Tools for microbial genome analysis

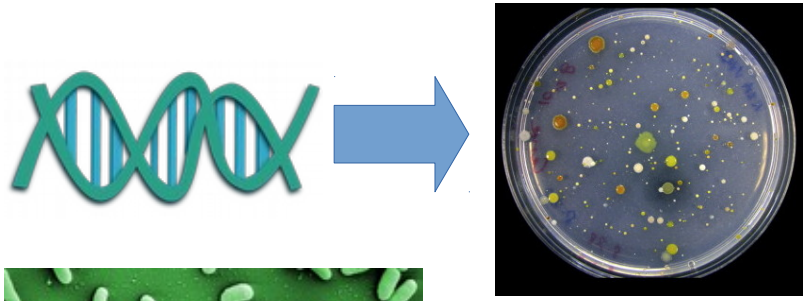
Aaron Weimann
EMBO course
April 6 ‘17

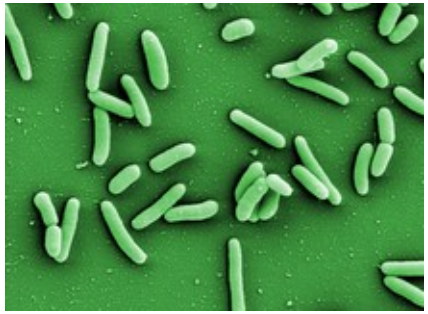
HELMHOLTZ
ZENTRUM FÜR
INFEKTIONSFORSCHUNG

About me

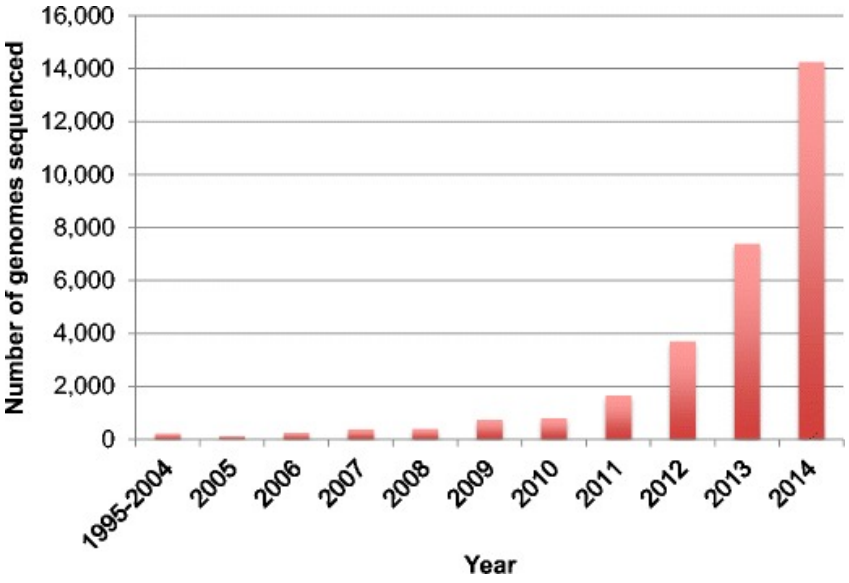
- Bioinformatics Post-doc
- Software for microbial geno2pheno inference
- *Pseudomonas aeruginosa* antibiotic resistance genomics

