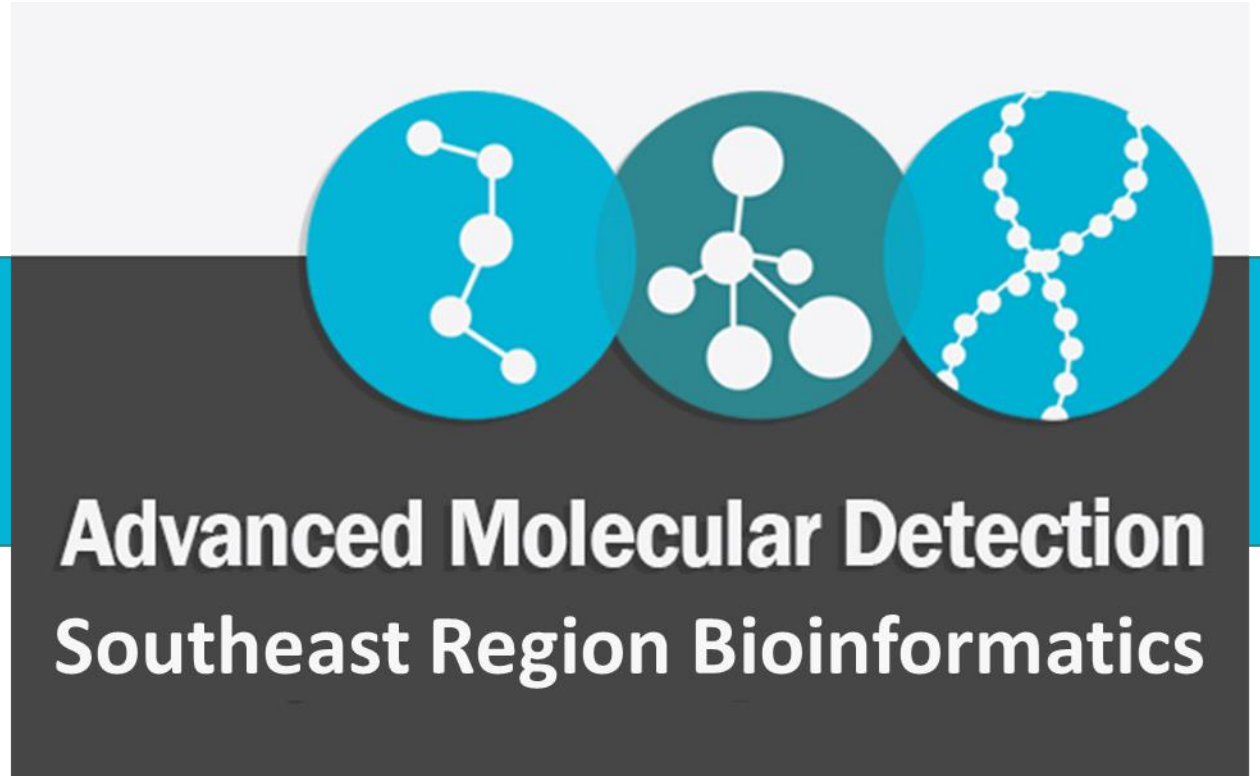


AMD Southeast Region Genomic Epidemiology
Training Series
Part 1: Introduction to Genomic Epidemiology
2/19/2024

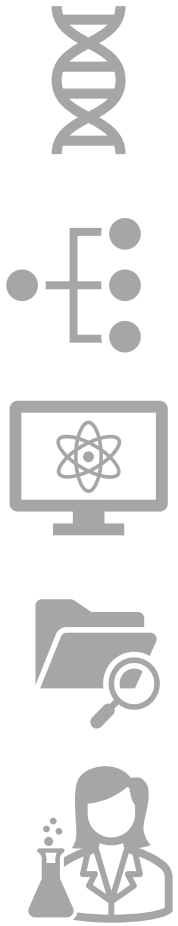
Outline

Welcome and Outline of Training Series

What is Genomic Epidemiology?

Benefits of Genomic Epidemiology

Genomic Workflow: Sequencing and Bioinformatics



- Introduction to Genomic Epidemiology (Today)
- Phylogenetic Trees (March 4th)
- Genomic Epidemiology Tools (March 18th)
- Case Studies (April 2nd)
- Retrospective Data Analysis and Current Capabilities (April 15th)

Advanced Molecular Detection
Southeast Region Bioinformatics

Our Bioinformatics Team

Supervisor
TBD (Funcitionally Molly and Callin)

Bioinformaticians
Sam Marcellus, MPH
Lakshmi Thsaliki, MS
Molly Mitchell, PhD
Yibo Dong, PhD
Tassy Bazile, MS

Data Analysts
Omer Tekin, MS
TBD

Genomic Epi Bioinformatician
Jianye Ge, PhD

Who am I?

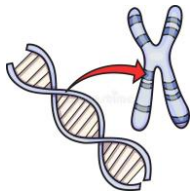
- BS Neurobiology, University of Iowa
- MPH in Epidemiology, University of Iowa
- Genetic Clinical Trials at Iowa Stead Family Children's Hospital
- APHL-CDC Newborn Screening Bioinformatics and Data Analytics Fellowship
- Texas Public Health Laboratory
- Florida BPHL

Advanced Molecular Detection
Southeast Region Bioinformatics

What is Genomics?

