Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Quran Wu Date

Identify chromatin state signature and its functional implications in nine cell lines

By

Quran Wu

Master of Science in Public Health

Biostatistics and Bioinformatics

_____

Zhaohui Qin, Ph.D

(Thesis Advisor)

_____

Tianwei Yu, Ph.D

(Reader)

Identify chromatin state signature and its functional implications in nine cell lines

By

Quran Wu

B.S.

Wuhan University

2015

Thesis Committee Chair: Zhaohui Qin, Ph.D

Reader: Tianwei Yu, Ph.D

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2017

Abstract:

The study provides a new way for including cis-regulatory elements into gene analysis. The genome is cut into small units and each unit is given a specific chromatin state according to its potential functions. The study analyzes the 15 chromatin states, which is defined by 9 histone marks, across 9 cell lines. The genome was cut into 200bp per unit and each unit was given a state according to their different responds to the histone marks. The study focuses on analyzing the chromatin states combinations and the combinations with length of 6 and length of 8 are emphasized. Among all the combinations, the combinations with repeated elements such as 11,7/7,11, 11,10/10,11 are relatively common. The study also uses GREAT system developed by Stanford University for gene set enrichment analysis aiming at regions of chromatin states combinations. The results show that combinations with repeated elements of 10,11/11,10 tends to have shared functions, while combinations with repeated elements of 11,7/7,11 and 13,8/8,13 tends to have unique functions. On the other hand, the shared functions for unique combinations within all cell lines are also emphasized. The functions related with ligase activity for combinations of repeated element of 10,11/11,10 appear in all cell lines with those combinations.

Identify chromatin state signature and its functional implications in nine cell lines

By

Quran Wu

B.S.

Wuhan University

2015

Thesis Committee Chair: Zhaohui Qin, Ph.D

Reader: Tianwei Yu, Ph.D

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2017

Introduction:

One of the major challenge in biology is to find the functions of the non-coding regions within the genome. The non-coding regions occupied 98% of the genome while the coding regions only take the place of 2% of the genome. It's obvious that we can't leave the non-coding regions as "junk regions" and take them as useless, therefore, ENODE (Encyclopedia of DNA Elements) project is launched by US National Human Genome Research Institute (NHGRI) for systematically analyzing those non-coding regions.

The ENCODE first choose 10% of the human genome for pilot study. They developed several techniques including CHIP-sequencing for detecting representative histone marks in order to define different chromatin states. Chromatin profiling provides a method that can find regulatory elements, given that chromatin has central toles in controlling DNA access and mediating regulatory signals(Barski et al., 2007). The functions of chromatin states include regulator binding, enhancer and repressor, transcriptional initiation and elongation.

The main purpose of this study is to analyze one of the chromatin states defining system developed by Jason Ernst in Manolis Kellis' Computational Biology Group. They used 9 histone marks that was found demonstrative in genome to define 15 chromatin states. The 9 marks are CTCF, H3K27me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac. As shown in Fig 1, the genome is cut into 200 bp per unit, and based on their different responds to the 9 marks, they were classified into 15 states(Ernst et al., 2011). The functions of those states are promotor, enhancer, transcriptional transition, elongation, etc. There are studies

found that the chromatin state distribution shows a highly dynamic landscape, with some specific difference across cell types showing a strong correlation between interacting functional elements(He et al., 2010; Johnson, Mortazavi, Myers, & Wold, 2007). The states with function of enhancer is used to locate target genes, predict activators and repressors, and find binding motifs that correlates with those regions. In this study, I consider to combine the states into 6 per group or 8 per group to see if those combination groups have any correlation with gene functions. If this correlation could be found, the new coding method for DNA will have practical meaning that could be applied to biological researches.

The chromatin states system gives a new way for coding human genome. Instead of using A,T,C,G, the 15 chromatin states could be applied for an alternative way of describing and constructing DNA sequence. Since the 15 states are functional based, this new way of DNA coding may give a straighter way of identifying functional regions.

Gene enrichment analysis is also a major part of the study. There are many tools for gene set enrichment analysis, among which GREAT developed by Stanford University was used in this study. GREAT performs well on analyzing cis-regulatory regions based on the measurement of binding regions across the genome. Unlike other gene set enrichment analysis tools that counts only the binding regions that are close to genes(Johnson et al., 2007; Khatri & Draghici, 2005; Mardis, 2007), GREAT can incorporate distal binding sites and use binomial test to control false positives(McLean et al., 2010). Other gene enrichment tools are gene based. Usually the inputs are a list of genes, and the tools will find the genes that are more common in the inputs

gene set than in the background genes. However, this is not accurate since by simply comparing gene sets with background genes, the genomic regions are not considered and hence there will be bias(Taher & Ovcharenko, 2009). In fact, the CRE region is much larger than gene regions, and the ignorance of them will result for loss of information(Lieberman-Aiden et al., 2009). The inputs for GREAT were the region of interest, and thus the CREs will be included in the gene enrichment analysis, which will give more accurate results.

