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

Lecture 2

Jan 9th, 2019

Border collies are known for their strong work ethic, even—it seems—when it comes to carrying tennis balls. MARK RAYCROFT/MINDEN PICTURES

Dog breeds really do have distinct personalities—and they're rooted in DNA

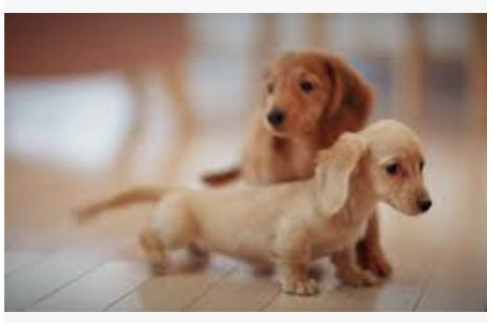
By **Elizabeth Pennisi** | Jan. 7, 2019 , 1:00 PM

American Kennel Club descriptions of dog breeds can read like online dating profiles: The border collie is a workaholic; the German shepherd will put its life on the line for loved ones. Now, in the most comprehensive study of its kind to date, scientists have shown that such distinct breed traits are actually rooted in a dog's genes. The findings may shed light on human behaviors as well.



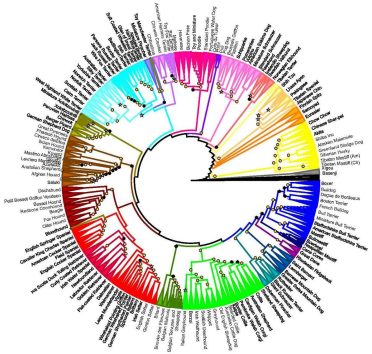






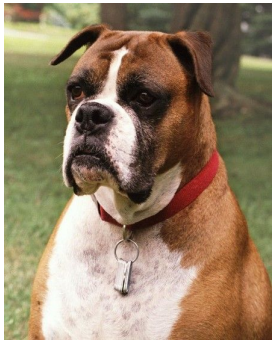
Key concepts from Monday



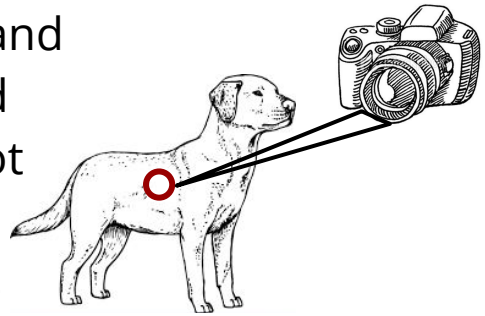
Breed genetics != looks

I'm 17% chow chow!

We have a dog genome that contains the full 2.8 billion base sequence for Tasha …



… but we don't usually put together whole genomes because it's hard and expensive. Instead we take a snapshot of single bases in the genome, SNPs



Terms from Monday

Genotype -- the two nucleotides at any position in the genome (one from each chromosome)

Phenotype -- a trait. Some examples are: breed, fur color, eye color, likelihood of a tumor being benign. Phenotypes are influenced by genotypes.

SNP -- single nucleotide polymorphism. It is a single base that differs between individuals.

Allele -- the possible bases at a given SNP site. We'll only deal with sites with 2 alleles.

Open questions from Monday

How do we figure out where the interesting SNPs are?

How do we actually figure out what the genotype is at a SNP location?

If we wanted to get a whole genome (like for Tasha) how would we do that?

How do the SNPs help us figure out breeds?

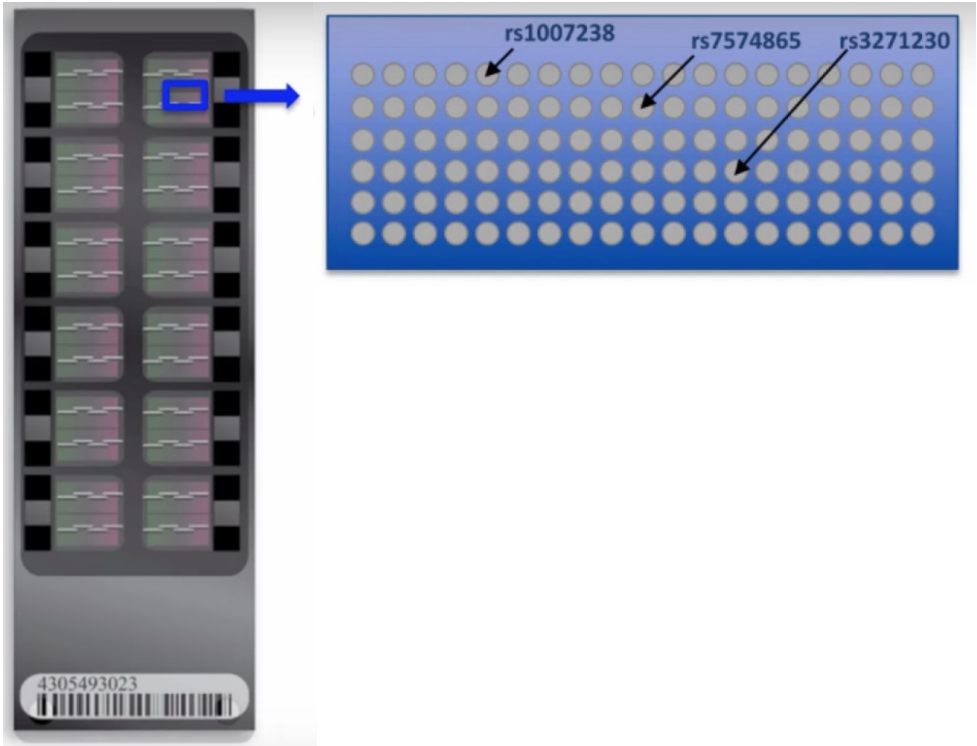
How do we find our mutt's breed makeups?

Reilly

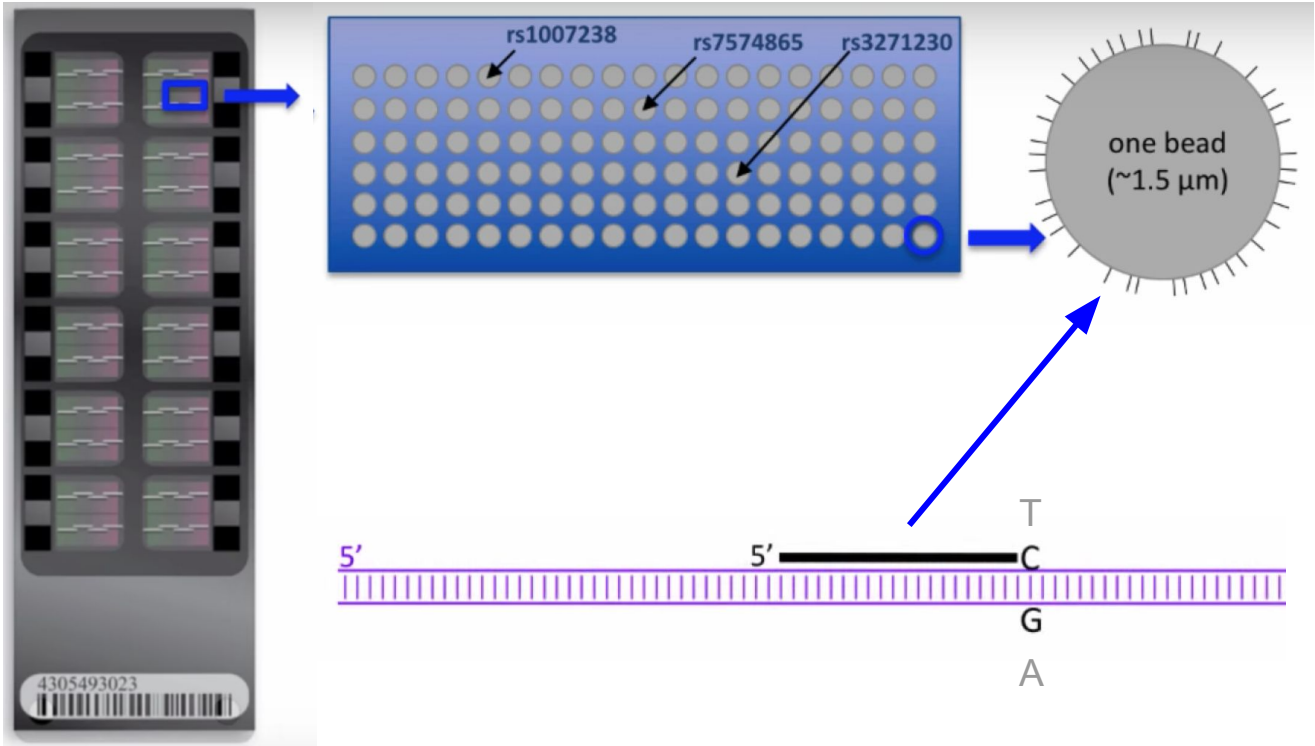
Clarence

Finch

Illumina Bead SNP Array

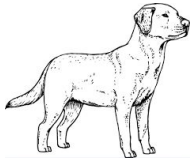


Illumina Bead SNP Array



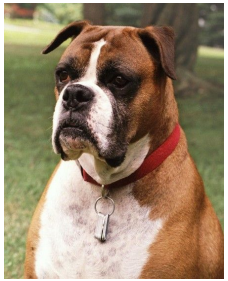
Illumina Bead SNP Array

SNP genotype: ??



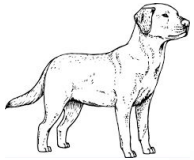
↓

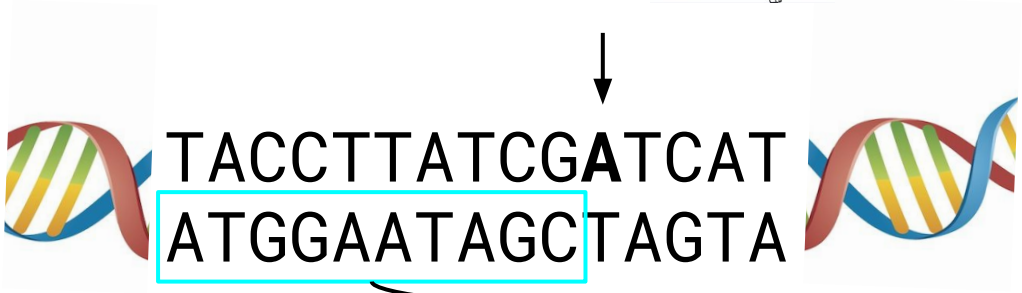
TACCTTATCG**A**TCAT

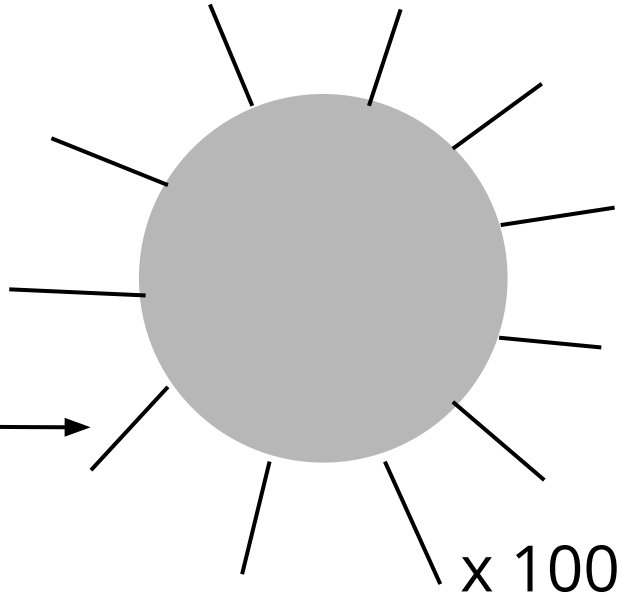


Illumina Bead SNP Array

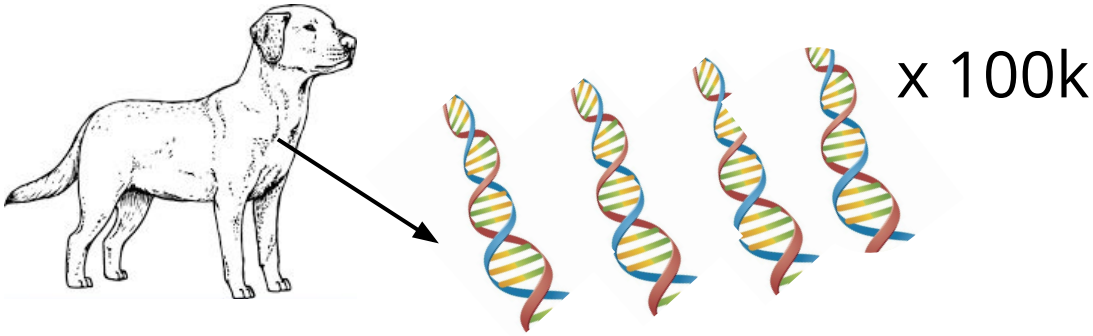
SNP genotype: ??



