“Analysis of a pan-genome from deep sequencing of 910 humans of Africa descent”, (Sherman et al., 2019)

Presentation to Statistical Genetics Seminar

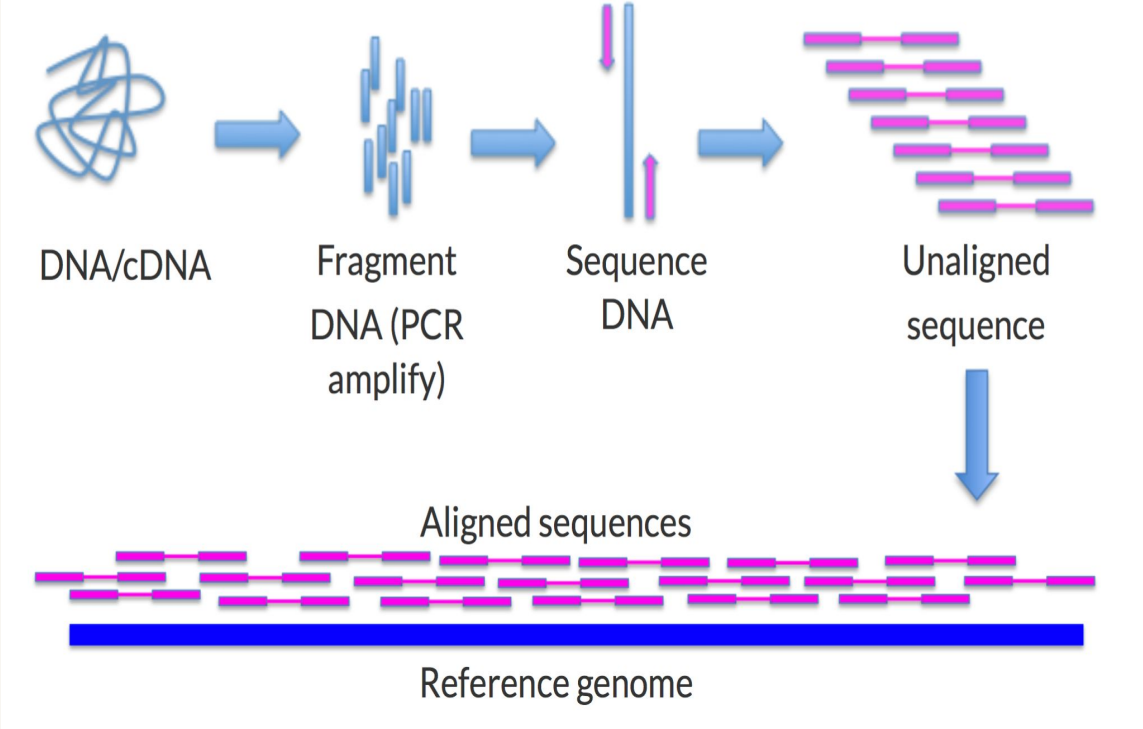
Seth Temple and Zorian Thornton

Background

2nd generation sequencing, sequence alignment, reference & pan-genomes.

What is a reference genome?

- A genome assembly that represents a single, idealized “template” genome for an organism in a species
- Built using bits of DNA sequences from multiple individuals in a group/clade
  - A *“mosaic”* of genetic information
- GRCh38 is most current build, released in 2014
  - GRC - Genome Reference Consurtium
  - h - human genome
  - 38 - assembly ID

