Introduction to Genome Versions

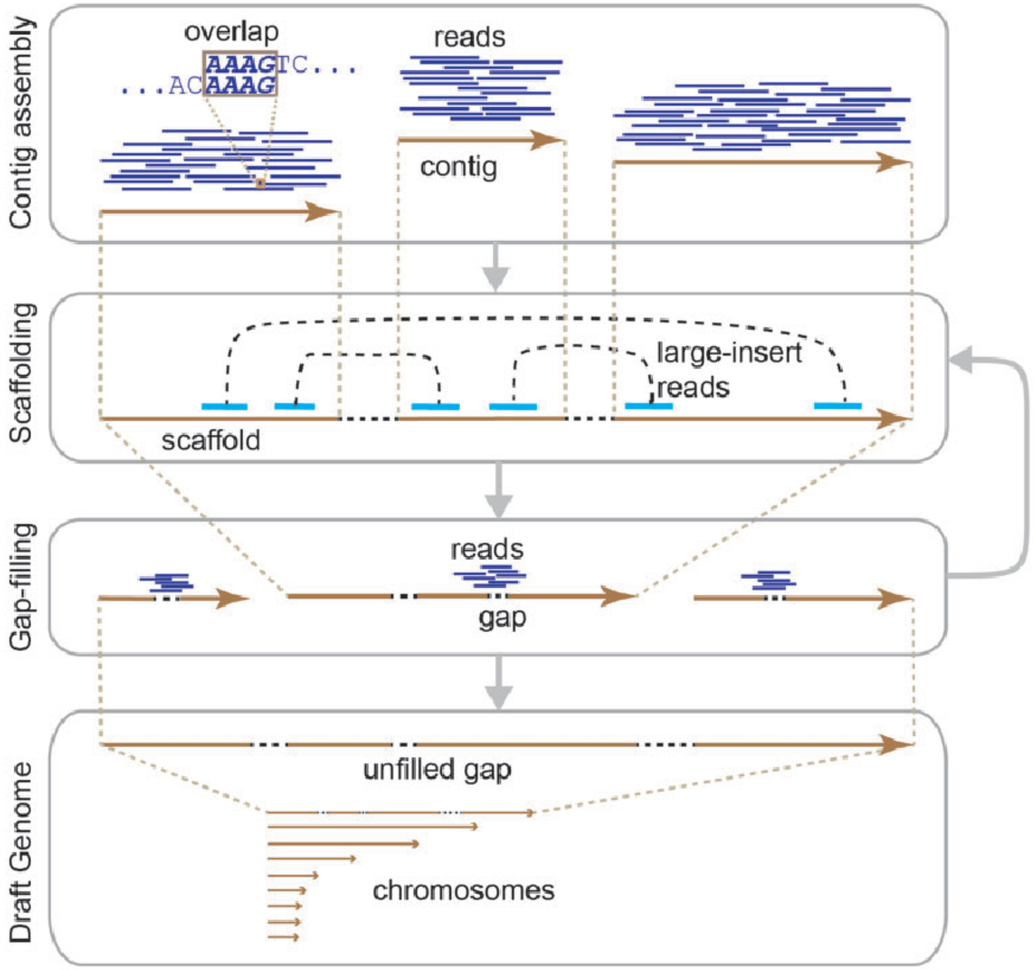
BIO392

September 24, 2020

The Reference Genome

- First Human Genome was finished in 2003.
- A result of the Human Genome Project.
- Sanger sequencing, multiple individuals.
- Currently maintained by the Genome Reference Consortium.

