Research Article
Volume 13:1, 2022

Journal of Biometrics & Biostatistics

ISSN: 2155-6180 Open Access

Chromosomal Distribution of Human Disease Genes

Fred Y. Ye[1]*, Lucy L. Xu[1,2] and Adam Y. Ye[3,4]

[1]*International Joint Informatics Laboratory (IJIL) & Jiangsu Key Laboratory of Data Engineering and Knowledge Service, School of Information Management, Nanjing University, Nanjing 210023, China*
[2]*Library of Nantong University, Nantong University, Nantong 226019, China*
[3]*Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA*
[4]*Program in Cellular and Molecular Medicine, Harvard Medical School, Boston, MA 02115, USA*

Abstract

It is necessary and useful to reveal the chromosomal distribution at whole-genome level. By studying the chromosomal distribution of human disease genes, we found that the known genes of single-gene diseases are significantly enriched in chromosome X, and depleted in 10p, 19q. And the genes of multi-gene diseases are significantly tended to be located in fewer chromosomes than random background. The chromosomal distribution at whole-genome level could promote future studies of human disease genes.

Key words: Chromosomal distribution • Human disease gene • Whole-genome

Introduction

Recent years, chromosome network is paid attention [1], with linking 3D structure [2]. Particularly, with considerations of human disease genes [3] and related issues [4], we try to explore the chromosome distribution at whole-genome level [5,6].

The disease-related genes are retrieved from two databases: OMIM [7], and KEGG [8]. The OMIM database is used to obtain the location information of genes on chromosomes by the OMIM Application Programming Interface. The KEGG database allows the KEGG Application Programming Interface to crawl data covering the link between human diseases and genes, which also allows retrieval of cross-references within the diseases and genes database in KEGG.

Here, we studied the chromosomal distribution of human disease genes. We separately studied single-gene diseases and multi-gene diseases. We found the genes of multi-gene diseases are significantly tended to be located in fewer chromosomes than randomly sampling whole-genome genes or genes of single-gene diseases.

Materials and Methods

Data collection

We retrieved the relationship between diseases and genes from KEGG on Dec 8, 2018. We retrieved human whole-genome protein-coding genes from UCSC hg38 refGene table annotated their chromosomal locations by hg38 ideogram table.

***Address for Correspondence:** Fred Y. Ye, International Joint Informatics Laboratory (IJIL) & Jiangsu Key Laboratory of Data Engineering and Knowledge Service, School of Information Management, Nanjing University, Nanjing 210023, China; E-mail: yye@nju.edu.cn*

Received: 09 June, 2021, Manuscript No. jbmbs-21- 33502; **Editor assigned:** 11 June, 2022, PreQC No. P- 33502; **Reviewed:** 17 January, 202, QC No. Q- 33502; **Revised:** 23 January, 2022, Manuscript No. R- 33502; **Published:** 31 January, 2022, DOI: 10.37421/2155-6180.2022.13.91

Statistical analysis

For single-gene diseases, we used two-tailed Fisher's exact test to compare the chromosomal distribution of disease-1 genes with whole-genome genes. For multi-gene diseases, we compared the distribution of m-chromosomal diseases among n-gene disease group with random background, which was done by simulation of random sampling n genes according to chromosome distribution of whole-genome all genes or disease-1 genes. For m-chromosomal diseases within n-gene disease group, expected percentage is calculated based on the simulated distribution of percentage, and empirical p-value is computed by comparing observed percentage with the simulated distribution. All the statistical analysis was done by R, version 3.5.1. P-value <0.05 is regarded as statistically significant.

Results

We retrieved 3462 disease genes related to 1747 human diseases from KEGG. Among them, 971 diseases (56%) are each related to only one gene, so we called such diseases as single-gene diseases. Similarly, 238 (14%) are two-gene diseases, 135 (8%) are three-gene diseases. We summarized the number of n-gene diseases in Table 1.

Since majority of diseases are single-gene diseases, we first studied the chromosomal distribution of their related genes, which we abbreviated as

Table 1. Number of n-gene diseases.

n	Number of Diseases	%
1	971	56
2	238	14
3	135	8
4	92	5
5	57	3
6	46	3
7	37	2
8	23	1
9	33	2
10	13	1
11-15	47	3
16-20	25	1
21-30	16	1
31-40	9	1
41-60	5	<1
Total	1747	100

'disease-1 genes'. As expected, disease-1 genes are more located in the long arms (q) of chromosomes. Compared to the background distribution of human whole-genome 28235 protein-coding genes, we found that disease-1 genes are statistically significantly enriched in chromosome X, and depleted in 10p, 19q (Table 2). The excess in chromosome X might be due to the detection bias of X-linked diseases. There is Loss of heterozygosis at 10p which refer to a specific type of genetic mutation, during which there is a loss of one normal copy of a gene or a group of genes [9]. In some cases, loss of heterozygosis can contribute to the development of cancer. Also, Chromosome 19 not only has the highest gene density among all human chromosomes but also carries a high density of repeat sequences [10]. Both of those may be related to depletion in 10p and 19q.