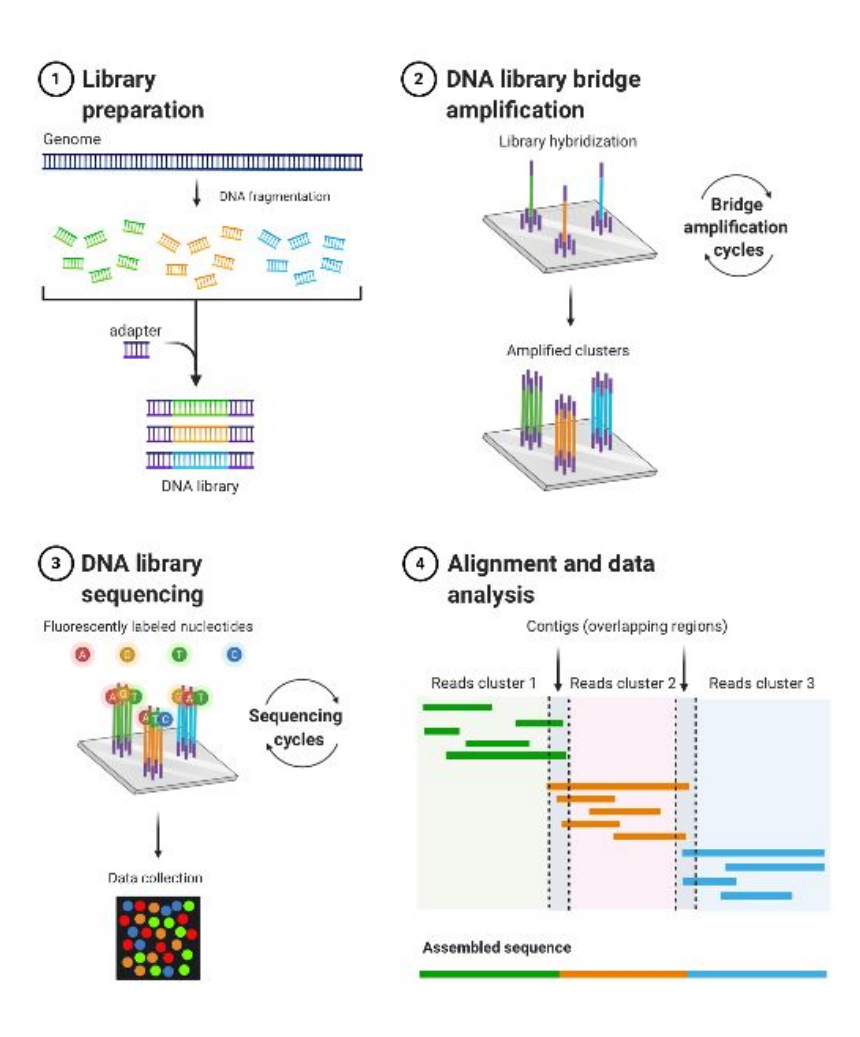


Source: Genomic Data course at Rockefeller University, https://rockefelleruniversity.github.io/Genomic_Data/

How is sequencing done?

Second generation sequencing

- Massively parallel sequencing technology that enables high-throughput sequencing at scale.
- A gross oversimplification:
  1. Break DNA up into small chunks
  2. Attach adapters to small chunks of DNA and attach to glass plate
  3. Make a lot of copies of DNA strands
  4. Attach fluorescent nucleotides to the DNA strands, and take pictures of the wells

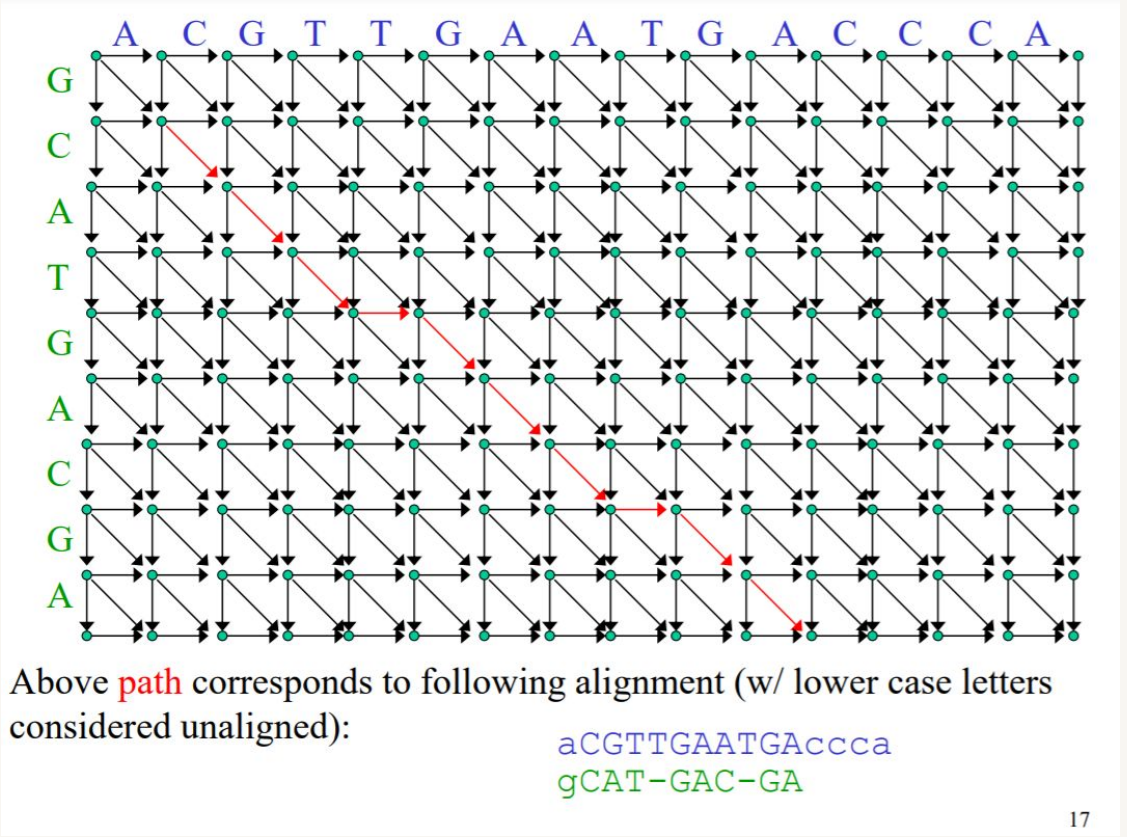


Reprinted from "Next Generation Sequencing (Illumina)", by BioRender, June 2020

Aligning similar sequences

Sequence alignment

- Following sequencing, we have a lot of small chunks of DNA but need to put them in the proper order
- Sequence alignment finds overlapping sequences (allowing for some mistakes) among these reads
- This process allows us to align smaller reads to longer sequences

