



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

Lecture 7
Jan 23rd, 2019



Dogs and their best friends



Human ancestry testing kits



Human ancestry testing kits



Ethnicity Regions

● Cameroon/Congo	13%
● Ireland/Scotland/Wales	14%
● Senegal	14%
● Benin/Togo	12%
● Scandinavia	11%
● Great Britain	9%
● Nigeria	8%
● +12 Other Regions	

Migrations

- North Carolina African Americans
- South Carolina African Americans
- Virginia & Southern States African Americans

AncestryDNA “White Paper”

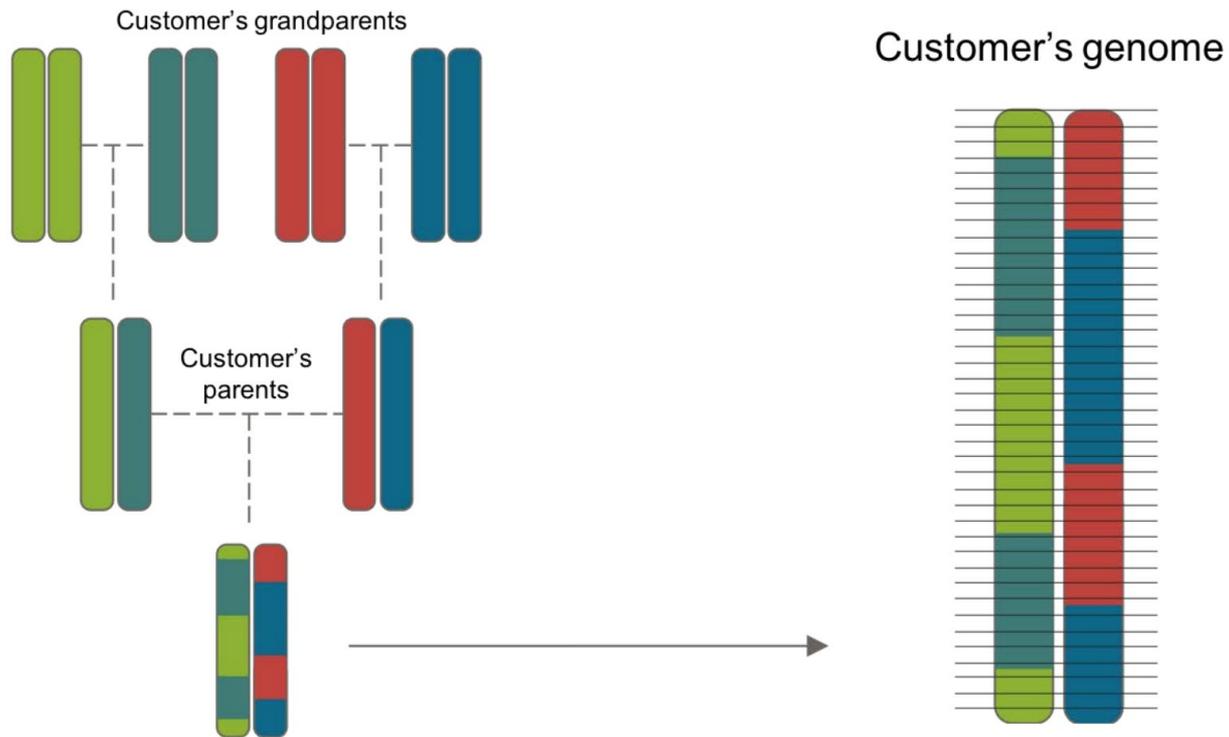
After establishing and **validating the reference panel**, the next step is to estimate a customer’s ethnicity by **comparing** over 300,000 **single nucleotide polymorphisms (SNPs) from his or her DNA to those of the reference panel.**

We assume that an individual’s DNA is a mixture of DNA from the 43 populations represented in the reference panel.

Because DNA is passed down from one generation to the next in long segments, **it is likely that the DNA at two nearby SNPs, or positions, in the genome was inherited from the same person and so comes from the same population.**

This means we can **get more accurate results by looking at multiple nearby SNPs together as a group**, or haplotype, instead of looking at each SNP in isolation.

AncestryDNA “White Paper”



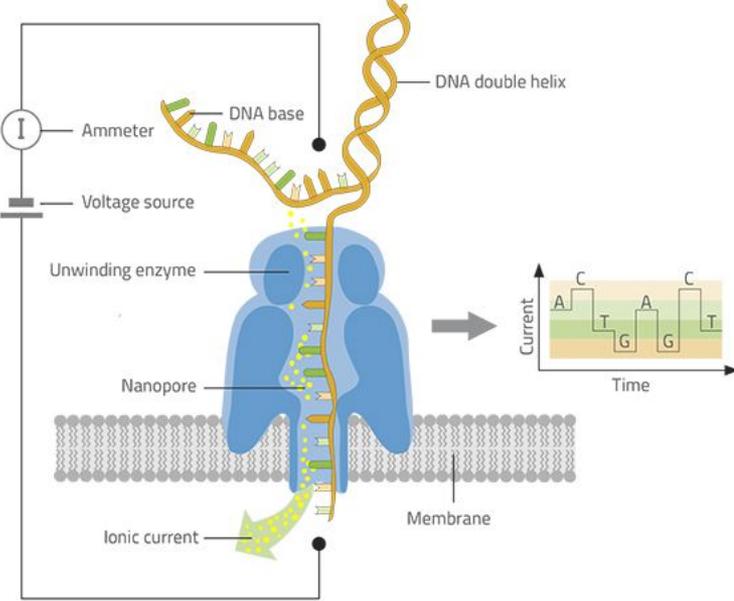
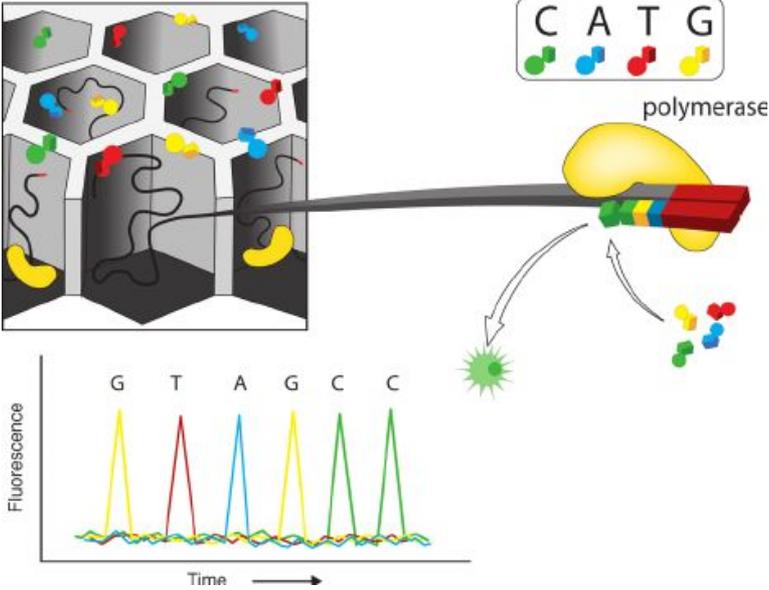
AncestryDNA “White Paper”

