Discovery of Structural Variants: Introduction

Alexander Schönhuth

UNIVERSITÄT BIELEFELD
Faculty of Technology

Bielefeld University

April 28, 2021

Organization

Organization

- Organization and introduction: *today*
- Full literature list available: *by April 30*
- How to present (brief): *May 5*
- How to write (brief): *May 12*
- **Presentations:** *from May 19:*
  - Each presentation 30-45 minutes
  - We can do two presentations per week, if this suits best
- **Technical Report:** *after presentation:*
  - Each report 8-15 pages
  - Optimally, report profits from feedback provided after presentation
  - Drafts can be submitted for discussion
  - Improving drafts based on feedback

Genetic Variants

Genetic Variants

Until 2006

Single nucleotide polymorphisms (SNPs)

```
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT   Reference
CCCAGCACTTTGGGAGGTCAAGGTGGGGGGAGGAAATAGCTTAAGCCCAGGAGT   New Genome
```

Genetic Variants

Until 2006

Single nucleotide polymorphisms (SNPs)

```
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT   Reference
CCCAGCACTTTGGGAGGTCAAGGTGGGGGGAGGAAATAGCTTAAGCCCAGGAGT   New Genome
```

From 2006

Structural Variants

Deletion

```
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT   Reference
CCCAGCACTTTGGGAGGCCAAGGTGGGAATTGCTTAAGCCCAGGAGT          New Genome
```

Insertion

```
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT   Reference
CCCAGCACTTTGGGAGTTATGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT   New Genome
```

Translocation

```
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT   Reference
CCCAGCACTTTGGGAGAGGTGGGGGGAGGAAATGCCATGCTTAAGCCCAGGAGT   New Genome
```

Further variations: inversions, duplications, ...

Somatic Variants

Somatic Variants

Cancer ↔ Control

Single Nucleotide Polymorphisms (SNPs)

CAGCATTGAAATA A GGCACAT C CGAA	Cancer Genome
CAGCATTGAAATA T GGCACAT C CGAA	Control Genome
CAGCATTGAAATA T GGCACAT G CGAA	Reference

Somatic SNP Germline SNP

Analysis of two samples necessary

Somatic Variants

Krebs ↔ Kontrolle

Deletions

CAGCATTGAAATA	----	GGCACAT	------	CGAA		Cancer Genome
CAGCATTGAAATA	TATA	GGCACAT	------	CGAA		Control Genome
CAGCATTGAAATA	TATA	GGCACAT	GCTGCT	CGAA		Reference
	Somatic Deletion		Germline Deletion			

Insertions

CAGCATTGAAATA	TATA	GGCACAT	GCTGCT	CGAA	Cancer Genome
CAGCATTGAAATA	----	GGCACAT	GCTGCT	CGAA	Control Genome
CAGCATTGAAATA	----	GGCACAT	------	CGAA	Reference
	Somatic Insertion		Germline Insertion		

How to Discover Genetic Variants?

Genetic Variants: Modes of Discovery

Re-Sequencing

- Sequence DNA of genome of interest
- Align resulting reads against reference genome
- Note down all differences

De Novo Assembly

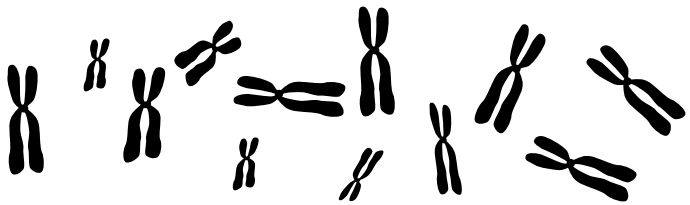
- Sequence DNA of genome of interest
- Connect resulting reads to form full-length genome
- Note down differences as per full-length comparison with reference genome

Somatic Variants

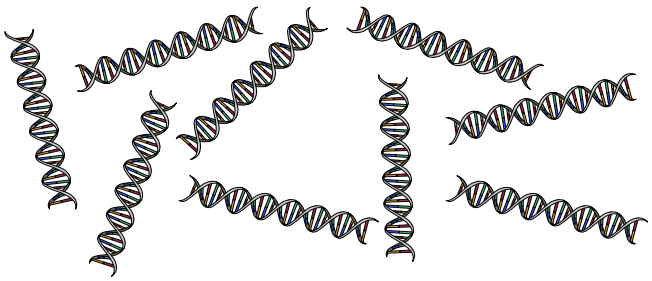
- Note down differences between cancer and control as well

