Supplementary Material

Z. Gáspári, C. Ortutay and G. Tóth:
Divergent microsatellite evolution in the human and chimpanzee lineages.
FEBS Letters 581(13): 2523-2526 (2007)

Supplementary Methods

Sequence data and extraction of SSRs

Genomic sequences for the 24 (1-22, X, Y) human and chimpanzee chromosomes were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens, Build 34 version 3; ftp://ftp.ncbi.nih.gov/genomes/P_troglodytes, Build 1 version 1; total lengths of the analyzed sequences are shown at our web site, http://bioinformatics.abc.hu/ssr). The sequence annotation provided by NCBI was used in all analyses.

Perfect microsatellites with a minimum total length of 12 base pairs (bp) and a repeat unit length of maximum 6 bp (criteria similar to those in [1]) were extracted using an in-house written Perl program. Imperfect microsatellites were identified by the program Tandem Repeats Finder (TRF) [2] using the following parameters: 2, 3, 7 for match score, mismatch and indel penalty, 80 and 10 for match and indel percentage, 500 for maximum period size, respectively, and a minimum allowed score of 24 in order to include 12-bp perfect repeats in the output. It is important to note that the term "imperfect repeat" is used here to denote repeats that may, but do not necessarily, contain insertions, deletions and/or mismatches relative to their perfect counterpart. Thus, perfect repeats constitute a subset of the imperfect repeats defined this way. Partial repeat units from the TRF results were trimmed in order to ensure that the number of units is always an integer (INT repeats, Supplemetary Scheme 1). This adjustment allowed a straightforward comparison of imperfect to perfect repeats with partial units not considered as part of the repeat.

Gttctagt**acgacgacgacgaccacgac**tct	TRF
Gttctagt**acgacgacgacgaccacg**actct	INT
Gttctagt**acgacgacgacgacg**accacgactct	PERF

Supplementary Scheme 1. Different interpretations of the same DNA sequence as an SSR. Tandem Repeats Finder allows mismatches and insertions/deletions (indels) with respect to the repeat unit (**acg**) and recognizes partial units (TRF). These partial units are removed to yield repeats with an integer number of units (INT). No mismatches, indels or partial units are allowed in perfect repeats (PERF).

Classification of SSRs

Both perfect and imperfect SSRs were categorized on the basis of their repeat unit, and also according to their sequence location, as exonic, intronic or intergenic repeats. Repeat units were classified as described before [1,2,3]. The class 'ACG', for example, contains all permutants and/or reverse complements of the trinucleotide ACG, i.e. ACG=CGA=GAC=CGT=GTC=TCG, and is named after the triplet that comes first in alphabetical order. In each case, the shortest possible repeat unit was considered.

Classification of repeats by sequence location was based on the CDS fields in the annotation using simple hierarchical rules. SSRs were considered exonic if they fell within the boundaries of a protein-coding exon, regardless of all other matching annotated features. Intronic SSRs were identifed by their location within the boundaries of an annotated intron but not of any exons. This scheme allows for the consideration of alternative splicing, as an SSR in a translated region is classified as exonic while an SSR between the first and last exons of a transcript that has never been translated is considered intronic. SSRs located in all other noncoding regions were considered intergenic, while repeats crossing exon–intron and coding–noncoding sequence boundaries were not categorized.

Comparison of perfect and imperfect SSRs

To identify imperfect repeats corresponding to perfect ones (i.e. to select perfect and imperfect repeats at identical loci), all imperfect repeats starting before and ending after each perfect one were selected. To allow for straightforward comparison, only imperfect–perfect pairs with identical repeat class and region (i.e. exonic, intronic and intergenic segments) were retained, this being necessary to avoid discrepancies caused by repeats spanning multiple regions. Total number and cumulative length of repeats per megabase are reported. These data can be found on our supplementary website (http://bioinformatics.abc.hu/ssr/).

Comparison of repeat distributions

To compare repeats in different regions and chromosomes, and also to assess the differences regarding perfect/imperfect repeats, we applied the chi-square contingency analysis that tests the equality of two observed distributions. Input data for the calculations corresponded to those presented in the columns of the online tables at the SSRDB web site. All repeat types contributing less than 5% to the total value were omitted by removing the

smallest data points iteratively until all the remaining ones constituted more than 5% of the total. Although this procedure led to unused data points that might carry relevant information, the statistical analysis can be considered more reliable if applied this way [4]. To avoid any bias imposed by the different total abundance of perfect and imperfect repeats, all data were normalized before analysis. Contingency analysis was applied to the number per megabase and cumulative length per megabase measures as reported on our supplementary website (http://bioinformatics.abc.hu/ssr/).