One set of chromosomes comes from Mom and the other from Dad. This means **there are two results for each position AncestryDNA analyzes, and those results must be interpreted to assign which DNA came from which set of chromosomes (this process is called phasing).**

AncestryDNA must consider what possible combinations of ethnicities might look like. **We create a genome-wide HMM where each possible ethnicity combination (or hidden state) is represented by a pair of populations in a window of the genome, and changes between windows that are next to each other are unlikely to change the state.**

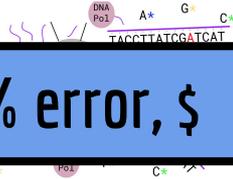
By applying these probabilities to the whole genome, we can **obtain a sequence of population assignments along a customer's genome.**

Long read sequencing (aka third generation)



Dog and human genomics research

SNP Arrays

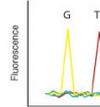


Single bases, < 0.1% error, \$

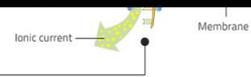
Second generation sequencing

Whole genome, reads
are ~250 bases
< 0.1% error rate, \$\$

Third generation sequencing



Whole genome, reads ~30k
base (Pacbio) to ~100k+ base
(Nanopore) average length
~10-15% error rate
\$\$\$\$ or lower throughput



Dog and human genomics research

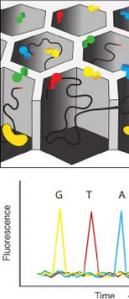
SNP Arrays



Second generation sequencing



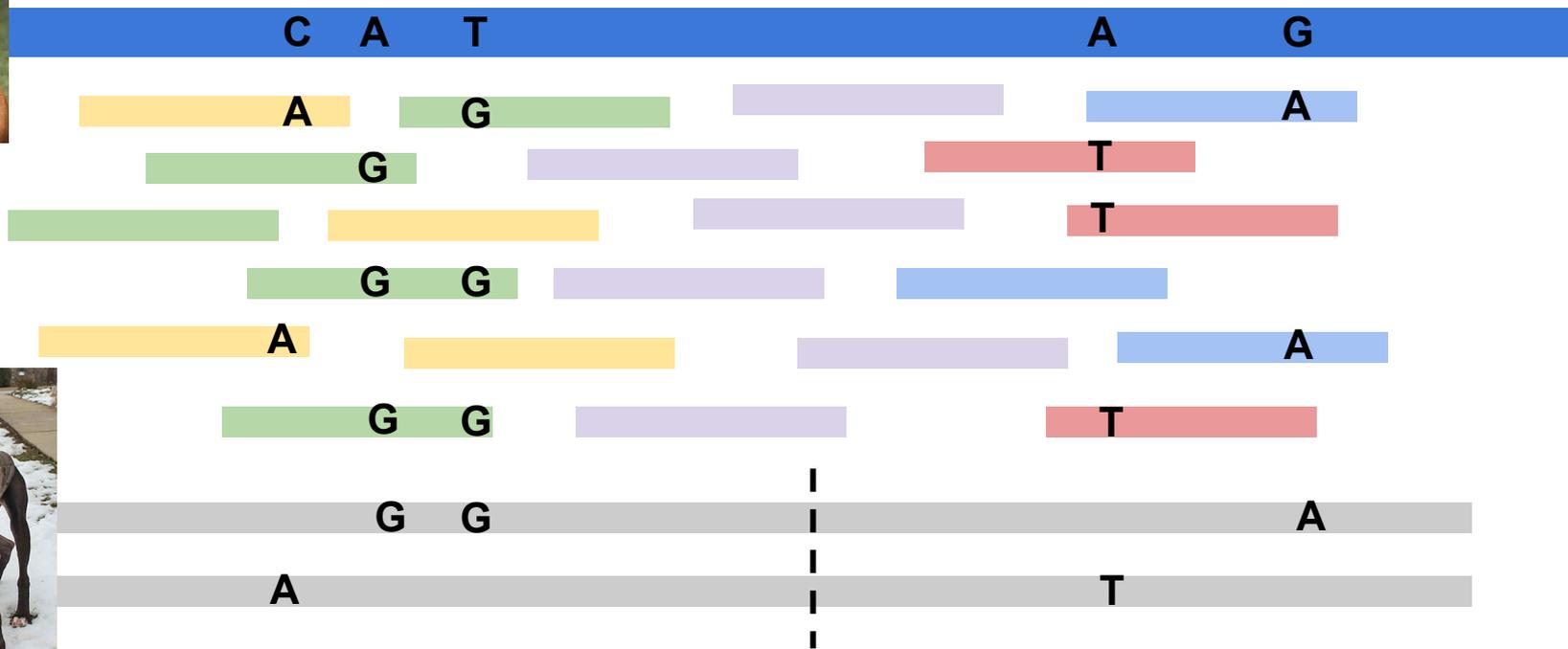
Third generation sequencing



Alignment and phasing



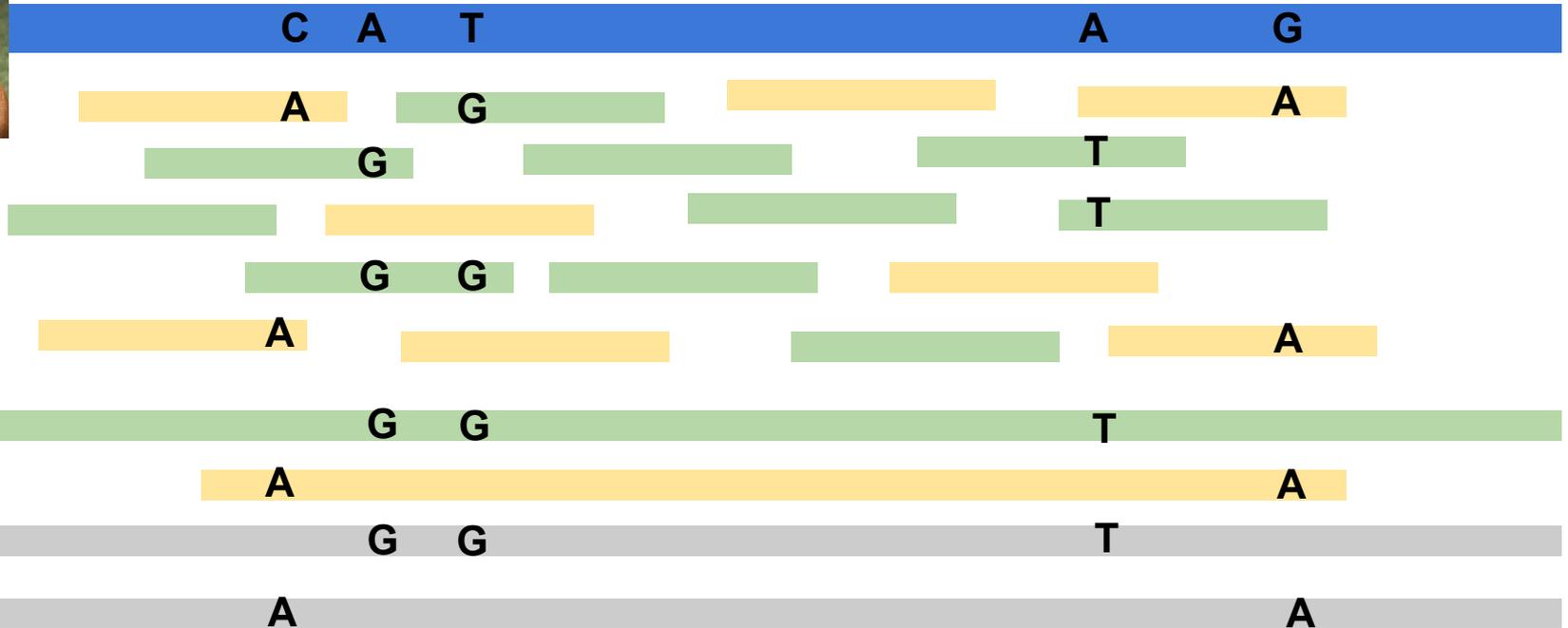
Tasha (Reference)



Third generation sequencing (long reads)

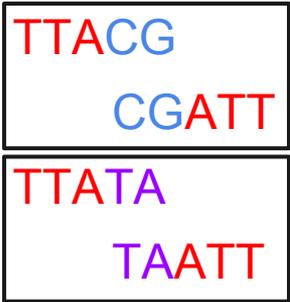


Tasha (Reference)



Repetitive sequences + short reads = ambiguity

GGATT



ATTAG

AGACC



GGATTACGATTATAATTAGACC
GGATTATAATTACGATTAGACC



Repeated sequence

Third generation sequencing (long reads)

GGATTACGA

TTACGATTATA

GATTATAATTAG

ATTAGACC



GGATTACGATTATAATTAGACC

Breed without \$\$\$ limitations

How might you want to look at breed,
assuming there's no cost limit?

Population reference genomes



Boxer reference



Corgi reference



Saint Bernard reference



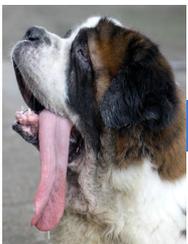
Population reference genomes



Boxer reference



Corgi reference



Saint Bernard reference



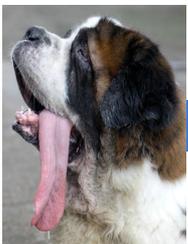
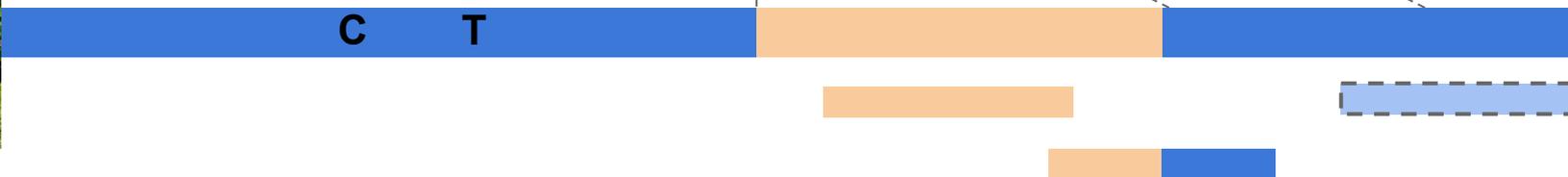
Population reference genomes