Further, we explored chromosomal distribution of genes related to n-gene diseases, which we abbreviated as 'disease-n genes'. For each disease, we studied how many chromosomes its related genes are located in. If its related genes are located in m different chromosomes, we defined the disease as a m-chromosomal disease. So, it is for sure that $m \leq n$ for an n-gene disease. We counted the number of m-chromosomal diseases for each n-gene disease group (Table 3). We found that most multi-gene ($n > 1$) diseases are multi-chromosomal ($m > 1$) diseases.

Table 2. Chromosomal distribution of whole genome genes and disease-1 genes.

Chromosome arm	Number of Genes	Percentage of Genes	Number of Disease-1 Genes	Percentage of Disease-1 Genes	Percentage Difference	Fisher's Exact Test p-value
1p	1495	5.30	40	4.83	-0.47	0.636
1q	1340	4.75	44	5.31	0.56	0.456
2p	687	2.43	17	2.05	-0.38	0.566
2q	1119	3.96	43	5.19	1.22	0.087
3p	708	2.51	22	2.65	0.15	0.736
3q	847	3.00	23	2.77	-0.23	0.836
4p	334	1.18	9	1.09	-0.10	1
4q	754	2.67	21	2.53	-0.14	0.913
5p	256	0.91	9	1.09	0.18	0.575
5q	1042	3.69	36	4.34	0.65	0.306
6p	825	2.92	17	2.05	-0.87	0.171
6q	624	2.21	16	1.93	-0.28	0.718
7p	455	1.61	13	1.57	-0.04	1
7q	877	3.11	26	3.14	0.03	0.919
8p	385	1.36	8	0.97	-0.40	0.443
8q	641	2.27	14	1.69	-0.58	0.340
9p	300	1.06	10	1.21	0.14	0.608
9q	799	2.83	23	2.77	-0.06	1
10p	276	0.98	2	0.24	-0.74	0.028
10q	830	2.94	22	2.65	-0.29	0.753
11p	584	2.07	20	2.41	0.34	0.458
11q	1091	3.86	30	3.62	-0.25	0.784
12p	384	1.36	11	1.33	-0.03	1
12q	1007	3.57	33	3.98	0.41	0.506
13p	0	0.00	0	0.00	0.00	1
13q	631	2.23	15	1.81	-0.43	0.474
14p	0	0.00	0	0.00	0.00	1
14q	934	3.31	19	2.29	-1.02	0.113
15p	0	0.00	0	0.00	0.00	1
15q	996	3.53	21	2.53	-0.99	0.149
16p	633	2.24	14	1.69	-0.55	0.339
16q	506	1.79	21	2.53	0.74	0.113
17p	460	1.63	16	1.93	0.30	0.486
17q	1105	3.91	44	5.31	1.39	0.046
18p	120	0.43	1	0.12	-0.30	0.269
18q	312	1.11	14	1.69	0.58	0.128
19p	714	2.53	17	2.05	-0.48	0.498
19q	1063	3.76	12	1.45	-2.32	1.6E-4
20p	262	0.93	13	1.57	0.64	0.067
20q	517	1.83	12	1.45	-0.38	0.509
21p	57	0.20	0	0.00	-0.20	0.414
21q	355	1.26	11	1.33	0.07	0.753
22p	2	0.01	0	0.00	-0.01	1
22q	640	2.27	18	2.17	-0.10	1
Xp	450	1.59	30	3.62	2.02	7.2E-5
Xq	702	2.49	42	5.07	2.58	2.9E-5
Yp	48	0.17	0	0.00	-0.17	0.648
Yq	67	0.24	0	0.00	-0.24	0.268

Table 3. Number of m-chromosomal diseases for n-gene disease group.

n m	1	2	3	4	5	6	7	8	9	10	11-15	16-20	21-30	31-40	41-60
1	971	42	11	2	2	0	1	0	1	0	0	1	1	0	0
2		196	30	6	2	0	0	0	0	1	0	0	1	0	0
3			94	34	7	4	1	0	0	1	0	0	0	0	0
4				50	15	4	1	0	0	0	1	0	0	0	0
5					31	21	7	2	2	2	1	0	0	0	0
6						17	18	5	2	0	2	1	0	0	0
7							9	12	18	4	4	0	0	0	0
8								4	8	4	6	0	0	0	0
9									2	0	16	6	1	0	0
10										1	12	3	2	0	0
11											3	5	2	0	0
12											2	5	0	0	0
13												1	2	0	0
14												2	3	2	0
15												1	1	1	0
16													3	1	0
17														2	2
18														1	2
19														2	0
20															0
21															1
Total	971	238	135	92	57	46	37	23	33	13	47	25	16	9	5
(m >1%)	-	82	92	98	96	100	97	100	97	100	100	96	94	100	100

Table 4. Percentage and statistical test of m-chromosomal diseases for n-gene disease group.

