Bioinformatics and traveling across europe

Beat Wolf

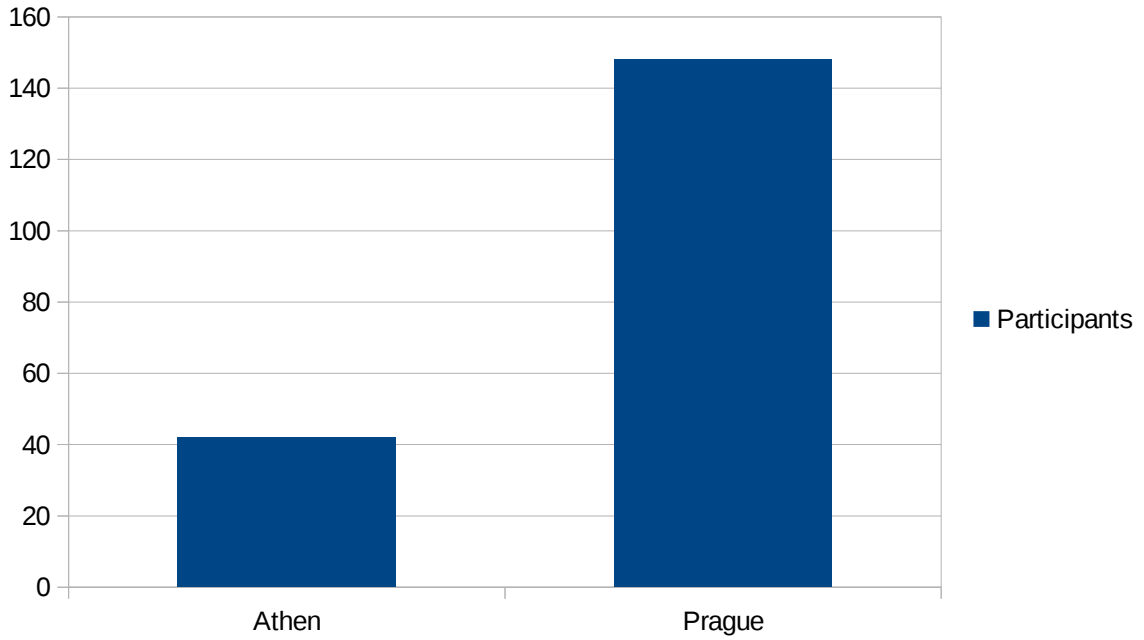
Overview

- Bioinformatics course in Prague and Athen
- Poster presentation in Leiden/Netherlands

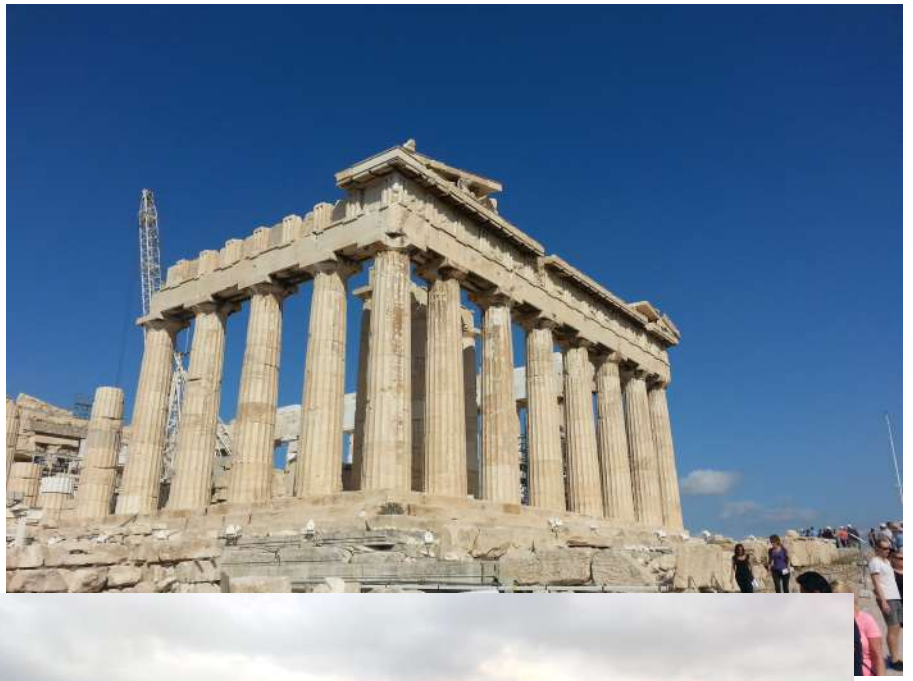
- European research project
- Organizes introductory courses for geneticists
  - NGS Course; next-generation sequencing in a diagnostic setting
- Topic, next generation sequencing in diagnostics
- I was asked to give a 45 minute course + practicals

NGS Course; next-generation sequencing in a diagnostic setting

- 2 courses have been held
  - 2014 Athens, 42 participants
  - 2015 Prague, 148 participants



Athens



Prague

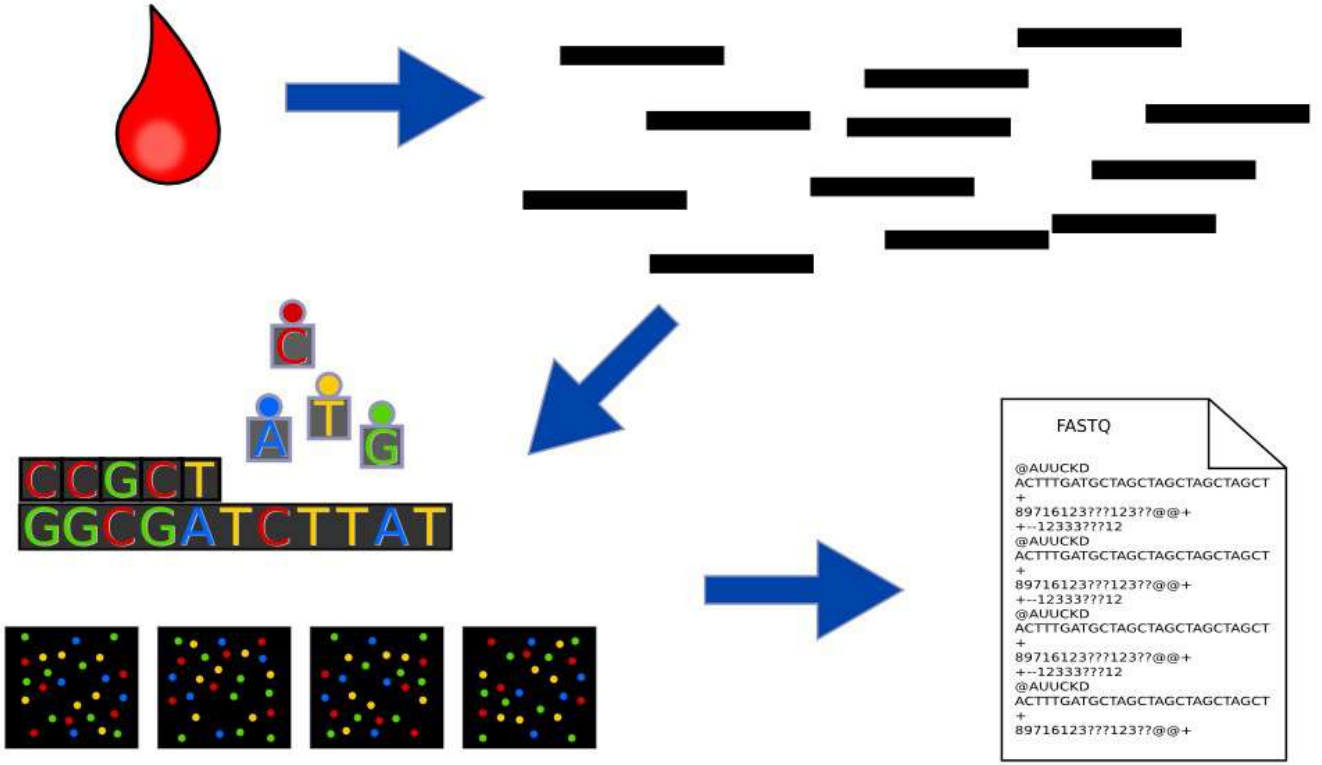
Sequence alignment vs assembly

Two ways to extract information from DNA sequencing data.

