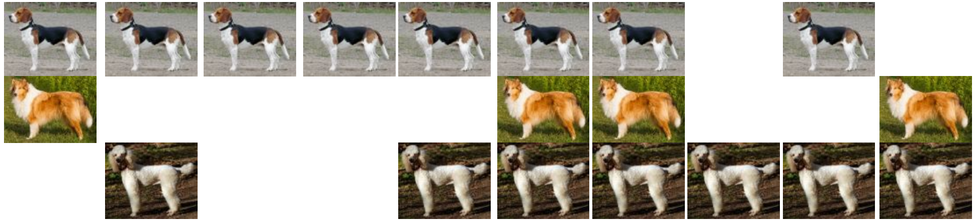
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T
Fido	A	C	G	T	T	C	G	A	C	T

Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T

Fido	A	C	G	T	T	C	G	A	C	T

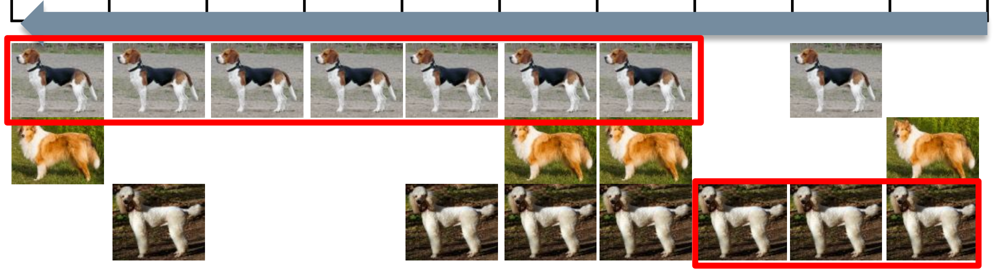


Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T

Fido	A	C	G	T	T	C	G	A	C	T

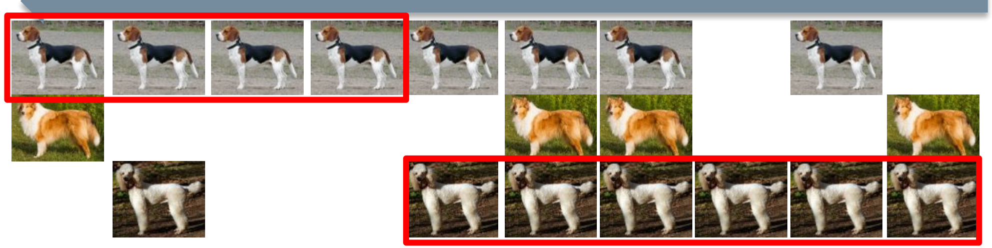


Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T

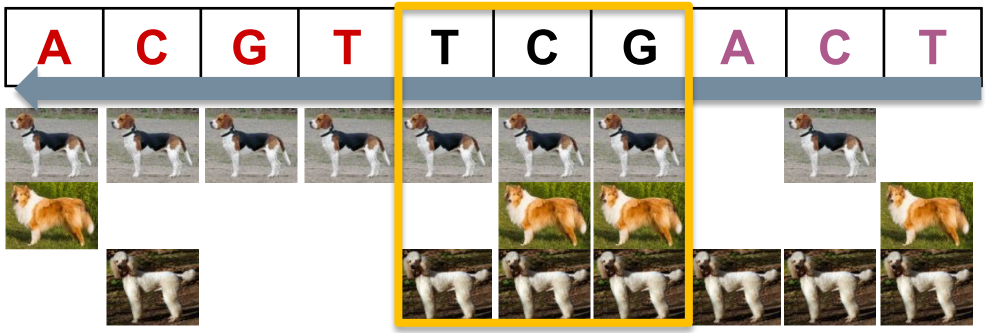
Fido	A	C	G	T	T	C	G	A	C	T



Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T



Fido	A	C	G	T	T	C	G	A	C	T

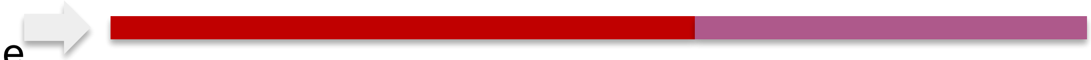
Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
Collie	A	G	T	G	G	C	G	T	A	T
Poodle	T	C	T	G	T	C	G	A	C	T

