Current Research Journal of Biological Sciences 2(5): 328-332, 2010
ISSN: 2041-0778

Submitted Date: December 23, 2009 Accepted Date: January 06, 2010 Published Date: September 15, 2010

Comparative Genomic Analysis and Evolutionary Study of CD4 Antigen

[1]Liaqat Ali, [1]Muhammad Ilyas, [2]Syed Saleem Shah, [1]Muhammad Ali,
[1]Syed Muhammad Ibrahim, [1]Shakeel Ahmad and [3]Inamullah
[1]National Centre of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan
[2]Department of Pharmacy, University of Balochistan, Quetta, Pakistan
[3]Department of Genetics, Hazara University Mansehra, Pakistan

Abstract: CD4 antigen is important to act as the primary receptor for the HIV. In order to understand the human CD4 gene and to determine its evolutionary aspects, we characterized this gene in detail in six different organisms. A comparative study was made of nucleotide length variations, intron and exon sizes and number variations, differential compositions of coding to non-coding bases, etc., to look for similarities/dissimilarities in the CD4 gene across all six taxa. Phylogenetic analysis showed the pattern found in other genes, as *Homo sapiens* and *Pan troglodytes* were placed in a single clade, and *Rattus norvegicus* and *Mus musculus* in another. We further focused on the two primates and aligned the amino acid sequences; there were small differences between humans and chimpanzees; both were more different from the other organisms.

Key words: CD4 antigen, CpG islands, genomics, *Homo sapiens*, *Pan troglodytes*, phylogentics

INTRODUCTION

Comparative analyses of genome sequences will be a major part to improve the health of individuals and society after the completion of Human Genome Project (Collins *et al*., 2003). It is the direct comparison of genomic information of one organism against that of another to gain a better understanding of how species evolved and to determine the function of genes and noncoding regions in genomes (Sivashankari and Shanmughavel, 2007). It is believed to be the important aspect of understanding the evolutionary relationship across different taxa. With the availability of genomic information of different organisms, it's now easy to compare them in different angles, which also help in finding the unknown genes and mutations in them that can affect the function. Such study can be done with looking at the homologous and conserved regions in the sequences of different taxa in addition to the gene length, number of introns exons, GC contents, CpG islands and such type of other information can be obtained, that can help in understanding the evolutionary effects on genes.

The CD4 antigen is a 55 kD membrane glycoprotein molecule in human blood present on the surface of 65% of human T cells, also known as T4 antigen. The most important property of the CD4 antigen is to act as the primary receptor for the AIDS virus where CD stands for "cluster of differentiation" (Clark *et al*., 1987). When the HIV virus attaches CD4 surface proteins, it diminishes the number of T cells, B cells, natural killer cells, and monocytes in host blood. The human immunodeficiency virus (HIV-1) infects T lymphocytes via an interaction between the virus envelope glycoprotein gp120 and the CD4 antigen of T helper cells (Schockmel *et al*., 1992). CD4 binds to relatively invariant sites on class II major histocompatibility complex (MHC) molecules outside the peptide-binding groove, which interacts with the T-Cell Receptor (TCR).

From the literature it was found that T4 RNA is expressed not only in T lymphocytes, but also in B cells, macrophages, and granulocytes (Maddon *et al*., 1987). It is also expressed in a developmentally regulated manner in specific regions of the brain, which makes CD4 gene very important as AIDS is concerned and demands detailed evolutionary genomic understanding. Therefore in this project, we used a comparative approach to investigate the importance of CD4 gene in comparison with in six taxa.

