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

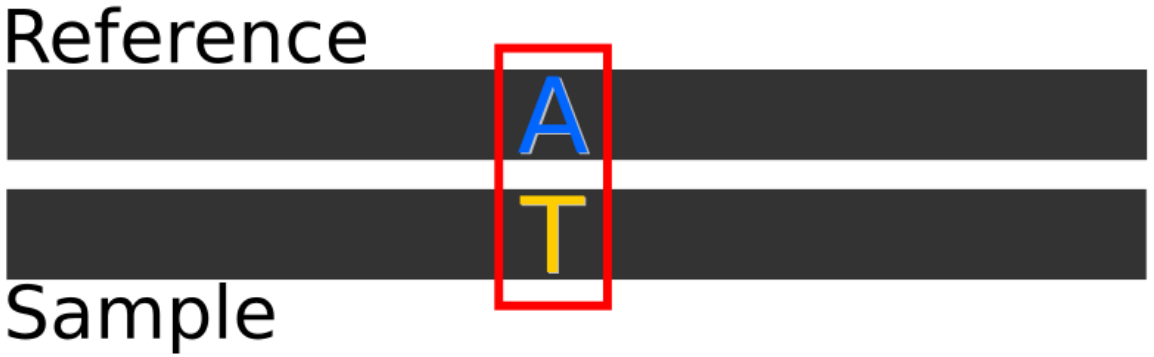
- Genetic variations
- De novo sequence assembly
- Reference based mapping/alignment
- Variant calling
- Comparison
- Conclusion

What are variants?

- Difference between a sample (patient) DNA and a reference (another sample or a population consensus)
- Sum of all variations in a patient determine his genotype and phenotype

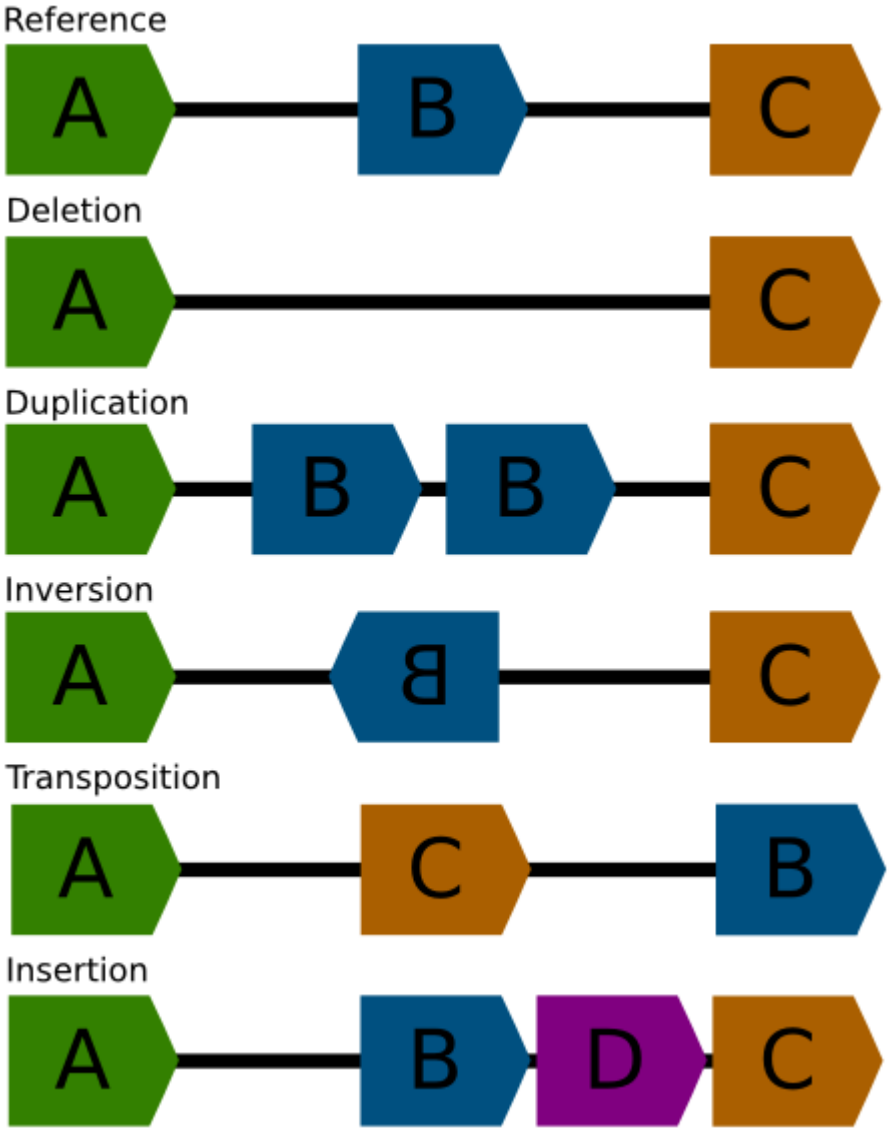
Variation types

- Small variations (< 50bp)
  - SNV (Single nucleotide variation)



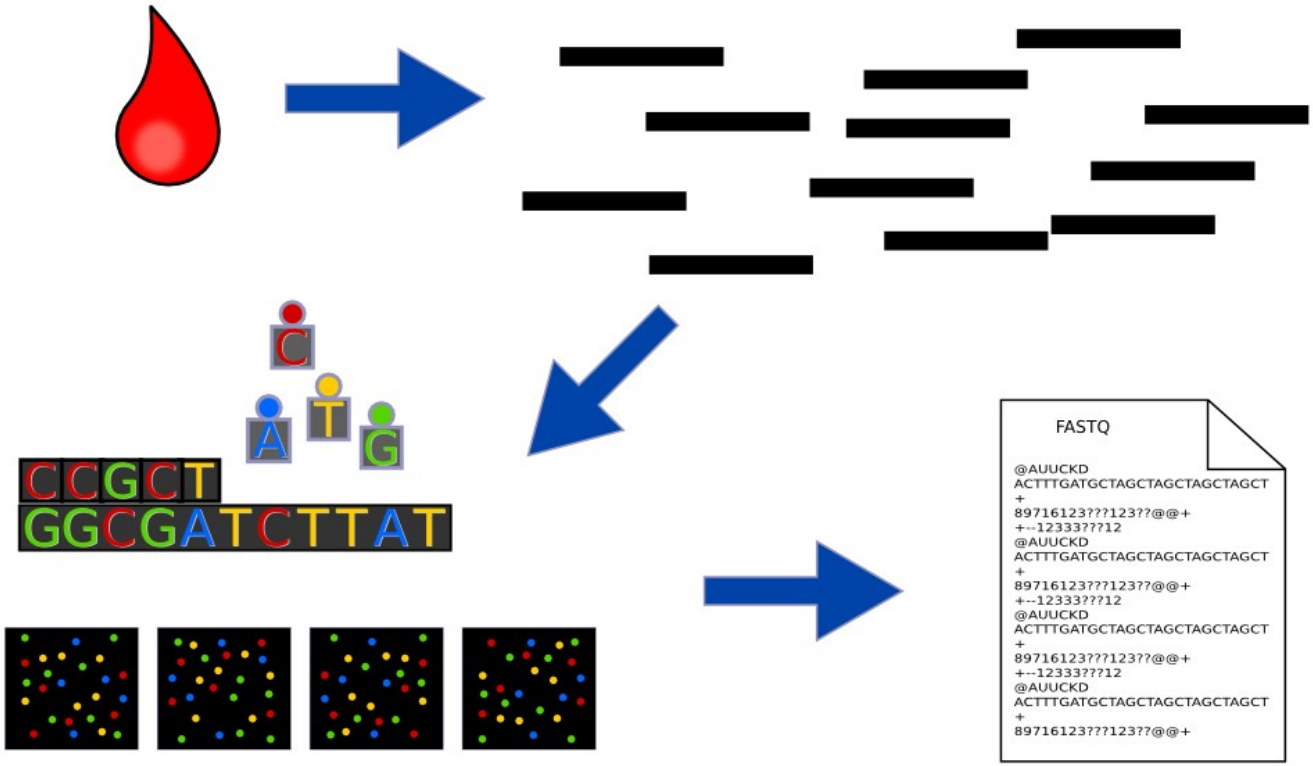
  - Indel (insertion/deletion)

