Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015)
Barcelona, Spain – July 13 - 14, 2015
Paper No. 314

An Efficient Method for Storing Human Genome Variations

Onur Çakırgöz, Süleyman Sevinç
Dokuz Eylül University, Department of Computer Engineering
Tınaztepe Buca, İzmir, Turkey
onurcakirgoz@cs.deu.edu.tr; suleyman.sevinc@cs.deu.edu.tr

Abstract –Storing human genome structurally is one of the fundamental issues in bioinformatics. Discovering the relationship between genotype and phenotype, in other words, discovering which genes, genetic elements, and variations are associated with particular diseases or traits requires storing large number of people's genome. Methods to be developed for the storage of human genome should take into account some particular issues. The first case we must consider is that the method should store all the necessary genomic data using minimal space. The second case is that the method should enable the studies investigating which genes, genetic elements, and variants are associated with particular diseases or traits without needing major data transformation. We have developed an efficient method to store human genome considering these two cases. In our study, we propose storing all the variations that exist in the genome, rather than storing the raw sequence. If we consider the total 46 chromosomes a human has, reference sequence consists of 6072607692 and 5976710698 base pairs for females and males respectively. To store this data, we need a space whose size is approximately 5,65 GB for females and 5,56 GB for males. On the other hand, raw sequence does not make any sense. Because for the diagnosis and the treatment of genetic diseases, clinicians need the variations that individuals have. To find the variations, this raw sequence should be aligned to the reference sequence by using well-known alignment algorithms. Instead of finding variations by aligning each time, storing the variations we get by aligning once is more reasonable. Necessary analyses were performed on variation data published by 1000 Genomes Project and the methodology we developed was tested using the results of the analyses. Based on our method, except for the structural variations, approximately 30,08 MB and 29,66 MB are sufficient to store all the variations which exist in the genome of any female or male respectively.

Keywords**:** storage method, personal genetic data, human genome variations, 1000 genomes project.

1. Introduction

Humans have 22 pairs of autosomal chromosomes and one pair of sex chromosomes. Females have two X chromosomes, while males have an X and a Y chromosome. That means we have 46 total chromosomes. And all these 46 chromosomes are called the human genome (Alberts B., et al., 2007). If we add up the lengths of the human reference chromosomes (The International Human Genome Sequencing Consortium and Web-4), we see that female genome consists of 6072607692 base pairs and male genome consists of 5976710698 base pairs. The difference between the length of female genome and the length of male genome arises from the difference of X chromosome and Y chromosome. But the values specified here are valid for the reference genome. Actually, the lengths of the individuals' genomes are different from each other (Levy S., et al., 2007). While some people's genomes are longer than the reference, others' genomes are shorter. Although there are differences between the lengths of the individuals' genomes, these differences are relatively very small. Therefore, on the average, the length of a person's genome can be considered as the length of the reference genome. When we consider human genome from computational perspective, we will see a string made up of letters A,G,C,T. And to store this data, we need a space whose size is approximately 5,65 GB for females and 5,56 GB for males.

Since Dna sequence is a string consisting of letters A, G, C, T, researches have intensively used textual data compression techniques to reduce the huge storage costs. Textual data compression techniques are basically classified under four main headings: Substitutional-Statistical methods, Grammar-based methods, Transformational methods, and Table compression methods. Among these

methods, Substitutional-Statistical methods have showed the greatest improvements. The basic logic of this class of methods is, as the name implies, combining the substitutional techniques and statistical techniques in order to improve the compression ratio. Biocompress 1 (Grümbach S., Tahi F., 1993) is the first example of Substitutional-Statistical methods. Later, many studies were published. Among them, the well-known studies are DNACompress (Chen X., et al., 2002) and DNAPack (Behzadi B., Fessant F.L., 2005). Although these studies combine the substitutional and statistical techniques to improve the compression performance, XM (Cao M.D., et al., 2007) a pure statistical compression method, yielded better results than DNAPack, the best performing of the Substitutional-Statistical methods that we analyzed. If XM is used to compress the human genome, approximately 1,20 GB and 1,18 GB will be sufficient to store female genome and male genome respectively.

