

PERSPECTIVES

OPINION

Fish as models for environmental genomics

Andrew R. Cossins and Douglas L. Crawford

Abstract | Fish offer important advantages for defining the organism–environment interface and responses to natural or anthropogenic stressors. Genomic approaches using fish promise increased investigative power, and have already provided insights into the mechanisms that underlie short-term and long-term environmental adaptations. The range of fish species for which genomic resources are available is increasing, but will require significant further expansion for the optimal application of fish environmental genomics.

The extensive advances during the past decade in our understanding of genes and genomes have yielded several new areas of research. One of these, environmental genomics, explores how the genome interacts with and integrates cues from the environment to produce both effects of environmental stress and adaptive responses to this stress (BOX 1). This has been achieved in some so-called ‘model’ species — notably yeast — in laboratory situations, but a far richer understanding comes from characterizing responses in species that are exposed routinely to environmental stress in natural situations and in an appropriate ecological context. Such organisms frequently possess more powerful responses to these selective pressures, and these responses are therefore easier to detect and analyse. Moreover, comparisons of natural populations or closely related species that show distinct environmental responses provide a means of identifying the underpinning

adaptations. Integrating this with a detailed knowledge of the underlying molecular mechanisms is an important challenge that will benefit greatly from the application of genomic techniques.

Fish offer many advantages for investigating the organism–environment interface. In particular, studying this taxonomic group provides access to populations of a wide variety of cultivated and wild species, the latter of which are naturally exposed to many kinds of environmental stress. Unfortunately, the full power of a well-resourced genomics toolkit is seldom found in the fish species that are commonly used in environmental genomics¹. However, the usefulness of these species in environmental science would be greatly enhanced by including genomics alongside more conventional genetic and physiological approaches. Studies that have overcome restricted genomic resources with the aim of defining genome-wide responses in fish are now beginning to yield insights. Here, we outline the advantages of using fish as model organisms for environmental genomics, describe applications so far of fish genomics to environmental issues and propose an agenda for the expansion and further development of fish genomics in an environmental and evolutionary context.

Why use fish?

As vertebrates, fish share most developmental pathways, physiological mechanisms and organ systems with humans. However, unlike humans, their bodies are immersed in water, which is in intimate physiological contact with

all bodily fluid compartments and tissues through the gills and gastro-intestinal system². Fish therefore have specific physiological problems, as variations in water conditions (including oxygen levels, temperature, salinity and water-borne pollutants) can have a direct and unavoidable impact on susceptible cells, tissues and organ systems. This close physiological relationship with the environment is more easily defined and its impact more readily studied than in terrestrial species. Fish are therefore sensitive sentinels of environmental challenge, particularly pollution.

Fish are also extraordinarily diverse in terms of numbers of species, body forms, lifestyles, physiologies and environmental conditions that they experience (BOX 2). They therefore offer a large number of models of adaptation and response to a wide variety of natural and anthropogenic environmental conditions. They also provide a rich source of naturally occurring genetic variation within and between species, which can be used to explore biological processes that are adaptively important³. This diversity allows responses to the environment to be mapped onto an extensive phylogeny that extends over 500 million years of history, providing insights into a range of evolutionary processes, from ancient to modern⁴. Moreover, fish genomes are more varied and ‘plastic’ in comparison with other vertebrates⁵, owing to more frequent genomic changes (through polyploidy, gene and chromosomal duplications, and gain of introns), such that gene redundancy seems to be widespread in fish genomes⁶. Fish therefore offer unparalleled opportunities for comparative analysis within a genomic context. Indeed, it is easy to identify fish species that are well-suited to the investigation of almost any environmental question, and the understanding that is gained from these studies can be greatly enhanced by using a comparative approach.

Finally, some fish species are experimentally tractable: some are easy and inexpensive to maintain and culture, some have large

brood sizes, and others can provide ISOGENIC and genetically manipulated lines that allow the roles of specific genes or genotypes to be directly addressed. These features make fish generally useful as models for investigating a range of biological problems^{7,8}. They are also the principal attributes that are required for genetic or genomic models, and fish have important representatives among sequenced organisms, notably the zebrafish (*Danio rerio*), medaka (*Oryzias latipes*) and two pufferfish species (*Takifugu rubripes* (formerly designated as *Fugu rubripes*) and *Tetraodon nigroviridis*). Access to these fish sequences is opening up new comparative approaches to gene function across the vertebrates⁹.

What has been achieved so far?

Genome-scale analyses are in their infancy, particularly in environmental sciences. However, there are already some examples of how environmental genomics approaches have increased our understanding of how organisms respond to changes in their environment. Work in fish has so far shed new light on two kinds of environmental response: short-term phenotypic responses and evolutionary responses.

Short-term responses. Transcript expression profiles have been used to examine the effects of oxygen concentration, temperature and toxicological exposure in bony fish. Here, we focus on insights provided by the first two of these.

In the first genome-scale analysis of a non-model fish species, Gracey *et al.*¹⁰ used cDNA microarrays to explore mechanisms of hypoxia adaptation in an estuarine goby, *Gillichthys mirabilis*, that is adapted to survival in hypoxic burrows to which it retreats at low tide. The arrays comprised ~1,600 amplified cDNAs isolated from the tissues of interest. This study showed that hypoxia had distinct effects on different tissues: in skeletal muscle, the expression of genes involved in protein synthesis and locomotion rapidly decreased, and during later stages of hypoxia, transcripts for cell growth and proliferation were downregulated in the liver. The liver also showed increased expression of genes involved in anaerobic metabolism, which might provide energy for survival during hypoxia. This study was important in showing how lack of genome sequence or of EST resources, neither of which is available for this species, does not necessarily impede progress, and that even modest collections of cDNA probes can yield penetrating overviews of transcriptional responses.

Similar patterns of mRNA expression have more recently been found during exposure of zebrafish embryos to hypoxia¹¹,

indicating that even model genomic species, not regarded as possessing well-developed hypoxia-resistance mechanisms, can show interpretable responses to environmental stressors. Therefore, studies of responses identified in wild species can be augmented using genomic models.

More recently, Gracey *et al.*¹² carried out a much larger multi-tissue analysis of responses to increasing cold in the common carp (*Cyprinus carpio*), which can withstand wide seasonal extremes of temperature. Gene-expression levels were compared when animals were either exposed to constant temperatures or after a single change from one temperature to another. This revealed a

core response across all tissues that showed strong similarity to that observed for exposure to many stressors in laboratory yeast. This response involved the upregulation of transcripts involved in stress responses, cytoskeletal reconstruction, protein catabolism, lipid metabolism and energy pathways. In addition, many tissue-specific responses were seen. A surprisingly large fraction of the genome was estimated to be involved in the phenotypic transition to the cold-adapted state and several important new candidate genes were identified for conventional physiological assessment. Similar results were seen by Ju. *et al.*¹³, who studied cold exposure in the catfish.

Box 1 | Defining the organism–environment interface

All organisms experience fluctuations in environmental conditions that can cause damage to cells, tissues and organs, and might result in injury and death.

The environment

The physical or abiotic environment consists of many natural factors that can function as stressors either individually or in combination. These include temperature, oxygen, salinity, desiccation and ultraviolet light, as well as anthropogenic factors such as heavy-metal, organic and thermal pollution. These can all affect living organisms in two fundamental ways, either by debilitation and ultimately death (which result in ‘resistance’ or tolerance effects), or at non-lethal levels of exposure by affecting normal life processes (resulting in ‘capacity’ effects).

Responses and adaptations

Environmental adaptations might be manifested as a result of both resistance and capacity effects. Therefore, animals might improve their resistance to the debilitating or damaging effects of stress. A good example is cold-hardening during prior exposure to cooler temperatures, a classical resistance adaptation. In addition, they might show compensation of biological rates for the direct effects of daily or weekly temperature fluctuations; a capacity adaptation. Some fish have innovative molecular response mechanisms that provide insights into the adaptive nature of the organism–environment interface (for example, production of antifreeze proteins).

Responses to abiotic stress occur over several distinct timescales. Some occur either immediately or within a few minutes, including behavioural or rapid physiological responses, such as colour changes or vascularization of the skin. Others might take days or weeks to occur and might involve changes in gene and protein expression, leading to longer-lasting but reversible changes in phenotype. These ‘acclimatization’ responses contrast with irreversible developmental changes, when stressful experiences during development can give rise to distinctive non-reversible phenotypes, a process termed ‘canalization’ or developmental programming. A good example of this is the effect of natural rearing temperatures on the number of vertebrae in larvae of the herring (*Clupea harengus*)⁴⁸, or the sex of turtles⁴⁹. These non-heritable phenotypic outcomes contrast with changes that occur in populations over many generations that involve heritable changes in the genotype during evolution.

Conformity and regulation

Sensitivity to abiotic stresses is fundamentally affected by the development of powerful homeostatic systems, especially in more complex higher organisms, and this affects the modes of response. ‘Conformers’ are organisms whose internal condition or ‘state’ varies with fluctuations in the external state. They are distinguished from ‘regulators’, which, by virtue of homeostatic systems, can maintain their internal state independently of external conditions. A good example of this is the thermoregulation of body temperature in mammals and birds. However, in most fish, a change in aquatic temperature results in a change in body temperature, so that all cells and all molecules experience and cope with the full force of temperature changes. ‘Conforming’ fish use different modes of adaptation to cold compared with ‘regulating’ mammals, although both involve molecular components. Alternatively, for osmotic and ion regulation, all vertebrates (including fish) rely on organ-level responses that effectively shield the cellular and molecular machinery from environmental fluctuations in salinity. Of course there are limits to this, and organ systems might fail under extreme stress, ultimately leading to death.

Podrabsky and Somero¹⁴ explored responses to a daily pattern of temperature changes that was more similar to that experienced in a natural setting in the killifish, *Austrofundulus limnaeus*. This species inhabits ephemeral ponds in Venezuela, with daily temperatures that routinely cycle over a 20°C range. Cycling versus constant temperatures differentially affected mRNA expression for heat-shock factors, cholesterol and fatty-acid biosynthesis, fatty-acid modification, carbohydrate metabolism and protein turnover. Evidently, responses to temperature can occur over just a few hours, rather than the longer-term responses shown in studies of other environmental model species, highlighting the benefits of using fish in studies of short-term responses to environmental stress.

Evolutionary responses. The responses to hypoxia and temperature described above occur on a physiological timescale and, being phenotypic in character, are generally reversible. Two other mechanisms that can have irreversible effects on patterns of gene expression are alterations in developmental programmes and evolved differences between genetically distinct populations or species. We focus here on how fish genomics can be used to provide important insights into the second of these two mechanisms.

Much of our understanding of how the environment affects vertebrate evolution, especially the evolution of enzymes^{15–17} and gene-expression patterns^{18,19}, has come from the

study of fish. For example, populations of fish of the same species that are subjected to lower environmental temperatures have enzymes that work better at lower temperatures than those expressed in populations from warmer water, owing to the presence of different allelic variants in different populations. STANDING GENETIC VARIATION of this kind is the raw material for evolution: natural selection of GENETIC DRIFT giving rise to NEUTRAL VARIATION effects a change in a population by altering the frequency of an allelic polymorphism, eventually FIXING one allele and excluding all others.

Environmental genomics studies using fish have begun to provide a greater understanding of evolutionary mechanisms, providing surprising new estimates of the extent of natural variation in transcript expression. For example, among European flounder (*Platichthys flesus*) populations found in polluted and unpolluted environments, a surprising 7% of the genes examined had different patterns of expression²⁰.

A good example of how fish environmental genomics has provided insights into adaptation to the local environment by natural selection comes from a study of contrasting populations of two closely related species of minnow of the genus *Fundulus* (one from a cold habitat, and the second from a warmer habitat; FIG. 1), both raised in a common environment. Expression patterns of 18% of approximately 1,000 genes were significantly different between individuals within the same population¹⁸. Similar variation in

gene-expression levels was also found between different populations of the same species, and although some of these differences are most readily explained by neutral variation, others might be adaptive — that is, they could have evolved by natural selection. Specifically, in contrast to what is expected under the NEUTRAL HYPOTHESIS, gene expression for 27 genes in one of these species was more similar to the other closely related species in the same environment than to individuals of the same species raised in different environments.

To determine whether these potentially adaptive changes in transcript expression affect biological functions, both cardiac metabolism and gene expression were analysed in 16 individuals from the two populations of *Fundulus heteroclitus*²¹. Differences in gene expression among individuals within populations explains most of the variation in cardiac metabolism; that is, we can predict the rate of substrate utilization from the patterns of gene expression. A provocative outcome of these investigations is that ‘the genes that matter’ — those in which variation in expression accounted for differences in cardiac physiology — vary among individuals. For example, among some individuals, the expression of genes required for oxidative phosphorylation accounted for much of the variation in cardiac physiology. However, in other individuals, differences in expression of glycolytic enzymes were more important. Therefore, the relationship between gene expression and phenotype might also vary among individuals. From an evolutionary point of view, these results indicate the existence of high levels of variation in gene expression between individuals that have functional consequences, and this might be important for evolutionary adaptations to environmental change. These functional genomic studies were possible owing to the divergence among individuals within and between outbred populations taken from the wild. These circumstances are more likely to occur with organisms such as fish — which are in intimate contact with their environment and have very large population sizes — than in other environmental model organisms that do not have these characteristics.

