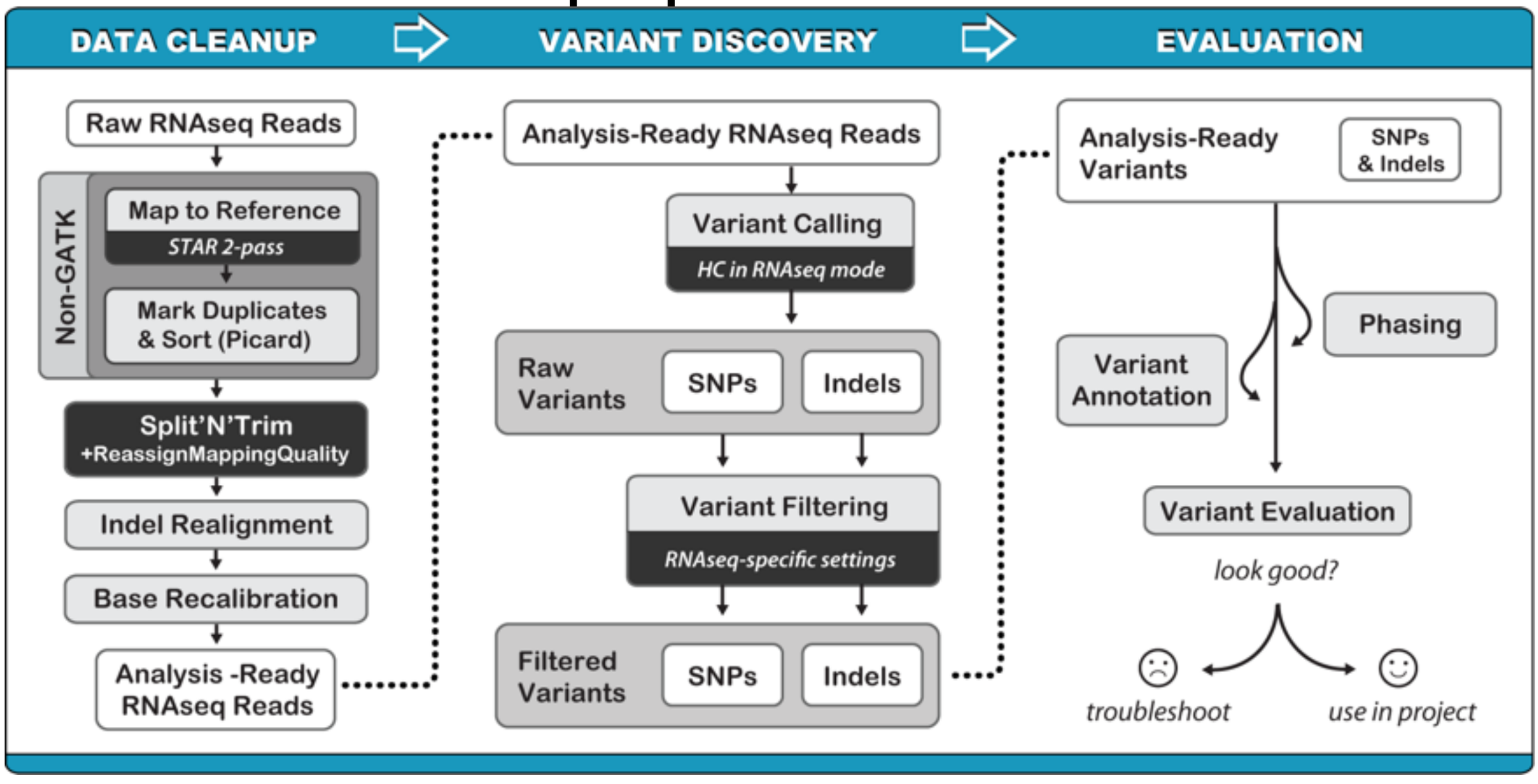
RNF: a general framework to evaluate NGS read mappers

Karel Břinda, Valentina Boeva, and Gregory Kucherov

International Workshop Algorithmics, Bioinformatics and Statistics for NGS data analysis
June 23, 2015