HELMHOLTZ
ZENTRUM FÜR
INFEKTIONSFORSCHUNG





Sequenced bacterial genomes

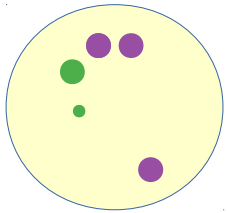


Land *et al.* 2015

- Requires fast and sophisticated analysis software

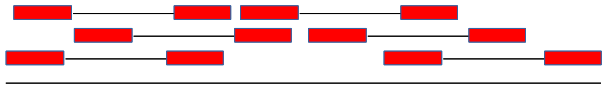
Bacterial genome analysis

Bacterial cell cultures

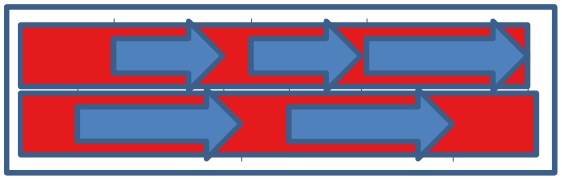


Shotgun genome sequencing

Assembly

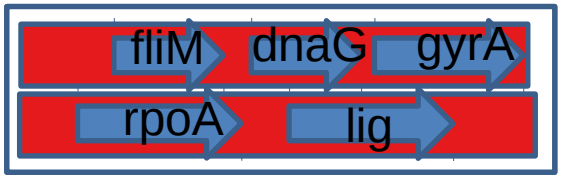


Determining gene coding regions

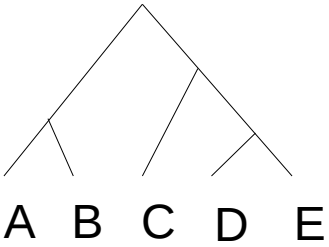


Bacterial genome analysis

Functional annotation



Phylogenetic tree inference



Microbial phenotype prediction

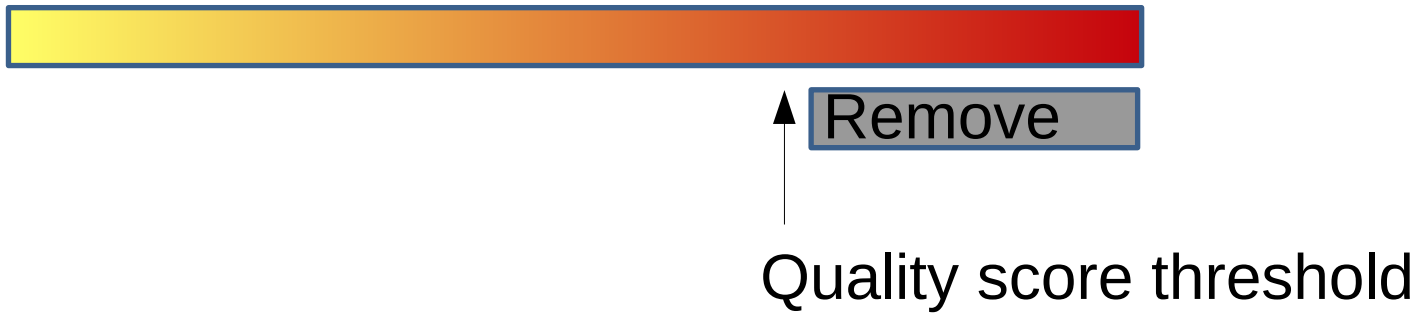
	A	B
Gram-negative		
Colistin-Polymyxin		
Growth in 6.5% NaCl		

Presence or absence of phenotype

Raw sequencing reads

- Cut sequencing adapters

- Quality trimming



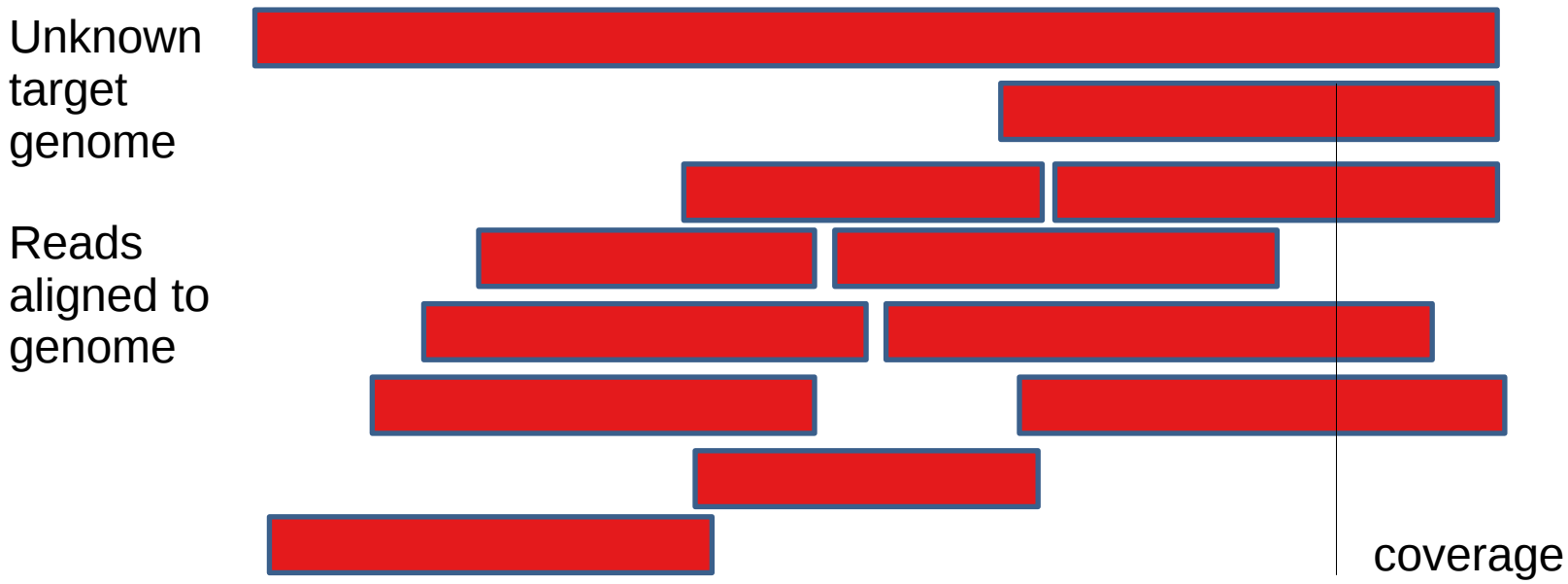
- Software: Trimmomatic (Bolger *et al.*, 2014)