Future goals

As post-genomic technologies mature, and with the provision of a much-expanded sequence resource for fish, we expect that genomics will prove useful in a growing list of specific investigations, including those listed in BOX 3. Here, we discuss some important general future aims of fish environmental genomics.

Box 2 | The diversity of fish

Currently, more than 28,000 species of fish have been identified — more species than for all other vertebrate groups combined.

- Fish have highly diverse body forms and lifestyles (from highly active to sedentary, hypometabolic and quiescent), and have correspondingly diverse physiologies.
- Fish range in size from the whale shark, *Rhincodon typus*, which weighs ~10,000 kg, to the dwarf goby, *Trimmatom nanus*, which weighs ~0.1 g, a size range of 10⁸.
- Fish genome sizes vary between 0.32 and 133 billion base pairs with, for most species, a consistent number of chromosomes.
- Divergence and radiation of fish species has taken place over 500 million years of evolution, and the evolutionary history of fish is well resolved and broadly accepted.
- Fish show diverse physiological responses to the environment that correspond to their phylogenetic grouping and evolutionary status. Extant species that belong to ancient groups provide representatives that allow assessment of evolutionary change over both short (thousands of years) and long (hundreds of millions of years) timescales.
- Fish inhabit an enormous range of environments: temperatures ranging from –1.89°C in polar oceans to +45°C in springs; salt levels ranging from fresh water, to brackish water, to sea water, to hypersaline waters; hypoxic and even anoxic waters; and pressures ranging from 1 Atm up to 1,000 Atm in abyssal environments.

Fishbase provides the most comprehensive searchable worldwide database of more than 28,000 listed species, as well as a glossary of over 6,000 terms. A useful introductory text on fish biology and diversity is provided by Helfman *et al.*⁵⁰.

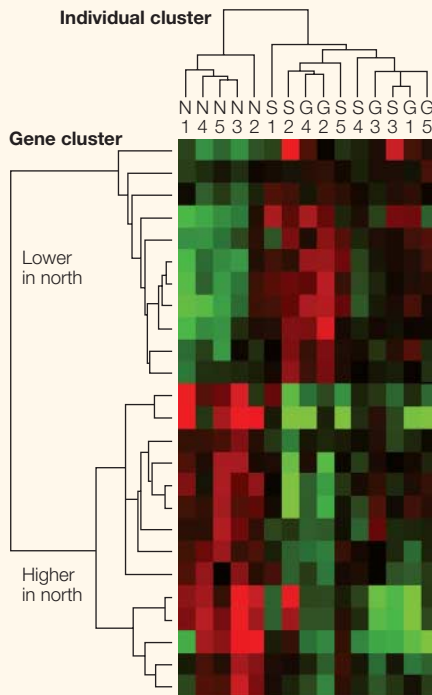


Figure 1 | Variation in gene expression within and between closely related species. These microarray analysis results reveal patterns of gene expression that can distinguish between different populations of closely related *Fundulus* minnows. The populations analysed were a *Fundulus heteroclitus* population from cold northern waters (labelled N), a population of the same species from warmer southern waters (labelled S), and a population of the sister species, *Fundulus grandis*, which was also from warm waters (labelled G). Red and green indicate values that are larger or smaller, respectively, than the overall mean. For the northern *F. heteroclitus* population, profiles for 25 genes were statistically different from both southern *F. heteroclitus* and *F. grandis*. The dendrogram on the left groups genes with similar patterns of expression among individuals. The dendrogram at the top groups specimens on the basis of similar patterns of expression for the 25 genes. This shows that the southern population of *F. heteroclitus* had a profile that was indistinguishable from that of *F. grandis*, whereas both of these populations could be distinguished from the northern population of *F. heteroclitus*. These data indicate that some patterns of gene expression exceed levels of variation that are expected under the neutral hypothesis, and seem to be an adaptation to the environment. Reproduced, with permission, from *Nature Genetics* REF. 18 © (2002) Macmillan Magazines Ltd.

Improved profiling of gene networks. In interpreting the results of microarray analyses, it is not uncommon to interpret the resulting gene list selectively by focusing on just a few transcripts. A less biased, more objective method is to categorize genes initially using

Box 3 | Potential future areas for fish environmental genomics research

- Salinity adaptation in seasonally migratory species, such as eels, flounder and salmonids.
- Thermal tolerance and compensation in seasonal and evolutionary adaptation.
- Hypoxia and anoxia resistance mechanisms in cyprinids, particularly the crucian carp (*Carassius carassius*) and goldfish (*Carassius auratus*), which are tolerant of anoxic environments.
- Metabolic depression in winter and dry seasons (aestivation and torpor).
- Embryonic developmental quiescence in response to desiccation, cold or lack of food (for example, in the killifish *Austrofundulus limnaeus*).
- Influence of stress on behaviour, social stress and hierarchical social interactions.
- Changes in developmental programmes and phenotypic plasticity in the face of environmental stress.
- Effects of pollutants on sex-determination mechanisms, endocrine disruption and fecundity.
- Understanding toxicological susceptibility and responses to toxins, local adaptation of tolerant populations and identification of refined ecotoxicological biomarkers for environmental regulation.

the GENE ONTOLOGY (GO) OR KEGG SCHEMES²² according to their molecular functions, the biological processes or metabolic pathways they participate in, or their cellular location. The distinctive features of an entire gene list can then be assessed using standard statistical techniques, such as the GO Matrix¹². However, all of these methods of annotation and interpretation are ultimately dependent on the accurate identification of genes, and this depends on the quality of both the homology searching methods and the sequence databases used.

Although the most heavily annotated high-quality vertebrate gene databases are those for humans and mice, there is an increasingly large amount of genome sequence data for fish. This can be used as a point of reference against which genes from fish species that lack a sequenced genome can be searched. However, there are complications that have a notable effect on the quality of gene identification in fish; a significant number of chromosomal and genomic duplication events have occurred during fish evolution, which has led to extensive PARALOGY among fish genes. Understanding the resulting diverse set of proteins within families is a problem that will only be solved by generating a more comprehensive set of well-annotated, full-length cDNAs for a few selected species, together with an improved understanding of the evolution of genome architecture, as discussed later.

Improved identification of important genes. Identifying exactly which genes control susceptibility to environmental stress remains challenging. Transcript screening might yield a large number of responding genes and advanced pattern-searching techniques can

yield useful overviews, but few genes stand out as being special except in the context of prior knowledge. Genome-wide expression-profiling approaches that are based on knowledge of the entire genome therefore need to be combined with other methods that allow gene function to be determined. Mutagenesis screens in zebrafish have proved important in developmental studies, where phenotypes are easy to score in embryos. Manipulating the expression of key genes using MORPHOLINOS or antisense RNA, or expressing more copies of genes transgenically, with subsequent phenotypic assessment, are alternative ways of demonstrating the roles of genes identified from genomic studies. However, conventional genetic analysis (QTL analysis) of genetically divergent populations is likely to provide the most specific means of linking genetic loci to environmentally relevant phenotypes, although resolving the resulting chromosomal region down to specific genes is still problematic and remains an important challenge for the future. The QTL approach has proved useful in identifying the genes responsible for morphological variation in sticklebacks²³ and can also be used in this and other fish species to identify loci that control environmentally adaptive traits. One particular trait that is likely to be amenable to genetic dissection is the heritable resistance to environmental dioxins in *F. heteroclitus*²⁴.

Assessing variations in environmental responsiveness. Assessment of direct responses of fish to environmental challenges represents the tip of the iceberg, and genomic screening approaches could also be used to address a far wider range of issues. Foremost among these is the scale of phenotypic variation between

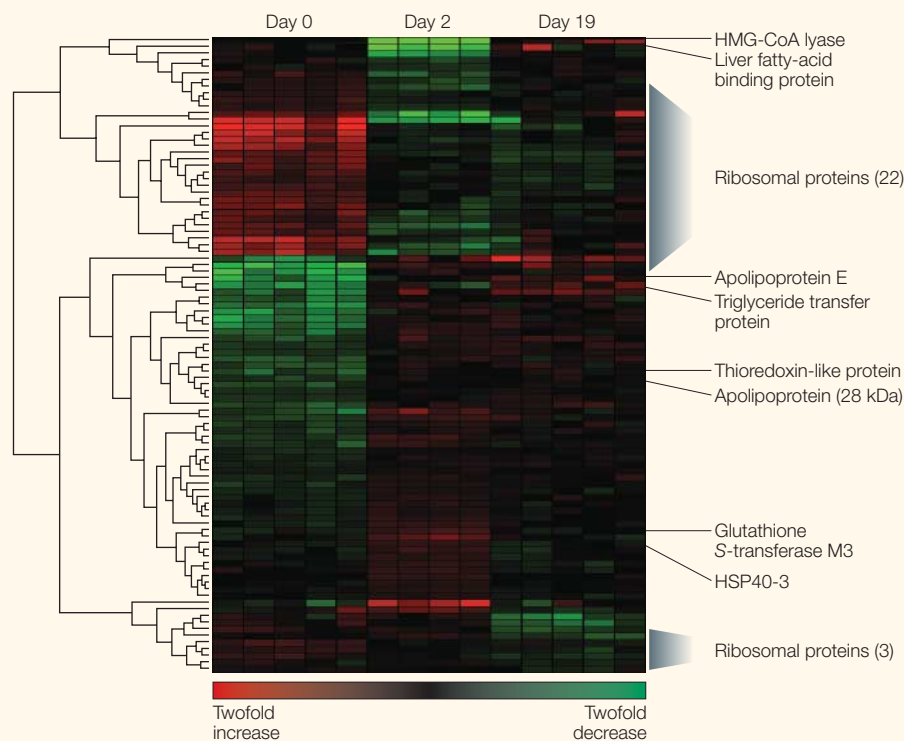


Figure 2 | Transcript profiling as a phenotyping tool. A group of common carp, *Cyprinus carpio*, were held in a large tank under standard conditions in the absence of environmental stressors. Individuals were sampled on days 0, 2 and 19. Hepatic mRNA expression was profiled using a *C. carpio* cDNA microarray; red and green indicate values that are larger or smaller, respectively, than the overall mean. Although the magnitude of changes in expression is small (typically < twofold), 96 out of 3,636 non-redundant cDNAs showed significant variation in expression between the days of sampling (determined using an analysis of variance test, with $p < 0.01$ and using a multiple testing correction). Interestingly, 26% of the cDNAs (25 out of 96) encoded ribosomal proteins, and others included genes that are involved in fatty-acid metabolism and the cell stress response. These indicate changes in protein and lipid metabolism, perhaps owing to minor day-to-day variations in a controlled feeding protocol. Transcript profiles allow specimens to be discriminated according to day of sampling, indicating just how sensitive this technique is to subtle differences in holding conditions (A.Y. Gracey and E.J. Fraser, personal communication). HMG-CoA lyase, hydroxymethylglutaryl-CoA lyase; HSP40-3, a heat shock protein.

individuals, populations and closely related species, and the relationship of this to environmental plasticity and adaptation. In addition, the comparative analysis of responses between the main taxonomic groups might explain why certain taxa have well-defined molecular and physiological responses associated with survival whereas others do not. The ability to make comparisons between fish species — for example, between those that can survive cooling to temperatures as low as 0°C and their sister taxa that die at temperatures below 20°C, or between those that can survive movement between fresh and sea water with those that cannot — is a key strength of fish environmental genomics.

The power of transcript screening to make such comparisons lies in the large number of independent probes used to characterize a particular tissue sample. When measured over large numbers of specimens, or across multiple conditions, the resulting large

dataset can be interpreted using powerful algorithms to discriminate not only patterns in the responses of genes, but also patterns in the responses provided by specimens. Just as transcript screening across thousands of probes can help diagnose cancer subtypes in humans, it can also discriminate between even subtly different phenotypes and explain the basis of these differences.

However, there are both advantages and disadvantages to such sensitivity. FIGURE 2 provides an example where specimens from a single group of common carp (*C. carpio*) that were held under constant conditions were sampled on three successive occasions, with microarray-based assessment of RNA expression in the liver. HIERARCHICAL ANALYSIS clustered the specimens according to the day of sampling, and FIG. 2 shows a minimal set of genes that discriminates between different days (A.Y. Gracey and E.J. Fraser, personal communication). These day-to-day variations in

expression probably relate to minor variations in husbandry in an otherwise controlled procedure, and clearly much greater care must be taken to control sources of inter-individual and technical variation in microarray experiments.

On the other hand, this ability to discriminate extremely subtle differences between specimens is likely to prove useful in identifying differences between populations. FIGURE 1 shows an example where genes that can be used to differentiate between distinct populations of a particular species can be readily identified by statistical analysis and pattern searching. One area in which such techniques are likely to prove useful is in investigating the exposure of fish to toxic chemicals in the environment, as the wide range of man-made pollutants present different physiological challenges and induce different gene responses. Provided transcript responses are distinctive enough, they can be used to assess both exposure and effect, and to predict the likely toxicological properties of untested pollutants.

