HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Assembly of MHC with hybrid data

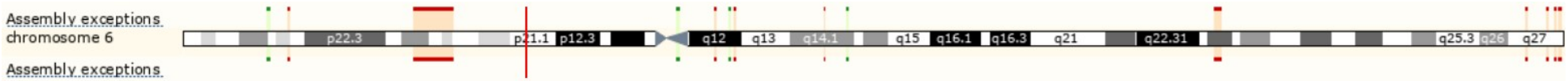
Dr. Torsten Houwaart

group of Dr. Alexander Dilthey
joint project with Prof. Dr. Birgit Henrich

Group Seminar
22.05.2019

MHC (Major Histocompatibility Complex)

Human genome: chromosome 6



ensembl.org

Length: 171 Mbp
MHC locus: ~ 28.7 – 33.4 Mbp

Goal

=> Get a better assembly of the haplotype of the MHC (Major Histocompatibility Complex) region for different cell lines

cell line	**origin**
QBL	dutch, blood (EBV-transformed lymphoblastoid)
SSTO	amish, blood
MANN/MOU	danish, blood
DBB	amish, blood
APD	not specified
MCF7	caucasian, breast adenocarcinoma
COX	south african, unknown
PGF	english, blood

Goal

=> Get a better assembly of the haplotype of the MHC (Major Histocompatibility Complex) region for different cell lines

cell line

QBL

SSTO

MANN/MOU

DBB

APD

MCF7

~~COX~~

~~PGF~~

Already assembled

Goal

=> Get a better assembly of the haplotype of the MHC (Major Histocompatibility Complex) region for different cell lines

cell line

QBL

SSTO

MANN/MOU

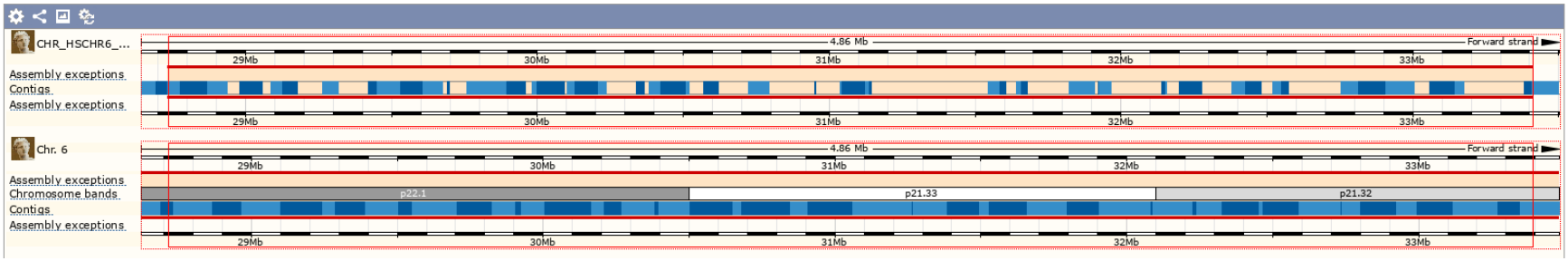
DBB

APD

MCF7

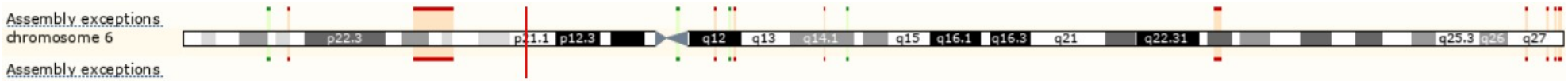
~~COX~~

~~PGF~~



Goal

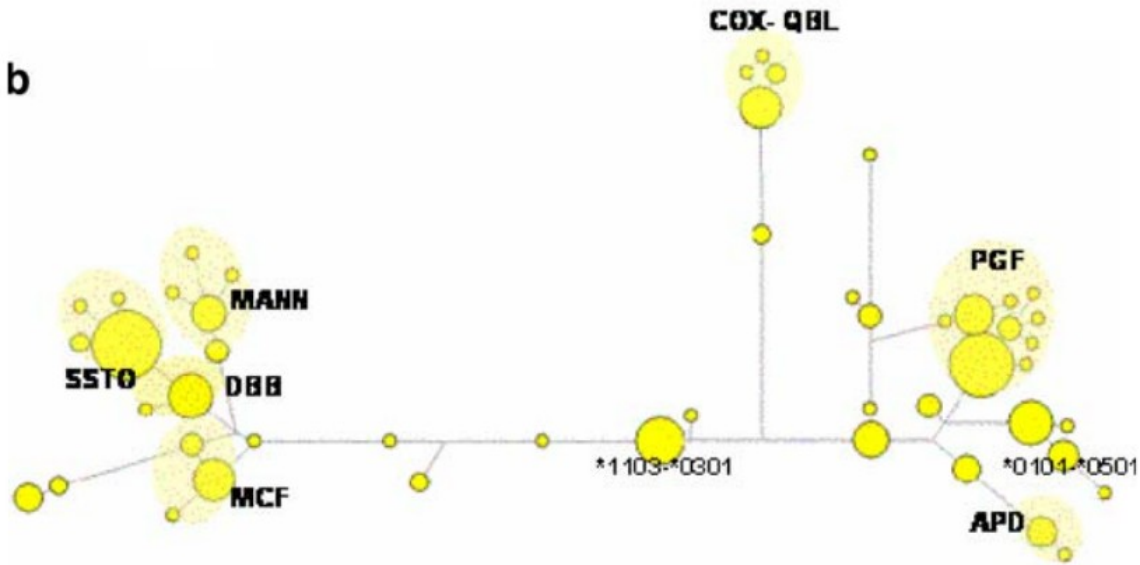
Human genome: chromosome 6



ensembl.org

Length: 171 Mbp
MHC locus: ~ 28.7 – 33.4 Mbp

"… were selected on the basis of conferring either protection against or susceptibility to two autoimmune diseases, type 1 diabetes and multiple sclerosis, and that represented common haplotypes in European populations"