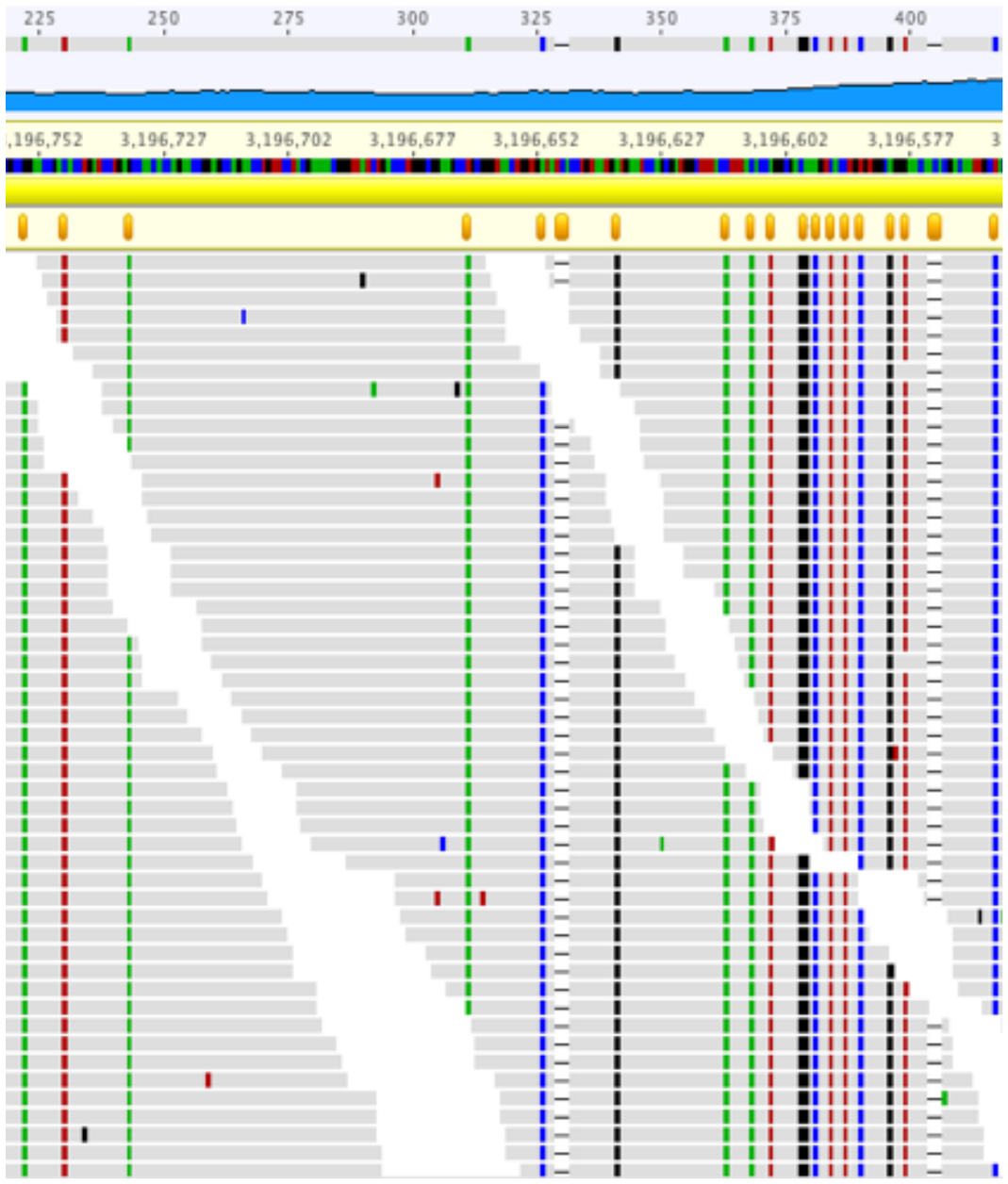
Example of a genomic pipeline



Source: GATK Guide, www.broadinstitute.org

Read mapping

- Critically affects results in genomic pipelines
- Various indexing schemes (FM-index vs. hash tables)
- Tradeoff between sensitivity vs. time and memory
- Modes of mapping: best mapping vs. all mapping
- For every read (or segment), the mapper reports:
  - coordinates in the reference sequence
  - edit operations (via CIGAR strings)
  - alignment score
  - mapping quality **(!!)** :
    MAPQ = −10 log10 Pr{mapping is wrong}



Read mapping

- **Very extensively studied task**

…new mappers published during last 10 days:

Hardware-Acceleration of Short-read Alignment Based on the Burrows-Wheeler Transform
H Waidyasooriya, M Hariyama - ieeexplore.ieee.org
Abstract—The alignment of millions of short DNA fragments to a large genome is a very important aspect of the modern computational biology. However, software-based DNA sequence alignment takes many hours to complete. This paper proposes an FPGA-based ...

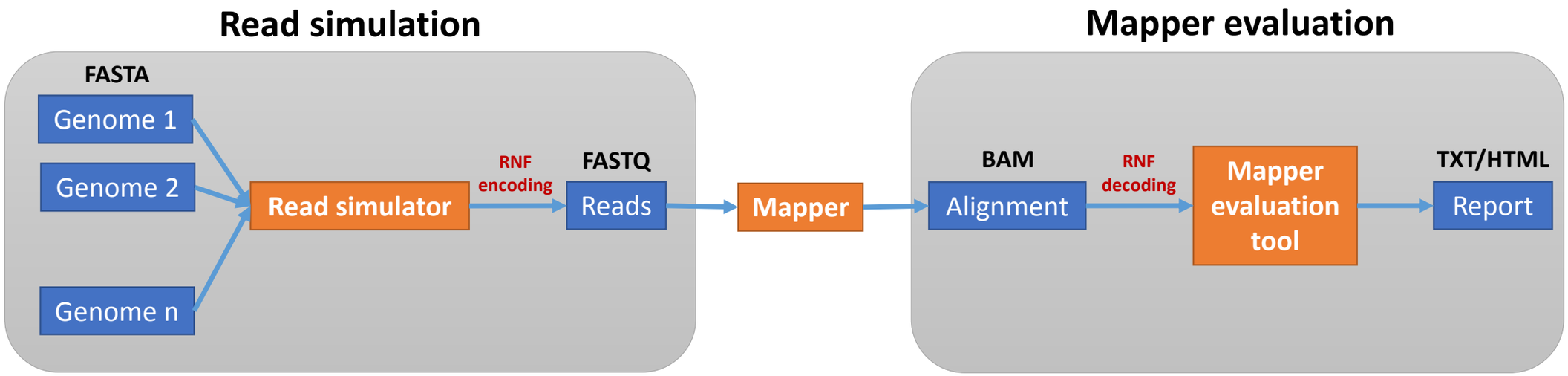
[PDF] **Fast** and **sensitive mapping** of **error-prone nanopore sequencing reads** with **GraphMap**
I Sovic, M Sikic, A Wilm, SN Fenlon, S Chen… - bioRxiv, 2015 - biorxiv.org
Abstract Exploiting the power of nanopore sequencing requires the development of new bioinformatics approaches to deal with its specific error characteristics. We present the first nanopore read mapper (GraphMap) that uses a read-funneling paradigm to robustly ...

