OPINION

Segmental duplications and the evolution of the primate genome

Rhea Vallente Samonte and Evan E. Eichler

Initial human genome sequence analysis has revealed large segments of nearly identical sequence in particular chromosomal regions. The recent origin of these segments and their abundance (~5%) has challenged investigators to elucidate their underlying mechanism and role in primate genome evolution. Although the precise fraction is unknown, some of these duplicated segments have recently been shown to be associated with rapid gene innovation and chromosomal rearrangement in the genomes of man and the great apes.

Single-base-pair mutation, sequence duplication and chromosomal rearrangement are the primary forces by which any genome evolves over time[1]. Mammalian species have undergone extensive chromosomal rearrangements that have reshaped their genomes and, in some cases, are believed to have led to speciation[2,3]. Both ancient and recent duplication events have been documented in mammalian genomes[4–6], although, until recently, it was thought that most major duplication events in the human genome were ancient in origin (>450 million years ago (Mya))[1,4]. The initial sequence reports of the human genome have challenged this idea[5,7,8].

It is now estimated that ~5% of our genetic material is composed of segmental duplications that have emerged during the past 35 million years of our species' evolution[5]. Many of these recently duplicated segments are located in regions that are hot spots of chromosomal and/or evolutionary instability, which indicates that there is likely to be an important link between the processes of chromosomal rearrangement and duplication. The abundance of segmental duplications, their central role in the emergence of genes with new function and their association with chromosomal instability, have important implications for primate genome organization and evolution. Previously, we summarized the organization and structure of these duplications, as well as their potential role in DOMAIN ACCRETION during vertebrate evolution[8,9]. In this Perspective, we discuss evidence that recent duplications have had a vital role in altering the genetic constitution of man and the great apes, both at the level of the gene and the genome.

Features of segmental duplications

Segmental duplications are large, nearly identical copies of genomic DNA, which range in size from 1 to >200 kb and are present in at least two locations in the human genome[5,8] (ONLINE TABLE 1). Segmental duplications, in contrast to whole-genome polyploidization events[1], originate from the duplicative transpositions of small portions of chromosomal material[4,6]. Segmental duplications contain both high-copy number repeats and gene sequences with intron–exon structures and, unlike other repeat classes, share no defining characteristics[10,11]. Their high sequence identity (90–100%) provides ample substrate for paralogous recombination events to occur and they have been identified on every human chromosome. The distribution of these segments among human chromosomes seems non-uniform (TABLE 1), with some chromosomes, such as the Y chromosome, showing peculiar enrichments for these types of duplication[12].

The identification and characterization of segmental duplications have been on the basis of both computational and fluorescence *in situ* hybridization (FISH) analysis of the human genome[5,7,8]. Both types of analysis, which only consider large blocks with 90–100% sequence similarity, indicate that duplicated segments comprise ~5% of the available sequence, and that they tend to be located in pericentromeric and subtelomeric regions[7]. *In silico* analysis has revealed a 3–5-fold enrichment of duplication within 100 kb and 1 Mb of telomeres and centromeres, respectively[8]. The segments are structurally very complex, with little or no unique sequence over hundreds of kilobases. Large intrachromosomal duplications (>200 kb) have been documented for many human chromosomes (15, 16, 17, 21 and 22). Detailed analyses have indicated that these larger blocks of duplicated material are, in fact, composed of smaller units or modules of duplications.

So far, there is no mechanism that can fully explain the frequency, organization and distribution of these segments in the human genome. The fact that most of the duplicated segments are interspersed throughout the genome, as opposed to being organized as tandem arrays, argues against unequal crossing-over during meiosis as the primary mechanism of dispersal. No evidence for short direct repeats at the sites of integration has been found, excluding double-stranded breakage followed by repair (typically associated with transposition events) as a mechanism[13]. L1-ELEMENT-mediated transduction also seems unlikely[14], as this process transposes only tracts of limited size (<1 kb) under experimental conditions. Furthermore, most segmental duplications that have been studied are greater than 10 kb in length and show no association with L1 repeat elements at their integration breakpoints. Owing to the

Table 1 | **Overview of segmental duplications in the human genome**

Chromosome	Intrachromosomal (%)	Interchromosomal (%)	All (%)
1	1.4	0.5	1.9
2	0.1	0.6	0.7
3	0.3	1.1	1.1
4	0.0	1.0	1.0
5	0.6	0.3	0.9
6	0.8	0.4	1.1
7	3.4	1.3	4.1
8	0.3	0.1	0.3
9	0.8	2.9	3.7
10	2.1	0.8	2.9
11	1.2	2.1	2.3
12	1.5	0.3	1.8
13	0.0	0.5	0.5
14	0.6	0.4	1.0
15	3.0	6.9	6.9
16	4.5	2.0	5.8
17	1.6	0.3	1.8
18	0.0	0.7	0.7
19	3.6	0.3	3.8
20	0.2	0.3	0.5
21	1.4	1.6	3.0
22	6.1	2.6	7.5
X	1.8	3.2	5.0
Y	12.1	16.0	27.4
Unknown	0.0	0.5	0.5
Total	2.0	1.5	3.3