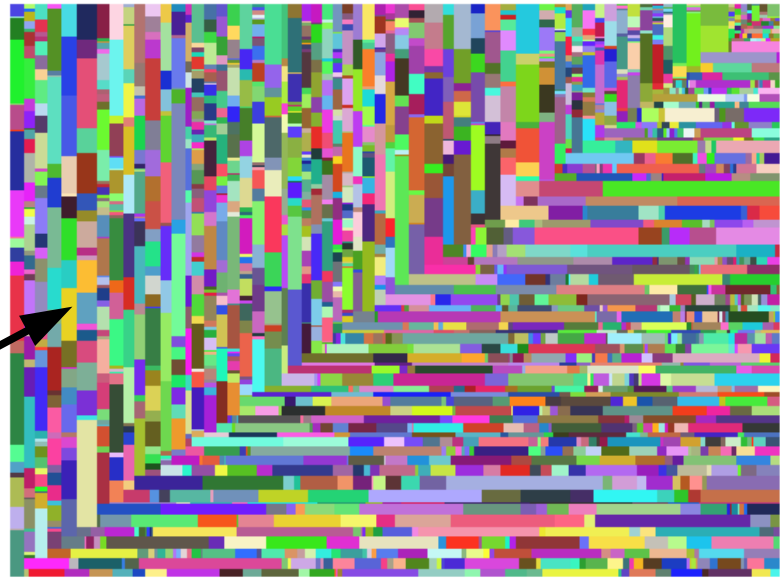


2008 "Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project" doi: 10.1007/s00251-007-0262-2

Short Read data

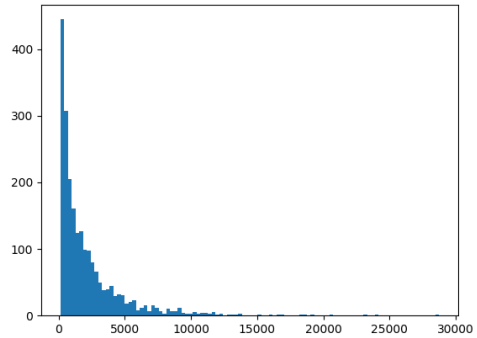
contig 1 ATGGAATCGTGTTG**CTCTCTCTCTCTCTCTCTCT**TAGGTCGCTCCAGTAG **contig 2**

The contig length distribution of the APD cell line



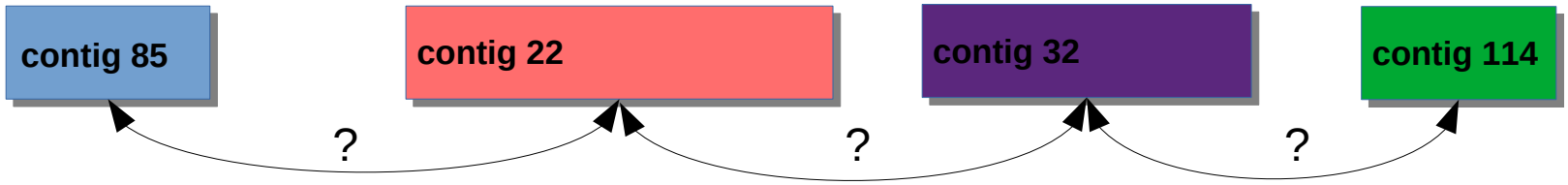
each rectangle
depicts size of one contig

APD
2219 contigs



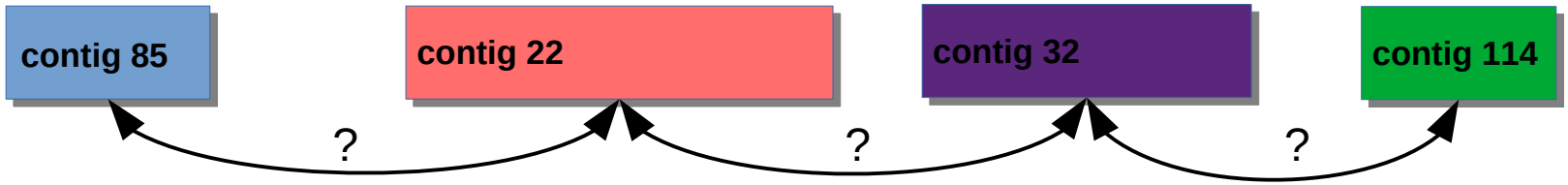
base pairs

Contig Arrangement Problem



- How are the contigs arranged (direction)?