Boxer reference



Corgi reference



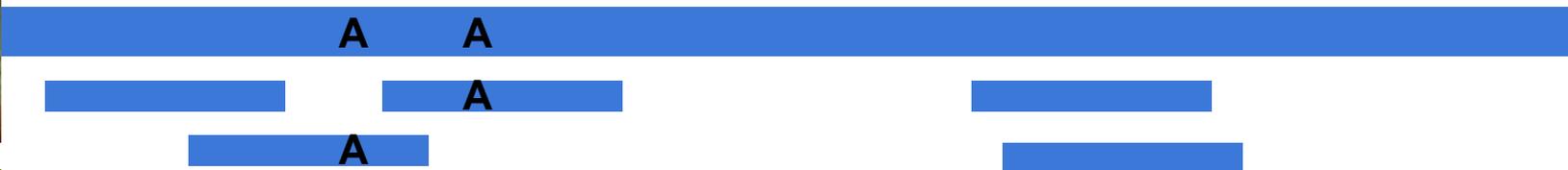
Saint Bernard reference



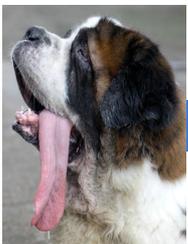
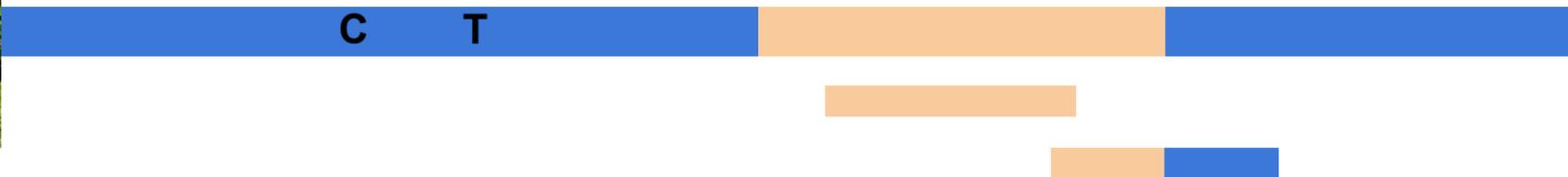
Population reference genomes



Boxer reference



Corgi reference



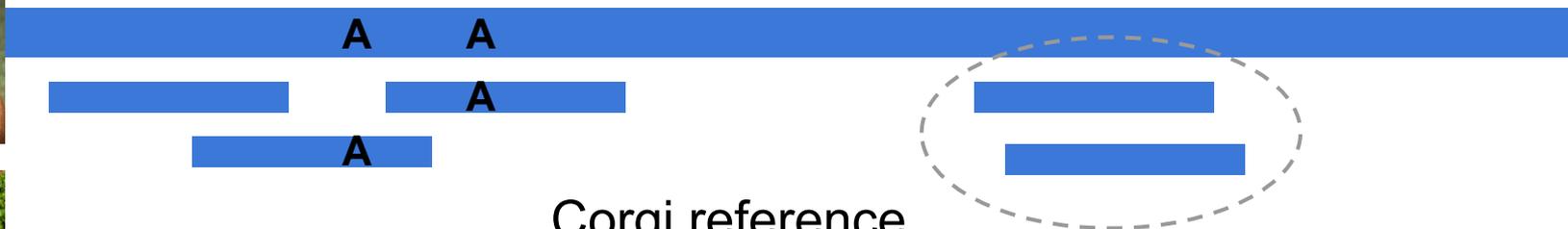
Saint Bernard reference



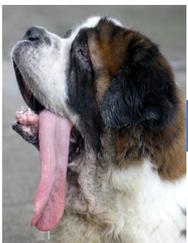
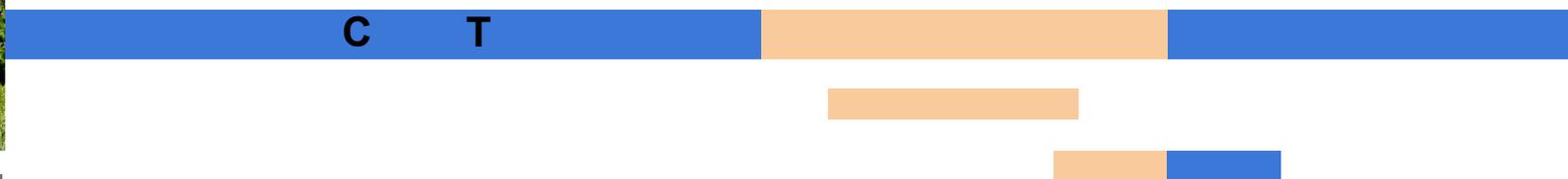
Population reference genomes



