A RESEARCH SUMMARY: 1) RNA BINDING PROTEINS, 2) *SELECTIVE CONSTRAINT ON COPY NUMBER VARIATION IN HUMAN PIWI-INTERACTING RNA LOCI*

By

DAVID WILLIAM GOULD

A thesis submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Computational Biology and Molecular Biophysics

written under the direction of

Dr. Kevin C. Chen

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2013

ABSTRACT OF THE THESIS

A RESEARCH SUMMARY: 1) RNA BINDING PROTEINS, 2) *SELECTIVE CONSTRAINT ON COPY NUMBER VARIATION IN HUMAN PIWI-INTERACTING RNA LOCI*

By DAVID WILLIAM GOULD

Thesis Director
Dr. Kevin Chen

The overall aim of my research dealt with the understanding of regulatory elements in various systems (most important, in humans) through two research projects 1) a study of RNA Binding Proteins in S. cerevisiae and 2) a study of piRNA in humans.

My first project involved the study of RNA Binding Proteins – thought to play a role in post-transcriptional translation in mammals. The algorithms miReduce and PhyloGibbs were used towards the prediction of binding sites for these proteins in S. cerevisiae. The putative binding sites found with the algorithms miReduce and PhyloGibbs warrant more extensive analysis, but further work needs to be done to determine the importance of secondary structure conservation inherent in many functional RNAs.

The second project examined the nature of piwi-interacting RNA (piRNA). piRNA are small noncoding RNA that are found in animals thought to act as regulatory elements in the germ-line. This study in particular considers possible forces of selection on piRNA through the analysis of their copy number variation in humans. Three human populations were included in the data used: Europeans, Yorubans, and Chinese/Japanese. Results from our methods support a hypothesis of negative selection on piRNA; they were presented in a publication co-authored by Dr. Kevin Chen and myself [11].

Table of Contents

II. List of Tables

III. List of Figures

I. Introduction – RNA Binding Proteins

RNA Binding proteins (RBPs) are thought to play a significant role in the gene interaction network of mammals in the form of post-transcriptional regulation [1]; as of 2008, relatively few RBPs had been systematically studied despite their hypothesized importance. In the summer of 2010 I began studying the prediction of RBP motifs and binding sites in S. cerevisiae. My project aimed to predict sequence specificity of RBPs, predict binding sites in the genome for RBPS, and analyze SNPs at the binding sites for evidence of correlation between sequence and gene expression variation.

II. RNA Binding Proteins

Prediction of Regulatory Elements

Towards the prediction of regulatory elements, [2] provides the algorithm miReduce, which bases the prediction on correlation between input sequences and input gene expression data. The algorithm returns statistically significant k-mers, where k-mers are sequences of fixed length k, with k determined by the user. In the particular case of sequence specificity for RBPs, 3' and 5' UTRs are hypothesized to be regions of interest [1]; we suggest that using 3' and 5' UTRs of S. cerevisiae [3] and RIP chip data [1] as the input for miReduce will predict RBP sequence specificity. miReduce outputs a list of statistically significant k-mers, sequences of fixed length k nucleotides, with k determined by the user. In this case we chose to search for 7-mers for the forty RBPs examined in the RIP chip data.

In order to validate the results of miReduce 7-mers were checked against motifs represented as Position Weight Matrices (PWMs) – matrices which contain the nucleotide distribution per site of a fixed length sequence. Two PWM sets, [1] and [4], base their sequence specificity prediction on computational methods distinct from miReduce. I scored 7-mers against each set of PWMs; in order to control for background noise, randomly selected 7-mers from the original UTRs were also scored against the PWM sets. Any cases in which miReduce suggested a significantly scoring 7-mer with respect to the background distribution was noted. The significantly scoring 7-mers yielded by miReduce for 3' UTRs had at least one such score for each RBP in a set of high confidence RBP sequences which [1] identified from the literature. We did not believe the 7-mers resulting from the 5' UTRs to be biologically relevant, and the

remainder of the analysis focuses on the 3' UTR predictions. As a final processing step to construct our own list of high confidence results, cases were noted when significantly scoring 7-mers from both [1] and [4] were identical for a given particular RBP.

To search for binding sites, we utilized a method from [5] which looked for evidence of selection on regulatory elements by comparing conserved elements from genome alignments of five yeast species. We downloaded the multiple alignment for these strains against S. cerevisiae from the UCSC and utilized an in house computer program to find all conserved 7-mers ranked by z-score [6]. Due to the low number of conserved instances in all alignments save for that between the two closest strains, only the global alignment between S. cerevisiae and S. paradoxus were included in the following analysis. I used the high-confidence list of 7-mers from the aforementioned processes and obtained conservation information for each predicted sequence. There were four RBPs which were significantly conserved across S. cerevisiae and S. paradoxus, with two of them being conserved in a large number of genes.