Tutorial

- FastQC – (Andrew, 2010)

Sequencing reads

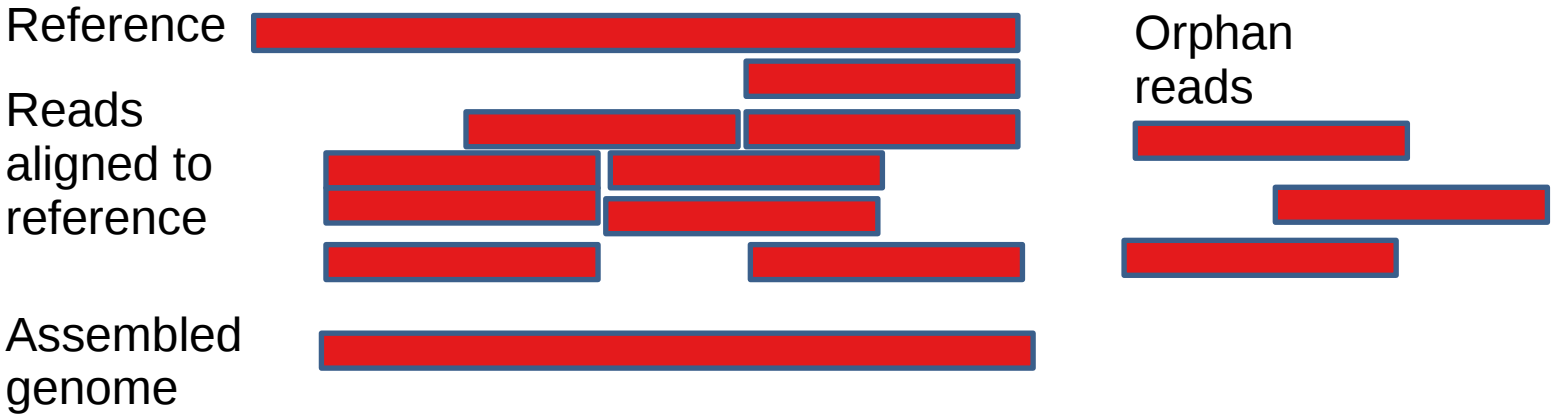
- How much of genome of interest covered?



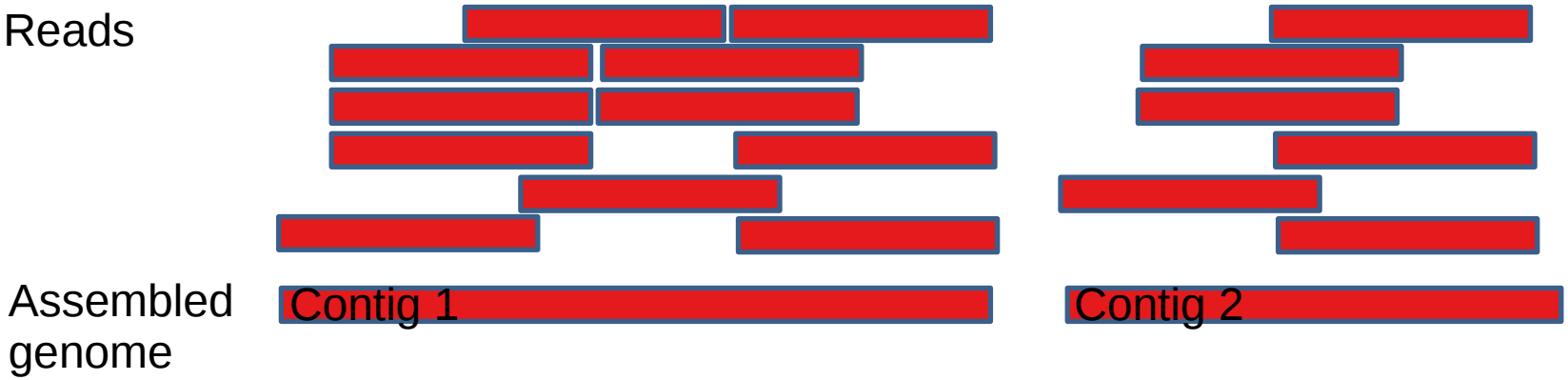
- Coverage / sequencing depth: mean number of reads covered at every base; 30x for bacteria

