The Challenges and Opportunities of Reference Free Genomic Data Analysis

Ted Kalbfleisch PhD

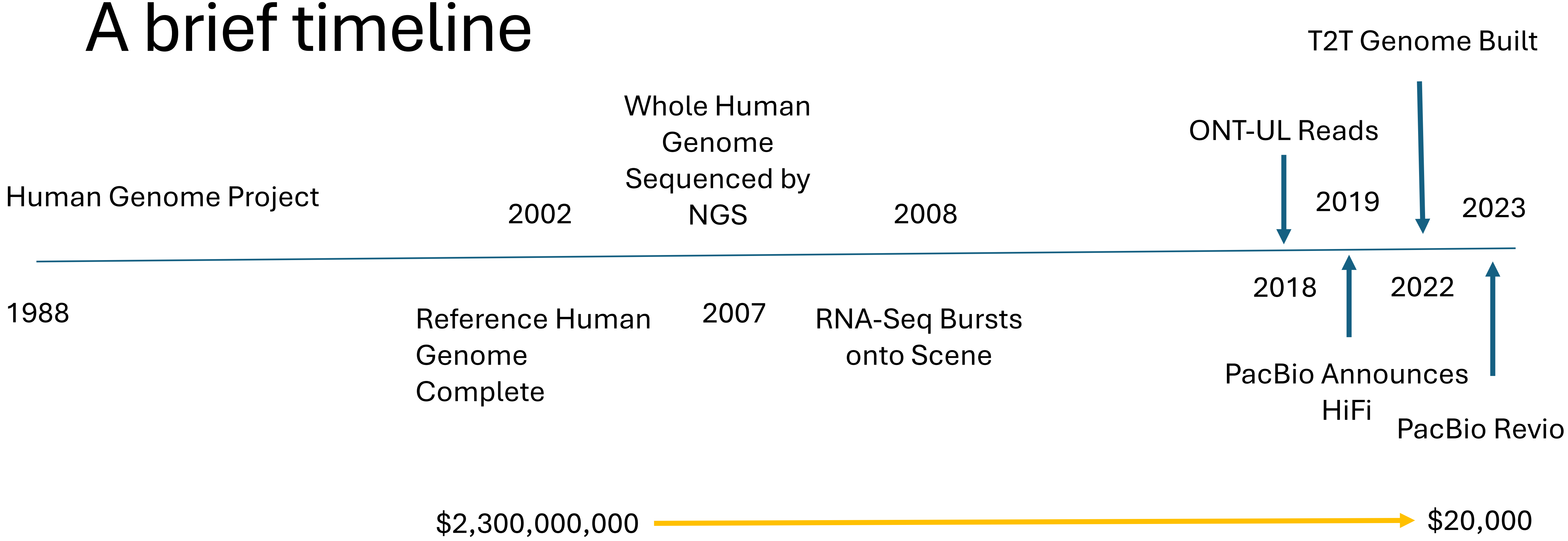
Associate Professor

Department of Veterinary Science

Martin-Gatton College of Agriculture, Food, and Environment

University of Kentucky

A brief timeline



Human Genome Project

1988

2002

Reference Human Genome Complete

Whole Human Genome Sequenced by NGS

2007

2008

RNA-Seq Bursts onto Scene

ONT-UL Reads

2018

2019

PacBio Announces HiFi

T2T Genome Built

2022

2023

PacBio Revio

$2,300,000,000 → $20,000

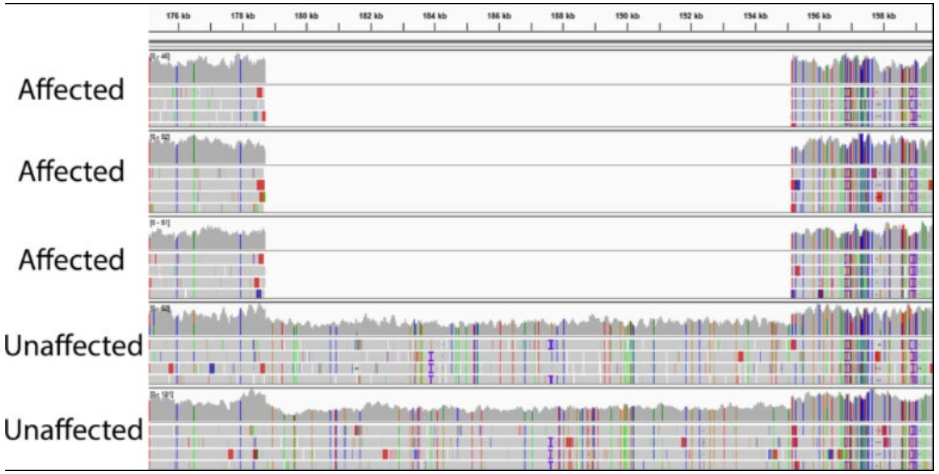
- We built genomes!
- We cataloged variation in species and between individuals
- We learned in detail and without bias how transcriptomes changed cell-type to cell-type and condition to condition

We made compromises

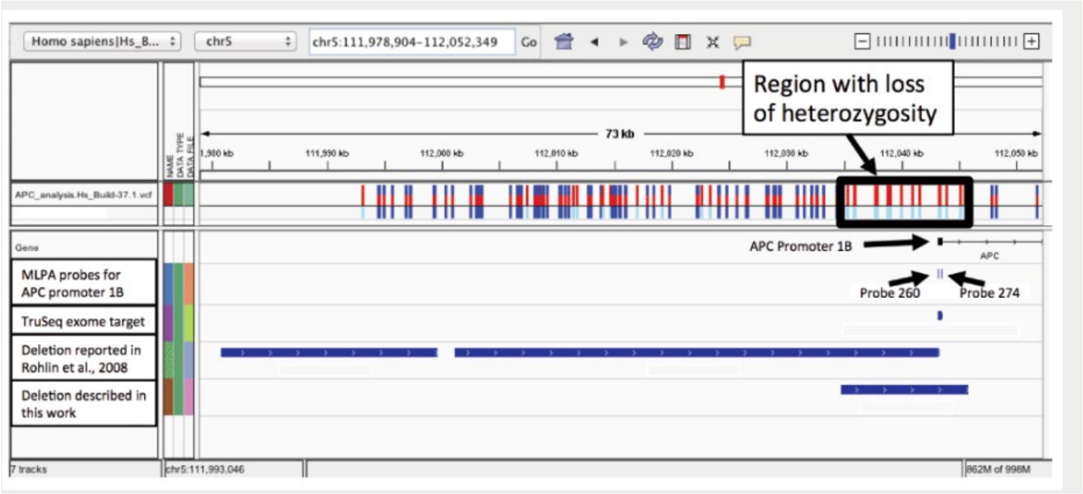
- Pseudo-haploid representation and interpretation of the genome
- Copy number variation, and haplotypes inferred as opposed to directly observed
- Why is this a big deal?
  - Biology works with the haplotypes.
  - Alleles on the same haplotype may have additive effects
  - Epigenetic imprinting in some genes results in different functions for maternal copies vs. paternal copies of a gene

But we were happy!

- Much of the genome that interests us is unique, readily assembled, and well characterized with short read data!



Hisey EA, Hermans H, Lounsberry ZT, Avila F, Grahn RA, Knickelbein KE, Duward-Akhurst SA, McCue ME, Kalbfleisch TS, Lassaline ME, Back W, Bellone RR. Whole genome sequencing identified a 16 kilobase deletion on ECA13 associated with distichiasis in Friesian horses. BMC Genomics. 2020 Nov 30;21(1):848. doi: 10.1186/s12864-020-07265-8. PMID: 33256610; PMCID: PMC7706231.



Kalbfleisch T, Brock P, Snow A *et al.* Characterization of an APC Promoter 1B deletion in a Patient Diagnosed with Familial Adenomatous Polyposis via Whole Genome Shotgun Sequencing [version 1; peer review: 2 approved]. *F1000Research* 2015, **4**:170 (https://doi.org/10.12688/f1000research.6636.1)

Bioinformatics retreat I once attended:

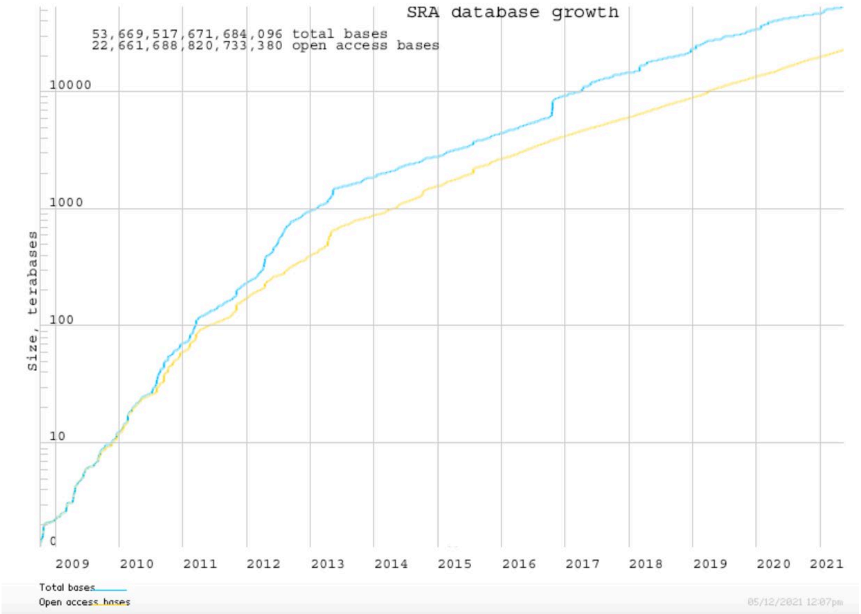
We could stop everything else we are doing right now and spend the next 10 years just analyzing the data we have.