Researchers do experimental studies to discover which genes, genetic elements, and variants are associated with particular diseases or traits. Especially, they aim to determine which variants are more common among people with the disease as compared to those without the disease. The most familiar examples of such studies are those associated with the drug Warfarin. These studies investigate how genetic variants contribute to differences in patients' responses to warfarin (Lindh J. D., 2005), (Schwarz U. I., et al., 2008), (Veenstra D. L., et al., 2005), (Millican E. A., et al., 2007). In order to achieve this goal, they stored genetic variants of many individuals. This is the most important common feature of these studies which concerns us. Namely, the fundamental elements of the studies investigating the relationship between genotype and phenotype are variations. But unfortunately such studies take too long and their costs are too high. Significant reduction of sequencing costs due to the recent advances in sequencing technology has enabled the realization of the 1000 Genomes Project (1000 Genomes Project Consortium). The 1000 Genomes Project is the first project that has performed the sequencing of a large number of individuals' whole-genomes. At the end of the project, the genetic variations of 2504 anonymous people from 26 populations around the world were published. Researchers can freely access these results. This saves researchers the time and expense of having to sequence their own samples.

The 1000 Genomes Project published the variation data of 2504 anonymous people as VCF (Web-1) and BCF (Web-2) files. The VCF (Variant Call Format) format is a tab delimited text file format for storing variations and individual genotypes (Web-3). Although VCF is widely used in the community, it has two substantial drawbacks. Because the file is text, it requires a lot of space on disk and is excessively slow to parse. BCF is a binary, compressed equivalent of VCF. A BCF file is composed of a series of compressed blocks of binary records. BCF files are faster and take up less space compared to VCF files. Because these two file formats are equivalent, both have some common shortcomings. They are not much convenient to store genetic data of a single person. On the other hand, while storing the genetic data of many people; low allele frequencies lead to so much redundant space usage. The cost of sequencing whole-genome is too high with older sequencing devices. However, older sequencing devices are still widely used in most clinics (Wheeler D. A., et al., 2008). Therefore, clinics using older devices sequence small regions most of the time. In that case, start and end positions of the regions should be stored too. But unfortunately, there is no field to store the start and end positions of the regions in both formats.

Structural approaches for storing and querying genomic data are Genomics Algebra (Hammer J., Schneider M., 2003) and GQL (Bafna V., et al., 2013). J. Hammer and M. Schneider have proposed an integrating approach that is based on two fundamental structures. These are Genomics algebra and genomics functions. While genomic algebra consists of genomic data types (e.g., genome, gene, protein, nucleotide), genomic functions consist of functions (e.g., translate, transcribe). They propose extending SQL by embedding these two structures into it. However, in order to extend SQL, data structures that will be running in the background should be created and added to the system. For instance, to add a new function such as get_variations( ) to SQL, necessary data structure should be added to the system too. The second study, GQL, is very similar to the first study. GQL is also based on genome query algebra and uses a standard SQL-like syntax. In fact, the main difference between the studies is that while the first study recommends making additions to existing database management systems, the second study recommends constructing the system from scratch. Naturally, GQL has also data-structures requirement.

These two studies have other common shortcomings too. The space requirements of the methods were not calculated. The methods were not tested on real personal genetic data.

2. Our Storage Method

We propose storing variations, rather than storing the raw sequence. Through the method we developed, the space requirement will be reduced considerably and the studies will be facilitated investigating which variants are associated with particular diseases or traits.

