Dimension reduction and covariation analysis with partial least squares

Introduction

Understanding the source of natural variation in complex traits

A major goal of genomics is to accurately quantify the covariation of two blocks of large data matrices (such as a chIPSeq, RNA Seq, DNA Seq, phenotype). With this, one could obtain a high-resolution understanding of how different 'omes (transcriptome, genome, microbiome, proteome, etc.) interact with one another to produce the effects being observed. This necessitates methods that allow for the measurement of how much (or how little) each variable in a large data set explains the variation in an observed phenotype to a high degree of nuance. The resulting data is very relevant to the study of the natural variation that we observe in biology, particularly towards understanding the basis of variation in complex traits.

The ways in which changes in genotype, channeled through development, lead to variation in organismal phenotype has long been fascinating for biologists. The morphological diversity produced by the information encoded in the genome of organisms is a prerequisite for evolution and is the basis of biodiversity. Natural variation in body pigmentation in the fruit fly (Dembeck et al. 2015), coat colour in the beach mouse (Linnen et al. 2009), and plate number in three spine sticklebacks (Colosimo et al. 2004) are some examples where we can map trait outcome to individual nucleotides, genes, or protein functions. There seems to be a straightforward and often mechanistic explanation for the trait observed. For example, an allele on the single gene, *Eda*, is responsible for the parallel evolution in the loss of armor plates seen in sticklebacks that have migrated from marine water to freshwater(Colosimo et al. 2005). For these examples, we have an unusually straight forward understanding of the ways in which genotype link directly to phenotype, or at least we can imagine a straight-forward experiment to measure if this is the case, for example a knock-out experiment. But the situation where large effect alleles cause a predictable change in phenotype is relatively rare. More common is the case that small changes to many genes produce small changes in different aspects of the phenotype being measured (Boyle, Li, and Pritchard 2017). This requires adopting analytical methods that allow for a holistic understanding of a system, rather than a 'one-gene-hypothesis', or even hypothesis-free framework (Lê Cao 2021). The results from holistic analysis can help identify genes of interest, and circle back to the one-gene-hypothesis through experimental work.

Partial Least Squares or Projection to Latent Components (PLS)

Partial Least Squares, or Projection to Latent Components (PLS), is a dimension reduction technique that can be applied in the context of studying variation in complex traits in various way (Boulesteix and Strimmer 2007). It can be used to integrate sample variables together, which can be used to reduce the number of dimensions of a dataset. It is well suited to handling any dataset where the number of variables (for example, gene counts) is far larger than the number of measurements (for example, number of individuals sequenced), or p >> n. Latent components, obtained by constructing linear combinations of variables through singular value decomposition, are constructed in a way that attempts to maximize the covariance between two blocks of data, or in the regression case, between the predictors and the response. This is unlike principal components analysis, which creates latent components based off maximizing axes of variation within a dataset, PLS maximizes axes of covariation between two data sets.

PLS is an active area of researcher, where new applications of PLS methods are continuously developed for use in different situations. Using the PLS algorithm rather than the ordinary least squares method, PLS regression allows for a regression analysis of two blocks of data. This is especially useful in cases where there is too much variance in the data to be able to analyze it through ordinary least squares regression. Two-block PLS (2BPLS) is an approach with a similar foundation as PLS regression, but the blocks of data are treated symmetrically, meaning that there is no X predictor and Y outcome, but the variation in both blocks of data is considered causal (Rohlf and Corti 2000). Sparse PLS is a version of PLS which involves feature selection, where variables which do not contribute enough to the response are penalized (Lê Cao et al. 2008). In all cases of PLS, the weight coefficients of the linear combinations of the original variables are stored in loading vectors, which give an idea of how much the original variables contribute to the response. Loading vectors associated with genes can be investigated to determine which genes are predicted to have the greatest effect on the phenotypic outcome, where the higher the absolute value of the loading vector, the greater the contribution of that variable.