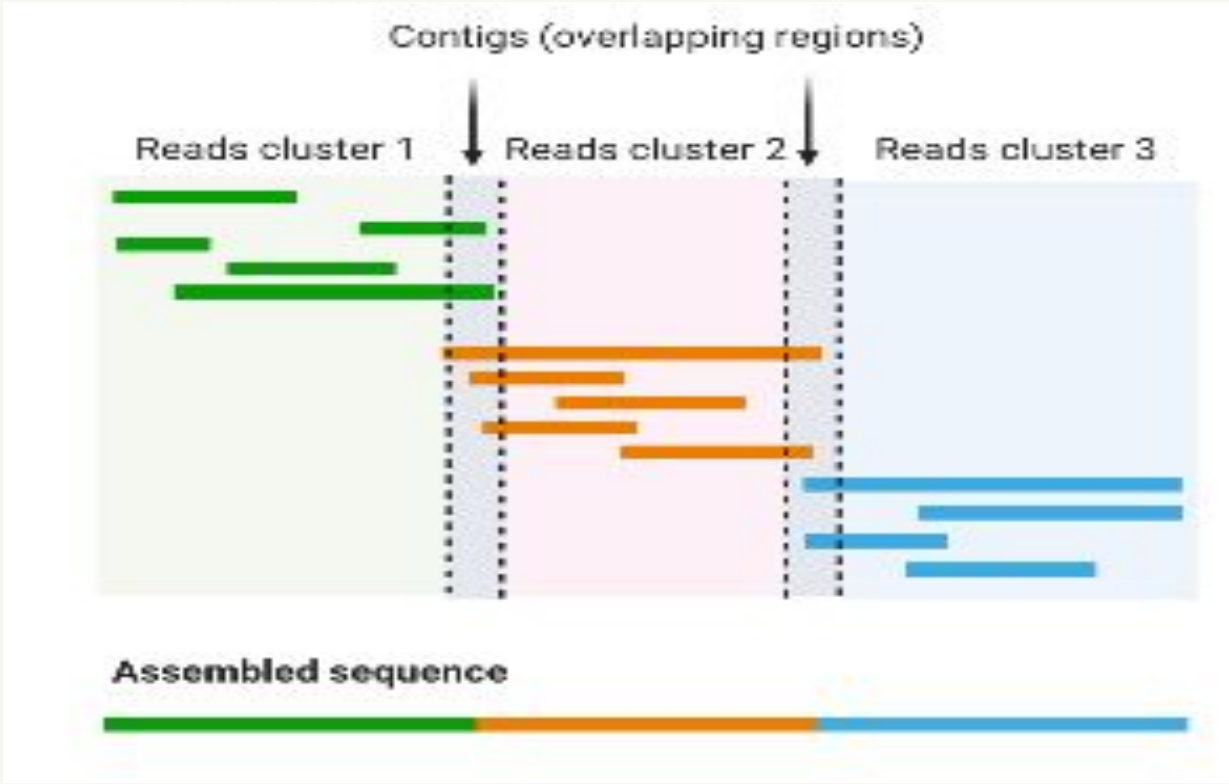


Source: Phil Green's GENOM 540 winter course

Putting the pieces together

Genome assembly

- Sequence assembly is the process of consolidating these smaller chunks of DNA into a longer sequence
- “Align” sequence overlaps to create *“contigs”*, longer stretches of continuous DNA
- Put contigs together to assemble full sequence



Reprinted from "Next Generation Sequencing (Illumina)", by BioRender, June 2020

What is a pan-genome?

- A collection of all DNA sequences that occur in a species.