Assembly

- Reference-based assembly

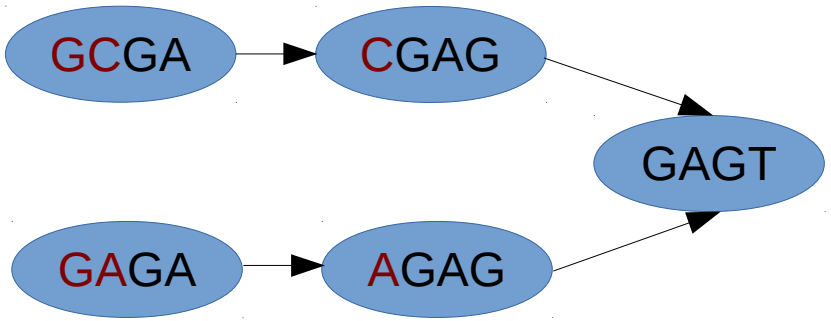


- De novo assembly



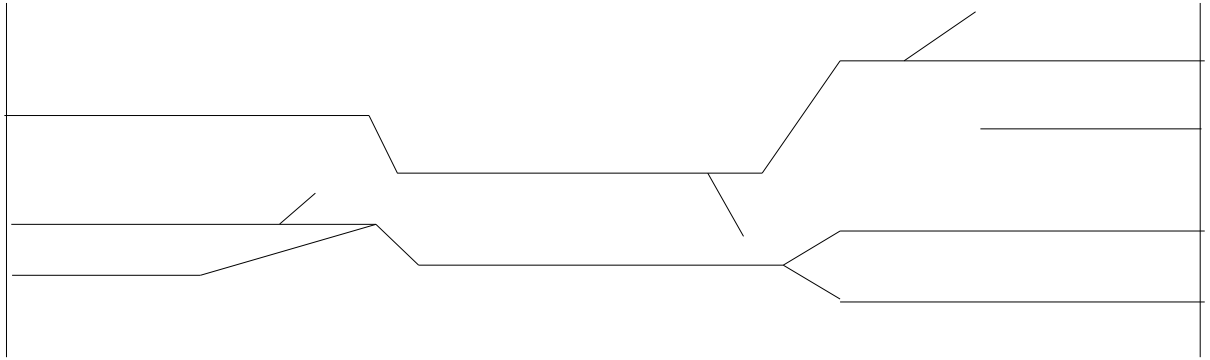
De Bruijn graphs

- Idea: split up reads into overlapping k-mers
- Example using 4-mers:
  - Read 1: GCGAGT
  - Read 2: GAGAGT

