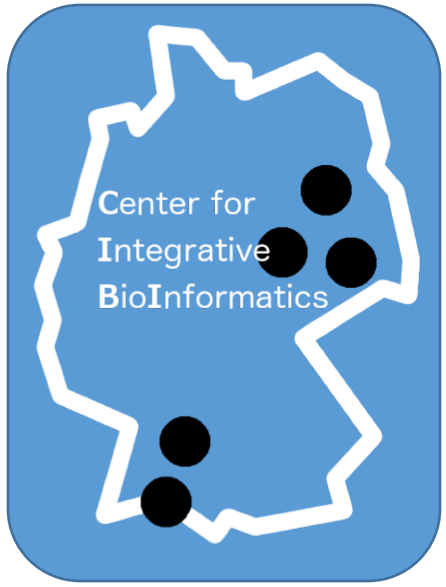


KNIME Summit 2018

OpenMS / SeqAn Workshop

René Rahn, Julianus Pfeuffer, Julian Uszkoreit, Alexander Fillbrunn

1997

Against a Whole-Genome?Shotgun

Philip Green

Genome Res. 1997 7: 410-417
Access the most recent version at doi:10.1101/gr.7.5.410

However, it is clear upon reflection that unmapped genomic reads are an extremely inefficient way to obtain biological information and are virtually useless for most purposes.

2013

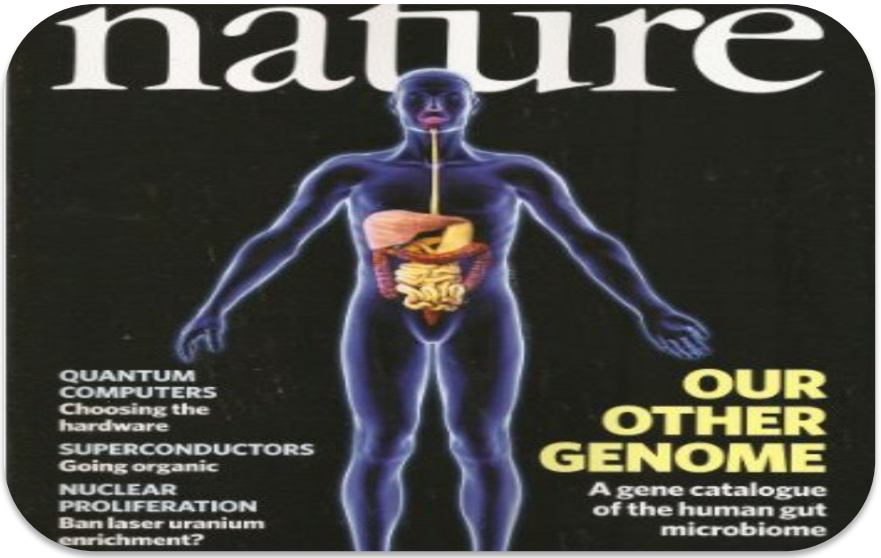
McKinsey Global Institute:
Disruptive technologies: Advances that will transform life, business, and the global economy (2013)

ECONOMIC IMPACT of NGS

In the applications we assessed, we estimate that next-generation genomics have a potential economic impact of $700 billion to $1.6 trillion per year by 2025…..