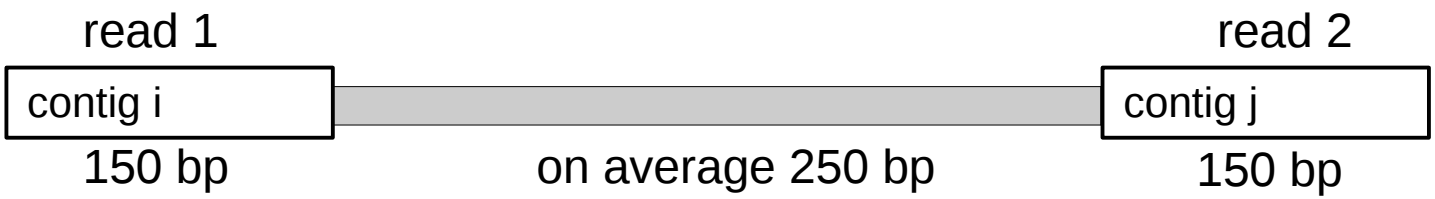
Contig Arrangement Problem

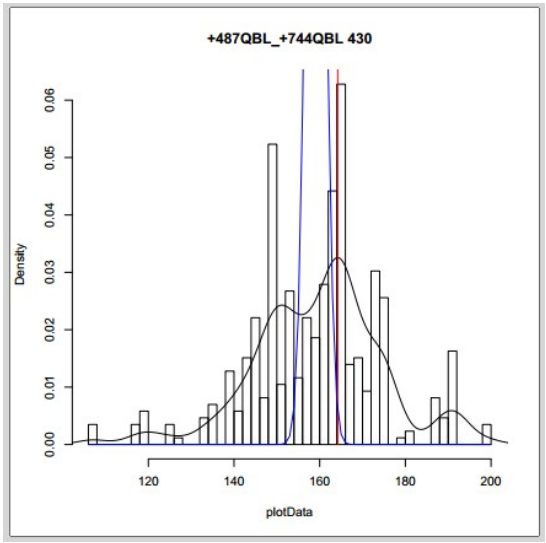


- How are the contigs arranged (direction)?

- Distances between the contigs?
- What is the sequence between the contigs?

Short Read data



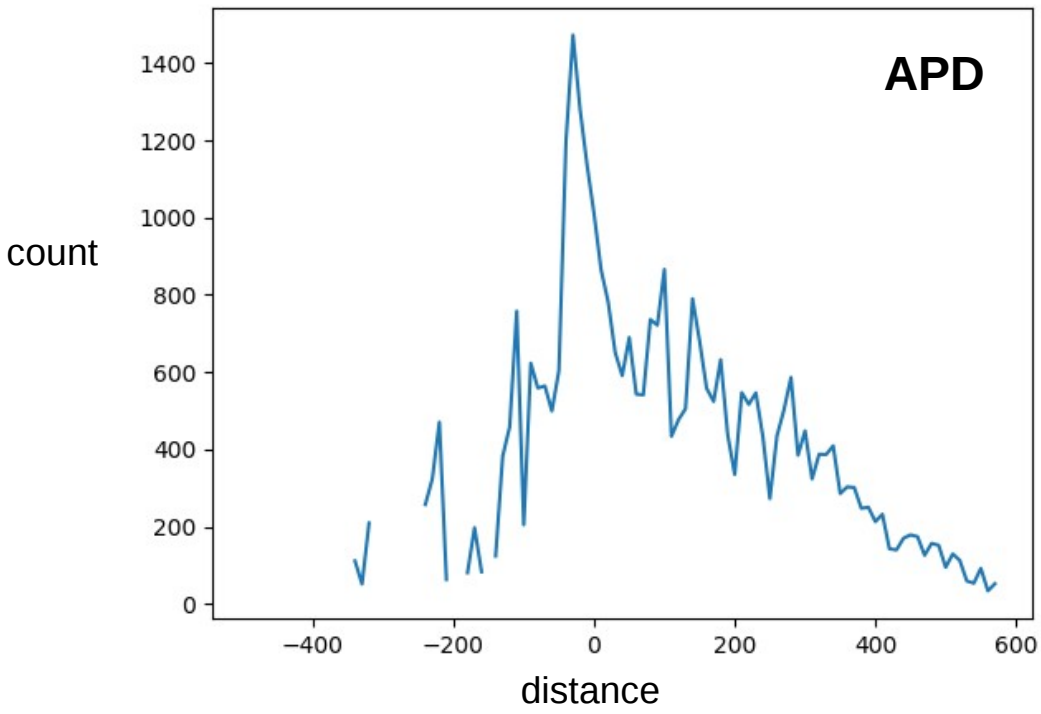


→ distance between 487QBL and 744QBL :

