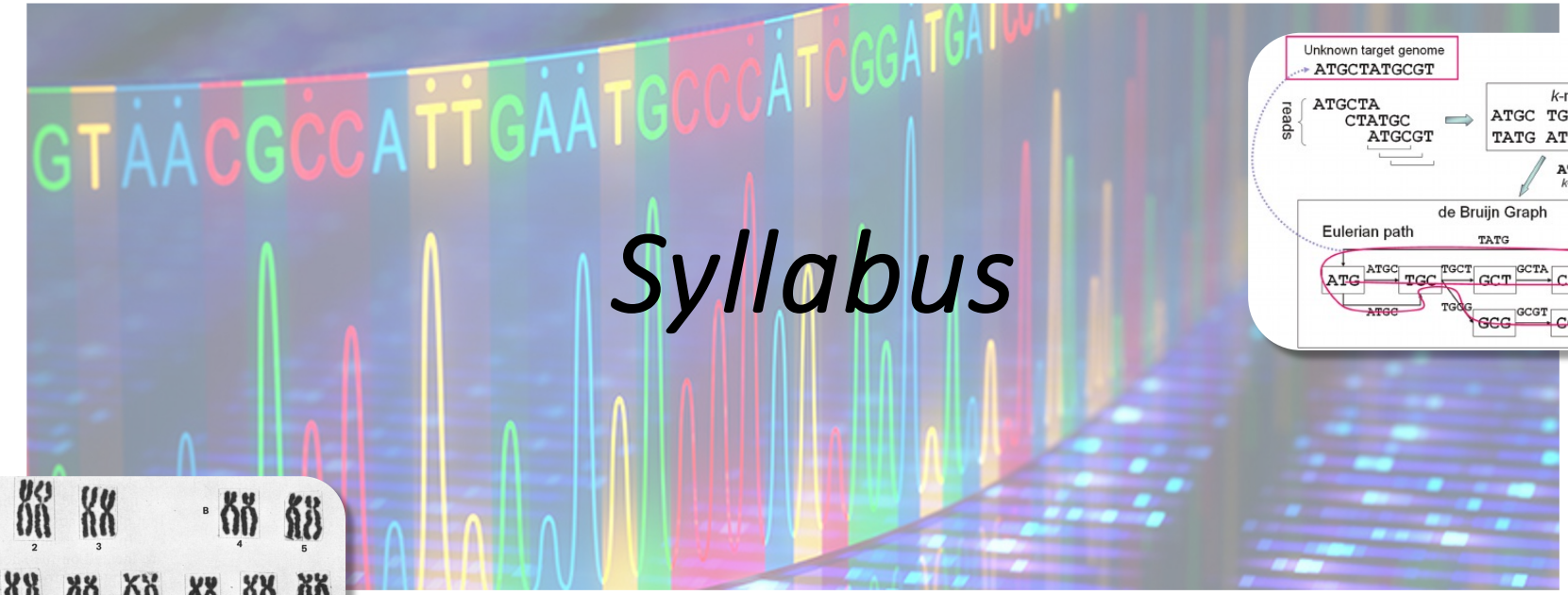
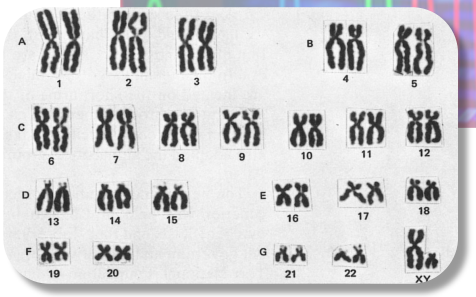
Genomics & Bioinformatics





BIOL 497, 597

Boise State University

Spring

Health & Safety

- **Our priority is your safety!**
- The following rules apply for in-person classes:
  - ✓You can use sanitizing wipes to clean surfaces (desk and keyboard) at beginning and end of class
  - ✓Students should not attend class in person if they have any of the listed infectious diseases within the Communicable Disease Policy
  - ✓For more information consult the BSU campus public health website

INSTRUCTOR

- **Name:** Sven BUERKI
- **Office:** Science Building, office 114
- **Email:** svenbuerki@boisestate.edu

Class location & meeting times

- **Location:** SCNC149 computer room
- **Lectures:** Wednesdays from 9:00-10:50 AM
- **Labs:** Fridays from 9:00-10:30 AM
- No office hours, but please contact me if you want to set an appointment

ETHOS

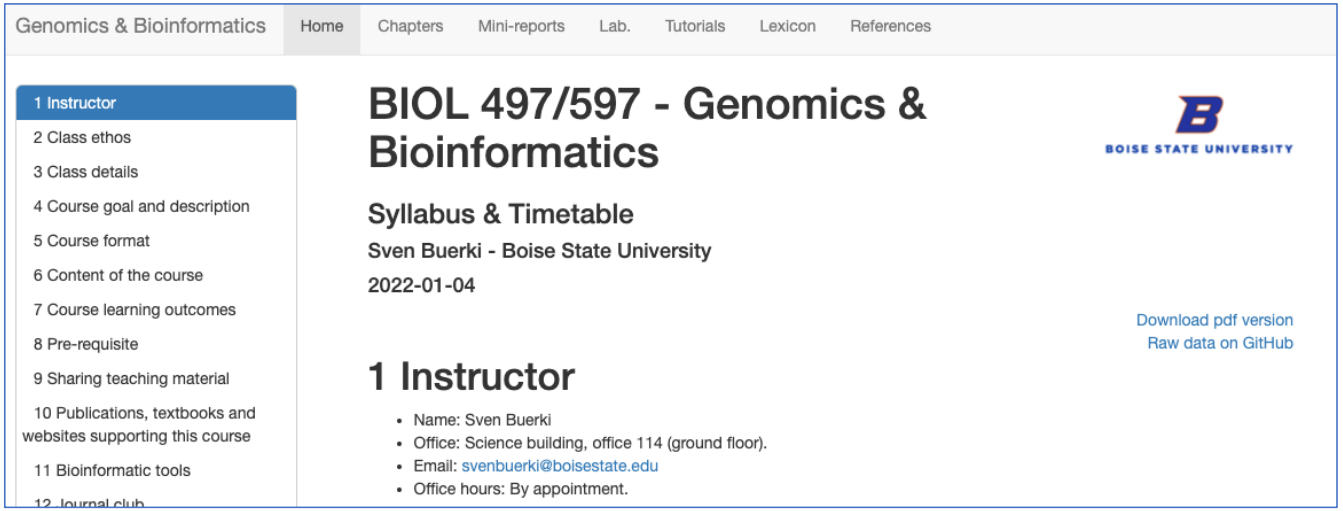
- Everyone here is smart; distinguish yourself by being kind

  Kindness in Science is an inclusive approach that fosters diversity, respect, wellbeing & openness leading to better science outcomes.