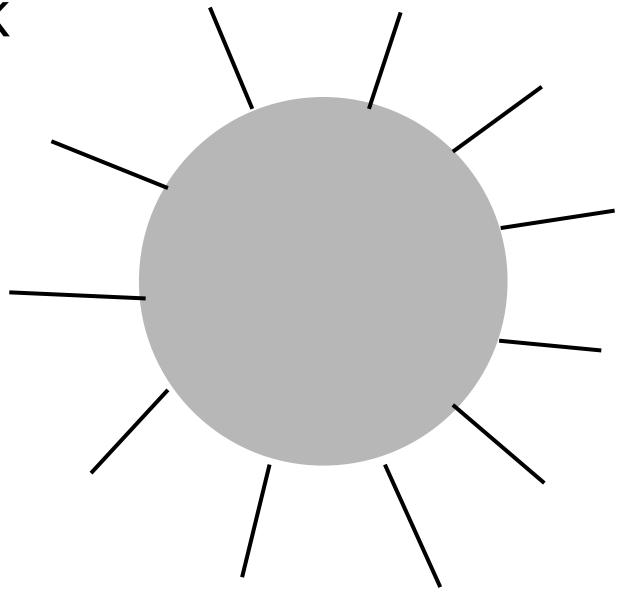




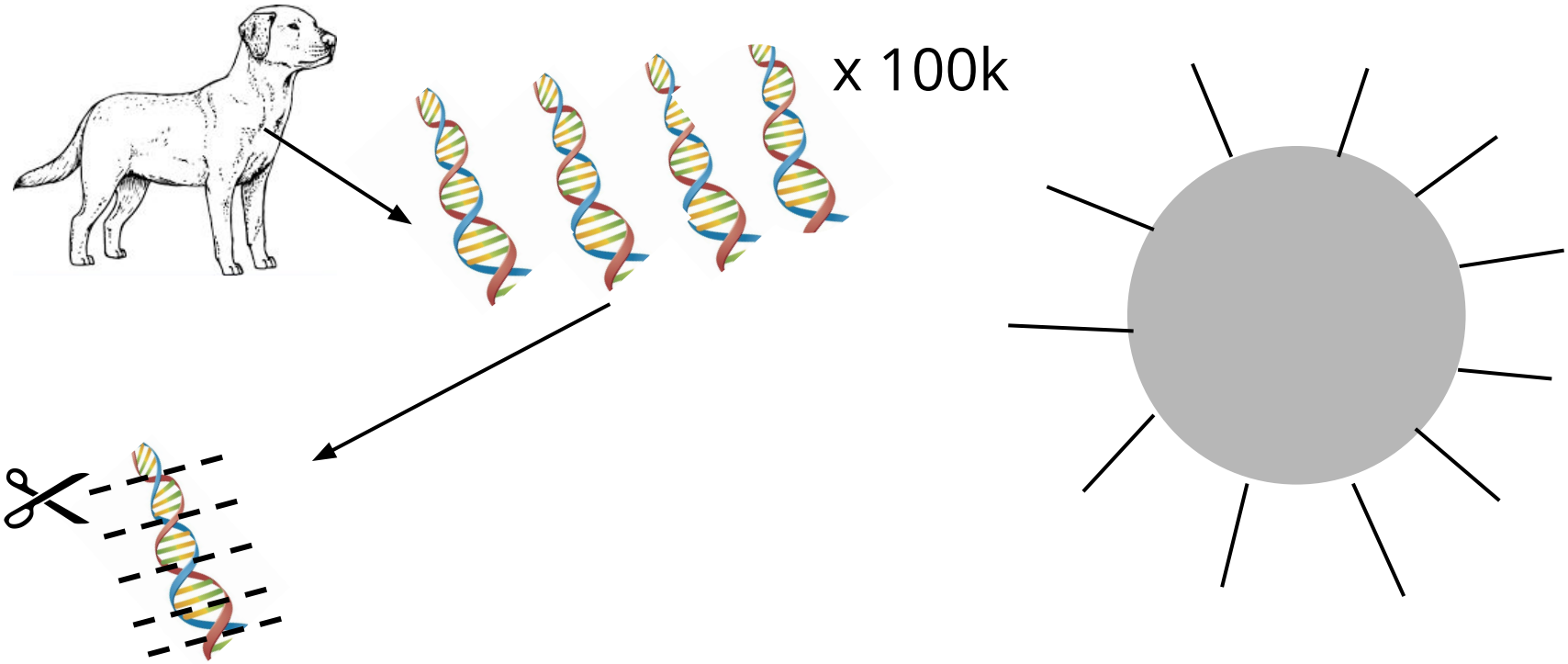
Illumina Bead SNP Array



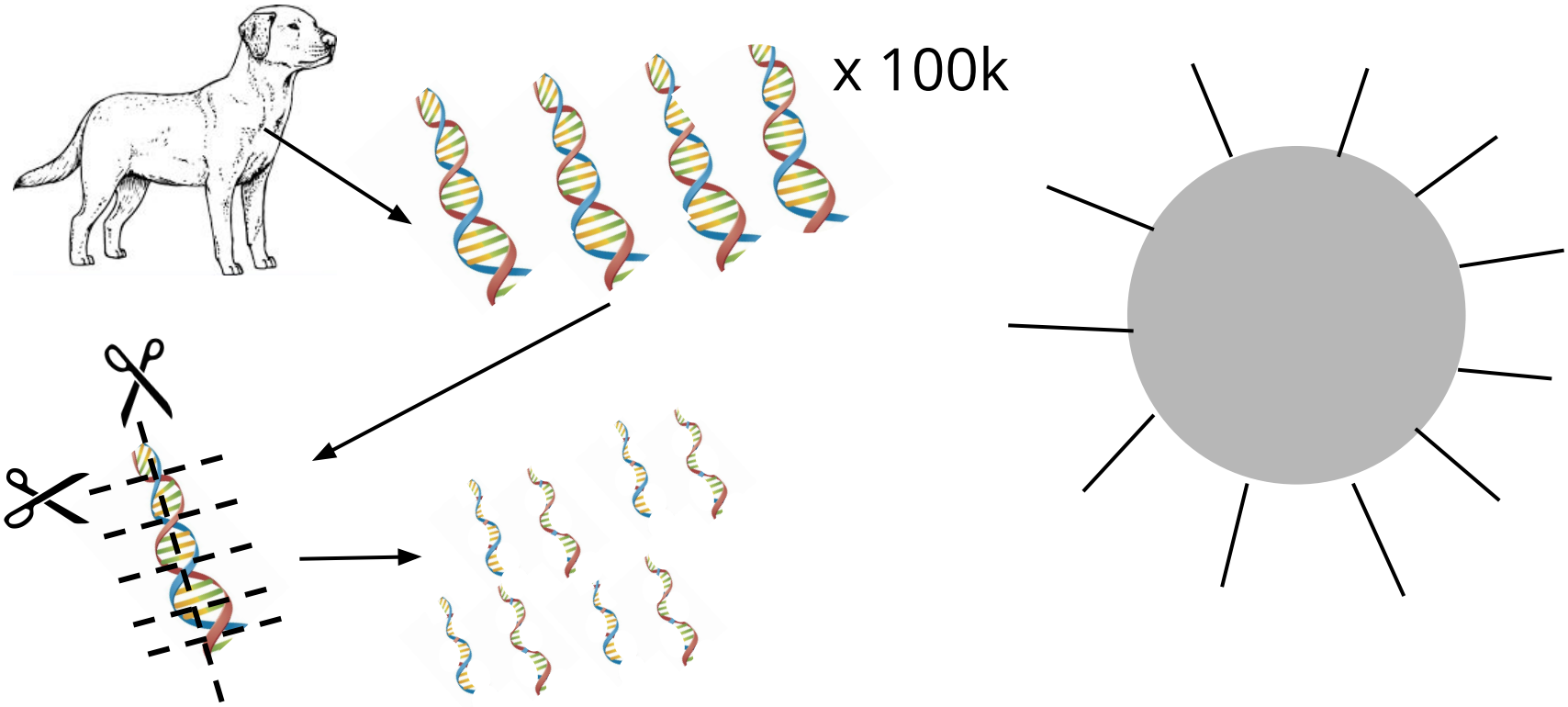
x 100k



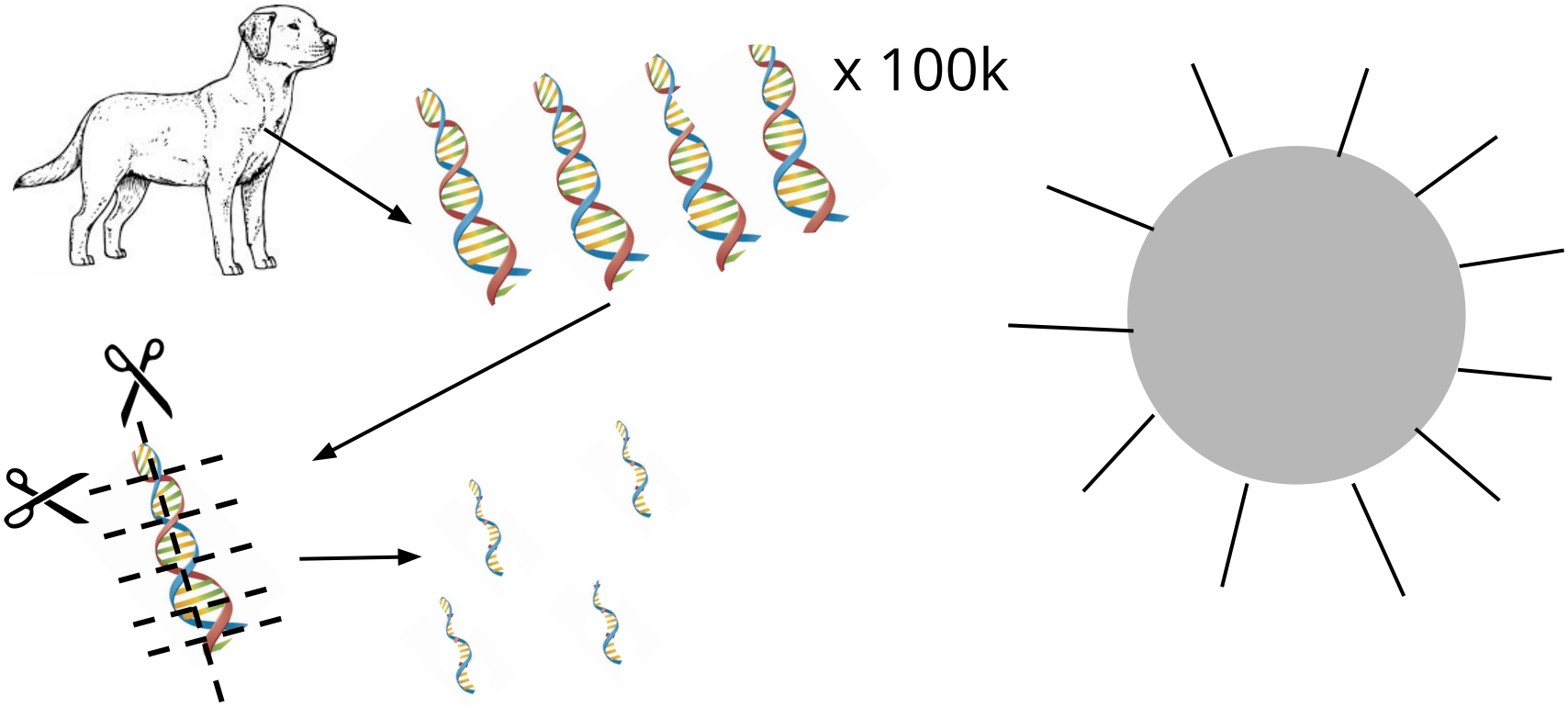
Illumina Bead SNP Array



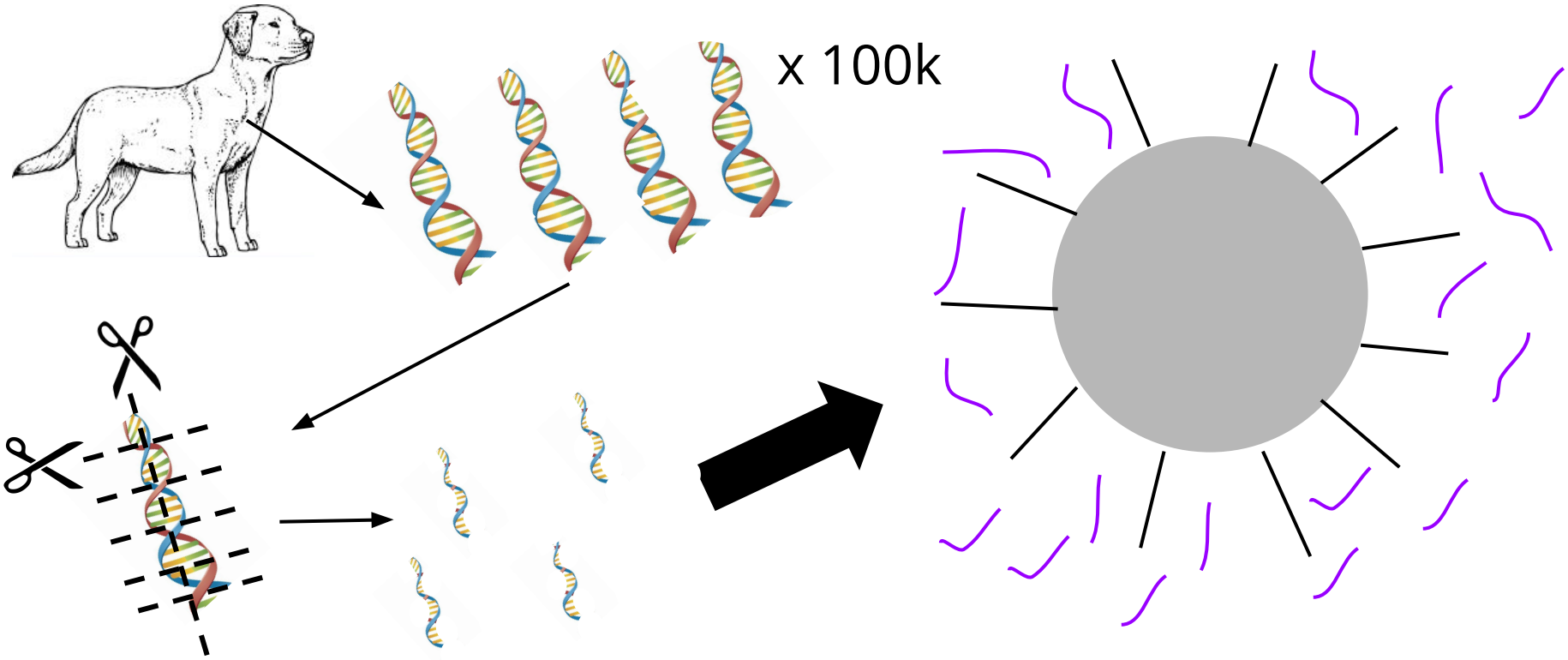
Illumina Bead SNP Array



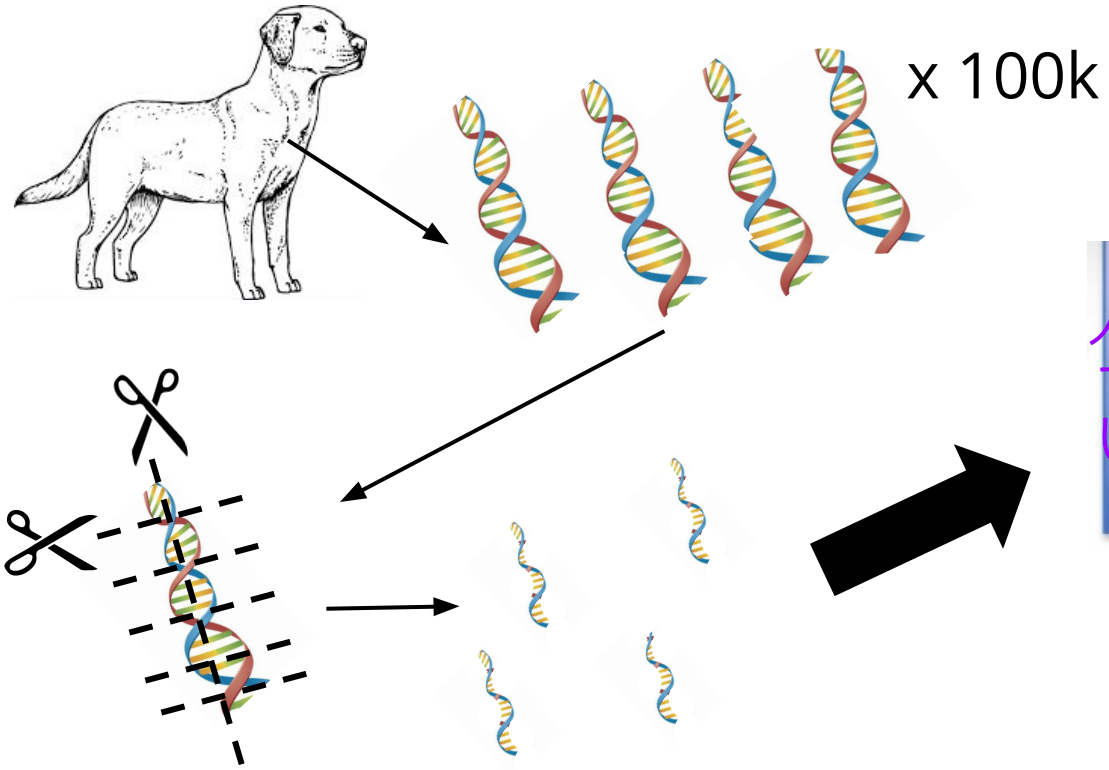
Illumina Bead SNP Array

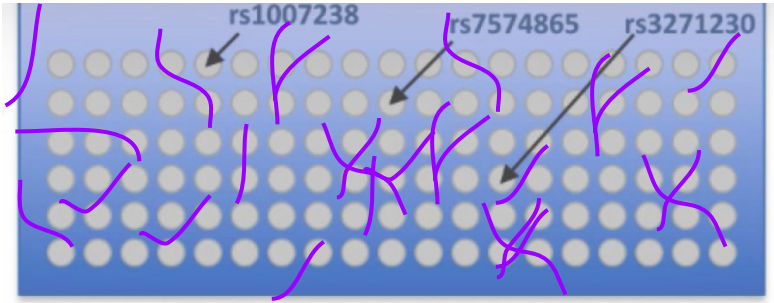


Illumina Bead SNP Array



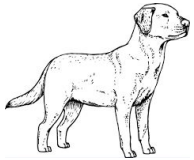
Illumina Bead SNP Array





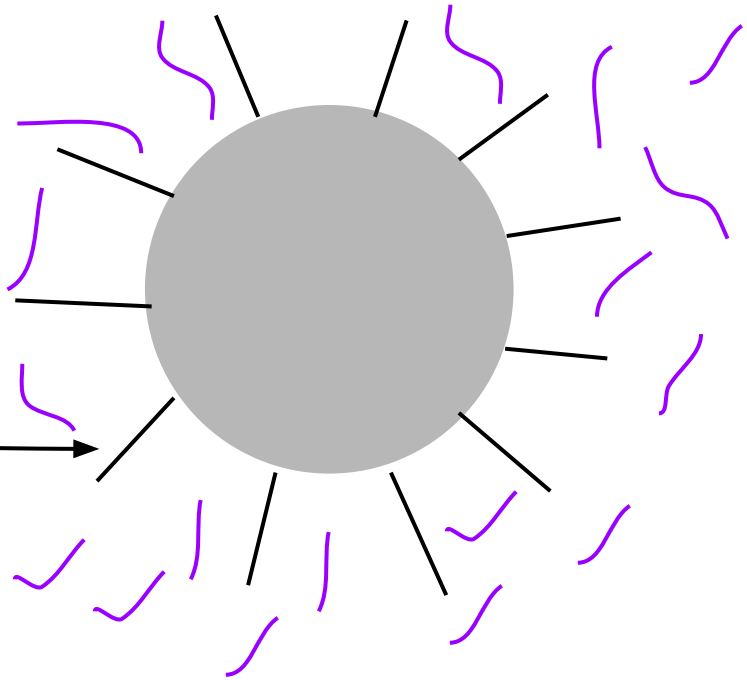
Illumina Bead SNP Array

SNP genotype: ??