- Data generation
- Genome reconstruction
- Variant detection

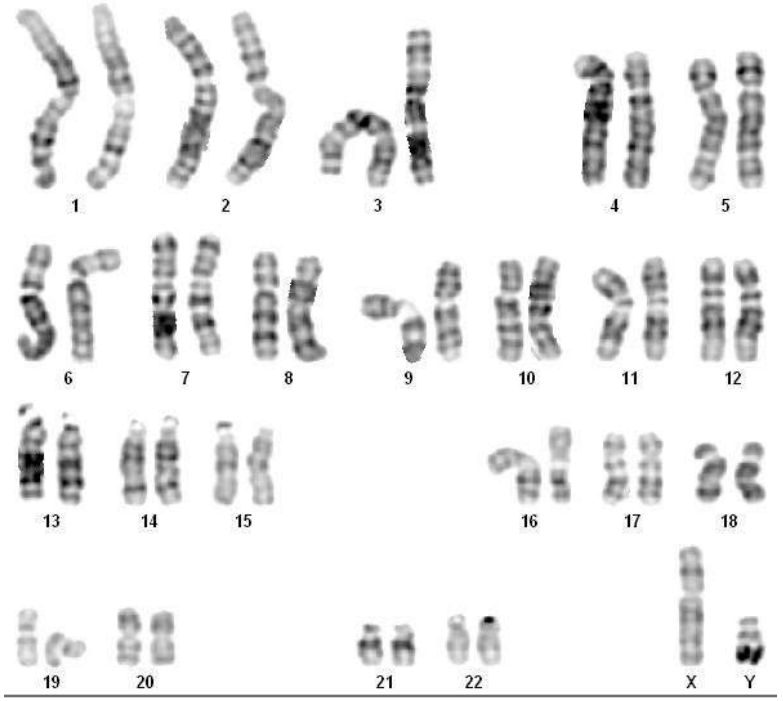
NGS data generation

NGS data generation



Genome reconstruction

- The problem:
  - Billions of short reads of unknown origin
  - Human genome consists of 3 billion bases (ACTG)



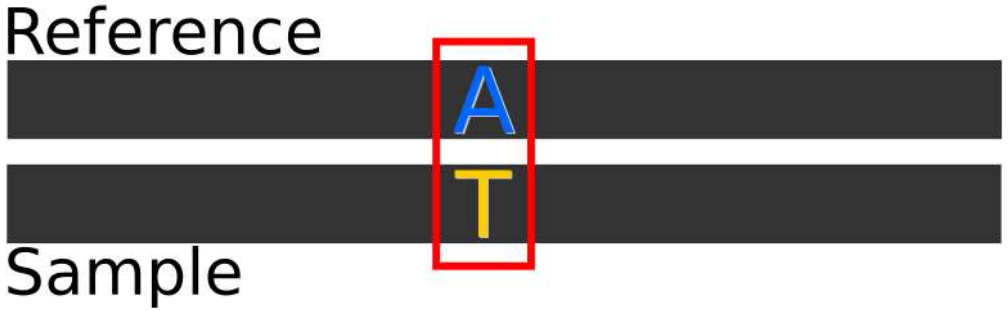
Raw data

- The reads contain errors
  - Base replacements
  - Bases deleted or new ones inserted
- Standardized sequencer output:

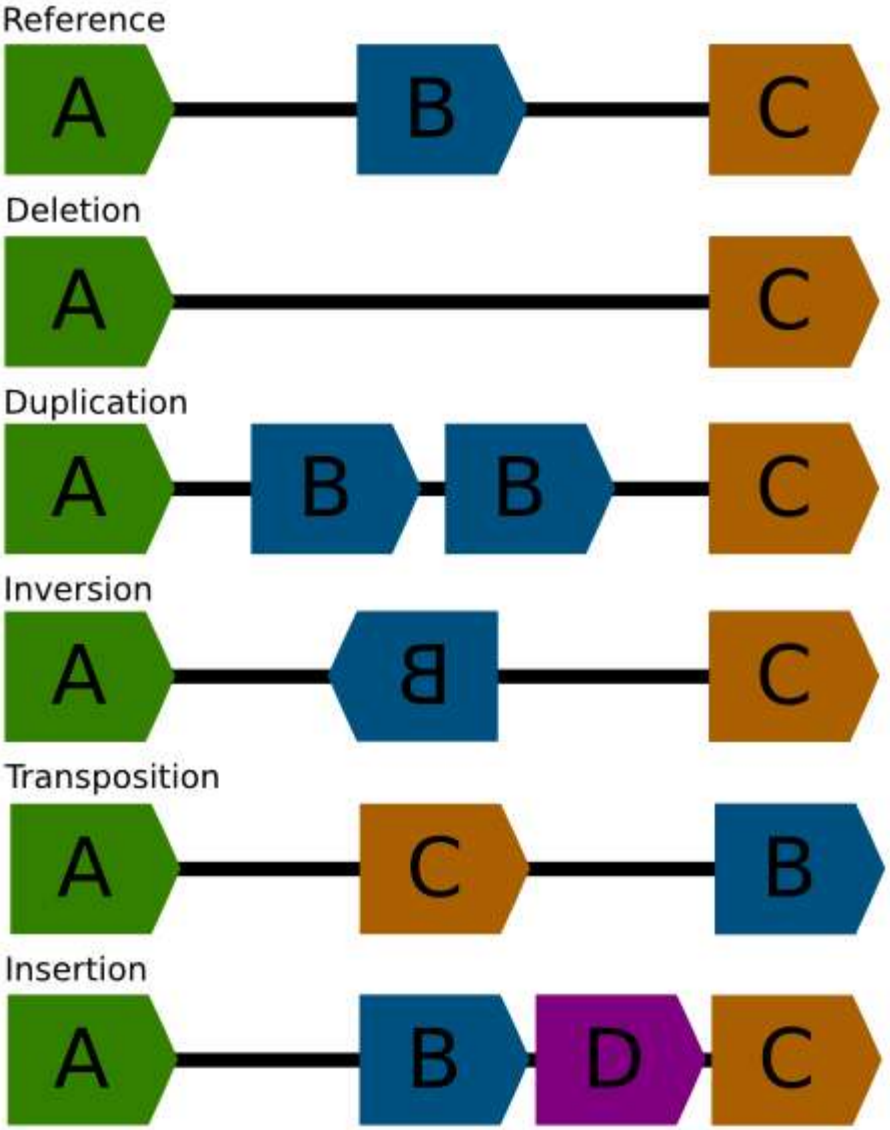
```
@9WV6Z:791:946
GCTCTTCCGATCTATGGATGCACCAAGATATATGACCCTGTCTGTGGGACTGATGGAA
+
774399222034221774399222034221774399222034221774399222034
```

Variants

- SNV



- Indel



Genome reconstruction

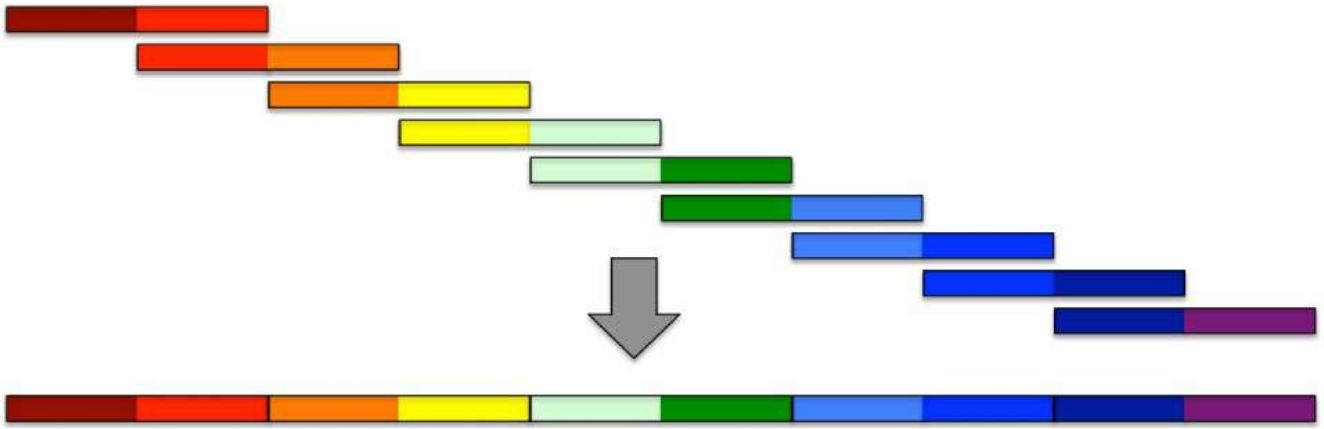
- Two possible approaches:
  - Sequence assembly
  - Sequence mapping
- Sequence assembly requires no prior knowledge, uses synergies between reads
- Sequence alignment uses known information about the sample to recreate the genome