MATERIALS AND METHODS

The nucleotide sequences of CD4 antigen of six different taxa i.e., Homo sapiens (Human) = NM_000616.3, Pan troglodytes (Chimp) = NM_001009043.1, Canis familiaris (Dog) = XM_850488.1, Bos taurus (Cow) = NM_001103225.1, Mus musculus (Mouse) = NM_013488.2, Rattus norvegicus (Rat) = NM_012705.1 were retrieved from the NCBI Genebank database (www.ncbi.nlm.nih.gov) during the month of December 2009. Gene length, number of introns and exons, GC contents, coding and non-coding regions and such other characters were obtained. Different bioinformatics tools and algorithms were used for analyzing the nucleotide sequences. CpG program (http://www.ebi.ac.uk/Tools/emboss/cpgplot) was used to predict the CpG islands and the interspersed repeats (SINEs, LINE's, LTR elements, Simple repeats) were identified by using the RepeatMasker (Smit and

Corresponding Author: Liaqat Ali, National Centre of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan

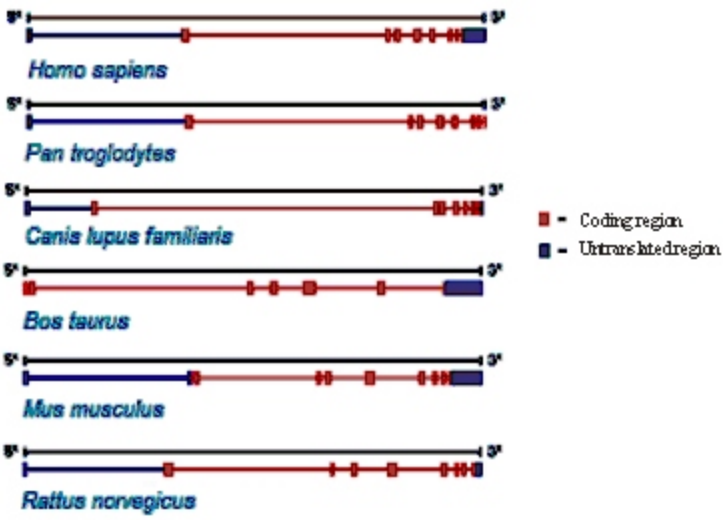

Fig. 1a: Coding and non-coding regions in different organisms

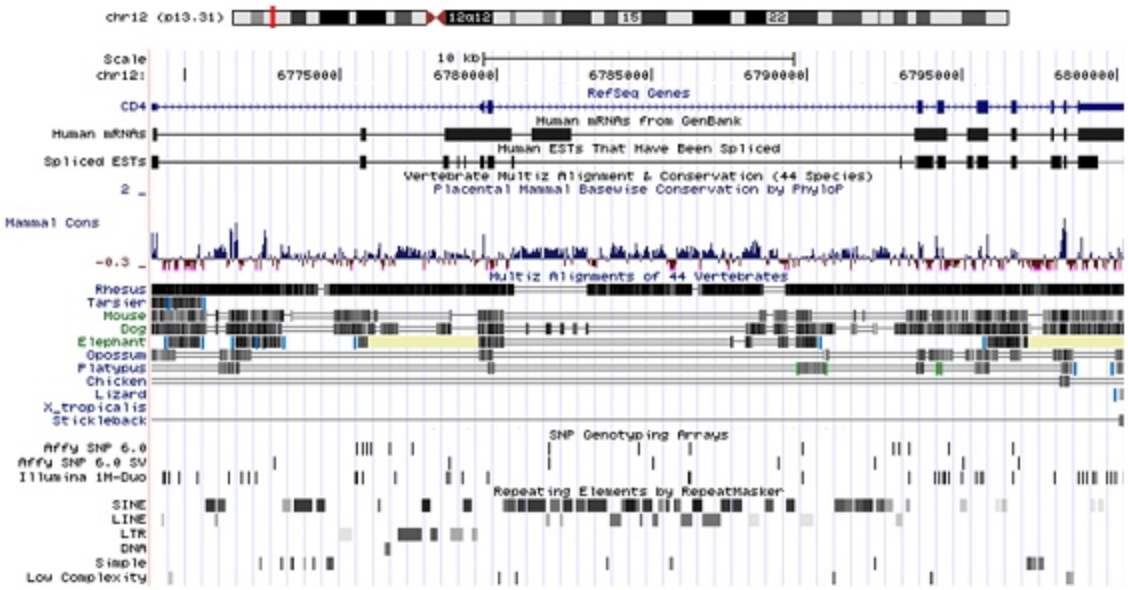

Fig. 1b: UCSC genome browser (Kent *et al*., 2002) view characterizing the Human CD4 gene

Green, 2004). Secondary structure of CD4 antigen was predicted to determine the effect of amino acids change, using SOPMA library (Geourjon and Deléage, 1994), which can be accessed by going to their freely available server (http://npsapbil.ibcp.fr/cgibin/npsa_automat.pl). For phylogenetic analysis, we considered only the amino acid sequences of the selected species. We used ClustalW (Thompson *et al*., 1994) for multiple sequence alignment with default settings and PHYLIP 3.5 (Felsenstein, 1981) was used to construct the neighbor joining phylogenetic tree.