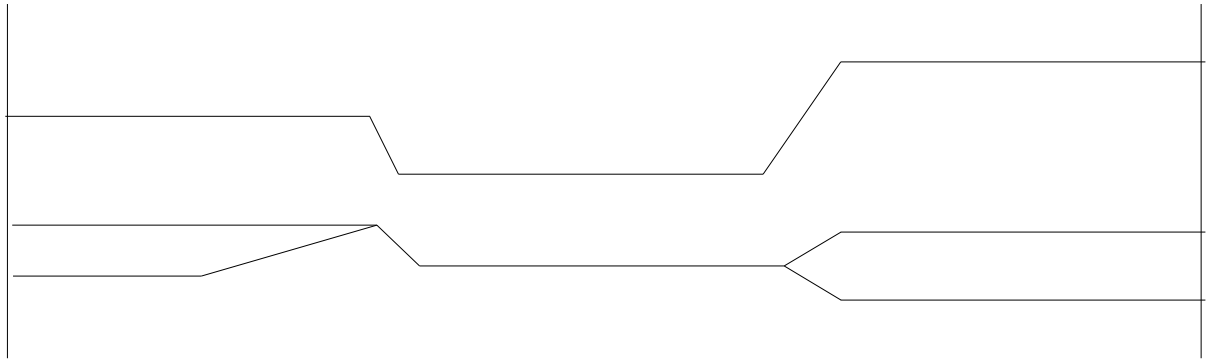


Sequencing errors

- Sequencing errors cause unconnected tips

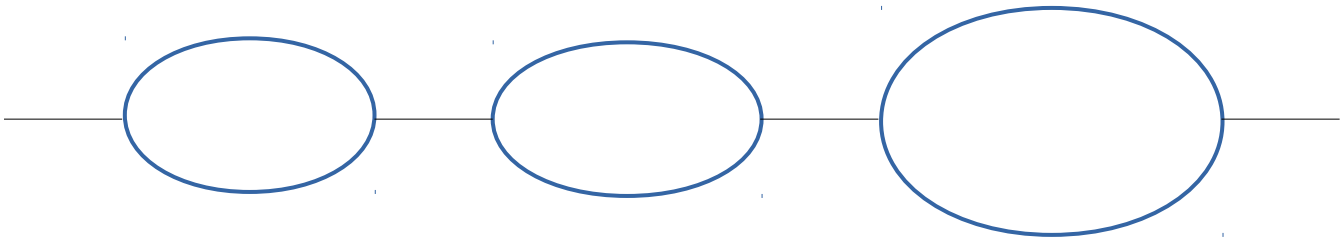


- Remove tips from the assembly

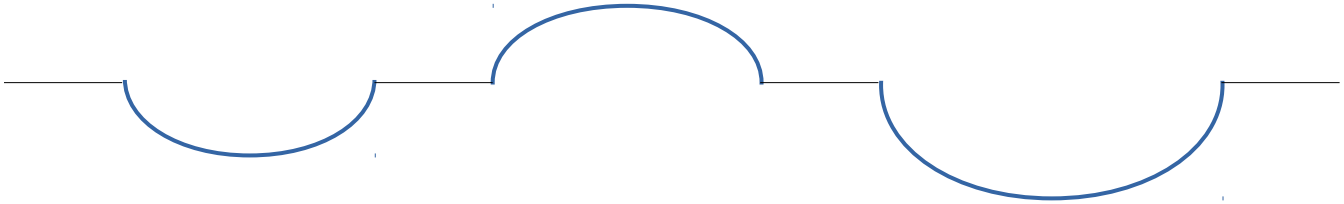


Repeats and sequence variation

- Sequence variation causes bubbles



- Assembler needs to decide for one path



Different assemblers

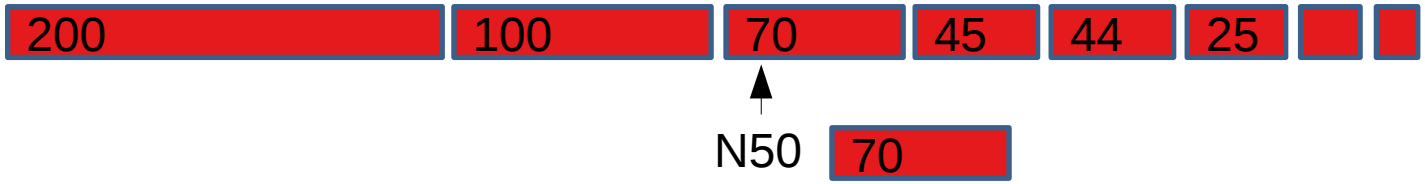
- A5-Assembler: Darling *et al.*, 2015
  - Integrated workflow: Read cleaning, contig assembly (IDBA-UD), scaffolding
- SPAdes: Bankevich *et al.*, 2012
  - Very reliable
- MegaHit: *Li et al.*, 2016
  - Very fast and reliable