Sequence assembly

- Construct assembly graph from overlapping reads

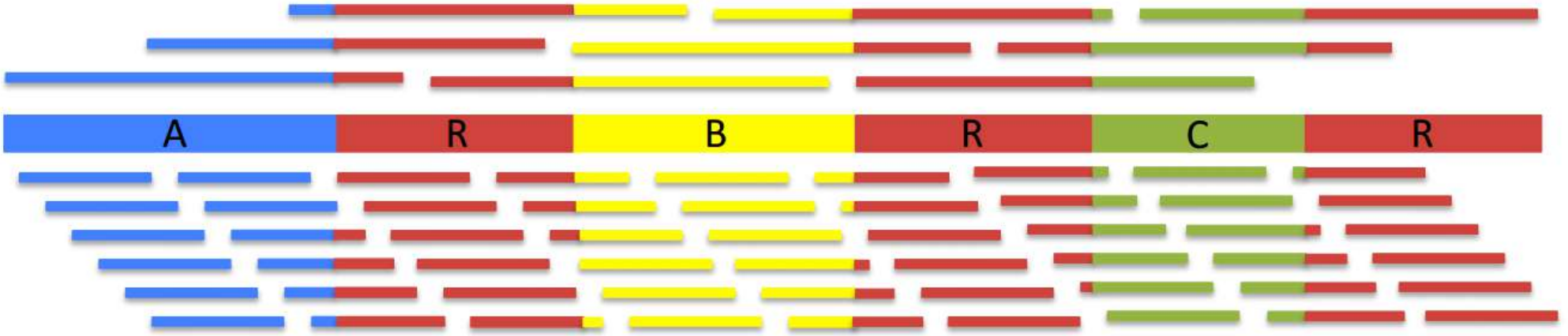
...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