Understanding the evolution of complex responses. Many key environmental responses at the cellular and sub-cellular levels clearly arose early in evolution, particularly those that confer protection of native states of proteins, nucleic acids and cellular membranes. Mechanisms of this type can be found in fish and other vertebrates, and these can be identified by comparing gene-expression responses in fish to those seen in model organisms from other distant taxonomic groups, such as *Caenorhabditis elegans*, *Drosophila melanogaster* and yeast¹². Nevertheless, the increasing complexity of multicellular organisms, and vertebrates in particular, has required the evolution of complex homeostatic systems that operate at the level of the whole organism. In principle, the properties of these organ systems, and especially of the homeostatic regulation offered by endocrine and neuroendocrine systems, are as amenable to genomic analysis as sub-cellular systems. This offers a means of profiling the physiological phenotype that relates to properties of the whole animal. This level of integration has not yet been achieved, but is likely to be crucially important in understanding responses to naturally occurring stressors, such as temperature change, and in understanding the evolution of these responses.

During the evolution of complexity, gene, chromosome and whole-genome duplications have had key roles in shaping morphological and developmental processes⁵. In fish, the obvious example is the role of the highly

Box 4 | Existing resources for fish genomics

Fish have contributed much to genomic sequence databases and represent a sizeable fraction of all vertebrate genomic information. Nevertheless, the number of species with sufficient resources to carry out intensive genomic investigations is small, and is almost entirely deficient for some important taxonomic groups. Existing genomic resources for fish are discussed below and the phylogenetic distribution of these resources is shown in FIG. 3.

Sequencing projects

Four genome-sequencing projects are underway for bony fish. The zebrafish genome currently has 5.7x coverage, but is proving difficult to assemble due to the use of pooled DNA from many polymorphic individuals. Two of the other species being sequenced, *Takifugu rubripes* (formerly designated as *Fugu rubripes*) and *Tetraodon nigroviridis*, are closely related members of the pufferfish family, selected, because of their very small genome sizes, as models for the human genome project. The current *T. rubripes* genome assembly covers ~95% of the non-repetitive fraction of the genome (~365 Mb)⁵¹, and the *T. nigroviridis* genome currently represents ~70% of the genome³⁴. The most recent project is the sequencing of the medaka (*Oryzias latipes*) genome, and others planned for the next few years include the sea lamprey (*Petromyzon marinus*) and the three-spined stickleback (*Gasterosteus aculeatus*).

Large-insert libraries

Production of large-insert (BAC, YAC and fosmid) libraries is a prelude to full genome sequencing, and also provides essential tools for genetic mapping, including marker-assisted selection and positional cloning of QTLs for environmentally interesting traits. The number of these libraries, mainly BAC libraries, is rapidly increasing. They cover almost all taxonomic groups of fish and biologically interesting species, as well as the species for which whole-genome sequencing is currently being carried out³². The US National Institutes of Health National Center for Biotechnology Information dbEST database lists a large number of other fish species that are not included in FIG. 3, with smaller numbers of EST sequences. Species that are likely to be given priority for developing genomic resources in the future include the gilthead seabream (*Sparus aurata*), paddlefish (*Polyodon spathula*), sea lamprey (*P. marinus*), blind cavefish (*Astyanax mexicanus*) and the shark *Ginglymostoma cirratum*.

EST resources

Over one million teleost EST sequences are currently listed in the dbEST database and the list is rapidly increasing. However, this resource is unevenly distributed throughout the fish phylogeny. Half the teleost EST resource, ~0.5 x 10⁶ sequences, has been provided for zebrafish, which reflects the well-funded research community for this species. The US National Institutes of Health have notably provided ~6,000 full-length cDNA sequences for zebrafish (see the link to the **Zebrafish Gene Collection** in the Online links box) which have considerable value for mapping non-overlapping EST sequences of homologues in other species. Salmonid fish are also well represented in the database, with more than 150,000 entries. Finally, there is a larger number of species with more modest but still useful EST collections of ~10,000 per species, including the common carp (*Cyprinus carpio*), the minnow (*Fundulus heteroclitus*), various tilapia species, the stickleback (*G. aculeatus*), the flounder (*Platichthys flesus*) and the goby *Gillichthys mirabilis*.

but the first — a straightforward expansion of genomic resources for fish species — is likely to have the most impact.

Expanding fish genomic resources. Undoubtedly, the key resources for genomic analysis in fish are the emerging full-genome sequences and the growing list of cDNAs. BOX 4 outlines current genome-sequencing projects and cDNA availability for fish. The fact that the species that are being sequenced are distributed across different fish taxa is fortunate (FIG. 3), as they will increasingly be used as reference points for analyses of gene and protein identity, and for exploring gene diversity across the full taxonomic range of fish.

EST collections²⁹ are essential for microarray-based studies, and these can now be built up from cDNA libraries by individual laboratories at relatively modest cost (BOX 4; FIG. 3). They can also be mapped against the many EST resources available for model vertebrate species, particularly zebrafish¹³, but also including humans and mice, to explore orthologous relationships and to co-locate non-overlapping ESTs. The advantage of custom production of cDNA libraries, clone collections and microarrays is flexibility. Efforts can be directed at specific questions and specific tissues, using subtractive techniques where appropriate to enrich libraries for genes that show different expression patterns between physiological states, experimental conditions, and even between populations or species.

The bedrock of microarray and EST provision for fish is likely to remain with academic research laboratories because, with the exception of zebrafish, there is insufficient demand to constitute a commercial opportunity. Most microarray production is currently based on the use of cDNA probes, but as sequence information increases in quality and coverage it is likely that oligonucleotide probes — currently commercially available only for zebrafish — will become an increasingly important option. Oligonucleotide probes provide increased specificity compared with cDNA probes, allowing isoforms, paralogues and different members of gene families to be distinguished, because they can be directed exclusively at non-conserved parts of a transcript. cDNA microarrays produced for one species are also useful for screening RNA from related species³⁰, and cross-species heterologous probes provide the most cost-effective means of screening species for which probes are not available, including many fish species. However, the use of such probes will certainly be at the expense of fidelity in discriminating between closely

duplicated genes of the *HOX* CLUSTER in generating morphological complexity²⁵. Duplication of genes and the subsequent diversification of function and genetic fixation might also have been important in the evolution of increasingly complex responses to environmental challenge. For example, in *C. carpio*, a recent (~15 Mya) duplication of a key cold-inducible gene, the acyl-CoA desaturase gene, has given rise to isoforms that differ in their physiological regulation²⁶, probably owing to the redistribution, or 'subfunctionalization', of several promoter responses of the ancestral gene among the two duplicates²⁷. Indeed, studies of the greater number of gene paralogues in fish compared with mammals have offered insights into gene function in higher organisms. For example, fish have several paralogues

of the *SOX* gene, each with a specialized pattern of expression that allows a dissection of responses that is not possible with the more pleiotropic mammalian models²⁸. Assessment of other replicated gene groups might generate insights into complex environmental responses.

An advanced genomics toolkit for fish

Improved availability of genomic resources and tools for functional genomics will be required for future advances in fish environmental genomics. As discussed previously, large-scale genome resources are restricted to just a few key fish species, and for most others resources are limited. Here, we suggest several options for meeting these needs. These options can be combined where appropriate,

related sequences, because they often contain consensus regions. Carefully designed oligoprobes might overcome this problem while still offering the potential for heterologous screening. In addition, the availability of custom-designed oligoarrays³¹ that contain several probes against each gene, from coding and non-coding regions, will greatly increase the options for achieving adequate hybridization between fish species.

Developing a comprehensive genome architecture for fish.

Despite the absence of DNA sequence data for most fish species, much can be achieved by relating genes and chromosomal locations of interest to the genomic model species that is phylogenetically most closely related. A similar physical mapping of genomes from different plant taxa has transformed plant science through comparative genomics^{32,33}, providing a direct means of establishing the likely SYNTENY for any gene or chromosomal location and greatly expanding the range of species for which positional cloning is an option. Recent comparative analyses of *T. nigroviridis*³⁴ and *T. rubripes*³⁵ with the human genome have provided broad insights into genome evolution and have highlighted conserved coding and non-coding features. Further studies of this type will become possible when the zebrafish genome is finally published. Just how useful the zebrafish physical map will be to species outside the *Cypriniformes* family remains to be seen because, until now, there has been little attempt to derive a comparative physical map of genomes from each of the main fish lineages. Doing so would not only provide a useful resource for positional cloning³⁶, but would also provide important new information on how genome architecture and gene complements have changed over evolutionary time and how this relates to lifestyle and environmental susceptibility.

Unfortunately, the necessary tools for physical mapping, such as RADIATION HYBRID PANELS (RH panels)³⁷, are available for few fish species, which currently include zebrafish (*D. rerio*), rainbow trout (*Oncorhynchus mykiss*), medaka (*O. latipes*) and more recently the sea bream (*Sparus aurata*). Production of RH panels is fraught with difficulty, and the newer HAPPY MAPPING method offers an easier alternative³⁸. Another complication, at least with some groups of fish, is the occurrence of large-scale duplications either of chromosomes or of whole genomes^{25,39}, making both physical and genetic mapping more difficult.

If altered gene expression is important for environmental responses, then knowing

which transcription factors effect these changes is the basis for understanding the upstream regulatory events. The identification and analysis of gene regulatory regions is aided in unsequenced species by the availability of large-insert BAC, YAC and FOSMID genomic libraries, especially when they are mapped onto a genome of the same species or that of the nearest model species. BAC libraries are available for a rapidly increasing number of

fish species (FIG. 3) and, through physical mapping, will form the most convenient means of comparative analysis.

Gene manipulation. Expanding the use of gene-manipulation techniques to a wider range of fish species will be important in exploring the function of candidate genes identified from screening studies and is an important aim for the future. The production

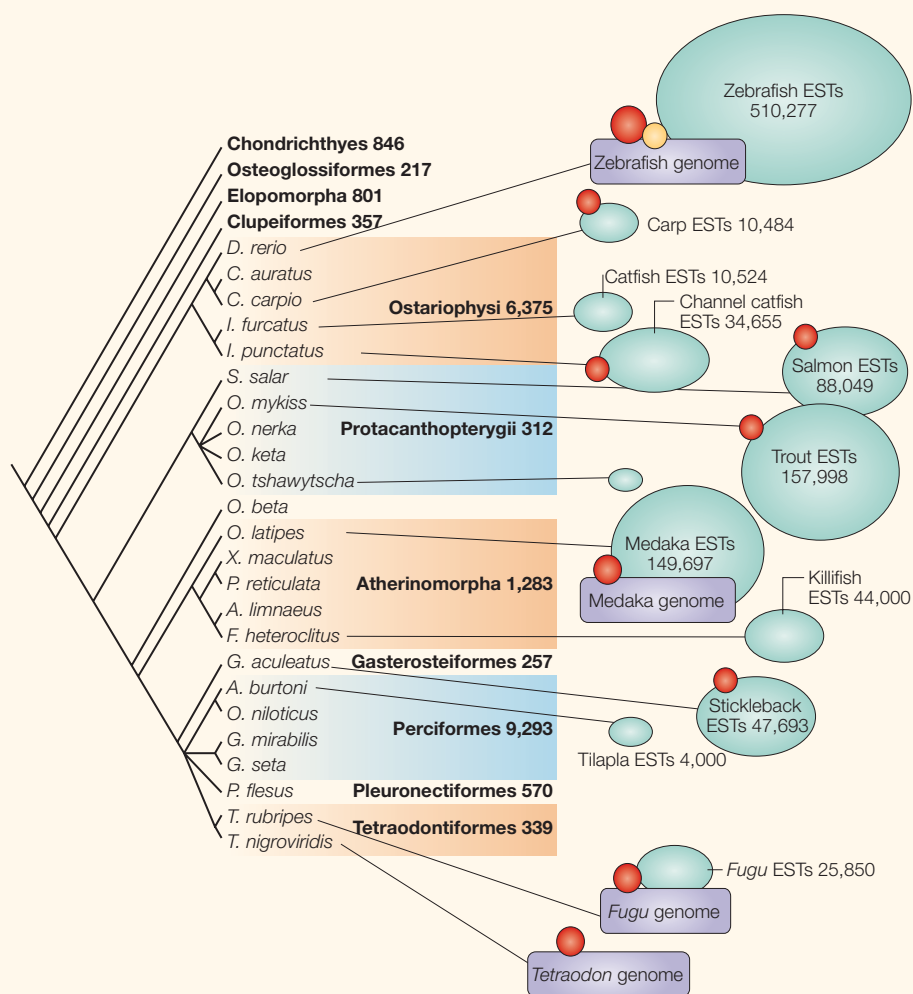


Figure 3 | Distribution of fish genomic resources. The distributions of genome-sequencing projects, ESTs (species with >1,000 submissions in the dbEST database), large-insert libraries and radiation hybrid panels are shown in relation to fish phylogeny. Fish comprise ~29,000 species that are distributed between more than 120 families, of which the largest are displayed, together with an indication of the number of species they contain. The cladogram illustrates a conventional phylogeny for fish, and the coloured boxes, ovals and circles to the right indicate genome-sequencing projects (in purple), ESTs (in green, where size indicates the number of ESTs in publicly available collections), large-insert libraries (mainly BACs, in red) and radiation hybrid panels (in yellow). Data for ESTs were taken from the dbEST database and by using Google to search web sites. Resources are rapidly increasing worldwide and this illustration is neither complete nor comprehensive. The cladogram is based on data from REF. 53 and was constructed by M. Berenbrink. *A. burtoni*, *Astatotilapia burtoni*; *A. limnaeus*, *Austrofundulus limnaeus*; *C. auratus*, *Carassius auratus*; *C. carpio*, *Cyprinus carpio*; *D. rerio*, *Danio rerio*; *F. heroclitus*, *Fundulus heroclitus*; *G. aculeatus*, *Gasterosteus aculeatus*; *G. mirabilis*, *Gillichthys mirabilis*; *G. seta*, *Gillichthys seta*; *I. furcatus*, *Ictalurus furcatus*; *I. punctatus*, *Ictalurus punctatus*; *O. beta*, *Opsanus beta*; *O. keta*, *Oncorhynchus keta*; *O. latipes*, *Oryzias latipes*; *O. mykiss*, *Oncorhynchus mykiss*; *O. nerka*, *Oncorhynchus nerka*; *O. niloticus*, *Oreochromis niloticus*; *O. tshawytscha*, *Oncorhynchus tshawytscha*; *P. flesus*, *Platichthys flesus*; *P. reticulata*, *Poecilia reticulata*; *S. salar*, *Salmo salar*; *T. nigroviridis*, *Tetraodon nigroviridis*; *T. rubripes*, *Takifugu rubripes*; *X. maculatus*, *Xiphophorus maculatus*.