Next Generation Sequencing

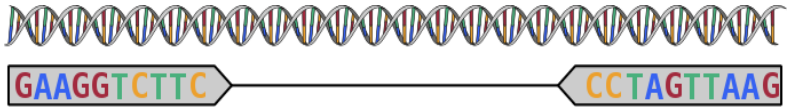
1. Extract Donor Genome DNA



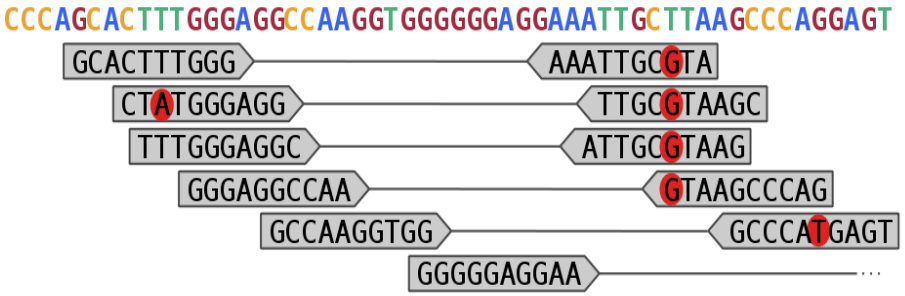
2. Break into fragments



3. Sequence fragments



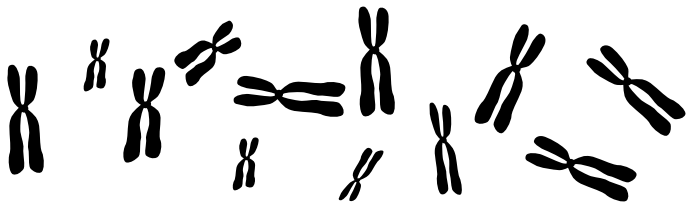
4. Map against reference genome



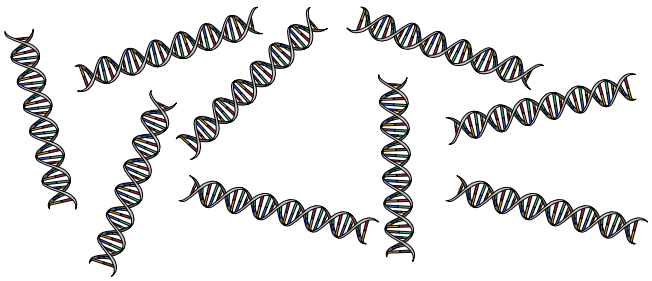
- For reference guided variant discovery, start from 4.
- For de novo assembly, start from 3.

Next Generation Sequencing

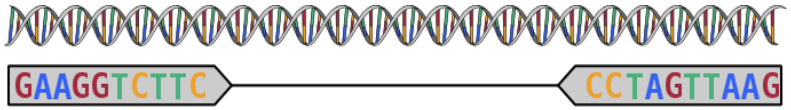
1. Extract Donor Genome DNA



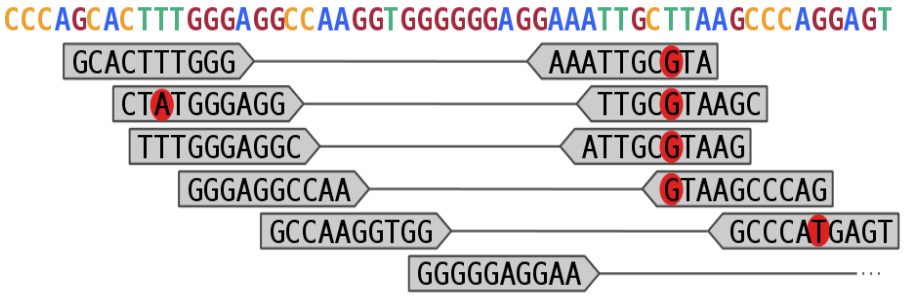
2. Break into fragments



3. Sequence fragments



4. Map against reference genome



- For reference guided variant discovery, start from 4.
- For de novo assembly, start from 3.

Re-Sequencing

Re-Sequencing: Variant Discovery

Evaluate signals emerging from aligned reads

SNP's and small indels

- Look at alignments of single reads with reference

Re-Sequencing: Variant Discovery

Evaluate signals emerging from aligned reads

SNP's and small indels

- Look at alignments of single reads with reference

Structural variants

- Variants may still yield signals in alignments directly
- Variants give rise to signals in paired-end alignments

Re-Sequencing: Variant Discovery

Evaluate signals emerging from aligned reads

SNP's and small indels

- Look at alignments of single reads with reference

Structural variants

- Variants may still yield signals in alignments directly
- Variants give rise to signals in paired-end alignments

