Breakout Session 4: Track B

Cloud Forward Data Sharing: “Limit Testing” with Long Reads at CARD

Dr. Michael Nalls
Lead, NIH CARD

National Institutes of Health
Center for Alzheimer's and Related Dementias

Cloud forward data sharing: "limit testing" with long reads at CARD

Mike A. Nalls, PhD
on behalf of the
NIH's Center for Alzheimer's and Related Dementias

DATATECNICA

Quick note …

Mike Nalls is a consultant and the Supervisory Lead for Advanced Analytics at NIH's CARD.

He is also the Managing Partner at DataTecnica LLC (DT), a data science and technology firm that competed for the contract to support this scope of work at CARD.

Additionally he is a scientific advisor and shareholder at Neuron23 Inc and Character Bioscience Inc.

National Institutes of Health
Center for Alzheimer's and Related Dementias

CARD overview

Data sharing context ...

DATATECNICA

CARD

Common issues with data.

Real impressions from researchers in the neurodegenerative disease space.

"It cost me $10k to download and reprocess the exomes from the platform. That's a barrier."

"I can't replicate that study because there is no code available."

"To build a comparable dataset it'll take >15 applications and weeks of munging/admin."

"Can you run this analysis for us, our university doesn't have the resources for compute?"

"Redundant data storage is costing us tens of thousands of dollars annually."

"I can't find any of the relevant data on this disorganized platform."

Our team designed the CARD data sharing ecosystem with the following priorities:

1. Bring the user to the data → **safety**
2. Sponsored compute + training → **inclusivity**
3. Close to real-time sharing → **fairness**
4. Improve navigation and documentation → **clarity**
5. Standardized and harmonized data → **interoperability**
6. No silos (funding scope or datatype) → **flexibility**
7. Single sign on → **accessibility***

*** = aspirational across silos to a degree**

Two platforms meet the standards for the CARD ecosystem we have laid out:

ADWB

CARD + Biowulf

AnVIL / Terra

This comes from systematic review of all available data sharing platforms.

We have developed tooling at CARD+UMC to make sharing data across these two platforms (and other repositories) as painless and seamless as possible.

DNASTACK is a newer addition to the ecosystem, strongly supporting interoperability / flexibility.

