Signatures of adaptation to environmental variability in the protein-protein interaction network

Csenge Petak
csenge.petak@uvm.edu

Abstract

Studying adaptation to variable environments is essential, not only to understand more about the theory of evolution but also to predict whether species will be able to adapt in time to the rapid global climate change we are currently experiencing. In this study, we wanted to find mutations putatively under selection related to environmental variability and see if these are located in genes with higher network centrality and level of expression. We sequenced the whole genome of a 140 purple sea urchins (Strongylocentrotus purpuratus) from 7 populations experiencing different environmental variability and calculated Bayenv factors from the SNP data. We found no linear relationship between Bayenv factor and centrality or level of expression, but specific genes were identified for further investigation with potential regulatory role (gli3 transcription factor and RNA exonuclease) and role in biomineralisation (fibrillin and calcium-binding protein). Moreover, we found that many genes with high Bayenv factor regarding to temperature variability also had high Bayenv factor regarding to frequency of pH under 7.8, which could indicate the existence of a general mechanism evolved to deal with highly variable environmental conditions. Overall, these results provide a great stepping stone to further studies.

Keywords: gene expression, protein-protein interaction network, gene regulation, adaptation, environmental variability, sea urchins

1 Introduction

In 1859 Charles Darwin gave us the fundamental mechanism for evolution which resulted in the birth of modern biology. That is, individuals with certain traits survive and reproduce more than others leading to an increase in the frequency of those traits in the population. Thanks to the work of countless scientists in the 20th century we understood how variation in traits originate and how it's inherited. These discoveries led to the Modern Synthesis that describes evolution by natural selection as a feedback loop between the generation of random genetic variation and selection. Even though this model of evolution explains a lot of life's diversity we observe in nature, there are still some essential questions in evolutionary biology that need to be answered. We now have a good understanding of how species adapt to static conditions, but how do species adapt to environmental variability and "remember" past selection to be prepared to reoccurring environmental stressors? The answer seem to be hidden in how genetic information translate into traits.

Traits are emergent properties of a very complex system of interacting gene products (i.e. proteins). Therefore, changes in the genome go through a highly complicated nonlinear mapping to changes in the traits of an organism. Just like a mutation from c to b in the sentence "I love my car and my coat" changes the meaning of the sentence differently depending on where the mutation happens, the effect of a mutation is highly dependent on the context as well. The context of a protein is other proteins in the cell. Proteins can interact with each other indirectly through gene expression regulation or directly through protein-protein binding. We can thus build an **interaction network** where the nodes represent proteins and edges represent either kinds of interaction (only regulatory interactions = gene regulatory network (GRN) [1], only protein-protein binding = protein-protein interaction network (PPIN)). Due to recent technological innovations we can build these networks quickly a cost effectively and weight the edges based on our confidence in the interaction. Many studies have shown that GRNs and PPINs can evolve through time. Furthermore, there are not only subject to evolution, but they themselves influence evolution as well. Many theoretical studies have hypothesised the central role of interaction networks in evolution, both in its rate and direction, and recent empirical studies are starting to back up these predictions.

Studying the evolution and structure of interaction networks and how species adapt to variable environments could helps us understand the fundamental workings of evolution and life on Earth, but there could be more tangible and applied consequences to our discoveries as well. For example, they could help us understand and predict whether species will be able to adapt to the rapid global change our planet is currently undergoing.

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Originally, I proposed to use a GRN, however, I found that there were only a few hundred interactions available in that network. Thus, I decided to include other kinds of interactions as well, which extended the number of interactions to 7.694.135.

In this study, my aim was to find mutations likely under selection related to environmental variability and answer the following questions:

- Are important mutations in genes with higher centrality?
- What genes have the most genetic variation? Do more central nodes have less variation as mutations can be more harmful there?
- Are nodes with higher expression levels less variable or under selection?

Purple sea urchins (Strongylocentrotus purpuratus) are excellent model organisms to address these questions as they have been frequently used for developmental studies for the past century and thus their GRNS and PPINs are well described. Furthermore, their life histories involve a highly mobile larval stage that allows dispersal across a highly heterogeneous landscape along the west coast of North America, thus populations with high gene flow can adapt to different local environmental conditions [20]. In addition, as global ocean temperatures and dissolved CO2 levels started to rise due to the current global climate change, weather stations have been placed throughout the habitat of this species that send publicly available up to date data.