My motivation is to understand the ways in which variation in gene expression covary with variation in trait shape and size, a step towards understanding the ways in which genotype produces a phenotype. The wing of *Drosophila melanogaster* is a system for modeling the development of complex traits. In this project I use two datasets, one for larval gene expression and one which quantifies the morphology of the adult wing. I analyse the covariation between variation in wing shape and variation in gene expression during development, using R to perform different methods for PLS. Three different PLS methods were used with the goal of identifying which genes have the highest loading values, or are predicted to have the highest effect on, wing shape. This includes 2BPLS, multivariate PLS regression, and sparse PLS. A sample of 372 genes known to be expressed in the wing during development (a 72 x 372 matrix) combined with shape data for the adult wing (a 72 x 96 matrix), the five genes with the highest absolute loading vector value were identified with each method, the variation in the expression of these genes is predicted to covary the greatest with the variation observed for wing shape.

Methods

Data Acquisition and Processing:

Although the association between genotype and phenotype is commonly studied by inducing mutations of large effects in an organism and studying the outcome, these mutations are not necessarily relevant to natural variation in these organisms, especially in *Drosophila* where many mutations are lethal unless balanced. The gene expression and shape data used for this project come from samples from the Drosophila Genetics Reference Panel (DGRP) (Mackay et al. 2012). The DGRP is a population of over 200 inbred Drosophila lines, and it is most used as a resource for studying common polymorphisms which affect complex trait variation. Therefore, studying the natural variation in the DGRP is a way to assess the genotype-phenotype relationship in a way that is more relevant to how complex traits have evolved (Dworkin et al. 2011).

Gene Expression in the developing wing tissue

Larval gene expression data was obtained from Jason Meezey and Yuxin Shi from the Meezey Lab in Cornell. The gene expression data represents single end reads, and compromises 108 samples, although a smaller subset of this was used. Sequences were trimmed using bbduk to remove adapter contamination and poor-quality reads. Quality control of the reads was performed with fastqc. The final dataset contained 372 genes for 72 samples.

Shape data for the adult wing

Measures of adult wing morphology were obtained from the Dworkin lab. This data was used in a previous study of wing variation in the DGRP (Pitchers et al. 2019), and has already been processed (shape coordinates have already been aligned in order to remove the effects of variation in location, orientation, and scale). This dataset contained 96 variables (48 landmarks with an x and y coordinate) for 72 samples.

Data Analysis:

All data was analyzed using packages available in R version 4.0.4.

Partial Least Squares

Two Block Partial Least Squares – Geomorph

In 2BPLS, the data in the blocks being analyzed are considered symmetric, meaning that the effect there is no assumed direction of effect (as in regression with an X predictor and Y response), and the variation in both blocks are considered (rather than in the X predictor alone). The function *two.b.pls()* from the geomorph package performs 2BPLS to investigate the degree of covariation between two blocks of data (Rohlf and Corti 2000). Originally designed to investigate modularity between blocks of shape data, any data matrix can be used.

A 3D array of landmark coordinates, representing shape, and a 2D matrix of gene counts were used for the 2BPLS analysis. A subset of the original gene expression data was obtained such that only 372 genes known to be expressed in the developing wing were considered. Only DGRP lines for which there existed both gene counts and wing shape data were considered, leaving a sample of 72 lines to analyze. Additionally, both males and females were included in the analysis without distinction between the sexes (although sex-level effects exist, variation between lines tends to be higher than variation between the sexes within a line). Only one replicate of each line was used rather than averaging between replicates (although between replicate variation was low). This was the dataset that was used for all PLS analyses, including PLS2 regression and sparse PLS.

Genes were then sorted by order of absolute PLS loading vector value, and the top 5 genes with the largest effect on covariation with wing shape were investigated with FlyBase (GO terms). This was done for all PLS analyses. Unique to *two.b.pls()*, a test value is also obtained from the function, which represents statistical significance (or lack thereof) of the covariation between gene expression and shape.