Example: Read Pair Signals

Discovering Indels

Reference genome

CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Discovering Indels

Reference genome

CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Sequenced genome

CCCAGCACTTTGGGAGGCCAAAAATTGCTTAAGCCCAGGAGT

Discovering Indels

Reference genome

CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Sequenced genome

CCCAGCACTTTGGGAGGCCAAAAATTGCTTAAGCCCAGGAGT

Fragment

Discovering Indels

Reference genome

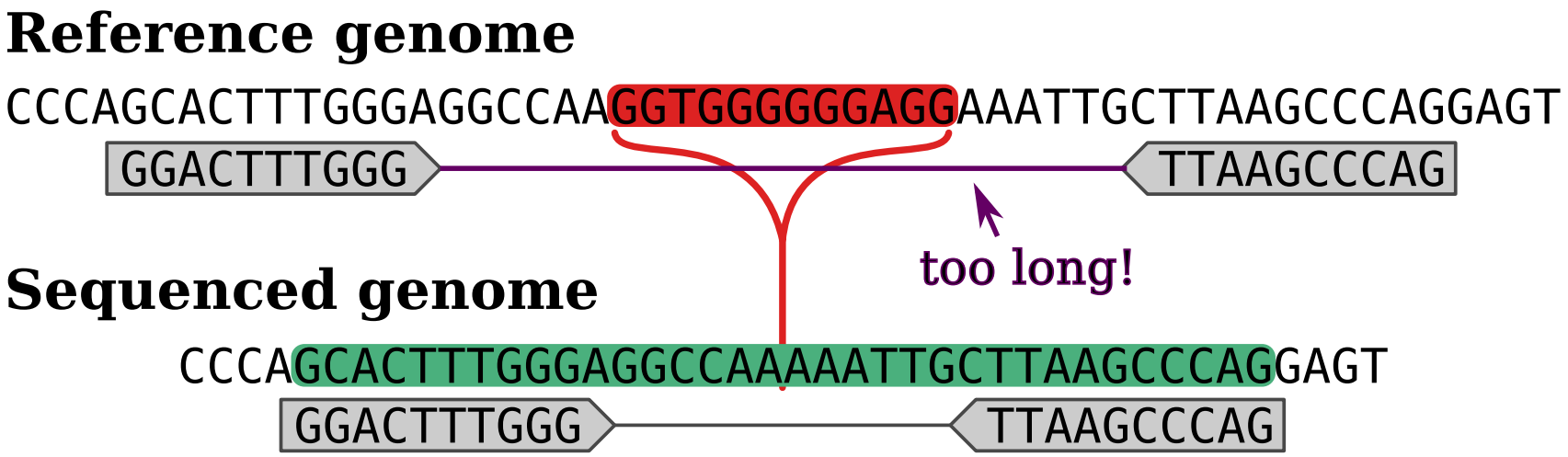
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Sequenced genome

