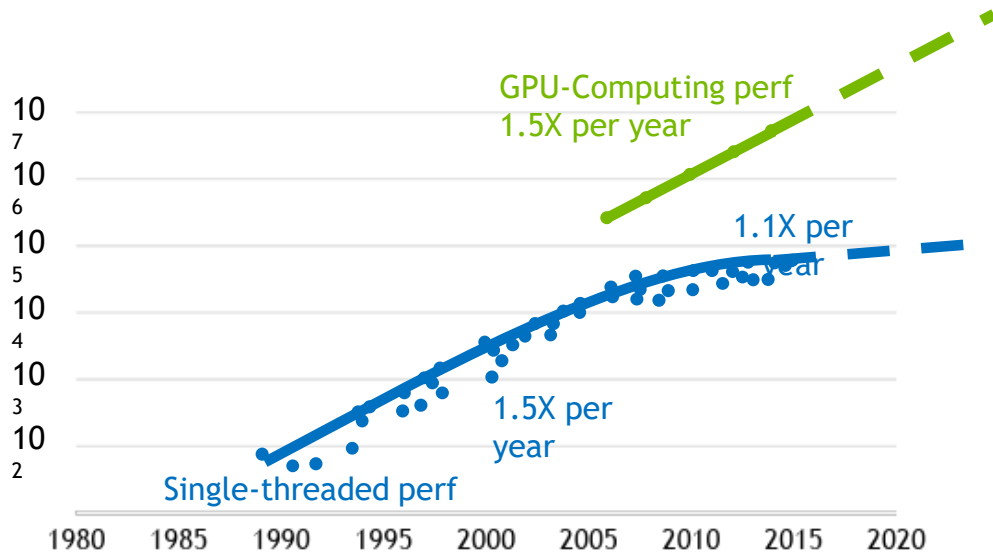
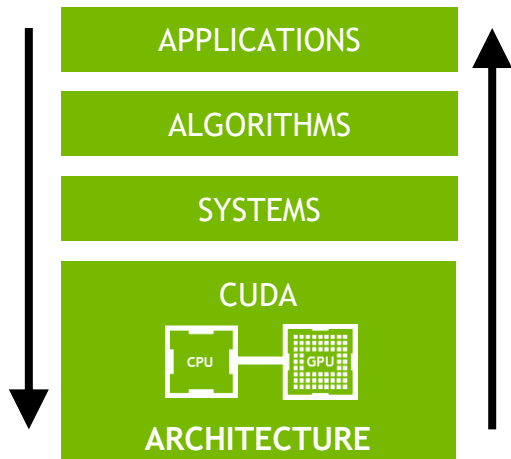




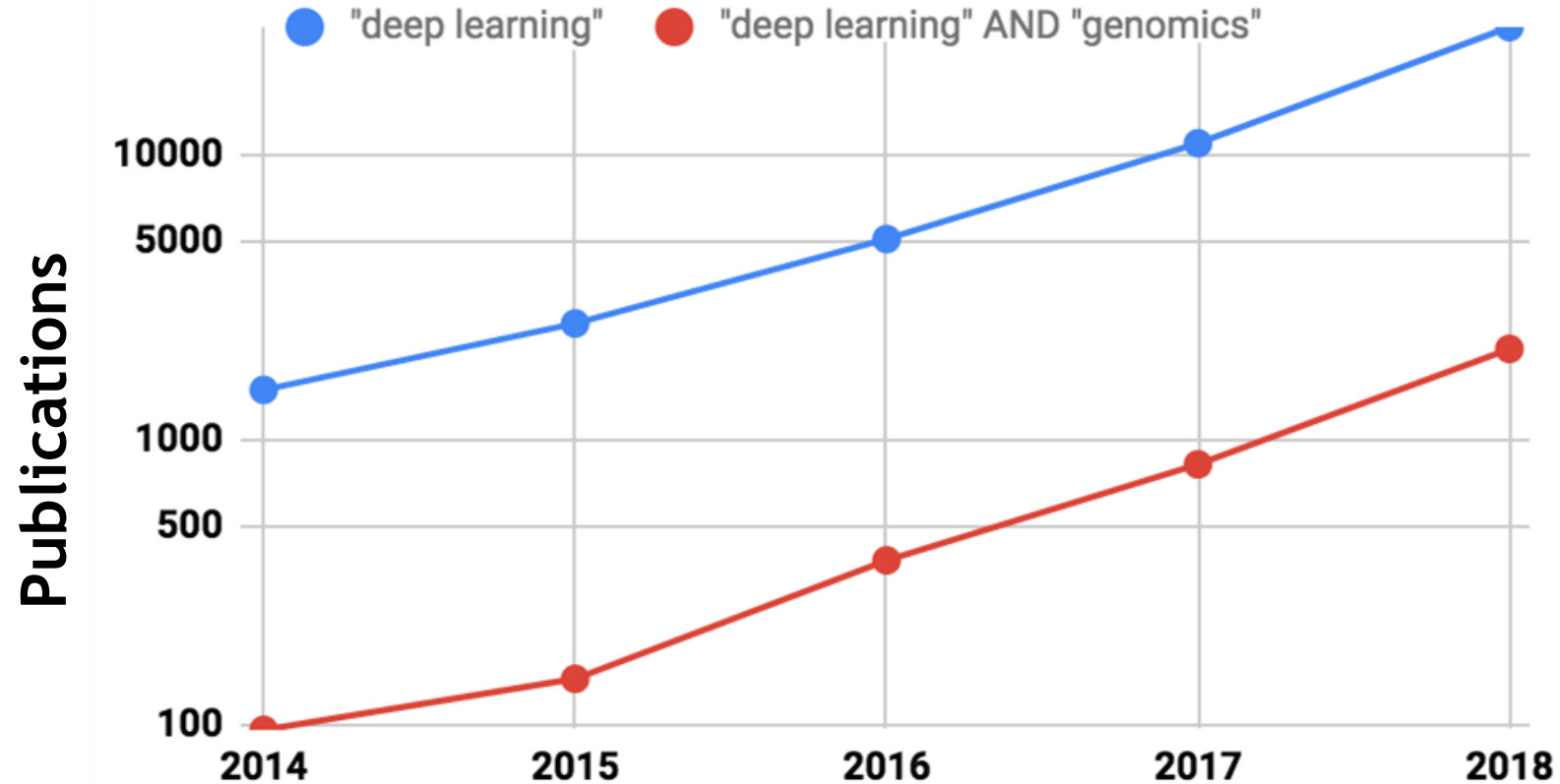
Accelerating Sequencing with GPU Computing and Deep Learning

Johnny Israeli

COMPUTE TRENDS



COMPUTE TRENDS

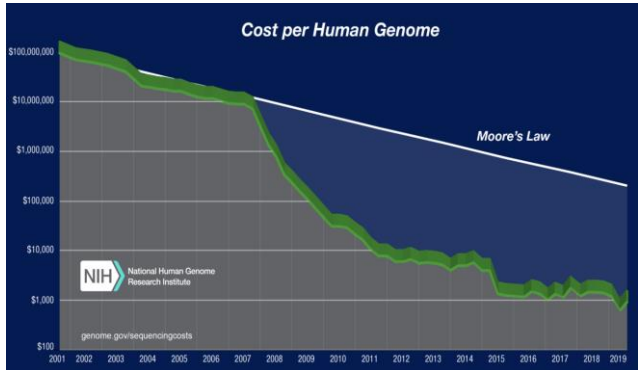


The background is a dark, almost black, field filled with a complex network of thin, glowing green lines. These lines connect various points, some of which are highlighted as bright green dots. The lines and dots create a sense of a dynamic, interconnected system, possibly representing a network or a data flow. The overall effect is futuristic and technological.

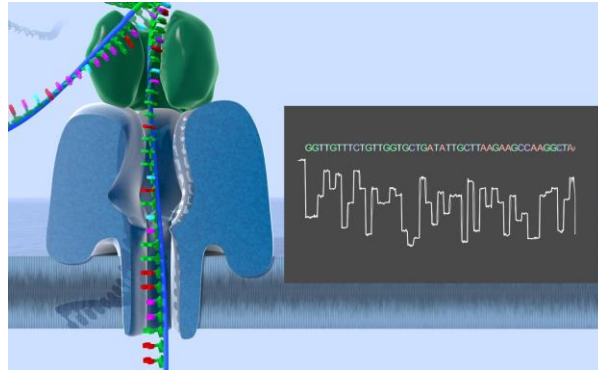
Sequencing Trends

SEQUENCING TRENDS

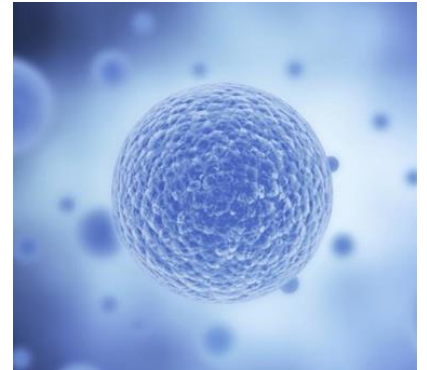
Sequencing Data Growing in Volume and Complexity