Tutorial

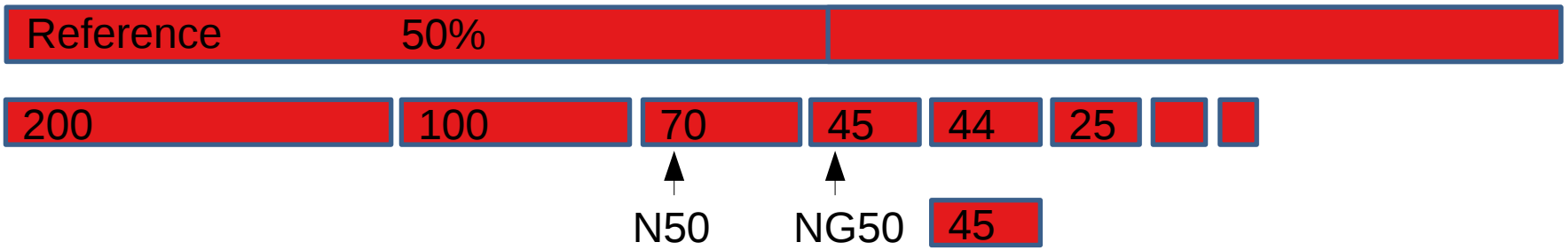
- MegaHit

Assembly quality

- N50: shortest contig length at 50% genome



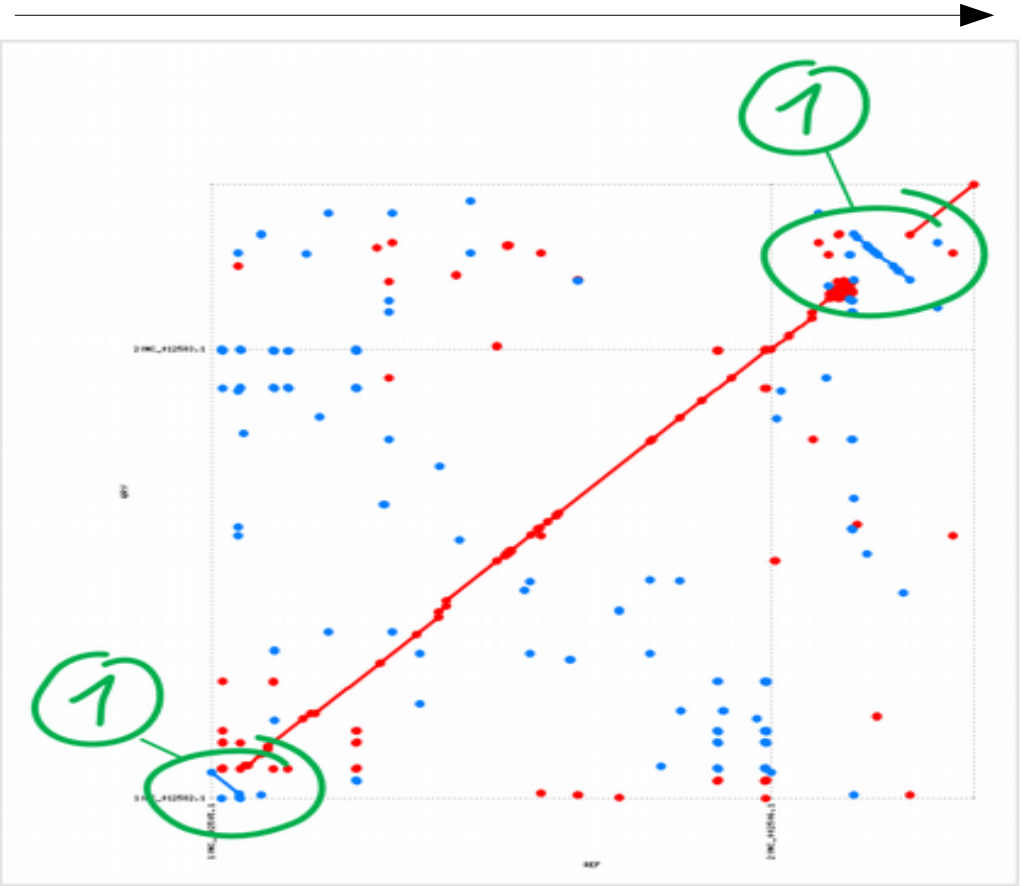
- NG50: shortest contig length at 50% of **reference** genome size



- Number of misassembled contigs
- Software: Quast, Gurevich *et al.* 2013

NUCMER alignment

Assembled genome



Reference genome

Delcher et al. 1999

- Whole genome alignment reveals two inversions

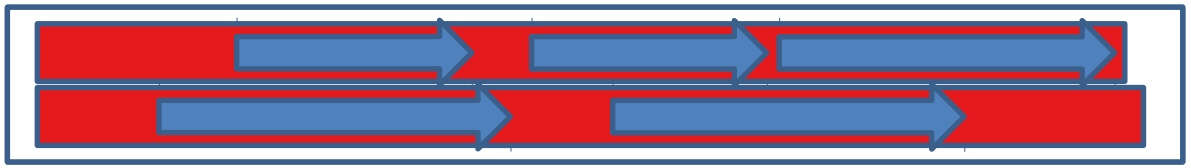
https://chunlab.wordpress.com/whole-genome-alignment-using-nucmer/

Tutorial

- Quast

Microbial genome annotation

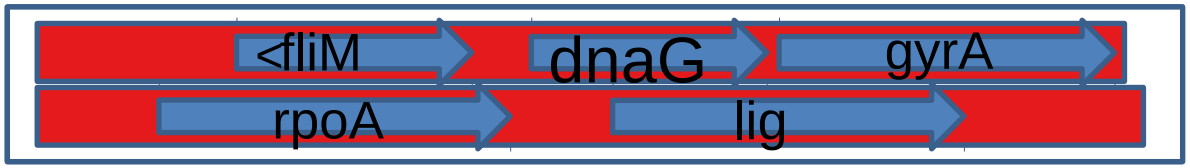
- Gene calling: identify coding regions



- Use discriminatory features: GC content, promotor regions, etc.
- Popular tools
  - geneMark (Lukashin *et al.* 1998) – the classic
  - Prodigal (Hyatt *et al.* 2010) – most popular today
  - FragGeneScan (Rho *et al.* 2010) – for short reads

Prokka

- Very fast microbial genome annotation
- Gene product assignment



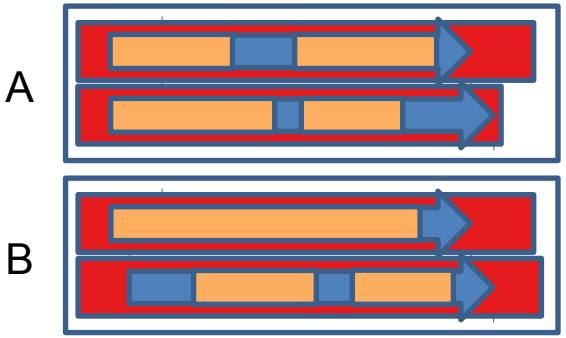
- Integrates various information sources
  - Hierarchical approach to search these
- Produces standard-compliant output files
- Includes Prodigal gene calling

Tutorial

- Prokka

In silico phenotyping

Protein family annotated genomes or genome bins



	A	B
Gram-negative		
Anaerobic		
Glucose fermenter		
Colistin-Polymyxin		
Growth in 6.5% NaCl		

Presence or absence of phenotype

"Sample A is **Gram-negative** and **anaerobic**. It's **susceptible to Colistin-Polymyxin**."