CCCAGCACTTTGGGAGGCCAAAAATTGCTTAAGCCCAGGAGT

GGACTTTGGG TTAAGCCCAG

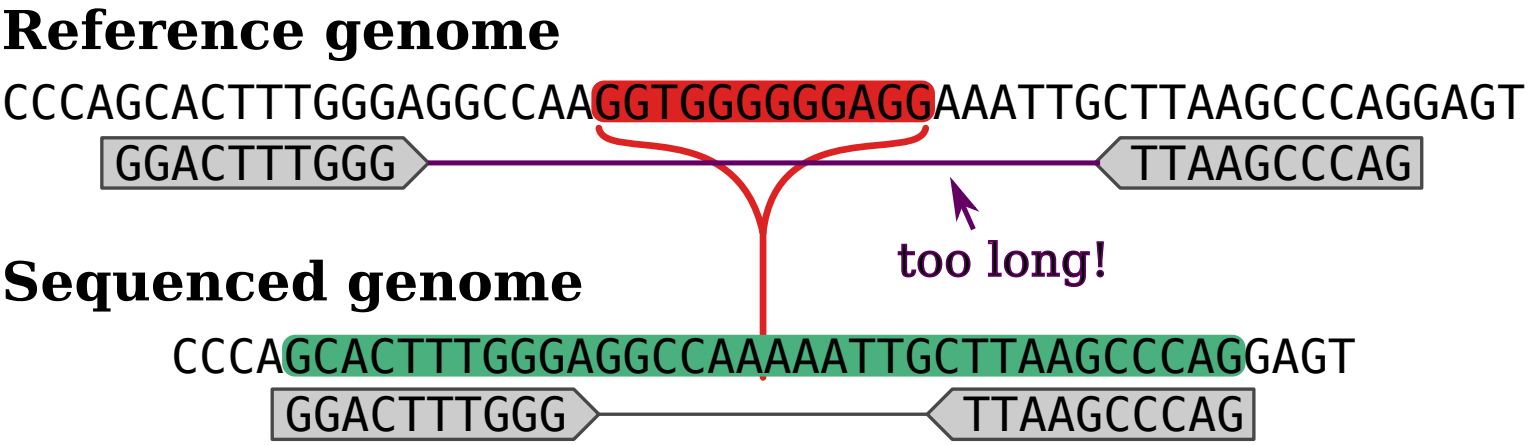
Discovering Indels



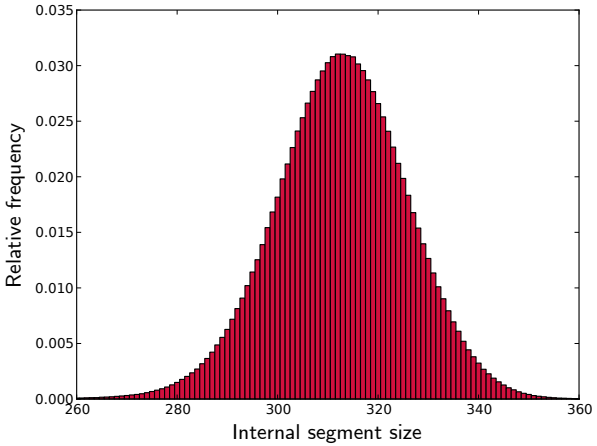
- Insertions: alignment length too short

Statistical Quantification

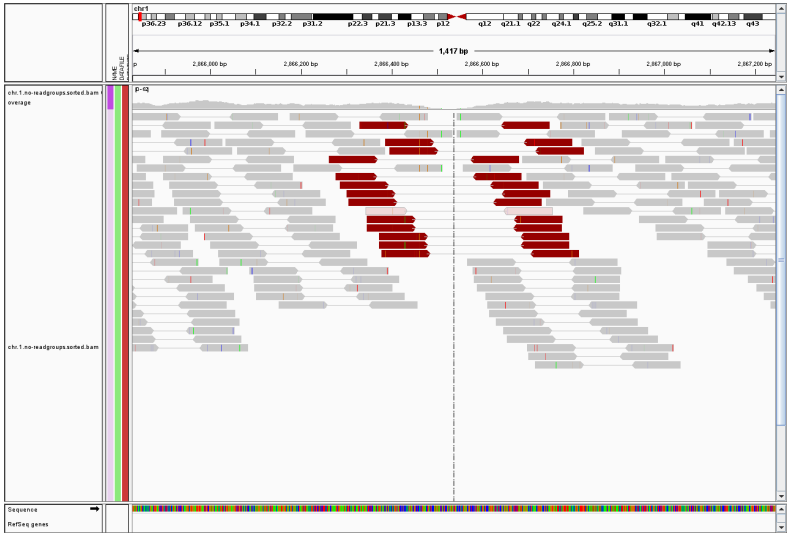
Fragment Length Distribution



- fragment length $\sim \mathcal{N}_{\mu,\sigma}$
- alignment length L: the greater $|L - \mu|$, the more significant



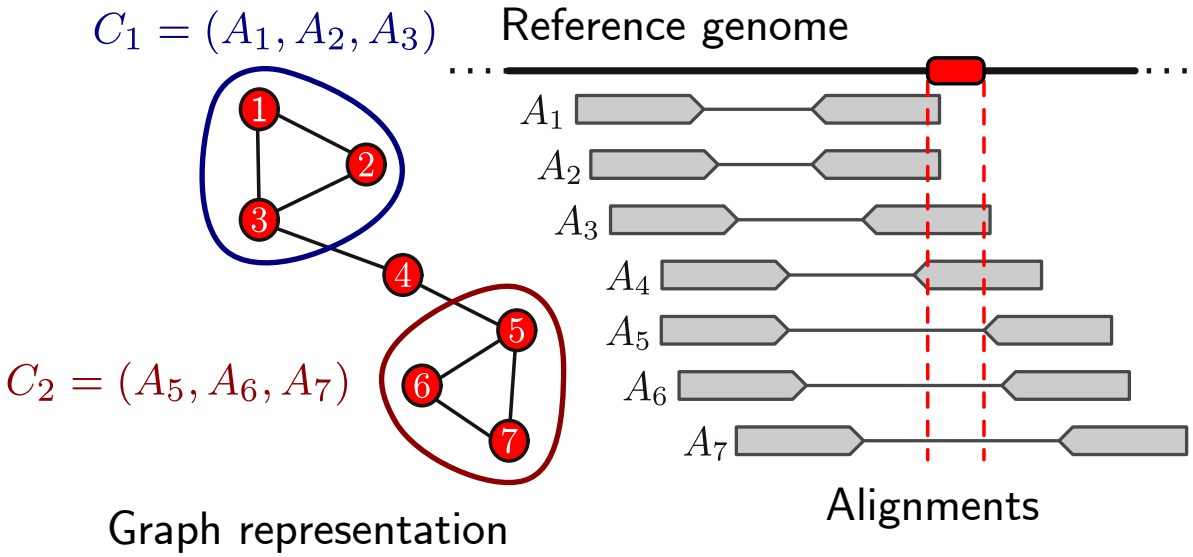
Discovering Twilight Zone Indels



IGV snapshot
Red alignment group: deletion

- Statistical significance increases
  - with increasing numbers of supportive alignments
  - with length of deletion
- Twilight Zone: Deletions of length 30 - 150 bp
  - $<$ 30 bp: discovered directly by alignment
  - $>$ 150 bp: single alignment significant

