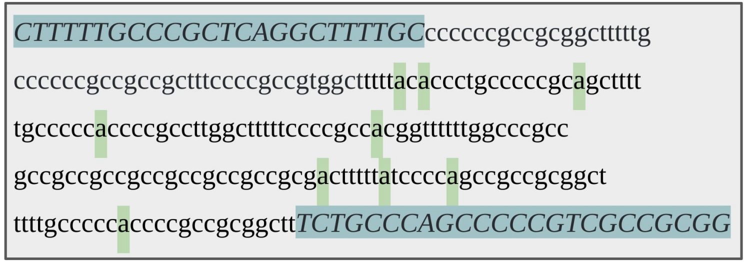
Machine Learning with “Tiled” Human Genomes using Microsoft Azure and Arvados



CTTTTTGCCCGCTCAGGCTTTTGCccccccgccgcggctttttg
ccccccgccgccgctttccccgccgtggctttttacaccctgcccccgcagctttt
tgccccca ccccgccttggctttttccccgccacggtttttggcccgcc
gccgccgccgccgccgccgccgcgactttttatccccagccgccgcggct
ttttgcccccaccccgccgcggcttTCTGCCCAGCCCCCGTCGCCGCGG

Sarah Wait Zaranek, Alexander Wait Zaranek
Curoverse Research

Agenda

Introduction to Genomics and Precision Medicine

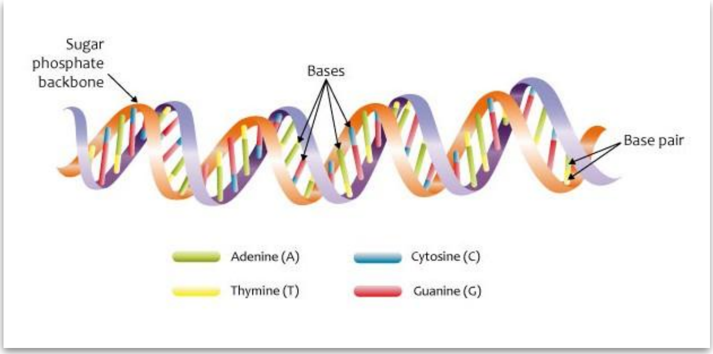
Arvados Platform

Basics and Benefits of Tiling

Machine Learning using Arvados and Tiled Data

Genomics 101

- Human DNA has 6 billion bases
  - Bases are the building blocks of DNA (A,G,C,T)



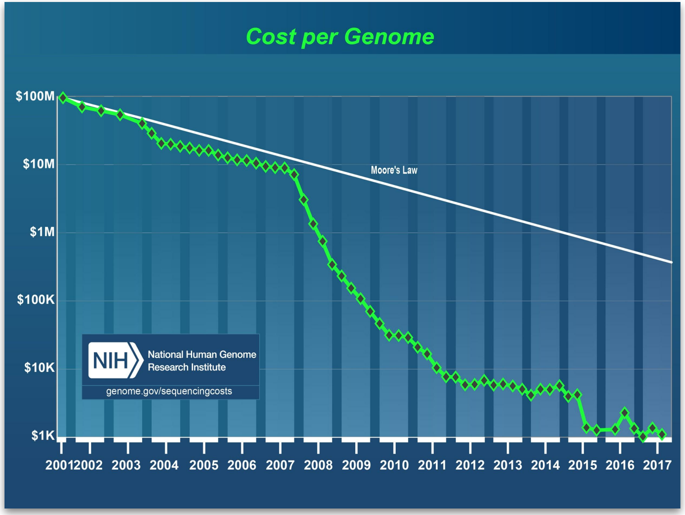
- DNA analysis can provide insights about health, behavior, and other traits
  - Large majority of DNA is shared across all humans
  - Genetic variations, or variants, are the differences
  - DNA sequencing identifies an individual's variants by comparing to a reference genome

