

because each siRNA for a given target has the same on-target activity but has different off-target activity. Observing the same phenotype with multiple independent siRNAs increases the confidence with which the phenotype can be ascribed to silencing of the target gene.

The full extent of the contribution of off-target gene regulation to phenotypic induction is not known, especially because there are currently no reports examining the effect of siRNAs on the proteome. Similar to miRNAs, the effect of siRNAs on off-target protein regulation might be even greater than the effect on off-target transcript silencing. Although the magnitude of off-target transcript silencing is generally lower than that of the on-target gene, small changes in the expression levels of some proteins, such as transcription factors, might translate to large effects on phenotype. Until proteomic analyses of siRNA experiments are performed, we will not know the full consequences of siRNA off-target activity.

References

- 1 Aza-Blanc, P. *et al.* (2003) Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. *Mol. Cell* 12, 627–637
- 2 Reynolds, A. *et al.* (2004) Rational siRNA design for RNA interference. *Nat. Biotechnol* 22, 326–330
- 3 Hsieh, A.C. *et al.* (2004) A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic Acids Res.* 32, 893–901
- 4 Ui-Tei, K. *et al.* (2004) Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.* 32, 936–948
- 5 Jackson, A.L. *et al.* (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.* 21, 635–637
- 6 Saxena, S. *et al.* (2003) Small RNAs with imperfect match to endogenous mRNA repress translation: implications for off-target activity of siRNA in mammalian cells. *J. Biol. Chem.* 278, 44312–44319
- 7 Scacheri, P.C. *et al.* (2004) Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1892–1897
- 8 Sledz, C.A. *et al.* (2003) Activation of the interferon system by short-interfering RNAs. *Nat. Cell Biol.* 5, 834–839
- 9 Bridge, A.J. *et al.* (2003) Induction of an interferon response by RNAi vectors in mammalian cells. *Nat. Genet.* 34, 263–264
- 10 Chi, J.-T. *et al.* (2003) Genomewide view of gene silencing by small interfering RNAs. *Proc. Natl. Acad. Sci. U. S. A.* 100, 6343–6346
- 11 Semizarov, D. *et al.* (2003) Specificity of short interfering RNA determined through gene expression signatures. *Proc. Natl. Acad. Sci. U. S. A.* 100, 6347–6352
- 12 Persengiev, S.P. *et al.* (2004) Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). *RNA* 10, 12–18
- 13 Pebernard, S. and Iggo, R. (2004) Determinants of interferon-stimulated gene induction by RNAi vectors. *Differentiation* 72, 103–111
- 14 Hughes, T.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19, 342–347

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.08.006

Genome Analysis

Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees

Tomàs Marquès-Bonet¹, Mario Cáceres^{2,3}, Jaume Bertranpetit¹, Todd M. Preuss^{3,4}, James W. Thomas² and Arcadi Navarro¹

¹Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Doctor Aiguader, 80 08003 Barcelona, Catalonia, Spain

²Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Atlanta, GA 30322, USA

³Division of Neuroscience, Yerkes National Primate Research Center, Emory University, 954 Gatewood Road, Atlanta, GA 30329, USA

⁴Department of Pathology and Laboratory Medicine, Emory University School of Medicine, 1364 Clifton Road, Atlanta, GA 30322, USA

The genomic DNA sequences of humans and chimpanzees differ by only 1.24%. Recently, however, substantial differences in gene-expression patterns between the two species have been revealed. In this article, we investigate the genomic distribution of such differences. Besides confirming previous findings about the evolution of sex chromosomes and duplications, we show that chromosomal rearrangements are associated with increased gene-expression differences in the brain and

that rearrangements can have both direct and indirect effects on the expression of linked genes. In addition, our results are consistent with a role for some rearrangements in the original speciation events that separated the human and chimpanzee lineages.

Although the idea that differences in gene regulation might be as important as coding sequence changes in determining the morphological, behavioral and cognitive differences between humans and other primates is not new [1], it has recently been strengthened by studies

Corresponding author: Arcadi Navarro (arcadi.navarro@upf.edu).

Available online 11 September 2004

unveiling numerous changes in gene-expression patterns between the two species [2–7]. These differences are particularly remarkable in the human brain, in which a trend towards increased gene-expression has been detected [3,4,7] that is consistent with most distinctive human traits being related to our cognitive capabilities.