Read Alignment Graph

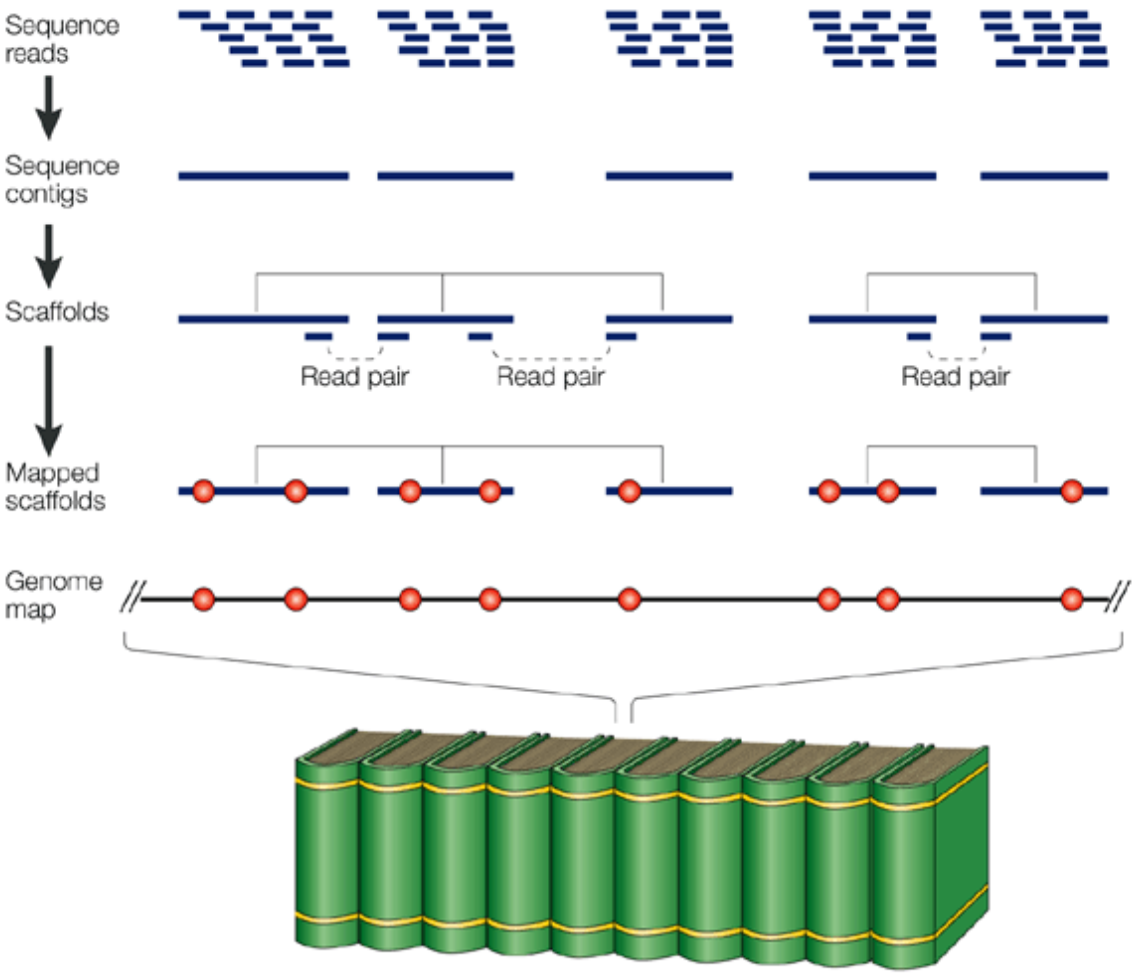


- Graph theoretical modeling:
  - Nodes: Paired-end alignments
  - Edges: Overlapping, length compatible alignments
  - Wanted: Groups of nodes sharing many edges

