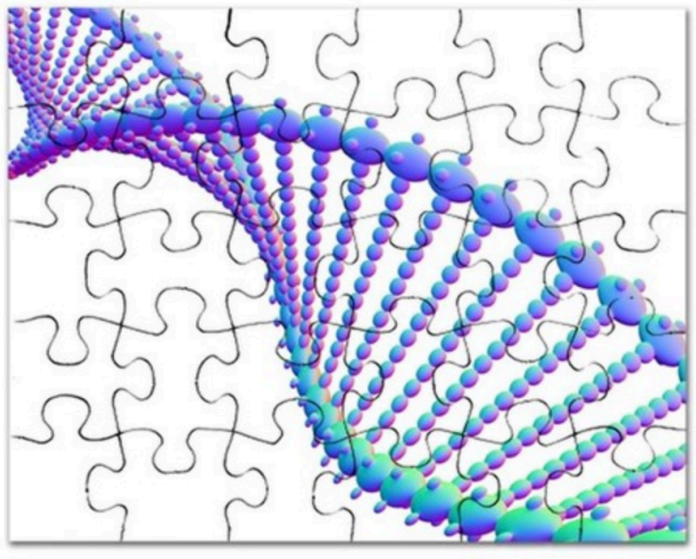
De novo assembly

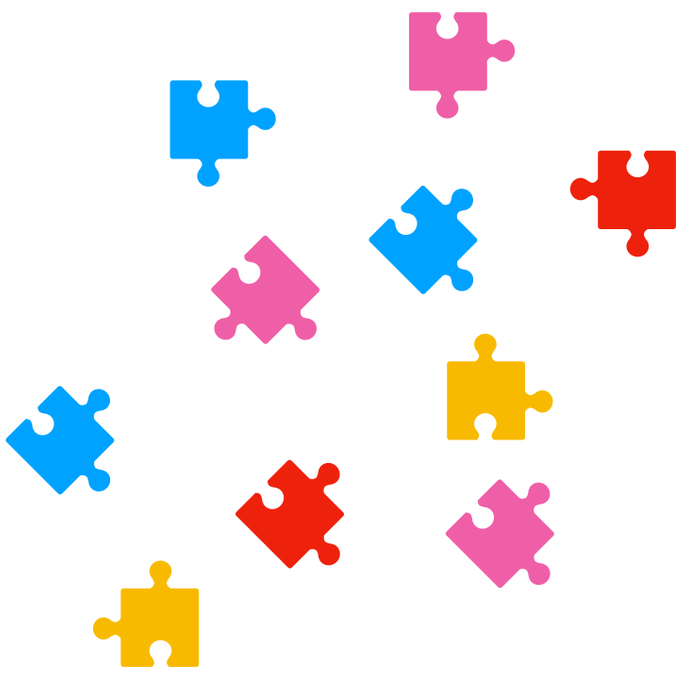
Dr Cristina Venturini

Genomic jigsaw puzzle

Reference-based jigsaw

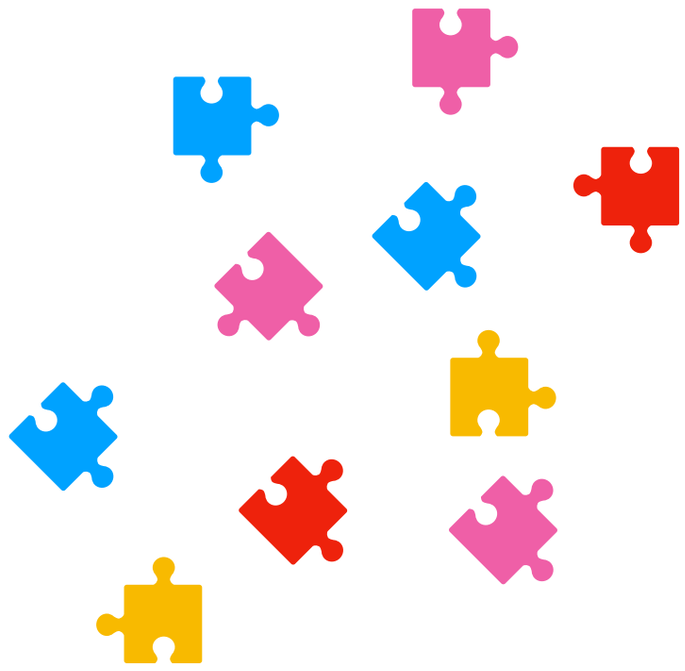
We have the box with the picture!!!