--Dr. Matt Roth

CuraGen Corporation, New Haven, CT

December 1996

What have we accomplished in the last 26 years?
1996-2010, A golden age for metaphors and memes



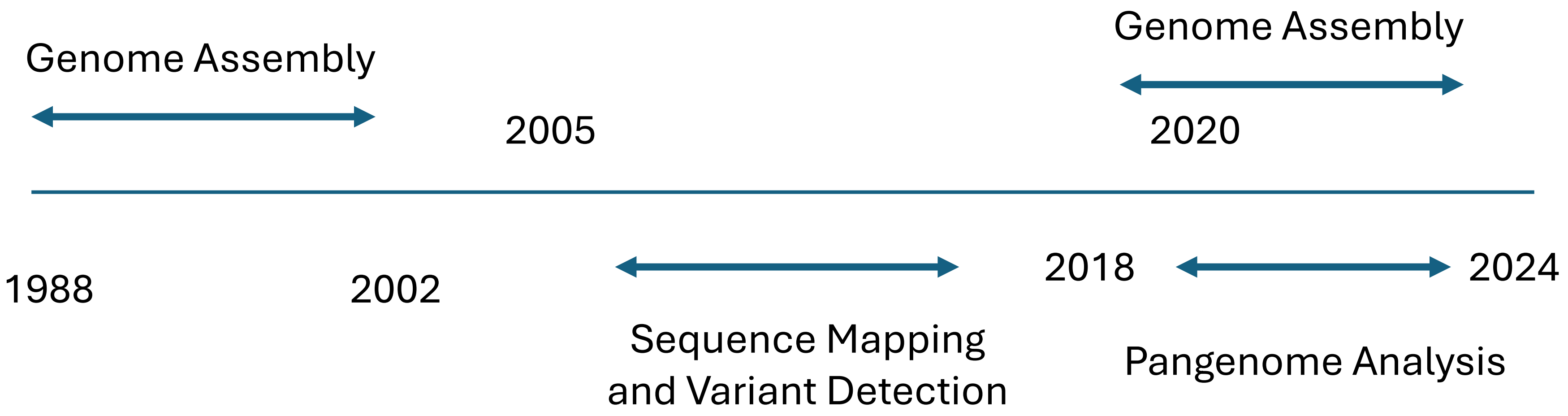
- Not enough:
  - Computing power
  - Band width
  - Storage
  - Good standards-based software
  - Expertise/Programmers/Database Developers
  - Money

Historically, what has been the role of RCD

Research

Computing ☑

Data ?



We don’t manage data well.

F indable

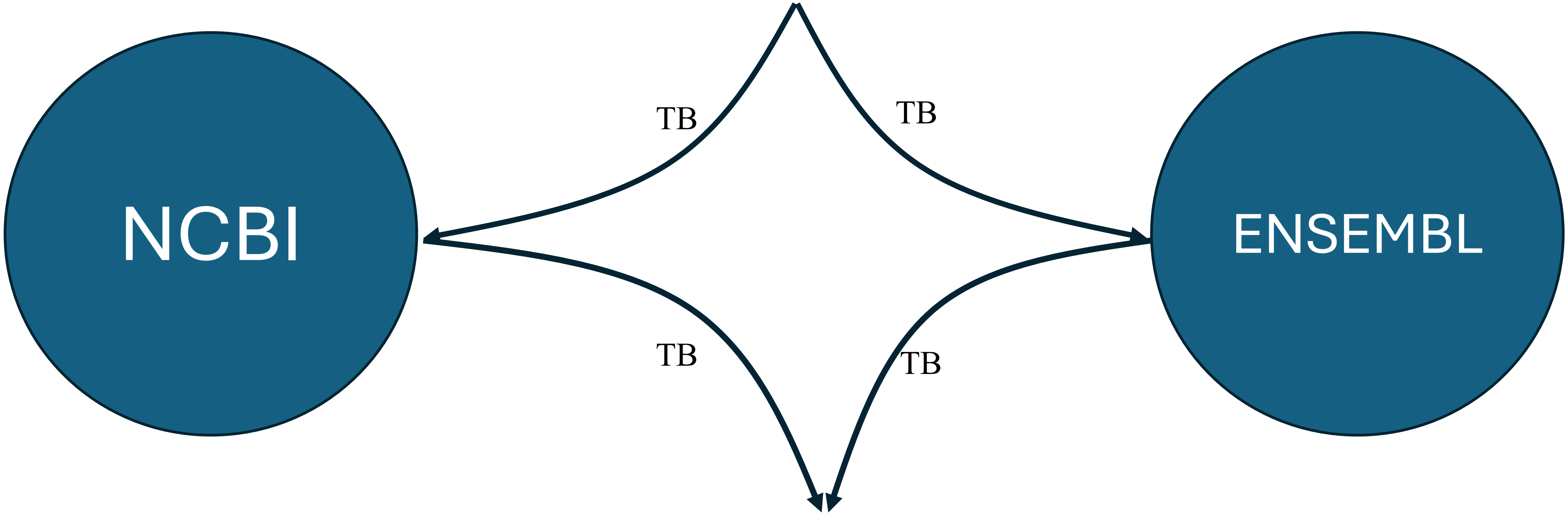
A ccessible

I nteroperable

R euseable

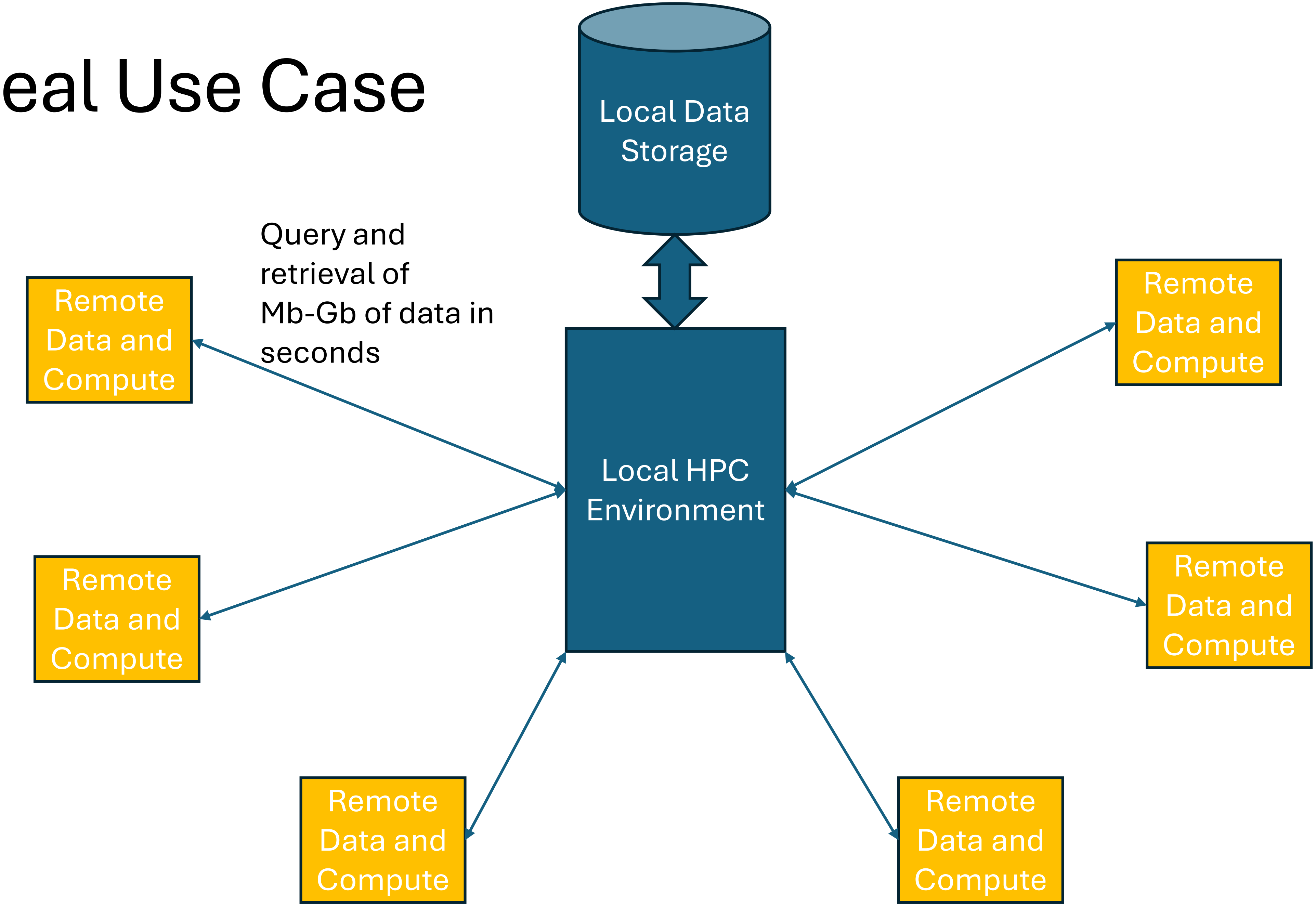
Data Management in the Life Sciences: Boil the Ocean

Raw, unprocessed sequence data

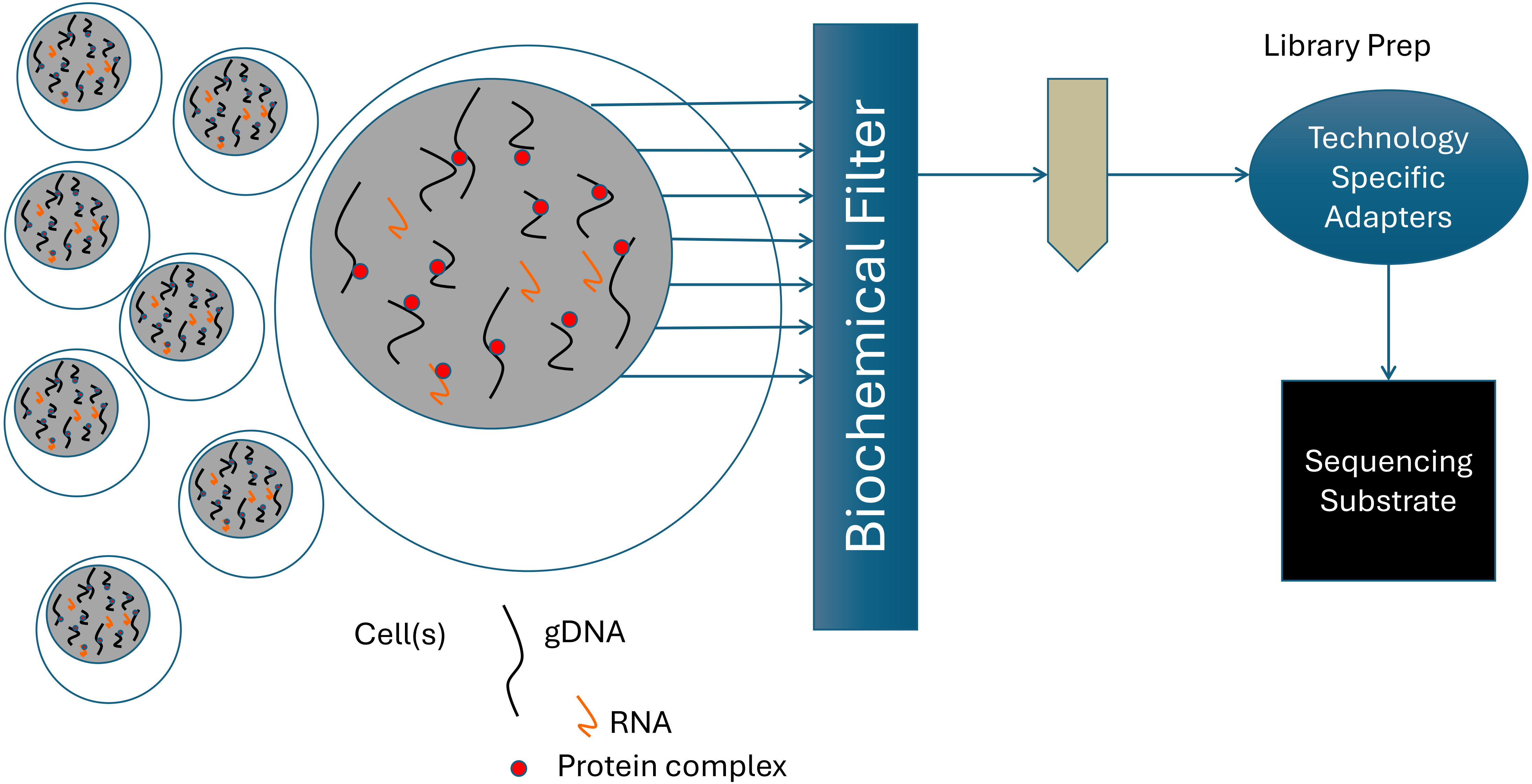


Download in whole and reprocess (weeks)!