The basic element of the method we developed is variation. There are two different variation formats. While the same format was designed for both substitution and insertion, a different format was designed for deletion. The cause that gives rise to this situation is while we should store the sequence in insertion and substitution, we don't need to store the sequence in deletion. As known, sequence is made up of nucleotide bases. There are four bases. These are Adenine, Guanine, Cytosine and Thymine. Apart from these, sequencing devices return "N" character for the bases which cannot be identified. Therefore, there are totally 5 characters. Three bits are sufficient to represent these 5 characters. But, since a byte is composed of 8 bits, we preferred using 4 bits to represent the bases. Naturally, one byte can store two characters. Accordingly, the formula "Ceiling(Length/2)" is used to compute the number of bytes adequate to store the sequence, where "Length" stands for the length of the sequence in the formula. The first field of both formats is "L_T". While higher 4 bits of L_T indicates the length of the sequence, lower 4 bits indicates the type of the variation. If the value of the higher 4 bits of L_T is 15, there are two alternatives. The length of the sequence might be either 15 or more than 15. Therefore, there is also an extra field "Len" in the case where the value of the higher 4 bits of L_T is 15. The field Len stores the length of the sequence. The last field of the format devised for substitution and insertion is "Alt" and this field is used to store the sequence. Since we don't need to store the sequence in deletion, there is no "Alt" field in the deletion format. You can attain the other details about the variation format from Table-1.

Table. 1. Variation Formats

Variation Type	Field		Description	Type
Substitution & Insertion	L_T		Length of the Sequence(Higher 4 bits) and Type of the Variation(Lower 4 bits)	Unsigned Byte
	If Length < 15	Alt	A byte array storing the sequence, byte[Ceiling(Length/2)]	Byte Array
	If Length = 15	Len	Length of the Sequence	Unsigned 16 Bit Integer
		Alt	A byte array storing the sequence, byte[Ceiling(Length/2)]	Byte Array
Deletion	L_T		Length of the Sequence(Higher 4 bits) and Type of the Variation(Lower 4 bits)	Unsigned Byte
	If Length < 15	-	Null, no information is stored	-
	If Length = 15	Len	Length of the Sequence	Unsigned 16 Bit Integer

The general format we developed can be seen from Table-2. The costs of sequencing whole-genome decreased considerably due to the recent progressions in sequencing technology. However, older

sequencing devices are still used extensively in most clinics. And, most of the time clinics using older devices sequence small regions distant from each other in accordance with the requirements. Considering this situation, the general format has been devised to store all or a portion of the variations that exist in any particular chromosome of an individual. The format can be defined shortly as it is composed of regions; a region is composed of records. The details of the general format are as follows: Since variations are determined by aligning a sequence to the reference sequence, the assembly of the human reference genome should be specified. Thus, the first two fields of the general format are used for that purpose.

These fields are "L_Hga" and "Hga". While L_Hga stores the length of the human genome assembly, Hga stores the human genome assembly. The third field is "N_Regions". N_Regions is used to store the number of discrete regions. The three fields ("S_Pos", "E_Pos" and "N_Rec") following the field N_Region define a region. S_Pos, E_Pos and N_Rec store the start position and end position of the region, and the number of records in the region respectively. There are two different Record formats, one for haploid calls and the other for diploid calls. Haploid format consists of two fields ("V_Pos", "Var") whereas diploid format consists of three ("V_Pos", "I_Byte", "Var"). The extra field "I_Byte" of the diploid format is used to represent four situations whose details are given in Table-2. For haploid calls, e.g. on Y, male nonpseudoautosomal X, only one allele value should be given. Accordingly, one variation is stored in the Var field of Haploid format. For diploid calls, e.g. on chromosome 1, female nonpseudoautosomal X, two alleles' values should be given. Here, based on the value of the field "I_Byte", either one variation or two variations can be stored in the Var field of Diploid format.

