Assessment of pQTL method performance reveals optimal proteogenomic approach to assess the impact of genetic variation on plasma protein levels

Guhan Ram Venkataraman[1]*, Harendra Guturu[1], Ryan W. Benz[1], Khatereh Motamedchaboki[1], Margaret K.R. Donavan[1], Matthijs B. De Geus[2], Sudeshna Das[2], Pia Kivisäkk[2], Steven E. Arnold[2], Serafim Batzoglou[1], and Asim Siddiqui[1]

[1]Seer, Inc., Redwood City, CA 94065, USA, [2]Massachusetts General Hospital, Boston, MA 02114, USA

The Proteograph™ Product Suite enables the robust detection of pQTLs in an unbiased manner

pQTLs identified in an Alzheimer's Disease cohort are enriched for complement and coagulation cascades and extracellular activities

Quantitative trait loci (QTL) analysis is a useful tool for understanding the genetic etiologies of the molecular mechanisms underlying human health and disease. While a vast body of expression QTL (eQTL) studies have provided important insights into the relationship between genetic variation and gene expression, studying the impact of genetic variation on protein levels could provide greater insights into complex human biology. Recent advancements in proteomics – unbiased, deep, and scalable assessment of the plasma proteome using physicochemically distinct nanoparticles coupled with liquid chromatography-mass spectrometry (LC-MS)[1] – have enabled high-resolution protein QTL (pQTL) analysis. While proteogenomics approaches have potential for a better understanding of human health through the identification of new disease biomarkers and drug targets or the development of new diagnostic tools to improve disease prediction, a comprehensive evaluation of existing computational tools employed to perform pQTL analysis should be performed to ensure reliable and sensitive results. Here, we used intensities from 5,061 proteins groups (acquired using the Proteograph™ Workflow and LC-MS measurements) and genotyping array data to compute pQTL associations in 184 plasma samples.

Methods

Genotype QC

From a starting dataset comprising 184 samples with genotyped data at 665,608 sites, we took variants and samples with a call-rate ≥ 0.9 for a total of 174 samples across 481,798 sites.

Protein quantifications

We acquired Proteograph™ workflow and LC-MS measurements across the same 184 samples and performed peptide and protein inference using DIA-NN 1.8[2]. We used a variety of spectral libraries for inference, including an *in silico* predicted library; fractionated plasma (DDA); project-specific cohort plasma (DDA); and project-specific cohort plasma (DIA).

Genotype Association Comparison Analysis

We used three programs; BOLT[3], PLINK[4], and REGENIE[5]; to compute associations between the genotypes and protein quantifications (i.e., pQTLs) and compared the number of pQTLs (total, *cis,* and *trans*) found amongst the three programs.

Gene-based functional enrichment testing and comparison to pQTL database

We used the online gProfiler tool to find functional enrichments amongst the pQTLs in our Alzheimer's Disease cohort.