Genome Assembly

Genome Assembly

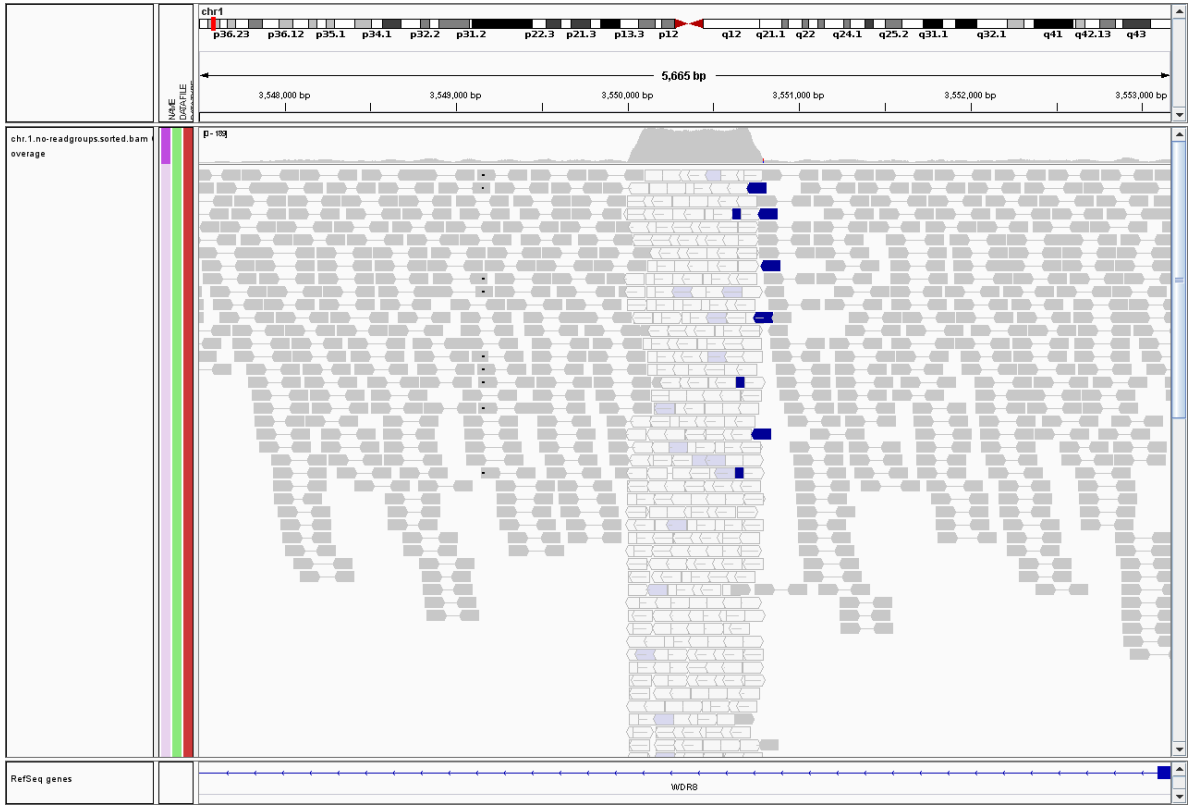
Step by Step



Nature Reviews | Genetics

From Contigs To Scaffolds: Why?

Repetitive Sequence



- Repetitive areas disturb the problem decisively
- Make contigs from reads exhibiting unique sequence
- Make scaffold of contigs, to bridge repetitive sequence

Generating Contigs

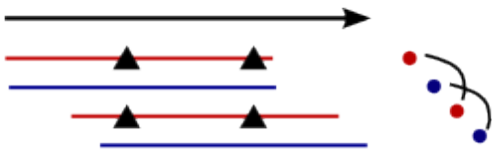
- De novo assembly programs usually deliver contigs
- Scaffolding requires data from additional, unconventional sources
- We will focus on contig generation in the following

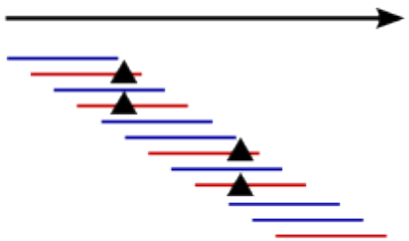
Generating Contigs

- De novo assembly programs usually deliver contigs
- Scaffolding requires data from additional, unconventional sources
- We will focus on contig generation in the following