Table. 2. General Format

Field			Description	Type
L_Hga			Length of the Human Genome Assembly	Unsigned Byte
Hga			Human Genome Assembly (ex. GRCh37)	Char[L_Hga]
N_Regions			Number of Regions	Unsigned 32 Bit Integer
List of Regions (n=N_Regions)				
S_Pos			Start Position of the Region	Unsigned 32 Bit Integer
E_Pos			End Position of the Region	Unsigned 32 Bit Integer
N_Rec			Number of Records in the Region	Unsigned 32 Bit Integer
List of Records (n=N_Rec)				
	If Haploid	V_Pos	Position of the Variation	Unsigned 32 Bit Integer
		Var	Variation	Variation
	If Diploid	V_Pos	Position of the Variation	Unsigned 32 Bit Integer
		I_Byte	For Diploid Genotypes, there are four situations: Only first allele has the variation, I_Byte=00000000 Only second allele has the variation, I_Byte=00000001 Both alleles have the same variation, I_Byte=00000010 The alleles have different variations, I_Byte=00000011	Unsigned Byte
		Var	If the value of I_Byte is 00000011, two variations	Variation
			Else, one variation	Variation

The results of the analyses made on the variation data published by 1000 genomes project are seen in Table-3, Table-4 and Table-5. The values in these three tables represent the averages of many samples. Table-3 shows the average lengths of the variations for each of the chromosomes. The lengths of the variations for the chromosomes 1 to 22 represent the averages of 2504 samples including females and males. On the other hand, the values of the X chromosome were calculated separately for females and males. Because the Y chromosome is found only in males, the values of the Y chromosome are averages

of only males. When Table-3 is examined, except for the one or two exceptions, values are seen to be very close to each other. For instance, the average lengths of deletions for each of the chromosomes (except chromosome Y) are closer to 3.

Table. 3. Average Lengths of Variations

Chromosome No	Insertion	Deletion	Substitution
1)	2,73693	2,98974	1,00134
2)	2,71074	2,99387	1,00126
3)	2,72172	2,87282	1,00117
4)	2,72560	2,90215	1,00102
5)	2,67391	2,96484	1,00098
6)	2,78769	3,00509	1,00133
7)	2,69781	2,95759	1,00116
8)	2,72771	2,93116	1,00106
9)	2,78807	3,02053	1,00120
10)	2,77815	2,95977	1,00145
11)	2,68479	2,94463	1,00103
12)	2,77193	2,92728	1,00187
13)	2,77706	2,98568	1,00119
14)	2,64973	2,97081	1,00114
15)	2,73844	2,97968	1,00124
16)	2,87496	3,07868	1,00121
17)	2,83552	3,01799	1,00143
18)	2,82122	2,98795	1,00122
19)	2,94230	3,08953	1,00137
20)	2,85493	3,11415	1,00144
21)	2,40971	2,77012	1,00129
22)	3,13896	3,31479	1,00125
X)F	2,65969	2,95998	1,00189
X)M	2,66046	2,98178	1,00179
Y)	1,26473	2,05990	1,02386

Table-4 shows the total number of variations existing in each chromosome of a person. The values in Table-4 were calculated to be compatible with the method we developed. When we look at the table carefully, we see that the values of the field “Only First allele” are very close to the values of the field “Only Second Allele”. Another surprising case is that the values of the field “Different” are very small when compared to the values of the field “Total”.

Table-5 shows the total number of variations a person has in the whole-genome. The values in Table-5 were obtained by adding the value of each chromosome in the relevant field of Table-4. The fields “Only First Allele” and “Only Second Allele” of Table-4 were combined as the field “Only One Allele” in Table-5. Since females have two X chromosomes as sex chromosomes, they have no haploid variations. But, males have one X chromosome and one Y chromosome. Thus, they have haploid variations. Similarly to Table-4, the values in Table-5 are average values. Naturally, any female or male might have more or less variations than the values represented in the table.