2 Related Work

It has been theoretically predicted that interaction networks have a central role in evolution. A big set back to evolutionary change is the issue of pleiotropy. Pleiotropy is defined as the case when a single gene/protein is involved in more than one trait [10]. This can be problematic since changing the protein so that is would perform better for one of its functions can make it worse for its other function. Modularity in the network can help with this problem. If nodes whose interactions result in a specific trait are separated into a module, changing a node or an edge in this module won't influence nodes in another module responsible for another trait, making it easier for evolution to search the space of optimal protein structure [16]. Also, the structure of the network that evolves due to specific environmental conditions can preserve information about selection, "remembering" and "learning" that way [23].

There have been some studies that investigated the interplay between evolution and centrality. However, some of these came to contradictory conclusions and none of them (to the best of my knowledge) answers the specific questions I'd like to address. All of these studies used either only one population, or compared across species and not populations (higher divergence time), used only a subset of the available network, or sequenced only specific genes of interest. For example, while [11] found that proteins with higher centrality evolve faster, [8] found the exact opposite. The latter argued that proteins that have a more central position evolve slower because they are essential for survival and thus strong purifying selection is acting on them. This seems to be the more commonly accepted relationship as [1] and [6] also found that more central proteins were more conserved and less variable, but argued that they evolve slower not because of how important they are but because greater proportion of the central proteins are directly involved in their function (e.g. domains involved in various kinds of protein binding).

In order to computationally detect signatures of selection many approaches have been used. One of the common ways involves finding alleles whose frequency co-vary with environmental variables. The Bayesian method "Bayenv" was developed exactly for this purpose [7] and it has been used to identify mutations putatively under selection in Tibetan poplar [24], common starlings [21] and patula pine [19], just to mention a few example studies published this year.

In this study, we used purple sea urchin whole genome data and 4 environmental variables: temperature mean, temperature variability, pH mean and frequency of pH under 7.8 to find signatures of selection using the Bayenv method described above. These variables were chosen as they highly influence the fitness of animals of this species, and they differ significantly between the populations we investigated.

3 Methods

Detailed steps are available at https://github.com/Cpetak/WGS. Please refer to the README.md file for guidance.

3.1 Data Collection

20 urchins were collected from 7 populations along the west coast of North America. See Appendix A for specific coordinates. Sample collection locations were chosen based on environmental data availability and such that there would be significant temperature and pH differences between collection points. Urchins were shipped overnight to the University of Vermont where DNA was extracted using the QIAGEN DNeasy Blood and Tissue Kit. High quality DNA was sent to be sequenced on NovaSeq S2 Flow Cell 150 x 150 bp on a single lane, and library was prepped using Nextera DNA Flex Small Genomes Library Prep.

3.2 Processing sequence data

The quality of sequencing was assessed using FastQC [2]. Since the reads were of high confidence, all sequences were kept for further analysis. Sequence alignment files were generated using the Burrows-Wheeler Alignment Tool (BWA), and specifically the MEM algorithm [15]. To assess the quality of the mapping, a package called samtools was used. The mapping rate was sufficiently high, on average 80% of the reads mapped to the reference genome (Spur v.5.0). Next, the Analysis of Next Generation Sequence Data (ANGSD) program was run with 3 different filtering settings to calculate genotype likelihoods and SNPs [13]. Based on the number

SNPs and overall quality, a specific filtering was used for further processing. See Appendix B for specifics. As a final quality check step, a Principle Component Analysis was generated using PCAngsd [18] based on a subset of the SNPs data (~ 30%). Figure 1 shows that there were 3 individuals that were very different from the rest. This could be because of misidentification of the species during urchin collection, thus these 3 individuals were removed from further analysis.

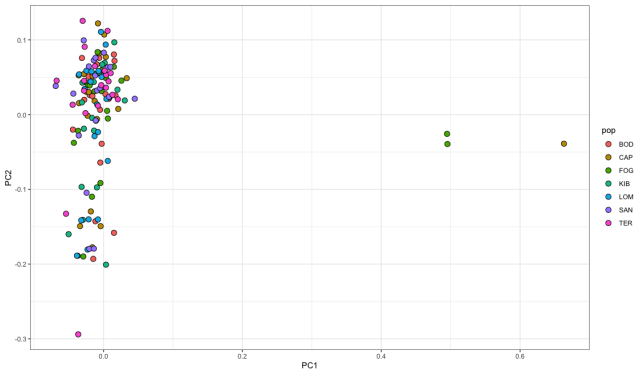


Figure 1. PCA of ~ 30% of the total SNPs identified by ANGSD. PC1: 1.25%, PC2: 0.78%. The 3 outliers were removed.