[HTML] **BitMapper**: an **efficient all-mapper based** on **bit-vector computing**
H Cheng, H Jiang, J Yang, Y Xu, Y Shang - BMC bioinformatics, 2015 - biomedcentral.com
Background As the next-generation sequencing (NGS) technologies producing hundreds of millions of reads every day, a tremendous computational challenge is to map NGS reads to a given reference genome efficiently. However, existing methods of all-mappers, which ...

How to show that some method is better than another one?

Evaluation of mappers

- **Direct approach** (with simulated reads):



- **Indirect approach** (with real reads and subsequent validation using Sanger sequencing):

"Did we detect more true variants?"

Situation

- Many read simulators exist:

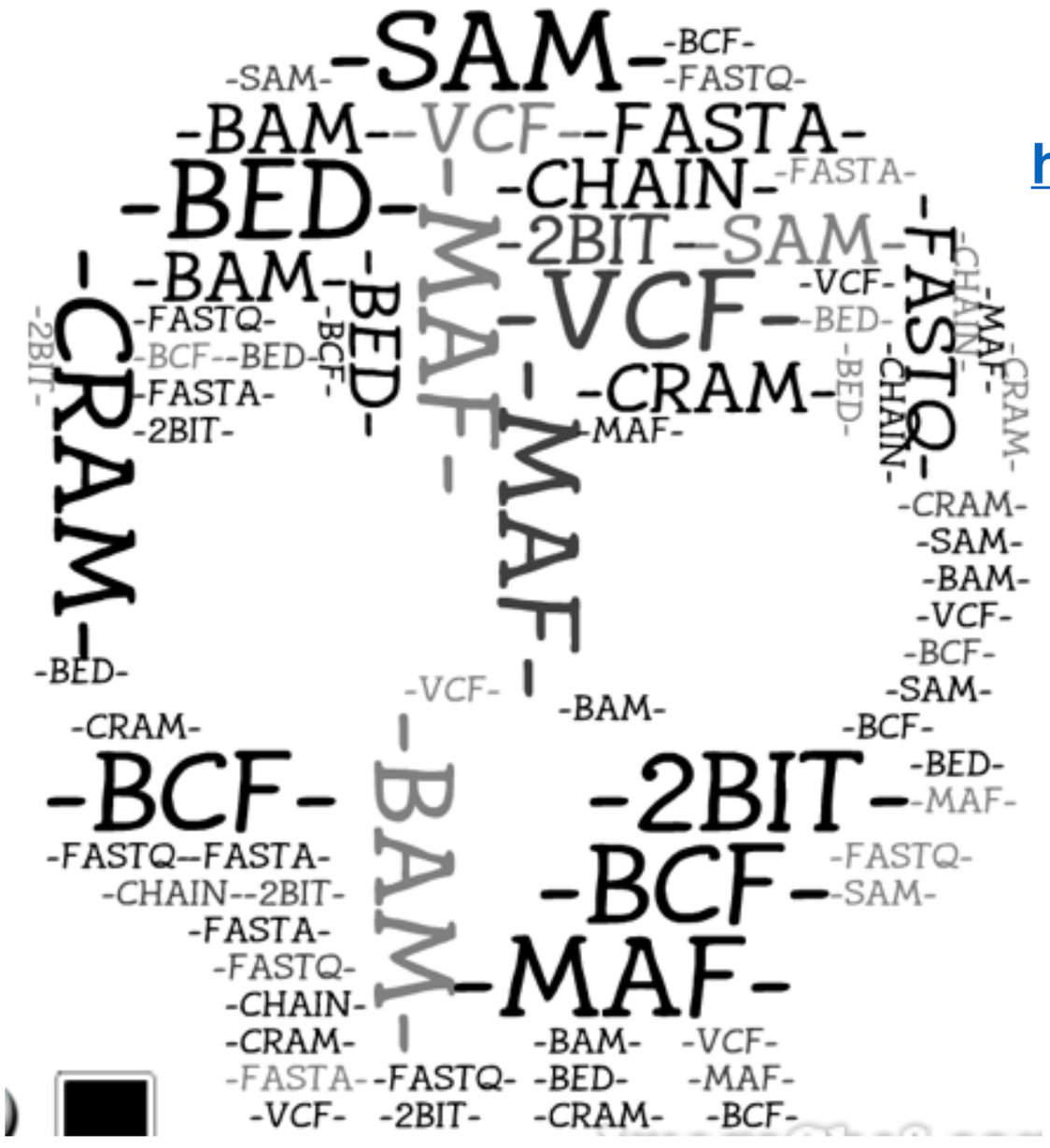
  Art, ***CuReSim***, *DNemulator,* ***DwgSim***, *FastqSim, FlowSim, GemSim,* ***Mason***, *MetaSim, PbSim, Pirs, Sherman, SimNgs, SimSeq, SInC, Wessim,* ***WgSim***, *XS, …*

- Each simulator uses own encoding of the original positions of reads in the genome