Methodology:

1. Datasets:

   The chromatin states datasets are downloaded from UCSC ENCODE (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHmm), which contains information from 9 cell lines (GM12878, H1hesc, HEPG2, Hmec, Hsmm, HUVEC, K562, Nhek, NHLF). The datasets have around 600,000 observations and 15 variables including chromatin states, start point, end point, etc, for each cell line.

2. Choose frequent chromatin states combinations.

   The first goal of the study was to choose the most common chromatin states combinations within each cell line. The 15 chromatin states are coded as 1-15, for which the corresponding chromatin states functions are shown in Table 1. For each cell line, I pasted all the chromatin states into a long string, within which I cut them into different length of combinations and counted the number of occurrence for each combination. The length of combinations I've done are 2, 3, 4, 5, 6, 8, 10, where length of 6 (for example, 7,11,7,11,7,11) and length of 8 (for example, 7,11,7,11,7,11,7,11) are used for further analysis considered its biological significance and number of occurrence. For each combination, the number of occurrence (frequency) and the percentage divided by the total number of combinations (percent) were calculated. Top 10 combinations with highest frequency and percent were picked out for further analysis.

   For the top 10 combinations, the "significance of occurrence" were calculated as:

$$s = \frac{P_{combination}}{\prod_{i=1}^{n} P_i}$$

   Where n is the length of the combination, $P_{combination}$ is the percentage of the occurrence of a single combination within one cell line, and P_i is the percentage of a single state within one cell

line. The denominator calculated the percentage of the combination when all chromatin states distributed randomly across the cell line.

Then I calculated the sum of length of each individual kind of combination within each cell line. The length of a single combination was denoted as the position of end point minus the position of start point.

3. Gene Enrichment analysis.

3.1 Identify chromatin states combinations with common gene functions across cell lines.

The gene enrichment analysis was done by using Great Stanford (http://bejerano.stanford.edu/great/public/html/). The inputs are chromosome number, start point, and end point for each chromatin state combination, and the outputs are gene functions. All those information could be acquired from original datasets came from UCSC ENCODE. The gene enrichment analysis was only done to top 10 chromatin state combinations within 9 cell lines with combination lengths equal to 6 and 8.

To determine the chromatin states combinations with common gene functions, a scoring system used to calculate "match score" for identifying the combinations was built. The system was built as follow:

a. For each kind of combination in a single cell line, if one function for that combination could be found a duplicate in another cell line within the same combination, then it earned one score. For example, if combination 10,11,10,11,10,11 has 10 gene enrichment functions in GM12878, and 5 of them can be also found in the functions of

10,11,10,11,10,11 in H1hesc, then 5 points were earned for combination 10,11,10,11,10,11. Then the sum of those points for the single combination in the single cell line is defined as “match score”.

b. The “match score” had limitations that the number of functions within each combination will influence the result. For example, assuming 10,11,10,11,10,11 had 10 functions, where 4 of them had match in other cell lines, and 11,7,11,7,11,7 had 4 functions, where 3 of them had match in other cell lines. Even though 11,7,11,7,11,7 should be more common than 10,11,10,11,10,11, the match score will tell the opposite story. Therefore, “match percentage” was created, denoted as “match score” divided by “total score”, where total score was calculated as assuming all other cell lines with the specific combinations had the entirely same functions. For example, if 10,11,10,11,10,11 in GM12878 had 8 functions, and there were another 5 cell lines has the same combination, then the total score should be 8 times 5 equal to 40. Finally, the average of “match percentage” should be calculated for each kind of combination and it should reflect the commonality of each combination.

3.2 Identify common chromatin state combination functions within each chromatin states combinations across cell lines.

The common combination functions were identified by its frequency of occurrence across all cell lines. For example, for combination 10,11,10,11,10,11, the function “acid-amino acid ligase activity” could be found in 4 cell lines, then the frequency of occurrence is 4. Here we didn’t need to consider any “percentage” since all kinds of

combinations were considered separately. Finally, for each kind of combination, we could select combination functions that happened the most across all cell lines.

Results:

1. Choose frequent chromatin states combinations

The frequency and percentage of chromatin states combinations had been calculated for length as 2, 3, 4, 5, 6, 8, 10. Among them, the combinations that consisted of 7, 6 or 6, 7, consisted of 11, 7 and 7, 11 are most frequently appeared. The top 10 most frequent chromatin states combinations for length of 6 and 8 as well as their "significance of occurrence" were shown in Table 2. The result shows that although frequent combination were not all the same across 9 cell lines, they shared lots of similarities. The repeat of 11, 7 was the most common kind of element seen in the top 10 combinations, either occupied part of the combinations or constituted the whole combinations. For combinations with length of 2, though it's too short for further gene enrichment analysis, 11,7 and 7,11 were the top 2 combinations for all 9 cell lines. Meanwhile, 7, 5 and 7, 6 were also frequently appeared. The repeat of 11, 10 and 8, 13 could also be seen, but they were always followed by a long repeat of themselves (11,10,11,10,11,10,11,10, or 8,13,8,13,8,13,8,13). The "significance of occurrence" could be explained as "how many times the combinations occurred compared to all chromatin states distributed randomly across the cell line". The "significance of occurrence" were all larger than 500, sometimes even larger than 10,000,000, showing that those combinations were not accidently occurred. The combinations with length of 8 always had larger "significance of occurrence" compared to length of 6, even though they had less frequency. Note that here

