PGB2022

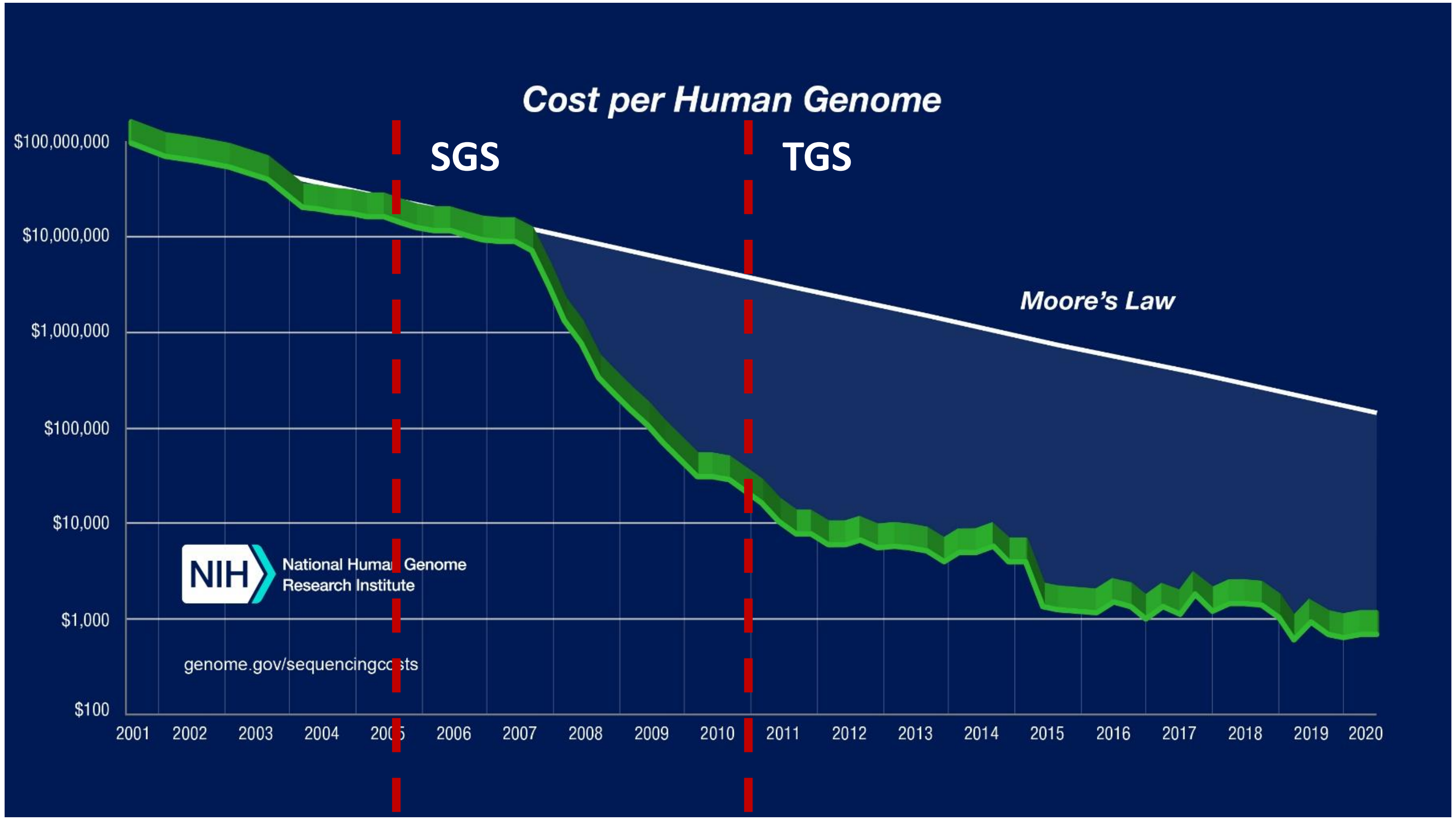
Dr. Agnieszka A. Golicz

agnieszka.golicz@agrar.uni-giessen.de

Genome sequencing technology

Sanger sequencing

- Older technology
- Reads ~1000 bp
- Very accurate



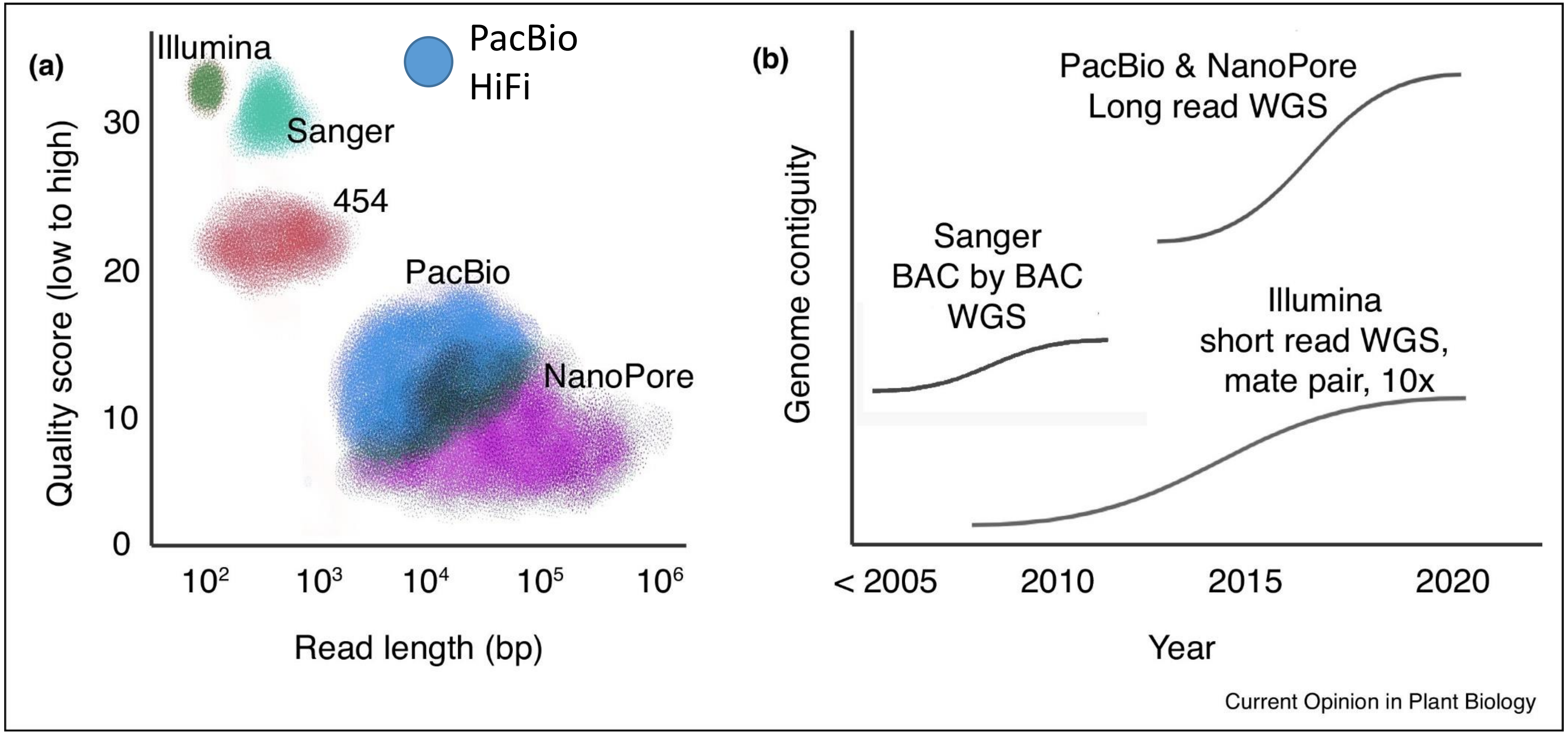
Second generation sequencing (Illumina)

