BIOZENTRUM
UNIVERSITAT
WÜRZBURG

Hes ·SO FRIBOURG FREIBURG
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

De novo genome assembly versus mapping to a reference genome

Beat Wolf

PhD. Student in Computer Science

University of Würzburg, Germany

University of Applied Sciences Western Switzerland

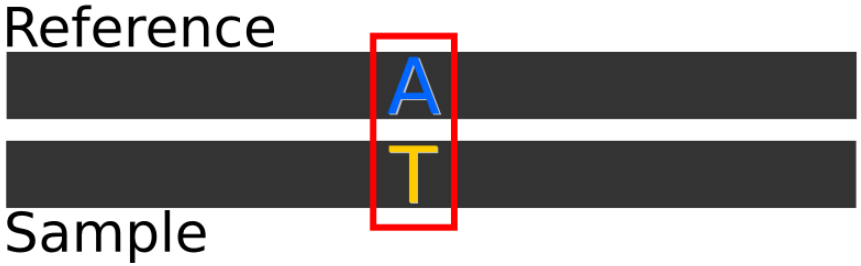
beat.wolf@hefr.ch

Outline

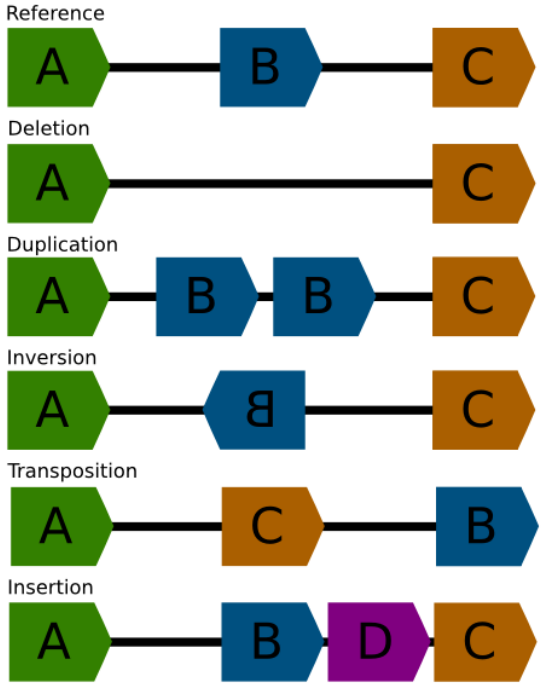
- Genetic variations & sequencing
- De novo sequence assembly
- Reference based mapping/alignment
- Comparison
- Conclusion

Variation types

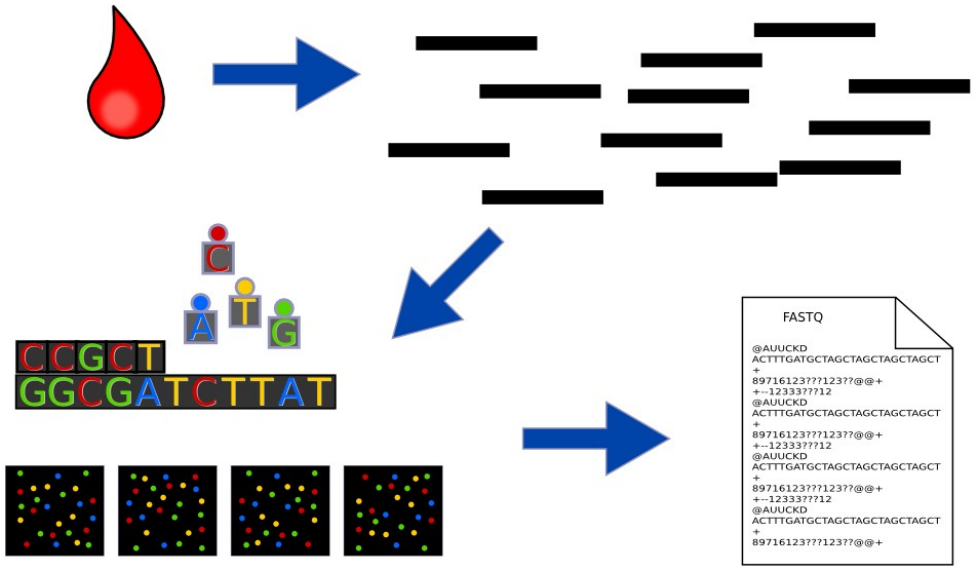
SNV (Single nucleotide variation)



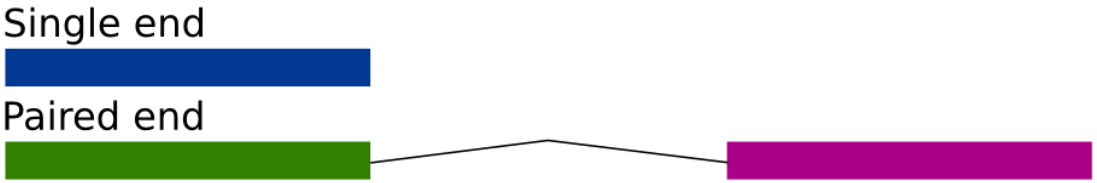
Structural variations



Sequencing technologies



- Different read lengths, 36 – 10'000bp (150-500bp is typical)
- Different sequencing technologies produce different data

