Bioinformatics for NGS analysis

From Reads to Variants

Tim Yu, MD, PhD

Division of Genetics, Boston Children's Hospital
Dept of Neurology, MGH
Harvard Medical School & the Broad Institute

2013 NSGC Annual Education Conference

Bioinformatics for NGS analysis

From Reads to Variants

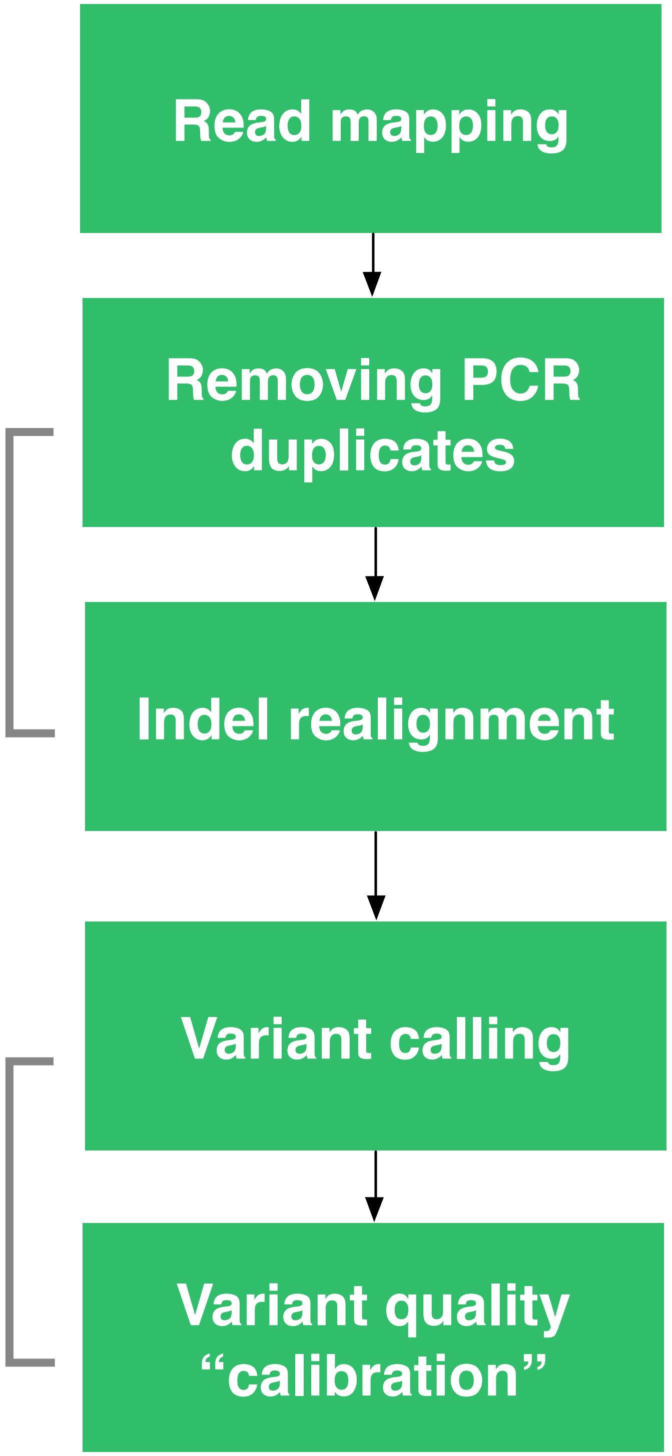
Tim Yu, MD, PhD

Disclosure:
Co-founder & Principal consultant,
Claritas Genomics

CLARITAS
GENOMICS

- Goal: arm you with concepts and vocabulary to understand how NGS data is analyzed, and to ask critical questions

A typical NGS processing pipeline



Mapping

Read mapping

BWA, Bowtie, Novoalign…

Cleaning

Removing PCR duplicates

Picard

Indel realignment

GATK, SRMA

Variant calling

Variant calling

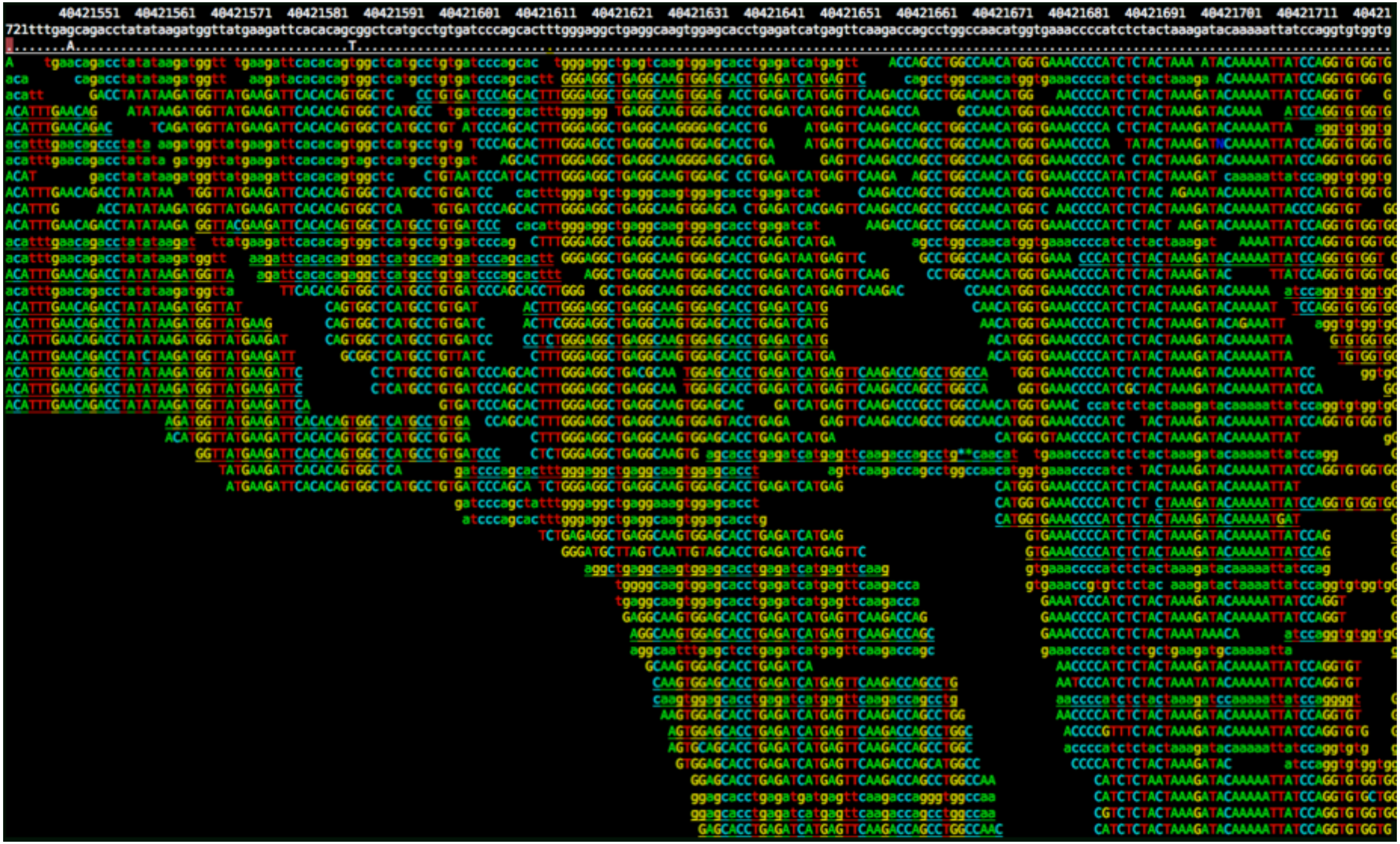
GATK, Dindel, etc.

Variant quality “calibration”

GATK, others

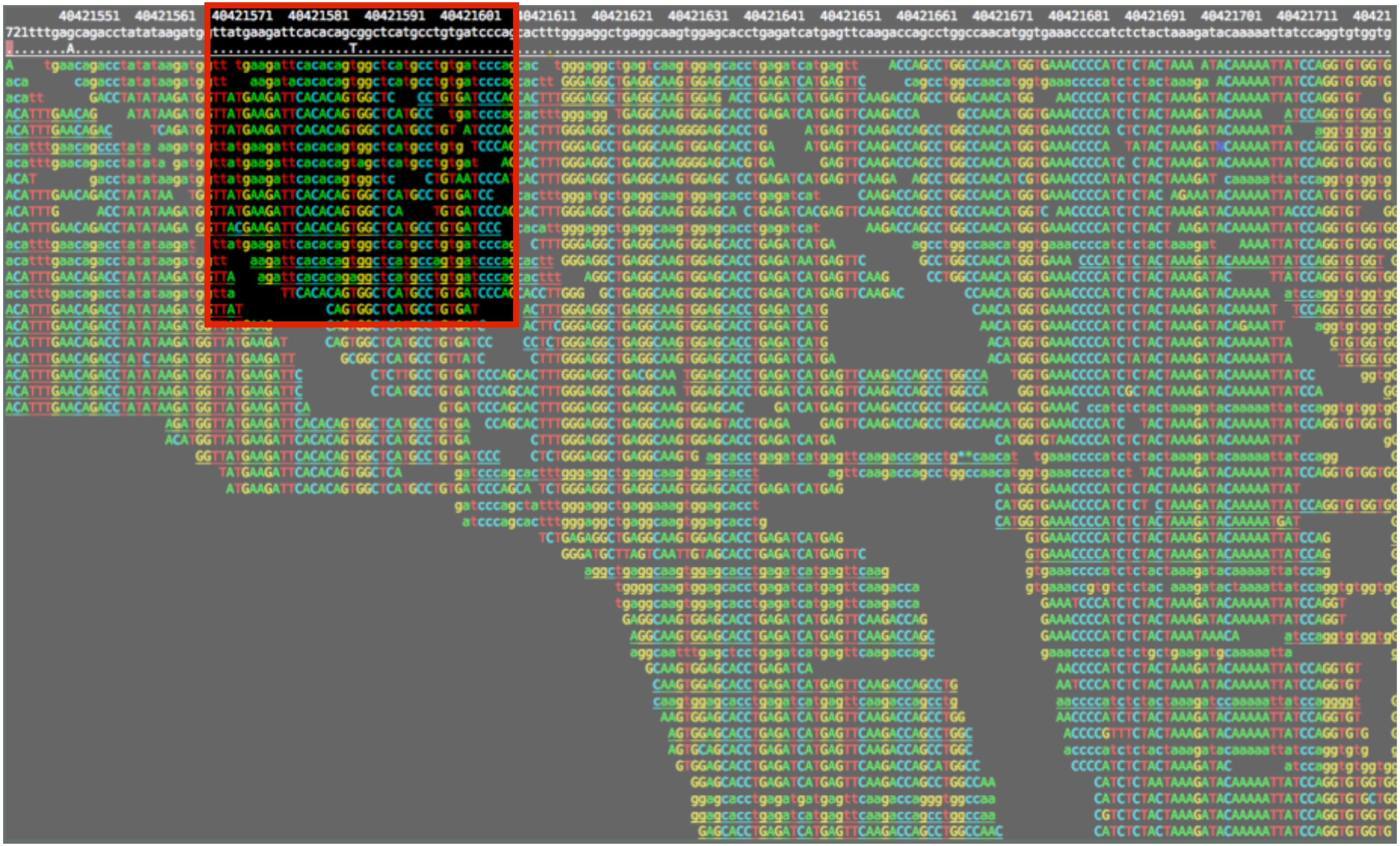
NGS analysis is, in principle, a two step process