Table. 4. Total Number of Variations Existing in each chromosome of a Person

Chr No	Only First Allele				Only Second Allele				Total of Only One Allele	Homozigot				Total of Homozigot	Different	Total
	Ins	Del	Sub	Str	Ins	Del	Sub	Str		Ins	Del	Sub	Str			
1)	6162	7627	90411	67	6181	7660	90330	69	208510	8461	9614	115386	47	133510	386	342408
2)	6340	7784	96197	66	6363	7819	96084	69	220726	8868	9865	121580	34	140348	406	361482
3)	5435	6634	82602	58	5458	6671	82626	60	189548	7266	8450	102479	25	118221	357	308126
4)	5541	6815	85865	60	5568	6850	85922	62	196686	7836	8698	112160	30	128725	391	325804
5)	4988	6022	73362	52	5012	6058	73424	54	168976	6569	6835	84997	29	98431	306	267713
6)	5141	6226	77843	62	5158	6259	77823	64	178580	6593	7254	94129	38	108016	365	286963
7)	4385	5511	69250	45	4401	5535	69221	47	158399	6225	6655	83000	31	95913	293	254606
8)	3824	4663	64683	43	3848	4693	64848	45	146650	4951	5703	77557	20	88232	271	235154
9)	3217	3890	50485	37	3224	3900	50432	38	115226	4029	4749	61058	20	69857	208	185292
10)	3842	4730	59153	38	3859	4756	59182	39	135603	5117	6277	73327	30	84753	251	220608
11)	3700	4476	57853	46	3719	4501	57849	48	132195	5729	6188	77224	18	89161	256	221613
12)	3869	4734	55426	43	3883	4761	55459	45	128224	5153	6068	69976	20	81219	255	209698
13)	2934	3514	42363	35	2948	3530	42429	37	97793	4499	4807	58879	14	68200	206	166200
14)	2615	3179	37922	29	2622	3188	37856	30	87444	3628	4078	47804	17	55529	159	143133
15)	2334	2865	34061	29	2340	2880	34063	30	78605	3181	3793	43104	13	50092	152	128850
16)	2254	2777	36656	22	2263	2795	36674	22	83466	2861	3414	44559	24	50859	149	134475
17)	2370	2980	31799	25	2379	2991	31797	25	74368	3111	3715	38051	12	44891	133	119394
18)	2139	2542	33201	24	2146	2551	33257	24	75887	3053	3424	43130	12	49621	151	125660
19)	1959	2490	27634	18	1968	2504	27671	18	64265	2382	2923	30924	10	36241	110	100617
20)	1636	1970	25718	14	1642	1984	25801	14	58784	2111	2291	29484	9	33896	96	92777
21)	1185	1408	17402	12	1188	1410	17394	13	40015	2108	2103	21844	7	26064	95	66175
22)	1081	1395	16290	13	1085	1403	16318	13	37601	1460	1681	18889	5	22036	60	59698
X)F	3216	3923	40882	33	3203	3939	40994	33	96227	4550	4924	51288	12	60776	161	157165
X)M	7830	8783	91898	115	-	-	-	-	108628	-	-	-	-	-	-	108628
Y)	16	17	757	1	-	-	-	-	792	-	-	-	-	-	-	792

Table. 5. Total Number of Variations Existing in the Whole-Genome of a Person

Gender	Diploid									Haploid			
	Only One Allele				Both Alleles				Different	Ins	Del	Sub	Str
	Ins	Del	Sub	Str	Ins	Del	Sub	Str					
Female	160625	196793	2414512	1770	109741	123509	1500829	477	5217	-	-	-	-
Male	154206	188931	2332636	1704	105191	118585	1449541	465	5056	7846	8800	92655	116

Based on the method we developed, Table-6 shows how much space each record takes up. A record consists of multiple fields. In order to find how much space any record takes up, we should add the cost of each field constituting the record. These fields and their costs can be seen from Table-1 and Table-2. When we look at Table-6, we see that the costs of the fields “Del” are constant. But, the costs of the other fields except the fields “Del” are variable. Since we should store the sequence in insertion and substitution, their costs are determined according to the length of the sequence. For the “Different” field of Table-6, total cost differs according to the type of each variation and according to the length of the sequence if the variation is an insertion or a substitution.