Structural variations



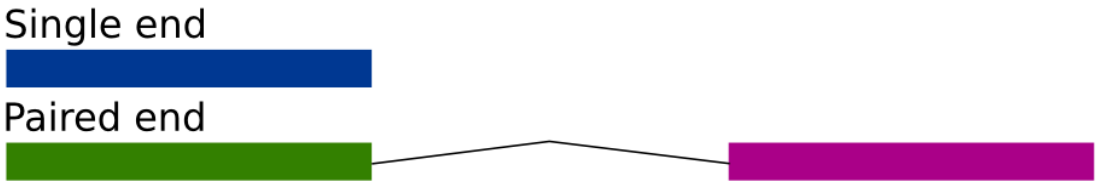
Sequencing technologies

- Sequencing produces small overlapping sequences



Sequencing technologies

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



And different kinds of errors

- Substitutions (Base replaced by other)
- Homopolymers (3 or more repeated bases)
  - AAAAA might be read as AAAA or AAAAAA
- Insertion (Non existent base has been read)
- Deletion (Base has been skipped)
- Duplication (cloned sequences during PCR)
- Somatic cells sequenced

Sequencing technologies

- Standardized output format: **FASTQ**
  - Contains the read sequence and a quality for every base

```
@9WV6Z:791:946
GCTCTTCCGATCTATGGATGCACCAAGATATATGACCCTGTCTGTGGGACTGATGGAA
+
774399222034221774399222034221774399222034221774399222034
```

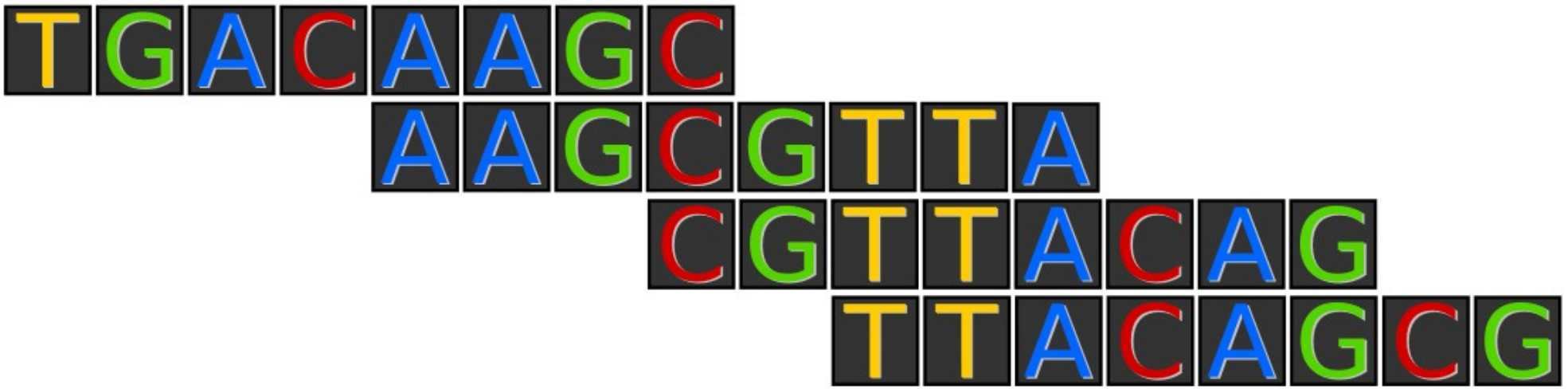
http://en.wikipedia.org/wiki/FASTQ_format

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
- Solution:
  - 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