1. Millions/billions of reads are mapped *en masse* to a reference genome



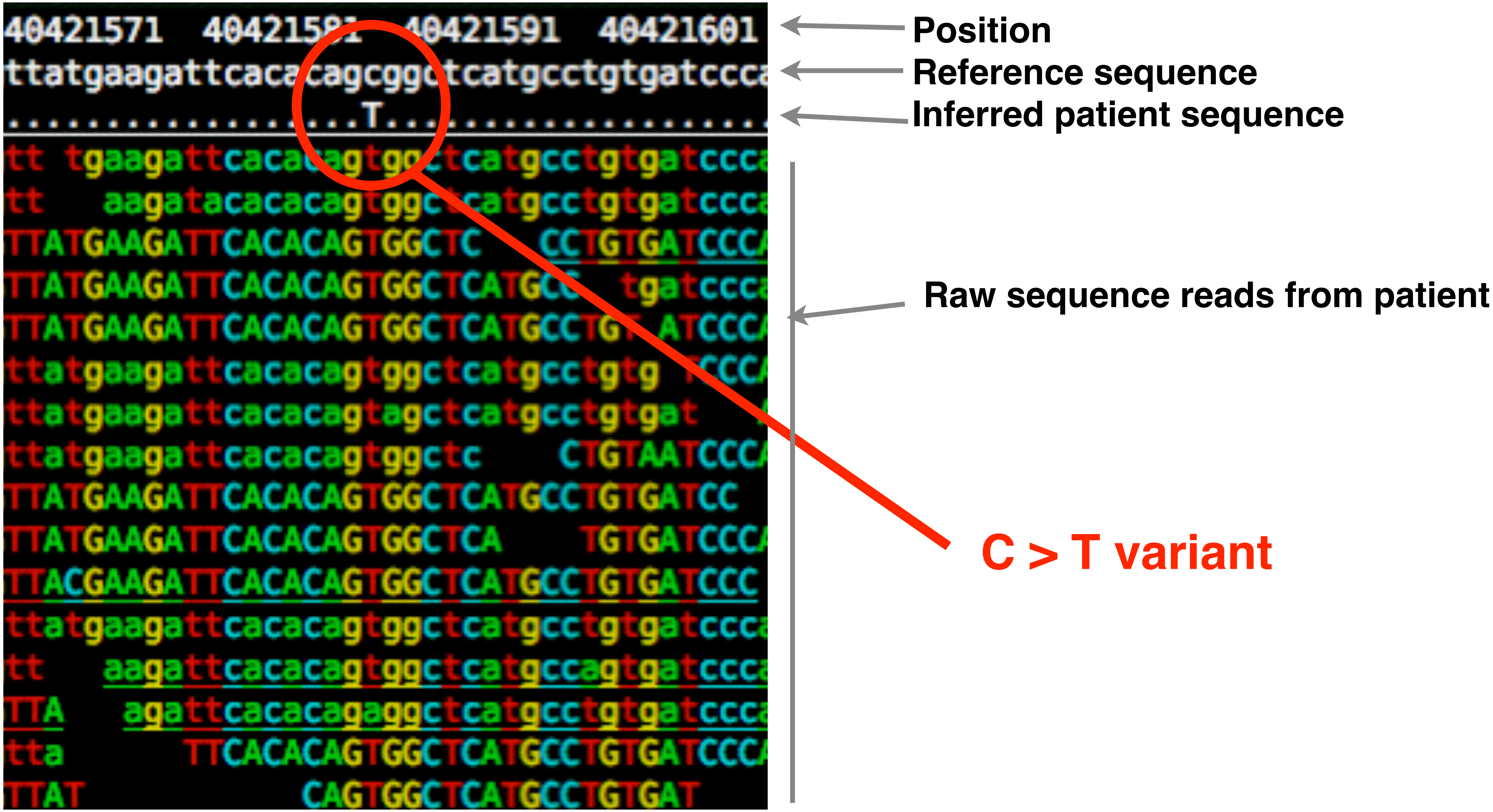
```
      40421551  40421561  40421571  40421581  40421591  40421601  40421611  40421621  40421631  40421641  40421651  40421661  40421671  40421681  40421691  40421701  40421711  40421
721tttgagcagacctatataagatggttatgaagattcacacagcggctcatgcctgtgatcccagcactttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatggtgaaaccccatctctactaaagatacaaaaattatccaggtgtggtg
........A..................................T.....................................................................................................................................
```

1. Millions/billions of reads are mapped *en masse* to a reference genome



```
        40421551  40421561  40421571  40421581  40421591  40421601  40421611  40421621  40421631  40421641  40421651  40421661  40421671  40421681  40421691  40421701  40421711  40421
721tttgagcagacctatataagatggttatgaagattcacacagcggctcatgcctgtgatcccagcactttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatggtgaaaccccatctctactaaagatacaaaaattatccaggtgtggtg
........A....................T.....................................................................................................................................................
```

2. Variants are detected when enough reads disagree with reference



```
40421571  40421581  40421591  40421601
ttatgaagattcacacagcggctcatgcctgtgatccca
..................T....................
tt tgaagattcacacagtggctcatgcctgtgatccca
tt   aagatacacacagtggctcatgcctgtgatccca
TTATGAAGATTCACACAGTGGCTC   CCTGTGATCCCA
TTATGAAGATTCACACAGTGGCTCATGCC  tgatccca
TTATGAAGATTCACACAGTGGCTCATGCCTGT ATCCCA
ttatgaagattcacacagtggctcatgcctgtg TCCCA
ttatgaagattcacacagtagctcatgcctgtgat   A
ttatgaagattcacacagtggctc    CTGTAATCCCA
TTATGAAGATTCACACAGTGGCTCATGCCTGTGATCC
TTATGAAGATTCACACAGTGGCTCA    TGTGATCCCA
TTACGAAGATTCACACAGTGGCTCATGCCTGTGATCCC
ttatgaagattcacacagtggctcatgcctgtgatccca
tt   aagattcacacagtggctcatgccagtgatccca
TTA   agattcacacagaggctcatgcctgtgatccca
tta      TTCACACAGTGGCTCATGCCTGTGATCCCA
TTAT           CAGTGGCTCATGCCTGTGAT
```

Position

Reference sequence

Inferred patient sequence

Raw sequence reads from patient

C > T variant

Complicating factors

- Mapping can be tricky
- Sequencing coverage is biased
- Not all variant calls are created equal
- Beyond SNPs and small indels

1. Mapping can be tricky

- **Easy**: Perfect matches to unique genomic regions A, B, and C

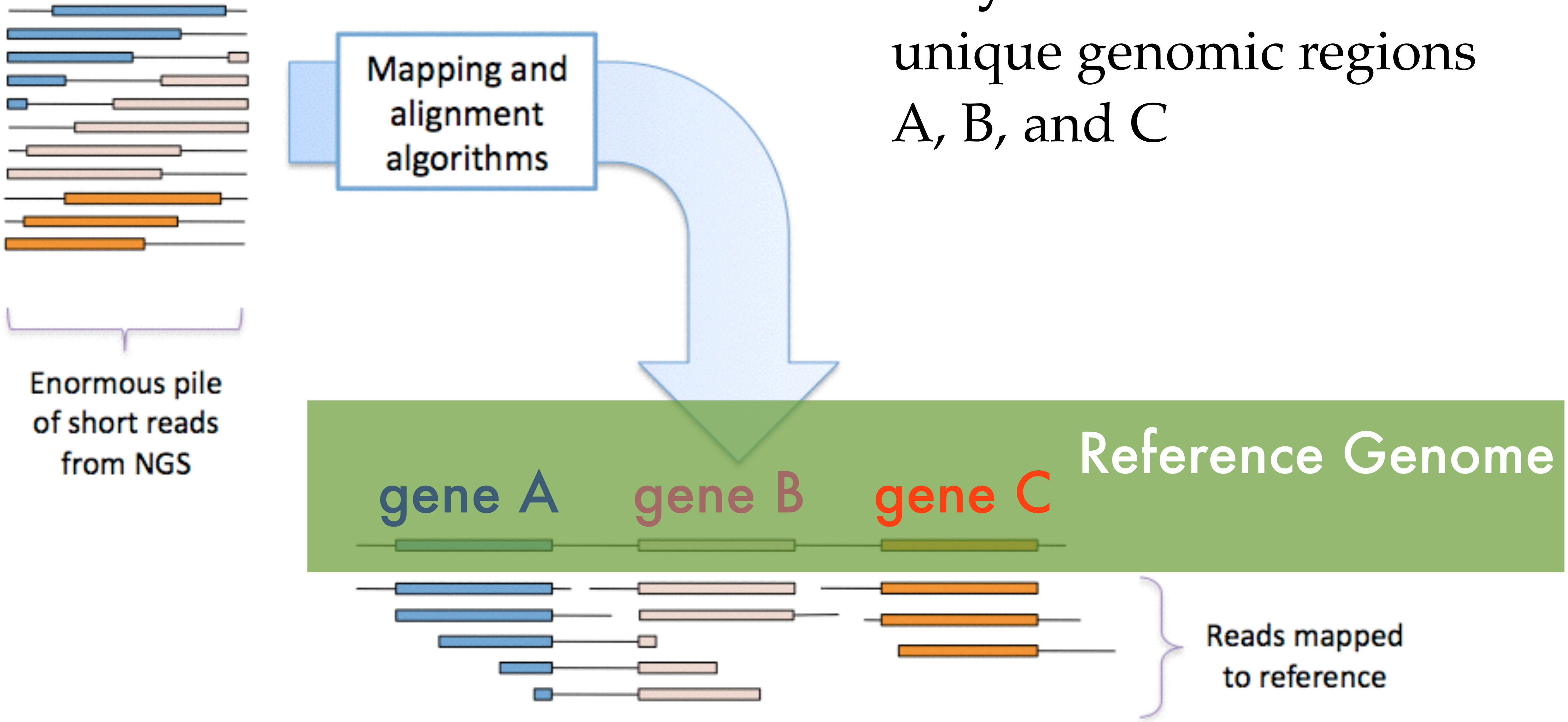
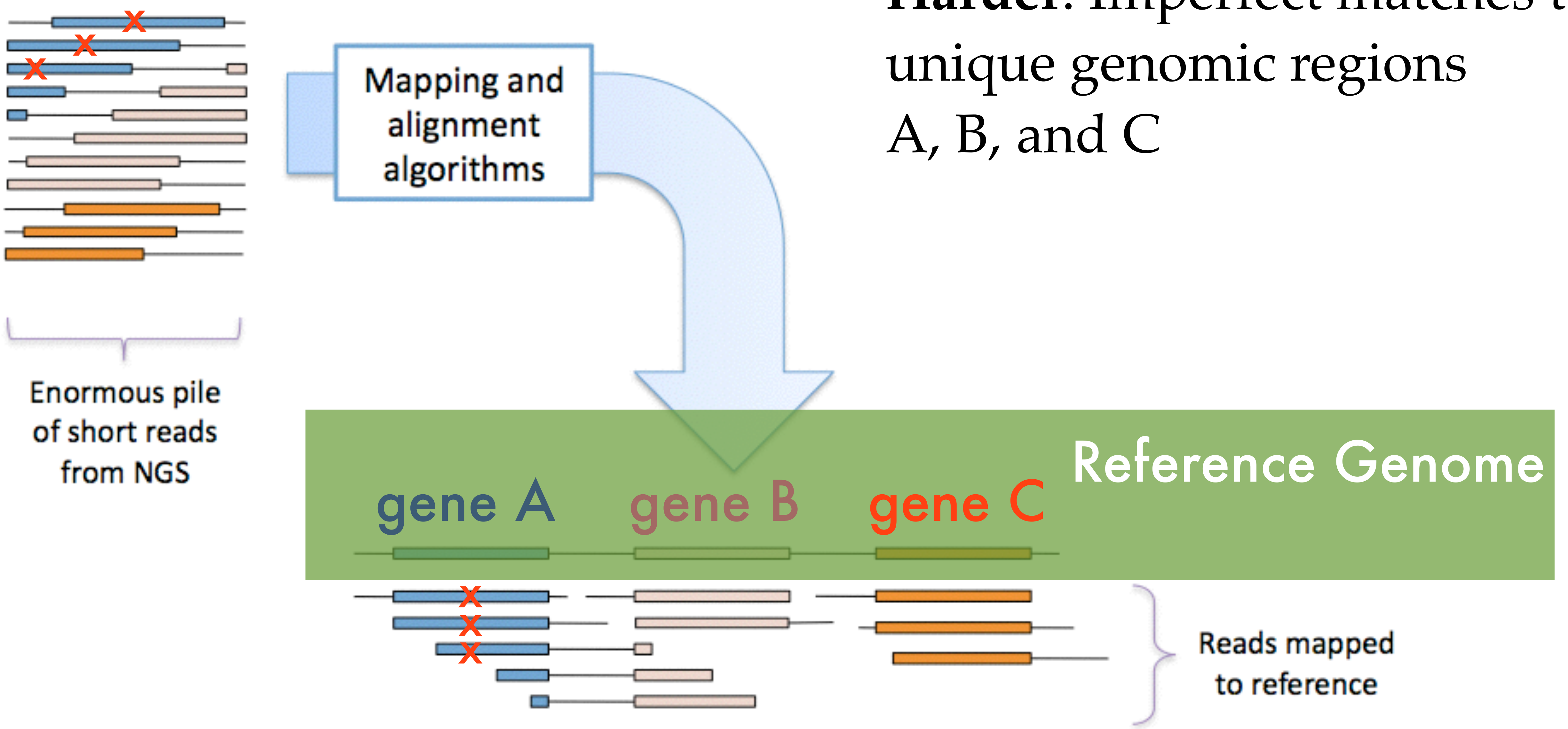