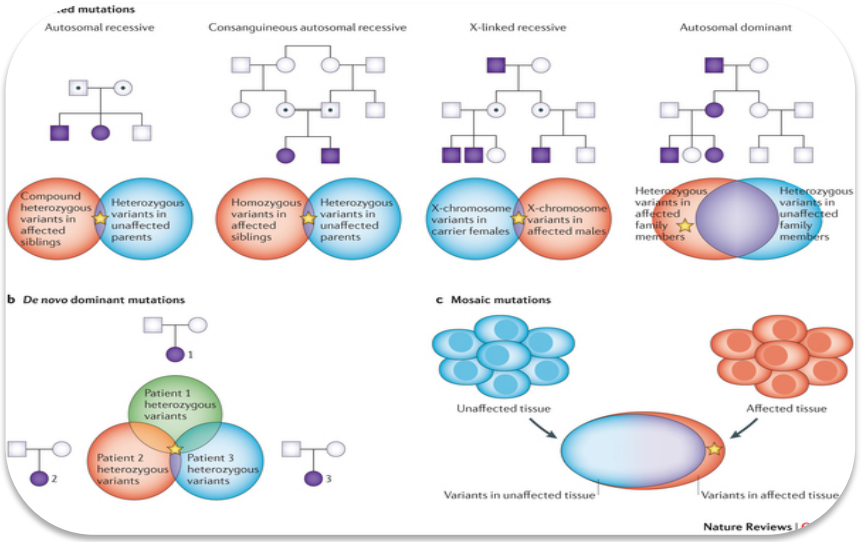
Apr

Agriculture



Metagenomics

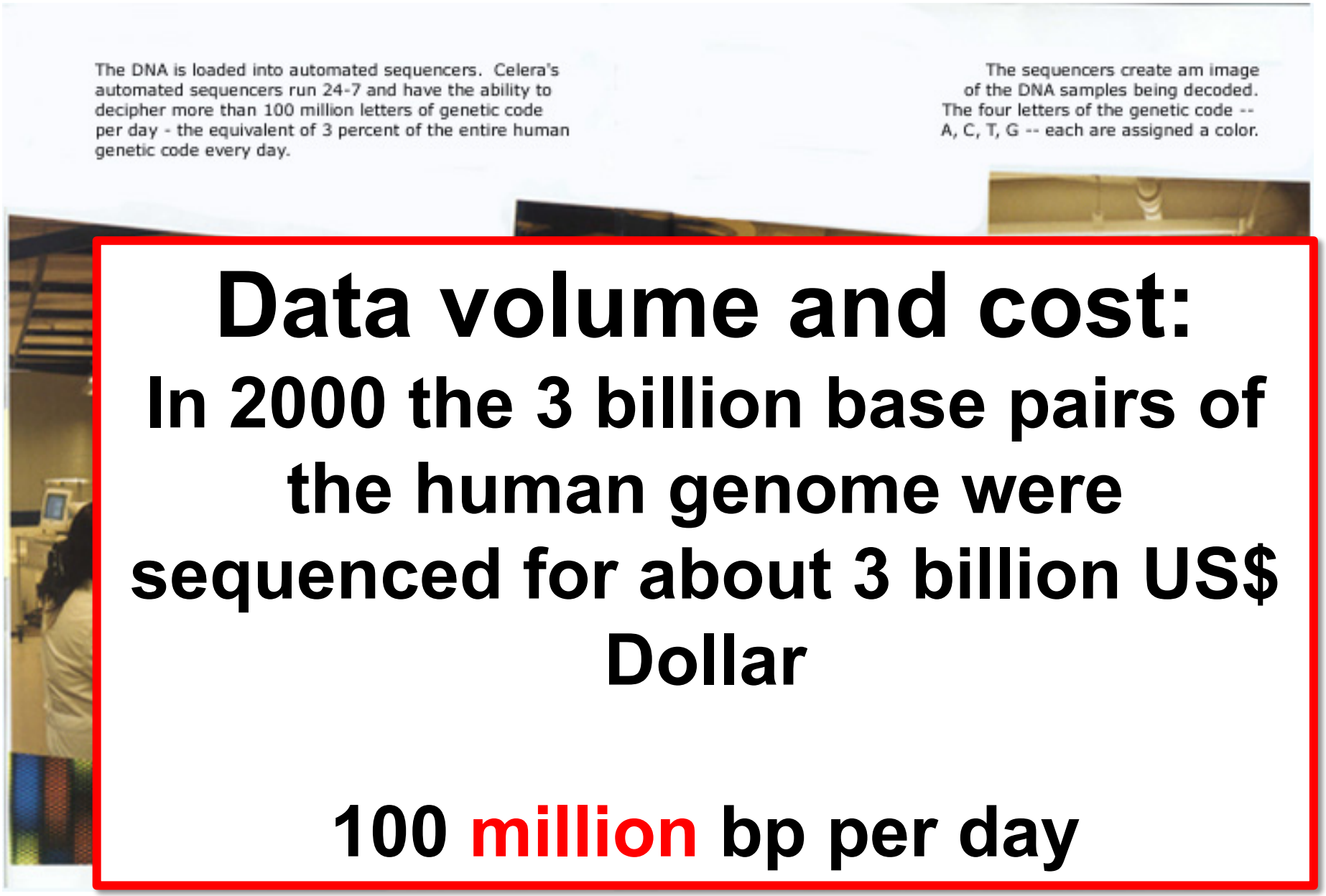
Cancer



Hereditary Diseases

Taken from:
http://www.nfcr.org/sites/default/files/images/GenomicProfiling2.jpg
http://ecx.images-amazon.com/images/I/51tztcMqlRL._SS500_.jpg

~ 13 years ago...



Data volume and cost:

In 2000 the 3 billion base pairs of the human genome were sequenced for about 3 billion US$ Dollar

100 million bp per day

Sequencing today...

Illumina HiSeq

400 billion bps per day

Within roughly ten years sequencing has become about 10 million times cheaper Pangenomics analyses possible

Sequencing earth?

Published online 23 August 2011 | Nature | doi:10.1038/news.2011.498
Corrected online: 24 August 2011

News

Number of species on Earth tagged at 8.7 million

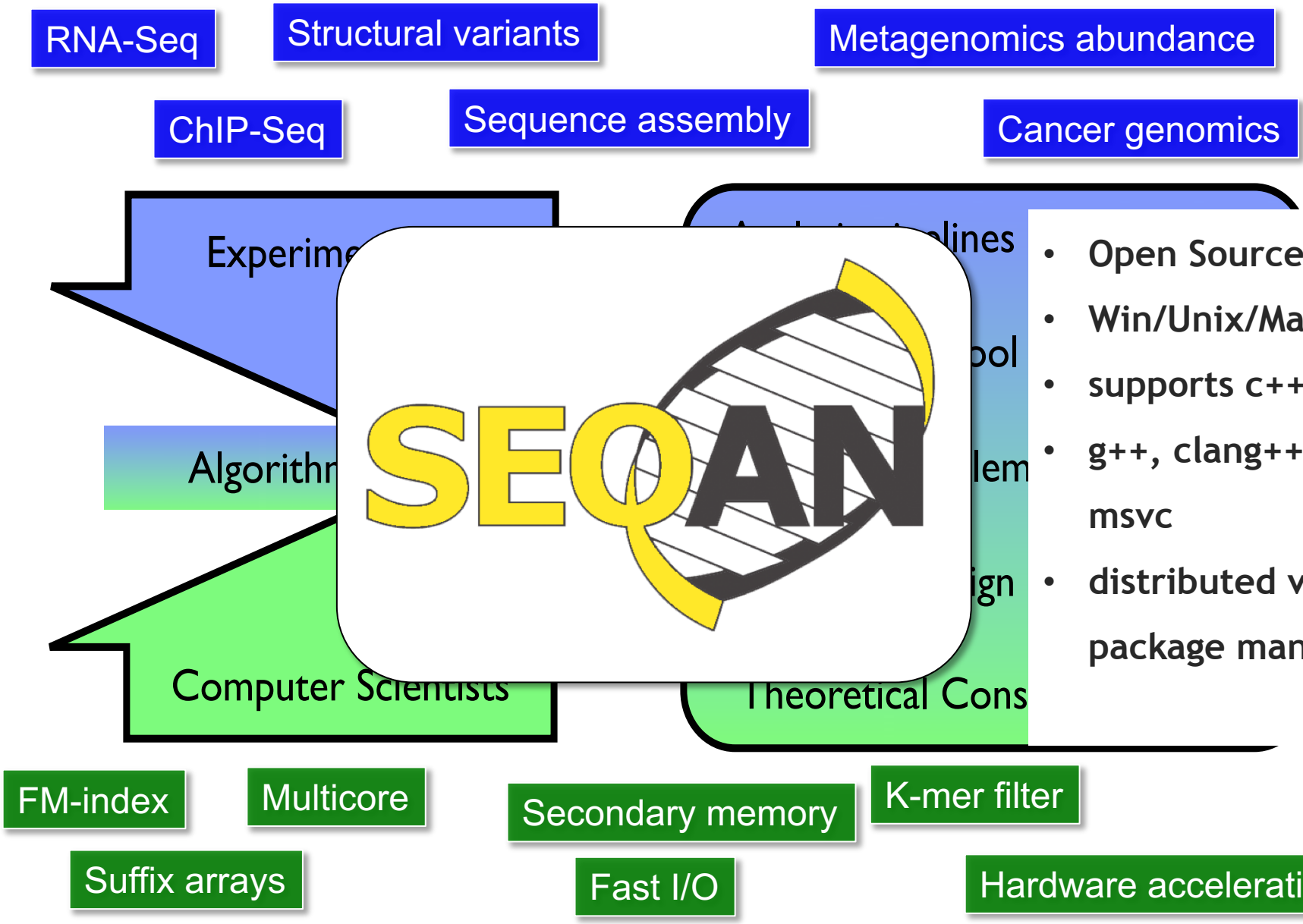
Most precise estimate yet suggests more than 80% of species still undiscovered.

