CARD long-read seq project using the cloud as only option

February 25, 2022

Cornelis Blauwendraat

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

General aims

Create an easy accessible structural variant reference dataset of Alzheimer's Disease and Related Dementias (ADRDs) including ~4000 individuals from multiple ancestries

With this data we expect to:

- Assess the role of structural variants in ADRDs
- Resolve complex regions of interest in ADRDs (*MAPT*, *GBA, APOE* etc)
- Assess the impact of structural variants on gene expression in health and disease
- Investigate methylation patterns across samples and diseases

Long read sequencing technology

- Comparison with short reads

  ~100-200bp vs <100kb reads

- Size of long read data

  30Gb vs ~1TB per human genome

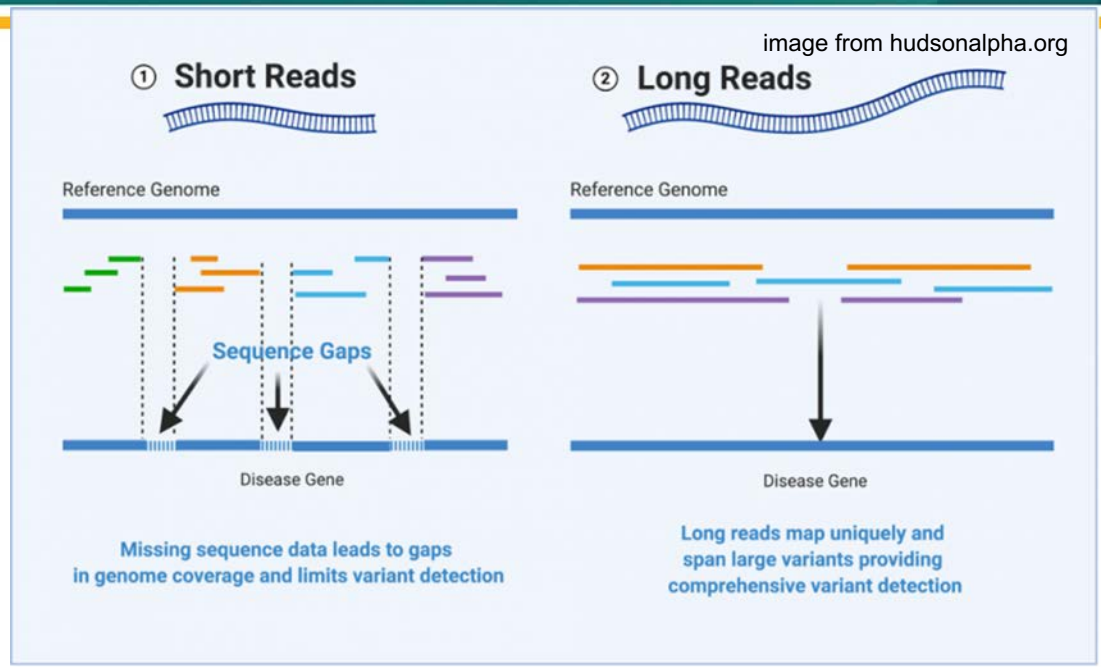
Collaborative workspace solutions needed

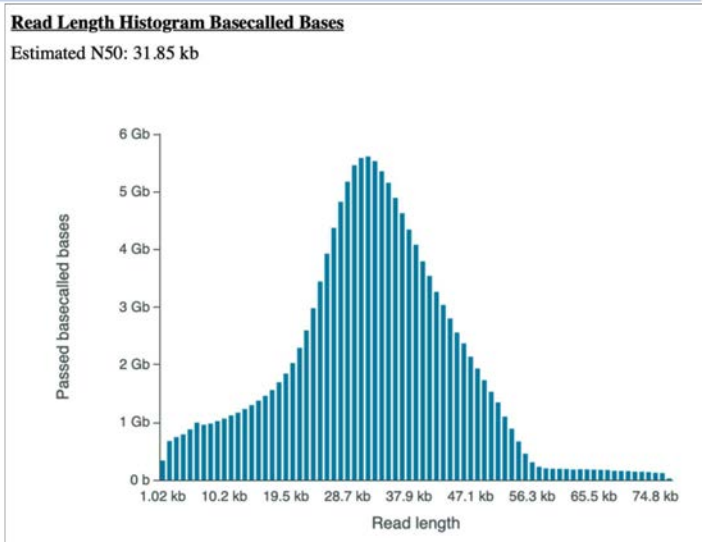
no one can and want to store this files locally