Mapping and alignment algorithms
Enormous pile of short reads from NGS
Reference Genome
gene A
gene B
gene C
Reads mapped to reference

Image credit: Broad GSA Platform

1. Mapping can be tricky


Mapping and alignment algorithms
Enormous pile of short reads from NGS
Reference Genome
gene A
gene B
gene C
Reads mapped to reference

- **Harder**: Imperfect matches to unique genomic regions A, B, and C

Image credit: Broad GSA Platform

1. Mapping can be tricky

- **Hardest:** Mapping to related genomic regions:
  - Gene families
  - Pseudogenes
  - Repeats / segmental dups
  - CNVs

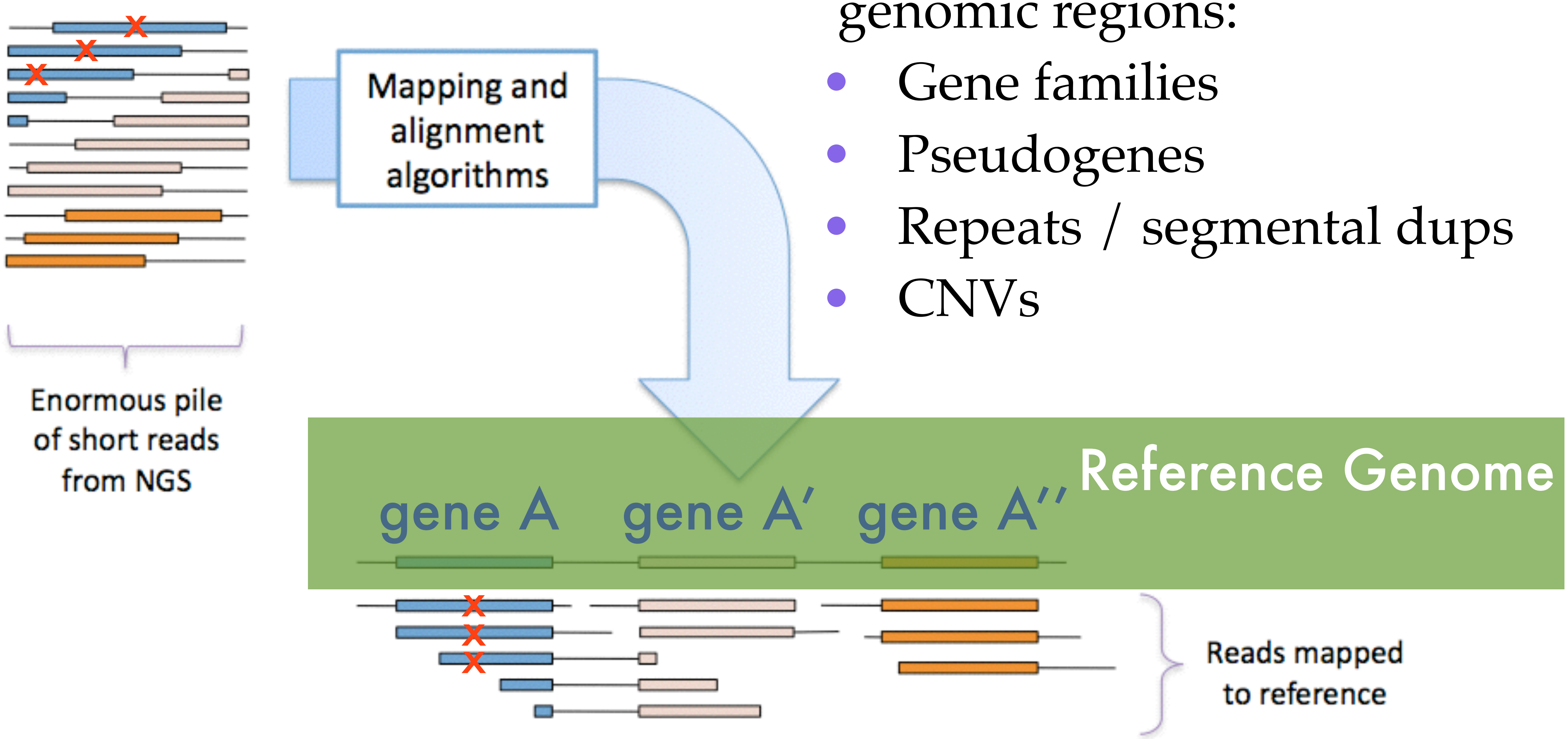
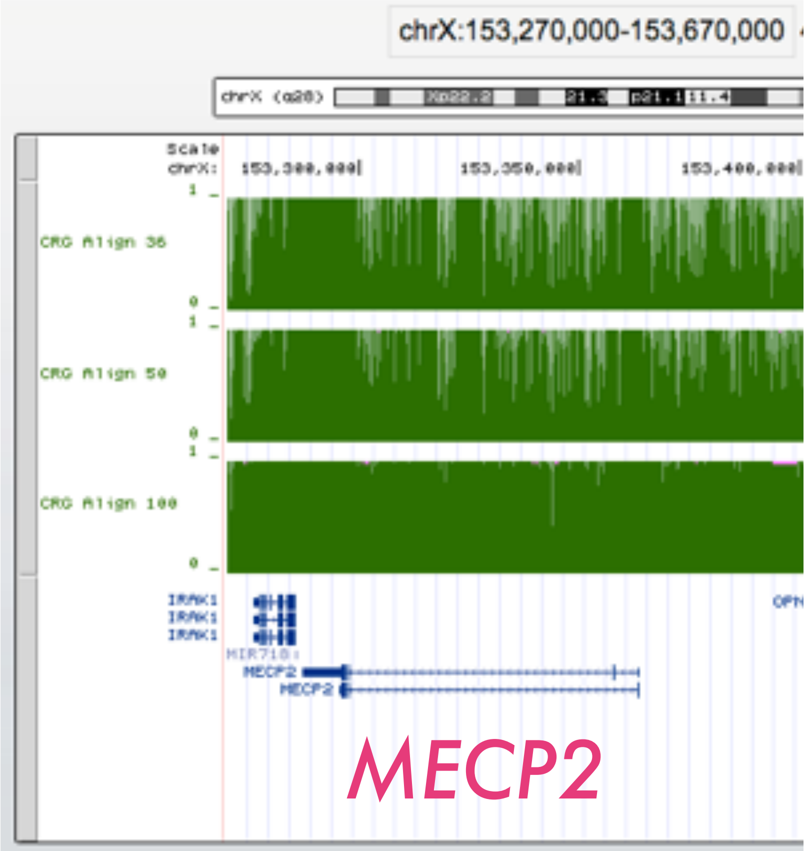

Enormous pile of short reads from NGS
Mapping and alignment algorithms
Reference Genome
gene A
gene A′
gene A′′
Reads mapped to reference

Image credit: Broad GSA Platform

Mapping confidence/mapability

- Mapping confidence is a prerequisite for good variant calls
- But mapability can vary quite a bit!



chrX:153,270,000-153,670,000

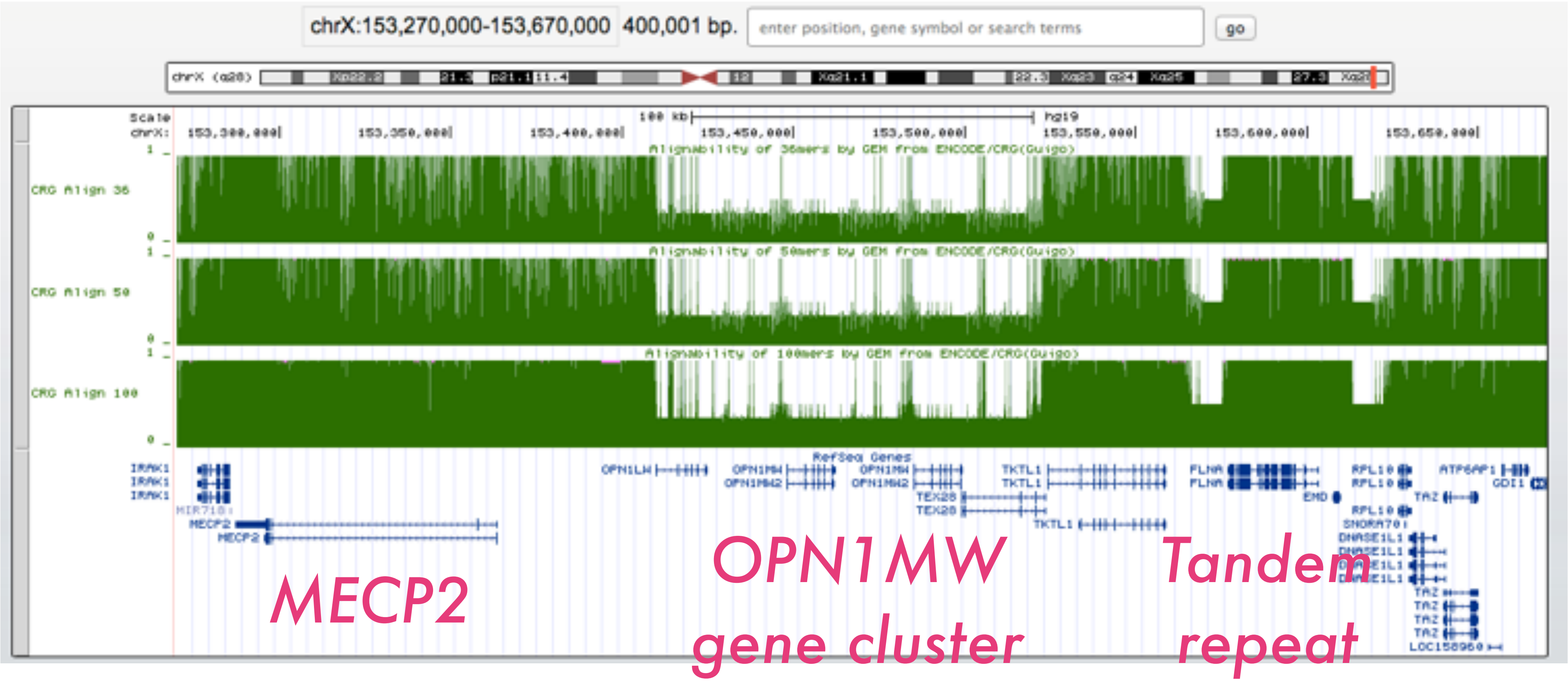
MECP2

Read length

- 36 bp
- 75 bp
- 100 bp

Mapping confidence/mapability

- Mapping confidence is a prerequisite for good variant calls
- But mapability can vary quite a bit!


chrX:153,270,000-153,670,000 400,001 bp.