Fido	A	C	G	T	T	C	G	A	C	T

One chromosome



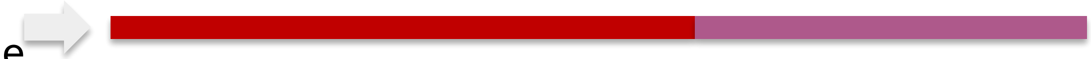
Represent ancestry by painting with the breed color

Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
	.82	.85	.74	.95	.90	.89	.81	.75	.91	.94
Poodle	T	C	T	G	T	C	G	A	C	T
	.58	.91	.93	.79	.84	.85	.92	.78	.86	.99
Fido	A	C	G	T	T	C	G	A	C	T

One chromosome



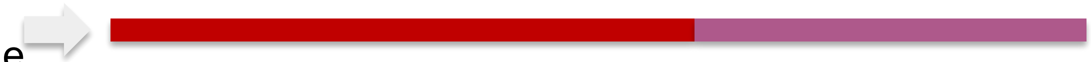
Represent ancestry by painting with the breed color

Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
	.82	.85	.74	.95	.90	.89	.81	.75	.91	.94
Poodle	T	C	T	G	T	C	G	A	C	T
	.58	.91	.93	.79	.84	.85	.92	.78	.86	.99
Fido	A	C	G	T	T	C	G	A	C	T

One chromosome



Represent ancestry by painting with the breed color

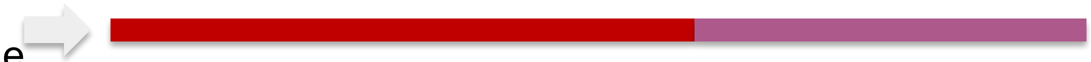
Slides adapted from the Jackson Laboratory

HMMs

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Beagle	A	C	G	T	T	C	G	T	C	A
	.82	.85	.74	.95	.90	.89	.81	.75	.91	.94
Poodle	T	C	T	G	T	C	G	A	C	T
	.58	.91	.93	.79	.84	.85	.92	.78	.86	.99
Fido	A	C	G	T	T	C	G	A	C	T

It seems unlikely we'd transition for one SNP and then transition back. HMMs account for this!