combinations with smaller frequency may have larger “significance of occurrence”, which caused by different occurrence rate of single chromatin state.

The sum of the length for each kind of combination showed that the combination with repeated element of 13,8 and 8,13 had the most sum of length while combinations with element of 5,7 and 7,5 usually had the least sum of length. As described in the background, although each “scan window” was 200 bp long, when the same state appeared constantly, they will be counted as one state. Therefore, it seemed that state 8 and 13 tended to have longer length than others, while state 5 tended to have shorter length.

2. Gene Enrichment Analysis.

   a. Chromatin states combinations with common functions.

      The result of the “scoring system” is shown in table 3. The mean “match percentage” and how many cell lines has the specific combination were presented in table 3. From table 3, we could find that 11,10,11,10,11,10 and 10,11,10,11,10,11 were the combinations that has the most common functions (match percentage = 0.46 and 0.47). They also appeared in 5 cell lines. However, although 11,7,11,7,11,7 and 7,11,7,11,7,11 appeared in 6 cell lines, which is the largest number of appearance across 9 cell lines, the “match percentage” was only 0.033 and 0.037, showing that the functions of those two combinations are quite identical across all cell lines. Also, combinations with 5,7 (11,7,5,7,11,7 and 7,11,7,5,7,11) tended to have larger “match percentage”, even though this element was not very common in all combinations. Meanwhile, combination 13,12,13,12,13,12, 13,8,13,8,13,8,

7,11,7,11,7,5, 7,11,7,11,7,6, 8,13,8,13,8,13 appeared in more than 1 cell line, but none of their functions had a match in other cell line.

For combinations with length of 8, the result was consistent with length of 6. 10,11,10,11,10,11,10,11 and 11,10,11,10,11,10,11,10 had the highest mean "match percentage". However, 11,7,5,4,5,7,11,7, 13,8,13,8,13,8,13,8, 7,11,7,11,7,5,4,5 were all appeared in 7 cell lines, which is not consistent with length of 6 (11,7,11,7,11,7 and 7,11,7,11,7,11 appeared in the most cell lines). Therefore, it seemed that the 11,7 element wouldn't repeat for many times. Those combinations had low mean "match percentage", showing that their functions were quite identical. Also, combinations 11,7,11,7,5,4,5,7, 11,7,11,7,5,7,11,7, 12,13,12,13,12,13,12,13, 13,12,13,12,13,12,13,12, 5,4,5,7,11,7,11,7, 6,7,11,7,11,7,11,7, 7,11,7,11,7,11,7,6, 7,11,7,11,7,5,4,5, 7,11,7,11,7,5,7,11, 7,5,4,5,7,11,7,11, 8,13,8,13,8,13,8,13 had no match functions across cell lines even though they appeared in more than one cell lines. Those results were quite consistent with length of 6 since most of them contains part of the 6-length combinations.

b. Common chromatin states functions.

The result of common chromatin states combination functions was shown in table 4. For combinations with length of 6, most of the functions belonged to 10,11,10,11,10,11; 11,10,11,10,11,10; 7,11,7,5,7,11, and 11,7,5,7,11,7, which matched the result above. Among them, "acid-amino acid ligase activity", "ligase activity, forming carbon-nitrogen bonds", "small conjugating protein ligase activity", "ubiquitin-protein ligase activity", "centrosome", "microtubule organizing center", "mitotic cell cycle", "small conjugating

protein ligase activity", and "ubiquitin-protein ligase activity" were functions that could be found in 5 cell lines within combination 10,11,10,11,10,11 and combination 11,10,11,10,11,10. Given that both of the combinations appeared in 5 cell lines also, it means that whenever the two combinations appeared, they might perform those functions. The other two combinations, however, had no functions that could be seen in more than 3 cell lines, even though those two combinations also appeared in 5 cell lines.