- Evaluation tools had to be explicitly compatible with a used read simulator

  CuReSim_eval, DwgSim_eval, Seq-Suite, WgSim_eval, …

Formats for NGS data



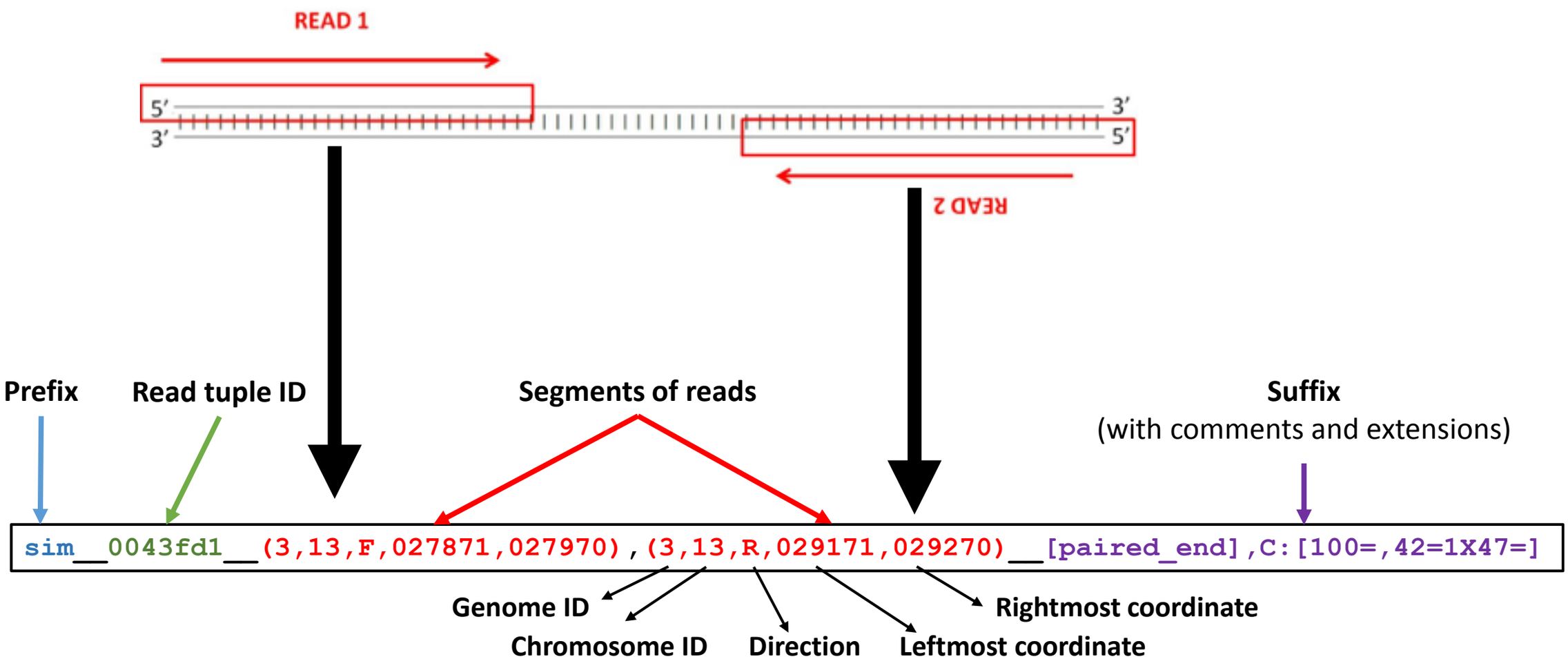
UCSC
https://genome.ucsc.edu/FAQ/FAQformat.html

HTS formats specifications
http://samtools.github.io/hts-specs/

Global Alliance for Genomics and Health
Genomics Data Working Group
File Format Tasks team
http://ga4gh.org/#/fileformats-team

Read Naming Format

Read tuple > **Read** > **Segment**



RNFtools

- **Associated software package** for RNF: http://karel-brinda.github.io/rnftools/

- Written using **SnakeMake** (Python-based Make-like software for scientific workflows, [Köster&Sven, Bioinformatics, 2012])

- All used programs are **automatically compiled** when they are requested

- Available on **PyPI** - installation:
  pip install rnftools

- Besides SnakeMake layer, also a command line layer with basic functionality is present

- **Components:**

  - **MIShmash:** a tool for simulating reads in RNF (calling existing simulators followed by read name transformation)

  - **LAVEnder:** a tool for evaluation of mappers using reads in RNF

RNFtools — MIShmash
(simulation of reads)