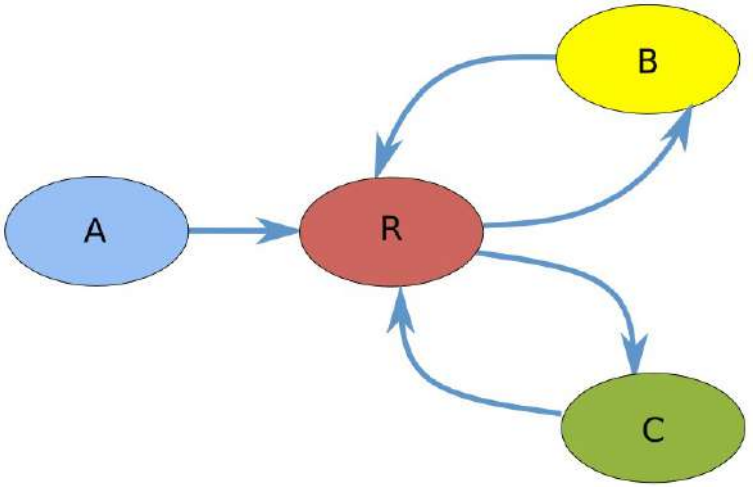
- Simplify assembly graph



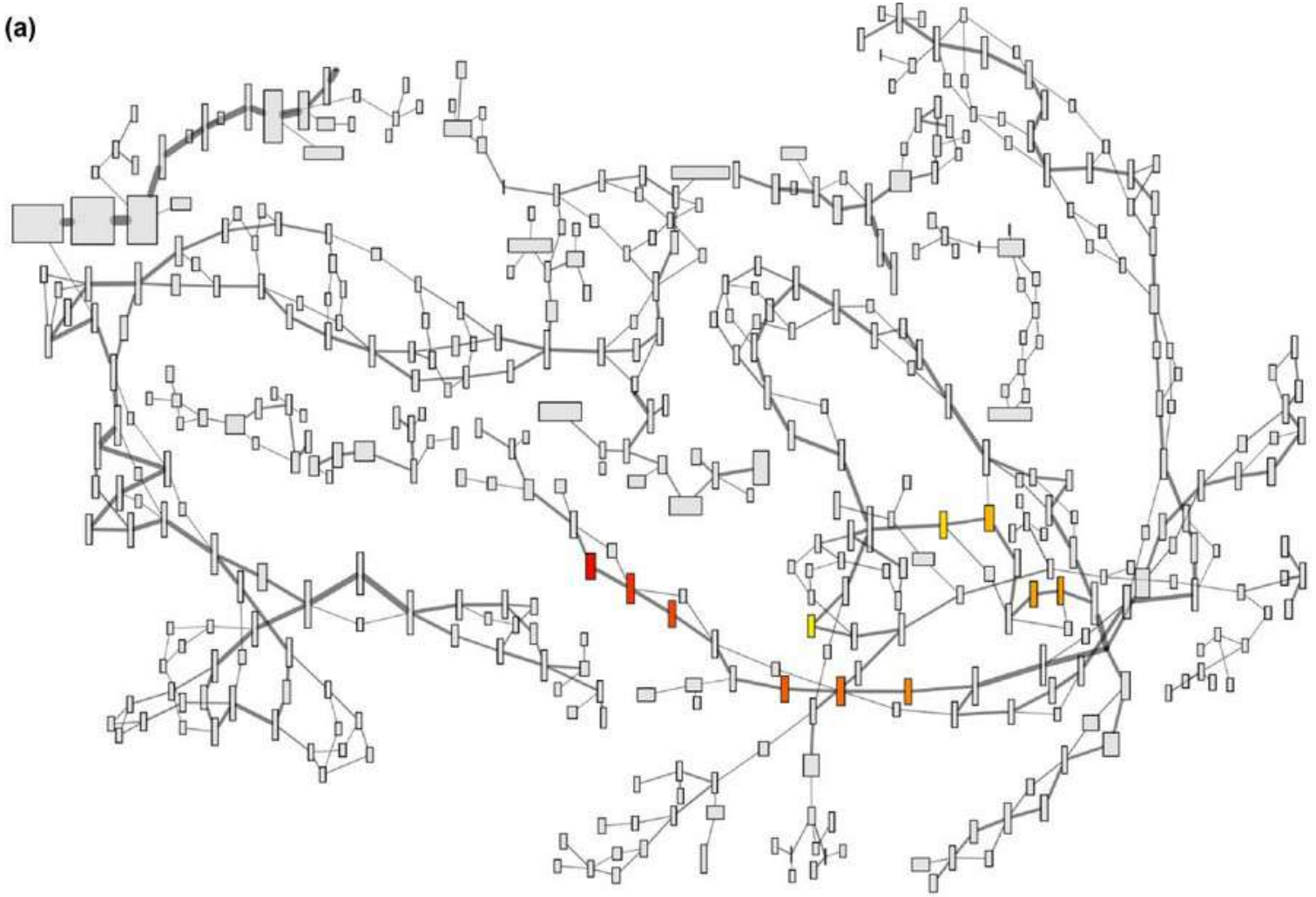
Sequence assembly

- Problem: repeated regions



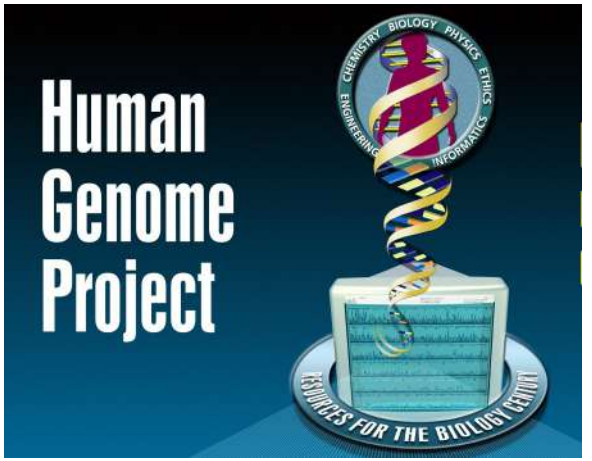


Sequence assembly



Human genome project

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



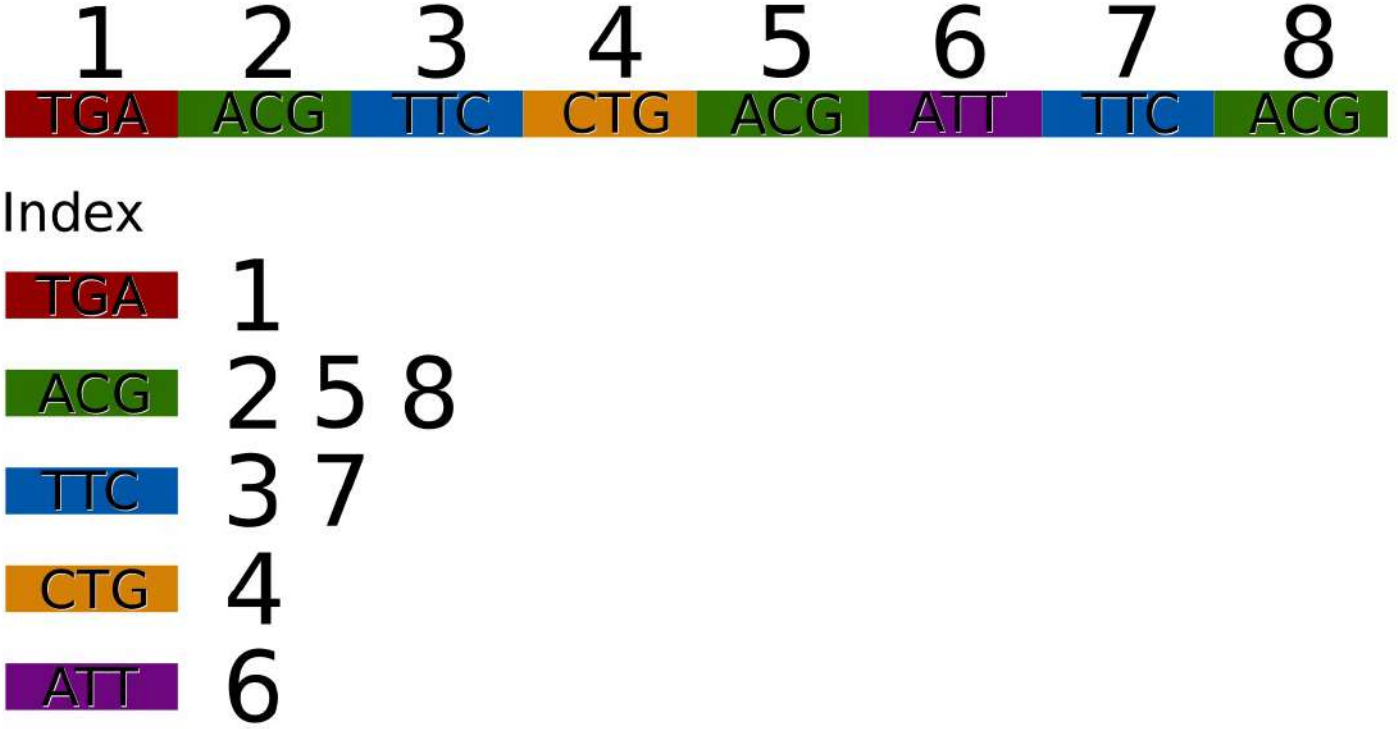
Sequence alignment

- Naive approach:
  - Evaluate every location on the reference

- Too slow for billions of reads on a big reference

Sequence alignment

- Speed up with the creation of a reference index

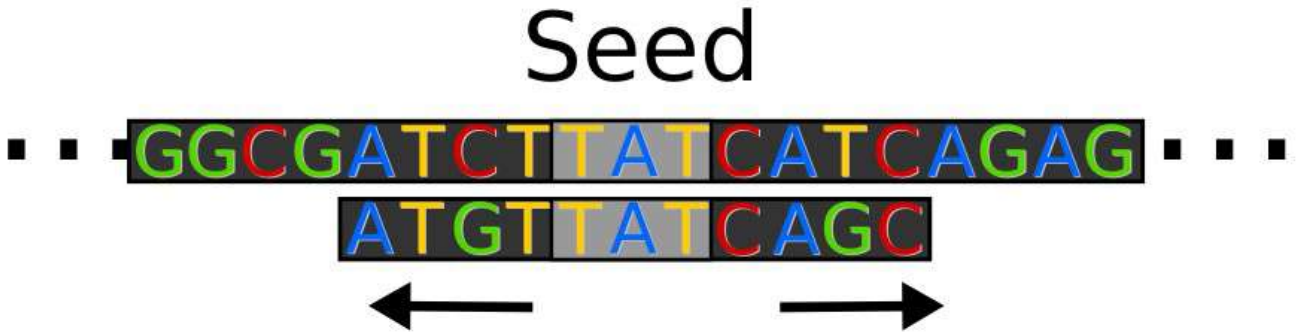


- Fast lookup table for subsequences in reference

Sequence alignment

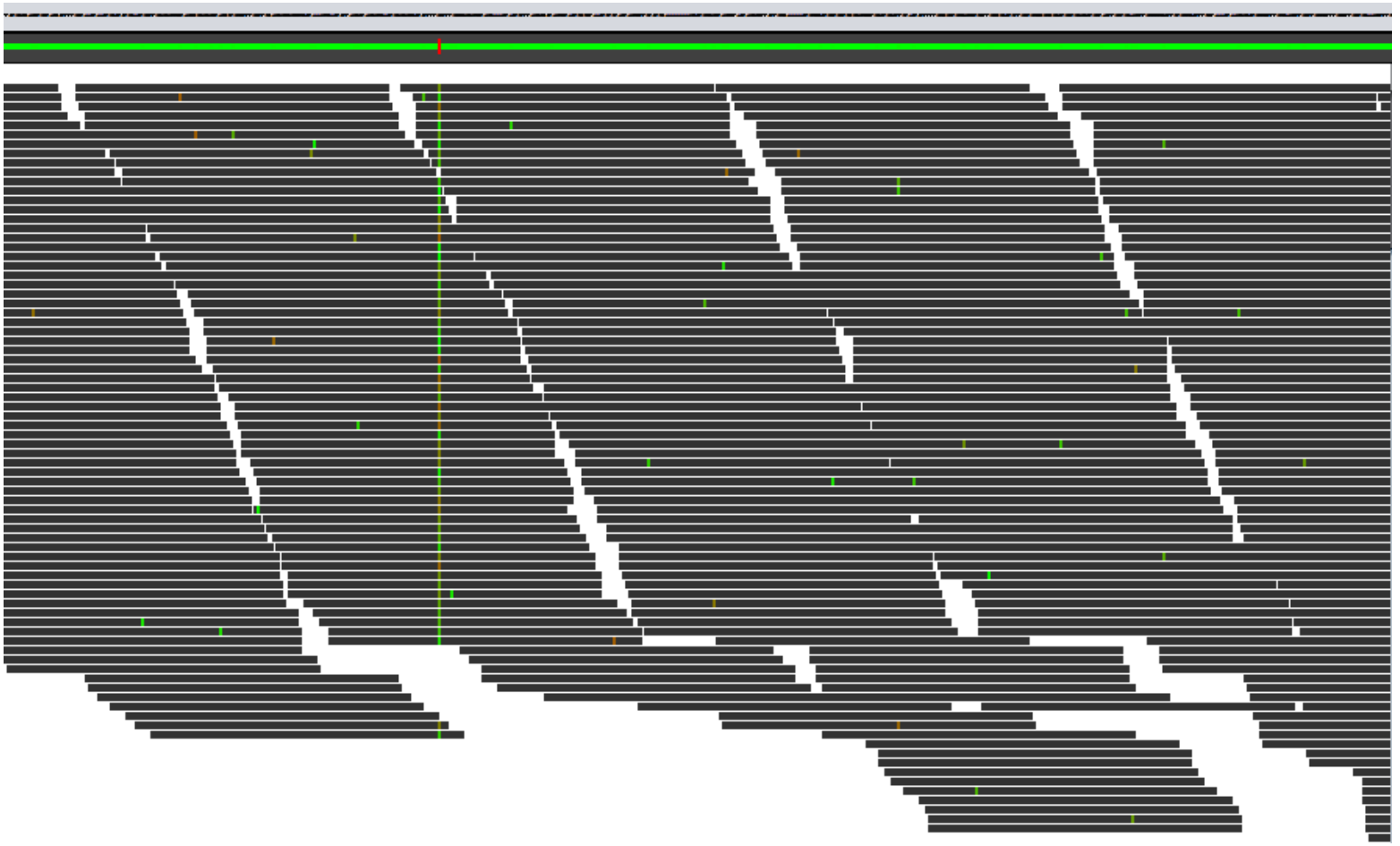
- Find all possible alignment positions
  - Called seeds