Decreasing Cost

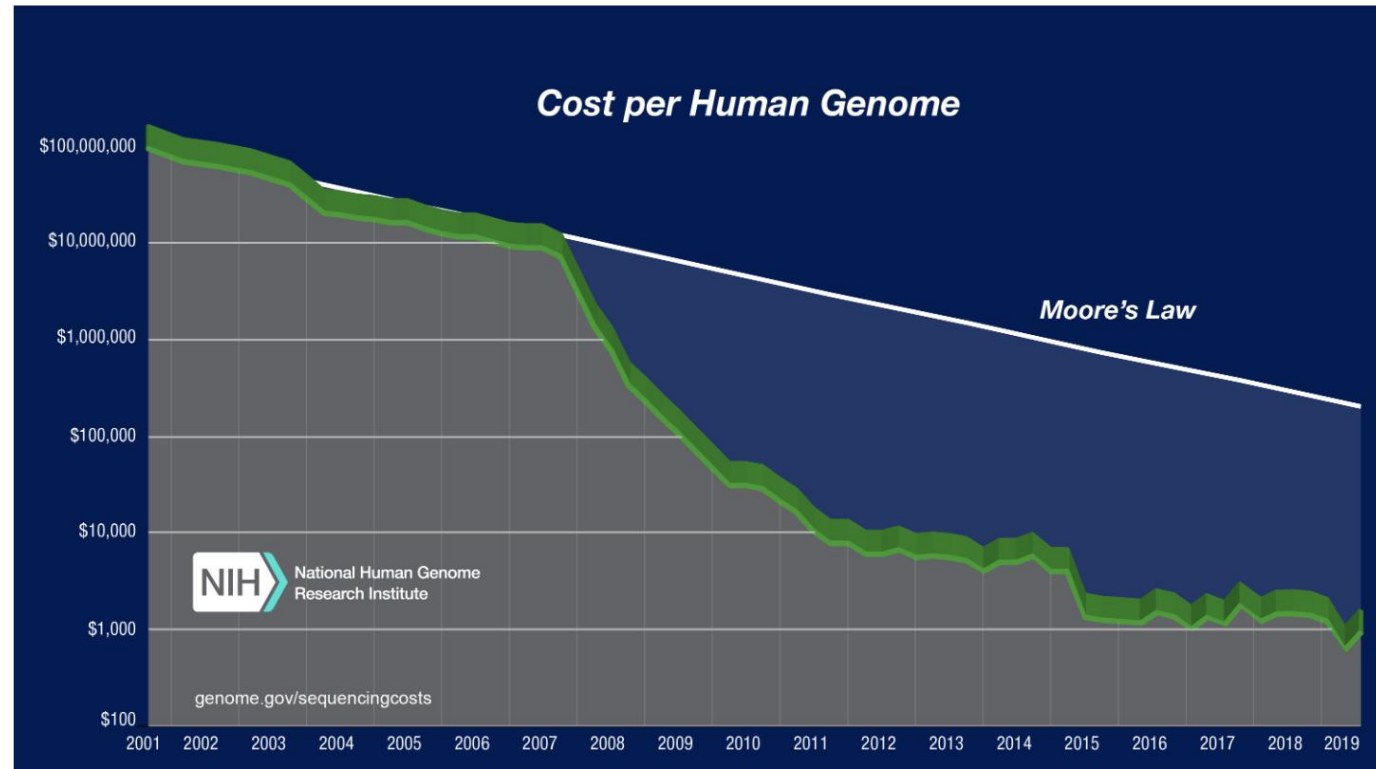


Increasing Read Length

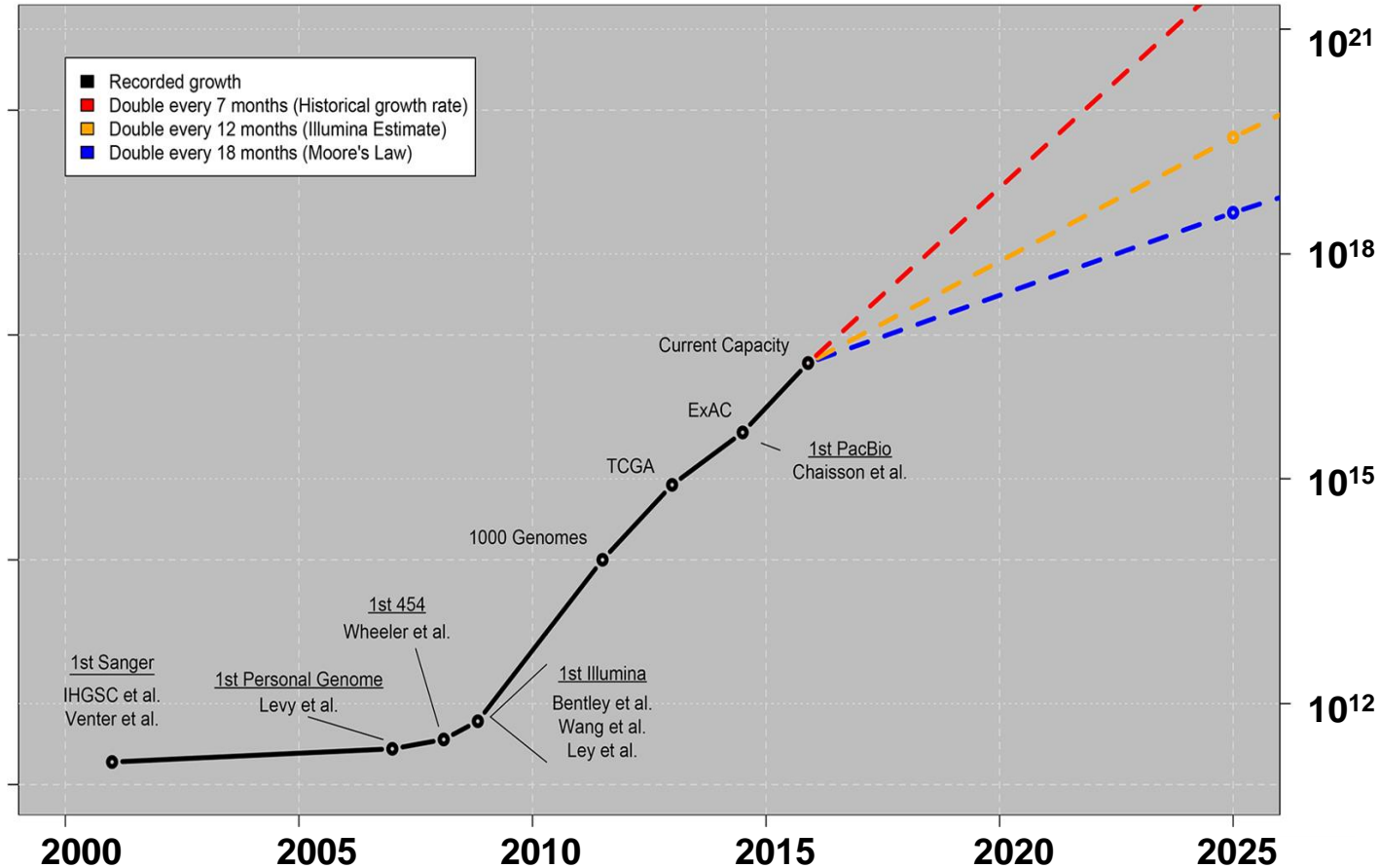


Rise of Single Cell
Data

SEQUENCING TRENDS

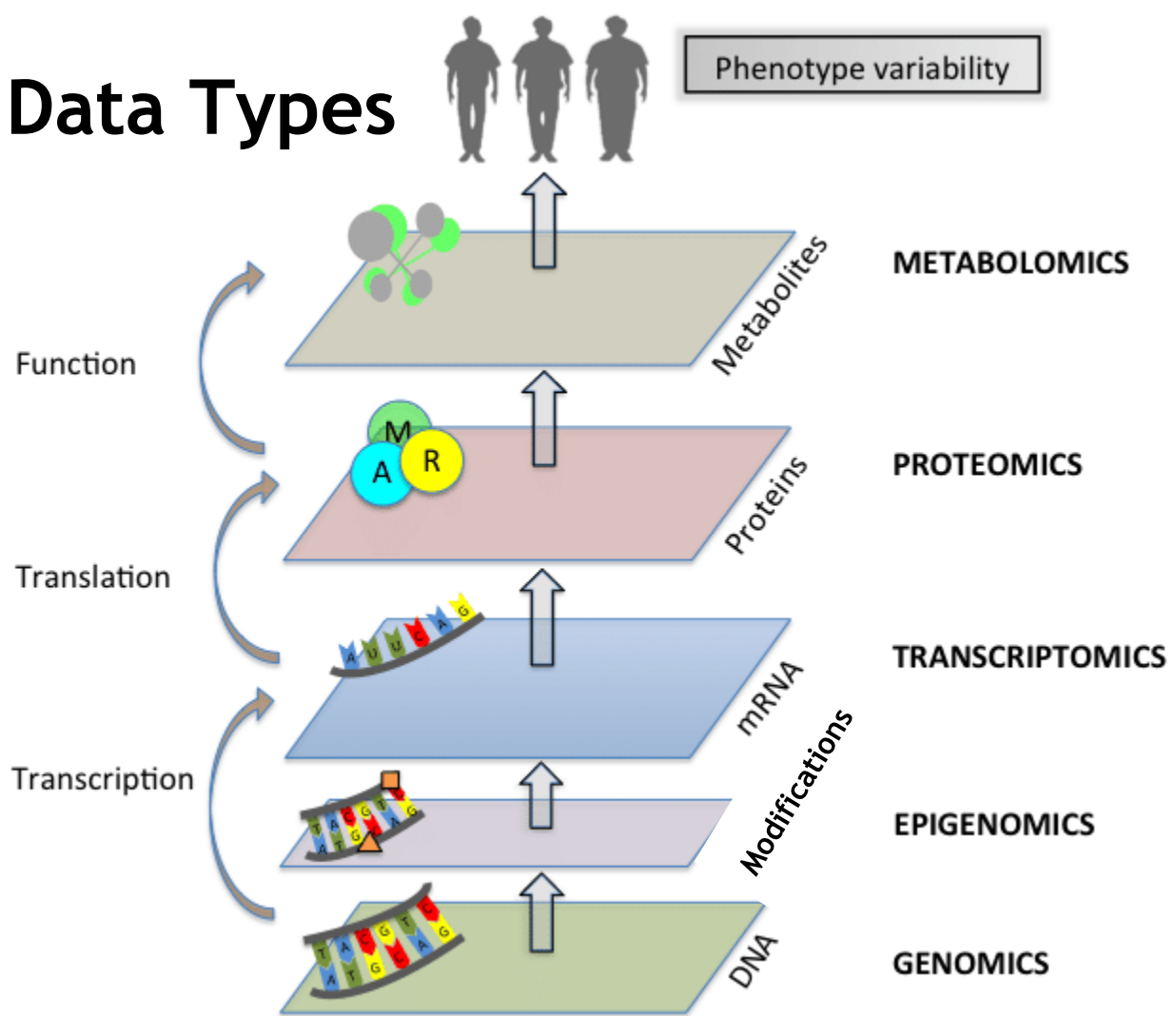


Worldwide Annual Sequencing Capacity



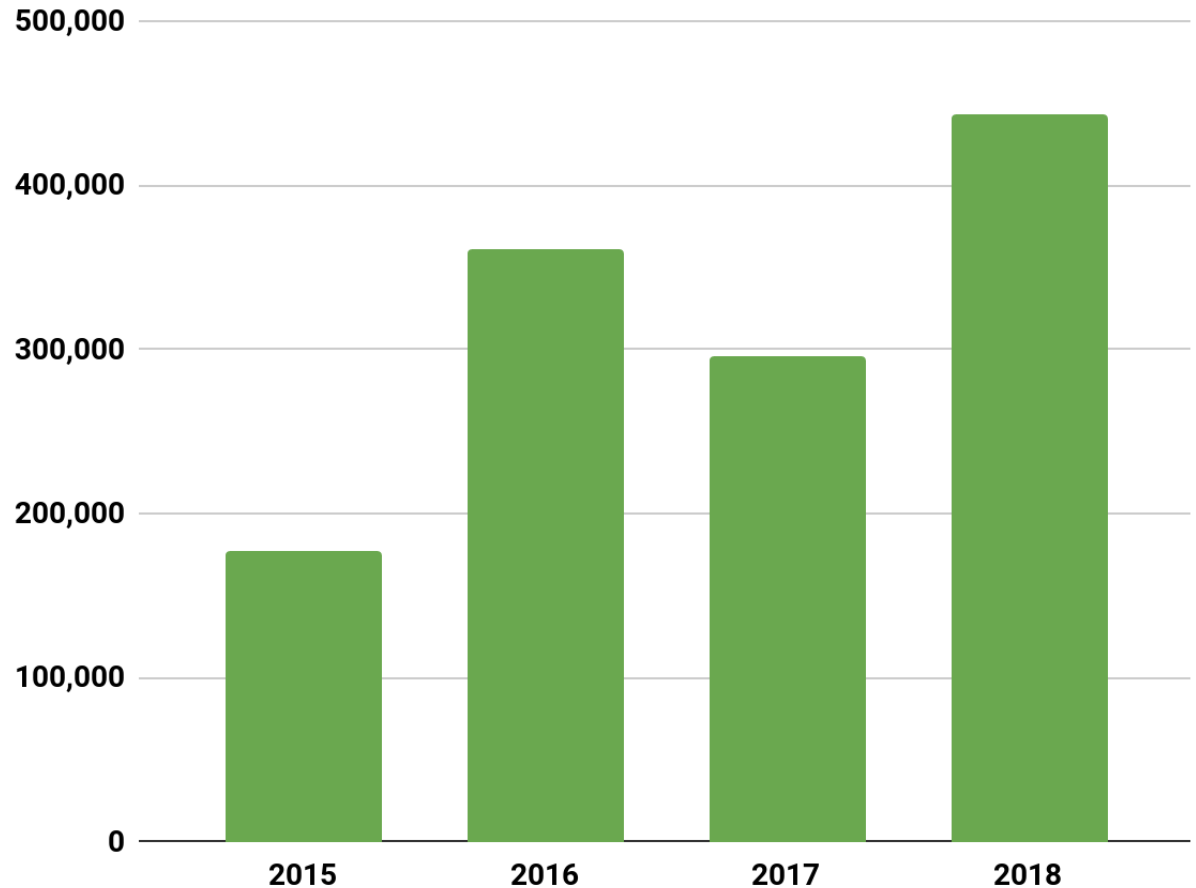
SEQUENCING TRENDS

Sequencing Data Types



SEQUENCING TRENDS: Genomics

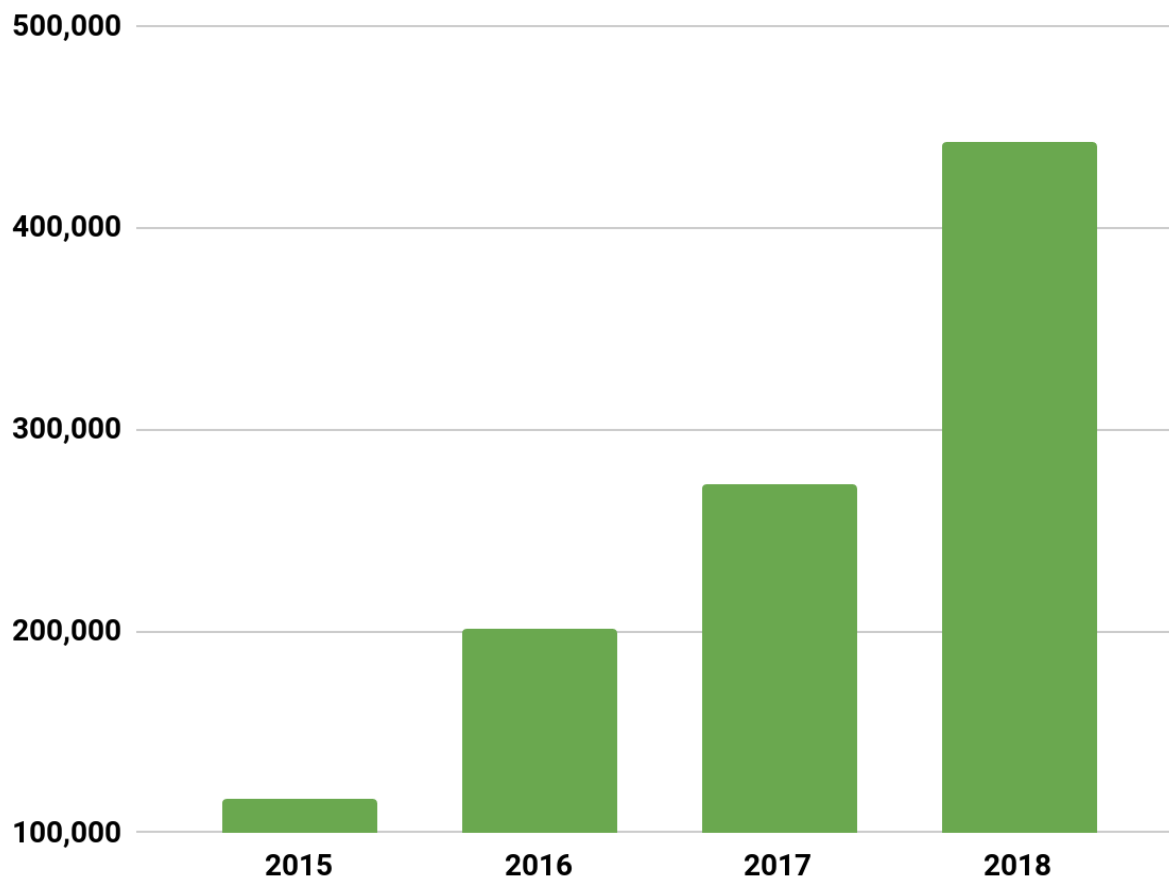
Whole Genomes Sequencing Experiments Annually*



*ENA Database

SEQUENCING TRENDS: Transcriptomics

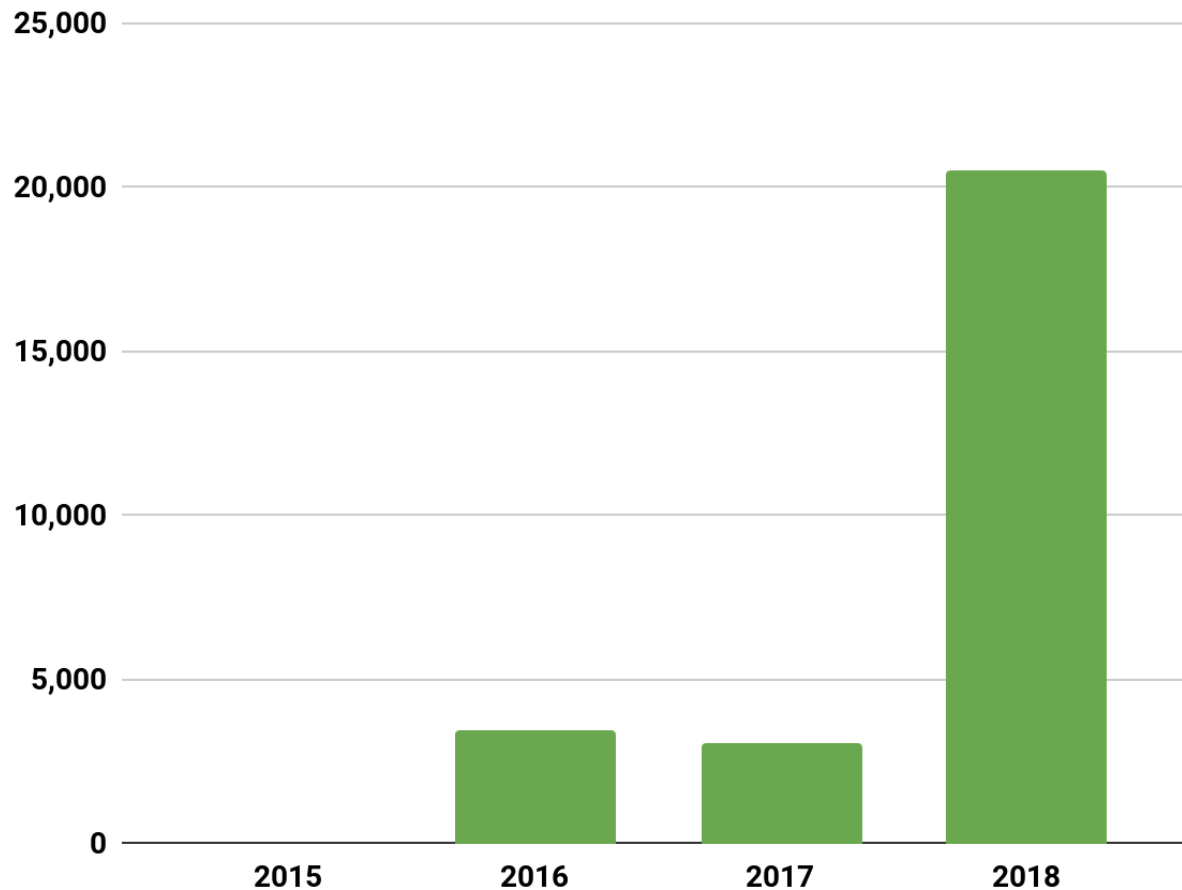
RNA-seq Experiments Annually*



*ENA Database

SEQUENCING TRENDS: Epigenomics

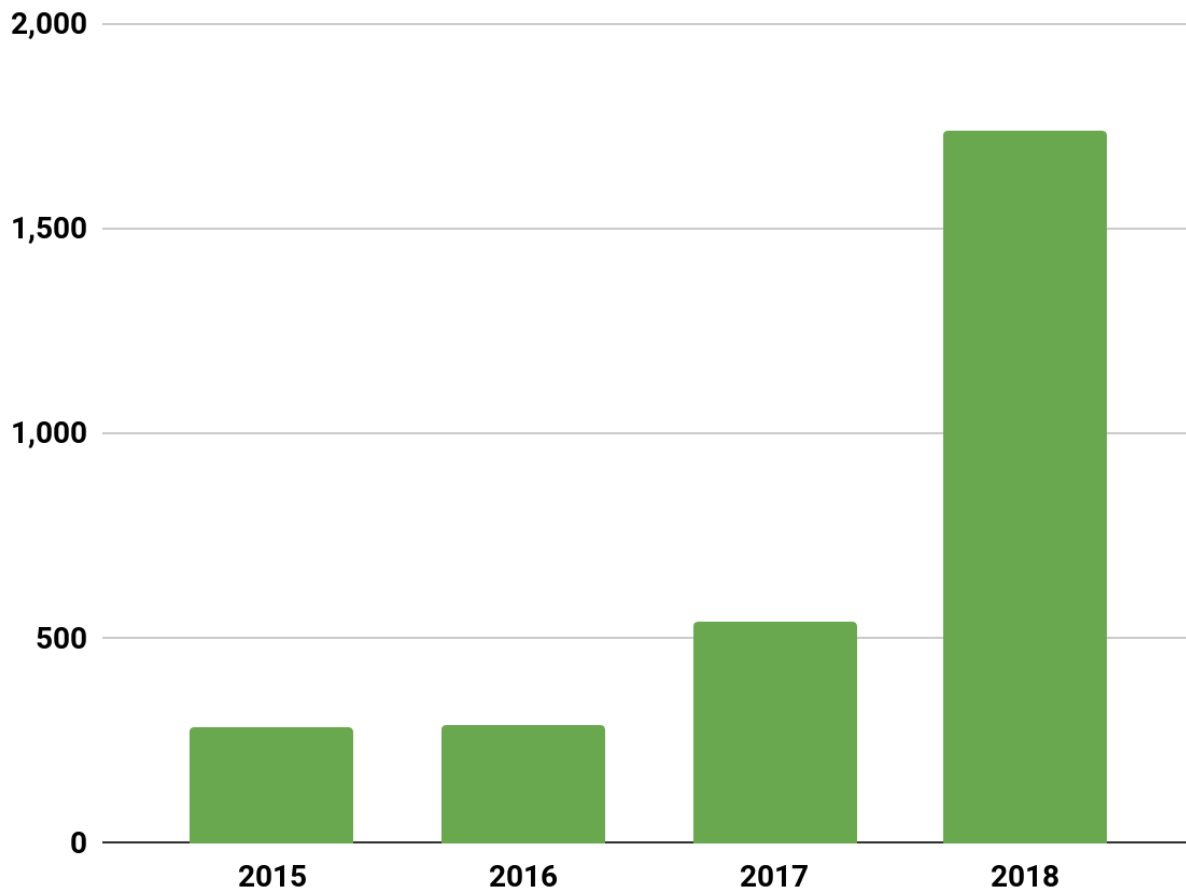
ATAC-seq Experiments Annually*



*ENA Database

SEQUENCING TRENDS: Nanopore Long Read Sequencing

MinION Experiments Annually*

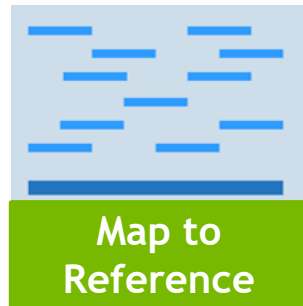


*ENA Database

The background of the slide is a dark, almost black, field. It is populated with numerous thin, light green lines that crisscross in various directions, creating a complex web-like pattern. Scattered throughout this network are several small, bright green circular dots. Some of these dots are slightly larger and more prominent than others. In the upper left quadrant, there are a few faint, larger, light blue circular shapes that appear to be out of focus or part of a different layer of the design. The overall aesthetic is technical and digital, suggesting a theme related to data, networks, or genetics.