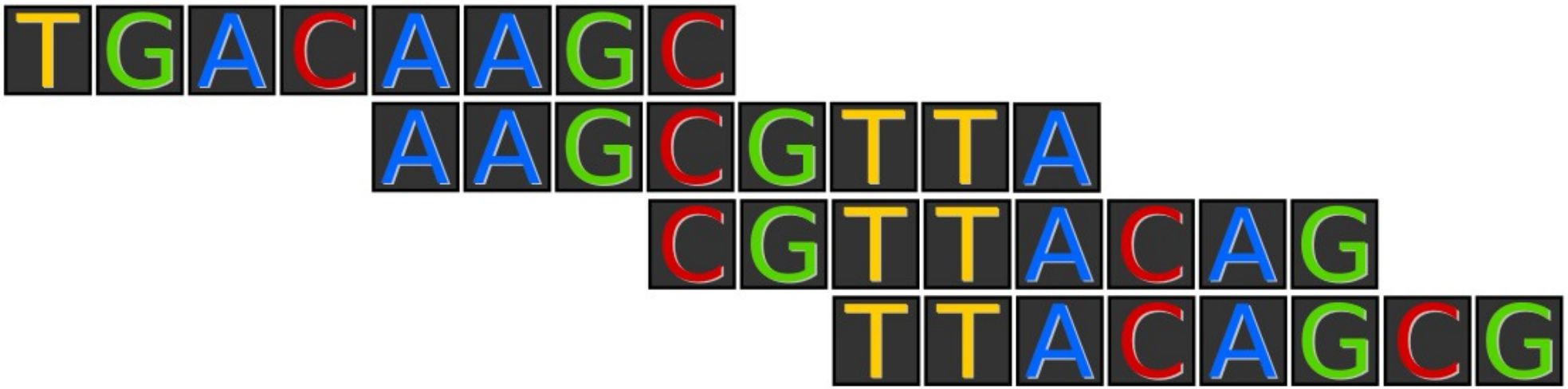


Recreating the genome

- The problem:
  - Recreate the original patient genome from the sequenced reads
    - For which we dont know where they came from and are noisy
- Solutions:
  - Recreate the genome with no prior knowledge using de novo sequence assembly
  - Recreate the genome using prior knowledge with reference based alignment/mapping

De novo sequence assembly

- Ideal approach
- Recreate **original genome** sequence through overlapping sequenced reads