- Reads ~200 bp
- More error prone
- But cheaper!

Technological limitation – we can not sequence (read) the whole chromosome at once

Sequencing read – a short sequence representing a fragment of the DNA

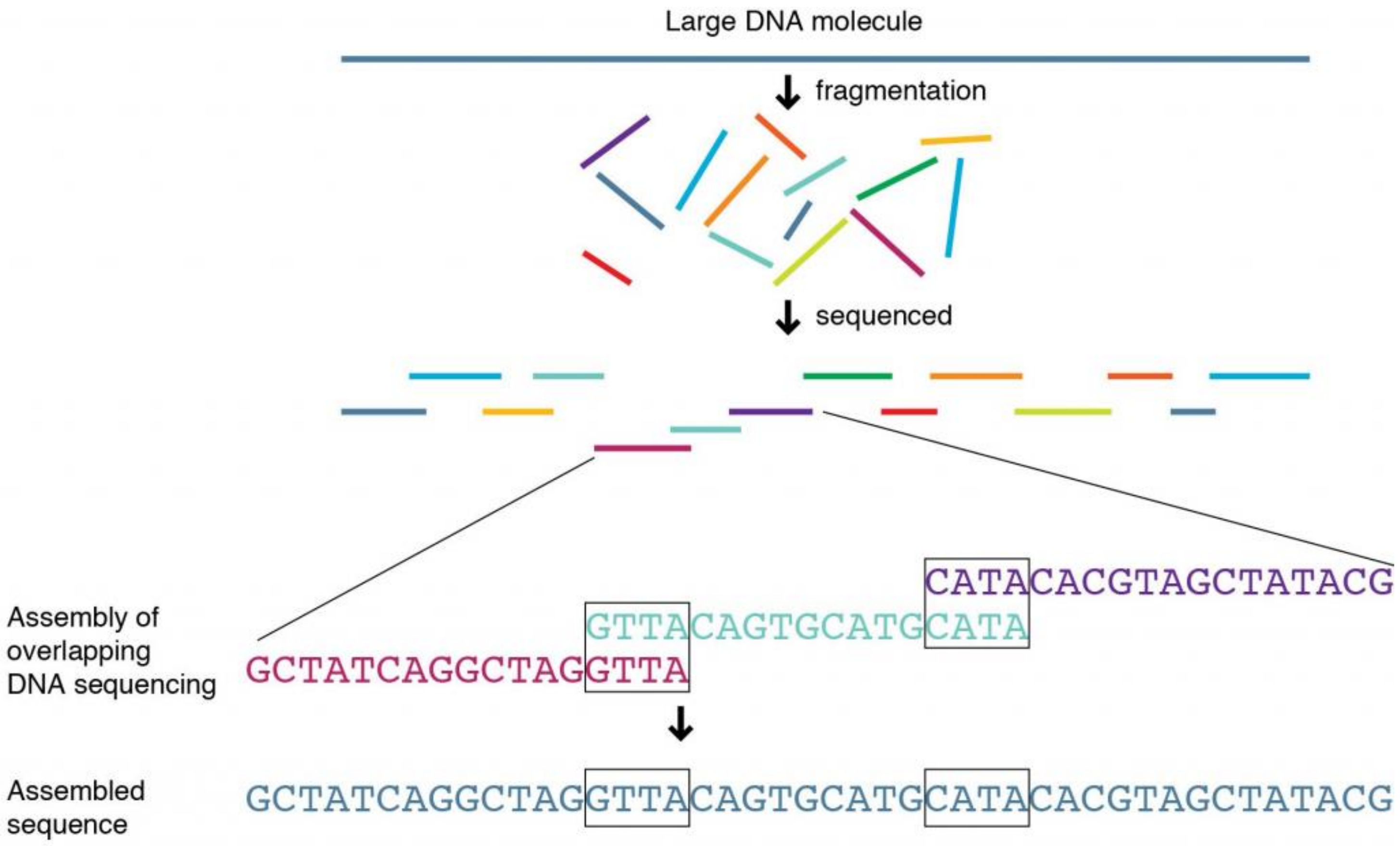
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data