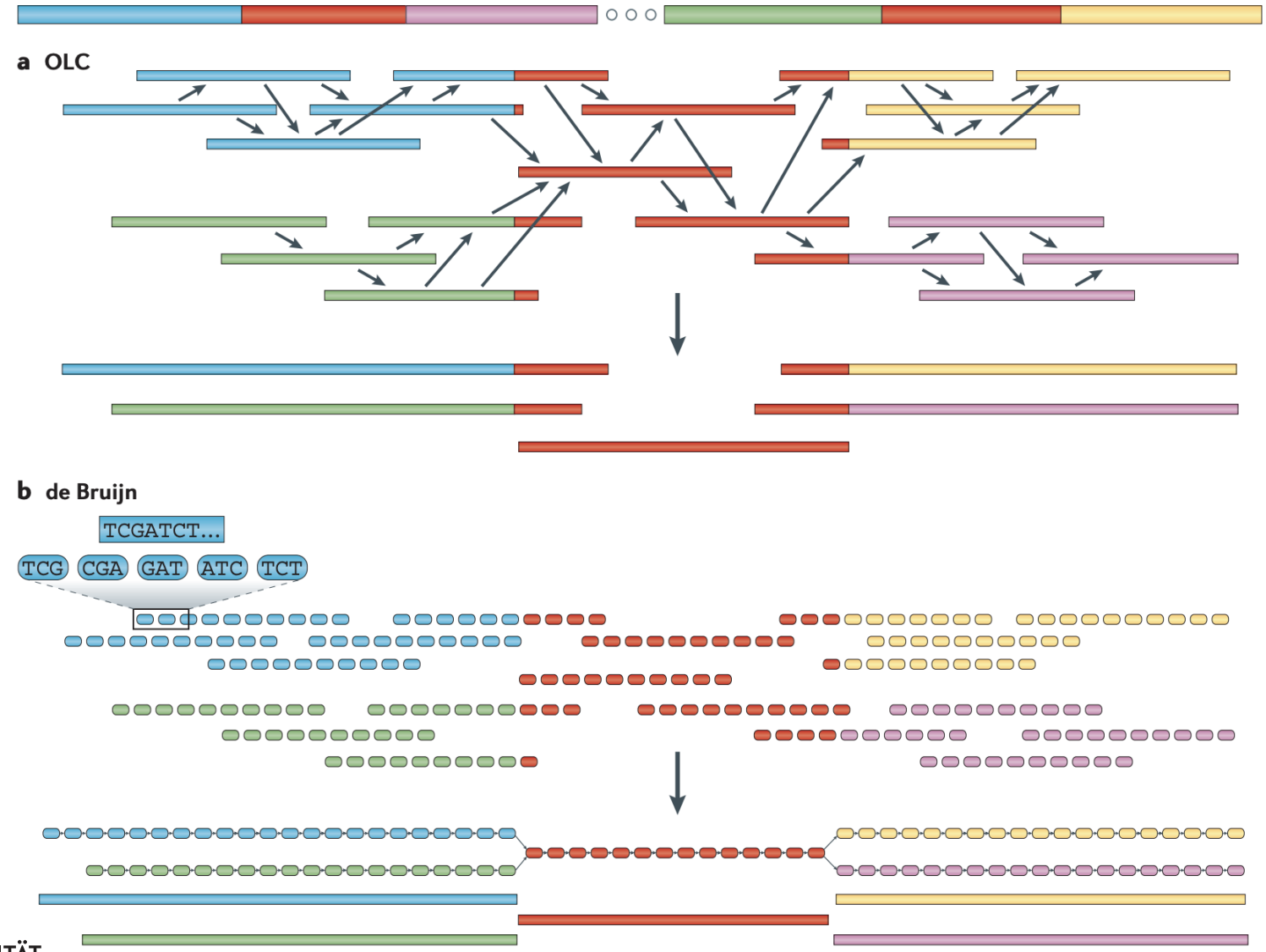
Generating Contigs: Assembly Paradigms

	Overlap graphs	De Bruijn graphs
nodes:	sequencing reads	length *k* substrings
edges:	approximate suffix-prefix overlaps	exact overlaps of length *k-1*





Assembly Paradigms



Genome Assembly Paradigms

Overlap-Layout-Consensus Paradigm

- Natural approach: identify overlapping fragments
- Seminal paper: [Kececioglu, Myers, 1995]
- In use at Celera for human genome assembly

Genome Assembly Paradigms

Overlap-Layout-Consensus Paradigm

- Natural approach: identify overlapping fragments
- Seminal paper: [Kececioglu, Myers, 1995]
- In use at Celera for human genome assembly

De Bruijn Graph Paradigm

- Construct *de Bruijn graph* from sequence fragments
- *"Cut pieces into even smaller pieces to solve the puzzle"*
- Seminal paper: [Pevzner, Tang, Waterman, PNAS 2001]
- Predominant paradigm for NGS based genome assembly

De Bruijn Graphs

De Bruijn Graphs

Genome: ACGATGCTCAGACCCCCCCCC

Short reads: ACGATGCTCAGA CTCAGACCC AGACCCC CCCCCCC

k-mers:

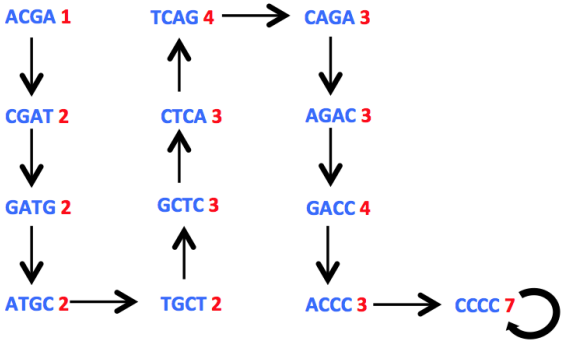
ACGAT
CGATG
GATGC
ATGCT
TGCTC
GCTCA
CTCAG
TCAGA