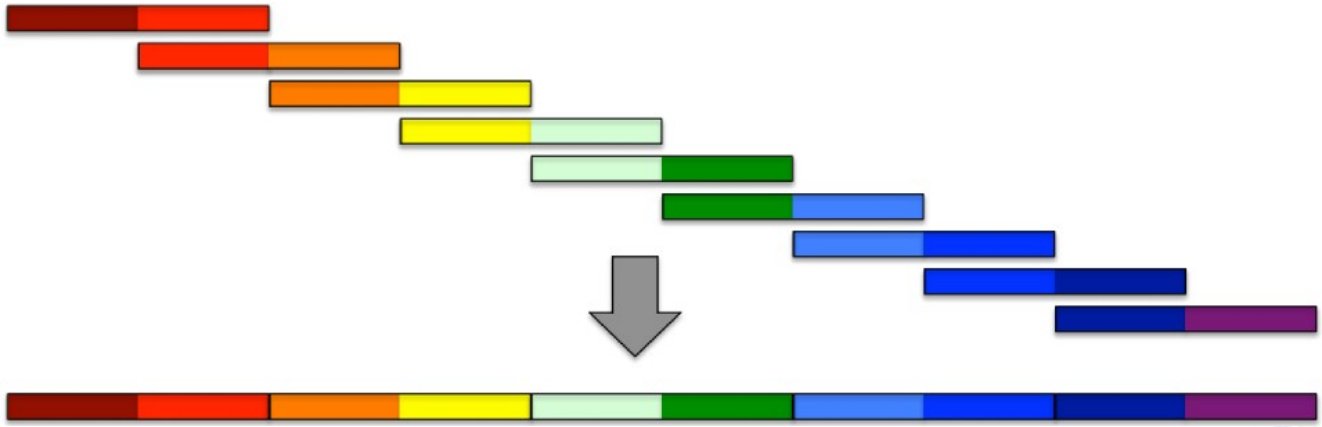


De novo sequence assembly

- Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

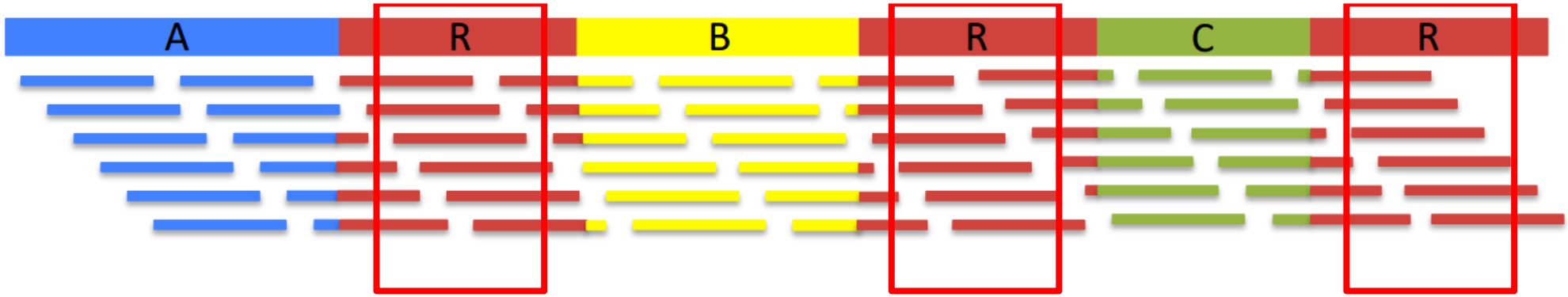
- Simplify assembly graph



Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly

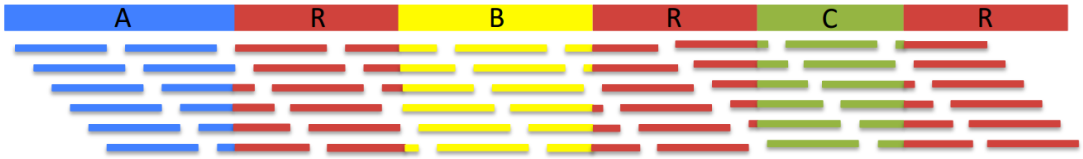
- Genome with repeated regions

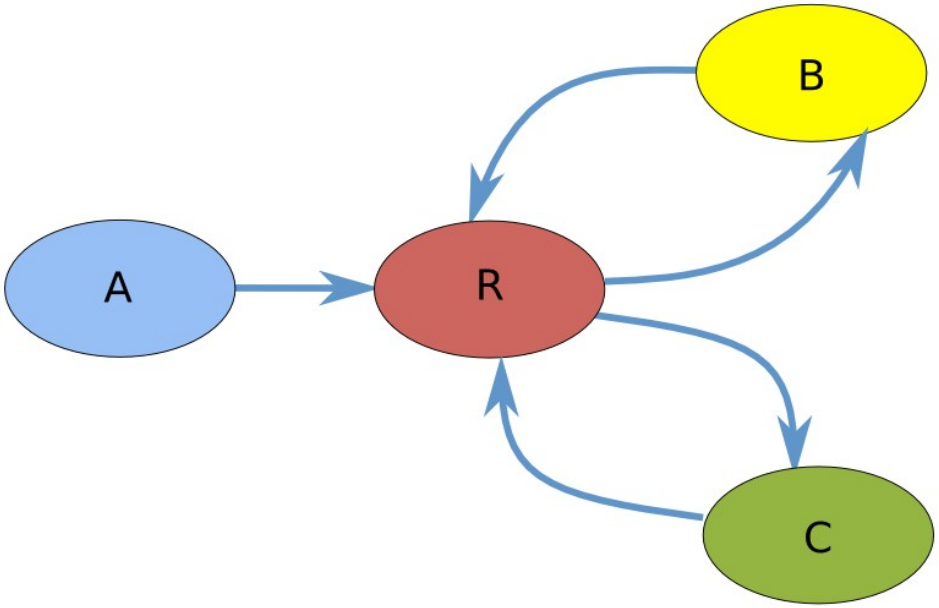


Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly

- Graph generation

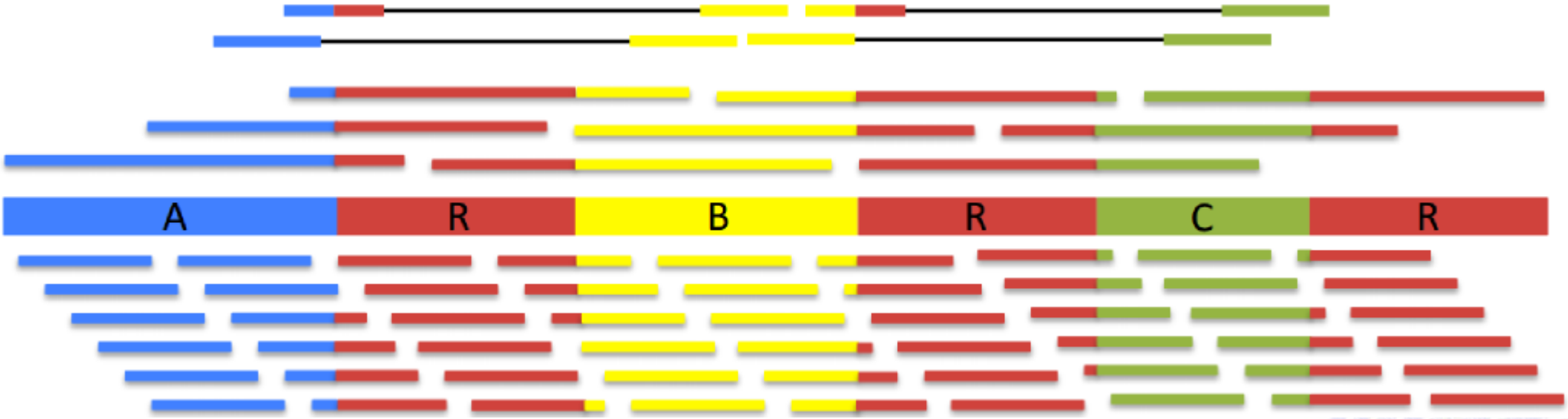




Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly

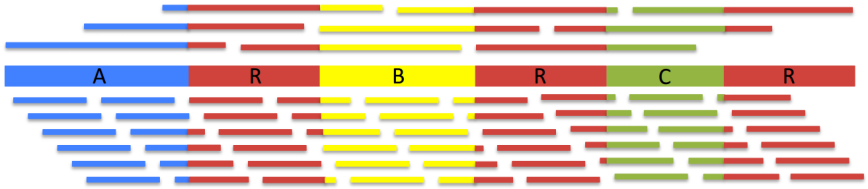
- Double sequencing, once with short and once with long reads (or paired end)