Genes

- Small sections of DNA
- DNA -> RNA -> Protein (Central Dogma)



Genome

- Complete set of genetic instructions
- Stored in long DNA molecules called chromosomes

Genomics

- Study of groups of genes
- Can study one organism or many organisms

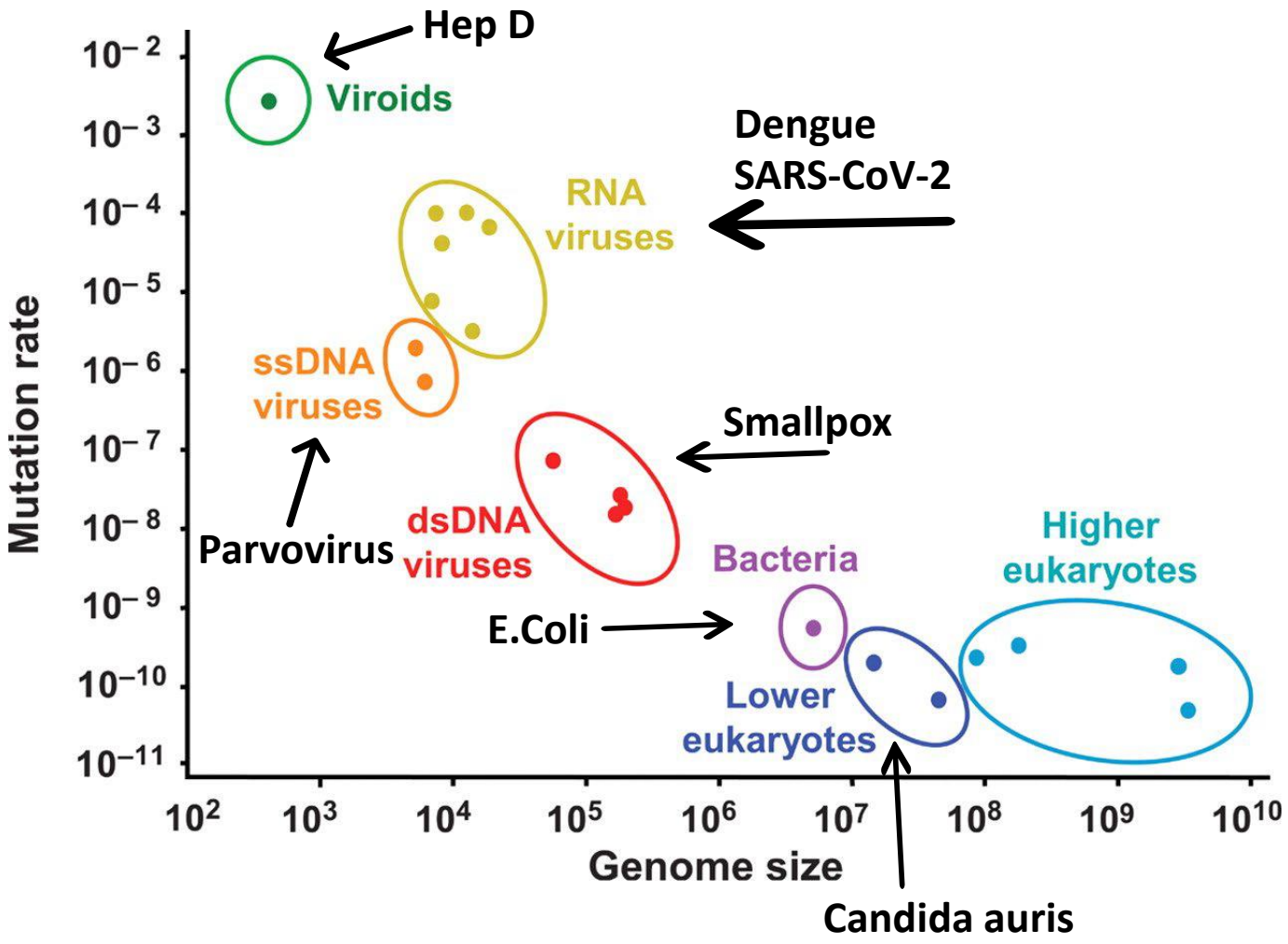
Genomic Epidemiology

- CDC Definition: "use of pathogen genomic data to determine the distribution and spread of an infectious disease in a specified population and the application of this information to control health problems"
- Functional Definition: Using genomic data to enhance traditional outbreak investigation and routine surveillance

Advanced Molecular Detection
Southeast Region Bioinformatics

https://www.nature.com/scitable/definition/genome-43/

Diversity of Pathogen Genomics



- All pathogens are different lengths
  - Human is 10^9
- Each require a different (but similar) preparation for sequencing
- Most can be sequenced on similar machines, but some require special long-read sequencing

Bedford.io
H.Blankenship

Whole Genome Sequencing (WGS)