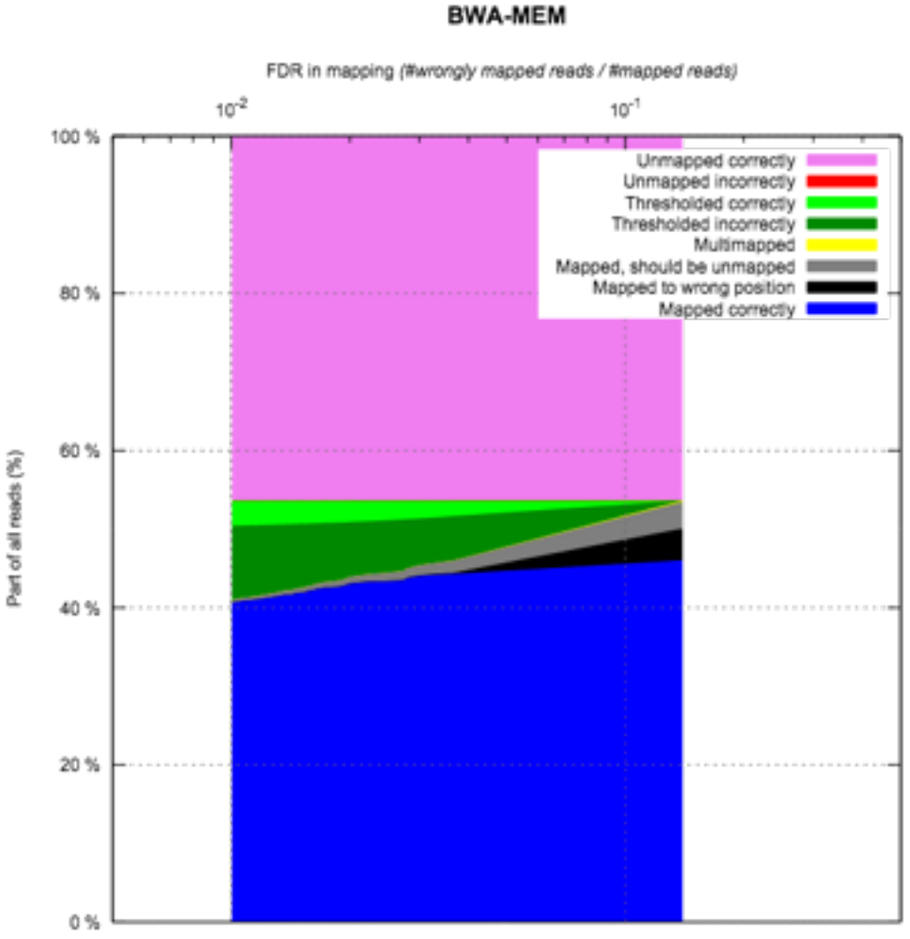
- **Simulating reads** by calling existing simulators + converting read names to RNF

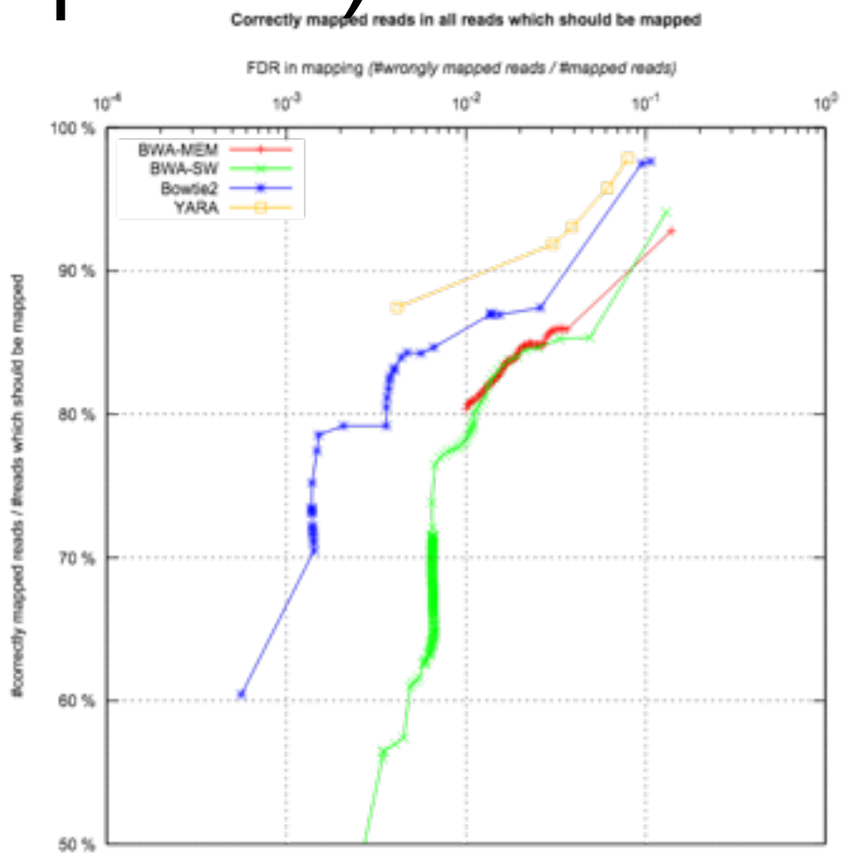
- **Supported simulators:** Art, CuReSim, Mason, DwgSim, WgSim

- Easy switching between them and combining reads from different genomes (e.g., for metagenomic simulations)

```
import rnftools
import smbl

rnftools.mishmash.sample("simple_example",reads_in_tuple=2)

rnftools.mishmash.ArtIllumina(
        fasta=smbl.fasta.EXAMPLE_1,
        number_of_read_tuples=10000,
        read_length_1=100,
        read_length_2=100,
)

include: rnftools.include()
rule: input: rnftools.input()
```

RNFtools — LAVEnder
(evaluation of mappers)





BWA-MEM [Back to main report]

Information about program - ROC table - Graphs

Information about program [Top of this page] [Back to main report]