In a step towards constructing a genomic map of human–chimpanzee gene-expression differences, we gathered gene-expression level data from the cerebral cortex, liver, heart and fibroblasts of humans and chimpanzees from published microarray studies [2,4,6]. To estimate expression divergence, the average absolute fold-change values (FC) of gene-expression levels for the genes expressed in a given tissue were calculated in both species as previously described [4]. To avoid bias caused by sequence differences [4,6,7], we excluded any probes that were not identical in both species (supplementary data online). The genomic positions of the genes in the arrays were determined by mapping to the human genome (NCBI build 34; <http://www.ncbi.nlm.nih.gov>).

Sex chromosomes and segmental duplications

An analysis pooling all tissues together showed that the levels of expression divergence differ significantly between chromosomes (Kruskal-Wallis, $df=23$, $P<0.001$; Figure 1). Two main factors that are known to influence sequence evolution and that, *a priori*, could also influence expression differences are the evolutionary dynamics of sex chromosomes relative to autosomes [8,9] and segmental duplications [10,11]. Table 1 shows the average absolute FC in the expression level for all tissues classified according to these factors. As predicted by classical population genetics models [8,9], and as previously shown for sequence divergence data [12,13], chromosomes X and Y differ markedly from autosomes in the accumulation of gene-expression differences. These differences are particularly striking in the brain. When comparing the FC values of all genes expressed in cortical tissue from the study by Caceres *et al.* [4], we found that chromosome X presents less expression divergence (FC=1.43) and chromosome Y presents more expression divergence (FC=2.14) than the autosomes (FC=1.51). These differences are significant. Segmental duplications are another factor contributing to the differential rates of evolution between chromosomes. Duplications are not randomly distributed across the genome [14] and their expression patterns have been shown to diverge rapidly [10,11].

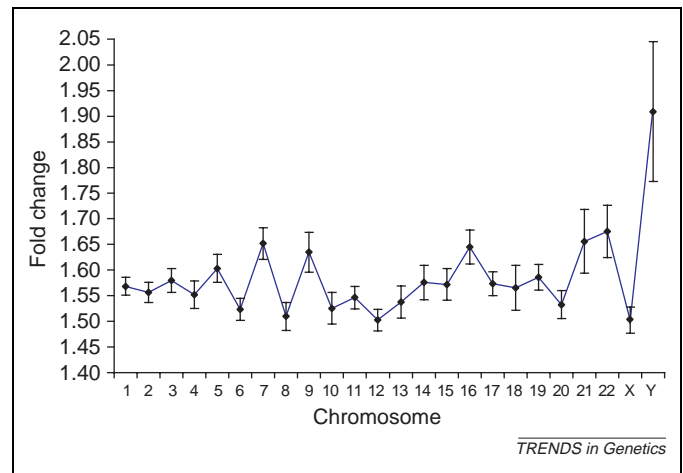


Figure 1. Average absolute fold change (FC) values between gene-expression levels of humans and chimpanzees in all tissues pooled together. Mean FC values and two standard errors (SEs) for each individual chromosome are shown.

A permutation test shows that duplicated genes present higher divergence in expression patterns in all tissues ($P<0.001$). After the removal of these genes (Table 1), differences in expression divergence between chromosomes X, Y and autosomes are still significant, because of their strong differences in the gene-expression divergence levels in the cortex (Table 1).

Box 1. Possible causes of an association between rearrangements and larger genetic divergence

Several mechanisms might contribute to an association between rearrangements and larger divergence. A new class of speciation models [16,20,21] suggests that chromosomal rearrangements might have a role in speciation processes, based on their recombination-reducing effect in heterokaryotypes [16,24]. Rearrangements segregating in an ancestral species would act as genetic barriers to gene flow between chromosomes with different organizations, which can eventually result in two daughter species with different chromosomal structures. Under these models, divergence time will be longer, and differentiation greater, in rearranged than in colinear chromosomes in the two species. Another group of alternative or complementary explanations postulates that rearrangements tend to occur, or be favored, in genomic regions that have undergone fast molecular evolution because they are regions of low functional constraint or contain clusters of genes under positive selection [16,18]. Also, it is possible that chromosomal rearrangements have direct positional effects (i.e. the establishment of rearrangements might induce changes in the expression patterns of associated genes [16,25]). Indeed, experimental evidence suggests that expression patterns can be altered around the breakpoints of chromosomal rearrangements [26–29].