National Institutes of Health
Center for Alzheimer's and Related Dementias

Today's focus

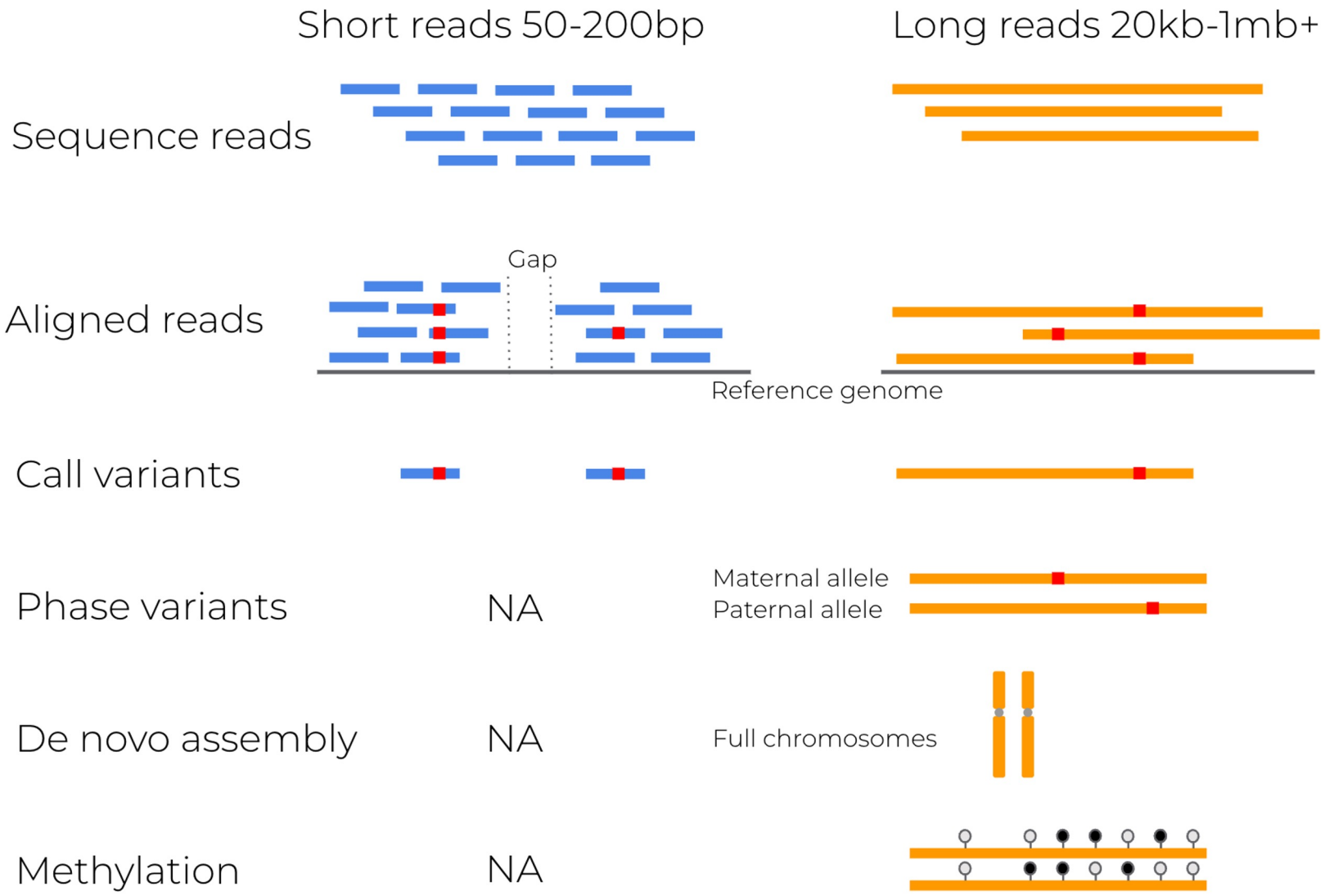
Testing limits …

Next 2 slides stolen in majority from / graciously donated by:
Cornelis Blauwendraat
Kim Billingsley
Pilar Alvarez

Long read sequencing

Structural variants are very understudied part of the genome and typically have a higher impact than “simpler” SNPs

Studying the impact of structural variants on genes and on disease at scale is now possible



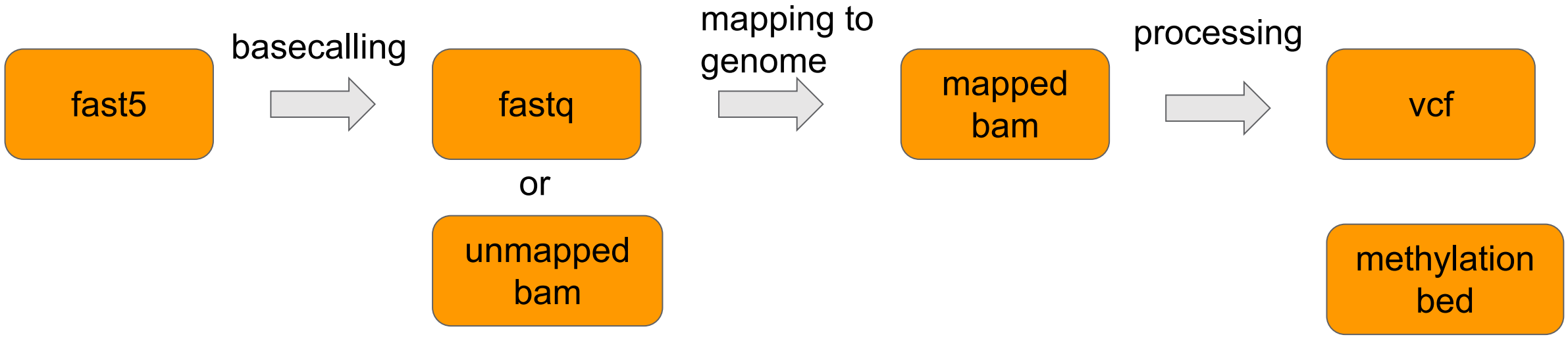
National Institute on Aging
Intramural Research Program

Data file types and sizes…

Raw data ~1TB

fastq/bam files ~100GB

vcf/bed files
<1GB



tools:

guppy

samtools
minimap2
winnowmap

CARD workflow
sniffles
modbamtools
modbam2bed

Data strategy

~1TB per sample = 1,498 CDs stacked 8 feet tall or 1 million e-books

466 out of thousands of ADRD and control samples processed

BIOWULF
HIGH PERFORMANCE COMPUTING AT THE NIH

AD Workbench

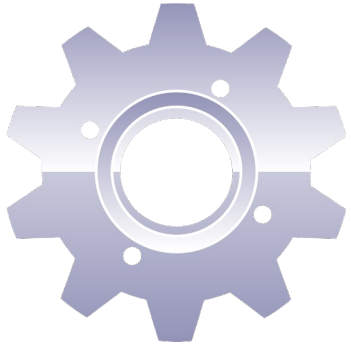
NIH STRIDES
Accelerating biomedical research

6 weeks of hybrid cloud data processing across GCP, Azure and local resources.