MECP2

OPN1MW gene cluster

Tandem repeat

Read length

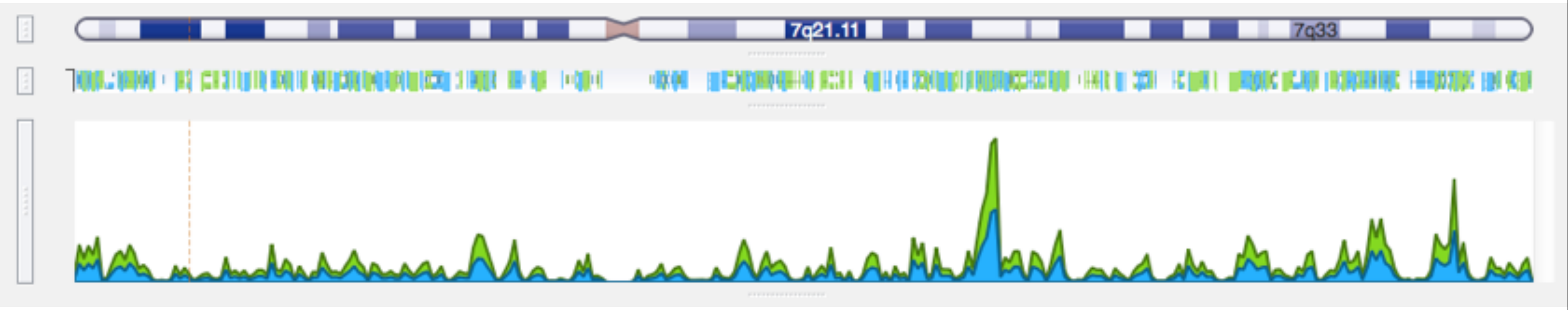
- 36 bp
- 75 bp
- 100 bp

Solutions to the mapability problem

- Longer reads
- Paired-end and mate-pair sequencing
- Better reference sequences (eg taking into account CN variable regions)

2. Sequencing coverage is biased

Read coverage on chr7 for a
a typical WES (whole exome sequencing) experiment



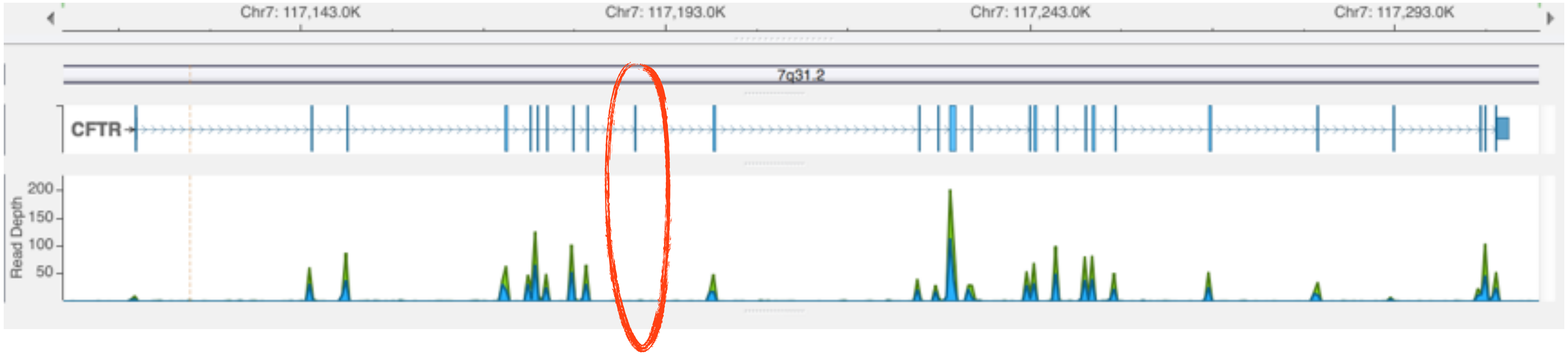
Refseq genes

Coverage

Contributing factors:

- intentional (eg, exome capture design)
- unintentional
  - PCR-related (eg, GC-rich regions)
  - mapability

…& may result in gaps in coverage (insufficient breadth)

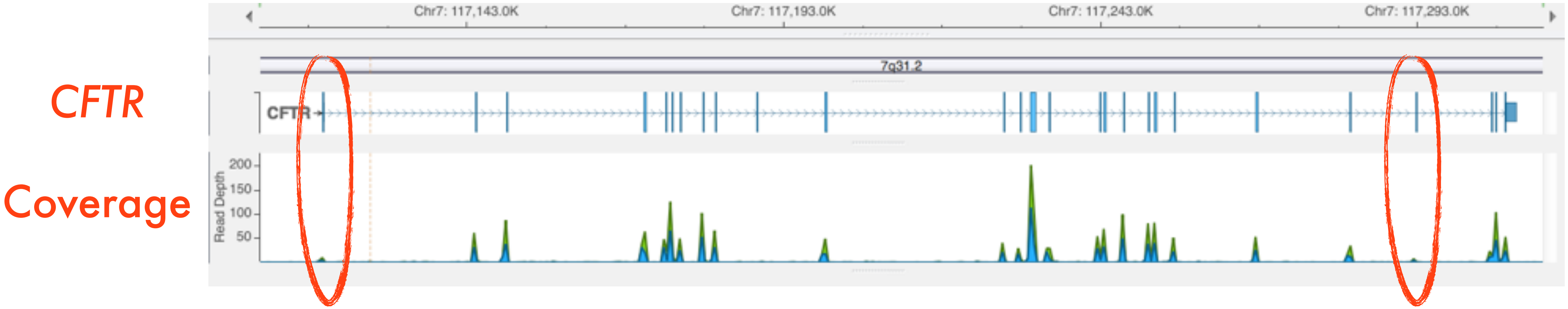


CFTR

Coverage

- **Example:** absent read coverage over *CFTR* exon 10
- **Consequence: variant dropout (false negatives)**

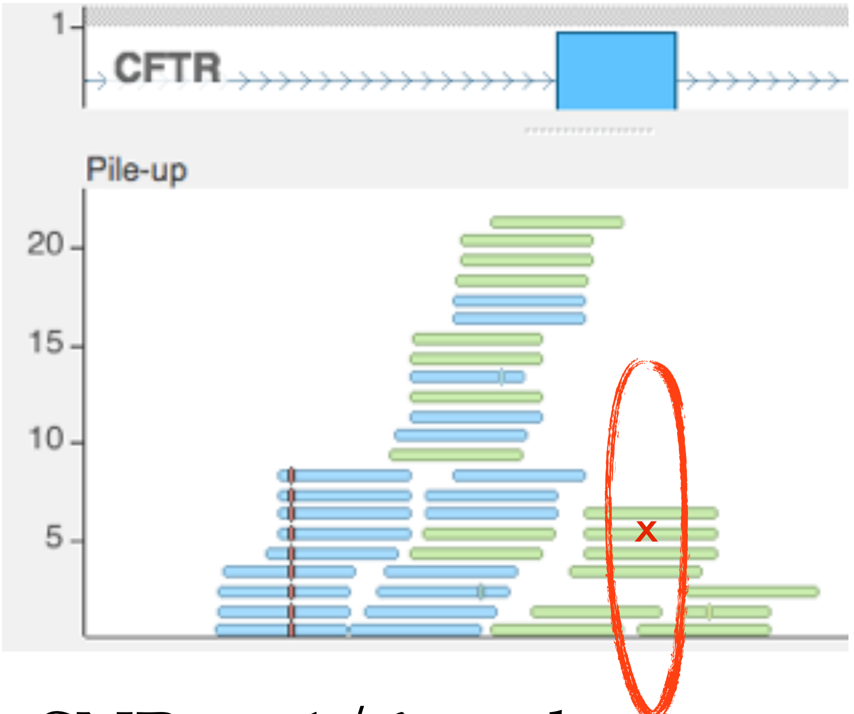
...or just inadequate coverage (insufficient depth)



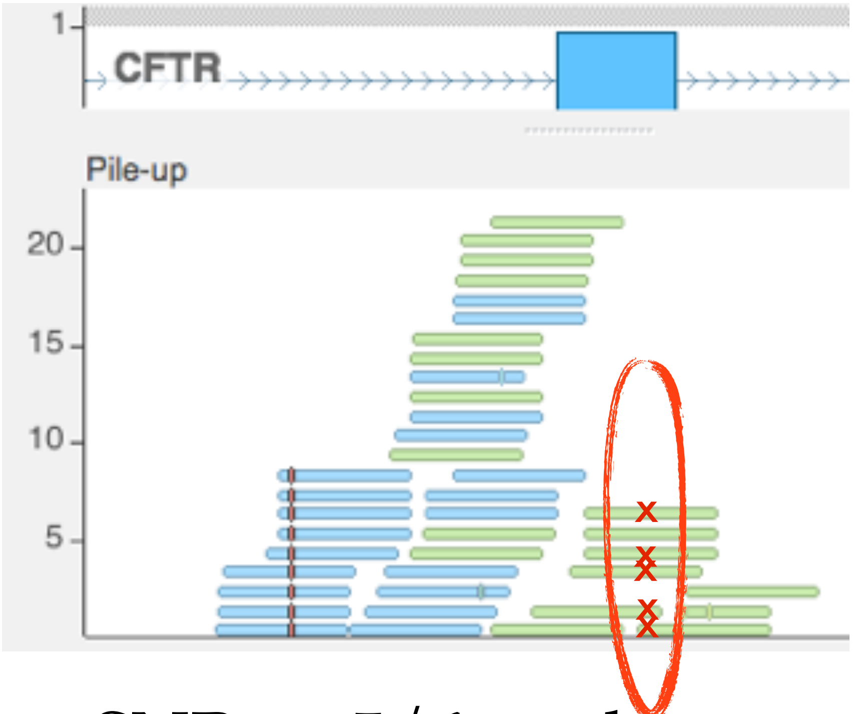
- **Example:** low read coverage over *CFTR* exons 1 & 24

...or just inadequate coverage (insufficient depth)