- WGS (Whole Genome Sequencing)
  - Genetic tests usually characterize only one gene (or just specific parts of one gene)
  - SNP arrays/microarrays are ~0.1% (or less) of a genome
  - WGS characterizes the entire genome

Precision Medicine

- Precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person."
  -- http://pged.org/

- Whole Genome Sequencing (WGS) is rapidly becoming more inexpensive (~$1,000) and accessible allowing precision medicine to become a reality



Genomics and Machine Learning

- Looking at relationship between genome and phenotype

- Phenotype: physical characteristic including visible characteristics like eye color, current health conditions, health history, and general behavior

- For drug discovery, target identification, discovery of new risk factors, diagnostics, personalized treatment, and discovery of protective variants

Challenges with Precision Medicine

- Scientists and physicians struggling to analyze these large, high-dimensional datasets
  - Many patients want access to and more control of their own data
  - Data are physically distributed and difficult to move
  - Analysis is time consuming and algorithmically challenging
  - Regulatory and/or legal barriers
  - Privacy concerns

- We created tiling and use Arvados running on Microsoft Azure to help with these large data challenges

Agenda

Introduction to Genomics and Precision Medicine

Arvados Platform

Basics and Benefits of Tiling

Machine Learning Results Arvados and Tiled Data

Arvados Platform

- An open source platform for managing and analyzing biomedical big data
- Runs anywhere
  - Supports running in the cloud (Azure, AWS, GCP) on as well as on premise
- Auto-scaling of compute resources
  - Scales compute resources dynamically on the cloud
- Large scale
  - Single cluster can store petabytes of data and use thousands of cores of compute simultaneously

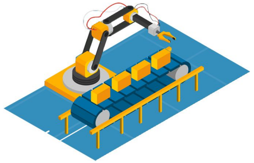
The Arvados Community

- Wide range of organizations including very large pharmaceutical companies, genomics startups, CROs, and universities
- Installations on 4 continents
- Largest single Arvados cluster manages well over a petabyte of data
- Routinely run computations that use many thousands of simultaneous CPU cores spread out over hundreds of machines

Arvados Core Components: Keep

- Guarantees retrieval of gigabytes to petabytes of files
  - Uses content addresses
  - Automatic deduplication
  - Very efficient data management
- Backed by object storage/S3 or traditional GNU/Linux filesystem
- Users and code manipulate *collections*
  - Virtual folders
  - Cheap to create, edit and delete
  - Allows for fine grain permission management

Keep File Manager



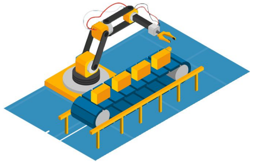
Crunch Workflow Manager

Microsoft Azure

Arvados Core Components: Crunch

- Ensures consistent reproducibility of complex computational workflows
- Maintains an automated provenance chain
- Jobs run inside Docker
- Inputs come from Keep, and outputs are stored in Keep
- Smart about job re-use

Keep File Manager



Crunch Workflow Manager

Microsoft Azure

Common Workflow Language (CWL)

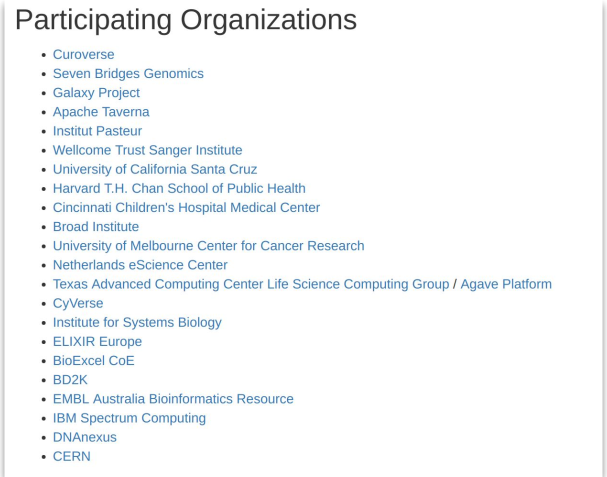
- Community developed open standard for describing computational data-analysis workflows