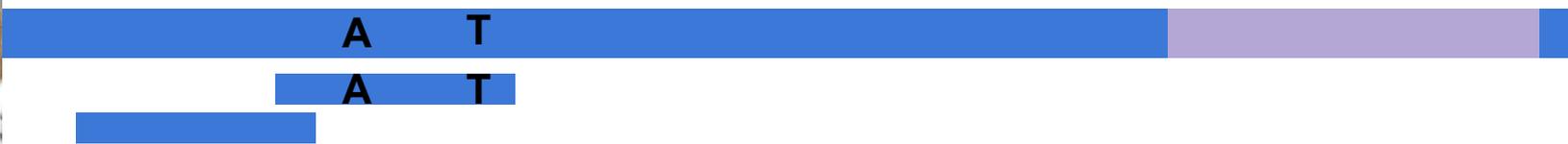
Boxer reference



Corgi reference



Saint Bernard reference



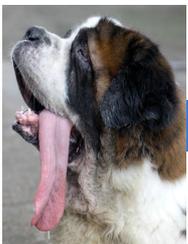
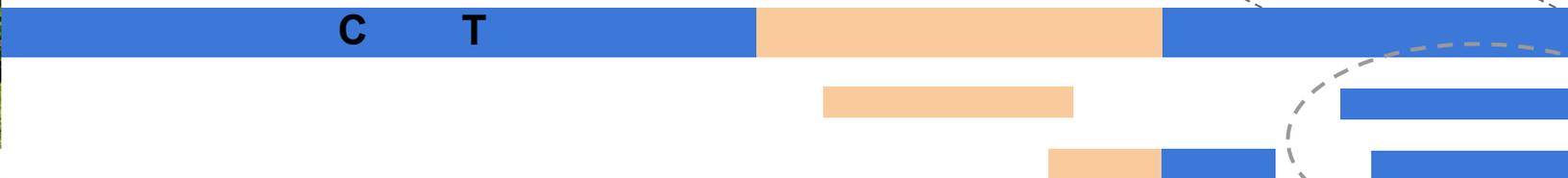
Population reference genomes



Boxer reference



Corgi reference



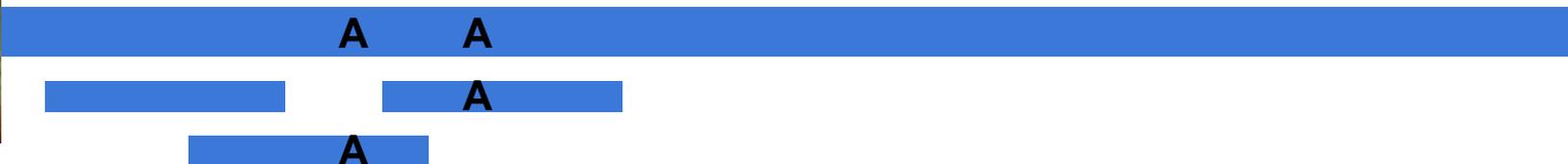
Saint Bernard reference



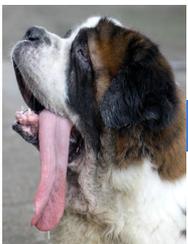
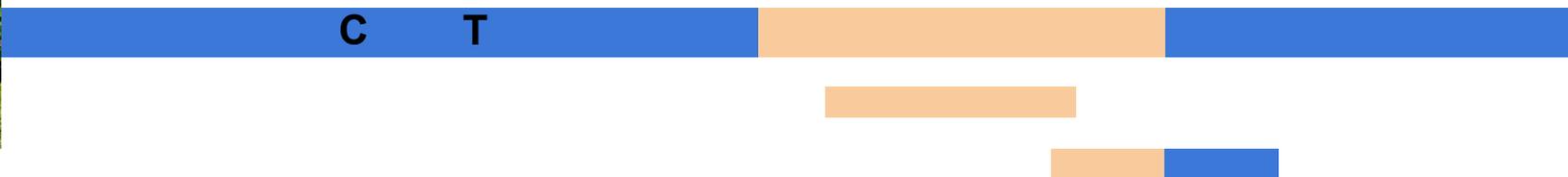
Population reference genomes



