TruCal Variant Recalibrator

SVAI Hackathon 2017
Team: Mutrakers
Problem Track: ranking dataset mutations

Team

Don Freed, Ph.D
Bioinformatics Scientist

Sentieon

Johns Hopkins School of Medicine

Annabelle Tang
Research Associate

GUARDANT HEALTH

University of California Santa Cruz

Alex Francis
Data Scientist

Quora

David Streid
Software Engineer

E*TRADE

Columbia University

Magdalena Matusiak, Ph.D
Postdoc Fellow Bioinformatics

Stanford

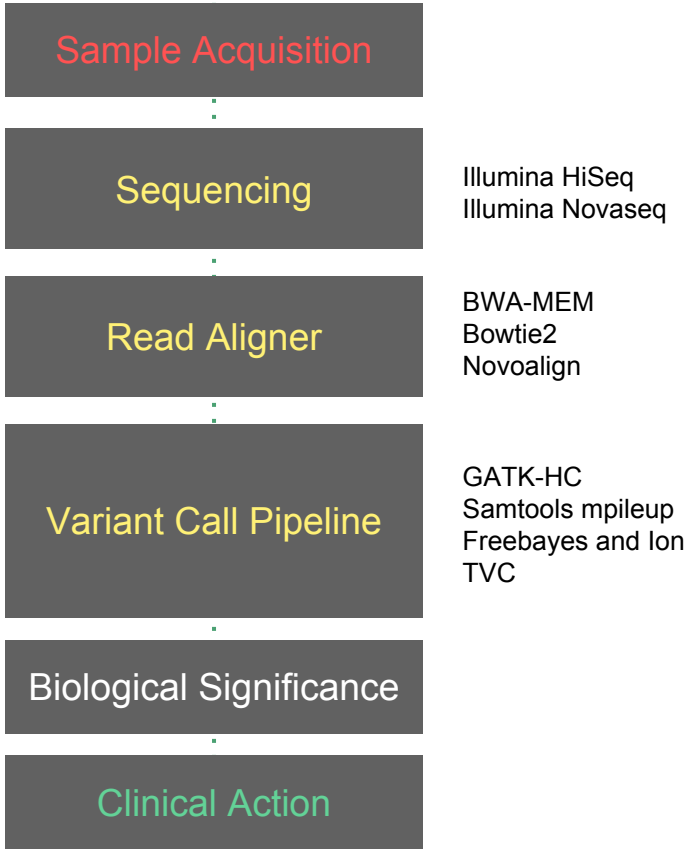
Chris Margono
Medical Student

UCSF

Introduction

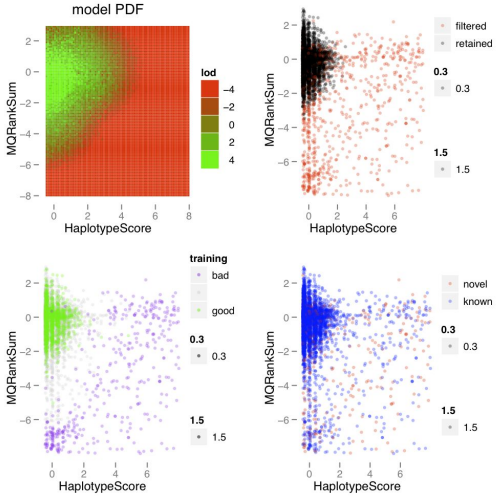
- Translating data to care requires:
  - proper processing
  - Proper interpretation
- **Errors** pose a risk to clinical application

- It has been demonstrated that different variant call methods can produce discordance outputs, indicating the need for careful interpretation of their results