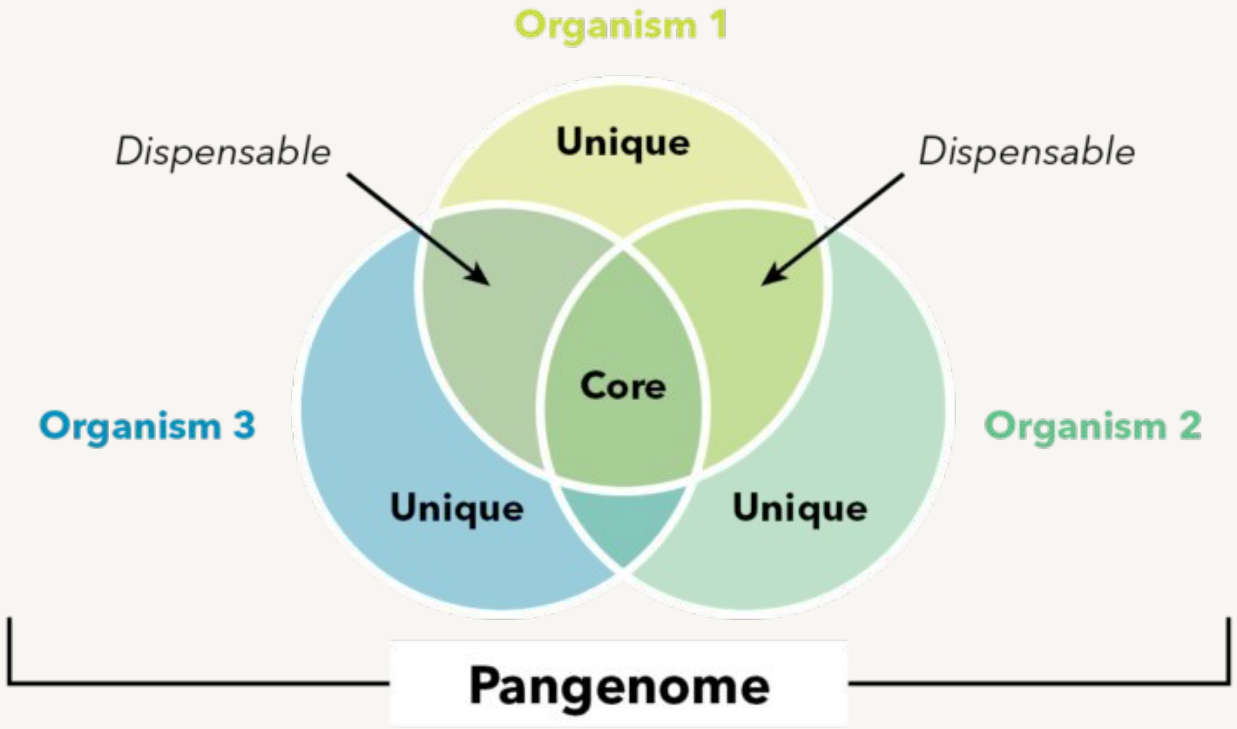


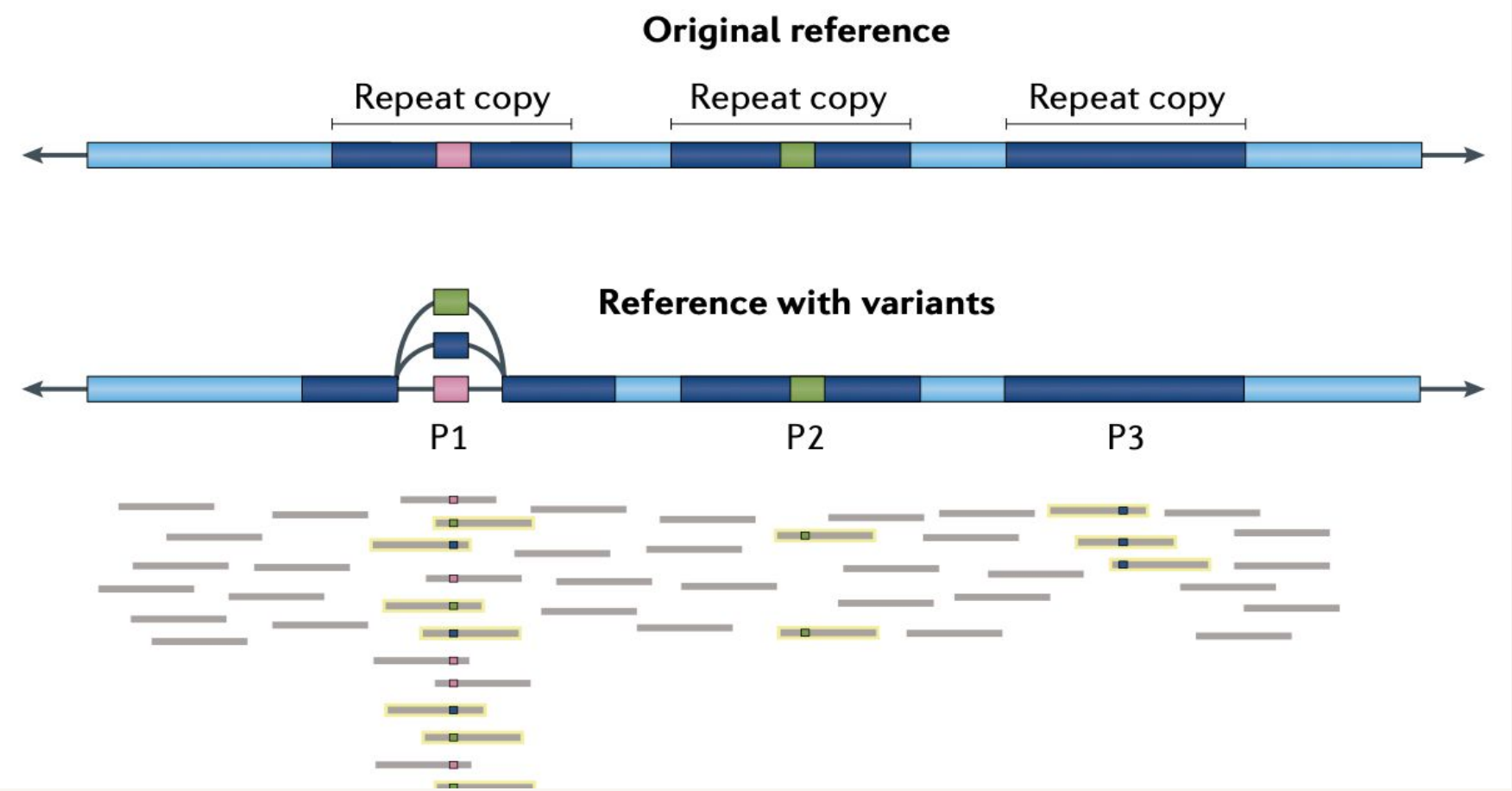
Figure from PacBio
https://www.pacb.com/blog/sequencing101pangenome/