The calculation was based on the finished sequence of September 2000 (http://genome.ucsc.edu/). The analysis excludes duplications with identities >99.5% to avoid artefactual duplicates caused by incomplete assembly of working draft sequence. There is some overlap between the interchromosomal and intrachromosomal sets. By these criteria, only 3.3% of the human genome is duplicated. However, estimates based on fluorescence *in situ* hybridization and computational analysis indicate that the final amount will be ~5%. This table is adapted with permission from REF. 5.

size of the duplications and the problems associated with sequencing these regions as part of the Human Genome Project[8], an eventual understanding of the mechanism will require specialized strategies that target these regions for completed sequencing. Comparative sequencing of these regions in closely related primates might also help to uncover the series of events that led to their creation, and to provide insight into the nature of integration sites before and after duplication. Once such information is obtained, it might become possible to design experiments that ultimately test the mechanism *in vitro*, as has been done to model L1-retrotransposition events[14].

Although the details about the molecular mechanism remain obscure, several different models have been proposed to explain why duplicated segments accumulate in subtelomeric and pericentromeric regions. One model indicates that regions near centromeres and telomeres might have a greater tolerance for the incorporation of new genetic material without adverse effects to the organism, perhaps owing to lowered gene density. However, recent genome-wide surveys render this explanation less plausible because large tracts of human sequence have been identified outside pericentromeric regions that are devoid of genes and yet show no increase in segmental duplication content[8,15]. Another explanation is that a much greater length of time is required to delete duplicated segments near centromeres and telomeres because of the suppression of recombination in these regions of the genome. Indeed, transposon 'graveyards' in the vicinity of centromeres have been reported for several organisms, including both *Drosophila* and *Arabidopsis*[16,17]. The enrichment is largely restricted to known classes of well-characterized, short mobile elements (retroposons and DNA transposons). In *Arabidopsis*, however, pericentromeric regions have been shown to be more recombinationally active than previously anticipated[17]. Furthermore, other than transposons, there is no evidence in these organisms that pericentromeric regions are particularly prone to accumulate recently duplicated genomic segments that originate from the nuclear genome. In fact, the proportion and size of recent (>90% sequence identity) segmental duplications in other sequenced animal genomes are markedly reduced when compared with the human data[5,8] (TABLE 2).

Alternatively, the unique sequence characteristics of pericentromeric and subtelomeric regions might contribute to the bias. It has been postulated that hyper-recombinogenic sequences, which are restricted to pericentromeric regions, preferentially target duplications to these regions. Unusual (A+T)-rich or (G+C)-rich minisatellite-like repeats often demarcate sites of duplication integration. For one of these repeat classes (a 3.0-kb segment of directly repeated CAGGG sequence motifs), the repeat was shown to be restricted to primate pericentromeric regions, to have existed in these regions before the arrival of the duplicated segments and to be present at the site of integration for at least seven unrelated segmental duplications[10,18,19]. These properties, and the sequence similarity of these elements to other known recombinogenic signals and G4 DNA[20,21], lend further support to a model of targeted integration[18,19]. It should be noted, however, that not all pericentromeric segmental duplications show the presence of such putative transposition integration signals. Furthermore, pericentromeric duplications represent only ~35% of all segmental duplications, the remainder being found in telomeric or euchromatic regions of the genome[8]. No unusual sequences have been reported at the junctions of duplications outside pericentromeric regions.

A common feature of both intrachromosomal and interchromosomal segmental duplications is their proximity to other apparently unrelated segmental duplications. With the exception of PROGENITOR LOCI, segmental duplication events seem to cluster in zones of duplication. Moreover, comparisons between these zones show that the precise junction and order of smaller duplicated segments is often conserved. Phylogenetic and comparative analyses of several regions strongly support a two-step model of duplication[11,22] (FIG. 1). An initial progenitor locus duplicatively transposes a copy to a chromosomal region that is accepting duplicated sequence. A series of such events creates a mosaic of duplicated segments that originate from

Table 2 | **Cross-species comparison for segmental duplications**

	Percentage of genome (%)		
Size of duplication (kb)	**Fly**	**Worm**	**Human (finished)***
>1	1.2	4.25	3.25
>5	0.37	1.50	2.86
>10	0.08	0.66	2.52

*This is an underestimate of the total amount of segmental duplication in the human genome because it only reflects duplication that is detectable within available finished sequence. A greater proportion of working draft sequence was found to contain duplicated sequence. The proportion of segmental duplications of >1 kb has been projected to be ~5%. This table is adapted with permission from REF. 5.

different regions of the human genome. Subsequent duplications of larger segments between these zones of duplication — either intrachromosomally or interchromosomally — create additional copies of the initial modules, in which the order of the duplicated segments is preserved. Several rounds of duplication help to explain the remarkable complexity and structure of most of these regions of the human genome.

