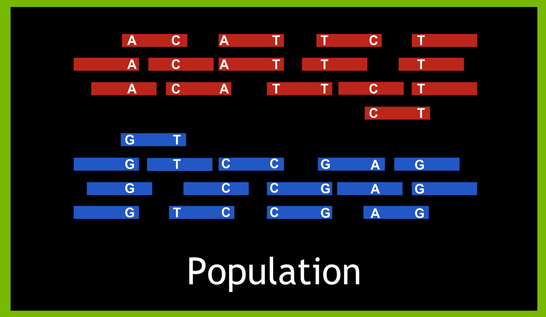
CLARA PARABRICKS

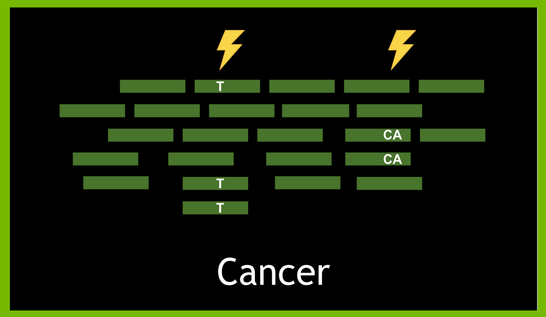
GREG ZYNDA - GZYNDA@NVIDIA.COM

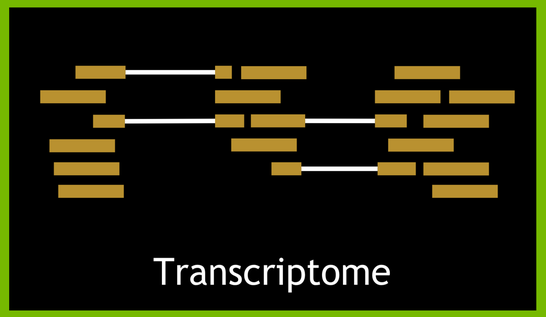
CLARA PARABRICKS FOR SECONDARY GENOMIC ANALYSIS

The Fastest Gold Standard in Sequencing Analysis

NVIDIA Clara Parabricks







Applications

Alignment	Variant Calling	Variant Processing
Preprocess	Joint Genotyping	Quality Checking

CUDA-X

EGX

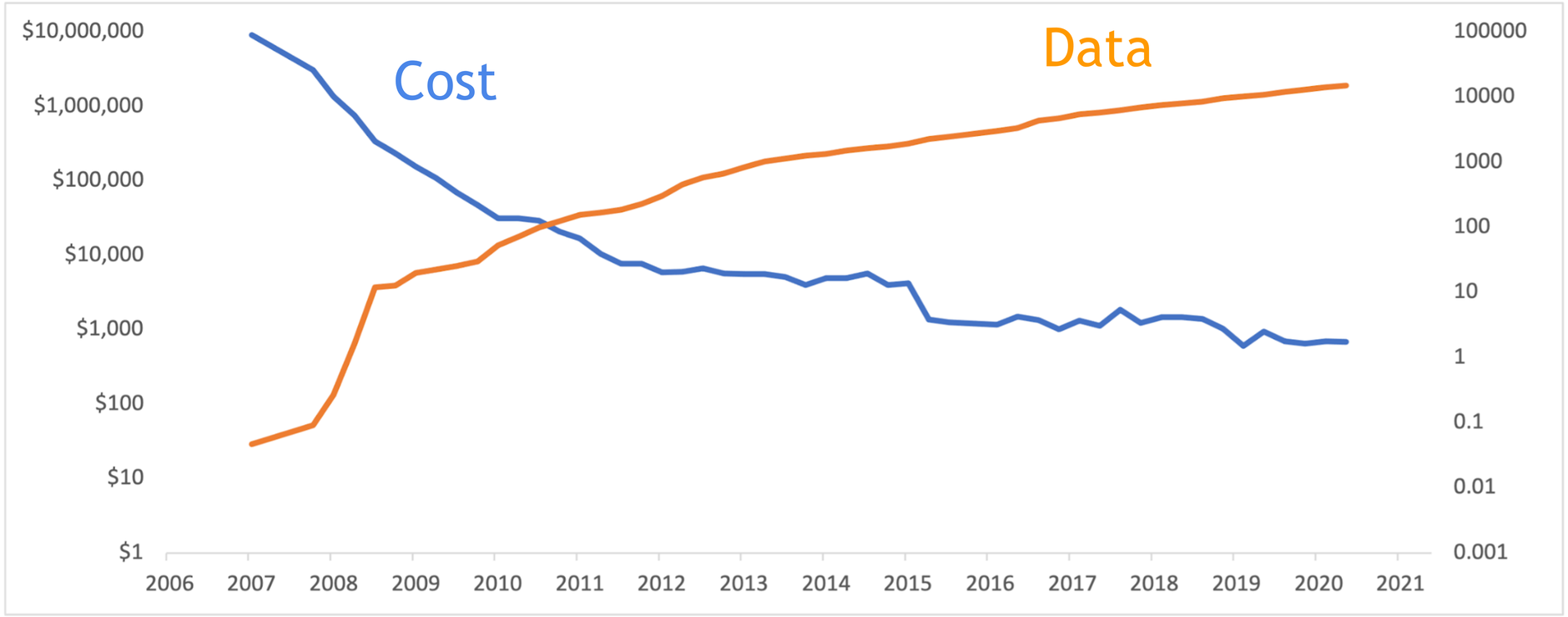
DGX

CLOUD

- Enterprise grade turn-key software
- Full Application Support
- Reference Mapping for DNA and RNA based applications
- GPU accelerated end-to-end workflows
- A full suite of variant callers

COMPUTATIONAL GENOMICS INFLECTION POINT

Cost decreasing, throughput increasing



1000 Genomes Project launches (2008)

100,000 Genomes Project launches (2013)

1,000,000 Genomes Initiative launches (2018)

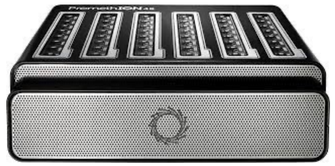
https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

NEXT GENERATION SEQUENCING WORKFLOWS

Highlighted green boxes are currently supported with Clara Parabricks

Short Read Platforms

Long Read Platforms



Alignment → Pre-processing → Variant calling

Alignment → Pre-processing → Variant calling

Phenotype to Genotype

Primary Analysis

On Instrument

Base calling turns signal into sequence data (reads)

Secondary Analysis

Near Instrument

Reference based: alignment-based applications (Dominant)
Reference free: de novo assembly of "New" genomes (Emerging)

Tertiary Analysis

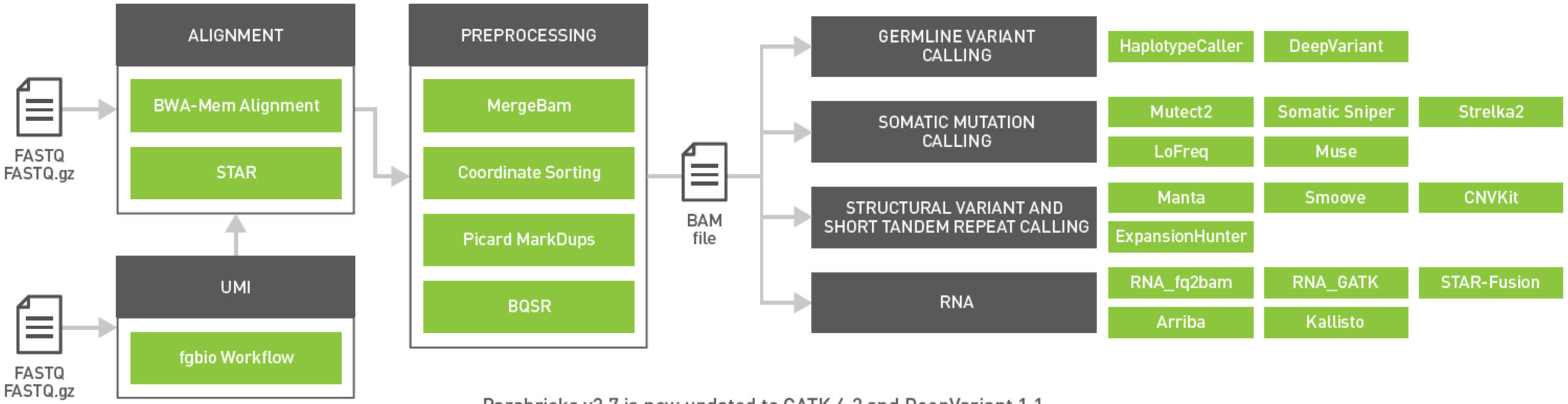
Off Instrument

Analyzing variations to assign annotations such as disease susceptibility or resistance