Why are pan-genomes useful?

- Most common approach for including divergent sequences in genetic analysis is to include them in the alignment to the reference.
- A few drawbacks claimed by Sherman & Salzburg:
  - Most popular aligners are not designed to handle variation and tend to just treat these variants as repeats
  - Space and time complexity of storing and searching the reference becomes an issue when adding divergent sequences
  - Alignment ambiguity increases as variant number grows
  - Including these sequences alone does not accurately represent underlying biology
- Capture vast amounts of genetic variation in a population

Sherman & Salzburg, *Nature Reviews, 2020*

Numerous variants induce alignment ambiguity



Reference vs. pan- genome

Reference genome

- A genome assembly that represents a single, idealized “template” genome for an organism in a species
- Built using the DNA of multiple individuals from a group/clade
  - A **“mosaic”** of genetic information
- GRCh38 is most current build, released in 2014

Pan-genome

- The union of all sequenced genomes in a given species clade.
- Also built using the DNA of multiple individuals from a group/clade
  - A **“sorting-bin”** of genetic information
- Hierarchy of shared genetic information:
  - Core pan-genome - genes present in all individuals
  - Cloud pan-genome - genes present in 2 or more individuals
  - “Accessory genome” - singletons

Methods

These are admixed!

In a supplementary analysis, the authors conclude that the contigs are more representative of African ancestry than European ancestry.

Can we conclude that this pan-genome is an "African" pan-genome?

Supplementary Table 6 | Cohorts of CAAPA samples.

Cohort	Number of Samples
African American (Atlanta)	50
African American (Baltimore-DC)	50
African American (Chicago)	50
African American (Detroit)	50
African American (Jackson, MS)	50
African American (Nashville)	48
African American (NYC)	48
African American (San Francisco)	50
African American (Winston-Salem)	50
Barbados	49
Brazil	47
Colombia	50
Dominican Republic	47
Gabon	34
Honduras	50
Jamaica	50
Palenque	34
Nigeria	50
Puerto Rico	53

Data was collected from 19 distinct cohorts across the Americas, the Caribbean, and Africa resulting in 910 analyzed samples.