162 – 166 bp

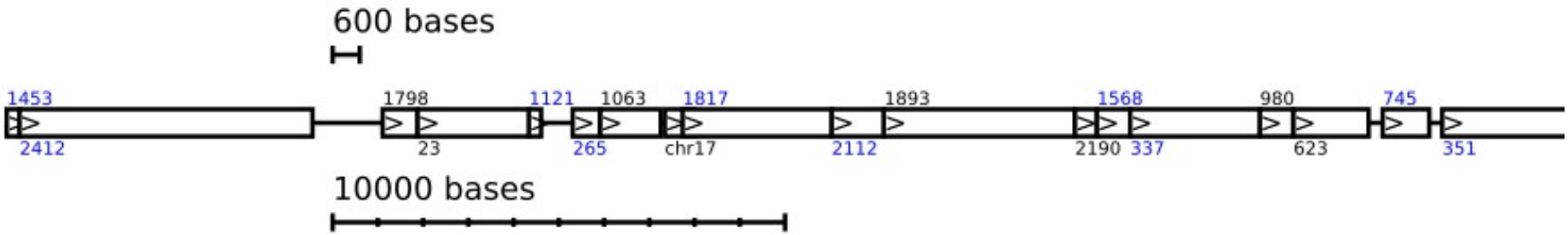
distances between all pairs of contigs summary

→ up to 600 bp



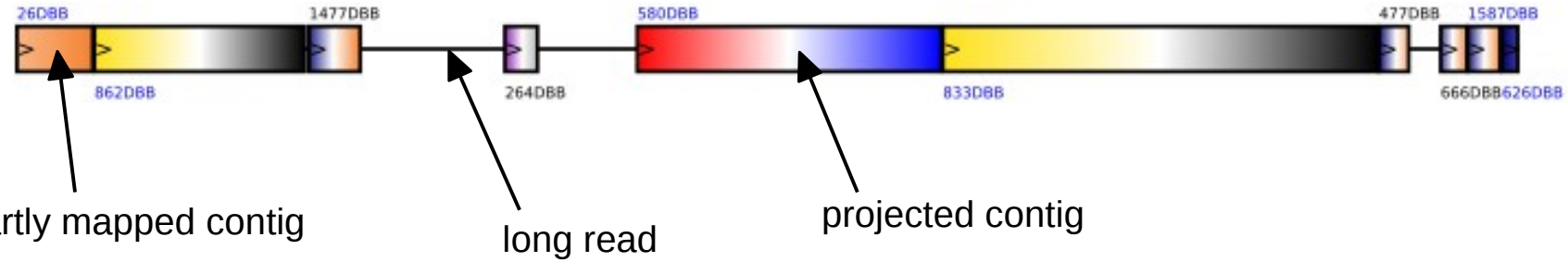
Long Read data

Data from 3rd generation sequencing **Nanopore**
→ - high error rate (15%), - expensive, + long



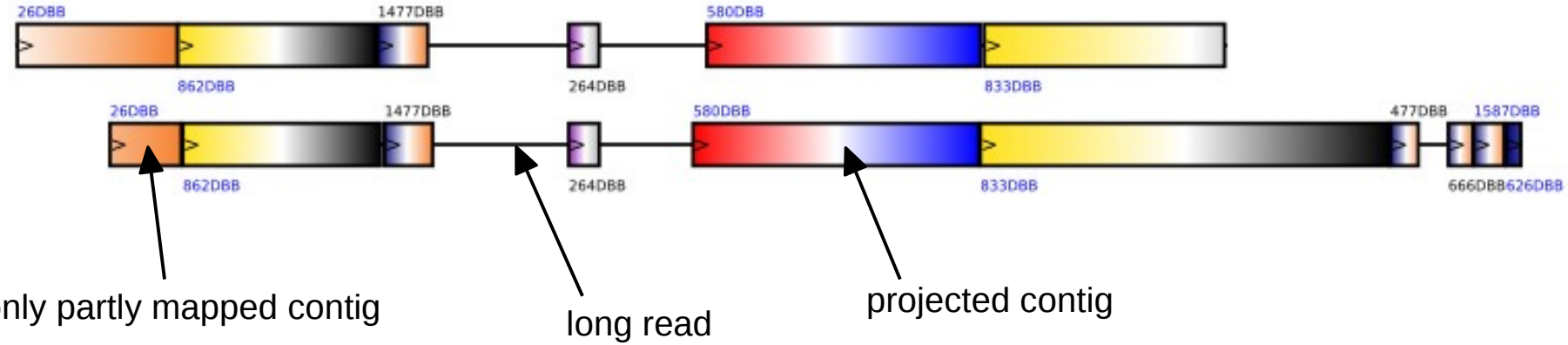
The Idea

long read – contig signature



The Idea – Aligning Signatures

long read – contig signatures fit together



Code Improvements

Modularized Scaffolds

Pseudolongreads
- consist of pieced together real longreads

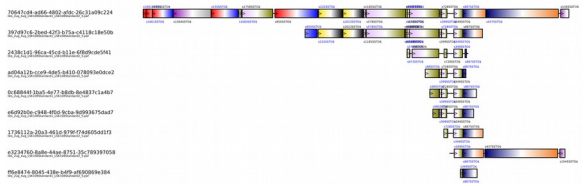
Pseudoalignments
- diminishes problem of paralogs greatly
- automatize assembly (in principle)

Colors

Modularized Scaffolds

less_naive_scaffolding

automatized assembly



alignment2SVG

clusters reads and draws them

class Longreads

- initialize with *.paf, *.erate fileS(!)
  - track origin of each longread
- alternative constructor from dictionary
  - allows merging of longreads
- pseudoalign_all
  - N^2 alignments between all N longreads

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Automatic Assembly

- always yields a result
- but problems can arise
  caused by “wrong” longreads: dsDNA, unexplainable long/short regions
  regions with high variability and/or low coverage