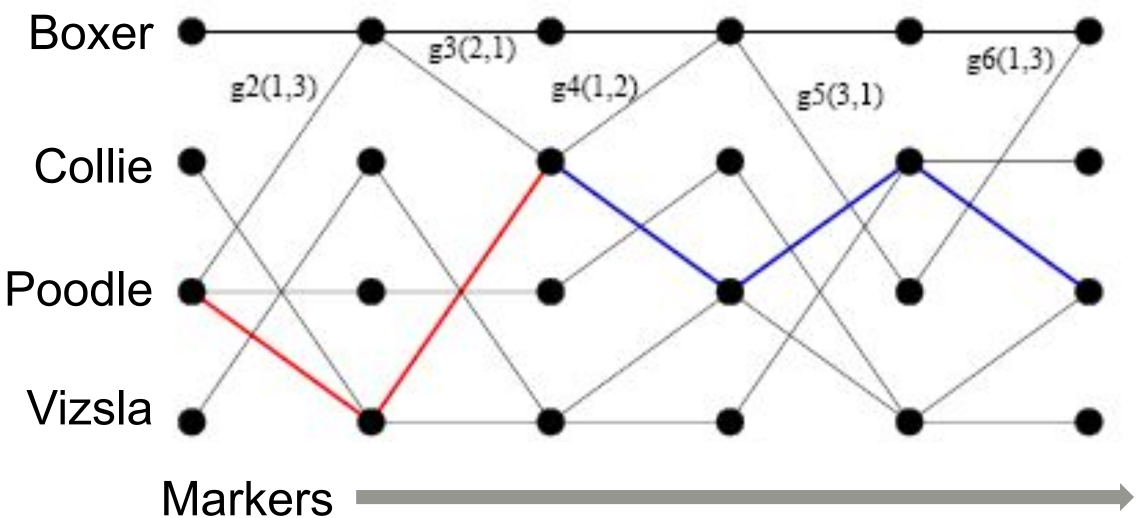
One chromosome



Represent ancestry by painting with the breed color

HMM: Viterbi Decoding

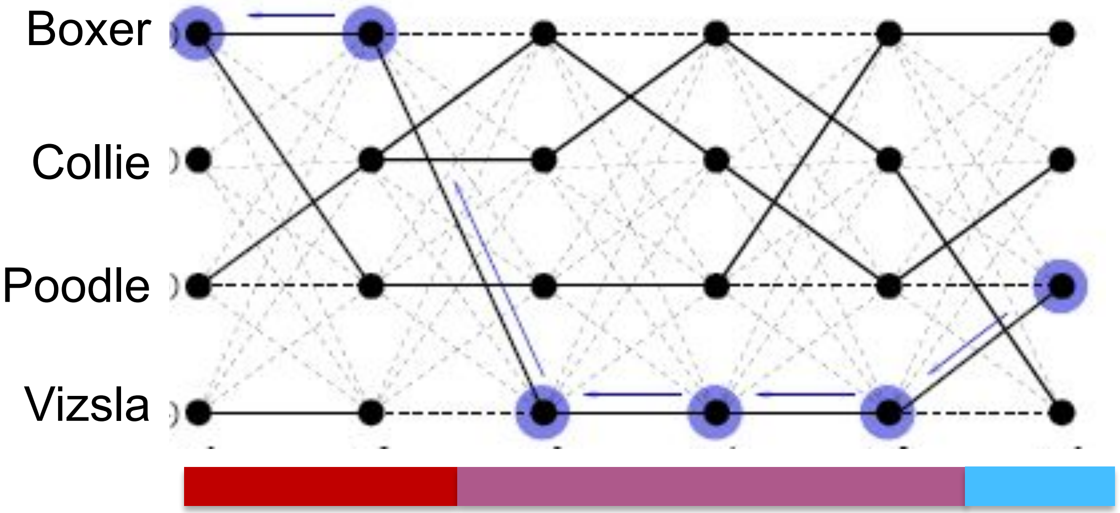
- Goal: Determine the most probable path through the data.
  - Translation: Determine the most probable breed along each haplotype. Maximize Pr(breed|data)



https://onlinecourses.science.psu.edu/stat857/node/203

HMM: Viterbi Decoding

- To determine the most probable path, we take into account probabilities of seeing a SNP given a breed, but we also consider the probability of transitioning breed.

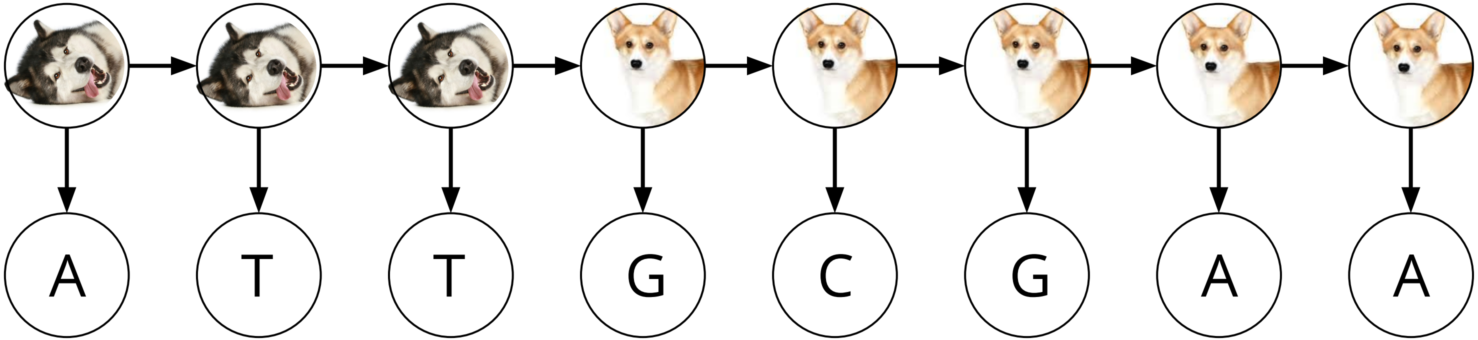


https://onlinecourses.science.psu.edu/stat857/node/203

Hidden Markov Models (HMMs)