Variant Calling

Variant Calling

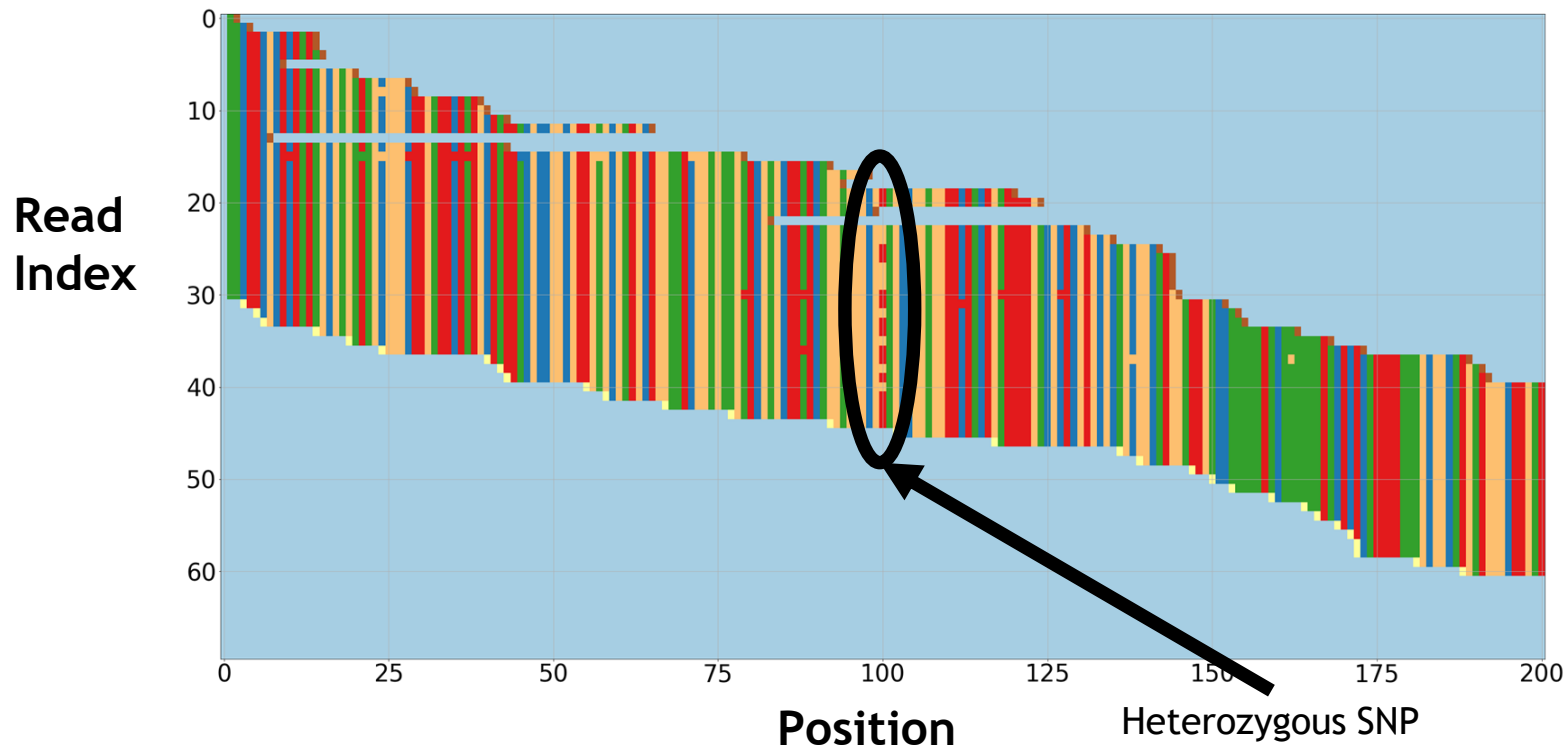


Reference	TGGATTTGAAAACGGAGCAAATGACTG
	TGGATTTGAAAACGGAGCAAATGACTG
Illumina	TGGATTTGAAAACGGAGCAAATGACTG
Reads	TGGATTTGAAAACAGAGCAAATGACTG
	TGGATTTGAAAACAGAGCAAATGACTG
	TGGATTTGAAAACAGAGCAAATGACTG
	TGGATTTGAAAACGGAGCAAATGACTG

Likely heterozygous variant

- Identify sites with potential mismatch
- True variants or instrument errors?
- SNPs or insertions or deletions?
- Heterozygous or homozygous variants?

Example Pileup Input Data



GATK Variant Calling Pipeline

Variant Calling Pipeline

Align to
Reference

Sort
Mark Duplicates
Calibrate

Call Variants

Joint Call

Filter Variants

Accelerated GATK Variant Calling Pipeline

Variant Calling Pipeline

Align to
Reference

Sort
Mark Duplicates
Calibrate

Call Variants

Joint Call

Filter Variants

Parabricks

Alignment

Preprocessing

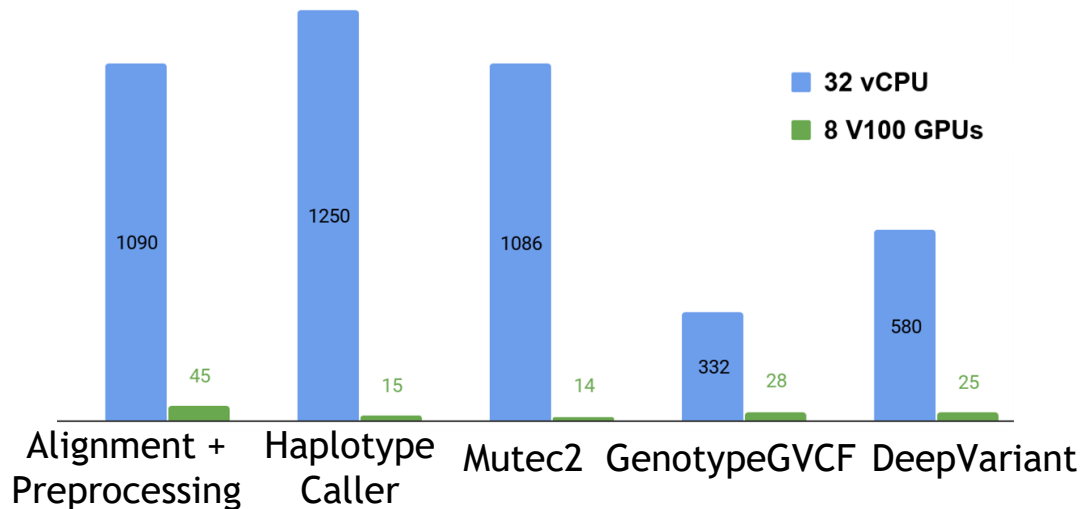
Variant Calling

Joint Genotyping

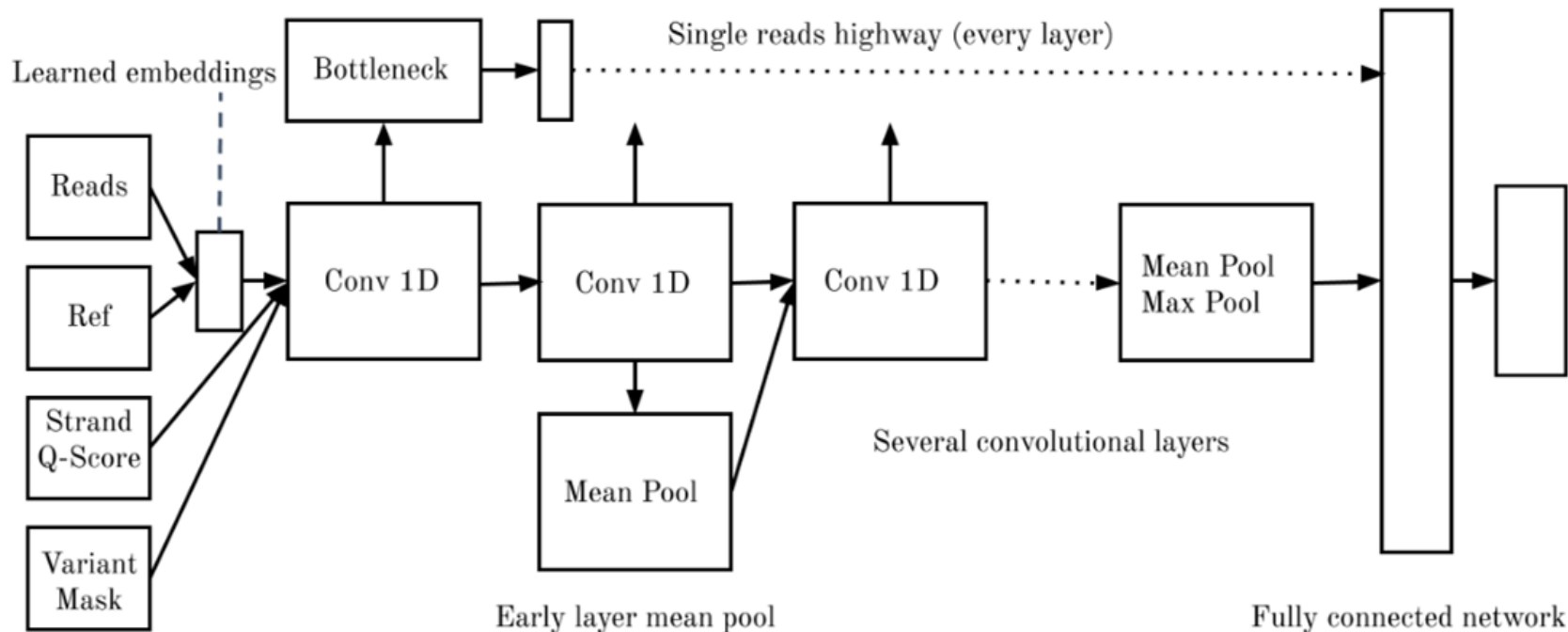
Variant
Processing

Accelerated Variant Calling Pipelines

Whole Genome Processing in Minutes

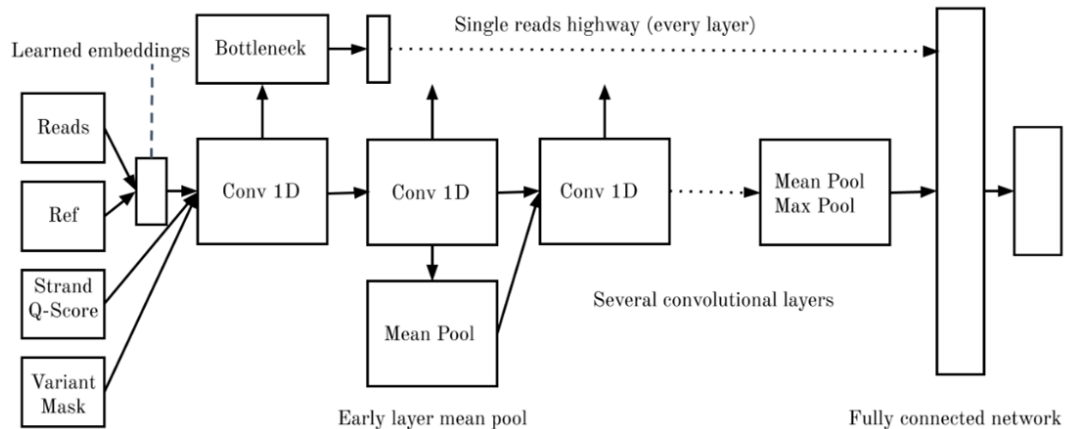


Deep Averaging Network (DAN)



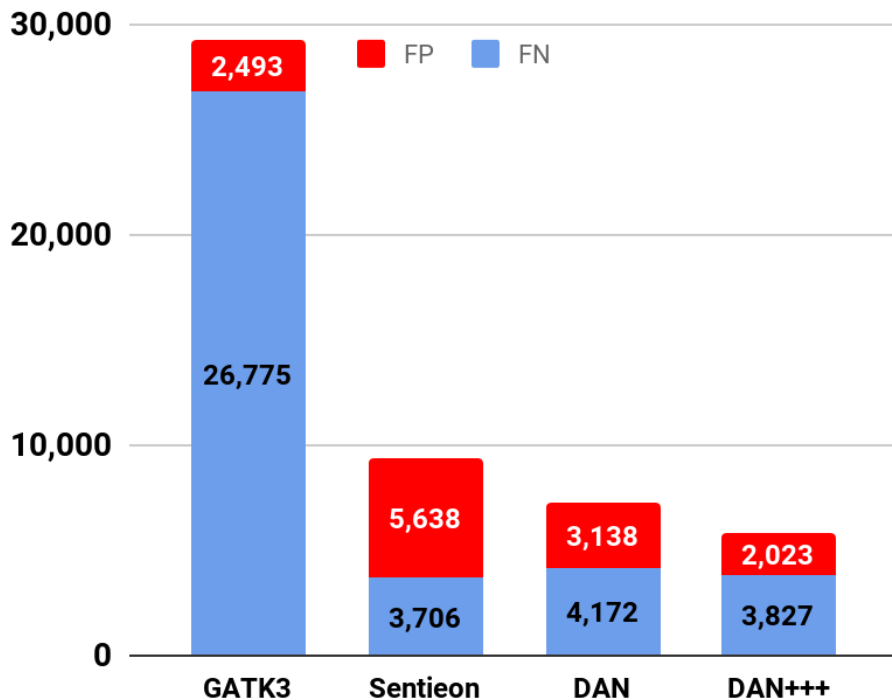
DAN Development

- PyTorch-based 1D model
- Learned embeddings of bases
- Encoding variant proposals
- Downsample easy variant candidates during training