Calculation of contingency values was done as follows:

Given two proportions of *m* categories (*Obs(1,1)*, *Obs(2,1)*, ..., *Obs(m,1)*, *Obs(1,2)*, *Obs(2,2)*, ..., *Obs(m,2)*), expected values are first calculated according to:

$$Exp(i,j) = \frac{Obs(i,x)Obs(x,j)}{Obs(xx)} \qquad \text{Eq.1}$$

where *Obs(i,x)* and *Obs(x,j)* are the corresponding row and column sums and *Obs(xx)* is the total sum of all observations (here, the normalized length/megabase data are used, thus *Obs(x,j)*=100 and *Obs(xx)*=200). The chi-square value (χ^2) is then computed via:

$$\chi^2 = \sum_{j=1}^{2}\sum_{i=1}^{m}\frac{[Obs(i,j) - Exp(i,j)]^2}{Exp(i,j)} \qquad \text{Eq. 2}$$

Finally, the chi-square value is converted to probability (*P*) by using *m*-1 as the degrees of freedom. The "null hypothesis" (i.e. the observed proportions are identical) is normally rejected if $P \leq 0.05$. However, prior normalization (conversion of the observed counts to percentages) renders the test very conservative where probability values below 0.5 give a strong indication that the two compared proportions differ.

We note that we do not interpret this test as a conventional statistical test but only as a measure of similarity between the two distributions compared. Thus numbers close to 1 indicate close similarity and lower values, primarily below 0.5, are interpreted as discrepancies between the distributions compared. We note that statistically significant differences are normally not expected when comparing perfect and imperfect repeats at the same loci since identical genomic sequences are used as the data source (see Schemes 1 and S1). Our approach is similar to the PRIDE method used for protein structure comparison [5].

Supplementary Results

Detailed results are available on our supplementary web site (http://bioinformatics.abc.hu/ssr/).

Trinucleotide SSRs on chromosomes 19, 21 and Y

We chose chromosomes 19, 21 and Y for detailed analysis. Chromosome 19 shows remarkable divergence in the perfect/imperfect repeat distributions in both species (Table 1), whereas chromosomes 21 and Y have the largest interspecies differences with respect to trinucleotide repeats (Table 2). Detailed comparison of a number of sequence features of the human and chimpanzee chromosome counterparts are already available for chromosomes 21 and Y [6,7]. Hereafter, the abbreviations HSA# and PTR# refer to the human and chimpanzee chromosomes, respectively, where # is replaced by the chromosome identifier, according to the standard numbering.

On HSA19 the discrepancy between perfect and imperfect repeats is clearly apparent (Supplementary Table 1a). For example, perfect AGG repeats are less than half as abundant as AAT repeats, but the two repeat classes have about the same overall length per megabase when their imperfect variants are taken into consideration. The trends are similar for PTR19 although the imperfect/perfect ratios of these repeats are different from those in humans (Table 2). The absolute abundance data in coding regions reveal a more marked difference, even though the distribution profiles (repeat class preferences) are similar. There are almost twice as many $(AGC)_n$, $(AGG)_n$, $(ATC)_n$ and $(CCG)_n$ sequences on HSA19 as on PTR19 (Supplementary Table 1a). The difference between CCG repeat content in introns and intergenic regions is detectable in both species, irrespective of whether perfect or imperfect repeats are considered. This is consistent with our earlier results [1]. HSA19 is known to have high interspersed repeat content and also exceptionally high gene density [8], although these features cannot be unambiguously linked to the observed SSR patterns, since they are not associated with it on other chromosomes (see below).

Chromosome 21 exhibits a trinucleotide repeat pattern distinctly different from that of chromosome 19, which is best seen in the relative scarcity of CCG repeats in the former (Supplementary Table 1b). On HSA21, there is considerable discrepancy between perfect and imperfect repeats in the coding regions and introns (Table 1), which can be explained by the relatively high abundance of AAG repeats in the former and ACT repeats in the latter regions (Supplementary Table 1b). In noncoding sequences ACC repeats are mainly responsible for

interspecies differences in imperfect repeat content, whereas in coding regions the abundance of AAG and AGG repeats is high.

