

## Letter to the Editor

### Selective Sweeps in the Human Genome: A Starting Point for Identifying Genetic Differences Between Modern Humans and Chimpanzees

Karl C. Diller, William A. Gilbert, and Thomas D. Kocher

Hubbard Center for Genome Studies and Graduate Program in Genetics, University of New Hampshire

Despite more than a century of interest in the evolution of humans from our close relatives the great apes, the genes responsible for phenotypic differences between humans and chimpanzees have remained elusive. Sequencing of the chimpanzee genome is expected to identify some 42 million nucleotide differences between humans and chimpanzee. How can we identify the small proportion of these differences which are the essential elements of being human? We have analyzed the draft human genome to find regions which may have experienced recent strong selection in the human line. Included in the identified regions are several genes for neural development and function, skeletal development, and fat metabolism. These observations provide a starting point in the search to identify the salient genetic differences between modern humans and our immediate hominid ancestors.

Strong directional selection for a favorable new allele can cause a "selective sweep." As the new mutant rises in frequency, adjacent chromosomal regions are also swept to fixation in a process sometimes called genetic hitchhiking (Maynard-Smith and Haigh 1974). These events can be recognized as regions of low nucleotide diversity. The size of this region depends both on the length of time required to bring the mutation to fixation in the population and the local recombination rate. With time, the record of the selective sweep is gradually erased by new mutations and the continual turnover of neutral variation. The average persistence of time for a human single nucleotide polymorphism (SNP) is  $4N$  generations, or about 1 Myr (assuming a population [ $N$ ] of 10,000 and a generation time of 25 years). So a region devoid of SNPs because of a selective sweep will regain 50% of its normal level of polymorphism in 1 Myr and 75% within 2 Myr. The current areas of low nucleotide diversity in the human genome may therefore reflect recent selective sweeps surrounding genes which are important for the evolution of modern humans some 200,000 years ago but are much less likely to be associated with the rise of *Homo erectus* 2 MYA.

We examined the data of the International SNP Map Working Group (2001) to identify areas of low nucleotide diversity in the human genome. In these data, the genome is divided into consecutive bins of 200,000 base pairs, and the density of SNPs is recorded from a

panel of 24 ethnically diverse individuals. Over the whole genome, 2.5% of the bins had nucleotide diversity ( $\pi$ ) less than  $2.0 \times 10^{-4}$ . Regions of low SNP density are most prevalent on the sex chromosomes: 89% of bins in the nonrecombining region of the Y and 15% of bins on the X chromosome have  $\pi < 2.0 \times 10^{-4}$ .

Our analysis focused on the autosomes. While any bin where  $\pi < 2.0 \times 10^{-4}$  is a likely place to look for a selective sweep, areas where we find two or more consecutive bins are of special interest because they may indicate an episode of strong selection. The 192 low-diversity autosomal bins are not distributed randomly. One would expect no more than five bins to be clustered with other bins by chance, but we found 31 bins clustered into 12 runs of two or more consecutive bins. This clumping is highly unlikely (Runs test,  $t = -10.57$ ,  $P \ll 10^{-16}$ ; see table 1) (Sokal and Rohlf 1981, pp. 782–784). There are two runs of four consecutive low-diversity bins, three runs of three consecutive bins, and seven runs of two consecutive bins. None of these runs are on chromosomes containing "recombination deserts" described by Yu et al. (Yu et al. 2001). Because of continuing changes in the assembly of the human genome, we used BLAST to match the sequences of the original bins to the current annotated assembly (February 2002) and scanned these chromosomal regions to identify genes which might be responsible for a selective sweep.

The run of four consecutive bins on chromosome 21 encodes two genes involved in neural development and function: SYNJ1, synaptojanin 1, is a presynaptic protein with a role in synaptic vesicle recycling and with a possible link to bipolar disorder (Saito et al. 2001); OLIG2, oligodendrocyte lineage transcription factor 2, specifies the development of oligodendrocytes (the myelin sheath) from the stem cell precursors of neurons and glial cells. It is required for motor neuron development and contributes to neural patterning (Zhou and Anderson 2002).