Genome plasticity

Microchromosomal repatterning. Attempts to reconstruct the evolutionary history of several segmental duplications show high levels of restructuring of primate chromosomes over the past 35 million years. Differences in copy number and location of duplicated segments have been observed for many chromosomes, particularly between the genomes of man and the great apes[19,22,23]. The effects, as expected, are most pronounced in regions near centromeres and telomeres[11,23], where accelerated rates of duplication and rearrangement have markedly altered the structure between species and in populations. STRUCTURAL POLYMORPHISMS of duplicated segments that range from 30 kb to 1 Mb have been identified in the pericentromeric regions of human chromosomal regions 16p11 and 15q11. In telomeric regions, the extent of structural variation and chromosomal reorganization is extraordinarily complicated, such that orthologous copies of duplicated regions might be found on different chromosomes depending on the individual or population. Similarly, the most proximally sequenced portion of chromosome 22 (REF. 24) (>400 kb) has recently been shown to be the result of a human-specific duplication event that involved the transposition of this segment from chromosome 14. This occurred after the separation of chimpanzees and humans from their common evolutionary ancestor.

More ancient, subtle architectural changes in the chromosomes of great apes have also been described that are generally consistent with phylogenetic relationships of the species (FIG. 2). Two copies of the Charcot–Marie–Tooth neuropathy type 1A repeat sequence (*CMT1A-REP*), for example, have been identified in both chimpanzees and humans, but only the single ancestral locus is present in gorillas, which indicates that the locus might have duplicated in a common ancestor of chimpanzees and humans[25]. Some duplicated loci (including the creatine transporter gene *SLC6A8* (solute carrier family 6, member 8) and the adrenoleukodystrophy gene *ABCD1* (ATP-binding cassette, sub-family D, member 1) at Xq28, and the Hs.135840 gene at 4q24) seem to be restricted to humans and the African apes, but are represented as a single copy among the orang-utans and Old World monkey species[10,22,26,27]. Other segmental duplications are found among all HOMINOID species, whereas others apparently arose in a common ancestor of Old World monkeys and apes. In summary, segmental duplications have occurred at many different times during primate evolution.

This continuum of segmental duplication events during recent primate evolutionary history is generally supported by *in silico* analysis of the human genome, which has used the sequence divergence of the duplicated segments to estimate their evolutionary age[5,8]. Both interchromosomal and intrachromosomal duplications that range from 90 to 100% sequence identity have been identified, indicating that segmental duplication has been a continuing process during the past 35 million years of evolution (FIG. 3). Overall, these studies indicate that segmental duplications have subtly and consistently restructured primate chromosomes (although possibly not at a constant rate) during evolution.

Chromosomal rearrangements

In human genetic disease, segmental duplications have been directly implicated in a growing list of recurrent chromosomal rearrangements[28]. Most recently, rearrangements that are associated with segmental duplications on chromosomes 7 and Y have been implicated in Williams–Beuren syndrome and infertility, respectively[29,30]. Several studies have shown that highly homologous sequences are predisposed to homologous, unequal recombination, which can lead to large-scale chromosome rearrangements, such as deletions, duplications, PARACENTRIC INVERSIONS and, possibly, translocations. Is it possible that the same segmental duplications have had a role in the structural rearrangements that distinguish primate chromosomes?

Comparative analysis of high-resolution G-banded chromosomes from the hominoid species shows that 18 out of the 23 chromosome pairs in modern man are virtually identical to those of a common hominoid ancestor[31]. Most chromosomal differences are PERICENTRIC INVERSIONS (chromosomes 1, 4, 5, 9, 12, 15 and 16 of humans and chimpanzees), although a paracentric inversion of chromosome 7 distinguishes chimpanzee and gorilla chromosomes. In some cases, both peri- and paracentric inversions are necessary to

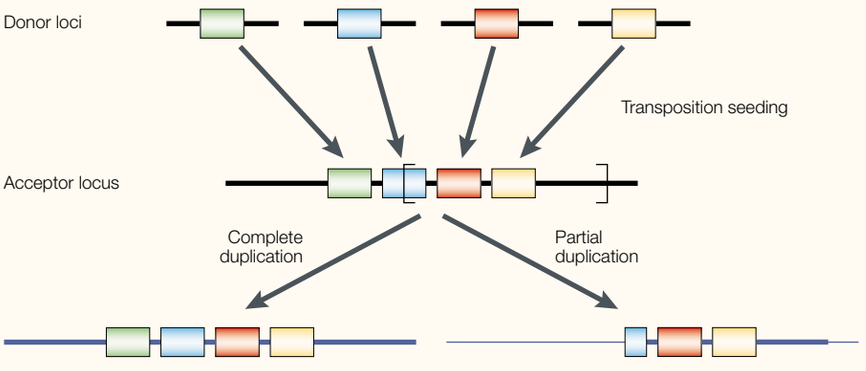


Figure 1 | **Model of segmental duplication.** Acceptor regions of the genome acquire segments of genomic material that range from 1–200 kb from disparate regions (donor loci) through a process of duplicative transposition. Events occur independently over time, which results in the formation of larger blocks of duplicated sequence that are mosaic in structure. Secondary events duplicate portions of this mosaic structure to other regions of the genome. Rearrangements (deletions and inversions) subsequently alter the structure of these regions.