For combination with length of 8, the functions that had most occurrence belonged to 10,11,10,11,10,11,10,11 and 11,10,11,10,11,10,11,10 also, which was consistent with length of 6. Among them, only "microtubule organizing center" for both combinations occurred in 5 cell lines, and "centrosome", "ligase activity", "microtubule cytoskeleton", "small conjugating protein ligase activity", "ubiquitin-protein ligase activity", and "mitotic cell cycle" occurred in 4 cell lines. Note that here all the functions except "microtubule organizing center" and "centrosome" happened only in combination 10,11,10,11,10,11,10,11, while those two functions happened in both two functions. This was different from length of 6 since in former both two combinations seemed to have same amount of common functions. I used to think the different form of repeated element (11,7,11,7,11,7 and 7,11,7,11,7,11, for example) should act equivalently, but here it seemed that combinations started from 10 became dominant. The other combinations had no functions that existed in more than 3 cell lines, which is the same as length of 6. However, combination 11,7,11,7,11,7,11,7 had two functions "regulation of cell junction assembly" and "regulation of focal adhesion assembly" that appeared in 3 cell lines while the

combination occurred only in 4 cell lines. Given that 11,7,11,7,11,7 had no such functions that was so common, those two functions were not consistent with length of 6 and might be used as an identification of combination 11,7,11,7,11,7,11,7.

Discussion:

As it was shown in the result, combinations with repeated element of 10,11 or 11,10 have most common gene enrichment analysis functions. From table 1, we know that state 10 is transcriptional elongation, state 11 is weak transcribed. The common functions are “acid-amino acid ligase activity”, “ligase activity, forming carbon-nitrogen bonds”, “small conjugating protein ligase activity”, etc., for both length of 6 and length of 8. It was obvious that most of the functions were related with ligase activity, which is essential for building proteins. It makes sense since state 10 is for elongation and state 11 is for transcribing, which is used in building and complete proteins, and thus ligase activity is necessary in those areas. In fact, it’s possible that ligase activity follows with state 10 only. Given that state 10 always comes with state 11, I can’t verify if it’s true.

Another interesting combination is the repeats of element 11,7 or 7,11. Given that state 11 is weak transcribed and state 7 is weak/poised enhancer, it’s not surprise that few common functions could be find. The enhancer is the regions that bind proteins to enhance the gene transcribed activities, and therefore they could be seen anywhere, which makes them have few common functions. It also explains why it always follows with the weak transcribed. However, functions “regulation of focal adhesion assembly” and “regulation of cell junction assembly” seems to be common among 11,7,11,7,11,7,11,7 combinations. Those functions are common only in length of 8 combinations

but not seen in length of 6 combinations. The reason for that is unclear.

Combinations with repeat elements of 8,13 or 13,8 should also gain some attention. The state 8 is insulator and state 13 is heterochromatin. The results show that they are longer than other states (with more repeats of the same state). Meanwhile, state 8 always comes with state 13 and state 13 always comes with state 8 also, showing that the insulator always follows with heterochromatin. Those combinations seldom have common functions across cell lines, showing that they may appear in any part of the genome.

Apart from the specific combinations described above, the fact that many combinations have repeated elements which were always constituted by two states is also interesting. The most common repeated elements are 11,7/7,11, 11,10/10,11, 8,13/13,8. The repeated elements showing that the same kind of functional regions tends to be constant and continuous. It may have another explanation that the repeated element should be seen as a single state, and the repeating regions are just a single functioning region.

The study also has some limitations. First of all, the classification of 15 states is simply determined by the different respond of 9 cell marks. Whether this kind of classification is representative enough is questionable. The 15 states are mostly promotors, enhancers, transition, and elongation. Although they filled the whole genome, they may simply represent part of the functions of their region, while some non-coding regions might be misrepresented. There are other studies which use 51 chromatin states instead of 15. Therefore, how to construct an overall applied chromatin state system is

essential. Also, the rules for constructing the chromatin states system should also be considered. In this study, the 15 states are constructed by 9 cell marks, but there may be some more appropriate ways to construct the chromatin states.

Acknowledgement:

I would like to thank the Department of Biostatistics and Bioinformatics in RSPH of Emory university for the guidance in the past two years. I would also especially to thank Dr. Zhaohui Qin for being my thesis advisor and Dr. Hongde Liu for helping me on my thesis. I would also like to thank Dr. Tianwei Yu for being my thesis reader and give me advices on thesis writings.

References:

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., . . . Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell, 129*(4), 823-837. doi:10.1016/j.cell.2007.05.009

Consortium, E. P., Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., . . . de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature, 447*(7146), 799-816. doi:10.1038/nature05874

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., . . . Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature, 473*(7345), 43-49. doi:10.1038/nature09906

He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., . . . Liu, X. S. (2010). Nucleosome dynamics define transcriptional enhancers. *Nat Genet, 42*(4), 343-347. doi:10.1038/ng.545

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502. doi:10.1126/science.1141319

Khatri, P., & Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics, 21*(18), 3587-3595. doi:10.1093/bioinformatics/bti565

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326*(5950), 289-293. doi:10.1126/science.1181369

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nat Methods, 4*(8), 613-614. doi:10.1038/nmeth0807-613

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., . . . Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol, 28*(5), 495-501. doi:10.1038/nbt.1630