- **Example:** low read coverage over *CFTR* exons 1 & 24



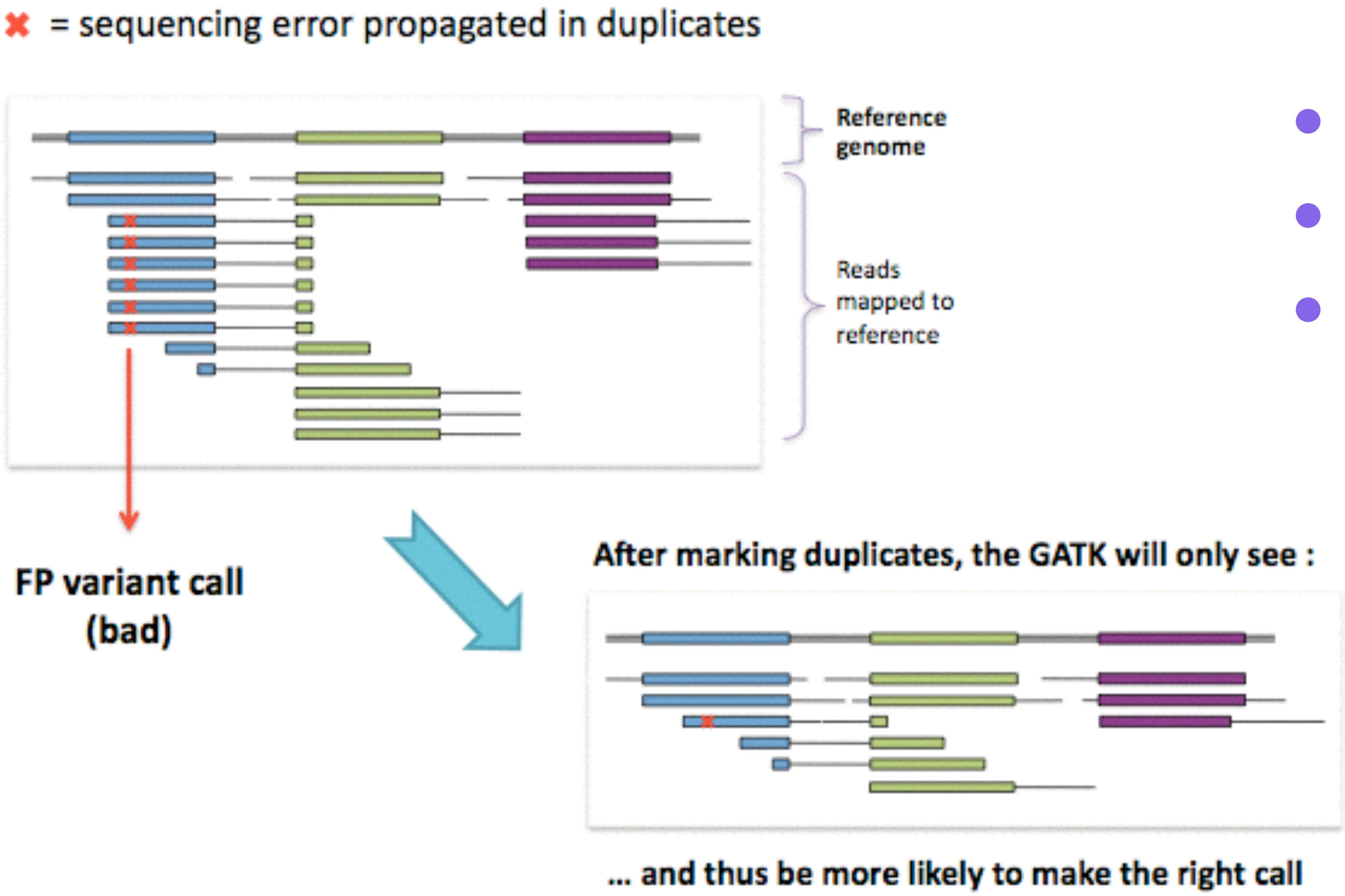
- SNP on 1/6 reads
- Is this a heterozygous variant?



- SNP on 5/6 reads
- Is this a heterozygous or homozygous variant?

- **Consequence:**
  Inaccurate genotyping in areas of low read depth

Excess coverage is sometimes a red flag


✖ = sequencing error propagated in duplicates
Reference genome
Reads mapped to reference
FP variant call (bad)
After marking duplicates, the GATK will only see :
... and thus be more likely to make the right call

- Caused by:
- PCR duplicates
- CN expansions

Image credit: Broad GSA Platform

- “Cleaning” alignments by **finding and removing PCR duplicates** evens out coverage, and reduces false positives

Depth and breadth are usually a tradeoff

- **Given fixed $$$: Depth or Breadth, choose one!**

Shallow & wide	Narrow & deep
more variants	fewer variants
less accurate genotypes	more accurate genotypes
e.g., “exomes” at 50-150X	e.g., “panels” at 500-1500X

- **Costs are gradually dropping so hopefully this tradeoff will become moot!**

Solutions to the coverage bias problem

- [Optimize mapability (longer reads, paired end sequencing, etc.)]
- Optimize library prep
    - Minimize PCR, or use PCR-free library prep methods
    - Normalize baits
- Informatically, find and remove PCR duplicates

3. Not all variant calls are created equal

- **We do quite well with SNPs (i.e., single base substitutions)**

- Calls are reliable: >99% concordance with chip-based SNP genotyping or other “truth sets”

3. Not all variant calls are created equal

- **But indels (i.e., small insertions or deletions) are significantly harder**
- It is computationally hard to map a 100bp read to the genome if you allow for gaps
- Sensitivity estimates vary hugely (50-90%), & 2-10X more false positives (compared to SNPs)

Example: calling around homopolymers

- Small insertions/deletions (especially near the ends) can trick mappers into misaligning with mismatches

10bp “T” homopolymer run

```
ref: TGACTCGTAACCAGGCTTTTTTTTTTGCGGCCGAA
```

Example: calling around homopolymers

- Small insertions/deletions (especially near the ends) can trick mappers into misaligning with mismatches

10bp “T” homopolymer run

```
ref: TGACTCGTAACCAGGCTTTTTTTTTTGCGGCCGAA
reads:   TCGTAACGAGGCTTTTTTTTTGCGGGC
                 AGGCTTTTTTTTTGCGGGCCGAA
      GACTCGTAACGAGGCTTTTTTTTTGC
               CGAGGCTTTTTTTTTGCGGGCCG
     TGACTCGTAACGAGGCTTTTTTTTTG
```

many single-bp mismatches?

Example: calling around homopolymers

- Small insertions/deletions (especially near the ends) can trick mappers into misaligning with mismatches

10bp “T” homopolymer run

```
ref: TGACTCGTAACCAGGCTTTTTTTTTTGCGGGCCGAA
reads:   TCGTAACGAGGCTTTTTTTTT^GCGGGC
                 AGGCTTTTTTTTT^GCGGGCCGAA
      GACTCGTAACGAGGCTTTTTTTTT^GC
               CGAGGCTTTTTTTTT^GCGGGCCGAA
     TGACTCGTAACGAGGCTTTTTTTTT^G
```

Local realignment reveals a hidden 1bp delT

Red flags that a variant may be suspicious

- In fact, raw indel calls are infested with false positives
- Statistics can be calculated that predict problematic variants:
  - Low read depth
  - Strand bias
  - Low mapping quality
  - Clusters of nearby variants
  - Nearby homopolymer run/other repeats

Variant Quality Scores

- “**Variant quality score**”: These statistics can be combined to derive a score that expresses the confidence in a particular call

Solutions for calling difficult variants

- Increase coverage
- **Main advice: Be aware that variant calling is imperfect**
  - SNPs pretty good
  - indels less so

- Investigational approaches:
  - Joint calling in large batches
  - Building custom references for specific difficult-to-catch variants

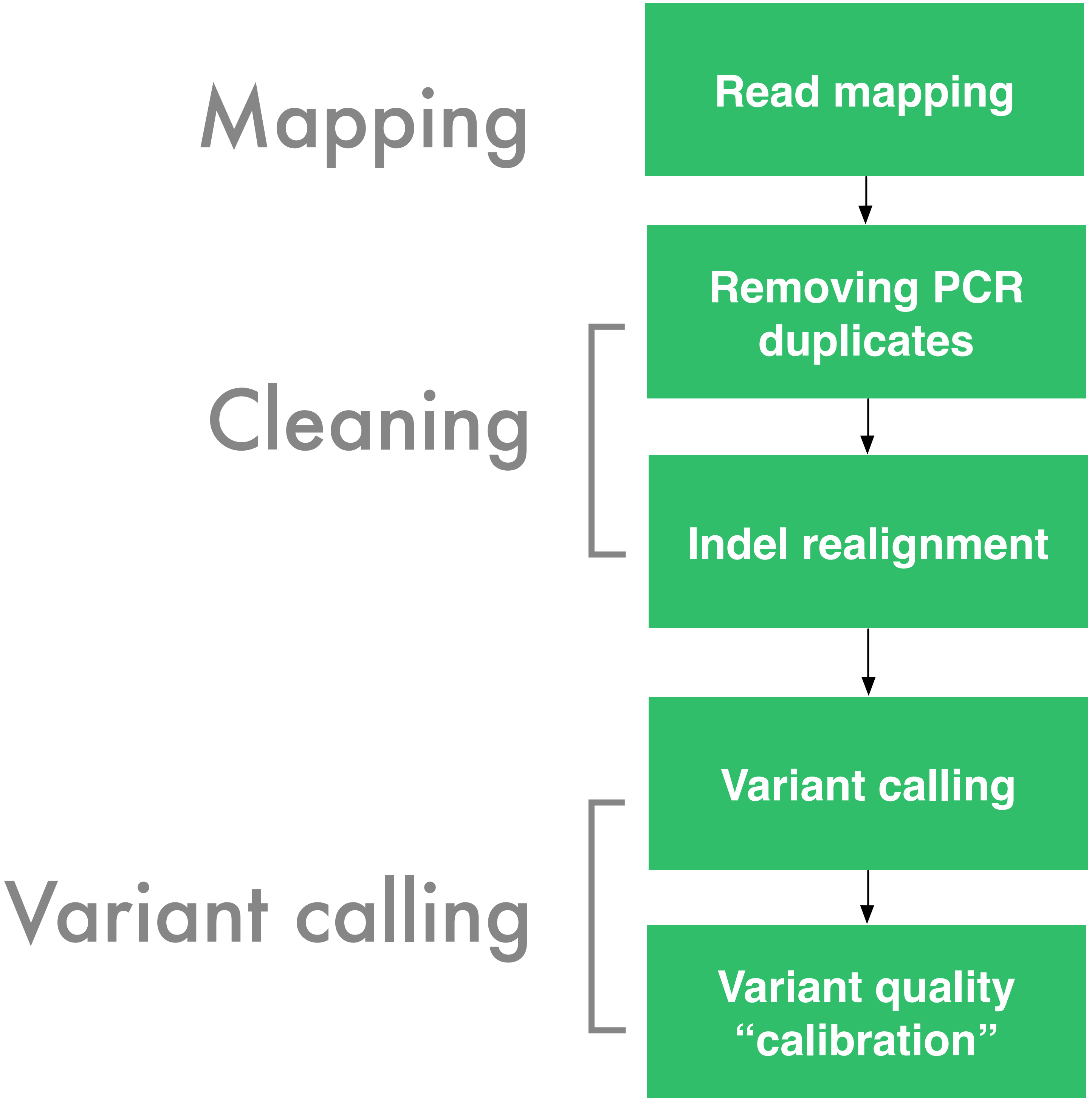
- Trust, but **verify!**

4. Beyond SNPs and small indels

- Algorithms for other variant classes are coming, but still largely investigational:
  - CNVs* and structural variants
  - Larger insertions (>20bp) or deletions (>50bp)
  - Repeat expansions/contractions
  - Transposable elements

Take home points

Take home points



Mapping

Read mapping

Cleaning

Removing PCR duplicates

Indel realignment

Variant calling

Variant calling

Variant quality “calibration”