Boxer reference



Corgi reference



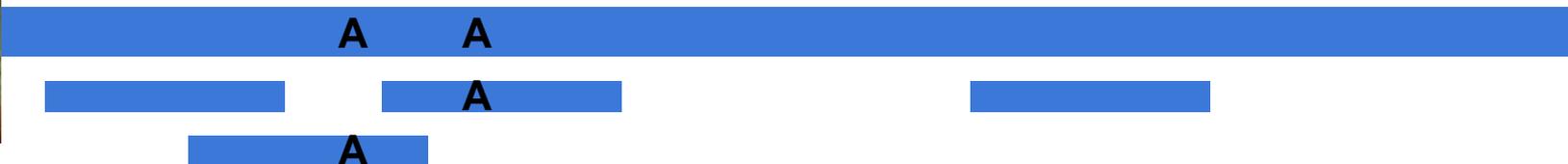
Saint Bernard reference



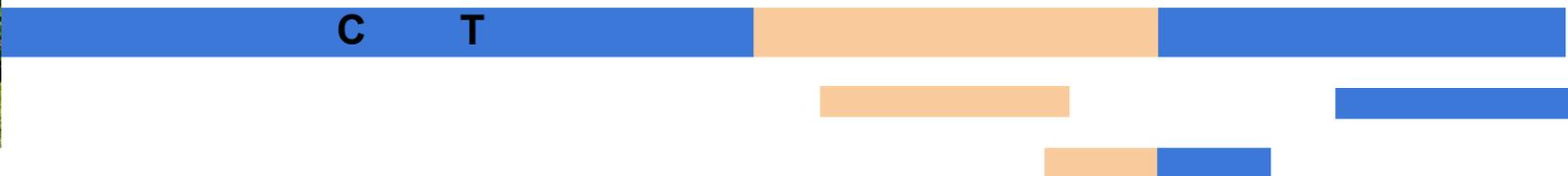
Population reference genomes



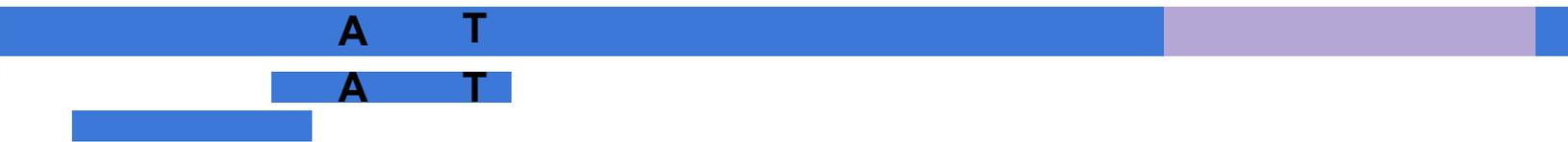
Boxer reference



Corgi reference



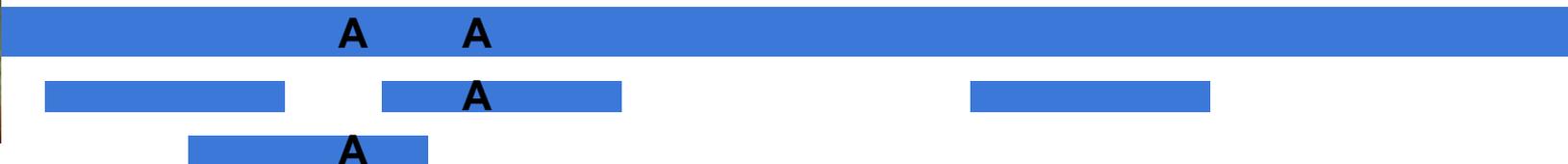
Saint Bernard reference



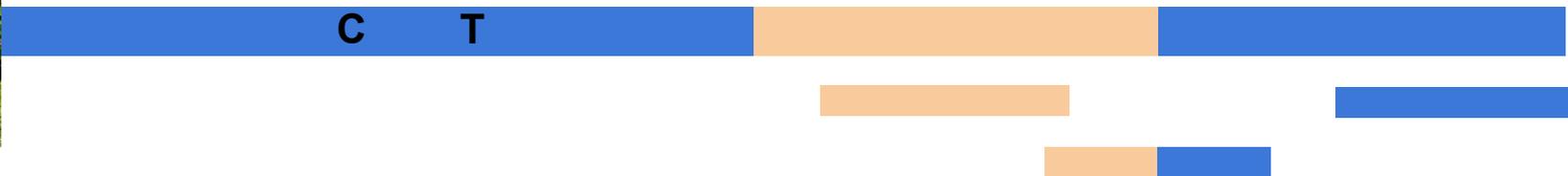
Population reference genomes



Boxer reference



Corgi reference



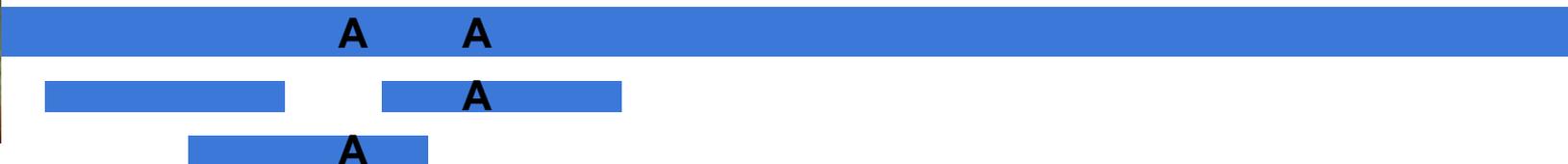
Saint Bernard reference



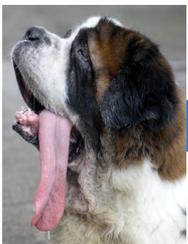
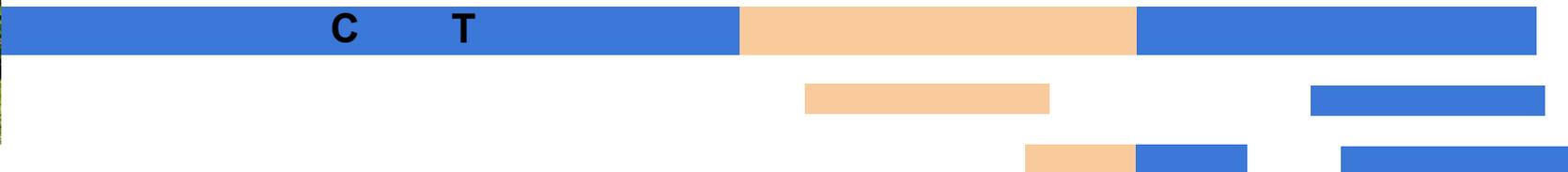
Population reference genomes



Boxer reference



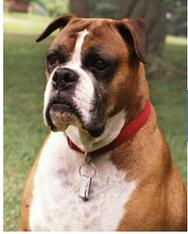
Corgi reference



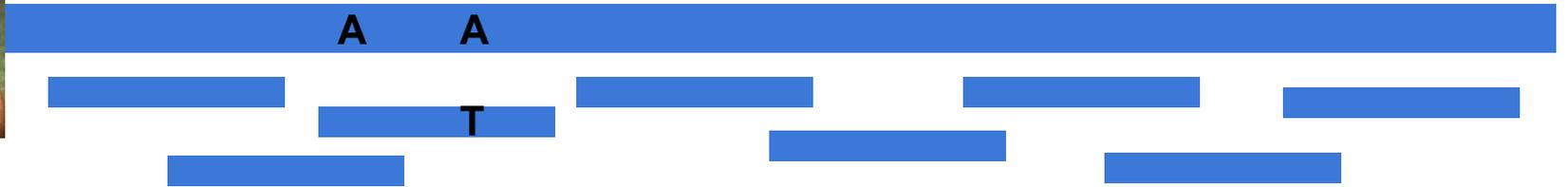
Saint Bernard reference



Population reference genomes



Boxer reference



Don't align, generally discarded

Might align, but won't help determine breed



Genomics population research

Letter | [OPEN](#) | Published: 19 November 2018

Assembly of a pan-genome from deep sequencing of 910 humans of African descent

Rachel M. Sherman , Juliet Forman, [...] Steven L. Salzberg 

Article | [OPEN](#) | Published: 24 November 2016

An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes

Yun Sung Cho, Hyunho Kim, Hak-Min Kim, Sungwoong Jho, JeHoon Jun, Yong Joo Lee, Kyun Shik