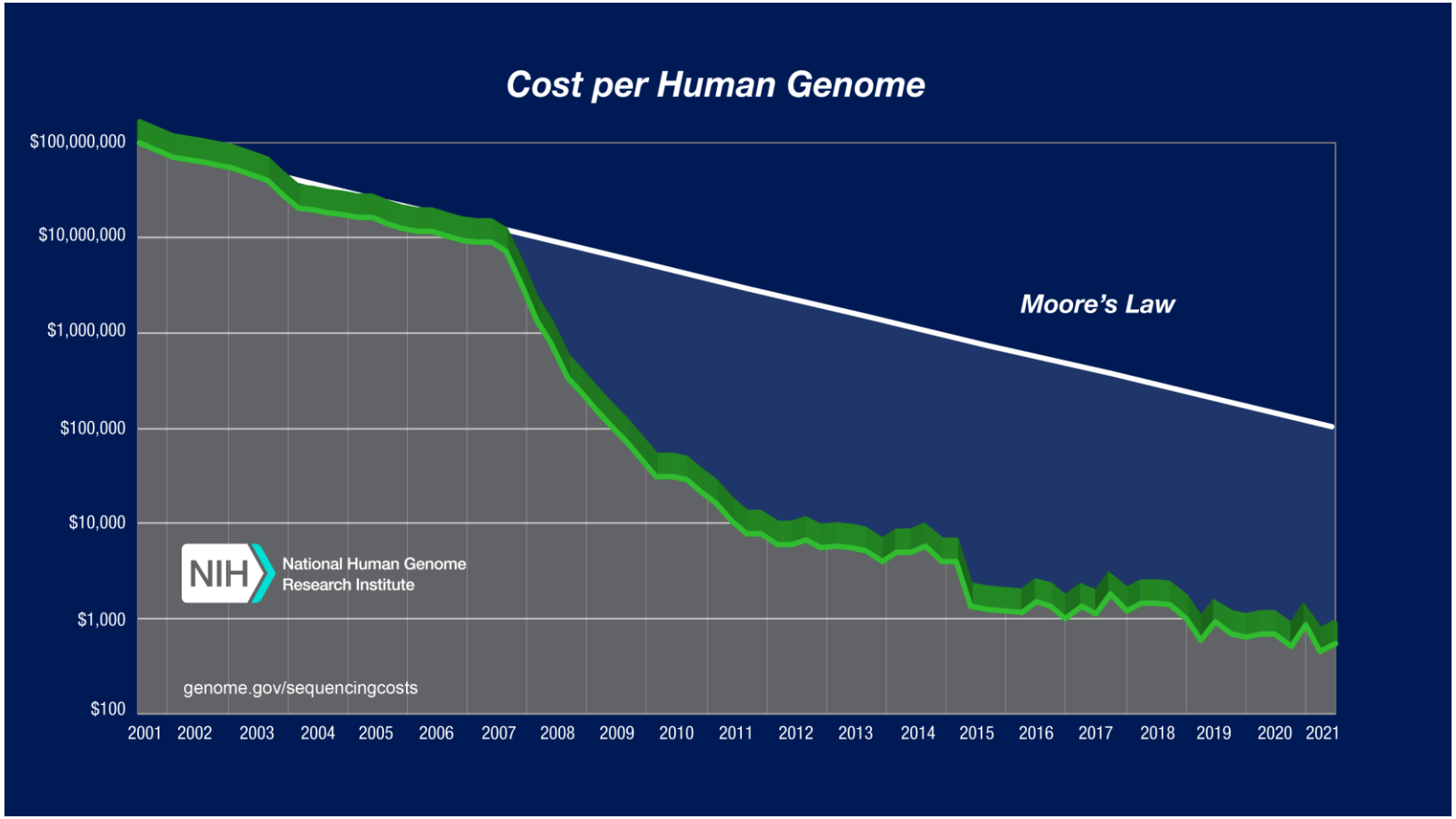
- All organisms have a unique genetic code made of nucleotide bases
  - A, T, C, and G
- WGS reads the entire genetic code of an organism
- Short or Long Read
  - Short read breaks the DNA in smaller pieces before sequencing
  - Long read is beneficial for hard to identify pathogens
  - Both are used and are beneficial in Public Health
- Collaboration between microbiologists and bioinformaticians

Advanced Molecular Detection
Southeast Region Bioinformatics

Image adapted from H. Blankenship

WGS Pricing

- Moore's Law-price of computing decreases at ~2.85%/month
- Sequencing price declines ~6.5%/month
- One of the fastest price declines in history
- Covid-19 increased funding for sequencing



Advanced Molecular Detection
Southeast Region Bioinformatics

https://www.humanprogress.org/the-fastest-learning-curve-in-history/#:~:text(Adjusting%20for%20the%20time%20price,Long%20live%20learning%20curves.
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Whole Genome Sequencing (WGS)

