

GO VS C++ VS JAVA FOR DNA SEQUENCING

PASCAL COSTANZA
CHARLOTTE HERZEEL

WHAT IS NEXT-GENERATION SEQUENCING?

CHEMICAL PROCESS





Intact DNA

(millions of molecules)

shearing

fragmentation

DNA fragments

make

library



sequence

the ends

GTTGAGCTTGCGTTTTTGGTACGCTGGACTTTGT
GTACTCGTGCGTGGGTTTGGTGTTTTGGT
ATGGTAGGCTGGACTTTGTAGGATACCCTGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GGGTTTATGGTACGCTGGACTTTGTAGGATACC
GAGCTTGCGTTATGGTACGCTGGACTTTGTAGGATACCT
GAGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCT

"reads"

Illumina HiSeq X image courtesy of Illumina, Inc.

WHAT IS NEXT-GENERATION SEQUENCING?

SOFTWARE PIPELINE



ACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCA
CAACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAG
CAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAG
CAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAG
GATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTG
CAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTG
CAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTG

reads

mapping

alignments

variant

calling

ACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCA
ACAAGATGTTTTGCC
CAAGACAGACCTGCCCTGTGCAG
CAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAG
CATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAG
GATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTG

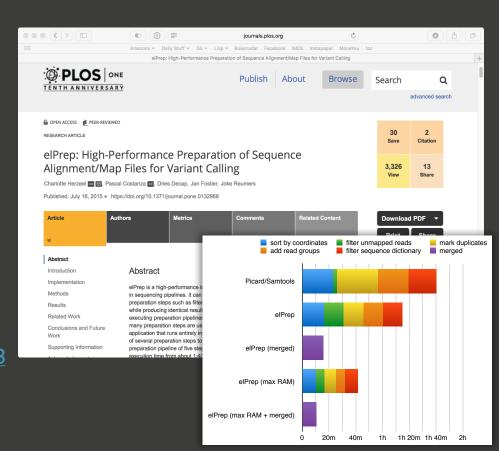
caagatgttttgcc actggccaagacctgccctgtgcag

Variants

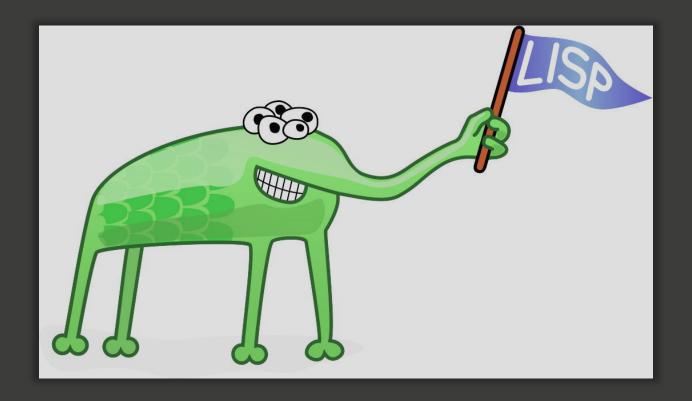
DECVT100 image by Jason Scott, CC BY 2.0, https://creativecommons.org/licenses/by/2.0/

ELPREP: A HIGH-PERFORMANCE TOOL FOR SEQUENCING

- High-performance tool for preparing SAM files for variant calling.
- Multi-threaded application that runs entirely in RAM and merges multiple steps to avoid repeated file I/O.
- Can improve performance by a factor of up to x10 compared to standard tools, and therefore remove hundreds of hours of computing time.
- Open-access publication in PLOS One, https://doi.org/10.1371/journal.pone.0132868



ELPREP: ORIGINAL IMPLEMENTATION IN COMMON LISP



Alien Lisp Mascot by Conrad Barski, M.D.

MODERN MEMORY MANAGEMENT

- Reference counting
 - C++ (std::shared_ptr)
- Concurrent garbage collection
 - Java
 - Go
- Input data: a whole-exome data set (13 GB compressed)
- Platform: Intel Xeon E5-2600 v4 (Broadwell)
 - 22 cores x 2 sockets = 88 threads

RESULTS

GNU g++ 6.3
Intel TBB 4.4
gperftools 2.5

13:38 mins @ 227.4 GB RAM

Java (JDK 1.8)
 ConcMarkSweepGC
 GIGC
 ParallelGC

15:05 mins @ 293.4 GB RAM 11:57 mins @ 358.1 GB RAM 11:07 mins @ 477.3 GB RAM

Go 1.7 default settings

10:20 mins @ 233.7 GB RAM

ELPREP: A HIGH-PERFORMANCE TOOL FOR SEQUENCING

- elPrep 3.0 implemented in Go
- Open-sourced (BSD) in September 2017
 https://github.com/exascience/elprep
- Pargo library for parallel programming in Go https://github.com/exascience/pargo