- Evaluate every seed



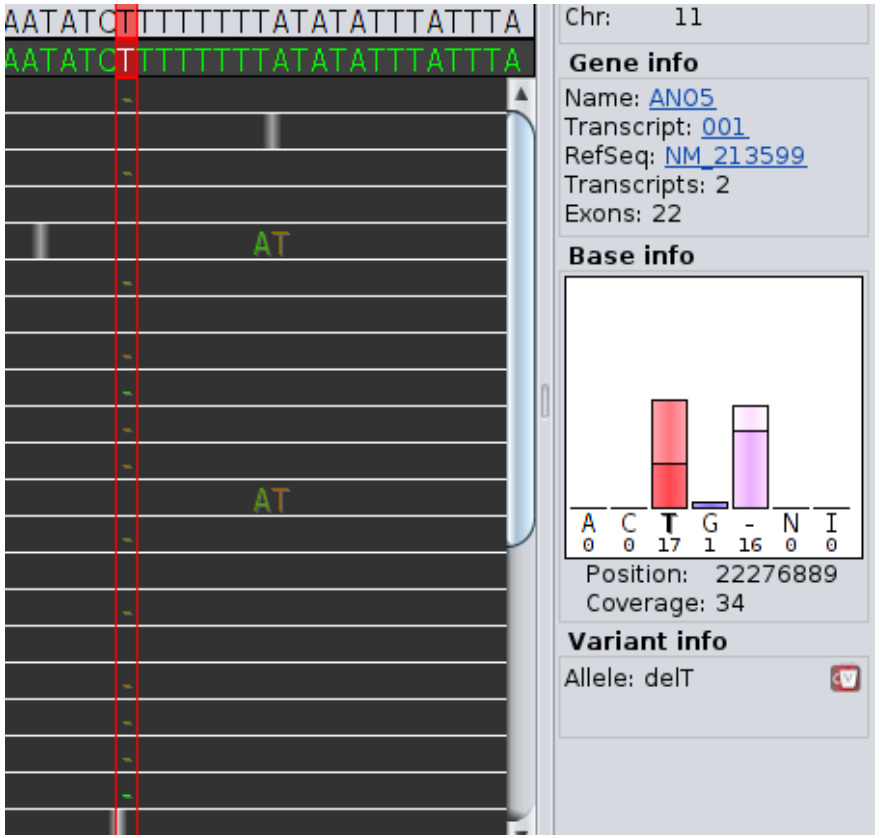
Result

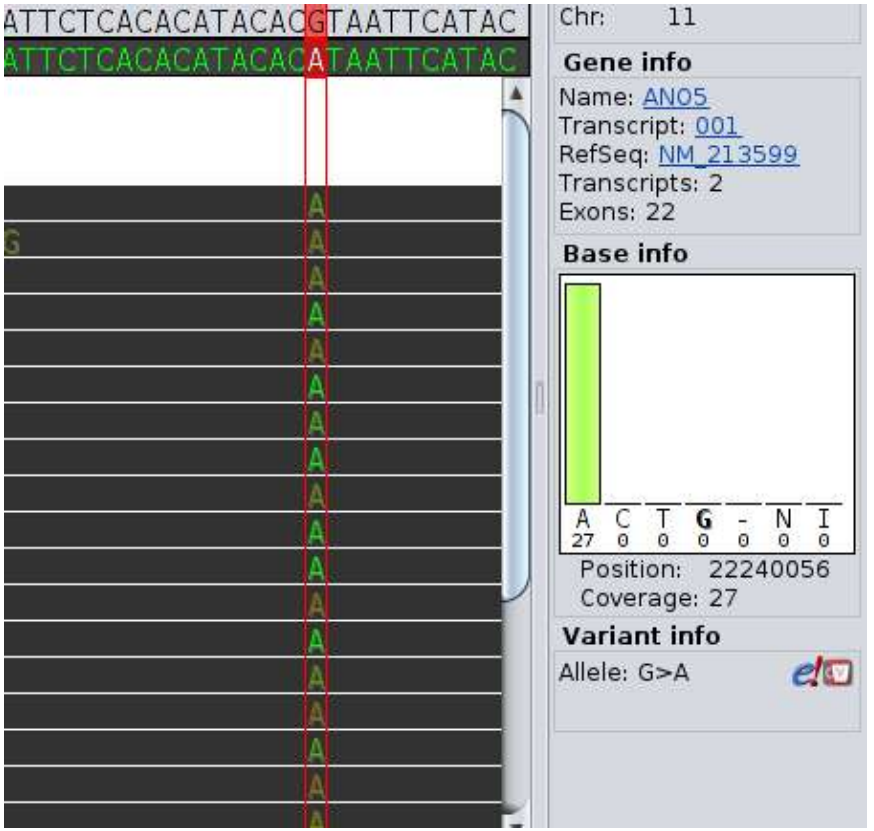
- Final result, an alignment file (BAM)



Variant calling

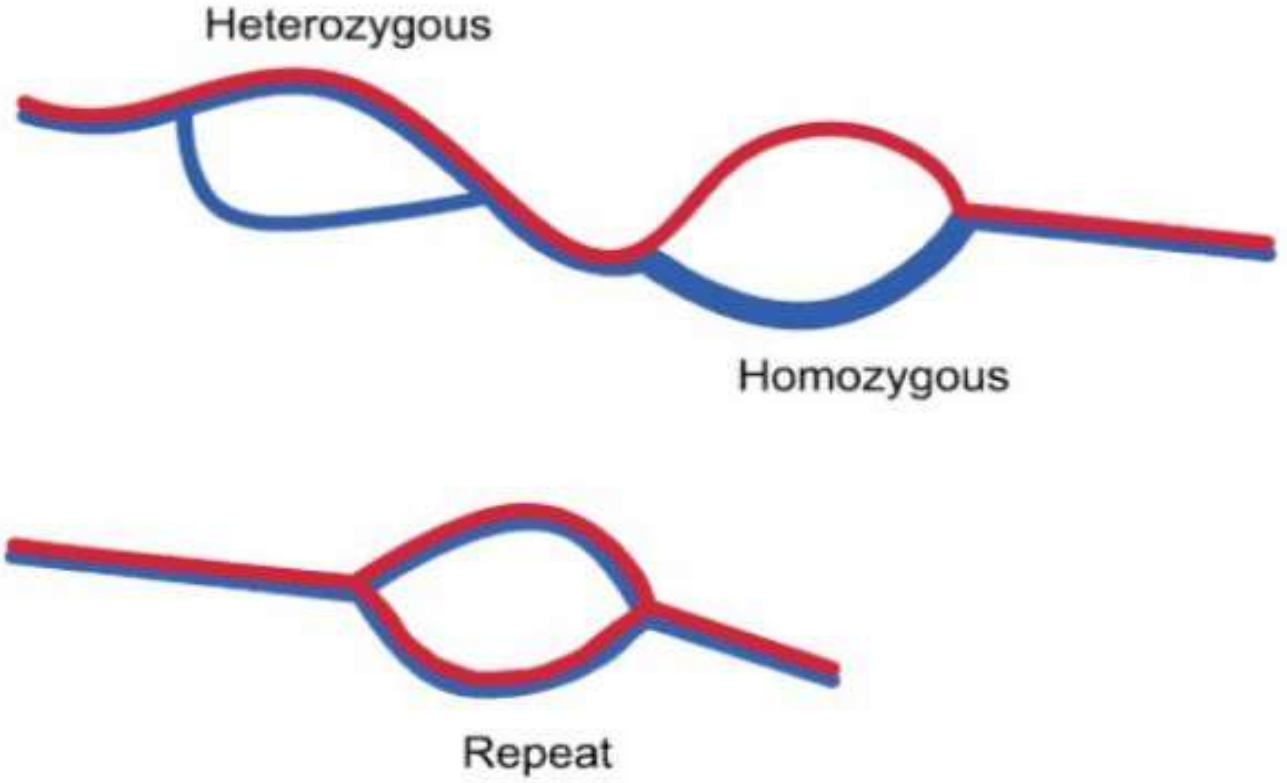
- Alingment based variant calling
  - Compare to reference, list differences