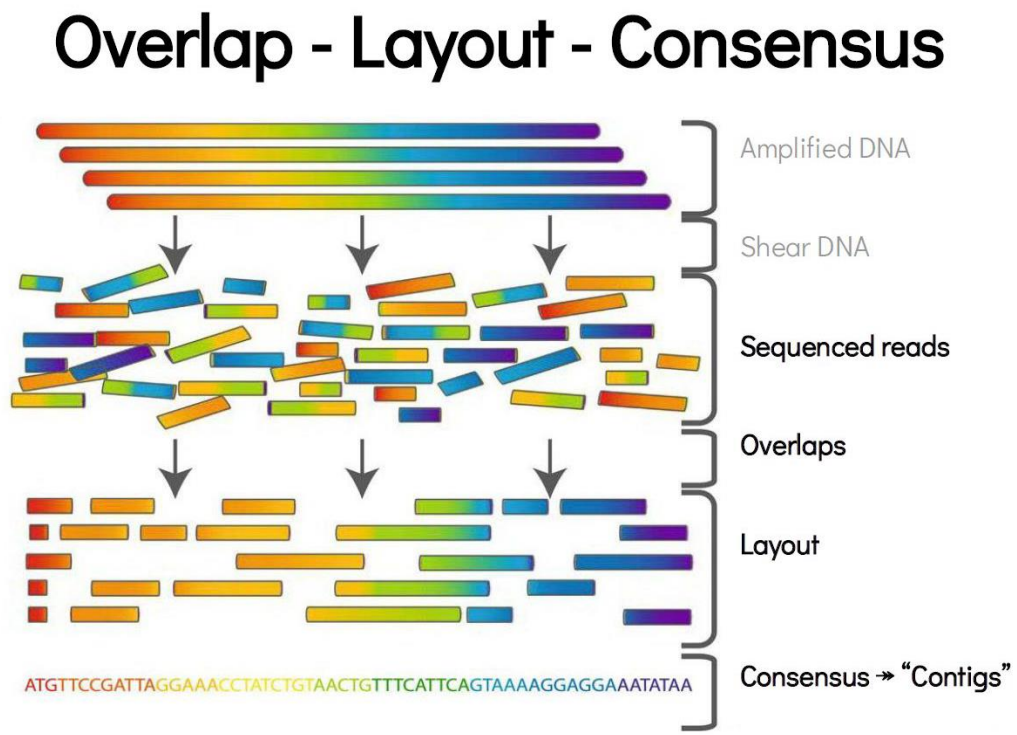
- Sanger Sequencing
  - Reads one molecule at a time, in order
  - Output like a streamer
- Next-Generation Sequencing (NGS)
  - Reads lots of little fragments at once
  - Massively parallel sequencing (MPS)
  - Output like confetti

Advanced Molecular Detection
Southeast Region Bioinformatics

Image adapted from H. Blankenship

Next-Generation Sequencing (NGS)

- Bioinformaticians choose an assembly method
  - De novo
    - "from the beginning"
    - Does not require a reference genome
    - Assembles all the reads it can
    - Can miss complex repeats
  - Reference based
    - Contigs (stretch of continuous sequence) assembled using reference as a map
    - Requires a quality reference genome
    - Highly diverse elements can be lost

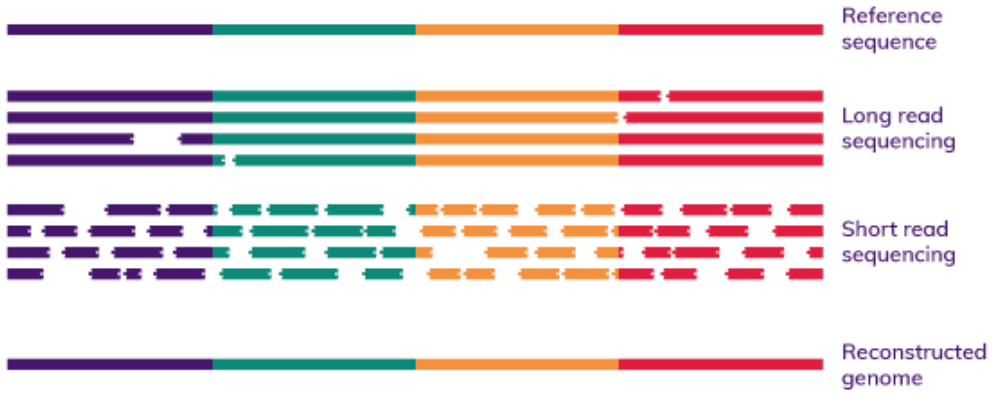


Advanced Molecular Detection
Southeast Region Bioinformatics

Adapted from H. Blankenship

Next-Generation Sequencing (NGS)

- Short Read
  - Breaks the DNA in smaller pieces before sequencing
  - Can have gaps between sequences
  - More common in Public Health
- Long Read
  - Long read is beneficial for hard to identify pathogens
  - Accuracy per read can be lower than short read
  - Takes longer than short read



Advanced Molecular Detection
Southeast Region Bioinformatics

Adapted from H. Blankenship
Image from H. Turner 6 Dec 2022

Additional Sequencing Terminology

- Enrichment
  - Targets specific regions of the genome for sequencing
- Metagenomics
  - AKA shotgun sequencing
  - Study of DNA isolated from a bulk sample, typically an environmental sample (like wastewater)

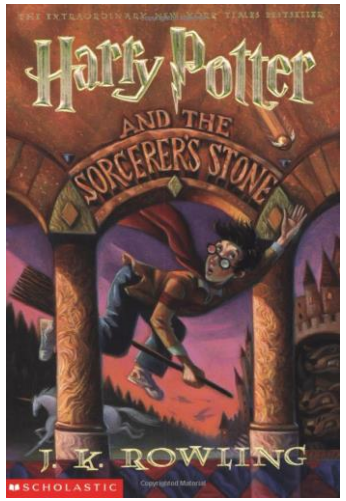
Advanced Molecular Detection
Southeast Region Bioinformatics

https://www.illumina.com/science/technology/next-generation-sequencing/beginners/glossary.html