NVIDIA

CLARA PARABRICKS INCLUDES MANY FAMILIAR TOOLS

GPU-Accelerated Bioinformatics Tools



Parabricks v3.7 is now updated to GATK 4.2 and DeepVariant 1.1

More tools available:

JOINT GENOTYPING	ImportGVCF, SelectVariants, GenotypeGVCF
QUALITY CHECKING	Alignment Summary, InsertSize, SequencingArtifact, GcBias, QualityScoreDistribution, WGSMetrics, RAWWGSMetrics, BaseDistributionByCycle, MeanQualityByCycle
VARIANT PROCESSING	VQSR, Variant Filtration, Select Variants, CNNScore Variants
VARIANT QC	Variants can be annotated, filtered, and QC'd using NVIDIA generated tools, including merging multiple VCFs

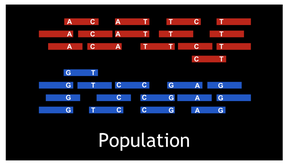
NOT ALL TOOLS SHOWN

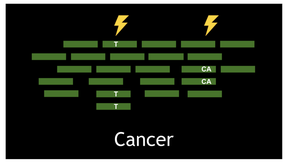
https://docs.nvidia.com/clara/parabricks/3.7.0/SoftwareOverview.html

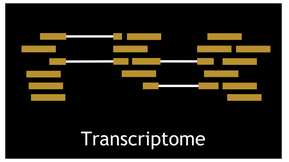
CLARA PARABRICKS ACCELERATED TOOLS

Release v3.7 now includes 50+ tools

New in v3.7







FastQ to BAM

Barcoded FastQ → fgbio Workflow* → Consensus FastQ

Normal FastQ → BWA-Mem Alignment → Coordinate Sort → Mark Duplicates → Apply BQSR

Tumor FastQs → BWA-Mem Alignment → Coordinate Sort → Mark Duplicates → Apply BQSR