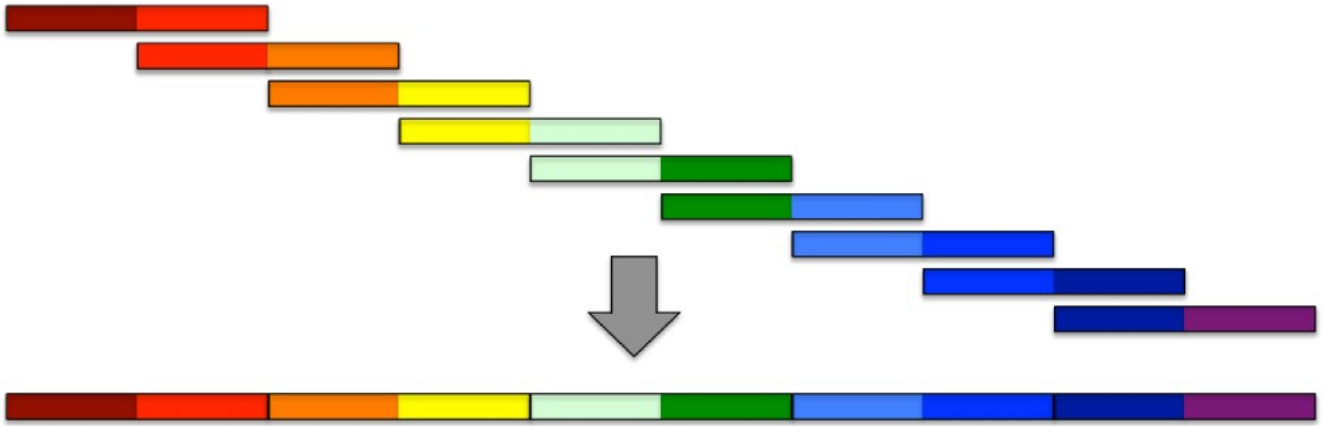


De novo sequence assembly

- Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

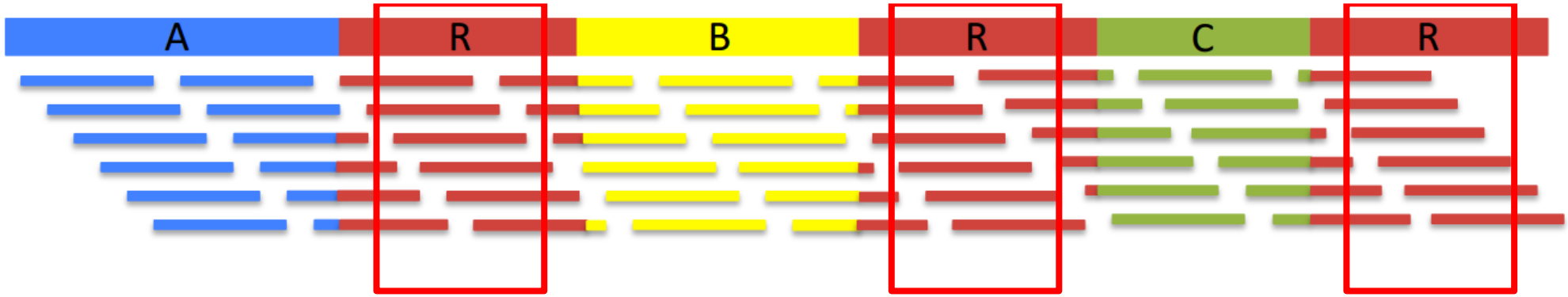
- Simplify assembly graph



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

De novo sequence assembly

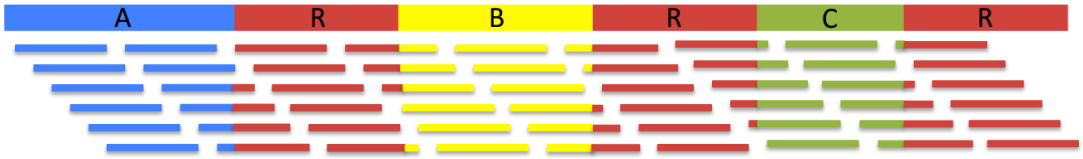
- Genome with repeated regions

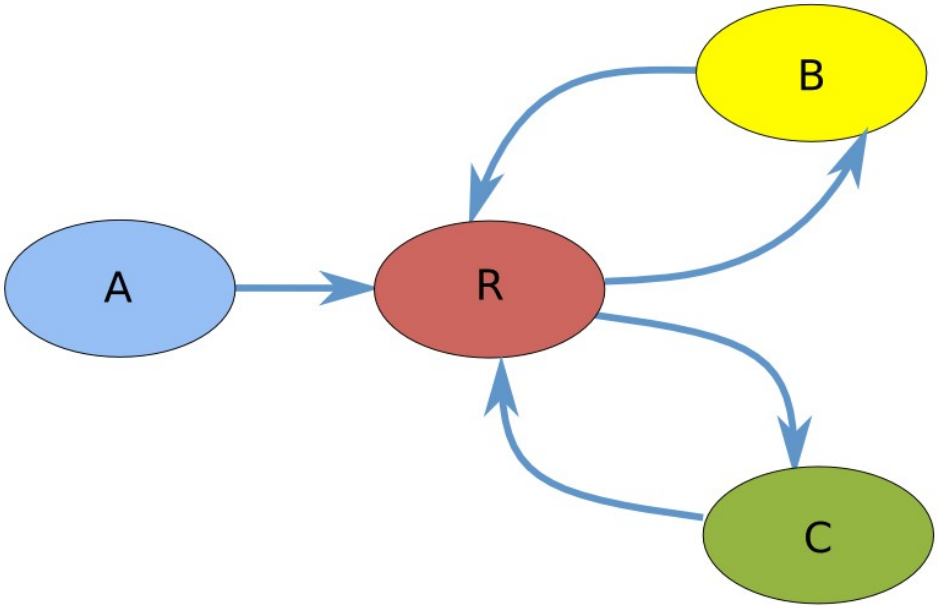


Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly

- Graph generation

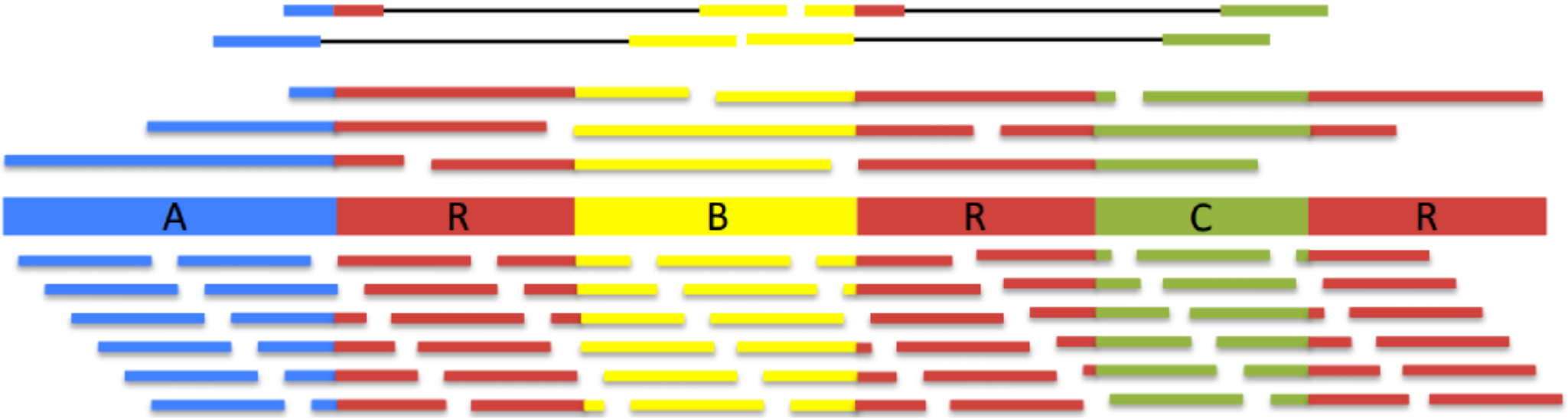




Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly

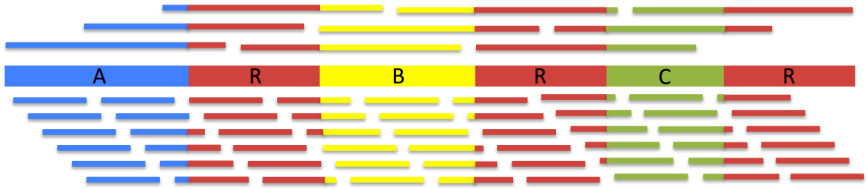
- Double sequencing, once with short and once with long reads (or paired end)

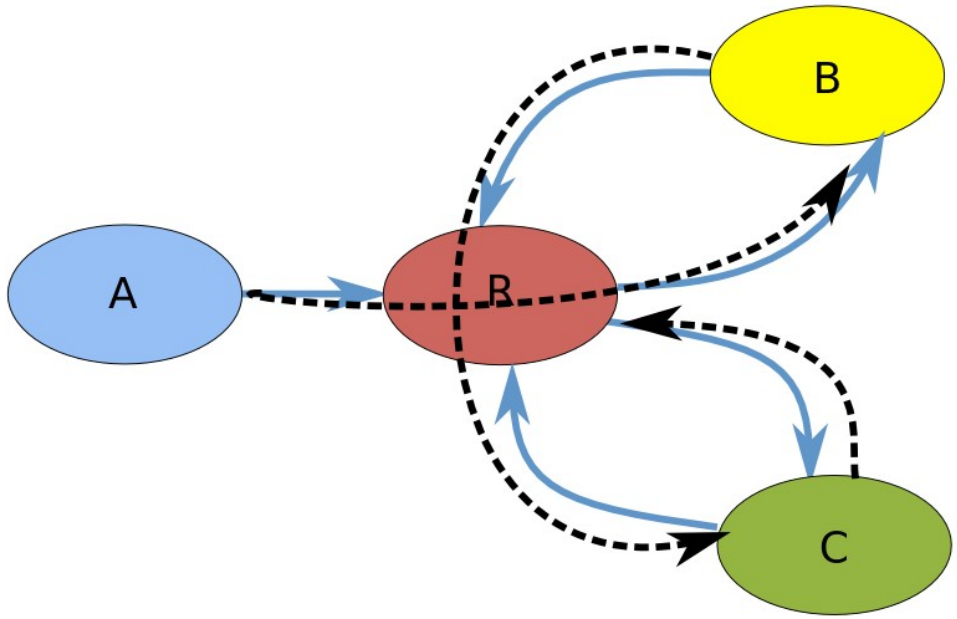


Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly

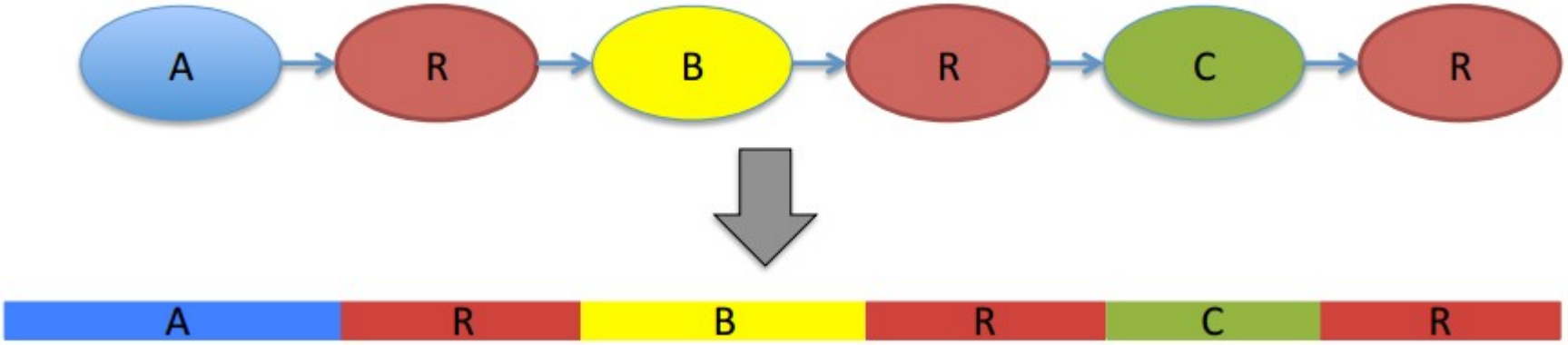
- Finding the correct path through the graph with:
  - Longer reads
  - Paired end reads