Current Opinion in Plant Biology

https://www.sciencedirect.com/science/article/pii/S1369526619301244

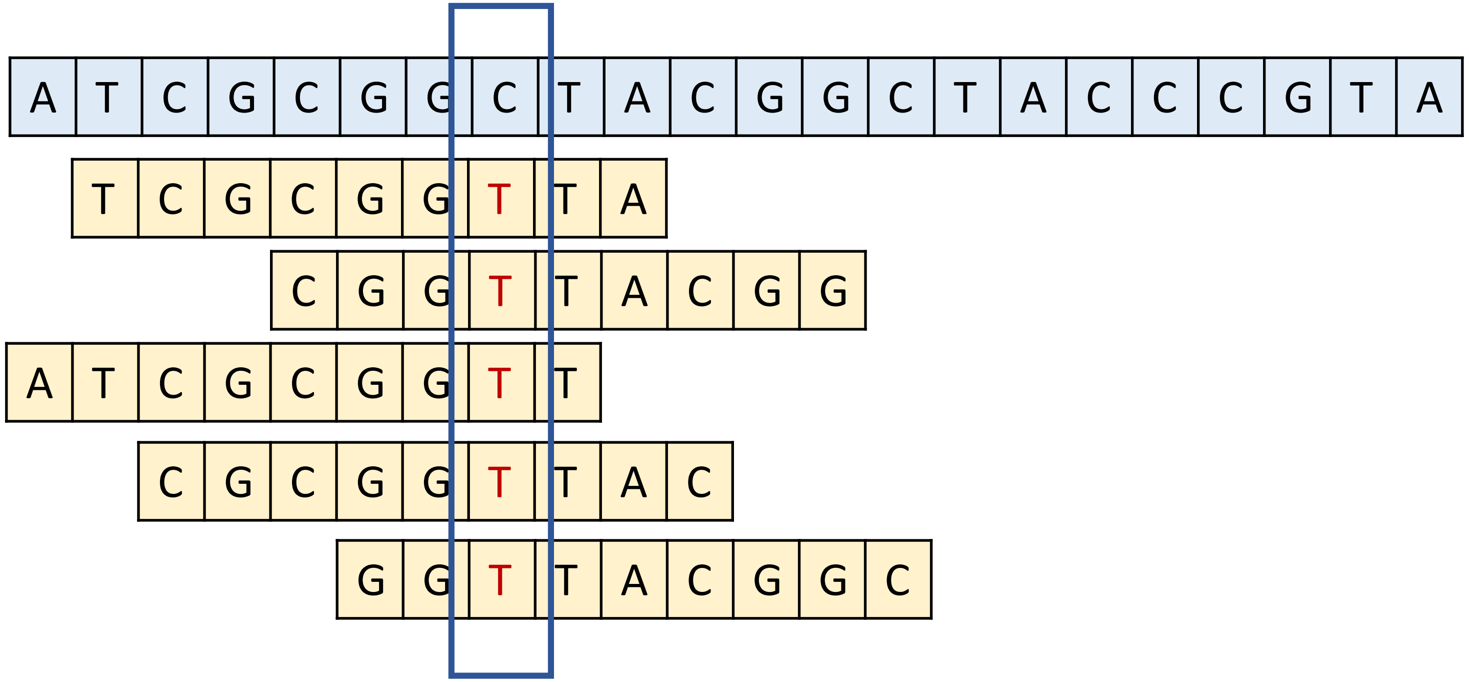
Genome sequencing



https://knowgenetics.org/whole-genome-sequencing/

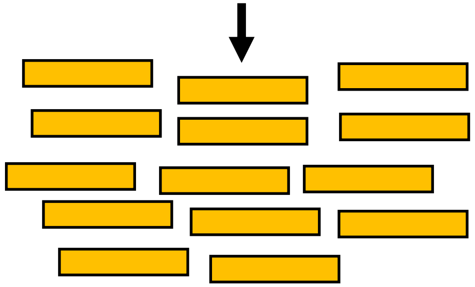
Read mapping

Reference sequence

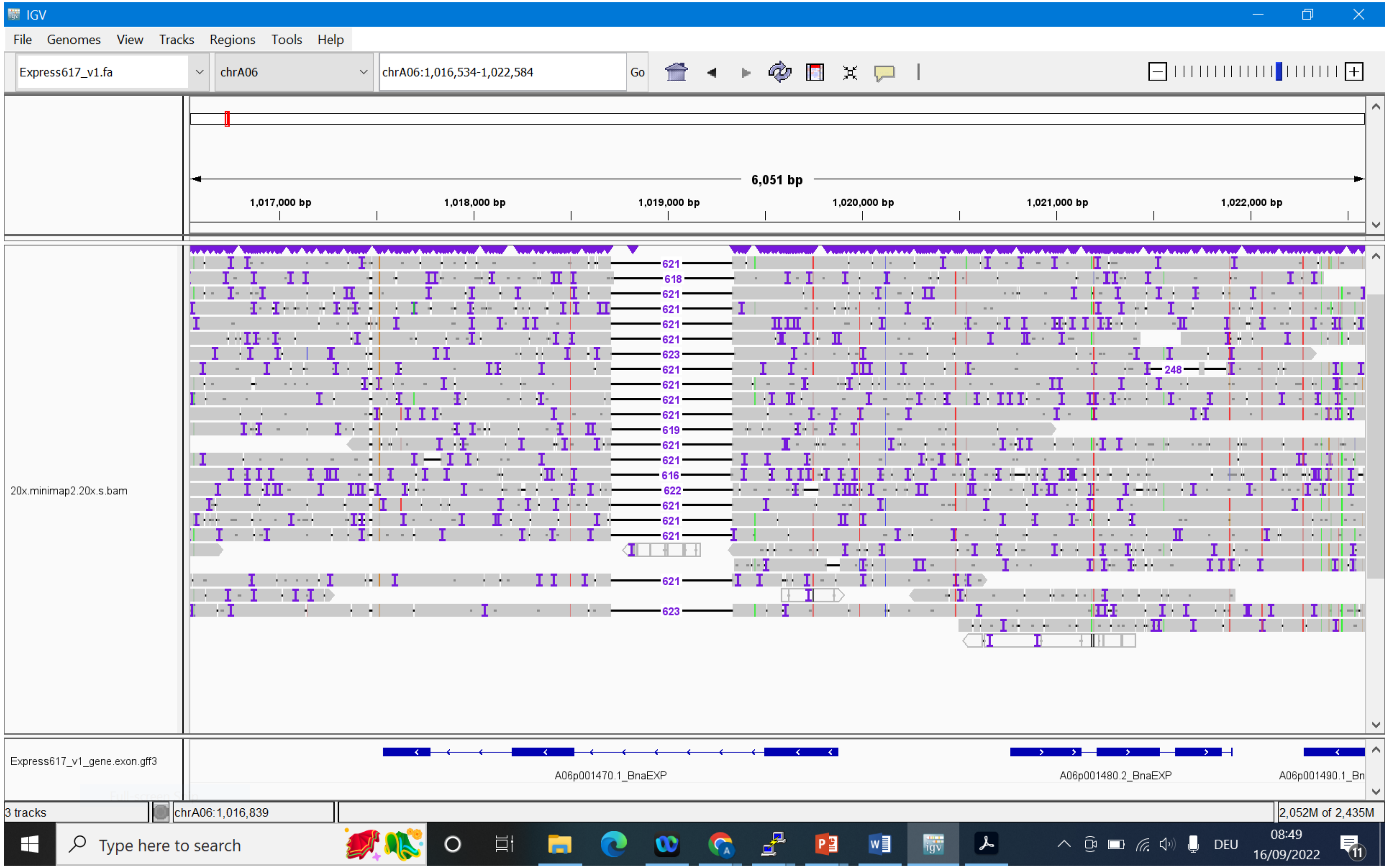


```
ATCGCGGCTACGGCTACCCGTA
 TCGCGGTTA
    CGGTTACGG
ATCGCGGTT
  CGCGGTTAC
     GGTTACGGC
```

