Pakistan J. Zool., vol. 50(2), pp 401-407, 2018. **DOI**: http://dx.doi.org/10.17582/journal.pjz/2018.50.2.401.407

Association of Types of Overlapping Genes with the Size of Chromosome in Human Genome

Syed Kashif Nawaz*, Faiza Zubair and Sidra Kanwal

Department of Zoology, University of Sargodha, Sargodha, Punjab, Pakistan

ABSTRACT

Overlapping genes are an integral component of the genome. The present study investigated the various types of overlapping genes in human genome. Study was completed using genome assembly GRCh 38/ hg38 data was accessed using NCBI map viewer for the identification of exact loci of overlapping genes. It was noticed that maximum number (635) of overlapping genes were present on chromosome 1. Chromosome 21 carried the minimum number (85) of overlapping genes. Overlapping genes were further categorized on the basis of their orientation. It was found that the human genome has 62% embedded genes, 17% convergent genes, 16% divergent genes and 5% uni-directional genes. These categories may further be differentiated as coding vs. coding, coding vs. non-coding and mixture of both types on the basis of overlapping gene structure. Our analyses revealed that human genome has various types of overlapping genes. The occurrence of these overlapping genes is not solely dependent on the size of the chromosomes.

Article Information
Received November 2015
Revised 13 March 2017
Accepted 28 June 2017
Available online 24 January 2018

Authors' Contribution
SKN planned and conducted the study. FZ and SK helped in data collection and analysis.

Key words
Convergent overlapping genes, Divergent overlapping genes, Embedded overlapping genes.

INTRODUCTION

Overlapping genes are described as genes whose genomic regions overlap to some extent. For these genes, the high proportion of the transcripts overlap partially or completely. Such genes are present in quick evolving genomes with high mutation rate such as viruses, bacteria and mitochondria (Luo *et al*., 2007).

Overlapping genes are found in microbial genomes, comprising one third of the complete genome. Characteristics (size, distribution *etc*.) of overlapping genes strongly support their role in the regulation of gene expression (Jhonson *et al*., 2004). Makałowska *et al*. (2004) found that overlapping genes are usually involved in controlling various steps of gene expression including transcription, mRNA (messenger ribonucleic acid) splicing and translation.

Human genome consists of around 3.2 billion base pairs. There are 20,000 to 25,000 protein coding genes in humans - one gene for 3,000,000 nucleotides. The distribution of gene is not random, and unexpectedly a large number of genes overlap in human genome (Veeramachaneni *et al*., 2004). The exact number of overlapping genes is unknown in human and almost other organisms (Boi *et al*., 2004). Overlapping genes are classified on the basis of genomic regions shared and the relative orientation of the genes. Gene overlapping is said to be tail to tail when overlap involves 3` region of both genes. Sharing of 5` region is known as head to head. If a genomic sequence of one gene is totally contained in another then this is classified as embedded gene. Most of the overlapping genes are involved in controlling various stages of gene expression (Makalowska *et al*., 2005).

Makałowska *et al*. (2007) reported 15 theoretical situations for the evolution of overlapping genes based on the study of human overlapping gene orthologs in fish and rodents. Analysis of these orthologs confirmed that numerous overlapping genes are not conserved among the vertebral genomes. There are many mechanisms which explain the origin of overlapping genes. According to Keese and Gibbs (1992), overlapping genes are a result of over expression of formerly existing nucleotide sequences during the course of evolution.

The presence of complementary transcripts has many functions like genomic organization and gene regulation in prokaryotes (Inouye and Delihas, 1988). In eukaryotes the effect of antisense RNA is still not clear. But a recent study has predicted its role in the gene expression. A large number of complementary transcripts have been reported in eukaryotes but very few of them have been examined experimentally. There are three mechanisms which explain how these transcripts regulate the gene expression: transcriptional interference, RNA masking, and double-stranded RNA (dsRN-A)-dependent mechanism (Munroe, 2004).