Lee Sweetlove

10^7 species x 10^8 genome size =>
earth genome has 10^{15} bps

10^4 Hiseqs can each sequence 10^{11} bps per day =>
earth genome at 10x in 10 days

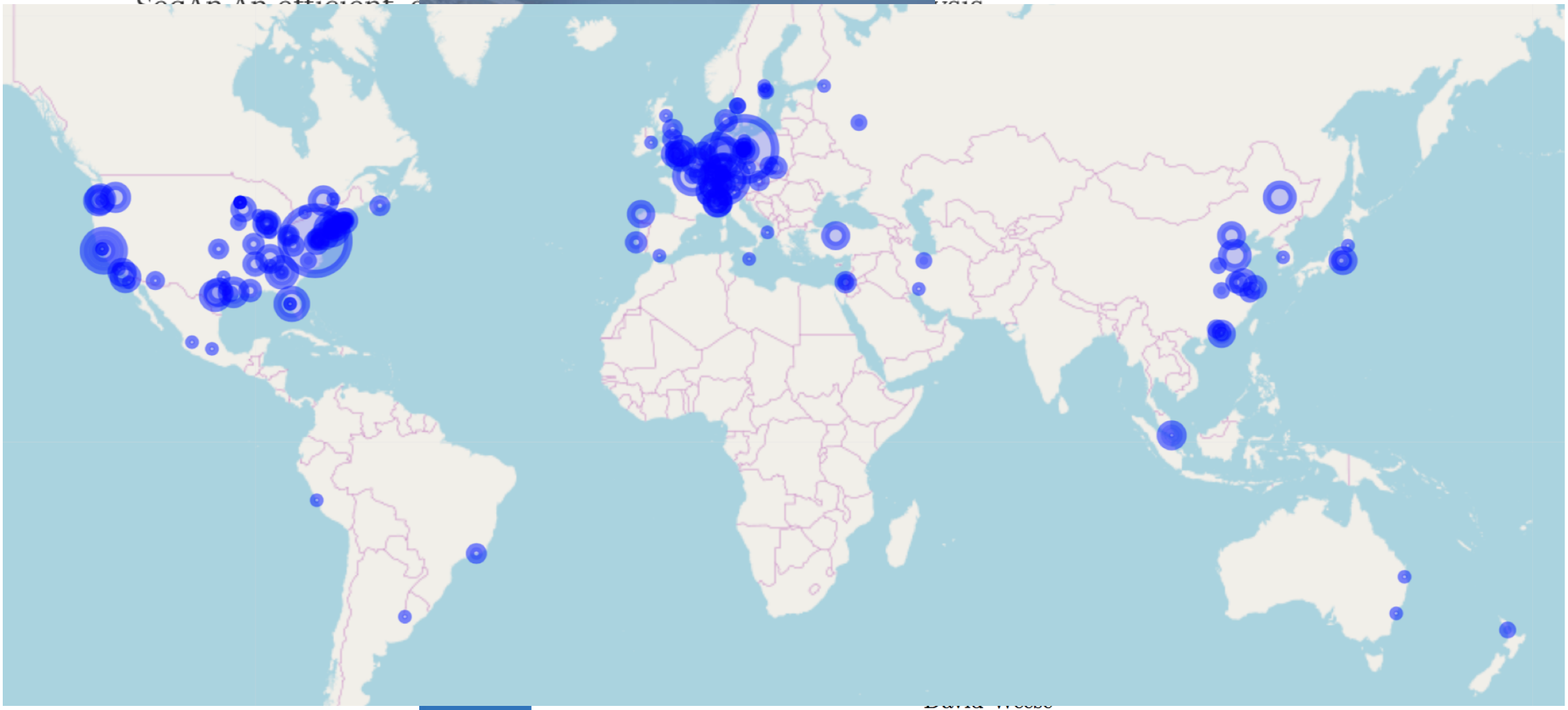
Software libraries bridge gap



- Open Source BSD-3
- Win/Unix/Mac
- supports c++11/14
- g++, clang++, icpc, msvc
- distributed via various package managers

SeqAn

SOFTWARE | OPEN ACCESS



SAP Innovation Center, Potsdam, Germany

RECOMB 2017, Hongkong

Constant time bidirectional indices

EPR-Dictionaries: A Practical and Fast Data Structure for Constant Time Searches in Unidirectional and Bidirectional FM Indices

Christopher Pockrandt[1,2(✉)], Marcel Ehrhardt[1], and Knut Reinert[1]

[1] Department of Computer Science and Mathematics, Freie Universität Berlin, Berlin, Germany
christopher.pockrandt@fu-berlin.de
[2] International Max Planck Research School for Computational Biology and Scientific Computation, Berlin, Germany
http://reinert-lab.de