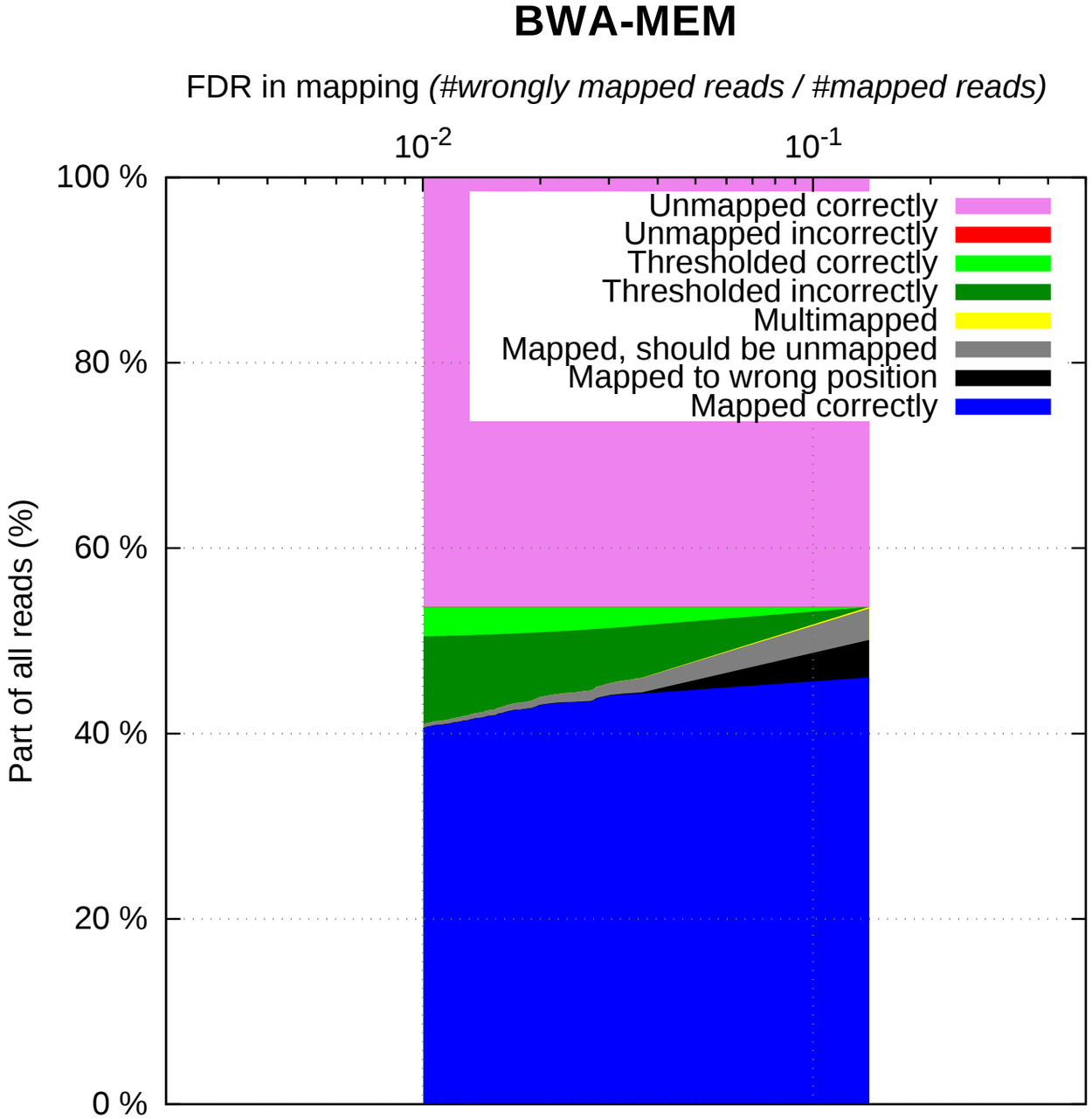
ID:	bwa
PN:	bwa
VN:	0.7.12-r1044
CL:	/Users/karel/.senbi/bin/bwa mem -t 1 /Users/karel/.senbi/fa/hg38.fa reads_1.fq