of transgenic zebrafish or medaka by micro-injection is now widely used in developmental research, and other more efficient methods might soon become available. These include the injection of gene constructs or even large-insert clones into fertilized eggs using a mega-nuclease to increase the frequency of chromosomal integration⁴⁰, which would allow complementation of gene function to be tested directly, even between species. Other emerging methods for genetic manipulation that might be of particular value in fish research include the use of embryonic stem cells, as are already used in mouse research⁴¹, and also the wider use of RNAi or morpholinos to suppress the expression of specific genes and explore environmentally relevant phenotypes⁴². Finally, cultured cell lines will provide an important means of investigating the expression of important genes and their encoded proteins. The range of continuous cell lines that is available for most fish species is either limited or non-existent, and the expansion of this range is therefore a key requirement for the future.

Translating environmental questions onto genomic model species. Despite the potential of the strategies described above, there is no doubt that, because of their experimental tractability, zebrafish and medaka offer significant advantages for the rapid identification of genes and the exploration of their functions. There is some scepticism about the usefulness of intensively inbred strains, because they are phenotypically different from their wild counterparts. However, laboratory populations can show relevant environmental responses and yield penetrating insights into underpinning mechanisms, a good example being hypoxia responses in zebrafish¹¹. Indeed, relating fish to more distantly related eukaryotes, such as the yeast *Saccharomyces cerevisiae*⁴³, allows identification of conserved elements of stress responses that are likely to be relevant to all fish species¹². Comparisons between distantly related taxa also allow the more convenient exploration of the phenotypic importance of candidate genes. For example, gene manipulation by mutation and RNAi in *C. elegans* has been used to test the extent to which the cold-inducible acyl-CoA desaturase gene, which has been well-studied in fish models, mediates acquired cold tolerance (P.A. Murray *et al.*, unpublished observations). Therefore, the combination of hypothesis generation in a fish followed by hypothesis testing in a model species offers a useful means of increasing understanding.

Conclusions

Genomic approaches using fish are in their infancy. Despite this, 4 fish genome sequences are almost complete, over 1.5 million fish ESTs are available in publicly accessible databases, and there are more than a dozen species for which >10,000 ESTs are already available. These resources have already led to several important advances: we have a much better understanding of how the environment affects gene expression^{10–14}, the conserved and differentiated features of genomic responses to specific environmental stresses are becoming clear and new candidate genes

are being identified¹², the influence of developmental programmes on environmental susceptibility²³ is being established and the significance of variation in gene expression within and between populations is being addressed^{18,21,44}.

Fish genomics undoubtedly benefits from the wide range of species and of environmental conditions experienced. However, realizing this benefit will require a significant expansion of species with genomic and genetic tractability. This will be achieved for several key species in the near future, including the killifish (*F. heteroclitus*), sheepshead minnow

Glossary

FIXATION

The increase in the frequency of a genetic variant in a population to 100%.

FOSMID

A low-copy vector for the construction of stable genomic libraries that uses the *Escherichia coli* F-factor origin for replication.

GENE ONTOLOGY

A hierarchical organization of concepts and nomenclature for molecular function, biological processes and cellular components. It constitutes a controlled vocabulary with orderly relationships between parent and daughter terms, onto which known genes are mapped, and provides a useful means of categorizing gene lists into functionally meaningful groupings.

GENETIC DRIFT

(Also known as random drift.) A phenomenon whereby the frequency of a gene in a population changes over time because the number of offspring born to parents that carry the gene is subject to chance variation.

HAPPY MAPPING

A simple method for ordering markers and determining the physical distances between them. It uses sub-haploid equivalents of randomly sheared DNA and requires the use of whole-genome amplification methods to carry out many PCR reactions.

HIERARCHICAL ANALYSIS

An organization or 'clustering' of elements that best describes the relationships between them. A tree diagram or dendrogram is frequently used to represent the results of a cluster analysis, with cases of greatest similarity being adjacent to each other.

HOX CLUSTER

A family of genes involved in directing the morphological development of the body during early stages of life. In vertebrates they are clustered together on defined chromosomes and are widely believed to have originated by extensive duplication. The order of developmental expression over time is related to the position on the chromosome.

ISOGENIC

Cells or organisms that are derived from inbreeding or by genetic manipulation, and have identical or almost identical genomes.

KEGG SCHEME

A database comprising a collection of graphical pathway maps for metabolism, regulatory processes and other biological processes.

MORPHOLINO

A non-degradable antisense oligonucleotide in which the sugar component is replaced by a morpholine ring structure. Morpholinos are currently used to block target-gene expression in zebrafish, *Xenopus laevis* and sea urchins. They bind stably to target mRNAs in order to block translation, and give more consistent phenotypes than traditional antisense oligonucleotides.

NEUTRAL VARIATION

Variation in protein sequence that is not selectively important.

NEUTRAL HYPOTHESIS

An evolutionary model that assumes that the selective advantage of the variation in a trait is insufficient to provide any fitness advantage. Changes in allele frequency are said to be the result of chance alone, or 'drift'.

PARALOGY

Paralogous genes show homology because they originated as a result of duplication of a single ancestral gene.

QTL

(Quantitative trait loci). A genetic locus that is identified through the statistical analysis of quantitative traits (such as height in plants or body weight in animals). These traits are typically affected by more than one gene and also by the environment.

RADIATION HYBRIDS

Cells produced by fusing irradiated donor cells that contain chromosomal fragments with recipient rodent cells to produce a panel of cell lines. Each of these cell lines contains fragments of donor chromosomes, which can be used to screen for physical linkage of genetic or physical markers. The resulting maps are indispensable tools for the positional cloning of candidate genes.

STANDING GENETIC VARIATION

The genetic variation or heterozygosity that occurs between individuals in laboratory or natural populations.

SYTENY

Collinearity in the order of genes (or of other DNA sequences) in a chromosomal region of two species.

(*Cyprinodon variegatus*), three-spined stickleback (*Gasterosteus aculeatus*), rainbow trout (*O. mykiss*), various salmon species, the flounder (*P. flesus*) and the common carp (*C. carpio*). However, broadening this to encompass other species, and generalizing this knowledge, will require a much greater coordination of what is currently a fragmented research effort. An important goal for allowing easier translation of research problems from one species to another and one taxon to another would be a broad understanding of genome architecture across all fish, as has been achieved in plants. The need for cooperation across national boundaries to achieve these aims has been recognized previously^{45,46}, but until the relevant national funding agencies provide a similar level and coherency of resources to that enjoyed in other research areas, progress will be slow.

The new post-genomic screening techniques and the availability of extensive DNA sequence data will undoubtedly generate a wealth of new mechanistic data on the nature of environmental responses, and on how this relates to and is limited by genotype. Charting how transcript and protein expression profiles vary between individuals, populations and species, and relating this to tolerance, might point to the principal determinants of adaptive phenotypes, a key goal of environmental genomics. However, one more immediate and practical benefit of fish environmental genomics is likely to be felt in the field of ecotoxicology, given the focused interest of both regulatory authorities and industry. In the short term, this will provide much more detailed descriptions of responses to toxic exposure⁴⁷.

In the longer term, a detailed knowledge of responses to known ecotoxicants will provide scientists with a useful means of predicting toxic responses of compounds or mixtures of compounds and of identifying compounds that are responsible for pollution events, thereby generating the data necessary for regulatory approval. In the future, 'ecotoxicogenomics' might inform environmental management strategies, and provide regulatory bodies with alternative, robust means of defining ecological effects, specifying standards and monitoring compliance. Delivering these outcomes in this and other areas of environmental genomics will require a clearer focus of research activity. It will also necessitate a much better connected research community that can coordinate its activities in terms of both genomic resources and data, and in developing the technologies required for genetic manipulation of environmental fish models.

Andrew R. Cossins is at the Liverpool Microarray Facility and Centre for BioArray Innovation, School of Biological Sciences, University of Liverpool, Crown Street, Liverpool L69 7ZB, United Kingdom.

Douglas L. Crawford is at the Centre for Marine Genomics, Rosenstiel School of Marine and Atmospheric Sciences, 4600 Rickenbacker Causeway, Miami, Florida 33149, USA.