Ideal Use Case



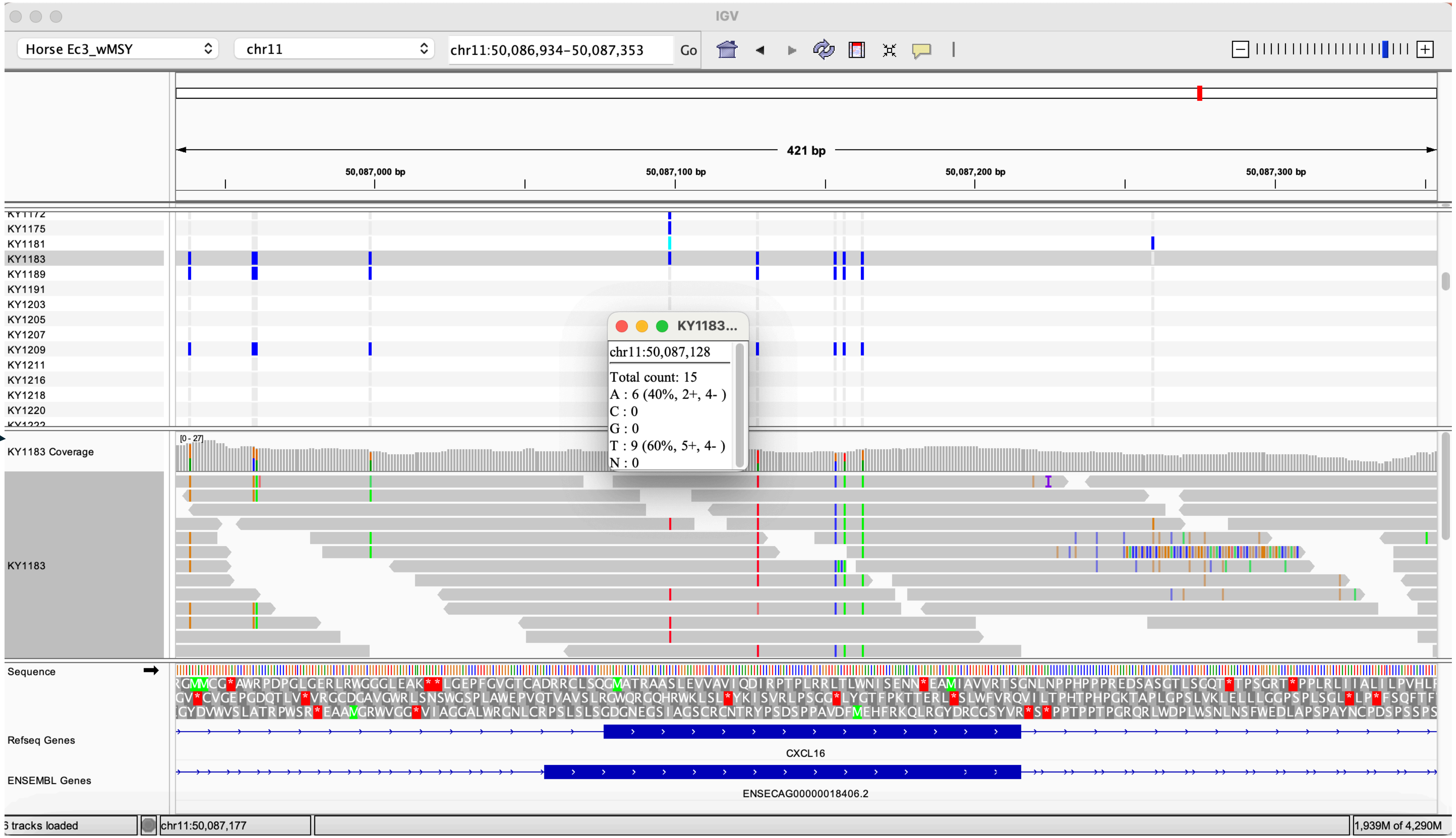
Unbiased sequencing



Why Do We Have Reference Genomes?

- Genomic context

```
@A00744:400:HHLN5DSX3:1:1101:1579:1000 1:N:0:GGACTTGG+CGTCTGCG
NAAGTCTCAGCAAGAGCTCAGCACACCTGTGTGGGCCAGGGAGCGCTACTGAGATCTGACCAAACCAGGAGAGACCTGTTCAC
GAGGGACAGGAGCTCAGACATGCCAAAGGGACCCGATACTTCCCCGTGGATCTCGTGAGCCATCCACA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF
@A00744:400:HHLN5DSX3:1:1101:2953:1000 1:N:0:GGACTTGG+CGTCTGCG
NAAACAGCAGTCTTTATACAAAGCTCAAAGATATTCAGATACCCTAAATATAGCCTGCAAAATATGCACACACACATGCACAC
ACACAGGCACACACACAGGCACCTTTGGATGCATGCCAGACCTGGTGAAACTACTCATGGTTTTCCAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFF
@A00744:400:HHLN5DSX3:1:1101:3766:1000 1:N:0:GGACTTGG+CGTCTGCG
NGGGTTTTGAGTCAGAAACACCTGCCTTTAGGTCCCACTTCTACCCTGTACTGGTTTTGTGACTCATGAGCTACTCTCACGGT
GCACATTTTAGGATACGTGACATCTTTCAATATCACACAGTTAAGAAGATTCAAAGCCCAGATTTGAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFF:F,FF:FFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF,FFFFF
```



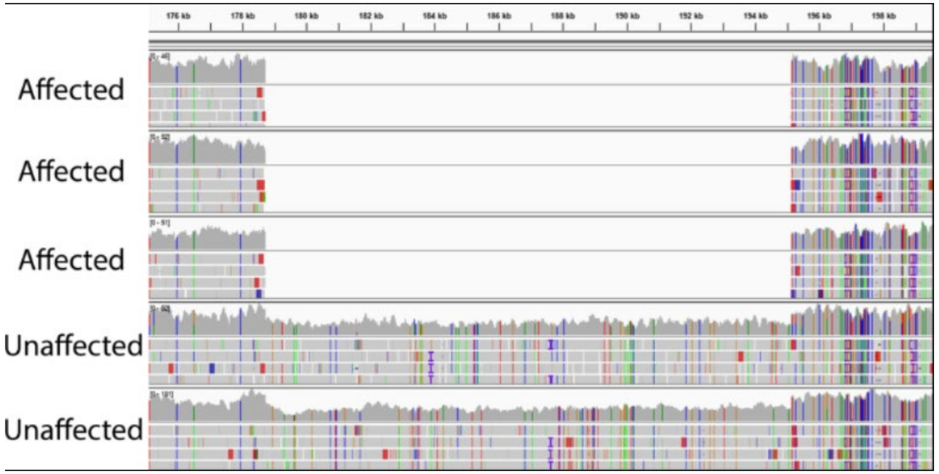
KY1183...

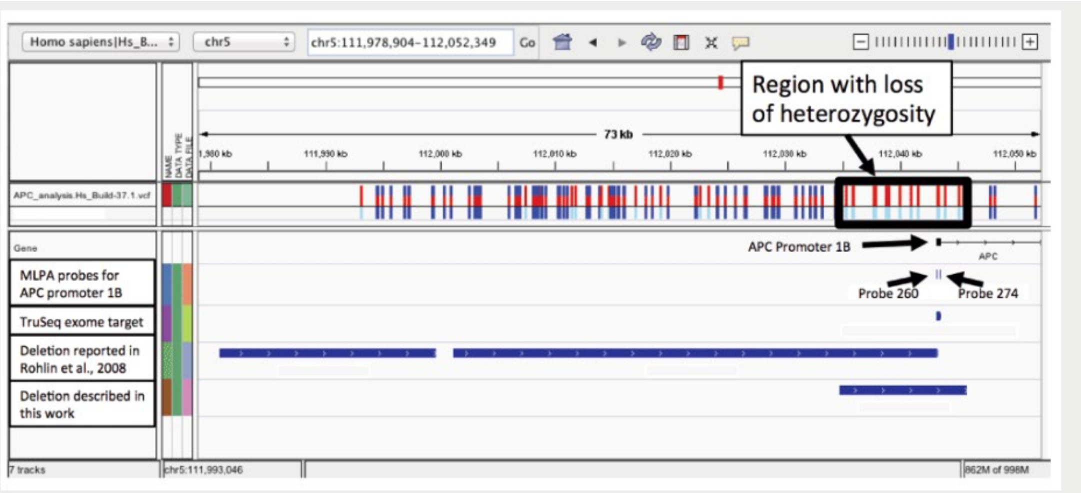
chr11:50,087,128

Total count: 15
A : 6 (40%, 2+, 4-)
C : 0
G : 0
T : 9 (60%, 5+, 4-)
N : 0

What are the Shortcomings of this Approach?