Table 1. Average absolute fold change (FC) values between gene-expression levels of humans and chimpanzees^a

Tissue	Total (SE)	Autosomes (SE)	Chromosome X (SE)	Chromosome Y (SE)	<i>P</i> value ^b	Segmental duplications (SE)
All tissues	1.569 (0.0057)	1.571 (0.0059)	1.502 (0.0252)	1.909 (0.1363)	<0.001	1.655 (0.0296)
Cortex ^c	1.505 (0.0110)	1.506 (0.0114)	1.431 (0.0365)	2.142 (0.2933)	0.004	1.585 (0.0531)
All tissues (excluding duplications)	1.565 (0.0059)	1.567 (0.0060)	1.510 (0.0268)	1.799 (0.1025)	<0.001	–
Cortex ^c (excluding duplications)	1.501 (0.0113)	1.504 (0.0116)	1.423 (0.0379)	2.020 (0.2071)	0.004	–

^aValues for autosomes, sex chromosomes and segmental duplications are shown. Tissues are pooled together by averaging the FC values of every gene in different tissues. Sex chromosomes and segmental duplications differ significantly from autosomes (see main text).

^bKruskal-Wallis *P*-value for the comparison of FC in autosomes, X chromosome and Y chromosome. All tests are performed with 2 *df*.

^cData are from Ref. [4].

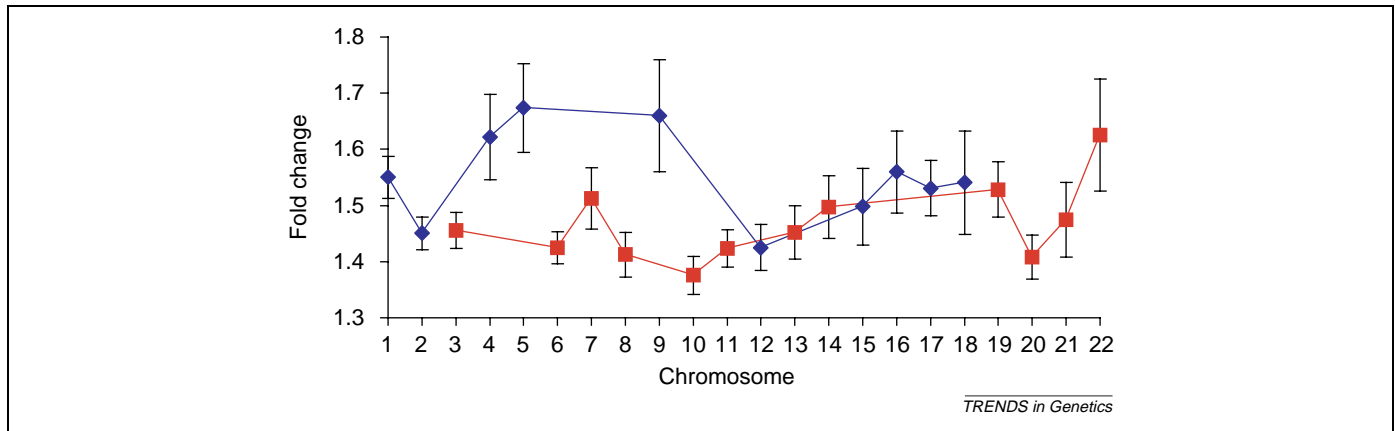


Figure 2. Average fold change (FC) values of colinear (red) and rearranged (blue) chromosomes between humans and chimpanzees in cortical samples given in Ref. [4]. The mean FC values and two standard errors (SEs) for each individual chromosome are shown.

In fact, both the effect of sex chromosomes and that of duplications are strongest in cortical samples (supplementary data online) and reinforce previous results [2–5], indicating that the human brain displays a distinctive pattern of gene expression compared with other tissues. Therefore, we will focus only on gene-expression patterns in the brain excluding genes located in segmental duplications and sex chromosomes.

