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](http://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 (!!):
    \[ \text{MAPQ} = -10 \log_{10} \Pr\{\text{mapping is wrong}\} \]
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 - ieeeexplore.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 - bioxiv.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):

  - **Read simulation**
    - FASTA
      - Genome 1
      - Genome 2
      - Genome n
    - Read simulator
      - RNF encoding
      - FASTQ
      - Reads
    - Mapper

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

  - **Mapper evaluation**
    - BAM
      - RNF decoding
      - Alignment
      - Mapper evaluation tool
      - TXT/HTML
        - Report

  “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

Prefix | Read tuple ID | Segments of reads | Suffix (with comments and extensions)

sim_0043fd1(3,13,F,027871,027970),(3,13,R,029171,029270)_[paired_end],C:[100=,42=1X47=]
**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)
RNFtools — LAVEnder
(evaluation of mappers)
Contamination test 1
(human & mouse)

BWA-MEM
FDR in mapping (#wrongly mapped reads / #mapped reads)

YARA
FDR in mapping (#wrongly mapped reads / #mapped reads)
Contamination test 2
(human & chimpanzee)

BWA-SW

FDR in mapping (#wrongly mapped reads / #mapped reads)
Future perspective

• The software is ready to use, you can use it in your 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

Web: http://karel-brinda.github.io/rnftools/