somalogic

pQTL Analysis: the latest tool in disease research

Four letters, four compounds – adenine (A), cytosine (C), guanine (G), and thymine (T) – are the basic code "letters" of our genetic material, DNA. As elegant as the genetic code is, it is also quite dense with information, leading to many questions about how genetic variations affect our biology or health. How can the answers be found amongst all that tightly packed information?

Advances in computing power, rapid sequencing of the genetic code and new emerging technologies have opened the door to the possibility of deconvoluting the genetic code into something meaningful, even unlocking the mystery to how phenotypes manifest. Initially, work focused on determining locations on chromosomes that may be influencing complex phenotypes, such as height, and led to the development of quantitative trait locus (QTL) analysis[1].

With advances in proteomics, it is now feasible to search for genetic variants in a population that influence how much of a particular protein is made[2]. If a dose response is seen with a set of genetic variants, then that location on the chromosome is considered a protein quantitative trait locus (pQTL). Finding these pQTLs, which can include protein single nucleotide polymorphisms (pSNPs), may be the best approach for translating the deluge of genetic information into meaningful biological insights.

pQTLs can be further stratified as "cis," "trans" or "cis-trans" pQTLs (Figure 1). This designation depends on the location of the pQTL in relation to the open reading frame of the targeted protein's gene. Several definitions exist that provide the cutoff for determining when a given pQTL is cis or trans. Generally, a cis-pQTL affects a local gene that encodes the protein.

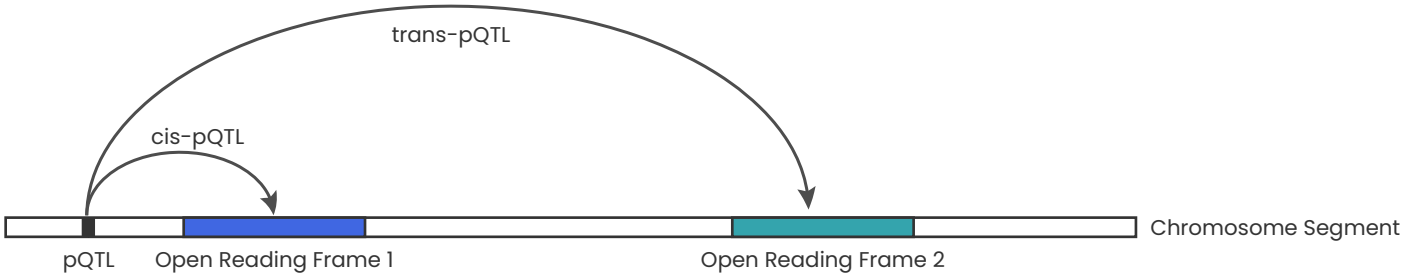

Figure 1. Designation of "cis," "trans" or "cis-trans" pQTL

A trans-pQTL is one that is quite a distance from the affected protein's gene. Instances have been reported where a cis-pQTL can act in trans, making it a "cis-trans-pQTL[3]." This means that the local gene and the distant gene are affected, and the protein levels will be affected.

pQTLs offer perhaps the best means of determining the contribution of DNA to our health through understanding the affected proteins. From our own experience and the growing body of literature, proteins offer a way of understanding the biology of disease and its progression. Seeing how protein concentrations rise and fall in real-time offers a wealth of information about chronic conditions including cardiovascular disease, respiratory disorders, diabetes, obesity, "unhealthy aging", etc. Trying to infer this type of information from mRNA levels is not equivalent[4–6].

What Can Proteomics Decrypt from the Genetic Code?

Only a few networks control the flow of information

Linking disease to a genetic origin is not a trivial endeavor, particularly if the disease is chronic and arises later in life. Unlocking how the body's information carried in blood is organized could make it easier. In a study of 5,457 Icelanders (over the age of 65) from the AGES Reykjavik study, researchers found that many proteins coursing through blood belonged to 27 networks (grouping based on non-random associations, such as sharing a functional relationship)[3]. No one source contributed entirely to a single network. The proteins contained within each network originated from many sources or tissues. The team's mRNA analysis could not recapitulate this observation. The hubs or lynchpins in these networks could be linked back to a variety of diseases and conditions, such as coronary heart disease, heart failure, type 2 diabetes, visceral and subcutaneous adipose tissue, and metabolic syndrome. Particularly of interest, the networks were found to be influenced by pQTLs, which may affect how the diseases manifest.