Remarks on miReduce

Though the results from a multiple linear regression scheme to predict RBP sequence specificity and binding sites has yielded compelling results, other methods of prediction may be superior for the prediction of RNA binding sites. matrixReduce [2] predicts the sequence specific binding affinity of a transcription factor in the form of a Position Specific Affinity Matrix (PSAM) and could easily be used for our analysis. Alternatively [8], using Drosophila as their organism of choice, develops probabilistic modeling upon which they build the algorithm Ahab. Ahab expands the work of previous

computational methods [9] in which the goal is to determine whether a sequence is more likely made by sampling from known weight matrices or a background distribution. Ahab in particular is an HMM constructed from PWMs of interest; in [8], they search the entire genome of Drosophila to find regulatory elements with specific PWMs for cis-regulatory modules – it was the first algorithm of its kind to have reasonable success in predicting regulatory elements in multi-cellular Eukaryotes.

Use of PhyloGibbs in RNA Binding Protein binding site prediction

At this point, the above analysis was redone except the algorithm PhyloGibbs was used instead of miReduce. PhyloGibbs uses a Bayesian Markov-chain Monte-carlo approach to discover those sequences which are likely to be transcription factor binding sites. The benefit is that this approach takes into account the phylogenetic relationship of the underlying species (whereas, with miReduce, we had to incorporate another filtering step in order to look for conservation across species). It was thought that looking for exact conservation - even though it was between S. cerevisiae and S. paradoxus, may have been too stringent a requirement, and resulted in putative RNA binding protein site 7-mers which were not enriched when compared to background 7-mers.

Methods

Using 3' UTRs and RIP chip data from S. cerevisiae we predict motifs for RBPs using two different approaches. The first involved the algorithm miReduce, which runs a multiple linear regression on S. cerevisiae UTRs and emits statistically significant k-mers (sequences of fixed length k). These k-mers were compared to motifs found in the literature, and a pairwise alignment of S. cerevisiae and a closely related species, S.

paradoxus, were searched for conserved instances. The second approach utilized a 7-way alignment of yeast UTRs and a phylogenetic tree as input to PhyloGibbs, which uses a Gibbs Sampler to search the space of putative regulatory motifs for binding sites which occur in UTRs. The putative binding sites from both approaches were searched for SNPs, and compared against a background distribution.

Results

From the list of 7-mers generated by miReduce, there were four which were significantly conserved across S. cerevisiae and S. paradoxus, with two of them being conserved in a large number of genes. The instances of these 7-mers in the 3'UTRs of S. cerevisiae and S. paradoxus were searched for SNPs and compared via a shuffling scheme against background 7-mers; the results of this method suggest that utilizing a multiple linear regression may not be the ideal way to identify putative binding sites for RBPs in S. cerevisiae.

The putative binding sites identified by PhyloGibbs have a SNP density which is lower than the background SNP density of the 3' UTRs; this suggests that a more careful examination of the PhyloGibbs motifs could identify multiple binding sites for RBPs.

Conclusion on RNA Binding Proteins

The putative binding sites produced by miReduce and PhyloGibbs are promising, but further work needs to be done to determine the importance of secondary structure conservation inherent in many functional RNAs. In addition to the algorithms used, there exists a variety of methods to predict binding specificity utilizing Stochastic Context Free Grammars (SCFGs) which take into account secondary structure. Implementations

include the program EvoFold, which takes a multiple alignment and phylogenetic tree as input to search conserved sequences for significantly scoring secondary structures.

Ultimately, I think that the project is worth pursuing (perhaps as a rotation project) - especially if you have someone who is capable of running and understanding the SCFG algorithms in the literature. A good paper to review is [10], which you originally sent me many months ago. Within, they mention secondary structure used to discover RNA binding protein target sites.

That said, the above paper [10] does a reasonable job of discovering putative RNA binding protein sites. Anything done would most likely only be complementing the work done in this paper. As such, it may not be worth the time and effort to pursue this project any further, given the amount of novel research you will get out of following it to completion.

For further information and data / code used, please see Dr. Kevin Chen.

III. Introduction -- piRNA and Copy Number Variation in Humans

The following data and the analysis thereof were used in the publication co-authored by Dr. Kevin Chen and myself, [11]. For any supporting information including computer scripts and datasets, please contact Dr. Kevin Chen.

piRNA are small noncoding RNA that are found in animals thought to act as regulatory elements in the germ-line. This study in particular considers possible forces of selection on piRNA through the analysis of their copy number variation in humans [12]. The data used, from Conrad et. al [13], contained copy numbers from three human populations: Europeans, Yorubans, and Chinese/Japanese (CEU, YRI, CHBJPT respectively).

IV. piRNA and Copy Number Variation in Humans

To begin the analysis I classified the CNV based on the elements they overlapped. For example, any coding region which (even partially) overlaps a CNV classifies it as a coding CNV. To the same end, any piRNA which overlaps a CNV classifies it as a piRNA CNV, etcetera.

Minor Allele Frequencies

Our first step involved examining the distribution of minor allele frequencies for the three different populations, separating CNV by the aforementioned classification scheme. It became evident that the information from the CHBJPT population contained too much noise to reliably analyze (see Figure 1). However, the CEU and YRI populations have distributions of minor allele frequencies which could indicate negative selection (Figure 2, 3; [11]).