RESULTS AND DISCUSSION

This study on the characterization and comparative analysis of CD4 antigen in six different organisms specialy focused on human and chimpanzee. CD4 is known to be the important gene responsible for interaction with HIV envelope glycoprotein gp120. A schematic representation of the detailed coding versus non-coding contents of this gene is shown in Fig. 1a, b. The variability in number and size of the gene, exons and introns in each taxon was observed in addition to the

Table 1: Characterization of CD4 gene in six organisms

Taxa	Acc No	Chr	Gene length	Exon length	No. of exons	Introns length	No. of introns	aa
Homo sapiens (Human)	NM_000616.3	12	31,326 bp	3,103 bp	10	28,223bp	9	458
Pan troglodytes (Chimp)	NM_001009043.1	12	31,712 bp	1,553 bp	10	30,159bp	9	458
Canis familiaris (Dog)	XM_850488.1	27	46,601 bp	1,656 bp	8	3,498bp	7	463
Bos taurus (Cow)	NM_001103225.1	5	16,356 bp	2,486 bp	8	19,996bp	7	395
Mus musculus (Mouse)	NM_013488.2	6	23,517 bp	3,095 bp	10	20,422bp	9	457
Rattus norvegicus (Rat)	NM_012705.1	4	25,527 bp	1,749 bp	10	23,778bp	9	457

difference in location of gene (Table 1). In four out of six taxa the number of exons was 10, including human and chimpanzee. The difference between human and chimpanzee CD4 gene was approximately 1.22%. 9 introns were present in both primates each. According to the previous studies that the common chimpanzee (Pan troglodytes) are human's closest evolutionary relatives (Goodman, 1999). Chimpanzees are thus especially focused by many scientists to teach us about humans, both in terms of their similarities and differences with human (Tarjei *et al*., 2005). The size of CD4 gene in each taxa and its respective exons and introns distribution is shown in (Fig. 2). Because of the major portion of introns in nucleotides of the CD4 gene, it was believed that the size of the gene is somehow depended on the size of the introns. The chimpanzee CD4 gene showed the genomic size of 31,712 bp which was 1.22% different from that of human CD4 gene i.e. 31,326bp. Human and chimpanzee have the same number of exons = 10 and introns = 9 (Table 1). The obvious difference between them is that the size of introns in human is 28,223 bp and in chimpanzee it is 30,159 bp. Total 277 Transcription Factor Binding Site (TFBS) were identified in upstream sequence of human CD4 which was conserved in chimpanzee. A TFBS in the promoter region of target genes in a sequence-specific way, but this contact can tolerate some degree of sequence variation (Mao and Zheng, 2006). Such information can help in predicting and verifying noncoding RNA genes is a hot issue in computational biology (Rivas and Eddy, 2001). The program used for finding TFBS was rVISTA 2.0 (Loots and Ovcharenko, 2004). Overall, not much difference was noted in the total gene size of H. sapiens and P. troglodytes, whereas other seven taxa were clearly different in all aspects.