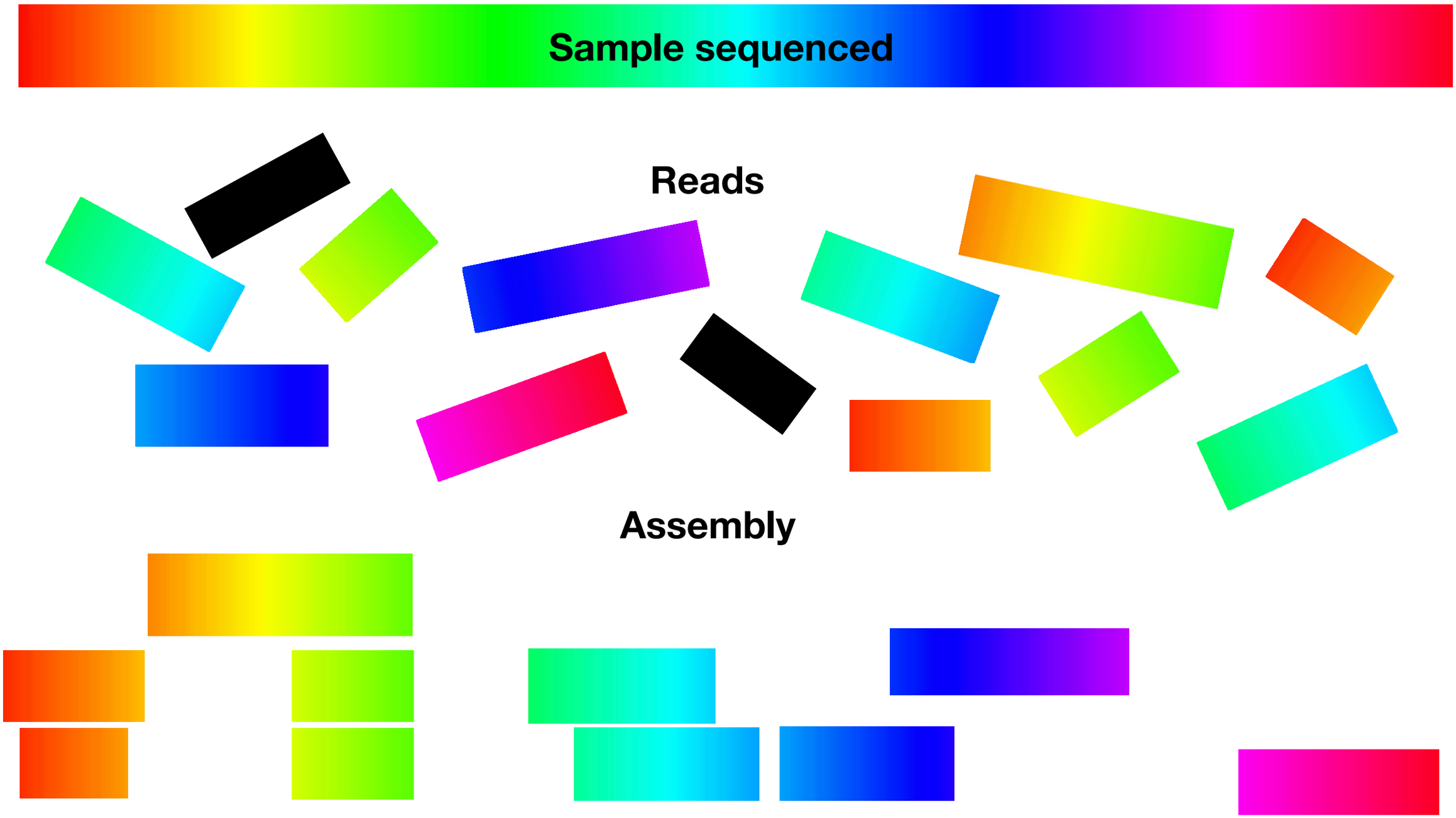
De novo jigsaw

?