Resequencing individuals



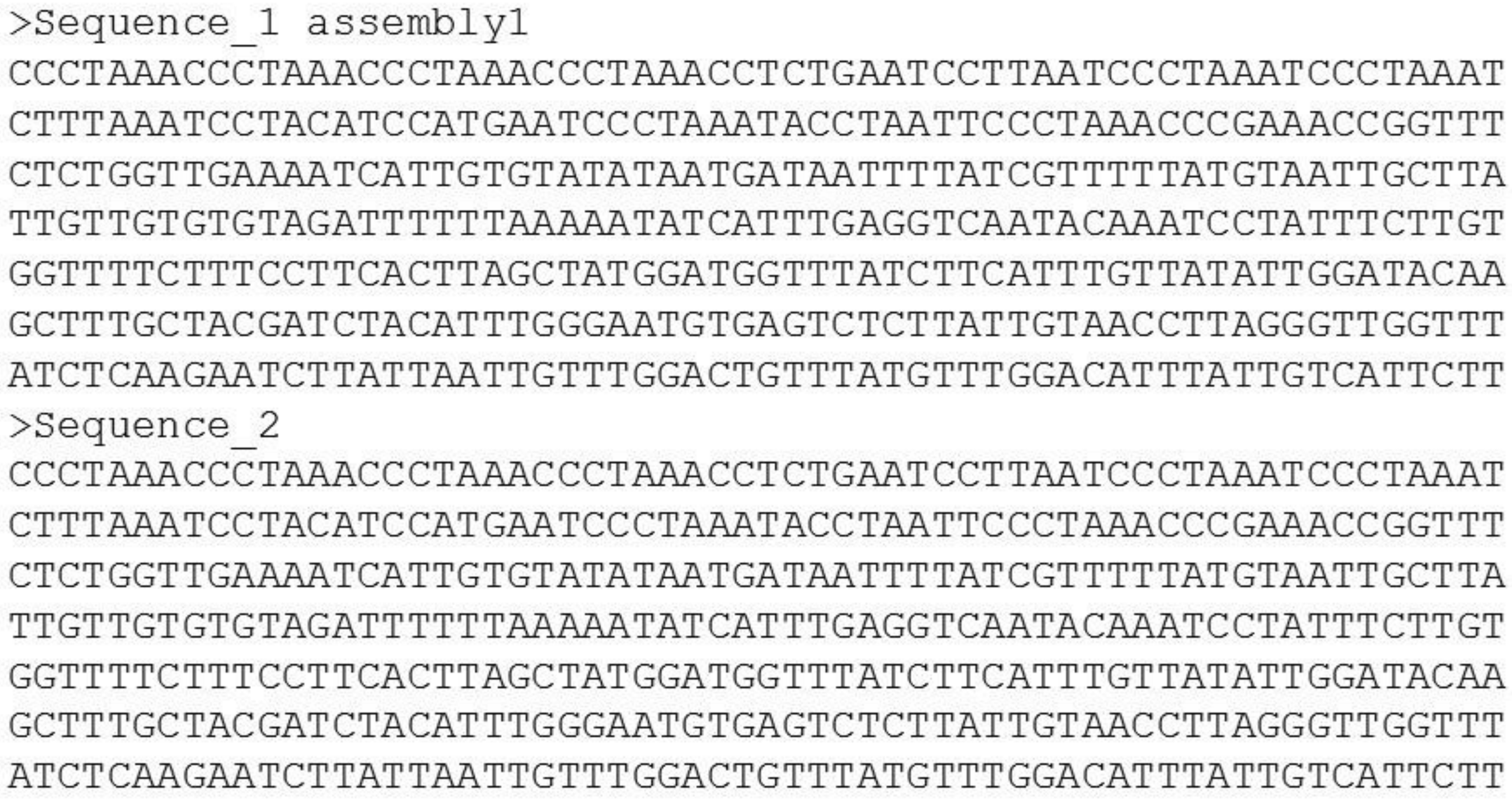
Read mapping



Bioinformatics file formats

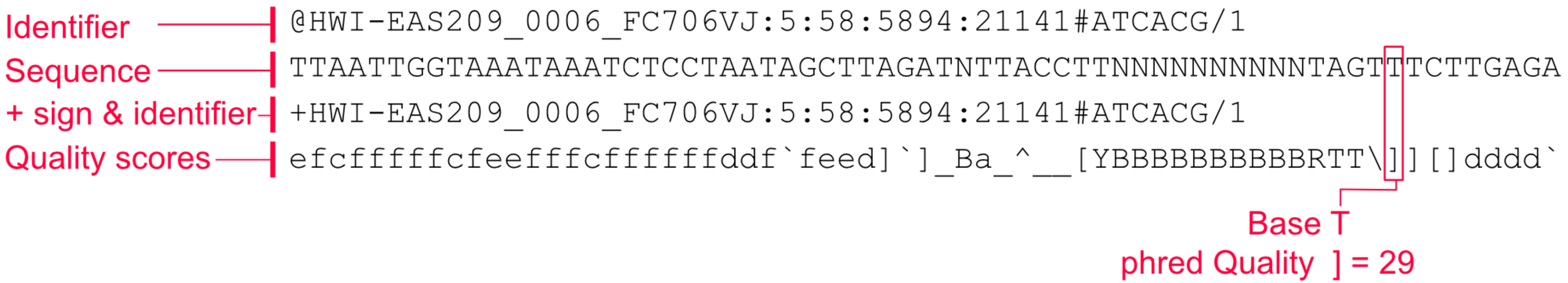
- Dealing with large quantities of data
- Automated processes
- Require standardized file formats
    - FASTA
    - FASTQ
    - SAM/BAM
    - VCF
    - GFF

FASTA file format



```
>Sequence_1 assembly1
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAATCATTGTGTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTA
TTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGT
GGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTT
ATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
>Sequence_2
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAATCATTGTGTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTA
TTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGT
GGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTT
ATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
```

