TISSUE ANTIGENS
IMMUNE RESPONSE GENETICS

Tissue Antigens ISSN 0001-2815

REVIEW ARTICLE

Nine things to remember about human genome diversity

G. Barbujani, S. Ghirotto & F. Tassi

Department of Life Sciences and Biotechnologies, University of Ferrara, Ferrara, Italy

Key words
admixture; F_{ST}; gene flow; genetic drift; human origins; human races; pharmacogenomics; population structure

Correspondence
Guido Barbujani
Dipartimento di Scienze della Vita e Biotecnologie
Università di Ferrara
Via Borsari 46
I-44121 Ferrara
Italy
Tel: +39 0532 455312
Fax: +39 0532 249761
e-mail: g.barbujani@unife.it

doi: 10.1111/tan.12165

Abstract

Understanding how and why humans are biologically different is indispensable to get oriented in the ever-growing body of genomic data. Here we discuss the evidence based on which we can confidently state that humans are the least genetically variable primate, both when individuals and when populations are compared, and that each individual genome can be regarded as a mosaic of fragments of different origins. Each population is somewhat different from any other population, and there are geographical patterns in that variation. These patterns clearly indicate an African origin for our species, and keep a record of the main demographic changes accompanying the peopling of the whole planet. However, only a minimal fraction of alleles, and a small fraction of combinations of alleles along the chromosome, is restricted to a single geographical region (and even less so to a single population), and diversity between members of the same population is very large. The small genomic differences between populations and the extensive allele sharing across continents explain why historical attempts to identify, once and for good, major biological groups in humans have always failed. Nevertheless, racial categorization is all but gone, especially in clinical studies. We argue that racial labels may not only obscure important differences between patients but also that they have become positively useless now that cheap and reliable methods for genotyping are making it possible to pursue the development of truly personalized medicine.

Introduction

Genetics is a fast-changing field. The first sequence of a whole human genome was completed in 2003, thanks to the efforts of 2800 scientists who worked for 13 years (1). The second and the third complete sequences took 4 years and, respectively, 31 (2) and 27 scientists (3). In 2010, only 2.3 days were necessary for sequencing a whole genome (4) while the costs had dramatically dropped, from 2.7 million to a few thousand dollars. As a result, by 2012 the number of individual genomes completely typed has exceeded 1000 (4), still growing very fast. In parallel, functional elements in the genome have been systematically mapped (5) providing new, precious insights into the processes of gene regulation and gene-to-gene interaction.

Many scientists believed or hoped that the availability of this enormous mass of data would immediately improve our ability to predict phenotypes and design new therapies. Unfortunately, this has not happened yet. As a matter of fact, we still fail to understand the causes of most common diseases, and we only have a vague idea of the genetic bases of normal variation for non-pathological traits. In a sense, this is hardly surprising. Indeed, many diseases have complex causes. It is out of discussion that genetic factors play a crucial role in their onset, but there is still a substantial gap between what we can currently do, estimating the genetic predisposition to develop a disease, and what we would like to do, i.e. deciphering the complex network of interactions between genetic and non-genetic predisposing factors (exposure to chemicals, diet, lifestyle, etc.), thus coming up with accurate predictions. In the meantime, however, a much clearer picture of human diversity has emerged, only partly confirming previous ideas based on the analysis of small portions of the genome. The new genomic data have cast a different light on both normal and pathological variation, and hence understanding exactly what we know about human genome diversity seems indispensable for a rational planning of new clinical studies, for interpreting their results, and for raising public awareness of science.

In this review, we discuss nine key points about human genome variation. We present results emerging from the study of different genetic markers and complete genome sequences, emphasizing the demographic features of human evolution that can explain the observed patterns. We also stress the importance of a proper use of this information in clinical practice, with a particular focus on racial categorizations as a

poor predictor of human biological diversity and its potentially negative effects upon clinical research.

Individual genetic diversity among humans is the lowest of all primates

The comparison of genetic variation in great apes and humans is crucial to deeply investigate the origins and the evolution of our species, not to mention the fact that it can help show the molecular bases of common human diseases (6). Complete genome sequences from primates, now available (6–8), have confirmed that we are evolutionarily very close to them and have provided us with quantitative measures of that closeness. We share with the genome of our closest living relatives (chimpanzee) more than 98% of the nucleotides, over an estimated haploid genome length close to 3 billion nucleotides. Thirty-five million single-nucleotide changes (and about 5 million insertion/deletion events) have been identified, corresponding to a mean rate of single-nucleotide substitutions of 1.23% between copies of the human and chimpanzee genome. Most of these changes, 1.06% over 1.23%, appear to be fixed between species, meaning that at these sites all chimpanzees share one allele, which is different from the one shared by all humans. However, the main genetic differences between humans and other Primates do not seem to depend on point mutations, but on gain or loss of entire genes (9) that have undergone copy-number changes large enough to suggest the influence of natural selection. These genomic regions are likely to be responsible for the key phenotypic changes in morphology, physiology, and behavioral complexity between humans and chimpanzees.

What also emerged from this picture is that humans are genetically less variable than any other primate. At the beginning of 2013, 65 million nucleotide sites have been shown to vary in humans (10), and this number is steadily increasing, as more complete genomes are being sequenced. Yet, a vast majority of these polymorphisms has a very limited distribution across the species. By contrast, much larger differences are observed between pairs of orangutans, gorillas, chimpanzees, and bonobos (11). The study of the genetic relationships among three geographically close populations of common chimpanzees has shown a level of differentiation higher than that found among continental human populations (12), and the global genetic diversity of the orangutan species has been found to be roughly twice the diversity of modern humans (7), although both chimpanzees and orangutans occupy a far more restricted geographical range than we do. Further studies will doubtless expand the list of polymorphic sites, but on average a pair of random humans is expected to share 999 of 1000 nucleotides (13, 14). Quite surprisingly, as we shall see in the following section, this average similarity reflects only in part the geographic distance between the subjects being compared.