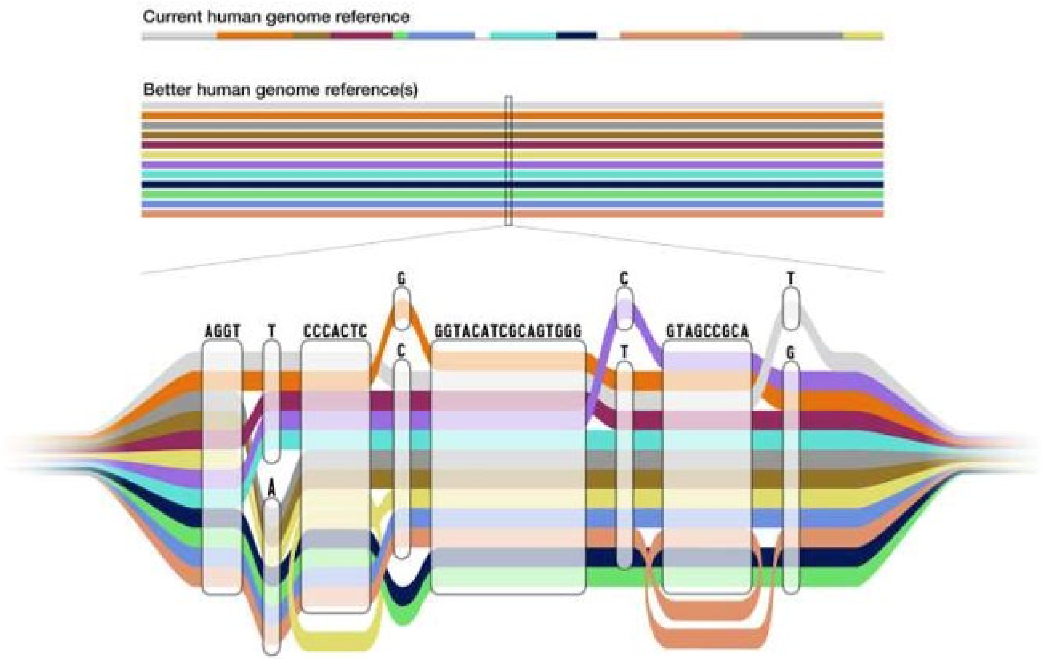
- The reference is generated from a single individual, and no single genome is comprehensive





The Approach that is Emerging:

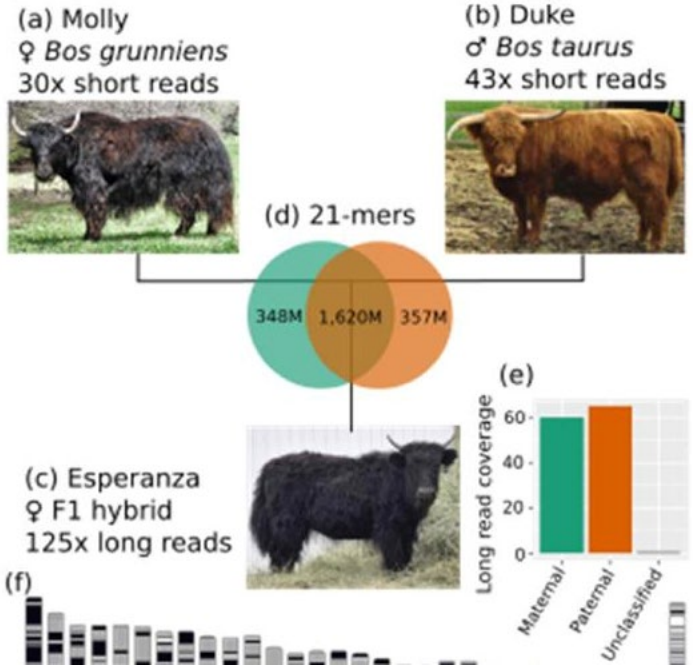
Pangenomes!

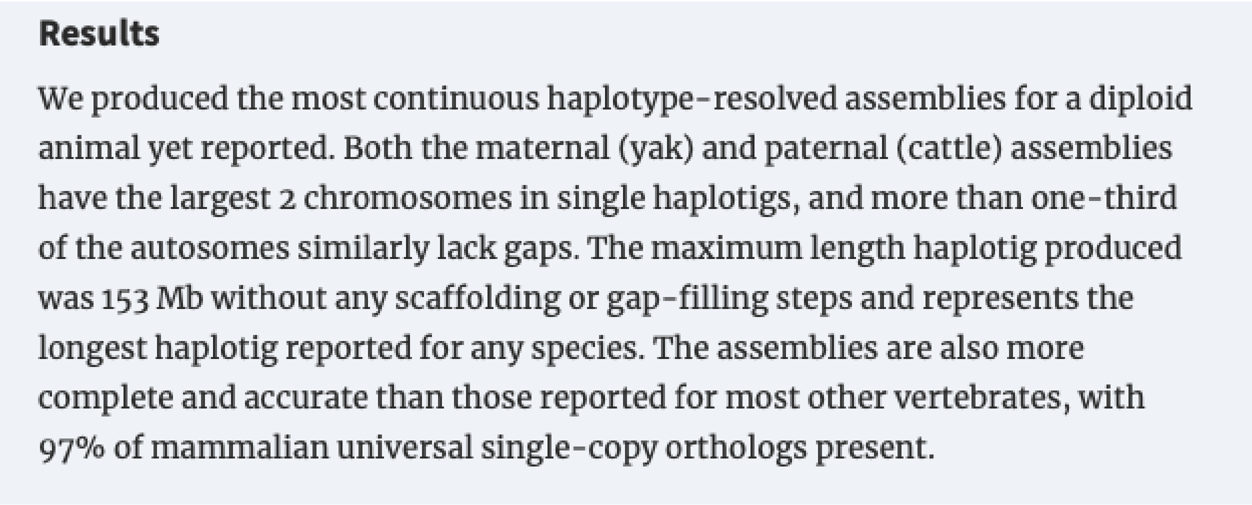

Current human genome reference
Better human genome reference(s)
G
C
T
AGGT
T
CCCACTC
GGTACATCGCAGTGGG
GTAGCCGCA
C
T
G
A

The new draft pangenome reference contains 47 genomes instead of just one and will provide a much better point of comparison than the traditional reference to find and understand the differences in our DNA. Credit: National Human Genome Research Institute

https://phys.org/news/2023-05-human-pangenome-enable-equitable-genomic.html

A great idea that worked out very well



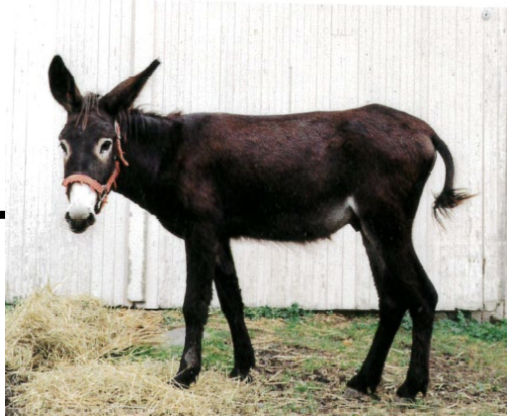


Results

We produced the most continuous haplotype-resolved assemblies for a diploid animal yet reported. Both the maternal (yak) and paternal (cattle) assemblies have the largest 2 chromosomes in single haplotigs, and more than one-third of the autosomes similarly lack gaps. The maximum length haplotig produced was 153 Mb without any scaffolding or gap-filling steps and represents the longest haplotig reported for any species. The assemblies are also more complete and accurate than those reported for most other vertebrates, with 97% of mammalian universal single-copy orthologs present.

Edward S Rice, Sergey Koren, Arang Rhie, Michael P Heaton, Theodore S Kalbfleisch, Timothy Hardy, Peter H Hackett, Derek M Bickhart, Benjamin D Rosen, Brian Vander Ley, Nicholas W Maurer, Richard E Green, Adam M Phillippy, Jessica L Petersen, Timothy P L Smith, Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle, *GigaScience*, Volume 9, Issue 4, April 2020, giaa029, https://doi.org/10.1093/gigascience/giaa029