McDonald-Kreitman test

What follows is a table containing the various numbers on unique/nonunique nucleotide counts for the various elements from the study. Please note: above, if any one base pair from a piRNA overlapped a CNV we classified the entire CNV as a piRNA overlap. In contrast for the Mcdonald-Kreitman test, my computer script counted overlaps on a base by base case. Note that there are 135,576,995 unique coordinates from the CNV dataset (not shown in Table 14).

Element	Total Basepairs	Unique Basepairs	Unique overlapping CNV
piRNA	764786	551879	66210
RNA Genes	942570	937441	75929
miRNA	63575	63451	3208
snoRNA/miRNA	107918	107202	9655
LINE	252646811	252342528	13556553

Table 1. *nucleotide counts for elements in study*

For our purposes, we have 92,414,015 intergenic coordinates – defined as any coordinate in a CNV which does not overlap one of the above functional elements. We back into this number by considering that there are 43,162,980 unique functional coordinates in all CNV: 135,576,995-43,162,980 = 92,414,015. With this information we conducted a McDonald-Kreitman test using CNV as our polymorphism data, and using estimated divergence. Roughly, the estimation of divergence data entailed sampling non-overlapping subsequences from 1) piRNA and 2) intergenic regions. These subsequences were aligned to 1) the human genome, and 2) the reference chimp genome; if a subsequence alignment count differed between humans and chimpanzees, the subsequence was considered diverged [11].

Matrix setup for the chi-squared test:

	[,1]	[,2]
[1,]	estimated piRNA bp diverged	piRNA overlapping CNV (Table 14)
[2,]	estimated intergenic bp diverged	intergenic bp overlapping CNV

The R output:

```
> matrix
        [,1]     [,2]
[1,]     41967    66210
[2,] 182486266 92414015
> chisq.test(matrix)

        Pearson's Chi-squared test with Yates' continuity correction

data:  matrix
X-squared = 36872.44, df = 1, p-value < 2.2e-16

> 182486266 / 92414015
D(int)/P(int) = 1.97466

> 41967/66210
D(piRNA)/P(piRNA) = 0.634
```

The chi-squared test shows that the two ratios are significantly different, which supports our hypothesis of negative selection for piRNA. This conclusion meshes with our interpretation above of the minor allele frequencies as being consistent with negative selection [11].

If we look at only those CNV which contain **no** functional elements, there are 41182769 polymorphic intergenic coordinates. The R output:

```
> matrix
        [,1]     [,2]
[1,]     41967    66210
[2,] 182486266 41182769
> chisq.test(matrix)

        Pearson's Chi-squared test with Yates' continuity correction
```

data: matrix
X-squared = 131688.2, df = 1, p-value < 2.2e-16

> 182486266/41182769
D(int)/P(int) = 4.431132

BLAT Analysis

To further strengthen our analysis, I next examined the derived allele frequencies for the various CNV by using BLAT [14] to align 400 basepair subsequences of the CNV to the chimp genome. Even using a score of 360 – allowing for 4 mismatches in an alignment, the vast majority (greater than ~85%) of the CNV have counts of 2.

The tables below show the number of CNV which were successfully aligned with the chimp genome using BLAT. These CNV were divided into gains (Table 2) and losses (Table 3) using the number of Chimp hits as the ancestral state. If the CN in humans was greater than that of the Chimp, it was considered a gain. Towards that end, if the CN in humans was less than that of the Chimp, it was considered a loss

Numbers represent CEU,CHBJPT,YRI sample sizes

BLAT Score Cutoff	RefSeq	Intergenic	Repeats	RNA Genes	piRNA
360	236, 93, 168	185, 90, 147	75, 25, 45	18, 14, 27	35, 24, 34
365	235, 93, 172	181, 89, 147	72, 24, 45	17, 13, 22	36, 24, 36
370	232, 93, 174	178, 86, 144	74, 26, 46	19, 13, 28	40, 26, 35
375	224, 88, 167	163, 75, 131	71, 23, 46	20, 12, 23	36, 21, 34
380	209, 88, 159	145, 71, 123	66, 24, 47	20, 11, 23	35, 21, 31
385	182, 80, 141	109, 55, 98	56, 23, 40	17, 8, 22	37, 20, 31
390	144, 62, 104	75, 41, 61	50, 17, 31	15, 6, 13	30, 13, 23

Table 2. *SAMPLE SIZES – GAINS*

Table 2. highlighted areas –

orange indicates the best p-values (From Table 4, 6)

yellow indicates the graph that we considered to have the most reasonable trade-off between sample size and lack of noise.