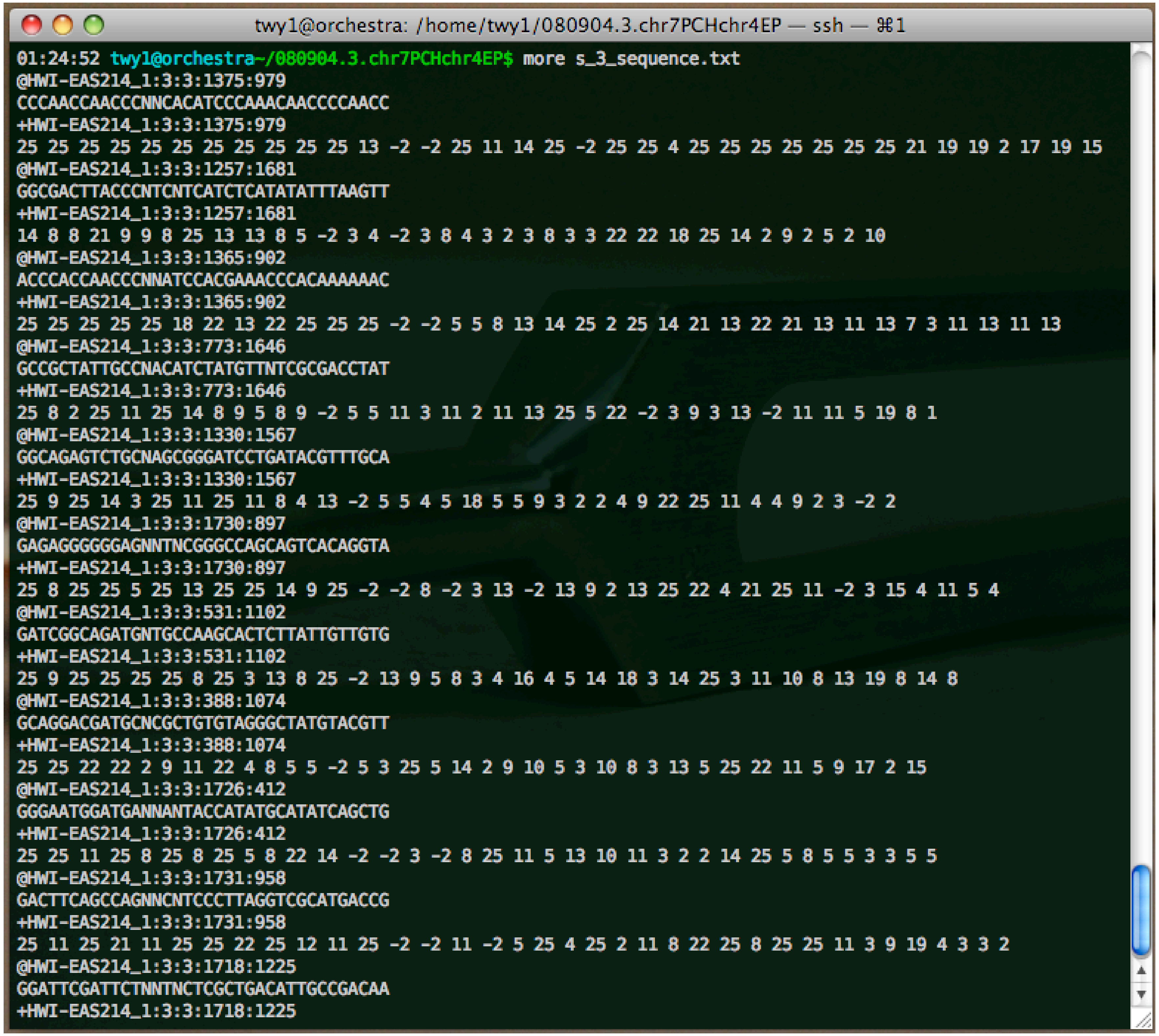
- A proper analytic pipeline mitigates many of the complications of NGS analysis

Take home points

- Ask not just about mean coverage, but coverage breadth and depth ("95% coverage at 30X")
- Ask for a list of coverage dropouts. *There is no such thing as a "whole" genome!*
- Weigh the pros and cons of maximizing breadth (exome) vs. depth (panels)
- SNPs are generally high quality, but it is still important to weigh variant quality and other red flags. Especially for indels, trust but verify
- Recognize that CNV, SV, larger indels, repeat expansion/contractions, and mobile elements are out of the scope of most clinical NGS pipelines

Questions

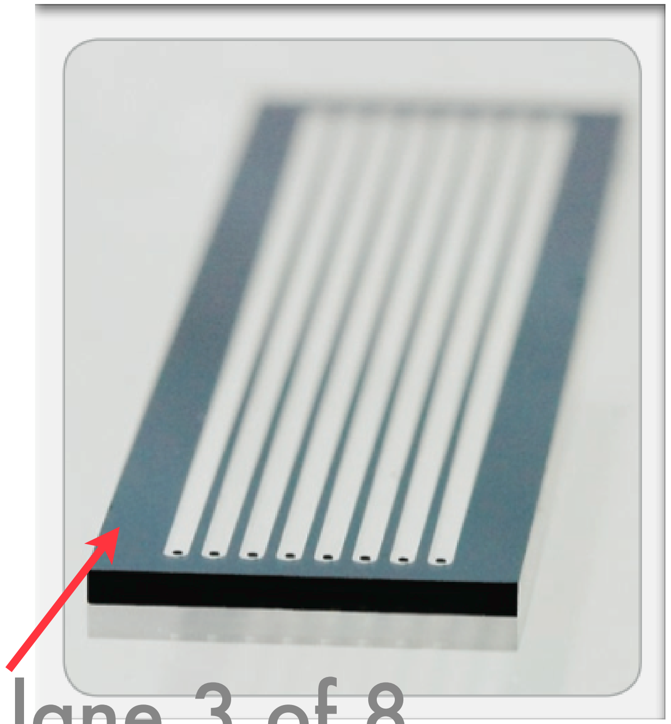
Sequence reads



twy1@orchestra: /home/twy1/080904.3.chr7PCHchr4EP — ssh — ⌘1

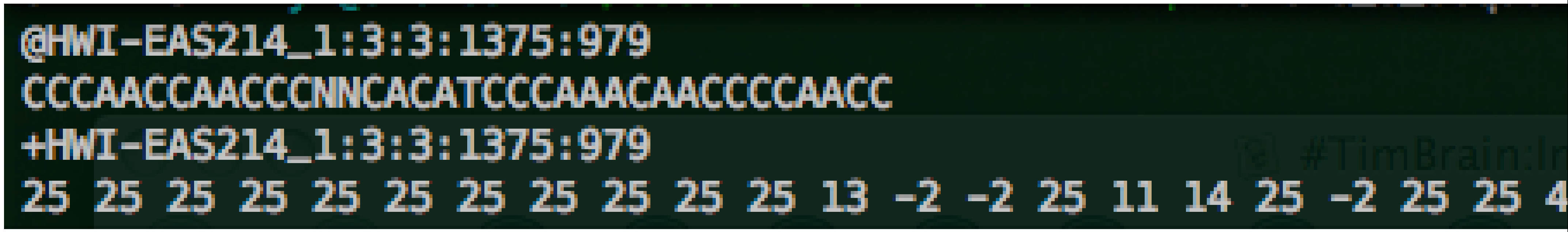
```
01:24:52 twy1@orchestra~/080904.3.chr7PCHchr4EP$ more s_3_sequence.txt
@HWI-EAS214_1:3:3:1375:979
CCCAACCAACCCNNCACATCCCAAACAACCCCAACC
+HWI-EAS214_1:3:3:1375:979
25 25 25 25 25 25 25 25 25 25 25 13 -2 -2 25 11 14 25 -2 25 25 4 25 25 25 25 25 25 25 21 19 19 2 17 19 15
@HWI-EAS214_1:3:3:1257:1681
GGCGACTTACCCNTCNTCATCTCATATATTTAAGTT
+HWI-EAS214_1:3:3:1257:1681
14 8 8 21 9 9 8 25 13 13 8 5 -2 3 4 -2 3 8 4 3 2 3 8 3 3 22 22 18 25 14 2 9 2 5 2 10
@HWI-EAS214_1:3:3:1365:902
ACCCACCAACCCNNATCCACGAAACCCACAAAAAAC
+HWI-EAS214_1:3:3:1365:902
25 25 25 25 25 18 22 13 22 25 25 25 -2 -2 5 5 8 13 14 25 2 25 14 21 13 22 21 13 11 13 7 3 11 13 11 13
@HWI-EAS214_1:3:3:773:1646
GCCGCTATTGCCNACATCTATGTTNTCGCGACCTAT
+HWI-EAS214_1:3:3:773:1646
25 8 2 25 11 25 14 8 9 5 8 9 -2 5 5 11 3 11 2 11 13 25 5 22 -2 3 9 3 13 -2 11 11 5 19 8 1
@HWI-EAS214_1:3:3:1330:1567
GGCAGAGTCTGCNAGCGGGATCCTGATACGTTTGCA
+HWI-EAS214_1:3:3:1330:1567
25 9 25 14 3 25 11 25 11 8 4 13 -2 5 5 4 5 18 5 5 9 3 2 2 4 9 22 25 11 4 4 9 2 3 -2 2
@HWI-EAS214_1:3:3:1730:897
GAGAGGGGGGAGNNTNCGGGCCAGCAGTCACAGGTA
+HWI-EAS214_1:3:3:1730:897
25 8 25 25 5 25 13 25 25 14 9 25 -2 -2 8 -2 3 13 -2 13 9 2 13 25 22 4 21 25 11 -2 3 15 4 11 5 4
@HWI-EAS214_1:3:3:531:1102
GATCGGCAGATGNTGCCAAGCACTCTTATTGTTGTG
+HWI-EAS214_1:3:3:531:1102
25 9 25 25 25 25 8 25 3 13 8 25 -2 13 9 5 8 3 4 16 4 5 14 18 3 14 25 3 11 10 8 13 19 8 14 8
@HWI-EAS214_1:3:3:388:1074
GCAGGACGATGCNCGCTGTGTAGGGCTATGTACGTT
+HWI-EAS214_1:3:3:388:1074
25 25 22 22 2 9 11 22 4 8 5 5 -2 5 3 25 5 14 2 9 10 5 3 10 8 3 13 5 25 22 11 5 9 17 2 15
@HWI-EAS214_1:3:3:1726:412
GGGAATGGATGANNANTACCATATGCATATCAGCTG
+HWI-EAS214_1:3:3:1726:412
25 25 11 25 8 25 8 25 5 8 22 14 -2 -2 3 -2 8 25 11 5 13 10 11 3 2 2 14 25 5 8 5 5 3 3 5 5
@HWI-EAS214_1:3:3:1731:958
GACTTCAGCCAGNNCNTCCCTTAGGTCGCATGACCG
+HWI-EAS214_1:3:3:1731:958
25 11 25 21 11 25 25 22 25 12 11 25 -2 -2 11 -2 5 25 4 25 2 11 8 22 25 8 25 25 11 3 9 19 4 3 3 2
@HWI-EAS214_1:3:3:1718:1225
GGATTCGATTCTNNTNCTCGCTGACATTGCCGACAA
+HWI-EAS214_1:3:3:1718:1225
```

Anatomy of a read



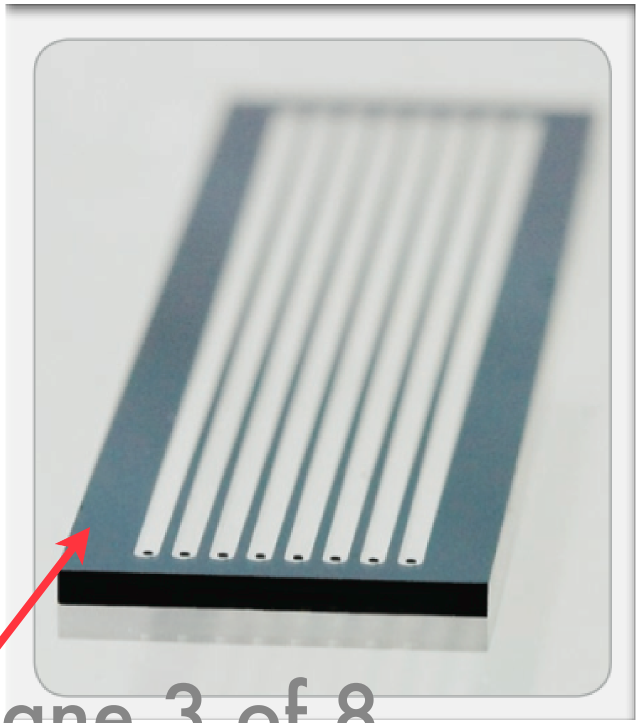
lane 3 of 8

identifier



```
@HWI-EAS214_1:3:3:1375:979
CCCAACCAACCCNNCACATCCCAAACAACCCCAACC
+HWI-EAS214_1:3:3:1375:979
25 25 25 25 25 25 25 25 25 25 25 13 -2 -2 25 11 14 25 -2 25 25 4
```