In order to infer the evolutionary position of each individual taxon, a neighbor joining phylogenetic tree was constructed, showing branch distances for the CD4 gene, considering all six taxa (Fig. 3). It's quite clear from the Figure that *P. troglodytes* and H. sapiens are closely related to each other at the CD4 gene, as are R. norvegicus and M. musculus. Its because the mouse genome is about 14% smaller than the human genome (2.5 Gb compared with 2.9 Gb) and approximately 40% of the human genome can be aligned to the mouse genome at the nucleotide level (Waterston *et al*., 2002). Further, it is clear that Bos taurus and Canis familiaris

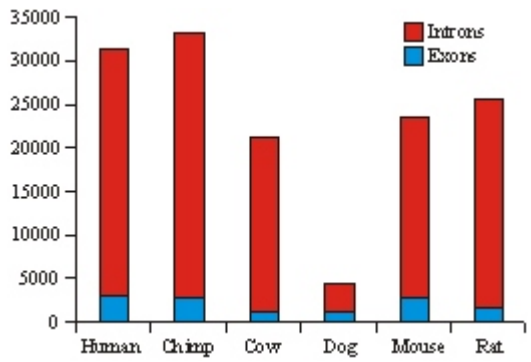

Fig. 2: length of exons and introns in different organisms

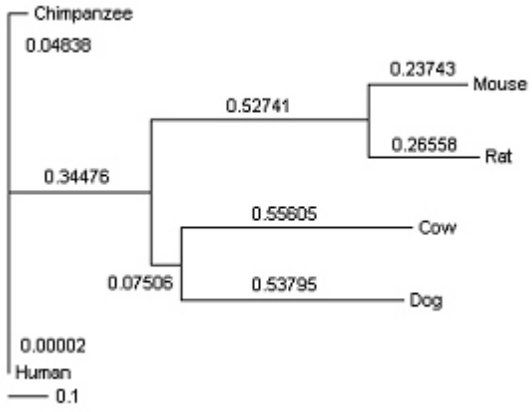

Fig. 3: Phylogenetic tree of HUMAN-CD4 with the other organisms

recently diverged from the *H. sapiens* - *P. troglodyte's* clade.

Given the importance of human CD4 for disease pathogenicity and the evolutionary closeness with mammalian species that we included in our study, we aligned the amino acid sequences of this gene in human and chimpanzee to locate variations in amino acid sequences. Amino acid changes in the CD4 polypeptide chain were also detected. It is apparent from molecular evolutionary studies in mammals that small changes in amino acid compositions between species can result in large phenotypic variation.

Secondary structure was predicted using SOPMA server, shown in the Fig. 4. It was observed that human