The Y chromosome has a unique evolutionary history [9,10], which is apparent, for example, from the existence of a human-specific region compared to PTRY [9]. We show here that these evolutionary differences are also reflected in the SSR content of HSAY and PTRY. In coding regions most repeat classes are not represented in the current genome assemblies, e.g. AAG and ACC repeats, which are common on other chromosomes, are missing (Supplementary Table 1c). The similarity between perfect and imperfect repeat distributions is low in the introns of both species (Table 1). As reflected in the low interspecies (*Homo–Pan*) contingency probability (Table 2), the repeat class contributing most to the large discrepancy between the perfect and imperfect repeat distributions differs between the human and chimpanzee genomes: it is AGG for the former and AAG for the latter (Supplementary Table 1c). These apparently similar repeats differ in their G+C content which might account for the observed interspecies variation. It has been observed earlier that the total number of simple repeats and low complexity regions is higher on HSAY than on PTRY, but the opposite trend is true for the alignable regions [9]. Our results indicate higher overall SSR density on HSAY, especially when considering imperfect repeats (Supplementary Table 1c).

As was speculated earlier [1], the differences in SSR patterns cannot be explained solely on the basis of the intrinsic properties of DNA sequence stretches like hairpin-forming potential. The decisive role of G+C content alone can also be ruled out, considering the markedly different abundance of repeats of identical G+C percentage (ACC, ACG, AGC and AGG) in the exonic, intronic and intergenic regions. Therefore other factors, such as repair mechanisms and overall evolutionary rate of specific regions, should be considered.

Supplementary Table 1

Abundance (cumulative length[a] per megabase) of corresponding perfect and imperfect trinucleotide repeats in different regions of the human and chimpanzee chromosomes 19 (*a)*, 21 (*b*) and Y (*c*).

Chromosome 19

a)	**ALL**		**CODING**		**INTRON**		**INTERGENIC**	
	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect
	Homo sapiens							
aac	223.88	*350.26*	5.60	*5.60*	222.44	*333.89*	205.54	*330.78*
aag	84.67	*234.19*	155.37	*302.34*	67.48	*162.45*	78.56	*238.61*
aat	438.58	*635.54*	0.00	*0.00*	459.60	*648.73*	378.02	*564.21*
acc	65.98	*124.90*	103.58	*176.83*	85.67	*169.98*	38.97	*64.12*
acg	0.54	*0.70*	7.00	*9.80*	0.00	*0.00*	0.00	*0.00*
act	4.67	*6.57*	0.00	*0.00*	7.58	*10.77*	2.47	*3.74*
agc	61.85	*84.36*	573.88	*874.81*	39.73	*47.06*	28.54	*36.64*
agg	179.37	*686.42*	666.26	*1567.20*	176.04	*666.83*	113.19	*566.74*
atc	63.08	*288.21*	83.98	*142.77*	46.10	*123.23*	46.30	*190.46*
ccg	67.86	*259.51*	487.10	*1898.93*	22.44	*78.29*	43.83	*190.63*
	Pan troglodytes							
aac	190.63	*288.00*	6.03	*7.54*	192.62	*282.09*	167.44	*261.33*
aag	61.74	*156.96*	120.67	*226.26*	48.03	*124.19*	58.62	*149.59*
aat	340.29	*489.26*	6.03	*6.03*	341.43	*489.80*	312.03	*453.53*
acc	44.84	*69.66*	101.06	*130.23*	43.93	*73.22*	36.02	*55.67*
acg	0.43	*0.62*	7.54	*10.56*	0.00	*0.00*	0.00	*0.00*
act	3.06	*4.46*	7.54	*7.54*	3.86	*6.64*	1.92	*2.45*
agc	45.99	*62.92*	381.63	*552.58*	34.88	*49.60*	26.93	*33.58*
agg	114.08	*279.27*	386.15	*860.80*	104.52	*233.45*	90.72	*236.20*
atc	35.77	*84.38*	30.17	*43.74*	35.48	*78.45*	31.77	*84.58*
ccg	38.54	*124.14*	276.04	*898.51*	22.93	*70.00*	30.60	*103.43*