**Correspondence to A.R.C.
e-mail: cossins@liverpool.ac.uk**

doi:10.1038/nrg1580

- Gracey, A. Y. & Cossins, A. R. Application of microarray technology in environmental and comparative physiology. *Annu. Rev. Physiol.* **65**, 231–259 (2003).
- Randall, D. J., Burggren, W. & French, K. *Animal Physiology: Mechanisms and Adaptations* (W. H. Freeman, New York, 2002).
- Oleksiak, M. F., Kolell, K. J. & Crawford, D. L. The utility of natural populations for microarray analyses: isolation of genes necessary for functional genomic studies. *Mar. Biotech.* **3**, S203–S211 (2001).
- Berenbrink, M., Koldkjaer, P., Kepp, O. & Cossins, A. R. Evolution of complex systems: oxygen secretion in fish. *Science* 18 March 2005 [epub ahead of print].
- Venkatesh, B. Evolution and diversity of fish genomes. *Curr. Opin. Genet. Dev.* **13**, 588–592 (2003).
- Robinson-Rechavi, M. *et al.* Euteleost fish genomes are characterised by expansion of gene families. *Genome Res.* **11**, 781–788 (2001).
- Powers, D. A. Fish as model systems. *Science* **246**, 352–358 (1989).
- Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Rev. Genet.* **5**, 288–298 (2004).
- Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. *Nature Rev. Genet.* **5**, 456–465 (2004).
- Gracey, A. Y., Troll, J. V. & Somero, G. N. Hypoxia-induced gene expression profiling in the euryoxic fish *Gillichthys mirabilis*. *Proc. Natl Acad. Sci. USA* **98**, 1993–1998 (2001).
- Ton, C., Stamatou, D. & Liew, C.-C. Gene expression profile of zebrafish exposed to hypoxia during development. *Physiol. Genomics* **13**, 97–106 (2003).
- Gracey, A. Y. *et al.* Coping with cold: an integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. *Proc. Natl Acad. Sci. USA* **101**, 16970–16975 (2004).
- Ju, Z., Dunham, R. A. & Liu, Z. Differential gene expression in the brain of channel catfish (*Ictalurus punctatus*) in response to cold acclimation. *Mol. Genet. Genomics* **268**, 87–95 (2002).
- Podrabsky, J. E. & Somero, G. N. Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish *Austrofundulus limnaeus*. *J. Exp. Biol.* **207**, 2237–2254 (2004).
- Hochachka, P. W. & Somero, G. N. *Biochemical Adaptation* (Princeton Univ. Press, New Jersey, 1984).
- Somero, G. N. & Hand, S. C. Protein assembly and metabolic regulation: physiological and evolutionary perspectives. *Physiol. Zool.* **63**, 443–471 (1990).
- Pierce, V. A. & Crawford, D. L. Phylogenetic analysis of glycolytic enzyme expression. *Science* **276**, 256–259 (1997).
- Oleksiak, M. F., Churchill, G. A. & Crawford, D. L. Variation in gene expression within and among natural populations. *Nature Genet.* **32**, 261–266 (2002).
- Schulte, P. M., Glemet, H. C., Flebig, A. A. & Powers, D. A. Adaptive variation in lactate dehydrogenase-B gene expression: role of a stress-responsive regulatory element. *Proc. Natl Acad. Sci. USA* **97**, 6597–6602 (2000).
- Williams, T. D., Gensberg, K., Minchin, S. D. & Chipman, J. K. A DNA expression array to detect toxic stress response in European flounder (*Platichthys flesus*). *Aquat. Toxicol.* **65**, 141–157 (2003).
- Oleksiak, M. F., Roach, J. L. & Crawford, D. L. Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. *Nature Genet.* **37**, 67–72 (2005).
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
- Shapiro, M. D. *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723 (2004).
- Nacci, D. E., Champlin, D., Coiro, L., McKinney, R. & Jayaraman, S. Predicting the occurrence of genetic adaptation to dioxin-like compounds in populations of the estuarine fish *Fundulus heteroclitus*. *Env. Toxicol. Chem.* **21**, 1525–1532 (2002).
- Amores, A. *et al.* Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**, 1711–1714 (1998).
- Polley, S. D. *et al.* Differential expression of cold-specific and diet-specific genes encoding two isoforms of the $\delta 9$ -acyl-CoA desaturase in carp liver. *Am. J. Physiol.* **284**, R41–R50 (2002).
- Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
- Chiang, E. F. L. *et al.* Two Sox9 genes on duplicated zebrafish chromosomes: expression of similar transcription activators in distinct sites. *Dev. Biol.* **231**, 149–163 (2001).
- Zweiger, G. & Scott, R. From expressed sequence tags to 'epigenomics': an understanding of disease processes. *Curr. Opin. Biotechnol.* **8**, 684–687 (1997).
- Renn, S. C. P., Aubin-Horth, N. & Hofmann, H. A. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics* **5**, 42 (2004).
- Baum, M. *et al.* Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res.* **31**, e151 (2003).
- Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Cereal genome evolution- grasses, line up and form a circle. *Curr. Biol.* **5**, 737–739 (1995).
- Gale, M. D. & Devos, K. M. Plant comparative genetics after 10 years. *Science* **282**, 656–659 (1998).
- Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957 (2004).
- Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
- Meyers, B. C., Scalabrin, S. & Morgante, M. Mapping and sequencing complex genomes: let's get physical! *Nature Rev. Genet.* **5**, 578–588 (2004).
- Schuler, G. *et al.* A map of the human genome. *Science* **274**, 540–546 (1996).
- Dear, P. H., Bankier, A. T. & Piper, M. B. A high-resolution metric HAPPY map of human chromosome 14. *Genomics* **48**, 232–241 (1998).
- Meyer, A. & Schartle, M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11**, 699–704 (1999).
- Epinat, J. C. *et al.* A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res.* **31**, 2952–2962 (2003).
- Fan, L. C., Alestrom, A., Alestrom, P. & Collodi, P. Development of cell cultures with competency for contributing to the zebrafish germ line. *Crit. Rev. Eukaryot. Gene Expr.* **14**, 43–51 (2004).
- Nasevicius, A. & Ekker, S. C. Effective targeted gene 'knockdown' in zebrafish. *Nature Genet.* **26**, 216–220 (2000).
- Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
- Whitehead, A. & Crawford, D. Variation in tissue-specific gene expression among natural populations. *Genome Biology* **6**, R13 (2005).
- Rayl, A. How to create a successful fish tale? *Scientist* **15**, 11–12 (2001).
- Clark, M., Crawford, D. L. & Cossins, A. Worldwide genomic resources for non-model fish species. *Comp. Func. Genomics* **4**, 502–508 (2003).
- Snape, J. R., Maund, S. J., Pickford, D. B. & Hutchinson, T. H. Ecotoxicogenomics: the challenge of integrating genomics into aquatic and terrestrial ecotoxicology. *Aquat. Toxicol.* **14**, 143–154 (2004).
- Johnston, I. A., Vieira, V. L. A. & Temple, G. K. Functional consequences and population differences in the developmental plasticity of muscle to temperature in Atlantic herring *Clupea harengus*. *Marine Ecol. Prog. Ser.* **213**, 285–300 (2005).
- Nilsson, G. E. & Lutz, P. L. Anoxia tolerant brains. *J. Cereb. Blood Flow Metab.* **24**, 475–486 (2004).

50. Helfman, G., Collette, B. & Facey, B. *The Diversity of Fishes* (Blackwell Science, Malden, Massachusetts, 1997).
51. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
52. Katagiri, T. *et al.* Construction and characterization of BAC libraries for three fish species; rainbow trout, carp and tilapia. *Anim. Genet.* **32**, 200–204 (2001).
53. Nelson, J. S. *Fishes of the world* (John Wiley and Sons, New York, 1994).

Acknowledgements

We thank M. Berenbrink for helpful discussions and comments and anonymous referees for comments. A.R.C. was supported by long-term funding from the UK Natural Environmental Research Council who also have supported the Liverpool Microarray Facility. D.L.C. was supported by grants from the US National Science Foundation Biocomplexity Programme and the US National Heart Lung and Blood Institute.

Competing interests statement
The authors declare no competing financial interests.

Online links

FURTHER INFORMATION

dbEST — Database of Expressed Sequence Tags: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html
Fishbase: www.fishbase.org
Gene Ontology: <http://www.geneontology.org>
KEGG — Kyoto Encyclopedia of Genes and Genomes: www.genome.jp/kegg
Medaka Genome Project: <http://dolphin.lab.nig.ac.jp/medaka>
Tetraodon Genome Browser: www.genoscope.cns.fr/externe/tetraodon
The Danio rerio Sequencing Project: www.sanger.ac.uk/Projects/D_rerio
The Fugu Genomics Project: <http://fugu.hgmp.mrc.ac.uk>
Zebrafish Gene Collection: <http://zgc.nci.nih.gov>
Access to this interactive links box is free online.

OPINION

The Human Genome Diversity Project: past, present and future

L. Luca Cavalli-Sforza

Abstract | The Human Genome Project, in accomplishing its goal of sequencing one human genome, heralded a new era of research, a component of which is the systematic study of human genetic variation. Despite delays, the Human Genome Diversity Project has started to make progress in understanding the patterns of this variation and its causes, and also promises to provide important information for biomedical studies.

The **Human Genome Diversity Project** (HGDP) provides a resource that is aimed at promoting worldwide research on human genetic diversity, with the ultimate goal of understanding how and when patterns of diversity were formed. It also has the added benefit of providing information that is likely to prove useful to several areas of biomedical research. Here, I provide an update on the HGDP, focusing on important progress since earlier reviews¹, the present status of the project² and how this resource could be developed most effectively in the future. I also discuss possible relations with the **International HapMap Project**^{3,4}, another large-scale study of human genome diversity. Despite having generally different aims, the two projects provide complementary resources, indicating that interactions between the two could prove mutually beneficial. Finally, I summarize some desirable future developments for the HGDP

and its potential to improve the understanding of the genetic structure of the human species and facilitate medical applications.

The HGDP — rationale and history

Founding of the HGDP. As early as the beginning of the twentieth century, the potential for genetic data to provide information on the history and geography of human populations was known from the study of proteins (BOX 1). However, until recently, the collection of such data remained largely a piecemeal endeavour. Indeed, it was not until the **Human Genome Project** (HGP) was in full swing that the idea of a large-scale systematic study of human genomic variation was raised⁵. Specifically, it was realized that renewable samples from well-chosen populations, for which any part of the genome could be examined, could greatly facilitate studies of the genetic geography and history of our species.

The foundation of the HGDP was prompted by discussions among geneticists interested in human evolution and population genetics⁶. The President of the **Human Genome Organisation** (HUGO) at that time, Sir Walter Bodmer, asked me to chair a committee to study the feasibility of a human genomic variation project. As the idea developed, this was named the Human Genome Diversity Project. The US National Institutes of Health (NIH) Institute for General Medical Science, the US National Science Foundation, and initially also

the US Department of Energy (which had previously financed mutation rate studies by my group), supported four symposia, between 1991 and 1994, that addressed the genetic and statistical issues, anthropological issues, general organization, and molecular and ethical issues related to the HGDP.

Overcoming initial difficulties. Political and ethical difficulties arose¹ in 1994, similar to those that marked the beginning of the HGP, but in the case of the HGDP they focused especially on the fear that indigenous people might be exploited by the use of their DNA for commercial purposes ('bio-piracy'). However, since its initiation, the HGDP has avoided commercial interests, and when the project was finally ready to be launched, it was made clear that DNA samples would be provided only to non-profit-making laboratories. The HGDP has always opposed the patenting of DNA, to allow the study of genetic variation for fundamental research purposes. Concern that HGDP data would feed 'scientific racism' was also expressed by naive observers, despite the fact that half a century of research into human variation has supported the opposite point of view — that there is no scientific basis for racism. Consequently, agencies that had financed the HGDP organizational symposia asked the US **National Research Council** (NRC) of the National Academy of Sciences (NAS) to convene a committee to study the feasibility and ethics of the project, similar to the evaluation of the HGP at a comparable time in its history.

From 1994 to 1997, while the NRC committee was organized, met and wrote its report, the HGDP took no major action. Instead, it prepared for the final stage of organization, in the hope and expectation that the NRC committee would give a positive response. In particular, at the first Cold Spring Harbor (CSH) Symposium on Human Evolution, held in October 1997, there was a meeting of several research workers who had collected cell lines from indigenous populations. At my request, they unanimously agreed to contribute cell lines to a central collection that would form the core of the HGDP.

Since its initiation, the organizers of the HGDP were convinced that the crucial first effort was to establish a collection of **LYMPHOBLASTOID CELL LINES** (LCLs) from many populations — rather than simply collecting DNA samples — for reasons of accuracy and renewability (BOX 2). The fact that LCLs had already been made from worldwide populations by researchers of human evolution also supported the validity of the approach, and

the donation of these lines to the HGDP made immediate funding unnecessary.

Uncertainties concerning such issues as strategies for collecting samples in a way that would facilitate anthropological or medical research, and the choice of the populations¹, were obviated by the sources of the cell lines, which were donated by researchers working on human evolution. However, as discussed later, it became clear that orientation of the collection towards anthropological interests also offered excellent chances of aiding medical research.

The recommendation of the NAS-NRC committee, made public⁷ at the end of 1997, was that the HGDP could proceed, with particular attention being paid to informed consent and related ethical issues. The NIH Institute of General Medical Sciences, a chief supporter throughout, has constantly followed and revised the ethical rules of the endeavour (BOX 3).

Two important problems had to be solved at the time when the NAS-NRC authorization finally came. The first was the question of where to house the HGDP collection. The Center for the Study of Human Polymorphism (CEPH) at the **Fondation Jean Dausset** in Paris agreed to house and distribute the collection. In 1984, the CEPH had initiated the international collaboration to genetically map the human genome⁸, which was built around a resource of LCLs from 40 large kindreds, and so had all the facilities needed for storing cell lines and distributing large numbers of DNA samples.

The second question concerned whether the collection was adequate for the intended research purposes. This could be decided only after all the LCLs had arrived in Paris. All five continents are represented in the collection, and all samples are from populations of anthropological interest — that is, those that were in place before the great diasporas started in the fifteenth and sixteenth centuries, when navigation of the oceans became possible. This choice was important, because these diasporas caused significant population admixtures, especially in the Americas but also in other continents. Only genetic knowledge of the original populations that contributed to these admixtures can disentangle the various genetic complexities that resulted, and the HGDP fulfils these criteria.

The HGDP collection was to include more than 1,000 cell lines; inevitably, there would be large gaps given the collection strategy. However, it was important to begin the project, as determining the success of the initial collection was thought to be essential for understanding if and how it would be worth expanding.

Box 1 | Some general principles of human genetic diversity studies

Gene-based studies of genetic variation

Data on human genetic variation, collected since 1919 on proteins^{31–34}, and more recently on DNA, have been widely used for the reconstruction of human history^{9,15,25}, calculating genetic distances between populations from their gene frequencies and averaging them for many genes. Mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY) are transmitted by only one parent (mothers and fathers, respectively), are haploid and do not undergo recombination. Their evolution by genetic drift is therefore four times more rapid than that of autosomal genes, and they have proved particularly informative about the time range that is most useful for studies of modern human evolution — the last 100,000–150,000 years.

Population trees versus haplotype genealogies

There are two general methods for the evolutionary analysis of genetic variation. Population tree analysis usually starts by calculating genetic distances between pairs of populations from the differences in frequencies of genetic variants, averaged over many genes. Trees describe the history of population splits, after which populations diverge (it is assumed) independently. However, migration reduces distances between populations, especially those that are geographically close, generating distortions of the trees²⁴.

The other method uses genealogies of haplotypes. They are reconstructed using information from individuals rather than populations, and the branchpoints of the trees that are generated correspond to specific mutations. These genealogies are not affected by migration, and those that are based on patterns of less mutable SNPs and on insertion or deletion mutations are particularly rigorous. The dates of branching points can be calculated on the basis of knowledge of mutation rates, assuming an absence of selection. Microsatellites have much higher mutation rates and are more useful for evaluating more recent branching dates.