Taher, L., & Ovcharenko, I. (2009). Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics, 25*(5), 578-584. doi:10.1093/bioinformatics/btp043

Appendix:

Fig 1: The chromatin states on genome:

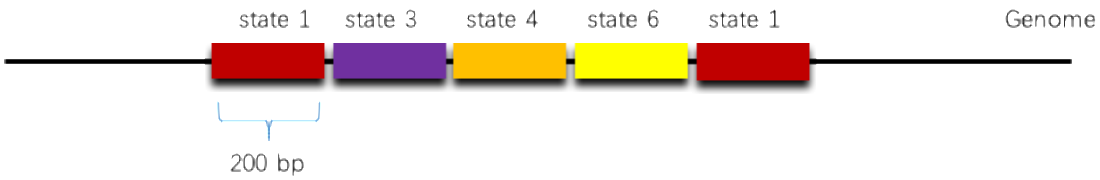

Table 1: 15 chromatin states and their corresponding functions.

States	Functions
State1	Active promoter
State2	Weak Promoter
State3	Inactive/poised promoter
State4	Strong enhancer
State5	Strong enhancer
State6	Weak/poised enhancer
State7	Weak/poised enhancer
State8	Insulator
State9	Transcriptional transition
State10	Transcriptional elongation
State11	Weak transcribed
State12	Polycomb-repressed
State13	Heterochromatin; low signal
State14	Repetitive/Copy Number Variation
State15	Repetitive/Copy Number Variation

Table 2: Summary of chromatin states combination frequency for 9 cell lines:

GM12878

Combinations	Mean frequency	Significance of occurrence
11,7,11,7,11,7	20.48	2495
7,11,7,11,7,11	20.17	2458
11,10,11,10,11,10	14.59	125201
10,11,10,11,10,11	14.54	124766
7,11,7,5,4,5	14.10	33547
5,4,5,7,11,7	13.81	32858
7,11,7,11,7,6	13.34	1936
7,5,4,5,7,11	12.95	30806
6,7,11,7,11,7	12.73	1847
11,7,5,4,5,7	12.61	29996

Combinations	Mean frequency	Significance of occurrence
11,10,11,10,11,10,11,10	4.72	13029818
10,11,10,11,10,11,10,11	4.71	12999830
11,7,11,7,11,7,11,7	4.29	40684
7,11,7,11,7,11,7,11	4.28	40632
7,11,7,5,4,5,7,11	3.50	648450
11,7,5,4,5,7,11,7	3.42	633346
7,11,7,11,7,5,4,5	3.34	618242
13,8,13,8,13,8,13,8	3.20	5127840
8,13,8,13,8,13,8,13	3.20	12921
5,4,5,7,11,7,11,7	3.08	12752

H1hesc

Combinations	Mean frequency	Significance of occurrence
11,7,11,7,11,7	53.97	1040
7,11,7,11,7,11	53.91	1039
8,13,8,13,8,13	24.72	11982
6,7,11,7,11,7	24.61	661
13,8,13,8,13,8	24.33	11789
7,11,7,11,7,6	24.17	650
11,7,11,7,6,7	20.67	556
7,6,7,11,7,11	20.15	542
7,11,7,6,7,11	19.36	520
11,7,6,7,11,7	19.22	517

Combinations	Mean frequency	Significance of occurrence
11,10,11,10,11,10,11,10	4.72	12851
10,11,10,11,10,11,10,11	4.71	12779
11,7,11,7,11,7,11,7	4.29	7910
7,11,7,11,7,11,7,11	4.28	7613
7,11,7,5,4,5,7,11	3.50	398232
11,7,5,4,5,7,11,7	3.42	394560
7,11,7,11,7,5,4,5	3.34	6691
13,8,13,8,13,8,13,8	3.20	6691
8,13,8,13,8,13,8,13	3.20	6596
5,4,5,7,11,7,11,7	3.08	6527

HEPG2

Combinations	Mean frequency	Significance of occurrence
10,11,10,11,10,11	19.57	218718
11,10,11,10,11,10	19.51	218071
13,12,13,12,13,12	12.20	82391
12,13,12,13,12,13	12.17	82244
11,7,11,7,11,7	12.02	5446
7,11,7,11,7,11	11.95	5414
5,4,5,7,11,7	11.38	187531
11,7,5,4,5,7	11.07	182398
7,11,7,5,4,5	10.91	179653
7,5,4,5,7,11	10.91	179653

Combinations	Mean frequency	Significance of occurrence
11,10,11,10,11,10,11,10	6.98	27003982
10,11,10,11,10,11,10,11	6.85	26499628
13,12,13,12,13,12,13,12	4.01	7927921
12,13,12,13,12,13,12,13	3.98	7863466
11,7,5,4,5,7,11,7	3.40	6651315
7,11,7,5,4,5,7,11	3.30	6459757
5,4,5,7,11,7,11,7	2.61	5118852
11,7,11,7,5,4,5,7	2.59	5065642
13,8,13,8,13,8,13,8	2.58	4902181
7,11,7,11,7,5,4,5	3.08	6527