Bioinformatics 2018 (accepted)

Accelerated Pairwise Alignment

Generic accelerated sequence alignment in SeqAn using vectorization and multi-threading

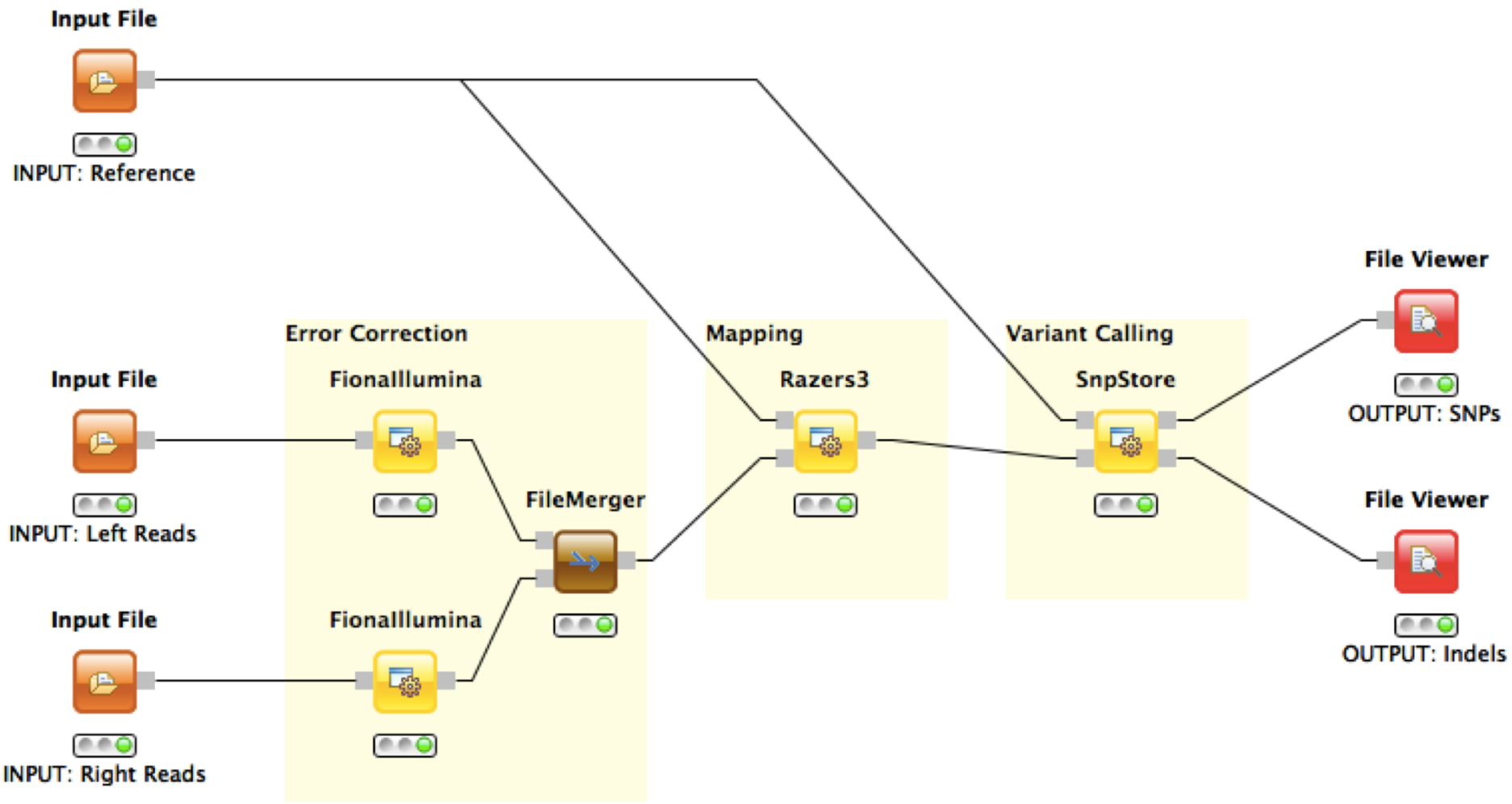
René Rahn [1,*], Stefan Budach [2], Pascal Costanza [3], Marcel Ehrhardt [1], Jonny Hancox [4] and Knut Reinert [1,2,*]

[1]Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany
[2]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany
[3]imec, Belgium and
[4]Intel Corporation (UK) Limited, United Kingdom

*Tel: +49 (0)30 838-72974; Fax: +49 (0)30 838-472974; Email: rene.rahn@fu-berlin.de