RNA-Seq FastQs → STAR Alignment → Coordinate Sort → Mark Duplicates

**8 steps of fgbio condensed to a single command*

Variant Calling

- HaplotypeCaller
- DeepVariant

SNPs & short InDels

- CNVkit
- Manta
- Smoove
- ExpansionHunter

InDels, CNVs & SVs

- Mutect2
- SomaticSniper
- Strelka2
- LoFreq
- MuSe

SNVs & short InDels

- STAR-Fusion
- Arriba

Fusions

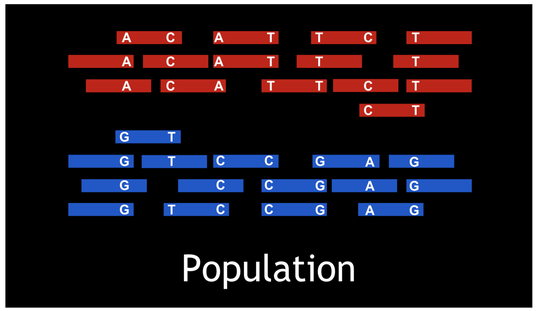
- Kallisto

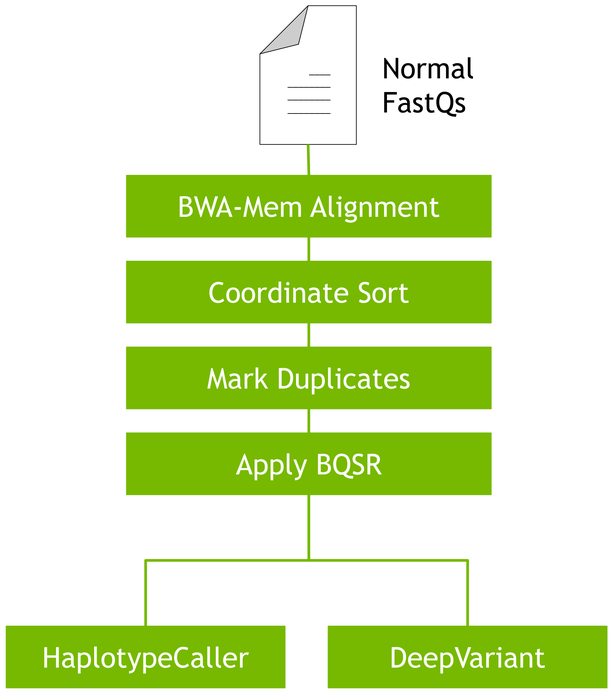
Quantification

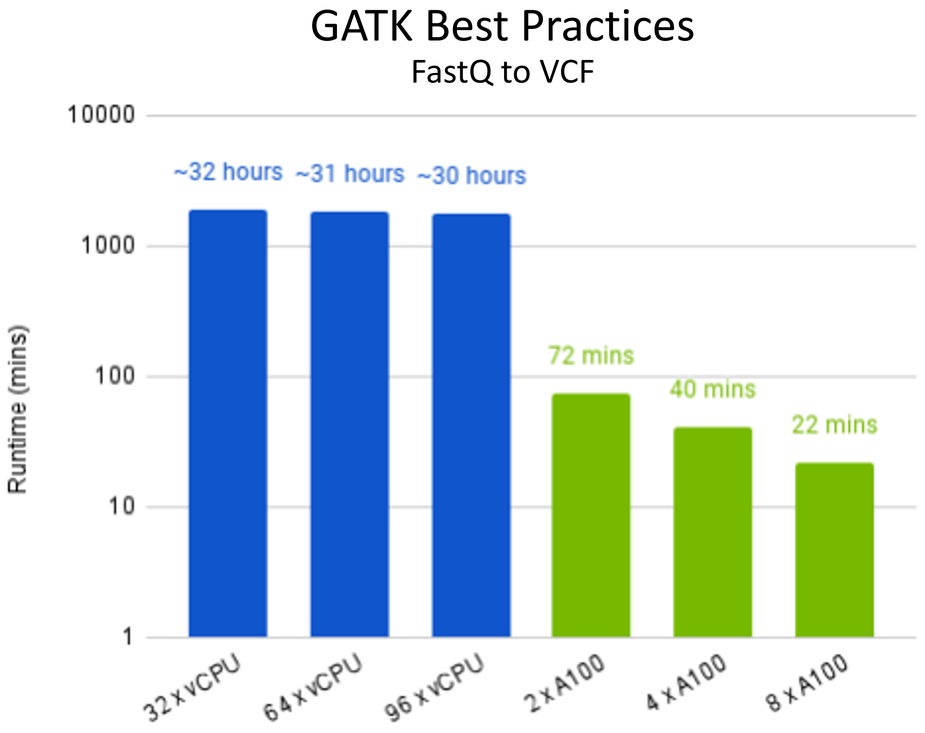
Recently added in v3.6 & v3.6.1

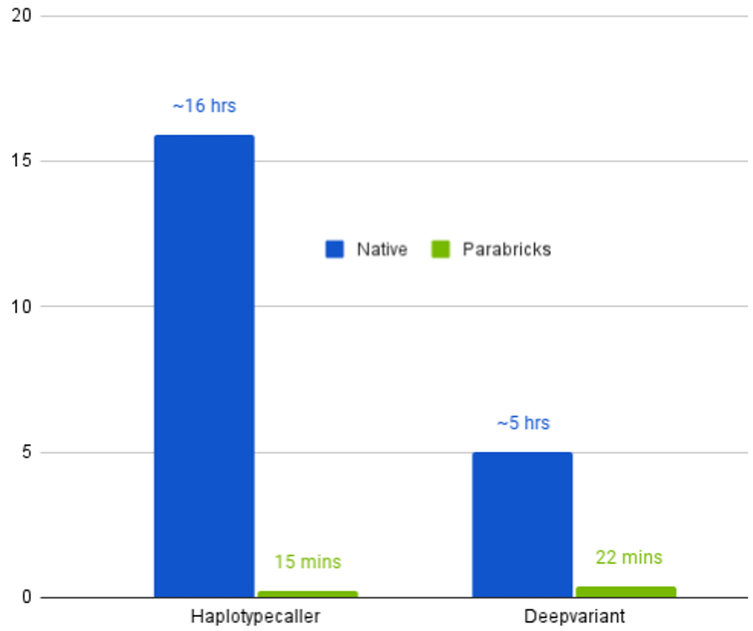
NVIDIA

CLARA PARABRICKS FOR GERMLINE WORKFLOWS









All Benchmarks are on 30X Whole Human Genome
Parabricks v3.7 is updated to GATK v4.2 and DeepVariant v1.1

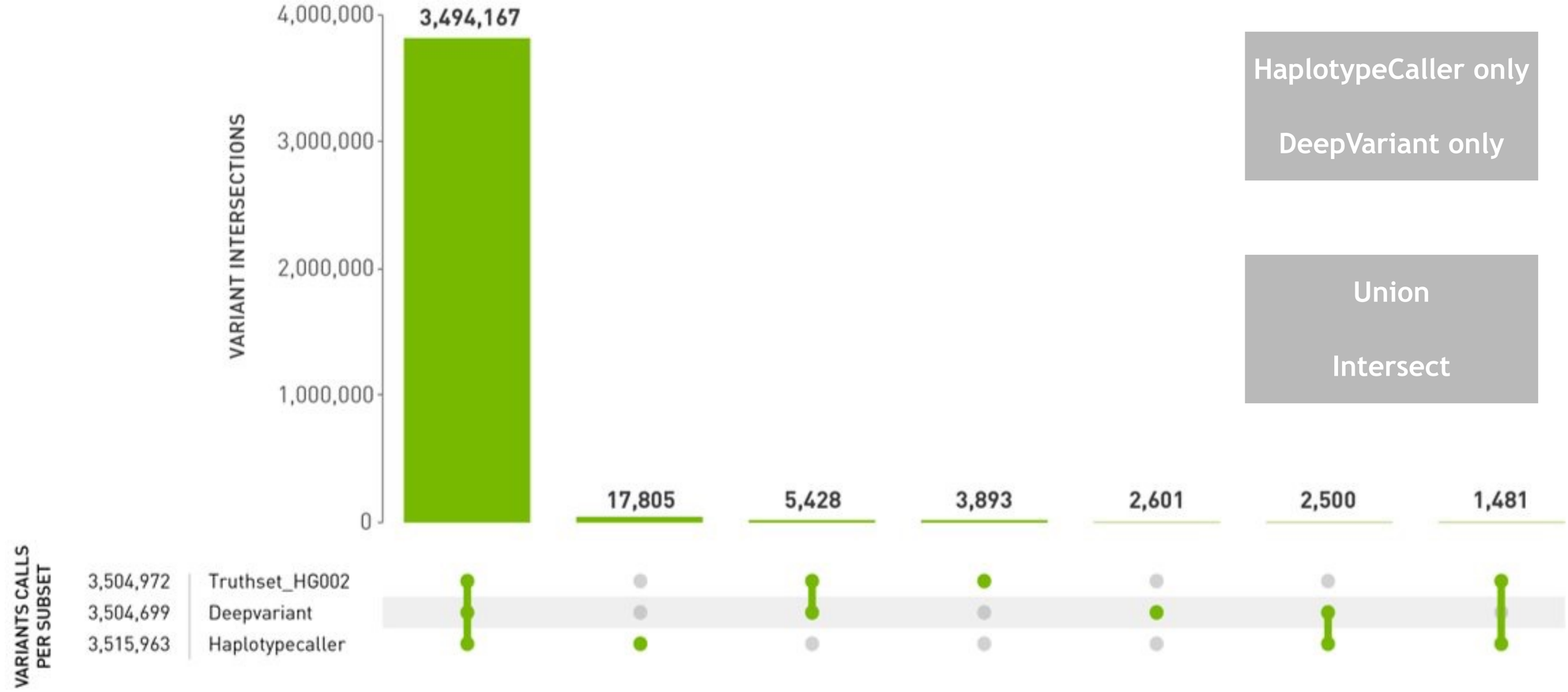
F1 scores

HaplotypeCaller: 0.9958
DeepVariant: 0.9991

MULTI-TOOL IMPLEMENTATION PROVIDES THE BEST RESULTS

Comparing germline calling between GATK Haplotype and Google's DeepVariant

	False Positives	False Negatives
HaplotypeCaller only	20,305	9,321
DeepVariant only	5,101	5,374
Union	22,906	3,893
Intersect	2,500	10,802