Based on the relative orientation of genes, these

* Corresponding author: kashifnawazshabbir@yahoo.com
0030-9923/2018/0002-0401 $ 9.00/0

overlapping genes are classified into types like convergent, divergent, embedded (one gene is contained in another) and uni-directional (both genes are in the same direction). Solda *et al*. (2008) confirmed that gene overlapping is always specie specific and in most cases it arises by gaining the terminal non coding exons. It has been reported that among 13,484 human-mouse orthologous genes, 10% are overlapping genes (Solda *et al*., 2008). Among these overlapping genes, majority reside on the opposite strand. Most of the overlapping genes on the same strand are of embedded type while the majority of overlapping genes on opposite strand are of convergent forms. The present study aimed to investigate the presence of different types of overlapping genes in human genome. The frequency of each category was also recorded by using Human Genome Assembly GRCh38/hg38.

MATERIALS AND METHODS

The position and relative direction of each overlapping gene was identified using human genome assembly GRCh 38/ hg38. The genome data was accessed by using human genome resource (https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml). Using tools of map viewer, the chromosomes were selected and reading of the sequence was performed after enabling the feature ofgraphics. The chromosomes were selected for the count of overlapping genes identified in the form of graphical presentation. The tools of map viewer designed for the dragging and zooming of the nucleotide sequences were used for the identification of the loci having overlapping genes. The information obtained through the manual reading of the chromosome sequence was used for the generation of secondary data showing the overlapping genes present on each chromosome. The data was stored in the word file for the arrangement of information about the overlapping genes present on the chromosomes. Additionally, the images were stored in power point slides according to the chromosomes.

Genes overlapping in tail to tail orientation or sharing their 3' end were named as convergently overlapping genes (Fig. 1A). Genes overlapping in head to head orientation or sharing their 5' end were named as divergent overlapping genes (Fig. 1B). Genes nested in other genes were categorized as embedded genes. In this type one gene is fully contained in the other (Fig. 1C). The genes in the human genome with the same orientation were termed as uni-directional (Fig. 1D).

We further sub categorized the genes on the basis of regions being shared by both genes. The present data was differentiated into the following sub-categories: (1) Coding vs. coding overlapping genes: Genes having their coding regions overlapping with each other are classified as coding vs. coding (Fig. 1E). (2) Coding vs. non coding overlapping genes: Gene pairs sharing their introns vs. exons were categorized as coding vs. non coding (Fig. 1F). (3) Mixed type overlapping genes: In mixed type both genes are sharing their exons as well as introns (Fig. 1G).

The whole data was saved according to the chromosome number. Count of various types of overlapping genes was completed manually.

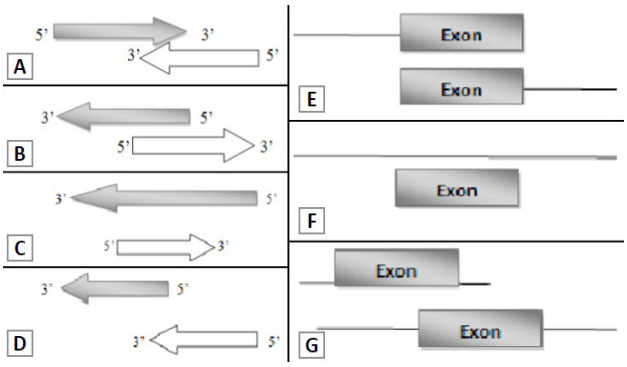

Fig. 1. Types of overlapping genes: A-D, the arrows show the orientation of genes; E-G, the line indicates the introns and boxes represent the exons.

RESULTS

Total number of overlapping genes in the human genome was 6305 (Table I). The maximum number of overlapping genes (635) was present on chromosome 1. Chromosome 21, as the smallest of all the chromosomes, had the lowest number of overlapping genes (85). The total number of embedded overlapping genes was 3896(62%). The chromosome with highest number of embedded type overlapping gene pairs is 1, with 386 (10%) of all overlapping gene pairs. Chromosome 22 has the smallest number of embedded type overlapping gene pairs (52) making only 1.3% of total embedded overlapping genes. The total number of convergent overlapping gene found in whole genome was 1093 (17%). The maximum number of convergent type overlapping gene pair 115 (10.5%) resides on chromosome 1. Chromosome 17, 11, 19 contain 75, 71, 71 convergent overlapping gene pairs respectively (~30% of all convergent type overlapping gene pairs). The chromosome with the least number of convergent overlapping gene pairs is 21(~1%). The total number of divergent overlapping genes is 1034 (16%). Chromosome 1 has maximum number of divergent type overlapping gene pairs (111) which makes about 11% of all overlapping genes. Despite the small size, chromosome 17 has the fourth highest count of divergent overlapping

gene pairs (6.28 %). Chromosome 21 and X has the least number of divergent overlapping genes (nearly 0.67% and 0.77%, respectively). The total number of uni-directional overlapping genes is 282 (5%). In difference to count of other types of overlapping genes, chromosome 2 has highest number of uni-directional overlapping gene pairs (16%). Chromosome 18 and Chromosome 21 has least number of uni-directional overlapping genes (0.35%). Chromosome 1 had the second largest number (8%) of uni-directional overlapping genes.