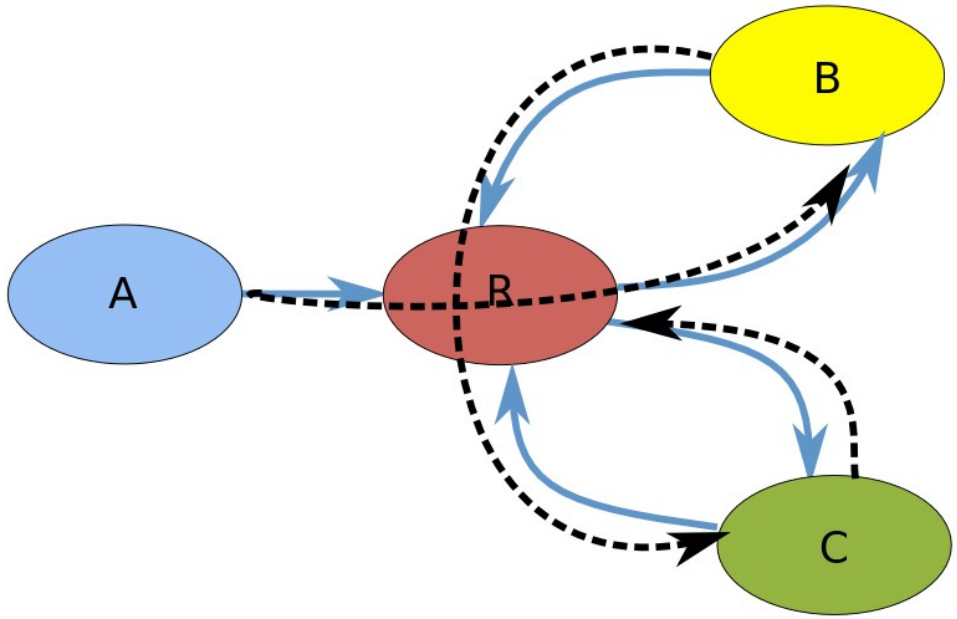


Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly

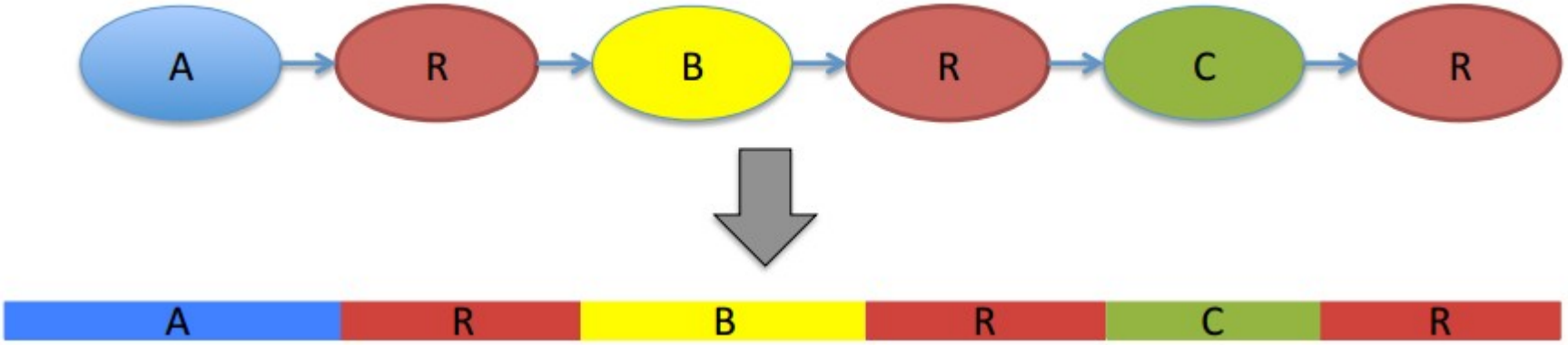
- Finding the correct path through the graph with:
  - Longer reads
  - Paired end reads





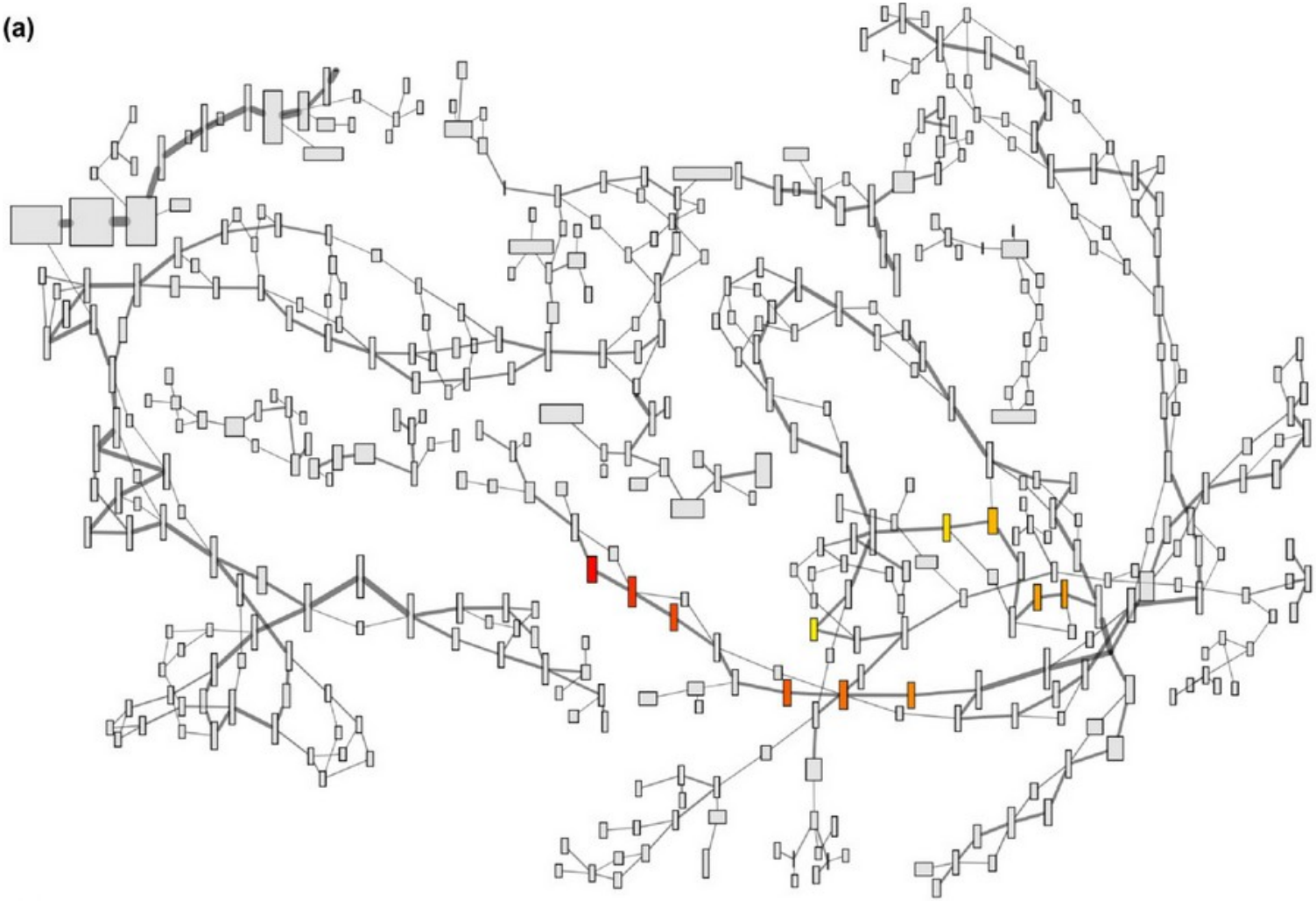
Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly



Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

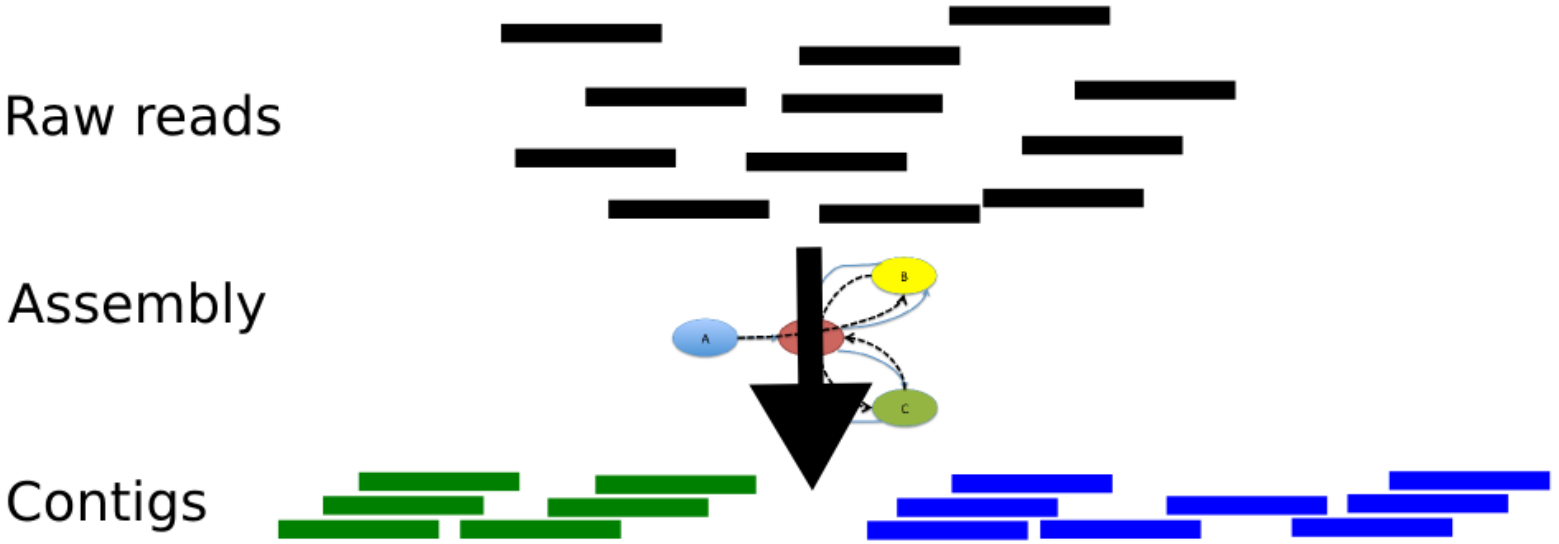
De novo sequence assembly



Modified from: EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data, Miller et al.

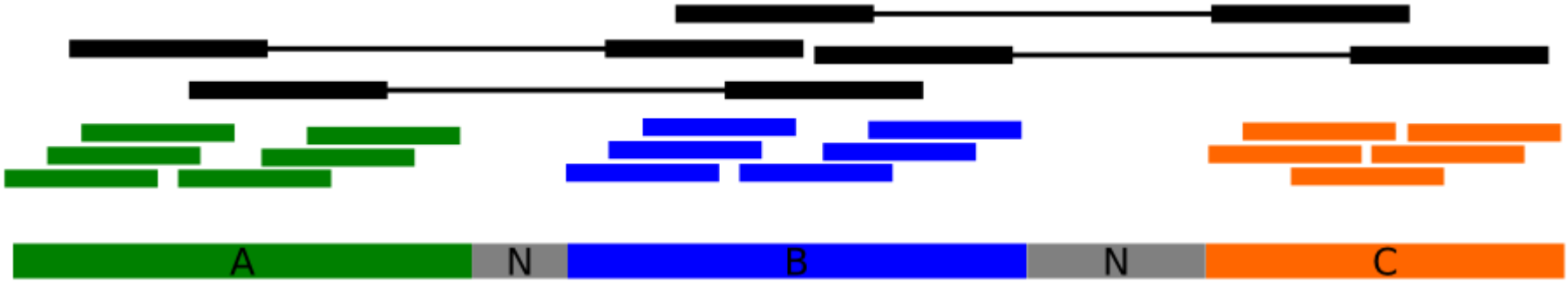
De novo sequence assembly

- Overlapping reads are assembled into groups, so called contigs



De novo sequence assembly

- Scaffolding
  - Using paired end information, contigs can be put in the right order



De novo sequence assembly

- Final result, a list of scaffolds
  - In an ideal world of the size of a chromosome, molecule, mtDNA etc.

Scaffold 1

Scaffold 2

Scaffold 3

Scaffold 4

De novo sequence assembly

- What is needed for a good assembly?
  - High coverage
  - High read lengths
  - Good read quality
- Current sequencing technologies do not have all three
  - Illumina, good quality reads, but short
  - PacBio, very long reads, but low quality

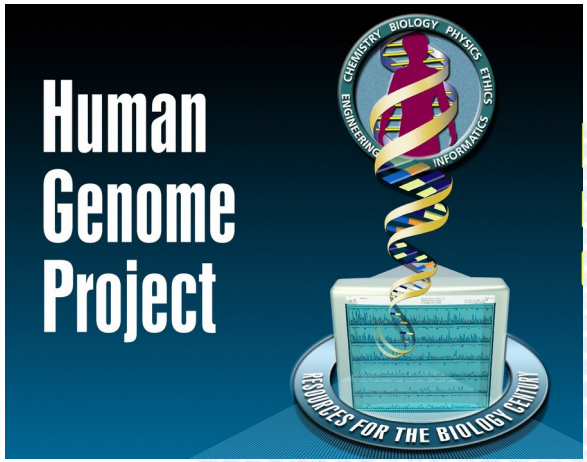
De novo sequence assembly

- Combined sequencing technologies assembly
  - High quality contigs created with short reads
  - Scaffolding of those contigs with long reads

- Double sequencing means
  - High infrastructure requirements
  - High costs

Human reference sequence

- Human Genome project
  - Produced the first „complete“ human genome
- Human genome reference consortium
  - Constantly improves the reference
    - GRCh38 released at the end of 2013



Reference based alignment

- A previously assembled genome is used as a reference
- Sequenced reads are independently aligned against this reference sequence
- Every read is placed at its most likely position
- Unlike sequence assembly, no synergies between reads exist