CTCAG
TCAGA
CAGAC
AGACC
GACCC

AGACC
GACCC
ACCCC

CCCCC
CCCCC
CCCCC

De Bruijn graph:



Assembled Contigs: ACGATGCTCAGACCCC

Advantages:

- Removes redundancy among reads
- Easy to construct and store

[Lu, Shen, Warren and Walter, 2016]

De Bruijn Graphs

Genome: ACGATGCTCAGACCCCCCCCC

Short reads: ACGATGCTCAGA CTCAGACCC AGACCCC CCCCCCC

k-mers:

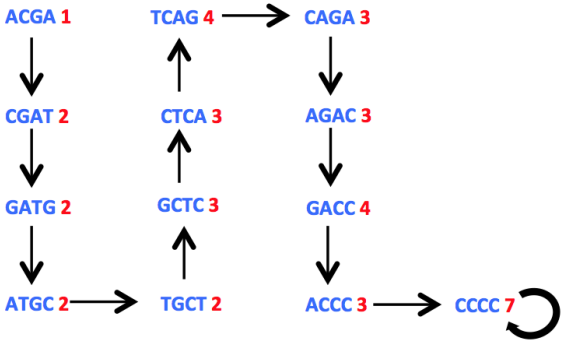
ACGAT
CGATG
GATGC
ATGCT
TGCTC
GCTCA
CTCAG
TCAGA

CTCAG
TCAGA
CAGAC
AGACC
GACCC

AGACC
GACCC
ACCCC

CCCCC
CCCCC
CCCCC

De Bruijn graph:



Assembled Contigs: ACGATGCTCAGACCCC

Advantages:

- Removes redundancy among reads
- Easy to construct and store

Issues:

- looses information about linked mutations
- works only well with error-corrected reads

[Lu, Shen, Warren and Walter, 2016]

De Bruijn Graphs

Summary

- Superior for NGS based consensus genome assembly
- For polyploid genome assembly: *colored de Bruijn graphs* (not treated here, but seem to have certain limitations)
- (Likely severe) limitations for polyploid genome assembly:
  - Error correction removes lowly frequent true mutations
  - Loose information about linked mutations

Overlap-Layout-Consensus

Overlap-Layout-Consensus (OLC)

- **Overlap**: Construct overlap graph from sequencing reads
- **Layout**: Compute *contigs*, longer stretches of overlapping reads using the overlap graph
- **Consensus**: Compute consensus sequence for contigs

OLC and Polyploid Genomes

Summary

Advantages

- Edges ☞ overlapping reads from the same haplotype
- Error correction *after graph construction*

Disadvantages

- Overlap graph construction time and space consuming
- Error correction: new ideas required

Overlap Graph Construction

Overlap Graph Construction

Without reference genome

For each pair of reads, determine

- how they optimally overlap
- whether the resulting overlap indicates that the two reads are from the same haplotype (statistically significantly likely!)

Naive approaches infeasible!

Literature

Literature References

Reviews

- *Structural Variant Discovery:* "Genome structural variation discovery and genotyping", by Alkan et al., Nature Reviews Genetics, 2011
- *De Novo Assembly:* "Genetic variation and the de novo assembly of human genomes", Chaisson et al., Nature Reviews Genetics, 2015
- *Long-range sequencing:* "Piercing the dark matter: bioinformatics of long-range sequencing and mapping", Sedlazeck, et al., Nature Reviews Genetics, 2018
- *Somatic Structural Variant Discovery:* "Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology", van Belzen et al., Nature Precision Oncology, 2021

Literature References

Papers

- *Varlociraptor:* Koester et al., Genome Biology, 2020 (somatic variants)
- *Delly:* Rausch et al., Bioinformatics, 2012 (structural variants, also somatic)
- *FreeBayes:* Garrison and Marth, arXiv:1207.3907, 2012 (haplotype-aware small variants)
- *Hifiasm:* Cheng et al., Nature Methods, 2021 (de novo assembly of "HiFi" reads)
- *MiniMap2:* Heng Li, Bioinformatics, 2018 (overlap graph construction)
- *Lancet:* Narzisi et al., Communications Biology, 2018 (somatic variants)
- *GRIDSS:* Cameron et al., Genome Research, 2017 (rearrangement type variants)
- *Sniffles:* Sedlazeck et al., Nature Methods, 2018 (structural variants from long reads)
- *NanoVar:* Tham et al., Genome Biology, 2020 (structural variants from long reads)