- Native workflow language for Arvados

- Designed to makes workflows portable and scalable across a variety of software and hardware environments

- Focused particularly on serving the data-intensive sciences (e.g. Bioinformatics, Astronomy)

COMMON WORKFLOW LANGUAGE

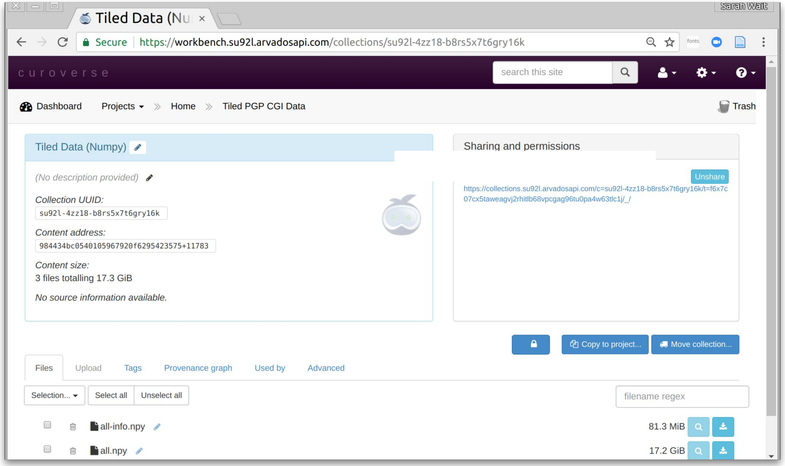


Participating Organizations

- Curoverse
- Seven Bridges Genomics
- Galaxy Project
- Apache Taverna
- Institut Pasteur
- Wellcome Trust Sanger Institute
- University of California Santa Cruz
- Harvard T.H. Chan School of Public Health
- Cincinnati Children's Hospital Medical Center
- Broad Institute
- University of Melbourne Center for Cancer Research
- Netherlands eScience Center
- Texas Advanced Computing Center Life Science Computing Group / Agave Platform
- CyVerse
- Institute for Systems Biology
- ELIXIR Europe
- BioExcel CoE
- BD2K
- EMBL Australia Bioinformatics Resource
- IBM Spectrum Computing
- DNAnexus
- CERN

Our Arvados Cluster on Microsoft Azure

- Stores and manages ~250 TiB of data
- Regularly run 100-200 simultaneous instances ranging from 1-20 cores, 3.50-140 GiB RAM (D1 v1 - D15 v2)
- Leverage 64 cores, 432.00 GiB (E64 v3) instances for larger scale debugging and prototyping
- “Cool” storage for costs savings
- Premier support was *very* responsive



Tiled Data (Numpy)

(No description provided)

Collection UUID:
su92l-4zz18-b8rs5x7t6gry16k

Content address:
984434bc0540105967920f6295423575+11783

Content size:
3 files totalling 17.3 GiB

No source information available.

Sharing and permissions

https://collections.su92l.arvadosapi.com/c=su92l-4zz18-b8rs5x7t6gry16k/t=f6x7c07cx5taweagvj2rhitlb68vpcgag96tu0pa4w63tlc1j/_/

all-info.npy	81.3 MiB
all.npy	17.2 GiB

Arvados cluster su92l running on Microsoft Azure

Agenda

Introduction to Genomics and Precision Medicine

Arvados Platform

Basics and Benefits of Tiling

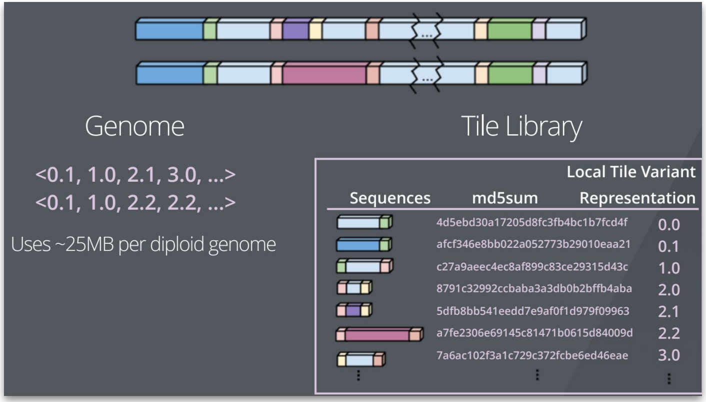
Machine Learning using Arvados and Tiled Data