Numbers represent CEU,CHBJPT,YRI sample sizes

Numbers represent CEU,CHBJPT,YRI sample sizes

BLAT Score Cutoff	**RefSeq**	**Intergenic**	**Repeats**	**RNA Genes**	**piRNA**
360	639, 369, 901	185, 599 1575,	278, 169, 445	25, 11, 27	38, 22, 42
365	635, 368, 893	1013, 596, 1573	279, 167, 443	24, 11, 27	36, 19, 40
370	631, 369, 881	986, 578, 1548	271, 162, 434	25, 12, 28	33, 16, 35
375	620, 360, 869	957, 555, 1510	269, 157, 426	23, 11, 27	33, 17, 37
380	605, 355, 844	907, 524, 1415	255, 151, 400	26, 13, 28	31, 18, 33
385	520, 321, 732	758, 455, 1176	219, 134, 345	21, 11, 22	32, 20, 31
390	370, 220, 521	536, 327, 785	149, 94, 237	15, 8, 18	24, 16, 26

Table 3. *SAMPLE SIZES – LOSSES* (CEU,CHBJPT,YRI)

Table 3. highlighted areas –

orange indicates the best p-values (From Table 5)

yellow indicates the graph that we considered to have the most reasonable trade-off between sample size and lack of noise.

BLAT Score Cutoff	**CEU p-value** Kolmogorov-Smirnov, Wilcoxon	**YRI p-value** Kolmogorov-Smirnov, Wilcoxon	**CHBJPT p-value** Kolmogorov-Smirnov, Wilcoxon
375	0.02778, 0.01320	0.4484, 0.6658	0.3543, 0.2051
380	0.07201, 0.0302	0.4484, 0.6658	0.2809, 0.1811
385	0.05538, 0.02138	0.3672, 0.6295	0.1499, 0.2248

Table 4. *P-values – piRNA versus all repeats, gains only*

Table 4. tests the derived allele frequency distribution differences between piRNA and repeated regions in the human genome (Figure _). The orange highlight has both tests showing as significant at the alpha = 0.03 level, though one should note that the sample

size for the piRNA is a bit low for the CHBJPT population (22 piRNA CN classified CNV – Table 2). This may not matter, because the graphs for the CHBJPT population are odd, and it is doubtful we will find any significant p-value.

BLAT Score Cutoff	**CEU p-value** Kolmogorov-Smirnov, Wilcoxon	**YRI p-value** Kolmogorov-Smirnov, Wilcoxon	**CHBJPT p-value** Kolmogorov-Smirnov, Wilcoxon
375	0.2147, 0.07644	0.01574, 0.005635	0.3079, 0.3434
380	0.4699, 0.2729	0.01574, 0.005635	0.4983, 0.4766
385	0.4673, 0.3005	0.0652, 0.02127	0.3781, 0.1599

Table 5. *P-values – piRNA versus all repeats, losses only*

The above shows cases where at least one p-value (either Wilcoxon or KS or both) is significant in terms of *losses* (Figure _).

BLAT Score Cutoff	**CEU p-value** Wilcoxon	**YRI p-value** Wilcoxon	**CHBJPT p-value** Wilcoxon
375	3.31E-003	5.93E-001	3.98E-003
380	1.15E-002	5.93E-001	1.69E-003
385	2.50E-002	5.52E-001	2.32E-003

Table 6. *P-values – piRNA versus intergenic CNV, gains only*

I have the same output as above, using different BLAT parameters. However, there is a tradeoff for allowing more liberal matching criteria. While one could obtain more CNV matches, it might allow highly repetitive regions in the *Chimp* genome to be overrepresented. In general, this has the effect of adding tails to the end of the loss distributions.

Additional Loss Information

The following tables show more a one sided Wilcoxon test comparing piRNA CNV to repeat CNV (Table 7), and intergenic (Table 8).

Wilcoxon one sided test, H_a = piRNA < repeats

BLAT Score Cutoff	**CEU p-value**	**YRI p-value**	**CHBJPT p-value**
375	0.07644	0.005635	0.3434
380	0.2729	0.005287	0.4766
385	0.3005	0.02127	0.1599

Table 7. *P-values piRNA versus repeats – losses*

Wilcoxon one sided test, H_a = piRNA < intergenic

BLAT Score Cutoff	**CEU p-value**	**YRI p-value**	**CHBJPT p-value**
375	0.00E+000	0.0005253	0.1815
380	0.1010	0.0004742	0.3476
385	0.1080	0.005243	0.07698

Table 8. *P-values piRNA versus intergenic – losses*

These results show that the median difference between the two populations is not zero – in fact, the median for the piRNA CNV distribution is significantly lower from the intergenic CNV distribution. The results shown are for comparison with

Bootstrapping gains

We wanted to check our results using bootstrapping. In the below tables, the cutoff indicates the frequency threshold used in the bootstrapping – that is, a cutoff of 0.60 means that any CNV overlap-gain with frequency above 0.60 could be sampled. A

sample size equivalent to the total number of piRNA for a given population was taken from the Repeat distribution and the Intergenic distribution of frequencies. The p-value indicates the number of instances (out of 30,000) in which the Repeat/Intergenic samples had counts of high frequency gains which were greater than the number of high frequency gains for piRNA.