The total number of coding vs. coding overlapping genes is 851 which is 14% of all overlapping genes (Table I). Chromosome 1 has the largest number of coding vs. coding overlapping genes (11%). Chromosome 17 has the second highest number of coding vs. coding overlapping gene pairs. Present observations also revealed that chromosome 21 has lowest number of coding vs. coding overlapping gene pairs, 2 out of 851. Total number of coding vs. non-coding type of genes is 1600 (25 % of all overlapping genes). Interestingly all coding vs. non coding overlapping genes belongs to embedded type of overlapping genes. According to the findings, a maximum number of the coding vs. non coding overlapping genes resides on chromosome 1 (12%). Chromosomes 14 contain least number of coding vs. non coding overlapping genes nearly 0.25% of all coded vs. non coded overlapping genes. Chromosome 6 and chromosome 8 have equal number of coding vs. non coding overlapping gene pairs. Total number of mixture type of overlapping genes found was 3854. It was observed that 61% of all overlapping gene pairs are of mixture type. Chromosome 1 had the highest number of these types of overlapping genes (about 9%). Together with chromosome 3 and 7, Chromosome 1 contributes 2% to all mixture overlapping genes. Interestingly chromosome X contains least number of mixture of overlapping gene regions.

Table I.- Count of overlapping genes on the basis of gene orientation and structure (Ch., Chromosome).

S No.	Ch. No.	Total genes	Overlapping genes on the basis of gene orientation				Overlapping genes on the basis of gene structure		
			Embedded genes	Convergent genes	Divergent genes	Unidirectional genes	Coding vs. coding	Coding vs. non coding	Mixture
1	1	635	386	115	111	23	94	189	352
2	2	252	118	50	40	44	33	69	150
3	3	415	252	64	80	19	52	82	281
4	4	238	143	38	49	8	18	59	161
5	5	322	215	46	54	7	37	63	222
6	6	330	182	63	67	18	36	71	223
7	7	437	319	51	57	10	51	97	289
8	8	233	149	43	32	9	30	71	132
9	9	288	190	46	41	11	42	87	159
10	10	301	193	48	51	9	30	53	218
11	11	375	233	71	64	7	65	109	201
12	12	159	61	51	34	13	40	31	88
13	13	166	115	21	23	7	19	43	104
14	14	122	55	31	25	11	42	4	76
15	15	222	145	36	35	6	22	62	138
16	16	270	153	65	45	7	39	65	166
17	17	415	264	75	65	11	68	113	234
18	18	113	85	13	14	1	6	40	67
19	19	315	176	71	49	19	57	87	171
20	20	201	120	34	38	9	27	17	157
21	21	85	68	9	7	1	2	30	53
22	22	92	52	21	13	6	14	23	55
23	X	117	83	14	8	12	10	56	51
24	Y	202	139	17	32	14	17	79	106
Total		**6305**	**3896**	**1093**	**1034**	**282**	**851**	**1600**	**3854**

Table II.- Count of convergent and divergent overlapping genes on the basis of shared gene regions.

S No.	Ch. No.	Convergent overlapping genes		Divergent overlapping genes	
		C vs. C	Mix.	C vs. C	Mix.
1	1	43	72	29	82
2	2	11	39	14	26
3	3	45	19	21	59
4	4	4	34	9	40
5	5	5	41	8	46
6	6	19	44	11	56
7	7	14	37	9	48
8	8	15	28	4	28
9	9	20	26	10	31
10	10	7	41	10	41
11	11	29	42	22	42
12	12	18	33	13	21
13	13	6	15	4	19
14	14	12	19	10	15
15	15	5	31	7	28
16	16	24	41	10	35
17	17	30	45	15	50
18	18	1	12	4	10
19	19	31	40	10	39
20	20	7	27	4	34
21	21	0	9	0	7
22	22	4	17	2	11
23	X	3	11	1	7
24	Y	3	14	4	28
Total		**356**	**737**	**231**	**803**

Ch., chromosome; C, coding; NC, non-coding; Mix., Mixture.