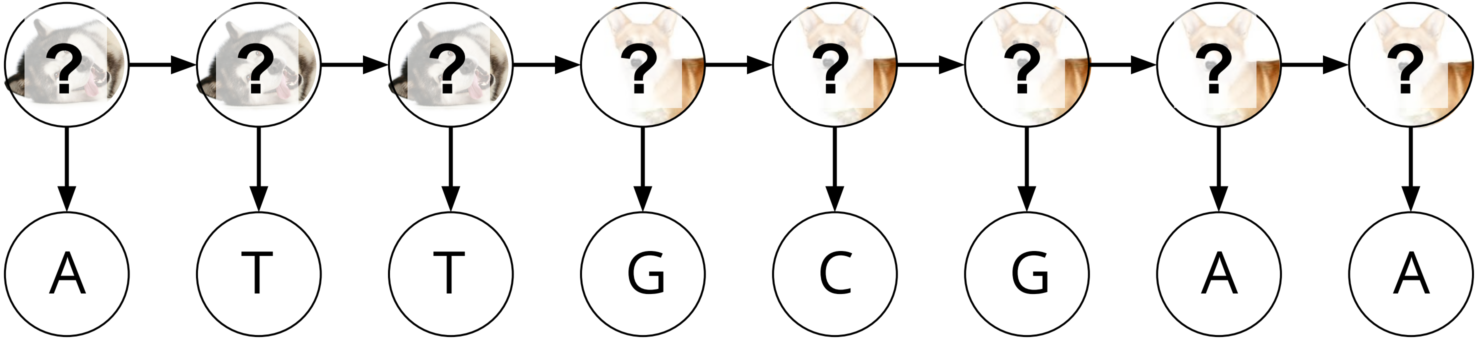
HMMs deal with data, which we call ***emissions***, and ***hidden states***, which is what we're trying to determine.

Emissions: SNPs
Hidden States: Breeds

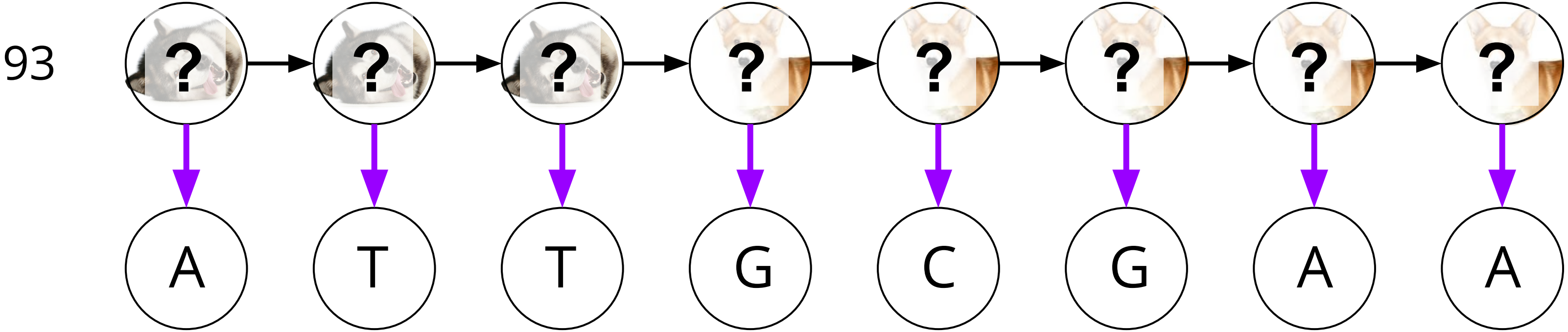


Hidden Markov Models (HMMs)