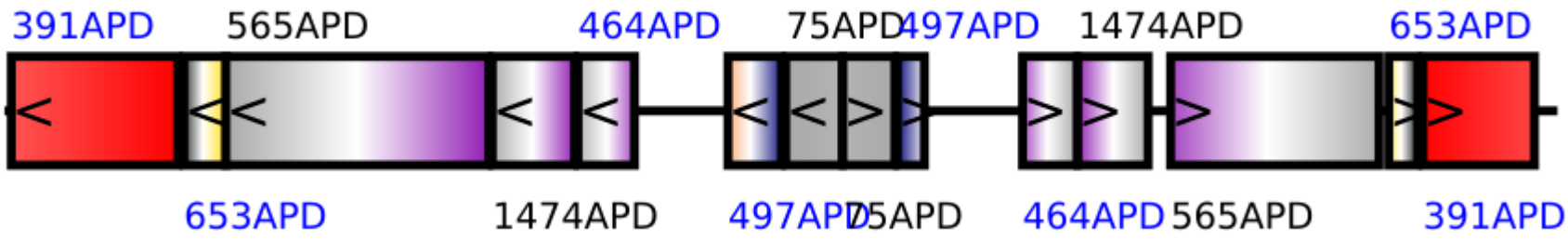
Semi-Automatic Curated Assembly

- takes a little longer
- guarantees there are no long range mistakes

Problematic Reads

“Fake Hairpins” - dsDNA that is seperated at the pore and sequenced in the same read?

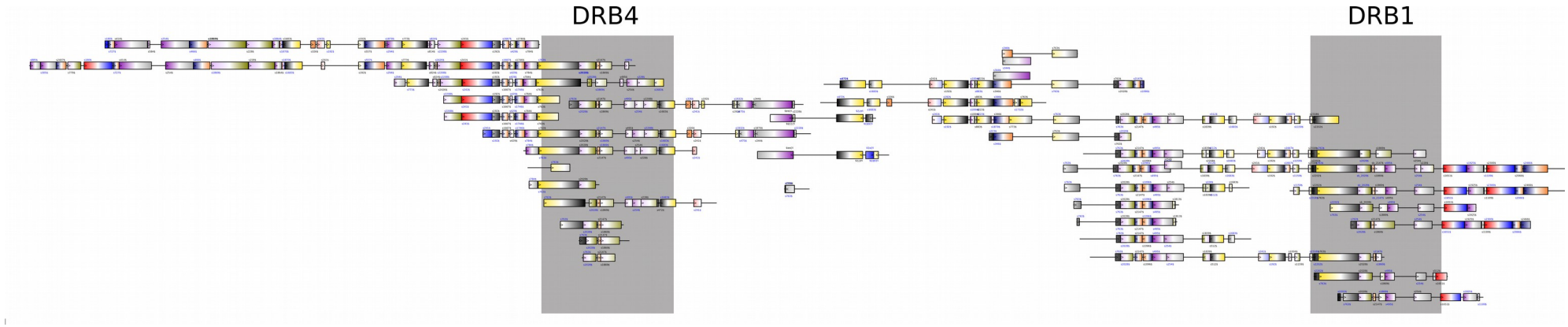
~1% of data



DRB region of cell line DBB

with genes DRB4 and DRB1
- paraloges
→ very similar sequence
→ very similar contig structure

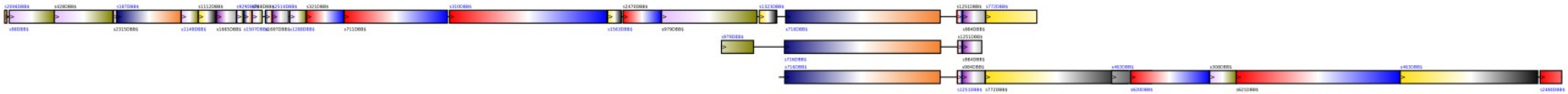
=> difficult to assemble, a strong case for manually curated assembly



HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Semi-Automatic Curated Assembly