Table. 6. Costs of the each record

Diploid					Haploid			
I_Byte != 00000011 (Only One Allele or Both Alleles)				I_Byte == 00000011 (Different)	Ins or Sub		Del	
Ins or Sub		Del						
Length < 15	Length = 15	Length < 15	Length = 15		Length < 15	Length = 15	Length < 15	Length = 15
V_Pos + I_Byte + L_T + Alt	V_Pos + I_Byte + L_T + Len + Alt	V_Pos + I_Byte + L_T	V_Pos + I_Byte + L_T + Len	In this case there are many options. Total Cost = V_Pos + I_Byte + cost of two different variations Total Cost = 5 + cost of two different variations	V_Pos + L_T + Alt	V_Pos + L_T + Len + Alt	V_Pos + L_T	V_Pos + L_T + Len
4 + 1 + 1 + Ceiling(Length/2)	4 + 1 + 1 + 2 + Ceiling(Length/2)	4 + 1 + 1	4 + 1 + 1 + 2		4 + 1 + Ceiling(Length/2)	4 + 1 + 2 + Ceiling(Length/2)	4 + 1	4 + 1 + 2
6 + Ceiling(Length/2)	8 + Ceiling(Length/2)	6	8		5 + Ceiling(Length/2)	7 + Ceiling(Length/2)	5	7

Space Requirement table (Table-7) is the table that the calculation is made for the space required to store all the variations that exist in the genome of any female or male. In order to fill the Space Requirement table, the values in Table-3, Table-5 and Table-6 were used. The values of the corresponding fields of Table-5 and Table-6 were multiplied firstly, and then all the values obtained were added. For the "Different" field of the Table-7, namely for the case where any person has different variations on both alleles of a particular chromosome, two different insertions whose lengths are approximately 3 were selected. Since the average lengths of insertions for each of the chromosomes (except chromosome 22 and chromosome Y) shown in Table-3 are between 2 and 3 (closer to 3), the result of the expression Ceiling(Length/2) is found as 2 for insertion. Similarly, since the average lengths of substitutions for each of the chromosomes shown in Table-3 are very closer to 1, the result of the expression Ceiling(Length/2) is found as 1 for substitution. When we look at this table, we see that approximately 30,08 MB and 29,66 MB are required to store all the variations which exist in the genome of any female or male respectively.

Table. 7. Space Requirement Table

Gender	Diploid							Haploid		
	Only One Allele			Both Alleles			Different	Ins	Del	Sub
	Ins	Del	Sub	Ins	Del	Sub				
Female	160625 * (6 + Ceiling(Length/2))	196793 * 6	2414512 * (6 + Ceiling(Length/2))	109741 * (6 + Ceiling(Length/2))	123509 * 6	1500829 * (6 + Ceiling(Length/2))	5217 * (5 + 3 + 3)	-	-	-
	160625 * 8	196793 * 6	2414512 * 7	109741 * 8	123509 * 6	1500829 * 7	5217 * 11			
	Total = 31549514 Bytes = 30810,07227 KB = 30,0879612 MB = 0,029382775 GB									
Male	154206 * (6 + Ceiling(Length/2))	188931 * 6	2332636 * (6 + Ceiling(Length/2))	105191 * (6 + Ceiling(Length/2))	118585 * 6	1449541 * (6 + Ceiling(Length/2))	5056 * (5 + 3 + 3)	7846 * 5 + Ceiling(Length/2)	8800 * 5	92655 * 5 + Ceiling(Length/2)
	154206 * 8	188931 * 6	2332636 * 7	105191 * 8	118585 * 6	1449541 * 7	5056 * 11	7846 * 7	8800 * 5	92655 * 6
	Total = 31105979 Bytes = 30376,93262 KB = 29,66497326 MB = 0,0289697 GB									

4. Conclusion

By using our proposed method, the space required to store all the variations except for the structural variations in the genome of a person is approximately 0.5 % of the space required to store the raw sequence of this person. This means that, on average, our method provides a space saving of approximately 99.5%. In addition, our method yielded better results than the accomplished compression methods. The space need of our method is 0,03 GB as opposed to 1,2 GB obtained by the accomplished compression methods. The format we designed can be converted into a complete file format with some minor additions. In this way, the deficiencies of the Vcf file format can be removed.