Reference based alignment

- Naive approach:
  - Evaluate every location on the reference

- Too slow for billions of reads on a big reference

Reference based alignment

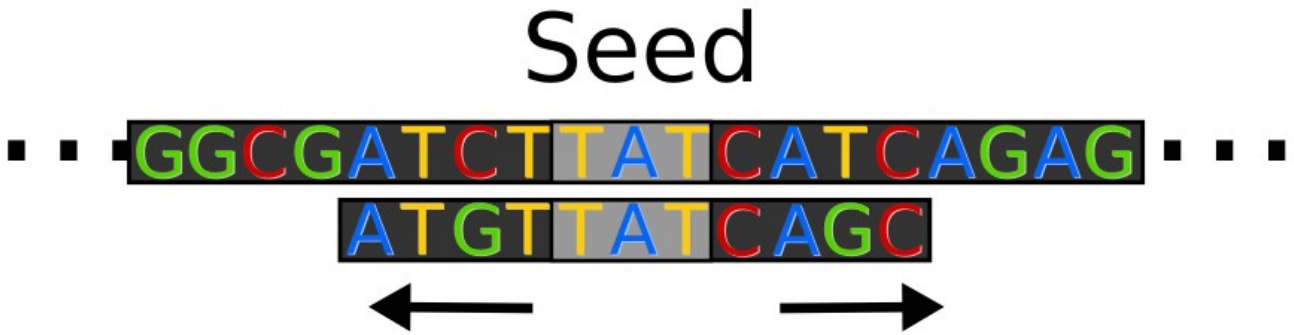
- Speed up with the creation of a reference index

- Fast lookup table for subsequences in reference

Reference based alignment

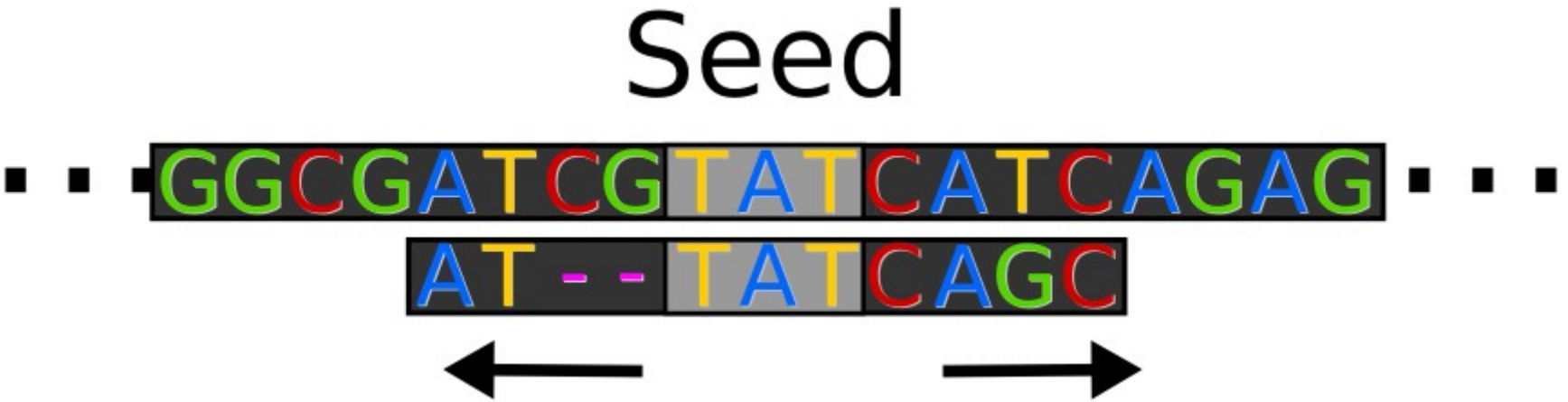
- Find all possible alignment positions
  - Called seeds

- Evaluate every seed



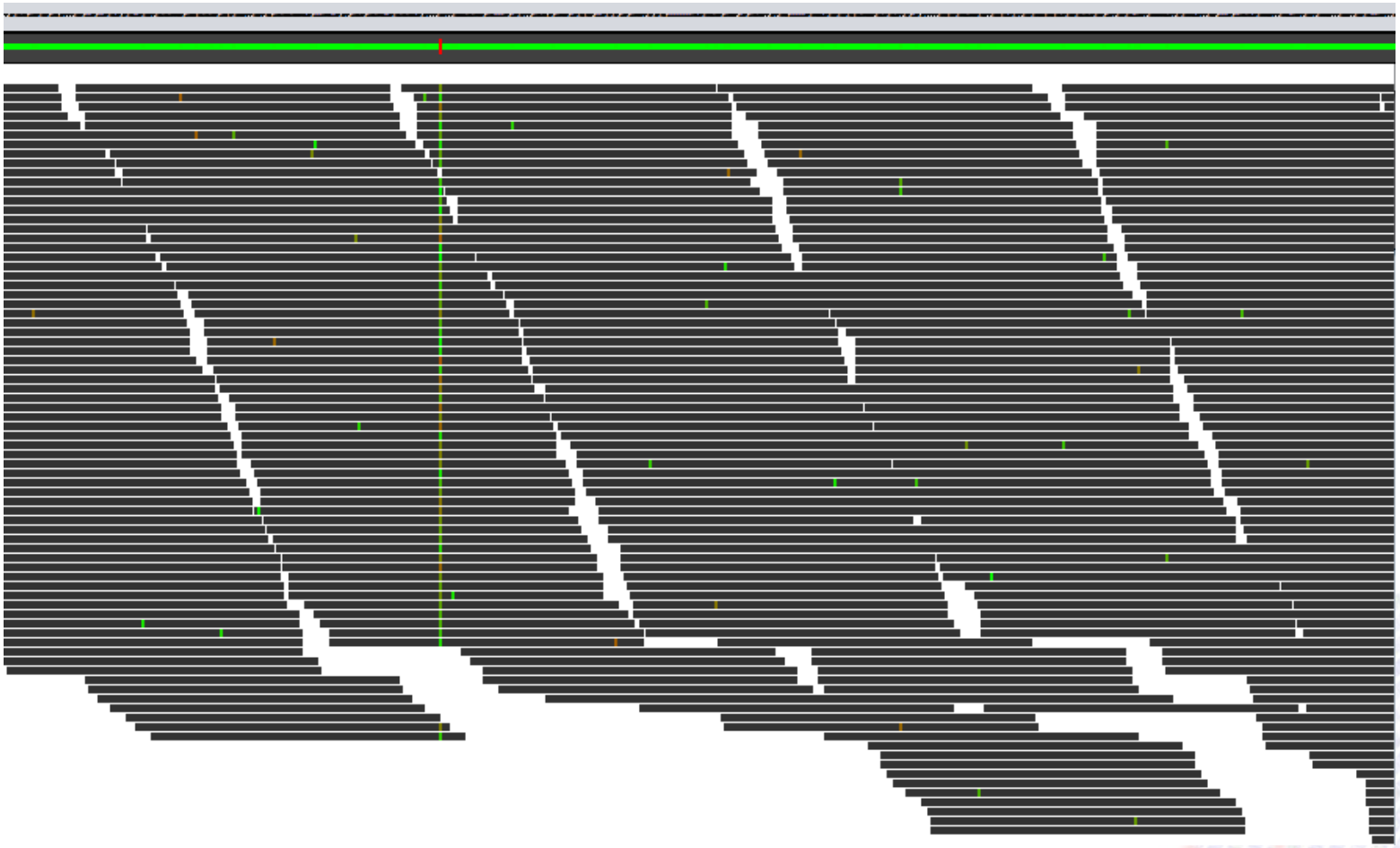
Reference based alignment

- Determine optimal alignment for the best candidate positions
- Insertions and deletions increase the complexity of the alignment