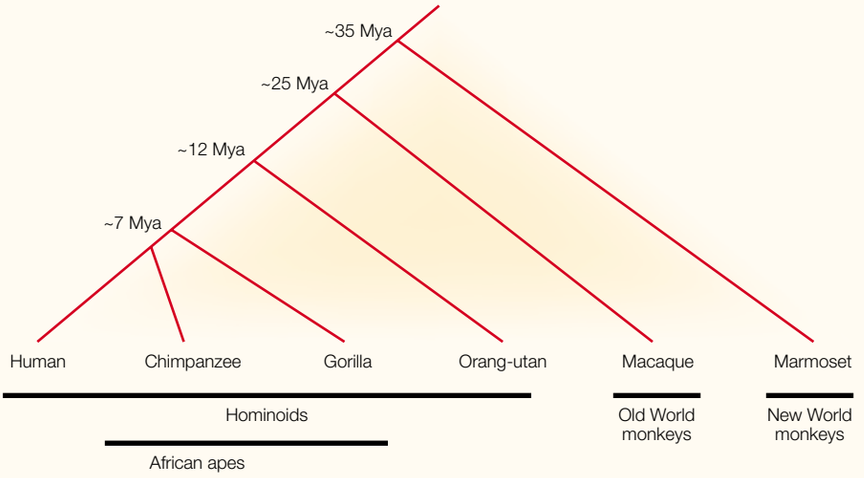


Figure 2 | **Primate phylogeny.** Generally accepted phylogeny of New World monkeys, Old World monkeys and hominoids (humans and apes). Estimated times of divergence are shown. Mya, million years ago.

account for differences in the cytological banding pattern (chromosome 16 of man and gorilla, and chromosomes 3 and 17 of man and orang-utan). Although less common than inversions, chromosome translocations are also observed cytogenetically, such as reciprocal translocation (human t(5;17) in gorilla)[31], band insertion (terminal band 20p13 on centromeric band 8q11.2 in orang-utan), and telomeric fusion (chromosomes 2p and 2q, with inactivation of the 2q centromere in man)[31–33]. The biological significance of these large-scale differences is unknown, but it has been postulated that such rearrangements create genetic barriers that lead to STASIPATRIC SPECIATION[2].

Historically, FISH has served as the main physical mapping tool for identification of chromosomal rearrangements among closely related species. Such straightforward procedures can be complemented by sequence and computational analyses, further refining the characterization of these chromosomal alterations. Recent comparative mapping and sequence analyses of specific homologous chromosomes in mammalian species have strongly implicated a duplication-driven mechanism for these evolutionary chromosomal rearrangements[34–38]. Sequence-level characterization of breakpoint regions in mouse chromosome 19 and the orthologous human regions has identified duplicated gene families in 10 out of 15 of these regions. Similarly, comparative mapping of human chromosome 7 and the orthologous mouse regions showed the presence of large low-copy repeats at the inversion breakpoints between man and mouse homologous chromosomes[35]. Two pericentric inversions between man and chimpanzee chromosomes 12 and 1 have been shown to contain large blocks of duplicated sequence near or precisely at the target site of the rearrangement[39]. Finally, recent comparative mapping across the gorilla translocation (human t(5:17)) has identified an ~250 kb sequence duplicated near both breakpoints of the rearrangement[37]. Although the cause-and-effect relationship for any of these associations has not been determined, one view is that duplications predispose primate chromosomes to rearrangements by providing templates for non-allelic homologous recombination events (FIG. 4). In this model, the segmental duplications would serve as the target sites for large chromosomal rearrangements to trigger speciation. Alternatively, as has been proposed by some authors, the segmental duplications might be products of the rearrangement process itself. If the latter is true, however, a large number of 'reverted' rearrangements are required to explain the current micropatterning of segmental duplications and the conservation of cytogenetic banding among human and great ape chromosomes.

Evolution of new function

From the perspective of the gene, there are several potential consequences of recent segmental duplications. The most likely outcome is that the duplicated genomic segment that harbours intron and exon structure is non-functional, which leads to the accumulation of unprocessed pseudogenes[1,40]. Most of the duplicated regions in the human genome are littered with the 'carcasses' of paralogous copies with no apparent function. Indeed, many of the duplicated segments appear to