3.3 Processing environmental data

Temperature and pH data was gathered from the supplementary materials of [4], available at **here**. Relevant pages were saved as text files, converted into csvs and processed using R. For each environmental variable of interest data from all timepoints was aggregated and mean and standard deviation was calculated. Then, all variables were scaled by subtracting the mean and dividing by the standard deviation. Since I couldn't find comparable temperature and pH data for San Diego, this population was dropped from further analysis. See Appendix C for data used as an input to the Bayenv program.

3.4 Setting up and running the Bayenv program

The latest version of the Bayenv program was used to find alleles whose frequency co-vary with the above mentioned environmental variables [7]. This program needs the SNPs in a specific input format, so the code available in the GitHub repository was used to transform the ANGSD output into the correct input. I tried running Bayenv with all the 18.196.369 SNPs, however, this program was not meant to be used with this amount of data (see manual available here), so I decided to drop SNPs that had missing data for at least one individual and then randomly sampled the rest to get down to 61.249 SNPs. The output of Bayenv was saved as a text file and imported in a jupyter notebook for further analysis.

3.5 Feature engineering

In order to get to a dataframe I could analyse that included information about what the Bayenv factor [2] for each of the 4 environmental variable was, which gene the SNP belonged to, how many neighbors the gene had, what its expression level was and how many SNPs the gene had, I imported files and manipulated data using mostly numpy and pandas. Code available here. Brief steps taken:

1. Assembly annotation information was gathered from NCBI. The gff file was first cleaned such that only columns containing chromosome, position and gene name information were kept, rows containing other than genes were dropped and a single gene ID was extracted from a sting of alternative gene names.
2. A for loop checked if the position of any SNP fell between the start and stop position of a gene.
3. A dataframe was created with SNP position, gene name and Bayenv factor. If the SNP was not found to be in a gene it was still kept but had missing data for gene name and other columns related to gene property.
4. Downloaded the whole interaction network information from the String database, specifically for our species of interest. After converting the old gene names the edge list was encoded in to the new gene names the annotation file was using, for each gene the weighted number of neighbors was calculated. Weights were calculated based on the confidence score assigned by the String database for the given edge in the network.
5. Next, betweenness centrality [3] was calculated for each node using networkx.
6. Gene expression data was given to me by the authors of [4] and it is available upon request. For each gene the total number of mRNA was calculated (summed over treatment groups).
7. Finally, for each gene the number of SNPs was calculated.

The generated dataframe containing Bayenv factors, weighted number of neighbors, expression level and SNP number was saved to a csv that was then explored in another jupyter notebook.

3.6 Exploration and Modeling

All columns were log transformed and 0's were replaced with NAs to compute and visualise a pairwise correlation matrix of all the variables. Betweenness centrality had a perfect correlation with number of neighbors and a very strong correlation with weighted neighbors, thus only weighted neighbors was used as a variable to avoid redundancy. Pairwise relationships were also plotted for visual inspection. Stronger correlations were further visualised and R-squared

[2] Magnitude of correlation of allele frequency with environmental variable.
[3] Betweenness centrality of a node v is defined as the sum of the fraction of all-pairs shortest paths that pass through v.

values were calculated using linear regression from sklearn. For all modeling, NAs were dropped. Sklearn's default parameters were used, unless otherwise specified. Code available here.

3.6.1 Multiple linear regression

Sklearn's linear regression with validation size of 0.2 was used. However, since the R-squared value was low and the residual plot was far from ideal, thus further models were tested.

3.6.2 Classification using logistic regression

Based on the above and the Bayenv manual available here, I decided to make Bayenv factor a categorical, rather than a quantitative variable. Bayenv factors higher than the 90th percentile were categorised as "1" and lower were categorised as "0". To have the same number of data points in each category, I randomly undersampled the "0" category. The classifier (sklearn's Logistic Regression classifier) had to learn based on the number of neighbors, level of expression and SNP count whether or not the SNP would have a high Bayenv factor with regards to any of the environmental variables. All the variables were scaled to make them comparable. Validation size was 0.1.

3.6.3 Classification using K-nearest neighbours

Next, the same classifier was set up but this time using sklearn's K-nearest neighbours classifier. Validation size was 0.2. After some testing (see Figure 2), 14 was chosen as the number of neighbors the algorithm should consider.