Chromosomal rearrangements

Since their separation from a common ancestor, six-to-seven million years ago, the genomes of humans and chimpanzees have diverged via chromosomal rearrangements, which are on a much larger scale than differences in nucleotide sequences or differences in gene expression. Although the extent of these differences awaits full-genome comparisons, ten major rearrangements can be detected in metaphase chromosomes [15]: they include nine pericentric inversions (human chromosomes 1, 4, 5, 9, 12, 15, 16, 17 and 18) and a fusion of two ancestral acrocentric chromosomes that produced human chromosome 2. Recently, an association between rates of

chromosomal and molecular evolution has been described in several species including humans and chimpanzees [16–21], although other studies have failed to find any significant relationship [22]. To investigate the existence of a relationship between chromosomal rearrangements and gene-expression divergence, we compared the FC values of rearranged chromosomes with those of colinear chromosomes. We detected an association between chromosomal evolution and expression level divergence in cortical tissue data from the study by Caceres *et al.* [4] (Figure 2 and Table 1 in the supplementary material online). Genes in rearranged chromosomes present larger expression differences than genes in colinear chromosomes (FC = 1.54 versus 1.46, permutation test, $P < 0.001$). Similar results can be obtained [5] using the cortex dataset from Ref. [2]. Remarkably, this association was not detected in other tissues (supplementary data online).

There are several possible mechanisms that might contribute to such striking association (Box 1). We first examined the possibility that rearrangements accumulate in regions of fast molecular evolution using data from macaques, the outgroup for the dataset studied. The

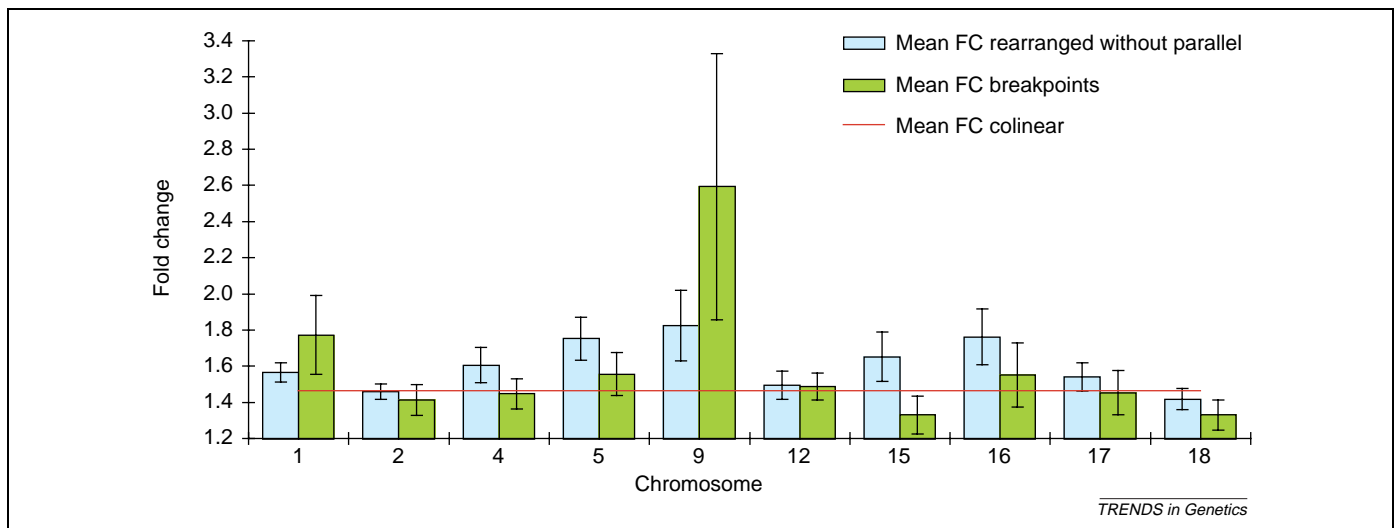


Figure 3. Average fold change (FC) values between humans and chimpanzees in cortical samples given in Ref. [4]. The mean values and two standard errors (SEs) are shown. The green bars represent FC values for the breakpoints of rearranged chromosomes. The blue bars represent FC values for the whole chromosome excluding ‘parallel’ genes (i.e. those genes that present more changes in the same lineage that harbors the new chromosomal arrangement, see main text for details). The average FC for colinear chromosomes is represented as a red line.