The information about convergent and divergent overlapping genes on the basis of shared gene regions is shown in Table II. Total number of coding vs. coding convergent genes was 356 (33% of all convergent type overlapping gene pairs). The Chromosome 21 had no coding vs. coding and convergent genes present on it. The maximum number is located on Chromosome 3 with 45 out of 356 coding vs. coding genes with convergent orientation. Mixture type of convergently overlapped genes was 737 in number. About 67% gene pairs were of mixture type among convergent overlapping genes. The maximum gene pairs with such pattern of overlapping genes reside on Chromosome 1. The minimum percentage is shown by Chromosome 21 (~1.2%). Together with Chromosome 6 and 17, Chromosome 1 contributes 22% to all such overlapping genes. The total number of coding vs. coding divergent overlapping genes was 231 approximately 22% of all divergent genes (Table II). Maximum number of coding vs. coding gene pairs resides on chromosome 1. Interestingly there is no gene pair with such overlap on Chromosome 21. It was estimated that mixture type overlapping gene pairs were 803 (78 % of all divergent genes). Mixture type overlapping gene pairs are found in much greater proportion than coding vs. coding gene pairs among divergent overlapping genes. Chromosome 1 has the maximum number of mixture type overlapping genes. Chromosome 2 is ranked as second for having mixture type of overlapping gene pairs.

Table III.- Count of embedded and unidirectional overlapping genes on the basis of shared gene regions.

S No.	Ch. No.	Embedded overlapping genes			Unidirectional overlapping genes	
		C vs. C	C vs. NC	Mix.	C vs. C	Mix.
1	1	14	189	183	8	15
2	2	5	69	44	3	41
3	3	8	82	162	4	15
4	4	5	59	79	0	8
5	5	23	63	129	1	6
6	6	5	71	106	1	17
7	7	24	97	198	4	6
8	8	11	71	67	0	9
9	9	10	87	93	2	9
10	10	8	53	132	5	4
11	11	12	109	112	2	5
12	12	9	31	21	0	13
13	13	9	43	63	0	7
14	14	17	4	34	3	8
15	15	8	62	75	2	4
16	16	5	65	83	0	7
17	17	20	113	131	3	8
18	18	1	40	44	0	1
19	19	11	87	78	5	14
20	20	16	17	87	0	9
21	21	2	30	36	0	1
22	22	8	23	21	0	6
23	X	5	56	22	1	11
24	Y	7	79	53	3	11
Total		**243**	**1600**	**2053**	**47**	**235**

Ch., chromosome; C, coding; NC, non-coding; Mix., Mixture.

The total number of embedded and uni-directional type of genes was 4178 (Table III). The total number of

coding vs. coding genes with embedded type overlap was 243 (6% of all embedded overlapping genes in human genome). Chromosome 7 had the highest number of this type of overlapping gene pairs (about 10%). Chromosome 18 had the lowest number of coded vs. coded overlapping gene pairs, one out of 243 (Table III). It was observed that coding vs. non coding overlapping genes comprises of 1600 gene pairs (approximately 41% of all embedded gene pairs). The maximum gene pairs of this category reside on Chromosome 1 (12%). Chromosome 1, together with chromosome 17 and 11 comprise 26% of all coded vs. non coding embedded overlapping genes. The least number of this category is located on Chromosome 14. It was observed that mixture embedded genes were 2053 (approximately 53% of all embedded genes). Maximum number of such overlapping was shown by Chromosome 7 that nearly comprised 10% of these genes. The least percentage (1%) was shown by Chromosome 12 and 22. Total number of coding vs. coding genes with uni-directional overlapping was 47 (approximately 17% of all uni-directional genes). There was no coding vs. coding genes with unidirectional overlapping in Chromosome 4, 8, 12, 13, 16, 18, 20, 21 and 22. The maximum number of this kind of overlapping gene pairs was on Chromosome 1 (17%). The total number of gene pairs with mixture type unidirectional overlapping was 235 (83% of all uni-directional overlapping genes) in human genome. Chromosome 2 has shown maximum number of such gene pairs among uni-directional overlapping genes (17%). Present results also revealed that chromosome 21 and 18 has lowest number of coding vs. coding overlapping gene pairs 1 out of 235. While Chromosome X and Y have shown the equal number of such overlapping gene pairs.