Discovering new pathways that contribute to disease risk

In one of the largest studies of its kind, a global collaboration led to interesting insights about trans-pQTLs contributing to a person's risk of disease susceptibility[7]. With nearly 12 million protein measurements from 3,301 individuals from the INTERVAL study (https://www.intervalstudy.org.uk/), the team found total 1,927 pQTLs, 89% of which they reported for the first time. Similar to the Iceland study, only 40% of the pQTLs were found to influence mRNA levels, reaffirming that mRNA levels do not correlate well with actual protein levels. Particularly of interest, the authors note that the trans-pQTLs identify new biological pathways that contribute to disease risk, which may be difficult to discover via other means.

Connecting genetics to disease outcomes with global implications

pQTLs can provide valuable insight into mechanisms of disease. Using data from 1,000 Germans who took part in the KORA study, Suhre et al. found a total of 539 pQTLs[8]. They further replicated 82% of these associations in samples from 338 people of Chinese or Arab descent who partook in the QMDiab study. For the trans-pQTLs, the authors noted interesting links to Alzheimer's disease, autoimmune disorders, cancer, cardiovascular disease etc., which could further the biological understanding of how these diseases manifest in people.

Technology Limits pQTL's Potential

Although pQTLs have tremendous potential, the technology used to find them is often limited. To find associated sites on chromosomes that may influence protein expression levels, the protein measurement technology must be able to measure a broad dynamic range of protein levels, look at thousands of proteins in a short time span and use a small sample volume. Although technologies exist for measuring proteins, many of them have limits that compromise pQTL identification.

Antibody based assays

Antibody based assays have been vital in unlocking the mysteries of the human body. Yet, they do present challenges. For use in multiplex assays, roughly only 30 to 50 antibodies at a time can be used due to cross reactivity[9]. New methodologies have marginally expanded the number, but other antibody shortcomings, such as batch to batch variation and lack of validation in some cases, still persist and can compromise the findings[10]. Even if the technology further improves, how many different assays need to be performed and how much sample would be needed to span the proteome (provided exquisitely specific antibodies even exist for all 20,000 proteins)?

Mass spectrometry

Mass spectrometry has long been the workhorse when it comes to analyzing proteins. With regards to proteomics, the use of the technology can lead to compromises on the type of information they can realistically acquire. Top-down (looking at whole proteins) methods are typically limited to proteins smaller than ~70 kDa, which excludes many proteins of interest[11]. Hence, the bottom-up (proteins are fragmented) approaches have become the favored means[11]. While the bottom-up approaches can be used to look at proteins of all sizes, it can be tricky to quantitate the plasma proteome, because ion suppression from abundant proteins reduces the dynamic range of protein concentrations that can be analyzed[11,12].

Protocols, such as depletion, do exist for dealing with this conundrum, but they are not without their own pitfalls. For example, depletion requires large amounts of starting material, making plasma proteomics challenging, and the act of depleting itself changes the proteomic make-up of a sample. In the end, a compromise will always have to be made with regards to the number of samples that can be analyzed in a high throughput manner, the number of proteins identified, the number of proteins quantitated and the ability to measure very low abundance proteins[11,12].

SomaScan® Assay

The SomaScan Assay is a proteomics tool that does not suffer from the same issues or force a researcher to compromise in the way other technologies do. As synthetically made compounds, SOMAmer® (Slow Off rate Modified Aptamers) reagents (protein binding agents on which the SomaScan Assay relies) are

free from the production variability seen in the production of antibodies, which require cells or animals. Also, researchers can easily measure a quarter of the translated proteome over a wide dynamic range of proteins (from femtomolar to micromolar) in thousands of samples.