ROUND OF INTRODUCTIONS

RESOURCES

https://svenbuerki.github.io/Genomics-Bioinformatics



Shared Google Drive

canvas

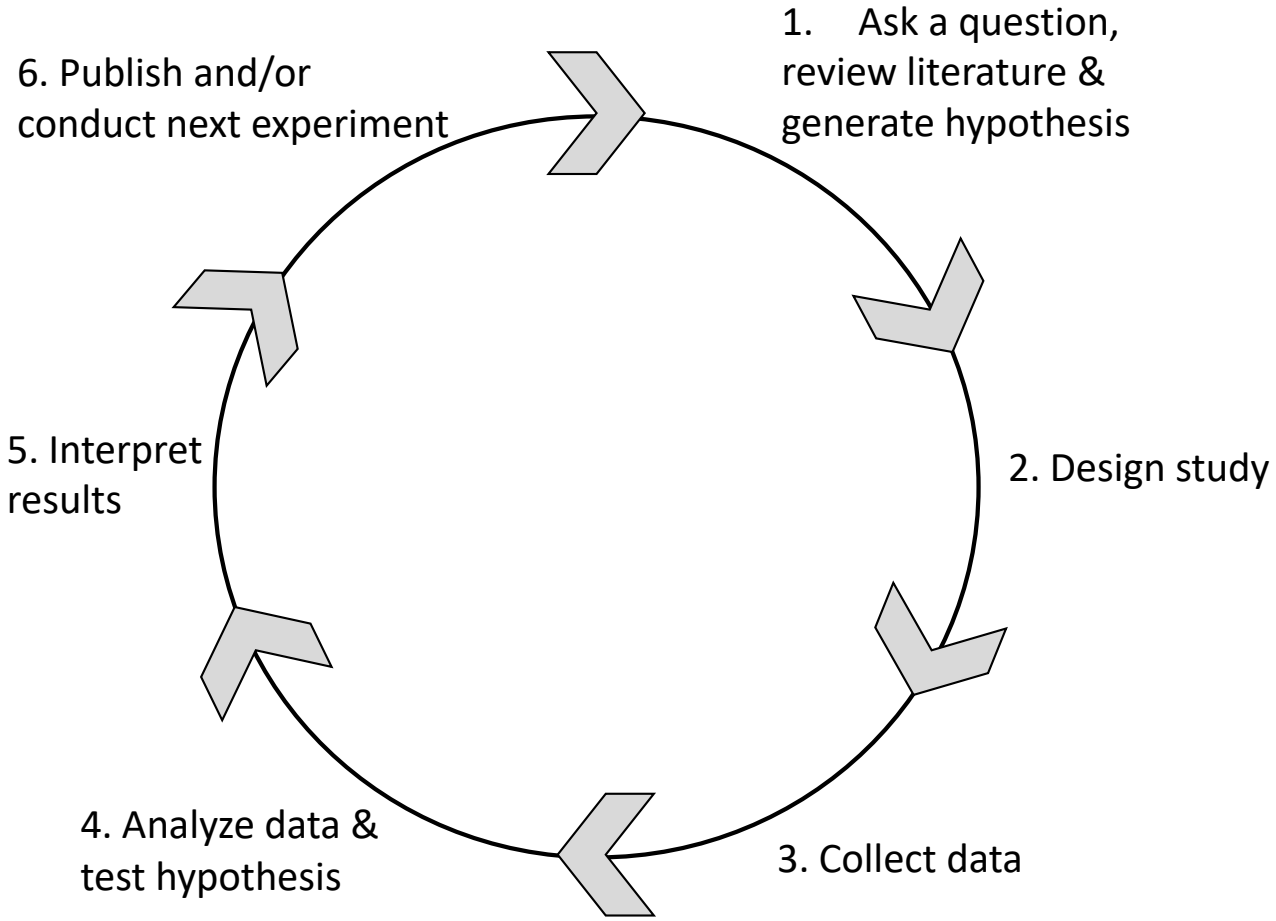
Course Goal & description

The goal is to provide students with the **theoretical and applied knowledge in genomics and bioinformatics to sequence, assemble and annotate genomes**, especially for non-model organisms.

COURSE FORMAT

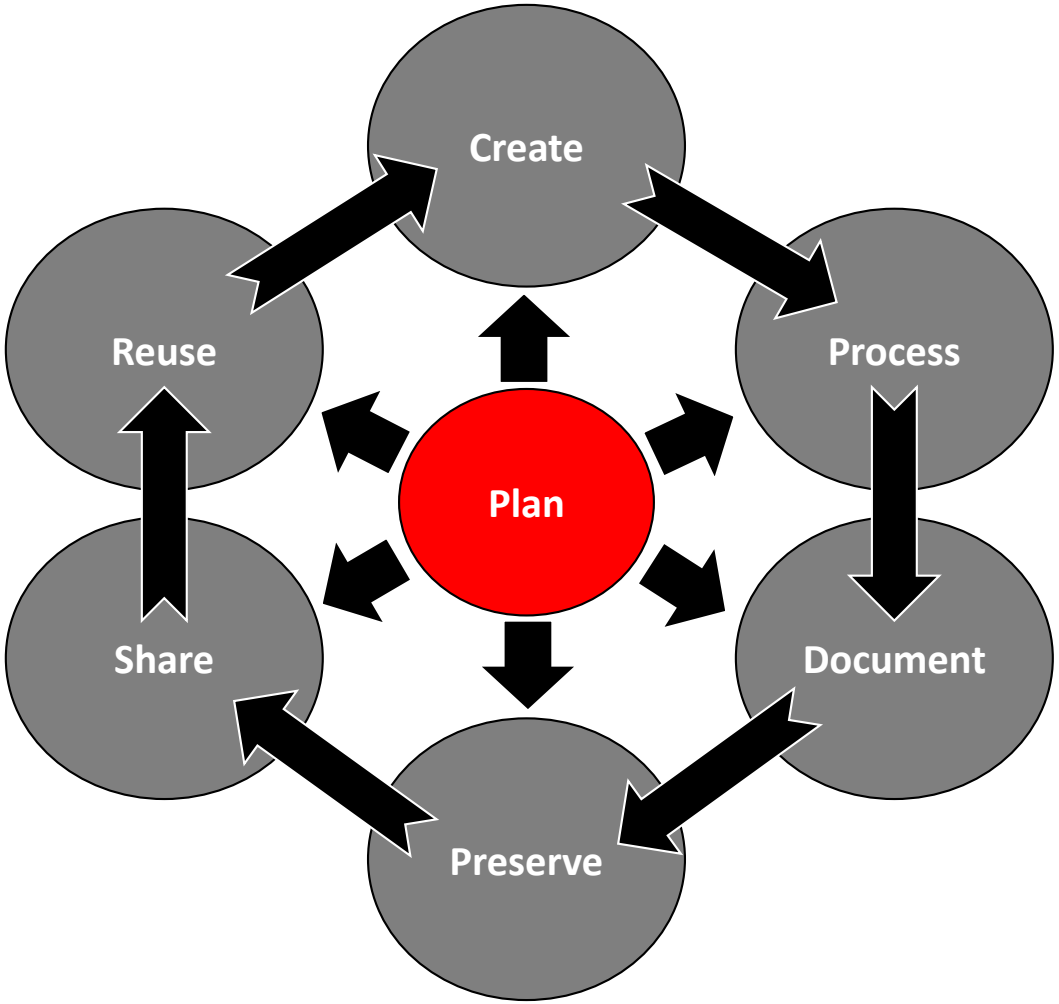
- This class provides a mixture of **lectures** together with more applied, **hands-on bioinformatic tutorials**
  - Tutorials are designed to support students mastering theoretical genomic concepts through their implementations in bioinformatics protocols
- The genomics field heavily relies on bioinformatic expertise
  - Work on computers running the **Linux operating system**
  - Opportunities to become familiar with the **bash/shell**, **Python** and **R** computing languages
- Lecture sessions will also serve as a platform to:
  - Work on graded mini-reports
  - Study genomic literature through a journal club