Basics of Tiling

- Abstracts a genome by partitioning it into overlapping shorter sequences (tiles)
- Tiles
  - Braced on either side by “tags” (24-mers)
  - Can have multiple variants, one for each sequence observed at a position
- Set of all positions and all tile variants = tile library
- Individual genome is then represented as an array referencing the tile library

*CTTTTTGCCCGCTCAGGCTTTTGC*cccccgccgcggctttttgcccccc
gccgccgctttccccgccgtggctttttacaccctgcccccgcagctttt
tgcccccaccccgccttggctttttccccgccacggtttttggcccgcc
gccgccgccgccgccgccgcgactttttatccccagccgccgcggct
ttttgcccccaccccgccgcggctt*TCTGCCCAGCCCCCGTCGCCGCGG*

Example tile where: *CTTTTTGCCCGCTCAGGCTTTTGC* is the 'start' or 'left' tag and *TCTGCCCAGCCCCCGTCGCCGCGG* is the 'end' or 'right' tag.



<0.1, 1.0, 2.1, 3.0, ...>
<0.1, 1.0, 2.2, 2.2, ...>

Uses ~25MB per diploid genome

Sequences	md5sum	Local Tile Variant Representation
	4d5ebd30a17205d8fc3fb4bc1b7fcd4f	0.0
	afcf346e8bb022a052773b29010eaa21	0.1
	c27a9aeec4ec8af899c83ce29315d43c	1.0
	8791c32992ccbaba3a3db0b2bffb4aba	2.0
	5dfb8bb541eedd7e9af0f1d979f09963	2.1
	a7fe2306e69145c81471b0615d84009d	2.2
	7a6ac102f3a1c729c372fcbe6ed46eae	3.0
⋮	⋮	⋮

Tiling Benefits

- Set of genomes can be represented as a numerical matrix
  - Can use “out of the box” machine learning (ML) and large data methods
- Represents full genome
  - Includes homozygous reference calls and both phases
  - Known if regions are confidently called as reference or have variants
  - Reference and sequencing technology independent
- Makes it possible to harmonize different studies
  - Genome, exome, microarray data, different sequencing technologies
- Compact and scalable
  - Human reference genome becomes ~10M tiles vs 3B bases
  - Stored in compact genome formatted (CGF) files, 30-50 MB per genome

Lighting

- Combination of:
  - Conceptual way to concisely think about genomes (tiling)
  - Internal representation of tiled genomes for efficient access
  - Software that performs tiling, manages access, and analyzes tiled data