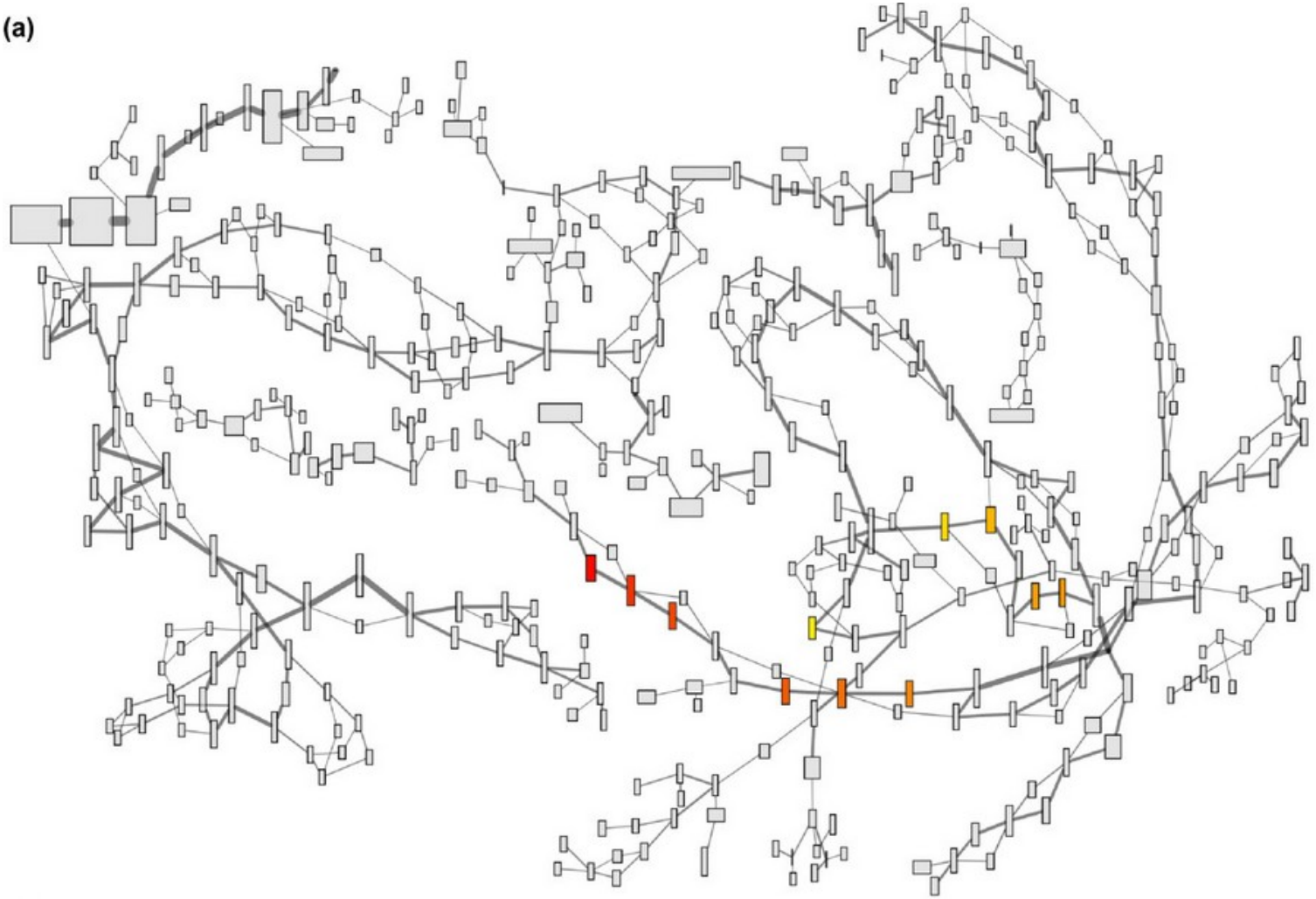
Modified from: De novo assembly of complex genomes using single molecule sequencing, Michael Schatz

De novo sequence assembly



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

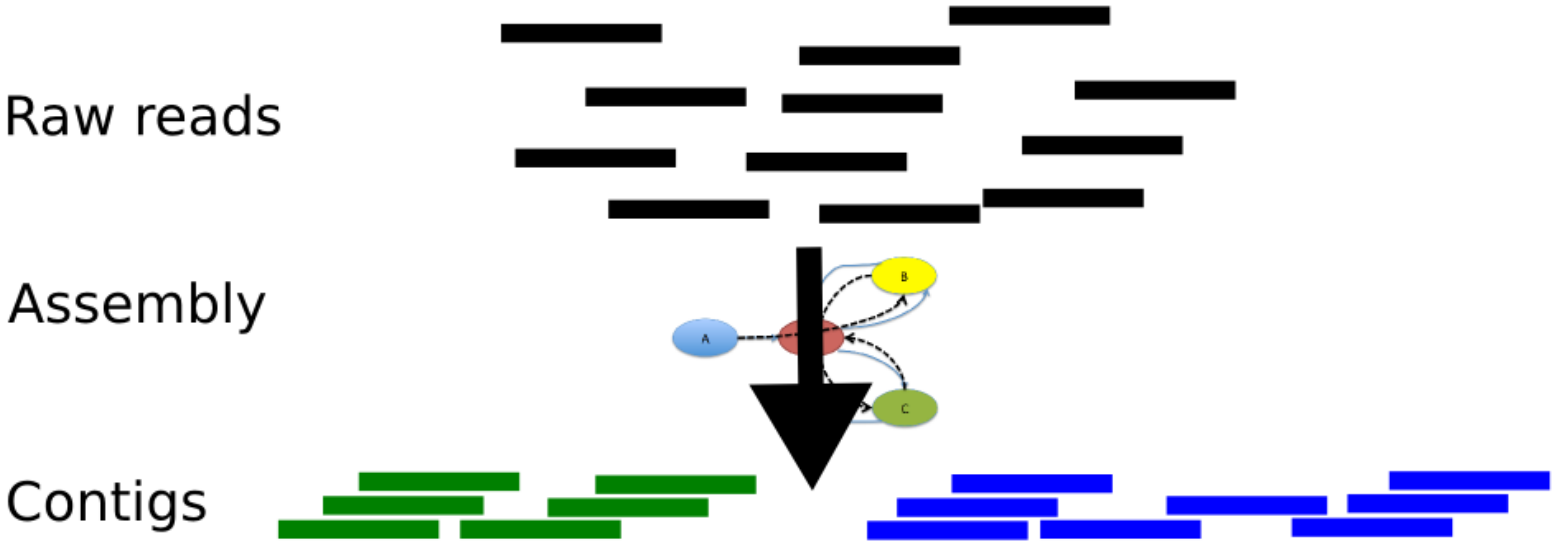
De novo sequence assembly



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

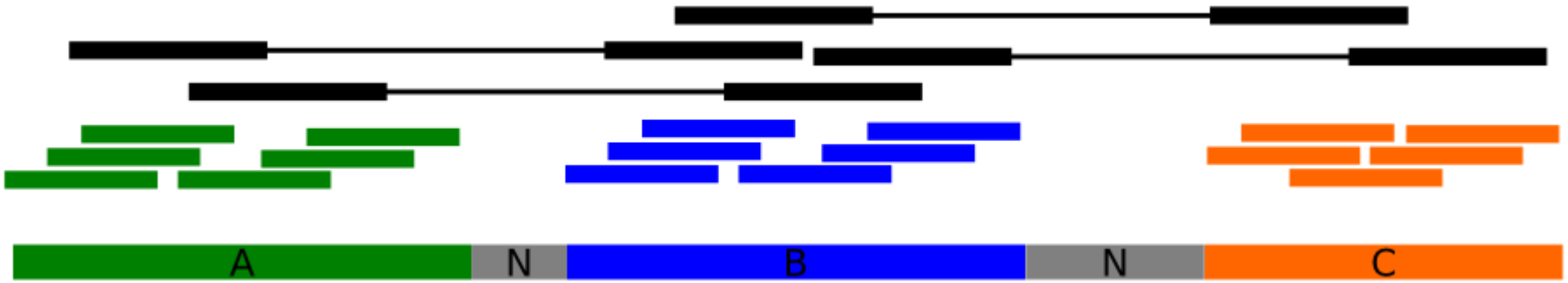
De novo sequence assembly

- 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

De novo sequence assembly

- Field of assemblers is constantly evolving
  - Competitions like Assemblathon 1 + 2 exist

    https://genome10k.soe.ucsc.edu/assemblathon
- The results vary greatly depending on datatype and species to be assembled
- High memory and computational complexity