HMEC

Combinations	Mean frequency	Significance of occurrence
11,7,5,7,11,7	29.99	1491
7,11,7,5,7,11	29.81	1482
5,7,11,7,11,7	21.81	1085
7,11,7,11,7,5	21.34	1061
7,5,7,11,7,5	21.00	1947
5,7,11,7,5,7	20.70	1919
7,11,7,11,7,11	20.66	551
11,7,11,7,11,7	19.95	532
7,5,7,11,7,11	19.19	954
7,5,4,5,7,11	18.95	13339

Combinations	Mean frequency	Significance of occurrence
5,7,11,7,5,7,11,7	6.45	28679
7,11,7,5,7,11,7,5	6.39	28389
7,11,7,5,4,5,7,11	6.26	211336
11,7,5,4,5,7,11,7	6.20	209318
7,11,7,5,7,11,7,11	5.75	13711
11,7,5,7,11,7,5,7	5.51	24499
7,5,7,11,7,5,7,11	5.51	24499
11,7,11,7,5,7,11,7	5.46	13024
11,7,5,7,11,7,11,7	5.42	12921
7,11,7,11,7,5,7,11	5.35	12752

Hsmm

Combinations	Mean frequency	Significance of occurrence
7,5,4,5,7,11	21.01	14935
11,7,5,4,5,7	20.41	14513
5,4,5,7,11,7	18.63	13246
11,10,11,10,11,10	18.07	28141
7,11,7,5,4,5	18.01	12808
10,11,10,11,10,11	17.92	27915
7,11,7,5,7,11	14.86	1244
11,7,5,7,11,7	14.44	1209
5,7,11,7,11,7	11.82	989
7,11,7,11,7,5	11.59	971

Combinations	Mean frequency	Significance of occurrence
10,11,10,11,10,11,10,11	5.98	1761962
11,10,11,10,11,10,11,10	5.98	1760362
11,7,5,4,5,7,11,7	5.68	234137
7,11,7,5,4,5,7,11	5.64	232345
7,5,4,5,7,11,7,5	3.59	275709
5,7,11,7,5,4,5,7	3.42	262362
7,5,4,5,7,11,7,11	3.30	136225
8,13,8,13,8,13,8,13	3.26	2066909
13,8,13,8,13,8,13,8	3.13	1984094
11,7,11,7,5,4,5,7	3.09	127487

HUVEC

Combinations	Mean frequency	Significance of occurrence
7,5,4,5,7,11	28.28	40233
11,7,5,4,5,7	28.22	40150
5,4,5,7,11,7	26.04	37047
7,11,7,5,4,5	25.47	36243
13,12,13,12,13,12	21.70	86909
12,13,12,13,12,13	21.66	86735
7,11,7,5,7,11	15.54	3800
11,7,5,7,11,7	15.23	3725
13,7,5,4,5,7	12.96	17853
5,7,11,7,11,7	12.86	3146

Combinations	Mean frequency	Significance of occurrence
11,7,5,4,5,7,11,7	8.21	989088
7,11,7,5,4,5,7,11	8.09	974687
12,13,12,13,12,13,12,13	6.64	6545270
13,12,13,12,13,12,13,12	6.54	6454140
5,7,11,7,5,4,5,7	5.92	1072673
7,5,4,5,7,11,7,5	5.85	1059867
7,5,4,5,7,11,7,11	4.59	552475
11,7,11,7,5,4,5,7	4.46	536765
5,4,5,7,11,7,11,7	4.36	525637
7,11,7,11,7,5,4,5	4.16	500763

K562

Combinations	Mean frequency	Significance of occurrence
7,11,7,11,7,11	50.78	2255
11,7,11,7,11,7	50.67	2250
6,7,11,7,11,7	18.91	1435
7,11,7,11,7,6	18.90	1434
5,7,11,7,11,7	18.83	2158
7,11,7,11,7,5	18.49	2119
12,13,12,13,12,13	18.15	52602
13,12,13,12,13,12	17.98	52098
5,4,5,7,11,7	15.30	18317
7,11,7,5,4,5	15.12	18100

Combinations	Mean frequency	Significance of occurrence
7,11,7,11,7,11,7,11	14.43	36691
11,7,11,7,11,7,11,7	14.35	36484
13,12,13,12,13,12,13,12	5.47	3650882
12,13,12,13,12,13,12,13	5.42	3618220
7,11,7,11,7,11,7,5	5.24	34368
6,7,11,7,11,7,11,7	5.22	22664
5,7,11,7,11,7,11,7	5.18	33976
7,11,7,11,7,11,7,6	5.03	21861
13,8,13,8,13,8,13,8	4.78	2551629
8,13,8,13,8,13,8,13	4.78	2551629

Nhek:

Combinations	Mean frequency	Significance of occurrence
7,5,4,5,7,11	26.20	15364
11,7,5,4,5,7	25.54	14978
7,11,7,5,4,5	24.13	14149
5,4,5,7,11,7	24.10	14132
10,11,10,11,10,11	17.61	92714
11,10,11,10,11,10	17.22	90692
8,13,8,13,8,13	16.08	22560
13,8,13,8,13,8	15.68	22001
7,11,7,5,7,11	15.10	1305
11,7,5,7,11,7	14.67	1267

Combinations	Mean frequency	Significance of occurrence
7,11,7,5,4,5,7,11	7.51	268669
11,7,5,4,5,7,11,7	7.34	262638
10,11,10,11,10,11,10,11	5.67	8427327
11,10,11,10,11,10,11,10	5.63	8362749
5,7,11,7,5,4,5,7	5.40	312801
7,5,4,5,7,11,7,5	5.26	304611
8,13,8,13,8,13,8,13	4.63	1179284
13,8,13,8,13,8,13,8	4.57	1162674
7,5,4,5,7,11,7,11	3.86	138128
11,7,11,7,5,4,5,7	3.68	131903

NHLF:

Combinations	Mean frequency	Significance of occurrence
13,8,13,8,13,8	30.52	8274
8,13,8,13,8,13	30.22	8193
7,11,7,11,7,11	18.99	921
11,7,11,7,11,7	18.97	920
7,11,7,5,7,11	16.36	1843
11,7,5,7,11,7	15.98	1800
10,11,10,11,10,11	14.59	55426
11,10,11,10,11,10	14.57	55371
11,7,11,7,5,7	13.59	1531
7,5,7,11,7,11	13.09	1475

Combinations	Mean frequency	Significance of occurrence
8,13,8,13,8,13,8,13	9.05	259147
13,8,13,8,13,8,13,8	9.02	258213
11,10,11,10,11,10,11,10	4.39	4250298
10,11,10,11,10,11,10,11	4.38	4239777
7,11,7,11,7,11,7,11	3.73	10784
11,7,11,7,11,7,11,7	3.68	10627
11,7,11,7,5,7,11,7	3.23	21656
7,11,7,11,7,5,7,11	3.18	21365
11,7,5,4,5,7,11,7	3.03	347274
7,11,7,5,4,5,7,11	3.01	344780

Table 3. Combinations with common functions across cell lines.

Combinations with length of 6

Combination	Mean match percentage	Occurrence rate
11,10,11,10,11,10	0.479	5
10,11,10,11,10,11	0.456	5
7,5,7,11,7,11	0.210	2
11,7,5,7,11,7	0.192	5
7,11,7,5,7,11	0.169	5
5,7,11,7,11,7	0.105	4
6,7,11,7,11,7	0.079	3
5,4,5,7,11,7	0.073	6
11,7,5,4,5,7	0.068	5
7,11,7,5,4,5	0.063	6
7,11,7,11,7,11	0.037	6
12,13,12,13,12,13	0.036	3
7,5,4,5,7,11	0.036	6
11,7,11,7,11,7	0.033	6
13,12,13,12,13,12	0	3
13,8,13,8,13,8	0	3
7,11,7,11,7,5	0	3
7,11,7,11,7,6	0	3
8,13,8,13,8,13	0	3
11,7,11,7,5,7	NA	1
11,7,11,7,6,7	NA	1
11,7,6,7,11,7	NA	1
13,7,5,4,5,7	NA	1
5,7,11,7,5,7	NA	1
7,11,7,6,7,11	NA	1
7,5,7,11,7,5	NA	1
7,6,7,11,7,11	NA	1

Combinations with length of 8

Combination	Mean match percentage	Occurrence rate
11,10,11,10,11,10,11,10	0.638	5
10,11,10,11,10,11,10,11	0.556	5
5,7,11,7,5,4,5,7	0.195	3
11,7,11,7,11,7,11,7	0.082	4
7,5,4,5,7,11,7,5	0.063	3
11,7,5,4,5,7,11,7	0.045	7
7,11,7,11,7,11,7,11	0.028	4
7,11,7,5,4,5,7,11	0.028	7
13,8,13,8,13,8,13,8	0.025	7
11,7,11,7,5,4,5,7	0	4
11,7,11,7,5,7,11,7	0	2
12,13,12,13,12,13,12,13	0	3
13,12,13,12,13,12,13,12	0	3
5,4,5,7,11,7,11,7	0	3
6,7,11,7,11,7,11,7	0	2
7,11,7,11,7,11,7,6	0	2
7,11,7,11,7,5,4,5	0	3
7,11,7,11,7,5,7,11	0	2
7,5,4,5,7,11,7,11	0	3
8,13,8,13,8,13,8,13	0	6
11,7,11,7,11,7,6,7	NA	1
11,7,11,7,6,7,11,7	NA	1
11,7,5,7,11,7,11,7	NA	1
11,7,5,7,11,7,5,7	NA	1
5,7,11,7,11,7,11,7	NA	1
5,7,11,7,5,7,11,7	NA	1
7,11,7,11,7,11,7,5	NA	1
7,11,7,5,7,11,7,11	NA	1
7,11,7,5,7,11,7,5	NA	1
7,11,7,6,7,11,7,11	NA	1
7,5,7,11,7,5,7,11	NA	1
7,6,7,11,7,11,7,11	NA	1

Table 4. Common functions from combinations across all cell lines.