TACCTTATCG**A**TCAT

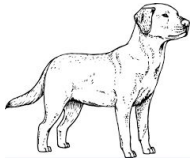
ATGGAATAGC



Illumina Bead SNP Array

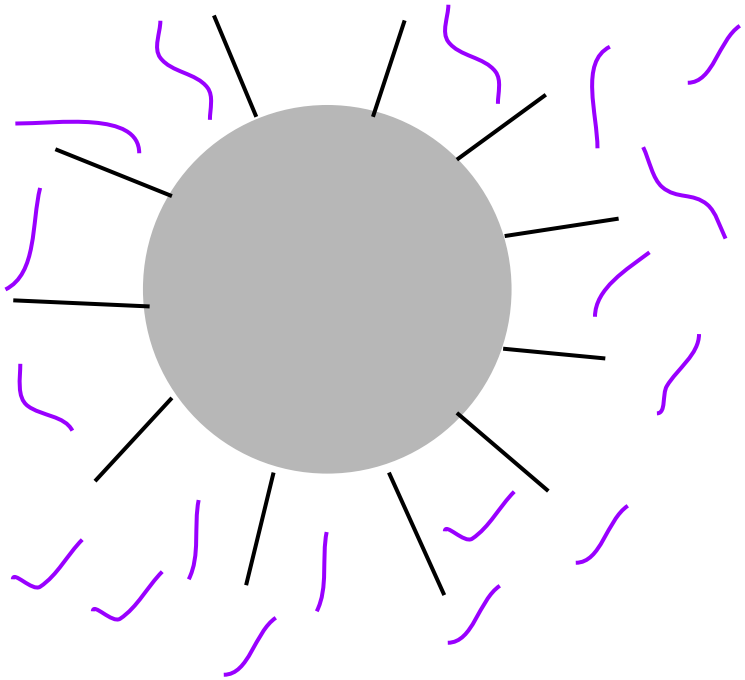
ATGGAATAGC

SNP genotype: AG



TACCTTATCGATCAT
ATGGAATAGCTAGTA

TACCTTATCGGTCAT
ATGGAATAGCCAGTA

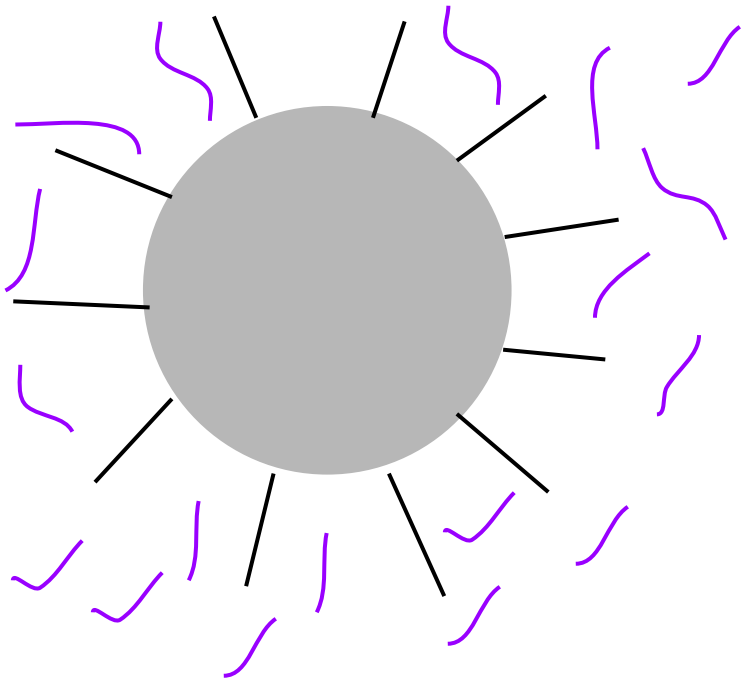


Illumina Bead SNP Array

ATGGAATAGC

TACCTTATCGATCAT

TACCTTATCGGTCAT

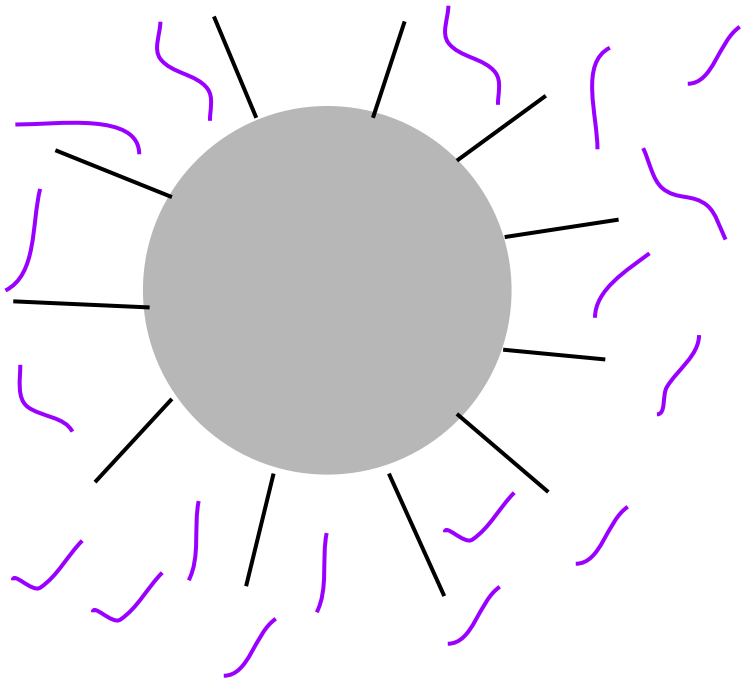


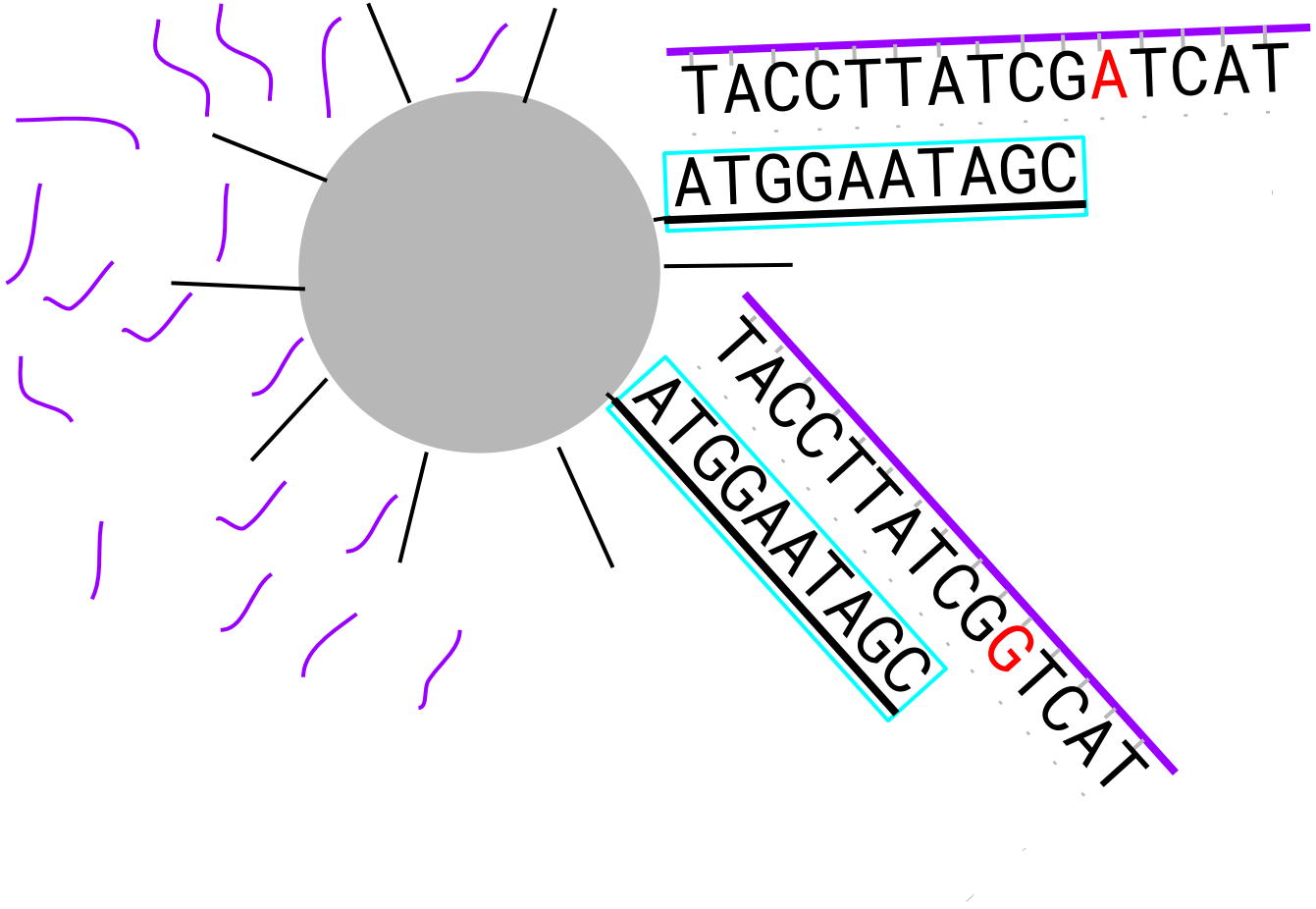
Illumina Bead SNP Array

ATGGAATAGC

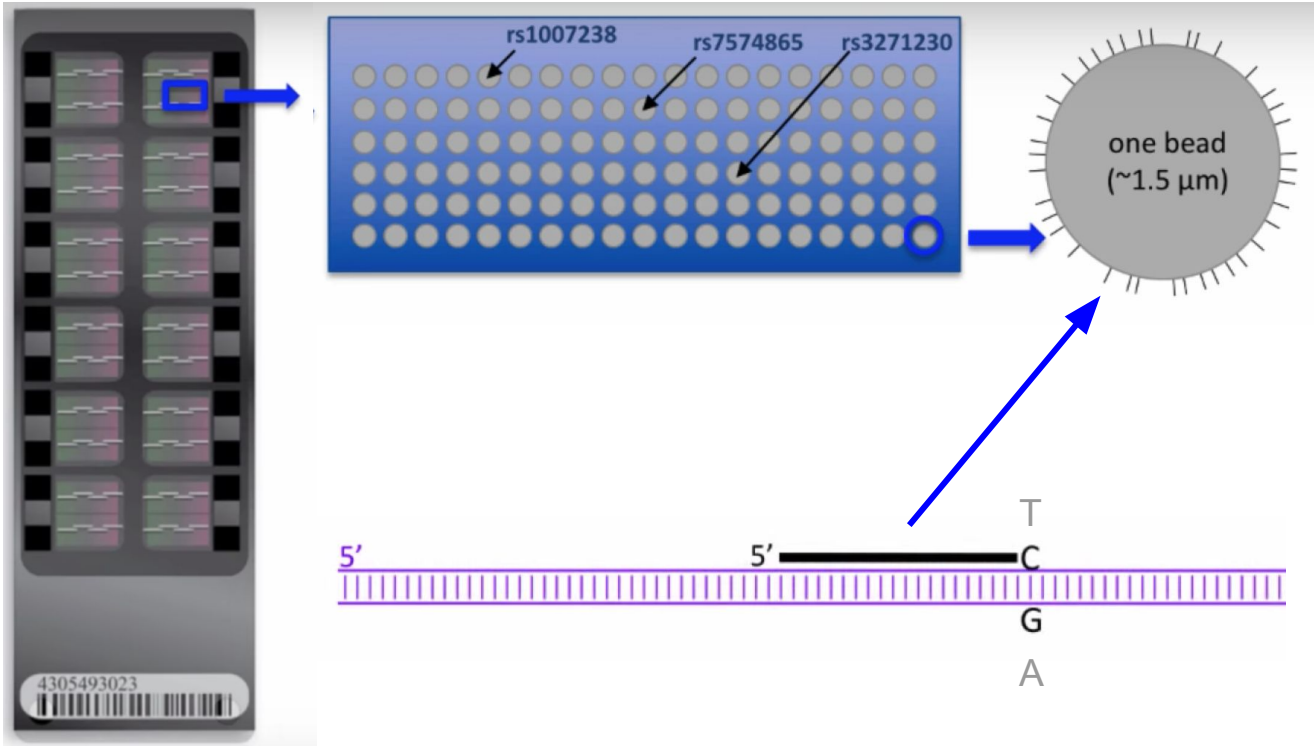
TACCTTATCGATCAT
ATGGAATAGC

TACCTTATCGGTCAT
ATGGAATAGC

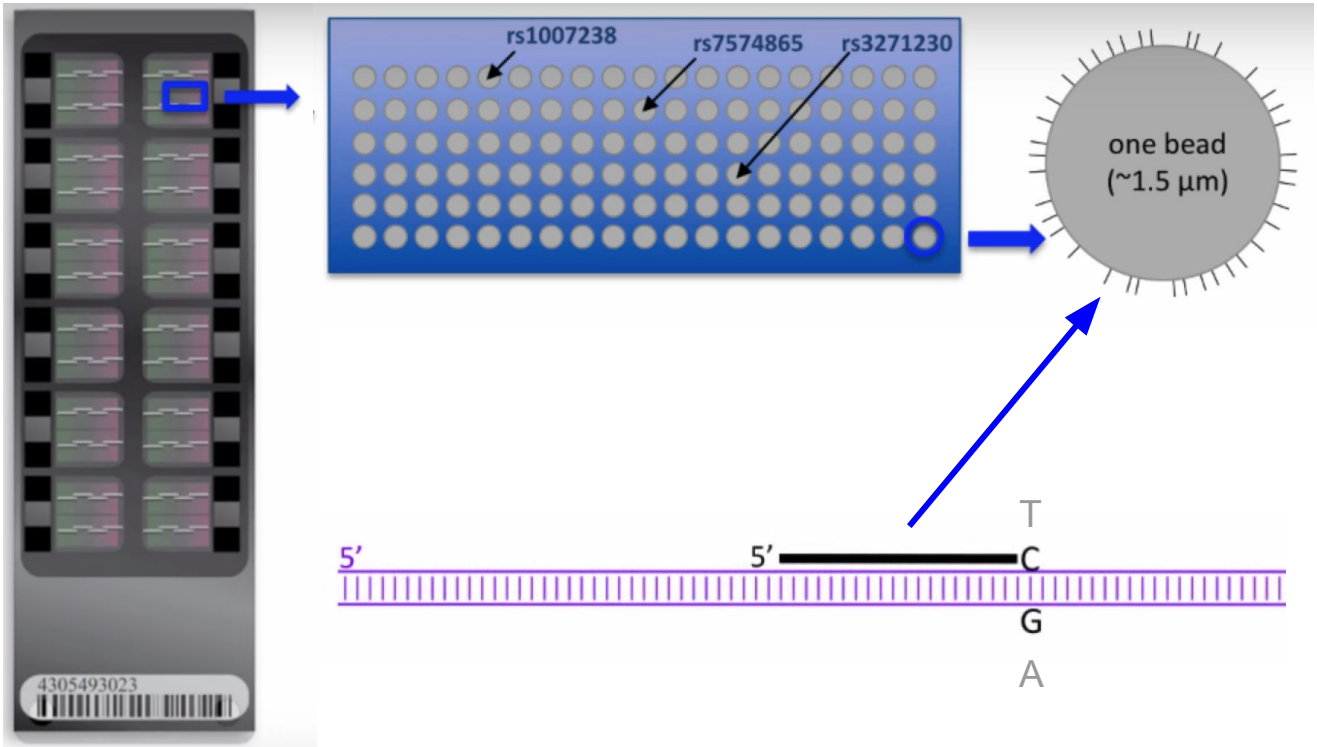


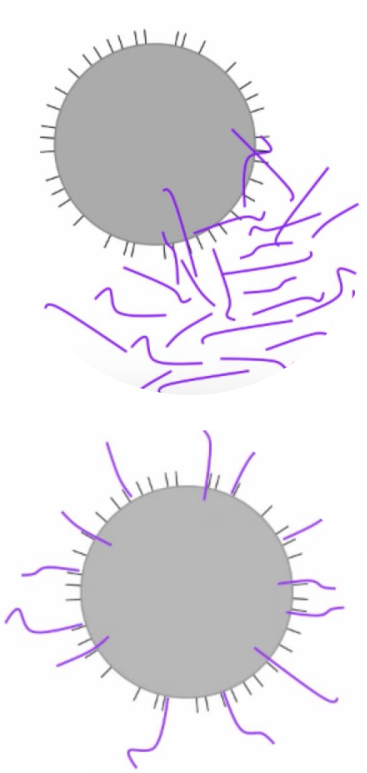

TACCTTATCGATCAT
ATGGAATAGC
TACCTTATCGGTCAT
ATGGAATAGC

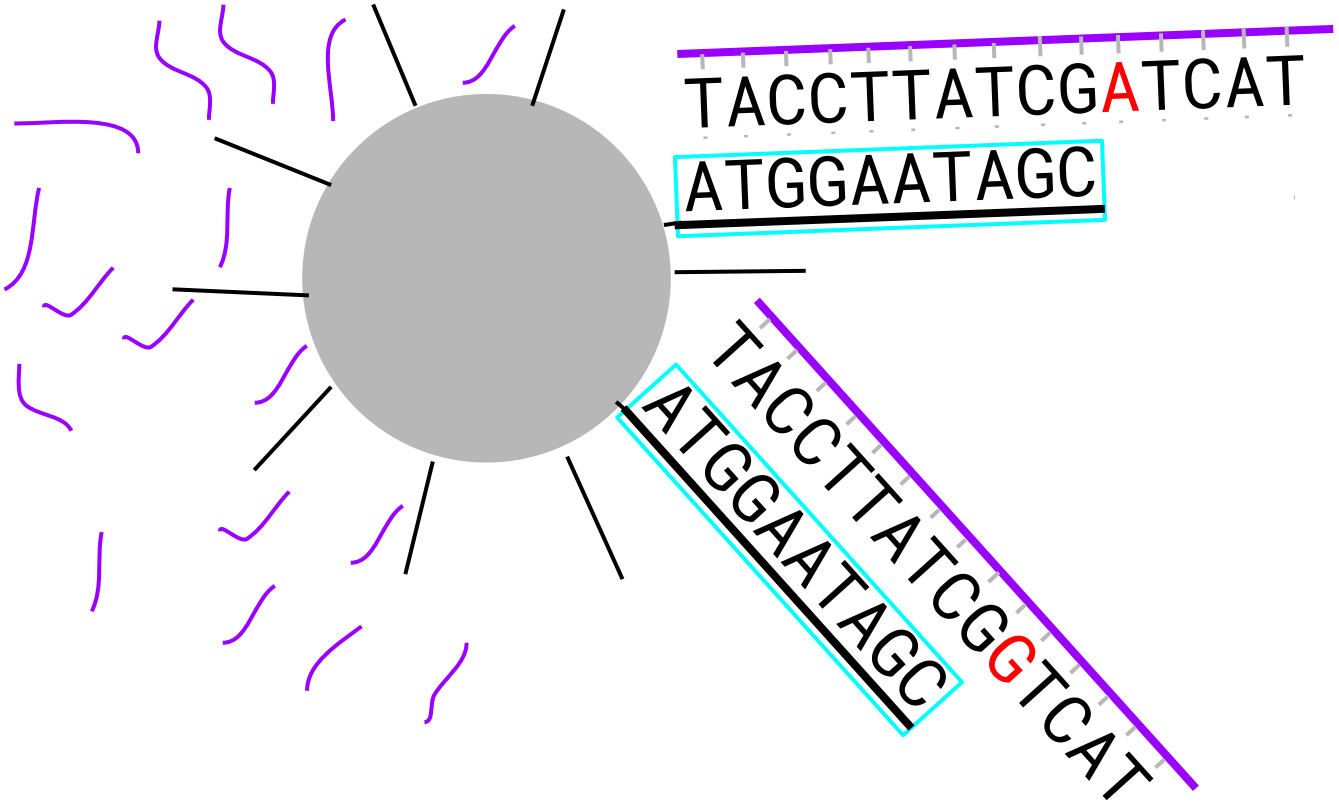
Illumina Bead SNP Array

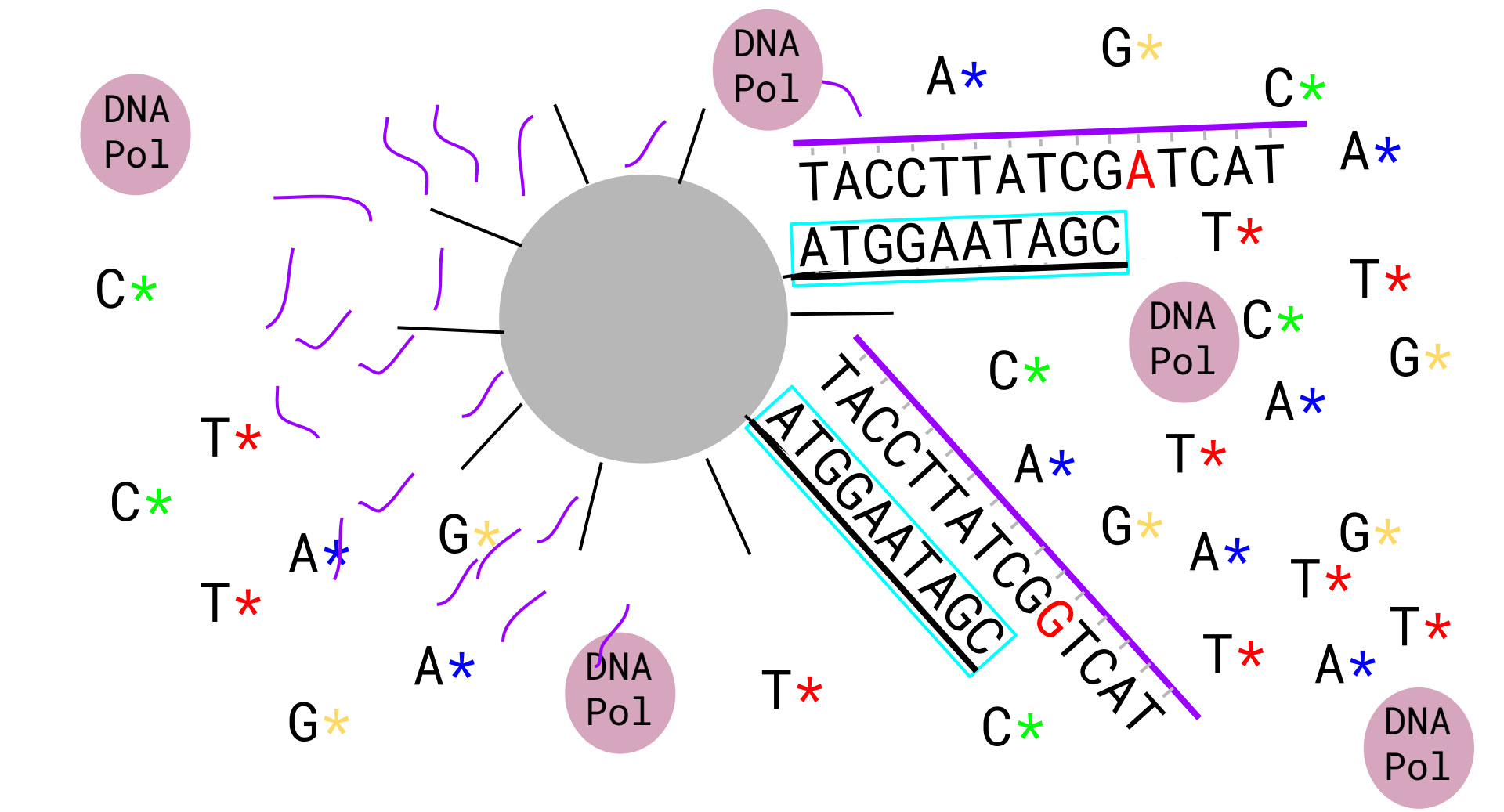


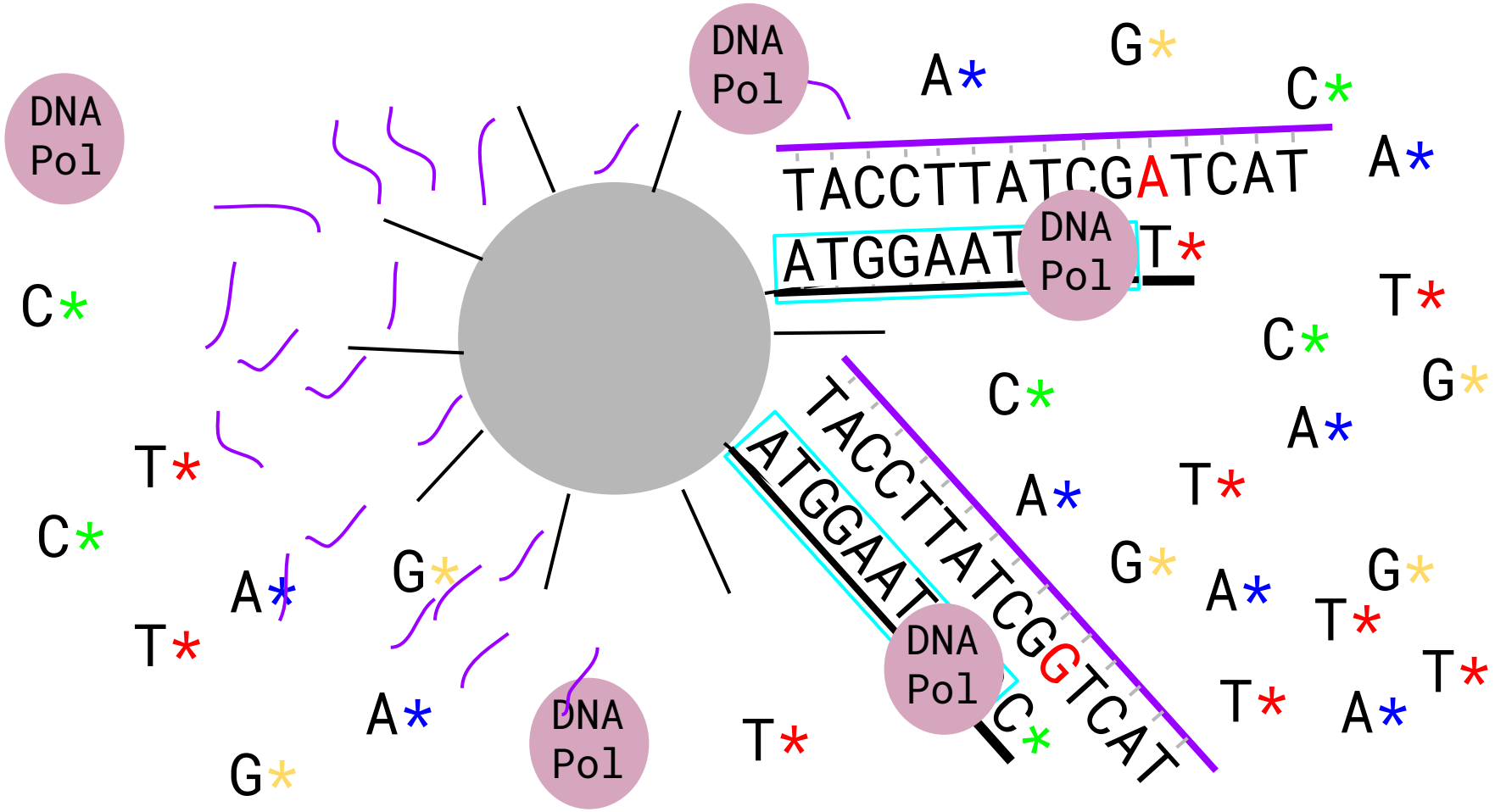
Illumina Bead SNP Array

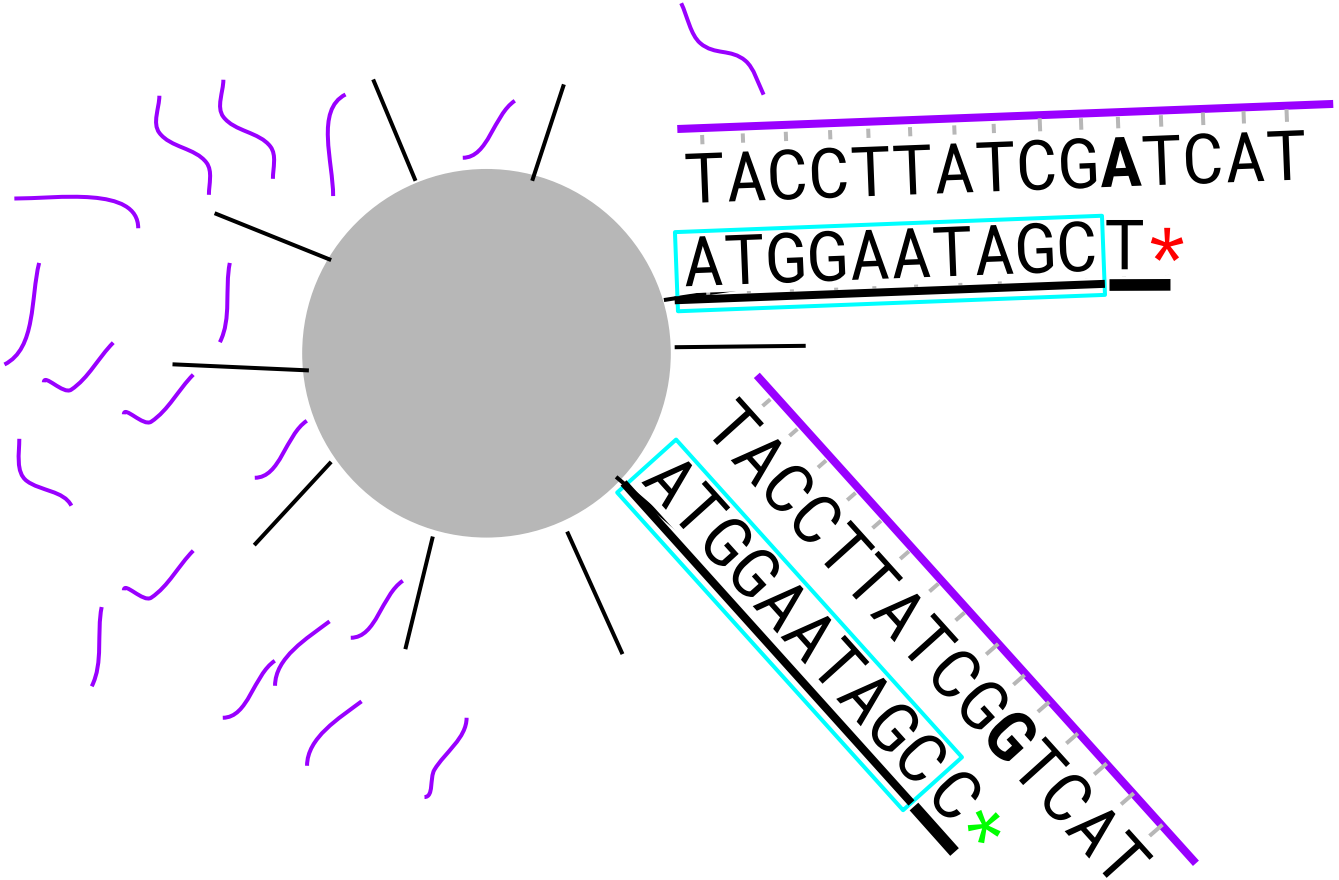




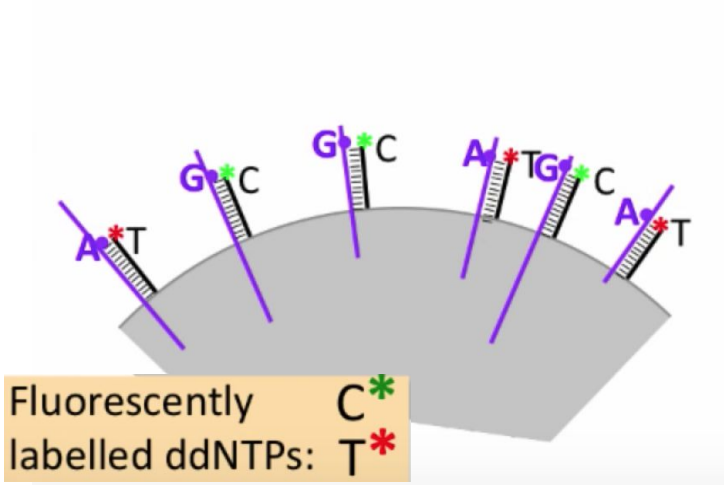

TACCTTATCGATCAT
ATGGAATAGC
TACCTTATCGGTCAT
ATGGAATAGC


DNA Pol
DNA Pol
DNA Pol
DNA Pol
DNA Pol
TACCTTATCGATCAT
ATGGAATAGC
TACCTTATCGGTCAT
ATGGAATAGC

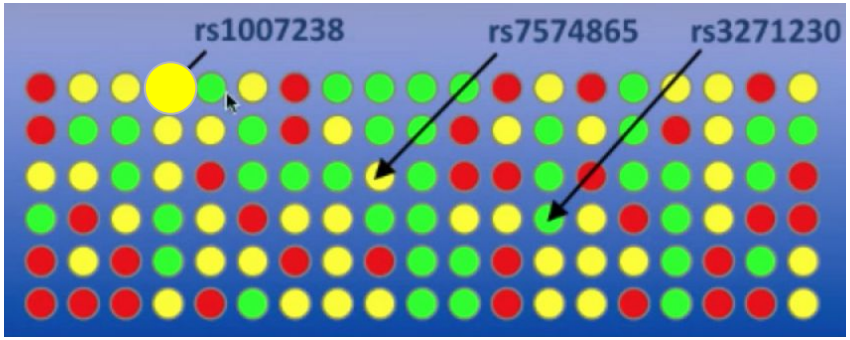

DNA Pol
DNA Pol
DNA Pol
DNA Pol
DNA Pol
TACCTTATCGATCAT
ATGGAAT
TACCTTATCGGTCAT
ATGGAAT

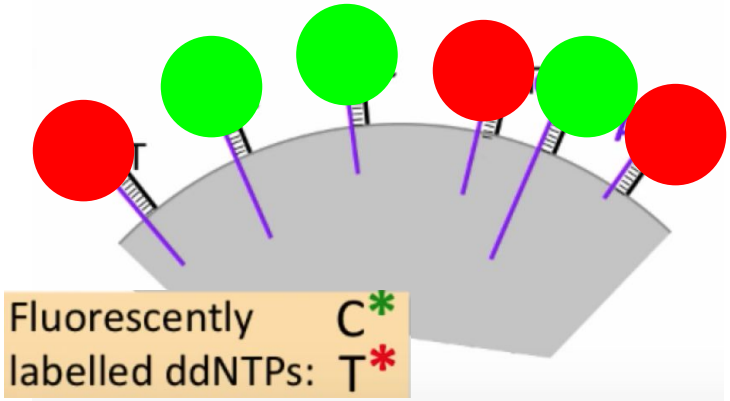

TACCTTATCGATCAT
ATGGAATAGCT*
TACCTTATCGGTCAT
ATGGAATAGCC*

Illumina Bead SNP Array



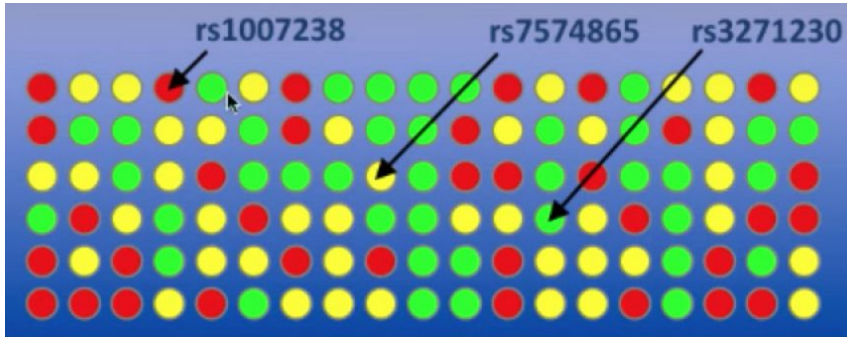