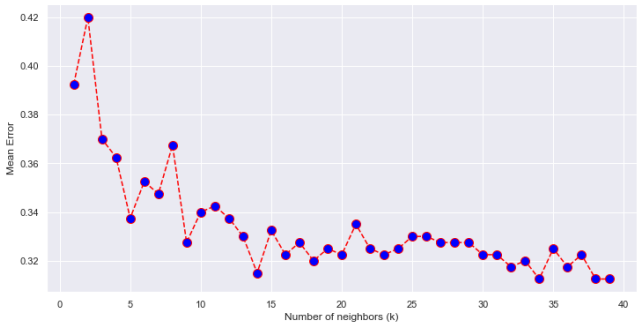


Figure 2. The K-nearest neighbours classifier was tested using values of k ranging from 1 to 40. 14 was chosen.

3.6.4 Support Vector Machine

Next, the same classifier was set up but this time using sklearn's SVM classifier. Validation size was 0.2. The classifier was tested with linear, poly (degree ranging from 1 to 5), RBF, and sigmoid kernels.

4 Results

Analysing the whole genome data, around 18 million positions were shown to be polymorphic (i.e. were SNPs). Based on a subset of 5 million a Principle Component Analysis was constructed, see Figure 1. As it is clearly shown in the figure, there was no clustering based on population. Out of the 18 million positions, 61,249 were tested for environmental correlations. Temperature mean and variability, and pH mean and frequency under 7.8 for each sampling location is shown in Figure 3.

	pH mean	pH freq
#1	gli3	RNA exonuclease 1
#2	noncoding	mucin-17
#3	noncoding	qgap4
#4	noncoding	noncoding
#5	noncoding	fibrillin-1

Table 1

	Temp mean	Temp var
#1	noncoding	mucin-17
#2	calcium-binding protein 1	RNA exonuclease 1
#3	RNA exonuclease 1	noncoding
#4	noncoding	noncoding
#5	uncharacterised protein	pgap4

Table 2

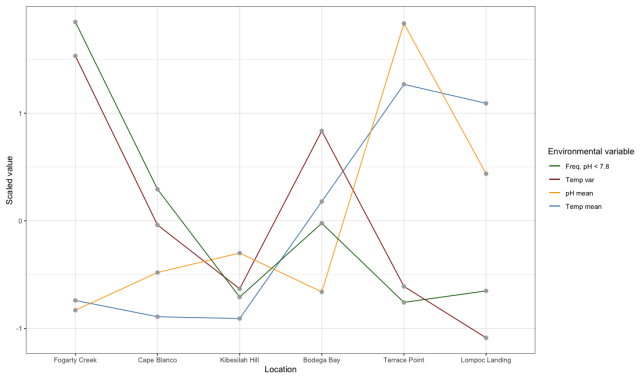


Figure 3. There was a strong negative or positive correlation between any two environmental variables (average correlation coefficient: 0.66 ± 0.15). Sampling locations are listed on the x-axis from North to South. pH and temperature variabilities were the highest in the North while means were highest in the South.

54% of SNPs fell within a gene, which is much higher than predicted by chance (~ 1% of the whole genome is protein coding). 33.01% of these genes were found in the interaction network downloaded from the String Database. For each environmental variable the 5 SNPs with the highest Bayenv factor were chosen and their location is shown in Tables 1 and 2.

Both the node degree distribution of the interaction network and the distribution of levels of gene expression showed power-low properties. See Appendix D.

There were strong positive correlations between some of the Bayenv factors of different environmental variables. There were weak positive correlations between Bayenv factors and SNP counts (average R-squared:0.1 ± 0.01, average slope:0.3 ± 0.02) and number of neighbors and expression level (R-squared: 0.23, slope: 0.2). See Figure 4.

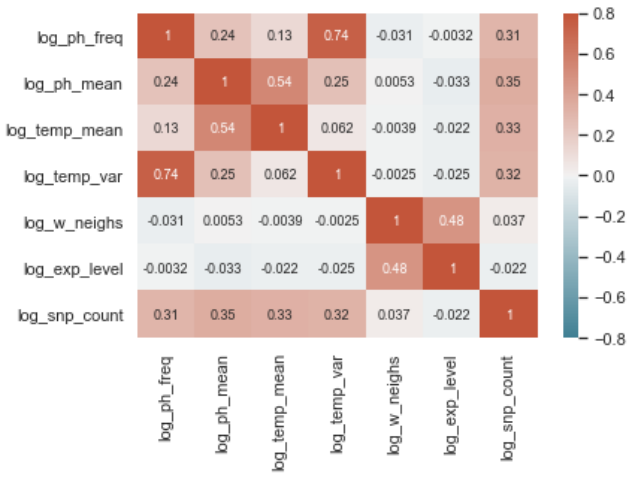


Figure 4. Pairwise correlation matrix of all log-transformed variables.