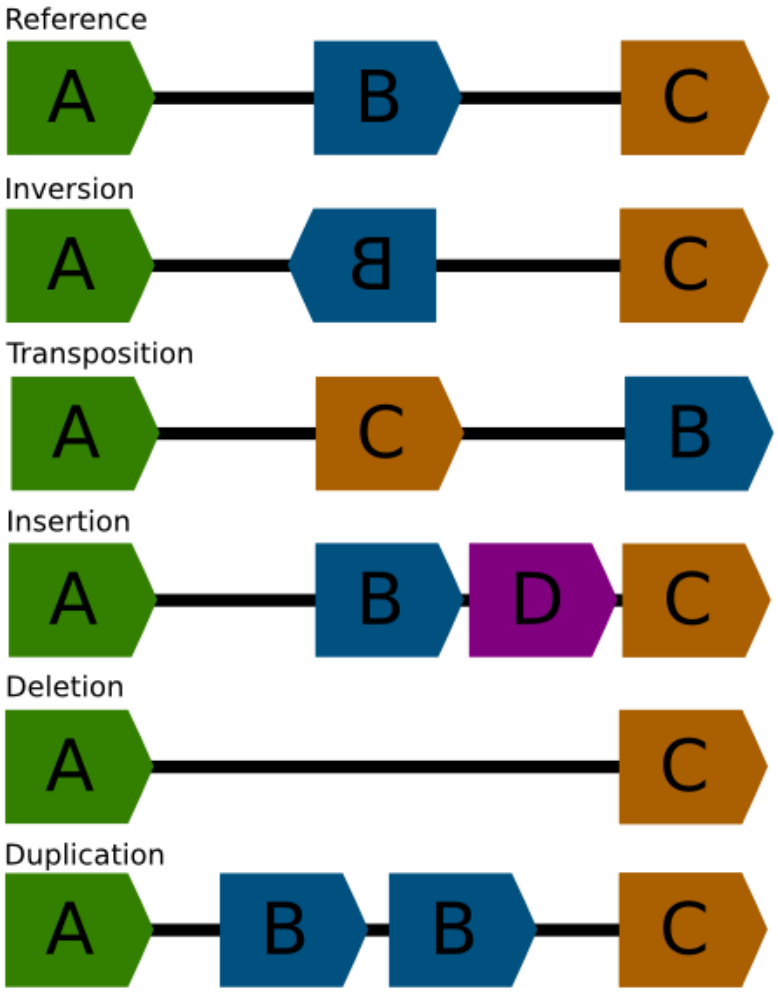
Reference based alignment

- Final result, an alignment file (BAM)



Alignment problems

- Regions very different from reference sequence
  - Structural variations
    - Except for deletions and duplications



Alignment problems

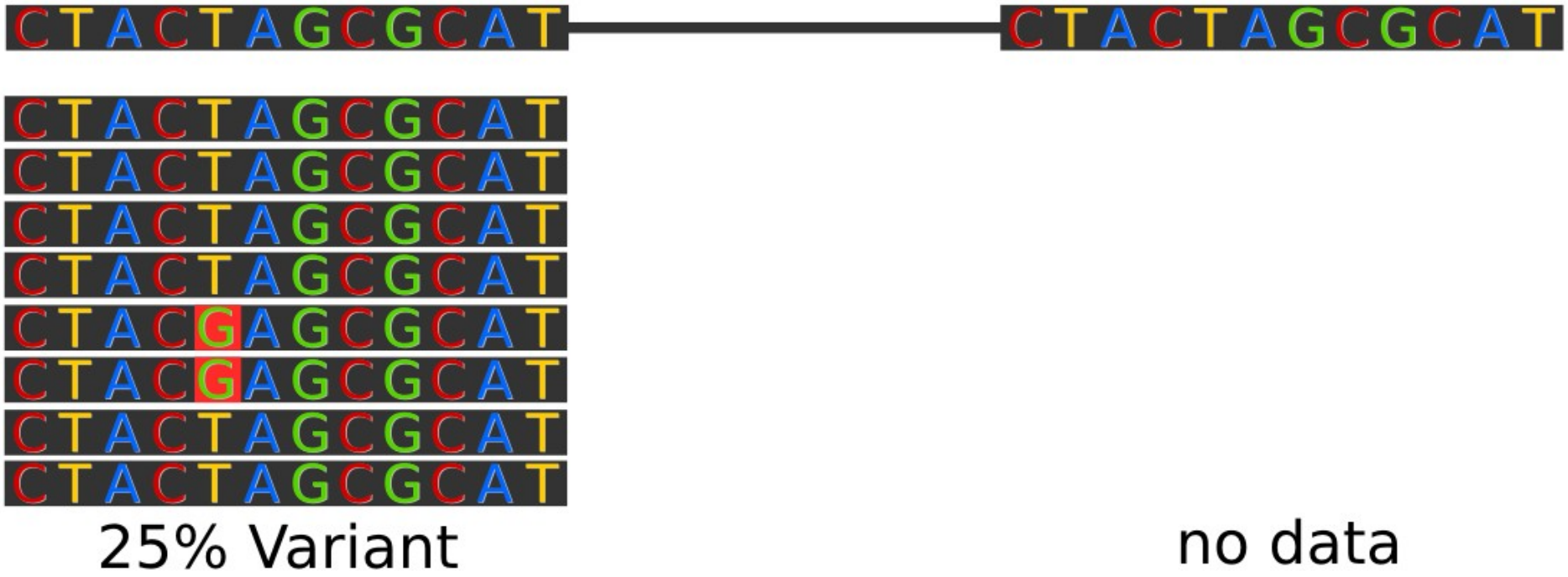
- Reference which contains duplicate regions
- Different strategies exist if multiple positions are equally valid:
  - Ignore read
  - Place at multiple positions
  - Choose one location at random
  - Place at first position
  - Etc.