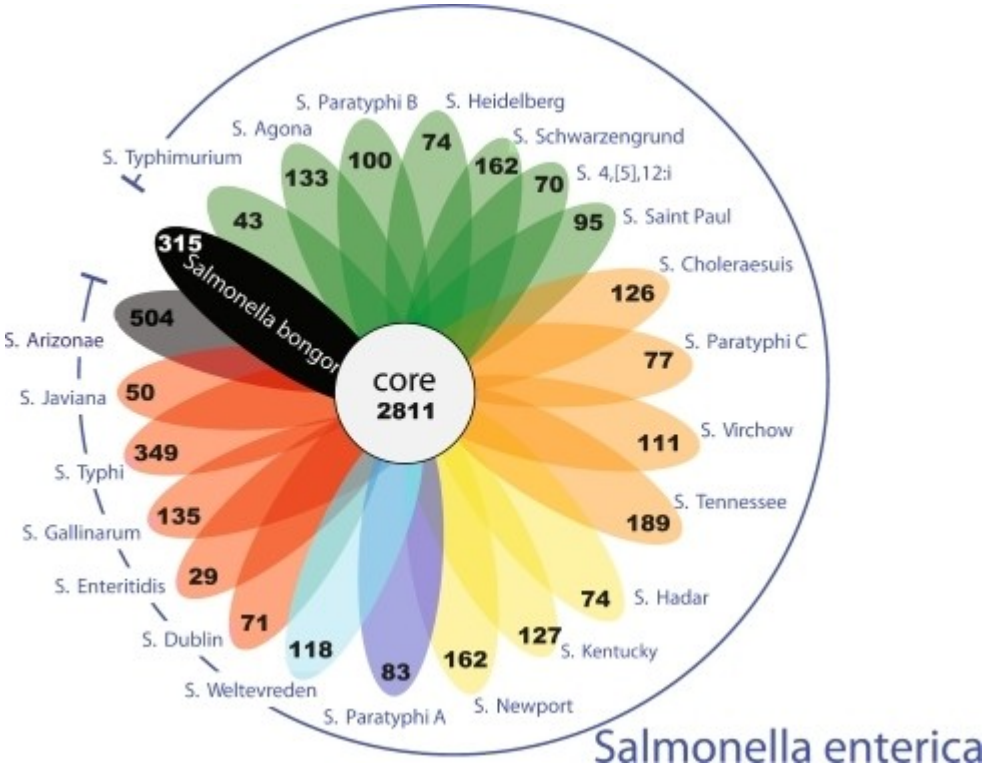
- Traitar: the microbial traitar analyzer (Weimann *et al.* 2016)
- Accurate prediction of 67 diverse traits solely based on genome information

Tutorial

- Traitar

The pan genome

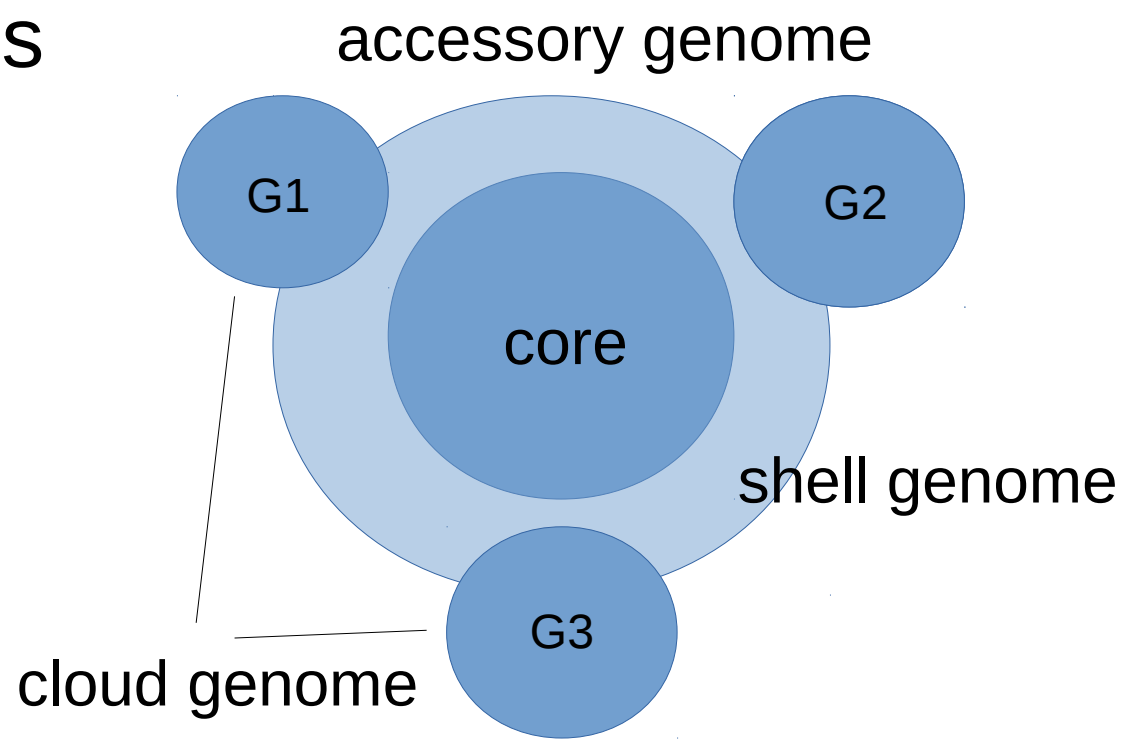
- Union of genes shared by genomes of interest
- Acquired genes related to abr, virulence. etc.