1. Collect good representatives (good long reads) of a section of the MHC
2. Connect them via contigs



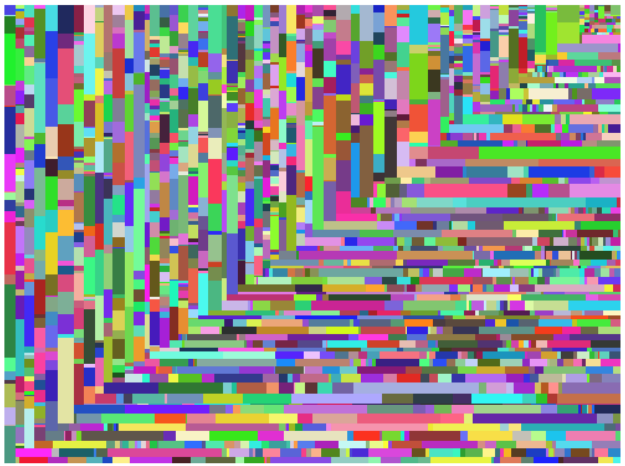
Assembly of MHC with hybrid data

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

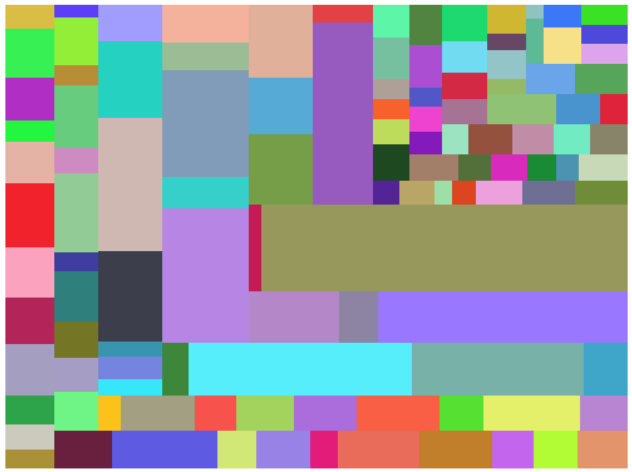
Semi-Automatic Curated Assembly

1. Collect good representatives of a section of the MHC
2. Connect them via contigs

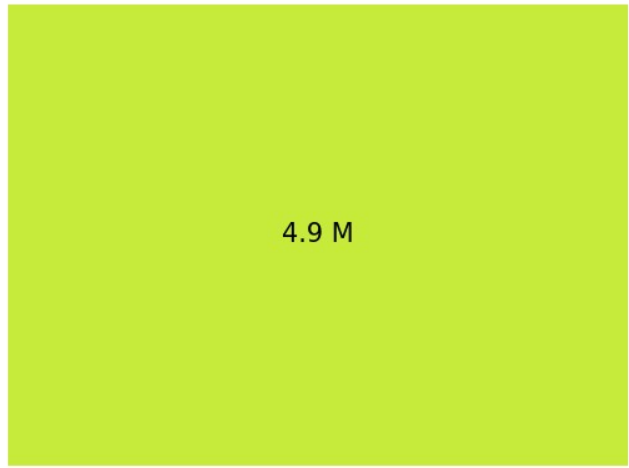
contigs



representative longreads



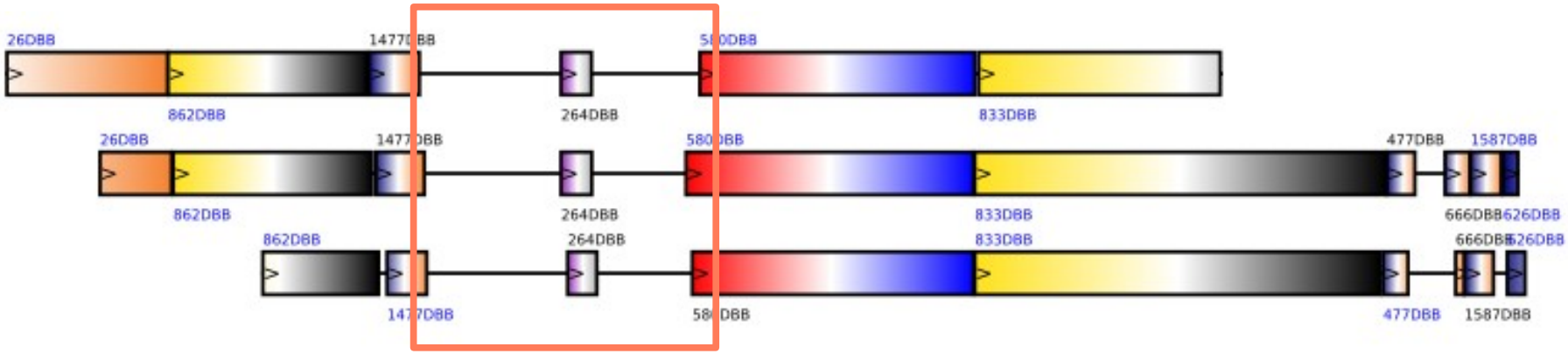
complete assembly



APD

DBB

Refinement / Polishing

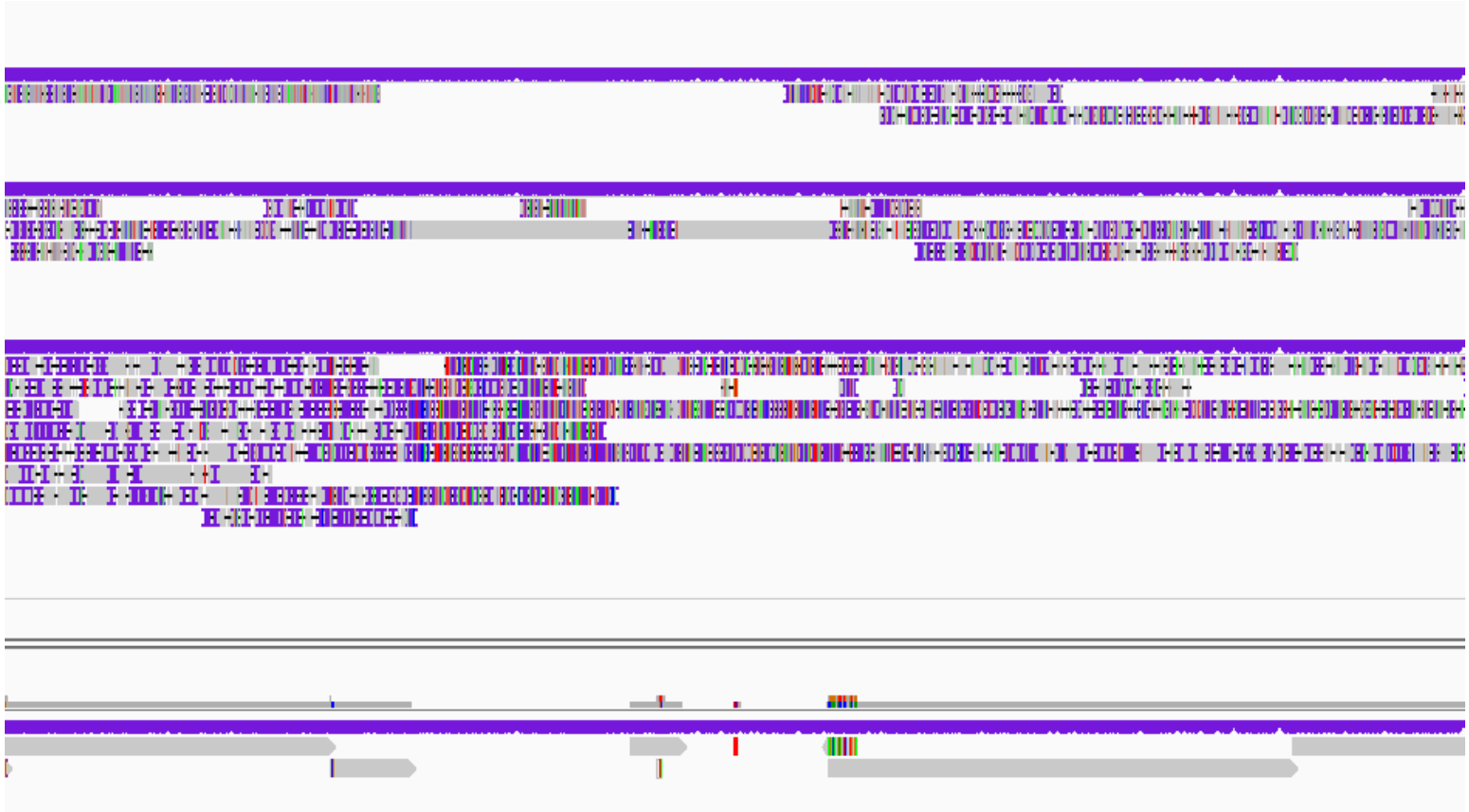


error rate of longreads, ~ 15%

Refinement / Polishing

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

data from three different flow cells



Refinement / Polishing

data from three different flow cells

→ the more colorful, the more alignment errors
→ where there are no errors: these are the representatives for the reference

Refinement / Polishing

Ref. built from contigs → less errors → sequencing errors