Genetic diversity between human populations is a small fraction of the species' diversity

Differences among populations are often summarized by F_{ST}, that is, the proportion of the global genetic diversity due to allele-frequency differences among populations (15). F_{ST} ranges from 0 (when allele frequencies are identical in the two populations) to 1 (when different alleles are fixed in the two populations) (for a review see Ref. 16).

Depending on the markers chosen, estimates of F_{ST} among major geographical human groups range from 0.05 to 0.13 (14). These figures mean that not only is the overall human genetic diversity the lowest in all primates but also the differences between human populations account for a smaller fraction of that diversity than in any other primate, i.e. between 5 and 13% of the species' genetic variance (17, 18). The remaining 90% or so represents the average difference between members of the same population. Different loci differ in their levels of diversity and so, for example, in 377 autosomal microsatellites (or STR, Short Tandem Repeat, markers), the differences among major groups constitute only 3–5% of the total genetic variance (19). By contrast, considering single-nucleotide polymorphisms (SNPs), the differences between continents can reach 13% (20). Recent global estimates over the whole genome from the 1000 Genomes Project suggest that the human F_{ST} could even be lower. Indeed, in analyses considering about 15 millions SNPs, 6 millions of them representing newly discovered variants, the mean values of $F_{ST} = 0.071$ between Europeans and Africans, $F_{ST} = 0.083$ between Africans and Asians, and $F_{ST} = 0.052$ between Asians and Europeans (4). This level of differentiation is less than one-third of what is observed in gorilla, $F_{ST} = 0.38$ (21) and chimpanzee, $F_{ST} = 0.32$ (6). The fact that human populations are more closely related than populations of the other primates suggests that in human evolution processes such as gene flow and admixture had a comparatively greater role than long-term isolation and differentiation.

In each individual, chromosomes are mosaics of DNA traits of different origins

When a mutation generates it, a new allele is in complete linkage disequilibrium with all the alleles that happen to lay on the same chromosome; with time, levels of linkage disequilibrium are reduced by recombination, but increase as a consequence of phenomena such as drift and admixture. The analysis of millions of SNPs over the genome has confirmed these theoretical expectations. Indeed, the combinations of alleles along the chromosome, or haplotypes, typically show blocks, namely regions of several kilobases in linkage disequilibrium, within which recombination has seldom or never occurred. The list of observed variants for every block represent the haplotype map of the human genome. Blocks vary in size across individuals and populations, depending on the relative historical weight of recombination (reducing their sizes) and drift or admixture

(increasing their size). Indeed, one of the clearest pieces of evidence supporting an African origin of humankind is the larger block size in Africans than in Europeans and Asians, a likely consequence of founder effects as small groups of Africans dispersed in the other continents (22). Information on the location and size of haplotype blocks is important for investigating the genetics of common diseases.

To understand how our genealogical history has shaped us, it is thus necessary to regard each genome as a mosaic of haplotype blocks, each with its own origin and history, brought together in the same cell by sexual reproduction. Although, as we shall see, very few genome fragments are found in a single continent (and even less so in single populations), the history of each such fragment can be inferred by comparing variation in different individuals and sometimes in different species (23). A spectacular illustration of this concept, and of how a single individual's genome may record a complex history of gene flow, is in Ref. 24.

The length of tracts assigned to distinct ancestries in an individual may be especially informative about the historical pattern of migration between populations, as well as about the time and mode of migration from one ancestral population into another. When two individuals from different parental populations mate, the first generation offspring inherits exactly one chromosome from each parental population. In subsequent generations, though, recombination events in admixed individuals generate mosaic chromosomes, essentially composed of segments having different ancestries. Intuitively, more recent admixture gives rise to longer ancestry blocks than older admixture. Thus, an excess of long blocks would indicate a recent increase in migration rate, while the opposite pattern would suggest recently reduced gene flow (25). This way, using a set of recently developed methods (26–31), it has become possible to infer with some accuracy the ancestry of many regions in individual genomes. By and large, these analyses suggest a very widespread impact of genetic admixture, a likely consequence of the absence of strong mating barriers between populations.

Allele sharing is the rule across continents

Sharing of polymorphisms across the world is extensive in humans. Jakobsson et al. (32) analyzed 525,910 SNPs and 396 CNV sites in 29 populations of five continents. They observed that 81.2% of the SNPs were cosmopolitan, i.e. occur, at different frequencies, in all continents. Less than 1% were specific to a single continent, and 0.06% were observed only in Eurasia. Combining the alleles in haplotypes, the fraction of cosmopolitan variants decreased to 12.4%, whereas 18% of the haplotypes appeared to be exclusively African. However, continent-specific haplotypes in the other four continents summed up to just 11% of the total. This small fraction of variants restricted to a single continent is in agreement with the results of a previous study of haplotype blocks. Gabriel et al. (22) sequenced 1.5 million bases of DNA in African, Asian, and European individuals: less than 2% of haplotype blocks appeared restricted to Asia, 2% appeared restricted to Europe, 25% were African specific, and the rest were shared among continents, with more than 50% occurring worldwide. Thus, with few exceptions, from the genomic standpoint, each of us can have either typically African, or generically human, features.

Several studies confirmed these results (19, 33, 34) and concurred in indicating that extensive allele and haplotype sharing across continents is the rule, not the exception, with variation within Africa exceeding that among other continents (33, 35–37). Classical population-genetics theory shows that these patterns of variation characterize species with weak or no reproductive barriers separating individuals in different groups (38). In short, it looks as though the rule for human populations is not to have independently evolved, but rather to have maintained connections through extensive gene flow. As a consequence, and as proposed by Frank Livingstone (39) on the basis of the extremely scanty data available in 1960s, genetic variation between populations tends to be continuous, without clear boundaries.