Chromosome 21

b)	**ALL**		**CODING**		**INTRON**		**INTERGENIC**	
	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect
	Homo sapiens							
aac	174.39	*269.51*	0.00	*0.00*	163.67	*244.18*	163.94	*257.22*
aag	52.97	*106.80*	215.97	*607.61*	41.48	*60.73*	51.50	*112.56*
aat	262.11	*387.45*	0.00	*0.00*	223.65	*318.51*	258.25	*388.19*
acc	71.54	*179.46*	241.89	*397.39*	52.64	*222.16*	72.10	*152.89*
acg	0.00	*0.00*	0.00	*0.00*	0.00	*0.00*	0.00	*0.00*
act	11.23	*40.18*	0.00	*0.00*	12.44	*101.46*	9.80	*15.43*
agc	46.07	*65.62*	587.45	*1529.09*	41.16	*50.52*	34.13	*40.95*
agg	69.77	*160.56*	276.45	*437.71*	68.91	*135.38*	60.31	*154.71*
atc	38.65	*57.69*	112.31	*120.94*	44.98	*76.04*	32.39	*45.67*
ccg	12.65	*39.47*	103.67	*152.62*	5.11	*22.01*	6.33	*29.74*
	Pan troglodytes							
aac	175.02	*266.80*	0.00	*0.00*	173.38	*268.49*	166.97	*255.57*
aag	50.40	*113.26*	34.15	*34.15*	38.42	*65.81*	53.34	*127.65*
aat	227.11	*346.65*	0.00	*0.00*	201.45	*284.86*	223.62	*349.33*
acc	37.70	*60.36*	213.44	*298.81*	40.09	*62.14*	33.74	*55.77*
acg	0.00	*0.00*	0.00	*0.00*	0.00	*0.00*	0.00	*0.00*
act	8.41	*21.40*	0.00	*0.00*	1.34	*1.34*	10.03	*25.83*
agc	37.11	*47.37*	418.34	*867.98*	35.75	*44.43*	31.46	*36.40*
agg	51.58	*124.03*	76.84	*179.29*	63.47	*115.37*	43.54	*121.50*
atc	32.48	*48.66*	34.15	*34.15*	29.40	*45.66*	31.11	*46.46*
ccg	11.02	*26.03*	179.29	*412.65*	10.02	*25.5*	7.41	*18.43*

Y chromosome

c)	**ALL**		**CODING**		**INTRON**		**INTERGENIC**	
	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect
	Homo sapiens							
aac	201.86	*311.92*	0.00	*0.00*	257.31	*385.96*	181.65	*282.21*
aag	138.57	*314.49*	0.00	*0.00*	47.40	*127.75*	133.78	*309.18*
aat	351.92	*542.86*	0.00	*0.00*	243.76	*377.84*	325.00	*498.51*
acc	15.52	*34.17*	0.00	*0.00*	10.83	*12.19*	13.89	*33.67*
acg	0.49	*0.49*	0.00	*0.00*	0.00	*0.00*	0.54	*0.54*
act	16.01	*30.39*	0.00	*0.00*	6.77	*8.12*	15.91	*31.24*
agc	23.83	*31.73*	534.43	*1102.25*	21.67	*26.63*	18.07	*23.11*
agg	86.64	*340.55*	400.82	*501.02*	173.34	*958.81*	63.92	*230.02*
atc	31.04	*202.96*	133.61	*167.01*	10.83	*136.78*	31.56	*208.62*
ccg	3.54	*9.12*	133.61	*367.42*	5.42	*5.42*	1.08	*4.86*
	Pan troglodytes							
aac	211.86	*336.35*	0.00	*0.00*	173.40	*249.85*	214.61	*341.25*
aag	166.50	*304.03*	0.00	*0.00*	47.55	*402.74*	180.15	*294.81*
aat	242.28	*381.89*	0.00	*0.00*	128.65	*199.50*	248.46	*395.56*
acc	22.27	*31.24*	0.00	*0.00*	0.00	*0.00*	22.97	*32.04*
acg	0.00	*0.00*	0.00	*0.00*	0.00	*0.00*	0.00	*0.00*
act	9.78	*18.11*	0.00	*0.00*	11.19	*16.78*	9.67	*18.34*
agc	19.29	*26.26*	254.27	*572.11*	11.19	*11.19*	17.83	*24.08*
agg	42.37	*80.13*	254.27	*317.84*	50.34	*122.13*	39.29	*70.53*
atc	11.95	*49.52*	0.00	*0.00*	22.37	*22.37*	10.88	*52.69*
ccg	1.90	*4.25*	0.00	*0.00*	0.00	*0.00*	0.00	*0.00*

[a] base pairs

Supplementary references

[1] Tóth, G., Gáspári, Z. and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10, 967-981.