The run of four consecutive low-diversity bins on chromosome 16 contains ABCC11 and ABCC12 (also

**Table 1**  
Runs Test for Distribution of Low-Diversity Bins on Human Autosomes

	Number of Adjacent Low Diversity Bins	Expected Occurrence	Actual Occurrence
4	.....	0	2
3	.....	0	3
2	.....	5 or fewer	7
1	.....	187 or more	161
Total number of low-diversity bins	.....	192	192
Total number of bins	.....	14,703	14,703

NOTE.—Runs test:  $t = -10.57$  ( $P \ll 10^{-16}$ ).

Key words: selective sweeps, human evolution, human genome, chimpanzee genome, genetic hitchhiking.

Address for correspondence and reprints: Karl C. Diller, Hubbard Center for Genome Studies, Environmental Technology Building, 430, University of New Hampshire, Durham, New Hampshire 03824. E-mail: karl.diller@unh.edu.

*Mol. Biol. Evol.* 19(12):2342–2345. 2002

© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

**Table 2**  
**Genes Found in Autosomal Low SNP Density Bins**

Gene	Chr	Bin	Name/Function
<b>Neural development and function</b>			
ABCC11	16	8	ATP-binding cassette transporter MRP8 multidrug resistance protein family (plasma membrane drug-efflux pump); convulsive disorders PKC and ICCA
ABCC12	16	8	ATP-binding cassette, subfamily C, member 12; convulsive disorders PKC and ICCA
AGRP	16	12	Hypothalamic agouti-related protein homolog; regulates body weight (fat) via central melanocortin; obesity; subcutaneous fat
DSCR1	21	8	Down Syndrome critical region gene 1; CNS development
EDN3	20	6	Endothelin 3; induces reversion of melanocytes to glia
GCMB	6	1	Glial cells missing homolog b ( <i>Drosophila</i> ); binary switch between neural and glial cell determination
GRM1	6	11	Glutamate receptor, metabotropic 1; neurotransmission
GRM3	7	4	Glutamate receptor, metabotropic 3; neurotransmission
ITSN1	21	6	Intersectin 1 (SH3 domain protein); brain (long form), Down Syndrome; in proliferating and differentiating neurons; synaptic vesicle recycling
KCND2	7	8	Potassium voltage-gated channel, Shal-related subfamily, member 2
NRD1	1	13	Nardilysin ( <i>N</i> -arginine dibasic convertase); neural development
OLIG2	21	2	Oligodendrocyte lineage transcription factor 2; glial cell development, myelin sheath, motor neurons
OSBPL9	1	13	Oxysterol binding protein-like 9; intracellular lipid receptor; brain, spleen
PCDHB1	5	12	Protocadherin beta 1; neural cell-cell recognition and interaction; synapse formation
PCDHB2	5	12	Protocadherin beta 2; neural cell-cell recognition and interaction; synapse formation
SEMA4F	2	4	Sema domain, (Ig), (TM) and short cytoplasmic domain, (semaphoring) 4F; postnatal brain and lung; can collapse retinal ganglion cell axons
SOD1	21	1	Superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult)); Down Syndrome neural "aging" damage