Illumina Bead SNP Array

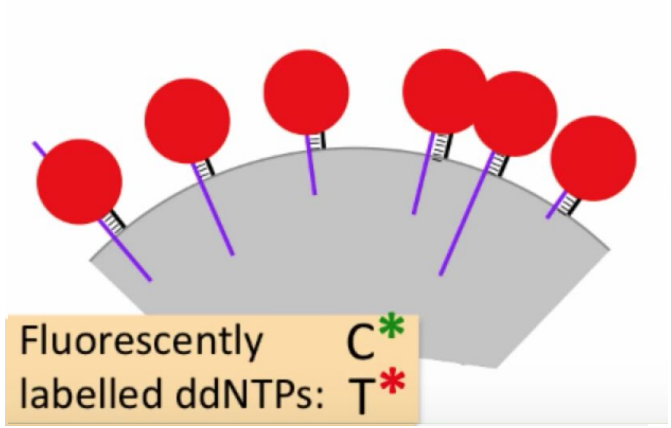




AG

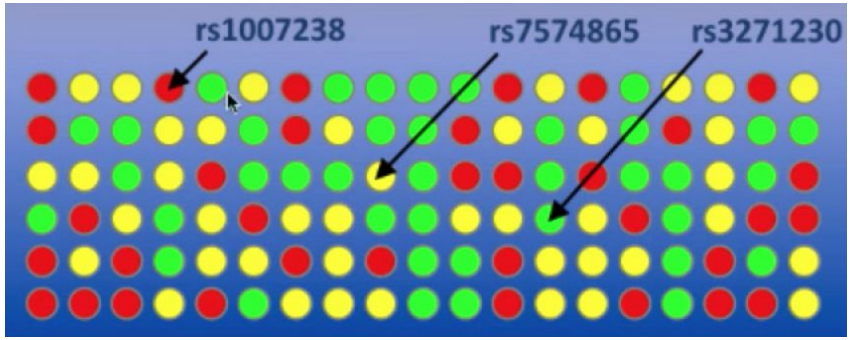
Illumina Bead SNP Array

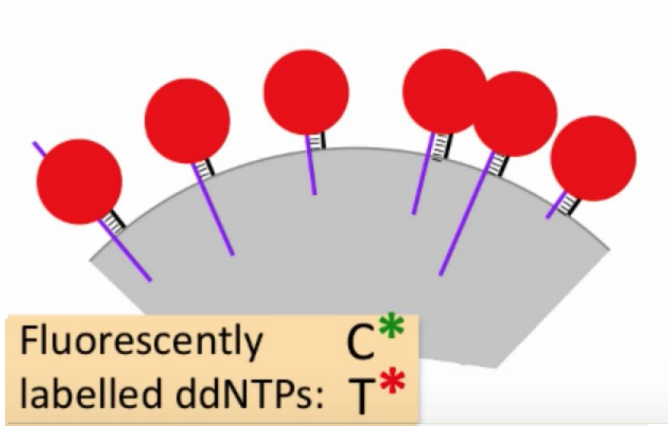




What's the genotype?

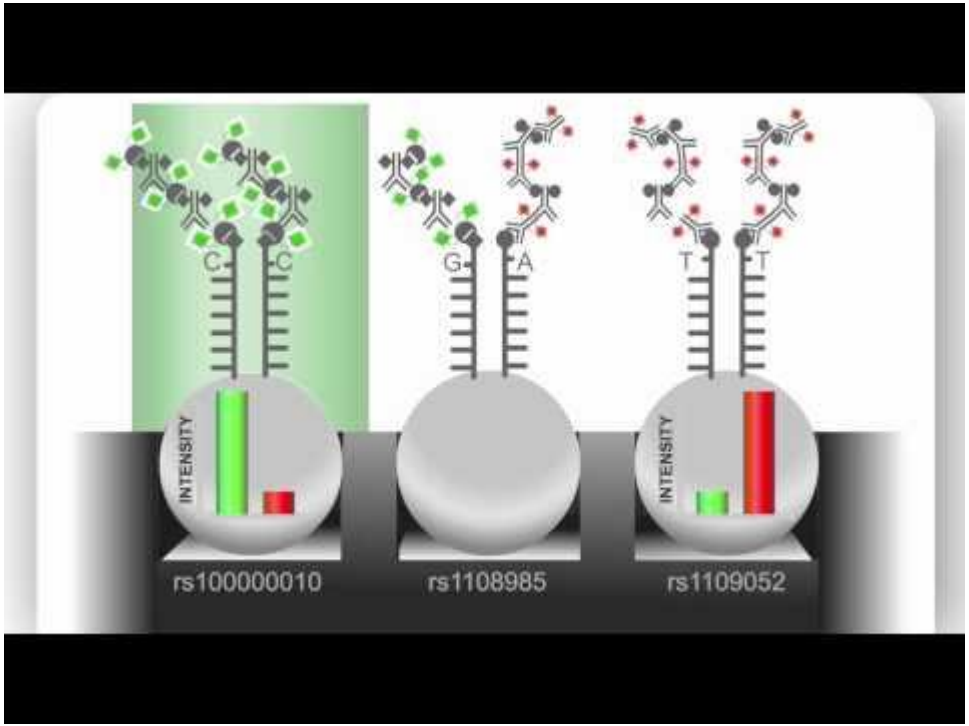
Illumina Bead SNP Array





AA

Illumina Bead SNP Array Recap



Take a break and then we'll talk about comparing dog's SNPs to one another

SNP array data for breed determination

Dog:						…
SNP1	AG	AA	GG	AA	AA	
SNP2	CC	CA	CC	CC	CC	
SNP3	AA	AA	TT	AA	AA	
SNP4	AT	TT	TT	TT	TT	
SNP5	GG	GG	GG	GG	GG	
SNP6	GG	CC	GG	GG	GG	
…						

We'll use purebred dogs to try and get an understanding of what each breed "looks like" on a SNP level

For our project, we'll have 6 dogs each from 93 different breeds!

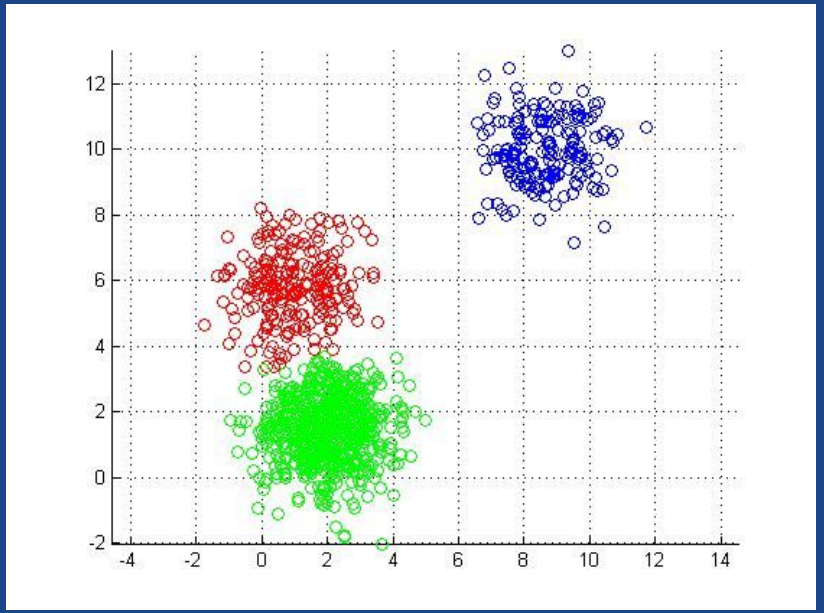
SNP array data for breed determination

Dog:						…
SNP1	AG	AA	GG	AA	AA	
SNP2	CC	CA	CC	CC	CC	
SNP3	AA	AA	TT	AA	AA	
SNP4	AT	TT	TT	TT	TT	
SNP5	GG	GG	GG	GG	GG	
SNP6	GG	CC	GG	GG	GG	
…						

We need to make sure the SNPs are helping us differentiate breed *before* we start comparing our mutt to the purebreds SNPs.

Clustering

Can we group similar dogs together using just their SNP data?



Breed A

Breed B

Breed C

How would you cluster these dogs?

How would you cluster these dogs?

How would you cluster these dogs?

How would you cluster these dogs?









Why cluster our dog data?

- To make sure the SNPs we have are actually capturing some breed “signature” that can differentiate breeds