93

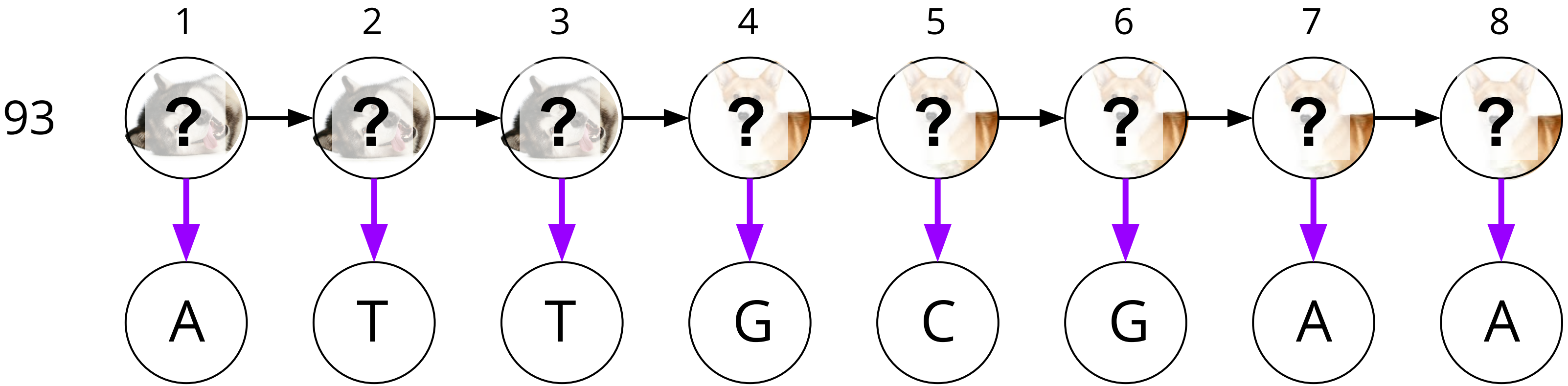


Hidden Markov Models (HMMs)



How likely is it I see “A” if the hidden state is a … husky? corgi? chow? Etc.

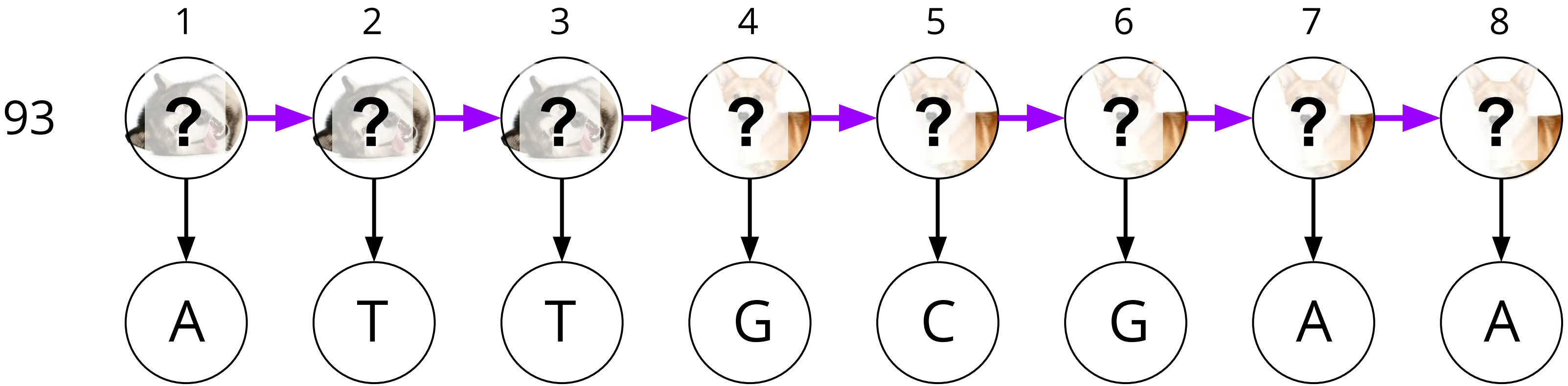
Hidden Markov Models (HMMs)



How likely is it I see “A” if the hidden state is a … husky? corgi? chow? Etc.