FASTQ file format



Identifier

Sequence

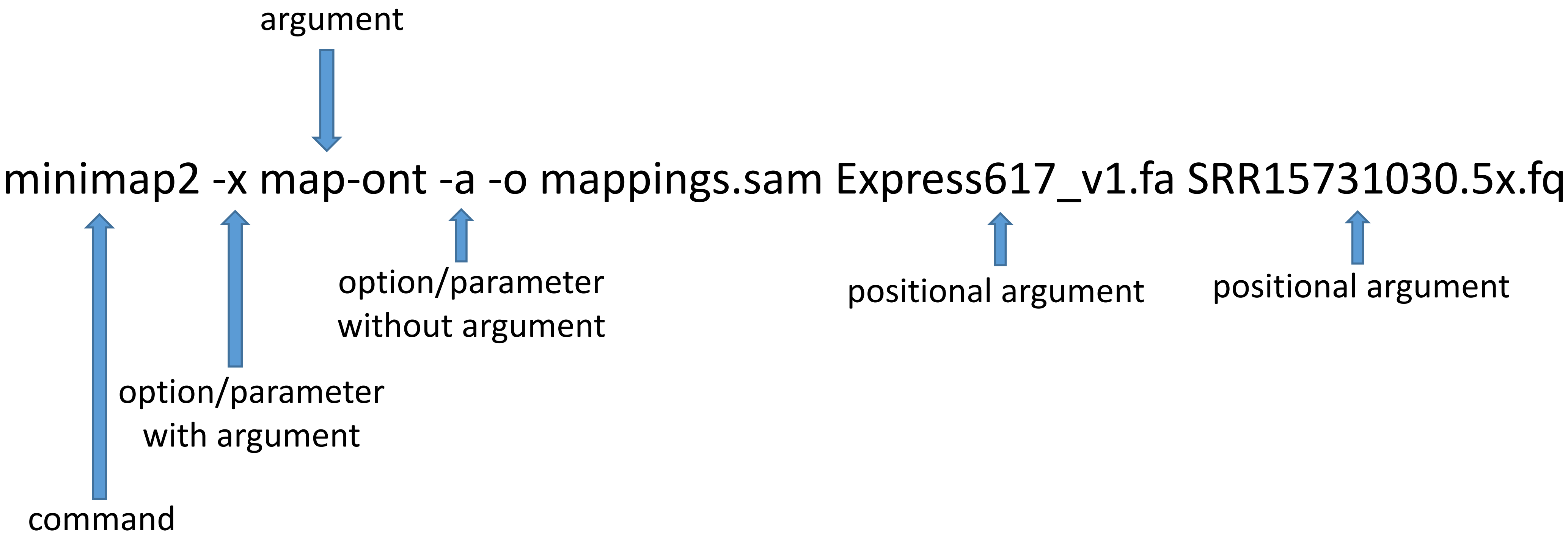
+ sign & identifier

Quality scores

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`
```

Base T
phred Quality] = 29

Using programs with command line options and arguments



argument