Need for centralized pipeline that is transparent and easy to use (also given the size and computation costs)

no one wants to process these files several times

image from hudsonalpha.org





General data workflow - external basecalling

Sequencing

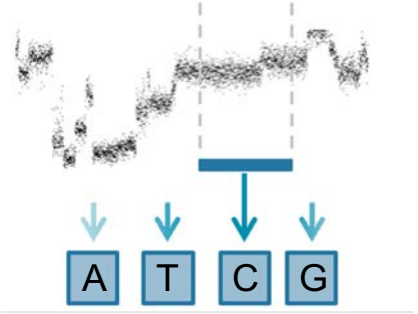
NIH National Institutes of Health
Center for Alzheimer's Disease and Related Dementias

Oxford NANOPORE Technologies

~1TB per sample

Locally

Processing



FAST5 (raw data)

FASTQ format

Analysing

Run pipelines including:

- Aligning to genome (hg38 and newer version of genome)
- Variant detection SNV
- Variant detection SV
- Phasing of Structural variants
- Localized assembly
- Whole-genome de novo assembly

Consortium level access

Broadly sharing with research community

Public access Roughly ~5000TB

Data processing (real consortium effort)

Pipelines are developed by leaders in the field and harmonized with other population long read studies already.

1. Mapping

NIH National Human Genome Research Institute

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

NIH NATIONAL CANCER INSTITUTE

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

2. Variant detection SNV

Baylor College of Medicine

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

3. Variant detection SV

Baylor College of Medicine

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

4. Methylation calling

JOHNS HOPKINS BIOMEDICAL ENGINEERING

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

5. Phasing of Structural variants

Baylor College of Medicine

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

6. Variant harmonisation

Baylor College of Medicine

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

7. Localized assembly

CHAN ZUCKERBERG INITIATIVE

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

8. Whole-genome de novo assembly

CHAN ZUCKERBERG INITIATIVE

NIH National Human Genome Research Institute

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Data storage/sharing/access

Data will be very large (raw estimates for 4000 samples).

Assuming ~1TB per sample (plus analysis ready files) => ~5000TB means:

$~100K storage cost per month on commercial cloud provider or
$~10K storage cost per month on commercial cloud provider Archive (cold storage)

Additionally clear need for consortium access prior to public release in order to jointly analyse (in the same space) data and harmonize pipelines (save costs). Rough costs of processing data on commercial cloud provider $~100 per sample (400K for full dataset).

Local storage is just not worth the investment anymore for long-read or any large scale sequencing projects, cloud is the future and key for success here. Data sharing wise => no one can and should be downloading these amounts of data.

Searching for a cloud provider/platform

Needs:

- Established costs efficient platform
- Safe data storage
- Consortium level access possible prior to public release
- Easy data access => Single sign on
- Bringing research to data and no versioning or data downloading/leakage
- Central authentication via dbGAP (established widely used method)
- Broad user base so data gets shared widely after made public
- Locality lock (for potential non-US samples)
- Cloud interoperability (able to tailor cloud option to user preferences)

Searching for a cloud provider/platform

Needs:

- Established costs efficient platform ✓
- Safe data storage ✓
- Consortium level access possible prior to public release ✓
- Easy data access => Single sign on ✓
- Bringing research to data and no versioning or data downloading/leakage ✓
- Central authentication via dbGAP (established widely used method) ✓
- Broad user base so data gets shared widely after made public ✓
- Locality lock (for potential non-US samples) ✗ *In development*
- Cloud interoperability (able to tailor cloud option to user preferences) ✗ *In development*

AnVIL

https://anvilproject.org/

General data workflow - hybrid environment

Sequencing

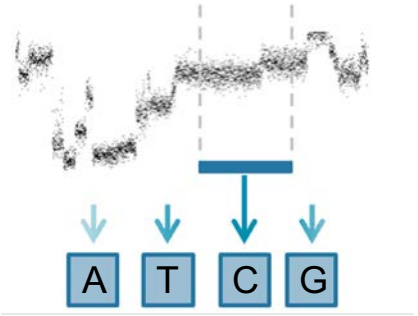
NIH National Institutes of Health
Center for Alzheimer's Disease and Related Dementias

Oxford NANOPORE Technologies

~1TB per sample

Locally

Processing



FAST5 (raw data)

FASTQ format

Analysing

Run pipelines including:

- Aligning to genome (hg38 and newer version of genome)
- Variant detection SNV
- Variant detection SV
- Phasing of Structural variants
- Localized assembly
- Whole-genome de novo assembly

Consortium level access

BIOWULF AT THE NIH

AnVIL

Broadly sharing with research community

Public access

Roughly ~5000TB

Questions/comments/ideas?