- We have a some unknown dogs and want to see if they seem similar to one another, or to known dogs
- To see if we find a pattern we weren’t aware of
    - Maybe dogs that cluster together all have some shared trait we weren’t measuring

Why cluster our dog data?

- **To make sure the SNPs we have are actually capturing some breed “signature” that can differentiate breeds**
- We have a some unknown dogs and want to see if they seem similar to one another, or to known dogs
- To see if we find a pattern we weren’t aware of
    - Maybe dogs that cluster together all have some shared trait we weren’t measuring

How would you cluster these dogs?

Clustering

Dog:	1	2	3	4	5	…
SNP1	A	A	A	A	A	
SNP2	C	G	C	G	G	
SNP3	T	A	T	A	A	
SNP4	A	T	A	A	A	
SNP5	G	G	G	C	C	
SNP6	G	G	G	C	C	

How would you cluster these 5 dogs?

Clustering

Dog:	1	2	3	4	5	...
SNP1	A	A	A	A	A	
SNP2	C	G	C	G	G	
SNP3	T	A	T	A	A	
SNP4	A	T	A	A	A	
SNP5	G	G	G	C	C	
SNP6	G	G	G	C	C	

Clustering

Dog:	1	2	3	4	5	...
SNP1	A	A	A	A	A	
SNP2	C	G	C	G	G	
SNP3	T	A	T	A	A	
SNP4	A	T	A	A	A	
SNP5	G	G	G	C	C	
SNP6	G	G	G	C	C	

What if there could only be two clusters?

Clustering

Dog:	1	2	3	4	5	...
SNP1	A	A	A	A	A	
SNP2	C	G	C	G	G	
SNP3	T	A	T	A	A	
SNP4	A	T	A	A	A	
SNP5	G	G	G	C	C	
SNP6	G	G	G	C	C	

Clustering

Dog:	1	2	3	4	5	6	…
SNP1	A	A	A	A	A	A	
SNP2	C	G	C	G	G	G	
SNP3	T	A	T	A	A	A	
SNP4	A	T	A	A	A	T	
SNP5	G	G	G	C	C	C	
SNP6	G	G	G	C	C	C	