De novo sequence assembly

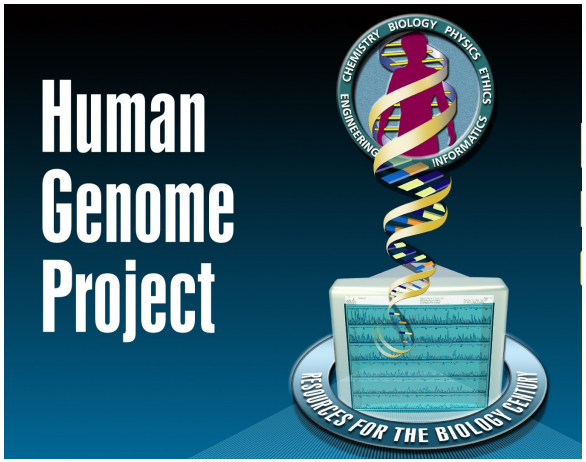
- Short list of assemblers
  - ALLPATHS-LG
  - Meraculous
  - Ray
- Software used by winners of Assemblathon 2:

  SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap, CrossMatch, Velvet, BLAST, and BLASR

- Creating a high quality assembly is complicated

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

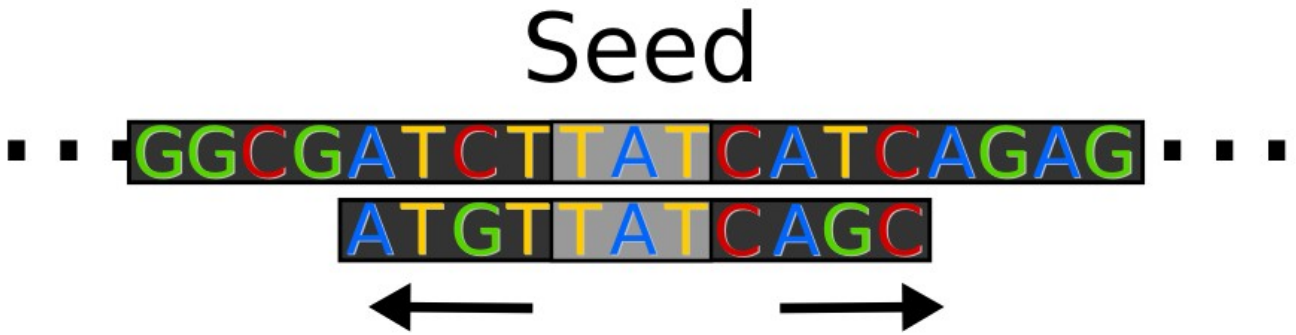
- Speed up with the creation of a reference index

- Fast lookup table for subsequences in reference

Reference based alignment

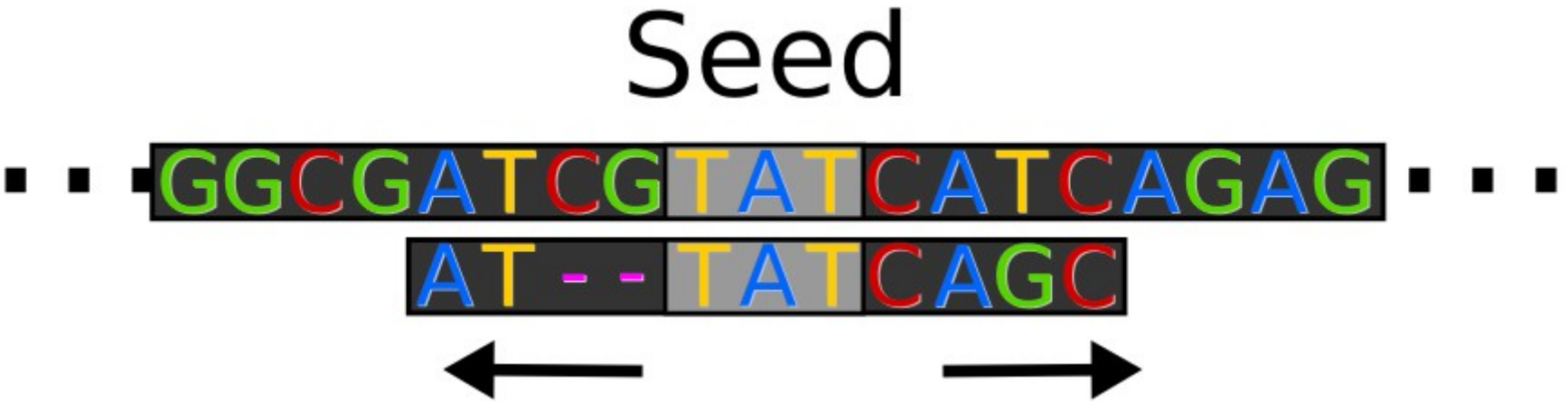
- Find all possible alignment positions
  - Called seeds

- 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

- Most common technique, dynamic programming
- Smith-Watherman, Gotoh etc. are common algorithms

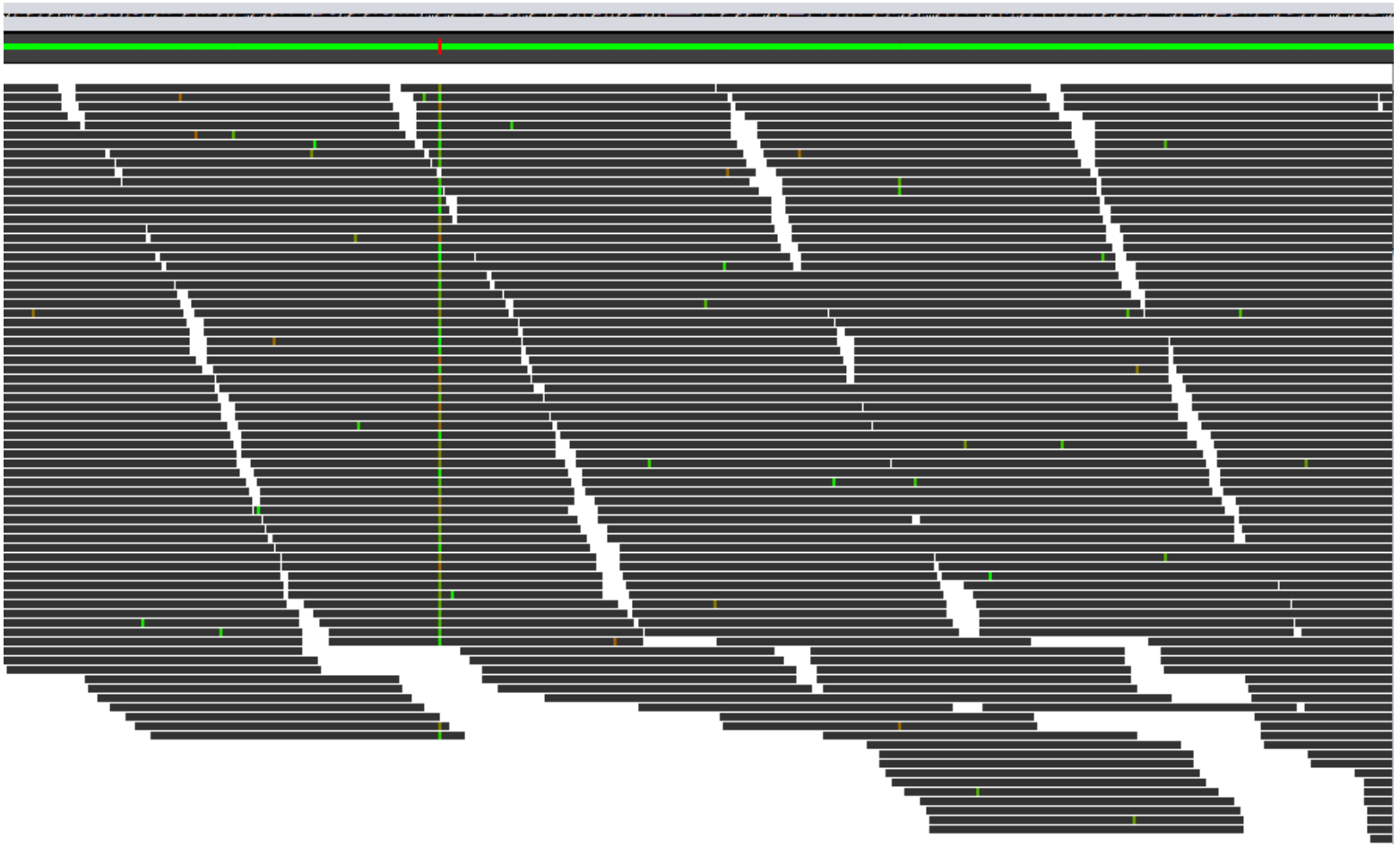
$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