SCIENTIFIC PROCESS



https://svenbuerki.github.io/EEB603_Reproducible_Science/

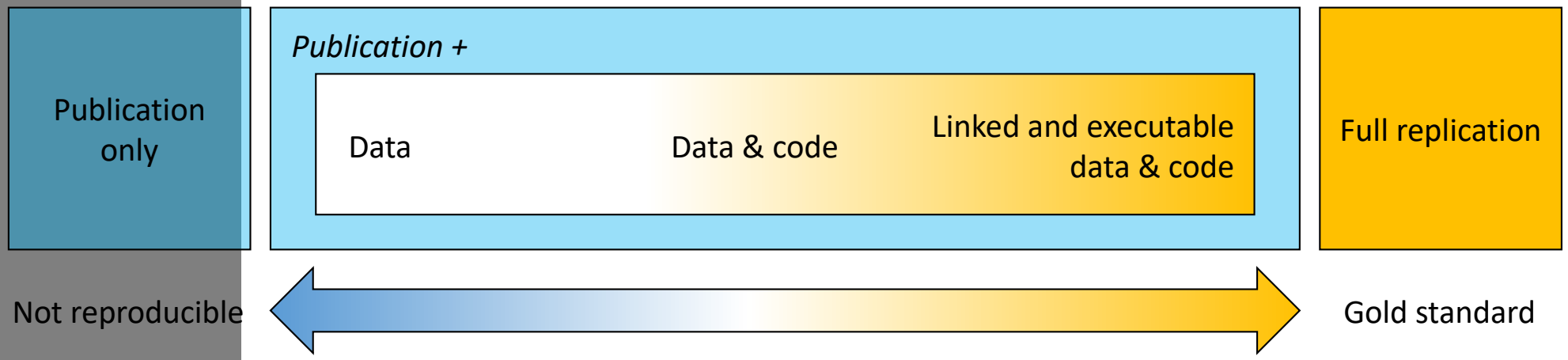
Data Life-Cycle



https://svenbuerki.github.io/EEB603_Reproducible_Science/

Reproducible Science

Based on a survey published in *Nature* (2016), 90% of the respondents said that **there is a reproducibility crisis in Science!**



https://doi.org/10.1038/533452a

Reproducible Science

Comment | Open Access | Published: 08 December 2015

Five selfish reasons to work reproducibly

Florian Markowetz

Genome Biology **16**, Article number: 274 (2015) | Cite this article

20k Accesses | **43** Citations | **492** Altmetric | Metrics

Abstract

And so, my fellow scientists: ask not what you can do for reproducibility; ask what reproducibility can do for you! Here, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

https://doi.org/10.1186/s13059-015-0850-7

Reproducible Science

SCIENCEINSIDER | POLICY

White House requires immediate public access to all U.S.-funded research papers by 2025

Policy is a blow to journal paywalls, but its impact on publishing is unclear

26 AUG 2022 · 2:20 PM · BY JEFFREY BRAINARD, JOCELYN KAISER

Main repository of genomic data

NIH National Library of Medicine
National Center for Biotechnology Information

doi: 10.1126/science.ade6076

Promote open-access and open-source resources

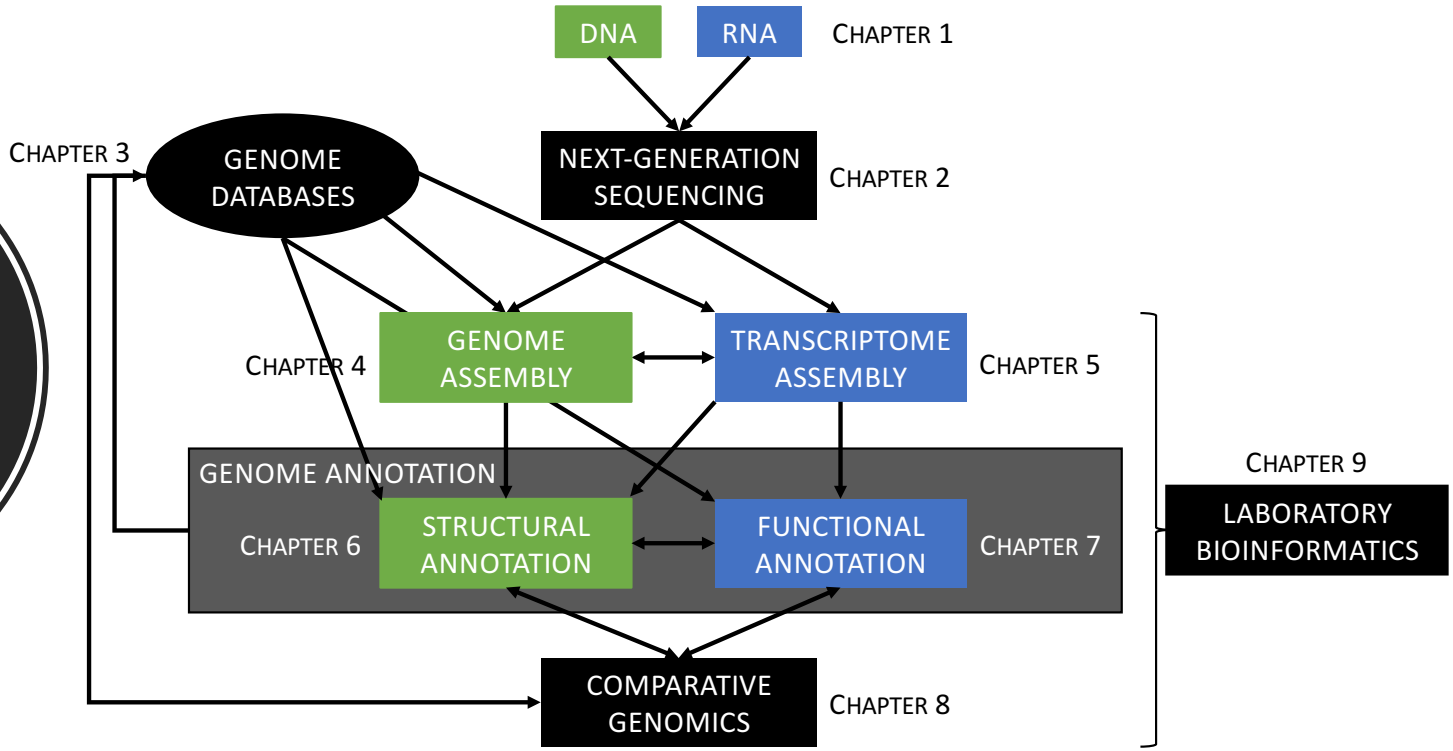
GRADING

Tests conducted during this course

- 3 individual mini-reports (2x 25 points and 1x 50 points, TOTAL: 100 points)
- 1 group lab report (150 points)
- 1 group lab presentation (50 points)

https://svenbuerki.github.io/Genomics-Bioinformatics/index.html#13_Assessments

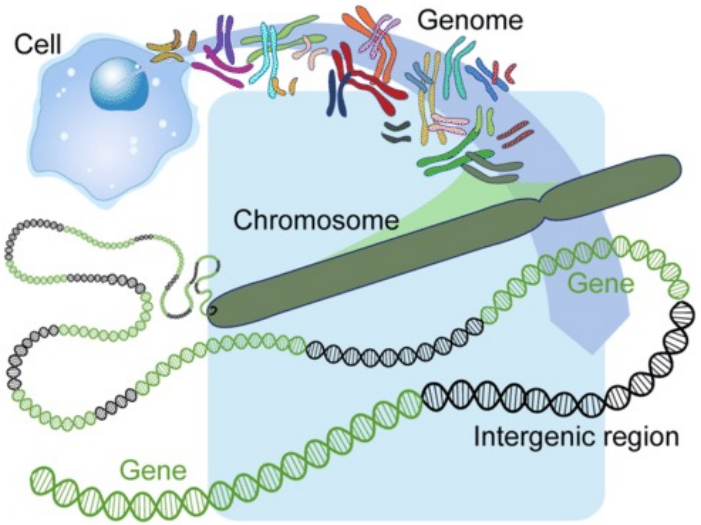
COURSE LEARNING OUTCOMES

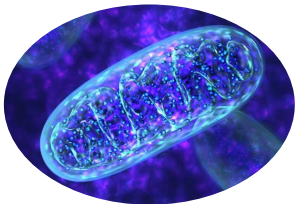


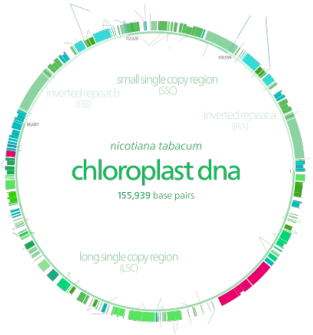
https://svenbuerki.github.io/Genomics-Bioinformatics/index.html#7_Course_learning_outcomes

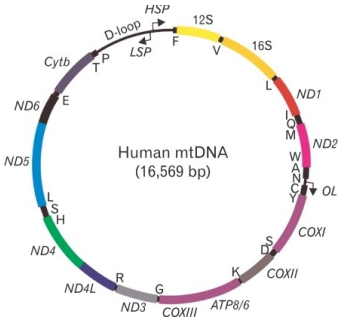
CHAPTER 1

Moving our mindset from the study of single genes (genetics) to the study of the entire genetic material in a cell (genomics).