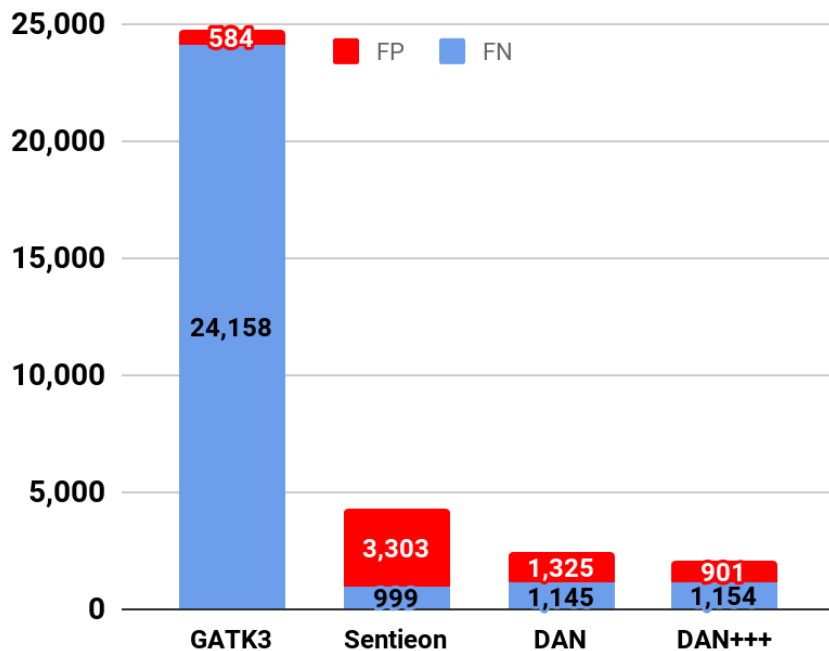
Variant Calling Errors

Total Errors on PrecisionFDA HG002

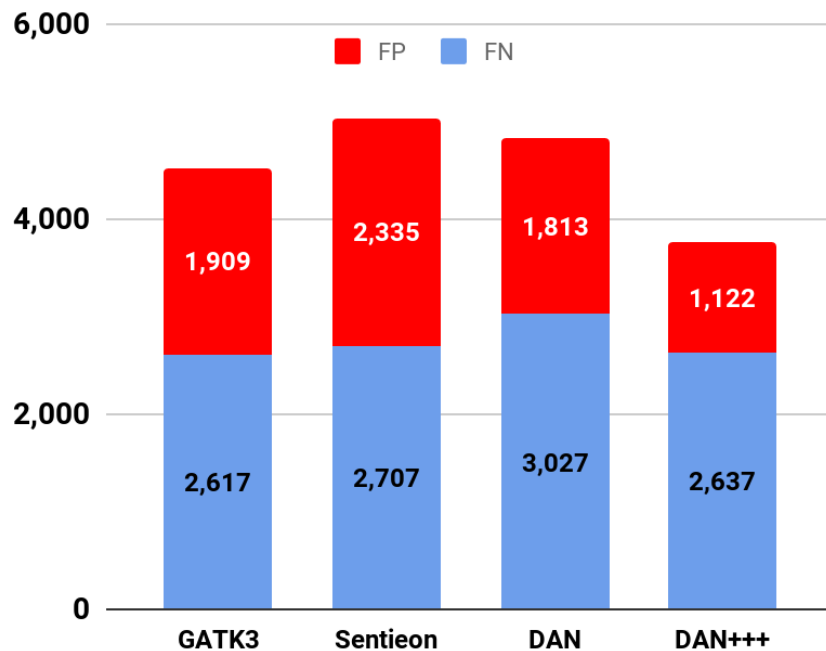


Variant Calling Error Breakdown

SNP Errors on PrecisionFDA HG002



Indel Errors on PrecisionFDA HG002



The background of the slide is a dark, almost black, field. It is populated with numerous thin, light green lines that crisscross the frame in various directions. Interspersed among these lines are several small, bright green circular dots. Additionally, there are a few larger, faint, out-of-focus blue and green circular shapes scattered across the background, giving it a sense of depth and complexity.