Combinations with length of 6.

Combination	Function	Occurrence Rate
10,11,10,11,10,11	acid-amino acid ligase activity	5
10,11,10,11,10,11	ligase activity, forming carbon-nitrogen bonds	5
10,11,10,11,10,11	small conjugating protein ligase activity	5
10,11,10,11,10,11	ubiquitin-protein ligase activity	5
11,10,11,10,11,10	centrosome	5
11,10,11,10,11,10	microtubule organizing center	5
11,10,11,10,11,10	mitotic cell cycle	5
11,10,11,10,11,10	small conjugating protein ligase activity	5
11,10,11,10,11,10	ubiquitin-protein ligase activity	5
10,11,10,11,10,11	cell cycle phase transition	4
10,11,10,11,10,11	cellular macromolecule catabolic process	4
10,11,10,11,10,11	centrosome	4
10,11,10,11,10,11	chromosome, centromeric region	4
10,11,10,11,10,11	DNA repair	4
10,11,10,11,10,11	microtubule organizing center	4
10,11,10,11,10,11	mitosis	4
10,11,10,11,10,11	mitotic cell cycle	4
10,11,10,11,10,11	mitotic cell cycle phase transition	4
10,11,10,11,10,11	protein ubiquitination	4
10,11,10,11,10,11	spindle	4
10,11,10,11,10,11	ubiquitin ligase complex	4
11,10,11,10,11,10	acid-amino acid ligase activity	4
11,10,11,10,11,10	ligase activity, forming carbon-nitrogen bonds	4
11,10,11,10,11,10	microtubule cytoskeleton	4
11,10,11,10,11,10	microtubule organizing center part	4
11,10,11,10,11,10	protein ubiquitination	4
11,10,11,10,11,10	spindle	4
11,10,11,10,11,10	ubiquitin ligase complex	4
11,10,11,10,11,10	ligase activity	4
11,10,11,10,11,10	protein catabolic process	4
11,10,11,10,11,10	protein modification by small protein conjugation	4
11,10,11,10,11,10	proteolysis involved in cellular protein catabolic process	4
11,10,11,10,11,10	ubiquitin-dependent protein catabolic process	4
7,11,7,5,4,5	platelet activation	4

Combinations with length of 8:

Combination	Function	Occurrence rate
10,11,10,11,10,11,10,11	microtubule organizing center	5
11,10,11,10,11,10,11,10	microtubule organizing center	5
10,11,10,11,10,11,10,11	centrosome	4
10,11,10,11,10,11,10,11	ligase activity	4
10,11,10,11,10,11,10,11	microtubule cytoskeleton	4
10,11,10,11,10,11,10,11	small conjugating protein ligase activity	4
10,11,10,11,10,11,10,11	ubiquitin-protein ligase activity	4
10,11,10,11,10,11,10,11	mitotic cell cycle	4
11,10,11,10,11,10,11,10	centrosome	4
10,11,10,11,10,11,10,11	centriole	3
10,11,10,11,10,11,10,11	microtubule organizing center part	3
10,11,10,11,10,11,10,11	ubiquitin ligase complex	3
10,11,10,11,10,11,10,11	modification-dependent macromolecule catabolic process	3
10,11,10,11,10,11,10,11	modification-dependent protein catabolic process	3
10,11,10,11,10,11,10,11	protein modification by small protein conjugation	3
10,11,10,11,10,11,10,11	protein modification by small protein conjugation or removal	3
10,11,10,11,10,11,10,11	proteolysis involved in cellular protein catabolic process	3
10,11,10,11,10,11,10,11	ubiquitin-dependent protein catabolic process	3
11,10,11,10,11,10,11,10	ligase activity	3
11,10,11,10,11,10,11,10	microtubule cytoskeleton	3
11,10,11,10,11,10,11,10	microtubule organizing center part	3
11,10,11,10,11,10,11,10	ubiquitin ligase complex	3
11,10,11,10,11,10,11,10	mitotic cell cycle	3
11,7,11,7,11,7,11,7	regulation of cell junction assembly	3
11,7,11,7,11,7,11,7	regulation of focal adhesion assembly	3
11,7,5,4,5,7,11,7	Fc-epsilon receptor signaling pathway	3
11,7,5,4,5,7,11,7	platelet activation	3
5,7,11,7,5,4,5,7	response to inorganic substance	3