There is a clear geographic structure in human genome diversity: any population can be shown to somehow differ from any other population

Although most human variation is found within populations, the proportion that lies between continents (summarized by F_{ST}) is still significantly greater than zero. Thus, it makes sense to ask whether individuals can be assigned with good statistical confidence to their population or continent of origin on the basis of their genotype. The answer may be yes; there is indeed a relationship between patterns of genetic variation and geographical ancestry. Several recent studies have used a likelihood-based approach, implemented in the software package structure (28), to identify genetic clusters and evaluate for every individual genotype the membership to each of the inferred clusters. Rosenberg et al. (19) showed that 52 globally distributed populations can be clustered into six groups, five of which correspond to major geographic regions and one to the Kalash of Pakistan. Similar results were obtained by Li et al. (40) analyzing 650,000 common SNPs in the same populations.

However, further attempts to identify major human groups by clustering genotypes have yielded inconsistent results. Different numbers of groups and different distributions of genotypes within such groups, were observed when different datasets were analyzed (30, 41–44). The inconsistencies in these results reflect a well-known feature of human diversity, that is, different genetic polymorphism are distributed over the world in a discordant manner (44). This variation reflects in part response to different environmental pressures (Refs 45–47) and in part the different impact of demographic history

upon different genomic regions (Refs 39, 48), but in both cases leads to differences in the apparent population clusterings. It comes as no surprise, then, that if we look back at the many racial catalogs compiled since the 17th century, and at more recent genomic analyses (compare Refs 19, 32, 34, Figure 1), the only point they seem to have in common is that each of them contradicts all the others (49, 50).

But within-population diversity is very large

Above and beyond the discordant geographic patterns of population diversity, a second factor makes it difficult, or impossible, to define, once and for good, the main genetic clusters of humankind; this factor is the high level of within-population diversity. Several studies show that the

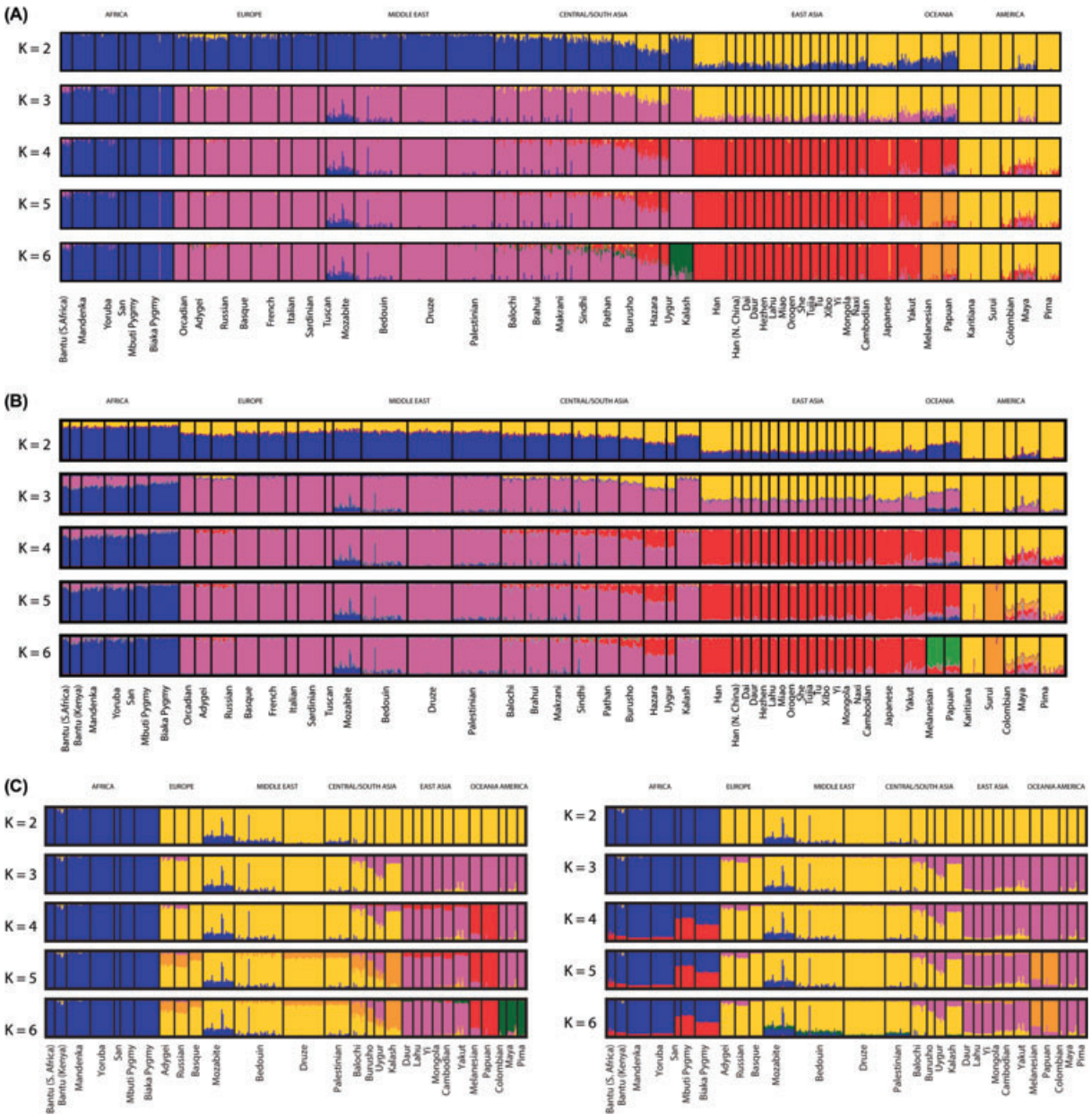

Figure 1 Comparison of four analyses of the global human population structure. Each individual genotype is represented by a thin vertical line partitioned into colored components representing inferred membership in K genetic clusters. Black lines separate individuals of different populations. Populations are labeled below each panel, with their regional affiliations above it. The analyses are based on different markers and samples: (A) 377 Short Tandem Repeat (STR) in 1056 individuals from 52 populations (19); (B) 993 STR and insertion/deletion polymorphisms in 1048 individuals from 53 populations (34); (C) 525,910 single-nucleotide polymorphisms (SNPs) (left panel) and 396 CNV sites (right panel) in 485 individuals from 29 populations (32).