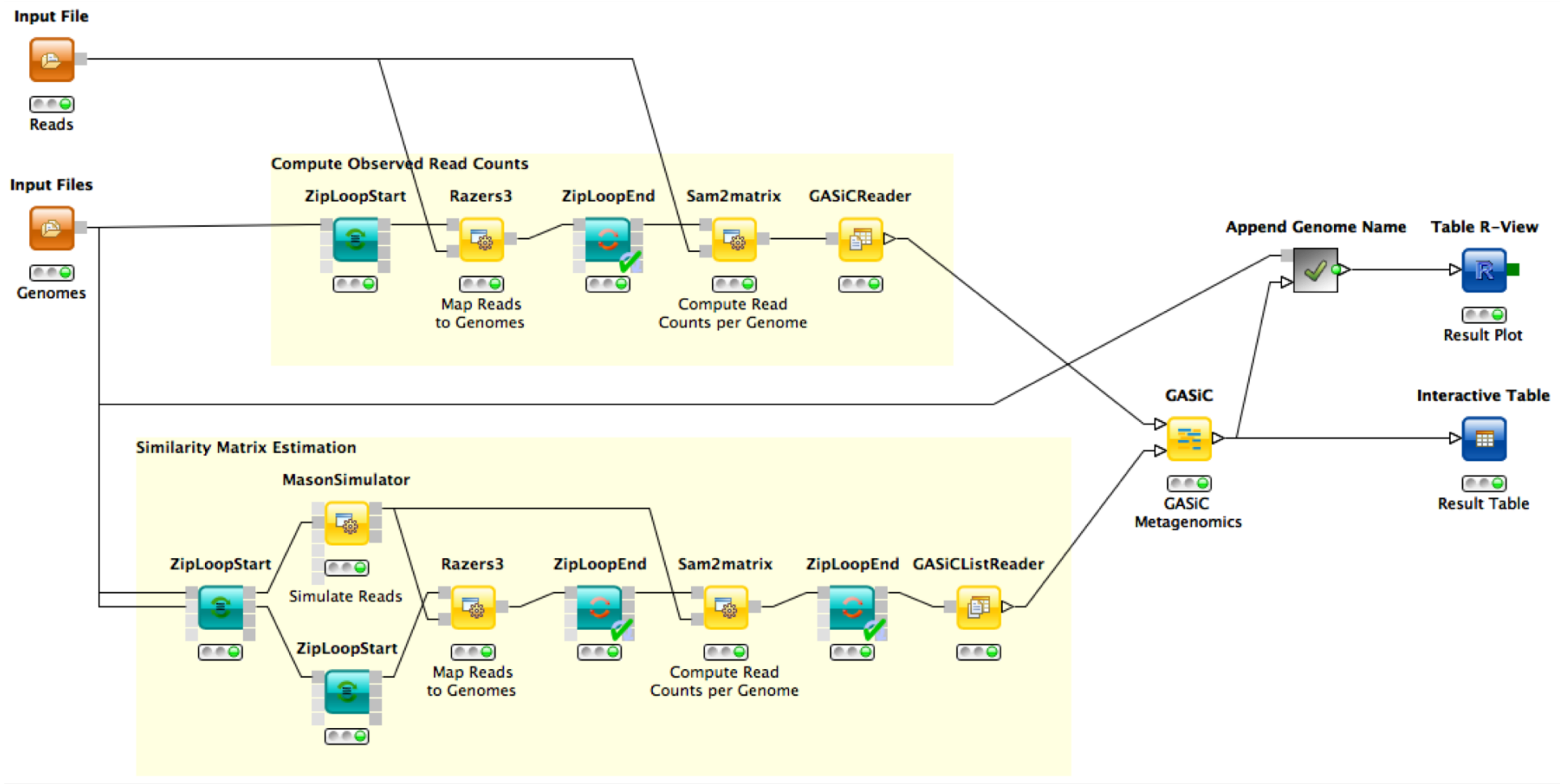
SeqAn Workflows

Variant Calling with prior Error Correction



SeqAn Workflows

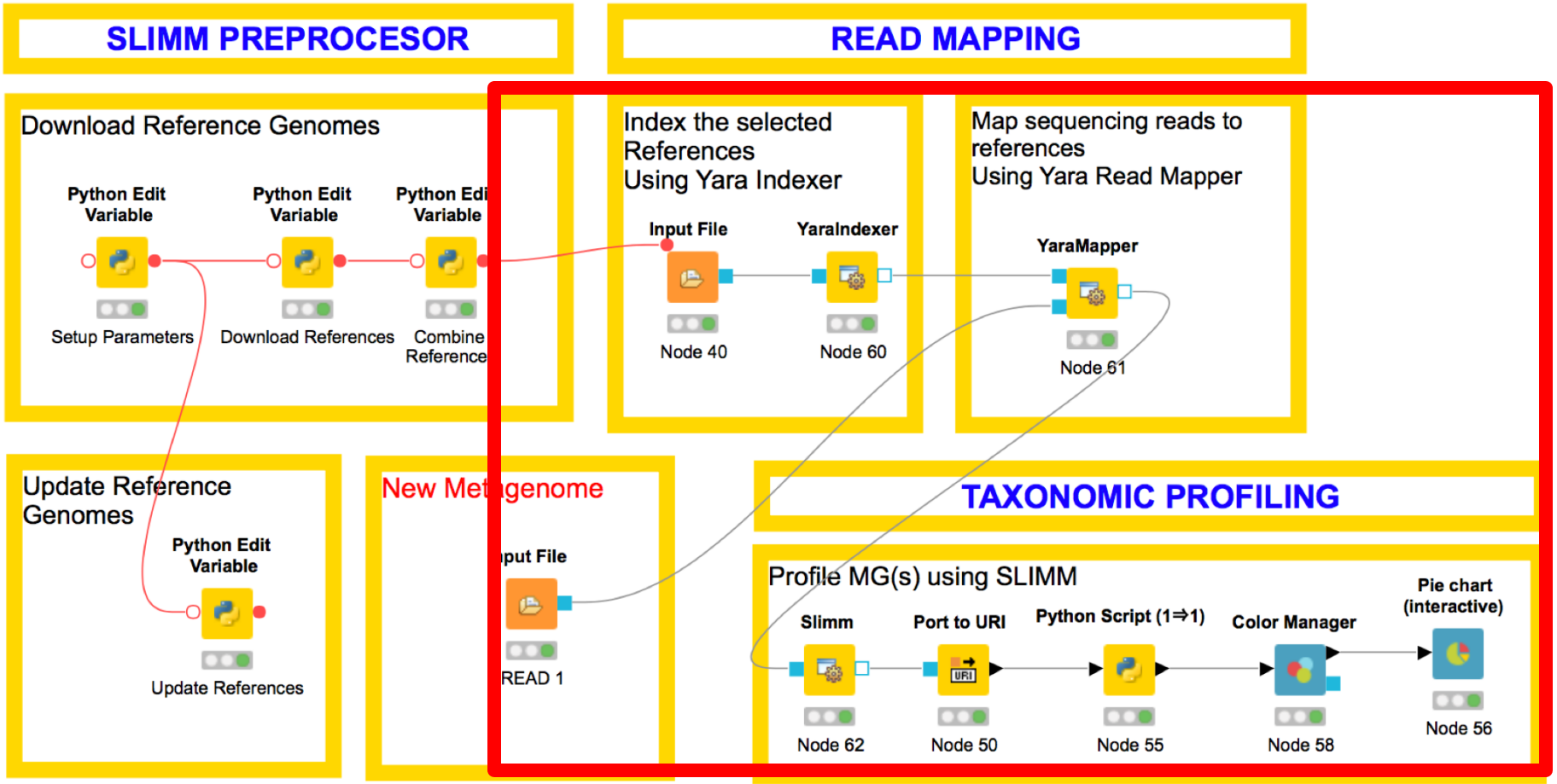
Metagenomics Workflow – GASiC*



* Lindner MS and Renard BY. *Metagenomic abundance estimation and diagnostic testing on species level*, Nucl. Acids Res. 2013, 41(1): e10, doi:10.1093/nar/gks803