Multiple linear regression: It was found the SNP count had the highest model coefficient, however, the R-squared value for the model was low, 0.1. Also, the residual plot was far from ideal, see Figure 5. Perfect prediction would be indicated by a horizontal line of points at 0. These results were the same when predicting Bayenv factors of any of the 4 environmental variables.

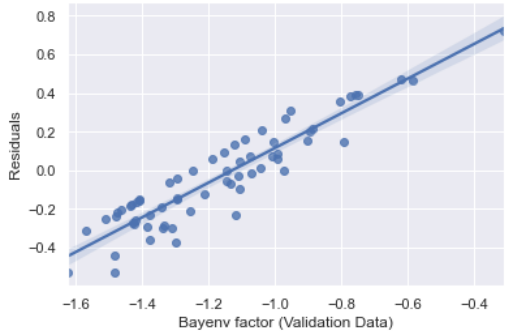


Figure 5. Residual plot of the multiple linear regression model.

Classification using logistic regression: Training accuracy: 0.683. Validation accuracy: 0.68. The highest model coefficient was SNP count again. The area under the ROC curve was 0.76.

Classification using K-nearest neighbours: Precision (tp / (tp + fp)): 0.67. Recall (tp / (tp + fn)): 0.665. The area under the ROC curve was 0.72.

Support Vector Machine:

RBF had the highest performance. Precision: 0.72. Recall: 0.71. The area under the ROC curve was 0.77. See Figure 6 for ROC curve.

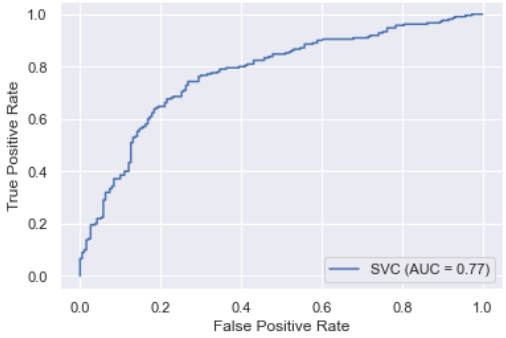


Figure 6. ROC (receiver operating characteristic curve) of the support vector machine classifier using the RBF kernel.

5 Discussion

In this study we were interested in finding mutations putatively under selection and see if these mutations occurred in genes central to the interaction network or had high levels of gene expression. We used the Bayenv method to find signatures of selection, however, for this method to work well there has to be high levels of gene flow between populations experiencing different environmental conditions, otherwise Bayenv would pick up differences between populations not necessary due to natural selection but rather drift independent populations experience. Figure 1 shows the lack of population structure in our sequencing data, indicating high levels of mixing of populations. This is due to the highly mobile life stage of the sea urchins - before developing into sessile adults they have a larval form in which they can actively move around but can also be carried with ocean currents for hundreds of kilometers. High levels of gene flow between populations of this species have been showed before with similar kinds of analysis [20].

For each environmental variable, the SNPs with the top 5 highest Bayenv factor was selected for further investigation (Tables 1 and 2). The gene which had an allele whose frequency correlated with pH mean the most was gli3, an important transcription factor. It regulates Hedgehog [9], which is involved in the establishment of left-right asymmetry during early sea urchin development [22]. There were several genes with allele(s) whose frequency highly correlated with the frequency of pH under 7.8; RNA exonuclease

1, mucin-17, qgap4 and fibrillin-1. RNA exonucleases generally cleave nucleotides one at a time from the end of an RNA molecule, however, this specific enzyme has not been studied in sea urchins. Mucin-17 is produced by epithelial tissues, and it has a high molecular weight and is heavily glycosylated. Their function is largely unknown in sea urchins [14]. Unfortunately, pgap4 is largely uncharacterised. Interestingly, exactly these 3 genes were the ones with the highest Bayenv factors with regards to temperature variability as well. Lastly, the most well known gene among the 4 mentioned above is fibrillin-1. Fibrillins have been hypothesized to have a structural role and are involved in biomineralisation [17]. Biomineralisation in sea urchins is known to be heavily influence by low pH [3]. Thus, looking into what SNP is in this specific gene could yield interesting discoveries. Finally, genes with the highest Bayenv factors regarding to temperature mean included the same RNA exonuclease mentioned above and a gene called SPARC-related modular calcium-binding protein 1. This gene is, again, not well studied in sea urchins but its homolog has been well characterised and based on those results it could be involved in biomineralisation as well [12]. However, further research is needed.