Principal-components analysis

This is an alternative approach that is especially useful for graphical purposes. PRINCIPAL COMPONENTS (PCs) are variables that summarize information about independent patterns that are present in a data matrix, which might be difficult to detect otherwise. PCs are ranked by the amount of information each of them reveals in the data that is being analysed. The first two or three might easily summarize more than 50% of the information that is contained in hundreds of gene frequencies²⁴. When single populations are plotted as dots on a Cartesian diagram using the first two PC values as coordinates, populations are clearly clustered by genetic similarity. If reciprocal migrations have been important in establishing similarity patterns after the principal historical separations of populations, the diagram of the first two PCs resembles their geographical map. Another mode of display is a geographical map of the values of a single PC (similar to a map of altitudes or ocean depths). This is especially useful for recognizing clines (areas of gradual variation) that occur owing to important, long-lasting migratory patterns that affect all genes equally. However, clines that are due to natural selection affect only one or a few genes. PC geographical maps have therefore been used for tracing demic expansions that started at different places and times in the same general area, and their centres of origin^{24,29}.

Current status and successes so far

The establishment of the HGDP collection, the list of populations included in it, and the conditions for obtaining DNA samples were announced in April 2002 (REF. 2). Laboratories that request samples must be non-profit-making and must send results of their studies to a CEPH database that will be made available to other researchers. One microgram of DNA — sufficient for hundreds of tests — is distributed to researchers at no cost other than shipping expenses, and larger amounts are supplied for a small charge that covers costs; however, cell lines are not distributed. The current collection consists of 1,064 cell lines from 52 populations around the world (FIG. 1). By July 2004, 56 laboratories had requested and obtained the collection.

Results from these laboratories have already been rewarding. The first researchers to use the HGDP collection, Rosenberg and colleagues⁹, genotyped each of the samples represented for 377 MICROSATELLITE loci that covered all autosomes. Their analysis of the structure of human populations was published at the end of 2002 and emphasized the importance of geographical isolation in determining genetic divergence (although other types of isolation were also observable), in agreement with the hypothesis that the divergence is mostly due to chance (random genetic drift). They also confirmed that genetic differences between populations are extremely small — in fact, smaller than those suggested by previous studies (BOX 4). Other research by the same group using the HGDP collection¹⁰ validated the estimation of early divergence

times in human evolution using microsatellites, whereas previous studies using fewer microsatellites indicated that these markers were generally useful only for dating more recent evolutionary events.

In addition, other studies that have used the HGDP resource have provided information on the usefulness of X-chromosome microsatellites¹¹; HAPLOTYPE frequencies and LINKAGE DISEQUILIBRIUM (LD) in folate-metabolism pathway genes¹²; evidence of recent positive selection at the lactase gene locus¹³ and analysis of seven Y-chromosome microsatellites¹⁴. Y chromosomes are not subject to recombination and therefore provide more information than other markers do about ancient evolutionary events¹⁵. Another study¹⁶ (discussed in more detail below) recently used HGDP samples to address general questions about sampling strategies and analysis of human genetic diversity.

The HGDP collection is the most complete worldwide human DNA collection that is available to not-for-profit researchers. The collection should prove to be an important resource for both human population genetics and evolutionary studies, as well as for biomedical studies, such as those described in the next section.

The HGDP and biomedical studies

It might be argued that the HGDP has no medical importance, because it provides no information on individual phenotypes — the only information provided about each sample is the population name, its geographical location in degrees of latitude and longitude, and the sex of each individual. However, this inference is wrong, as highlighted when, in 2003, the *Lancet* award for the year's best biomedical paper¹⁷ went to the first paper published with the HGDP data⁹. The HGDP collection is valuable not only for medical studies, but also for the study of other phenotypes, as described in later sections. It is true, however, that for most of these purposes, it would be important to increase the numbers of both populations and individuals that are currently represented.

One example of the potential use of the HGDP collection in biomedical research, particularly in countries where clinical surveys are not available or are difficult to carry out, is to estimate the incidence of recessive diseases, which are often unknown or underestimated, even if they are relatively frequent. If samples were taken from 50 individuals for each population, the detection of a single heterozygote for a mutant that is responsible for a recessive disease would indicate, according to the HARDY-WEINBERG EQUILIBRIUM, that the incidence

of individuals that express the disease should be about 1 in 10,000 (REF. 18). Most samples for populations in the current collection contain fewer than 50 individuals. However, predictions might be made more precise, not only by increasing the sample of phenotypically normal individuals from the population, but also by pooling information from genetically similar neighbouring populations.

Probably the most important medical application of the HGDP is the inexpensive provision of small but adequate and reliable control samples for ASSOCIATION STUDIES, which are increasingly being used to identify genetic variants that contribute to inherited diseases. These studies compare groups of unrelated patients from a defined population with an ancestrally similar control group¹⁹. Appropriate control samples are not easy to obtain, but could be provided by samples from collections such as the HGDP.

Another potential biomedical application is in examining the contributions of environmental factors to complex human disease.

Such analysis is usually done at the level of individuals. However, this can also be done at the population level and, in some cases — for example, for the influence of climatic and ecological factors and culture-specific customs — this approach can prove highly informative, as shown by its application to physical anthropology. In a recent successful study of this kind, data on cranial morphometry from several populations were compared with HGDP microsatellite data on genetic variation⁹ and with a database on climate. Significant correlations were found between specific cranial adaptations, patterns of genetic variation and climatic variables. One conclusion was that BRACHYCEPHALIZATION is the principal result of adaptation to extreme cold²⁰. The microsatellites examined in this study are unlikely to be located within the genes that are responsible for the observed correlations; however, they might be closely linked with genes that are responsible for some of the relevant phenotypes. Candidate genes responsible for specific phenotypes,

Box 2 | Cell lines as the basis of the HGDP collection

Ideally, the promotion of studies of human genetic diversity on the basis of comparative analysis of DNA sequences should ensure that the DNA is available to many researchers worldwide, without fear of exhausting the supply. DNA can be amplified in the laboratory in two ways: chemically, by the POLYMERASE CHAIN REACTION (PCR), and biologically, by growth of specific cell lines. Which of these approaches was more appropriate for the initiation of the Human Genome Diversity Project? Two main factors must be taken into consideration: accuracy and renewability.

Accuracy

Mitosis has been perfected in nature as a highly reliable method of DNA replication. There are indications that mutation rates are under the control of natural selection and might increase or decrease when necessary. In principle, after many growth cycles, mutations could also accumulate in cell lines that are grown in the laboratory, but we can minimize sources of error by keeping back-up subcultures of each cell line at low temperatures. The first cultures^{35,36} made for anthropological aims (in 1984) in central Africa and Bougainville are still in use and are part of the HGDP collection.

In the second half of the 1980s PCR was introduced. Direct methods of amplifying DNA by PCR have improved over the past several years, but in the absence of direct evidence, it is difficult to state how accurate they are in comparison with the replication of DNA in cells by mitosis. Rates of spontaneous mutation that take place during DNA replication of gamete-forming cells are low, on the order of 10^{-9} per nucleotide site a year, although this is difficult to estimate accurately. At the time that the HGDP was established, it seemed safer to continue to rely on cellular mitosis for DNA amplification, rather than risking the introduction of potentially higher error rates from *in vitro* methods. This is especially a source of concern for polymorphisms that occur in repetitive DNA — for example, microsatellite polymorphisms — which are much more prone to mutation than are SNPs.

Renewability

The advent of PCR means that studies are possible with much smaller samples of DNA. Nonetheless, practical experience shows that no matter how large the initial DNA samples taken are, they will eventually run out. As the aims of the HGDP dictate that many samples are taken from usually remote populations, returning to collect more DNA from the same individuals would be difficult and eventually impossible. With our increasing knowledge of genomic variation and the decreasing cost of DNA genotyping, it is becoming clear that it is preferable to study genome-scale datasets: for these purposes, renewable cell lines are still the best option. In the past, opinions differed on this point⁷, but it is notable that a major project such as the HapMap uses cell lines whenever possible³.

including medically important ones, can therefore be inferred and tested by more direct approaches. This will become more useful when the more detailed genetic information on microsatellites and SNPs that is expected from the HGDP collection becomes available, and with increased numbers of populations being available for study.

The HGDP and HapMap

Another potential biomedical application of the HGDP is its future interaction with the HapMap project^{3,4}, an ambitious research project that is aimed at solving problems of identifying the genetic determinants of complex diseases. Here, I briefly outline the aims of the HapMap project and discuss why its goals and those of the HGDP complement each other, with potentially important benefits.

The HapMap approach. During the last 50 years, substantial progress has been made in using LINKAGE MAPPING to identify genes responsible for inherited disorders that follow monogenic Mendelian patterns of inheritance. However, these disorders are responsible for a relatively small fraction of clinical cases. Most human diseases are believed to involve many genes, probably interacting in complex, non-additive ways, with potentially important environmental components.

For several reasons, the linkage approach has proved largely unsuccessful for determining the genetic basis of complex disease. The ideal approach would be to resequence the whole genome in patients and control cases to identify causal genetic variants. However, genome sequencing remains an expensive procedure, making this approach impractical. Full sequencing could be avoided by using SNPs as markers for disease-associated variants, as SNPs close to disease-related genes are likely to be transmitted with the disease. However, there is approximately 1 SNP every 1,000 nucleotides in the human genome, and a full study would require the testing of millions of SNPs per individual.

The HapMap plans to use knowledge about LD to increase the efficiency of SNP-based mapping. If haplotypes are sufficiently stable and of sufficient length, they could be used to reduce the number of SNPs that would need to be genotyped. For example, if haplotypes contained 20 SNPs on average, only 1 of these would be needed to tag a haplotype³ — the expected reduction in cost is on the order of the average number of SNPs per haplotype.

The HapMap project requires the identification of haplotypes and of at least one useful SNP in each of them. This project has already begun, and the first stage is nearing

Box 3 | Addressing ethical, legal and social issues in the HGDP

In the collection of the lymphoblastoid cell lines (LCLs) from worldwide populations, the Human Genome Diversity Project (HGDP) was acutely concerned with ethical, legal and social issues. Making sure that the needs of confidentiality and anonymity were properly addressed, that informed consent was obtained, that subjects were aware of the possible uses of the data and conforming with the legal needs of each country were the main concerns. All of the cell lines contributed to the collection were therefore reviewed to make sure that they had been collected in an ethical and legal manner¹. The US National Academy of Sciences National Research Council report⁷ provided general guidelines for this process. The background of each cell line that had been previously collected was reviewed to determine whether it was collected with informed consent for use in studies of human history or evolution. Only cell lines that complied with this were included in the HGDP resource. A protocol for confidentiality protection for donors of samples was also established. Both this and the vetting of cell lines were subsequently reviewed by an ethics advisory committee that was approved by the US National Institutes of Health Institute for General Medical Science. Other information was collected by the various researchers who contributed to the collection, but the only information that remained attached to each cell line concerned ethnic and geographical origin (in degrees of latitude and longitude), and sex.

completion. There are uncertainties about the usefulness of the HapMap data^{21–23}, because crossover events are not equally distributed, and the physical length of haplotypes and their SNP content vary greatly in different parts of the genome. Furthermore, populations vary in local LD and haplotype structure²³. There are also mechanisms of chromosome reshuffling other than crossing over that might interfere with the application of the method. Nevertheless, the project currently offers the best hope of tackling the problem of complex disease, other than a hypothetical, enormous and rapid decrease in sequencing costs.

The HapMap project will eventually cover slightly fewer than 100 individuals from each of three populations: one from Utah, with northern European ancestry, a Yoruba population from Nigeria and one from eastern Asia (Chinese and Japanese individuals). Haplotypes will be surveyed in these different populations, between which LD patterns might well vary. The extent of LD is known to be smaller in most African populations for evolutionary reasons — they have existed at least twice as long as other modern human populations and LD regions have had more time to decay because of the accumulation of crossovers. By contrast, all non-Africans — and a fraction of Africans — originate from a small population, probably of eastern African origin, that started expanding ~50,000 years ago and spread rapidly, first to Asia, and from there to other continents. The consequences of this demographical bottleneck are clearly visible in the genomes of modern humans^{15,24}, and explain their limited genetic variation and peculiar LD patterns.

Complementary, not competing. Both the HapMap and HGDP projects involve sampling

human genetic variation, but have different aims and little, if any, overlap at this stage. However, despite these differences, the two projects can only complement and reinforce each other.

The crossover events that have defined haplotypes occurred during the evolution of modern humans. Modern human history is characterized by early dispersions of an initially small population, and early rare crossovers are apparent in the geographical distribution patterns of haplotypes. Studies that use samples from HGDP populations to genotype pairs of SNPs that tag adjacent haplotypes will allow a description of the geographical distribution of the haplotypes that are defined in the HapMap project. This will help to determine the time and place of origin of rare crossovers that occurred early in human evolution, generating haplotypes that later spread to specific parts of the world and fundamentally affected the haplotype structure of human populations. As explained below, this knowledge could add a roughly equal amount to the power of current methods for evolutionary analysis. It could also be useful for the aims of the HapMap project, by increasing its analytical power in populations other than the three it has chosen to study.

Reconstructing human evolution. To clarify the way that HapMap data could complement those of the HGDP to understand the origins of human genetic diversity, it is important to appreciate that modern methods for reconstructing human evolution are based on two concurrent approaches, historical and geographical.