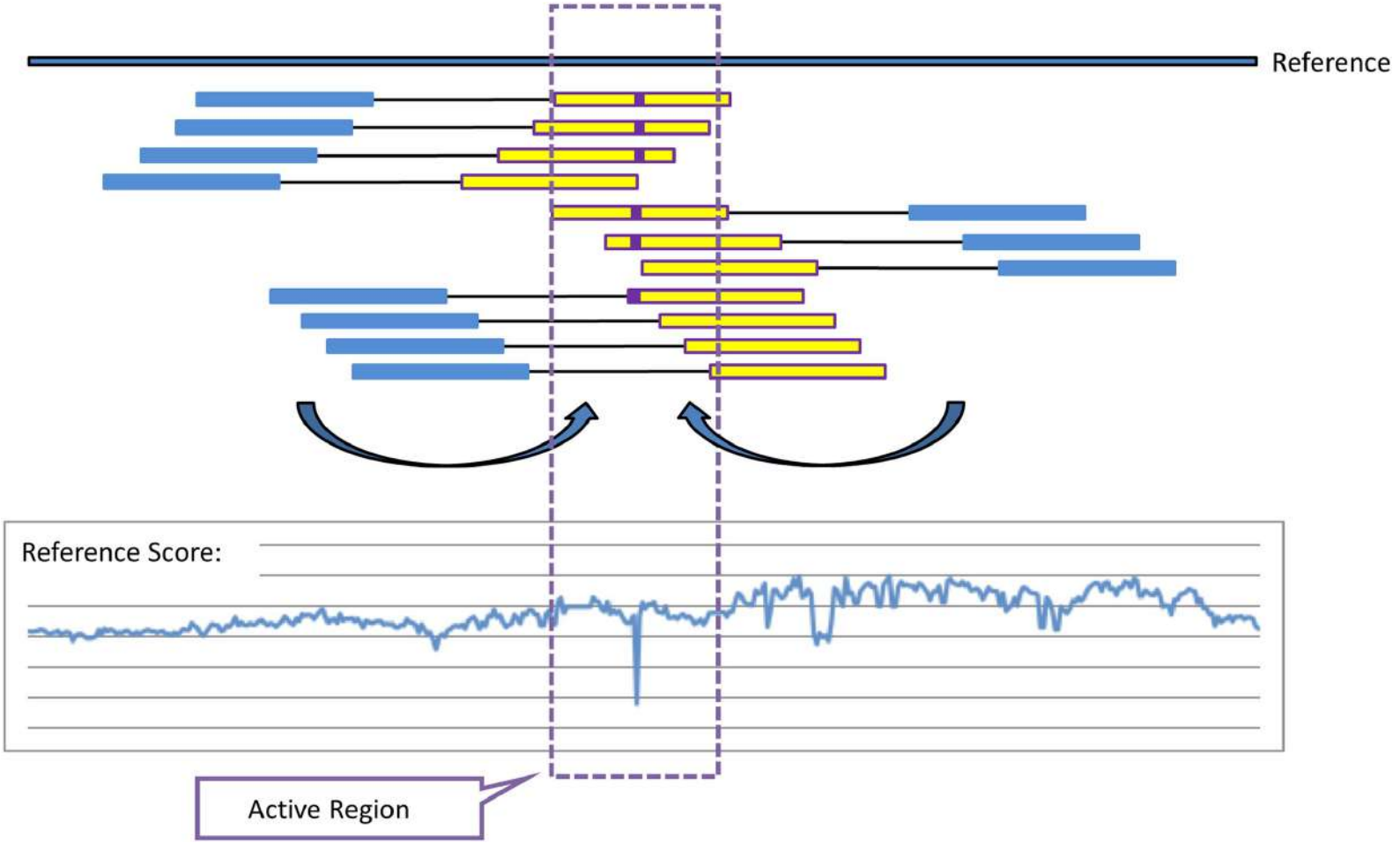
Variant calling

- Assembly based variant calling
  - One possibility, compare de Bruijn graphs



Variant calling

- Hybrid method
  - Start with alignment, improve with local assembly



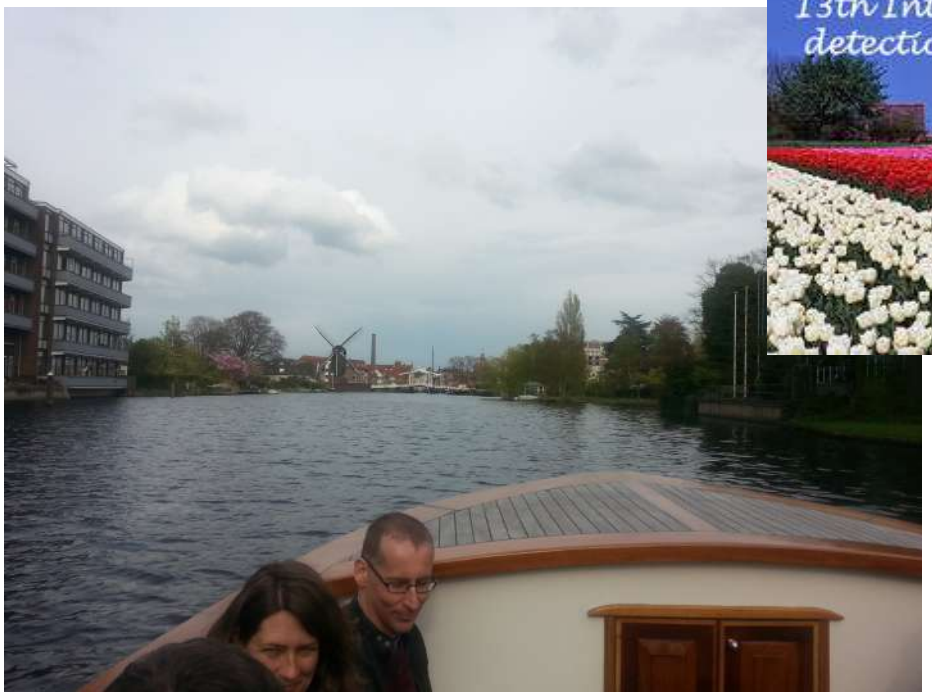
Future

- Sequence assembly will replace sequence alignment
- Sequence assembly recreates the original genome, in contrast to alignment
- Current sequencing technologies are not yet ready for it, but will soon

13th International Symposium on Mutation in the Genome: detection, genome sequencing & interpretation

- Technology - sequencing
- Technology - variant detection
- Technology - smaller, faster, cheaper
- From variant to function (or Functional analysis)
- Genome projects
- Diagnostics
- Knowledge from sharing
- Applications
- Bioinformatics - Variant Effect Prediction
- Databases and knowledge resources
- & more...

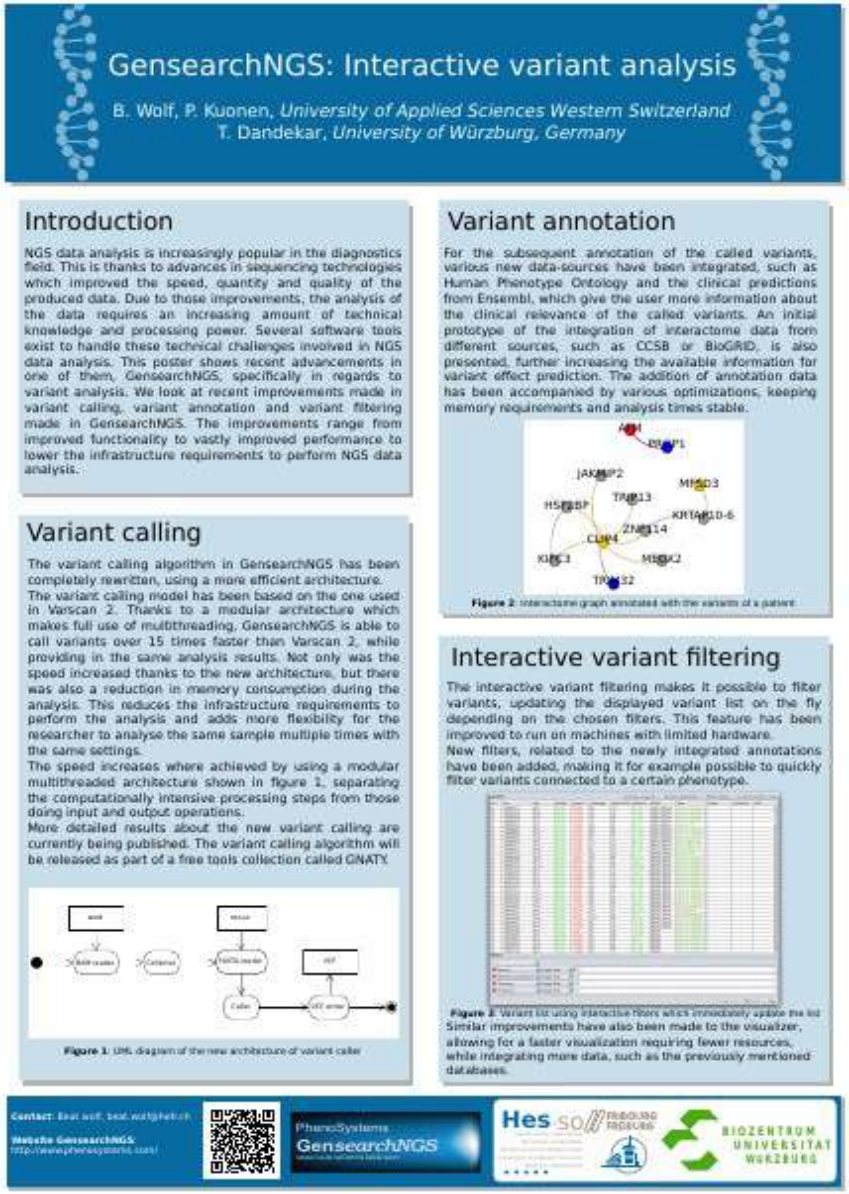
Leiden




13th International Symposium on Mutation in the Genome:
detection, genome sequencing & interpretation
27 - 30 April 2015 - Leiden, The Netherlands

Variant calling

- Poster presentation
  - Faster variant calling
  - Improved annotations
  - Interactive variant filtering



GensearchNGS: Interactive variant analysis

B. Wolf, P. Kuonen, *University of Applied Sciences Western Switzerland*
T. Dandekar, *University of Würzburg, Germany*

Introduction

NGS data analysis is increasingly popular in the diagnostics field. This is thanks to advances in sequencing technologies which improved the speed, quantity and quality of the produced data. Due to those improvements, the analysis of the data requires an increasing amount of technical knowledge and processing power. Several software tools exist to handle these technical challenges involved in NGS data analysis. This poster shows recent advancements in one of them, GensearchNGS, specifically in regards to variant analysis. We look at recent improvements made in variant calling, variant annotation and variant filtering made in GensearchNGS. The improvements range from improved functionality to vastly improved performance to lower the infrastructure requirements to perform NGS data analysis.

Variant calling

The variant calling algorithm in GensearchNGS has been completely rewritten, using a more efficient architecture.
The variant calling model has been based on the one used in Varscan 2. Thanks to a modular architecture which makes full use of multithreading, GensearchNGS is able to call variants over 15 times faster than Varscan 2, while providing in the same analysis results. Not only was the speed increased thanks to the new architecture, but there was also a reduction in memory consumption during the analysis. This reduces the infrastructure requirements to perform the analysis and adds more flexibility for the researcher to analyse the same sample multiple times with the same settings.
The speed increases where achieved by using a modular multithreaded architecture shown in figure 1, separating the computationally intensive processing steps from those doing input and output operations.
More detailed results about the new variant calling are currently being published. The variant calling algorithm will be released as part of a free tools collection called GNATX.