- Leverages CWL pipelines on Arvados running on Microsoft Azure

Agenda

Introduction to Genomics and Precision Medicine

Arvados Platform

Basics and Benefits of Tiling

Machine Learning using Arvados and Tiled Data

Big Questions for Machine Learning

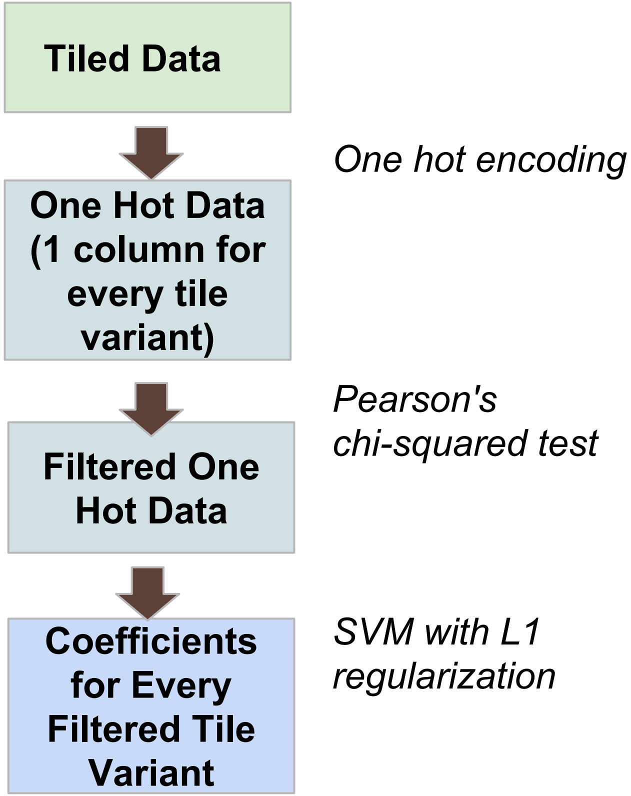
- Can we extract insights from WGS data using machine learning and large data techniques?
- How much more powerful are WGS than SNP arrays at detecting important variants?
- Can we create a “simple” model that be a base for comparison when testing more complex models or new algorithms?

Test Cases with PGP (Personal Genome Project)

- Started in 2005 at Harvard (now global)
- Provide freely available scientific resources that bring together genomic, environmental, and human trait data donated by volunteers
- Great source of consented, openly available genetic, and phenotype data
  - Tiled 200+ whole genomes
  - Focused machine learning on known Mendelian traits

The Harvard
Personal Genome Project
we are open for science

Machine Learning Model



Tiled Data

One hot encoding

One Hot Data (1 column for every tile variant)

Pearson's chi-squared test

Filtered One Hot Data

SVM with L1 regularization

Coefficients for Every Filtered Tile Variant

- Kept positions where at least 90% of tiles were “confidently called”
- For phenotypes studied, 1-5% of tile variants kept using Pearson's chi-squared test
- Linear SVM classifier with l1 penalty and class weights (scikit-learn)
- Optimum value of the penalty parameter found using 10 fold cross-validation

Results for Eye Color Classifier

- Initial work binned data into blue and not blue, ignoring hazel
- Yielded accuracy of **0.95 ± 0.08**
- Highest coefficient corresponded to tile located on Chromosome 15
  - Tile variant contains known SNP in OCA2/HERC2 region (**rs12913832**) linked to eye color

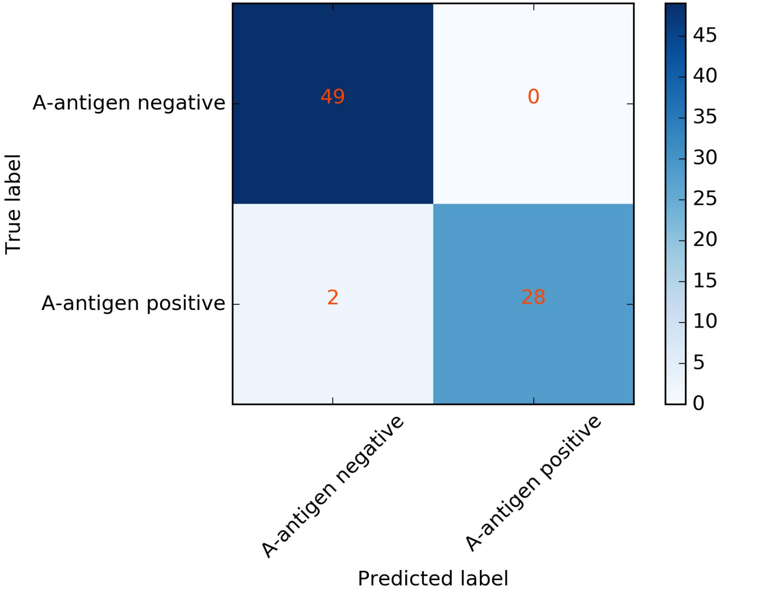
Image from PGP phenotype survey used to self-identify eye color