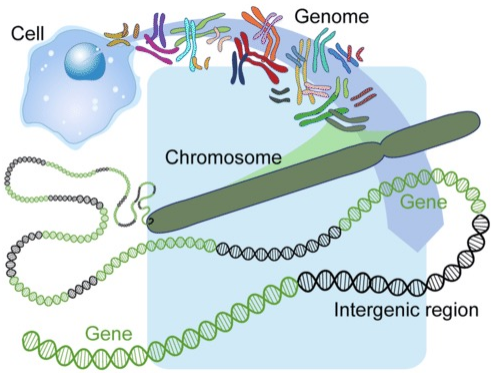


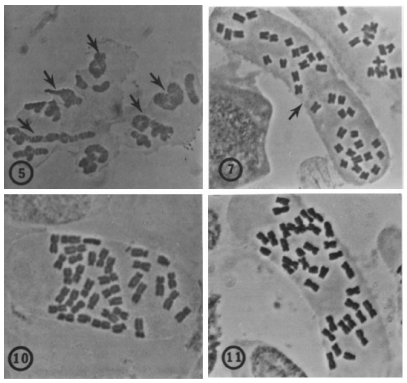


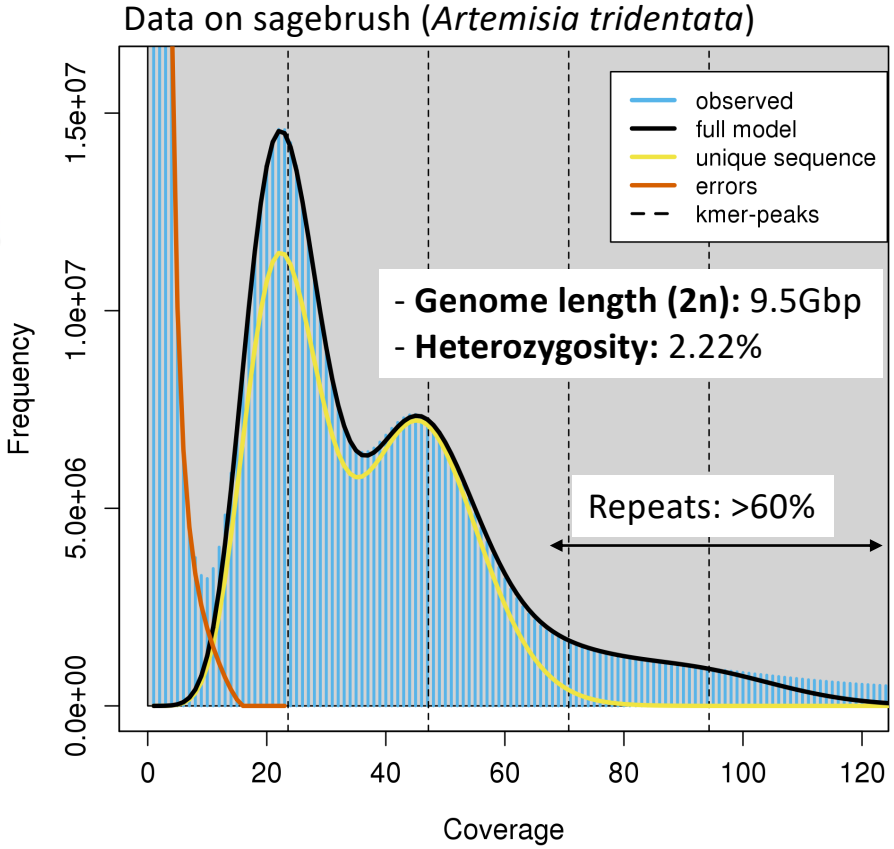


Chapter 1

Appreciate that **eukaryotic genomes contain extensive repetitive regions** of several different kinds. This provides a challenge for genome assembly!







Group Lab report

CHAPTER 1

Know the basic dogma that DNA is transcribed to RNA, which is translated to protein.