PLS2 Regression - pls

For this project, PLS regression was also considered. We lose the nuance of considering the variation in phenotype in a model of the covariation between change in gene expression and change in trait, but we can conduct the analysis from the angle that variation in gene expression is driving the response seen in variation of wing shape. PLS regression analyses in the case of a multivariate Y are sometimes referred to as PLS2. There are several algorithms used to solve the regression equation in PLS2, the one that was considered for this project was Statistically Inspired Modification of PLS (SIMPLS). This is the algorithm used in the *simpls.fit()* function from the package pls. In SIMPLS, the PLS is treated as an optimization problem, which can lead to better computational efficiency and be relatively interpretable. This is in contrast to the nonlinear iterative partial least squares algorithm, which is the algorithm used in the following sparse PLS analysis.

Sparse PLS - mixOmics

In data of high dimensions, it is common to have many measures which have minimal contribution to the predictive power of the model, and these measures also contribute to the potential for the model to overfit the data (perform very well on the training data and much more poorly on the testing data). Sparse PLS, where sparsity refers to a model with few non-zero parameters, is an application of PLS where the variables that do not contribute ‘enough’ to the outcome are penalized. MixOmics is an R package that contains several functions for various types of PLS, including sparse PLS. In the function *spls()*, sparsity of the model is achieved through a Lasso penalization of the PLS loading vectors when computing the singular value decomposition.

Results

Two-Block Partial Least Squares

FB gene ID	Gene Name	Loading vector value (absolute)	Associated GO Term(s)
FBgn0020618	Rack 1	5.764493e-01	Locomotory behavior, **wing disc development**, cuticle development, oogenesis
FBgn0011726	twinstar	5.164867e-01	Actin filament depolymerization, border follicle cell migration, compound eye development, imaginal disc-derived leg segmentation, meiotic cytokinesis
FBgn0020238	14-3-3ε	4.086474e-01	**Wing disc dorsal/ventral pattern formation,** pole cell migration, positive regulation of growth, microtubule sliding, DNA damage checkpoint
FBgn0283508	Narrow	1.470915e-01	**Regulation of imaginal disc-derived wing size**
FBgn0011760	Cut up	1.149775e-01	Autophagy, **imaginal disc-derived wing morphogenesis**, microtubule anchoring at centrosome, positive regulation of neuron remodeling, **wing disc development**

Table 1A: Gene ID, gene name, absolute loading vector value, and associated GO terms for the 5 genes with the highest absolute loading vector value as distinguished by two block partial least squares analysis with Geomorph.

r-PLS	0.364
Effect size (Z)	0.645
P-value	0.26

Table 1B: r-PLS (correlation coefficient between the scores of projected values on the first singular vectors of left and right blocks of data), the multivariate effect size, and P-value (empirically calculated from the resampling procedure, tests the hypothesis that there is significant covariation between the two blocks of data) values from *two.b.pls()*

PLS Regression with SIMPLS (component 1)

FB gene ID	Gene Name	Loading vector value (absolute)	Associated GO Term(s)
FBgn0020618	Rack 1	12013.278	Locomotory behavior, **wing disc development**, cuticle development, oogenesis
FBgn0011726	twinstar	10763.650	Actin filament depolymerization, border follicle cell migration, compound eye development, imaginal disc-derived leg segmentation, meiotic cytokinesis
FBgn0020238	14-3-3ε	8516.264	**Wing disc dorsal/ventral pattern formation,** pole cell migration, positive regulation of growth, microtubule sliding, DNA damage checkpoint
FBgn0283508	Narrow	3065.407	**Regulation of imaginal disc-derived wing size**
FBgn0011760	Cut up	2396.145	Autophagy, **imaginal disc-derived wing morphogenesis**, microtubule anchoring at centrosome, positive regulation of neuron remodeling, **wing disc development**

Table 2: Gene ID, gene name, absolute loading vector value, and associated GO terms for the 5 genes with the highest absolute loading vector value from the first component as distinguished by PLS2 regression with Pls.