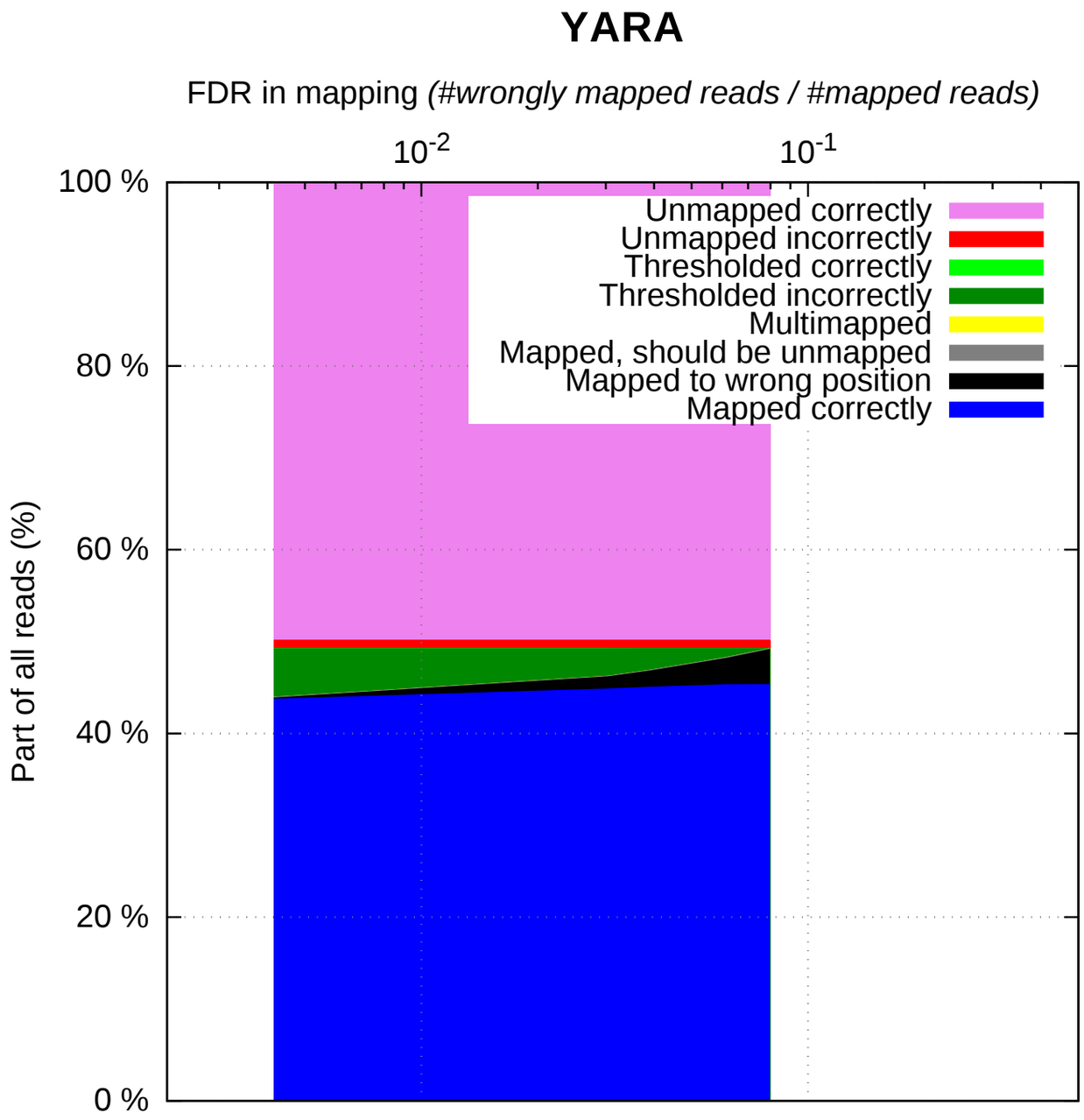
ROC table [Top of this page] [Back to main report]

M: **mapped correctly** (all segments of tuple are mapped once and correctly), **w**: **mapped to wrong position** (at least one segment was mapped to a wrong position), **m**: **mapped but should be unmapped** (at least one segment was mapped but the read should not be mapped), **P**: **multimapped** (read should be mapped but it at least one segment was mapped several times and all segments were mapped correctly at least once), **U**: **unmapped correctly** (all segments of the read were correctly marked as unmapped), **u**: **unmapped but should be mapped** (at least one segment was mapped but entire read should be unmapped), **T**: **thresholded correctly** (read shoud not be mapped), **t**: **thresholded incorrectly** (read should be mapped), **x**: **unknown** (read is probably not reported by mapper)

