

CAGI slide pack guidelines

I agree to my slides being used according to the following terms: _____ *Yanran Wang* _____

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Predicting the risk of Venous Thromboembolism with exome data

Yanran Wang

Bromberg Lab at Rutgers University



decoding the genomic blueprint of life
bromberg lab
bioinformatics of molecular function
genomic variation | microbes and -biomes



RUTGERS

What is VTE and AF?

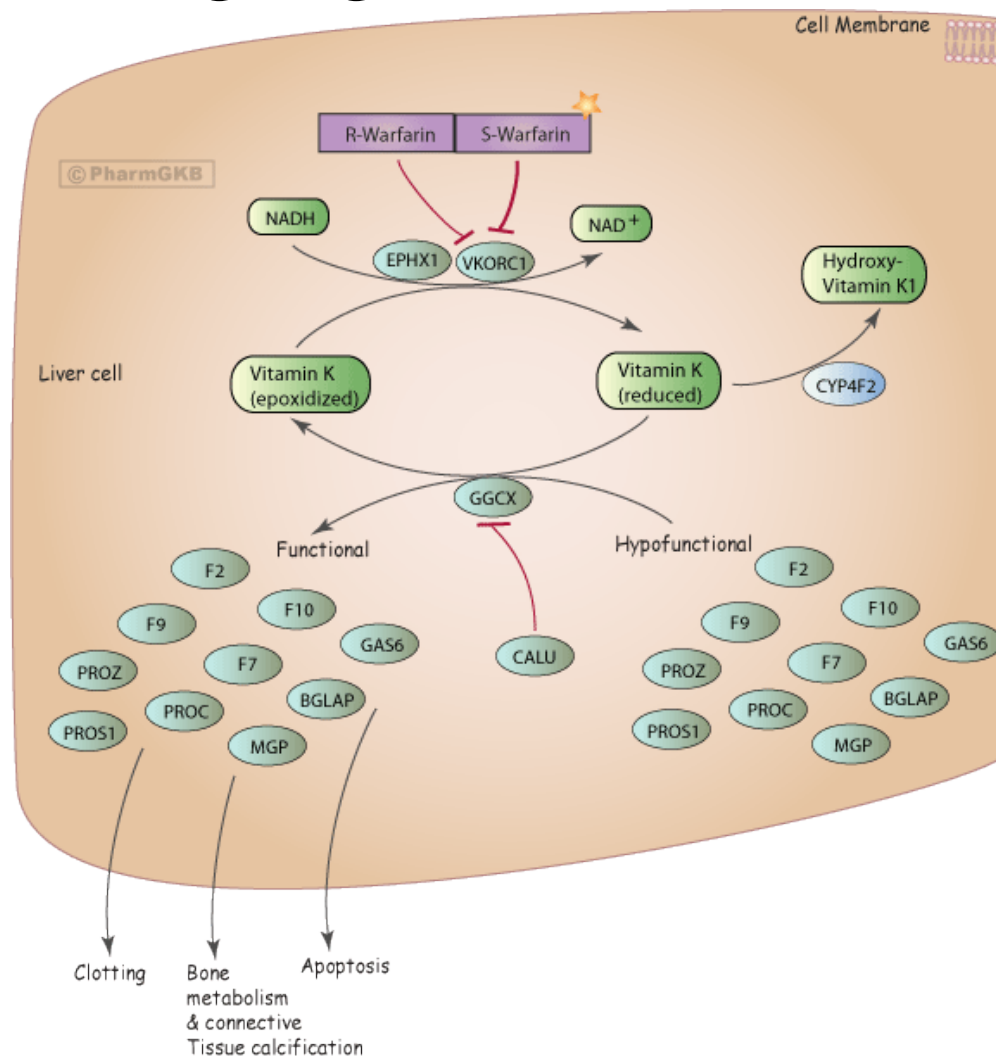
- Venous thromboembolisms (VTE)
 - Deep vein thrombosis (DVT)
 - Blood clot in the large deep veins
 - It usually occurs in the deep veins of the legs or hips
 - Pulmonary embolism (PE)
 - PE occurs when a deep vein clot breaks off and flows to the lungs
 - It is the most serious complication of DVT
- Atrial fibrillation (AF)
 - AF is an irregular heart rate
 - It can increase risk of stroke, heart failure and other complications
 - It can lead to blood clots forming in the heart that may circulate to other organs and lead to blocked blood flow (ischemia)

What is VTE and AF?

- Venous thromboembolisms (VTE)
 - Deep vein thrombosis (DVT)
 - **Blood clot** in the large deep veins
 - It usually occurs in the deep veins of the legs or hips
 - Pulmonary embolism (PE)
 - PE occurs when **a deep vein clot** breaks off and flows to the lungs
 - It is the most serious complication of DVT
- Atrial fibrillation (AF)
 - AF is an irregular heart rate
 - It can increase risk of stroke, heart failure and other complications
 - It can **lead to blood clots** forming in the heart that may circulate to other organs and lead to blocked blood flow (ischemia)

What is it with Warfarin?

- Warfarin is an anticoagulant drug with very narrow therapeutic index
- Warfarin acts as a VKORC1 (Vitamin K epoxide Reductase Complex subunit1) inhibitor
- Some coagulation factors are involved in the downstream



What affects Warfarin dosage?

- Indication for Warfarin therapy
- VKORC1 and CYP2C9 polymorphisms
- Age, height, weight, interacting drugs



What is the challenge?

Data: 103 African American individuals



- A jointly called VCF file of exomes is provided
- All individuals are prescribed Warfarin for anticoagulation
- Most of the VTE individuals in the dataset are “unprovoked” VTE
- A clinical co-variate file is also provided (including gender, age, height, weight, whether on aspirin, whether on amiodarone, warfarin dose)

What is the challenge?

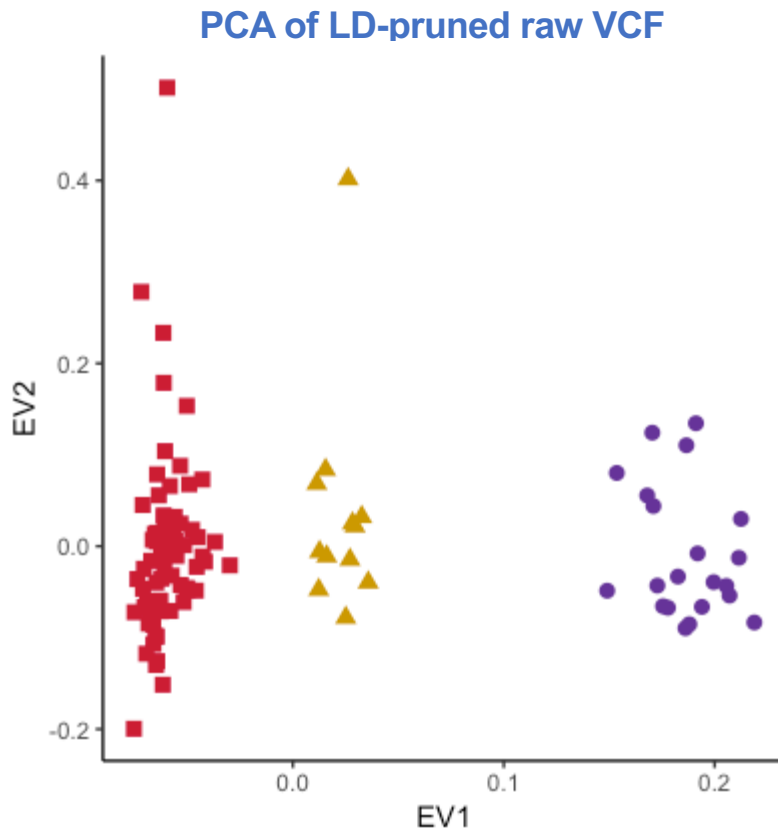
Data: 103 African American individuals

- African American population
 - higher incidence of VTE than European population
- Known genetic risk alleles in European descent individuals
 - F5 p.R506Q
 - F2 p.G20210A
- Known European risk alleles are barely found in African American population

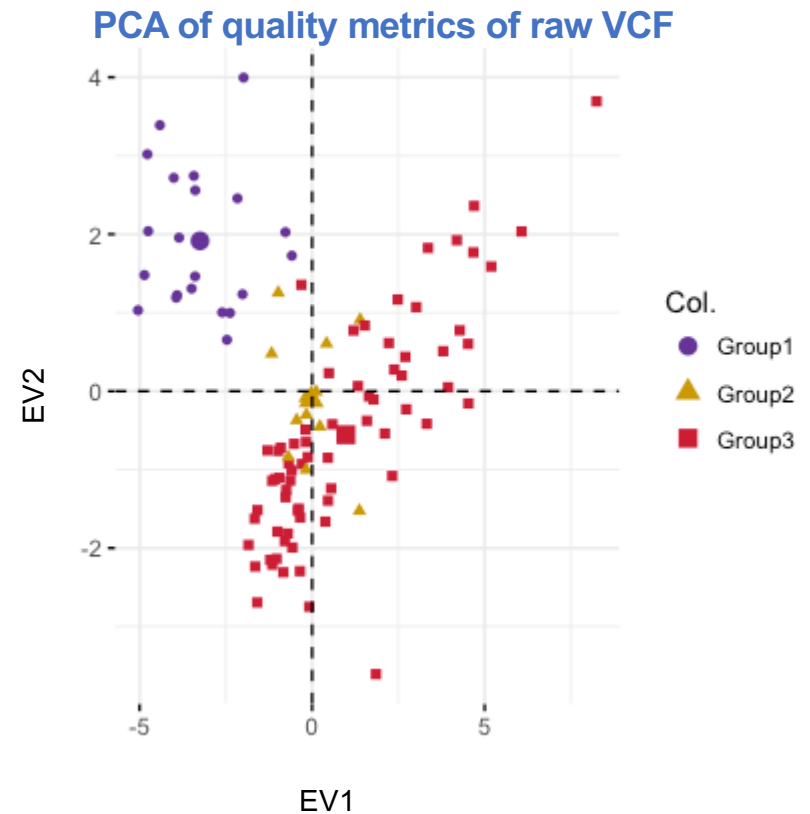
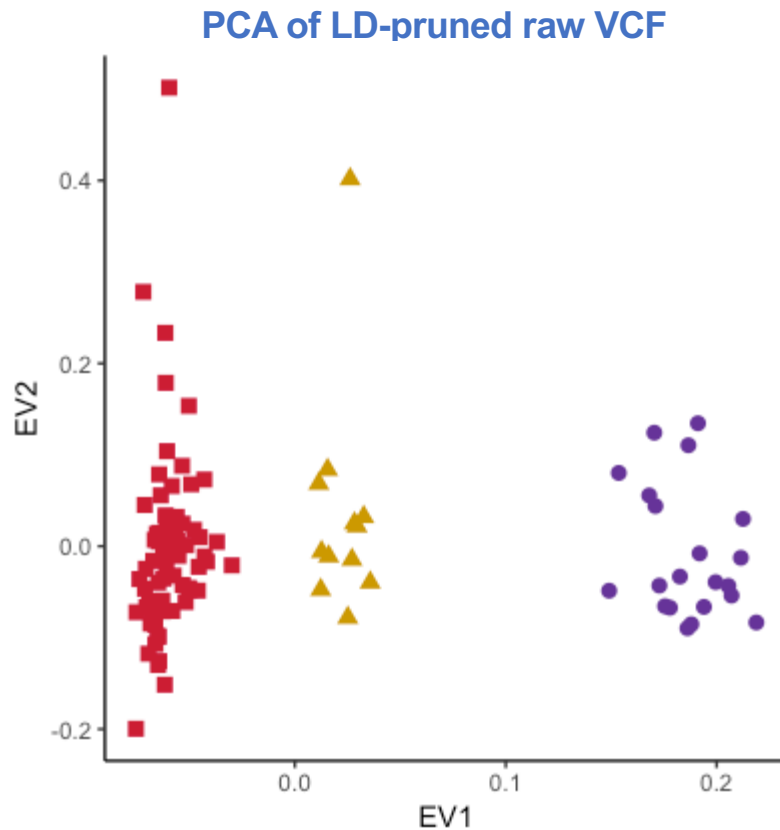
A first look of the data

Variant number	602,598
Call rate	98.68%
Ti/Tv ratio	2.19757
PASS ratio	84.27%

A first look of the data



A first look of the data

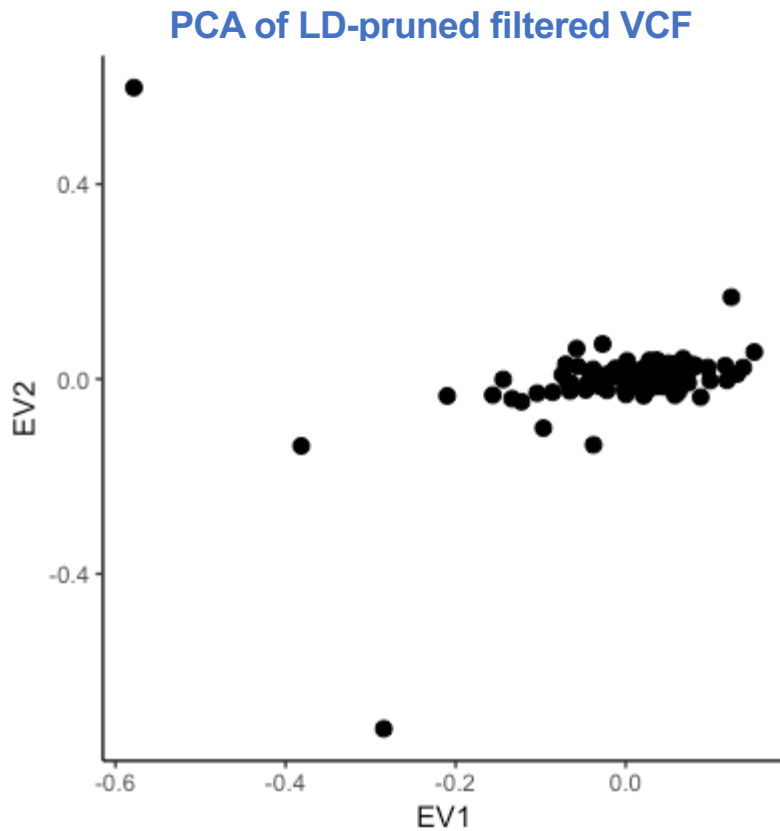


Filter the VCF

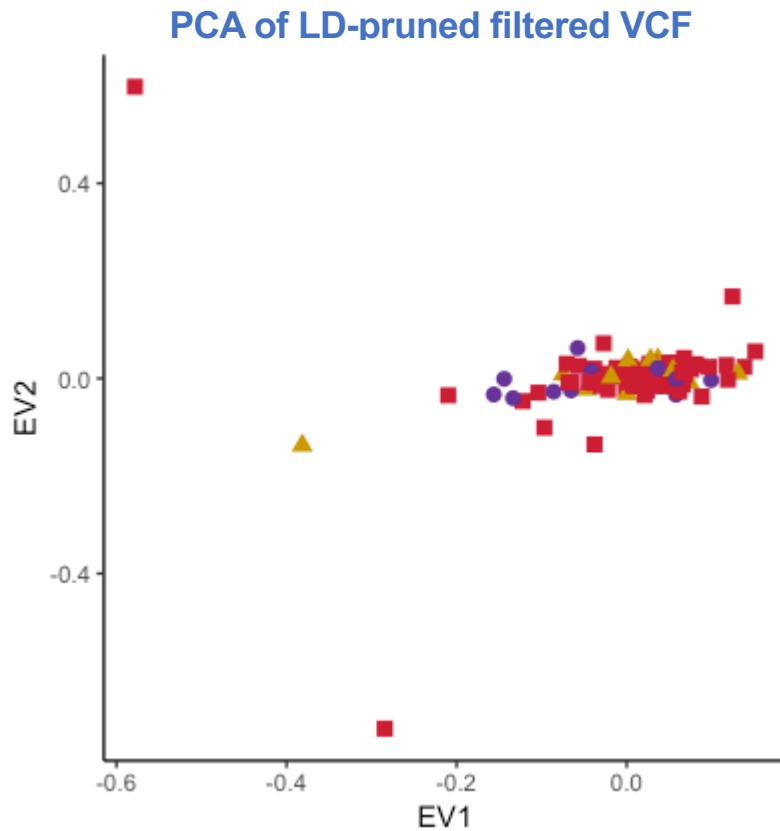
Retaining only PASS loci by VQSR (Variant Quality Score Recalibration)

	Before filtering	After filtering	Further filtering
Variant number	602,598	507,813	...
Call rate	98.68%	98.94%	...
Ti/Tv ratio	2.20	2.48	...
PASS ratio	84.27%	100%	...

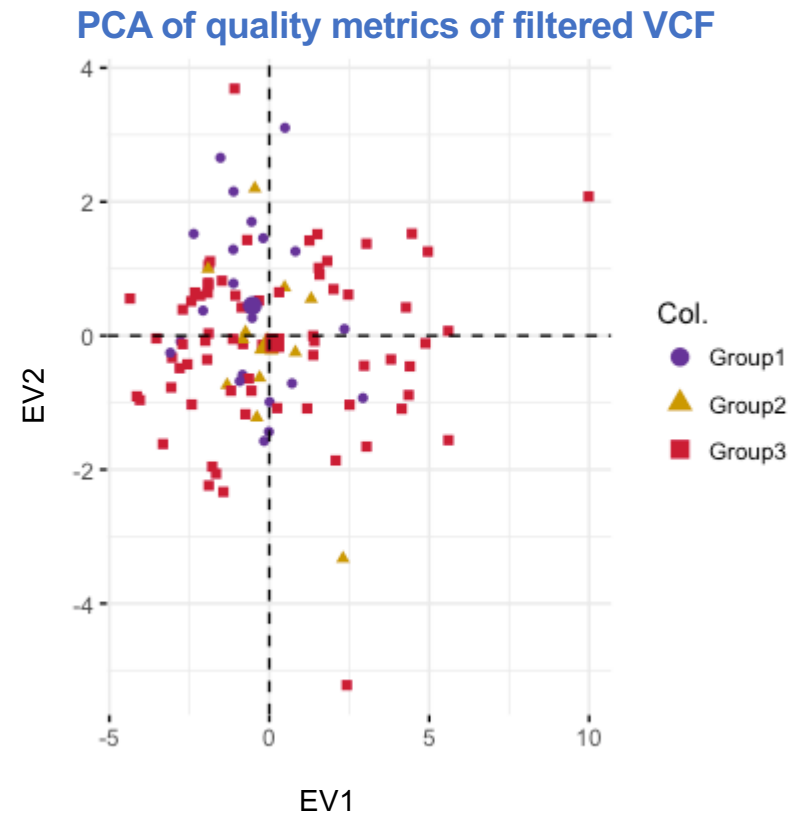
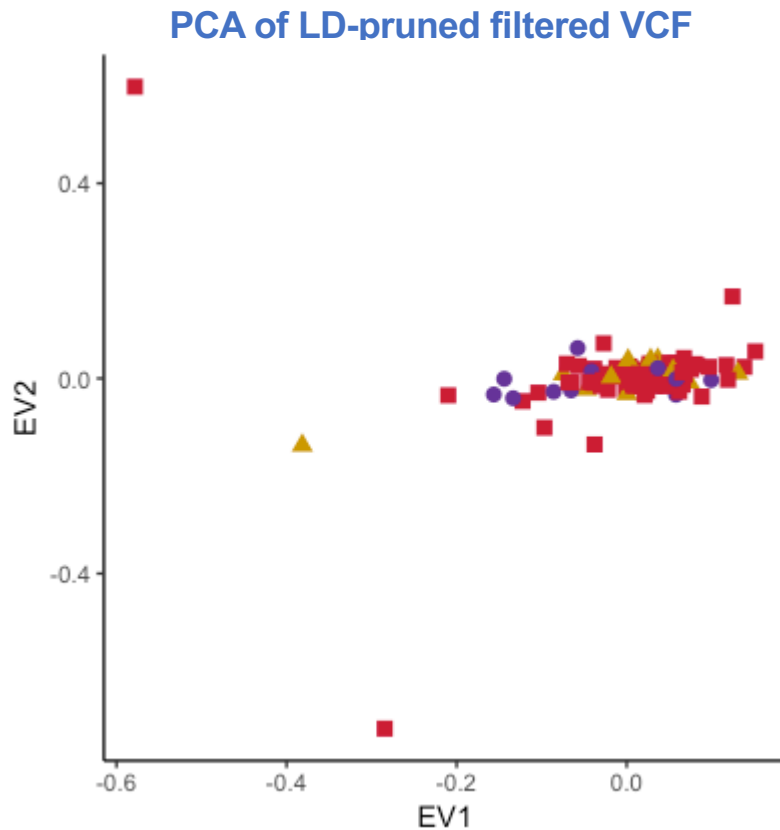
Filtered VCF



Filtered VCF



Filtered VCF



What can I find about VTE?

1. Papers from data provider
 - 1a. Data description paper
 - 1b. PROS1 V510M mutation paper
2. Genome-wide association studies about VTE
3. Known gene-disease relationships

(1a) Data description paper

Table 1

Clinical demographics of the discovery cohort and replication cohort

Covariate	Discovery mean/% (all) (n = 103)	Discovery mean/% (>49 mg/week) (n = 58)	Discovery mean (<35 mg/week) (n = 45)	Replication cohort mean/%
Dose	49.16 mg/wk	66.38 mg/wk	26.97 mg/wk	43.80 mg/wk
Gender	68.00% female	67.2% female	68.89% female	61.39% female
Age	56.75 y	50.38 y	64.96 y	57.58 y
Height	168.20 cm	168.83 cm	167.40 cm	169.60 cm
Weight	92.37 kg	96.43 kg	87.13 kg	92.05 kg
Aspirin	26.21%	12.07%	44.44%	—
Amiodarone	0.97%	0%	2.22%	4.30%
Phenytoin	0.97%	0%	2.22%	—
Carbamazepine	1.94%	3.45%	0%	—
Rifampin	2.91%	3.45%	2.22%	—
Indication (VTE)	64.71%	82.75%	40.91%	47.04%

Daneshjou, Roxana, et al. "Genetic variant in folate homeostasis associated with lower warfarin dose in African Americans." *Blood* (2014): blood-2014.

(1b) One SNP with increased risk of VTE in AA

- PROS1 V510M as the causative mutation (OR = 4.62, p = 0.0041 in one study)

Daneshjou, Roxana, et al. "Population-specific single-nucleotide polymorphism confers increased risk of venous thromboembolism in African Americans." *Molecular genetics & genomic medicine* 4.5 (2016): 513-520.

(2) GWAS reports about VTE

- Three intragenic SNPs:
 - LEMD3 rs138916004 (OR = 3.2, $p = 1.3e-8$)
 - LY86 rs3804476 (OR = 1.8, $p = 2e-8$)
 - LOC100130298 rs142143628 (OR = 4.5, $p = 4.4e-8$)

Heit, John A., et al. "Identification of unique venous thromboembolism-susceptibility variants in African-Americans." *Thrombosis and haemostasis* 117.4 (2017): 758.

(3) Known gene-disease relationships

DisGeNET

“It is a discovery platform containing collections of genes and variants associated to human diseases.”

(3) Known gene-disease relationships

Keyword - “**VTE**”: *F5, F2, FGA, PROC, PLAT, SERPINC1, TNF, SERPIND1*

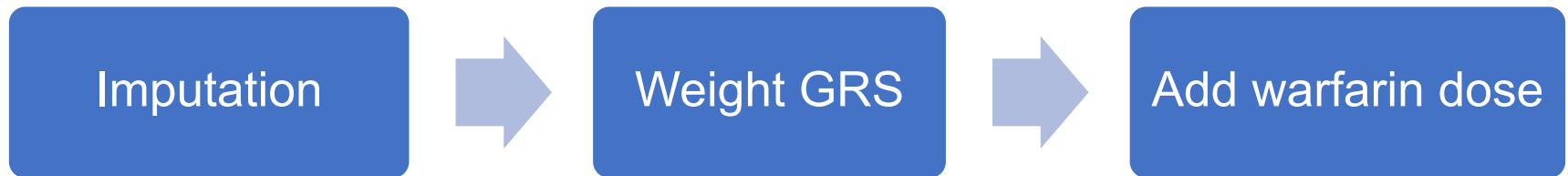
Keyword - “**DVT**”: *F5, F2, PROS1, PROC, SERPINC1, THBD, F9, AKT1*

Keyword - “**PE**”: *FGA, F2, MTHFR, SERPINC1, THBD, PROS1, PROC, PLAT, CBS, AGGF1, AKT1, PTEN, TBXA2R, DAB2IP, KLKB1, KCNJ5, PLAUI, XDH, CAT, MERTK, EPO*

My submission plans

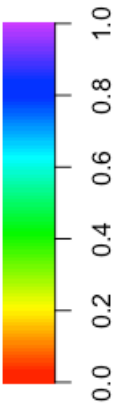
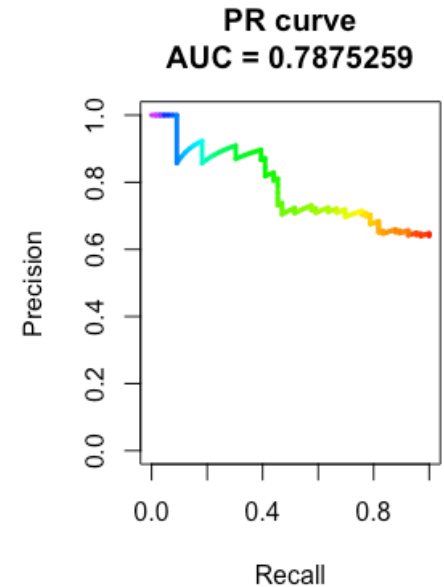
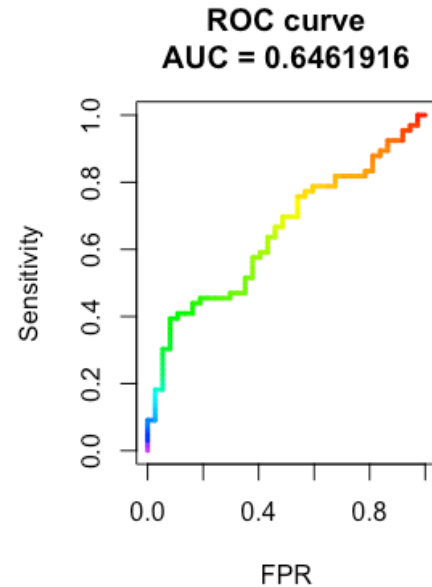
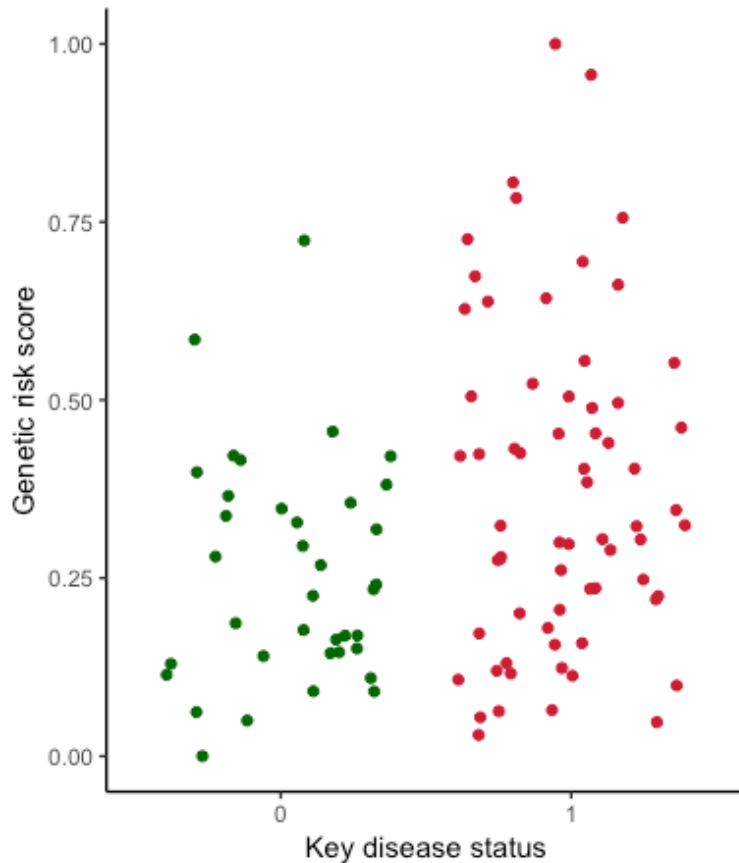
- **Numerical predictions**
 - Genetic risk score (GRS) approach (Submission 1)
- **Categorical predictions (clustering)**
 - All VTE-related variants (Submission 2)
 - Selected VTE-related variants (Submission 3)
 - Selected VTE-related genes (Submission 4)

Submission 1 – GRS



$$r_i = \sum_{j=1}^m w_j x_{ij} \quad .$$

Submission 1 – GRS



With a cutoff of 0.5
Accuracy: 50.49%
Precision: 89.47%
Recall: 25.76%

Submission 1 – GRS

Possible reasons for this poor performance:

- Known SNPs are not covered by exome sequencing
- Imputation in AA population is different from imputation in European population
 - imputation accuracies are very low
- Not enough GWAS results
- Did not use nongenetic factors
- etc.

Submission 2 – All VTE-related variants

Keyword - “**VTE**”: *F5, F2, FGA, PROC, PLAT, SERPINC1, TNF, SERPIND1*

Keyword - “**DVT**”: *F5, F2, PROS1, PROC, SERPINC1, THBD, F9, AKT1*

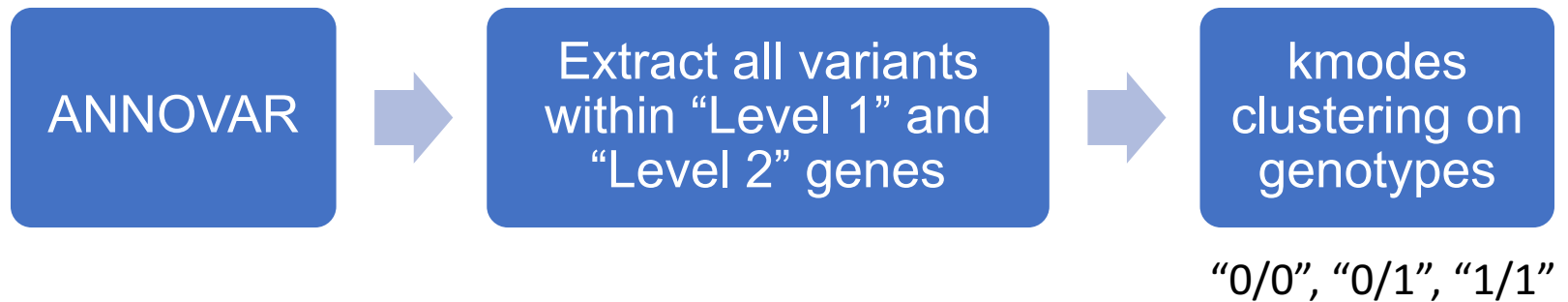
Keyword - “**PE**”: *FGA, F2, MTHFR, SERPINC1, THBD, PROS1, PROC, PLAT, CBS, AGGF1, AKT1, PTEN, TBXA2R, DAB2IP, KLKB1, KCNJ5, PLAUI, XDH, CAT, MERTK, EPO*

“Level 1” – *F5, F2, FGA, PROC, PLAT, SERPINC1, TNF, SERPIND1*

“Level 2” – *PROS1, THBD, AKT1*

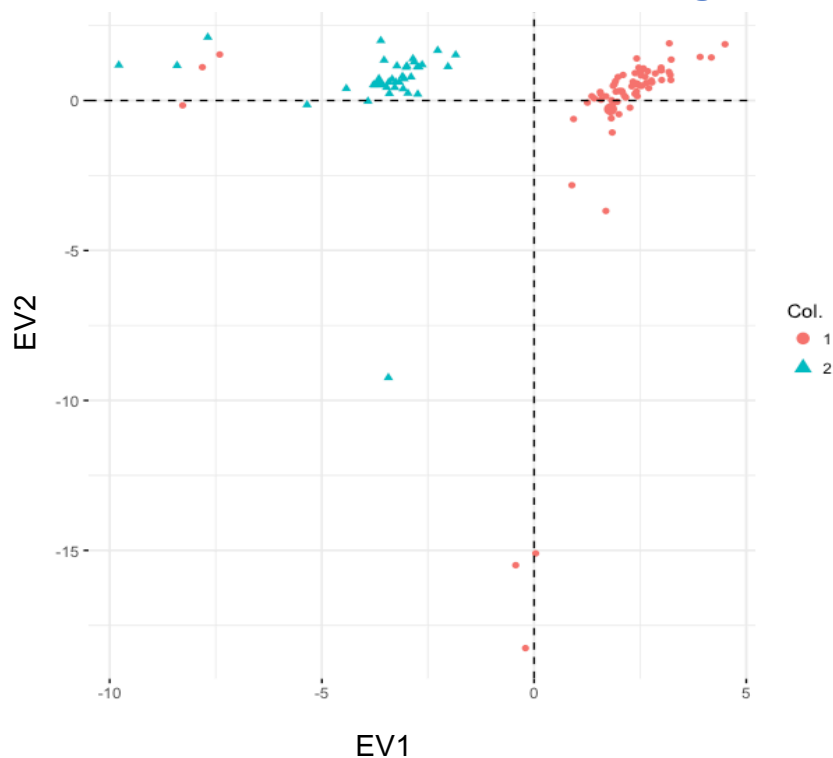
“Level 3” – *F9, MTHFR, CBS, AGGF1, PTEN, TBXA2R, DAB2IP, KLKB1, KCNJ5, PLAUI, XDH, CAT, MERTK, EPO*

Submission 2 – All VTE-related variants



Submission 2 – All VTE-related variants

PCA on variants from 11 VTE-related genes



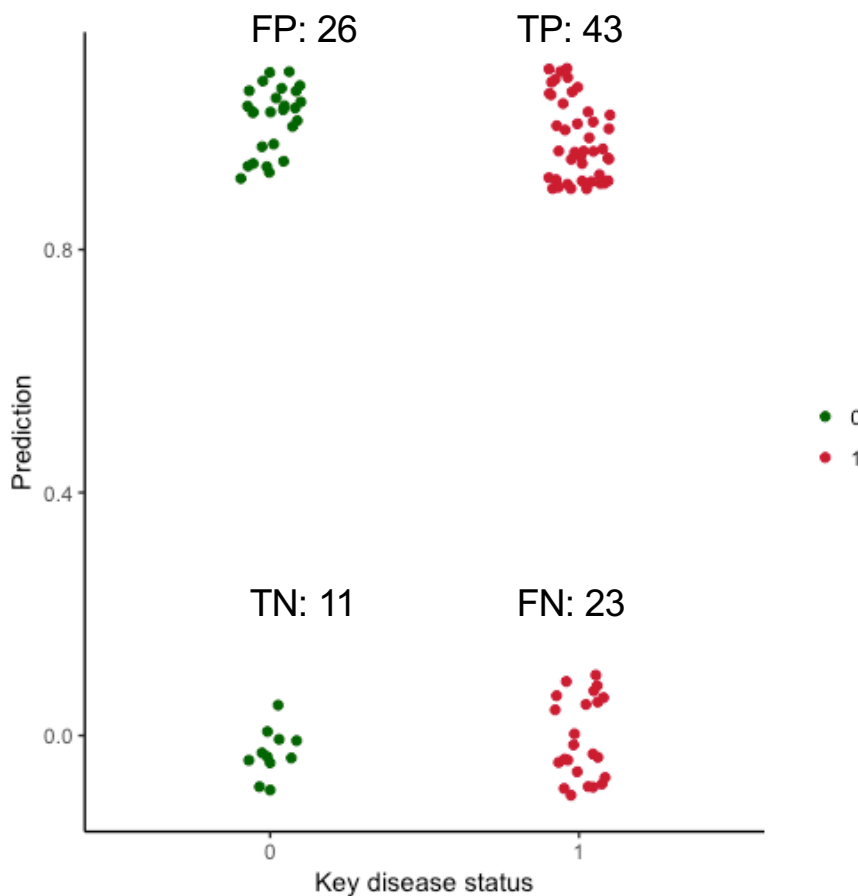
- 69 individuals (group1, red) and 34 (group2, blue) individuals
- Variance explained by EV1 and EV2: 16.7%

Submission 2 – All VTE-related variants

Accuracy: $(43+11)/103 = 52.4\%$

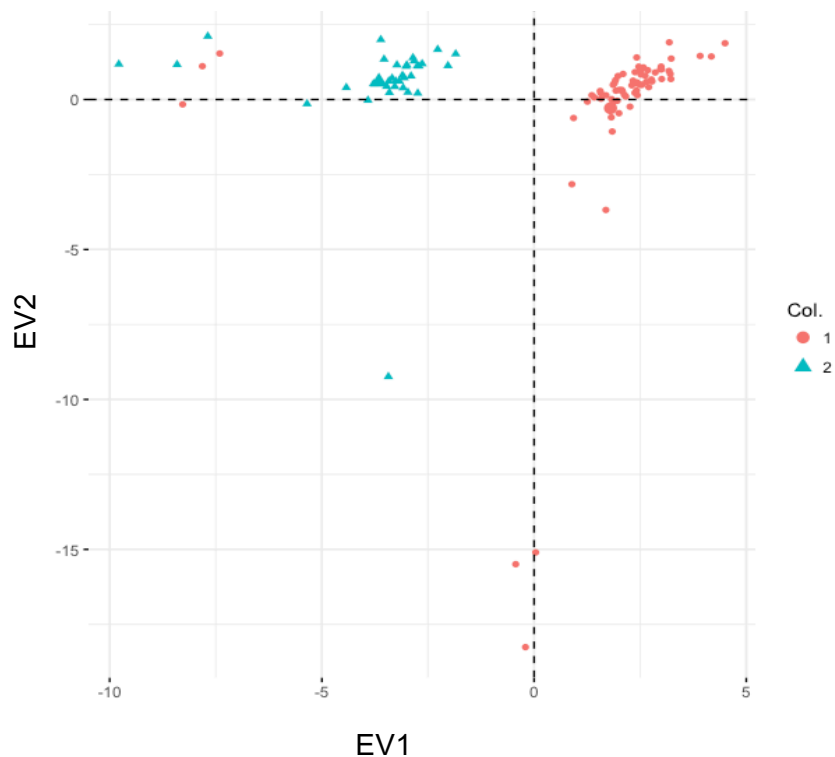
Precision: $43/(43+26) = 62.3\%$

Recall: $43/(43+23) = 65.2\%$

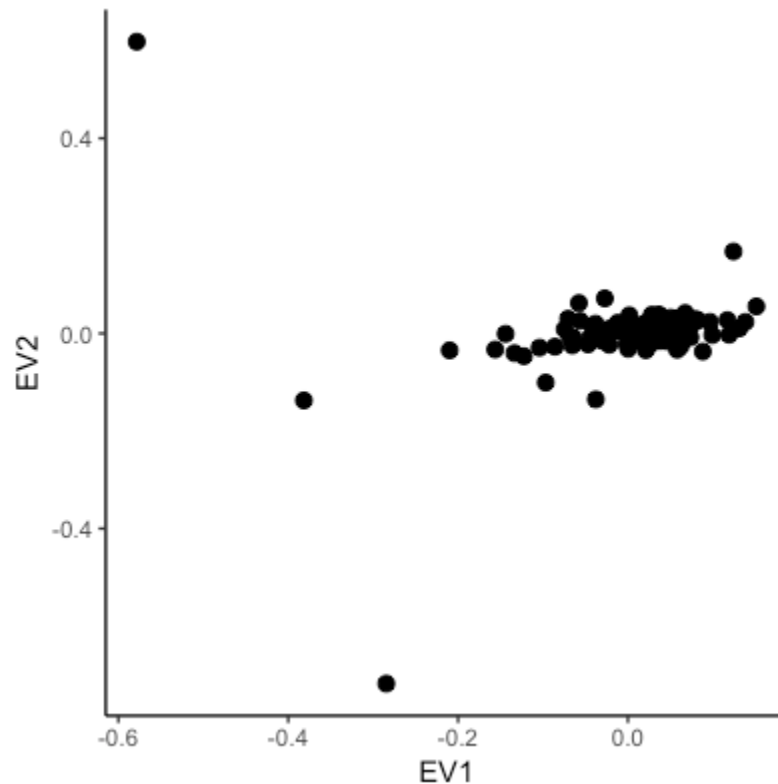


Submission 2 – All VTE-related variants

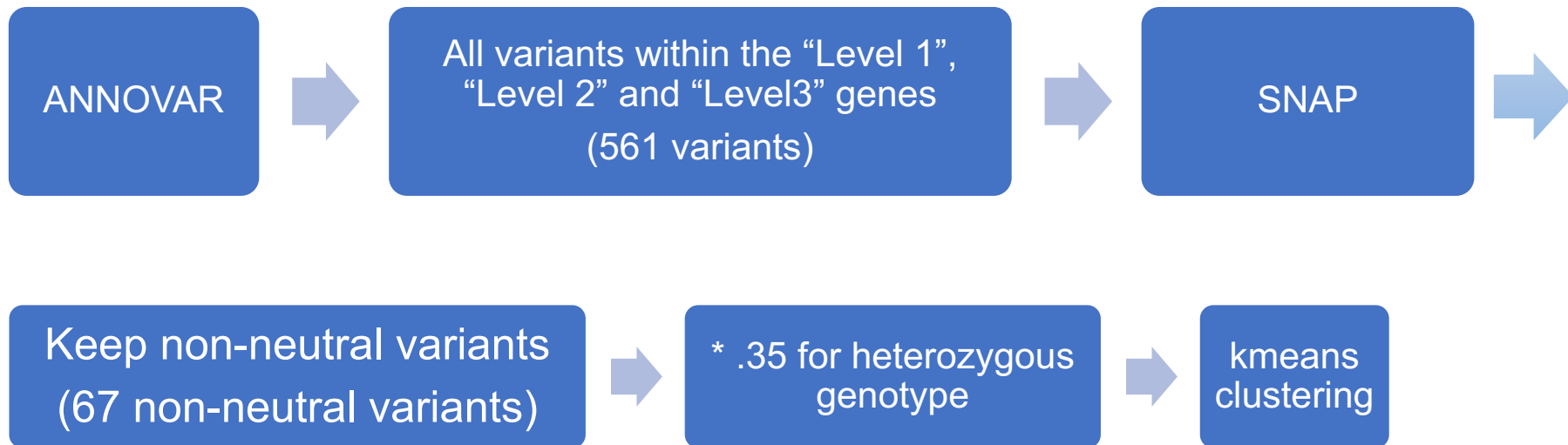
PCA on variants from 11 VTE-related genes



PCA of LD-pruned filtered VCF



Submission 3 – Selected variants

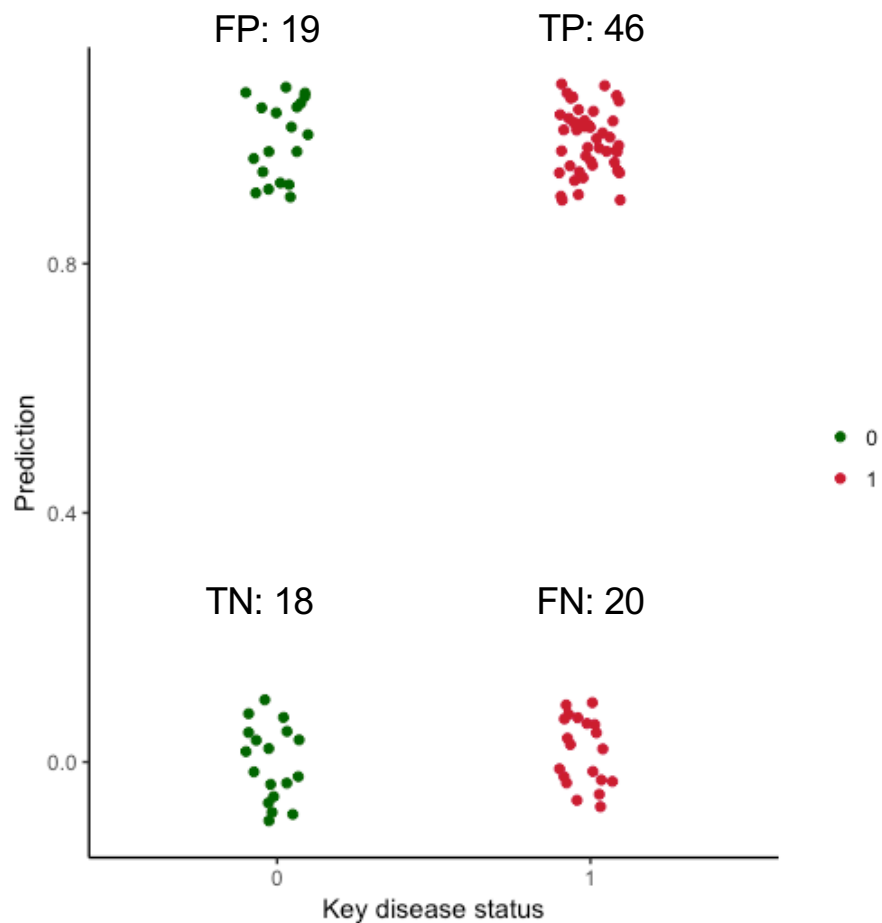


Submission 3 – Selected variants

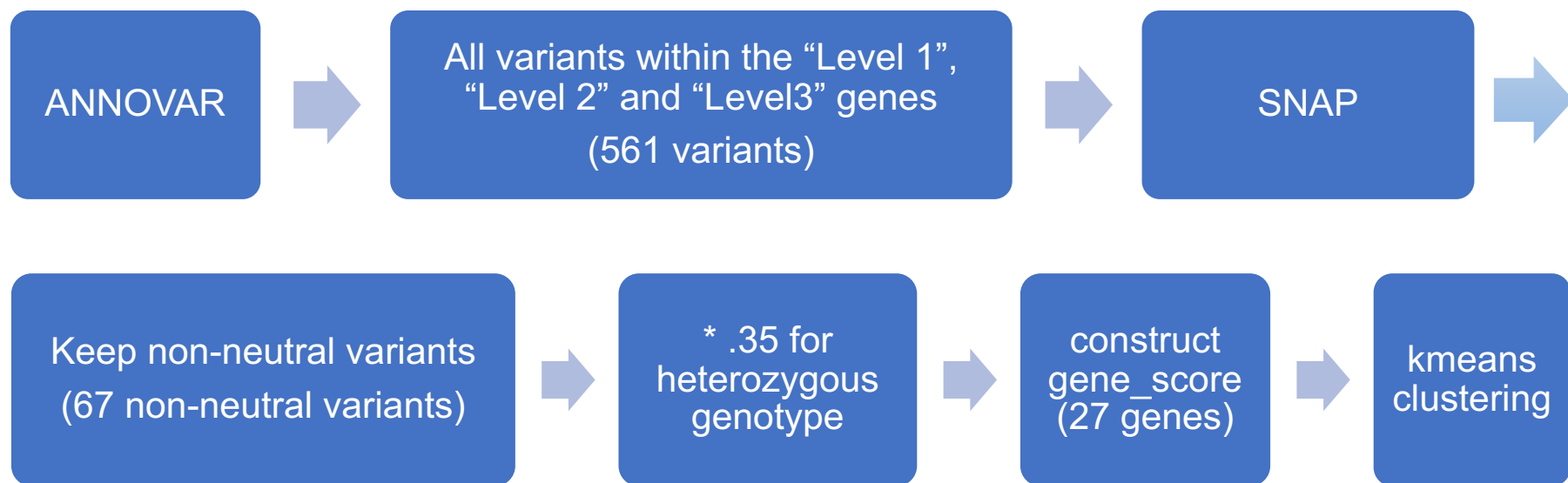
Accuracy: $(46+18)/103 = 62.1\%$

Precision: $46/(46+19) = 70.8\%$

Recall: $46/(46+20) = 69.7\%$



Submission 4 – Selected genes



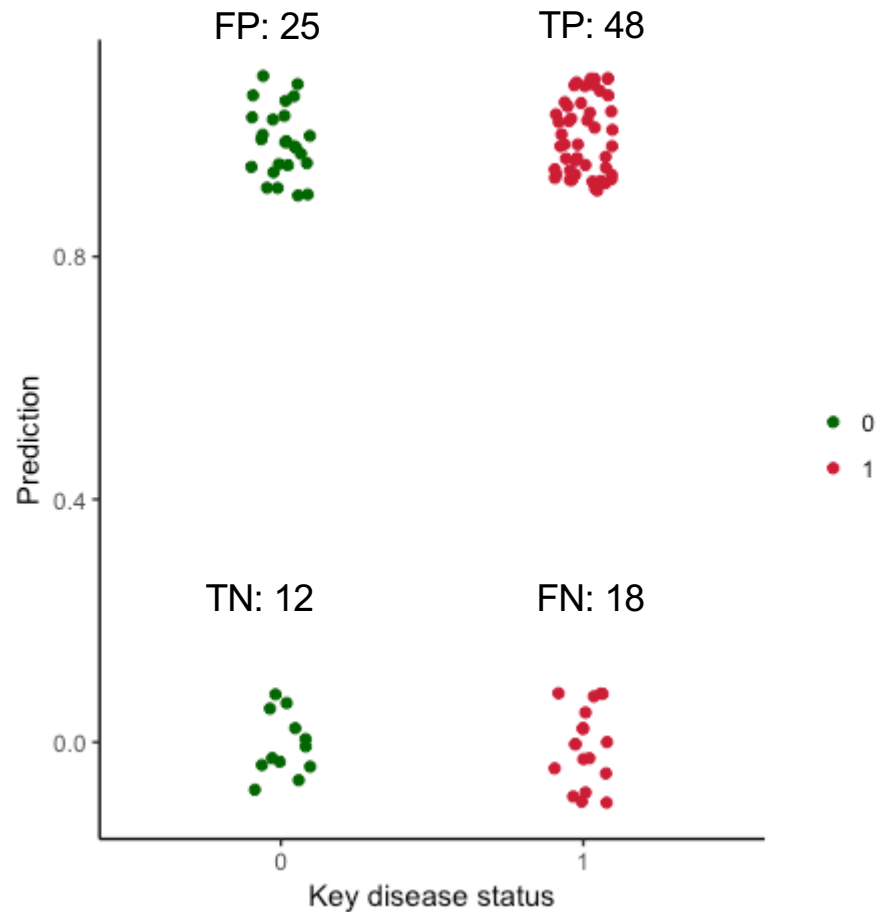
$$gene_score = \prod_i^N het_i \times (1 - score_i)$$

Submission 4 – Selected genes

Accuracy: $(48+12)/103 = 58.3\%$

Precision: $48/(48+19) = 65.8\%$

Recall: $48/(48+20) = 72.7\%$



Some thoughts

- Differentiating VTE from healthy instead of from AF
 - ~70% DVT has PE; AF is concomitant with PE (Keller et al. 2015)
 - AF patients have higher risk of both DVT and PE (Enga et al. 2015)
 - VTE and AF has similar pathophysiologic basis (Shariff et al. 2012)
 - F2, F3, F5, F7 are also recorded in DisGeNET for “AF”
- Relationships between Warfarin dosage and VTE risk
- Clinical application

Acknowledgement

CAGI Organizer

Dr. Steven Brenner

Dr. John Moulton

Dr. Gaia Andreoletti

and other organizers

Data provider

Dr. Russ Altman

Dr. Roxana Daneshjou

Bromberg Lab

Dr. Yana Bromberg

Dr. Maximilian Miller

Dr. Chengsheng Zhu

Yannick Mahlich

Zishuo Zeng

Nicholas Lusskin