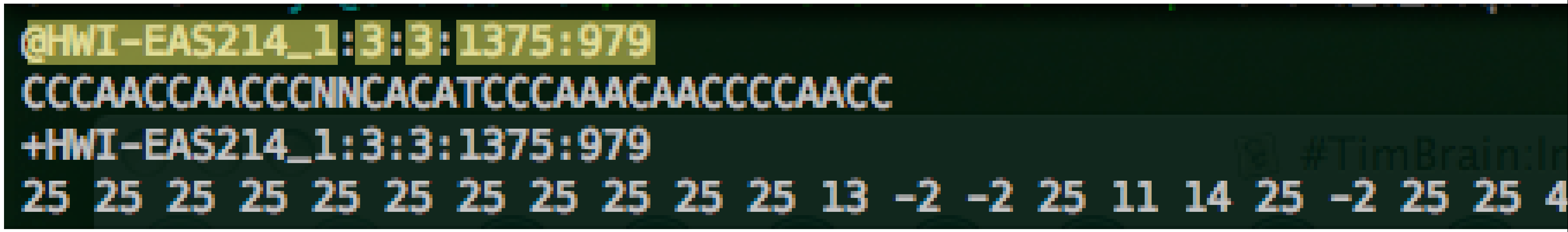
Anatomy of a read



lane 3 of 8

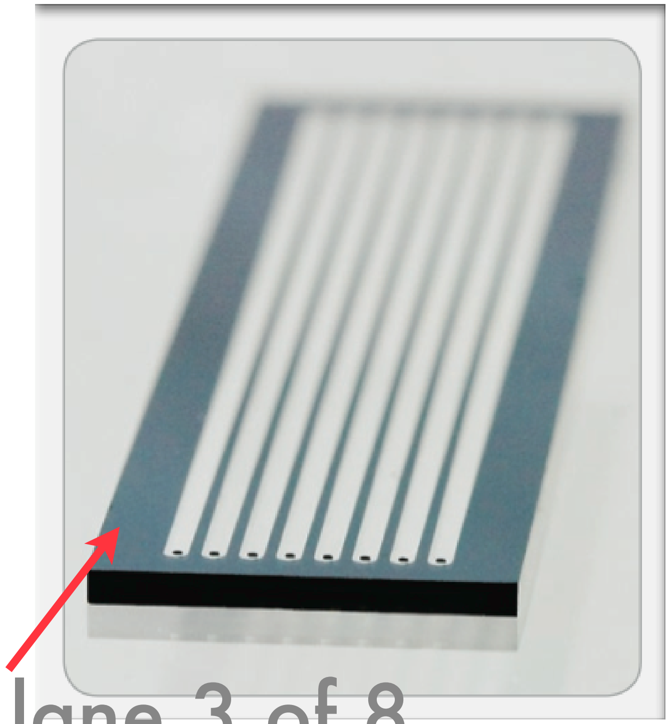
machine | lane | tile | X:Y

identifier



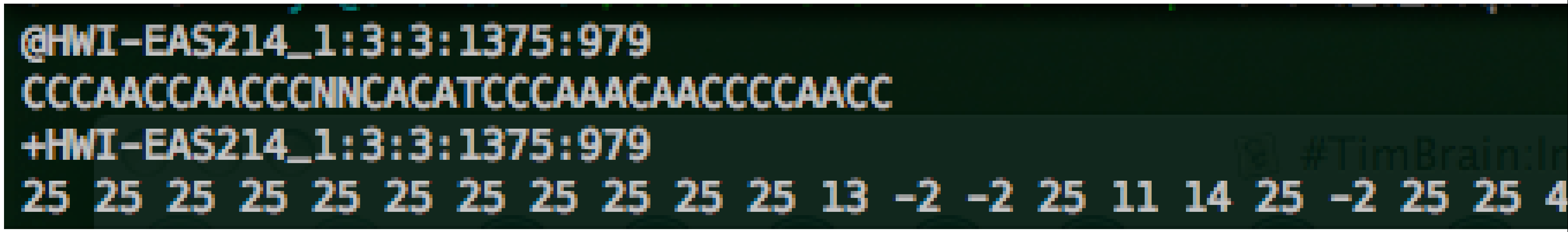
```
@HWI-EAS214_1:3:3:1375:979
CCCAACCAACCCNNCACATCCCAAACAACCCCAACC
+HWI-EAS214_1:3:3:1375:979
25 25 25 25 25 25 25 25 25 25 25 13 -2 -2 25 11 14 25 -2 25 25 4
```

Anatomy of a read



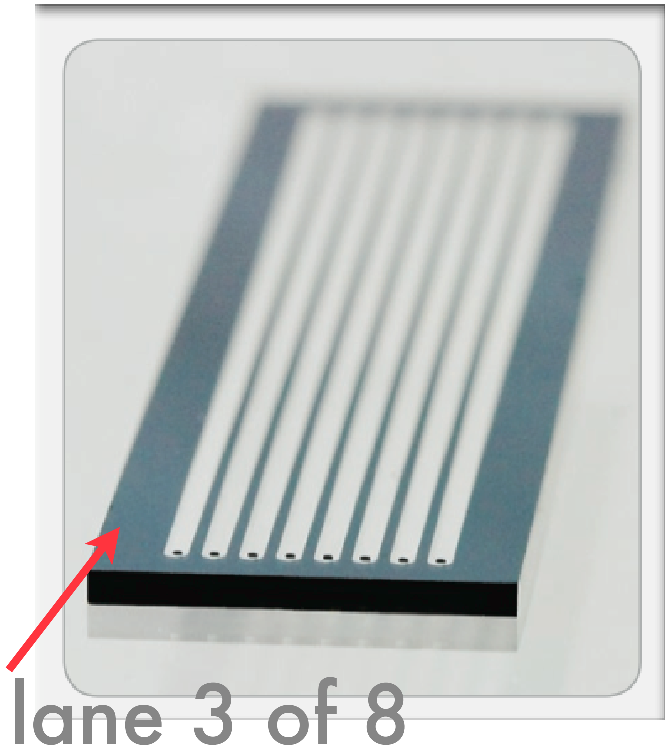
lane 3 of 8

sequence

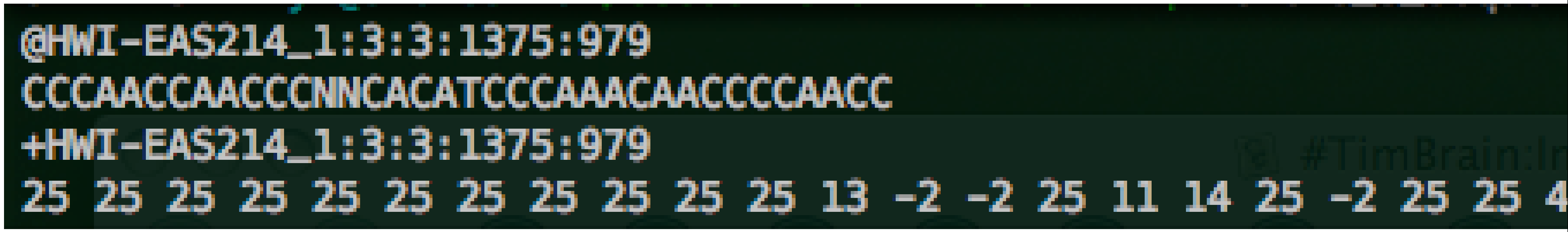


```
@HWI-EAS214_1:3:3:1375:979
CCCAACCAACCCNNCACATCCCAAACAACCCCAACC
+HWI-EAS214_1:3:3:1375:979
25 25 25 25 25 25 25 25 25 25 25 13 -2 -2 25 11 14 25 -2 25 25 4
```

Anatomy of a read

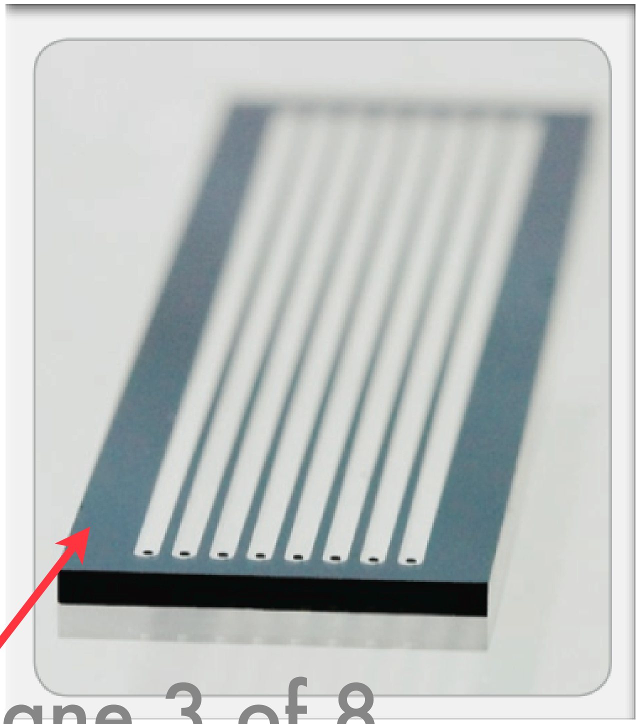


identifier
(again!)



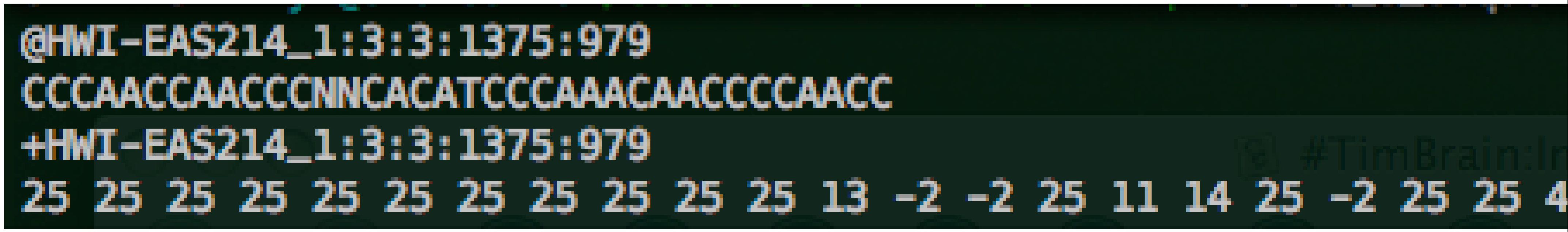
```
@HWI-EAS214_1:3:3:1375:979
CCCAACCAACCCNNCACATCCCAAACAACCCCAACC
+HWI-EAS214_1:3:3:1375:979
25 25 25 25 25 25 25 25 25 25 25 13 -2 -2 25 11 14 25 -2 25 25 4
```