First, the historical reconstruction of the genealogy of mutations (the sequence of their occurrence) answers the ‘when’ questions, concerning the timing of evolutionary

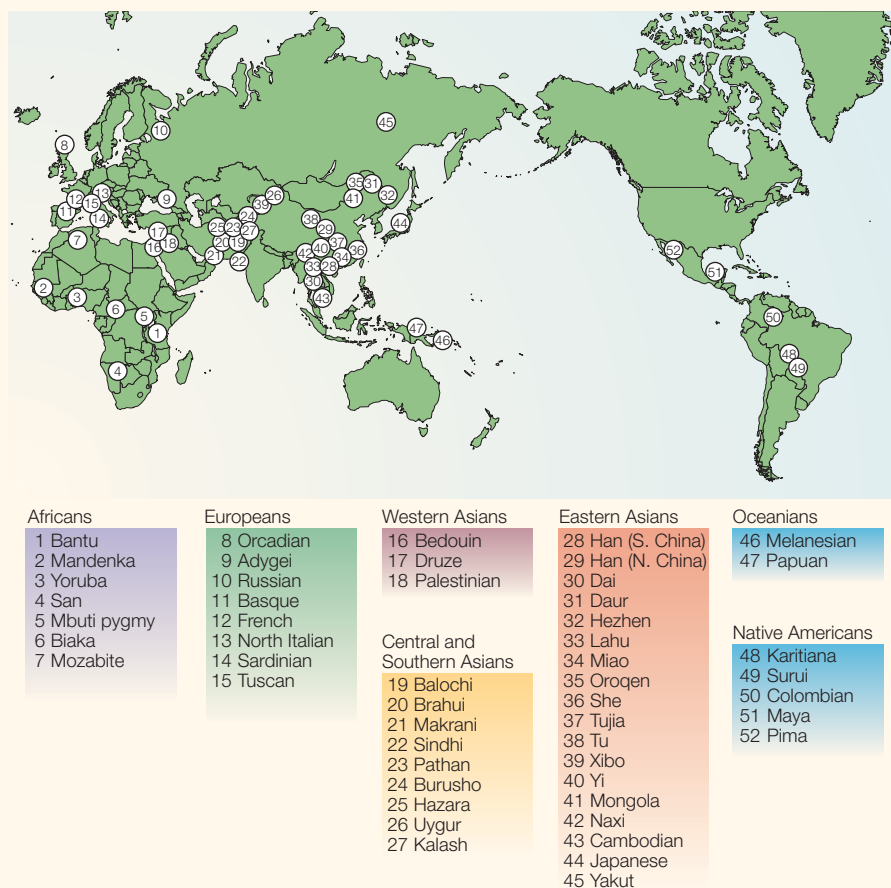


Figure 1 | **Populations that are included in the Human Genome Diversity Project collection.** A geographical distribution of the 52 populations that are represented in the Human Genome Diversity Project collection and that were used in the analysis by Rosenberg *et al.*⁹ described in BOX 4.

events. This allows the reconstruction of genealogies of extant haplotypes in the form of bifurcating trees, initiating from a common ancestor (marked by the first mutation in the species that formed modern humans)^{15,25}. These genealogies are different for every haplotype, but two haplotypes that have proved especially useful are mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY; BOX 1). They give similar genealogies¹⁵, and provide independent dates that are reasonably consistent with archaeological data. The numbers of mutations that occurred between successive splits in haplotype genealogies are used to evaluate lengths of branches in genealogies.

Knowledge of mutation rates allows an estimation of the times at which each branching occurred, including the first branching that defines the most recent common ancestor of the species for a particular haplotype. The date of the first branching is different for mtDNA and NRY (from the latest unpublished estimates, these are ~160,000 and ~100,000 years ago, with standard errors

of ~14% and ~20%, respectively). These differences are probably the consequence of variations in offspring numbers, which are greater for males, owing to polygamy.

There are several other methods of reconstructing human evolution^{24,25}. One uses distances between population pairs, calculated from the averages of large numbers of gene frequencies estimated in the populations, and is known as population tree analysis (PTA; BOX 1). Another method is based on identifying correlations between population separations and archaeological events for which the dates are known. There is good agreement between these different approaches, making the genetic reconstruction of the evolution of modern humans in the last 100,000 years fairly robust.

In contrast to this historical approach, geographical maps of the frequencies of mutations indicate 'where' they might have occurred^{26,27}. The times and places of the occurrence of mutations form the phylogeography of the species, and allow the reconstruction of the migratory paths that were taken during expansions of modern human

populations in the last 100,000 years^{15,26}.

Finally, the 'why' questions concern the processes of drift and natural selection that have affected haplotype frequencies. In the presence of natural selection, questions arise concerning the mechanism of adaptation that was involved and its identification at the molecular, biochemical, anatomical, physiological and pathological levels. The study of drift is best carried out at the whole-genome level, and that of natural selection is restricted to specific genes or interactions between genes. Both must include cultural and demographical histories of individual populations.

The HGDP facilitates these studies of the origins of human genetic diversity by providing samples of genomic DNA from populations across the world. From these, the haplotype and polymorphism data that are crucial for these studies can be obtained.

New insights from crossover events. The future availability of HapMap data will provide a potential new approach to human evolution, which could be used for reconstructing the history and geography of haplotypes by studying the times and places of the occurrence of rare, early crossovers. The evolutionary picture arising from these investigations would be complementary to that obtained from studies of the mutations that have created the current patterns of human genetic diversity. The interpretation of history is fragile because, in contrast to experimental science, no repetition of the 'experiment' is possible. But multidisciplinary approaches to the same sequence of historical events, as well as the comparison of evolutionary histories obtained from different parts of the genome and for different genetic mechanisms, provide an analogue of a repeat of the same history. They provide further opportunities to test interpretations, and can greatly strengthen our confidence in the conclusions.

These two approaches to evolution — by studying mutations or crossovers — might have similar statistical power in terms of numbers of events. We can infer from the total length of human chromosome linkage maps that more than thirty crossover events occur per meiosis, which is in the same range as the rate of SNP mutations that arise per genome per generation. It is true that the process of crossing over is likely to induce local mutations, so that the two approaches are probably not entirely uncorrelated. But it seems likely that most mutations occur independently from crossing over, and can therefore supply generally independent evolutionary evidence, although better knowledge of the relation

Box 4 | Analysis of the genetic structure of human populations using the HGDP resource

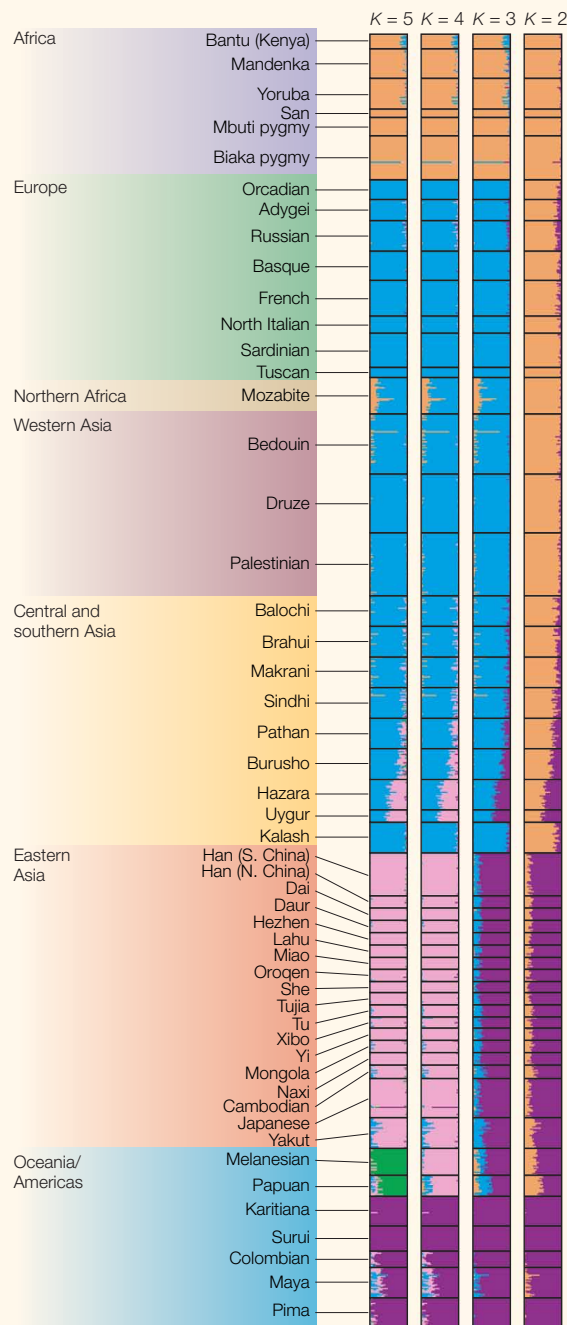
The first published study to use the Human Genome Diversity Project (HGDP) collection analysed data from 1,056 individuals from 52 populations, spanning all continents, for 377 microsatellite polymorphisms¹⁸. The study used the program Structure³⁷, which separates individuals into a number of clusters (K) that are chosen to maximize the variation between the clusters. The figure shows the results for $K = 2$ to $K = 5$, with clusters distinguished by arbitrary colours. Each vertical segment represents an individual, within which the different colours represent the proportion of admixture of the individual in terms of the clusters that represent the results. Admixture is the simplest explanation for the patterns seen, but other explanations are possible, and further analysis would be necessary to substantiate any chosen clustering pattern.

Forcing all data into 2 clusters ($K = 2$) separates all individuals and populations by longitude — with the west (Africa, Europe, and central, southern and western Asia) in one group (orange shading), and the east (eastern Asia, Oceania and the Americas) in another (purple shading). This is consistent with the beginning of the large expansion out of eastern Africa that began ~50,000 years ago and was directed predominantly eastwards. A three-cluster representation ($K = 3$) splits the western cluster of the $K = 2$ partition into sub-Saharan Africa in one group (orange shading) and Europe, western Asia and northern Africa in another (blue shading). Note that only one northern African population (Mozabites) is represented in the panel, but other results indicate that this conclusion can be extended to almost all of northern Africa. The $K = 4$ partition splits the Americas (purple shading) from eastern Asia and Oceania (pink shading), and $K = 5$ separates eastern Asia (pink shading) from Oceania (green shading).

The proportion of variation among the 52 HGDP populations estimated in REF. 9 is extremely small compared with that from within populations. Analysis of variance using F_{st} values (which are a common measure of gene frequency variation) indicates that, from these results, only 5–7% of the world genetic variation is accounted for by differences between the 52 populations⁹. Differences between the 5 principal groups account only for 3–5%, and differences between populations that are within a major group account for 2–4%. Clearly, the fraction of variation among individuals within populations — that is, the residual 93–95% — is by far the most important portion. This estimate is definitely larger than that observed previously from protein data, as well as from DNA polymorphisms, (including microsatellites that were tested previously³⁸), which was 85% (REF. 39). However, microsatellites are expected to give lower F_{st} variances than SNPs because they have lower gene frequencies, having on average many more alleles^{24,38}. Moreover, earlier estimates³⁹ were obtained with fewer (14), more heterogeneous population clusters than the 52 HGDP populations. Grouping the 52 HGDP populations data to mimic the 14 clusters used in REF. 38 decreased the estimate of the variation within populations to 89.8%.

Although the variation observed among the 52 HGDP populations with the microsatellite set used⁹ is smaller than any earlier estimate, it allows us to reconstruct¹¹ a population history that is consistent with the standard model of human evolution^{10,15}.

Reproduced, with permission, from REF. 9 © (2002) American Association for the Advancement of Science.



between the two processes would be beneficial. Using the LD approach, differences between populations or individuals based on LD comparisons would replace those that are based on population allele frequency or individual SNP difference. So, with increasing knowledge of HapMap tags, crossover history and geography might provide useful additional information to the evolutionary history of the human species that can be investigated using data from projects such as the HGDP.

Implications for the HapMap. An increased knowledge of LD patterns might also allow the application of the HapMap approach to medical problems within populations other than the three now being studied. LD blocks in some parts of the genome are likely to be useful in some parts of the world and not in others, because early crossovers might have destroyed LD of some haplotypes in some regions of the world and not in others⁹. It would therefore

seem that cooperation between the HGDP and the HapMap is highly desirable, even at this early stage.

The future of the HGDP

The main limitations of the current HGDP collection are that the present list of populations is small and does not evenly cover the inhabited regions of the world. In addition, the number of individual samples (1,064) is small, although this has already

allowed studies that have provided some interesting conclusions, as described above.

The main future requirement for the project is clearly to increase the number of cell lines, especially from areas that are now insufficiently represented. There is currently a greater concentration of samples from countries that have pioneered the idea of collecting cell lines from different ethnic groups and making them available for research — such as Israel, Pakistan and China. By contrast, India and Polynesia are not represented at all, and Europe, northern Asia, the Americas and Oceania have limited representation. Population samples in the collection contain an average of 20 individuals — about the size that was decided on as a compromise between suggestions from the first HGDP symposium. However, the sample size per population varies from 1 to 50 individuals. Populations from Pakistan and China represent various ethnic groups, and their samples are of 10 individuals, which is small by most criteria. One solution is to pool data from neighbouring populations that are sufficiently genetically similar.