Trio sequencing a mule



Less than a day on the UK HPC

30X Illumina Short read data on both sire and dam

Thoroughbred dam x Donkey sire

2.5Gb Donkey contig N50: 35.6 Mb
2.6Gb Thoroughbred contig N50: 43.5 Mb

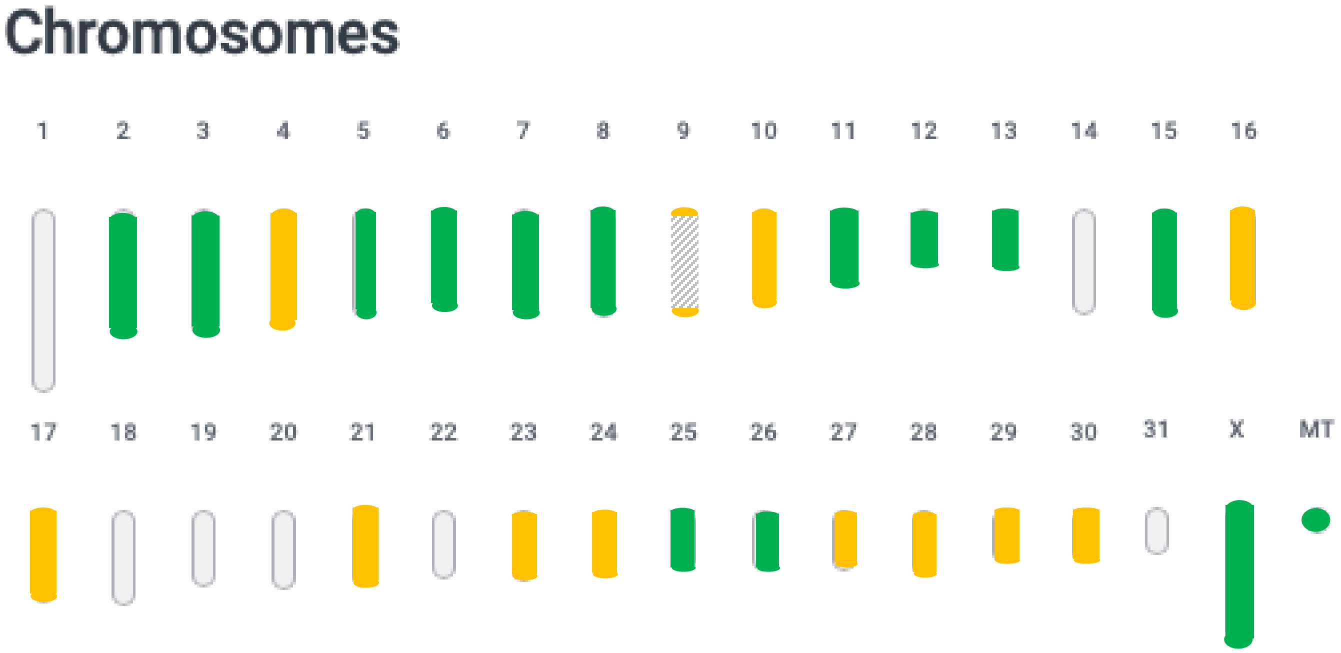
Less than a day on the UK HPC

No Illumina Short read data on either the sire or dam

Thoroughbred dam x Donkey sire

5.2Gb Contigs, contig N50: 38.3 Mb

Current Status of Horse/Donkey T2T Effort



🟩 T2T, gapless

🟨 T2T, with gaps

⬜ Telomere on
one end

UofL (PacBio)
Elizabeth Hudson
Melissa Laird-Smith PhD

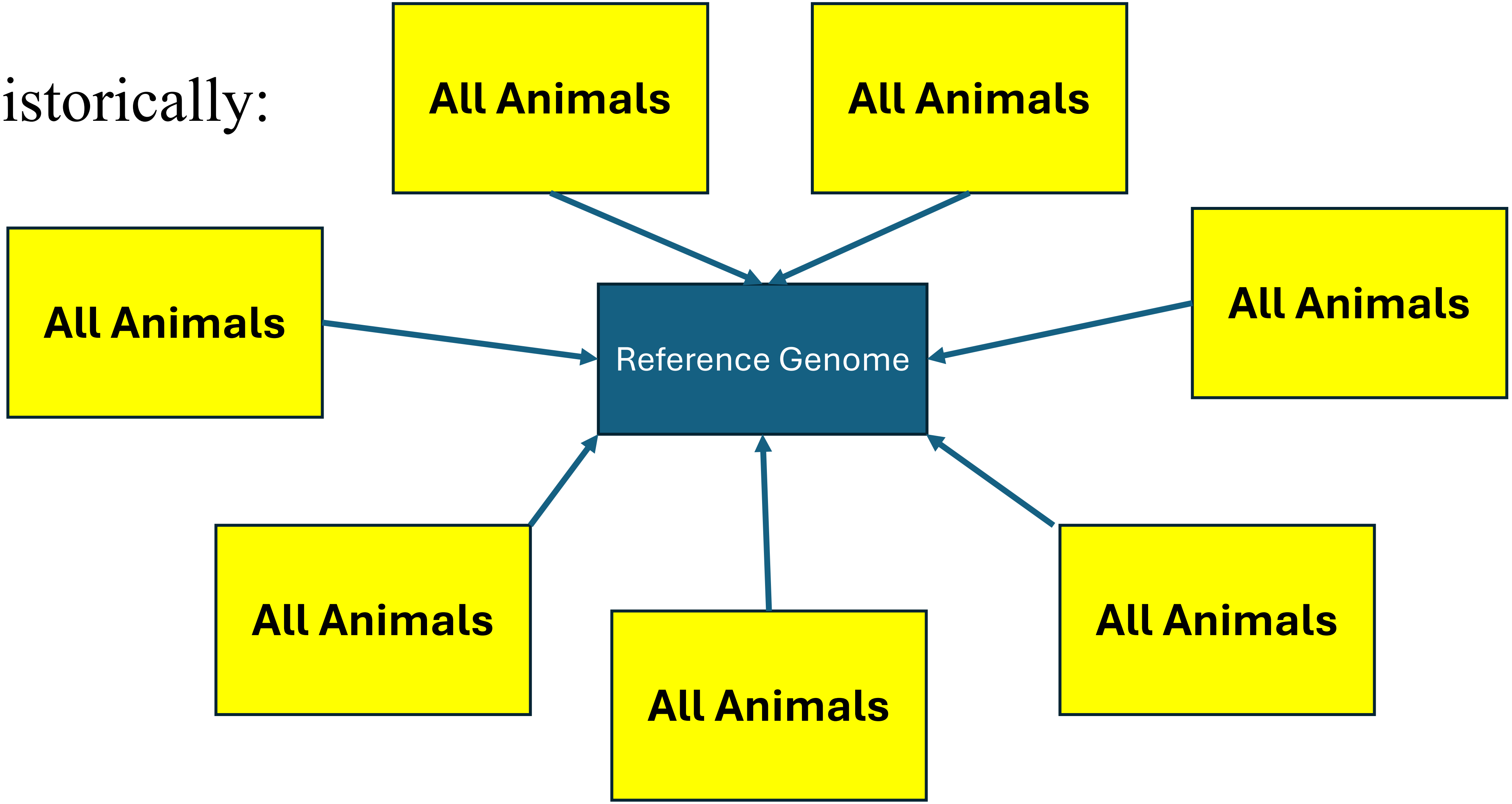
NHGRI (ONT)
Brian Walenz MS
Brandon Pickett PhD
Sergey Koren PhD
Adam Phillippy PhD

Assembly work
Mr. Kai Li
Graduate Student
University of Kentucky

6 days and 8 hours with 32 nodes on the MCC

Annotation:

Historically:



Tissues Collections

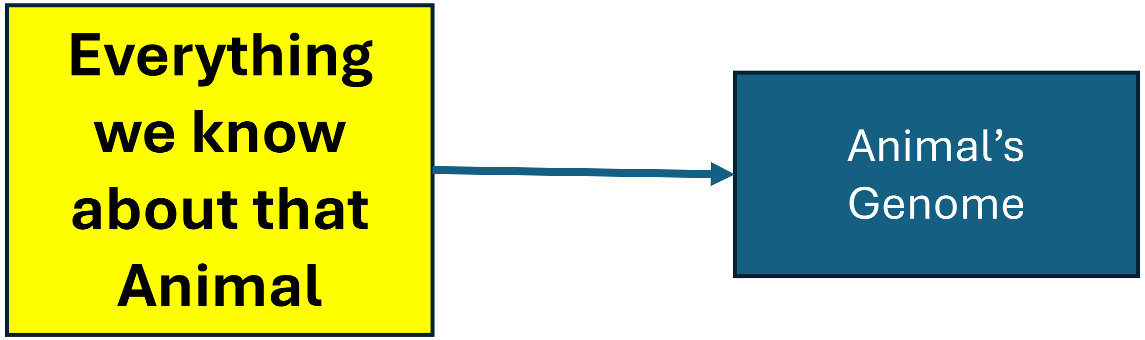
142 tissues were collected from this mule and banked.

Tissues Available for Analysis.

Credit: Doug Antczak, Don Miller
Photo credit: John Enright

Annotation:

Where are we moving?

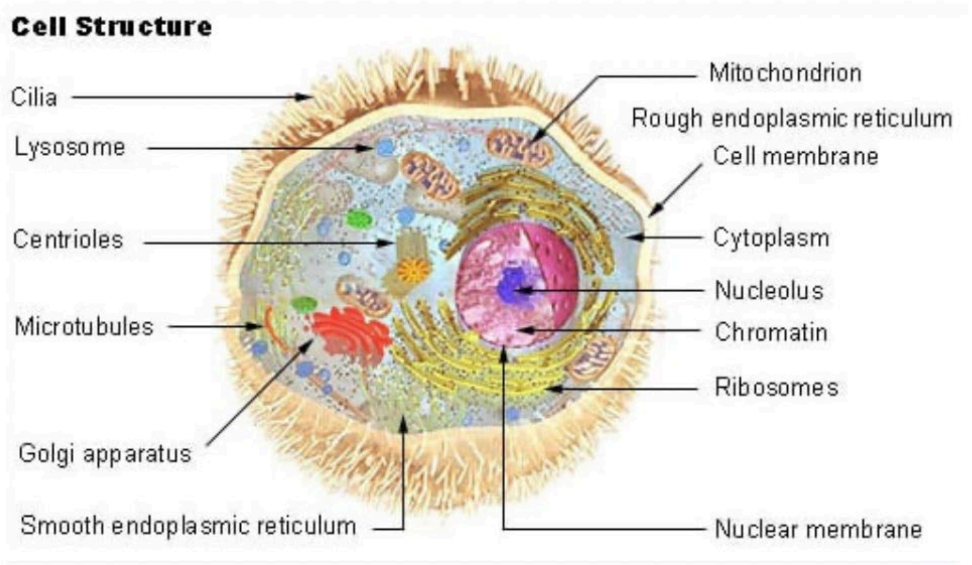


Everything we know about that Animal

Animal’s Genome

My genome knows nothing of reference genomes or their annotation, and functions just fine.

- We need to let the genome tell us its story based on the physical and chemical environment it creates.



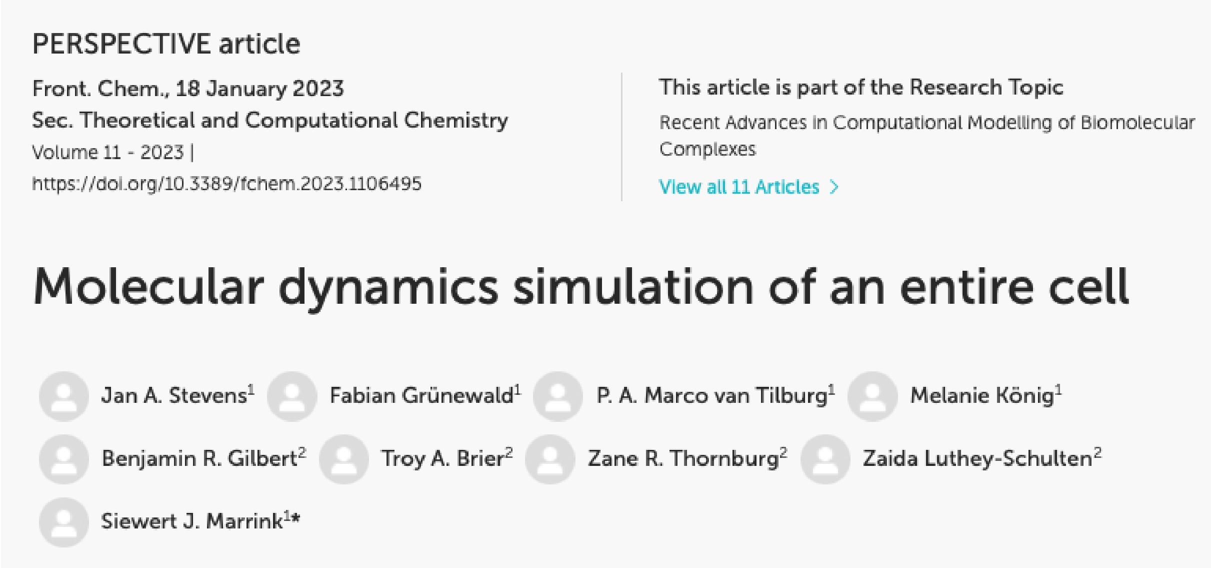
Human cell

Animals across breeds will differ both structurally and compositionally

Reference genomes will be incomplete, and otherwise have errors

https://app.achievable.me/study/usmle-step-1/learn/cell-and-molecular-biology-fundamentals

Our Cells Know Nothing of Reference Genomes or Annotation



PERSPECTIVE article

Front. Chem., 18 January 2023
Sec. Theoretical and Computational Chemistry
Volume 11 - 2023 |
https://doi.org/10.3389/fchem.2023.1106495

This article is part of the Research Topic
Recent Advances in Computational Modelling of Biomolecular Complexes
View all 11 Articles >

Molecular dynamics simulation of an entire cell

Jan A. Stevens[1] Fabian Grünewald[1] P. A. Marco van Tilburg[1] Melanie König[1] Benjamin R. Gilbert[2] Troy A. Brier[2] Zane R. Thornburg[2] Zaida Luthey-Schulten[2] Siewert J. Marrink[1]*

JCVI-syn3A
543 kbp dsDNA
493 genes

Simulation
60,887 soluble proteins
2,200 membrane proteins
1.3 million lipids
1.7 million metabolites
14 million ions
446 million water beads
Total: ~six billion atoms