largest fraction of genetic diversity in our species is due to differences between individuals of the same population, rather than to differences between populations. The pioneer analysis of this topic remains Lewontin's (18) study of protein polymorphisms from 17 loci in worldwide populations. In that study, 85.4% of total human diversity appeared due to individual variation within populations, with barely 6.3% representing the average difference between major population groups. This finding was commented with disbelief by some (51), but has been consistently replicated in protein and then DNA studies (52). The results of the analysis of 109 nuclear autosomal restriction fragment length polymorphism (RFLP) and microsatellite loci were extremely similar to Lewontin's: the variance component within population was 84.4% (17). The observation that each population harbors a large share of the global human diversity, replicated in ever-larger studies of nuclear data (19, 39, 53) means that random members of different continents tend differ, on average, just slightly more than members of the same community. On the basis of a large assemblage of microsatellite markers, Rosenberg showed that the mean proportion of alleles differing in random pairs of individuals worldwide (0.651) exceeds by 5% the mean difference for pairs from the same continent (0.618) (34).

Today, developments in DNA sequencing technology allow us to compare completely sequenced genomes. Ahn et al. (54) observed that two US scientists of European origin, namely James Watson (11) and Craig Venter (2), share fewer SNPs (461,000) than either of them shares with a Korean scientist, Seong-Jin Kim (569,000 and 481,000, respectively) (Figure 2). Of course, this does not mean that, on average, people of European origin are genetically closer to Asians than to other Europeans. However, it does show that patterns of genetic resemblance are far more complicated than any scheme of racial classification can account for. On the basis of the subjects' physical aspect, a physician would consider Venter's DNA, and not Kim's, a better approximation to Watson's DNA. Despite ideological statements to the contrary (55, 56) racial labels are positively misleading in medicine, and wherever one is to infer individual genome characteristics.

Differences between Africans are greater than between people of different continents

From a genetic standpoint, Africa is not just another continent. Paleontological data clearly indicate that anatomically modern *Homo sapiens* emerged there (57), and genetic evidence corroborates this view, showing that compared with populations from other continents, African populations have the highest level of genetic diversity at most loci (reviewed in Ref. 58). The analysis of high-quality genotypes at 525,910 SNPs in a worldwide sample of 29 populations, revealed that Africa shows the largest number of unique alleles, i.e. alleles specific to a single continent, and that in many cases the alleles found out of Africa represent a subset of the African alleles (32). In

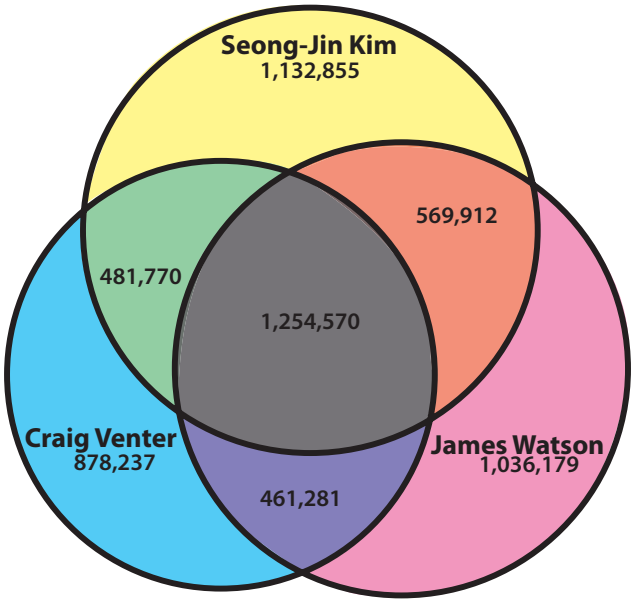

Figure 2 Venn diagram of single-nucleotide polymorphism (SNP) alleles in Seong-Jin Kim's, Craig Venter's and James Watson's genomes. Figures within the intersections are numbers of shared alleles between individuals. Modified and redrawn from Ref. 54.

the first survey of the 1000 Genomes project, populations with African ancestry contributed the largest number of variants and contained the highest fraction of novel variants roughly twice as many as in the populations of European ancestry (4). In a study of 117 megabases (Mb) of exomic sequences, the average rate of nucleotide substitutions between two hunter-gatherers from the Kalahari Desert was 1.2 per kb, compared to an average of 1.0 per kb between European and Asian individuals (35).

Gene diversity declines as a function of distance from Africa

Several measures of genetic diversity are patterned in space, with a maximum in Africa and decreasing values, respectively, in Eurasia, the Americas, and Oceania (40, 48, 59). On the contrary, linkage disequilibrium is minimal in African populations, and increases at increasing distances from there (32, 60), and the average length of haplotype blocks has a minimum in Africa around 10 kb and is close to 50 kb in Eurasia (22). All these findings are consistent with the expected consequences of an expansion of our species outside Africa, by means of dispersals of rather small groups of founders that then rapidly populated all the world (48, 61). The most likely origin of these migrational processes is East Africa (61, 62), and in fact, the geographic distance from East Africa along probable colonization routes is an excellent predictor of the genetic diversity of human populations (59). Because only a small part of the African population migrated

out of Africa, only part of Africa's genetic variation moved with them, which explains why genetic variation found in non-African populations can largely be regarded as a subsample of African variation (58, 60). Because the other continents were peopled at a relatively recent time, only few mutations are geographically restricted to these continents, i.e. those mutations that arose after the human expansion out of Africa (Figure 3).