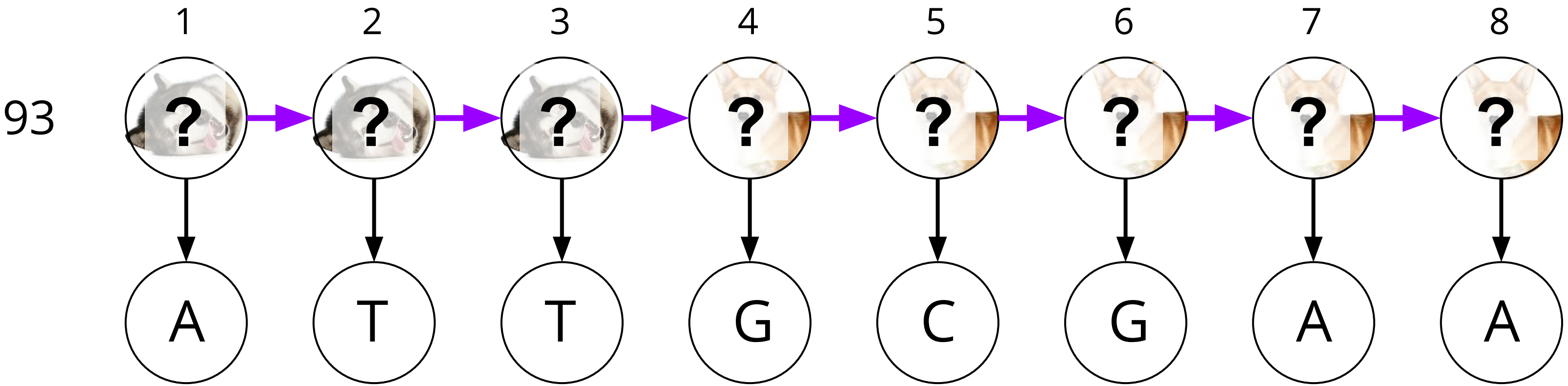
Emission probabilities: $P(A_1|\text{husky})$ $P(A_1|\text{corgi})$ … $P(T_2|\text{husky})$ $P(T_2|\text{corgi})$ … ••• $P(\textit{allele}_n|\text{husky})$ $P(\textit{allele}_n|\text{corgi})$ …

Hidden Markov Models (HMMs)



If the current breed is husky, how likely is it the breed at the next SNP site is … husky? corgi? chow? etc

Hidden Markov Models (HMMs)



If the current breed is husky, how likely is it the breed at the next SNP site is … husky? corgi? chow? Etc

Transition probabilities: Because we know we have linked regions inherited together, intuitively $P(\text{husky}_i|\text{husky}_{i-1}) > P(\text{corgi}_i|\text{husky}_{i-1})$

Hidden Markov Models (HMMs)

How do we get transition probabilities?

Based on what we know, we can intuit that:

1. Probability breed_A --> breed_B is the same regardless of breed (A != B)

2. It seems like it's a higher probability that breed_A --> breed_A.

So we don't need transition probabilities for all breeds --> all breeds!

Hidden Markov Models (HMMs)

How do we get transition probabilities?

We know two SNPs are more likely to be in the same “chunk” if they are nearby one another. We have centiMorgan distances between all our SNPs.

Hidden Markov Models (HMMs)

How do we get transition probabilities?

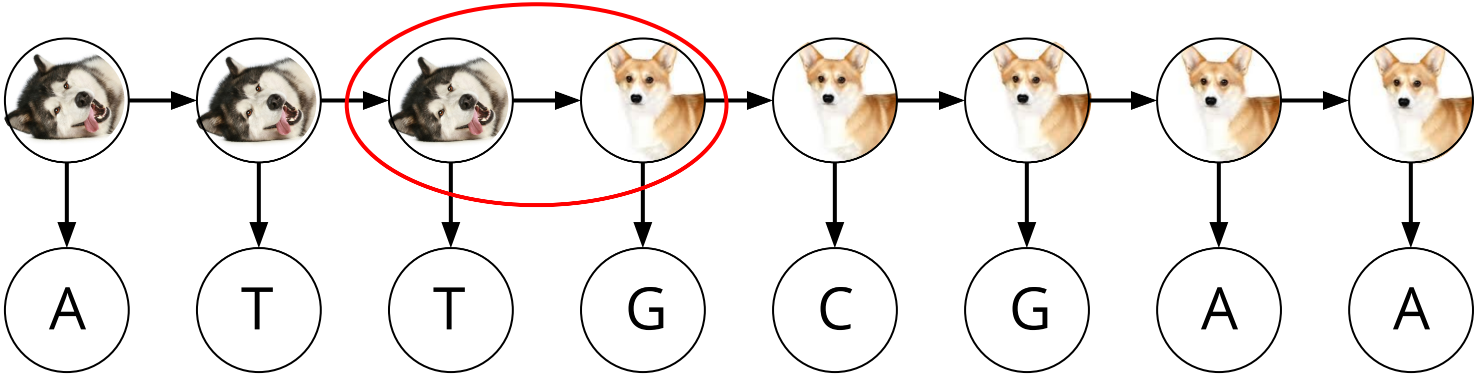
1. Probability breed_A --> breed_B is the same regardless of breed (A != B)
2. It seems like it’s a higher probability that breed_A --> breed_A.
3. We know two SNPs are more likely to be in the same “chunk” if they are nearby one another. We have centiMorgan distances between all our SNPs.