Figure 2

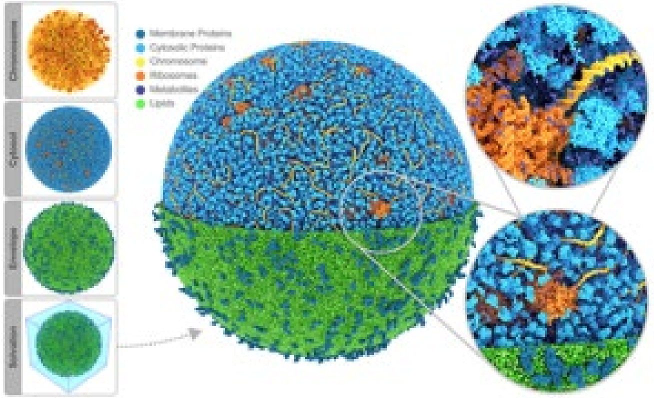
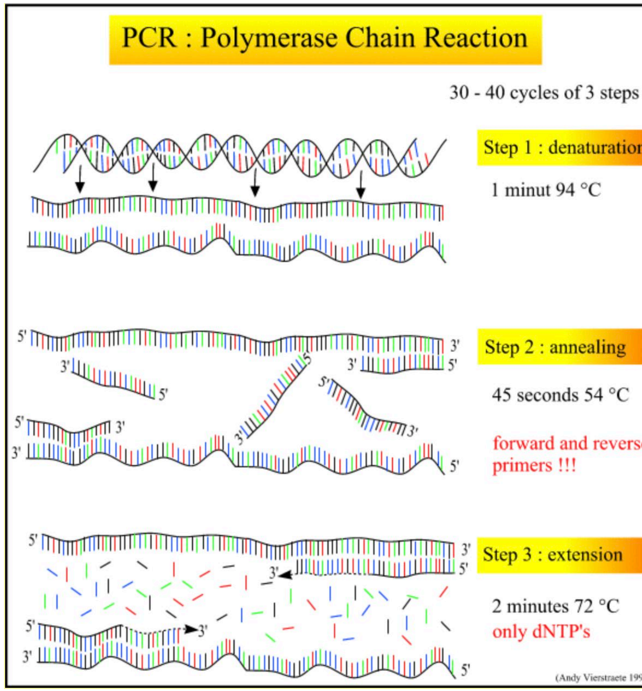
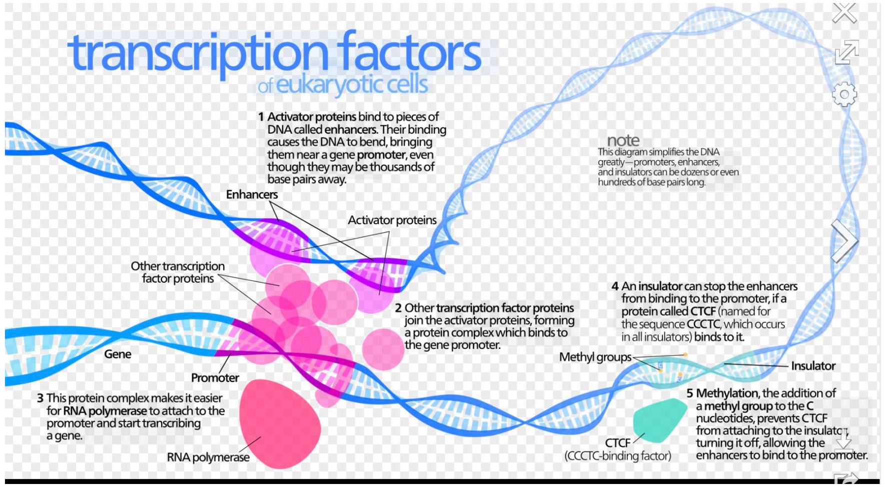


FIGURE 2. Whole-cell Martini model of JCVI-syn3A. The four stages of cell building are shown on the side. The final system contains 60,887 soluble proteins (light blue), 2,200 membrane proteins (blue), 503 ribosomes (orange), a single 500 kbp circular dsDNA (yellow), 1.3 million lipids (green), 1.7 million metabolites (dark blue), 14 million ions (not shown) and 447 million water beads (not shown) for a total of 561 million beads representing more than six billion atoms. Image rendered with Blender (Blender Online Community, 2022).

Query Biological Data the Way Biology Does

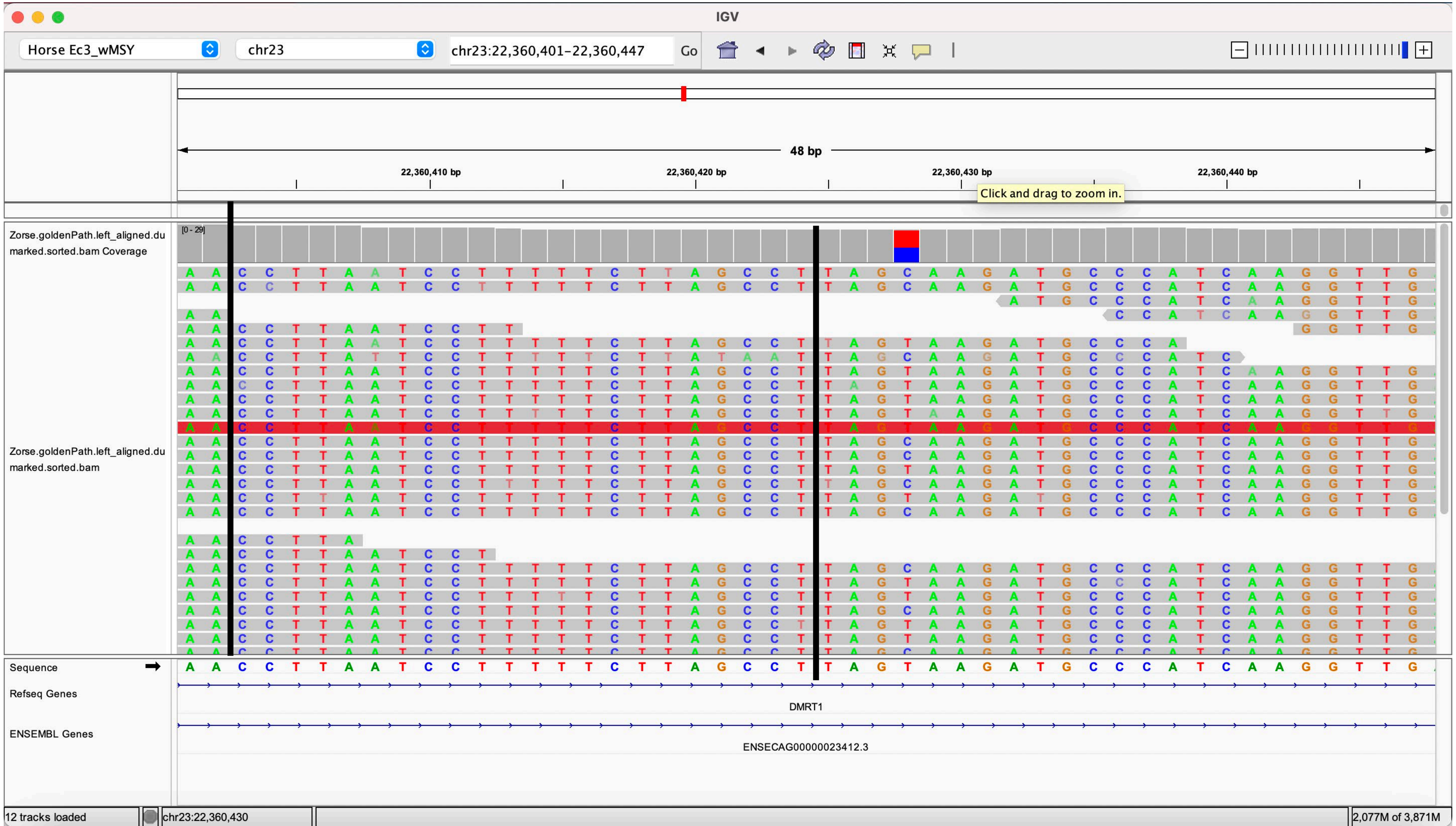
- Biology does not use accession numbers!
- Sequence identity/complementarity/composition
- Identity/complementarity
  - miRNA (microRNAs)
  - siRNA (small interfering RNAs)
  - PCR (Polymerase Chain Reaction)
  - CRISPR CAS-9
- Composition
  - Transcription factors




transcription factors
of eukaryotic cells
1 Activator proteins bind to pieces of DNA called enhancers. Their binding causes the DNA to bend, bringing them near a gene promoter, even though they may be thousands of base pairs away.
Enhancers
Activator proteins
Other transcription factor proteins
2 Other transcription factor proteins join the activator proteins, forming a protein complex which binds to the gene promoter.
Gene
Promoter
3 This protein complex makes it easier for RNA polymerase to attach to the promoter and start transcribing a gene.
RNA polymerase
note
This diagram simplifies the DNA greatly—promoters, enhancers, and insulators can be dozens or even hundreds of base pairs long.
4 An insulator can stop the enhancers from binding to the promoter, if a protein called CTCF (named for the sequence CCCTC, which occurs in all insulators) binds to it.
Methyl groups
Insulator
5 Methylation, the addition of a methyl group to the C nucleotides, prevents CTCF from attaching to the insulator, turning it off, allowing the enhancers to bind to the promoter.
CTCF
(CCCTC-binding factor)

What is a k-mer?

A DNA sequence of k-length, such as a 22-mer
CCTTAATCCTTTTTCTTAGCCT contained 23 times in this genome


IGV
Horse Ec3_wMSY
chr23
chr23:22,360,401-22,360,447
Go
48 bp
22,360,410 bp
22,360,420 bp
22,360,430 bp
22,360,440 bp
Click and drag to zoom in.
Zorse.goldenPath.left_aligned.dumarked.sorted.bam Coverage
[0 - 29]
Zorse.goldenPath.left_aligned.dumarked.sorted.bam
Sequence
A A C C T T A A T C C T T T T T C T T A G C C T T A G T A A G A T G C C C A T C A A G G T T G
Refseq Genes
DMRT1
ENSEMBL Genes
ENSECAG00000023412.3
12 tracks loaded
chr23:22,360,430
2,077M of 3,871M

de Bruijn Graphs

Schatz et al.