Bioinformatics Terminology

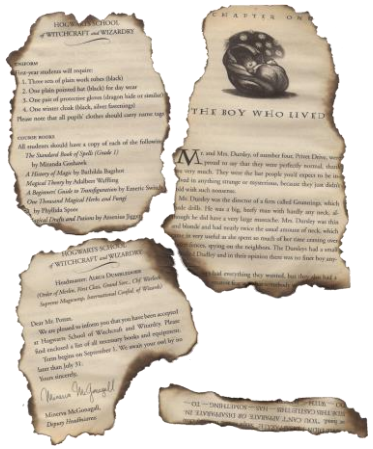
- Alignment (AKA Read Mapping)
  - Matching sequencing fragments to a reference genome
- Assembly
  - Putting the sequencing fragments into the correct order

Organism DNA

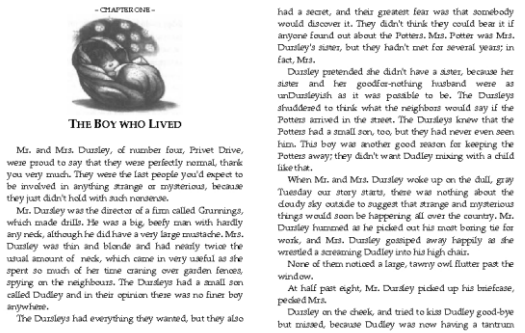


Library Prep

Assembly



Alignment



CHAPTER ONE

THE BOY WHO LIVED

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs.

Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that.

When Mr. and Mrs. Dursley woke up on the dull, gray Tuesday our story starts, there was nothing about the cloudy sky outside to suggest that strange and mysterious things would soon be happening all over the country. Mr. Dursley hummed as he picked out his most boring tie for work, and Mrs. Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair.

None of them noticed a large, tawny owl flutter past the window.

At half past eight, Mr. Dursley picked up his briefcase, pecked Mrs.

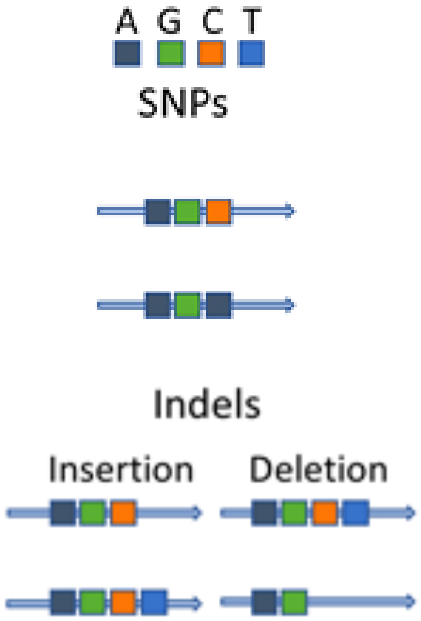
Dursley on the cheek, and tried to kiss Dudley good-bye but missed, because Dudley was now having a tantrum

1

Definitions from NIH

Bioinformatics Terminology

- Annotation
  - Identifying the location of genes and determining what they do
- Single Nucleotide Polymorphism (SNP, read:SNIP)
  - Change in one single nucleotide (A, T, G, or C) from the reference sequence
- INDEL
  - Insertion or deletion of nucleotides from the sequence



Advanced Molecular Detection
Southeast Region Bioinformatics

Definitions from NIH
Image from CSC Blog

Comparative Genomics Terminology

- Comparative Genomics
  - Comparing the genome of one organism to another individual of the same species to understand transmission patterns, evolutionary changes, and gene function
- Phylogenetic Tree
  - Diagram depicting the evolution of different genes and/or species from a common ancestor
  - More to come in a future module
- SNP Matrix
  - Chart comparing the number of SNP differences between pairs of all the samples analyzed (also more to come later)

Advanced Molecular Detection
Southeast Region Bioinformatics

WGS and Epi Collaboration

- Monitor emergence of new pathogen strains
- Compare trends pre- and post-vaccination campaigns
- Gather evidence to determine case relatedness to clusters
- Identify antibiotic resistant genes