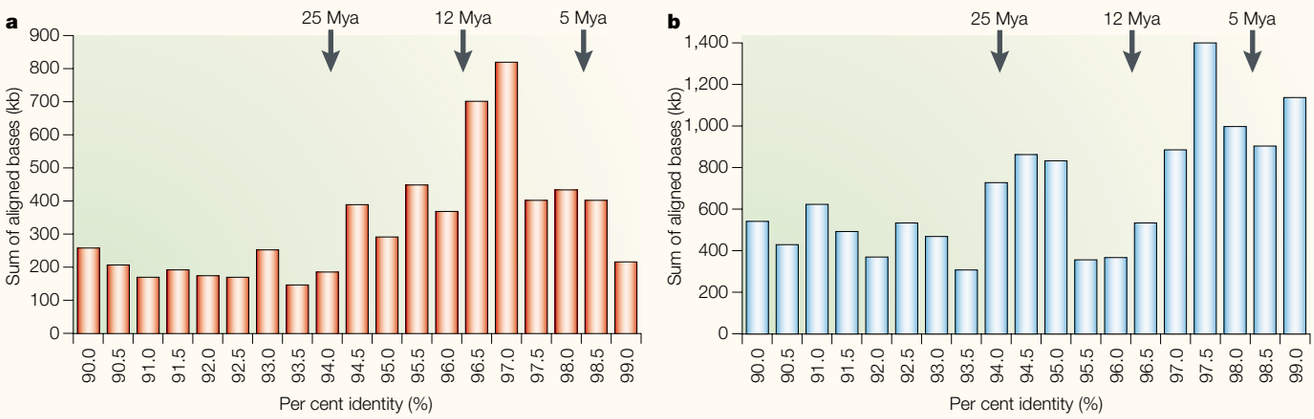


Figure 3 | **Sequence similarity among human segmental duplications. a** | Aligned intrachromosomal and **b** | interchromosomal duplications. The total number of kilobases aligned (kb) was calculated on the basis of all possible pairwise alignments from 563 Mb of finished human genome sequence (90–99.5% sequence identity). Assuming a model of neutral mutation, the degree of sequence identity corresponding to different evolutionary ages is shown, on the basis of the comparison of divergence of non-coding intron sequence between man and chimpanzee (5 Mya), man and orang-utan (12 Mya), and man and baboon (25 Mya). Most of the detected segmental duplications occurred between 1 and 12 Mya. Mya, million years ago.

contain only partial gene structure, and are missing 5′ exons, internal exons or sufficient regulatory machinery necessary to drive expression. Such duplicates are essentially 'dead on arrival'. However, after careful comparison of expressed sequence tags (EST) and genomic sequence databases, it has become apparent that many duplicated segments are expressed, sometimes in a tissue-specific fashion[41], although the transcripts frequently contain premature stop codons. Empirical and theoretical data indicate that such transcripts are non-functional[40,42]. It is also notable that dozens of examples have now been cited in which the transcripts are chimeric; transcription proceeds across adjacent duplicons that originated from different regions of the genome[9]. The production of such mosaic transcripts, which incorporate exons from different genes, raises the possibility of a mechanism akin to exon shuffling[43]. It is possible, therefore, that the increased complexity observed for the human proteome[5,6] might, in part, be due to the process of segmental duplication.

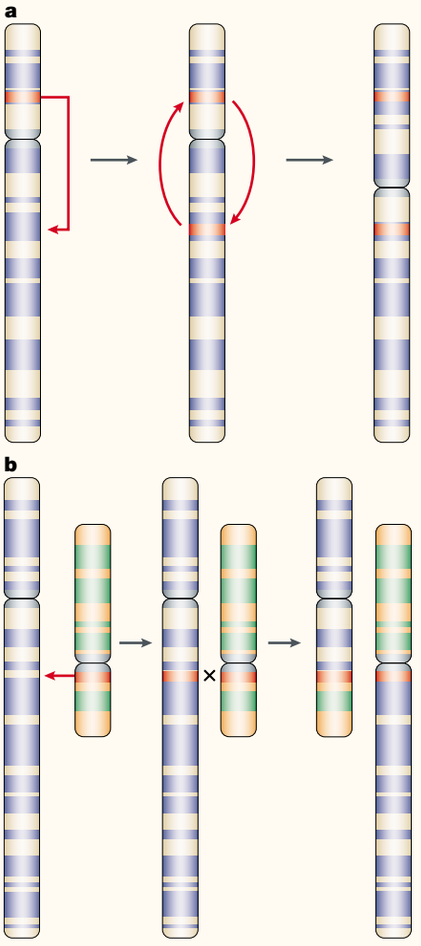


Figure 4 | **Duplication-driven chromosomal rearrangements. a** | A hypothetical pericentric inversion that occurs as a result of a segmental duplication, which creates two large blocks of homologous sequences. Subsequent aberrant recombination events between these two blocks lead to the formation of an inverted region of the chromosome that might serve as a genetic speciation barrier. **b** | A hypothetical reciprocal translocation that occurs as a result of an interchromosomal duplication, followed by non-allelic homologous recombination.

One of the most unlikely but important outcomes of segmental duplication is the evolution of new function. Gene duplication followed by positive selection has been postulated to be one of the primary forces responsible for achieving the proteome diversity and morphological complexity of vertebrates[1]. It has been argued that duplicating a copy of a gene or genomic segment effectively allows that copy of the gene to evolve unencumbered by the selective constraints imposed on its progenitor. Under such circumstances, a protein might emerge with a slightly modified or improved function relative to its precursor. Most of these 'successful' events were thought to be relegated to antiquity (>450 Mya), when genome polyploidization was still a viable possibility in our vertebrate ancestor. However, considering the abundance of segmental duplications that have emerged in our recent primate ancestor (<35 Mya), is it possible that new genes have been created through this pathway of duplication?