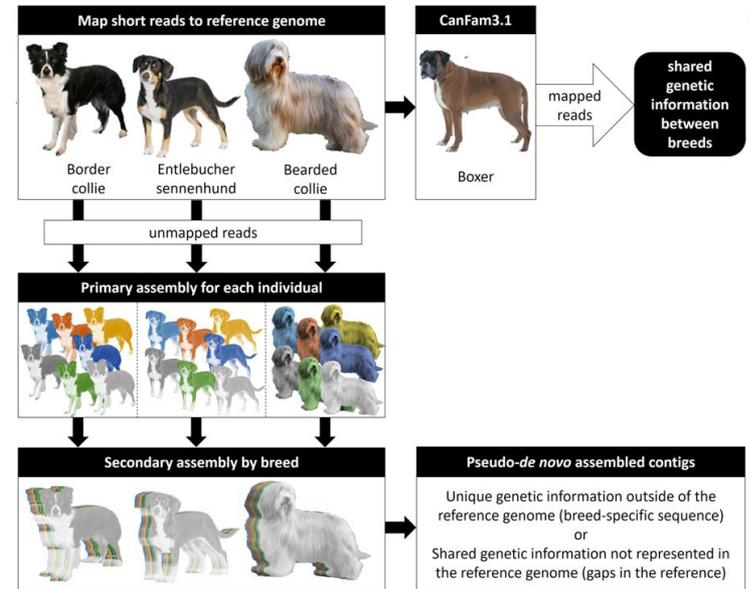
Article | [OPEN](#) | Published: 30 June 2016

Long-read sequencing and *de novo* assembly of a Chinese genome

Lingling Shi, Yunfei Guo [...] Kai Wang 

Article | [OPEN](#) | Published: 18 July 2018

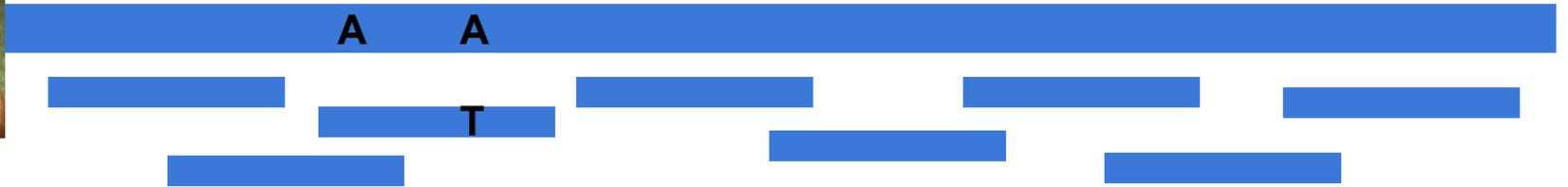
Assembly and Analysis of Unmapped Genome Sequence Reads Reveal Novel Sequence and Variation in Dogs



Population reference genomes

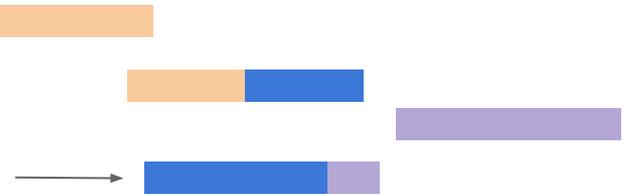


Boxer reference

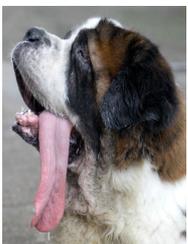


Don't align, generally discarded

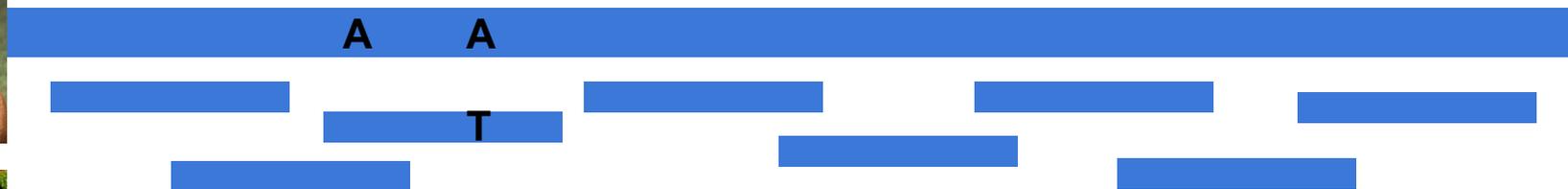
Might align, but won't help determine breed