Figure 1: UML diagram of the new architecture of variant caller

Variant annotation

For the subsequent annotation of the called variants, various new data-sources have been integrated, such as Human Phenotype Ontology and the clinical predictions from Ensembl, which give the user more information about the clinical relevance of the called variants. An initial prototype of the integration of interactome data from different sources, such as CCSB or BioGRID, is also presented, further increasing the available information for variant effect prediction. The addition of annotation data has been accompanied by various optimizations, keeping memory requirements and analysis times stable.

Figure 2: Interactome graph annotated with the variants of a patient

Interactive variant filtering

The interactive variant filtering makes it possible to filter variants, updating the displayed variant list on the fly depending on the chosen filters. This feature has been improved to run on machines with limited hardware.
New filters, related to the newly integrated annotations have been added, making it for example possible to quickly filter variants connected to a certain phenotype.

Figure 3: Variant list using interactive filters which immediately update the list

Similar improvements have also been made to the visualizer, allowing for a faster visualization requiring fewer resources, while integrating more data, such as the previously mentioned databases.

Contact: [illegible]

Website GensearchNGS: [illegible]

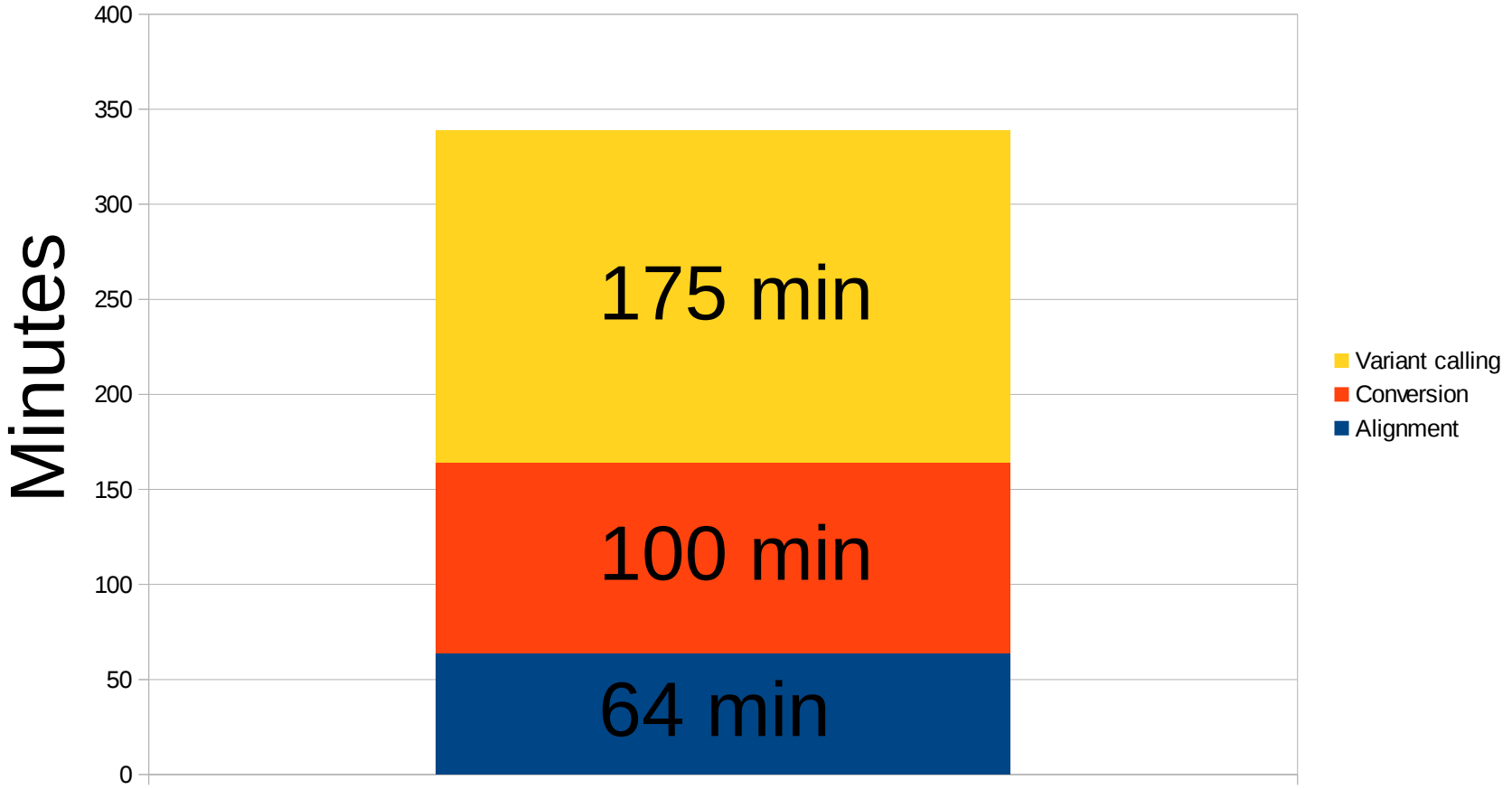
PhenoSystems GensearchNGS

Hes so

Biozentrum Universität Würzburg

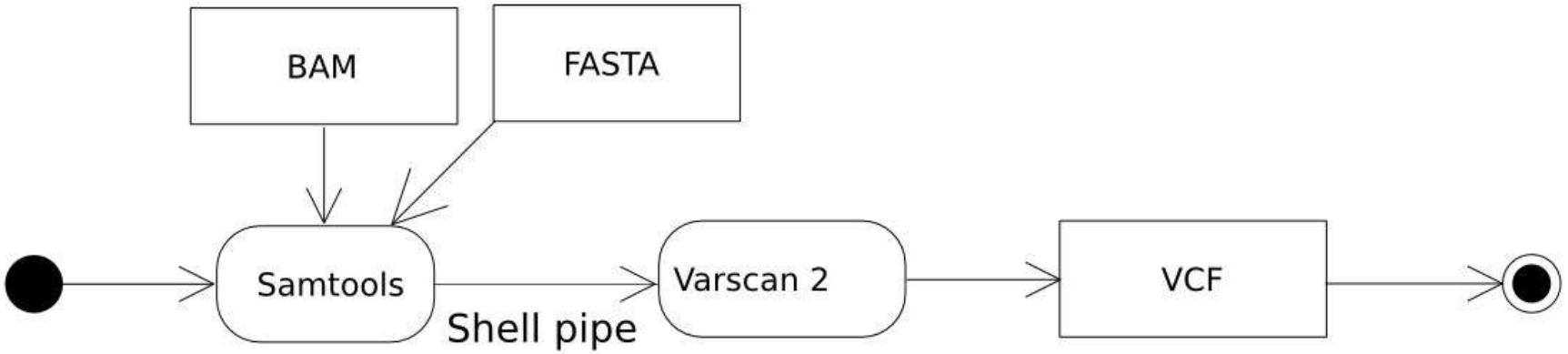
Current variant calling

- Varscan 2 performance



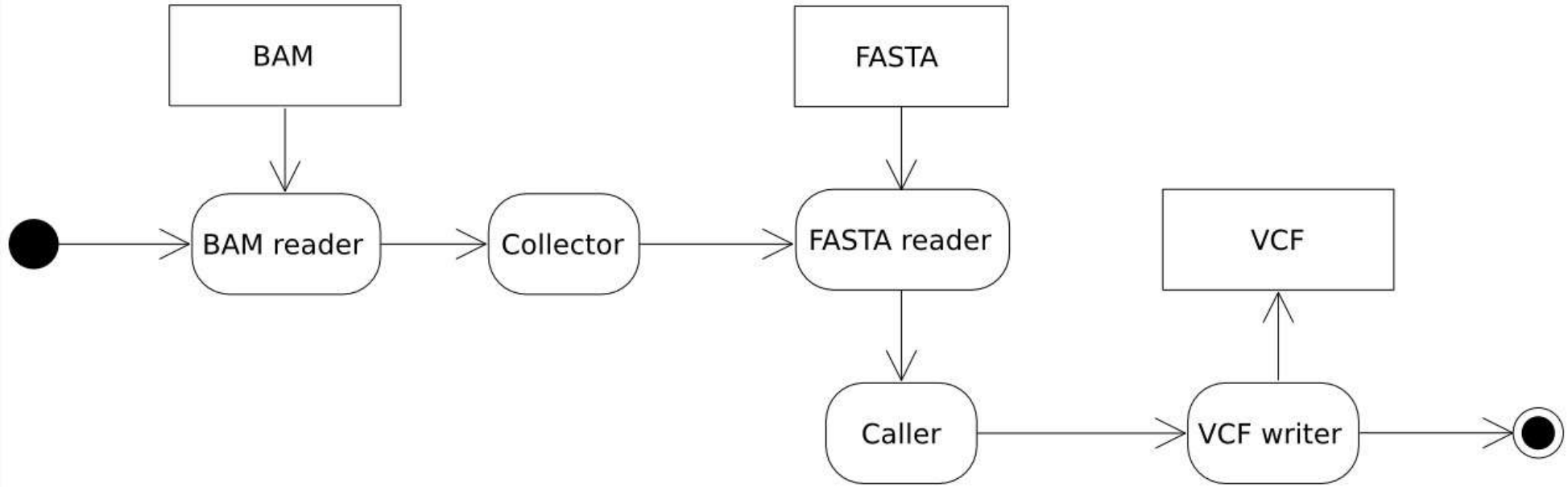
Current variant calling

- Varscan 2 architecture

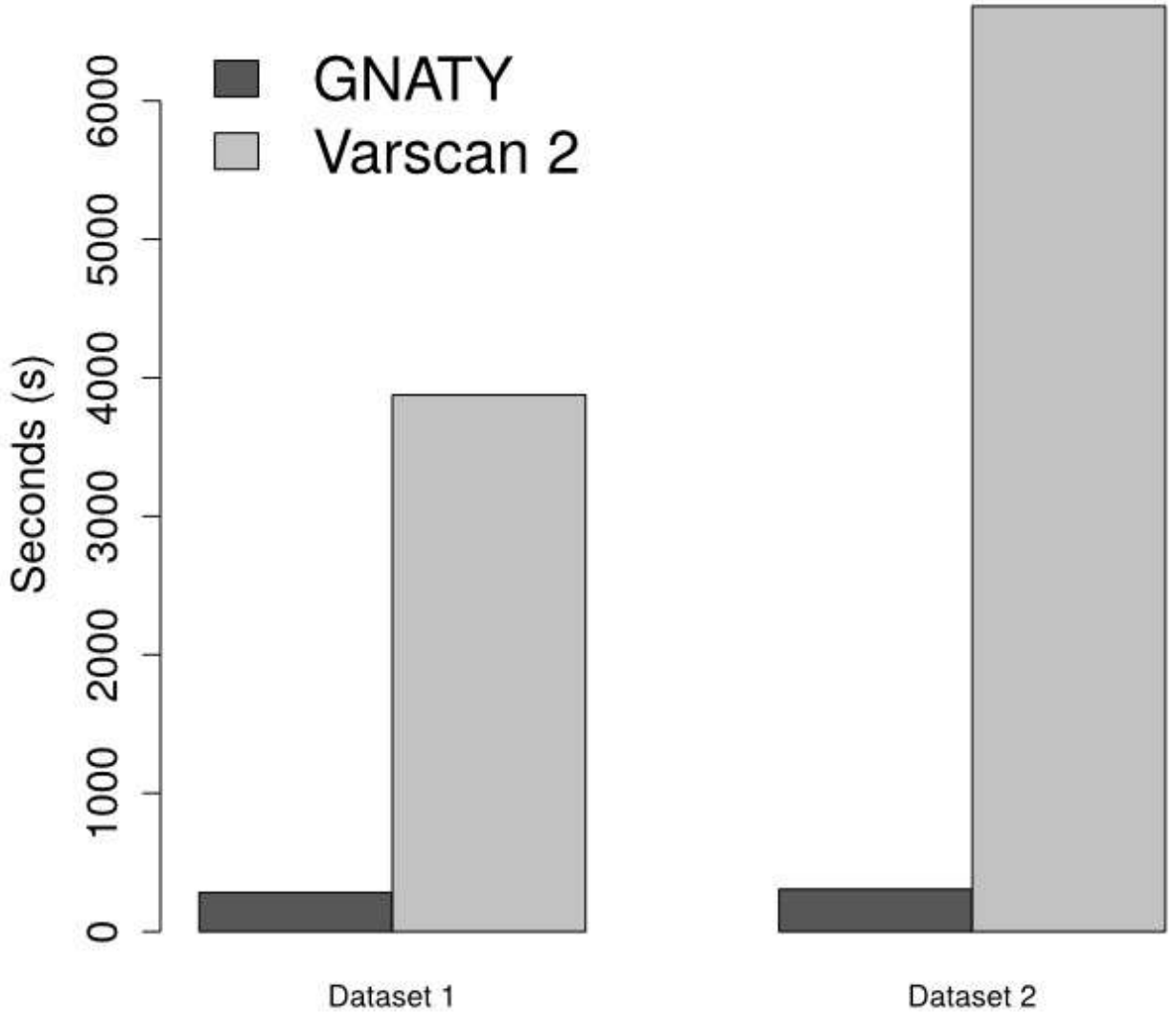


Variant calling architecture

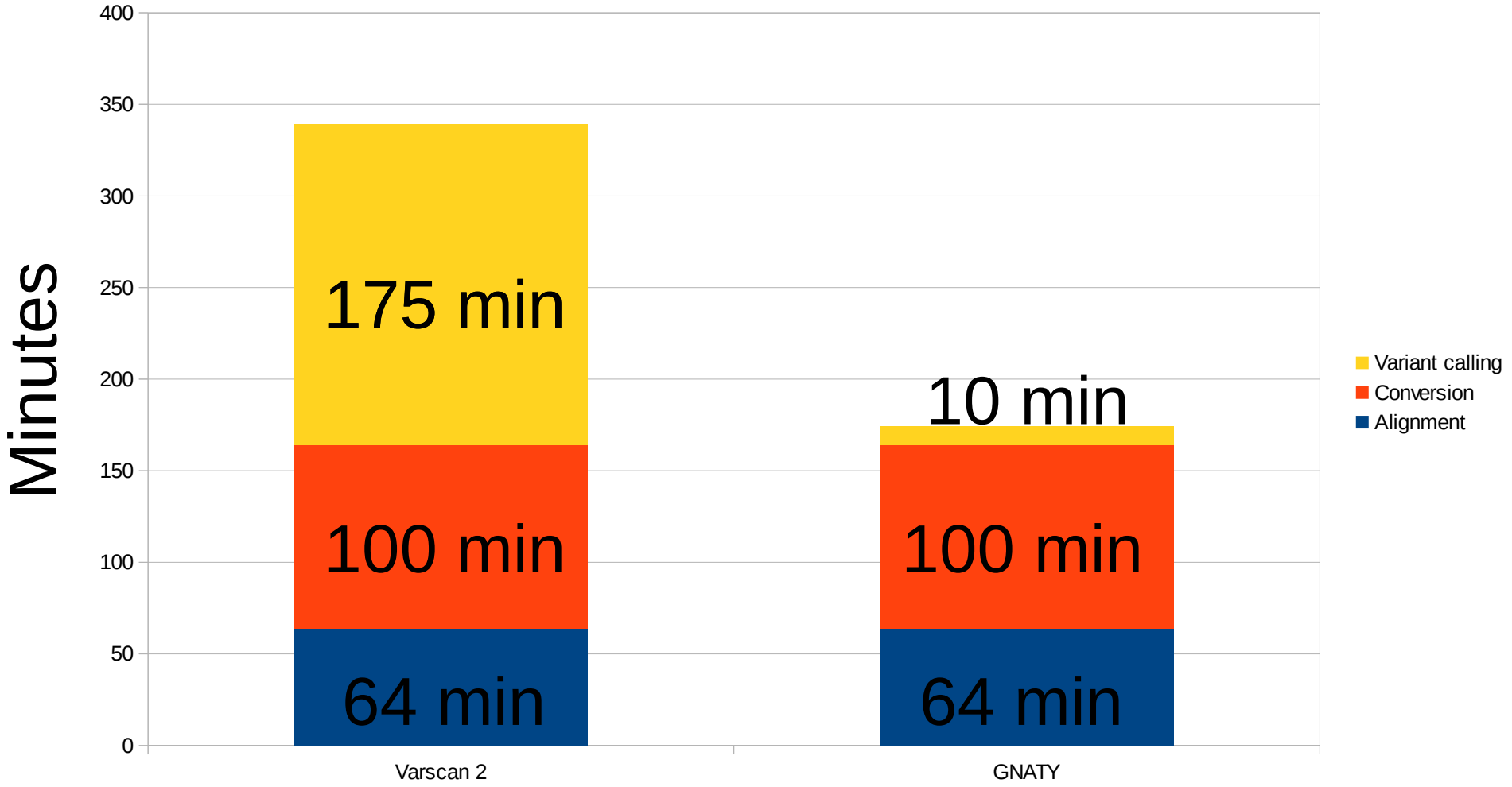
- Stream based approach



Benchmarks


GNATY
Varscan 2
Seconds (s)
0
1000
2000
3000
4000
5000
6000
Dataset 1
Dataset 2

Benchmarks


Minutes
400
350
300
250
200
150
100
50
0
175 min
100 min
64 min
10 min
100 min
64 min
Varscan 2
GNATY
Variant calling
Conversion
Alignment

Conclusion

- Informatics are an essential part of modern genetics
- The requirements of genetics push informatics to its limits, in terms of algorithms and infrastructure
- There are many opportunities for a computer scientist to have a meaningful impact on the field

Questions?