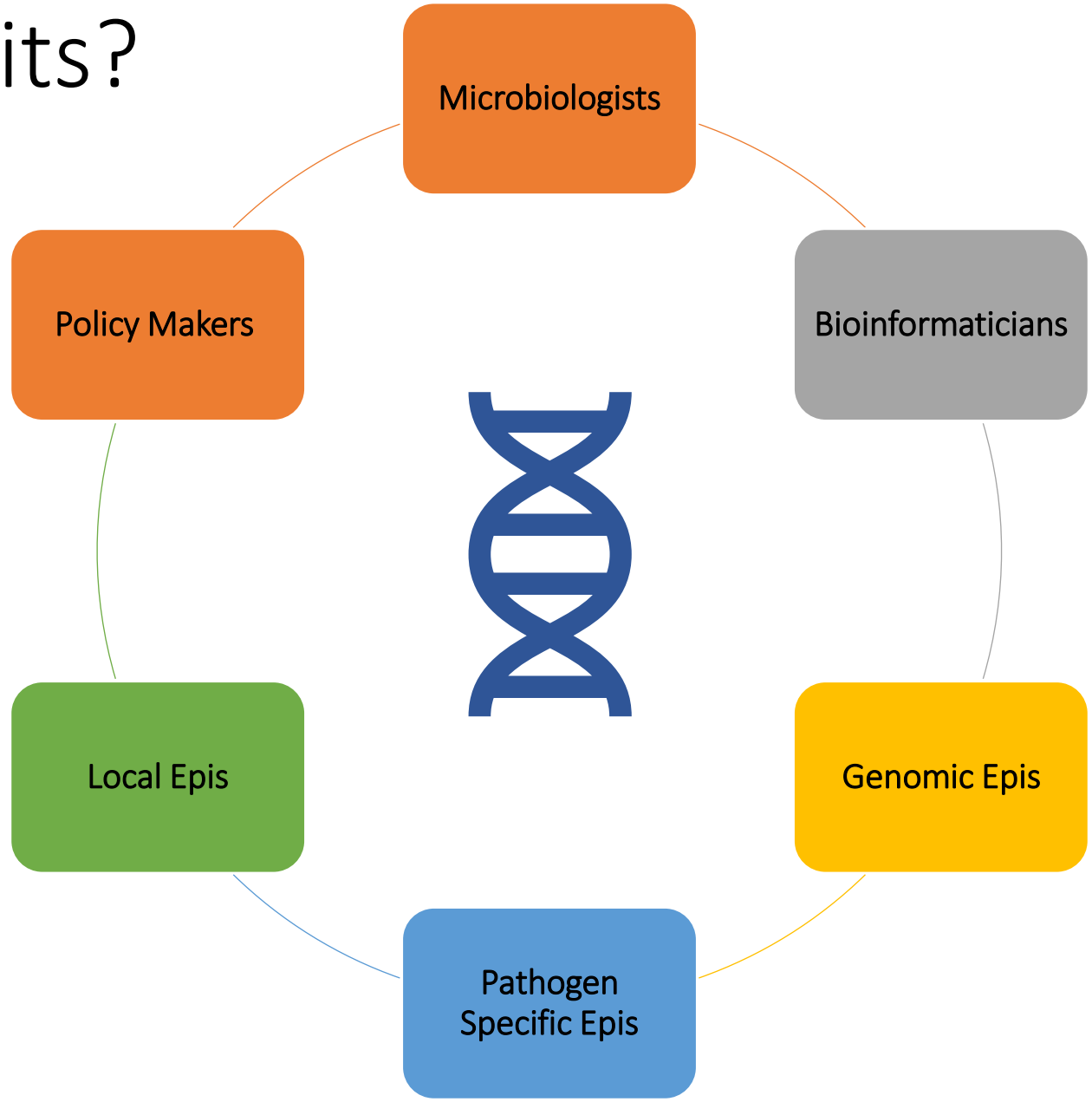
Advanced Molecular Detection
Southeast Region Bioinformatics

WGS and Epi

Advanced Molecular Detection
Southeast Region Bioinformatics

https://www.washingtonpost.com/graphics/2020/health/coronavirus-genetic-code/

Who Benefits?



Advanced Molecular Detection
Southeast Region Bioinformatics

Microbiology Workflow

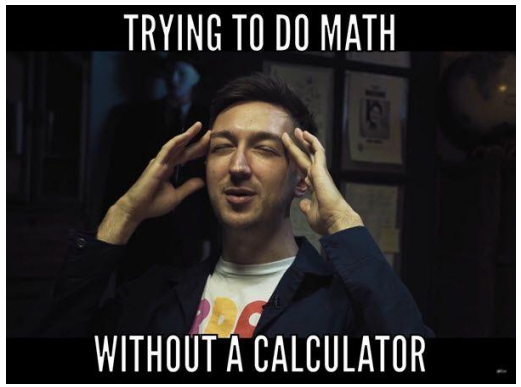
- Pre-Processing
  - Verify the sample is positive
  - Check if the sample is quality enough for sequencing

- DNA Extraction
  - Keep the DNA and RNA
  - Get rid of everything else
  - Takes 1-4 hours

Advanced Molecular Detection
Southeast Region Bioinformatics

Microbiology Workflow

- Library Prep
  - Molecular Origami
  - Get the DNA into the right shape for sequencing
  - Takes 4-8 hours
- Sample Pooling
  - All the math!
  - Determine what and how many sequences can be sequenced together
- Run Prep
  - Enter the sample information into the sequencer
  - Load reagents on the machine
  - Takes 1-3 hours



Advanced Molecular Detection
Southeast Region Bioinformatics

Adapted from B. McGruder Rawson 2023

Microbiology Workflow

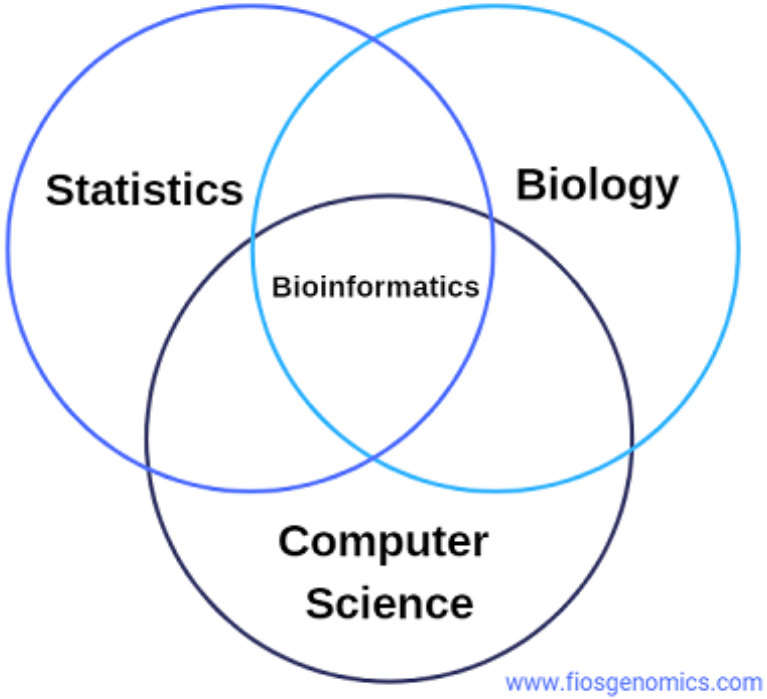
- Sequencing
  - Takes 18-45 hours depending on size of genome and sequencing platform
  - Fingers crossed nothing goes wrong