Racial pharmacogenomics is not a step toward individual pharmacogenomics

Despite all we have seen so far, the belief that race is a reasonably good descriptor of human biological diversity is all but gone, and so is the idea that a racial categorization of patients is part of a good clinical and scientific practice. On 3 May 2013, a PubMed search using the terms 'human races' yielded 141,245 items, nearly all of them from medical journals, and this number increases at a rate of more than 20 articles per day.

The basic tenet underlying these studies is that racial categorization, although occasionally inaccurate, remains indispensable for assessing risk factors in medical and pharmaceutical research. According to Gonzalez-Burchard et al. (55): (a) reproductive barriers and endogamy have given rise to a structured human population; (b) although these barriers are mainly geographic and social, they caused genetic divergence of racial and ethnic groups; (c) as a consequence, the human population tree has major branches corresponding to five major racial groups, as defined in the US 2000 census, with secondary branches associated with indigenous populations. For all these reasons, ignoring racial background would create disadvantages to the very people that this approach means to protect (55).

Statement (a) is obviously correct; mating is not random across the whole human species, and so genetic differences exist between populations. What has proved wrong is the idea that these differences subdivide humankind in a set of recognizable genetic clusters (statement b), and it seems, at best, naïve to maintain that these clusters correspond to those in the US census classification (statement c), if only because the US census classification changes every decade. Indeed, between 2000 and 2010, races recognized in the US census have grown from 5 to infinite (15 plus 'other races'), with Hispanics or Latinos classified in a 16th group, defined as 'origin' rather than 'race' for reasons that escape us. Clearly, folk concepts change following social changes

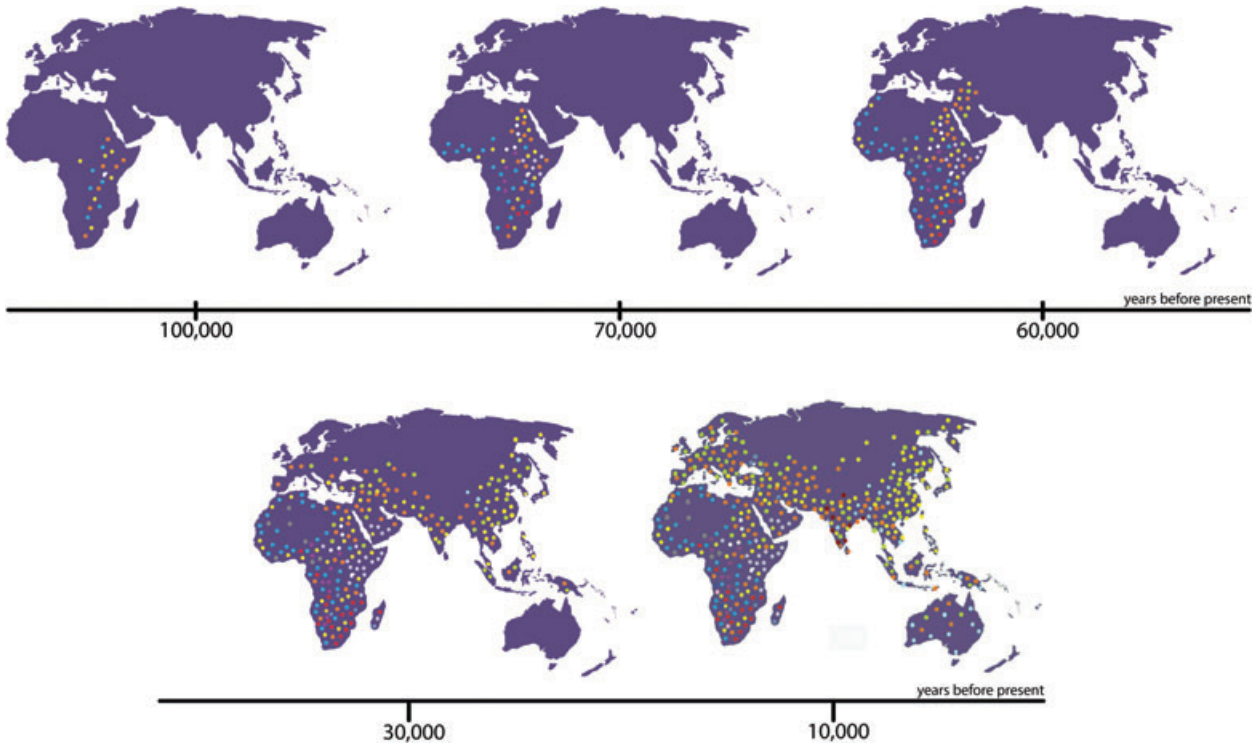

Figure 3 A highly schematic view of the evolution of human biodiversity in the last 100,000 years. Dots of different colors represent different genotypes, the distribution of which roughly corresponds to archeological evidence on human occupation of different regions. Dots of new colors appear in the maps in the course of time (e.g. red and violet in Africa at 70,000 BP, Burgundy in India at 10,000 BP), representing the effect of mutation. Because only part of the African alleles (yellow, orange and light green dots) are carried into Eurasia by dispersing Africans from 60,000 years bp (Refs 48 and 61), diversity in modern Eurasian populations is largely a subset of African diversity. Modified and redrawn from Ref. 14.

that are unrelated with biology, and hence are unsuitable for scientific purposes. One may add that differences other than those recognized in the United States may be relevant to people of different cultures (63); just as an example, in apartheid South Africa Japanese were classified as white and Chinese as colored (see Ref. 50 for more examples).