SNP: A single-nucleotide polymorphism, is a variation in a single nucleotide (e.g. A-> G) that occurs at a specific position in the genome

Blood Type Classification

- A antigen: accuracy **0.98 ± 0.05**
  - 8 non-zero coefficients
  - Top tile variants located in the ABO gene
  - Contains an indel, **rs782134971** (rs149092047) associated with blood type

- B antigen: accuracy **0.97 ± 0.05**
  - 5 non-zero coefficients
  - Top tile variants located in ABO gene
  - Contains a SNP, **rs505922**, associated with blood type



Indel: short polymorphism that corresponds to the addition or removal of a small number of bases in a DNA sequence

Alzheimer's Project

- Understand risk factors, discover protective variants, discover new possible drug targets using WGS of a large cohort

- With UPenn and IBM as part of an NIA project

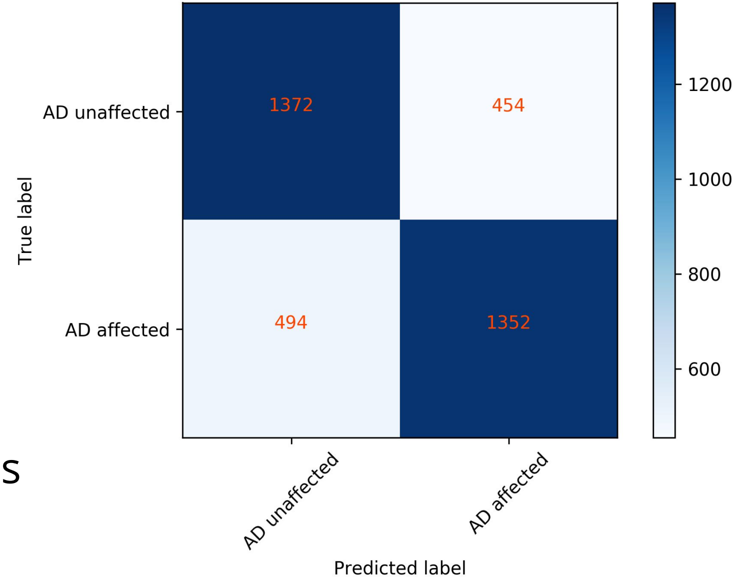
- 4000+ whole genomes and phenotypes available from ADNI and ADSP, **40+TB** *(310 GB in tiled arrays)*

- Approved to obtain an additional 4,000+ whole genomes from TOPMed / MESA

ADNI: Alzheimer's Disease Neuroimaging Initiative
ADSP: Alzheimer's Disease Sequencing Project
MESA: Multi-Ethnic Study of Atherosclerosis

Machine Learning Results

- Linear SVM: 74 (+/- 3)% Accuracy
  - Not determined entirely by genetics
  - Performs better than existing models [Escott-Price, et al., 2015]
- ~1500-2000 non-zero coefficients
  - From a possible ~200 million tile variants
  - Using p-values, reduced to ~900 tile variants
- Important tile variants mapped to genes and genetic variants
  - Gene list consistent with GWAS results
  - Novel genes



Accuracy from 10-fold cv 0.74 (+/- 0.03)

Future Work

- Test existing machine learning models on different cohorts

- Expand machine learning
  - Explore alternative filter (e.g very large scale ReliefF)
  - Include phenotype data (e.g. ethnicity) and variant interactions (non-linear models)

- Scale machine learning models to ~100,000 genomes

Summary

- Machine learning work made possible by tiling, Arvados and Microsoft Azure
- Extract insights from thousands of genomes (WGS) using scalable, reproducible machine learning techniques
- Gain better understanding of the power and possibility of WGS for detecting important variants
- Same techniques shown for AD work can be used for different data and phenotypes