There were some strong correlations between Bayenv factors calculated for the different environmental variables. This was expected as the environmental variables themselves were highly correlated. E.g. since populations that experience higher temperature variability also experience an increased pH variability the allele frequency that correlates with one would also largely correlate with the other. Nevertheless, still, there were some alleles whose frequency correlated with one but not the other. Furthermore, interestingly even though temperature mean and pH mean, and temperature variability and frequency of pH under 7.8 had similar correlation coefficients, Bayenv factors regarding to the latter two were much more highly correlated. Also, as we have seen in the previous paragraph, the top 3 genes were the same for temperature variability and frequency of pH under 7.8. This could indicate that these are genes that are important in environmental variability in general. Future studies should investigate these genes more closely.

When investigating the relationship between Bayenv factors and expression level, number of neighbors and number of SNPs, only the third had a weak positive correlation. An extra line of evidence comes from the multiple linear regression and logistic regression classifier, where the coefficient was the largest for SNP count. This could indicate that often more than 1 mutation is needed for a fitness advantage, however, further investigation is needed.

One of the strongest correlations was found between level of expression and number of neighbors in the interaction network. This is not surprising given that strong correlations have been found between network centrality and protein essentiality [8], and level of expression and protein essentiality [5]. However, there was no linear relationship between centrality or level of expression and Bayenv factor. Acquiring a beneficial mutation in a central gene can result in a large fitness advantage, but at the same time a deleterious mutation might have lethal consequences. Thus, it is predicted that central genes evolve slower, with lower genetic variation to fuel evolution, and there has been some experimental evidence to this [1, 6, 8, 11]. Therefore, I predicted that most important mutations will be found in less central genes. Contrary to this expectation the correlation coefficient was found to be very low and intriguingly the top 2 genes with the highest Bayenv factors that had associated network and gene expression information both had a large number of neighbors and level of expression.

6 Future work

One of the strengths of this study was the availability of different types of large datasets. 18 million SNPs were identified from whole genome data, 30,000 genes had gene expression data available, and a network with 7 million edges was obtained. Still, the biggest limitation of this study was the lack of information, as only 2.4% of all SNPs found in genes had greater than 0 number of neighbors and level of expression. Thus, in a future study, further information about these metrics should be gathered.

Protein-protein interaction networks could have different properties from gene regulatory networks [11], thus separate analysis might be necessary in a future study. There are also additional methods to identify signatures of selection, like calculating Fixation index (FST) and reduced nucleotide diversity. These additional metrics could increase the probability that our Bayenv factors actually signify evolutionary important mutations.

Many high Bayenv factor SNPs fell outside of coding regions. However, those could still have important functional roles, e.g. they could be part of so called enhancer regions essential for gene regulation and normal development. This is a very interesting further direction to go into.

Finally, there could be some nonlinear interactions between Bayenv factor, centrality and level of expression, which was indicated by the high performance of the support vector machine using RBF kernel. I am planning to collaborate with experts to apply neural networks to this problem, followed by permutations tests to identify variables of interest.

7 Conclusion

While there is still a lot to be done, these preliminary investigations yielded some interesting results. Although there were no clear linear relationships between Bayenv factor and network centrality or level of expression, the latter two correlated with each other as they are both indicative of

protein essentiality. Moreover, Bayenv factors regarding different environmental variables had stronger than expected correlations and specific genes were identified for further investigation, including genes with potential regulatory role (gli3 transcription factor and RNA exonuclease) and role in biomineralisation (fibrillin and calcium-binding protein).

8 Acknowledgements

I would like to acknowledge my PhD advisor, Dr. Melissa Pespeni, and Dr. Reid Breannan, Mackenzie Kerner and Emily Shore for help in caring for the sea urchins and DNA extractions. Furthermore, I would like to thank people in the Raimondi Lab at the University of California, Santa Cruz and at the Menge and Gravem Lab at the Oregon State University for collecting and shipping live urchins.

References

[1] David Alvarez-Ponce. 2012. The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. *BMC evolutionary biology* 12, 1 (2012), 192.

[2] Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. 2012. FastQC. Babraham Institute.

[3] Maria Byrne and Susan Fitzer. 2019. The impact of environmental acidification on the microstructure and mechanical integrity of marine invertebrate skeletons. *Conservation physiology* 7, 1 (2019), coz062.