$$T = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ G & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow \end{pmatrix}$$

http://en.wikipedia.org/wiki/Smith-Waterman_algorithm

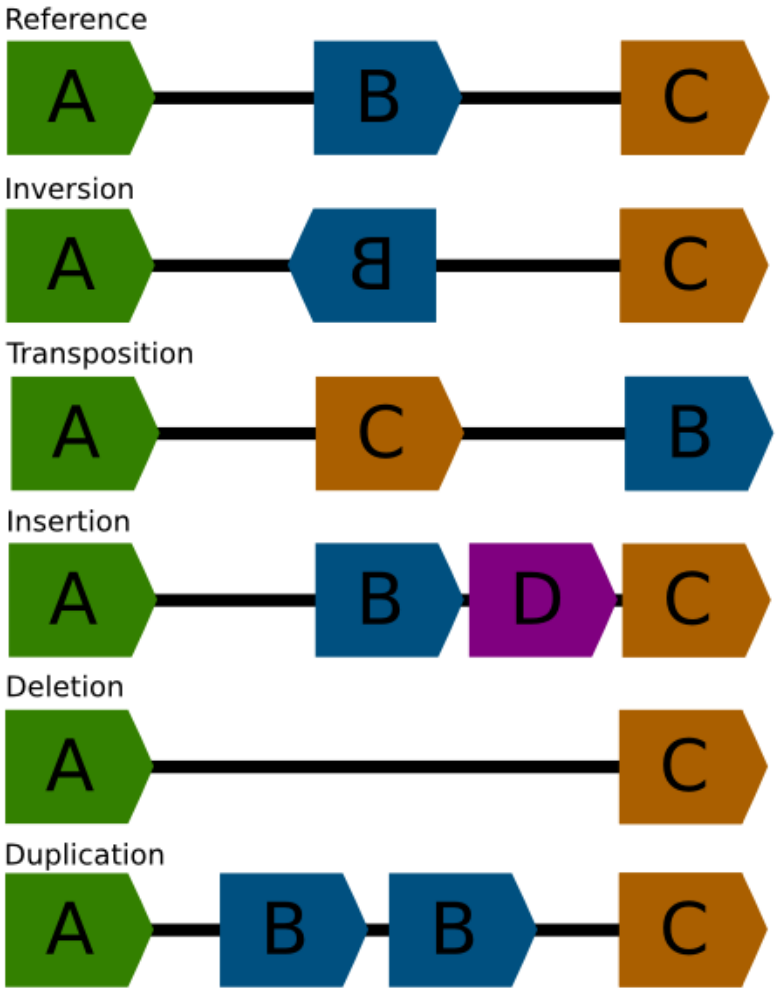
Reference based alignment

- Final result, an alignment file (BAM)



Alignment problems

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



Alignment problems

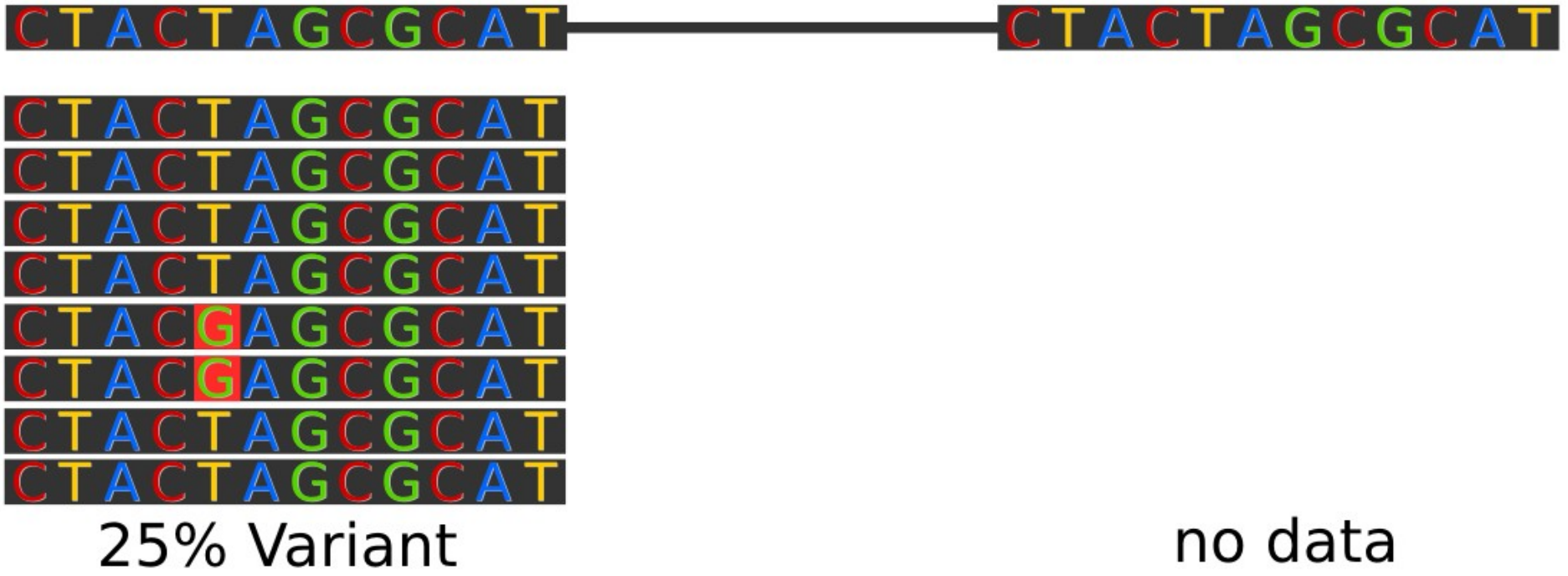
- Reference which contains duplicate regions
- Different strategies exist if multiple positions are equally valid:
  - Ignore read
  - Place at multiple positions
  - Choose one location at random
  - Place at first position
  - Etc.