[2] Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573-580.

[3] Jurka, J. and Pethiyagoda, C. (1995). Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* 40, 120-126.

[4] Townend, J. (2004). Practical Statistics for Environmental and Biological Scientists, John Wiley & Sons Ltd, Chichester, England.

[5] Carugo, O. and Pongor, S. (2002). Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J. Mol. Biol.* 315, 887-898.

[6] The International Chimpanzee Chromosome 22 Consortium (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429, 382-388.

[7] Kuroki, Y., Toyoda, A., Noguchi, H., Taylor, T.D., Itoh, T., Kim, D.S., Kim, D.W., Choi, S.H., Kim, I.C., Choi, H.H., Kim, Y.S., Satta, Y., Saitou, N., Yamada, T., Morishita, S., Hattori, M., Sakaki, Y., Park, H.S. and Fujiyama, A. (2006). Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat. Genet.* 38, 158-167.

[8] Grimwood, J., Gordon, L.A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., Aerts, A., Altherr, M., Ashworth, L., Bajorek, E., Black, S., Branscomb, E., Caenepeel, S., Carrano, A., Caoile, C., Chan, Y.M., Christensen, M., Cleland, C.A., Copeland, A., Dalin, E., Dehal, P., Denys, M., Detter, J.C., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Georgescu, A.M., Glavina, T., Gomez, M., Gonzales, E., Groza, M., Hammon, N., Hawkins, T., Haydu, L., Ho, I., Huang, W., Israni, S., Jett, J., Kadner, K., Kimball, H., Kobayashi, A., Larionov, V., Leem, S.H., Lopez, F., Lou, Y., Lowry, S., Malfatti, S., Martinez, D., McCready, P., Medina, C., Morgan, J., Nelson, K., Nolan, M., Ovcharenko, I., Pitluck, S., Pollard, M., Popkie, A.P., Predki, P., Quan, G., Ramirez, L., Rash, S., Retterer, J., Rodriguez, A., Rogers, S., Salamov, A., Salazar, A., She, X., Smith, D., Slezak, T., Solovyev, V., Thayer, N., Tice, H., Tsai, M., Ustaszewska, A,. Vo, N., Wagner, M., Wheeler, J., Wu, K., Xie, G., Yang, J., Dubchak, I., Furey, T.S., DeJong, P., Dickson, M., Gordon, D., Eichler, E.E., Pennacchio, L.A., Richardson, P., Stubbs, L., Rokhsar, D.S., Myers, R,M,, Rubin, E.M. and Lucas, S.M. (2004). The DNA sequence and biology of human chromosome 19. *Nature* 428, 529-535.

[9] Rozen, S., Skaletsky, H., Marszalek, J.D., Minx, P.J., Cordum, H.S., Waterston, R.H., Wilson, R.K. and Page, D.C. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423, 873-876.

[10] Skaletsky, H., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825-837.

Full author lists of references 10, 19, 20, 22 and 24 in the manuscript

[10] Kuroki, Y., Toyoda, A., Noguchi, H., Taylor, T.D., Itoh, T., Kim, D.S., Kim, D.W., Choi, S.H., Kim, I.C., Choi, H.H., Kim, Y.S., Satta, Y., Saitou, N., Yamada, T., Morishita, S., Hattori, M., Sakaki, Y., Park, H.S. and Fujiyama, A. (2006). Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat. Genet.* 38, 158-167.