The evolution of trichromatic colour vision is a relatively recent innovation that was achieved through duplication of the opsin genes in a common ancestor of Old World monkeys and apes[44,45]. Mutation of the duplicated X-chromosome gene facilitated detection of a wider spectrum of visible light, which is thought to have been an important asset to our primate ancestor that helped it to distinguish yellow and red fruits against green foliage[44]. Similarly, the evolution of the anti-pathogen function of the eosinophil cationic protein (ECP) occurred as a result of duplication and mutation of an eosinophil-derived neurotoxin gene ~31 Mya[46]. The most significant mutational changes resulted in an enrichment of arginines, which supposedly altered the function of this duplicate early in its evolution. In both these examples, rapid evolution of new function occurred in concert with tandem duplication events.

Rapid evolution of primate genes has also been documented for interspersed segmental duplications. A 20-kb segment of chromosome 16, termed LCR16a, recently (12–5 Mya) proliferated, creating 15–30 copies that are dispersed throughout 15 Mb of the short arm of human and chimpanzee chromosome 16. A novel hominoid gene family (termed morpheus, after the Greek god of dreams who could change into many different human forms) was discovered in half of the human duplicates. Surprisingly, the exonic regions of this gene family showed accelerated rates of mutation when compared with intronic

Glossary

DOMAIN ACCRETION
The evolution of larger multidomain proteins by the addition of DNA segments that encode distinct structural domains.

G4 DNA
G-quartet or quadruplex DNA structure formed *in vitro* by DNA oligonucleotides with repeats that contain three or more consecutive guanines. In the mammalian genome, such regions (for example, telomeres, rDNA and immunoglobulin heavy-chain segments) have specialized recombination properties.

HOMINOID
A primate superfamily that includes the great ape species and humans (hominids).

L1 ELEMENT
A family of long, interspersed repeat elements (LINE1) that is still actively retrotransposing in the mammalian genome.

NEGATIVE SELECTION
A process in which the effective rate of synonymous change exceeds that of amino-acid replacement between homologous genes. It can occur when most non-synonymous changes in the gene are selectively deleterious and decrease the fitness of the species.

PARACENTRIC INVERSION
A structural chromosome alteration that results from breakage, inversion and reinsertion of a fragment of a chromosomal arm.

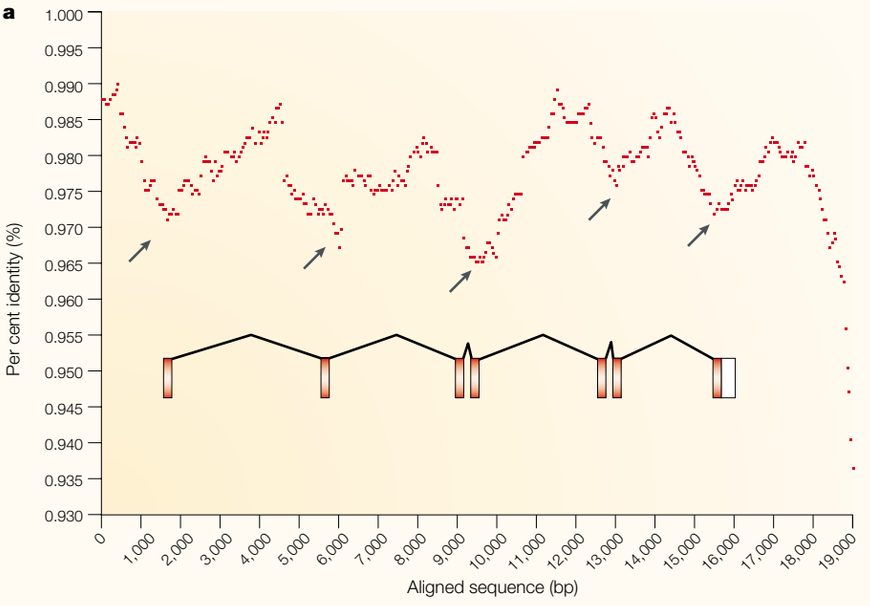
PERICENTRIC INVERSION
A structural chromosome alteration that results from breakage, inversion and reinsertion of a fragment that spans the centromere.

POSITIVE SELECTION
A process in which the effective rate of amino-acid replacement exceeds that of synonymous change between homologous genes. It can occur when non-synonymous changes in the gene are selectively advantageous and increase the fitness of the species.

PROGENITOR LOCUS
Ancestral locus from which the first segmental duplication is generated.

STASIPATRIC SPECIATION
Emergence of a new species as a consequence of chromosomal rearrangement and genetic isolation due to reduced fecundity and/or fertility of the hybrid species.