Alignment problems

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

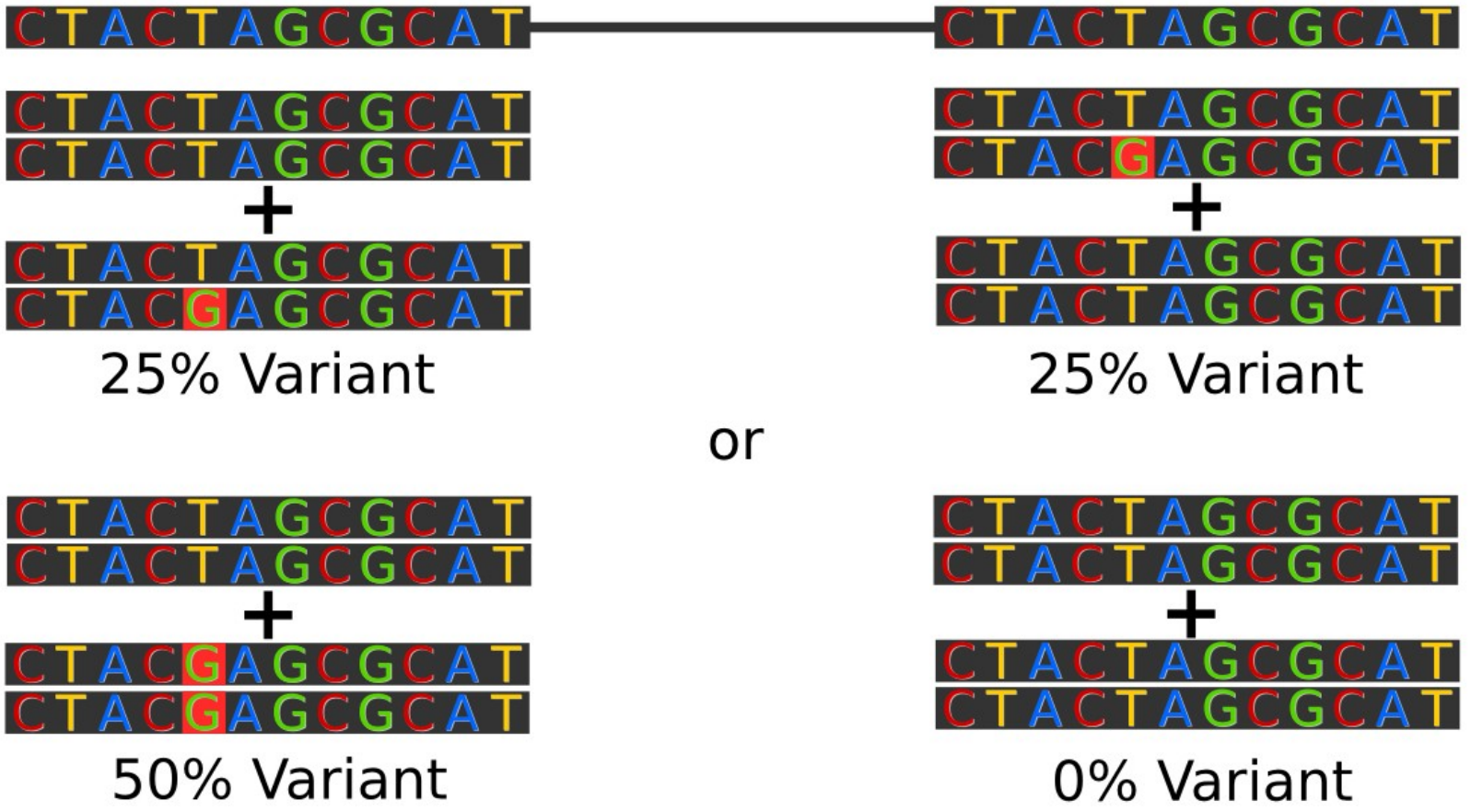
Alignment problems

- 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

Reference based alignment

- Popular aligners
  - Bowtie 1 + 2 (http://bowtie-bio.sourceforge.net/)
  - BWA (http://bio-bwa.sourceforge.net/)
  - BLAST (http://blast.ncbi.nlm.nih.gov/)
- Different strengths for each
  - Read length
  - Paired end
  - Indels

A survey of sequence alignment algorithms for next-generation sequencing. Heng Li & Nils Homer, 2010

Assembly vs. Alignment

- Hybrid methods
  - Assemble contigs that are aligned back against the reference, many popular aligners can be used for this

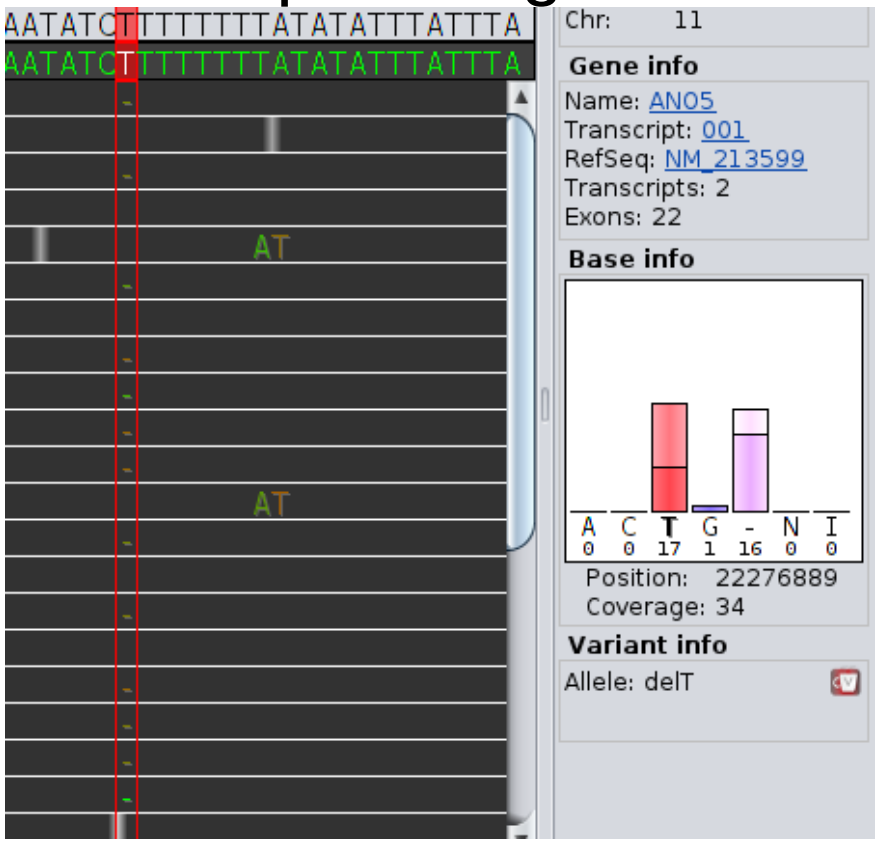
  - Reference aided assembly

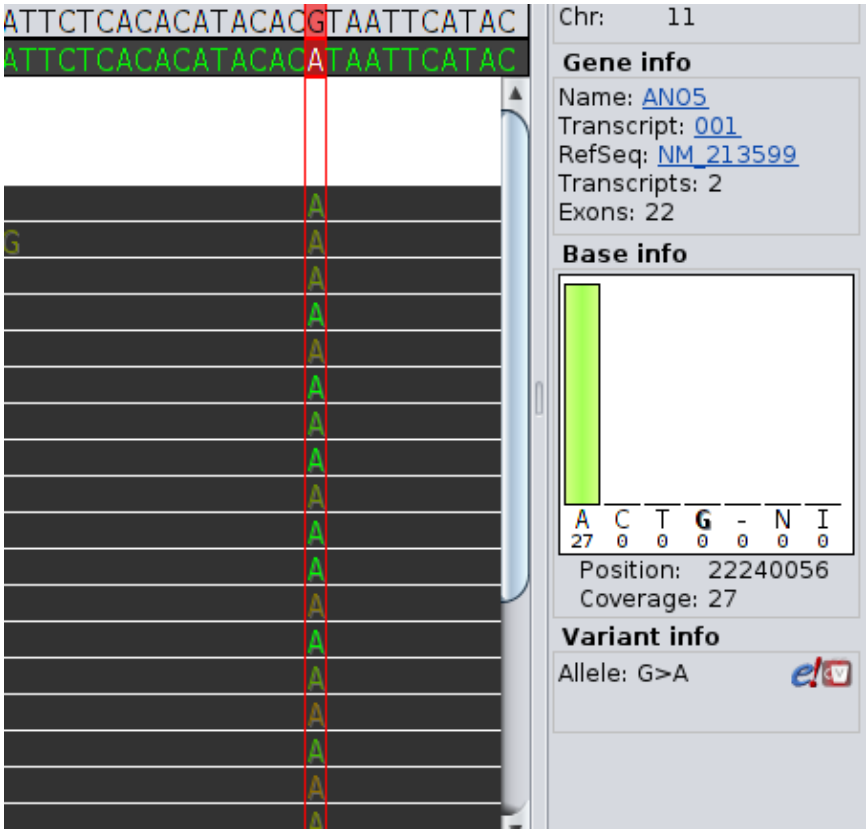
Variant calling

- Difference in underlying data (alignment vs assembly) require different strategies for variant calling
  - Reference based variant calling
  - Patient comparison of de novo assembly
- Hybrid methods exist to combine both approaches
  - Alignment of contigs against reference
  - Local de novo re-assembly