[19] Martin, J., Han, C., Gordon, L.A., Terry, A., Prabhakar, S., She, X., Xie, G., Hellsten, U., Chan, Y.M., Altherr, M., Couronne, O., Aerts, A., Bajorek, E., Black, S., Blumer, H., Branscomb, E., Brown, N.C., Bruno, W.J,, Buckingham, J.M., Callen, D.F., Campbell, C.S., Campbell, M.L., Campbell, E.W., Caoile, C., Challacombe, J.F., Chasteen, L.A., Chertkov, O., Chi, H.C., Christensen, M., Clark, L.M., Cohn, J.D., Denys, M., Detter. J.C., Dickson, M., Dimitrijevic-Bussod, M., Escobar, J., Fawcett, J.J., Flowers, D., Fotopulos, D., Glavina, T., Gomez, M., Gonzales, E., Goodstein, D., Goodwin, L.A., Grady, D.L., Grigoriev, I., Groza, M., Hammon, N., Hawkins, T., Haydu, L., Hildebrand, C.E., Huang, W., Israni, S., Jett, J., Jewett, P.B., Kadner, K., Kimball, H., Kobayashi, A., Krawczyk, M.C., Leyba, T., Longmire, J.L., Lopez, F., Lou, Y., Lowry, S., Ludeman, T., Manohar, C.F., Mark, G.A., McMurray, K.L., Meincke, L.J., Morgan, J., Moyzis, R.K., Mundt, M.O., Munk A.C., Nandkeshwar, R.D., Pitluck, S., Pollard, M., Predki, P., Parson-Quintana, B., Ramirez, L., Rash, S., Retterer, J., Ricke, D.O., Robinson, D.L., Rodriguez, A., Salamov, A., Saunders, E.H., Scott, D., Shough, T., Stallings, R.L., Stalvey, M., Sutherland, R.D., Tapia, R., Tesmer, J.G., Thayer, N., Thompson, L.S., Tice, H., Torney, D.C., Tran-Gyamfi, M., Tsai, M., Ulanovsky, L.E., Ustaszewska, A., Vo, N., White, P.S., Williams, A.L., Wills, P.L., Wu, J.R., Wu, K., Yang, J., Dejong, P., Bruce, D., Doggett, N.A., Deaven, L., Schmutz, J., Grimwood, J., Richardson, P., Rokhsar, D.S., Eichler, E.E., Gilna, P., Lucas, S.M., Myers, R.M., Rubin, E.M. and Pennacchio, L.A. (2004). The sequence and analysis of duplication-rich human chromosome 16. *Nature* 432, 988-994.

[20] Grimwood, J., Gordon, L.A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., Aerts, A., Altherr, M., Ashworth, L., Bajorek, E., Black, S., Branscomb, E., Caenepeel, S., Carrano, A., Caoile, C., Chan, Y.M., Christensen, M., Cleland, C.A., Copeland, A., Dalin, E., Dehal, P., Denys, M., Detter, J.C., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Georgescu, A.M., Glavina, T., Gomez, M., Gonzales, E., Groza, M., Hammon, N., Hawkins, T., Haydu, L., Ho, I., Huang, W., Israni, S., Jett, J., Kadner, K., Kimball, H., Kobayashi, A., Larionov, V., Leem, S.H., Lopez, F., Lou, Y., Lowry, S., Malfatti, S., Martinez, D., McCready, P., Medina, C., Morgan, J., Nelson, K., Nolan, M., Ovcharenko, I., Pitluck, S., Pollard, M., Popkie, A.P., Predki, P., Quan, G., Ramirez, L., Rash, S., Retterer, J., Rodriguez, A., Rogers, S., Salamov, A., Salazar, A., She, X., Smith, D., Slezak, T., Solovyev, V., Thayer, N., Tice, H., Tsai, M., Ustaszewska, A,. Vo, N., Wagner, M., Wheeler, J., Wu, K., Xie, G., Yang, J., Dubchak, I., Furey, T.S., DeJong, P., Dickson, M., Gordon, D., Eichler, E.E., Pennacchio, L.A., Richardson, P., Stubbs, L., Rokhsar, D.S., Myers, R,M,, Rubin, E.M. and Lucas, S.M. (2004). The DNA sequence and biology of human chromosome 19. *Nature* 428, 529-535.