Jacobsen *et al. 2011*

Identify the pan genome

- Cluster homologous protein sequences
- Use synteny and operon information
- Software: Roary (Page *et al.* 2015)



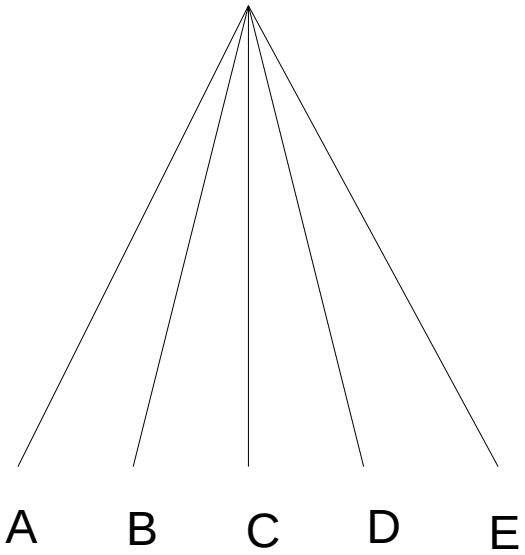
Tutorial

- Roary

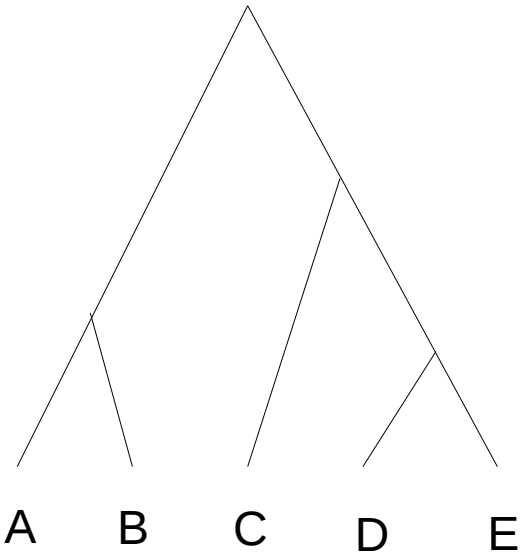
Phylogenetic trees

- Resolve the branching order of lineages

Unresolved phylogeny



Fully resolved phylogeny



- Applications: bacterial evolution, transmission history, etc.

Data for phylogenetic inference

- Varying sites (SNPs) in DNA or proteins
- Multiple sequence alignment

Samples	Characters																		
Sample 1	G	G	C	C	T	A	G	T	A	T	A	G	T	C	G	A	G	A	C
Sample 2	G	G	A	C	G	A	G	G	A	T	A	G	T	C	C	A	G	G	C
Sample 3	G	G	C	C	T	A	G	G	A	T	A	G	A	C	C	A	G	G	C
Sample 4	G	G	A	C	G	A	G	T	A	T	T	G	A	C	G	A	G	A	C
Sample 5	G	G	C	C	T	A	G	T	A	G	T	G	T	C	G	A	G	A	C

- Often use marker genes e.g. 16S or core genome concatenated sequence alignment

Algorithms for phylogenetic inference

- Statistical or mathematical methods to infer order of taxa

	Optimality criterion	Clustering algorithm
Characters	Parsimony Maximum likelihood Bayesian inference	
Distances	Minimum Evolution Least Squares	UPGMA Neighbor joining

Software for tree inference

- RAxML - Randomized Axelerated Maximum Likelihood (Stamatakis *et al.* 2006)
  - Very reliable
- FastTree approximately-maximum-likelihood phylogenetic trees *(*Price *et al. 2009)*
  - Reliable and very fast
- A lot more options out there

Tutorial

- FastTree

Further topics

Assembly of long-read sequencing reads

- Hybrid assembly

- Variation analysis
  - Short read mapping using programs like BWA
  - SNP calling, structural variations
- A lot of further things

Summary

- Complete workflow from sequencing read to function
- A lot of great microbial genomics software out there
- Try them out – redo our tutorial; use Traitar for your genomes if you like ;-)
- Get in touch if you have some traits and genome data and you need a genotype-phenotype model
- Twitter: @aaron_weimann
- Email: weimann@hhu.de