SeqAn Workflows

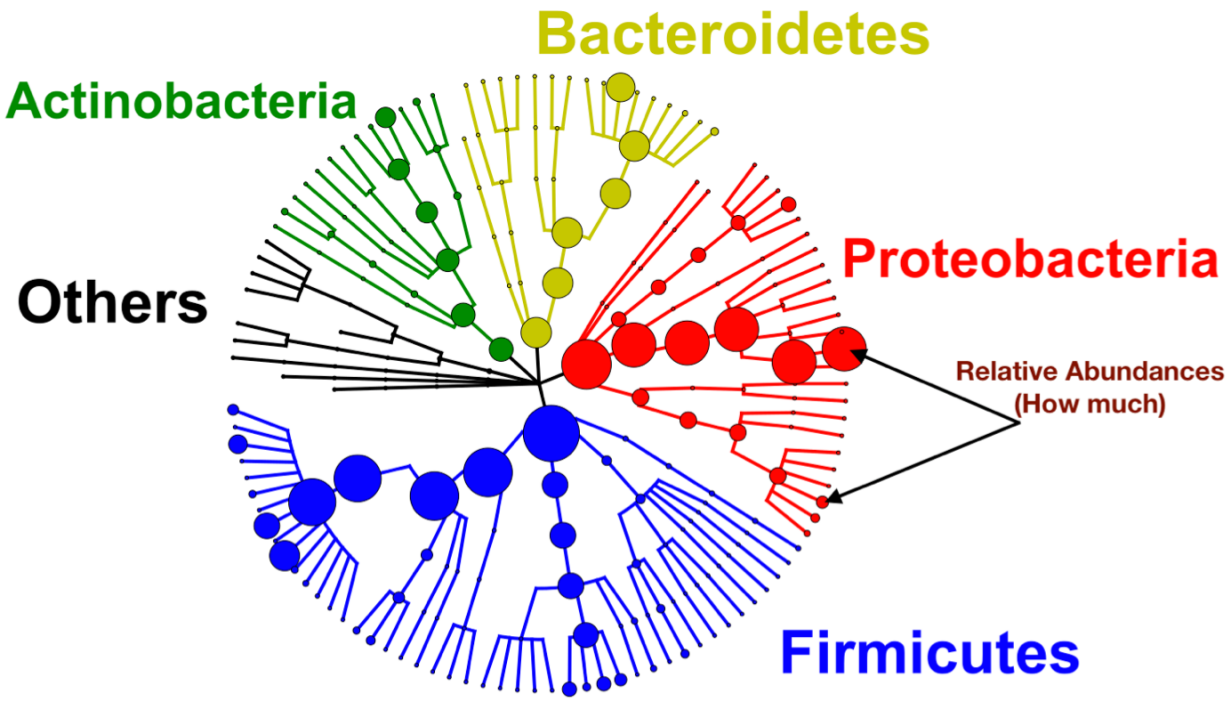
Metagenomics Workflow – SLIMM*



Dadi TH, Renard BY, Wieler LH, Semmler T, Reinert K. (2017) SLIMM: species level identification of microorganisms from metagenomes. PeerJ 5:e3138 https://doi.org/10.7717/peerj.3138

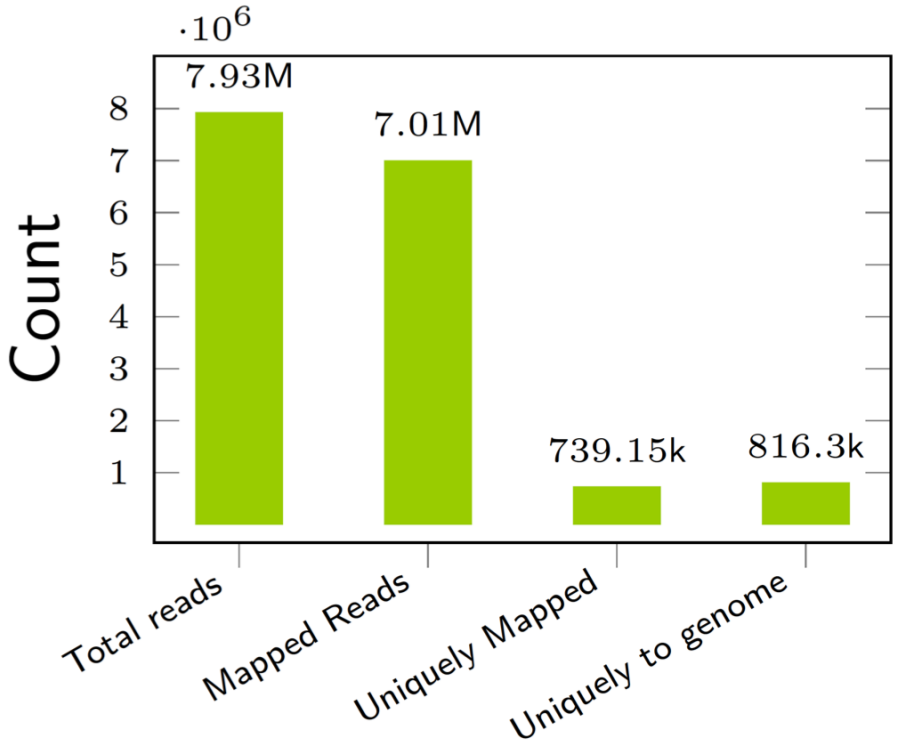
Taxonomic Profiling
(Who and how much?)