		Background Gene Set:		Whole-genome all Genes			Disease-1 genes		
n	m	Observed Disease Number	Observed Percentage (O)	Expected Percentage (E)	log2 (O/E)	Empirical p-value	Expected Percentage (E)	log2 (O/E)	Empirical p-value
2	1	42	17.8	5.1	1.82	1.3E-12	5.5	1.69	2.2E-11
	2	194	82.2	94.9	-0.21	1.3E-12	94.5	-0.20	2.2E-11
3	1	11	8.2	0.3	4.77	5.4E-13	0.37	4.49	4.4E-12
	2	30	22.4	14.4	0.64	0.013	15.5	0.53	0.031
	3	93	69.4	85.3	-0.30	3.5E-6	84.2	-0.28	2.4E-5
4	1	2	2.2	0.021	6.72	1.8E-4	0.025	6.49	2.4E-4
	2	5	5.5	1.8	1.58	0.027	21.7	1.34	0.049
	3	34	37.4	25.6	0.54	0.016	27.3	0.45	0.034
	4	50	54.9	72.5	-0.4	3.6E-4	70.5	-0.36	1.8E-3
5	1	2	3.5	2.3E-5	10.6	8.4E-7	3.2E-5	10.1	1.6E-6
	2	2	3.5	0.23	3.91	8.0E-3	0.30	3.55	0.013
	3	7	12.3	5.6	1.13	0.040	6.5	0.91	0.098
	4	16	28.1	36.1	-0.36	0.27	37.6	-0.42	0.17
	5	30	52.6	58.1	-0.14	0.42	55.5	-0.08	0.69

To answer whether the m-chromosomal distribution can be simply explained by randomly sampling genes in different chromosomes, we compared the observed distribution with simulated result of random background. For n=2, 3, 4, or 5, we found that there is a significant trend that the multi-gene diseases are tended to located to fewer m chromosomes, no matter using whole-genome all protein-coding genes or disease-1 genes as the background (Table 4). This result indicates that the genes related to the same disease are actually tended to be located in the same chromosome, suggesting that the chromosomal distribution of genes may be associated with their functional relationship. When m=1, n=4, one is 15q13.3 micro deletion syndrome. It is a genetic disorder caused by the deletion of several genes on chromosome 15, it is also known as a micro deletion syndrome or a contiguous gene deletion syndrome. Affected individuals exhibited a complex pattern of behavioural abnormalities, most notably hyperactivity, attention problems, withdrawal, and externalizing symptoms, as well as impairments in functional communication [11]. The other one is diffuse pan bronchiolitis; it may be located in the short arm of human chromosome 6, which is a disease of obscure aetiology that is traditionally associated with people in East Asians, including Japanese, Koreans and Chinese [12]. When m=1, n=5, X-Linked Mixed Deafness (DFN3) is a rare condition, characterized by hearing loss to varying degrees, dizziness and vertigo. The disease is transmitted in an X-linked recessive manner. That is to say, if a mother is a carrier of the recessive hearing loss gene, then the gene is passed on to half of the sons, and half of the daughters [13]. Another one is Wolf-Hirsch horn syndrome, 4p deletion, growth restriction and characteristic facial features variable congenital anomalies Introduction. It is first clinically described in 1961 by Hirsch horn and subsequently in 1965 by Wolf, is the first example of a classic human chromosomal deletion syndrome [14]. As if some genes are mutated or deleted, it may be easy to cause a disease of multi-gene on one chromosome. The same seems to be true for the X chromosome [15,16].

Discussion and Conclusion

With using a bioinformatics model, we revealed that multi-pathogenic genes distribute in multi-chromosomes, when we map pathogenic n-genes (n>1) onto chromosomes based on the datasets from OMIM and KEGG. The research reveals that pathogenic n-genes are linked to m-chromosomes, which distribute following negative exponential, behind random pattern of single pathogenic gene *vs*. single chromosome. Also the gene distribution near 1/3 genes located in p arm and about 2/3 in q arm looks a natural distribution. The distribution did not fit power model, we need more models to verify the pathogenic n-genes mapping to m-chromosomes.