Another question is whether the HGDP should focus in the future on individuals as the unit of sampling, or whether the emphasis should remain on sampling populations. These alternatives were considered at the start of the project. This is worth considering again now because of a recent paper¹⁶ emphasizing that, for interpreting human genome diversity,

attention to clines (gradual variations of populations in space) is preferable to focusing on clades (which results from a history of sharp population separations). Clines are certainly common for many genes²⁴, which is one reason to criticize the use of the distinction of races in humans, as already emphasized by Charles Darwin. Different methods of analysis emphasize either one or the other interpretation. The construction of trees by PTA (REF. 28) forces population data into clades, whereas multivariate principal component analysis (BOX 1), especially if displayed as geographical maps^{24,29}, tends to turn them into clines. Which interpretation is more accurate? This depends on local history and genetic geography, and understanding these factors should be an important aim of any study. There are geographical, linguistic and social barriers between populations, and the history of population separations — when these are sharp and not greatly altered by later migration — tends to generate discontinuities. When these discontinuities are real, population trees might be meaningful. DEMIC EXPANSIONS create strong clines for many genes, with clear centres of origin, but their multiplicity in a particular area (probably a common occurrence) generates a local mixture of clades and clines (BOX 1). Natural selection often creates clines of individual genes; individual migration tends to create clines for all genes, but group migrations to remote areas create new clades³⁰.

Different methods of analysis might emphasize an interpretation in terms of clines or clades, but the important question with respect to the HGDP is whether the sampling method used so far has irreversibly affected the analysis of data provided by the collection. Undoubtedly, data collection that involves gross or fine clustering of individuals into populations will strongly affect the perception of clines or clades¹⁶. Therefore, should the choice of the sampling unit for the HGDP be changed in the future from populations to individuals collected at uniform distances, as in one original suggestion? It is logistically more efficient to collect population samples, and it is certainly better than collecting random individuals at specified distances. It is also necessary, especially for the study of recessive alleles, to test whether random mating conditions and the absence of natural selection are satisfied, which might be difficult to do with individually based collections. Ignoring the social realities of populations also seems dangerous. For example, a naive sampling that is based on geographical distances between individuals in New York city would only generate a badly biased history of the whole world. Because of the narrow geographical range of most migration at the level of individuals, the similarity of geographically close populations is so strong^{15,24} that it seems reasonable to continue sampling small, well-defined populations of obvious anthropological or medical interest. Therefore, it seems reasonable to proceed in the direction followed so far by the HGDP. It is also comforting to notice that the sizes of populations sampled by the HGDP are small enough that there has been a substantial reduction in inter-population variance compared with all estimates from earlier population collections (BOX 4).

The danger of forcing cladistic interpretations¹⁶ using HGDP data seems remote, especially when the collection will be increased sufficiently in the future to remove major discontinuities in the present geographical distribution of populations. Moreover, the tendency of humans to cluster into social groups has important social and medical implications that might be lost if sampling was carried out at regular geographical intervals, rather than from small social groups. However, it is certainly interesting to test various random sampling schemes, and this might be made possible when national DNA collections become available.

Bearing in mind these considerations, strategies for extending the HGDP include asking teams of scientists that are collecting blood samples for anthropological or medical purposes to donate a fraction of their blood

Glossary

ADMIXTURE

The mixture of two or more genetically distinct populations.

ASSOCIATION STUDIES

A method for localizing genes that are responsible for specific diseases by comparing the DNA of a selected set of patients who are believed to carry the same mutation/s because of their ancestral origin, with that of unrelated healthy controls from the same population.

BRACHYCEPHALIZATION

An increase in the breadth to length ratio of the skull.

DEMIC EXPANSIONS

Processes of substantial demographical growth causing geographical expansions of a population. These are made possible by innovations that affect production of food, such as agro-pastoral economies and/or other improved technologies (for example, transportation, hunting and other weapons).

HAPLOTYPE

A set of genetic markers that show complete or nearly complete linkage disequilibrium; that is, they are inherited through generations without being changed by crossing-over or other recombination mechanisms.

HARDY-WEINBERG EQUILIBRIUM

A classical mathematical principle in population genetics used for testing random mating. It gives the expected frequencies of genotypes for a gene after one generation of random mating if the parental allele frequencies are known.

LINKAGE DISEQUILIBRIUM

The tendency for markers that are physically close to each other on the same chromosome to be transmitted to the progeny together, as there is a low probability that they will be split through recombination.

LINKAGE MAPPING

Mapping genes by typing genetic markers in families to identify regions that are associated with disease or trait values that occur within pedigrees more often than is expected by chance. Such linked regions are more likely to contain a causal genetic variant.

LYMPHOBLASTOID CELL LINES

Lymphoblastoid cell lines are obtained from B lymphocytes, a fraction of white cells from blood that can be grown indefinitely in the laboratory after special treatment of the cells with Epstein–Barr virus.

MICROSATELLITES

Microsatellites are tandem repeats of short nucleotide sequences (2–6 bases). They have a large number of alleles compared with SNPs, owing to a much higher mutation rate.

samples for making cell lines, continuing the tradition that built the HGDP collection. Sample donations from countries that are likely to generate national collections might allow an inexpensive enlargement of the present HGDP collection. National DNA and cell-line collections, known as biobanks, are being planned or established in Canada, Estonia, Iceland, Italy, Norway and the United Kingdom, and more are likely to follow. These biobanks could donate LCLs (or blood samples from which LCLs could be made), representing a minute fraction of their collection, ideally forming a geographically (and if necessary, linguistically and ethnically) stratified sample of the country. The only sample currently included in the HGDP that corresponds to this description is from France, and was generated by the CEPH at my suggestion.

From an ethical point of view, studies of human population genetics and evolution have generated the strongest proof that there is no scientific basis for racism, with the demonstration that human genetic diversity between populations is small, and perhaps entirely the result of climatic adaptation and random drift^{9,15,24}. It is to be hoped that the fears that were associated with the analysis of human variation have largely disappeared, and expansion of the study of human genetic diversity can become more efficient and scientifically satisfactory. The HGDP can help to achieve this aim. However, this project has survived with little support until now, and will need an increased level of funding. Its potential uses in medicine, science and social problems such as racism are sufficiently important that the project should be continued and expanded.

L. Luca Cavalli-Sforza is at the Genetics Department, Stanford Medical School, Stanford, California 94305, USA.

e-mail: cavalli@stanford.edu
doi:10.1038/nrg1579

1. Greely, H. T. Human genome diversity: what about the other human genome project? *Nature Rev. Genet.* **2**, 222–227 (2001).
2. Cann, H. M. *et al.* A human genome diversity cell line panel. (letter) *Science* **296**, 261 (2002).
3. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–795 (2003).
4. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Rev. Genet.* **5**, 467–475 (2004).
5. Cavalli-Sforza, L. L. How can one study individual variation for three billion nucleotides of the human genome? *Am. J. Hum. Genet.* **46**, 649–651 (1990).
6. Cavalli-Sforza, L. L., Wilson, A. C., Cantor, C. R., Cook-Deegan, R. M. & King, M.-C. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics* **11**, 490–491 (1991).
7. Committee on Human Genome Diversity, National Research Council. *Evaluating Human Genetic Diversity* (US National Academy of Sciences, Washington DC, 1997).

8. Dausset, J. *et al.* Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577 (1990) (in French).
9. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
10. Zhivotovskiy, L. A., Rosenberg, N. A. & Feldman, M. W. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72**, 1171–1186 (2003).
11. Ramachandran, S., Rosenberg, N. A., Zhivotovskiy, L. A. & Feldman, M. W. On the robustness of the inference of human population structure. *Hum. Genomics* **1**, 87–97 (2004).
12. Shi, M., Caprau, D., Romitti, P., Christensen, K. & Murray, J. C. Genotype frequencies and linkage disequilibrium in the CEPH Human Diversity Panel for folate pathway genes *MTHFR*, *MTHFD*, *MTFR*, *RFLJ* and *GCP2*. *Birth Defects Res. A* **67**, 545–549 (2003).
13. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
14. Macpherson, M. J., Ramachandran, S., Diamond, L. & Feldman, M. W. Demographic estimates from Y-chromosome microsatellite polymorphisms: analysis of a worldwide sample. *Hum. Genomics* **1**, 345–354 (2004).
15. Cavalli-Sforza, L. L. & Feldman, M. W. *Biology as history: population genetic approaches to modern human evolution.* *Nature Genet.* **33**, 266–275 (2003).
16. Serre, D. & Paabo, S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685 (2004).
17. Horten, R. *et al.* Read all about it: the *Lancet's* paper of the Year, 2003. *Lancet* **362**, 2101–2103 (2003).
18. Cavalli-Sforza, L. L. & Bodmer, W. *The Genetics of Human Populations* (Freeman, San Francisco, 1971; Dover, New York, 1999).
19. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
20. Roseman, C. C. Detecting inter-regionally diversifying natural selection of modern human cranial form using matched molecular and morphometric data. *Proc. Natl Acad. Sci.* **101**, 12824–12829 (2004).
21. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
22. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
23. McVean, G. A. T. *et al.* The fine scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
24. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, New Jersey, 1994).
25. Cavalli-Sforza, L. L. The DNA revolution in population genetics. *Trends Genet.* **14**, 60–65 (1998).
26. Underhill, P. A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62 (2001).
27. Edmonds, C. A., Lillie, A. S. & Cavalli-Sforza, L. L. Mutations arising in the wave front of an expanding population. *Proc. Natl Acad. Sci. USA* **101**, 975–979 (2004).
28. Cavalli-Sforza, L. L. & Edwards, A. W. F. Analysis of human evolution. *Genet. Today Proc. 11th Int. Congress Genet.* **3**, 923–933 (1964).
29. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. L. Synthetic gene frequency maps in Europe. *Science* **201**, 786–792 (1978).
30. Cavalli-Sforza, L. L. Some current problems in human population genetics. *Am. J. Hum. Genet.* **25**, 82–104 (1973).
31. Hirsfeld, L. & Hirsfeld, H. Essai d'application des methodes au probleme des races. *Anthropologie* **29**, 505–537 (1919) (in French).
32. Race, R. R. & Sanger, R. *Blood Groups in Man* (Blackwell Scientific, Oxford, 1975).
33. Pauling, L., Itano, A. H., Singer, S. J. & Wells, I. C. Sickle cell anemia, a molecular disease. *Science* **110**, 543–548 (1949).
34. Harris, H. *The Principles of Human Biochemical Genetics* 3rd edn (Elsevier; North Holland Biomedical Press, Amsterdam, 1980).
35. Cavalli-Sforza, L. L. *et al.* DNA markers and genetic variation in the human species. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 411–417 (1987).
36. Cavalli-Sforza, L. L. (ed.) *African Pygmies* (Academic, Orlando, 1986).
37. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
38. Bowcock, A. M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).
39. Barbujani, G. *et al.* An apportionment of human DNA diversity. *Proc. Natl Acad. Sci. USA* **84**, 4516–4519 (1987).

Acknowledgements

This work has been made possible by donors of blood samples and cell lines to the Human Genome Diversity Project (HGDP) and the Center for the Study of Human Polymorphism (CEPH). The collaboration with CEPH has been a decisive contribution. Support for preparing the first African cell lines in the Stanford laboratory in 1984–1985 came initially from the Lucille P. Markey Trust, with later additions from a National Institutes of Health Institute for General Medical Science programme and the HGDP–CEPH initiative from the Ellison Medical Foundation. H. Cann, M. Feldman, H. Greely and M.-C. King are thanked for suggesting improvements to the manuscript.

Competing interests statement

The author declares no competing financial interests.

 **Online links**

FURTHER INFORMATION

- Fondation Jean Dausset — CEPH:**
http://www.cephb.fr/ceph_presentation.html
- International HapMap Project:** <http://www.hapmap.org>
- Human Genome Diversity Project:**
<http://www.stanford.edu/group/morrinst/hgdp.html>
- Human Genome Organisation:**
<http://www.gene.ucl.ac.uk/hugo>
- Human Genome Project:**
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
- Marcus Feldman's laboratory:**
<http://charles.stanford.edu/pubs.html>
- National Research Council:**
<http://www.norveco.com/html/lab/NAS-NRC.htm>
- Noah Rosenberg's web site:**
<http://www.cmb.usc.edu/people/noahr/projects.html>
- Stanford Human Population Genetics Laboratory:**
<http://hppl.stanford.edu>
- Access to this interactive links box is free online.**

ERRATUM

EMERGING TECHNOLOGIES FOR GENE MANIPULATION IN *DROSOPHILA MELANOGASTER*

Koen J. T. Venken and Hugo J. Bellen

Nature Reviews Genetics **6**, 167–178 (2005); doi:10.1038/nrg1553

In this article the Cre recombinase was incorrectly defined as a cyclic AMP-response element. This correction has been made to the online enhanced text and PDF version of this review.