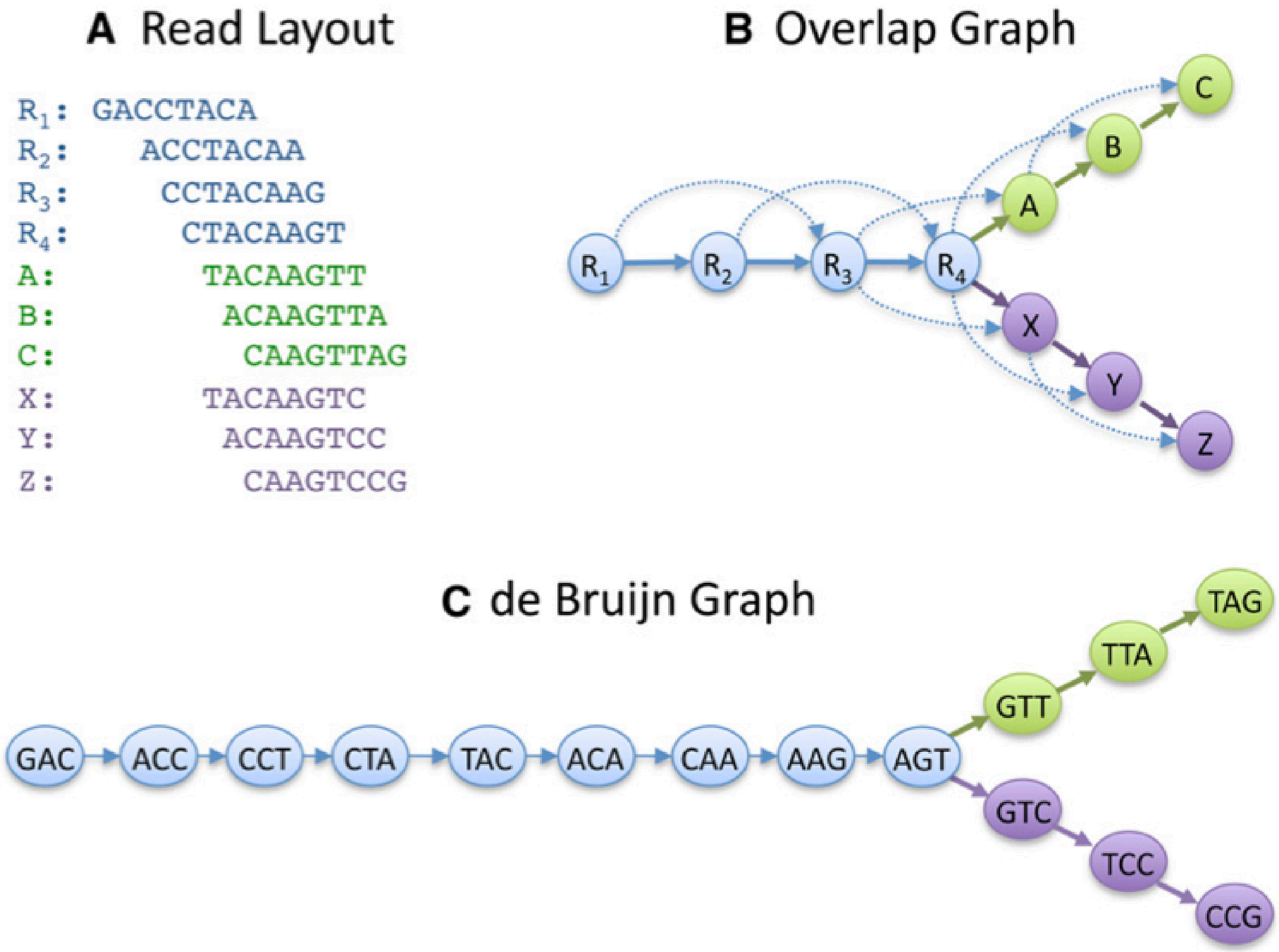
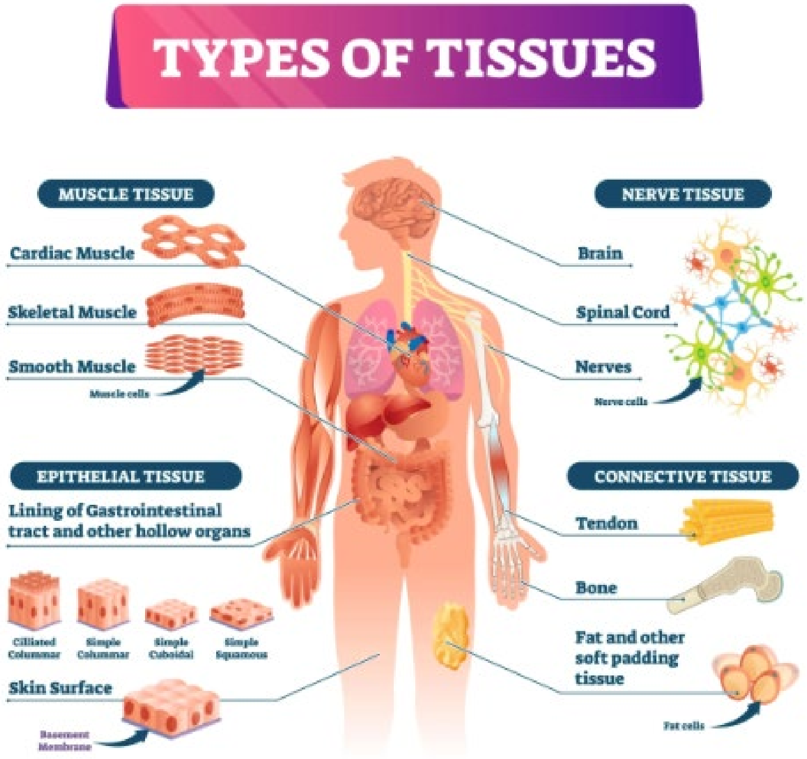
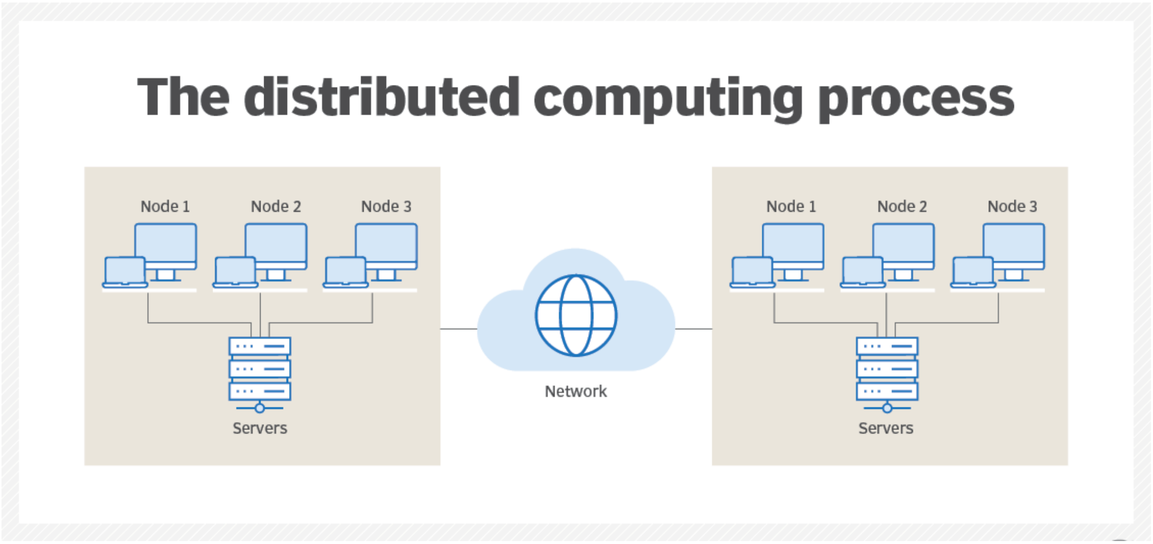


Figure 2. Differences between an overlap graph and a de Bruijn graph for assembly. Based on the set of 10 8-bp reads (*A*), we can build an overlap graph (*B*) in which each read is a node, and overlaps >5 bp are indicated by directed edges. Transitive overlaps, which are implied by other longer overlaps, are shown as dotted edges. In a de Bruin graph (*C*), a node is created for every *k*-mer in all the reads; here the *k*-mer size is 3. Edges are drawn between every pair of successive *k*-mers in a read, where the *k*-mers overlap by *k* – 1 bases. In both approaches, repeat sequences create a fork in the graph. Note here we have only considered the forward orientation of each sequence to simplify the figure.

Schatz,MichaelC.,ArthurL.Delcher,andSteven L.Salzberg."Assemblyoflargegenomesusingsecond-generationsequencing."*Genomeresearch*20.9 (2010):1165-1173.

Distributed model for data storage and computing: The network is the computer*!