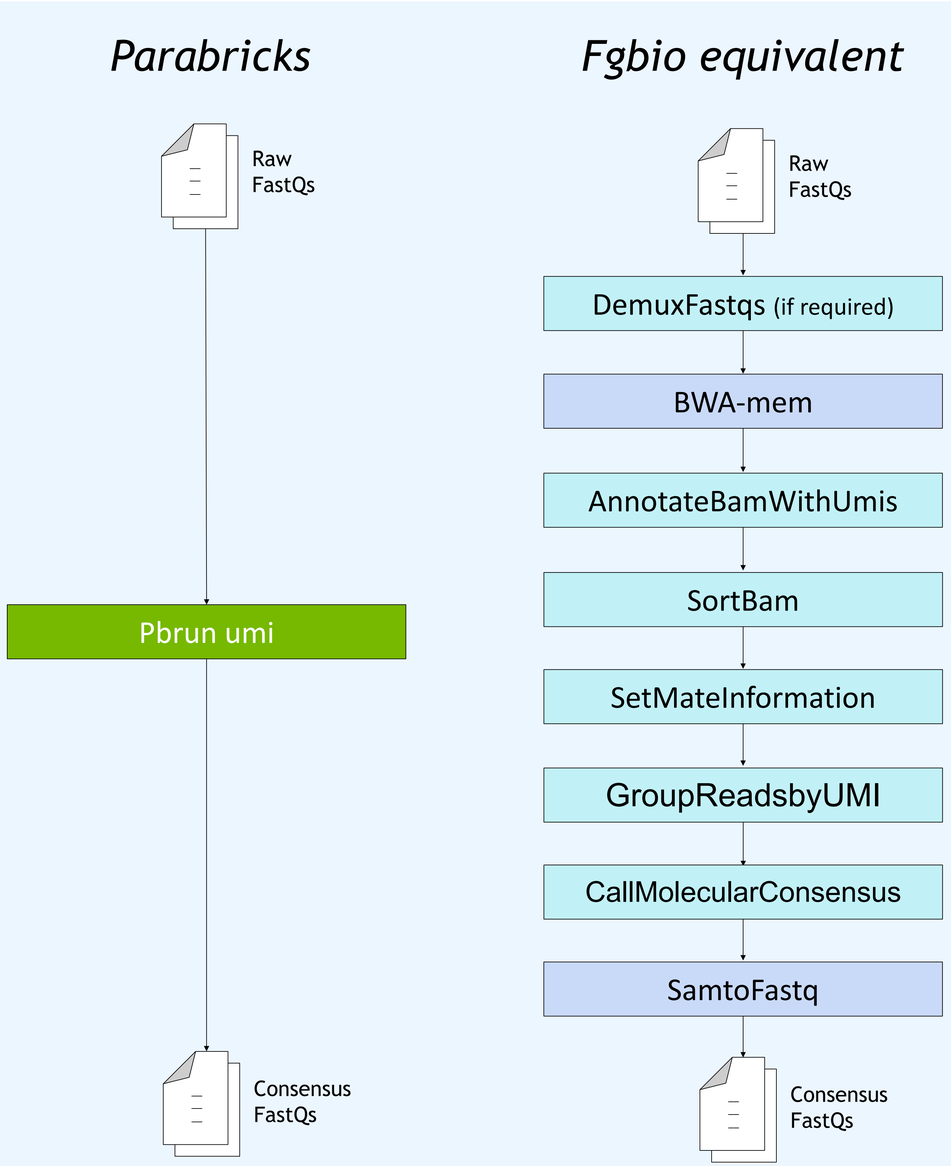


HG002 Target genome, analyzing only high confidence regions

NVIDIA

SIMPLIFYING & ACCELERATING GENE PANEL WORKFLOWS

NEW UPSTREAM TOOL SUITE: Molecular Barcodes or Unique Molecular Identifiers (UMIs)



A simplified workflow with **10x** acceleration

Multi input capability:

- UMIs still within the DNA read
- UMIs extracted to FastQ read names
- UMIs extracted to separate FastQ file

Run the entire workflow with a simple command:

```
pbrun umi \
    --inputs R1.fastq R2.fastq \
    --read-structures 8M92T 92T8M
```

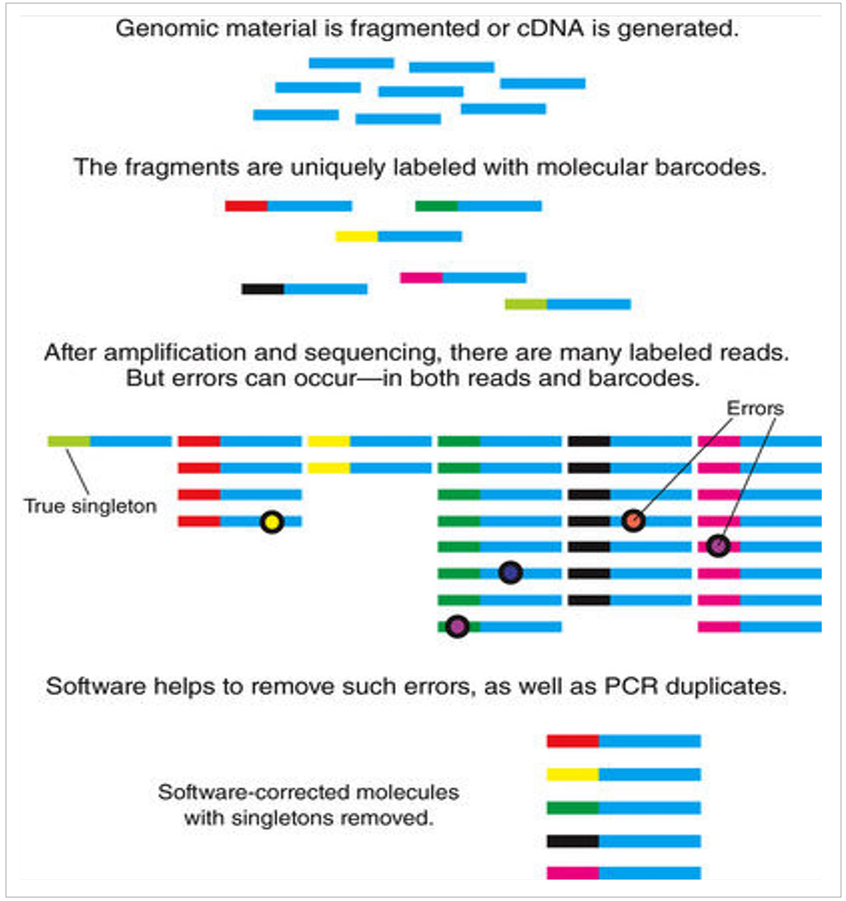
Fgbio= set of command line tools by Fulcrum Genomics to perform bioinformatic/genomic data analysis.

NVIDIA.

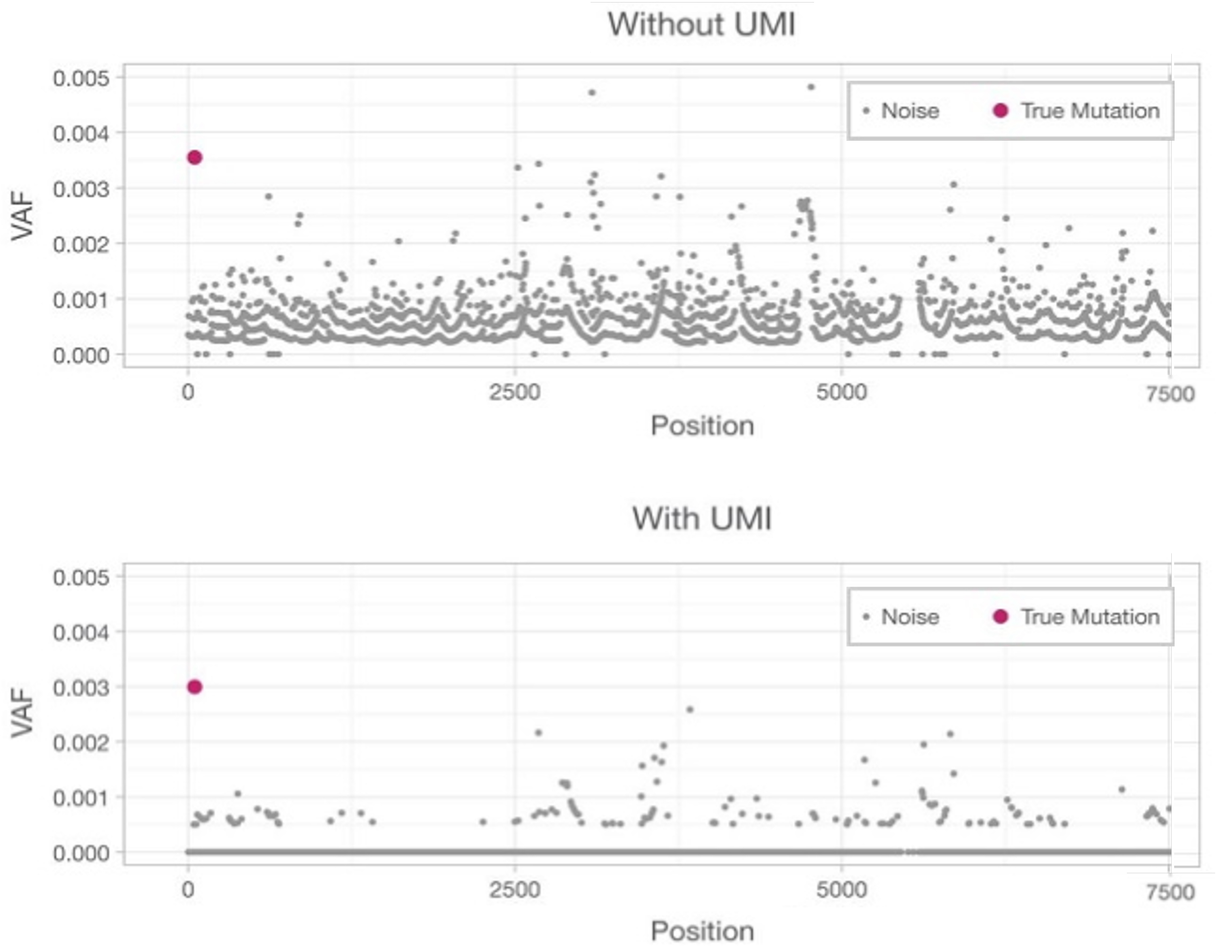
SUPPORTING GENE PANELS WORKFLOWS

Molecular Barcodes or Unique Molecular Indices (UMIs)