STRUCTURAL POLYMORPHISM
A large (usually greater than a few kilobases) chromosomal rearrangement (deletion, duplication or inversion) that is inherited and is polymorphic in a species. If such polymorphisms are cytogenetically visible, they are termed 'heteromorphisms'.



b

Gene	Exon 2	Exon 4
HSA8	VINTLADHRHRGTDFGG-----L--------------LLIITVFLRSYKFAISLCTSYLCV	GIKIVLEDIFTLWRQVETKVRAKIRKMKVTTKVNRHDKINGKRKTAKE
HSA11	...S..VY...E....VGVRDHPGQHGKTPSPQKLDNLII..IG...R.T.N.LF...C...	.N..G.K.VI..R.H.....R.....I.....RK
HSA5	.....H.....-----.----------.HV.IA.PT...VV.T.WIV..W.	.N..G.K.VI..R.H.....R.....I.H.....RK
HSA4	.....S..H.....-----.----------.HV.IA.PT...VV.T.WIV..W.	.N..G.K.VI..R.H...-.....R...-.I.H...-.....RK
HSA7	.....-----.----------.....T..W.	.....
HSA14	.....E....-----.----------II..I...GR...T.LF..I....	....G....C...K.A....Q.....K...H.Y.....
HSA6	.....H.....-----.----------.HV.IA.PT...VV.T.WIV..W.	.N....K.VI..R.H.....R.....I.H.....RK
PTR7	.....-----.----------.G..IE.....VV.T.W.V..W.	.N..GRK.VI..R.H.K.....H.....H.....K
PTR15	.....L.....-----F----------.....R....S.	.N..GRK.VI..R.H.KR.....H.....I.H.....K
PTR4	.....-----.----------.R..IA.....VVVT.W.V..W.	.N..G.K..I....H.....H.....I.H.....K
PTR12	.....L.....-----F----------...S.....S.	.....A.....K.....H.....T..
PTR16	.....VYH..E...C.GVRDHPGQHGKTPSPQKLDNLII..IG...P.T.T.LF..N....	.N..G.K.VI.P..H..R.F....H.R.....I.H.....RK
HKL1	..HS.P.DF.P....C.-----.----------IV..I...GISTL..F.WKTS...	...FGR..R..S..HM.A....EV..--..G...S.Y....Q....E.
HKL8	...SPP..F.P....C.-----.----------IV..I...GISTL..F.WKTS...	....G...L..SQ.HM.A....EVH.--..KN..S.Y....H.....K
CEA	...S.P..FQP.P..F.-----.----------IA..I...GISTLG.F.WKTSFG.	....GR..L..S..YM.A....VVH.--.....S.YQ.H.Q....TE.
PCR	...D.P..FQ..P..F.-----.----------RA..I...GISTLG.F.WKTSFG.	....A...L..SRSFM.ARA..EVH.--..R...S.YQ....Q....TEK
PHA	...S.P..FQP.P..F.-----.----------IA..I...GISTLG.F.WKASFG.	....G...L..SQ.HM.A....EVH.--..RN..S.Y....H.....K

Figure 5 | **Positive selection for the morpheus gene family among primates. a** | Sequence alignment of two human copies of LCR16a. Per cent identity was based on the number of substitutions observed in a 1,500 bp window, sliding every 50 bp. The intron–exon structure of one member of the morpheus gene family (NPIP, nuclear-pore-complex-interacting protein; AF132984) is shown (exons, red bars). The troughs in sequence identity correspond conspicuously to the position of the exons, which indicates that these might have been subject to rapid evolution. **b** | Multiple sequence alignment of encoded proteins corresponding to exons 2 and 4 of the morpheus gene family. Most of the nucleotide changes (95%) result in amino-acid changes. The sequences represent 14 genes from various species (CEA, *Cercopithecus aethiops*; HKL, *Hylobates klossii*; HSA, *Homo sapiens*; PCR, *Prebytis cristata*; PHA, *Papio hamadryas*; PTR, *Pan troglodytes*). For human sequences, only duplicates that have corresponding expressed sequence tags are translated. Conserved residues are shaded red. Gaps are denoted by (–), identical amino acids by (.). Reproduced with permission from REF. 50 © (2001) Macmillan Magazines Ltd.

regions (FIG. 5). Comparison of putative protein-encoding exons revealed that most of these changes (>95%) resulted in amino-acid changes. Consequently, average coding sequence divergence (15–20%) among copies between human and chimpanzee far exceeds the amount of intronic sequence divergence (1–2%). Analysis of this gene family among hominoids shows that the main episode of enhanced amino-acid replacement occurred after the separation of human, chimpanzee and gorilla lineages from the orang-utans (30-fold accelerated rate of amino-acid replacement when compared with neutrally evolving DNA).

Two aspects of the evolution of the LCR16a duplication are instructive. First, the analysis reveals an unprecedented degree of evolutionary plasticity between the human genome and that of its closest relatives. The recent dispersal of LCR16a to several sites on chromosome 16 is reminiscent of regional remodelling of chromosomes by transposons — a phenomenon that has been documented in other species[47,48]. The conditions that lead to the susceptibility of chromosome 16p to 'duplicon' invasion, or the predisposition of some species (orang-utan and chimpanzee) to have duplicated copies that 'escape' to other chromosomes, remains to be determined. For example, comparative maps of chromosome 16 among the hominoids do not indicate the presence of rearrangements, such as inversions, that would account for the distribution of this segment[31]. If the density of intrachromosomal segmental duplications is an indicator of such a dispersal event, similar regional remodelling might be expected for hominoid chromosomes 5, 7, 15, 17 and Y (TABLE 2). Limited analysis of the Y chromosome does support the view that this chromosome has been subject to rapid evolutionary turnover[49]. These data emphasize, once again, that primate chromosomal evolution might be a much less static process than previously anticipated.