SYNJ1	21	3	Synaptojanin 1; presynaptic protein; possible link to bipolar disorder
<b>Other development</b>			
BMPR2	2	11	Bone morphogenic protein receptor, type II
CDKN2C	1	11	Cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4); controls cell cycle G1 progression; spermatogenesis; inhibits tumorigenesis
GCS1	2	3	Glucosidase I, defect in a neonate with severe generalized hypotonia and dysmorphic features
GDF5	20	4	Growth differentiation factor 5 (cartilage-derived morphogenetic protein-1)
MAK	6	1	Male germ cell-associated kinase; spermatogenesis
NDRG3	20	5	NDRG family member 3; related molecule NDRG1; cell differentiation and growth arrest, especially in the peripheral nervous system
PAPA-1	2	3	Polydactyly, postaxial, type A1; hand morphology
PARD6A	16	13	Tax interaction protein 40; par-6 homolog alpha; a potential effector of Cdc42; cell polarization and movement
PPT3	8	6	NP_066280 pituitary tumor-transforming 3, may function in spermatogenesis
PTK7	6	4	PTK7 protein tyrosine kinase 7; cell proliferation and differentiation during development; colon cancer
RTKN	2	3	Rhotekin, regulation of cytoskeletal organization and the determination of cell polarity; brain, kidney
SLA2	20	5	Src-like adaptor 2 "engagement of antigen receptors can lead to either cell proliferation and differentiation or anergy and apoptosis"
SRF	6	4	Serum response factor (c-fos serum response element-binding transcription factor) for growth factor pathways
TEKT2	1	7	Tektin 2; testicular tektin B1-like protein; spermatogenesis
TRIM37	17	7	Tripartite motif-containing 37; developmental patterning, oncogenesis, mulibrey nanism
<b>Miscellaneous function</b>			
APOL5	22	1	Apolipoprotein L, 5
APOL6	22	1	Apolipoprotein L, 6
ARCN1	11	6	Archain 1; highly conserved; disrupted in leukaemia
ASK	7	5	Activator of S phase kinase
ATP5O	21	7	Mitochondrial ATP synthase, O subunit
AUP1	2	3	Ancient ubiquitous protein 1; cellular metabolism
B2M	15	1	Beta-2-microglobulin
CEP2	11	4	Cdc42 effector protein 2
CEP2	20	4	Centrosomal protein 2
CORO1A	15	2	Coronin, actin binding protein, 1A
CPNE1	20	4	Copine I, membrane associated protein
CTCF	16	13	CCCTC-binding factor (zinc finger protein); 11 zinc finger transcriptional repressor; transcriptional repressor; x-inactivation
DOK1	2	3	Docking protein 1; p62 protein downstream of tyrosine kinases; chronic myelogenous leukemia; mitogenic signalling
EEF1G	11	2	Eukaryotic translation elongation factor 1 gamma; expressed at higher levels in 7 of 9 pancreatic tumors than in normal tissues
EIF2C1	1	6	Eukaryotic translation initiation factor 2C, 1; Golgi endoplasmic reticulum protein 95 kDa ( <i>Homo sapiens</i> )
EIF3S1	11	1	Eukaryotic translation initiation factor 3, subunit 1 (alpha, 35kD)
FADS2	10	2	Fatty acid desaturase 2
FAF1	1	10	FAS-associated factor 1; apoptosis