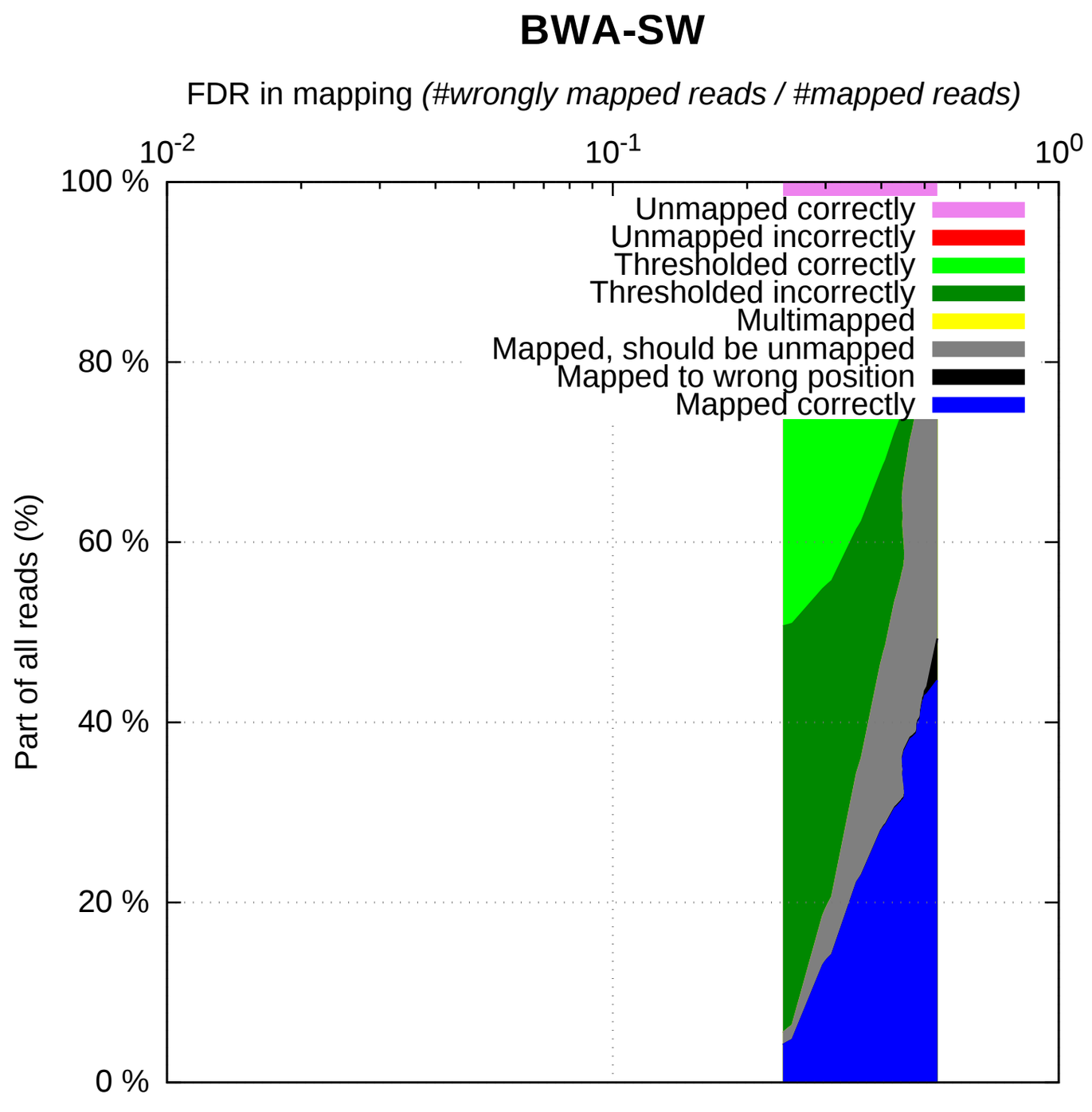
q	mapped	%	M	%	w	%	m	%	P	%	unmapped	%	U	%	u	%	T	%	t	%	x	%	sum	prec. (%)
0	107085	53.54	91878	45.94	8084	4.04	6766	3.38	357	0.18	92915	46.46	92896	46.45	19	0.01	0	0.00	0	0.00	0	0.00	200000	85.799
1	91774	45.89	88388	44.19	252	0.13	3086	1.54	48	0.02	108226	54.11	92896	46.45	19	0.01	3970	1.99	11341	5.67	0	0.00	200000	96.311
2	91564	45.78	88329	44.16	235	0.12	2961	1.48	39	0.02	108436	54.22	92896	46.45	19	0.01	4104	2.05	11417	5.71	0	0.00	200000	96.467
3	91372	45.69	88248	44.12	225	0.11	2868	1.43	31	0.02	108628	54.31	92896	46.45	19	0.01	4205	2.10	11508	5.75	0	0.00	200000	96.581
4	91182	45.59	88176	44.09	209	0.10	2769	1.38	28	0.01	108818	54.41	92896	46.45	19	0.01	4307	2.15	11596	5.80	0	0.00	200000	96.703
5	90955	45.48	88070	44.03	187	0.09	2675	1.34	23	0.01	109045	54.52	92896	46.45	19	0.01	4406	2.20	11724	5.86	0	0.00	200000	96.828
6	90714	45.36	87941	43.97	169	0.08	2587	1.29	17	0.01	109286	54.64	92896	46.45	19	0.01	4500	2.25	11871	5.94	0	0.00	200000	96.943
7	90458	45.23	87774	43.89	158	0.08	2512	1.26	14	0.01	109542	54.77	92896	46.45	19	0.01	4578	2.29	12049	6.02	0	0.00	200000	97.033
8	90138	45.07	87541	43.77	141	0.07	2444	1.22	12	0.01	109862	54.93	92896	46.45	19	0.01	4648	2.32	12299	6.15	0	0.00	200000	97.119
9	89810	44.91	87295	43.65	126	0.06	2379	1.19	10	0.01	110190	55.09	92896	46.45	19	0.01	4715	2.36	12560	6.28	0	0.00	200000	97.200
10	89275	44.64	86820	43.41	123	0.06	2323	1.16	9	0.00	110725	55.36	92896	46.45	19	0.01	4772	2.39	13038	6.52	0	0.00	200000	97.250
11	89054	44.53	86668	43.33	115	0.06	2262	1.13	9	0.00	110946	55.47	92896	46.45	19	0.01	4833	2.42	13198	6.60	0	0.00	200000	97.321
12	88950	44.48	86625	43.31	114	0.06	2203	1.10	8	0.00	111050	55.52	92896	46.45	19	0.01	4893	2.45	13242	6.62	0	0.00	200000	97.386
13	88859	44.43	86584	43.29	114	0.06	2154	1.08	7	0.00	111141	55.57	92896	46.45	19	0.01	4943	2.47	13283	6.64	0	0.00	200000	97.440
14	88733	44.37	86516	43.26	111	0.06	2101	1.05	5	0.00	111267	55.63	92896	46.45	19	0.01	4998	2.50	13354	6.68	0	0.00	200000	97.501

Contamination test 1
(human & mouse)





Contamination test 2
(human & chimpanzee)



Future perspective

- The software is ready to use, you can use it in you projects
  http://karel-brinda.github.io/rnftools/
  http://rnftools.readthedocs.org
- Support of RNF in software
  - Plugging RNF directly into existing read simulators
  - RNF-based evaluators
- Extend the concept of RNF from mappers to taxonomic sequence classifiers

Thank you for your attention

Valentina Boeva

Gregory Kucherov

Publication: K.B., V.B., G.K. RNF: a general framework to evaluate NGS read mappers, *arXiv*:1504.00556 [q-bio.GN], 2015.
Web: http://karel-brinda.github.io/rnftools/