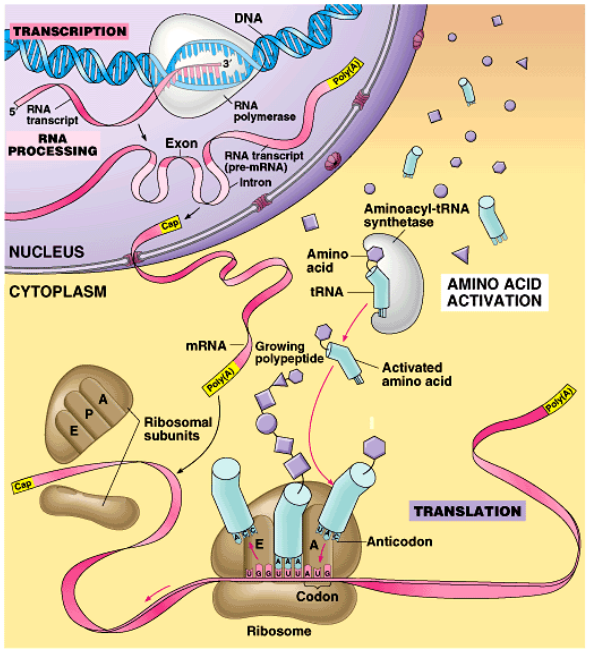
GENOME

↓ Transcription

TRANSCRIPTOME
RNA copies of the active protein-coding genes

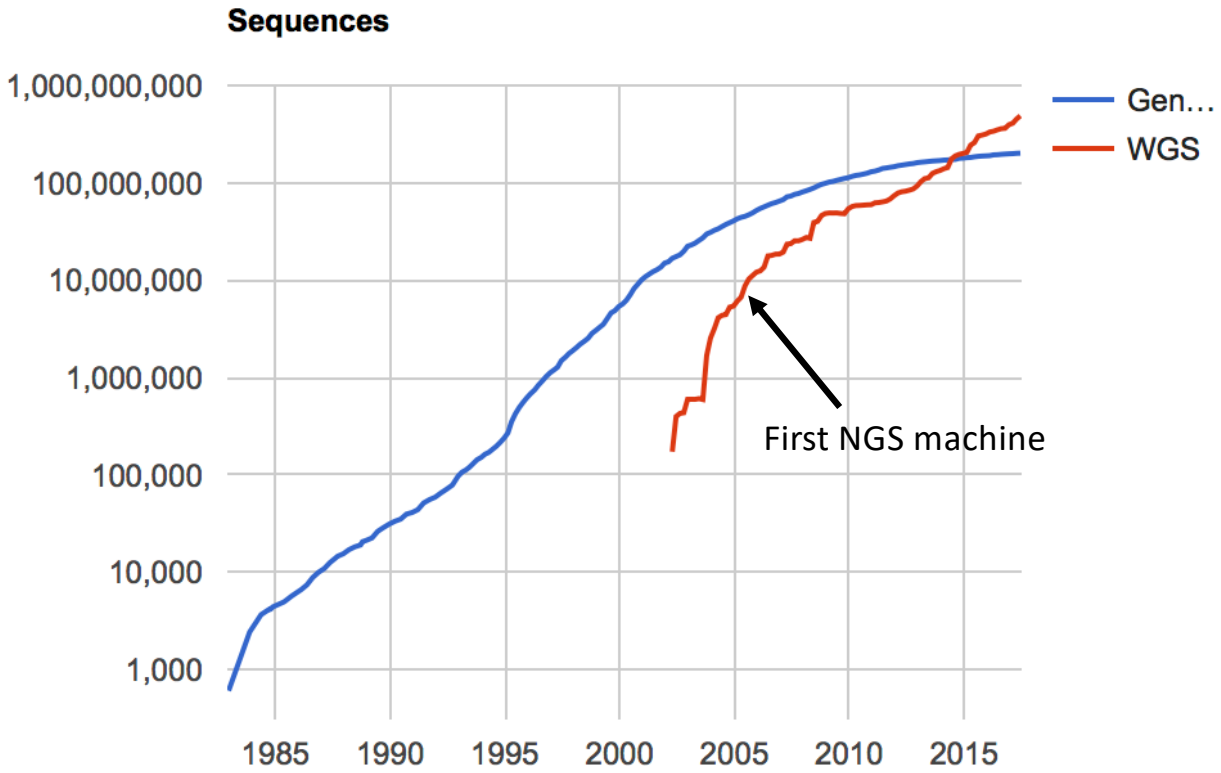
↓ Translation

PROTEOME
The cell's repertoire of proteins



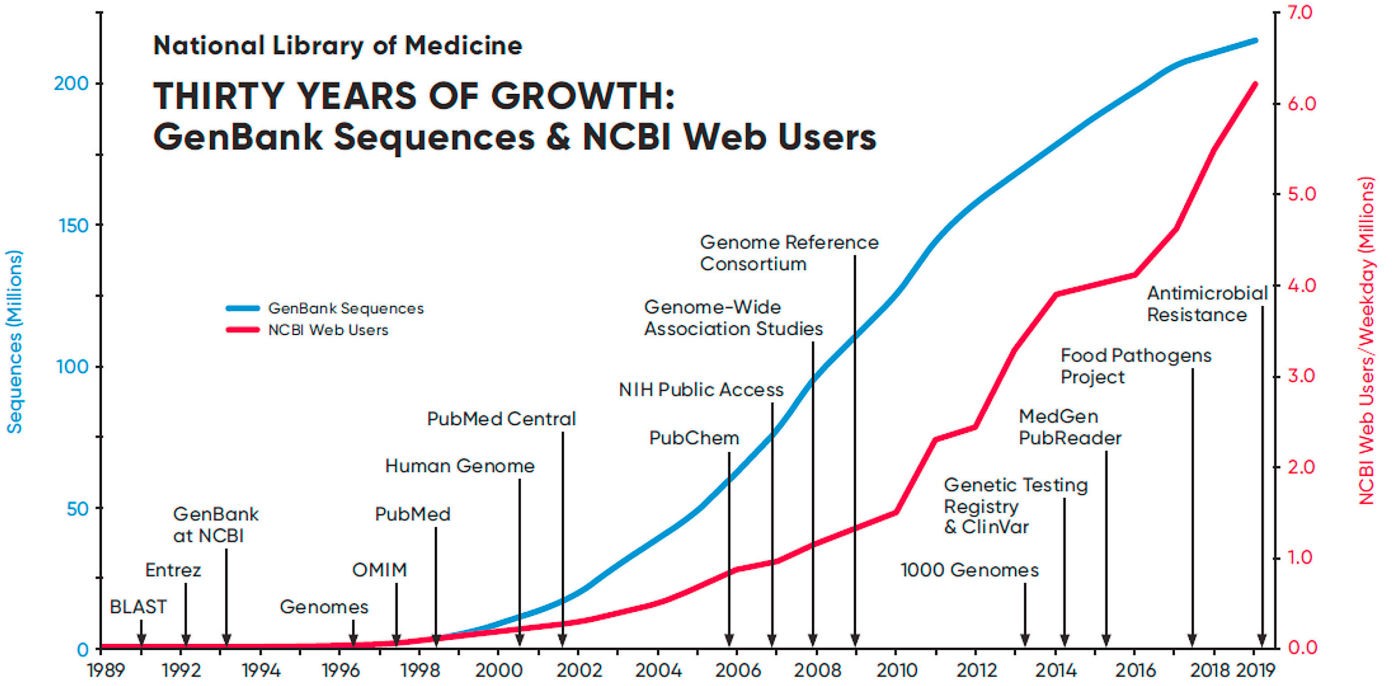
Chapter 1

Perfect moment to study genomics, a plethora of data have been generated in the last 15 years.



CHAPTER 1

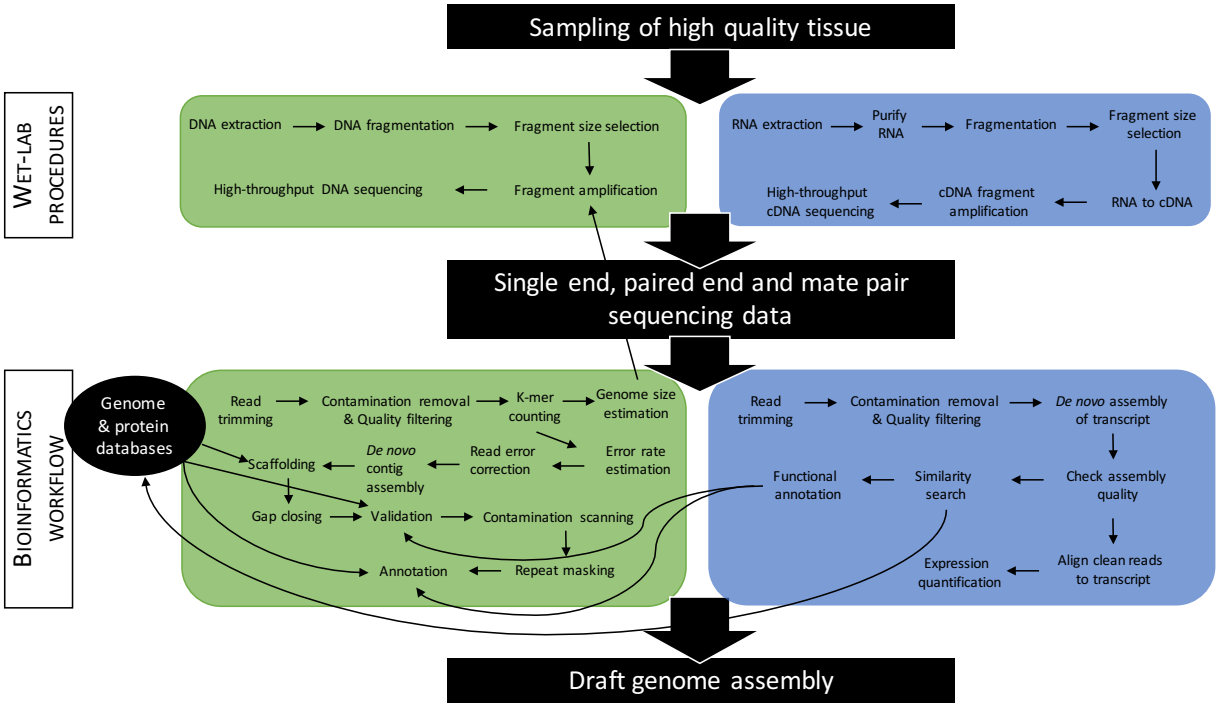
Perfect moment to study genomics, a plethora of data have been generated in the last 15 years.