You'll notice that in all cases the Intergenic p-values are much better than those attained from the Wilcoxon/KS tests above. Additionally the p-values are significant for repeats in CEU and CHBJPT (at 0.60) and for YRI at a cutoff of 0.80.

An estimated cutoff of 0.80 or 0.85 is reasonable.

	CEU	**CEU**	**CHBJPT**	**CHBJPT**	**YRI**	**YRI**
Cutoff	*Repeats*	*Intergenic*	*Repeats*	*Intergenic*	*Repeats*	*Intergenic*
0.6	0.000633	0	0.0214	0.0001	0.1208	0.0375
0.7	0.00186	0.0	0.2513	0.0032	0.0741	0.0309
0.8	0.00596	0.0	0.0695333	0.00056	0.02026	0.0096
0.85	0.0139	0.000066	0.029466	0.00003	0.0907333	0.0445
0.9	0.0022	0.0	0.071533	0.000166	0.101733	0.03583
0.95	0.0029666	0.035833	0.274	0.00406	0.0577	0.01386

Table 9. *estimated P-values – bootstrapping (gains)*

Bootstrapping Losses

Bootstrapping for losses used the same scheme for the previous bootstrapping on gains. That is, 30,000 subsamples were taken at the various cutoffs, and the p-value is indicative of the number of times that there were **more** low frequency repeats/intergenic CNV than low frequency piRNA.

It appears that this results in better p-values for CEU, but for the YRI (even though at all cutoffs we see significant results) the results from the Wilcoxon test(s) are better.

	CEU	**CEU**	**CHBJPT**	**CHBJPT**	**YRI**	**YRI**
Cutoff	*Repeats*	*Intergenic*	*Repeats*	*Intergenic*	*Repeats*	*Intergenic*
0.40	0.177	0.0653	0.3378	0.2437	0.0233	0.016866
0.30	0.2165666	0.083933	0.38703	0.349366	0.01866	0.0120
0.20	0.028733	0.01013	0.14466	0.13276	0.042866	0.014233
0.15	0.0223	0.0072	0.1432	0.10903	0.01666	0.004866
0.10	0.031333	0.0202	0.383733	0.198233	0.02166	0.0038
0.05	0.079466	0.047133	0.172	0.0806	0.0477	0.0153

Table 10. *estimated P-values – bootstrapping (losses)*

V. Appendix

Figures

The following graphs were created through an in-house computer script coded by myself. They were subsequently used in [11].