The chromosomal distribution of human disease genes resembles a valuable topic for further studies. The present results of pathogenic n-genes (n >1) onto chromosomes are reflecting on the background of the whole-genome, which could stimulate further exploration in future research.

Author Contributions

L. L. X. collected and processed data and wrote the paper, A. Y. Y. processed the data and designed the research and wrote the paper, and F.Y.Y initiated the idea and wrote the paper.

Additional Information

Competing interests

The authors declare no competing interests.

Acknowledgements

We acknowledge the financial support from the National Natural Science Foundation of China Grant No 71673131.

References

1. Sarnataro, Sergio, Andrea M. Chiariello, Andrea Esposito, and Antonella Prisco et al. “Structure of the human chromosome interaction network”. *Plos One*(2017):1-15. Google Scholar CrossRef Indexed in
2. Dekker, Job, and Leonid Mirny. “The 3D genome as moderator of chromosomal communication.” *Cell*164(2016):1110-1121.Google Scholar CrossRef Indexed in
3. Goh, Kwang-Il, Michael E. Cusick, David Valle, and Barton Childs et al. “The human disease network.” *Proc Natl Acad Sci U S A*104(2007):8685-8690. Google Scholar CrossRef Indexed in
4. Hu J, Zhang Y, Zhao L and Frock RL et al. “Chromosomal loop domains direct the recombination of antigen receptor genes.” *Cell*163(2015):947–959. Google Scholar CrossRef Indexed in
5. Kruglyak, Leonid. “Prospects for whole-genome linkage disequilibrium mapping of common disease genes.” *Nature Genetics*22(1999):139-144. Google Scholar Indexed in
6. Altshuler, David, Mark J. Daly, and Eric S. Lander. “Genetic mapping in human disease.” *Science* 22(2008):881-888. Google Scholar CrossRef Indexed in
7. McKusick, Victor A. “Mendelian inheritance in Man and its online version, OMIM.” *Am J Hum Genet* 80(2007):588-604. Google Scholar CrossRef Indexed in
8. Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, and Yoko Sato et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs.” *Nucleic Acid Res* 45(2017):D353-D361. Google Scholar CrossRef Indexed in
9. Sengelov, Lisa, Mariann Christensen, Hans von der Maase, and Thomas Horn et al. “ Loss of heterozygosis at 1p, 8p, 10p, 13q, and 17p in advanced urothelial cancer and lack of relation to chemotherapy response and outcome.” *Cancer Genet Cytogenet*123(2000):109-113. Google Scholar CrossRef Indexed in
10. Han, Fengyan, Lei Zhang, Chaoyi Chen, and Yan Wang et al. “GLTSCR1 Negatively Regulates BRD4-Dependent Transcription Elongation and Inhibits CRC Metastasis.” *Adv Sci (Weinh)* 6(2019):1-18. Google Scholar CrossRef Indexed in
11. Ziats, Mark N., Robin P. Goin-Kochel, Leandra N. Berry, and May Ali, et al. “The complex behavioral phenotype of 15q13.3 micro deletion syndrome.” *Genet Med* 18(2016):1111-1118. Google Scholar CrossRef Indexed in
12. Keicho, Naoto, and Minako Hijikata. “Genetic predisposition to diffuse pan bronchiolitis.” *Respirology* 6(2011):581-588. Google Scholar CrossRef Indexed in
13. Dahl, Niklas, Jocelyn Laporte, Lingjia Hu, and Valery Biancalana et al. “Deletion mapping of X-linked mixed deafness (DFN3) identifies a 265-525-kb region centromeric of DXS26.” *Am J Hum Genet* 56(1995):999–1002. Google Scholar, Indexed in
14. Hirschhorn, Kurt, Herbert L. Cooper, and I. Lester Firschein. “Deletion of short arms of chromosome 4-5 in a child with defects of midline fusion.” *Humangenetik* 1(1965):479-482. Google Scholar CrossRef Indexed in
15. Clauset, Aaron, Shalizi Cosma Rohila and Newman Mark EJ. “Power-laws in empirical data.” *SIAM Review* 51(2009):661–703. Google Scholar CrossRef Indexed in
16. Newman Mark EJ. “Power laws, Pareto distributions and Zipf's law.” *Contemporary Physics*46(2005):323-351. Google Scholar CrossRef Indexed in

How to cite this article: Ye, Fred Y., Lucy L. Xu and Adam Y. Ye. “Chromosomal Distribution of Human Disease Genes.” J Biom Biostat 13 (2022): 91