Anatomy of a read



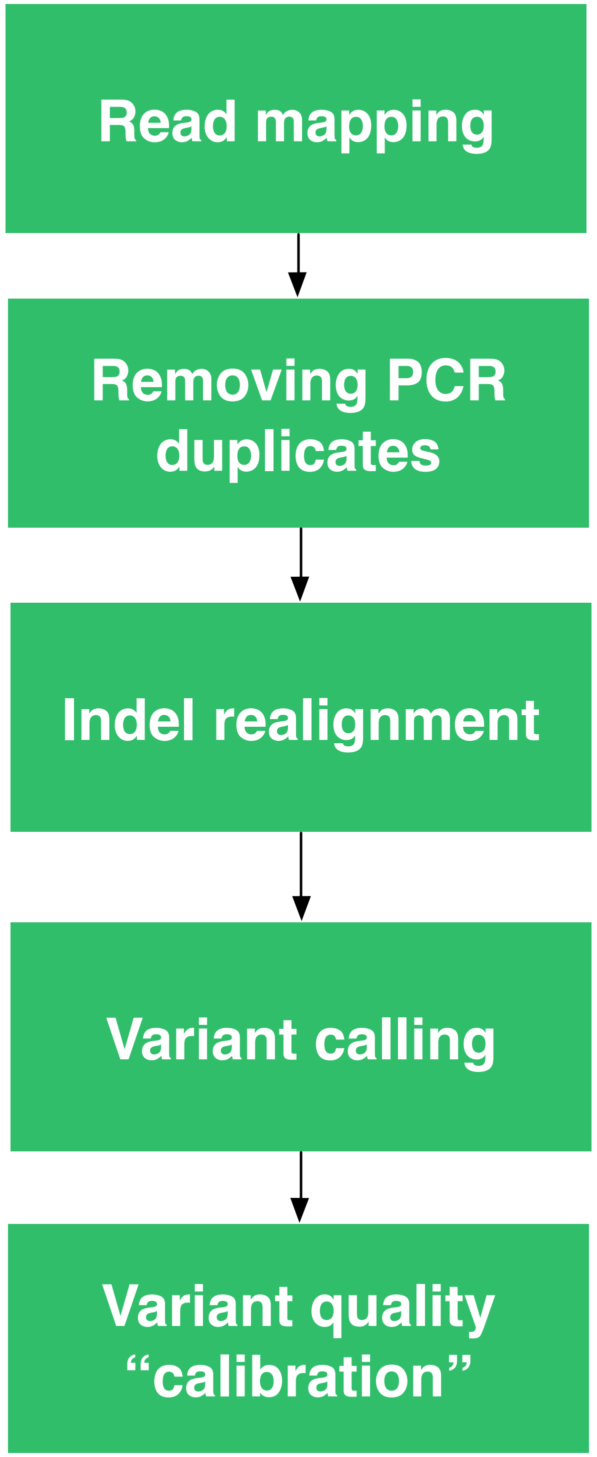
lane 3 of 8

base qualities (higher=better)



```
@HWI-EAS214_1:3:3:1375:979
CCCAACCAACCCNNCACATCCCAAACAACCCCAACC
+HWI-EAS214_1:3:3:1375:979
25 25 25 25 25 25 25 25 25 25 25 13 -2 -2 25 11 14 25 -2 25 25 4
```

Questions to ask



Read mapping

Removing PCR duplicates

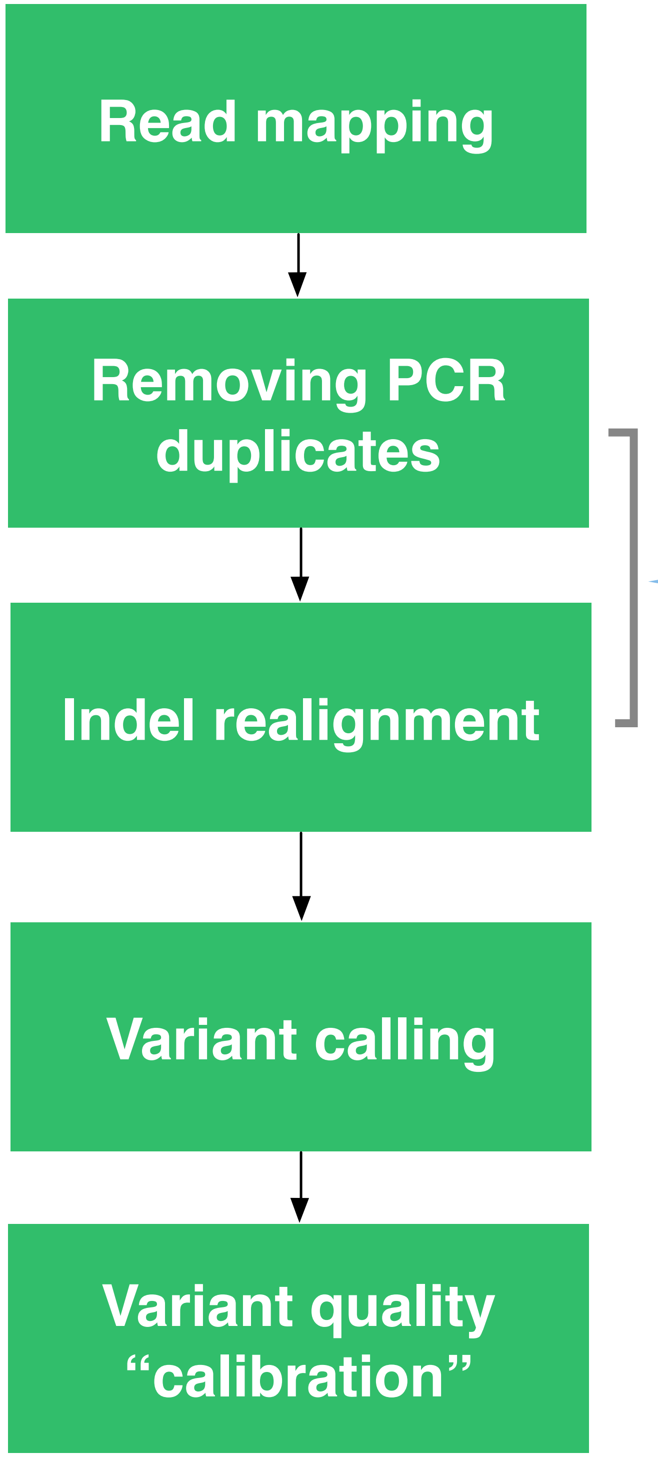
Indel realignment

Variant calling

Variant quality "calibration"

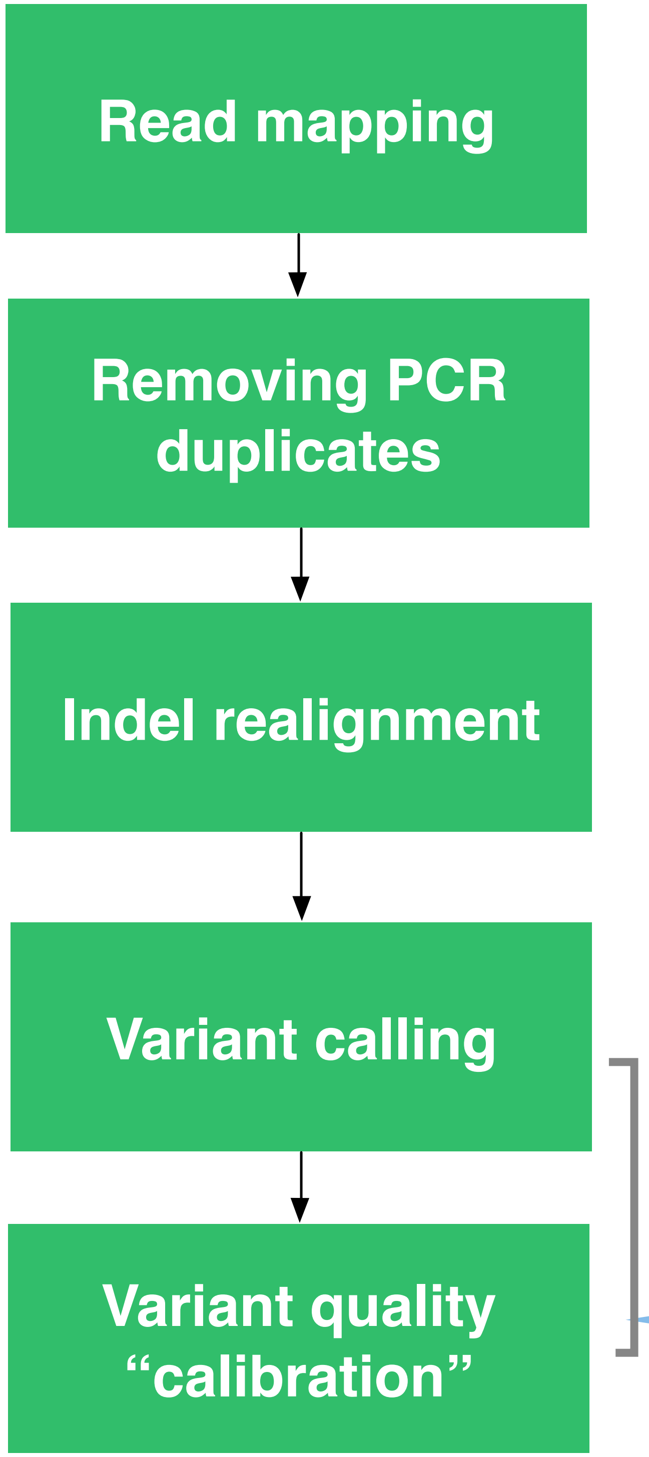
- Was sufficient breadth & depth of coverage achieved?
  - "85-95% coverage at >30X"
- What regions were missed?

Questions to ask



- Was appropriate cleaning performed?

Questions to ask



Read mapping

Removing PCR duplicates

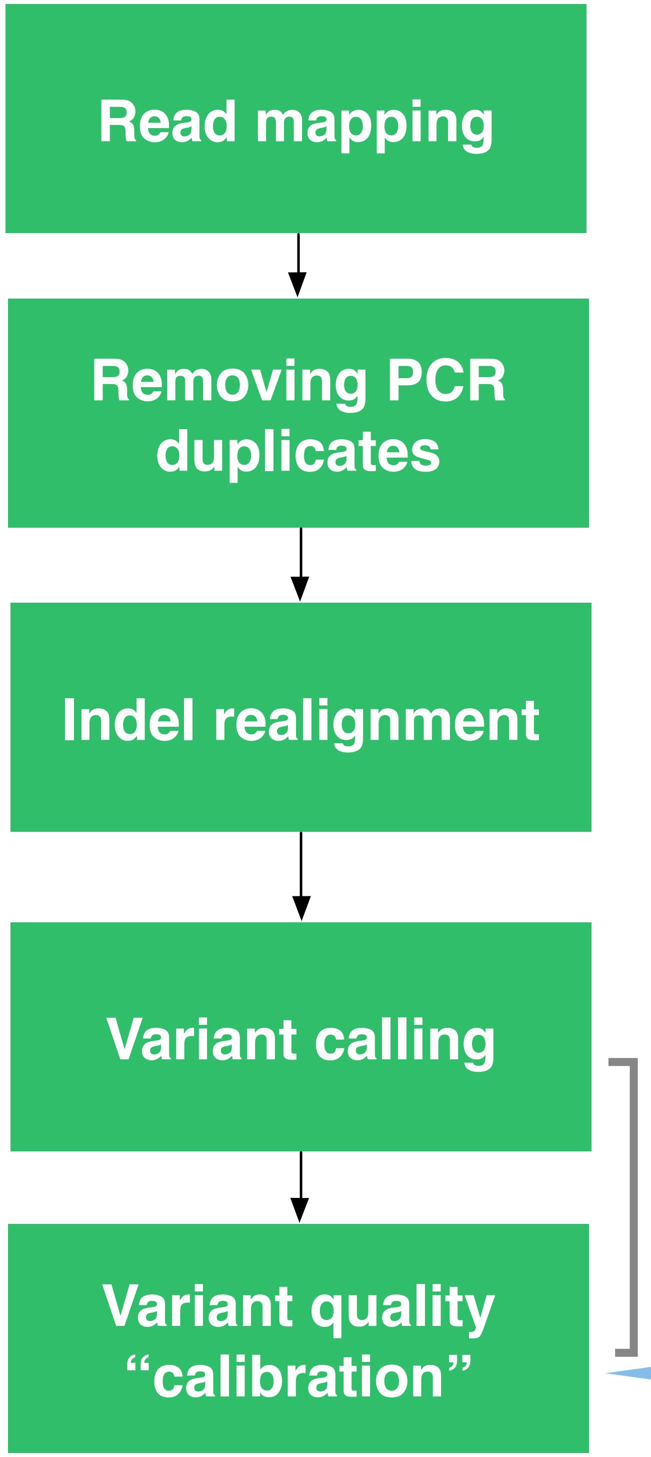
Indel realignment

Variant calling

Variant quality "calibration"

- Are the #s of variants called reasonable?
- Especially indels
- Is the percentage of "known SNPs" reasonable (98% in dbSNP)?

Questions to ask



- Is the variant quality score sufficiently high to be believed?
- What was used to confirm these variants?