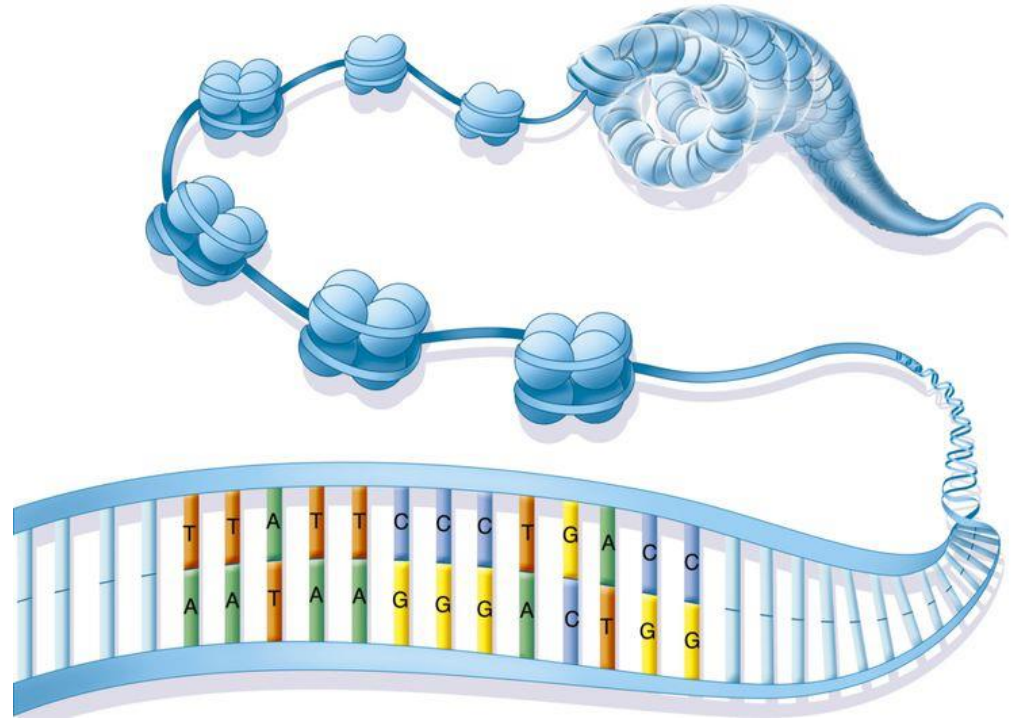
Atac Sequencing

DNA: Open And Closed

Closed DNA inactive

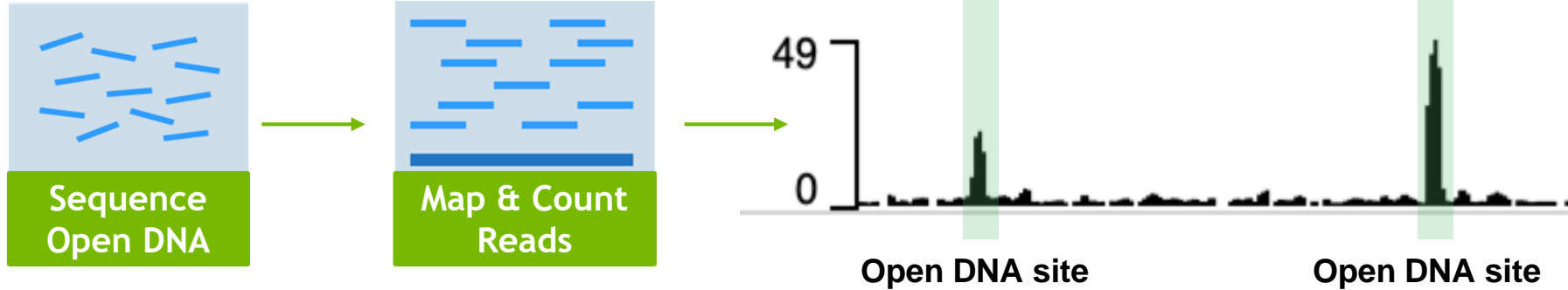
Open DNA active

Open DNA changes affect
development & disease



Atac Sequencing

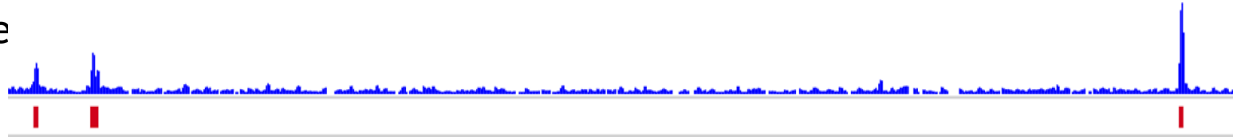
Mapping Open DNA Sites



Atac-seq Limits

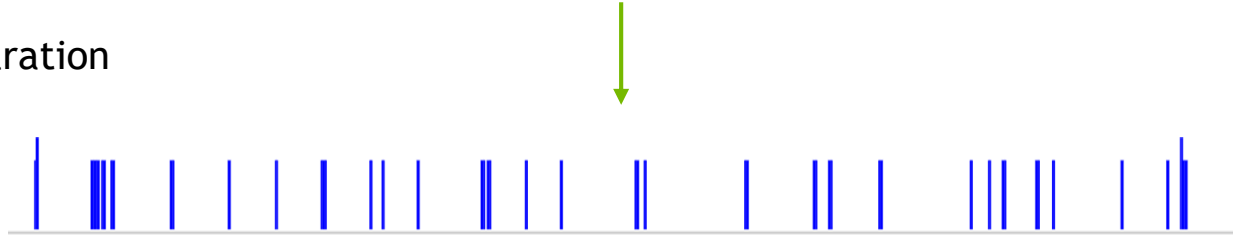
Atac-seq signal degrades in due

- Less sequencing



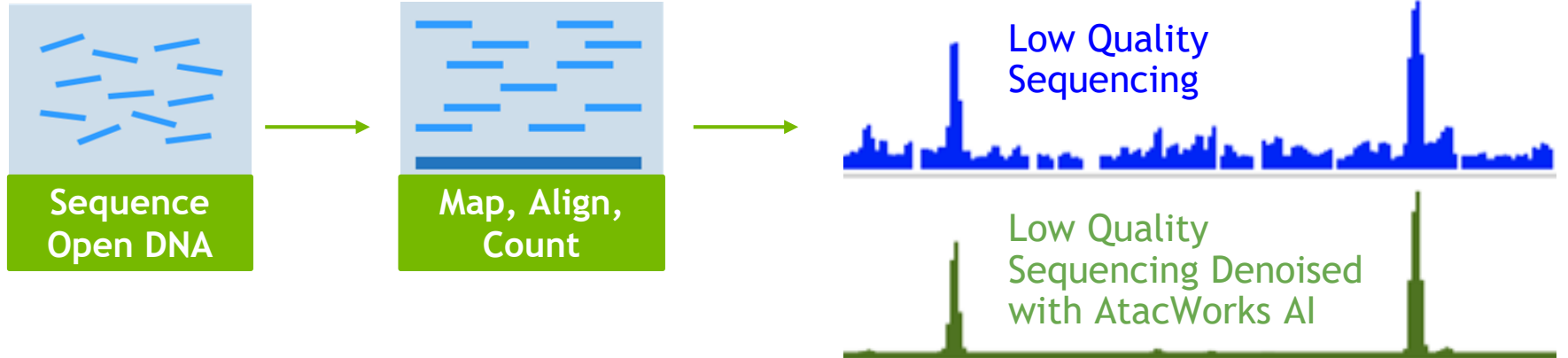
- Low quality sample preparation

- Small cell populations



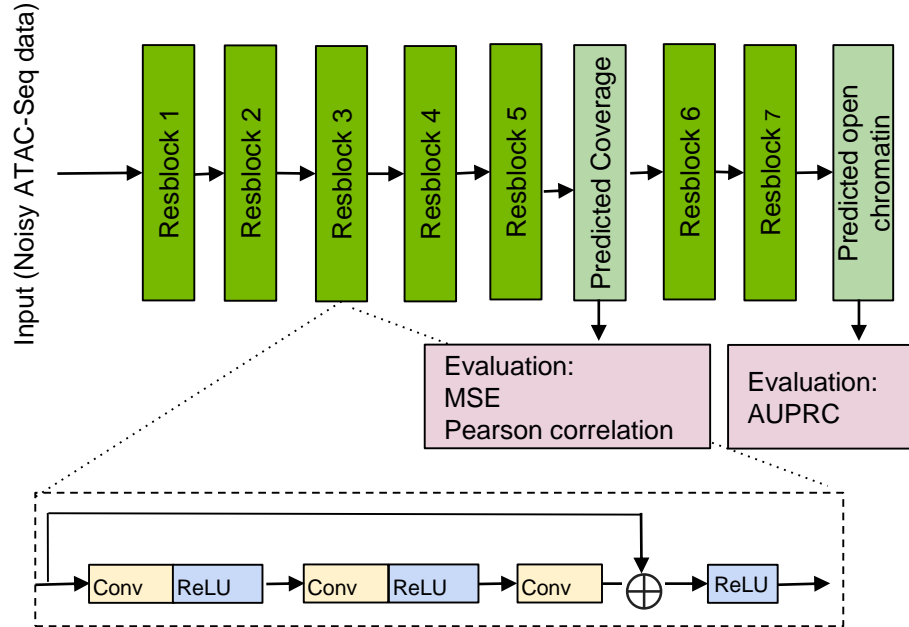
AtacWorks SDK

AI-Denoised ATAC-seq Data Processing



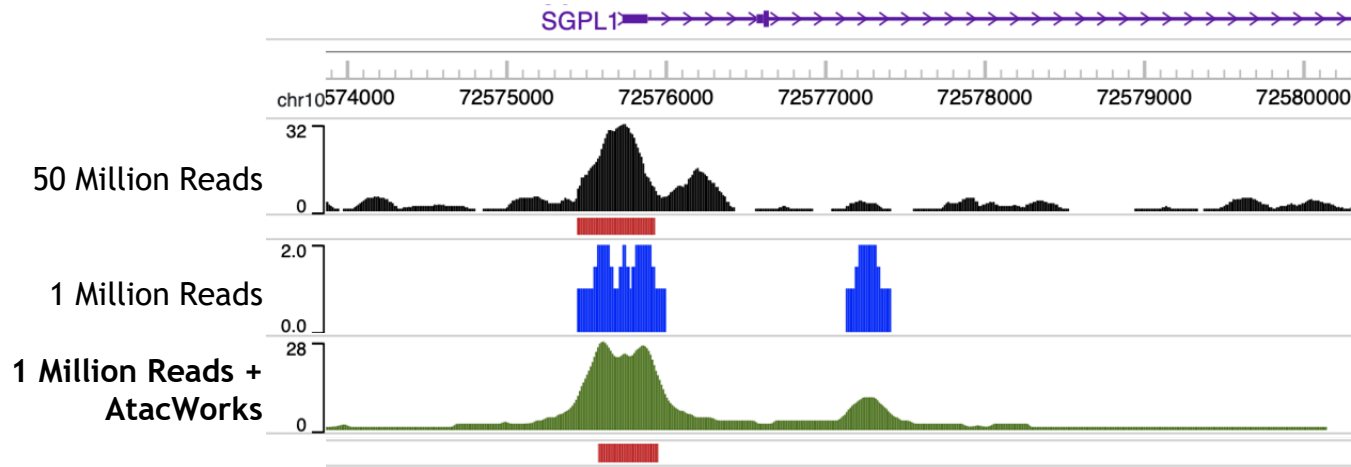
AtacWorks Model

Denoising + Open Chromatin Identification



Denoising Low Sequencing Data

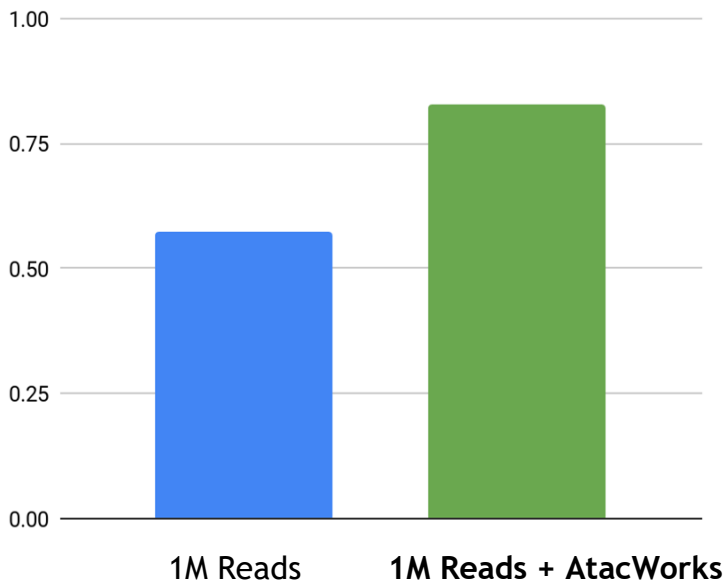
AtacWorks identifies open chromatin from low-coverage data



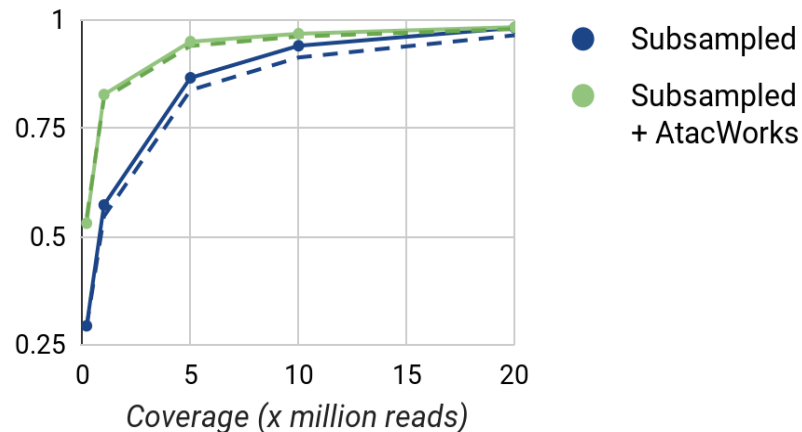
Genome-wide Sequencing Reduction

AtacWorks Reduces Sequencing Requirements 3x

Pearson Correlation with clean (50 M read) data

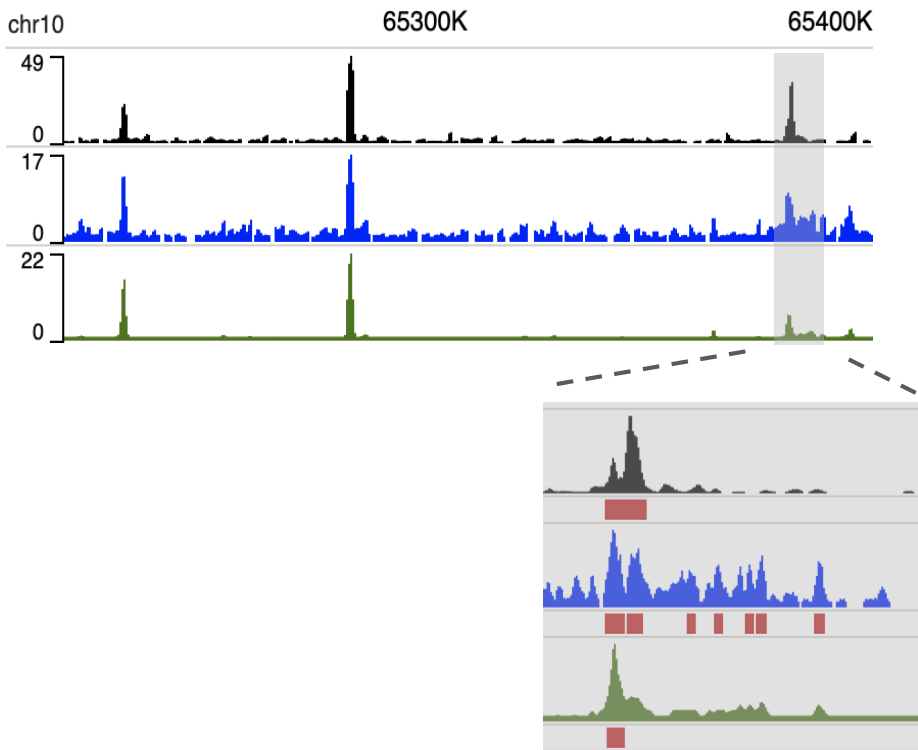
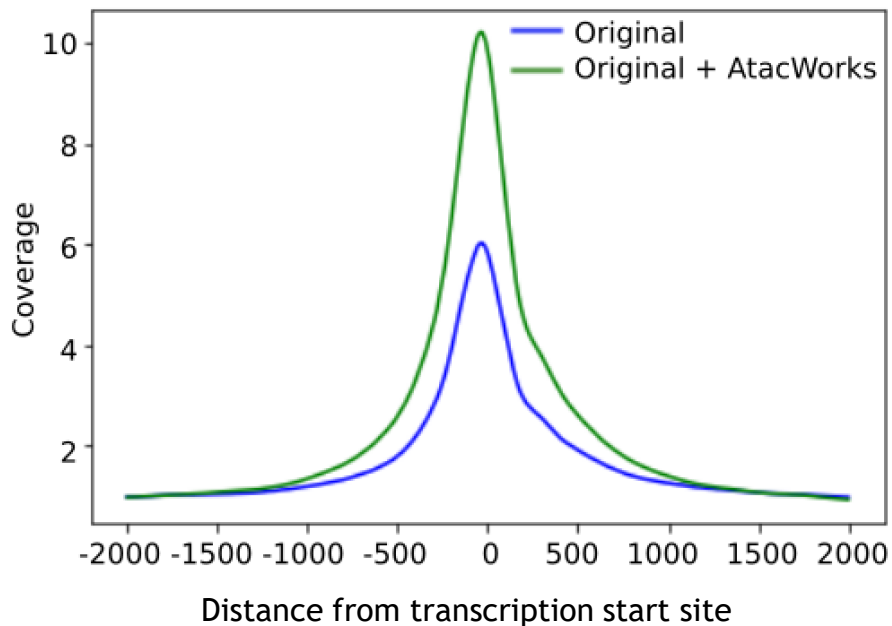


Pearson correlation with clean data (50M reads)