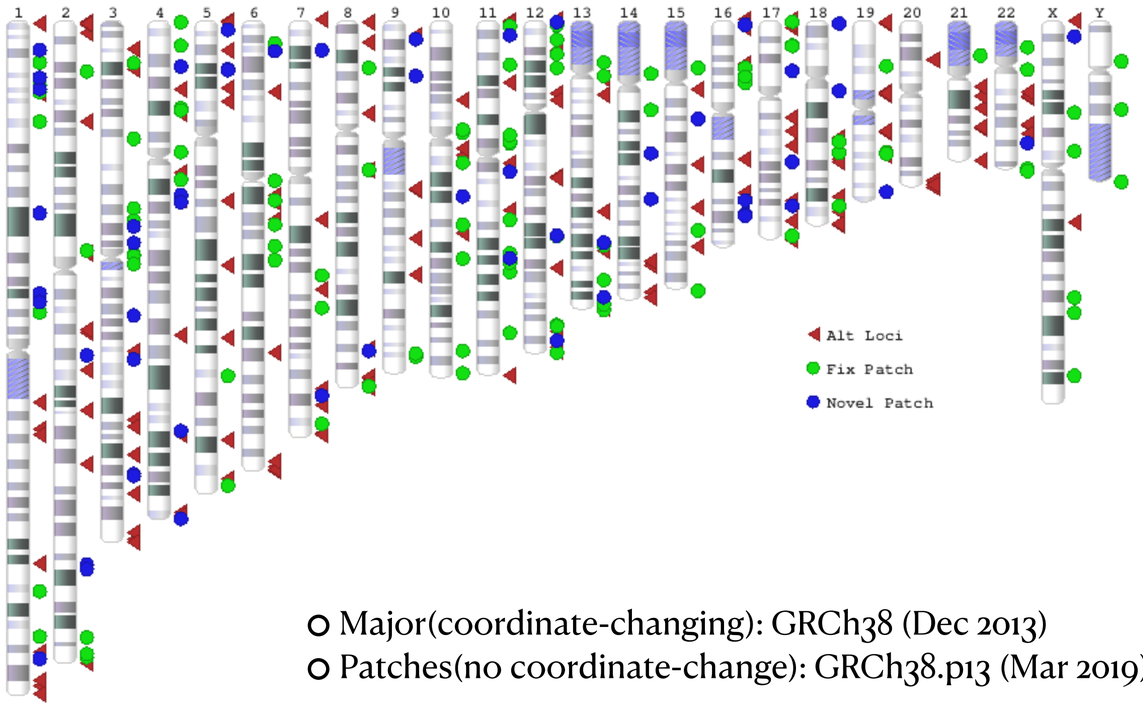
General workflow of the de novo assembly of a whole genome



Sohn, Jang-il and Jin-Wu Nam. “The present and future of de novo whole-genome assembly.”

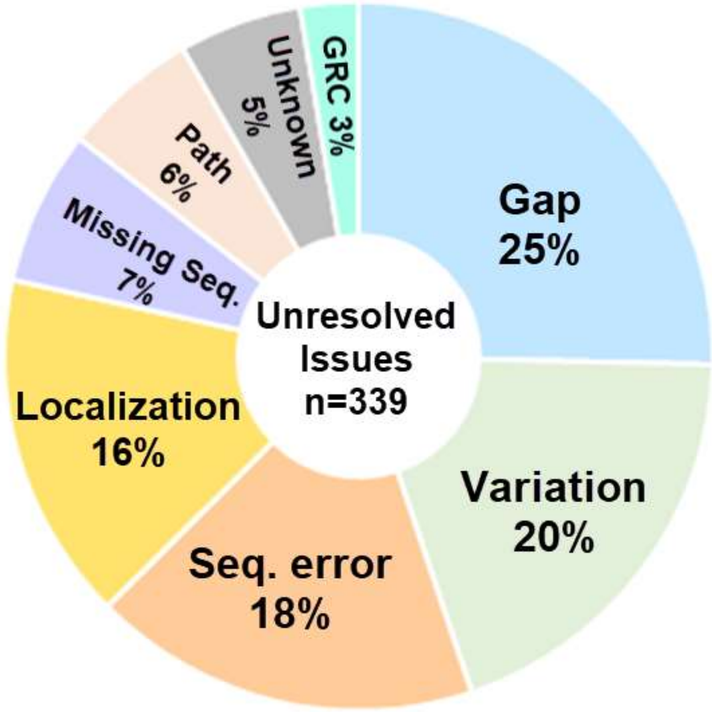
A continuous evolution

Reference assembly updates



- Major(coordinate-changing): GRCh38 (Dec 2013)
- Patches(no coordinate-change): GRCh38.p13 (Mar 2019)