https://doi.org/10.1016/j.gfs.2020.100411

CHAPTER 2

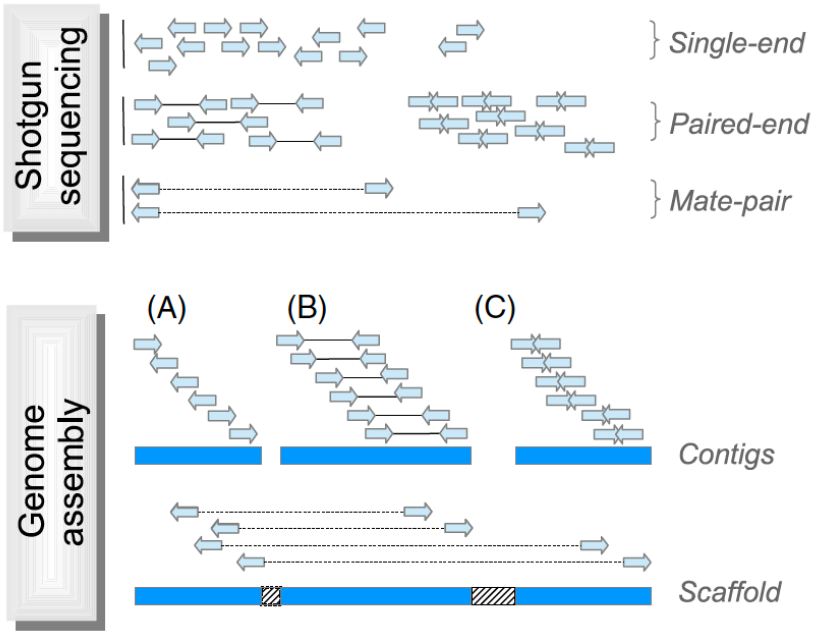
Independently on the approach used to produce a whole-genome sequence, all projects share the same major steps:



Lexicon on website

CHAPTER 2

Learn the next-generation sequencing (NGS) jargon necessary to assemble and annotate genomes.



Mini-report 1

CHAPTER 2

Introduction to NGS platforms:

Chapter 3

Understand importance of computer science in:

- Producing raw sequence data (e.g. base-calling)
- Creating databases in molecular biology
- Archiving and curation of data
- Distributing data via the Internet
- Creating information-retrieval tools to allow effective mining of the data for research application

Mini-report 2

CHAPTER 3

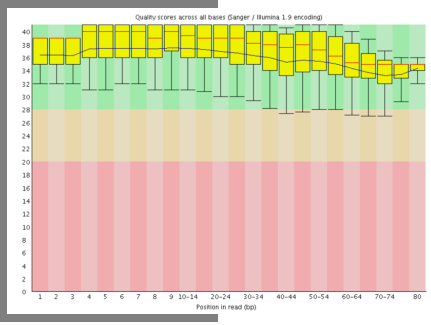
Gain knowledge on major molecular biology databases, which are key to genome assembly and annotation:

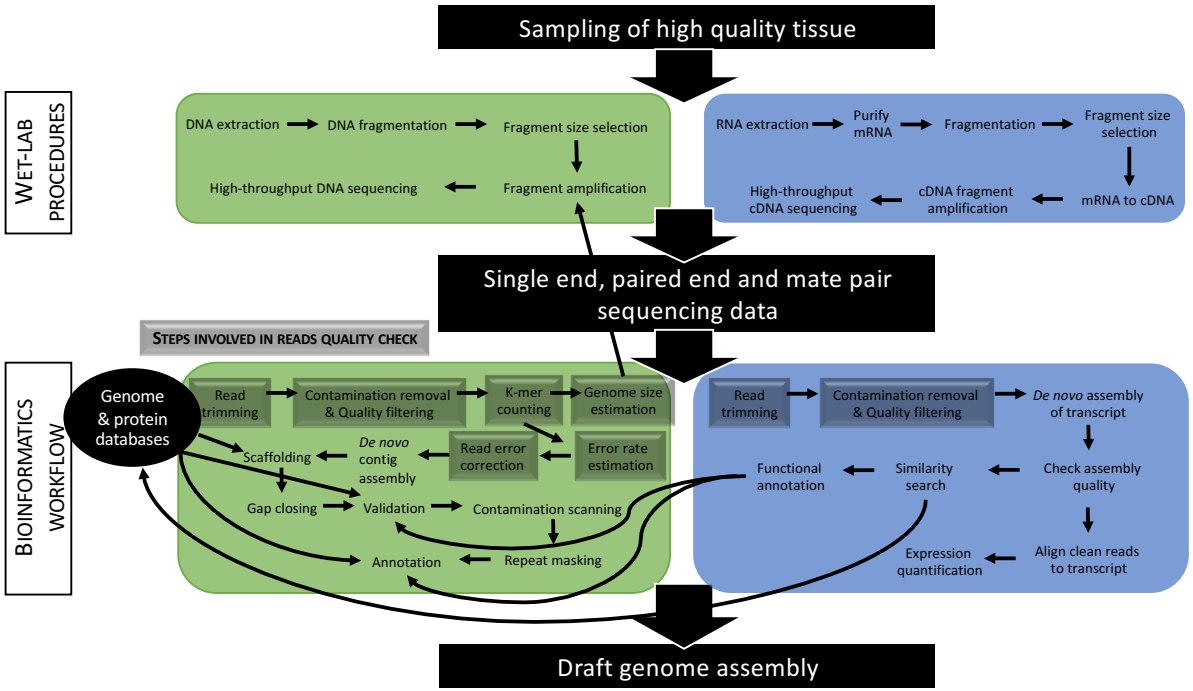
- Nucleic acid sequences databases (e.g. NCBI)
- Protein sequences databases (e.g. Swissprot)
- Gene ontology databases
- Metabolic pathways databases (e.g. KEGG)
- Specialized annotated genome portals

NCBI
National Center for
Biotechnology Information

CHAPTER 4

Learn how to conduct reads quality checks on raw NGS data.

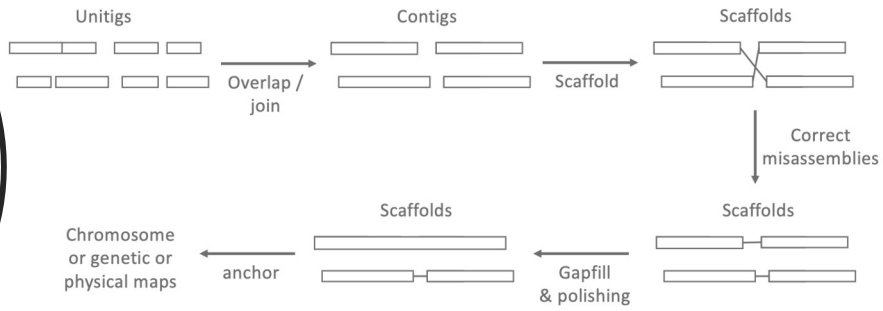




Chapter 4

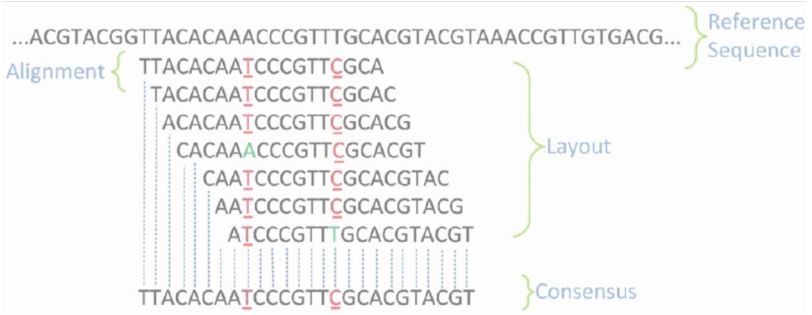
Understand key steps involved in producing a genome assembly and best strategies to get there.