All this notwithstanding, the scientific debate on race still becomes heated on occasions. One was the patenting of the first, and so far only, drug approved by the US Food and Drug Administration for a specific racial group, BiDil. Taylor et al. (64) found that adding isosorbide dinitrate and hydralazine to standard therapy for heart failure increases survival among black patients with advanced heart failure. Critics remarked that BiDil was tested only in self-defined black patients, comparing treatments with the drug and with the placebo, but not in other groups (65); that the degree of correspondence between self-identified race and any components of the patients' genome was unknown (66, 67); and that social and economic factors probably contributing to high blood pressure were overlooked as a result of oversimplified assumptions about the existence of racial differences (65, 68). By contrast, supporters of BiDil stressed that, no matter how inaccurate was the science behind it, BiDil did save lives (69) and, more recently, that racial medicine might be a useful first step toward personalized medicine (70). Both sides accused each other to be blinded by social or political considerations that have nothing to do with science.

As a matter of fact, patenting of BiDil resulted in the resurrection of the claim that humans are naturally subdivided in biological races (71), in many cases supported by improper analyses of data available at the HapMap web site (72). Generated as part of the International HapMap Project (73), this website contains information on four populations (Nigerian Yoruba, Americans of European origin from Utah, Chinese from Beijing and Japanese from Tokyo), chosen because their well-known differences would facilitate discovery of new polymorphisms. Certainly, the HapMap samples do not provide, and are not meant to provide, a faithful description of human genome variation, but this detail, by no means secondary, was often overlooked. As a consequence, many studies based on HapMap data concluded that there are differences between Africans, Asians, and Europeans, (to nobody's surprise), but then mistook these results as evidence that indeed there are three distinct genomic clusters in the human species (see, among many examples, (74–77)). As we have seen ((32)) that is simply not true.

Still, in clinical as well as in other kinds of studies on humans, we need names to define populations and subjects. Lee et al. (78) proposed a set of guidelines on the usage of terms referring, explicitly or implicitly, to racial or ethnic categories. After stating that there is no scientific evidence supporting a biological subdivision of humans in distinct racial or ethnic groups, they urged researchers to describe how individuals were assigned category labels and to explain why samples with such labels were included in the study. They also recommended to abandon the use of race as a proxy for biological similarity, to focus on the individual rather than the group, and to avoid deterministic connections between genes and phenotype, especially when communicating to the broad non-specialist public. However, only seldom were these wise recommendations put in practice. Actually, in a large analysis of medical papers published thereafter, Ali-Khan et al. (79) found that no authors using categories such as 'race', 'ethnicity' or 'ancestry' cared to discuss the meaning of these concepts in the studied context.

Recent technical progress has dramatically reduced genotyping costs, making it possible to obtain cheap and extensive information on individual genotypes. In the future, this large amount of genetic information will likely make it possible to target drugs on specific biomarkers, so that individuals who can benefit from treatment will be identified unambiguously through their genotype, rather than through biologically inaccurate and often highly subjective racial or ethnic definitions. As for the present, Ng et al. (80) examined six drug-metabolizing genes in J. Craig Venter's and James Watson's complete genome sequences. Although both subjects identify themselves as Caucasians, they show a set of differences of clinical relevance at loci involved in drug metabolism. Venter has two fully functional **1A* alleles at the *CYP2D6* locus, and an extensive metabolizer phenotype for β-Blockers, antiarrhythmics, antipsychotics and some antidepressants; conversely, Watson is homozygous for the *CYP2D6*10* allele (common in East Asian populations, but not among Europeans), and has a decreased metabolizing activity for the same class of drugs. Doctors would not guess this and other differences by simply looking at the subjects' physical aspect. Ng et al. (80) concluded that to attain truly personalized medicine, the scientific community must leave behind simplistic race-based approaches, and look instead for the genetic and environmental factors contributing to individual drug reactions. Far from being a necessary step toward personalized medicine, racial medicine is clearly showing, on top of its long-known lack of theoretical bases, its practical irrelevance.

Conclusions and future outlooks

In clinical as well as in other kinds of studies, we need names to define populations and subjects. Therefore, the question is not whether people should or could be categorized, but how to do it. From a social standpoint, the word race is so loaded with social and political implications that avoiding it seems just reasonable. However, from the scientific standpoint, the problem is not to replace it with a more elegant synonym. Whatever term one uses to define a group of people, be it population, ethnic group, or even race, both the authors and the readers must understand that there is no deterministic connection between being part of such groups and carrying a

certain genotype or phenotype. Races are a component of our psychological and social world, and as such their importance should not be dismissed, but are scientifically ambiguous to say the least, and in scientific communication ambiguities should be kept to a minimum.

To reduce the possibility of misunderstandings, Lee et al. (78) proposed a set of guidelines on the usage of terms referring, explicitly or implicitly, to racial or ethnic categories. After stating that scientific evidence does not support a biological subdivision of humans in distinct racial or ethnic groups, they urged researchers to describe how individuals were assigned category labels and to explain why samples with such labels were included in the study. They also recommended to abandon the use of race as a proxy for biological similarity, to focus on the individual rather than the group, and to avoid deterministic connections between genes and phenotype, especially when communicating to the broad non-specialist public. However, only seldom are these wise recommendations put in practice. Actually, in a large analysis of medical papers published thereafter, Ali-Khan et al. (79) found that no authors using categories such as 'race', 'ethnicity' or 'ancestry' cared to discuss the meaning of these concepts in the studied context.