Extract UMI > group by UMI > consensus reads

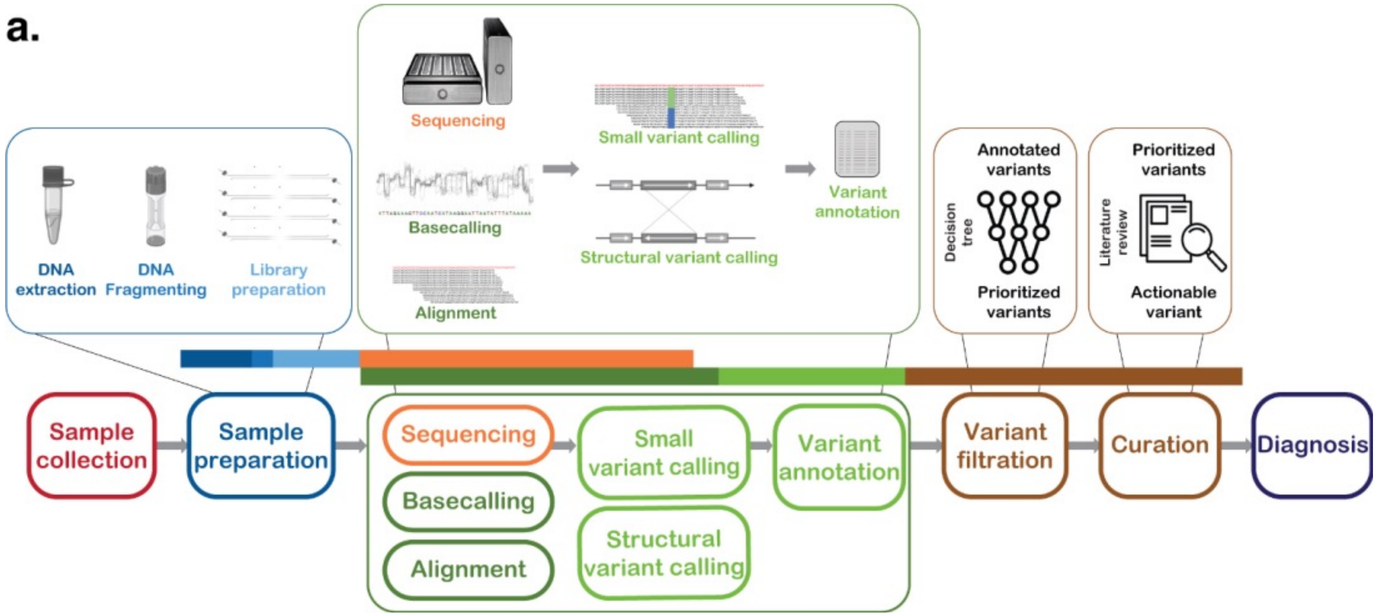


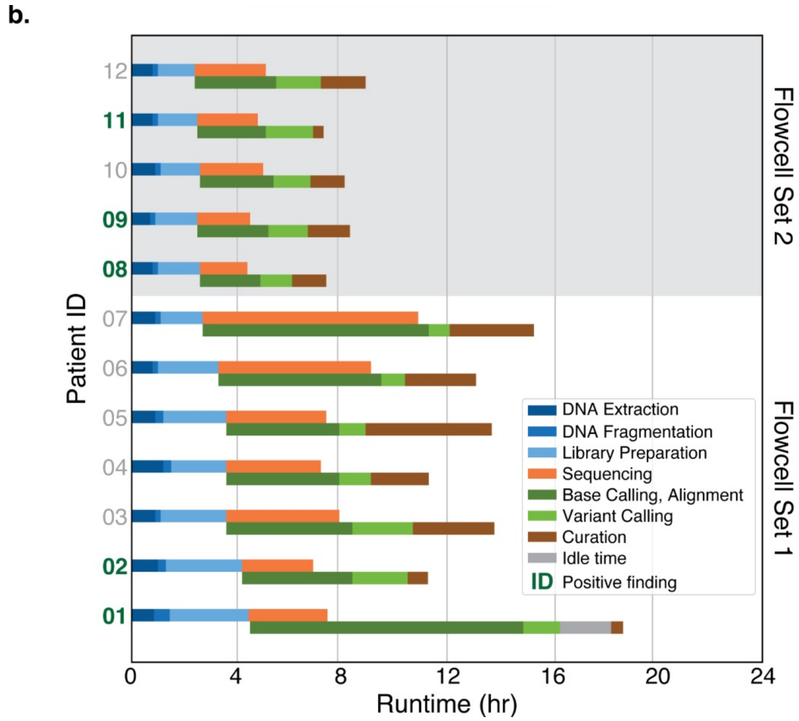
UMI de-duplicating for ultra low frequency variants



SETTING A WORLD RECORD FOR FASTEST DNA SEQUENCING TECHNIQUE

Stanford, NVIDIA, ONT, Google and UCSC Diagnoses Rare Diseases in under 7.5 Hours



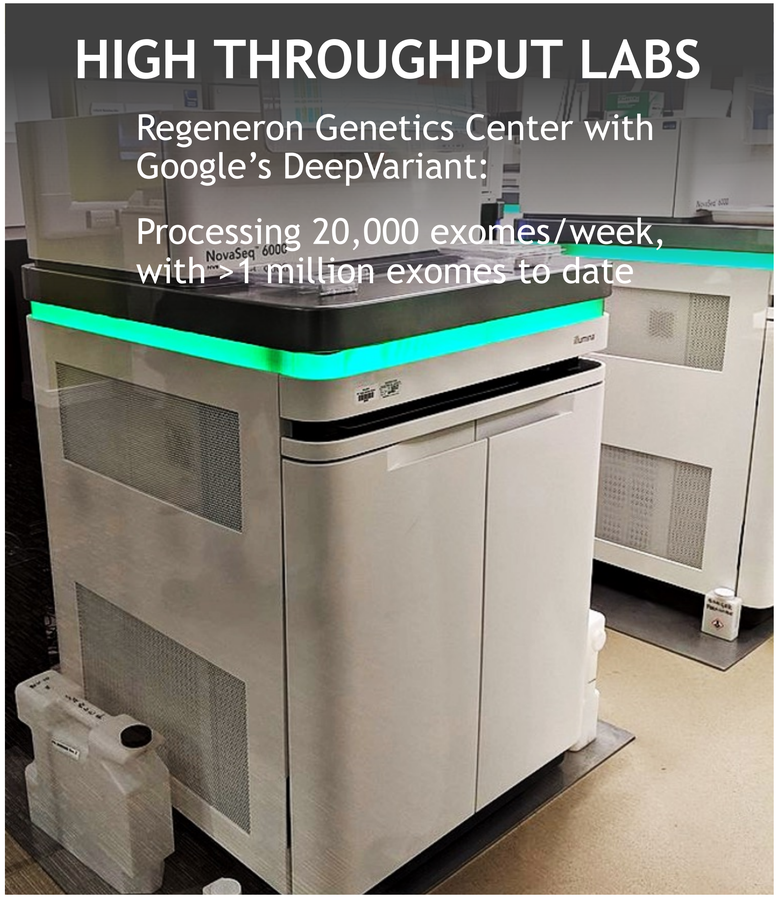


- 12-33% of Neonatal/Pediatric ICU hospitalizations are due to a primary genetic disease
- Whole Human Genome Sequencing diagnostic test
  - Typically takes 6-8 weeks
  - “Rapid” turnaround time is 2-3 days
  - Stanford took 7.5 hours
- Informed clinical decisions including a heart transplant
- Utilized Clara Parabricks GPU-Accelerated version of DeepVariant
- Base Calling in ONT Sequencing Instrument was also accelerated with NVIDIA GPUs
- Published in the New England Journal of Medicine