[4] Tyler G Evans, Melissa H Pespeni, Gretchen E Hofmann, Stephen R Palumbi, and Eric Sanford. 2017. Transcriptomic responses to seawater acidification among sea urchin populations inhabiting a natural pH mosaic. *Molecular Ecology* 26, 8 (2017), 2257–2275.

[5] Gang Fang, Karla D Passalacqua, Jason Hocking, Paula Montero Llopis, Mark Gerstein, Nicholas H Bergman, and Christine Jacobs-Wagner. 2013. Transcriptomic and phylogenetic analysis of a bacterial cell cycle reveals strong associations between gene co-expression and evolution. *BMC genomics* 14, 1 (2013), 450.

[6] Hunter B Fraser, Aaron E Hirsh, Lars M Steinmetz, Curt Scharfe, and Marcus W Feldman. 2002. Evolutionary rate in the protein interaction network. *Science* 296, 5568 (2002), 750–752.

[7] Torsten Günther and Graham Coop. 2013. Robust identification of local adaptation from allele frequencies. *Genetics* 195, 1 (2013), 205–220.

[8] Matthew W Hahn and Andrew D Kern. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution* 22, 4 (2005), 803–806.

[9] Sarah J Hatsell and Pamela Cowin. 2006. Gli3-mediated repression of Hedgehog targets is required for normal mammary development. *Development* 133, 18 (2006), 3661–3670.

[10] Jonathan Hodgkin. 2002. Seven types of pleiotropy. *International Journal of Developmental Biology* 42, 3 (2002), 501–505.

[11] Richard Jovelin and Patrick C Phillips. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome biology* 10, 4 (2009), R35.

[12] Anne Koehler, Sherwin Desser, Belinda Chang, Jacqueline MacDonald, Ulrich Tepass, and Maurice Ringuette. 2009. Molecular evolution of SPARC: absence of the acidic module and expression in the endoderm of the starlet sea anemone, Nematostella vectensis. *Development genes and evolution* 219, 9-10 (2009), 509–521.

[13] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. 2014. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics* 15, 1 (2014), 356.

[14] Tiange Lang, Gunnar C Hansson, and Tore Samuelsson. 2007. Gel-forming mucins appeared early in metazoan evolution. *Proceedings of the National Academy of Sciences* 104, 41 (2007), 16209–16214.

[15] Heng Li. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

[16] Dirk M Lorenz, Alice Jeng, and Michael W Deem. 2011. The emergence of modularity in biological systems. *Physics of life reviews* 8, 2 (2011), 129–160.

[17] Carmel McDougall, Ben J Woodcroft, and Bernard M Degnan. 2016. The widespread prevalence and functional significance of silk-like structural proteins in metazoan biological materials. *PloS one* 11, 7 (2016), e0159128.

[18] Jonas Meisner and Anders Albrechtsen. 2018. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* 210, 2 (2018), 719–731.

[19] Pablo Peláez, Alfredo Ortiz-Martínez, Laura Figueroa-Corona, José Rubén Montes, and David S Gernandt. 2020. Population structure, diversifying selection, and local adaptation in Pinus patula. *American Journal of Botany* (2020).

[20] Melissa H Pespeni, Thomas A Oliver, Mollie K Manier, and Stephen R Palumbi. 2010. Restriction Site Tiling Analysis: accurate discovery and quantitative genotyping of genome-wide polymorphisms using nucleotide arrays. *Genome biology* 11, 4 (2010), R44.

[21] Katarina C Stuart, Adam PA Cardilini, Phillip Cassey, Mark F Richardson, William B Sherwin, Lee A Rollins, and Craig DH Sherman. 2020. Signatures of selection in a recent invasion reveal adaptive divergence in a highly vagile invasive species. *Molecular Ecology* (2020).

[22] Jacob F Warner, Esther L Miranda, and David R McClay. 2016. Contribution of hedgehog signaling to the establishment of left–right asymmetry in the sea urchin. *Developmental biology* 411, 2 (2016), 314–324.

[23] Richard A Watson and Eörs Szathmáry. 2016. How can evolution learn? *Trends in ecology & evolution* 31, 2 (2016), 147–157.

[24] Chenfei Zheng, Lizhi Tan, Mengmeng Sang, Meixia Ye, and Rongling Wu. 2020. Genetic adaptation of Tibetan poplar (Populus szechuanica var. tibetica) to high altitudes on the Qinghai–Tibetan Plateau. *Ecology and evolution* 10, 20 (2020), 10974–10985.