DISCUSSION

The actual number of overlapping genes in human is still uncertain because the total number of protein coding genes were estimated to be 32000 in 2001 (Venter *et al*., 2001) and subsequently estimated to be 22,000 in 2004 (Makalowska *et al*., 2004). The number of overlapping genes and the relative percentage was not identified in those studies. We noticed that the 6305 overlapping gene pairs were present in human genome. Among these, the percentages for the convergent, divergent, embedded and uni-directional overlapping genes were different. Embedded type overlapping genes comprised 62% of all overlapping genes in humans. The percentage for convergent, divergent and uni-directional was 17%, 16% and 5% respectively. The trend of these overlapping genes corresponds to size of chromosome to some extent such as Chromosome 1 being the largest of all, has the maximum number of overlapping genes (635) while the least number of these gene pairs were located on chromosome 21 (the smallest chromosome). However variation exists among the size of chromosome and the number of overlapping genes. The previous study conducted by Galante *et al*. (2007) reported 25–27% for the convergent, 27–30% for the divergent, and 43–48% for the embedded types of overlaps. The comparison with the current findings indicates the difference in percentages for the different types of overlapping genes. In the present study, the percentage of convergent, divergent and embedded types of overlaps were 17%, 16% and 62%, respectively. Additionally, the present study also provides the information about the unidirectional overlapping genes. It was found that 5% unidirectional overlapping genes are present in the genome assembly under observation. The difference in findings of two studies may be defined in terms of scanning method used for the identification of overlapping genes. Source of genomic data was also different in the study reported by Galante *et al*. (2007). Genomic build No. 35 was used in that study whereas the present study is based on the latest genomic build (GRCh 38/ hg 38).

We further identified the overlapping genes on the basis of genomic regions shared by gene pairs. The percentage of overlapping genes varied greatly for this classification. We found that most of the overlapping genes of human genome are of mixture type, covering 61% of all overlapping genes. Coding vs. coding and coding vs. non coding overlapping genes encompasses 14% and 25% overlapping genes, respectively.

In above categories (based on genomic region being shared) the maximum overlapping genes were found on chromosome 1. While the least number of coding vs. coding, coding vs. non coding and mixture type overlapping genes were observed on chromosomes 21, 4 and X, respectively. Sanna *et al*. (2008) reported 51 coded vs. coded gene pairs (8.3 %) out of 615 overlapping genes. Discordantly, according to the present observation, a total of 851 coding vs. conding overlapping genes (14%) were observed. The difference in the findings may be due to the use of different sources for genomic data analysis. The report of Sanna *et al*. (2008) is based on the ENSEMBLE version 44. The current work was completed using the information from latest genomic assembly accessed through NCBI. We further analyzed the convergent overlapping genes on the basis of genomic regions being shared by both. The ratio of coding vs. coding gene pairs among the convergent overlapping genes is 33% with maximum number on chromosome 3 and least on chromosome 21. This trend may also correspond to the length of chromosomes. Secondly, mixture gene pairs among convergent overlapping genes accounts for 67% of all such gene overlap. We found no

convergent overlapping genes that belong to coded vs. non coded category.

The percentage of coding vs. coding genes among divergent overlapped genes is 22%. Most of the divergently coded vs. coded overlapping genes were found on chromosome 1. This type of overlapping gene was not found on chromosome 21. While the mixture gene pairs among divergently overlapping genes are clearly more common because this type comprises 78% of all divergent overlapping genes. Again maximum and minimum number was observed on chromosome 1 and 21, respectively.

We also calculated the percentage of coding vs. coding, coding vs. non coding and mixture pairs among embedded overlapping genes. The most common type among embedded overlapping genes is mixture genes that nearly comprises 53% all embedded genes. It was observed that maximum number of such gene reside on chromosome 7 and least were identified on chromosome 12 and 22. While percentage of coded vs. coded and coded vs. non coded types among embedded overlapping genes is 6% and 41%, respectively. The maximum number of coded vs. coded and coded vs. non coded genes among embedded overlapping genes were identified on chromosome 7 and 1, respectively.