Their bioinformatics pipeline

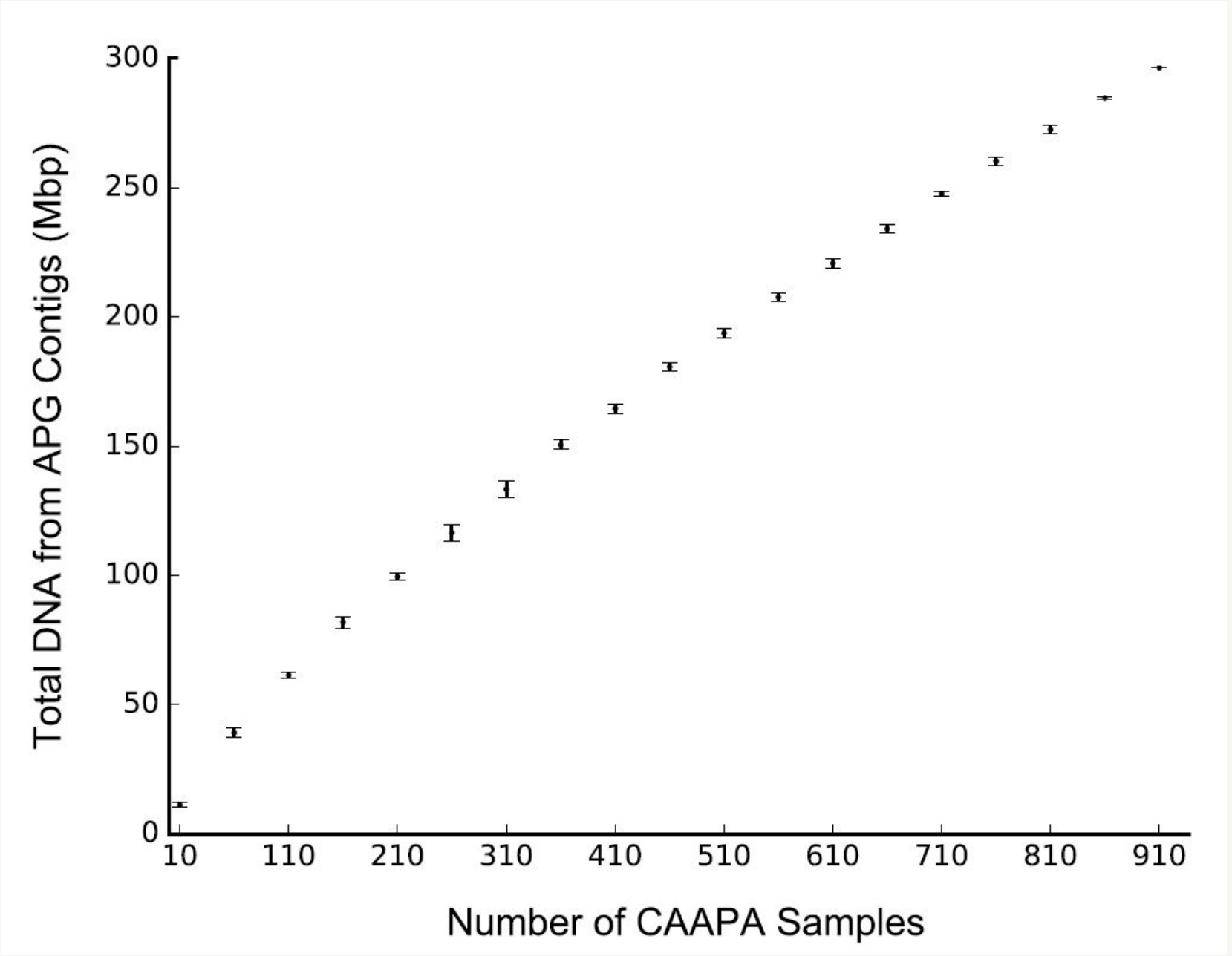
Three main steps:

1. Align all reads against GRCh38 & assemble reads that don't align with GRCh38 into contigs
2. Take non-aligned contigs from step 1, and try to align to reference
3. Cluster similar sequences together

Overall, this approach enables the discovery of 10% more DNA in the "African pan-genome" than GRCh38, a good deal of which appear within protein-coding genes.

Subsample (and bootstrap) to discover trend for APG contig discovery

Any hypotheses for what this figure says about the APG contigs and the variation they describe?



Results

Interesting statistics

- **296.5 Mb** of novel DNA distributed across **125,715 sequences**
- Resolved location of 1,548 APG contigs
  - 387 intersect known genes
- 42,207 APG contigs (120.7M Mb) aligned to Korean or Chinese assembly
- Subsampling method suggests a linear trend between APG contigs and sample size

Table 1 | Novel sequences in the African pan-genome

	Number of sequence contigs	Total length (bp)	Bases with no alignment to GRCh38 (<80% identity)	Longest contig (bp)
Two ends placed	302	667,668	431,656	20,732
One end placed	1,246	3,687,028	1,866,699	79,938
Unplaced	124,167	292,130,588	202,629,979	152,806
Total	125,715	296,485,284	204,928,334	152,806
Non-private only	33,599	80,098,092	50,044,650	152,806

Table 2 | African pan-genome contig presence/absence statistics

	Number of contigs	Mean number of insertions per individual	Mean number individuals per insertion
Two ends placed	302	120 (39.7%)	363 (of 910)
One end placed	1,246	212 (17.0%)	155 (of 910)
Unplaced	124,167	527 (0.4%)	4 (of 910)
Total	125,715	859 (0.7%)	6 (of 910)
Non-private only	33,599	758 (2.2%)	21 (of 910)