Despite all these problems, there is no doubt that recent genomic research has spectacularly improved our understanding of how humans differ, and of the demographic processes that generated human diversity. However, the attempt to convert that basic knowledge into clinical applications has been less successful. Genetics developed as a science in which data were scanty and hard to produce, and sophisticated methods had to be devised to draw inferences from the limited body of empirical evidence. Thanks to the new sequencing technologies, data have been generated on a previously unimaginable scale, but this has somewhat reversed the problem; what we seem to miss now is an intellectual framework allowing us to make complete sense of this enormous mass of information. Genome-wide association studies have shown that genetic differences account for a substantial fraction of variation among individuals, for both normal and pathological traits; we have learned that common variants predispose to, but not necessarily cause, common disease; we know less about the possible effect of rare variants, which need be better investigated, but are also difficult to recognize. However, so far only seldom has all this resulted in substantial clinical advance (81). We often conclude our papers and our talks claiming that we need more data, but it is not clear exactly what could be achieved by further expanding datasets already including thousands of cases and controls. Rather, it seems that now we need better ideas on how genetic variants and factors in the environment interact in causing the onset of disease. Only by shifting from the identification of polymorphisms associated with increased or decreased disease risk to the development of predictive models, which could then be tested against the data, genetic studies will be able to produce progress in disease treatment. For that purpose, a deep understanding of patterns of genome diversity is a necessary precondition, but just a precondition.

Acknowledgments

The research leading to this article has received funding from the European Research Council, under the European Union's 7th Framework Programme (FP7/2007-2013)/ERC Grant Agreement N 295733 (Langelin project).

Conflict of interests

The authors have declared no conflicting interests.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* 2003: **422**: 835–47.
2. Levy S, Sutton G, Ng PC et al. The diploid genome sequence of an individual human. *PLoS Biol* 2007: **5**: e254.
3. Bentley DR, Balasubramanian S, Swerdlow HP et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008: **456**: 53–9.
4. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010: **467**: 1061–73.
5. Dunham I, Kundaje A, Aldred SF et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012: **489**: 57–74.
6. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005: **437**: 69–87.
7. Locke DP, Hillier LW, Warren WC et al. Comparative and demographic analysis of orang-utan genomes. *Nature* 2011: **469**: 529–33.
8. Scally A, Dutheil JY, Hillier LW et al. Insights into hominid evolution from the gorilla genome sequence. *Nature* 2012: **483**: 169–75.
9. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics* 2007: **177**: 1941–9.
10. Lachance J, Vernot B, Elbers CC et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 2012: **150**: 457–69.
11. Kaessmann H, Wiebe V, Weiss G, Paabo S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat Genet* 2001: **27**: 155–6.
12. Bowden R, MacFie TS, Myers S et al. Genomic tools for evolution and conservation in the chimpanzee: pan troglodytes ellioti is a genetically distinct population. *PLoS Genet* 2012: **8**: e1002504.
13. Wheeler DA, Srinivasan M, Egholm M et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008: **452**: 872–6.
14. Barbujani G, Colonna V. Human genome diversity: frequently asked questions. *Trends Genet* 2010: **26**: 285–95.
15. Wright S. Genetical structure of populations. *Nature* 1950: **166**: 247–9.

16. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* 2009: **10**: 639–50.
17. Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* 1997: **94**: 4516–9.
18. Lewontin R. The apportionment of human diversity. *Evol Biol* 1972: **6**: 381–98. New York: Appleton-Century-Crofts.
19. Rosenberg NA, Pritchard JK, Weber JL et al. Genetic structure of human populations. *Science* 2002: **298**: 2381–5.
20. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* 2005: **15**: 1468–76.
21. Stone AC, Griffiths RC, Zegura SL, Hammer MF. High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc Natl Acad Sci U S A* 2002: **99**: 43–8.
22. Gabriel SB, Schaffner SF, Nguyen H et al. The structure of haplotype blocks in the human genome. *Science* 2002: **296**: 2225–9.
23. Paabo S. The mosaic that is our genome. *Nature* 2003: **421**: 409–12.
24. Henn BM, Botigue LR, Gravel S et al. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* 2012: **8**: e1002397.
25. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003: **164**: 1567–87.
26. Patterson N, Hattangadi N, Lane B et al. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 2004: **74**: 979–1000.
27. Price AL, Helgason A, Palsson S et al. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet* 2009: **5**: e1000505.
28. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000: **155**: 945–59.
29. Sankararaman S, Kimmel G, Halperin E, Jordan MI. On the inference of ancestries in admixed populations. *Genome Res* 2008: **18**: 668–75.
30. Tang H, Quertermous T, Rodriguez B et al. Genetic structure, self-identified race/ethnicity, and confounding in case–control association studies. *Am J Hum Genet* 2005: **76**: 268–75.
31. Zhu X, Zhang S, Tang H, Cooper R. A classical likelihood based approach for admixture mapping using EM algorithm. *Hum Genet* 2006: **120**: 431–45.
32. Jakobsson M, Scholz SW, Scheet P et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008: **451**: 998–1003.
33. Hinds DA, Stuve LL, Nilsen GB et al. Whole-genome patterns of common DNA variation in three human populations. *Science* 2005: **307**: 1072–9.
34. Rosenberg NA. A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum Biol* 2011: **83**: 659–84.
35. Schuster SC, Miller W, Ratan A et al. Complete Khoisan and Bantu genomes from southern Africa. *Nature* 2010: **463**: 943–7.
36. Yu N, Chen FC, Ota S et al. Larger genetic differences within africans than between Africans and Eurasians. *Genetics* 2002: **161**: 269–74.
37. Zietkiewicz E, Yotova V, Gehl D et al. Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. *Am J Hum Genet* 2003: **73**: 994–1015.
38. Wright S. Isolation by distance. *Genetics* 1943: **28**: 114–38.
39. Livingstone FB. On the non-existence of human races. *Curr Anthropol* 1963: **3**: 279–81.
40. Li JZ, Absher DM, Tang H et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008: **319**: 1100–4.
41. Bamshad M, Kivisild T, Watkins WS et al. Genetic evidence on the origins of Indian caste populations. *Genome Res* 2001: **11**: 994–1004.
42. Cooper RS, Kaufman JS, Ward R. Race and genomics. *N Engl J Med* 2003: **348**: 1166–70.
43. Romualdi C, Balding D, Nasidze IS et al. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* 2002: **12**: 602–12.
44. Wilson JF, Weale ME, Smith AC et al. Population genetic structure of variable drug response. *Nat Genet* 2001: **29**: 265–9.
45. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. *Nat Genet* 2008: **40**: 340–5.
46. Hancock AM, Witonsky DB, Alkorta-Aranburu G et al. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet* 2011: **7**: e1001375.
47. Hancock AM, Witonsky DB, Ehler E et al. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A* 2010: **107** (Suppl 2): 8924–30.
48. Liu H, Prugnolle F, Manica A, Balloux F. A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet* 2006: **79**: 230–7.
49. Barbujani G. Human races: classifying people vs understanding diversity. *Curr Genomics* 2005: **6**: 215–26.
50. Madrigal L, Barbujani G. Partitioning of Genetic Variation in Human Populations and the Concept of Race. In: Crawford MH, eds. *Anthropological Genetics: Theory, Methods and Applications*: *Cambridge University Press*, *2007*, *19–37*.
51. Edwards AW. Human genetic diversity: Lewontin's fallacy. *BioEssays* 2003: **5**: 798–801.
52. Jorde LB, Watkins WS, Bamshad MJ et al. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* 2000: **66**: 979–88.
53. Long JC, Li J, Healy ME. Human DNA sequences: more variation and less race. *Am J Phys Anthropol* 2009: **139**: 23–34.
54. Ahn SM, Kim TH, Lee S et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* 2009: **19**: 1622–9.
55. Burchard EG, Ziv E, Coyle N et al. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* 2003: **348**: 1170–5.
56. Risch N, Burchard E, Ziv E, Tang H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* 2002: **3**:comment2007.