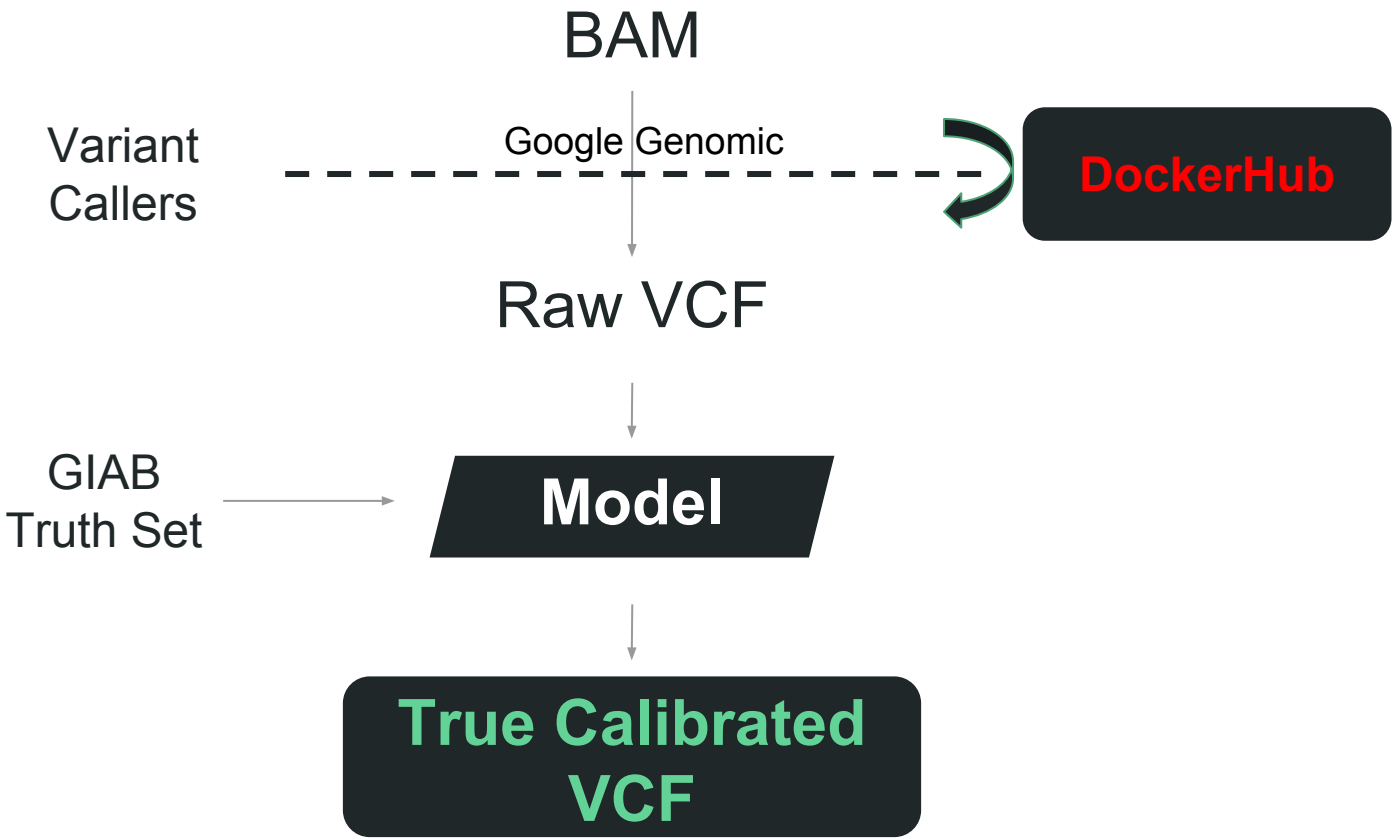
Purpose

- Previous work
  - used unsupervised learning to sort true mutations and errors from variant call outputs
  - no gold-standard truth set existed at that time