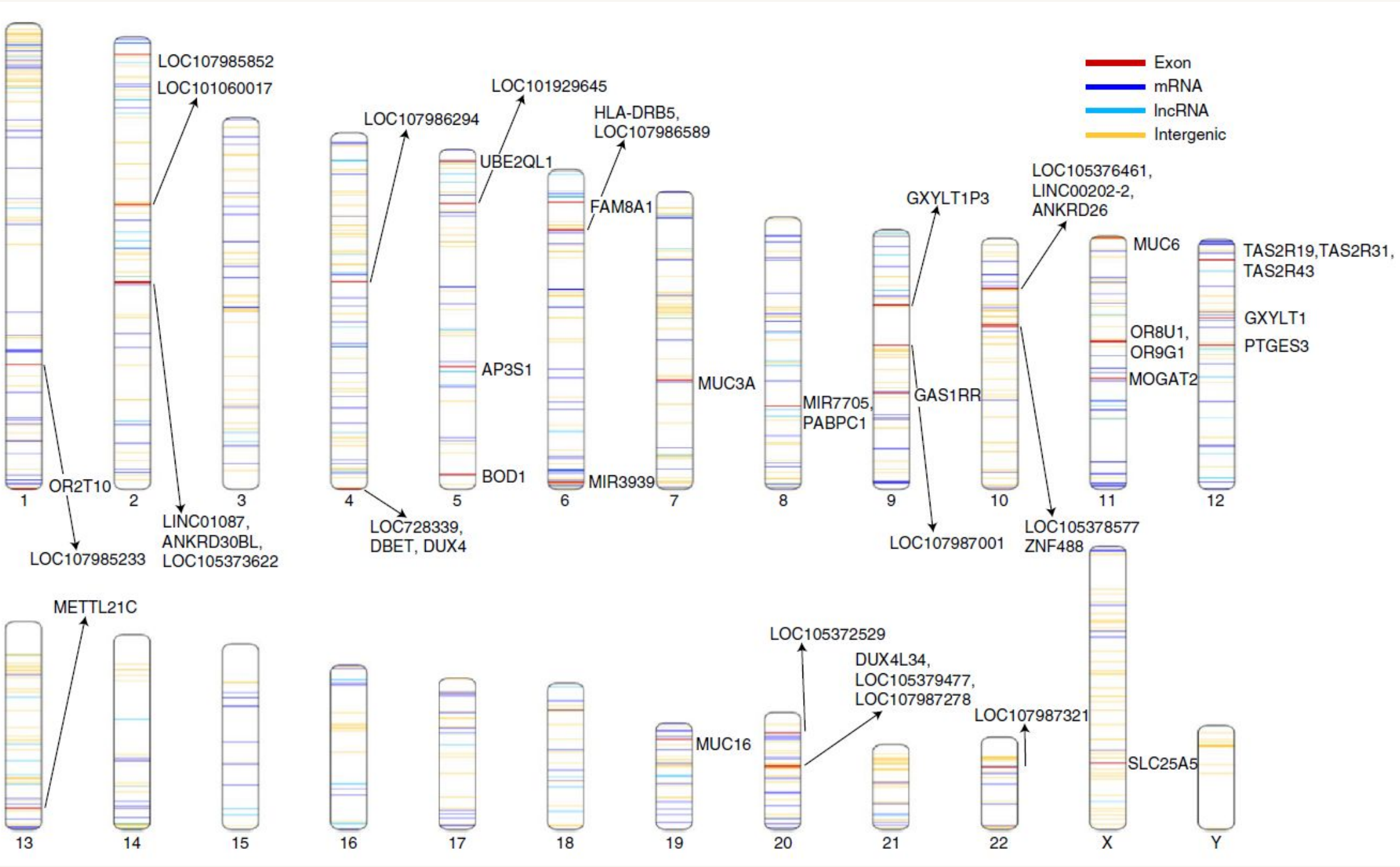

Exon
mRNA
lncRNA
Intergenic
LOC107985852
LOC101060017
OR2T10
LOC107985233
LINC01087,
ANKRD30BL,
LOC105373622
LOC107986294
LOC101929645
UBE2QL1
HLA-DRB5,
LOC107986589
FAM8A1
AP3S1
BOD1
MIR3939
LOC728339,
DBET, DUX4
MUC3A
MIR7705,
PABPC1
GXYLT1P3
GAS1RR
LOC107987001
LOC105376461,
LINC00202-2,
ANKRD26
LOC105378577
ZNF488
MUC6
OR8U1,
OR9G1
MOGAT2
TAS2R19,TAS2R31,
TAS2R43
GXYLT1
PTGES3
METTL21C
MUC16
LOC105372529
DUX4L34,
LOC105379477,
LOC107987278
LOC107987321
SLC25A5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

Table 3 | Comparison of African pan-genome contigs to the Chinese and Korean genomes