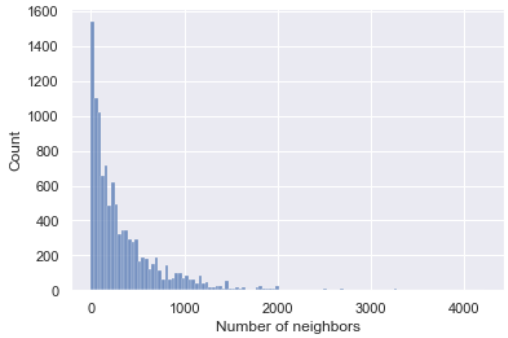
A Coordinates of urchin collection

Listed from North to South.

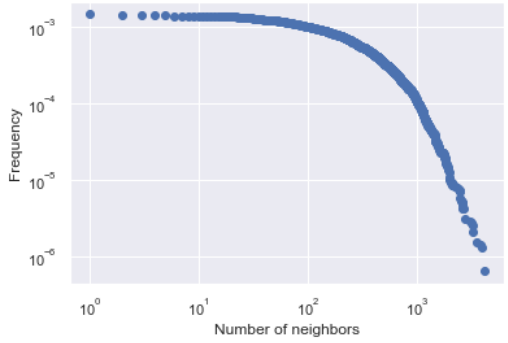
Location	Coordinates
Fogarty Creek	44.840000, -124.060000
Cape Blanco	42.840000, -124.570000
Kibesilah Hill	39.60412, -123.78887
Bodega Bay	38.3182, -123.07365
Terrace Point	36.94841, -122.06457
Lompoc Landing	34.718893, -120.609027
San Diego	32.66638889, -127.26138889

B ANGSD filtering parameters

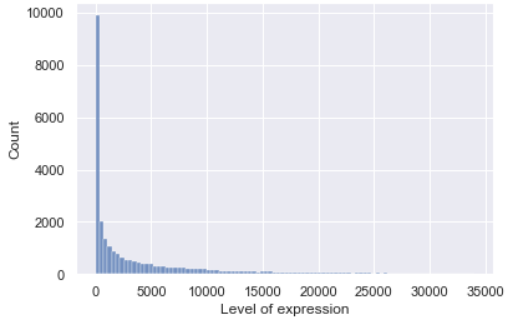
See Figure 7.



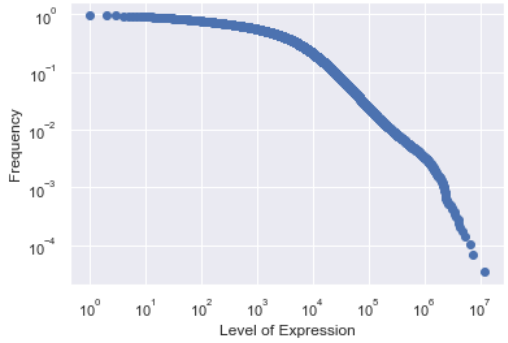
(a)



(b)



(c)



(d)

Figure 8

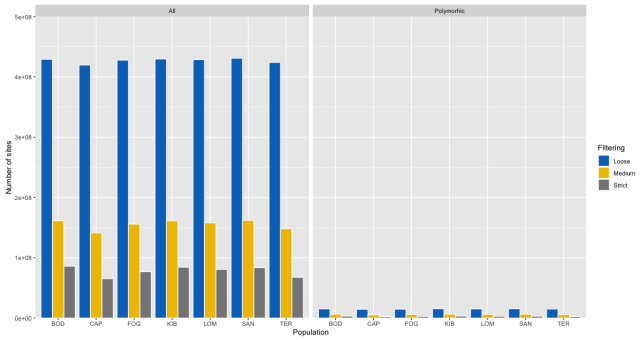


Figure 7. Number of all and polymorphic sites output by ANGSD genotype likelihood calculation with 3 different filtering. Loose: min 50% of individuals have to have data with min 2 coverage; Medium: min 85% of individuals have to have data with min 3 coverage; Strict: min 85% of individuals have to have data with min 4 coverage. For further analysis medium strictness was used.

C Environmental data

ph_freq	ph_mean	temp_mean	temp_var
-0.6512	0.4374	1.0915	-1.0872
-0.7580	1.8335	1.2685	-0.6095
-0.0227	-0.6603	0.1795	0.8351
-0.7086	-0.2993	-0.9083	-0.6316
0.2935	-0.4804	-0.8913	-0.0390
1.8472	-0.8307	-0.7399	1.5324

D Power-low distributions

See Figure 8.