de novo assembly



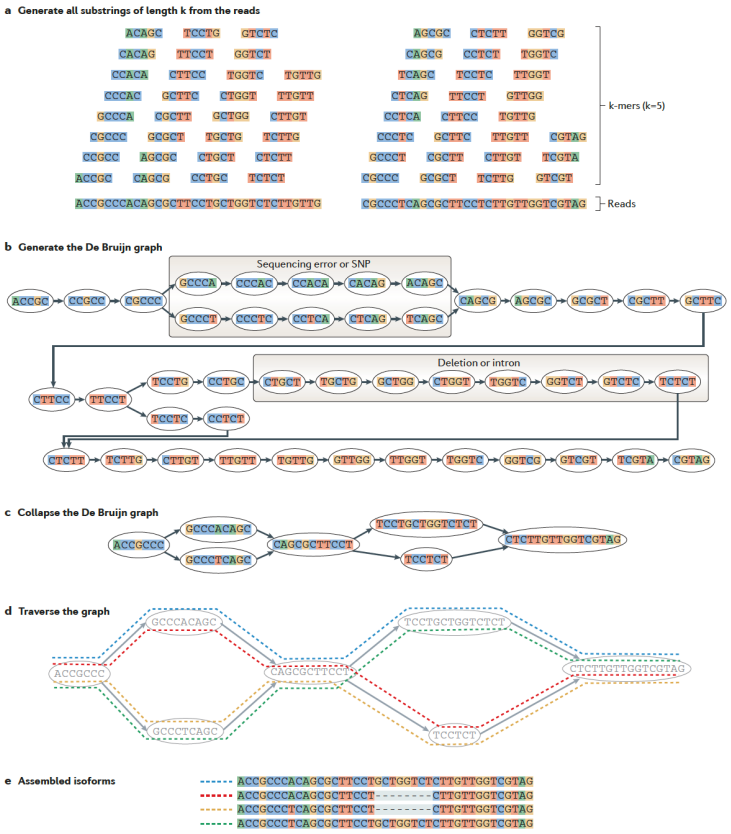
vs.

Referenced-based assembly



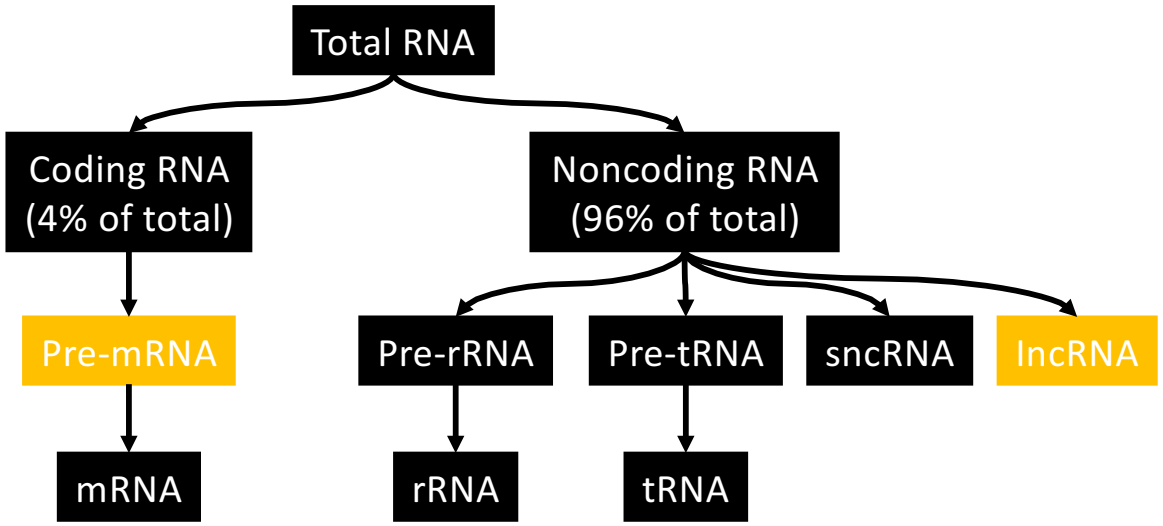
Chapter 4

Study de Bruijn graph procedure for *de novo* genome assembly.



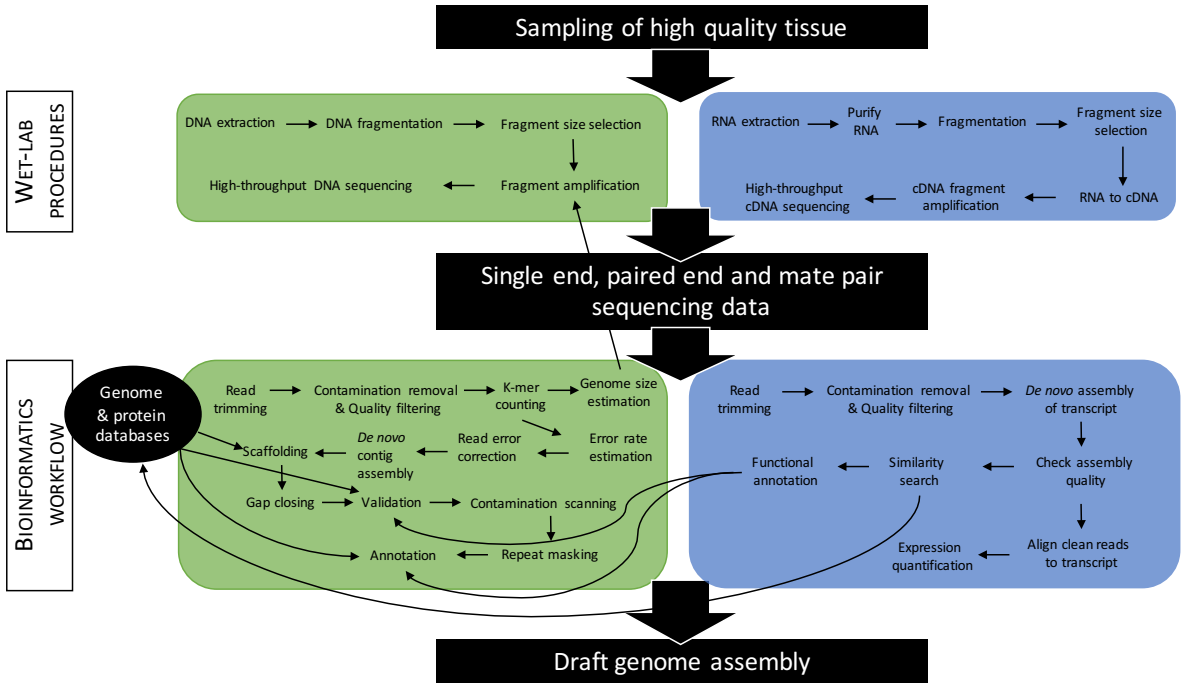
CHAPTER 5

Learn about the different types of RNA molecules in cell.



CHAPTER 5

Adapt wet-lab and bioinformatics workflow according to targeted RNAs.



CHAPTER 5

Study the link between the transcriptome and proteome via the genetic code.