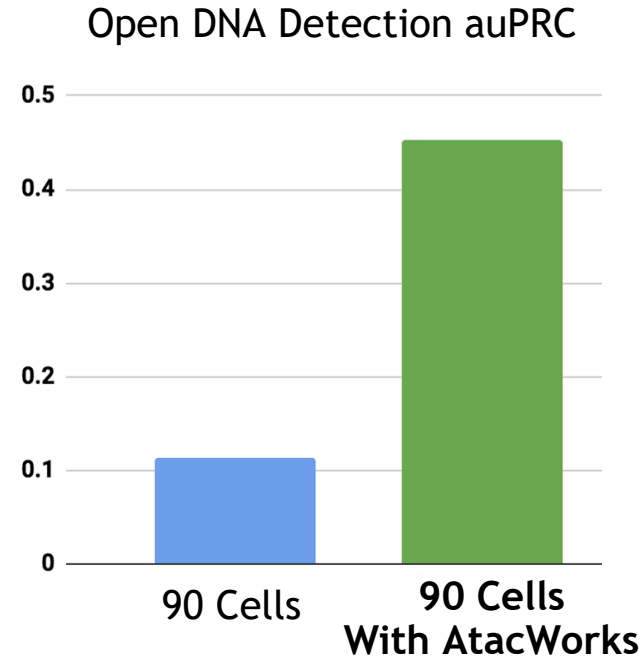
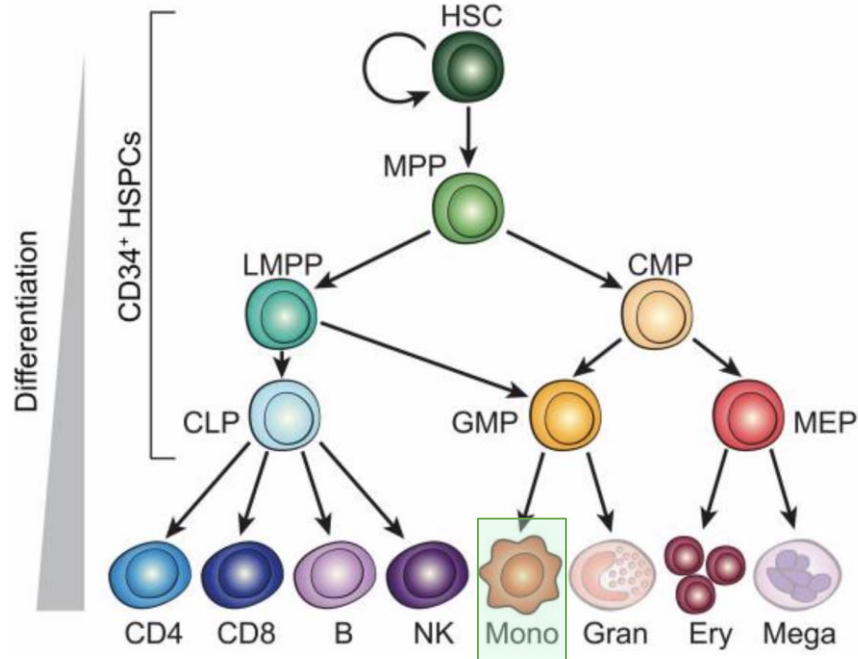
Denoising Low Quality Sample

AtacWorks improves signal-to-noise ratio in low quality samples



Denoising Single Cell Atac-seq Data

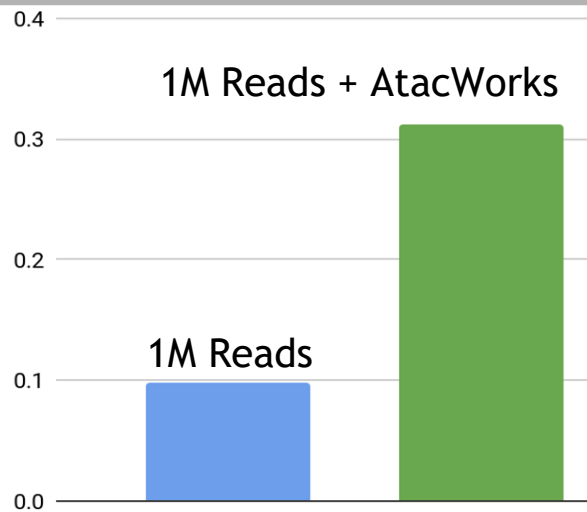
AtacWorks Improves Open DNA Detection From Few Cells



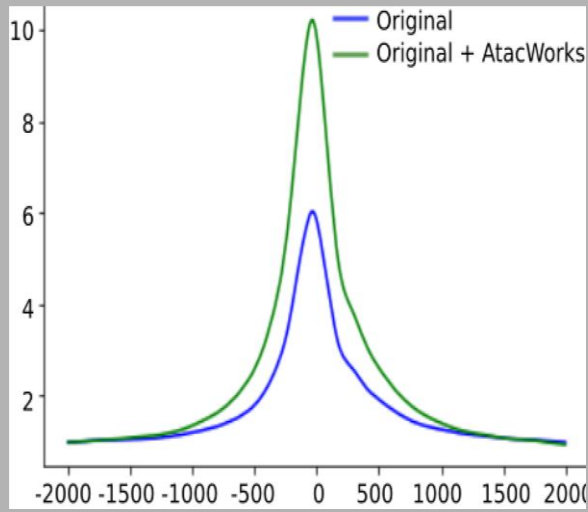
AtacWorks SDK

SDK on Clara Genomics: <https://github.com/clara-genomics/AtacWorks>

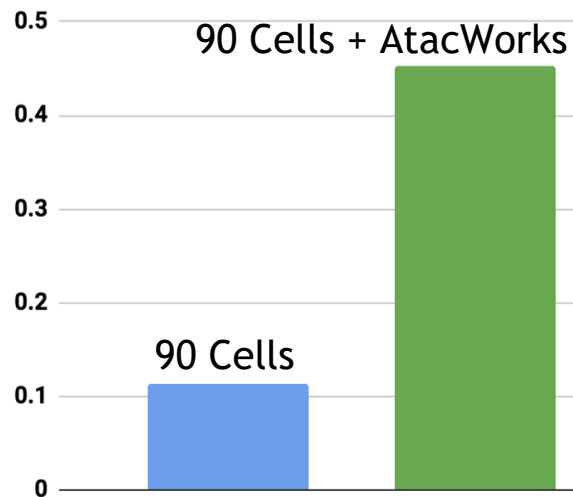
AtacWorks Preprint: <https://www.biorxiv.org/content/10.1101/829481v1>



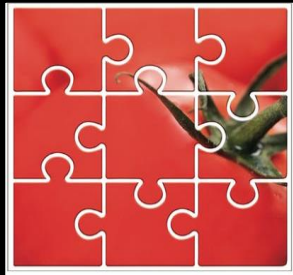
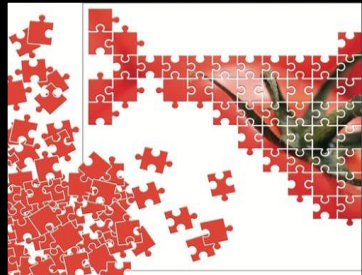
Reduce Sequencing Cost