Sparse PLS (component 1)

FB gene ID	Gene Name	Loading vector value (absolute)	Associated GO Term(s)
FBgn0043364	Cabut	0.57108801	Dpp signaling, dorsal closure, **wing disc dorsal/ventral pattern formation**
FBgn0033033	Scarface	0.57056076	Dorsal closure morphogenesis, imaginal disc morphogenesis, germ-band shortening
FBgn0041184	Socs36E	0.46552686	Border follicle cell migration, cell differentiation, compound eye pigmentation, haltere development,

			imaginal disc-derived wing vein morphogenesis
FBgn0001297	Kayak	0.36231903	DNA binding, DNA-binding transcription factor activity, protein binding
FBgn0020493	Dad	0.01812652	**Imaginal disc-derived wing morphogenesis, morphogenesis of larval imaginal disc epithelium, cell migration, anterior/posterior specification of the imaginal disc**

Table 3: Gene ID, gene name, absolute loading vector value, and associated GO terms for the 5 genes with the highest absolute loading vector value from the first component as distinguished by sparse PLS with mixOmics

Discussion

Three variations of a PLS analysis were applied to a dataset of *Drosophila melanogaster* gene counts from the developing wing a quantification of adult wing morphology.

2BPLS analysis and PLS2 regression analysis identified the same top candidate genes, in order of absolute loading value these are *Rack 1, 14-3-3ε, twinstar, narrow, and cutup* (Table 1A and 2). Interestingly, these methods identified the same genes which may suggest the assumed direction of causality is not very important. *Rack 1* is expressed in all developmental stages in many tissues in *Drosophila melanogaster*, including the developing wing. Rack 1 homozygotes are often lethal, but it has been found that of the wings of the survivors are blistered, abnormally small, or fail to unfold (Kadrmas et al. 2007). *14-3-3ε* is a necessary gene for several stages of development outside of wing development, but in the wing it contributes to dorsal-ventral boundary formation (Bejarano et al. 2008). *Twinstar* is another gene that is implicated in many aspects of development, particularly for its role in actin remodeling. In the wing, this is necessary for the development of wing hairs and in their orientation (and planar cell polarity patterning in general) (Blair et al. 2006). *Narrow* contributes to pathway which modulates Dumpy (an apical extracellular-matrix protein that shapes the wing) distribution in the wing (Ray et al. 2015). *Cut up* encodes a subunit of the cytoplasmic Dynein, it is another gene that contributes to several aspects of development, with mutants showing abnormal wing-veins (Ghosh-Roy et al. 2004).

The sparse PLS identified a completely different set of genes. This might suggest that 2BPLS and PLS2 regression overfit the model, and that once genes with less predictive power are removed a different set of genes becomes relevant. The genes identified by sparse PLS were *cabut, scarface, socs36E, kayak, and dad* (Table 3). *Cabut* positively regulates the expression of *STAT92E, spalt*, and *optomotorblind* genes in wing imaginal discs which are involved in wing patterning and cell proliferation (Belacortu et al. 2012). *Scarface* modulates the corss-talk between the Egfr, bsk/JNK, and dpp signal transduction pathways, the products of which are known to contribute to wing shape by contributing to growth regulation, cell survival, and developmental patterning generally in the imaginal discs. *Socs36E*, among other roles in development, is a negative regulator of EGFR pathways via JAK-STAT. The inhibition of *kayak* has been shown to lead to a reduction in both wing and eye disc size (Hyun et al. 2006). *Dad* transcripts have previously been found to be associated with wing shape (Dworkin et al. 2011).

It should be apparent that there are no single genes which are responsible for wing shape, which has been found by several previous studies. Genes do not target shape directly, but rather the developmental processes which alter shape (for example, through wing vein formation). The genes which have been found to be implicated in wing shape through this analysis are genes which play critical roles at several stages in development and often in several tissues, not just the wing. This suggests what has already been purported by others, which is that complex traits are shaped in through an omnigenic model (as opposed to a polygenic model) (Boyle, Li, and Pritchard 2017). The omnigenic model is a hypothesis that posits that gene regulatory networks are interconnected such that all genes end up having at least some minor effect on seemingly distantly related genes or gene products. This would make sense with the results of this project as the genes identified have many effects on several aspects of development, not only the wing. In the future, all the gene counts should be used rather than a subset of genes expressed during wing development to further investigate the breadth of genes that are even mildly associated with wing shape development. Although it can be difficult to parse what is a significant effect and what may be noise in the data.