Now what would the two clusters be?

Clustering

Dog:	1	2	3	4	5	6	…
SNP1	A	A	A	A	A	A	
SNP2	C	G	C	G	G	G	
SNP3	T	A	T	A	A	A	
SNP4	A	**T**	A	A	A	**T**	
SNP5	G	G	G	C	C	C	
SNP6	G	G	G	C	C	C	

Clustering

Dog:	1	2	3	4	5	6	…
SNP1	A	A	A	A	A	A	
SNP2	C	G	C	G	G	G	
SNP3	T	A	T	A	A	A	
SNP4	A	**T**	A	A	A	**T**	
SNP5	G	G	G	C	C	C	
SNP6	G	G	G	C	C	C	

Cluster1: Dogs 2, 4, 5, 6

Cluster2: Dogs 1, 3

Clustering Algorithms: *k-means*

Unsupervised learning; we just need SNPs, no other info necessary

k-means aims to create *k* clusters, each centered around a mean value

How can we get values (and means) from SNPs?

Clustering Algorithms: *k-means*

Dog:	1	2	3	4	5	…
SNP1	A	A	A	A	A	
SNP2	C	G	C	G	G	
SNP3	T	A	T	A	A	
SNP4	A	T	A	A	A	
SNP5	G	G	G	C	C	
SNP6	G	G	G	C	C	

How can we turn this into numbers?

Clustering Algorithms: *k-means*

Dog:	1	2	3	4	5	…
SNP1	A	A	A	A	A	
SNP2	C	G	C	G	G	
SNP3	T	A	T	A	A	
SNP4	A	T	A	A	A	
SNP5	G	G	G	C	C	
SNP6	G	G	G	C	C	

A = 1

C = 2

G = 3

T = 4

Clustering Algorithms: *k-means*

Dog:	1	2	3	4	5	…
SNP1	1	1	1	1	1	
SNP2	2	3	2	3	3	
SNP3	4	1	4	1	1	
SNP4	1	4	1	1	1	
SNP5	3	3	3	2	2	
SNP6	3	3	3	2	2	

Let’s think of these as dimensions. So now we could plot these points (dogs) in 6D.

Clustering Algorithms: *k-means*

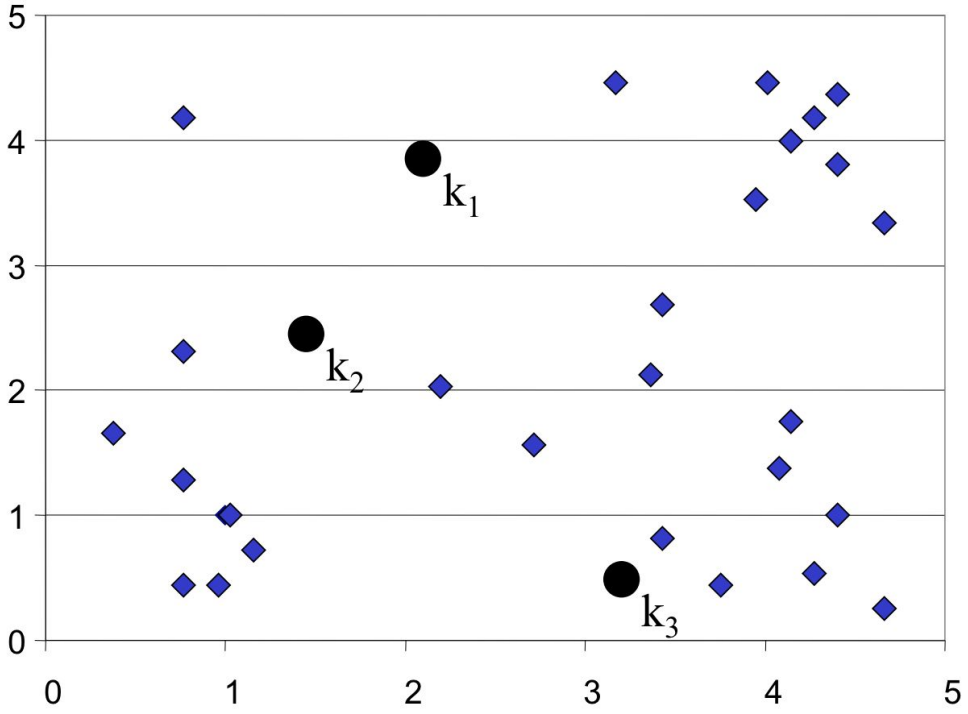
Dog:	1	2	3	4	5	…
SNP1	1	1	1	1	1	
SNP2	2	3	2	3	3	
SNP3	4	1	4	1	1	
SNP4	1	4	1	1	1	
SNP5	3	3	3	2	2	
SNP6	3	3	3	2	2	

Actually could plot in 5D, as SNP1 doesn't give us any useful info.