ENABLING GENOMICS AT INDUSTRIAL SCALE

>1 million whole exome datasets analyzed

HIGH THROUGHPUT LABS



Regeneron Genetics Center with Google's DeepVariant:

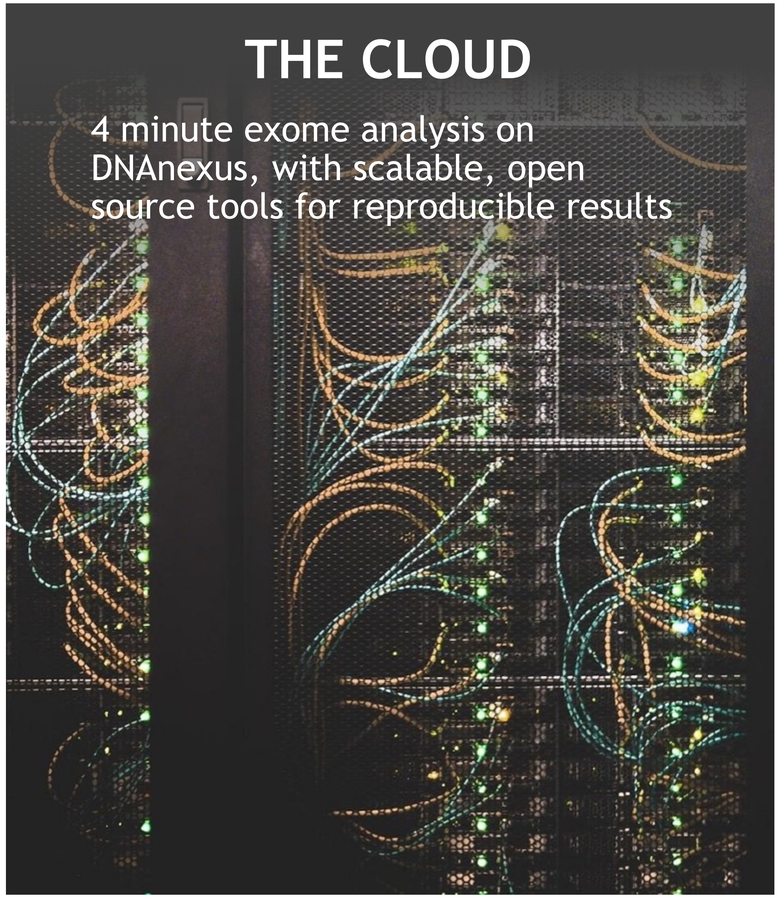
Processing 20,000 exomes/week, with >1 million exomes to date

BIOBANKS

UK BioBank: ~500,000 UK genomes

A foundation for breakthroughs in prevention, diagnosis & treatment.

THE CLOUD



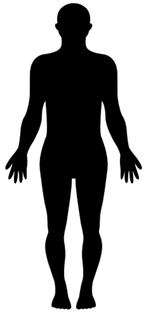
4 minute exome analysis on DNAnexus, with scalable, open source tools for reproducible results

[Learn More about Parabricks Accelerating Research and Discovery on UK Biobank's Research Analysis Platform]

NVIDIA

THE IMPACT OF AN AT-RICH GENOME ON ANALYSIS TIME

Malaria Project with Assistant Prof. Giovanna Carpi, Purdue University



Human

Homo sapien

Malaria

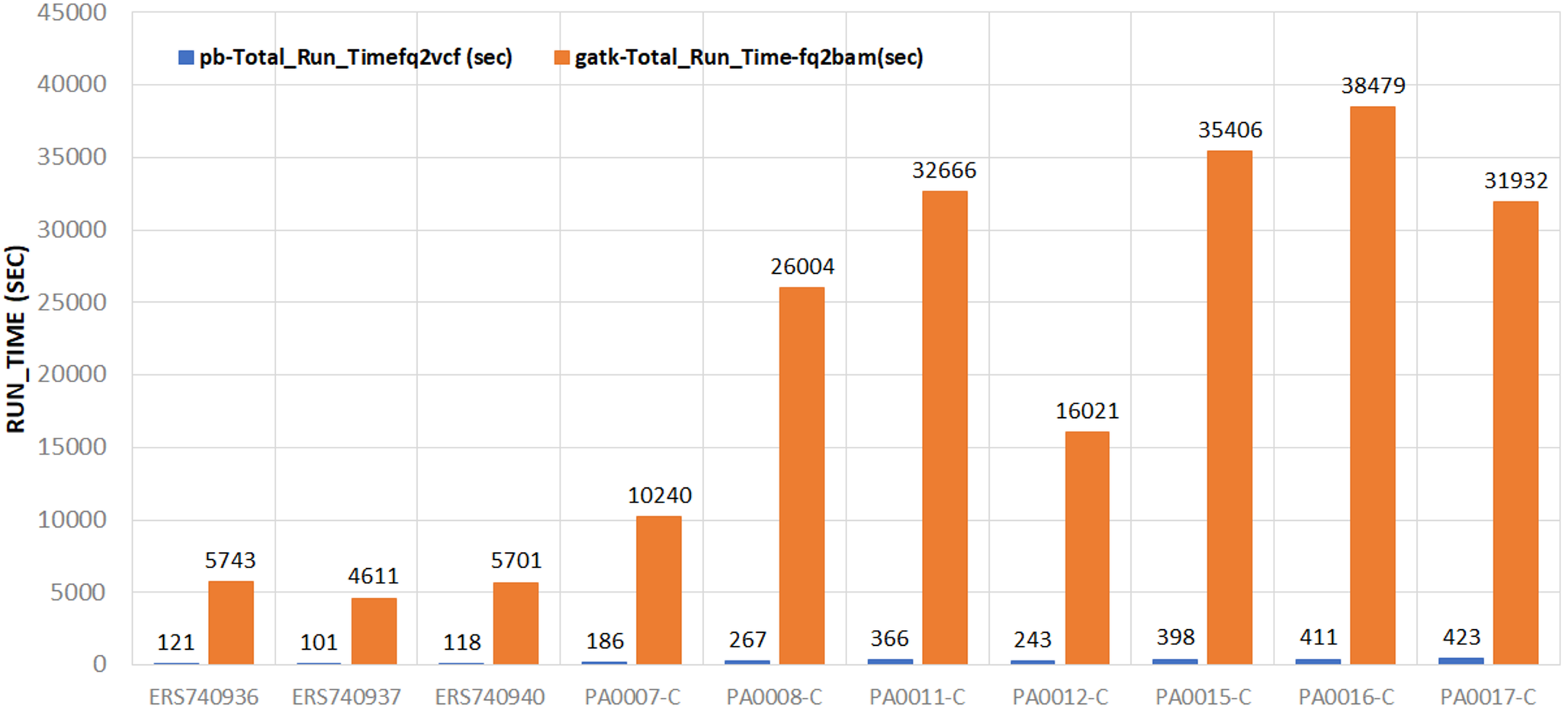
Plasmodium falciparum

Chromosome	46	14
Size	6,200 megabases	23 megabases
GC content	41%	19%
WGS CPU analysis time	~20-30 hours	~6 hours