```
minimap2 -x map-ont -a -o mappings.sam Express617_v1.fa SRR15731030.5x.fq
```

command

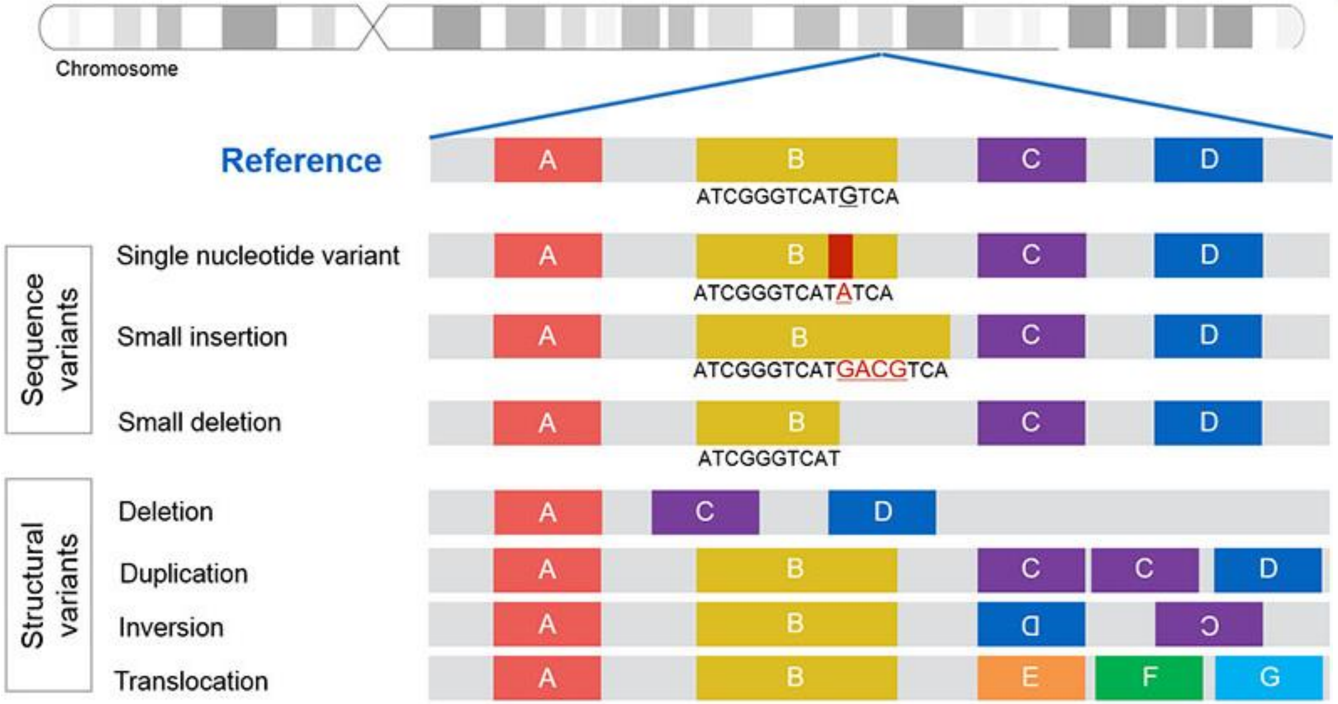
option/parameter
with argument

option/parameter
without argument

positional argument

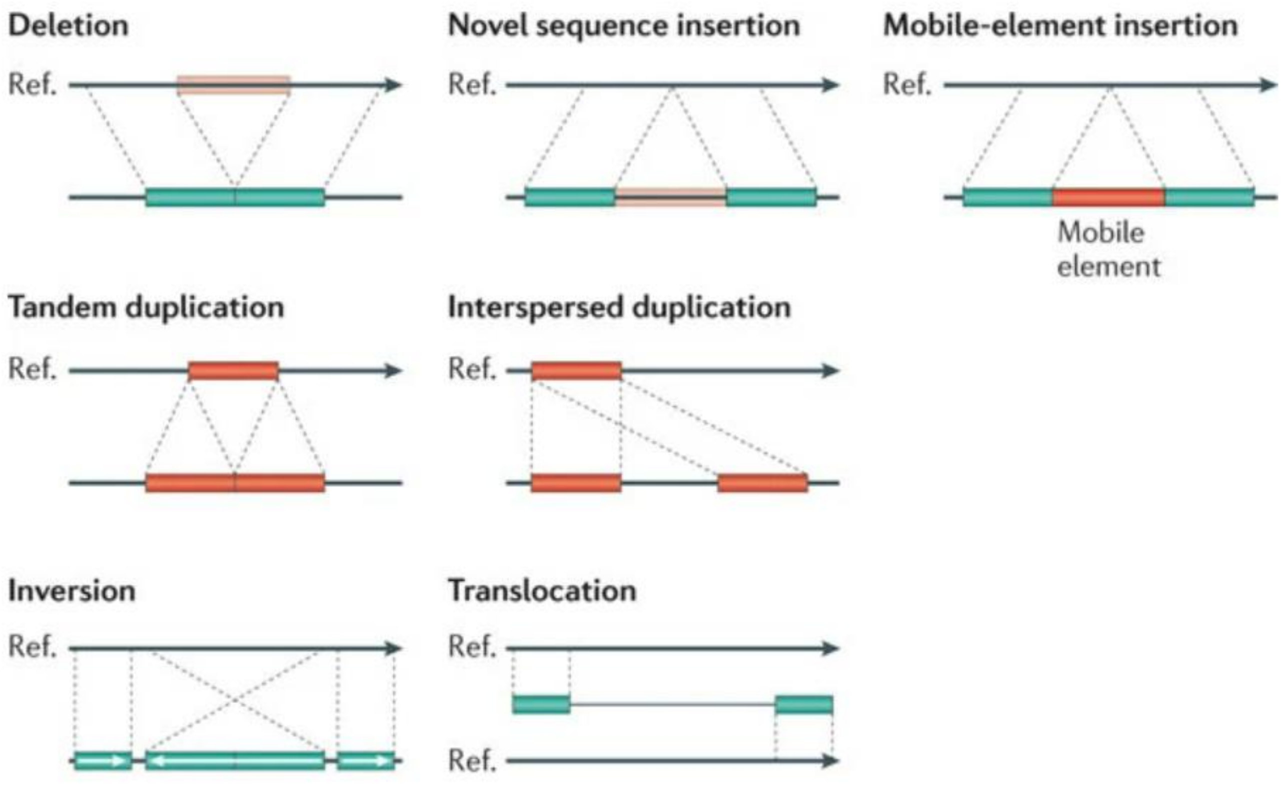
positional argument

Comparing genomes – sequence variants



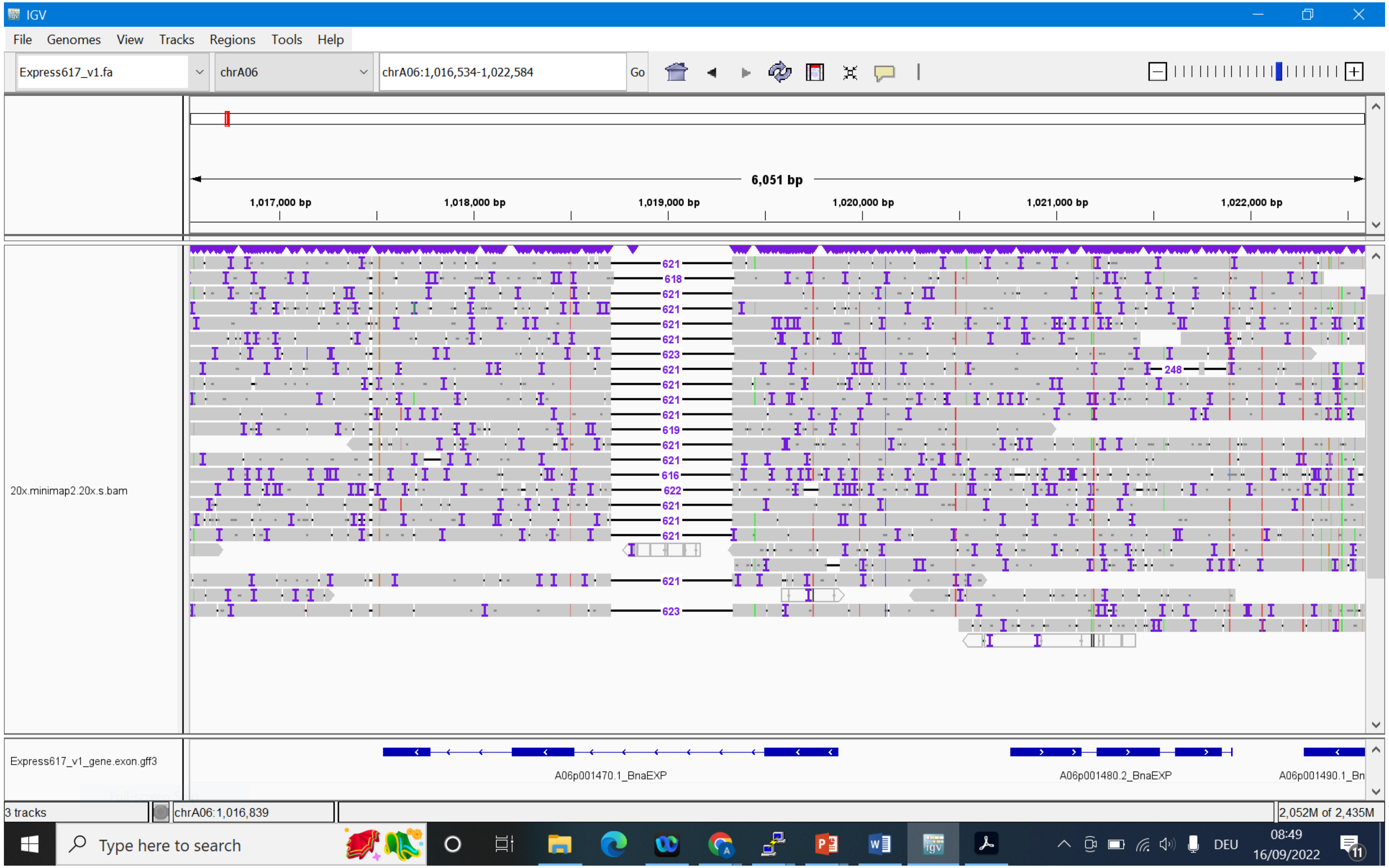
