Analyzing Large Scale Genomic Data Using the Google Cloud Platform

Cuiping Pan, Ph.D.

Palo Alto Veterans Institute for Research
VA Palo Alto; Cuiping@stanford.edu

460 Human Genomes from MVP Pilot

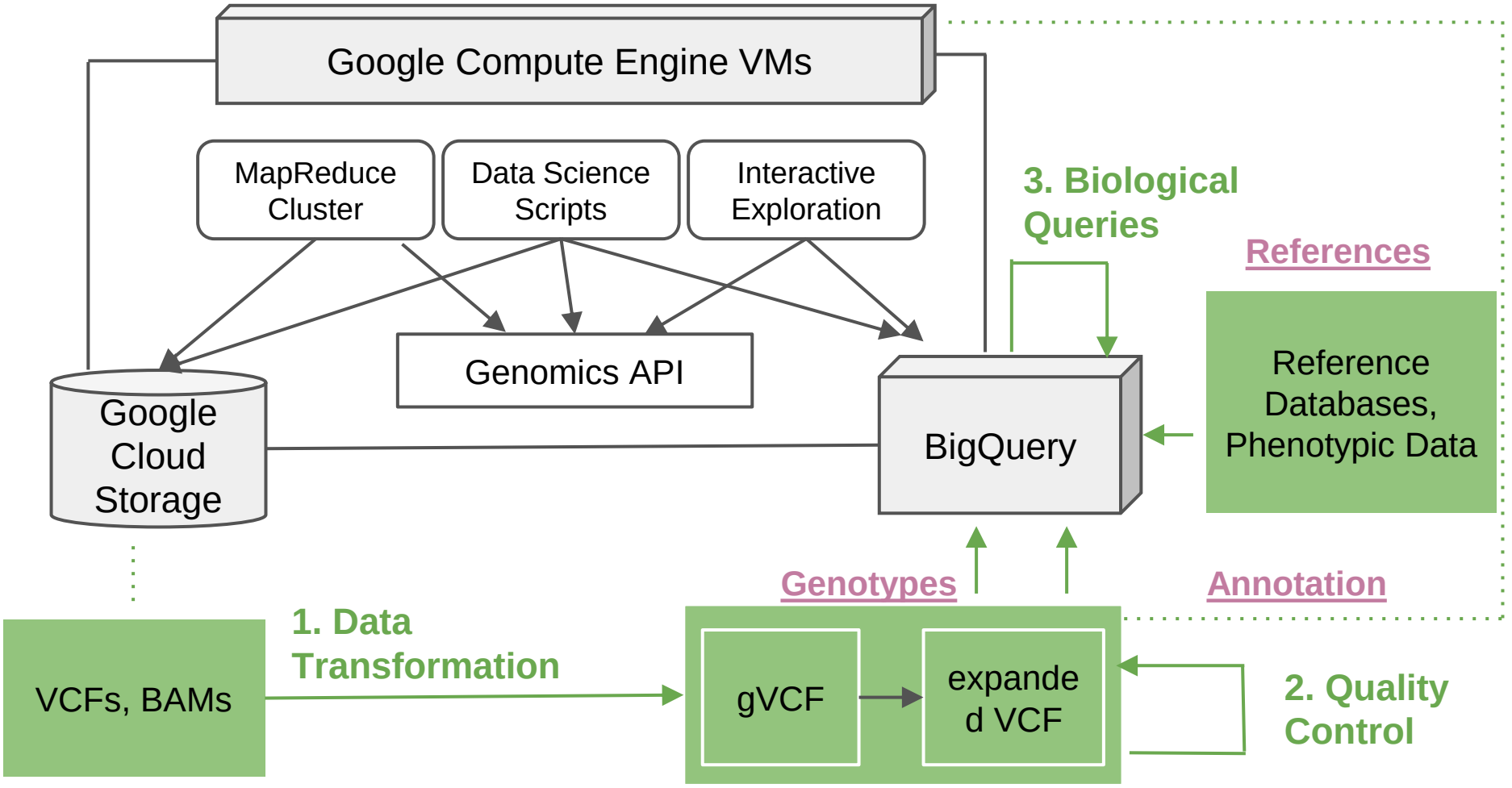
One Genome

- 3 billion DNA base pairs
- 3 million single nucleotide variation
- 50,000 private variants

460 Genomes

- 48T pre-processed data
- 26M single nucleotide variation
- 5M short insertion and deletion

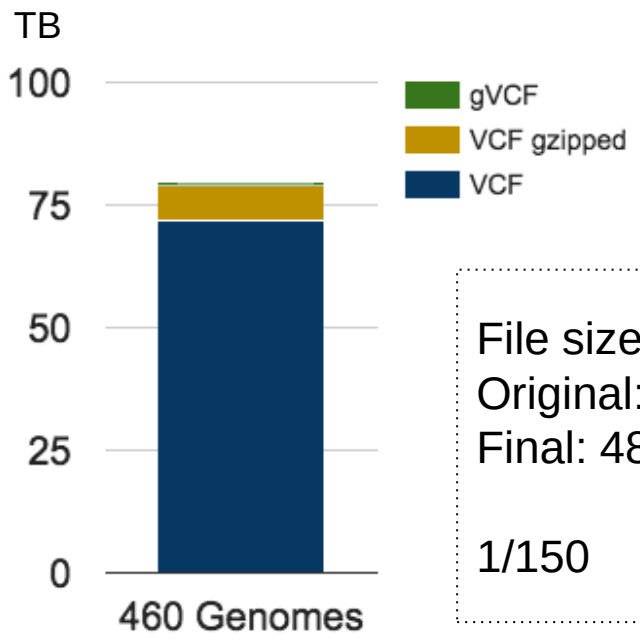
GCP and Our Workflow



Data Transformation and BigQuery Schema

gVCF:

- filter/tag variants by quality
- collapse consecutive REFs

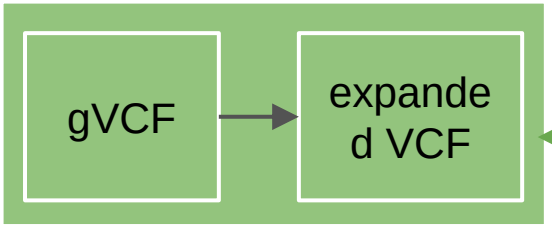


File size:
Original: 72T
Final: 481G

1/150

reference_name	STRING
start	INTEGER
end	INTEGER
reference_bases	STRING
alternate_bases	STRING
call	RECORD
-- call.call_set_id	STRING
-- call.call_set_name	STRING
-- call.genotype	INTEGER
-- call.FILTER	STRING
-- call.QUAL	FLOAT
...	

Multi-level Quality Controls



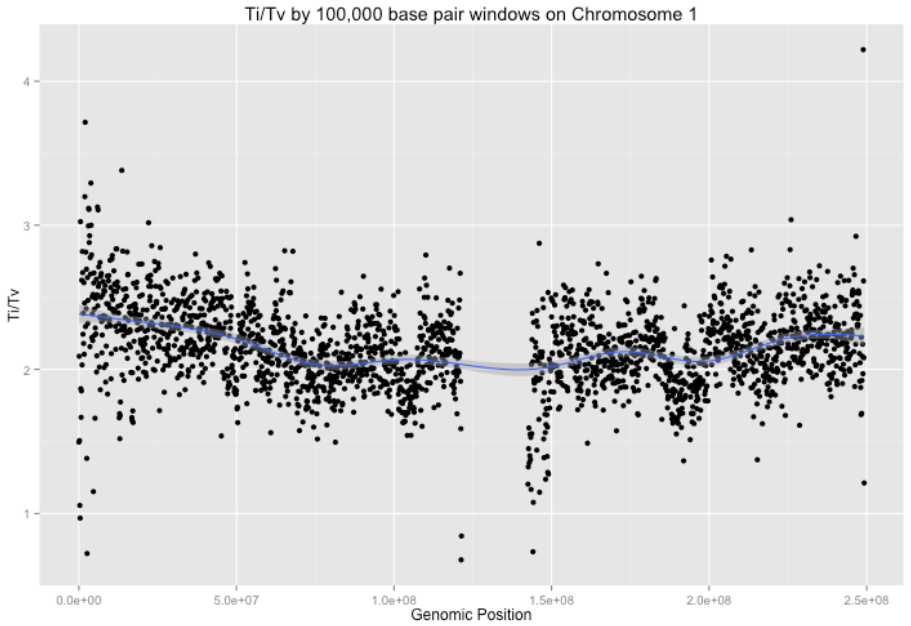
2. Quality Control

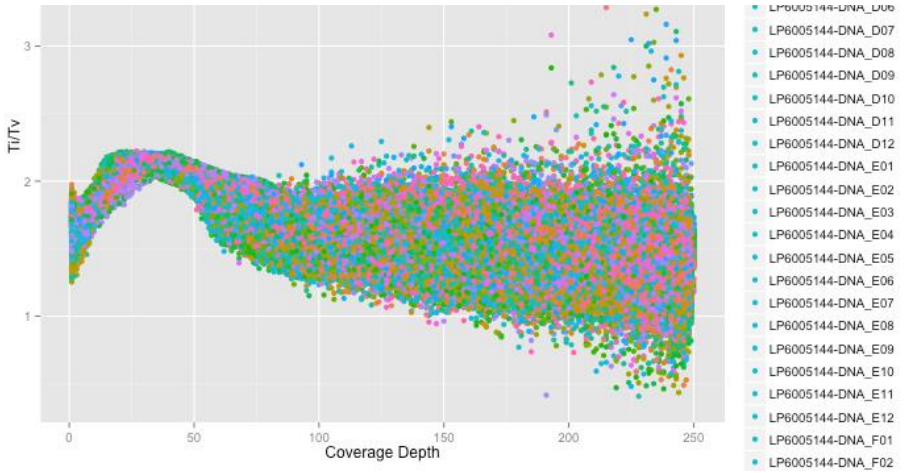
Sample Level QC

1. batch effect
2. concordance to SNP array data
3. sex inference
4. ethnicity inference
5. genome similarity: IBS
6. missingness rate
7. singleton rate
8. heterozygosity rate
9. inbreeding coefficient

Variant Level QC

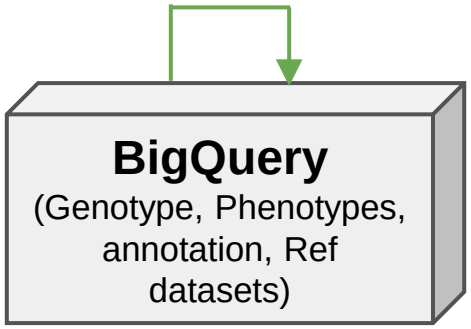
1. remove variants on blacklisted genes
2. heterozygous haplotype
3. missingness rate
4. Hardy-Weinberg Equilibrium
5. Ti/Tv by depth
6. Ti/Tv by alternate allele counts



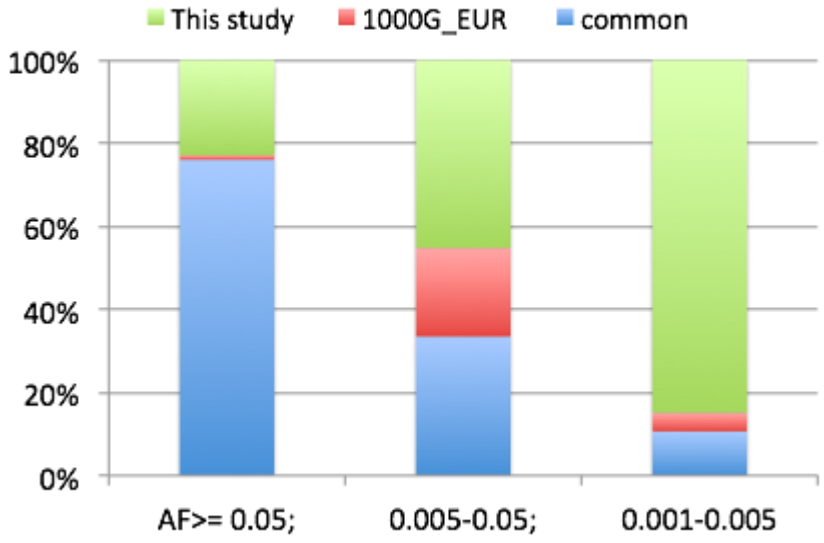


Biological Queries

3. Biological Queries



- Querying Genotypes
- Querying Biological Features
- Querying Medical Information

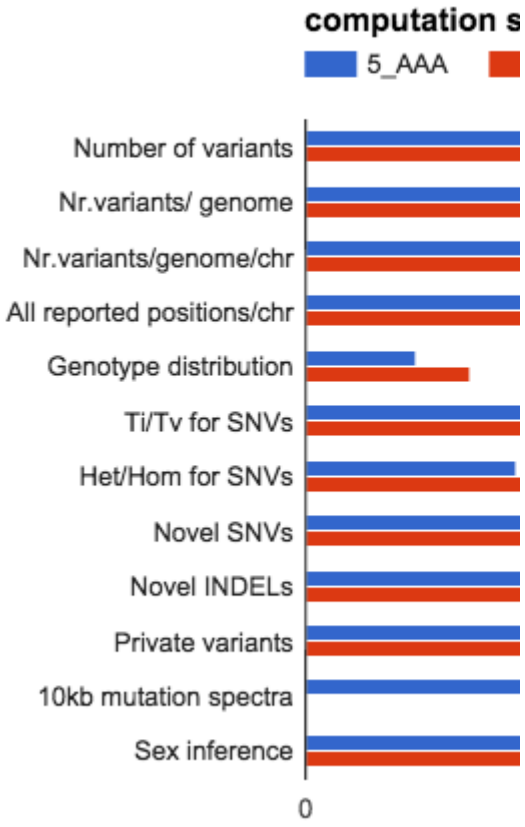


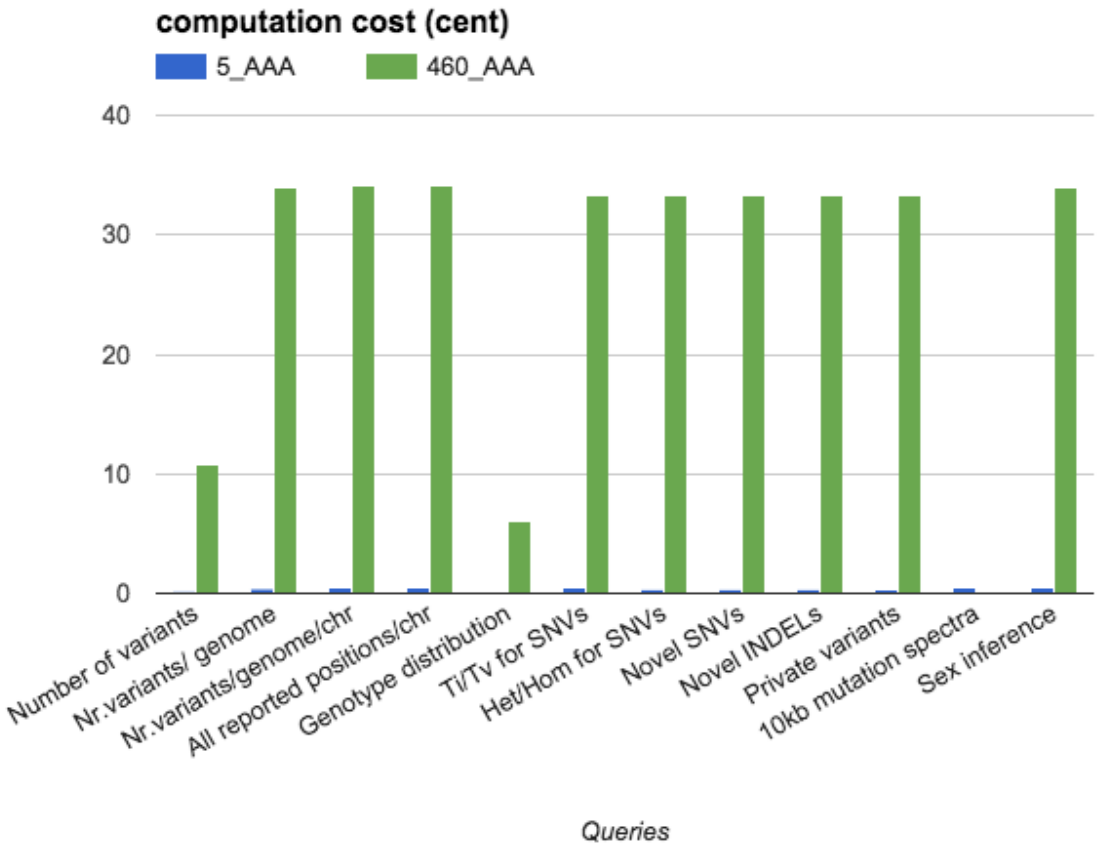
VKORC1 Genotype (-1639G>A, rs9923231)	*CYP2C9*1/*1*	*CYP2C9*1/*2*	*CYP2C9*1/*3*	*CYP2C9*2/*2*	*CYP2C9*2/*3*	*CYP2C9*3/*3*
GG	5-7	5-7	3-4	3-4	3-4	0.5-2
GA	5-7	3-4	3-4	3-4	0.5-2	0.5-2
AA	3-4	3-4	0.5-2	0.5-2	0.5-2	0.5-2

PharmGKB
Pharmacogenomics. Knowledge. Implementation.

5-7 mg/day: 288 subjects; 3-4 mg/day: 150 subject; 0.5-2 mg/day: 21 subjects

Scalability and Affordability of GCP





Acknowledgements

VA Palo Alto
Philip Tsao
Alicia Deng

Stanford University
Gregory McInnes
Somalee Datta
Isaac Liao
Denis Salins
Michael Snyder

Google Genomics
Nicole Deflaux
Jonathan Bingham
Elmer Garduno
David Glazer
Samuel Gross

Poster #1w

Email: cuiping@stanford.edu