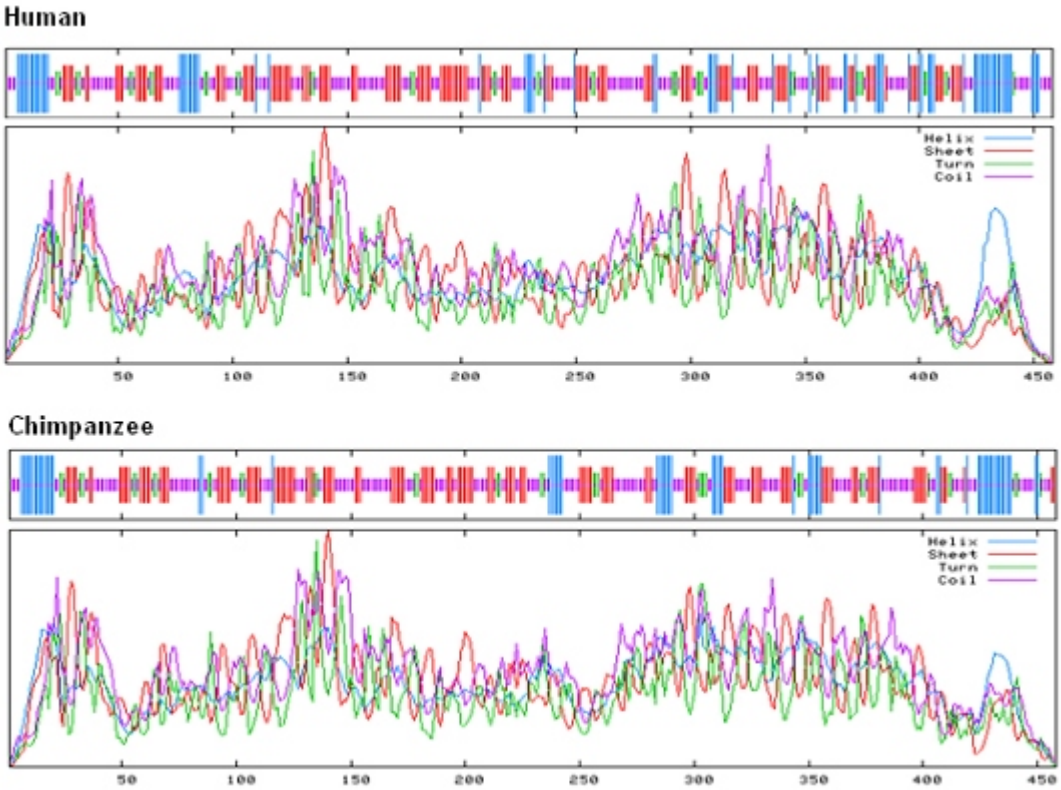

Fig. 4: Secondary structure of DENV-3 glycoprotein E.

Table 2: Percentage helix, beta turn, coils, total interspersed repeats, indels percentages and GC level in Human and Chimp CD4 gene

Protein seq.	Alpha helix (%)	Beta turn (%)	Random coil (%)	del (%)	Ins (%)	GC level (%)	Total int rep (%)
Human	17.25	8.52	42.14	4.4	5.4	54.56	10.76
Chimpanzee	14.19	8.30	46.07	3.4	5.9	54.67	10.85

protein structure has 17.25% alpha helix, 8.52% beta turns and 42.14% were coils showing a small change comparative to chimpanzee (Table 2). The result was approximately similar in both the human and chimpanzee's sequences. As there was no big difference in them, it was thought that the conserved regions maintained the special structure of CD4 in both primates.

CONCLUSION

In this study we present a thorough comparative genomics analysis and evolutionary relationship of the CD4 gene among the sequenced genomes of Homo sapiens, Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus. Specifcaly focused on Homo sapiens and Pan troglodytes. The CD4 gene in both primates showed small differences but both were more different from the other organisms. The analysis of the CD4 gene in the genomes of all selected taxa, constitute a source for future functional genomic studies.

ACKNOWLEDGMENT

We would like to express our thanks to our colleagues in Bioinformatics laboratory NCEMB Lahore, who encouraged us by showing interest in our work. We are also thankful to NCEMB for facilitating us to do this study.

REFERENCES

Clark, S.J., W.A. Jefferies, A.N. Barclay, J. Gagnon and A.F. Williams, 1987. Peptide and nucleotide sequences of rat CD4 (W3/25) antigen: evidence for derivation from a structure with four immunoglobulin-related domains. Proc. Natl. Acad. Sci. USA., 84:1649.

Collins, B.M., G.J. Praefcke, M.S. Robinson and D.J. Owen, 2003. Structural basis for binding of accessory proteins by the appendage domain of GGAs. Nat. Struct Biol., 10(8): 607-613.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol., 17: 368-376.

Geourjon, C. and G. Deleage, 1994. SOPM: A self optimised prediction method for protein secondary structure prediction. Protein Eng., 7: 157-164.

Goodman, M., 1999. The genomic record of humankind's evolutionary roots. Am. J. Hum. Genet., 64: 31-39.

Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin and T.H. Pringle, *et al*., 2002. The human genome browser at UCSC. Genome Res., 12: 996-1006.

Loots, G.G. and I. Ovcharenko, 2004. rVISTA 2.0: Evolutionary analysis of transcription factor binding sites. Nucleic Acid Res., 32: W217-21.

Maddon, P.J., S.M. Molineaux, D.E. Maddon, K.A. Zimmerman, M. Godfrey, F.W. Alt, L. Chess and R. Axel, 1987. Structure and expression of the human and mouse T4 genes. Proc. Nat. Acad. Sci., 84: 9155-9159.

Mao, L. and W.J. Zheng, 2006. Combining comparative genomics with de novo motif discovery to identify human transcription factor DNA-binding motifs. BMC Bioinformatics, 7(Suppl 4): S21.

Rivas, E. and S.R. Eddy, 2001. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2: 8.