- Send to the Bioinformaticians

Advanced Molecular Detection
Southeast Region Bioinformatics

What is Bioinformatics

- Interdisciplinary field that combines biology, computer science, and statistics to develop methods and software tools for understanding biological data, including the collection and analysis of large, complex genetic datasets (e.g., next-generation sequencing data)



Advanced Molecular Detection
Southeast Region Bioinformatics

Adapted from S. Schmedes 2023

Public Health Bioinformatics Skillset

- Genetics, molecular biology, evolutionary biology, microbiology, virology, epidemiology, computer science
- Hybrid lab and bioinformatics experience
- Programming/coding
- High-performance computing, cloud computing
- Software development
- Statistics, data science, machine learning
- Science and technical communication

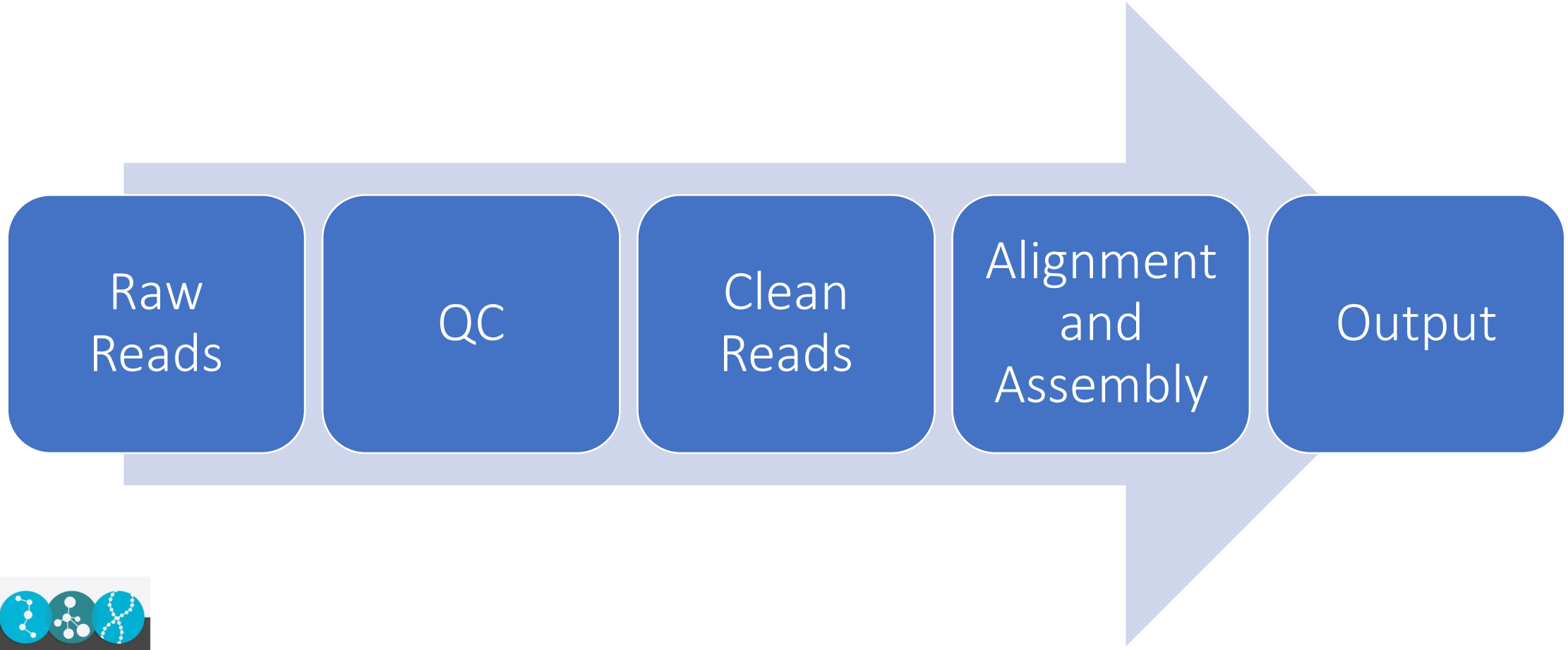
Advanced Molecular Detection
Southeast Region Bioinformatics

Bioinformatics Lab Tasks

- Develop and maintain analytic capability
  - Pipeline development
  - Data analysis
- Data management and reporting
- Genomic surveillance
- Support outbreak investigations
- Training and workforce development

Advanced Molecular Detection
Southeast Region Bioinformatics

Bioinformatics Workflow



Advanced Molecular Detection
Southeast Region Bioinformatics

Deliverables

- SNP differences (used to determine relatedness in cluster analysis)
  - Matrix, right, for a CRPA Outbreak
- Unknown pathogen identification
  - Species and subspecies identification
- Serotyping and subtyping
- Antimicrobial resistance detection
- Phylogenetic trees
  - More to come in future module