Ref. built from longread-representatives → more errors → seqeuncing errors, errors in the reference

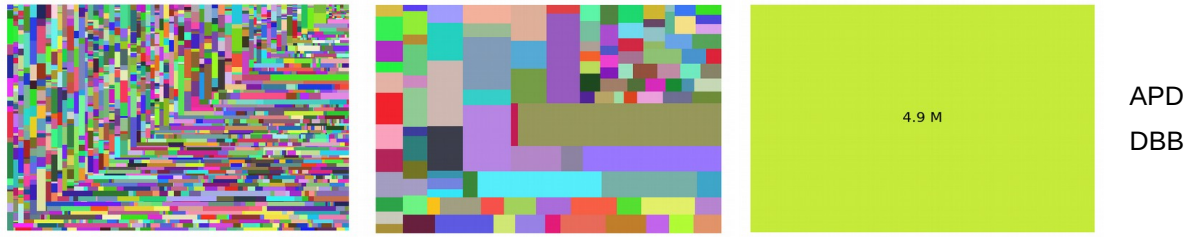
Refinement / Polishing

Taking a representative → Multiple sequence alignment ; Polish with Illumina WGS data

Nanopolish Seqan Samtools Pileup

Summary and Outlook

- Assembly with hybrid data successful



- Finish Missing Cell Lines
    - SSTO, longreads quite short makes assembly tedious, in progress
    - MANN, cells at BMFZ
    - MCF7, cell line with cooperation partner
    - QBL, does not grow, extracted DNA available

- Enhance Sequence Between Contigs
    1. multiple sequence alignment with Seqan
    2. nanopolish (default coverage: 20x, heuristics suitable for lower coverage?)
    3. samtools pileup consensus

- Graph Based Reference Project

Acknowledgments

UKD Universitätsklinikum Düsseldorf

Microbiology Institute

Dr. Alexander Dilthey

Alona Tyshaieva

Dana Belick

Prof. Dr. Birgit Henrich

all others ...