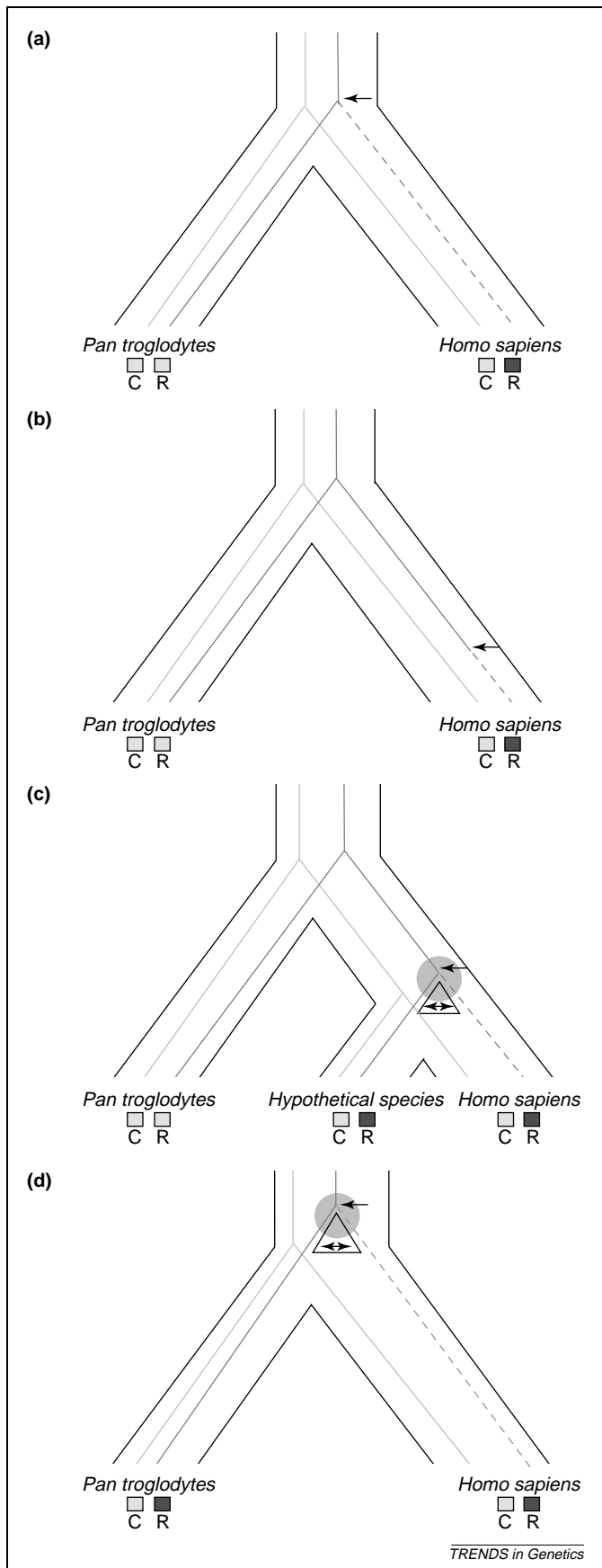


Figure 4. A schematic representing the direct (positional) and indirect (recombination-mediated) effects of rearrangements on gene-expression patterns in humans and chimpanzees. The figure represents the evolution of two kinds of chromosomes (the collinear chromosomes are represented by lighter lines and the

sequence divergence with respect to humans and chimpanzees ($\sim 4\%$) makes the expression measurements in this species less reliable, but it still provides information about the degree of change in each lineage. Regions rearranged between humans and chimpanzees do not show larger expression divergence from macaques (results not shown). In addition, we tested for the existence of clusters of rapidly diverging genes using the approach by Williams and Hurst [23]. No clustering was detected in the genome as a whole or when we separated colinear and rearranged chromosomes in two groups. Thus, the possibility that the rearranged regions studied happen to contain a large number of rapidly diverging genes is not supported by current data (supplementary online data).