- Taxonomic profiling is a process of generating qualitative and quantitative information about a composition of a given microbial community.



Taxonomic Profiling (major challenges)

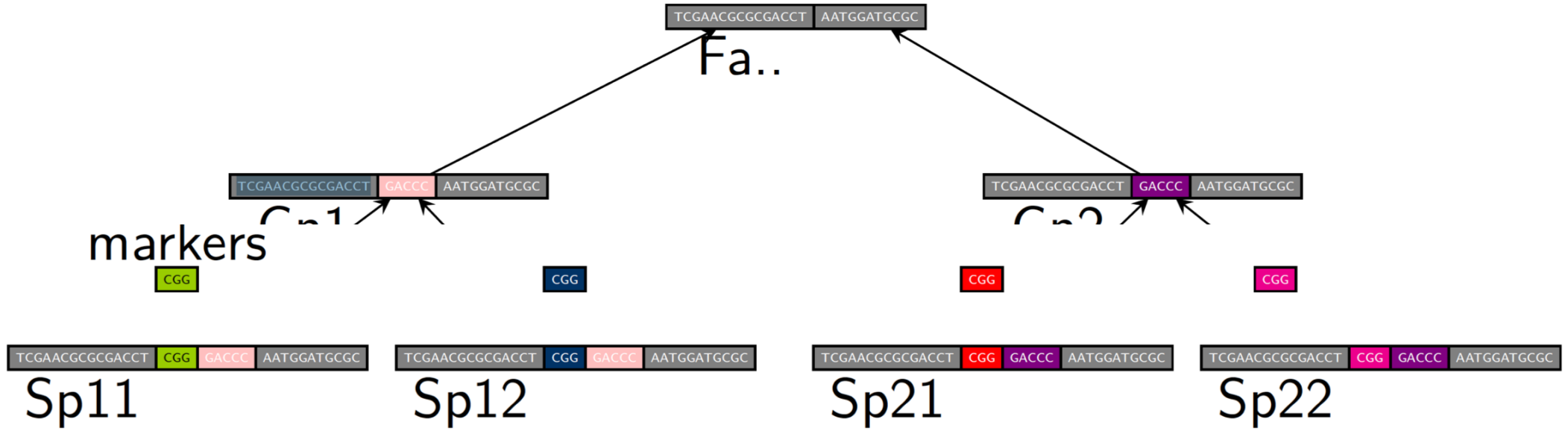
- Shared (homologous) regions of genome sequences across multiple microorganisms



- **Range of variation in the abundance of individual groups**

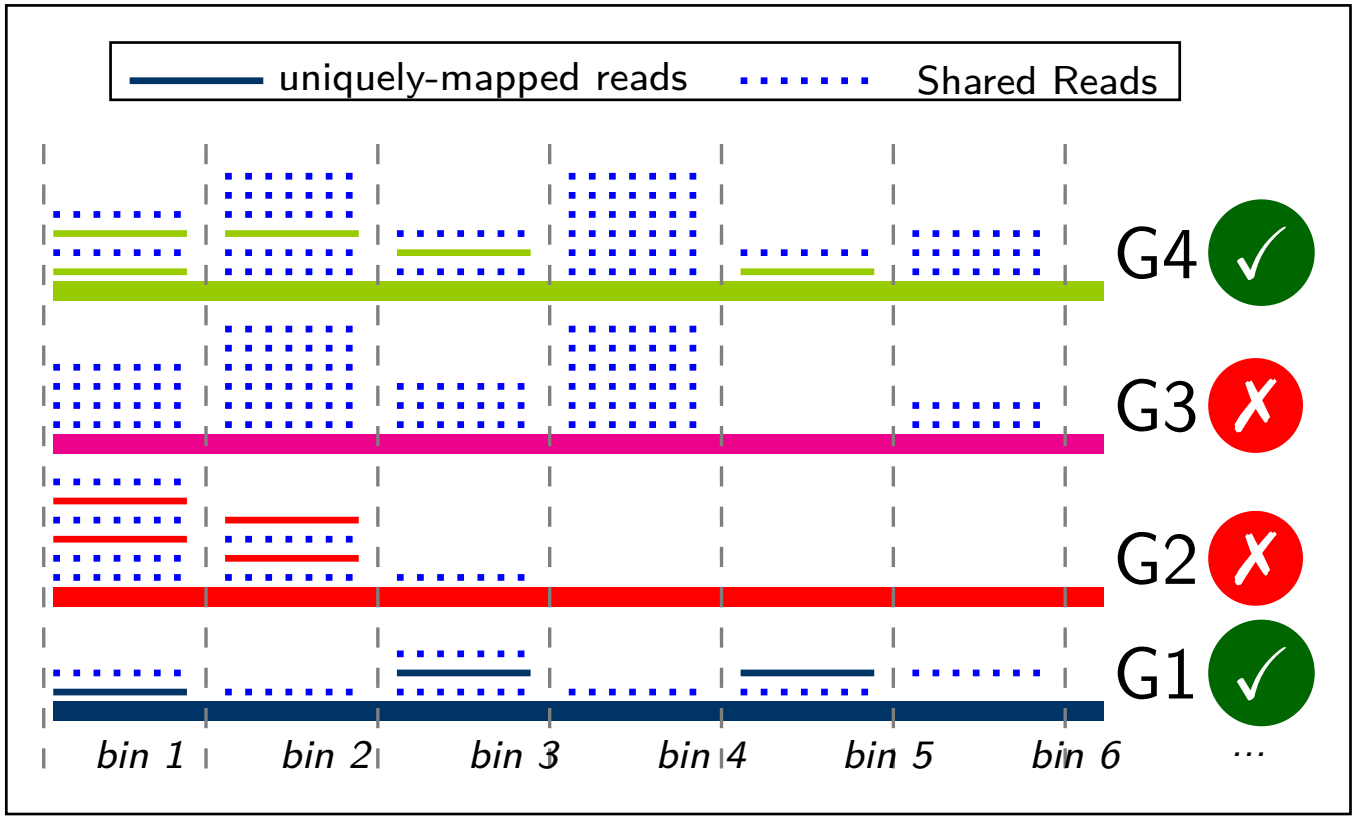
How existing methods try to resolve this ...