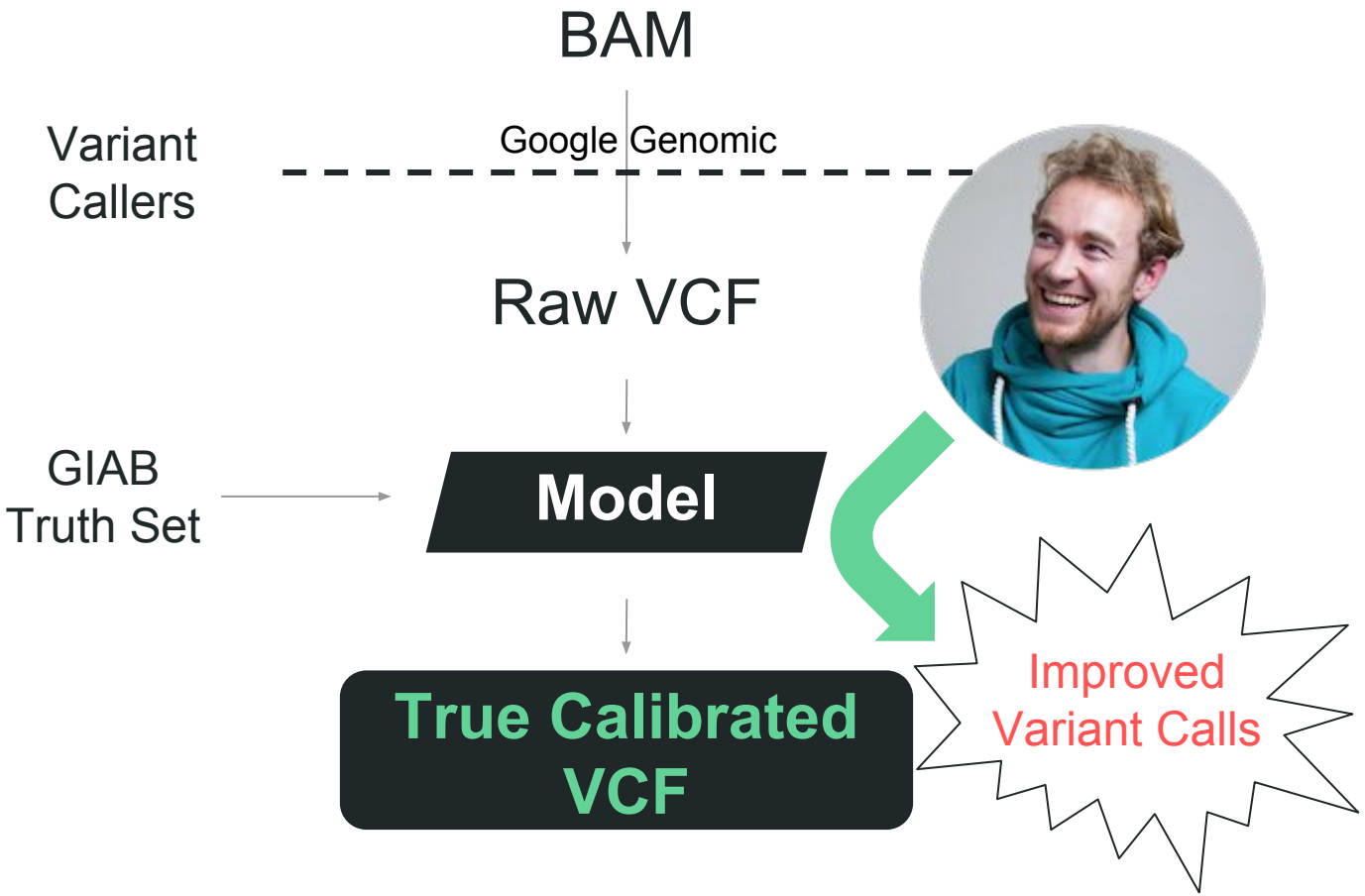


- Recent advancements (2015)
  - Genome in a Bottle Consortium (GIAB) Released high-confidence variant calls
- **Goal:**
  - **improve precision** (positive predictive value) of variant call outputs using **supervised learning** with **GIAB data**

Approach


BAM
Variant Callers
Google Genomic
DockerHub
Raw VCF
GIAB Truth Set
Model
True Calibrated VCF

Impact


BAM
Variant Callers
Google Genomic
Raw VCF
GIAB Truth Set
Model
True Calibrated VCF
Improved Variant Calls

Methods: Data Transformations

$$X = \begin{bmatrix} \overbrace{x_{11} \dots x_{1d}}^{\text{AD}} & \overbrace{}^{\text{QUAL}} \\ \vdots \ddots \vdots \\ x_{n1} \dots x_{1d} \end{bmatrix}$$

From the "True VCF," utilize a "featurizer" to obtain a data matrix for training

- *Impute missing features (messy data!) using the median of all features.*
- *Labels ("real variants" boolean) are compiled from Genome in a Bottle project*

$$n \approx 10^6 (\text{the number of reads/"mutations" in the VCF}), d = 21$$

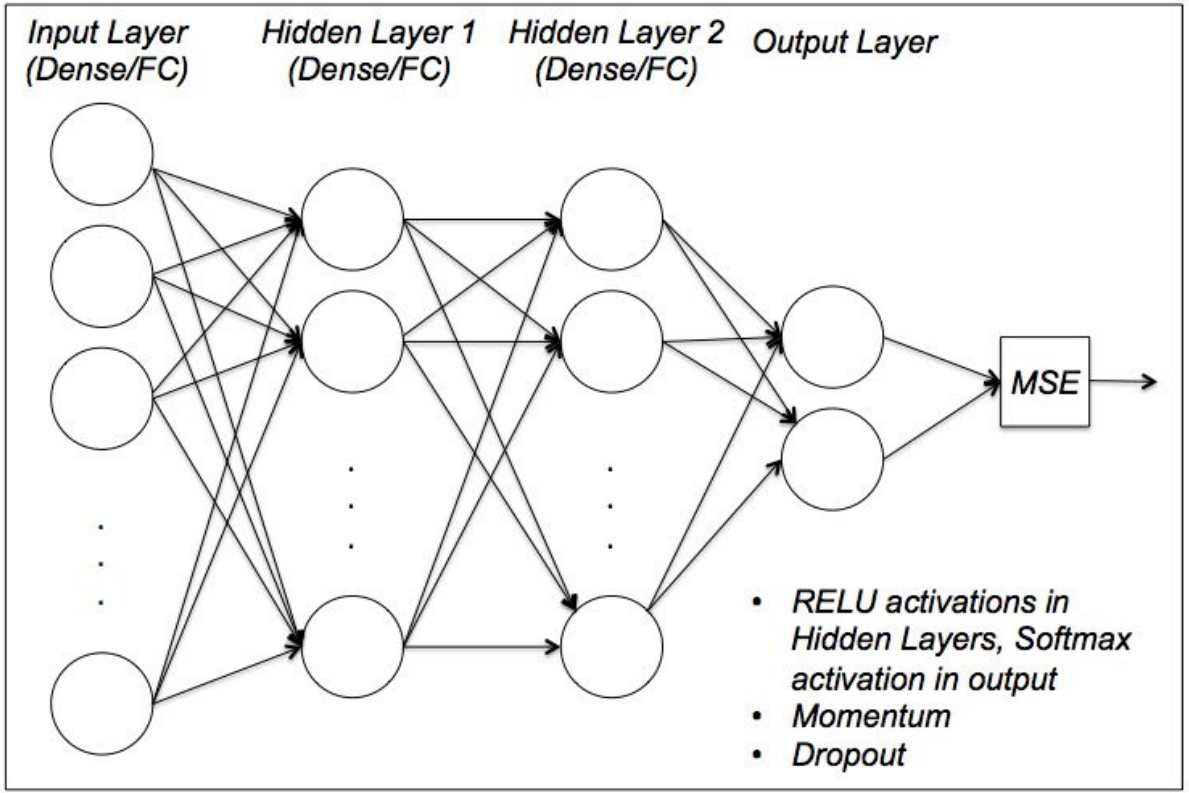
$$\text{features} = \{$$

AD, GQ, DP, GT, ExcessHet, AC, BaseQRankSum, FS, AF, MLEAC, AN, SOR, MQ, QD, DP, ClippingRankSum, MQRankSum, ReadPosRankSum, QUAL, is_snp

$$\}$$

Methods: Machine Learning

- Train three-layer neural network using Keras
- Produces a “probability calculator” from the output of the softmax activation in the final layer



Results

- Check out our model!

https://github.com/SVAI/MutRackers