Schockmel, G.A., C. Somoza, S.J. Davis, A.F. Williams and D.J. Healey, 1992. Construction of a binding site for human immunodeficiency virus type 1 gp120 in rat CD4. J. Exp. Med., 175(1): 301-304.

Sivashankari, S. and P. Shanmughavel, 2007. Comparative genomics - A perspective. Bioinformation, 1(9): 376-378.

Smit, A.F.A., R. Hubley and P. Green, 2004. Repeatmasker open-3.0. Unpublished

Tarjei, S.M., W.H. LaDeana, E.E. Evan, C.Z. Michael, B.J. David, Y. Shiaw-Pyng, E. Wolfgang, H. Ines, K. Lindblad-Toh, K.A. Tasha, A. Nicoletta, B. Peer, B. Jonathan, L.C. Jean, C. Ze, T.C. Asif, J. Pieter de, D.D. Kimberley, C.F. Catrina, L.F. Lucinda, G. Yoav, G. Gustavo, G. Sante, A.G. Tina, H. Toshiyuki, E.H. Karen, H. Xiaoqiu, Hongkai Ji, K.W. James, K. Mary-Claire, J.K. Edward, K.L. Ming, L. Ge, C. Lopez-Otin, D.M. Kateryna, M. Orna, R.M. Elaine, M. Evan, L.M. Tracie, E.N. William, O.N. Joanne, P. Svante, J.P. Nick, S.P. Craig, S.P. Katherine, P. Kay, S.P. Xose, R. David, R. Mariano, R. Kate, R. Maryellen, J.R. Daniel, F.S. Stephen, F.A.S. Arian, M.S. Scott, S. Mikita, T. James, T. David, T. Eray, V. Ajit, V. Gloria, V. Mario, W.W. John, C.W. Michael, K.W. Richard, S.L. Eric, H.W. Robert, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 437: 69-87.

Thompson, J.D., D.G. Higgins and T.J. Gibson, 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res., 22(22): 4673-4680.

Waterston, R.H., K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M.R. Brent, D.G. Brown, S.D. Brown, C. Bult, J. Burton, J. Butler, R.D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A.T. Chinwalla, D.M. Church, M. Clamp, C. Clee, F.S. Collins, L.L. Cook, R.R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K.D. Delehaunty, J. Deri, E.T. Dermitzakis, C. Dewey, N.J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D.M. Dunn, S.R. Eddy, L. Elnitski, R.D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G.A. Fewell, P. Flicek, K. Foley, W.N. Frankel, L.A. Fulton, R.S. Fulton, T.S. Furey, D. Gage, R.A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T.A. Graves, E.D. Green, S. Gregory, R. Guigo, M. Guyer, R.C. Hardison, D. Haussler, Y. Hayashizaki, L.W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D.B. Jaffe, L.S. Johnson, M. Jones, T.A. Jones, A. Joy, M. Kamal, E.K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W.J. Kent, A. Kirby, D.L. Kolbe, I. Korf, R.S. Kucherlapati, E.J. Kulbokas, D. Kulp, T. Landers, J.P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D.R. Maglott, E.R. Mardis, L. Matthews, E. Mauceli, J.H. Mayer, M. McCarthy, W.R. McCombie, S. McLaren, K. McLay, J.D. McPherson, J. Meldrim, B. Meredith, J.P. Mesirov, W. Miller, T.L. Miner, E. Mongin, K.T. Montgomery, M. Morgan, R. Mott, J.C. Mullikin, D.M. Muzny, W.E. Nash, J.O. Nelson, M.N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M.J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K.H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C.S. Pohl, A. Poliakov, T.C. Ponce, C.P. Ponting, S. Potter, M. Quail, A. Reymond, B.A. Roe, K.M. Roskin, E.M. Rubin, A.G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M.S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J.B. Singer, G. Slater, A. Smit, D.R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J.P. Vinson, A.C. Von Niederhausern, C.M. Wade, M. Wall, R.J. Weber, R.B. Weiss, M.C. Wendl, A.P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R.K. Wilson, E. Winter, K.C. Worley, D. Wyman, S. Yang, S.P. Yang, E.M. Zdobnov, M.C. Zody and E.S. Lander, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420: 520-562.