We then inspected how much of the association between expression divergence and chromosomal rearrangements can be attributed to local effects on genes close to breakpoints. To that end, we classified the genes on rearranged chromosomes as near or far from known breakpoints [15]. Genes were considered near a breakpoint if they were within 2 Mb of the breakpoint or if they were in the same cytological sub-band as the breakpoint. After excluding these genes, rearranged chromosomes still present significantly larger differences in gene expression levels (Figure 3 and Table 1 in the supplementary material online; permutation test, $P=0.002$). Overall, genes close to breakpoints present the largest expression divergence between humans and chimpanzees. However, the great variation between breakpoints suggests that some rearrangements induced local expression changes, whereas others did not (Table 1 supplementary material online).

Direct versus indirect effects of rearrangements

Little is known about how alterations in expression patterns that are directly induced by rearrangements could extend along a chromosome and, thus, we cannot be sure that their positional effects have been properly assessed by the previous tests. Moreover, models of recombination-mediated effects of rearrangements [16,24] also predict larger divergence around breakpoints, where gene-flow between different chromosomal arrangements has the strongest reduction. These questions can be evaluated by means of a phylogenetic test using an outgroup. The test is

rearranged chromosomes are represented by darker lines) and the speciation event separating the two lineages. Branches with increased expression divergence are marked with dark squares below each kind of chromosome for each species. A black arrow represents the generation of a chromosomal rearrangement and a triangle recombination mediated effects of rearrangements that affect both the ancestral and the new chromosomal organizations. (a) Direct effects of a rearrangement around the time of split of the two lineages. Humans maintained the new chromosomal structure and genes linked to the rearrangement would display increased divergence only in the human branch. (b) Direct effects of a rearrangement that occur further down the human lineage. Increased divergence would only be displayed by the human branch. (c) Recombination-mediated effects of rearrangements further down the human lineage. If a chromosomal rearrangement triggers the speciation process separating a hypothetical species from humans, this results in increased divergence in the newly restructured and the standard chromosomes, but it cannot be distinguished from the previous two scenarios in comparisons between humans and chimpanzees. (d) Recombination-mediated effects of a rearrangement around the time of the human-chimpanzees split. If a rearrangement triggered the separation of humans and chimpanzees, the branch leading to chimpanzees should also experience increased divergence. This scenario could, in principle, be detected using another species (e.g. macaques) as an outgroup by means of the test proposed in the main text.

based on the different consequences that would result from a rearrangement becoming established at different points in the phylogeny of two species. First, expression differences that are the direct effects of a new rearrangement should occur only in the lineage harboring this rearrangement (Figure 4a,b). A similar pattern should be observed whenever the rearrangement occurs after the split of the two daughter species, even if the differences accumulated because a given rearrangement took part in a recent speciation process (Figure 4c). A second scenario can be produced if a rearrangement that was present at the time of the split of the two lineages did not directly cause changes in gene expression, but rather had indirect effects by reducing recombination during the period in which the two lineages were becoming separate species (perhaps even taking part in the speciation process itself). In this case one should expect to find gene-expression changes in both daughter lineages (Figure 4d) because the rearranged and ancestral chromosome segments will be isolated from each other for a longer period of time and will accumulate more differences than colinear chromosomes.

To distinguish between these two scenarios, we mapped chromosomal rearrangements and gene-expression changes onto the tree of human and chimpanzee evolution, again using macaques as the outgroup. There are three major rearrangements that are unique to the human lineage (on chromosomes 1, 2 and 18), and seven that are unique to chimpanzees (on chromosomes 4, 5, 9, 12, 15, 16 and 17) [15]. Using this information, we eliminated from our analyses every gene in a rearranged chromosome that changes in 'parallel' with the rearrangement (i.e. every gene that presents more divergence in the lineage harboring the rearrangement than in the lineage with the ancestral chromosome). After removing these genes, we still found a highly significant association between rearrangements and expression differences in the cortex ($FC=1.59$ in rearranged versus 1.46 in colinear chromosomes, permutation test $P<0.001$, Figure 3, Table 1 in the supplementary material online). This association must be due to indirect effects of rearrangements that took place while they were segregating in the same population, rather than because of direct positional effects. Interestingly, this association is strongest at the breakpoints ($FC=1.69$), which suggests that recombination-protective effects of breakpoints might be at least as important as their direct positional effects. The signature of indirect (recombination-mediated) effects is especially apparent in, for example, chromosome 5.