[22] Dunham, I., Hunt, A. R., Collins, J.E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K.N., Beasley, O., Bird, C.P., Blakey, S., Bridgeman, A.M., Buck, D., Burgess, J., Burrill, W.D., Burton, J., Carder, C., Carter, N.P., Chen, Y., Clark, G., Clegg, S.M., Cobley, V., Cole, C. G., Collier, R.E., Connor, R.E., Conroy, D., Corby, N., Coville, G.J., Cox, A.V., Davis, J., Dawson, E., Dhami P.D., Dockree, C., Dodsworth, S.J., Durbin, R.M., Ellington, A., Evans, K.L., Fey, J.M., Fleming, K., French, L., Garner, A.A., Gilbert, J.G.R., Goward, M.E., Grafham, D., Griffiths, M.N., Hall, C., Hall, R., Hall, G., Hall-Tamlyn, G., Heathcott, R.W., Ho, S., Holmes, S., Hunt, S.E., Jones, M.C., Kershaw, J. Kimberley, A., King, A., Laird, G.K., Langford, C.F., Leversha, M.A., Lloyd, C., Lloyd, D.M., Martyn, I.D., Mashreghi-Mohammadi, M., Matthews, L., McCann, O.T., McClay, J., McLaren, S., McMurray, A.A., Milne, S.A., Mortimore, B.J., Odell, C.N., Pavitt, R., Pearce, A.V., Pearson, D., Phillimore, B.J., Phillips, S. H., Plumb, R.W., Ramsay, H., Ramsey, Y., Rogers, L., Ross, M.T., Scott, C.E., Sehra, H.K., Skuce, C.D.,

Smalley, S., Smith, M.L., Soderlund, C., Spragon, L., Steward, C.A., Sulston, J.E., Swann, R.M., Vaudin, M., Wall, M., Wallis, J.M., Whiteley, M.N., Willey, D., Williams, L., Williams, S., Williamson, H., Wilmer, T.E., Wilming, L.C., Wright, L., Hubbard, T., Bentley1, D.R., Beck, S., Rogers, J., Shimizu, N., Minoshima, S., Kawasaki, K., Sasaki, T., Asakawa, S., Kudoh, J,. Shintani, A., Shibuya, K., Yoshizaki, Y., Aoki, N., Mitsuyama, S., Roe, B.A., Chen3, F., Chu, L., Crabtree, J., Deschamps, S., Do, A., Do, T., Dorman, A., Fang, F., Fu, Y., Hu, P., Hua, A., Kenton, S., Lai, H., Lao, H.I., Lewis, J., Lewis, S., Lin, S.-P., Loh, P., Malaj, E., Nguyen, T., Pan, H., Phan, S., Qi, S., Qian, Y., Ray, L., Ren, Q., Shaull, S., Sloan, D., Song, L., Wang, Q., Wang, Y., Wang, Z., White, J., Willingham, D., Wu, H., Yao, Z., Zhan, M., Zhang, G. Chissoe, S., Murray, J., Miller, N., Minx, P., Fulton, R., Johnson, D., Bemis, G., Bentley, D., Bradshaw, H., Bourne, S., Cordes, M., Du, Z., Fulton, L., Goela, D., Graves, T., Hawkins, J., Hinds, K., Kemp, K., Latreille, P., Layman, D., Ozersky, P., Rohlfing, T., Scheet, P., Walker, C., Wamsley, A., Wohldmann, P., Pepin, K., Nelson, J., Korf, I., Bedell, J.A., Hillier, L., Mardis, E., Waterston, R., Wilson, R., Emanuel, B.S., Shaikh, T., Kurahashi, H., Saitta, S., Budarf, M.L., McDermid, H.E., Johnson, A., Wong, A.C.C., Morrow, B.E., Edelmann, L., Kim, U.J., Shizuya, H., Simon, M.I., Dumanski, J.P., Peyrard, M., Kedra, D., Seroussi, E., Fransson, I., Tapia, I., Bruder, C.E. and O'Brien, K.P. (1999). The DNA sequence of human chromosome 22. *Nature* 402, 489-495.

[24] Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S.F., Latrielle, P., Leonard, S., Mardis, E., Maupin, R., McPherson, J., Miner, T., Nash, W., Nguyen, C., Ozersky, P., Pepin, K., Rock, S., Rohlfing, T., Scott ,K., Schultz, B., Strong, C., Tin-Wollam, A., Yang, S.P., Waterston, R.H., Wilson, R.K., Rozen, S. and Page, D.C. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825-837.