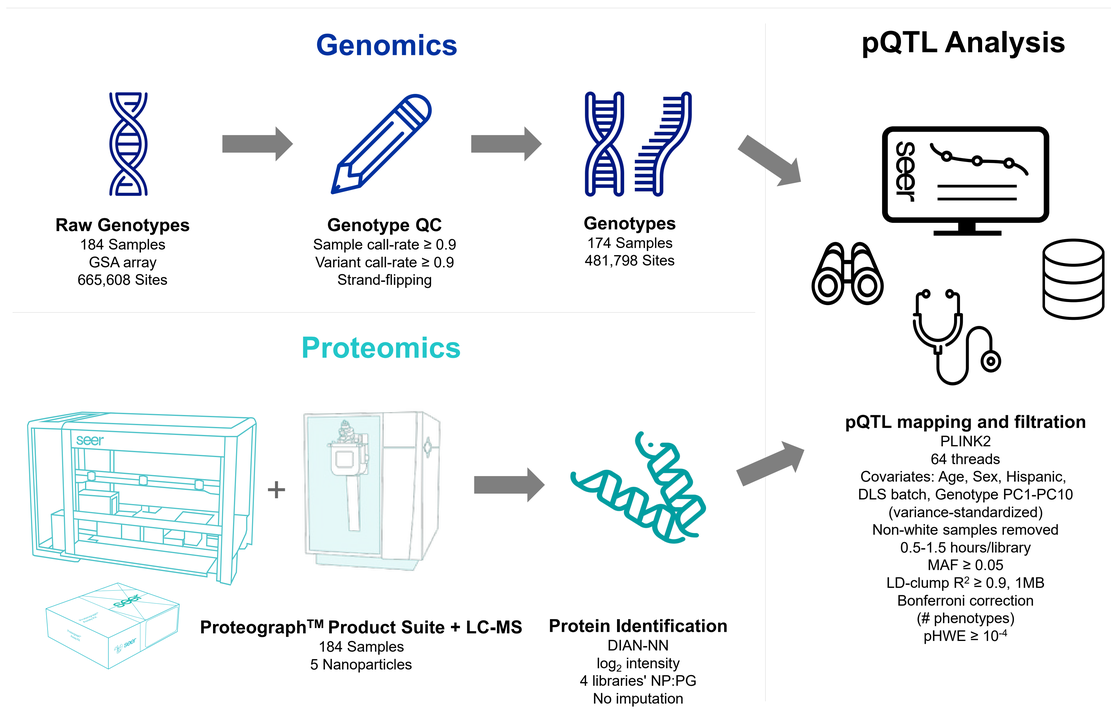

Figure 1. pQTL Analysis Workflow.

In our pQTL analysis workflow, 184 human plasma samples from a balanced Alzheimer's Disease (AD) cohort were genotyped on the GSA array. The same samples were processed using the Proteograph™ workflow and a panel of five proprietary engineered nanoparticles. After processing, an LC-MS analysis was conducted on the digested peptides. These data were then interpreted by the DIA-NN data analysis pipeline using four spectral libraries: the *in silico* predicted library; the fractionated plasma (DDA) library; the DDA project-specific cohort plasma library; and the DIA project-specific cohort plasma library. First, we selected the optimal genetic association program as per its performance on the *in silico* predicted spectral library. Then, we compared the number and spread of pQTLs across various libraries. Finally, we performed protein-based enrichment testing on the maximal set of pQTLs from these libraries to link the genetic etiology of the diseases to function.

Program Benchmarking Results

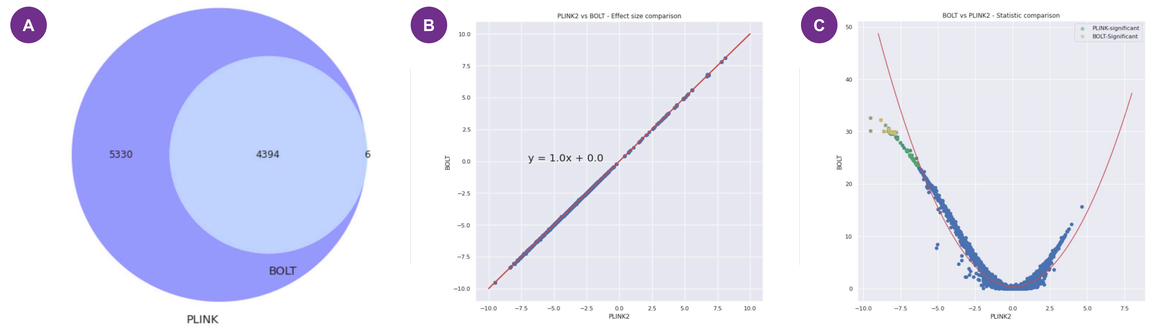

Figure 2. Summary of Findings from Program Benchmarking Analysis.

A) Comparison of PLINK and BOLT Associations. REGENIE was found to be virtually identical in association effect sizes and association p-values to BOLT, but slower; for the sake of simplicity, we dropped REGENIE from the analysis. BOLT finds almost a strict subset of the associations (SNP, PG pairs) that PLINK finds. **B) Comparison of PLINK and BOLT Effect Sizes.** Across the association "hits" that are shared between the two programs, the effect sizes are identical. **C) Comparison of PLINK and BOLT Statistics.** PLINK and BOLT statistics exhibit a parabolic relationship; while the PLINK statistic is derived from the *t*-distribution (with degrees of freedom corresponding to the number of samples in the analysis), the BOLT statistic is derived from the X^2 distribution. BOLT finds a lower number of associations. We found PLINK to be twice as fast as BOLT; thus, we proceeded forward with PLINK for the inter-library analysis.