The SomaScan Assay simultaneously measures several times more content than top competing technologies and offers significantly better precision, which is reflected in the lower coefficients of variation (%CV).

In a direct comparison of 912 common analytes, 90% of intra-assay CVs, and 97.5% of inter-assay CVs, are better than reported by the competition. SomaScan Assay's median intra-assay variation (within-run) and inter-assay variation (between runs) CV are 3.5% and 3.2%, respectively. The closest competitor reports CVs of 11% within-run and 22% between-runs. The SomaScan Assay offers better precision within and across runs, in addition to simultaneous detection of thousands of additional protein analytes.

SomaScan Assay and the Reproducible pQTL Hunt

For the pQTL search, the SomaScan Assay checks off many (if not all) of the requirements needed to find meaning in the genetic code and more people are beginning to use the technology as a result. In the literature, a simple search in PubMed for "pQTL" yields 46 references. Looking at just the ones related to human health, roughly a quarter of them used SomaScan technology to decrypt the genome. Table 1 illustrates a sampling of the pQTL bounty found using the SomaScan Assay.

Table 1. Sample Compilation of Research Using the SomaScan Technology to Find pQTLs

Reference	Number of People in Study	Number of Proteins	Total Number of pQTLs	Number of cis-pQTLs	Number of trans-pQTLs	PubMed ID
(Yao et al., 2018)[13]	3301	3620	55	23	33	30111768
(Hess et al., 2018)[14]	512	1129	15 (baseline); 20 (dieting)			29619113
(Carayol et al., 2017)[15]	494	1129	56 (baseline); 3 (dieting)	34 (baseline)	22 (baseline); 3 (dieting)	29234017
(Di Narzo et al., 2017)[16]	51	1128	41 (FDR=.1); 3424 (FDR=.5)	41 (FDR=.1); 3424 (FDR=.5)		28129359
(Sasayama et al., 2017)[17]	133	1126	446 to 1001 (cis-only analysis)	421 (580 additional in cis-only analysis)	25	28031287
(Lourdusamy et al., 2017)[18]	96	813	2106	2106		22595970
(Suhre et al., 2017)[19]	1000	1124	532(539)	384	148	28240269

With the pQTL hunt, reproducibility is always a concern. Emilison et al. sought to evaluate the reproducibility of pQTLs identified using the SomaScan Assay amongst different studies (which had different demographics) and other pQTL hunts using different proteomics technologies, such as mass spectrometry or immunoassays[3]. Although the team excluded proteins found on the X-chromosome, the comparisons were quite remarkable. For mass spectrometry derived data, Emilsson et al. reproduced 87.5 to 100% of the cis-pQTLs[3,19,20]. For cis-pQTLs found using immunoassays, 62.5 to 73.9% were reproduced[3,21,22]. The team also was able to reproducibly find 75.7 to 88.3% of cis-pQTLs and 72.8 to 84.5% of trans-pQTLs found via SomaScan technology[3,7,8].

Conclusion

Genomes pack in quite a bit information. The understanding about how biological pathways work in concert and contribute to our health or health trajectory can be broadened by extracting useful information from it. One useful way harnesses the power of proteomics to find pQTLs that can offer that wanted insight. Yet, the quality of pQTLs is hugely dependent on the quality of proteomic technology used to find them. The SomaScan Assay offers the quality and ability to deliver the data needed to find pQTLs that decrypt the genome.

With the pQTL hunt, reproducibility is always a concern. Emilison et al. sought to evaluate the reproducibility of pQTLs identified using the SomaScan Assay amongst different studies (which had different demographics) and other pQTL hunts using different proteomics technologies, such as mass spectrometry or immunoassays[3]. Although the team excluded proteins found on the X-chromosome, the comparisons were quite remarkable. For mass spectrometry derived data, Emilsson et al. reproduced 87.5 to 100% of the cis-pQTLs[3,19,20]. For cis-pQTLs found using immunoassays, 62.5 to 73.9% were reproduced[3,21,22]. The team also was able to reproducibly find 75.7 to 88.3% of cis-pQTLs and 72.8 to 84.5% of trans-pQTLs found via SomaScan technology[3,7,8].