The covariation between the two blocks of data is not statistically significant, with a P-value of 0.26 (Table 1B), and that the two blocks of data are weakly correlated. This is in agreeance with other studies (Dworkin et al. 2011) and suggests that there is not necessarily strong evidence for changes in gene expression to be linked with subtle changes in phenotype – at least not at the resolution of this project. On the other hand, this suggests that we might not be looking at the whole picture when considering only gene expression in the study of complex traits. In the future, it might be interesting to investigate the interplay between different 'omes and phenotype, including the transcriptome, proteome, metabolome, and microbiome. The 2BPLS analysis suggests that variation in gene expression covaries to some extent with the variation in wing shape, given an effect size of 0.645, but it is not the whole story. 2BPLS does not have to be restricted to only two blocks, and so it would be feasible to add more data to the analysis. Furthermore, there are methods of PLS which can account for group structure within data, known as group PLS. This would allow for the addition of gene pathways to the analysis.

Conclusion

Through 2BPLS, PLS2 regression, and sparse PLS regression, this project has identified 10 genes for which gene expression covaries with wing shape to some degree. The genes identified have roles in several stages of development and in several tissues, not only the wing, suggesting support for an omnigenic model of wing shape. There is not a strong covariation between gene expression and wing shape, suggesting that gene expression alone cannot account for complex trait variation, and an avenue for future study of complex trait variation could include the proteome, microbiome, and gene pathways and regulatory networks.

References

Bejarano, Fernando, Carlos M Luque, Héctor Herranz, Georgina Sorrosal, Neus Rafel, Thu Thuy Pham, and Marco Milán. 2008. "A Gain-of-Function Suppressor Screen for Genes Involved in Dorsal–Ventral Boundary Formation in the Drosophila Wing." *Genetics* 178 (1): 307–23. https://doi.org/10.1534/genetics.107.081869.

Belacortu, Yaiza, Ron Weiss, Sebastian Kadener, and Nuria Paricio. 2012. "Transcriptional Activity and Nuclear Localization of Cabut, the Drosophila Ortholog of Vertebrate TGF-β-Inducible Early-Response Gene (TIEG) Proteins." *PLOS ONE* 7 (2): e32004. https://doi.org/10.1371/journal.pone.0032004.

Blair, Adrienne, Andrew Tomlinson, Hung Pham, Kristin C. Gunsalus, Michael L. Goldberg, and Frank A. Laski. 2006. "Twinstar, the Drosophila Homolog of Cofilin/ADF, Is Required for Planar Cell Polarity Patterning." *Development* 133 (9): 1789–97. https://doi.org/10.1242/dev.02320.

Boulesteix, Anne-Laure, and Korbinian Strimmer. 2007. "Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data." *Briefings in Bioinformatics* 8 (1): 32–44. https://doi.org/10.1093/bib/bbl016.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." *Cell* 169 (7): 1177–86. https://doi.org/10.1016/j.cell.2017.05.038.

Colosimo, Pamela F., Kim E. Hosemann, Sarita Balabhadra, Guadalupe Villarreal, Mark Dickson, Jane Grimwood, Jeremy Schmutz, Richard M. Myers, Dolph Schluter, and David M. Kingsley. 2005. "Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles." *Science (New York, N.Y.)* 307 (5717): 1928–33. https://doi.org/10.1126/science.1107239.

Colosimo, Pamela F., Catherine L. Peichel, Kirsten Nereng, Benjamin K. Blackman, Michael D. Shapiro, Dolph Schluter, and David M. Kingsley. 2004. "The Genetic Architecture of Parallel Armor Plate Reduction in Threespine Sticklebacks." *PLOS Biology* 2 (5): e109. https://doi.org/10.1371/journal.pbio.0020109.