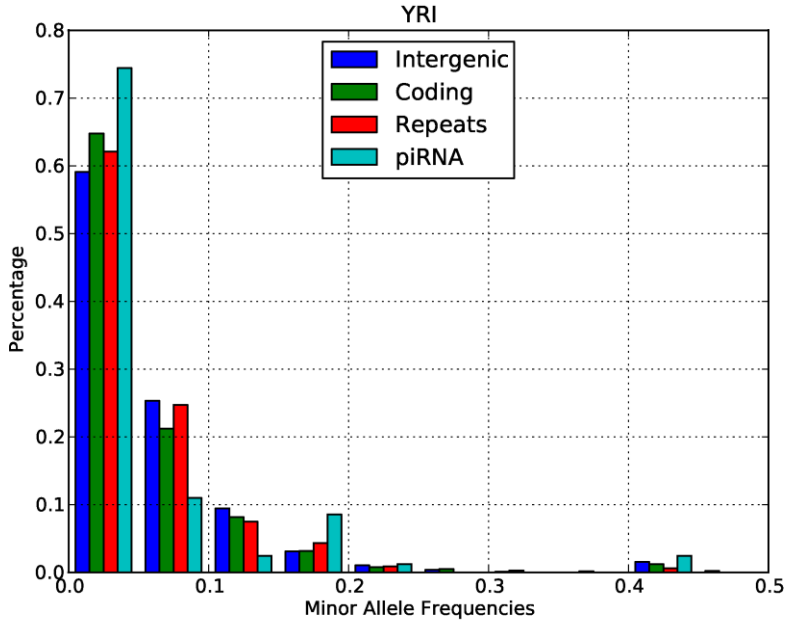

Figure 1

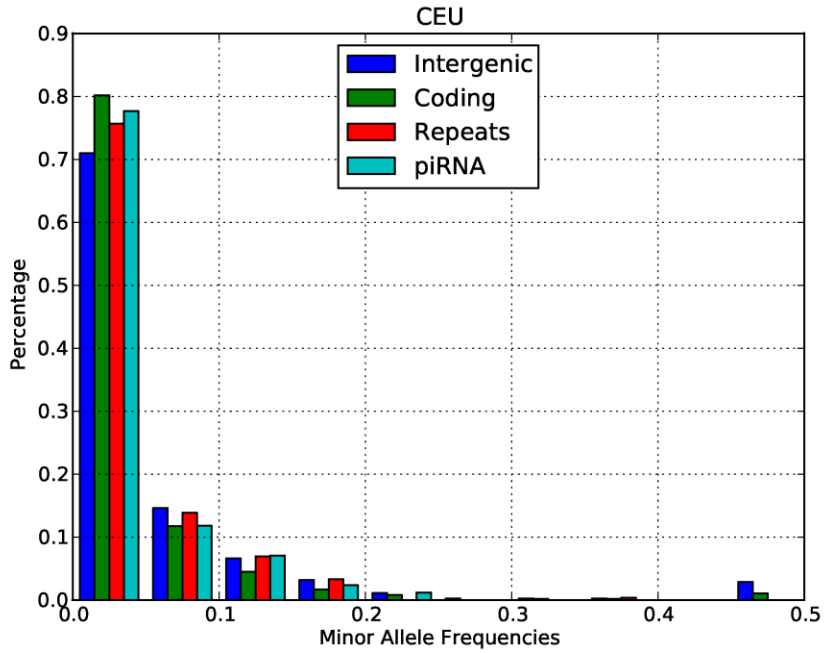

Figure 2

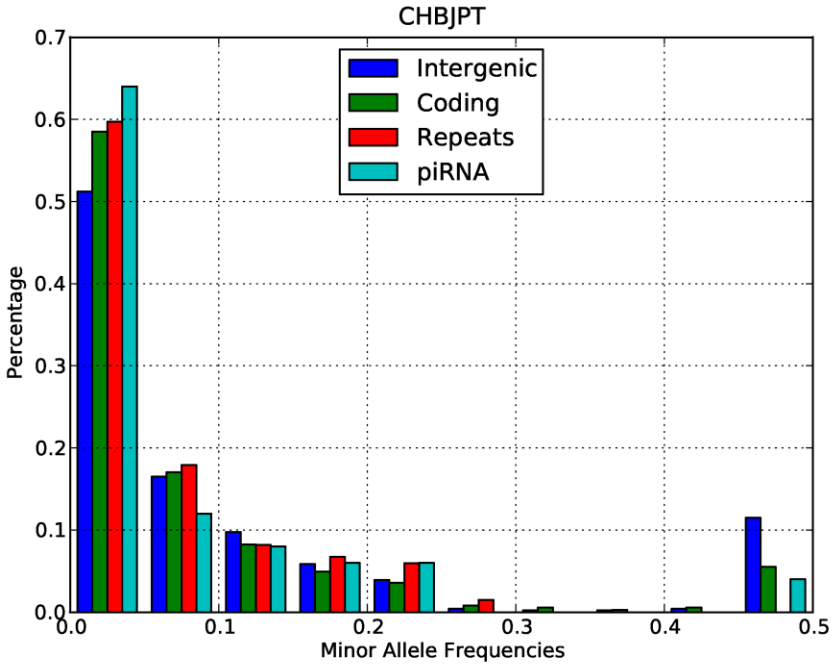

Figure 3

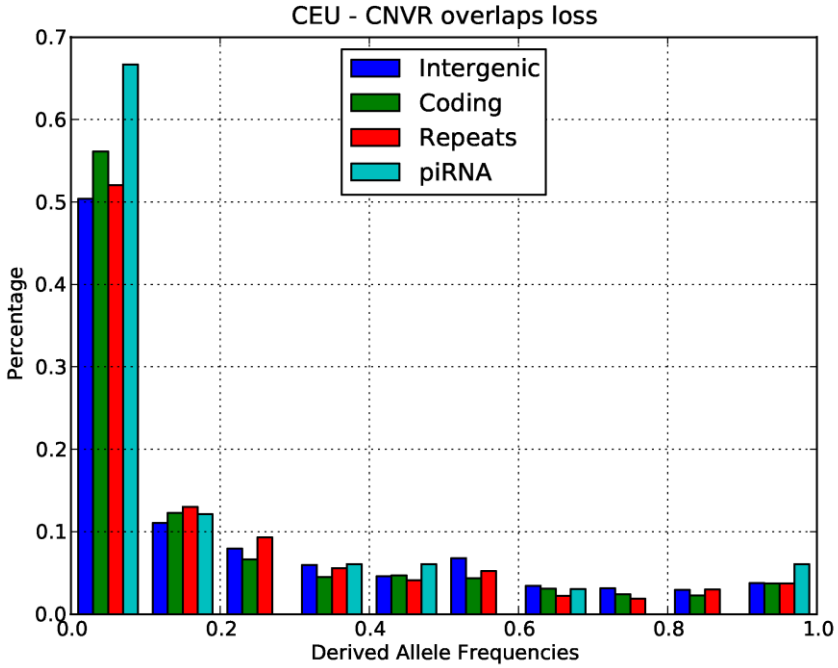

Figure 4

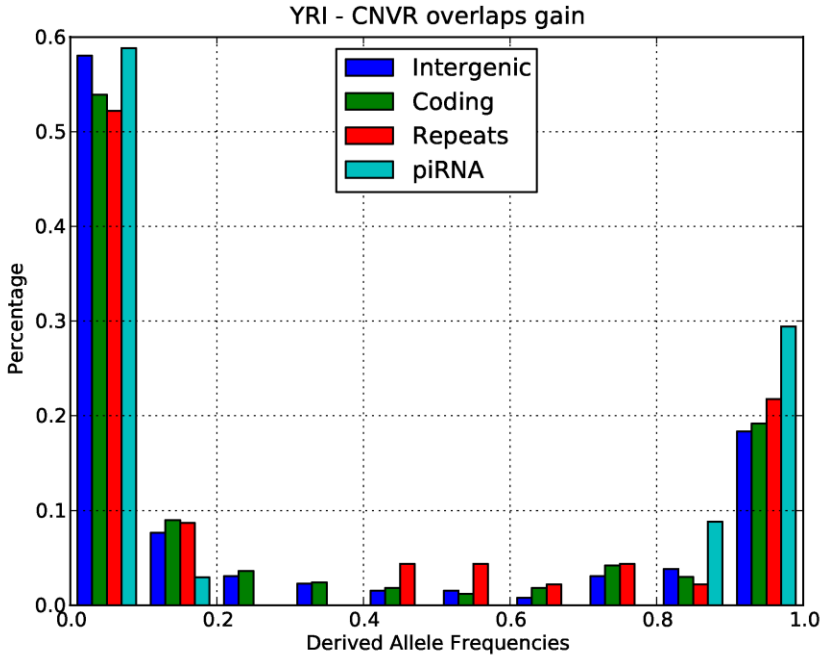

Figure 5

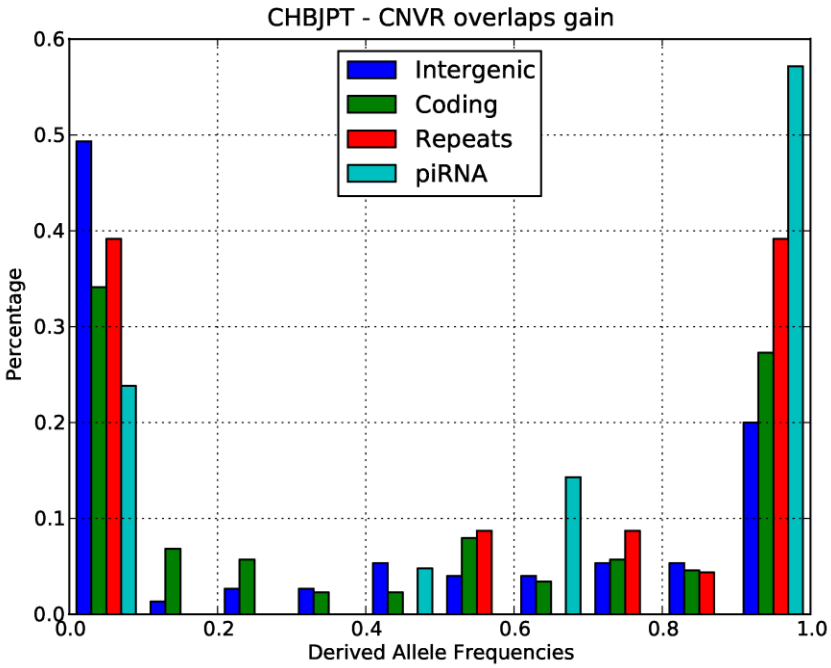

Figure 6

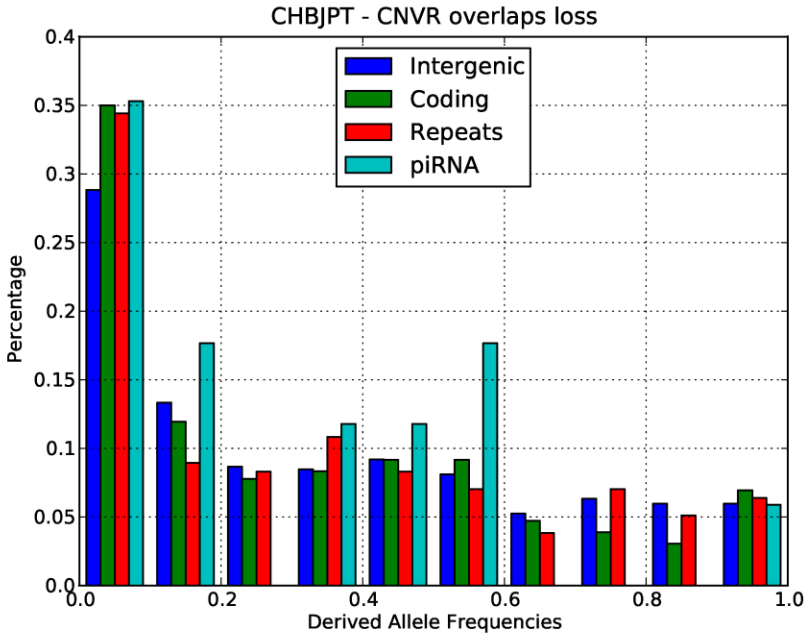

Figure 7

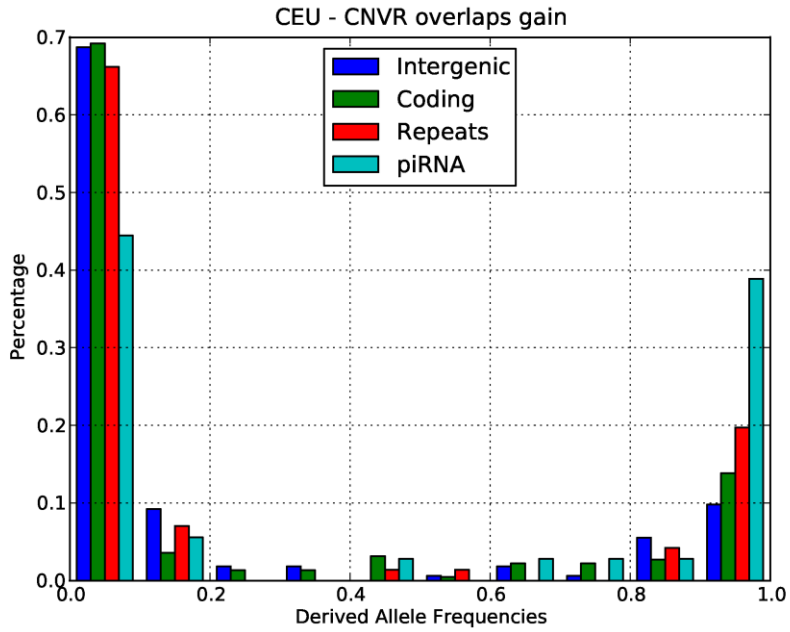

Figure 8

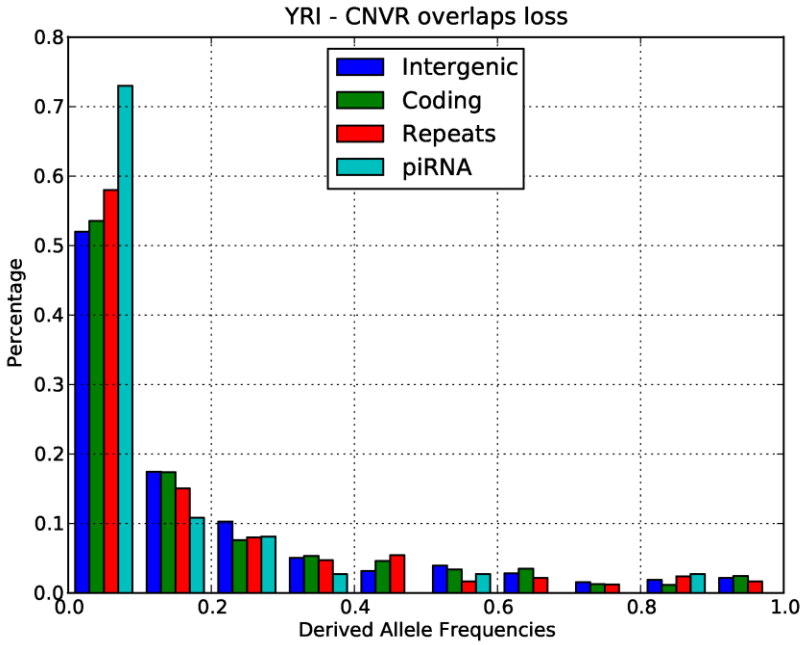

Figure 9

VI. Bibliography

[1] Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D. & Brown, P. O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol. 6, e255 (2008).

[2] Bussemaker, H.J. , Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. Nat. Genet. 27, 167–171. <http://bussemaker.bio.columbia.edu/software/REDUCE/>.

