IndexTools

A toolkit for accelerating genomic analysis using NGS index formats

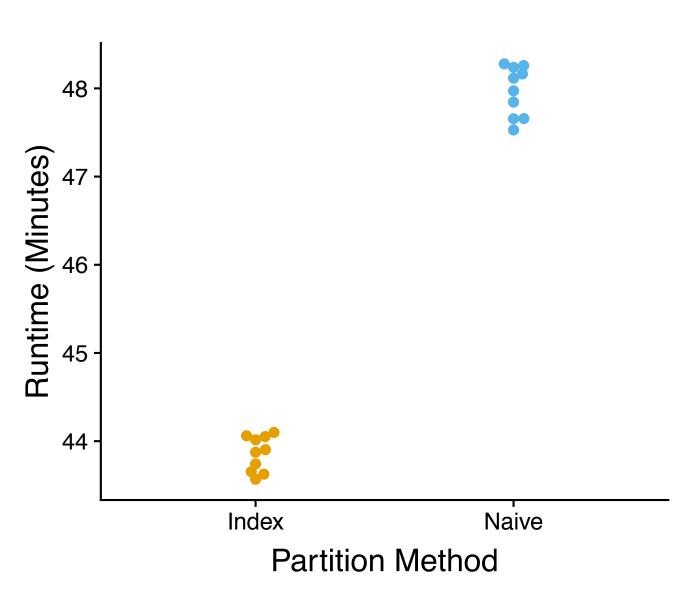
♣ John Didion, Yih-Chii Hwang, Steve Osazuwa, Aleksandra Zalcman

INTRO

- Variant calling is embarrassingly parallel, can be accelerated by a "scatter-gather" approach.
- Optimal parallelization requires N
 data partitions of approximately
 equal size (N is multiple of # cores).
- BAM index files (.bai) contain a coarse-grained representation of read density across the genome.
- IndexTools leverages index files for ultra-fast genomic applications.

RESULTS

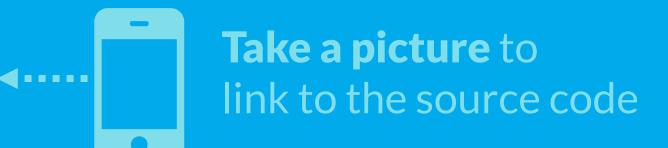
Compared to naïve parallelization, index-based partitioning of a 9.7 Gb Exome BAM file for GATK4 variant calling results in a **9**% speedup (mean 44 vs 48 minutes).

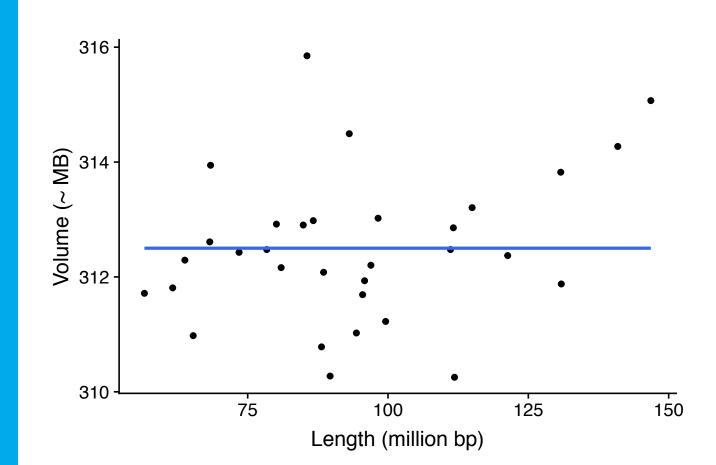




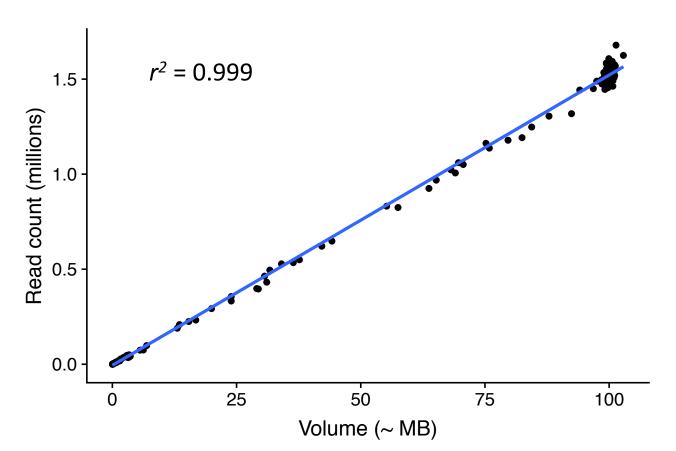
BAM index files can be used to estimate uniform partitions for parallelization, leading to faster variant calling.







Index-based partitions fall within a relatively narrow range of "volumes" (arbitrary unit correlated with uncompressed data size) despite spanning a wide range of genomic sizes.



Partition volumes are highly correlated with actual read counts.

FUTURE IMPROVEMENTS

- Testing on larger and diverse BAM files.
- Streaming support for multi-instance parallelization on cloud or HPC.
- Fast calling of large CNVs
- Coverage profiling
- Built-in task management (replacement for xargs)

IndexTools is written in Python 3 and built on several open-source libraries:



ngsindex: library for parsing NGS
index files



xphyle: automatic handling of compressed file I/O



AutoClick: automatically generate CLIs from type annotations