Relational databases have been used successfully in many areas so far. But unfortunately, there are several challenges confronted when using relational databases for storing personal genetic data. In order to store all the variations existing in the genome of a person in relational database, millions of rows are required. In practice this is almost impossible. However, based on the general format we developed, variations of each chromosome can be stored in type of "varbinary (MAX)". In this way, only 23 rows would be used for each person.

References

1000 Genomes Project Consortium (2010). A Map of Human Genome Variation from Population-Scale Sequencing. *Nature*, 1061-1073.

Alberts, B., et al. (2007). *Molecular Biology of the Cell*. New York, USA: Garland Science.

Bafna, V., et al. (2013). Abstractions for Genomics. *Communications of the ACM, 56*(1), 83-93.

Behzadi, B., & Fessant, F.L. (2005). DNA Compression Challenge Revisited: A Dynamic Programming Approach. In *CPM* (pp. 190-200). Springer.

Cao, M.D., et al. (2007). A Simple Statistical Algorithm for Biological Sequence Compression. *Proceedings of the IEEE Data Compression Conference*, 43–52.

Chen, X., et al. (2002). Dnacompress: Fast and Effective DNA Sequence Compression. *Bioinformatics, 18*(12), 1696-1698.

Grümbach, S., & Tahi, F. (1993). Compression of DNAsequences. *Proceedings of the IEEE Data Compression Conference*, 340–350.

Hammer, J., & Schneider, M. (2003). Genomics Algebra: A New, Integrating Data Model, Language, and Tool for Processing and Querying Genomic Information. *Proceedings of the 2003 CIDR Conference*, 5–8.

Herman, D., Peternel, P., Stegnar, M., Breskvar, K., & Dolzan, V. (2006). The Influence of Sequence Variations in Factor VII, Gammaglutamyl Carboxylase and Vitamin K Epoxide Reductase Complex Genes on Warfarin Dose Requirement, *Thromb Haemost, 95*, 782-787.

Levy, S., et al. (2007). The Diploid Genome Sequence of an Individual Human. *PLoS Biology, 5*(10), e254.

Lindh, J.D., Lundgren, S., Holm, L., Alfredsson, L., & Rane, A. (2005). Several-Fold Increase in Risk of Overanticoagulation by CYP2C9 Mutations. *Clin. Pharmacol. Ther.*, *78*, 540-550.

Millican, E.A., et al. (2007). Genetic-Based Dosing in Orthopedic Patients Beginning Warfarin Therapy. *Blood*, *110*, 1511-1515.

Schwarz, U.I., et al. (2008). Genetic Determinants of Response to Warfarin during Initial Anticoagulation. *New England Journal of Medicine*, *358*(10), 999-1008.

The International Human Genome Sequencing Consortium (2004). Finishing the Euchromatic Sequence of the Human Genome. *Nature*, *431*, 931–945.

Veenstra, D.L., et al. (2005). CYP2C9 Haplotype Structure in European American Warfarin Patients and Association with Clinical Outcomes. *Clin. Pharmacol. Ther.*, *77*, 353-364.

Wheeler, D.A., et al. (2008). The Complete Genome of an Individual by Massively Parallel DNA Sequencing. *Nature*, *452*(7189), 872-876.

Web sites:

Web-1: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ consulted 20 Mar. 2015.

Web-2: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/bcf_files, consulted 20 Mar. 2015.

Web-3: https://github.com/samtools/hts-specs, consulted 20 Mar. 2015.

Web-4: http://jul2012.archive.ensembl.org/Homo_sapiens/Location/Chromosome?r=1:1-1000000, consulted 20 Mar. 2015.