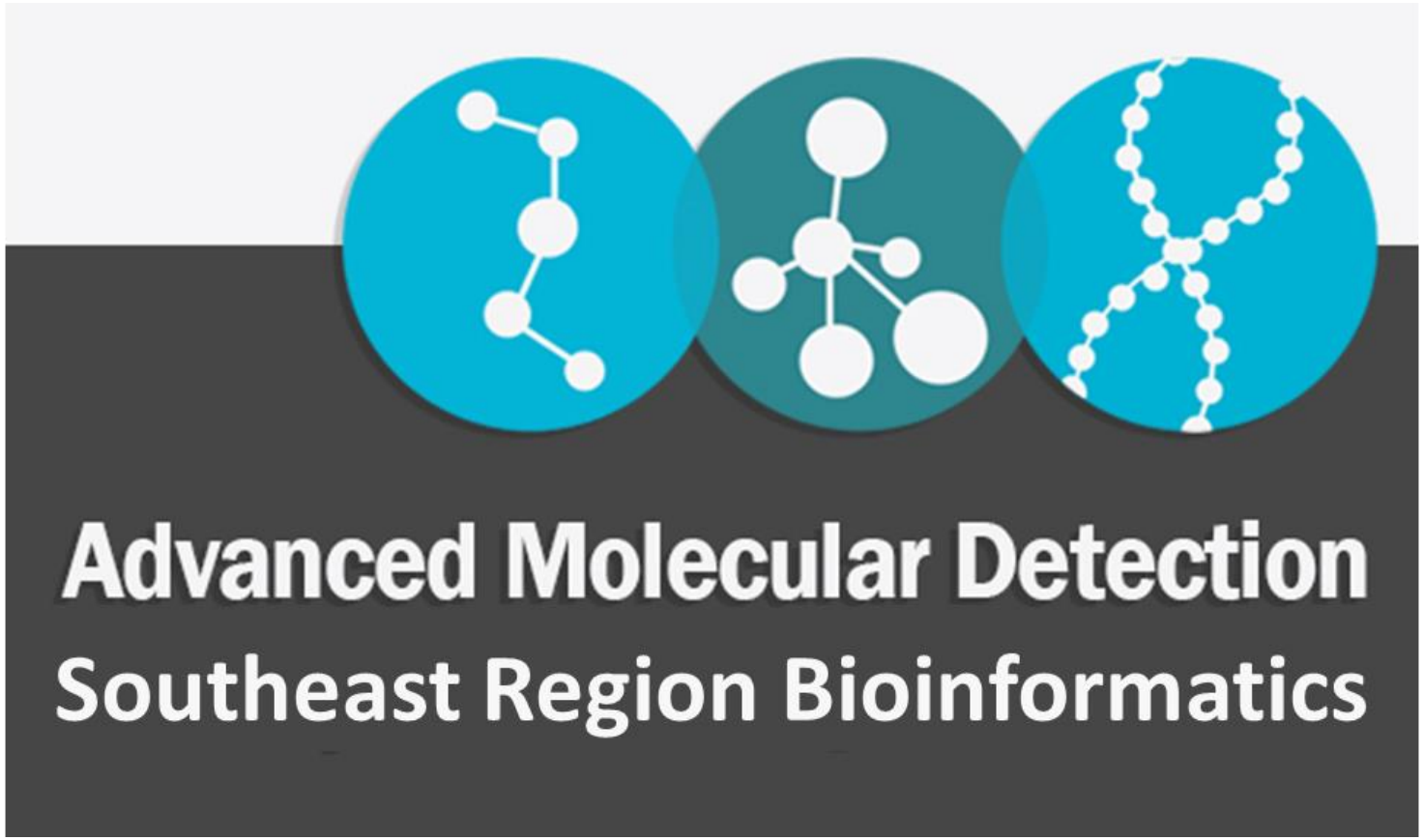
Source	Case A	Case B	Case D	Case F	Case G	Room X_HW	Room X _Bath
Case A	-	6	0	3	4	8	9
Case B	6	-	6	9	10	4	5
Case D	0	6	-	3	4	8	9
Case F	3	9	3	-	7	11	12
Case G	4	10	4	7	-	13	13
Room X_HW	8	4	8	11	13	-	7
Room X_ Bath	9	5	9	12	13	7	-

Advanced Molecular Detection
Southeast Region Bioinformatics

Summary

- Genomics is the study of groups of genes in one or many organisms
- Pathogens all have different genome lengths and levels of complexity
- Next Generation Sequencing is the most common sequencing method for pathogen sequencing
  - Bioinformaticians turn confetti into readable books
- Genomic Epidemiology uses genomic information to enhance outbreak investigations and routine surveillance
- Many groups benefit from genomic epidemiology

Advanced Molecular Detection
Southeast Region Bioinformatics



Questions?

Bphl-sebioinformatics@flhealth.gov

TBD
Lead Bioinformatician & Supervisor
TBD@flhealth.gov

Molly Mitchell, PhD
Bioinformatician
Molly.Mitchell@flhealth.gov

Lakshmi Thsaliki, MS
Bioinformatician
Lakshmi.Thsaliki@flhealth.gov

Sam Marcellus, MPH
Bioinformatician
Samantha.marcellus@flhealth.gov