[3] Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science published online, doi:doi: 10.1126/science.1158441 (1 May 2008).

[4] Shalgi R, Lapidot M, Shamir R, Pilpel Y. 2005. A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs. Genome Biol. 6:R86

[5] Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.

[6] Chen K, Rajewsky N. Deep conservation of microRNA-target relationships and 3' UTR motifs in vertebrates, flies, and nematodes. Cold Spring Harb Symp Quant Biol 2006; 71:149–156.

[7] Liti G, Carter DM, Moses AM, Warringer J, Parts L, et al. (2009) Population genomics of domestic and wild yeasts. Nature 458: 337–341.

[8] Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics 3: 30.

[9] Bussemaker HJ, Li H, Siggia ED: Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc Natl Acad Sci USA 2000, 97:10096-100

[10] Li X, Quon G, Lipshitz HD, Morris Q: Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA 2010;16:1096-1107

[11] Gould, David W., Sergio Lukic, and Kevin C. Chen. "Selective Constraint on Copy Number Variation in Human Piwi-Interacting RNA Loci." *PloS one* 7.10 (2012): e46611.

[12] Brennecke J, Aravin A, Stark A, Dus M, Kellis M, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128: 1089–1103. doi: 10.1016/j.cell.2007.01.043.

[13] Conrad D, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704–712. doi:10.1038/nature08516.

[14] Kent, W. James. "BLAT—the BLAST-like alignment tool." *Genome research*12.4 (2002): 656-664.