Improve Sample Quality



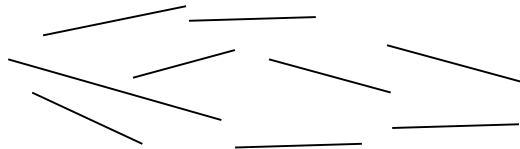
Increase Single Cell Resolution



Genome Assembly

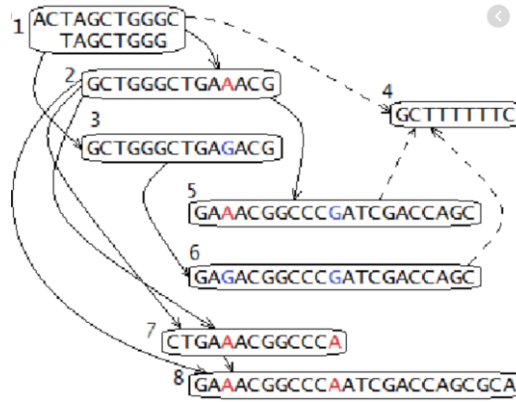
Long Read De Novo Assembly

Step 1: Mapping to detect overlaps between reads



ACTCGGTCATTCGTGCTTTATC
GCGTTATCGTCTACTTCGT

Step 2: Overlap graph traversal to generate draft genomes



Step 3: Error correction to polish genomes

Draft genome

Original reads

TGACTTCA
TCACGTCA
TGACGGCA
TGACGTCA
TGTCGCCA
AGACGTCA
TGACGTCA
consensus

Genome Assembly Workflow

Genome Assembly Pipeline

Overlap

Assemble

Align

Polish

DL Polish

MiniMap

MiniASM

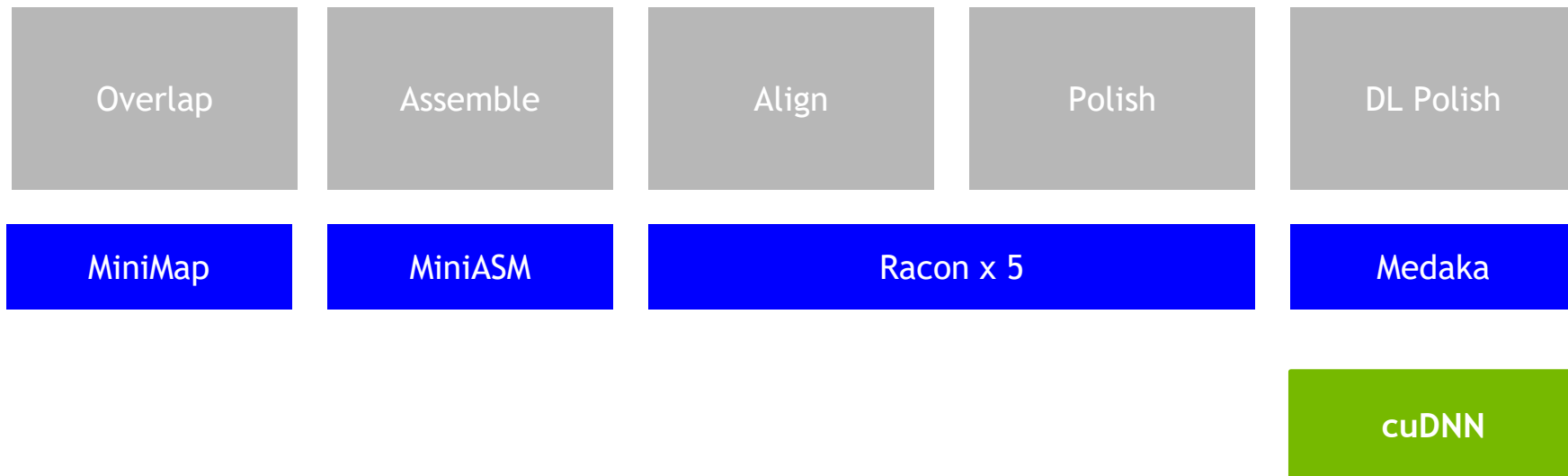
Racon x 5

Medaka

Accelerated Genome Assembly Workflow

Before ClaraGenomicsAnalysis

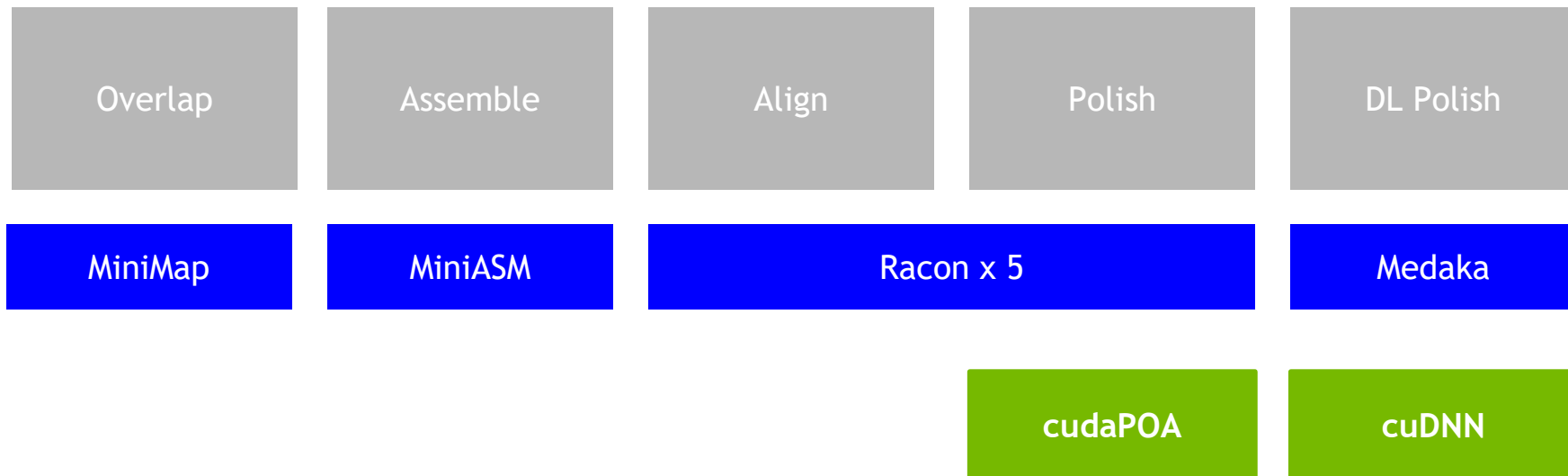
Genome Assembly Pipeline



Accelerated Genome Assembly Workflow

ClaraGenomicsAnalysis 0.1

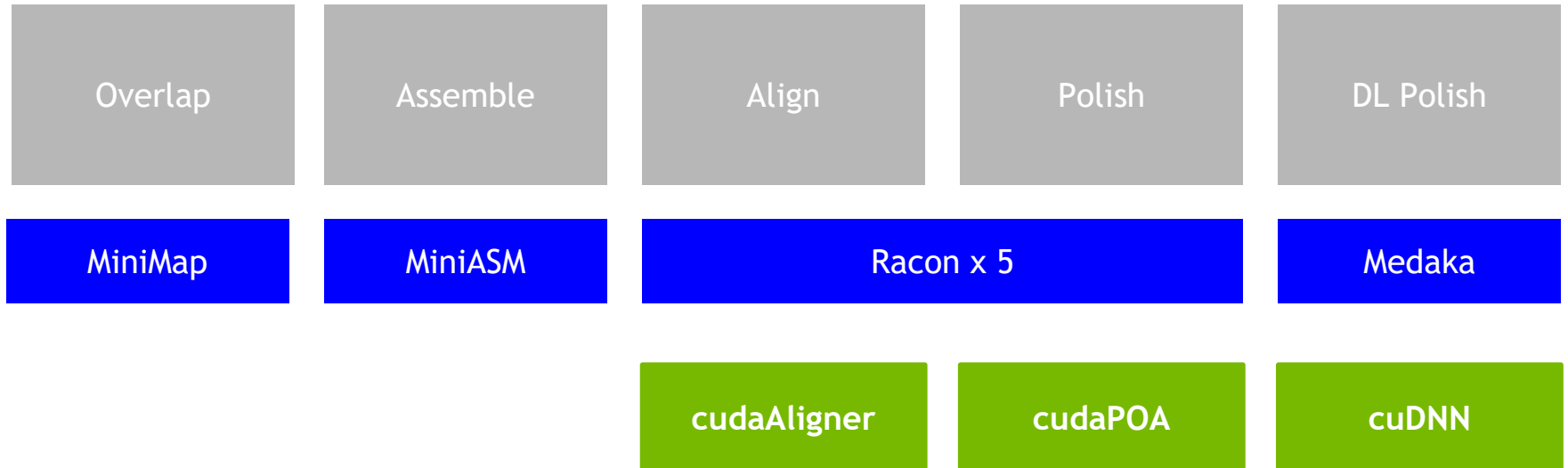
Genome Assembly Pipeline



Accelerated Genome Assembly Workflow

ClaraGenomicsAnalysis 0.2

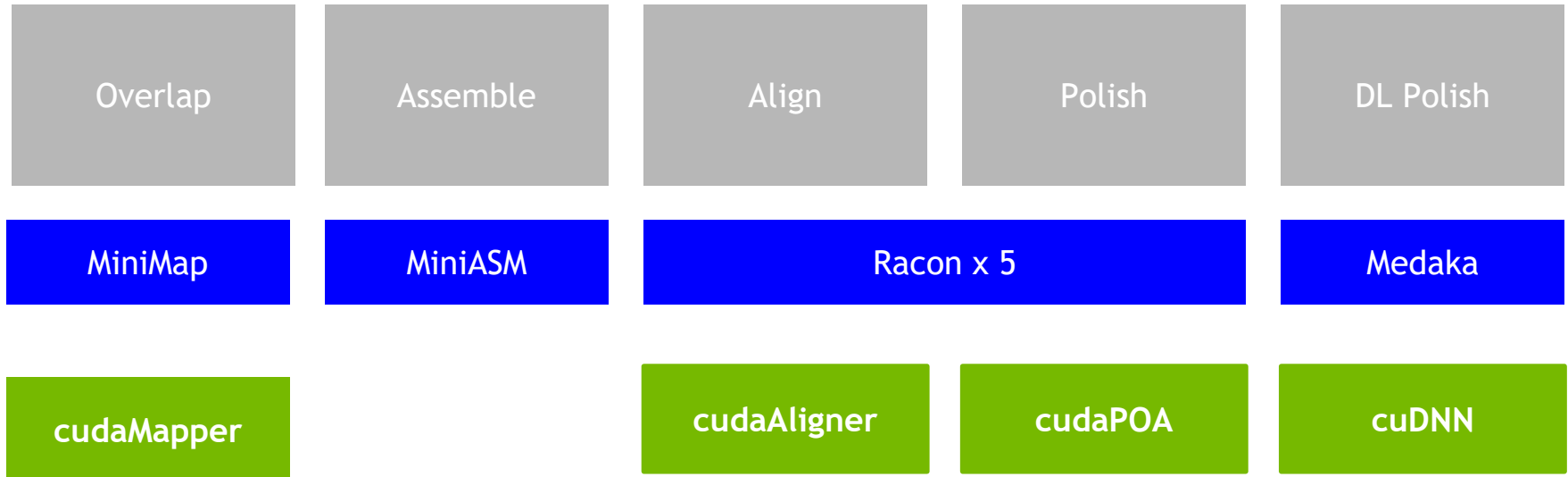
Genome Assembly Pipeline



Accelerated Genome Assembly Workflow

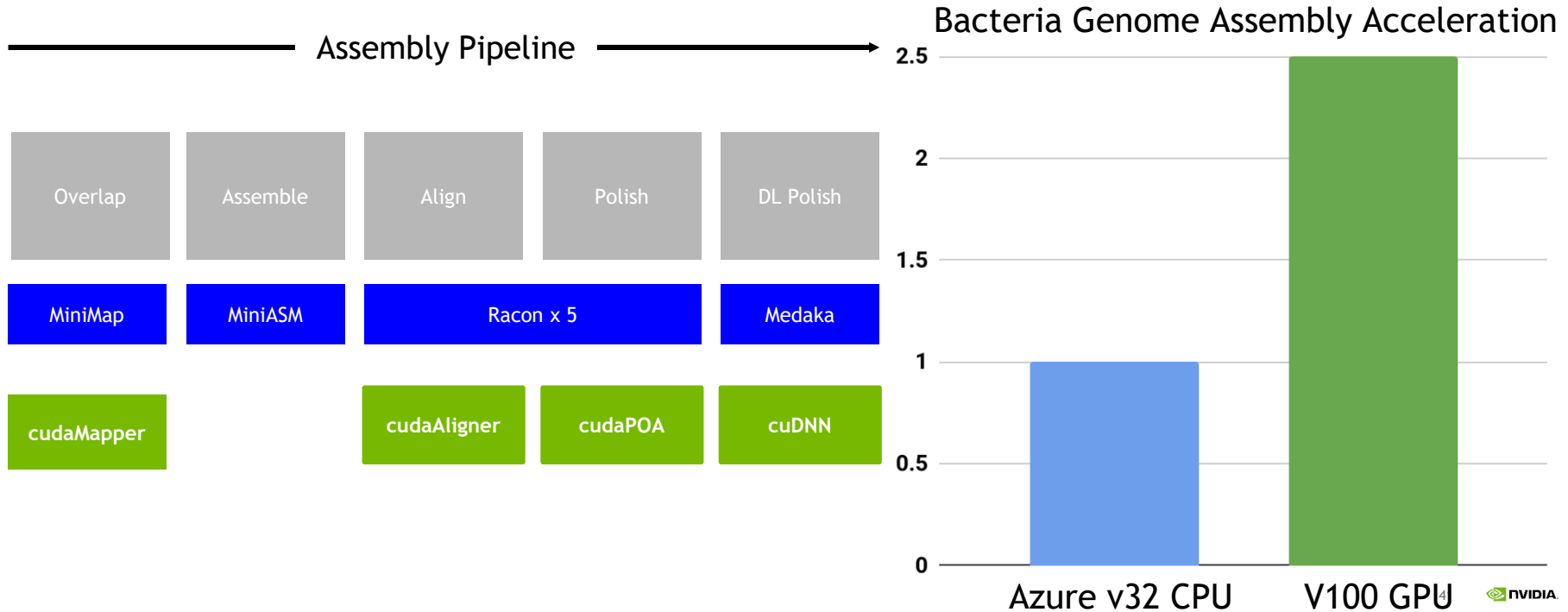
ClaraGenomicsAnalysis 0.3

Genome Assembly Pipeline



ClaraGenomicsAnalysis SDK

Enabling Accelerated Genome Assembly



CLARA GENOMICS SW

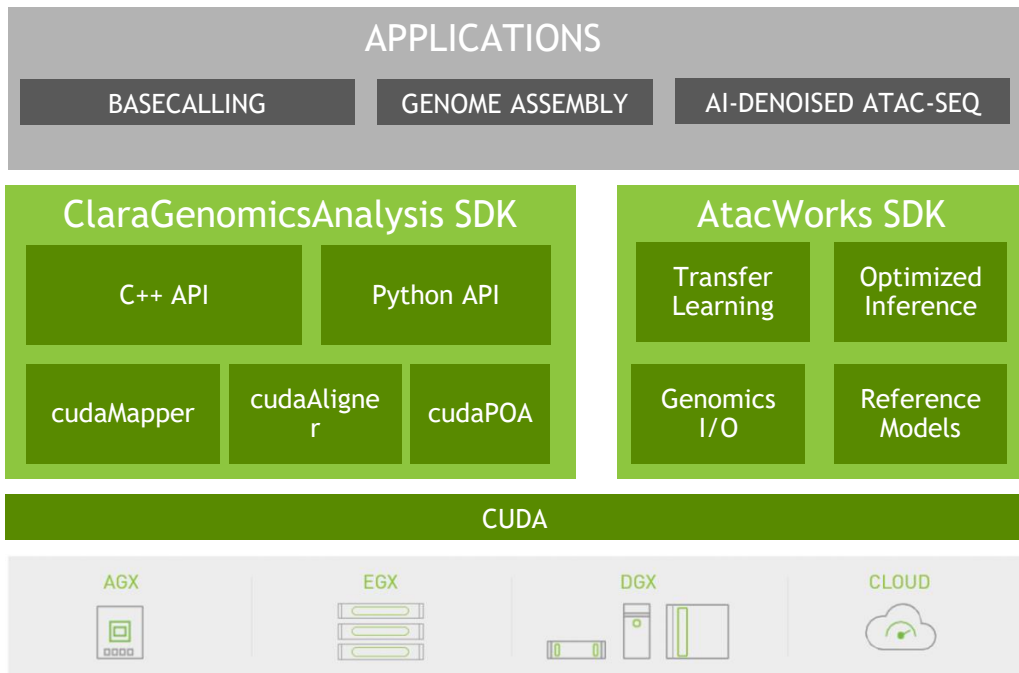
Open Source CUDA-Accelerated Sequencing Analysis Tools

Reference Applications

Integration with 3rd Party
Applications and Workflows

C++ and Python APIs

CUDA Accelerated HPC and
Deep Learning Modules



Useful Links

- Parabricks: <https://www.parabricks.com>
- ClaraGenomicsAnalysis
 - SDK on GitHub: <https://github.com/clara-genomics/ClaraGenomicsAnalysis>
 - C++ API Examples: [cudapoa](#), [cudaaligner](#)
 - Python API Examples: [cudapoa](#), [cudaaligner](#)
- AtacWorks
 - SDK on GitHub: <https://github.com/clara-genomics/AtacWorks>
 - AtacWorks Preprint: <https://www.biorxiv.org/content/10.1101/829481v1>
- 3rd party integrations:
 - Racon: <https://github.com/lcb-science/racon>
 - Raven: <https://github.com/lcb-science/raven>
 - Bonito: <https://github.com/nanoporetech/bonito>
- Additional GPU Accelerated Genomics Applications:
 - Kipoi Model Zoo: <https://ngc.nvidia.com/catalog/containers/hpc:kipoi>
 - SigProfiler: <https://github.com/AlexandrovLab/SigProfilerExtractor>



Accelerating Sequencing with GPU Computing and Deep Learning

Johnny Israeli