Library	*In silico* Predicted Library	Fractionated Plasma - DDA	Project-specific Cohort Plasma - DDA	Project-specific Cohort Plasma - DIA (Match between runs)
Number of Total NP:PGs	22752	14648	20940	20889
Number of Total PGs	5061	2976	4269	4198
Number of Processed NP:PGs	13625	12180	16457	18799
Number of Processed PGs	3787	2837	3980	4136
Bonferroni threshold	1.32E-11	1.76E-11	1.26E-11	1.21E-11
pQTLs (NP:PG)	69	81	95	110
pQTLs (PG)	56	63	70	70
SNPs	54	60	64	64
PGs	25	31	33	29
cis-pQTLs (PG)	22	27	35	40
cis SNPs	22	26	35	37
cis PGs	12	18	17	18
trans-pQTLs (PG)	34	36	35	30
trans SNPs	32	34	30	27
trans PGs	15	14	17	12

Table 1. pQTL Comparison Across Libraries.

We used PLINK with age, sex, Hispanic ancestry, batch, and the first ten genetic principal components to compute pQTLs across various libraries. We found that project-specific libraries had decreased sparsity as compared to generic library-free or standard timsTOF libraries, leading to less protein drop-out from analysis (and thus more pQTLs). We found that generally, *cis*- and *trans*-pQTLs were found at equal rates, post-Bonferroni correction.

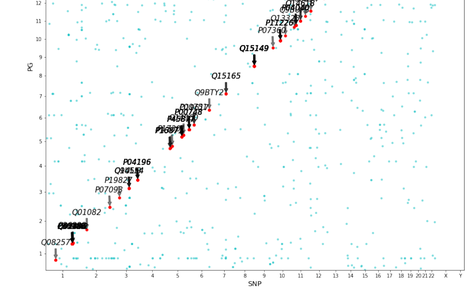

Figure 2. pQTL Distribution Across the Genome for Project-specific Cohort Plasma Library (DIA).

X-axis depicts SNP chromosome and position; *y*-axis depicts protein group (PG) chromosome and position. *Cis*-pQTLs are shown in red; *trans*-pQTLs are shown in blue.

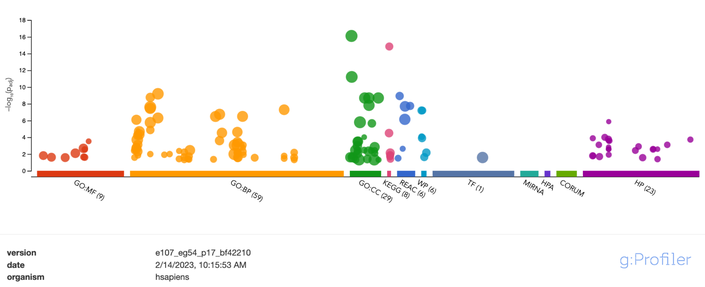

Figure 3. Gene-based Functional Enrichment Testing Results (gProfiler).

We find that the pQTLs found across the libraries are enriched for several complement, coagulation, and extracellular terms across several ontologies, with top terms: "GO:CC: extracellular region" (p = 7.87 x 10^{-17}), "KEGG: complement and coagulation cascades" (p = 1.42 x 10^{-15}), "GO:CC: extracellular space" (p = 6.18 x 10^{-12}), "GO:BP: response to external stimulus" (p = 6.39 x 10^{-10}), and "REAC: complement cascade" (p = 1.17 x 10^{-9}). These term enrichments are in line with previous research on Alzheimer's linking complement systems and their extracellular regulators to clinical expression of the disease[6].

Conclusions

- We found REGENIE to be slower than BOLT but the two to be equivalent in performance and results, whereas PLINK reported 2X faster results than BOLT for pQTL analysis.
- Project-specific libraries produce protein quantification results that are less sparse than standard libraries, leading to increased pQTL densities.
- Our samples being enriched for Alzheimer's Disease patients may have led to distinct pQTLs and pathways as compared to the INTERVAL dataset, which comes from blood donors (and thus, relatively healthy individuals).

References

[1] Blume et al. *Nat. Comm.* (2020) [2] Demichev, et al. *Nat. Methods.* (2020)
[3] Loh, et al. *Nat. Genet.* (2015) [4] Purcell et al. *AJHG.* (2007)
[5] Mbatchou, et al. *Nat. Genet.* (2021) [6] Dalakas et al. *Nat Rev.* (2020)

Acknowledgements

This work was supported by Research Grant No. 5R44AG065051-02 from the National Institute of Health.

Publications

Seer, Inc., Redwood City, CA 94065, USA | *gvenkataraman@seer.bio

Research funded by SBIR grant