Alignment problems

- Example situation
  - 2 duplicate regions, one with a heterozygote variant

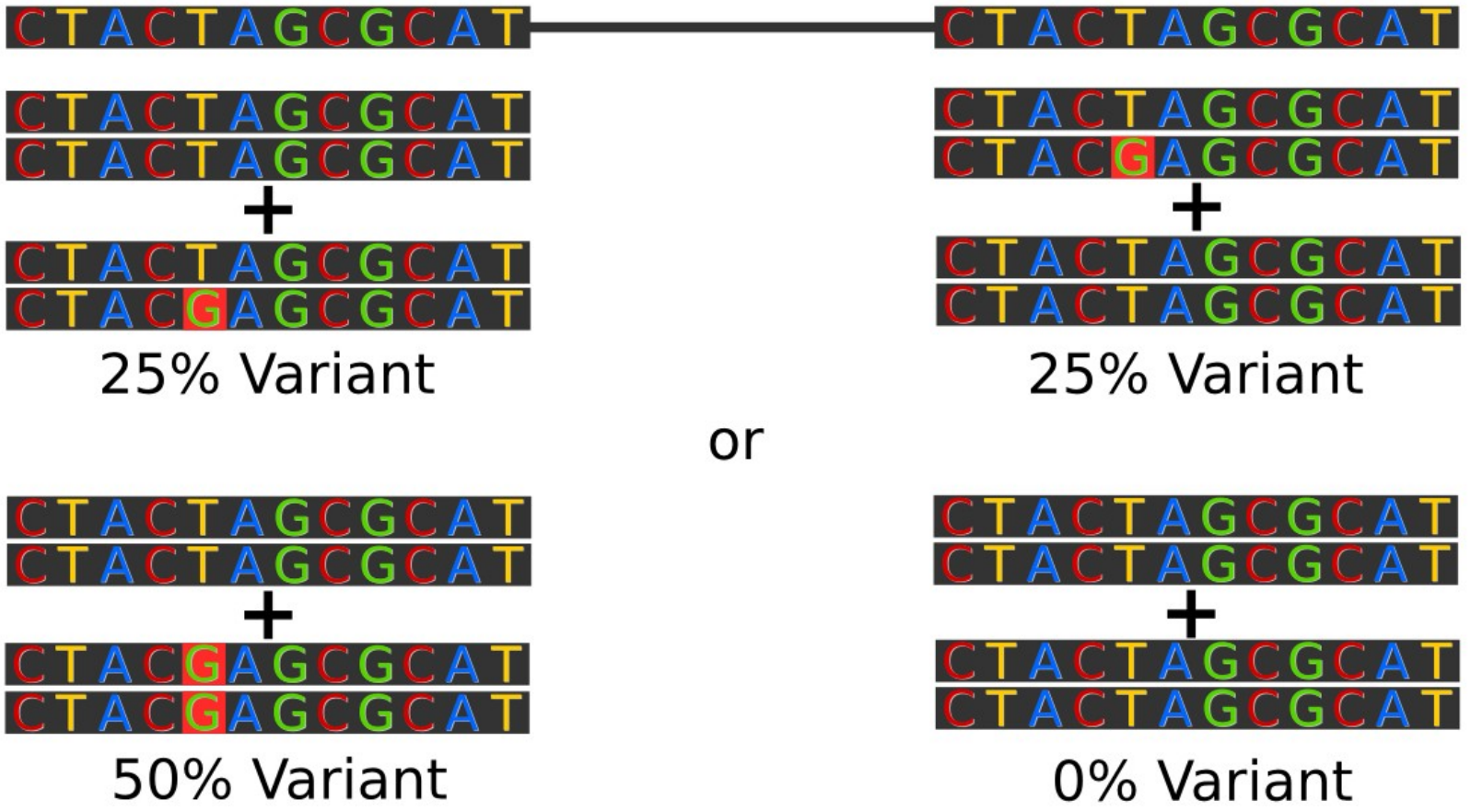
Alignment problems

- Map to first position



Alignment problems

- Map to random position



Alignment problems

- To dustbin

deletion

deletion

Dustbin

- Sequences that are not aligned can be recovered in the dustbin
  - Sequences with no matching place on reference
  - Sequences with multiple possible alignments
- Several strategies exist to handle them
  - De novo assembly
  - Realigning with a different aligner
  - Etc.
- Important information can often be found there

De novo vs. reference

- Reference based alignment
  - Good for SNV, small indels
  - Limited by read length for feature detection
  - Works for deletions and duplications (CNVs)
    - Using coverage information
  - Alignments are done “quickly“
  - Very good at hiding raw data limitations
  - The alignment does not necessarily correspond to the original sequence
  - Requires a reference that is close to the sequenced data

De novo vs. reference

- De novo assembly
  - Assemblies try to recreate the original sequence
  - Good for structural variations
  - Good for completely new sequences not present in the reference
  - Slow and high infrastructure requirements
  - Very bad at hiding raw data limitations

De novo vs. reference

- Unless necessary, stick with reference based alignment
  - Easier to use
  - More tools to work with the results
  - Easier annotation and comparison
  - Current standard in diagnostics
  - Can still benefit from de novo alignment through local de novo realignment
  - Analyze dustbin if results are inconclusive

Conclusion

- Reference based alignment is the current standard in diagnostics
- Assemblies can be used if reference based alignment is not conclusive
- Assembly will become much more important in the future when sequencing technologies are improved

BIOZENTRUM
UNIVERSITAT
WÜRZBURG

Hes ·SO FRIBOURG FREIBURG
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

Thank you for your attention

beat.wolf@hefr.ch

Further resources

Next Generation Variant Calling:
http://blog.goldenhelix.com/?p=1434
De novo alignment:
http://schatzlab.cshl.edu/presentations/
Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly:
http://www.nature.com/nbt/journal/v29/n8/abs/nbt.1904.html

3Gb TEST