- Prepare non overlapping reference catalog (MetaPhlAn, GOTTCHA, mOTUs)
  - Unable to detect low abundance organisms.
- Assign shared reads to their LCA
  - Most of the information goes down to the upper levels.

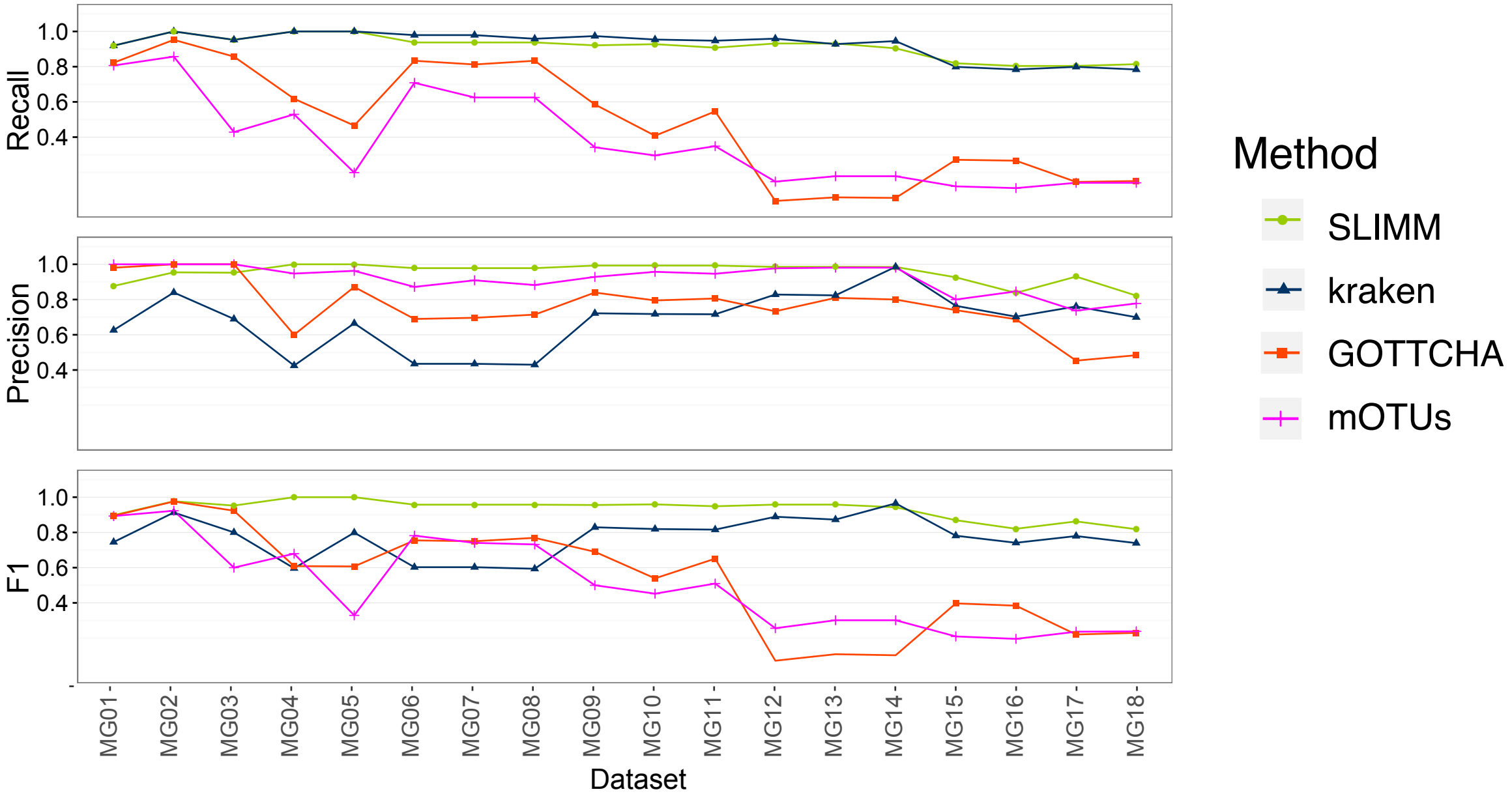


SLIMM - Method

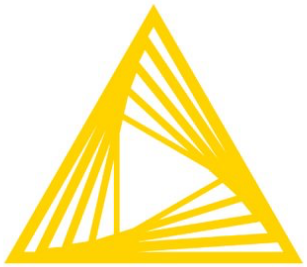
- Collect information about genomes from mapping results
- Bin reads a
  1. Shared I
  2. Uniquel
- Discard un scape using quantile ba
- Recalculat
- Assign rea es at a given rank



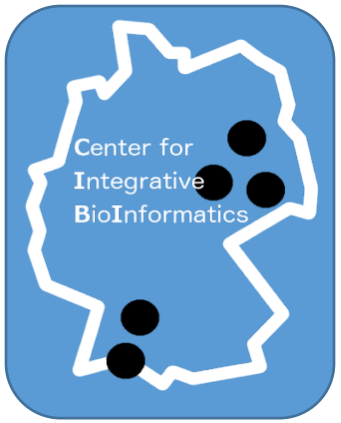
Precision, Recall and F1-Score

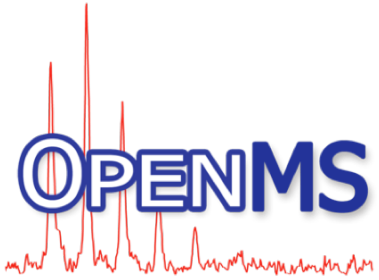


Hands On



ftp://ftp.mi.fu-berlin.de/pub/SeqAn/knime_summit/2018/





Thank you for your attention!