UC Denver

Dr. Paul Norman

University of Cambridge

Dr. James Traherne

BMFZ

Dr. Tobias Lautwein

HHU Algorithmische Bioinformatik

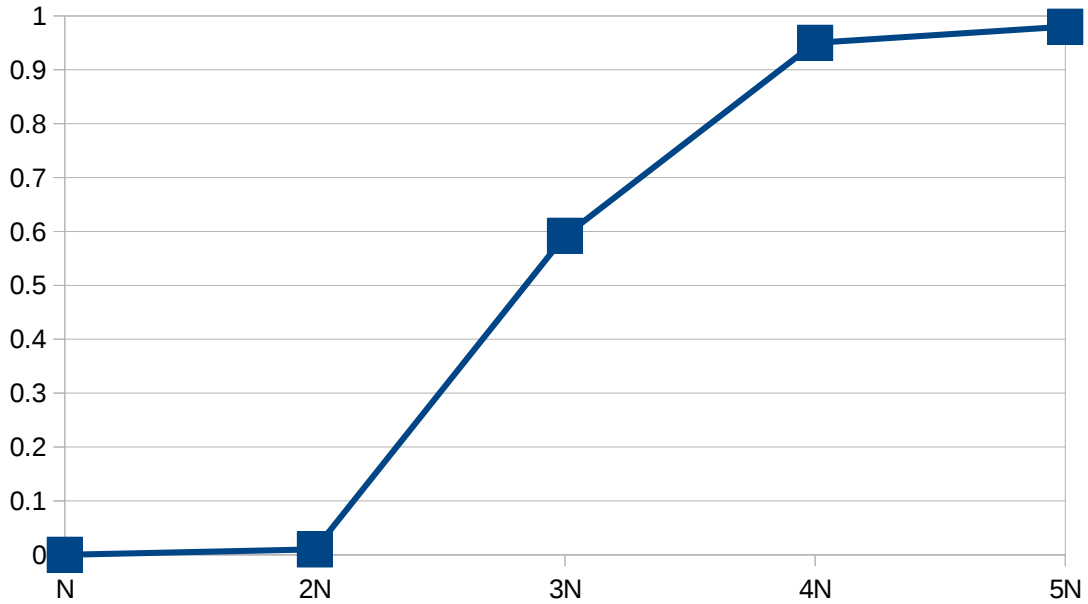
Prof. Dr. Gunnar Klau

Cost Estimation

simulate_read_distribution.py

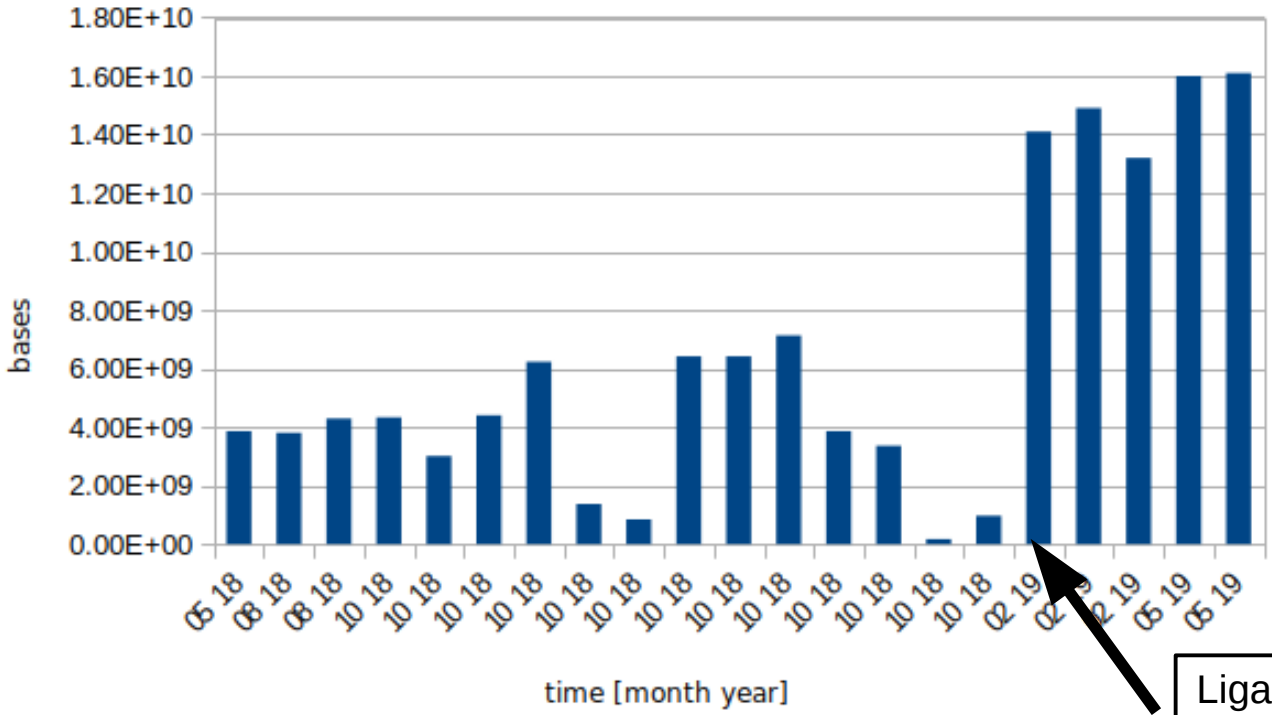
N = 39,000 reads → MHC covered completely 0 times

N ~ 1.5 flowcells
4N ~ 6 flowcells ~ 3600 - 6000€



https://github.com/DiltheyLab/ContigAnalysisScripts/

Quantity Improvements of Nanopore at BMFZ



=> 2-3 Flowcells
per cellline

Ligation Sequencing Kit 109
Flowcell revD
shearing at 75kb with Megaruptor

weighted histogram