**Table 2**  
**Continued**

Gene	Chr	Bin	Name/Function
FER1L4.....	20	4	Fer-1-like 4 ( <i>Caenorhabditis elegans</i> )
GCNT2.....	6	1	Glucosaminyl ( <i>N</i> -acetyl) transferase 2
GOLPH2.....	9	3	Golgi membranephosphoprotein 2 upregulated upon viral infection
GTPBP2.....	6	5	GTP-binding protein 2
HBOA.....	17	5	Histone acetyltransferase
HSD11B2.....	16	12	Hydroxysteroid (11-beta) dehydrogenase 2; converts cortisol and corticosterone to their 11-keto analogs; salt-sensitive hypertension
ITGAL.....	9	1	Integrin, alpha L (now on Chr 16)
KCNE1.....	21	8	Potassium voltage-gated channel, Isk-related family, member 1; heart arrhythmia
KCNE2.....	21	8	NinK-related peptide-1; human minK-related peptide; heart arrhythmia
LGALS12.....	11	3	Lectin, galactoside-binding, soluble, 12; cancer, cell cycle reg.
LOXL3.....	2	2	Lysyl oxidase-like 3
MAP3K5.....	6	8	Mitogen-activated protein kinase kinase kinase 5
MAPK8IP3.....	16	6	Mitogen-activated protein kinase 8 interacting protein 3
MCFP.....	7	5	Mitochondrial carrier family protein
MRPL53.....	2	3	Mitochondrial ribosomal protein L53
NOX3.....	6	12	NADPH oxidase 3
PAIP2.....	5	10	Poly(A)-binding protein-interacting protein 2; inhibits translation
PCMF.....	2	5	Potassium channel modulatory factor
PHB.....	17	5	Prohibitin; mutations in breast cancer
POLA2.....	11	4	Polymerase (DNA-directed), alpha (70kD)
POLH.....	6	5	Polymerase (DNA-directed)
PSMD5.....	9	9	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 5
RBM12.....	20	4	RNA-binding motif protein 12
RBM9.....	22	1	RNA-binding motif protein 9
REQ.....	11	4	Requiem, apoptosis response zinc finger gene; required for apoptosis
RPA40.....	6	5	RNA polymerase I subunit
RUNX1.....	21	10	"Runt-related transcription factor 1 (acute myeloid leukaemia 1; aml1 oncogene)"
SDBCAG84.....	20	4	Serologically defined breast cancer antigen 84
SLC22A7.....	6	4	Solute carrier family 22 (organic anion transporter), member 7
SON.....	21	6	DNA-binding protein
SPAG4.....	20	4	Sperm associated antigen 4
SULT1A3.....	16	2	Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 3
TCF8.....	5	3	Transcription factor 8 (represses interleukin 2 expression)
TEM5.....	8	4	Tumor endothelial marker 5
TJP4.....	6	5	Tight junction protein 4 (peripheral)
TMSB10.....	2	5	Thymosin, beta 10
UBE2D2.....	5	11	First exon of ubiquitin-conjugating enzyme E2D 2
UNRIP.....	12	1	Unr-interacting protein; breast cancer

known as MRP8 and MRP9). These genes encode members of the multidrug resistance protein group of the ATP-binding cassette transporter superfamily. They are most closely related to ABCC5, which is a transporter of nucleosides in a variety of tissues (Dean, Rzhetsky, and Allikmets 2001). ABCC11 is expressed in most tissues, whereas expression of ABCC12 is limited to testis, ovary, and prostate. Their chromosomal location identifies them as candidates for two inherited forms of convulsive disorders (PKC and ICCA) (Tammur et al. 2001).

Eighteen of the 89 named genes found in the low SNP density bins were involved in neural development and function. Related in function to OLIG2, we have GCMB, a binary switch between neural and glial cell determination and EDN3 which induces reversion of melanocytes to their bipotential neural crest stem cell precursors allowing them to develop either into glial cells or melanocytes (Dupin et al. 2000). In runs of two and three consecutive low-diversity bins on chromosome 21, in the area involved with Down Syndrome, there are genes for central nervous system development and function (DSCR1, ITSN1). Also of note are nardi-

lysin (NRD1) which in early mouse development is expressed almost exclusively in neural tissue (Fumagalli et al. 1998); members of the protocadherin beta family which are involved in synapse formation and neuron-neuron recognition and interaction; and two glutamate receptors (GRM1/3) involved in neurotransmission.

At least fifteen other genes on our list are involved in general skeletal development, including a bone morphogenic protein receptor (BMPR2) and a cartilage-derived morphogenic protein (GDF5). We find dysmorphic anatomic features from the Down Syndrome genes in trisomy 21, and from mutations in glucosidase (GCS1), polydactyly (PAPA-1), and TRIM37 (the syndrome of mulibrey nanism [Perpheentupa et al. 1973]).