	Best GRCh38 alignment is 80–90% identical with 50–80% coverage		Best GRCh38 alignment is <80% identical or <50% coverage		Total	
	Contigs	Length (bp)	Contigs	Length (bp)	Contigs	Length (bp)
Matches Chinese only	1,625	2,898,106	7,607	25,475,277	9,232	28,373,383
Matches Korean only	2,242	3,989,277	15,635	48,642,664	17,877	52,631,941
Matches both	5,385	9,720,662	9,713	29,981,048	15,098	39,701,710
Total	9,252	16,608,045	32,955	104,098,989	42,207	120,707,034

Contigs with a better alignment to the Chinese or Korean assemblies than to GRCh38. Alignments to the Chinese and Korean assemblies were required to have ≥90% identity and ≥80% coverage to be considered. Lengths shown are the sums of the contig lengths, not the alignment lengths.

Questions

Interesting questions

- Reference gaps:
  - Why have we not filled in the gaps in GRCh38?
  - To what extent have long reads addressed this issue?
  - Does the African pan-genome address this issue?

- Multiple reference genomes:
  - What are the advantages of multiple reference genomes? This study found that APG contigs aligned better to East Asian references.
  - How would we incorporate data from many populations if they are aligned to different reference genomes?
  - How were the Chinese and Korean reference genomes assembled?

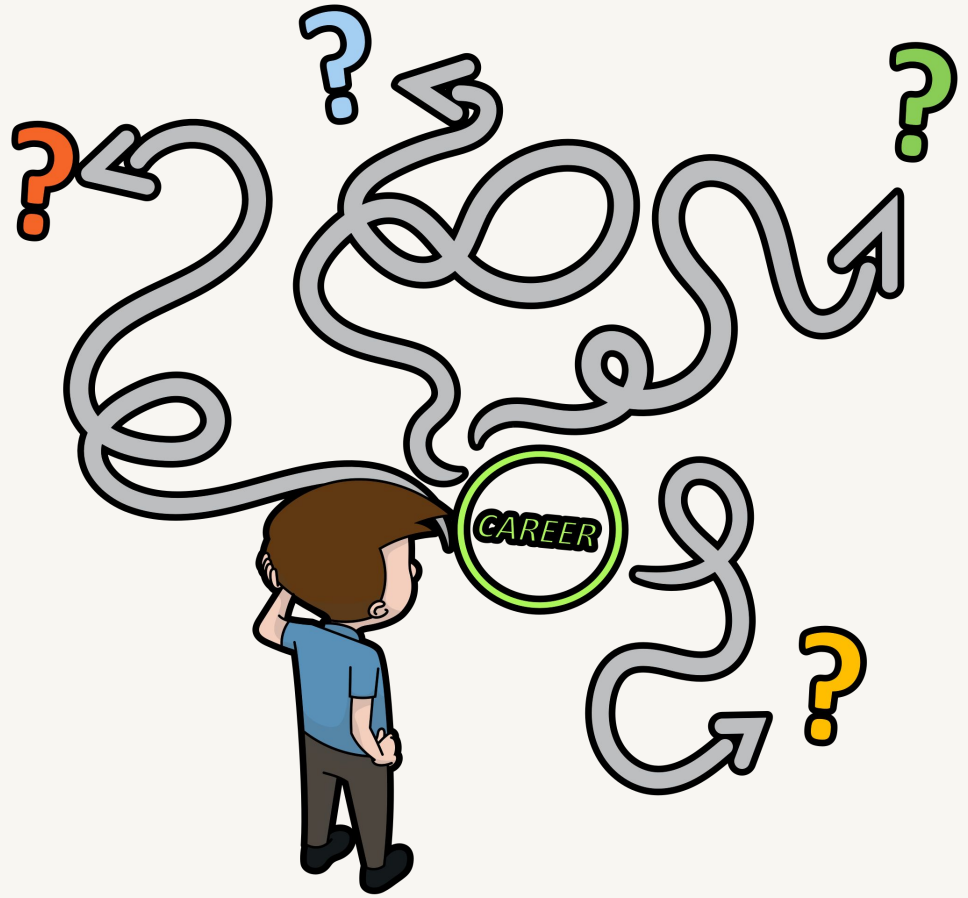
More interesting questions

- Biological relevance:
  - How do we interpret these findings, say ~ 300 Mb novel genetic material, given what we know about the underlying biology?
  - Are these APG contigs just different alleles that did not align to the reference genome?

- Applications:
  - How can we use the pan-genome to find variation of biological importance?
  - How much of this novel genetic material is due to singleton variation (likely structural)? Can we make use of singleton variation?

Increasingly speculative questions

- How long is the human genome?
- Is it time to change the reference genome?
- Would the linear trend in Supplementary Figure 4 continue indefinitely if we collected more samples?



The End