https://www.sanger.ac.uk/resources/downloads/protozoa/plasmodium-falciparum.html

FAST ANALYSIS- 4 MINUTES WITH CLARA PARABRICKS VS. 6 HOURS

Malaria Project with Assistant Prof. Giovanna Carpi, Purdue University



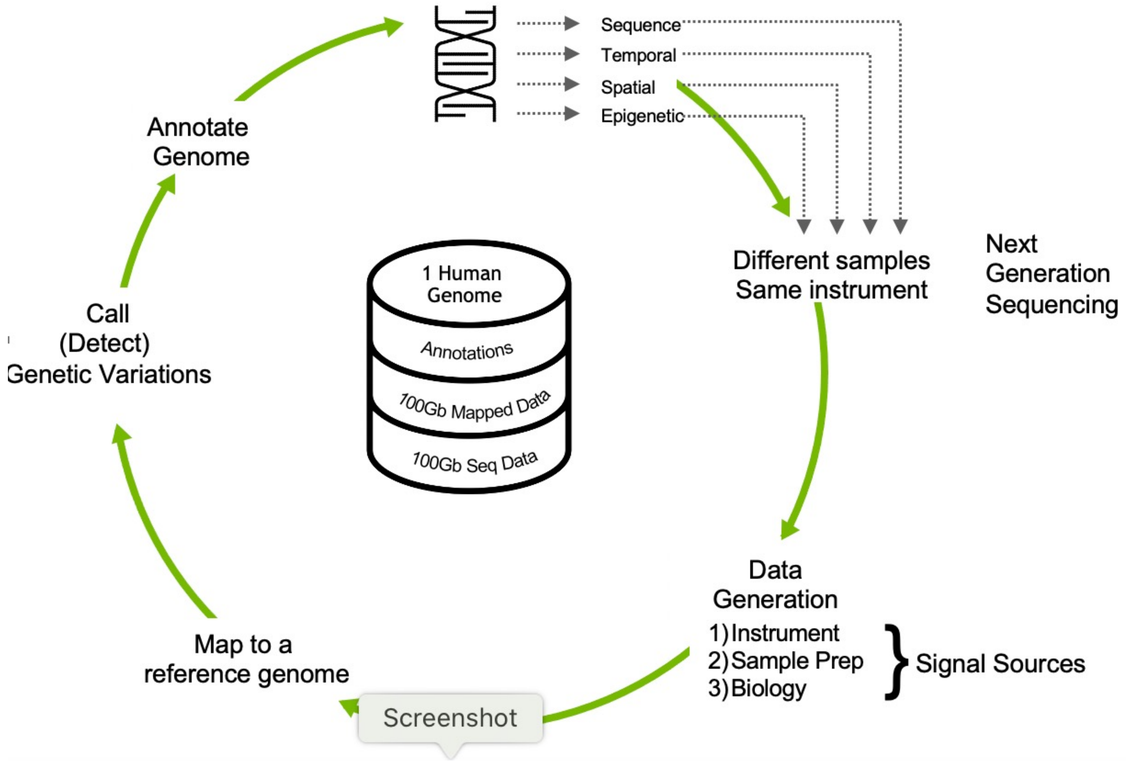
20,680 seconds (~6 hours) to 263 sec

TAKE ADVANTAGE OF ARCHIVED DATA & ALIGNING TO NEW REFERENCES

10x accelerated BAM2FASTQ tool for converting BAM/CRAM archived data back to FASTQ

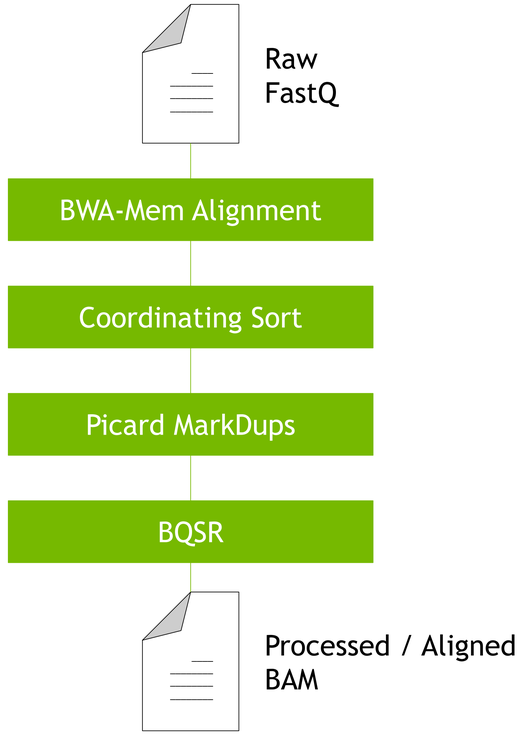
Whole human genome data set:

- 100Gb for FASTQ (sequencing reads)
- 100Gb for BAM (mapped reads)



Update referenced mapped data

- >10x acceleration of BAM2FASTQ
- Restore data back to FASTQ
- Update to most current reference GRCh38



NVIDIA

PARABRICKS LICENSING AND ENTERPRISE SUPPORT

Annual per GPU (on Prem) or 10k hour (cloud) License includes:

License Support		Full access to all pipelines in the NVIDIA Clara Parabricks software suite Access to different versions of the same tool No limitations on the number of genomes analyzed or number of users
NVIDIA Enterprise Support		Available via email or phone today. Web Portal for optimal customer experience under development
Access to Experts		Join the Forum: https://forums.developer.nvidia.com/c/healthcare/Parabricks/290 Direct access to NVIDIA expertise for timely resolution of issues