Population reference genomes

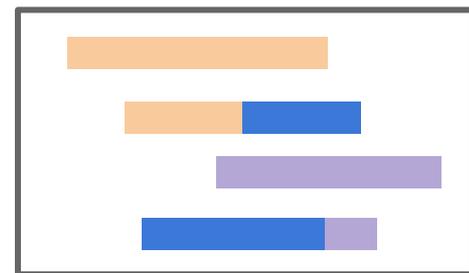


Boxer reference



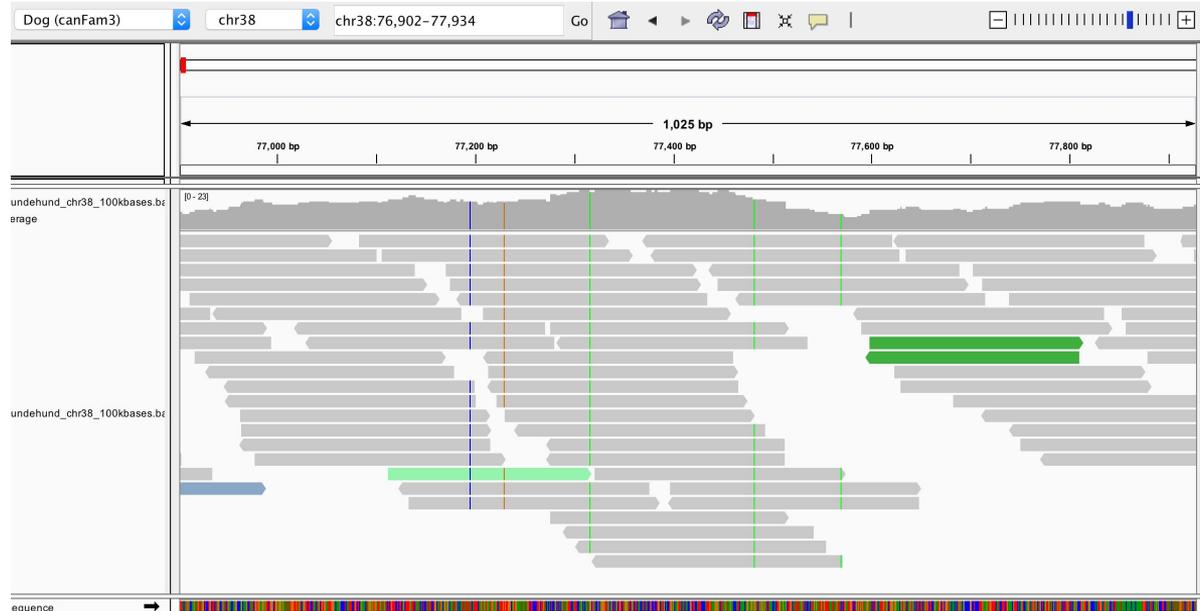
Pseudo-*de novo* assembled contigs

Unique genetic information outside of the reference genome (breed-specific sequence)



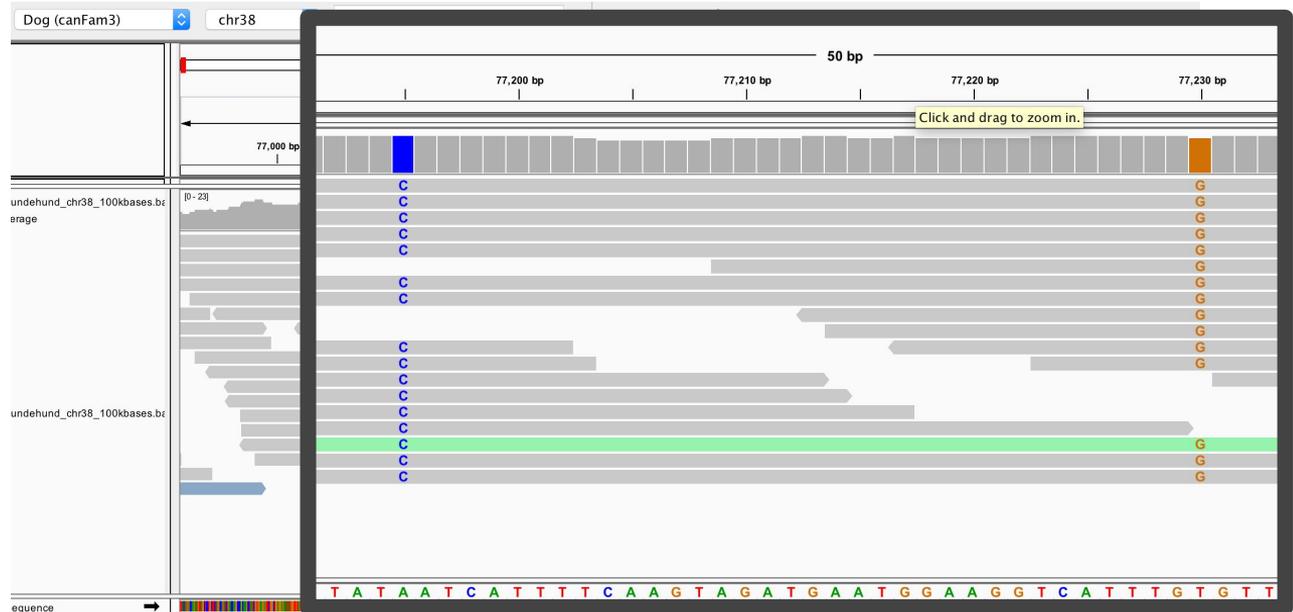
Mini Assignment: Using IGV (Integrative Genomics Viewer)

IGV is a tool that lets you visualize alignments from second or third generation sequencing to a reference genome



Mini Assignment: Using IGV (Integrative Genomics Viewer)

IGV is a tool that lets you visualize alignments from second or third generation sequencing to a reference genome



Logistics

For those of you who are finished with the project, you'll start exploring second generation sequencing data, aligned to Tasha's genome, guided by the IGV exploration worksheet on Piazza. Otherwise, keep working on part 2/3!

Friday: Mini lecture where as a class we'll recap what we've learned, how you go from a mutt to it's breed makeup, and go over what breeds Clarence, Reilly, and Finch actually are.