Concluding remarks

The results presented here confirm that, when considering gene-expression divergence, sex-linked genes evolve under different pressures than autosomes and that duplicated genes tend to evolve faster than single-copy genes. Furthermore, we have shown that rearranged chromosomes have accumulated greater differences in brain gene-expression patterns between humans and chimpanzees than colinear chromosomes. Our results therefore suggest several conclusions. First, there is an association between rearrangements and higher divergence in gene-expression levels in the brain. Second,

positional effects of rearrangements on the expression of genes located close to the breakpoints might exist, but only for certain rearrangements. Third, certain rearrangements might have exerted their influence on gene-expression levels by inhibiting recombination in an ancestral population around the time of the split of the two lineages. These observations are consistent with the model of chromosomal speciation suggested by Navarro and Barton [16,24]. Recent studies using the same datasets did not detect a significant association between chromosomal rearrangements and expression divergence [22] although, in our view, these analyses were somewhat limited (supplementary online data). Of course, the possibility remains that rearrangements and rapid expression divergence are linked to a third, as yet unknown, factor [16–18]. Also, the intrinsic noise associated with gene-expression data and the obvious shortcomings of cytological information make it necessary to undertake more detailed and thorough comparisons of the sequence and expression differences in humans with those of chimpanzees. These studies will shed light on these crucial questions, and will perhaps answer how and when we became human.

Acknowledgements

We thank A. Andrés, N. Barton, F. Calafell, J. Castresana, D. H. Geschwind, C. Lalueza and O. Lao for helpful discussions. We also thank C. Barlow and D. J. Lockhart for their contribution to the generation of the array data that was the basis of this study. We are indebted to L. Grossman, J. Hacia, M. Karaman, P. Khaitovich, S. Paabo and M. Uddin for making available the other array datasets used, and D. J. Lockhart for developing the BullFrog program. A.N. is supported by the Ramón y Cajal Program (Spanish government) and M.C. is supported by a postdoctoral EMBO fellowship. This research was funded by a grant from the Ministerio de Ciencia y Tecnología (Spain, BOS 2003–0870) to A.N., a joint project from Genoma España–Genoma Canada to A.N. (JLI/038) and a James S. McDonnell Foundation grant to T.M.P.

Supplementary data

Supplementary data associated with this article can be found at [doi:10.1016/j.tig.2004.08.009](https://doi.org/10.1016/j.tig.2004.08.009)

References

- King, M.C. and Wilson, A.C. (1975) Evolution at two levels: molecular similarities and biological differences between humans and chimpanzees. *Science* 188, 107–116
- Enard, W. *et al.* (2002) Intra- and interspecific variation in primate gene expression patterns. *Science* 296, 340–342
- Gu, J. and Gu, X. (2003) Induced gene expression in human brain after the split from chimpanzee. *Trends Genet.* 19, 63–65
- Caceres, M. *et al.* (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13030–13035
- Khaitovich, P. *et al.* (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* 14, 1462–1473
- Karaman, M.W. *et al.* (2003) Comparative analysis of gene-expression patterns in human and african great ape cultured fibroblasts. *Genome Res.* 13, 1619–1630
- Hsieh, W.P. *et al.* (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165, 747–757
- Hurst, L.D. and Ellegren, H. (1998) Sex biases in the mutation rate. *Trends Genet.* 14, 446–451
- Crow, J.F. (2000) A new study challenges the current belief of a high human male:female mutation ratio. *Trends Genet.* 16, 525–526