We can calculate probabilities from this!

Hidden Markov Models (HMMs)

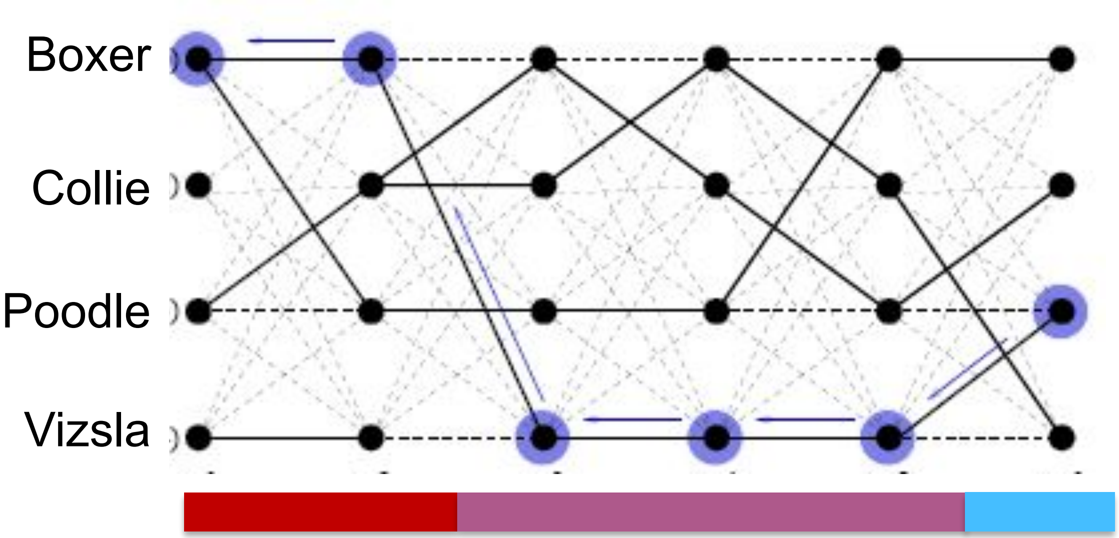
How do we get transition probabilities? Another way would to *train* the HMM on a labeled mutt.

If we have a mutt and we know what it's ancestral segments are, we can examine that data to determine how likely breed transitions are to occur at different cM distances.



HMM: Viterbi Decoding

1. Examine all possible *hidden state* paths (breed assignments)
2. Use *emission* and *transition probabilities* to choose the path that maximizes the probability of the entire sequence (Viterbi)



https://onlinecourses.science.psu.edu/stat857/node/203

Final HMM Notes

The way we calculate using the probabilities assumes that the state (breed) at a given SNP is only dependent on the state (breed) of the SNP before it.

HMMs are used for a lot of other biology applications, including gene finding in bacteria.

To learn about them in more detail (and code your own!), take Computational Genomics (EN 601.439/639) with Ben Langmead in Fall 2019!

Project Logistics

Today: More data exploration (continue part 1 and/or part 2)
Wed/Fri: Finding Clarence, Reilly, and Finch's breeds
Next week: Concept exploration (no coding, but you'll need laptops)

Part 1 due Wednesday, Jan 16.
Part 2 due Friday, Jan 18.

Please turn in your code and question answers to rsherman@jhu.edu and include EN.601.147 in the subject line.

Make sure both your names are on your writeups!