http://www.labspaces.net/blog/1627/When_Whole_Genome_Sequencing_Doesn___t_Give_Us_the_Whole_Genome

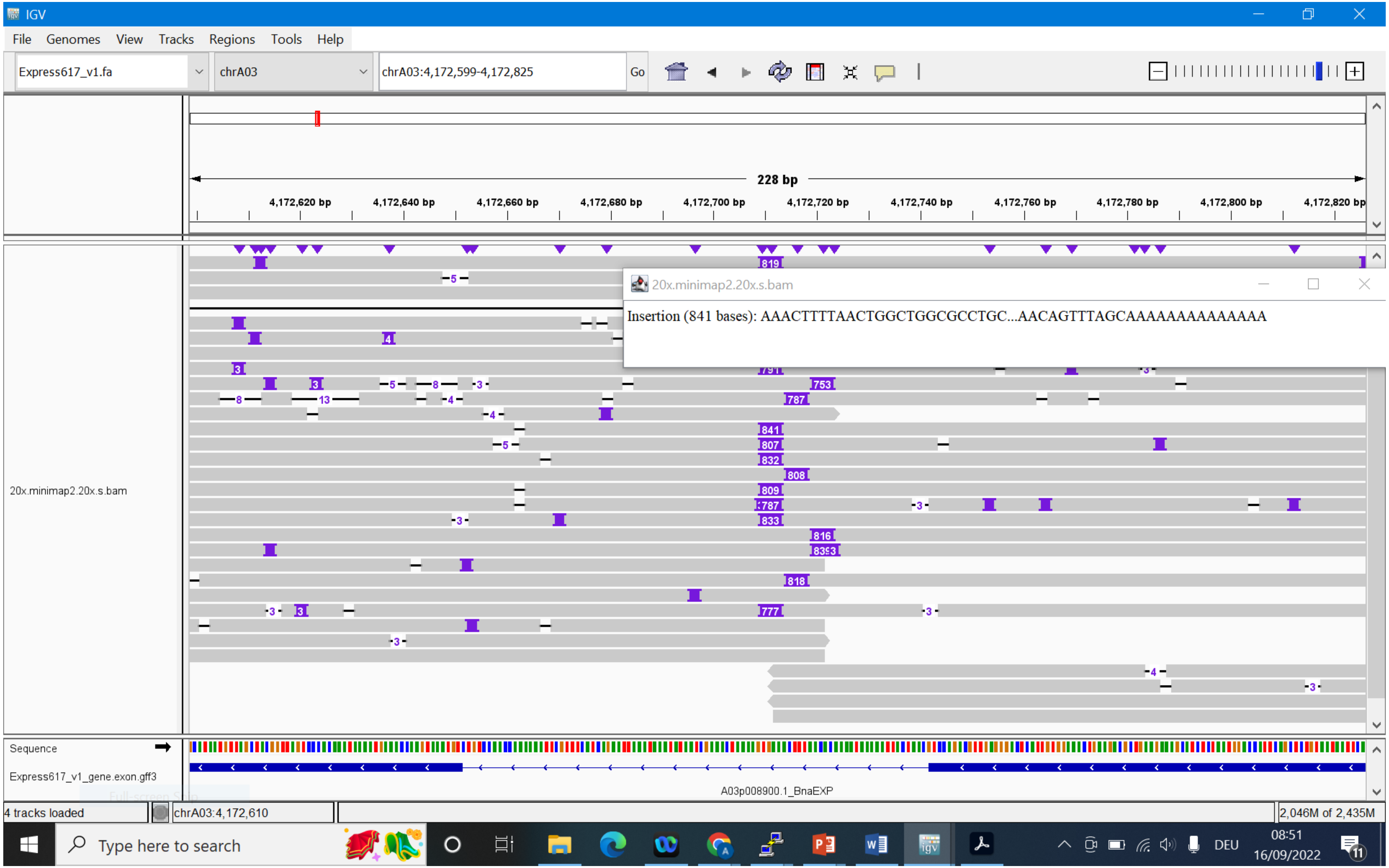
Types of structural variants



Nature Reviews | Genetics

Looking for structural variants with long reads





20x.minimap2.20x.s.bam

Insertion (841 bases): AAACTTTTAACTGGCTGGCGCCTGC...AACAGTTTAGCAAAAAAAAAAAAAA

Structural variant report

Pick one variant from SV.mRNA.overlap.tsv
Please note the ID of the variant and send it to me including your name in the email.

1. Short introduction - what are structural variants. Why are they important?
2. How was the structural variant identified?
3. Describe the structural variant. Include an image from IGV.
4. What are the possible consequences of this structural variant? Does it interrupt a gene, which part of the gene?

Please include at least 5-10 references.