ADDI Alzheimer's Disease Data Initiative

Terra AnVIL dbGaP GENOTYPES and PHENOTYPES

Derived data is shared across multiple compute enabled access points using uniform ACL. Low activation energy with analysis ready data.



QC
Reduce
Harmonize

Raw data = cold

Derived data = hot

National Institute on Aging
Intramural Research Program

Resource allocation test
Processing raw data on ADDI

Compute:

- A100-SMX(40GB)
- MIG2 mode (for supper accuracy model SUP)
- 80GB RAM
- 4TB SSD
- 100 N DNA+meth ~ $30K
- Compute cost split between AnVIL/GCP and ADWB/Azure for speed!

Storage:

- Reduce to just derived data for hot storage (mapped BAM + VCF +BED)
- Requester pays cold storage for raw or just rerun assay if thinking longest term storage

Benchmarking

We've included a benchmarking app based on our tests developed and led by CARD / DTi's **Syed Shah**!

Also special thanks to **Justin Pierpoint** (ADDI/Arihidia) and **Mukta Phatak** (ADDI) for their generosity, patience and not getting mad when we may have let some Azure GPUs on fire.

Amazing team effort to get this done across ADWB, CARD, Terra/AnVIL and also biowulf.

lrseqapp.shinyapps.io/lrapp/

Inbox/IM | Calendar | Drive | GitHub | reddit | Twitter | TwitList | basecamp | GCP | GenoTracker | ADDI - Portal | F1TV

Overview | TL;DR | A100 GPU Specs | A100 GPU Performance

GPU Benchmarking for Long Read Data Sets

Guppy performance on **A100-SMX(40GB) GPU on GCP** (please see GPU specs tab for details) in **MIG2 mode (for supper accuracy model SUP)**:

- **Guppy6.1.2 Basecalling in MIG2 mode is roughly $47.25 for average RNA samples at 1.2 TB and $130.73 for average DNA samples at 1.2 TB. Please see Fig 1 (A100 Performance Tab for details).**
- **Guppy6.1.2 Basecalling+Methylation calling in MIG2 mode is roughly $147.9 for average RNA samples at 1.2 TB and $272.18 for average DNA samples at 1.2 TB. Please see Fig 2 (A100 Performance Tab for details).**
- **Guppy6.1.2 basecalling+Methylation calling in MIG2 mode for 10 RNA and 97 DNA samples cost a total of $34290.86. Please see Figure 3 and 4 (A100 Performance Tab for details) for detailed calculations.**
- **Methylation (5mC) calling (using Guppy basecalled fastq files) with minimap2 based alignment costs $52.25 for average RNA samples at 1.2 TB and $145.73 for average DNA samples at 1.2 TB.**

Please note: All $ estimates are for persistent use discount rates on GCP

MIG2 Mode:
SSD: 4TB (for sample sizes <= 3TB), **Please adjust SSD size according to you sample size.** A general rule of thumb is to allocate an additional ~1TB for guppy results.
Cost/hr: 3.75 (**for single A100 with persistent discount rate**)
Memory: 20GB/MIG
RAM: 85GB

Guppy Settings used:
chunks_per_runner: 768
disable_pings
compress_fastq
read_batch_size: 250000
q: 25000
model: Basecalling/Methcalling: dna_r9.4.1_450bps_modbases_5mc_cg_sup_prom.cfg and --bam_out flag for methcalling

National Institutes of Health
Center for Alzheimer's and Related Dementias

Thanks for having me!

Nothing is possible without help from your friends:

Hampton Leonard, Hirotaka Iwaki, Faraz Faghri, Dan Vitale, Kristin Levine, Ziyi Li, Syed Shah, Michael Ta, Andy Henrie, Lietsel Jones Shannon Ballard, Anant Dadu, Chelsea Alvarado, Nicholas Johnson, Zhenbang Wu, Gracelyn Hill, Shahroze Abbas, Cory Weller (DTi)

Andrew Singleton, Cornelis Blauwendraat, Kim Billingsley, Pilar Alvarez, Mat Koretsky, Mary Makarious, Sara Bandres-Ciga, Nikki Washeka, Julia Stadler, Mark Cookson, Michael Ward, Dan Ramos, Andy Qi, Julie Lake, Lana Sargeant, Laurel Screven, Caroline Pantazis (CARD / LNG)

Jennie Larkin, Mette Peters, Mike Griswold, Caroline Worley Solzberg, Juliana Acosta-Uribe, Jen Yokoyama, Nick Cochran, James Olzmann, Martin Kampmann (external friends)

Entire DTi, CARD, LNG, DEMON, ASAP, CZI, BD2 and GP2 teams!

Mike A. Nalls, PhD (mike@datatecnica.com)

@mike_nalls

DATATECNICA