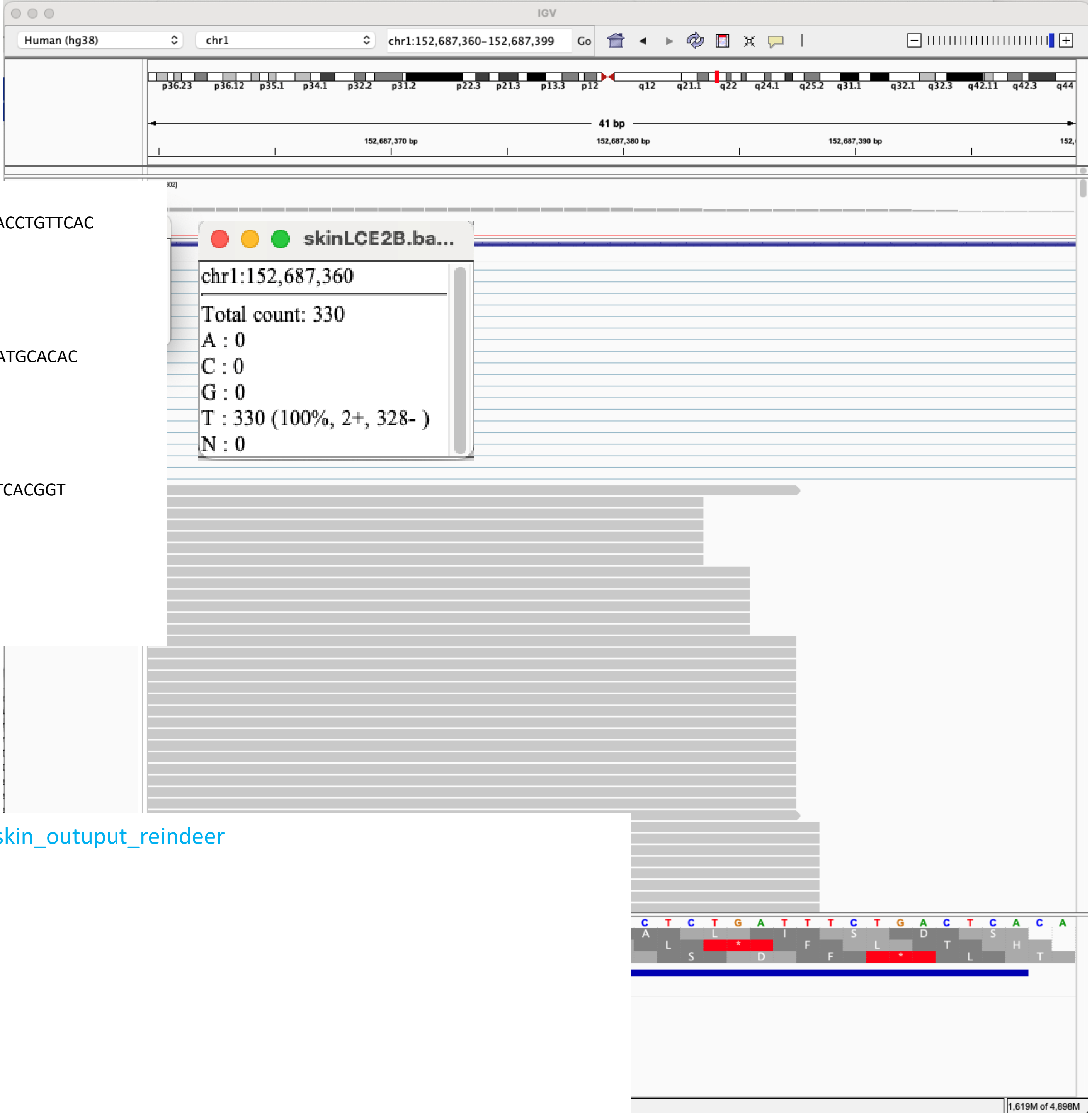




*”The Network is the Computer”, Jon Gage, Sun Microsystems, 1984

Could we just query the raw data?

```
@A00744:400:HHLN5DSX3:1:1101:1579:1000 1:N:0:GGACTTGG+CGTCTGCG
NAAGTCTCAGCAAGAGCTCAGCACACCTGTGTGGGCCAGGGAGCGCTACTGAGATCTGACCAAACCAGGAGAGACCTGTTCAC
GAGGGACAGGAGCTCAGACATGCCAAAGGGACCCGATACTTCCCCGTGGATCTCGTGAGCCATCCACA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFF
@A00744:400:HHLN5DSX3:1:1101:2953:1000 1:N:0:GGACTTGG+CGTCTGCG
NAAACAGCAGTCTTTATACAAAGCTCAAAGATATTCAGATACCCTAAATATAGCCTGCAAAATATGCACACACACATGCACAC
ACACAGGCACACACACAGGCACCTTTGGATGCATGCCAGACCTGGTGAAACTACTCATGGTTTTCCAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFF
@A00744:400:HHLN5DSX3:1:1101:3766:1000 1:N:0:GGACTTGG+CGTCTGCG
NGGGTTTTGAGTCAGAAACACCTGCCTTTAGGTCCCACTTCTACCCTGTACTGGTTTTGTGACTCATGAGCTACTCTCACGGT
GCACATTTTAGGATACGTGACATCTTTCAATATCACACAGTTAAGAAGATTCAAAGCCCAGATTTGAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFF:F,FF:FFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF,FFFFF
```



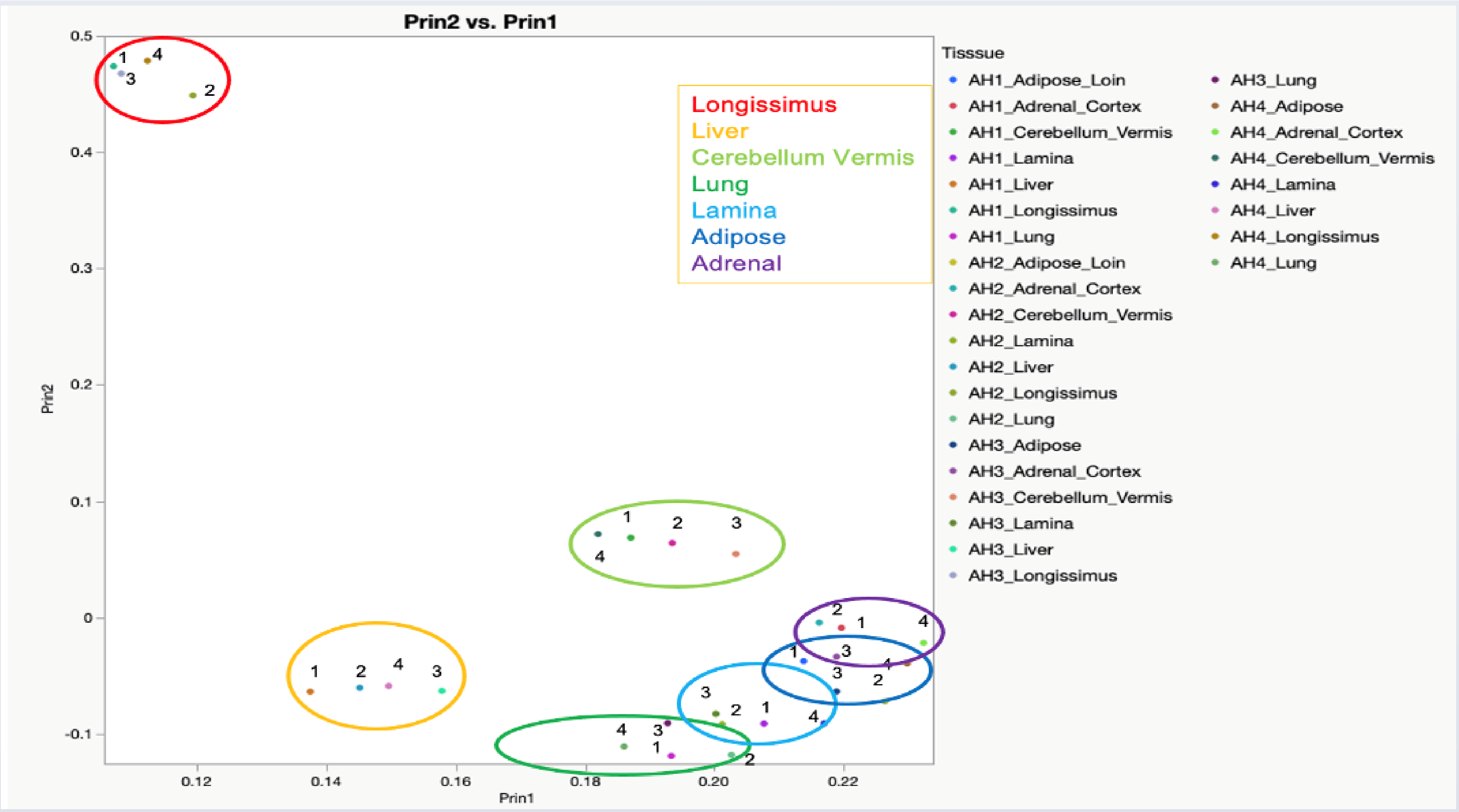
chr1:152,687,360

Total count: 330
A : 0
C : 0
G : 0
T : 330 (100%, 2+, 328-)
N : 0

```
>LCE2B
TTACCTCATGTTATAATAAAG
```

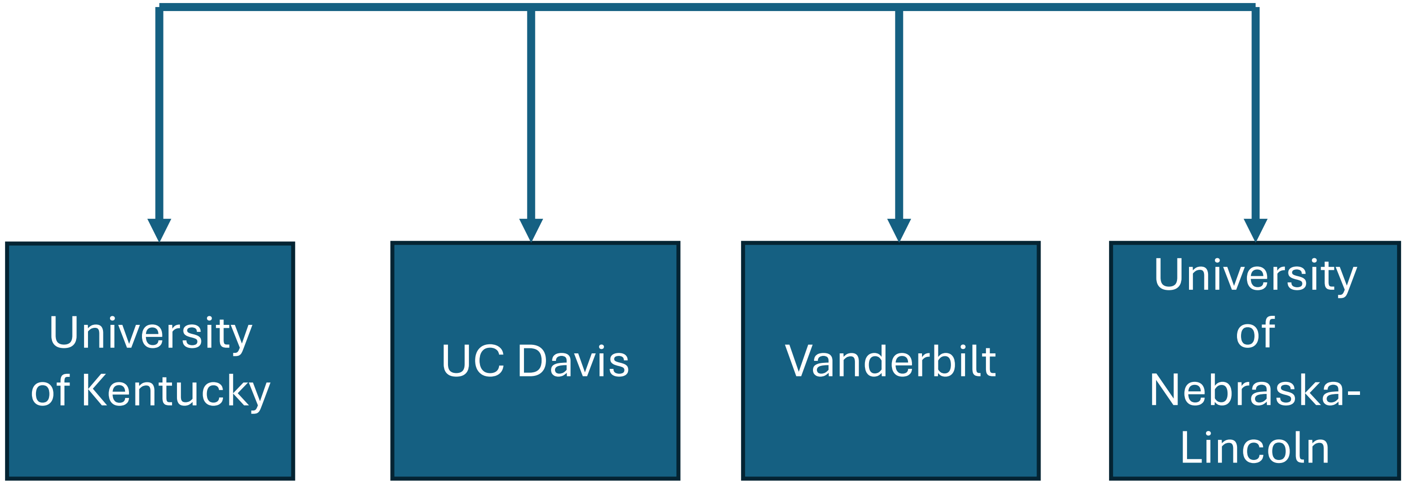
```
rdeer-client query -q rnaSeqQuery.fa SRR4422571_skin_outuput_r4422571_skin_outuput_reindeer
seq_name SRR4422571_skin
LCE2B 330


real 0m0.054s
user 0m0.044s
sys 0m0.008s
```

Can we easily differentiate datasets by tissue types?



Long view: A distributed network of interfaces allowing for the rapid query of all HTS datasets

In minutes, we could know of samples/individuals with an interesting genotype, or samples with a specific expression profile for use in hypothesis generating work.



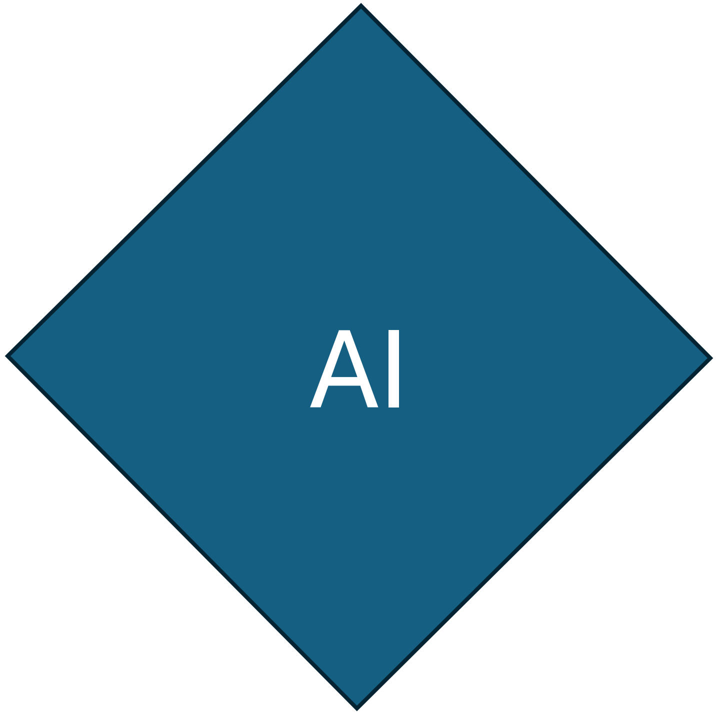
All queries are independent of a reference!

Process once, query forever!

What Data Structures will we feed AI?

How is one of these
things not like the others?

Population level Data
Accurate Phenotype Information
Genetic Information?



AI

What genetic elements are rigidly conserved

What genetic elements are linked to dysfunction

Students in the Lab

Kai Li

Lauren Johnson

Julia Ciosek

Dr. Nahla Hussien

Xiomara Arias

Agricultural Genome to Phenome Initiative (AG2PI):USDA-NIFA 2022-70412-38454.

This presentation was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH