Dear Dr. Tang,

Congratulations! The scientific review of your paper proposal, *Whole genome analysis of venous thromboembolism (VTE)*, has been approved by the reviewers in the TOPMed Publications Committee.

The next step for full approval of this paper proposal is the selection of the data sets you would like to use, and the subsequent approval of those selections by the designated data set contacts. That process is described here: [https://www.nhlbiwgs.org/paperproposals/data-sets](https://www.nhlbiwgs.org/paperproposals/data-sets)

Although you have completed the scientific review process for your paper proposal, you may also find this overview of the entire process useful: [https://www.nhlbiwgs.org/paperproposals/about](https://www.nhlbiwgs.org/paperproposals/about)

You can begin analysis as soon as the data sets are approved and sequence and phenotype data are in hand (including harmonized phenotypes for cross-cohort analyses).

We encourage your collaboration with other TOPMed studies and your participation in relevant TOPMed working groups in carrying out these studies.

We ask that you include the TOPMed authorship banner, which is under development, in your authorship plans.

Please see also the TOPMed Publications Policy: [https://www.nhlbiwgs.org/nhlbi-trans%28E2%80%90omics-precision-medicine-topmed-publication-and-data-access-policy](https://www.nhlbiwgs.org/nhlbi-trans%28E2%80%90omics-precision-medicine-topmed-publication-and-data-access-policy)

All submitted TOPMed WGS paper proposals can be viewed on the DCC web site: [https://www.nhlbiwgs.org/paperproposals](https://www.nhlbiwgs.org/paperproposals)

Thank you very much for taking the time to write these proposals. Your efforts on behalf of TOPMed are very much appreciated.

Best wishes,

The TOPMed Publications Committee

Scientific reviewers’ anonymized comments:

- **Review Date:** 07/08/16  
  **Review Status:** Approve  
  **Review Comments:** No concerns

- **Review Date:** 07/08/16  
  **Review Status:** Approve  
  **Review Comments:** No concerns

- **Review Date:** 07/07/16  
  **Review Status:** Approve  
  **Review Comments:** No concerns
Scientific review approval for your TOPMed paper proposal, and next steps

- Review Date: 07/05/16
  Review Status: Approve
  Review Comments: no concerns

- Review Date: 07/05/16
  Review Status: Approve
  Review Comments: Looks good.

- Review Date: 07/02/16
  Review Status: Conflicted
  Review Comments: Not actually conflicted--there is no selection for ex officio. If voting would approve.

- Review Date: 07/01/16
  Review Status: Approve
  Review Comments: none

- Review Date: 07/01/16
  Review Status: Approve
  Review Comments: I would be careful with the word 'causal' and instead focus on 'functional.'
Whole genome analysis of venous thromboembolism (VTE)

**Aims:** Conduct whole genome association analyses of VTE in the TOPMed Program.

1. Interrogate candidate regions of the genome to identify rare genetic variations associated with VTE risk. We will utilize the WGS data to identify causal variants in loci previously associated with the risk of VTE, including FGG, F11, PROCR, F8, TSPAN15, and SLC44A2.

2. Interrogate agnostically all other regions of the genome to identify common and rare genetic variation associated with VTE risk. For both aims, we will conduct analyses of annotated features of the WGS data based on protein coding and regulatory functions, as well as analyses of structural variants, sliding windows, noncoding RNA genes, and single nucleotide variants (SNVs).

**Were all of these aims included in the original study's grant application?:** Yes

**Proposal co-authors (Last, First):**

- Boerwinkle, Eric
- Chen, Ming-Huei
- O'Donnell, Christopher
- de Andrade, Mariza
- Heckbert, Susan
- Johnsen, Jill
- Johnson, Andrew
- Kooperberg, Charles
- Pankow, James
- Pankratz, Nathan
- Psaty, Bruce
- Puurunen, Marja
- Reiner, Alex
- Smith, Nicholas
- Informatics Research Center representative TBD
Phenotype trait domains/clinical data:

VTE Working Group

Will your proposed analyses include all genomes sequenced through the overall TOPMed program, or a subset?:

Subset of genomes

If subset, please specify:

Subset of TOPMed studies that have adjudicated VTE cases that include TOPMed VTE (ARIC, CHS, HVH, and Mayo), FHS, and WHI

Brief description of project analysis plans or lookup requested (½-1 ½ page):

Twin and family studies suggest a strong genetic influence in VTE risk (heritability = 0.5 to 0.6) with possible contribution from gene by environmental interactions. Of the nine genes/loci that have been reported for VTE in the GWAS studies, the most significant variants in FGG, F11, ABO, TSPAN15, PROC and F8 were intronic, intergenic, or in coding regions but not translated, bearing no known functional relevance in influencing protein structure. Moreover, these variants are not in strong LD with common, non-synonymous variants included on the GWAS platforms. We think that whole genome sequencing (WGS), which comprehensively captures regulatory regions and non-coding RNA genes, will markedly enhance genomic discovery for VTE.

We will participate with the TOPMed Phenotype Harmonization and Data Coordinating Center to harmonize the VTE phenotype and prepare datasets for analysis. VTE cases from the ARIC + VTE project (Atherosclerosis Risk in Communities Study, Cardiovascular Health Study, Heart and Vascular Health Study, and Mayo VTE Study) will be analyzed together with ARIC controls as these cases and controls were frequency matched by race/ethnicity. Women’s Health Initiative cases and controls will be analyzed together as they include participants from up to 40 clinical centers and were selected from both the hormone therapy trial and observational study. We will first stratify the analysis by hormone therapy trial status and then in the trial participants the analysis will be further stratified by treatment assignment. The association results from the three strata will then be pooled by meta-analysis. Framingham cases and controls will be analyzed together as they were selected from multi-generational pedigrees in a single community. To combine results across studies, we will use a fixed-effect meta-analysis approach and evaluate heterogeneity between studies. If significant heterogeneity is detected, a random effects model will be rerun to obtain accurate effect estimates.

Since novel common (MAF > 5%) or low-frequency variants (MAF = 1%-5%) might be detected in regions that are not tagged well or present on the GWAS imputation panels, especially those outside of exons, we will evaluate these variants individually, using a log-additive model, for association with VTE. Within the genomic subset of non-coding variation, the primary focus will be on those variants
that fall within regulatory elements (e.g., as delineated by ENCODE). We will conduct single variant analyses using a logistic regression model that will control for age, sex, and ancestry after stratifying by race/ethnicity. We will also explore other methods such as the recently published approach that controls for population structure and relatedness when correlating single variants with binary traits by using the generalized linear mixed model association test (GMMAT). We will also include a weighting scheme in our analyses based on predicted functional impact (e.g., CADD score).

Several statistical methods have been proposed to detect rare variant associations using sequencing data, including the group-based collapsing methods, also known as burden tests, where all rare variants in a genomic region are collapsed into a single variable. Several extensions of this test were proposed by incorporating the weights of the variant MAF, variable threshold, the adaptive summation that distinguishes the variant effect direction. Another popular approach is the kernel-based methods such as sequence kernel association test (SKAT) and SKAT-O that allow variants in the same region to have effects in opposite directions and still both contribute to the association. For instance, a sliding window approach across the genome aggregates rare variants within a physical window and relates them to the trait of interest (VTE) by multiple methods. Based on experience with simulated and real data, windows may be defined as 4 kb in length and begin at position 0 bp for each chromosome, with a skip length of 2 kb. For instance, the T5 statistic counts the number of variant alleles across all variants in a given window with minor allele frequency MAF < 5%. These counts are then used in logistic regression models. The T5 test is powerful when all variants in the defined window have the same magnitude and direction of effect, while SKAT-O will be performed to allow for differences in the direction and magnitude of the effect sizes for the included variants. In addition to a sliding window approach, annotated features, such as enhancer elements defined by DNase hypersensitivity or chromatin immunoprecipitation (ChIP-seq) experiments, non-coding RNAs, protein coding genes, as well as entire pathways, may be used as the unit of analysis and inference in aggregate rare variant tests. We will also pursue gene set analyses to detect any enrichment of rare functional variants in a set of genes of interest, and will focus on the four pathways known to be relevant to VTE: anticoagulant, procoagulant, fibrinolytic and innate immunity.

<table>
<thead>
<tr>
<th>How may this analysis overlap other TOPMed projects:</th>
<th>No overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are your current plans for publication and authorship:</td>
<td>Co-authors from each participating study providing VTE cases or controls, the Data Coordinating Center, Informatics Research Center, Sequencing Centers, and any other interested member of the TOPMed VTE working group will be included.</td>
</tr>
<tr>
<td>Will data be shared across institutions? Explain.:</td>
<td>Yes, analyses will be done in at least two different institutions, so data from participating cohorts will need to be downloaded from the dbGAP exchange area for analysis.</td>
</tr>
<tr>
<td><strong>Will individual-level data be distributed to investigators outside TOPMed? Explain:</strong></td>
<td>No</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Study or Studies associated with this proposal:</strong></td>
<td>FHS VTE WHI</td>
</tr>
<tr>
<td><strong>Scientific review status:</strong></td>
<td>Pending review</td>
</tr>
<tr>
<td><strong>Data Request Status:</strong></td>
<td>Datasets not yet requested.</td>
</tr>
<tr>
<td><strong>Overall Status:</strong></td>
<td>In Progress</td>
</tr>
</tbody>
</table>