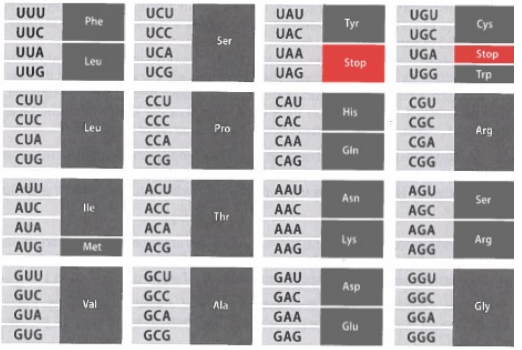
GENOME

↓ Transcription

TRANSCRIPTOME
RNA copies of the active protein-coding genes

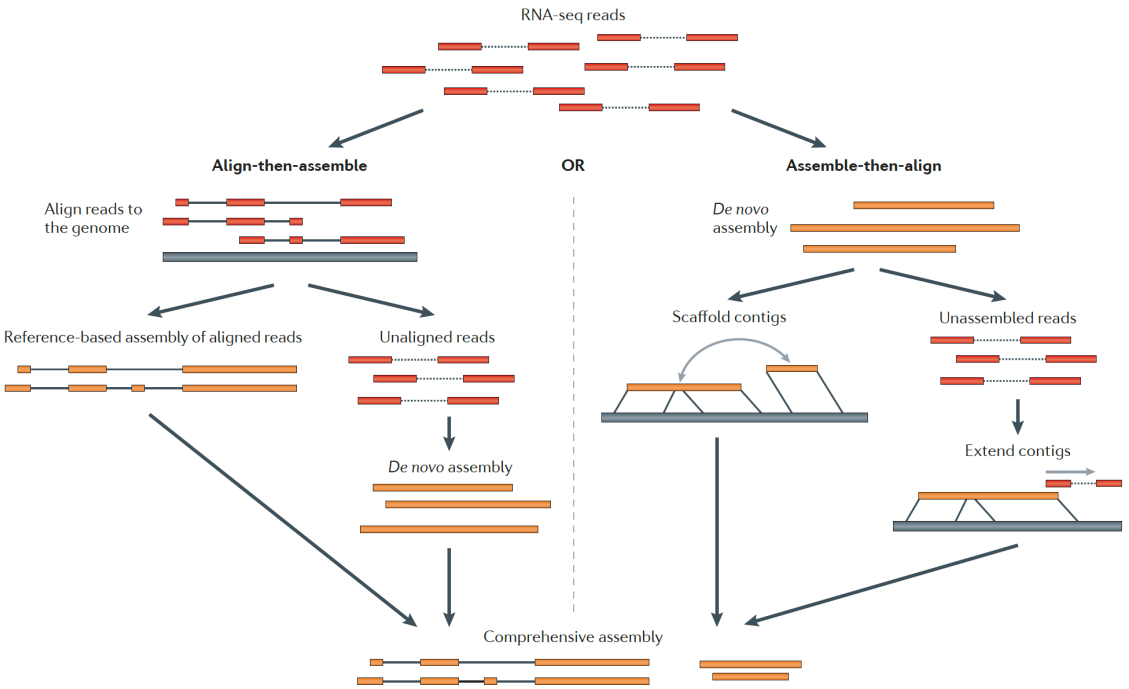
↓ Translation ←

PROTEOME
The cell's repertoire of proteins



Chapter 5

Understand key steps involved in **producing a transcriptome assembly** and best strategies to get there.



Chapter 6

- Learn what we exactly mean by the term "genome annotation".
- Know the following key stages of the structural genome annotation process:
  - Repeat identification
  - Evidence alignment (map transcriptome on genome)
  - *Ab initio* ("from the beginning”) gene prediction
  - Evidence-driven gene prediction (use external info to improve prediction of gene annotations)

CHAPTER 7

Be aware of challenges to obtain accurate data on gene functions.

Research article | Open Access

Massive parallel sequencing of mRNA in identification of unannotated salinity stress-inducible transcripts in rice (*Oryza sativa* L.)

Hiroshi Mizuno †, Yoshihiro Kawahara †, Hiroaki Sakai, Hiroyuki Kanamori, Hironobu Wakimoto, Harumi Yamagata, Youko Oono, Jianzhong Wu, Hiroshi Ikawa, Takeshi Itoh and Takashi Matsumoto

†Contributed equally

BMC Genomics 2010 **11**:683
https://doi.org/10.1186/1471-2164-11-683
Received: 20 April 2010 | Accepted: 2 Dece

A study on the response of rice to salt stress discovered 649 genes that were missing from the rice annotation!

CHAPTER 7

Study and compare available pipelines to conduct automated genome annotations.

MAKER

Last Software Update

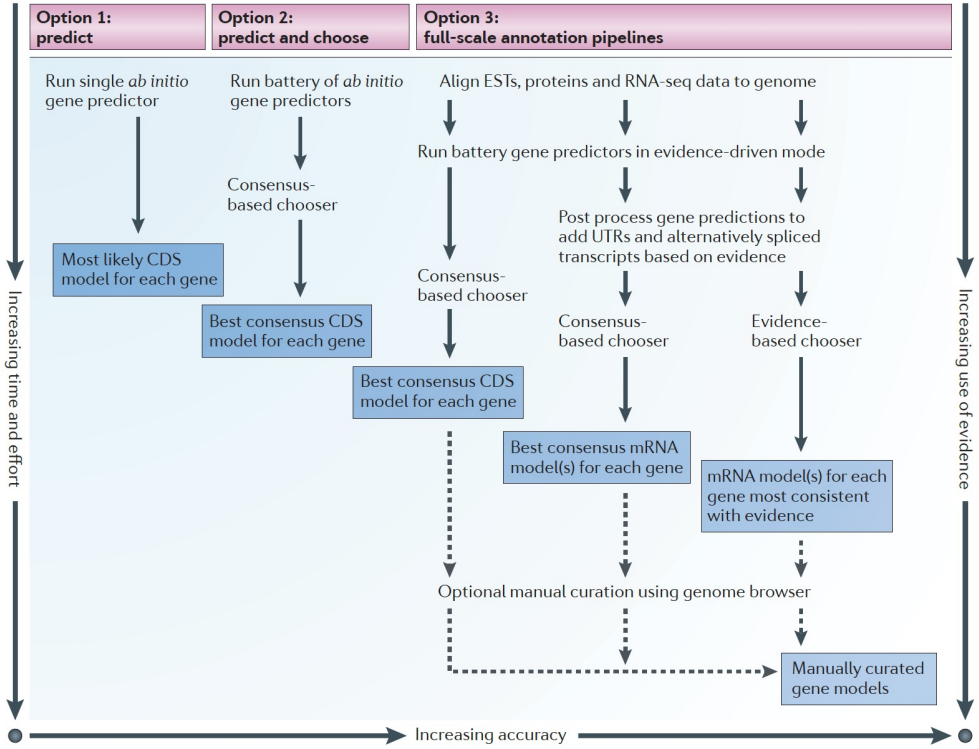
v2.31.9 (Dec 16, 2016)

Overview

MAKER is a portable and easily configurable genome annotation pipeline. Its purpose is to allow smaller eukaryotic and prokaryotic genome projects to independently annotate their genomes and to create genome databases. MAKER identifies repeats, aligns ESTs and proteins to a genome, produces ab-initio gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values. MAKER is also easily trainable: outputs of preliminary runs can be used to automatically retrain its gene prediction algorithm, producing higher quality gene-models on seusequent runs. MAKER's inputs are minimal and its ouputs can be directly loaded into a GMOD database. They can also be viewed in the Apollo genome browser; this feature of MAKER provides an easy means to annotate, view and edit individual contigs and BACs without the overhead of a database. MAKER should prove especially useful for emerging model organism projects with minimal bioinformatics expertise and computer resources.

Chapter 7

Review approaches to assess annotation quality.



Chapter 8

- Learn to read and present a scientific paper reporting new genomic or transcriptomic data.
- Become an expert on a specific area of comparative genomics and share it with your peers.