Variant calling

- Reference based variant calling
  - Compare aligned reads with reference





Variant calling

- Common reference based variant callers:
  - GATK
  - Samtools
  - FreeBayes

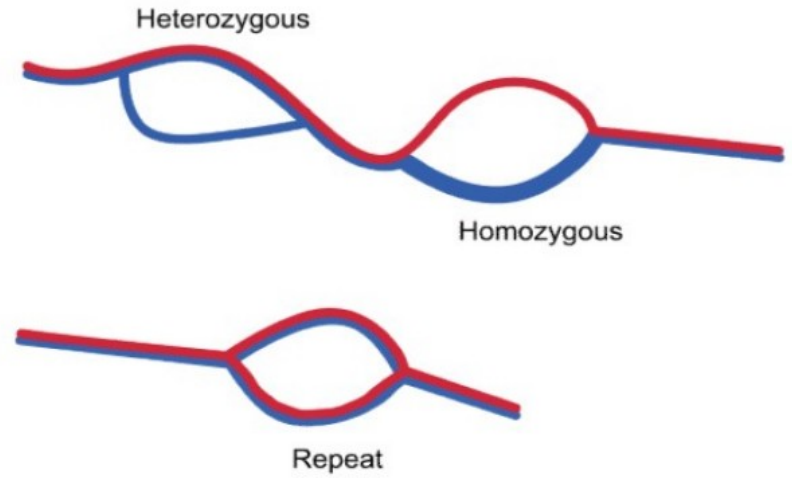
- Works very well for (in non repeat regions):
  - SNVs
  - Small indels

Variant calling

- De novo assembly
  - Either compare two patients
    - Useful for large structural variation detection
    - Can not be used to annotate variations with public databases
  - Or realign contigs against reference
    - Useful to annotate variants
    - Might loose information for the unaligned contigs

Variant calling

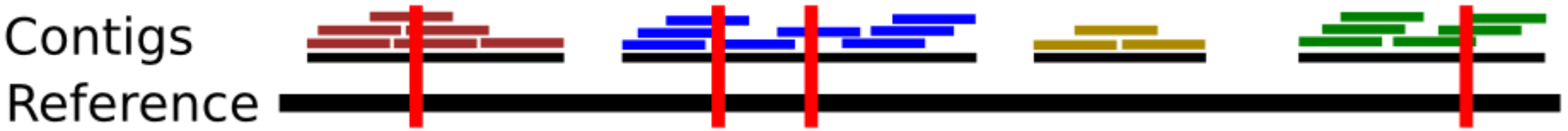
- Cortex
  - Colored de Bruijn graph based variant calling



- Works well for
  - Structural variations detection

Variant calling

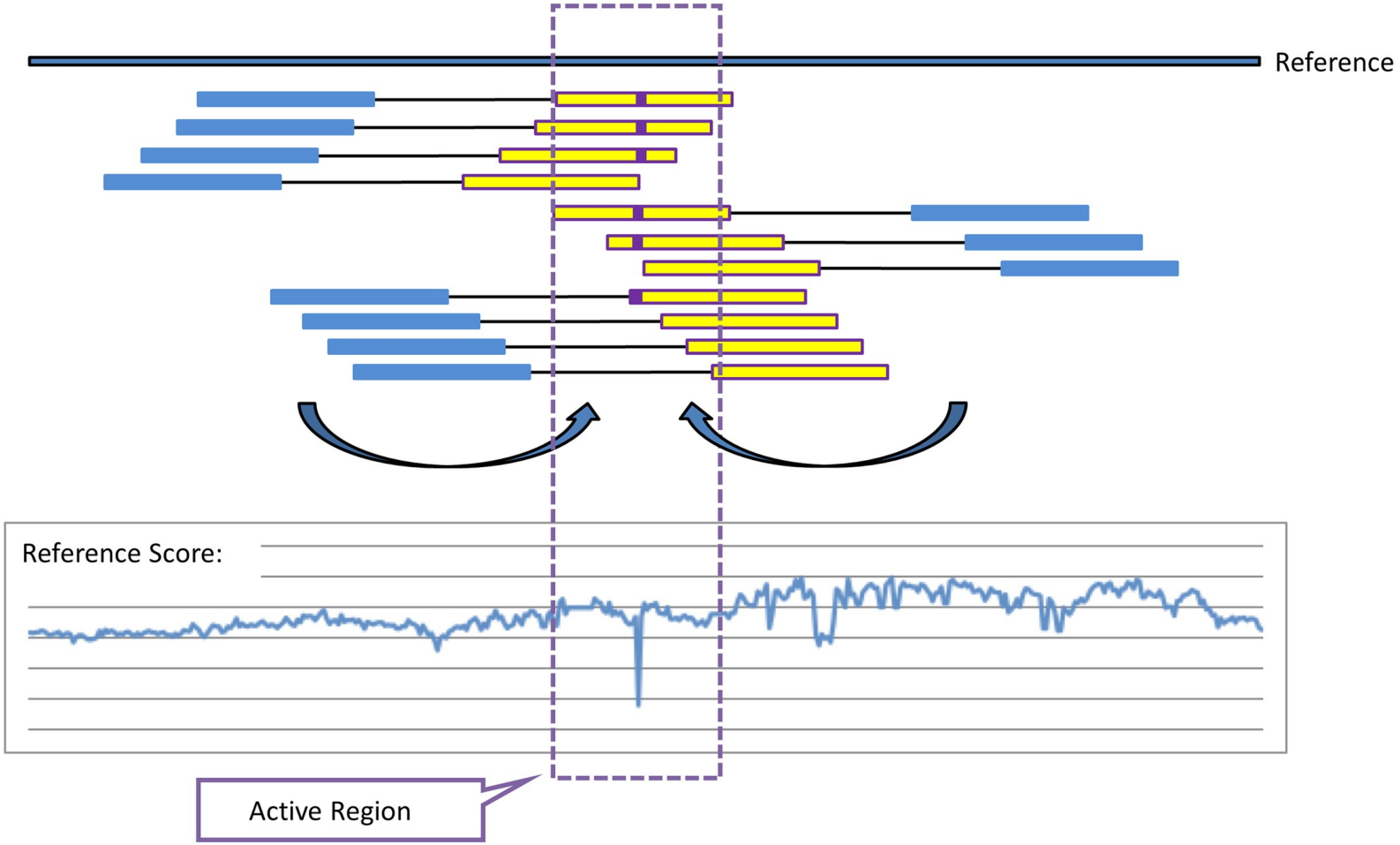
- Contig alignment against reference
  - Using aligners such as BWA
  - Uses standard reference alignment tools for variant detection
  - Helpful to „increase read size“ for better alignment
  - Variant detection is done using standard variant calling tools



Variant calling

- Local de novo assembly
  - Used by the Complete Genomics variant caller
- Every read around a variant is de novo assembled
- Contig is realigned back against the reference
- Final variant calling is done

Variant calling



Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads, Paolo Carnevali et al., 2012

Variant calling

- Local de novo realignment allows for bigger features to be found than with traditional reference based variant calling
- Faster than complete assembly

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

Other uses

- Transcriptomics, similar problematic to DNAseq
  - If small variations and gene expression analysis is done, alignment against reference is used
  - If unknown transcripts/genes are searched, de novo assembly is used
    - Used to detect transcripts with new introns, changed splice sites
    - Is able to handle RNA editing much better than alignment
    - Different underlying data (single strand, non uniform coverage, many small contigs)

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