In uni-directional overlapping genes, coding vs. coding and coding vs. non coding pairs accounts for 17% and 83%, respectively. Uni-directional genes are found in less number in human genome. Coding vs. coding gene pairs among uni-directional overlapping genes are absent in chromosome 4, 8, 12, 13, 16, 18, 20, 21 and 22. While least number of mixture type overlapping gene pairs were identified on chromosome 21.

The present study suggests that overlapping genes are also present in human genome. Their different types may be identified in humans. Number of various types of overlapping genes is not solely dependent on size of chromosomes. Some chromosome may lack some specific types of overlapping genes in human genome.

CONCLUSION

The present study concludes that the different types of overlapping genes are unevenly distributed in the whole genome. The size of a human chromosome does not reflect the estimate of the count of overlapping genes located on it. The chromosome, smaller in size may have higher number of a specific type of overlapping genes as compared to the chromosome with the larger size.

Statement of conflict of interest

Authors have declared no conflict of interest.

REFERENCES

Boi, S., Solda, G. and Tenchini, M.L., 2004. Shedding Light on the dark side of the genome: overlapping genes in higher eukaryotes. *Curr. Genom*., **5**: 509-524. https://doi.org/10.2174/1389202043349020

Galante, P.A., Vidal, D.O., de Souza, J.E., Camargo, A.A. and de Souza, S.J., 2007. Sense antisense pairs in mammals: functional and evolutionary considerations. *Genome Biol*., **8**: 14-21. https://doi.org/10.1186/gb-2007-8-3-r40

Inouye, M. and Delihas, N., 1998. Small RNAs in the prokaryotes: A growing list of diverse roles. *Cell*, **53**: 5-7. https://doi.org/10.1016/0092-8674(88)90480-1

Johnson, Z.I. and Chisholm, S.W., 2004. Properties of overlapping genes are conserved across microbial genomes. *Genome Res*., **14**: 2268-2272. https://doi.org/10.1101/gr.2433104

Keese, P.K. and Gibbs, A., 1992. Origins of genes: "big bang" or continuous creation? *Proc. natl. Acad. Sci. U.S.A*., **89**: 9489-9493. https://doi.org/10.1073/pnas.89.20.9489

Luo, Y., Fu, C., Zhang, D.Y. and Lin, K., 2007. BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. *BMC Bioinform.,* **8**: 266. https://doi.org/10.1186/1471-2105-8-266

Makalowska, I., Chiao-Feng, L. and Makalowski, W., 2004. Overlapping genes in vertebrate genomes. *Biomed. Sci*., 27: 429-430.

Makałowska, I., Lin, C.F. and Hernandez, K., 2007. Birth and death of gene overlaps in vertebrates. *BMC Evolut. Biol*., **7**: 193. https://doi.org/10.1186/1471-2148-7-193

Makalowska, I., Lin, C.F. and Makalowski, W., 2005. Overlapping genes in vertebrate genomes. *Comput. Biol. Chem*., **29**: 1-12. https://doi.org/10.1016/j.compbiolchem.2004.12.006

Munroe, S.H., 2004. Diversity of antisense regulation in eukaryotes: Multiple mechanisms, emerging patterns. *J. cell. Biochem*., **93**: 664-671. https://doi.org/10.1002/jcb.20252

Sanna, C. R., Wen-Hsiung, L. and Zhang, L., 2008. Overlapping genes in the human and mouse genomes. *BMC Genom*., **9**: 169. https://doi.org/10.1186/1471-2164-9-169

Soldà, G., Mikita, S., Paride , P., Silvia, B., Alessandro, G., Ermanno, R., Peer, B., Maria, L.T. and Francesca, D.C., 2008. Non-random retention of protein-coding overlapping genes in Metazoa. *BMC Genom.,* **9**: 174. https://doi.org/10.1186/1471-

2164-9-174

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R. and Makalowska, I., 2004. Mammalian overlapping genes: The comparative perspective. *Genome Res*., **14**: 280-286. https://doi.org/10.1101/gr.1590904

Venter, J.C., Adams, M. and Myers, E.W., 2001. The sequence of the human genome. *Science*, **291**: 1304-1351. https://doi.org/10.1126/science.1058040