For ease of understanding, we'll work in 2D to walk through the algorithm.

K-means Clustering: Initialization

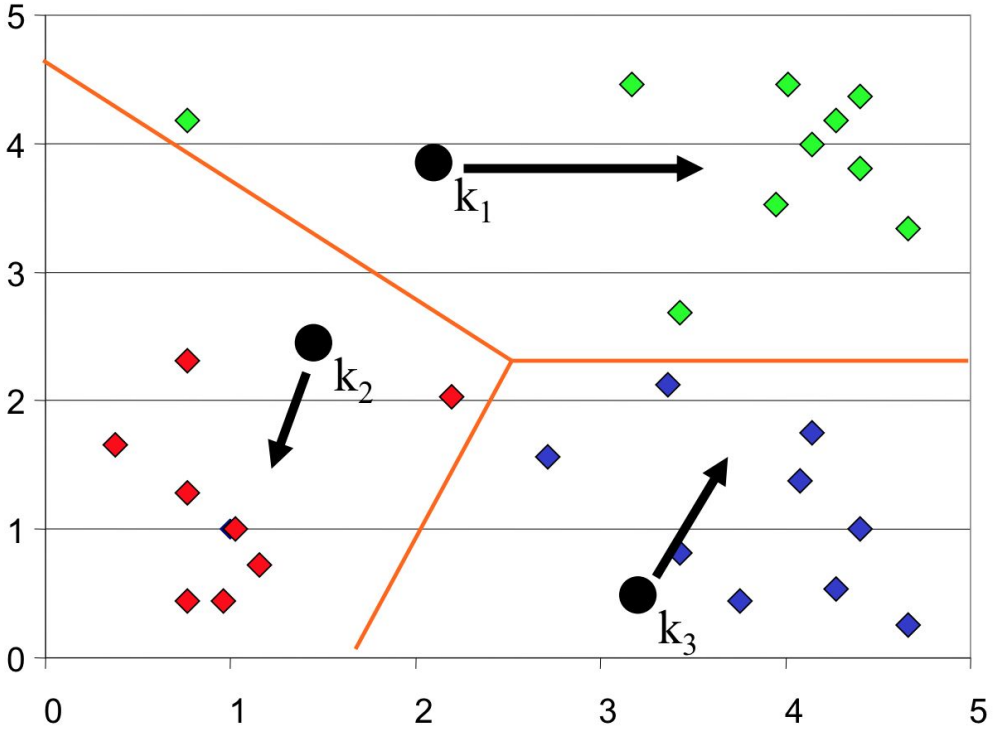
Decide K, and initialize K centers (randomly)



Slides from CMU

K-means Clustering: Iteration 1

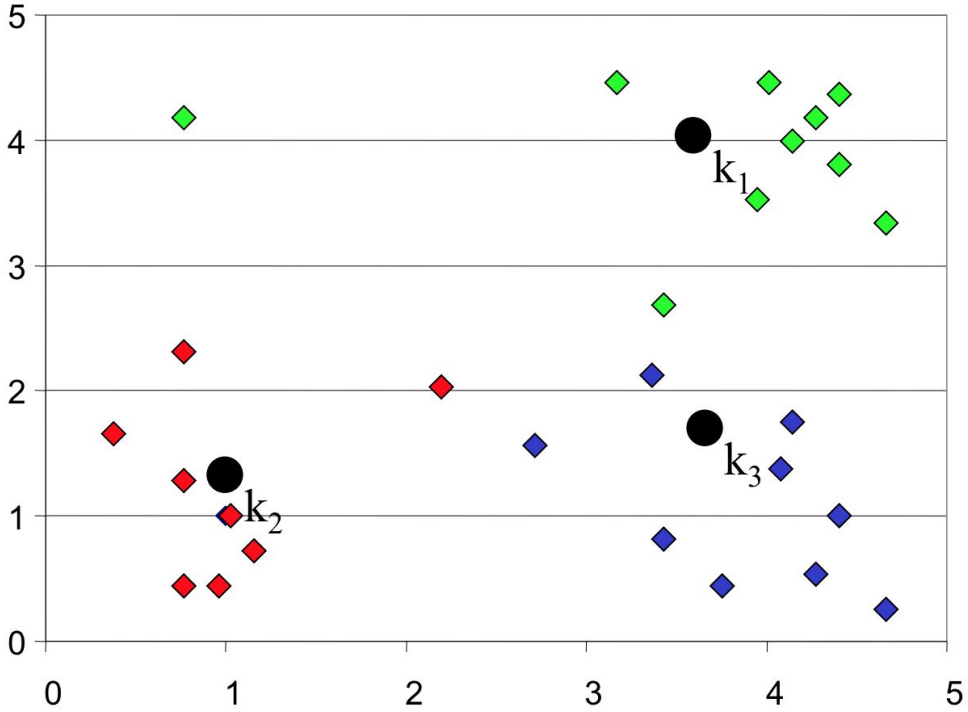
Assign all objects to the nearest center.
Move a center to the mean of its members.



Slides from CMU

K-means Clustering: Iteration 2

After moving centers, re-assign the objects…

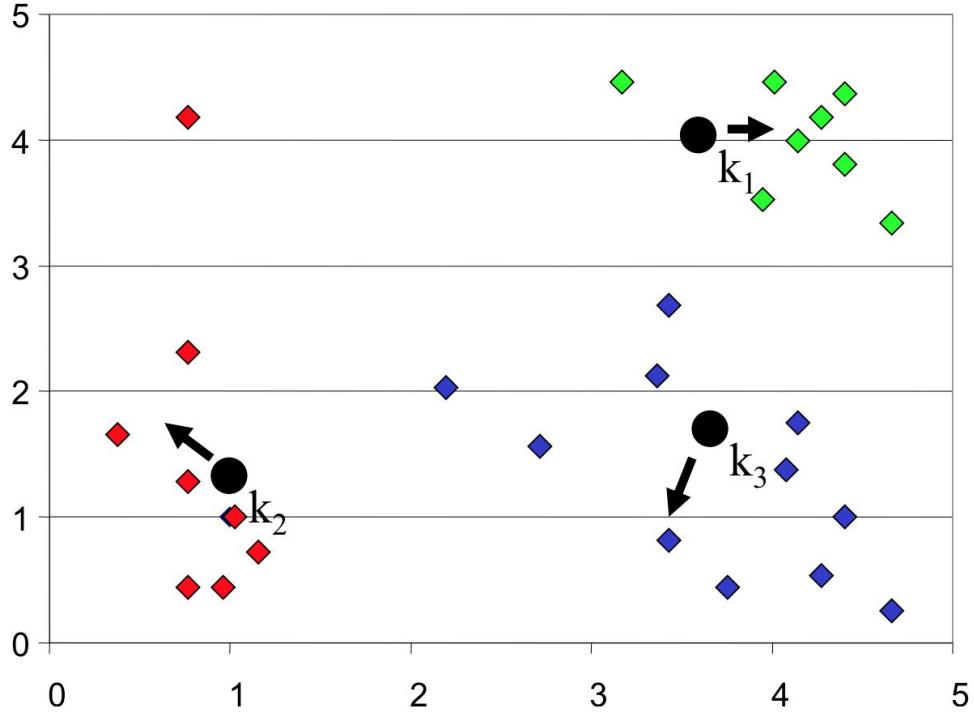


Slides from CMU

K-means Clustering: Iteration 2

After moving centers, re-assign the objects to nearest centers.

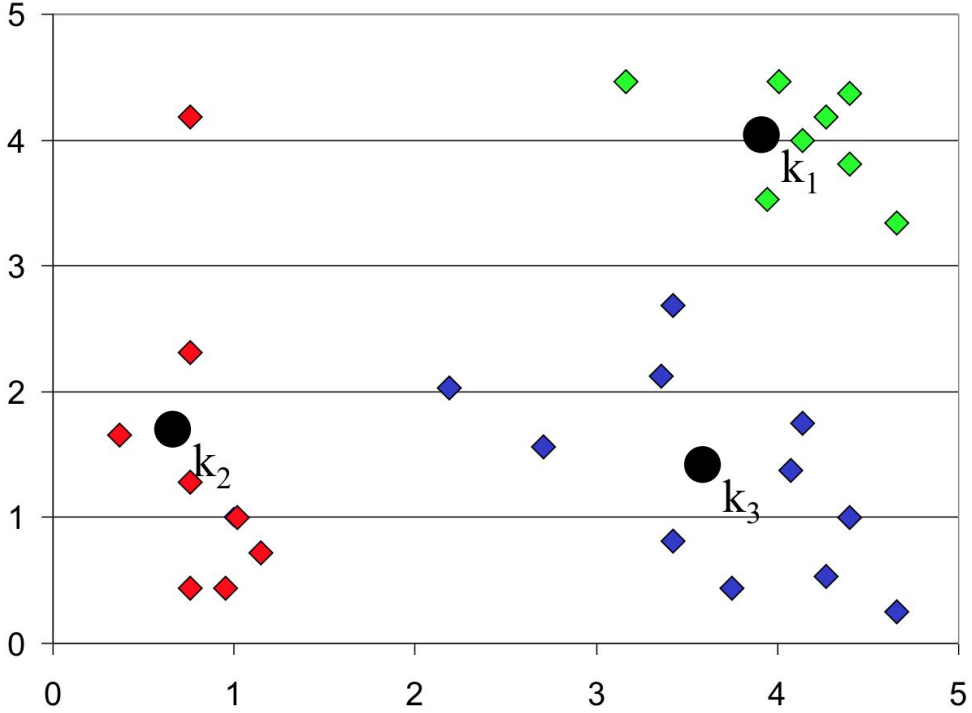
Move a center to the mean of its new members.



Slides from CMU

K-means Clustering: Finished!

Re-assign and move centers, until …
no objects changed membership.



Slides from CMU

Strengths and weaknesses of *k*-means

Strengths:

Easy to implement
Relatively interpretable

Weaknesses:

Choosing *k* can be hard
Initialization can greatly affect the final results

Final thoughts on *k*-means clustering

We can use *k*-means on purebred dogs to make sure they cluster together.

Could *k*-means help us figure out what breeds Reilly, Clarence, and Finch are?

How might we go about doing this?



Take a break and then we'll talk about how SNP sites get chosen

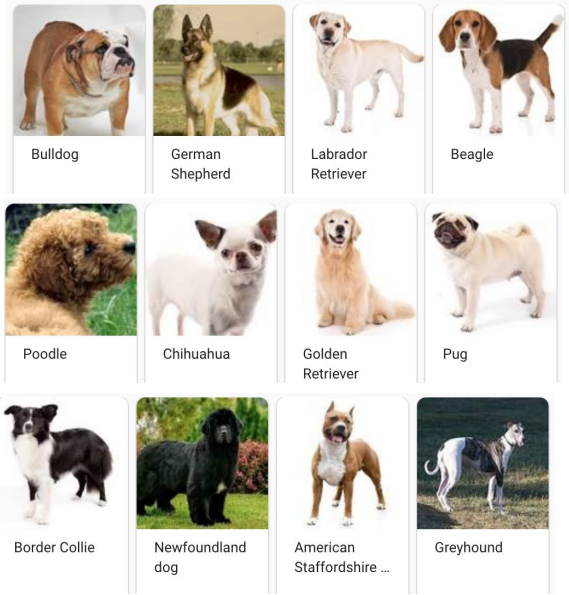
2.5 million recorded dog SNPs ---›

~170k in Illumina canine SNP Chip

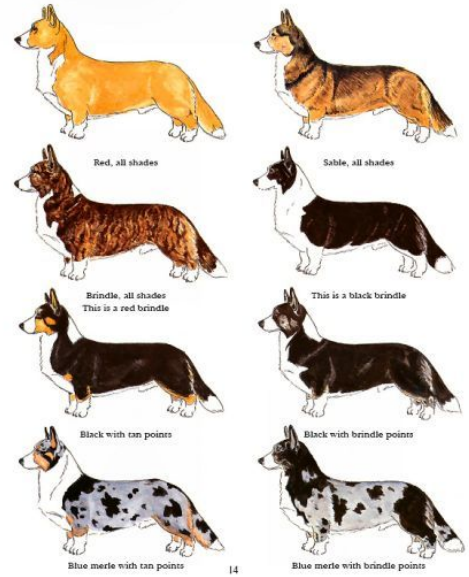
How do we get there?

Choosing SNPs of interest

What do we care about?