STARTING POINTS

Clara Parabricks Home Page

https://www.nvidia.com/en-us/clara/genomics

Getting Started Video

https://www.youtube.com/watch?v=AQltyCwPgU0

Request a Free Eval

On Prem or Cloud

https://www.nvidia.com/en-us/clara/genomics/parabricks-free-trial

Access Parabricks AMI on AWS

https://aws.amazon.com/marketplace/pp/prodview-apbngojlskcyq

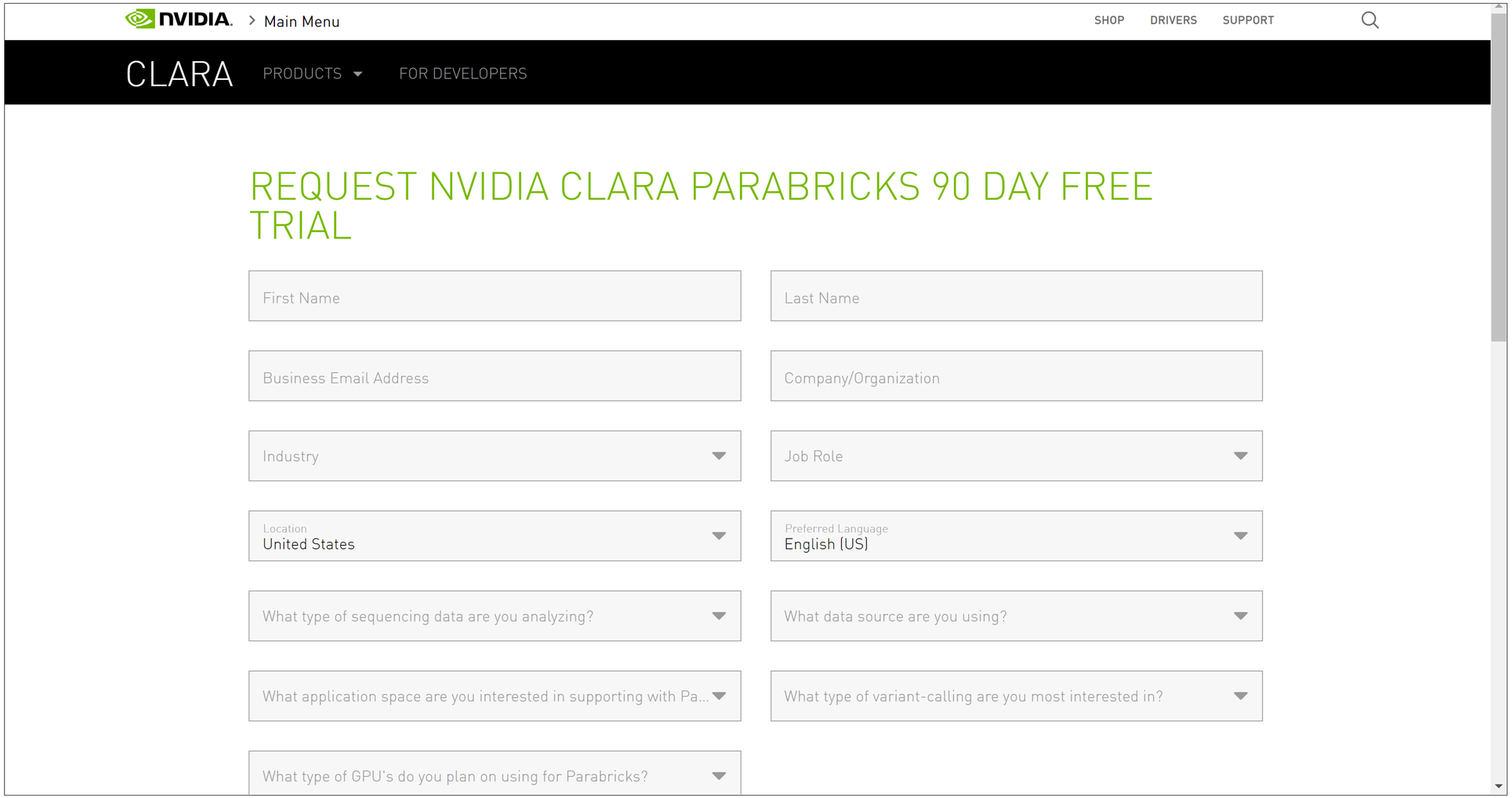
Documentation

https://docs.nvidia.com/clara/parabricks/3.7.0/index.html

How do you try Parabricks?

PARABRICKS HAS A 90-DAY FREE TRIAL LICENSE

Fill out the request form and you will receive a welcome email



https://www.nvidia.com/en-us/clara/genomics/parabricks-free-trial/

MAKING AN ACCOUNT ON NGC

The Parabricks welcome email has users create an NGC account

NVIDIA.

CLARA PARABRICKS PIPELINES EVAL

Thank you for requesting access to NVIDIA Clara Parabricks Pipelines.

A member of the Parabricks Support team (Parabricks-Support@nvidia.com) will reach out within the next 72 hours, with a license key for your free 90-day evaluation to Parabricks Pipelines.

To help get you started with quick-start guides, please visit our documentation page.

If you have any questions, please reach out to us on our developer forums.

Thank you,
The NVIDIA Parabricks team

INSTALLING PARABRICKS

Parabricks comes as a .tar.gz file

99% of people use the traditional Docker/Singularity Installation

There exists a Flexera license server option

There exists a bare metal installation using Conda

```
// Unzip the package
tar xzvf parabricks.tar.gz


// Run the installer
sudo parabricks/installer.py <installation_args>


// Verify the installation
pbrun version
```

https://docs.nvidia.com/clara/parabricks/3.7.0/GettingStarted.html#installing-the-software

TESTING PARABRICKS

Sample data is provided to test Parabricks

Running on AWS g4dn.8xlarge
(1 T4 GPU, 32 vCPU, 128 GB Memory)

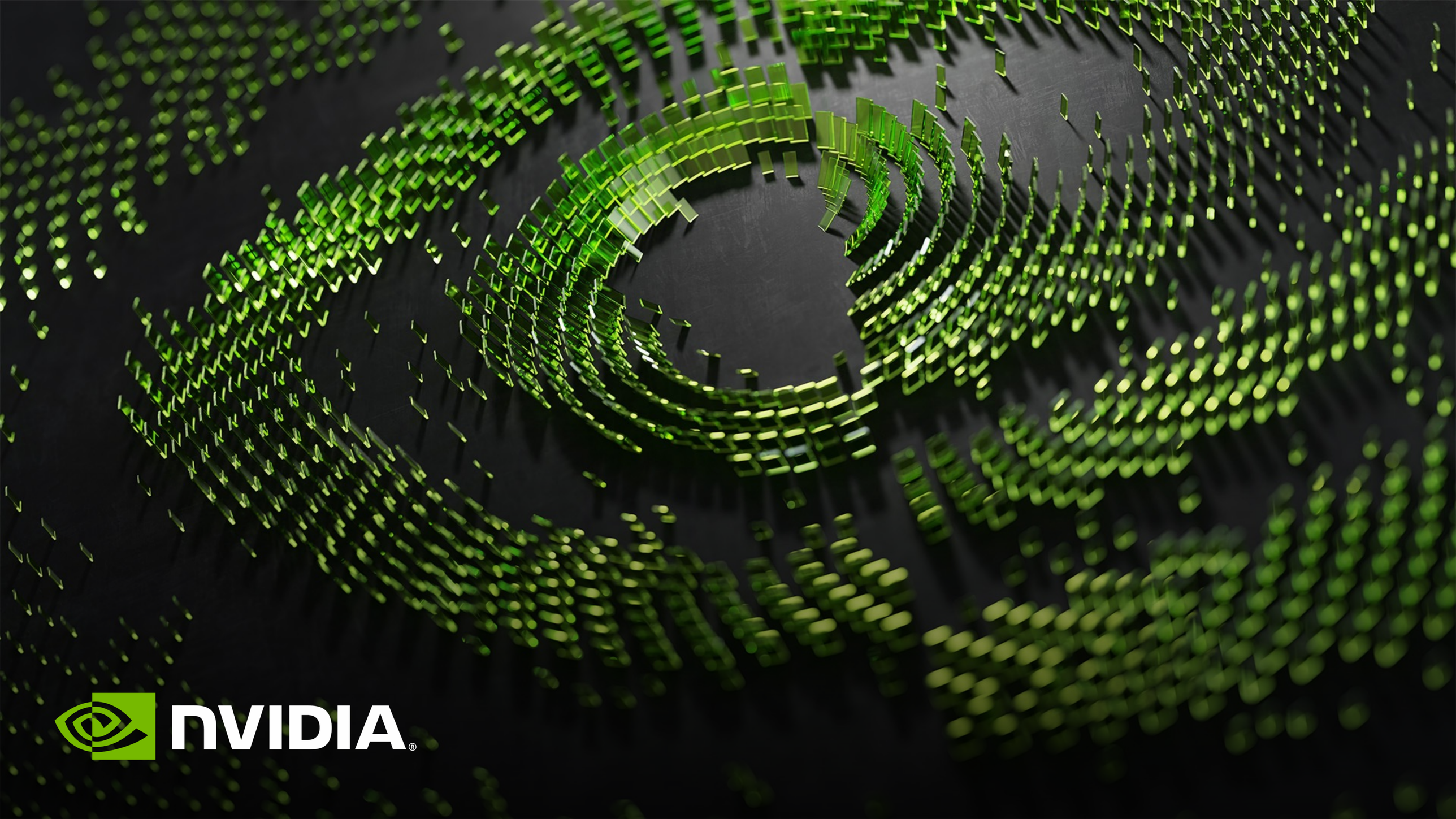
Fq2bam: ~6 minutes
HaplotypeCaller: ~6 minutes

```
// Download sample data
wget -O parabricks_sample.tar.gz \
https://s3.amazonaws.com/parabricks.sample/parabricks_sample
```

```
// Run fq2bam
pbrun fq2bam \
  --ref Homo_sapiens_assembly38.fasta \
     --in-fq sample_1.fq.gz sample_2.fq.gz \
     --out-bam output.bam

// Run HaplotypeCaller
pbrun haplotypecaller \
     --ref Homo_sapiens_assembly38.fasta \
     --in-bam output.bam \
     --out-variants variants.vcf
```

https://docs.nvidia.com/clara/parabricks/3.7.0/Tutorials.html


NVIDIA