Unresolved issues



Exercise

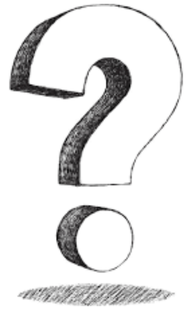
Most species have more than one versions of the reference genome.

Please find out:

1. The name and time of the latest version for Human, Mouse and E Coli.
2. The name and time of the first version for Human, Mouse and E Coli.
3. How many reference genomes were released in total for Human, Mouse, and E Coli.

Challenges

- For human genomic data generated before 2013 , it was for sure based on an older version.
- Most data generated a few years ago were still using the old assembly.
- It creates a big pitfall when using genomic data



Can you imagine some troubles it may bring to your study or research?

Exercise

What do you think of the difference between genome versions?

1. Find out the difference in chromosome length between the latest patch of hg38 and the last patch of hg19. The name and time of the first version for Human, Mouse and E Coli.

2. With your favorite gene, find out its position in hg38 and hg18.

Introduction to the UCSC Genome Browser

BIO392

September 24, 2020

About the USCS Genome Browser

- Hosted by the University of California, Santa Cruz.
- A graphical visualization tool for genomic data
- A broad collection species and annotations, along with a large suite of tools.
- There are other similar websites/tools.

The USCS Genome Browser

In this short introduction, we will learn:

1. To use the basics functions.
2. To switch between genome assemblies.
3. To use build-in annotation tracks.

Exercise

1. Show gene TP53 in the genome browser.
2. Where is this gene? (chromosome, cytoband, and exact start and end positions)
3. How many isoforms does it have?
4. How many exons does it have?
5. What the size of its longest exon? (roughly)
6. Find the three closest genes in upstream and downstream, respectively.

Exercise

1. Switch to hg19 and find TP53.
2. What is the start and end positions?
3. Switch to zebrafish, can you find TP53?
4. Switch to Fruitfly, can you find TP53?

Exercise

1. Switch to hg19 and find TP53.
2. What is the start and end positions?
3. Switch to zebrafish, can you find TP53?
4. Switch to Fruitfly, can you find TP53?

Exercise

Exploring annotation tracks

1. Compare human TP53 with other species, how similar are they?
2. Find out the conservation regions of TP53
3. Find out the frequent mutations of TP53 in cancer
4. Does it reveal anything? Is it what you expected?

Introduction to Genome Liftover

BIO392

September 24, 2020

About Liftover

- Convert data between different genome versions.
- Best strategy is to re-align the original data to the target genome version.
  - Availability of the data
  - Computational workload
- A practical solution is to convert using a map table.
  - Information lost
- Available tools

Exercise

Liftover with UCSC Genome Browser

1. Down-lift: TP53 from hg38 to hg19
2. Up-lift: TP53 from hg19 to hg38
3. Cross-species-lift: TP53 from human to mouse
4. Multi-step-lift: TP53 from hg38 to hg 18

Exercise

Liftover with UCSC Genome Browser

1. Liftover multiple positions with a BED file.
2. Lift a larger range and interpret the result.
3. Limitations of the liftover.

After-class Exercise

(I) More on liftover

1. Convert other files to the BED format.
2. Use the liftOver program locally.
3. Use segment_liftover

(II) A simple research

1. Find three other genome resources.
2. Write about their unique features and use cases.