One long-noted basic anatomical difference between humans and chimpanzees is the human subcutaneous layer of fat which is lacking in other primates (Wood-Jones 1929, p. 309). In one of our three-bin runs we find the gene *AGRP*, agouti-related protein homolog, which regulates body weight, obesity, and fat distribution, including the layer of subcutaneous fat. Other lipid-related genes are an intracellular lipid receptor (*OSBPL9*) and 2 apolipoprotein genes (*APOL5/6*).

Although multiple adjacent bins of low SNP density suggest recent selective sweeps, isolated bins may indicate relatively ancient selective sweeps or instances of weaker selection. In total, the 192 autosomal bins with low SNP density contain 18 genes related to neural development and function, 15 other genes relating to structural development or growth factors, and 56 other genes of miscellaneous function (see table 2). There are also 470 hypothetical genes whose functions may be important to human evolution.

Our analysis is based on data from the draft sequence of the human genome and a relatively small sampling of human genetic diversity. Therefore, we can expect that a few of the low-density bins are artifacts of a low depth of coverage and therefore a low power to detect SNPs; we cannot rule out the possibility that a few gene-rich bins might have low SNP density because of background (negative) selection; and we must expect a few bins to have low SNP density purely by chance because of the inherently stochastic nature of mutation and drift. Nevertheless the majority of these bins fulfill the criteria of likely selective sweeps, and they provide a principled and plausible starting point in the search for functionally significant differences between the genomes of modern humans and chimpanzees. Important next steps are the sequencing of homologous regions of the chimpanzee genome and further assessment of human variation in these candidate regions.

#### LITERATURE CITED

- DEAN, M., A. RZHETSKY, and R. ALLIKMETS. 2001. The human ATP-binding cassette (ABC) transporter superfamily. *Genome Res.* **11**:1156–1166.
- DUPIN, E., C. GLAVIEUX, P. VAIGOT, and N. M. LE DOUARIN. 2000. Endothelin 3 induces the reversion of melanocytes to glia through a neural crest-derived glial-melanocytic progenitor. *Proc. Natl. Acad. Sci. USA* **97**:7882–7887.
- FUMAGALLI, P., M. ACCARINO, A. EGEO et al. (12 co-authors). 1998. Human NRD convertase: a highly conserved metalloendopeptidase expressed at specific sites during development and in adult tissues. *Genomics* **47**:238–245.
- MAYNARD-SMITH, J., and J. HAIGH. 1974. The hitchhiking effect of a favorable gene. *Genet. Res.* **23**:23–35.
- PERPHEENTUPA, J., S. AUTIO, S. LEISTI, C. RAITTA, and L. TUUTERI. 1973. Mulibrey nanism, an autosomal recessive syndrome with pericardial constriction. *Lancet* **2**:351–355.
- SAITO, T., F. GUAN, D. F. PAPOLOS, S. LAU, M. KLEIN, C. S. FANN, and H. M. LACHMAN. 2001. Mutation analysis of SYNJ1: a possible candidate gene for chromosome 21q22-linked bipolar disorder. *Mol. Psychiatry* **6**:387–395.
- SOKAL, R. R., and F. J. ROHLF. 1981. *Biometry: the principles and practice of statistics in biological research*. W. H. Freeman and Co, San Francisco.
- TAMMUR, J., C. PRADES, I. ARNOULD et al. (12 co-authors). 2001. Two new genes from the human ATP-binding cassette transporter superfamily, ABCC11 and ABCC12, tandemly duplicated on chromosome 16q12. *Gene* **273**:89–96.
- THE INTERNATIONAL SNP MAP WORKING GROUP. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**:928–933.
- WOOD-JONES, F. 1929. *Man's place among the mammals*. Edward Arnold, London.
- YU, A., C. ZHAO, Y. FAN et al. (11 co-authors). 2001. Comparison of human genetic and sequence-based physical maps. *Nature* **409**:951–953.
- ZHOU, Q., and D. J. ANDERSON. 2002. The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. *Cell* **109**:61–73.
- NARUYA SAITOU, reviewing editor

Accepted August 26, 2002