Dembeck, Lauren M, Wen Huang, Mary Anna Carbone, and Trudy F C Mackay. 2015. "Genetic Basis of Natural Variation in Body Pigmentation in Drosophila Melanogaster." *Fly* 9 (2): 75–81. https://doi.org/10.1080/19336934.2015.1102807.

Dworkin, Ian, Julie A. Anderson, Youssef Idaghdour, Erin Kennerly Parker, Eric A. Stone, and Greg Gibson. 2011. "The Effects of Weak Genetic Perturbations on the Transcriptome of the Wing Imaginal Disc and Its Association With Wing Shape in Drosophila Melanogaster." *Genetics* 187 (4): 1171–84. https://doi.org/10.1534/genetics.110.125922.

Ghosh-Roy, Anindya, Madhura Kulkarni, Vikash Kumar, Seema Shirolikar, and Krishanu Ray. 2004. “Cytoplasmic Dynein–Dynactin Complex Is Required for Spermatid Growth but Not Axoneme Assembly in Drosophila.” *Molecular Biology of the Cell* 15 (5): 2470–83. https://doi.org/10.1091/mbc.e03-11-0848.

Hyun, Joogyung, Isabelle Bécam, Constantin Yanicostas, and Dirk Bohmann. 2006. “Control of G2/M Transition by Drosophila Fos.” *Molecular and Cellular Biology* 26 (22): 8293–8302. https://doi.org/10.1128/MCB.02455-05.

Kadrmas, Julie L., Mark A. Smith, Stephen M. Pronovost, and Mary C. Beckerle. 2007. “Characterization of RACK1 Function in Drosophila Development.” *Developmental Dynamics* 236 (8): 2207–15. https://doi.org/10.1002/dvdy.21217.

klecao. n.d. “Webinar: MixOmics in 50 Minutes | MixOmics.” Accessed May 3, 2021. http://mixomics.org/2019/09/webinar-mixomics-in-50-minutes/.

Lê Cao, Kim-Anh, Debra Rossouw, Christèle Robert-Granié, and Philippe Besse. 2008. “A Sparse PLS for Variable Selection When Integrating Omics Data.” *Statistical Applications in Genetics and Molecular Biology* 7 (1): Article 35. https://doi.org/10.2202/1544-6115.1390.

Linnen, Catherine R., Evan P. Kingsley, Jeffrey D. Jensen, and Hopi E. Hoekstra. 2009. “On the Origin and Spread of an Adaptive Allele in Deer Mice.” *Science (New York, N.Y.)* 325 (5944): 1095–98. https://doi.org/10.1126/science.1175826.

Mackay, Trudy F. C., Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F. Ayroles, Dianhui Zhu, Sònia Casillas, et al. 2012. “The Drosophila Melanogaster Genetic Reference Panel.” *Nature* 482 (7384): 173–78. https://doi.org/10.1038/nature10811.

Pitchers, William, Jessica Nye, Eladio J. Márquez, Alycia Kowalski, Ian Dworkin, and David Houle. 2019. “A Multivariate Genome-Wide Association Study of Wing Shape in Drosophila Melanogaster.” *Genetics* 211 (4): 1429–47. https://doi.org/10.1534/genetics.118.301342.

Ray, Robert P., Alexis Matamoro-Vidal, Paulo S. Ribeiro, Nic Tapon, David Houle, Isaac Salazar-Ciudad, and Barry J. Thompson. 2015. “Patterned Anchorage to the Apical Extracellular Matrix Defines Tissue Shape in the Developing Appendages of Drosophila.” *Developmental Cell* 34 (3): 310–22. https://doi.org/10.1016/j.devcel.2015.06.019.

Rohlf, F. James, and Marco Corti. 2000. “Use of Two-Block Partial Least-Squares to Study Covariation in Shape.” *Systematic Biology* 49 (4): 740–53. https://doi.org/10.1080/106351500750049806.