- 10 Gu, X. (2003) Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet.* 19, 354–356
- 11 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–65
- 12 Chen, F.C. and Li, W.H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68, 444–456
- 13 Ebersberger, I. *et al.* (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* 70, 1490–1497
- 14 Bailey, J.A. *et al.* (2002) Recent segmental duplications in the human genome. *Science* 297, 1003–1007
- 15 Yunis, J.J. and Prakash, O. (1982) The origin of man: a pictorial legacy. *Science* 215, 1525–1530
- 16 Navarro, A. and Barton, N.H. (2003) Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science* 300, 321–324
- 17 Navarro, A. *et al.* (2003) Response to comment on ‘Chromosomal speciation and molecular divergence – Accelerated evolution in rearranged chromosomes’. *Science* 302, 988
- 18 Lu, J. *et al.* (2003) Comment on ‘Chromosomal speciation and molecular divergence – Accelerated evolution of genes in rearranged chromosomes’. *Science* 302, 988
- 19 Vieira, J. *et al.* (2001) Evidence for selection at the fused1 locus of *Drosophila americana*. *Genetics* 158, 279–290
- 20 Rieseberg, L.H. *et al.* (2000) Hybridization, introgression and linkage evolution. *Plant Mol. Biol.* 42, 205–224
- 21 Noor, M.A.F. *et al.* (2001) Chromosomal inversions and the persistence of species. *Proc. Natl. Acad. Sci. U. S. A.* 98, 12084–12088
- 22 Zhang, J. *et al.* (2004) Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Res.* 14, 845–851
- 23 Williams, E.J.B. and Hurst, L.D. (2000) The proteins of linked genes evolve at similar rates. *Nature* 407, 900–902
- 24 Navarro, A. and Barton, N.H. (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution Int. J. Org. Evolution* 57, 447–459
- 25 Hey, J. (2003) Speciation and inversions: chimps and humans. *BioEssays* 25, 825–828
- 26 Spitz, F. *et al.* (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* 113, 405–417
- 27 Phippard, D. *et al.* (2000) The *sex-linked fidget* mutation abolishes *Brn4/Pou3f4* gene expression in the embryonic inner ear. *Hum. Mol. Genet.* 9, 79–86
- 28 Tanimoto, K. *et al.* (1999) Effects of altered gene order or orientation of the locus control region on human-globin gene expression in mice. *Nature* 398, 344–348
- 29 Puig, M. *et al.* (2004) Silencing of a gene adjacent to the breakpoint of a widespread *Drosophila* inversion by a transposon-induced antisense RNA. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9013–9018

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.08.009

Analysis of the centromeric regions of the human genome assembly

M. Katharine Rudd and Huntington F. Willard

Institute for Genome Sciences and Policy, Duke University, Durham, NC 27710, USA

The sequence of the human genome is not yet complete, and major gaps remain at the centromere region of each chromosome, which is comprised of repetitive α satellite DNA. In this article, we describe the sequences in the vicinity of the centromere that are included in the current genome assembly, analyze the ~7 Mb of α satellite that have been assembled thus far and anticipate the nature of the sequences that remain to be accounted for.

The centromere of most complex eukaryotic chromosomes is a specialized locus comprising repetitive DNA that is responsible for chromosome segregation during mitosis and meiosis [1,2]. Normal human centromeres consist of megabases of α satellite DNA, a repeat family containing ~171-bp monomers [3]. These monomers can be arranged either in a highly homogeneous, multimeric organization or in a more heterogeneous monomeric form that lacks this higher-order periodicity [4–6]. Despite their obvious functional significance, centromeric regions and their constituent α satellite sequences were largely omitted by the Human Genome Project because of their repetitive nature and the expected paucity of genes [7]; the reported

assemblies [8,9] of each chromosome arm thus end an uncertain distance from the functional centromere [10]. Although such regions are often considered to be difficult to sequence, in fact it is the assembly, not the sequencing itself, which presents a challenge because of the high degree of sequence homogeneity among many hundreds or thousands of copies of a given repeated sequence.

Alpha satellite organization and function

Alpha satellite DNA has been identified at every human centromere [5,6]; however, among reported chromosome assemblies, the amount and type of α satellite varies. There are two major types of α satellite DNA: higher-order and monomeric [4,5] (Figure 1). Higher-order α satellite DNA consists of ~171-bp monomers organized in arrays of multimeric repeat units that are highly homogeneous (typically 97–100% identical); by contrast, monomeric α satellite DNA lacks any higher-order periodicity, and its monomers are only on average ~70% identical [11]. In addition to their different sequence organization, monomeric and higher-order α satellite DNA also differ in their functionality. Higher-order α satellite DNA was shown to be associated with centromere function on the basis of genomic [10,12], biochemical [13,14] and artificial

Corresponding author: Huntington F. Willard (hunt.willard@duke.edu).

Available online 11 September 2004