57. Rightmire GP. Out of Africa: modern human origins special feature: middle and later Pleistocene hominins in Africa and Southwest Asia. *Proc Natl Acad Sci U S A* 2009: **106**: 16046–50.
58. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* 2008: **9**: 403–33.
59. Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. *Curr Biol* 2005: **15**: R159–60.
60. Tishkoff SA, Goldman A, Calafell F et al. A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am J Hum Genet* 1998: **62**: 1389–402.
61. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 2005: **102**: 15942–7.
62. Manica A, Amos W, Balloux F, Hanihara T. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 2007: **448**: 346–8.
63. Santos RV, Fry PH, Monteiro S et al. Color, race, and genomic ancestry in Brazil: dialogues between anthropology and genetics. *Curr Anthropol* 2009: **50**: 787–819.
64. Taylor AL, Ziesche S, Yancy C et al. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med* 2004: **351**: 2049–57.
65. Brody H, Hunt LM. BiDil: assessing a race-based pharmaceutical. *Ann Fam Med* 2006: **4**: 556–60.
66. Crawley L. The paradox of race in the Bidil debate. *J Natl Med Assoc* 2007: **99**: 821–2.
67. Hoover EL. There is no scientific rationale for race-based research. *J Natl Med Assoc* 2007: **99**: 690–2.
68. Garrod JZ. A brave old world: an analysis of scientific racism and BiDil. *Mcgill J Med* 2006: **9**: 54–60.
69. Petsko GA. Color blind. *Genome Biol* 2004: **5**: 119.
70. Wolinsky H. Genomes, race and health. Racial profiling in medicine might just be a stepping stone towards personalized health care. *EMBO Rep* 2011: **12**: 107–9.
71. Kahn J. BiDil: false promises: faulty statistics and reasoning have lead to the first "racial medicine". *Genewatch* 2005: **18**: 6–9 18.
72. Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. *Genome Res* 2005: **15**: 1592–3.
73. International HapMap Consortium, Frazer KA, Ballinger DG et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007: **449**: 851–61.
74. Kishi T, Ikeda M, Kitajima T et al. No association between prostate apoptosis response 4 gene (PAWR) in schizophrenia and mood disorders in a Japanese population. *Am J Med Genet B Neuropsychiatr Genet* 2008: **147B**: 531–4.
75. O'Donnell PH, Dolan ME. Cancer pharmacoethnicity: ethnic differences in susceptibility to the effects of chemotherapy. *Clin Cancer Res* 2009: **15**: 4806–14.
76. Sanoff HK, Sargent DJ, Green EM, McLeod HL, Goldberg RM. Racial differences in advanced colorectal cancer outcomes and pharmacogenetics: a subgroup analysis of a large randomized clinical trial. *J Clin Oncol* 2009: **27**: 4109–15.
77. Zhang W, Ratain MJ, Dolan ME. The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform Biol Insights* 2008: **2**: 15–23.
78. Lee SS, Mountain J, Koenig B et al. The ethics of characterizing difference: guiding principles on using racial categories in human genetics. *Genome Biol* 2008: **9**: 404.
79. Ali-Khan SE, Krakowski T, Tahir R, Daar AS. The use of race, ethnicity and ancestry in human genetic research. *Hugo J* 2011: **5**: 47–63.
80. Ng PC, Levy S, Huang J et al. *Genetic variation in an individual human exome*. *PLoS Genet* 2008: **4**: e1000160.
81. Need AC, Goldstein DB. Whole genome association studies in complex diseases: where do we stand? *Dialogues Clin Neurosci* 2010: **12**: 37–46.