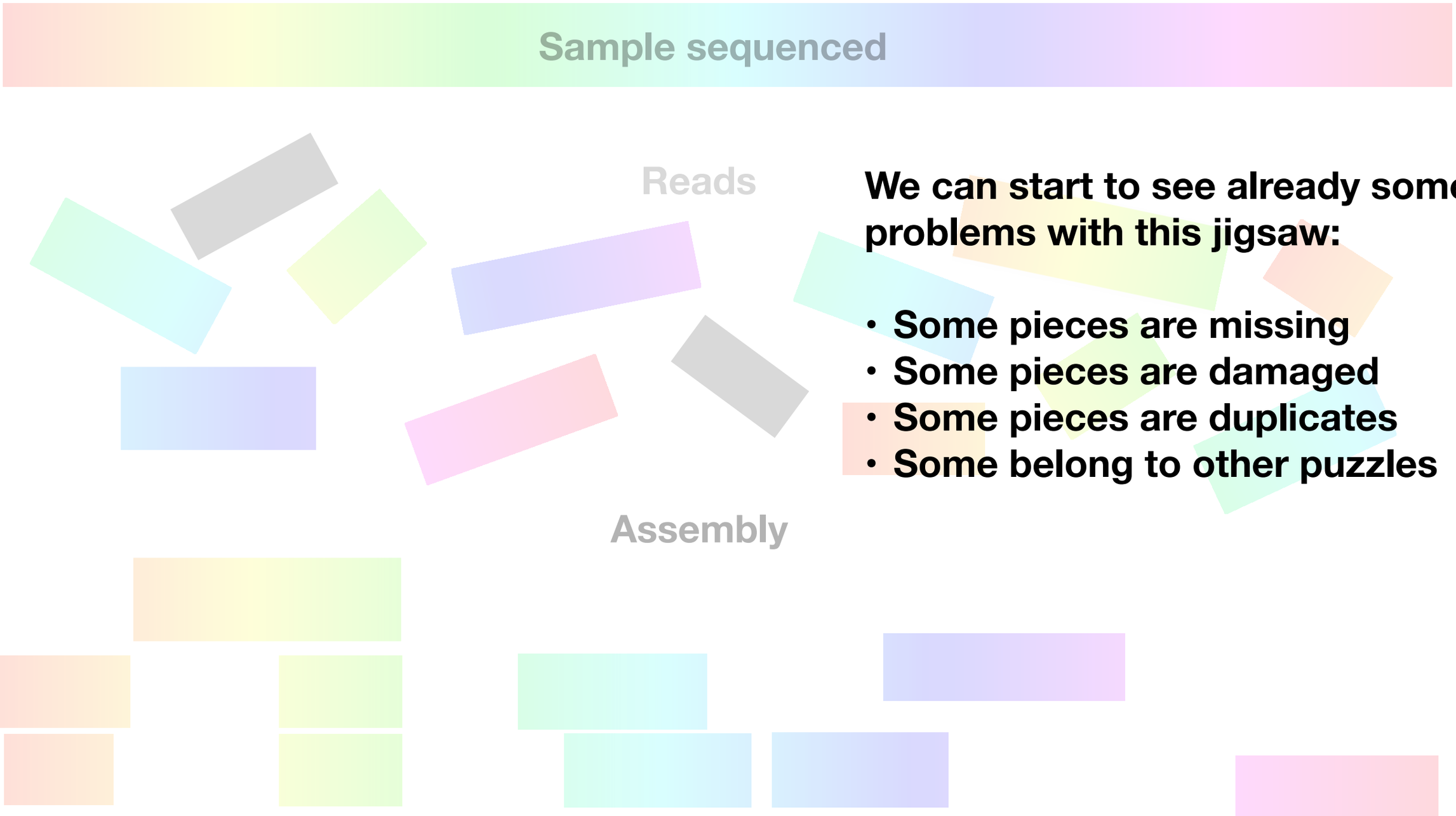
What is de novo?

- De novo assembly: the process of reconstructing sample sequence(s) without any guide reference(s)



What is de novo?

- De novo assembly: the process of reconstructing sample sequence(s) without any guide reference(s)



We can start to see already some problems with this jigsaw:

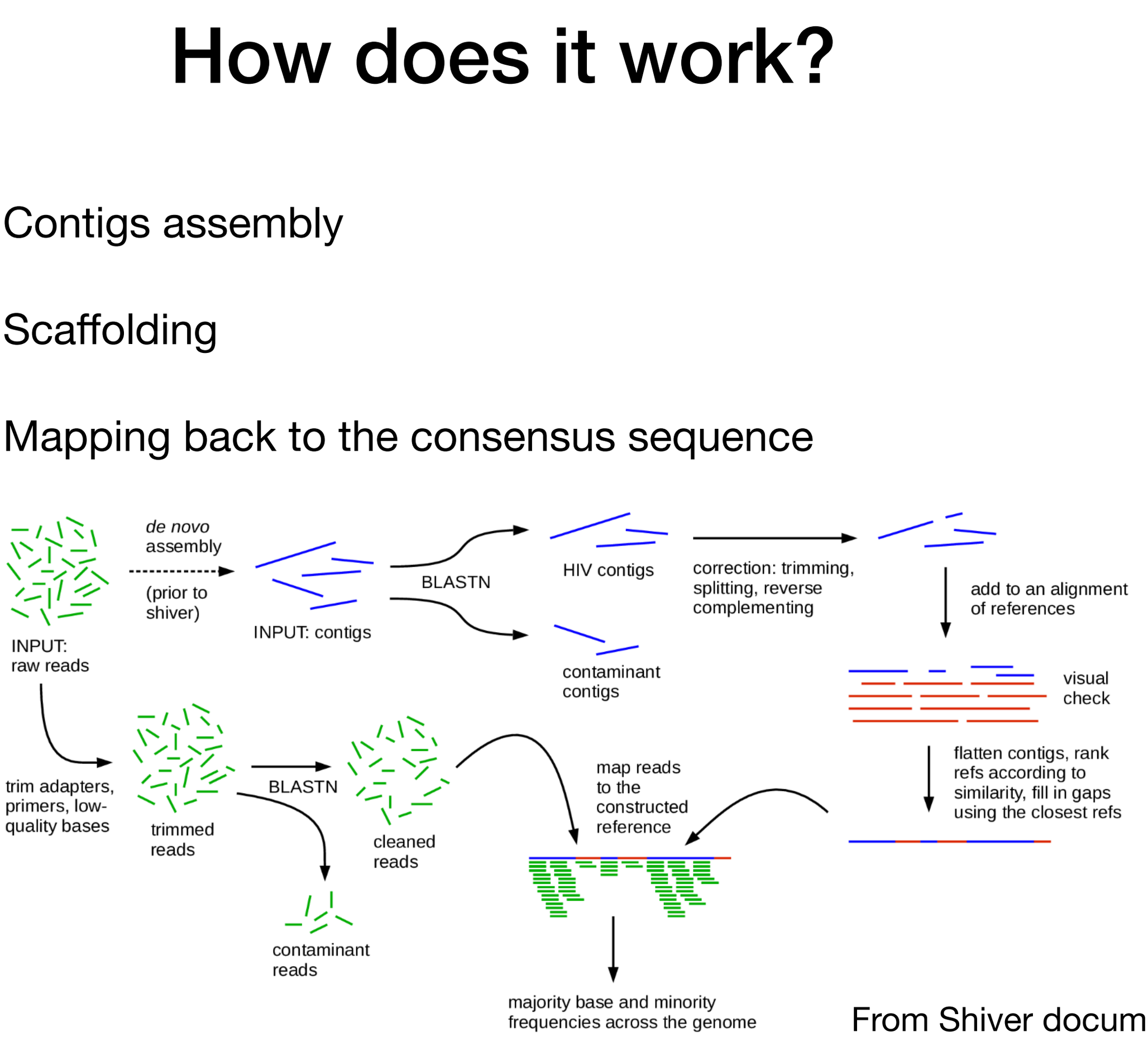
- **Some pieces are missing**
- **Some pieces are damaged**
- **Some pieces are duplicates**
- **Some belong to other puzzles**

Why do we do de novo assembly?

- **We don’t have a suitable reference for this virus:**
  - Poorly studied/new (2019-nCoV)
  - Hard to sequence
  - Highly divergent species (many RNA viruses)
- **We don’t know which virus we are looking at:**
  - Metagenomic analysis
  - Unknown aetiological agent of disease

How does it work?

- Contigs assembly
- Scaffolding
- Mapping back to the consensus sequence



From Shiver documentation

De Brujin graph approach

Reads

k=3 (this can vary!!)

ATTA: ATT —> TTA

GATT: GAT —> ATT

TACA: TAC —> ACA

TTAC: TTA —> TAC

GAT

ATT

TTA

TAC

ACA

GATTACA

How does it work?

MAY
RCE
YOU
MAY
TH_
BE_
_WI
_BE
AY_
_YO
_WI
CE_
_YO
HE_
FOR
ITH

MAY
MAY
AY_

_WI
_WI
ITH
TH_

FOR
RCE
CE_

THE

_YO
_YO
YOU

_BE
BE_

MAY_ | _WITH_ | FORCE_ | THE | _YOU | _BE_

Contigs

MAY_ _WITH_ FORCE_ THE _YOU _BE_

MAY_ THE N FORCE_ _BE_ _WITH_ _YOU

Scaffolding

MAY_THE_FORCE_BE_WITH_YOU

Reference

- Contigs: continuos stretches of sequence

- Scaffolds: created by joining contigs together using additional information about the relative position and orientation of the contigs with reference to a genome.

Ingredients for a good assembly

- **Coverage:** high coverage is required
- **Read length:** this is a problem for repeats - reads must be longer than the repeats to correctly form an assembly graphs

- **Quality**

Softwares

- SPAdes: St.Petersburg genome assembler. Toolkit containing various assembly pipeline

- IVA: iterative virus assembler. Denovo assembler designed to assemble virus genomes that have no repeat sequences, using Illumina pairs sequenced from mixed populations at extremely high and variable depth

- Shiver: tool for mapping paired-end short reads to a custom reference sequence constructed using de novo assembled contigs

Practical

- SPAdes: contigs

- SHIVER: for scaffolding/gap filling/mapping back to HIV genome