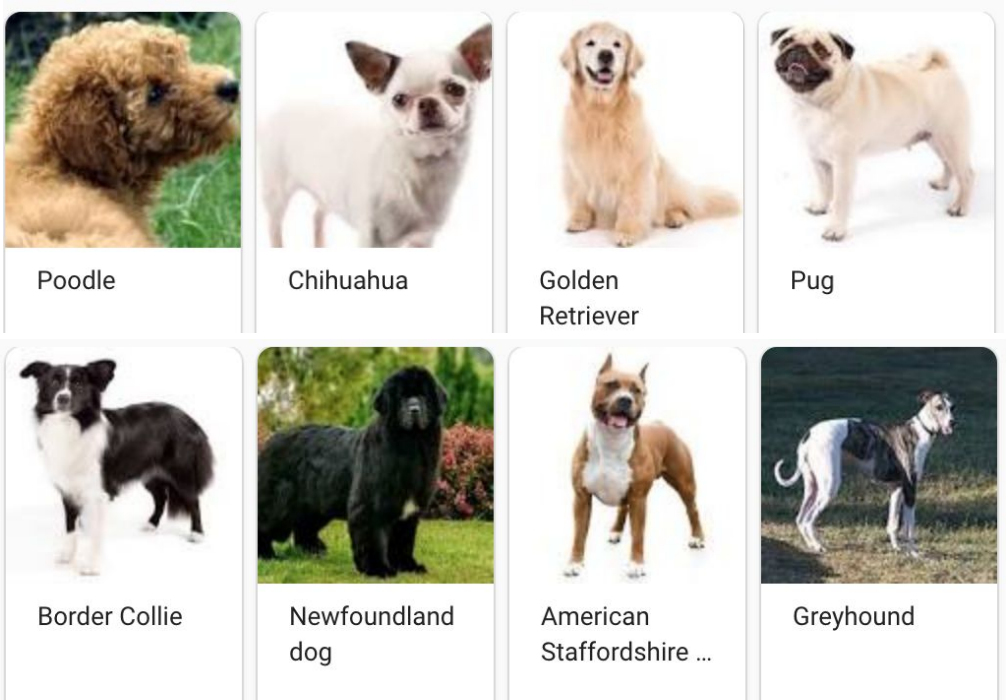
Breed



Coat color

Disease (albinism)

Choosing SNPs of interest - informativeness



Is a SNP informative?

Let’s say our goal is to tell these 8 breeds apart.

Does nose color help?

Choosing SNPs of interest - informativeness

Is a SNP informative?

What if instead we want to differentiate albino and non albino?

Does nose color help?

Choosing SNPs of interest - informativeness

Is a SNP informative?

What if instead we want to differentiate albino and non albino?

Does nose color help?

Maybe! Not many examples, but seems promising.

Choosing SNPs of interest - informativeness

Is a SNP informative?

What if instead we want to differentiate albino and non albino?

Does nose color help?

But what if this breed just happens to have pink noses, and it's not an albino trait?

Choosing SNPs of interest - informativeness

Is a SNP informative?

What if instead we want to differentiate albino and non albino?

Does nose color help?

Having the same breed in each category is more informative!

Choosing SNPs of interest - correlation

Is a SNP informative *for a given trait*?

Say we want to differentiate albino and non albino.

Does nose color help?

Choosing SNPs of interest - correlation

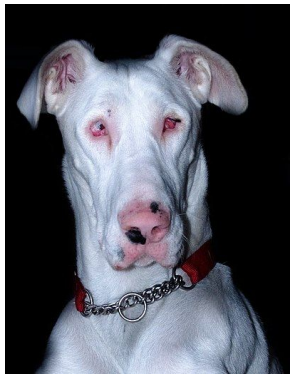
Is a SNP informative *for a given trait*?

Say we want to differentiate albino and non albino.

Does nose color help?

Seems less helpful now because we can see it's not 100% correlated with the trait we care about.

Choosing SNPs of interest - correlation



Double merle coat

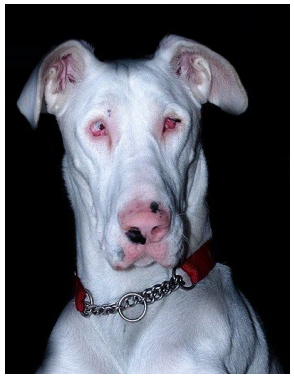
Albinism

Is a SNP informative *for a given trait*?

Say we want to differentiate albino and non albino.

Does nose color help?

Choosing SNPs of interest - correlation



Double merle coat

Albinism

Is a SNP informative *for a given trait*?

Say we want to differentiate albino and non albino.

Does nose color help?

Not here, because it's not correlated with our trait of interest!

Choosing SNPs of interest - redundancy

Imagine that:

All brown dogs have curly tails.
All curly tailed dogs are brown.



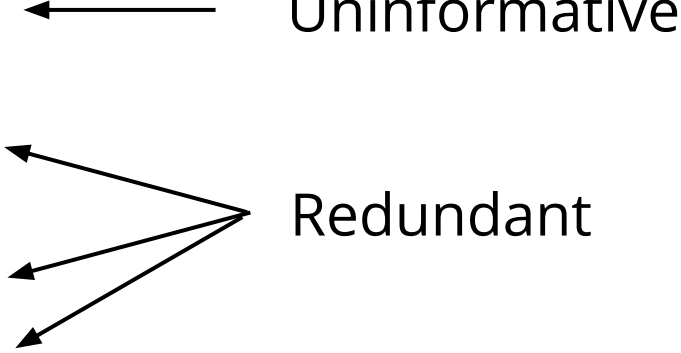
Let's say we want to differentiate some dogs from one another:




Does knowing curly vs not curly *and* brown vs not brown help?

Choosing SNPs of interest

Dog:	1	2	3	4	5
SNP1	A	A	A	A	A
SNP2	C	G	C	G	G
SNP3	T	T	T	A	A
SNP4	A	T	A	A	A
SNP5	G	G	G	C	C
SNP6	G	G	G	C	C



Choosing SNPs to maximize total information

Let’s say we want to differentiate four dogs from each other, and we decide one of our “snapshots” should be nose color (pink or black).

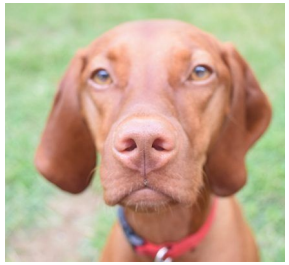
Choosing SNPs to maximize total information

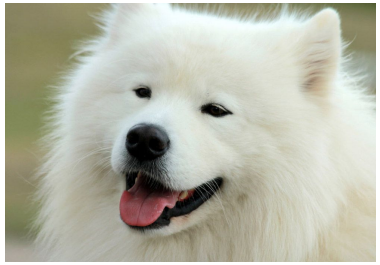
Is nose color more informative if one of our four dogs has a pink nose, or if two of them have pink noses?

Choosing SNPs to maximize total information

Let’s say we choose two traits to try to differentiate the four dogs.

1. Nose color (pink or black)
2. Coat color (brown or white)

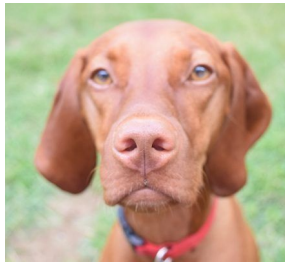
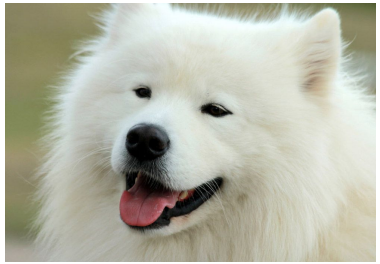





Choosing SNPs to maximize total information

Let’s say we choose two traits to try to differentiate the four dogs.

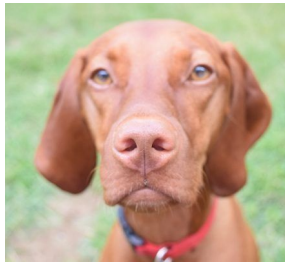
1. Nose color (pink or black)
2. Coat color (brown or white)

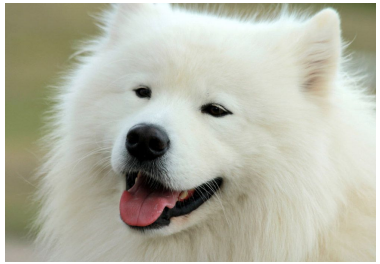
1. P 2. B	1. P 2. W	1. B 2. W	1. B 2. B
			

Choosing SNPs to maximize total information

Let’s say we choose two traits to try to differentiate the four dogs.

1. Nose color (pink or black)
2. Coat color (brown or white)



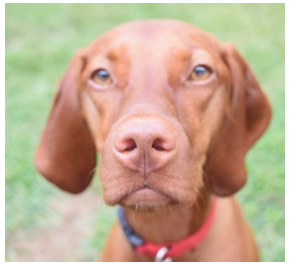



Choosing SNPs to maximize total information

Let’s say we choose two traits to try to differentiate the four dogs.

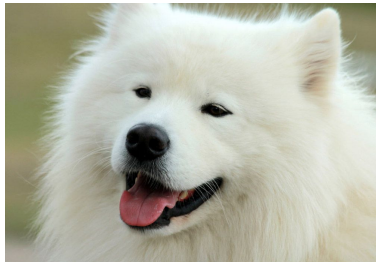
1. Nose color (pink or black)
2. Coat color (brown or white)

1. P
2. B



1. B
2. W

1. B
2. W



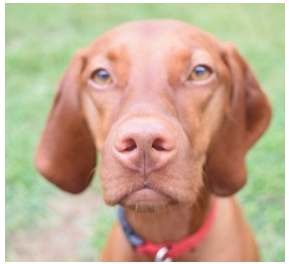
1. B
2. B

Choosing SNPs to maximize total information

Let’s say we choose two traits to try to differentiate the four dogs.

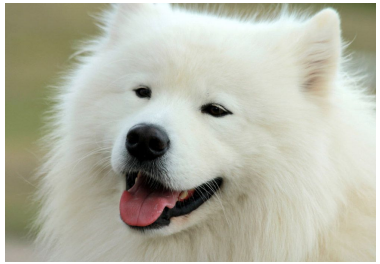
1. Nose color (pink or black)
2. Coat color (brown or white)

1. P
2. B



1. B
2. W

1. B
2. W



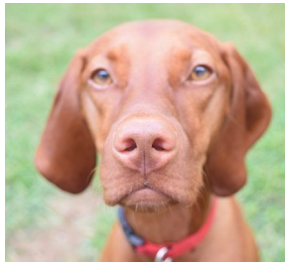
1. B
2. B

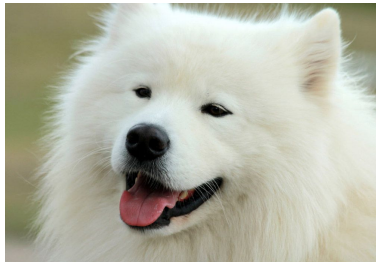
Choosing SNPs to maximize total information

Let’s say we choose two traits to try to differentiate the four dogs.

1. Nose color (pink or black)
2. Coat color (brown or white)

We’d need a third trait!



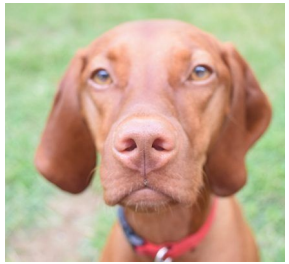



Choosing SNPs to maximize total information

This goes back to how many SNPs are theoretically needed. For 4 dogs, and traits with 2 options, we need 2 traits since for $2^x \geq 4$ combinations, $x \geq 2$.

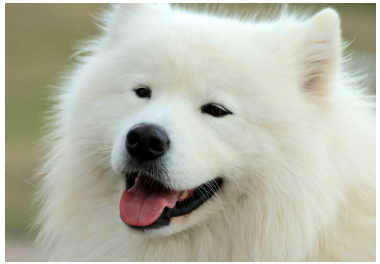
The closer to 50/50 each trait is, the more likely we see all combinations.

1. P
2. B



1. B
2. W

1. B
2. W



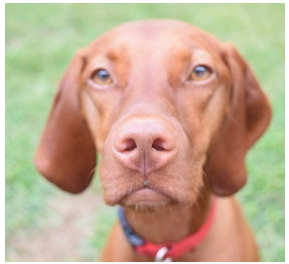
1. B
2. B

Choosing SNPs to maximize total information

This goes back to how many SNPs are theoretically needed. For 4 dogs, and traits with 2 options, we need 2 traits since for $2^x \geq 4$ combinations, $x \geq 2$.

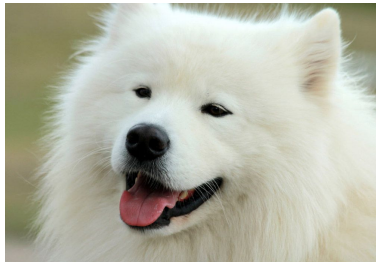
The closer to 50/50 each trait is, the more likely we see all combinations.

1. P
2. B



1. B
2. W

1. B
2. W



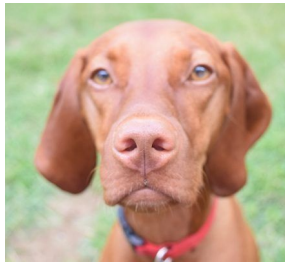
1. P
2. B

Choosing SNPs to maximize total information

This goes back to how many SNPs are theoretically needed. For 4 dogs, and traits with 2 options, we need 2 traits since for $2^x \geq 4$ combinations, $x \geq 2$.

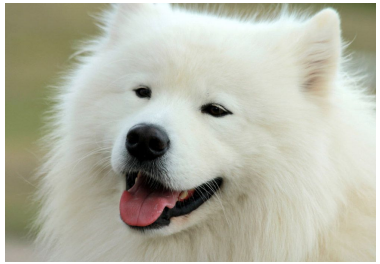
The closer to 50/50 each trait is, the more likely we see all combinations.

1. P
2. B



1. B
2. W

1. B
2. W



1. P
2. B

Redundancy!

Choosing SNPs to maximize total information

We refer to the % of the time we see the less common variant in the population as the ***minor allele frequency (MAF)***.

Maximizing total information by selecting SNPs with the largest MAFs can be useful if we don't know what traits we want to look at.

> More than 90% of the SNPs on the CanineHD BeadChip are polymorphic across the samples that were interrogated for product quality testing (Table 2). The average minor allele frequency (MAF) across all 26 breeds is 0.23, while the breed-specific MAFs range from 0.13 to 0.21 (Table 2).

For Friday

Make sure you have a partner!

Bring your laptop (one per group is fine)!

Discussion on commercial sequencing

Discussion Questions

Given the issues of owner/vet misunderstanding of results, do you think commercial dog sequencing for medical markers should be allowed? Why or why not?

Aside from medical decision making, what other problems might arise from unregulated sequencing?

Are there any ways in which sequencing for breed determination only might be problematic?