Sun et al. sought to reproduce their findings from SomaScan data using immunoassay-based data for 163 pQTLs[7]. In the replication, only 106 were found with 81% cis and 52% trans. The authors postulated that the low replication rate for trans-pQTLs may involve assay differences that may affect the presentation of the protein's epitope site.

References

1. Miles, C. & Wayne, M. Quantitative Trait Locus (QTL) Analysis. Nature Education 1, 208 (2008).
2. Horvatovich, P., Franke, L. & Bischoff, R. Proteomic studies related to genetic determinants of variability in protein concentrations. J Proteome Res 13, 5-14, doi:10.1021/pr400765y (2014).
3. Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. Science, doi:10.1126/science.aaq1327 (2018).
4. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 13, 227-232, doi:10.1038/nrg3185 (2012).
5. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? Nature 547, E19-E20, doi:10.1038/nature22293 (2017).
6. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell 165, 535-550, doi:10.1016/j.cell.2016.03.014 (2016).
7. Sun, B. B. et al. Genomic atlas of the human plasma proteome. Nature 558, 73-79, doi:10.1038/s41586-018-0175-2 (2018).
8. Suhre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun 8, 14357, doi:10.1038/ncomms14357 (2017).
9. Ellington, A. A., Kullo, I. J., Bailey, K. R. & Klee, G. G. Antibody-based protein multiplex platforms: technical and operational challenges. Clin Chem 56, 186-193, doi:10.1373/clinchem.2009.127514 (2010).
10. Voskuil, J. Commercial antibodies and their validation. F1000Res 3, 232, doi:10.12688/f1000research.4966.2 (2014).
11. Timp, W. & Timp, G. Beyond mass spectrometry, the next step in proteomics. Sci Adv 6, eaax8978, doi:10.1126/sciadv.aax8978 (2020).
12. Pappireddi, N., Martin, L. & Wuhr, M. A Review on Quantitative Multiplexed Proteomics. Chembiochem 20, 1210-1224, doi:10.1002/cbic.201800650 (2019).
13. Yao, C. et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat Commun 9, 3268, doi:10.1038/s41467-018-05512-x (2018).
14. Hess, A. L. et al. Analysis of circulating angiopoietin-like protein 3 and genetic variants in lipid metabolism and liver health: the DiOGenes study. Genes Nutr 13, 7, doi:10.1186/s12263-018-0597-3 (2018).
15. Carayol, J. et al. Protein quantitative trait locus study in obesity during weight-loss identifies a leptin regulator. Nat Commun 8, 2084, doi:10.1038/s41467-017-02182-z (2017).
16. Di Narzo, A. F. et al. High-Throughput Characterization of Blood Serum Proteomics of IBD Patients with Respect to Aging and Genetic Factors. PLoS Genet 13, e1006565, doi:10.1371/journal.pgen.1006565 (2017).
17. Sasayama, D. et al. Genome-wide quantitative trait loci mapping of the human cerebrospinal fluid proteome. Hum Mol Genet 26, 44-51, doi:10.1093/hmg/ddw366 (2017).
18. Lourdusamy, A. et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Hum Mol Genet 21, 3719-3726, doi:10.1093/hmg/dds186 (2012).
19. Johansson, A. et al. Identification of genetic variants influencing the human plasma proteome. Proc Natl Acad Sci U S A 110, 4673-4678, doi:10.1073/pnas.1217238110 (2013).
20. Liu, Y. et al. Quantitative variability of 342 plasma proteins in a human twin population. Mol Syst Biol 11, 786, doi:10.15252/msb.20145728 (2015).
21. Kim, S. et al. Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. PLoS One 8, e70269, doi:10.1371/journal.pone.0070269 (2013).
22. Enroth, S., Johansson, A., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. Nat Commun 5, 4684, doi:10.1038/ncomms5684 (2014).

somalogic

SL: 00000464 Rev 2: 2021-01
PQTL White Paper