The second important lesson learned from the LCR16a amplification is that the evolution of this duplication was accompanied by strong POSITIVE SELECTION. Unlike most other examples of gene duplication, there is no evidence from the morpheus gene family for conservation of ancestral function. In fact, the genic segments in many respects seem to have evolved from the 'ether' of non-functional DNA or DNA under weak NEGATIVE SELECTION[50]. How can this occur? One hypothesis might be that the genes encoded by this gene family are not essential for viability of the organism, but offer a huge selective advantage so that once a favourable mutation occurs, it effectively sweeps through the population. Genes under extreme sexual selection or genes involved with male reproduction have shown such

strong positive Darwinian selection[51–53]. Alternatively, the genes might have a role in adaptation similar to genes involved with the immune system or associated with predator–prey responses[51,52].

A common theme of recent genome project analyses is the large class (~30%) of genes that lack significant sequence similarity to genes in other organisms. It has been proposed that these genes[54,55] might be important in helping organisms to evolve and adapt to a specific ecological niche. It is tempting to speculate that the rapid evolution observed for the morpheus gene family and, perhaps, genes derived from other segmental duplications might have been crucial in helping man to adapt to his environment.

Conclusions

It has been more than 25 years since the seminal observation of King and Wilson that the limited amount of protein variation between chimpanzee and human was not consistent with the extreme phenotypic variation between the species[56]. At that time, it was thought that regulatory differences mediated either by structural rearrangements or single base-pair changes would most probably explain the phenotypic differences[57]. However, the recent data concerning the organization of segmental duplications in primate genomes have provided us with a new perspective on the structural differences between the genomes.

The organization and architecture of segmental duplications have two very broad implications in our understanding of the evolution and function of our genome. Particular regions of the genome have experienced extraordinary rates of evolutionary turnover, which result in considerable structural change between closely related primate species. This finding challenges our rather static idea of primate chromosomal evolution on the basis of cytogenetic data and indicates that non-uniform rates of genomic mutation might exist. Second, the process provides the opportunity for the duplication and transposition of genes into new chromosomal environments that allows them to evolve unencumbered by selective constraint. The recent origin of segmental duplications provides an ample substrate for both the alteration of existing genes and the birth of new ones. It is, therefore, not implausible that new genes with altered functions have emerged that distinguish man and great apes both at the phenotypic and genotypic level.

Recently, there has been considerable discussion about the value of comparative sequencing of great ape genomes[58]. If the mechanism of segmental duplication is to be understood, it is essential that genomic resources are developed for these organisms, and that the relevant regions are sequenced. More importantly, correlation of these hypervariable regions of the genome with phenotypic differences might be essential in understanding the nature of the human condition.

Rhea Vallente Samonte and Evan E. Eichler are at the Department of Genetics and Center for Human Genetics, School of Medicine and University Hospitals of Cleveland, Case Western Reserve University, Cleveland, Ohio 44106, USA. Correspondence to E.E.E. e-mail: eee@po.cwru.edu

DOI: 10.1038/nrg705

1. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
2. White, T. D. J. *Modes of Speciation* (W. H. Freeman, San Francisco, California, 1973).
3. O'Brien, S. J. & Stanyon, R. Phylogenomics. Ancestral primate viewed. *Nature* **402**, 365–366 (1999).
4. Lundin, L. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16**, 1–19 (1993).
5. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
6. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
7. Cheung, V. G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. The BAC Resource Consortium. *Nature* **409**, 953–958 (2001).
8. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
9. Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
10. Eichler, E. E. *et al.* Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* **5**, 899–912 (1996).
11. Horvath, J., Schwartz, S. & Eichler, E. The mosaic structure of a 2p11 pericentromeric segment: a strategy for characterizing complex regions of the human genome. *Genome Res.* **10**, 839–852 (2000).
12. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945 (2001).
13. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H. Jr. Isolation of an active human transposable element. *Science* **254**, 1805–1808 (1991).
14. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
15. Hattori, M. *et al.* The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405**, 311–319 (2000).
16. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
17. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286**, 2468–2474 (1999).
18. Eichler, E., Archidiacono, N. & Rocchi, M. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9**, 1048–1058 (1999).
19. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* **9**, 2029–2042 (2000).
20. Davis, M., Kim, S. & Hood, L. DNA sequences mediating class switching in α-immunoglobulins. *Science* **209**, 1360–1365 (1980).
21. Sen, D. & Gilbert, W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications in meiosis. *Nature* **334**, 364–366 (1988).
22. Horvath, J. *et al.* Molecular structure and evolution of an α/non-α satellite junction at 16p11. *Hum. Mol. Genet.* **9**, 113–123 (2000).
23. Jackson, M. S. *et al.* Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications, and unstable sequences with homologies to telomeric and other centromeric locations. *Hum. Mol. Genet.* **8**, 205–215 (1999).
24. Bailey, J. A. *et al.* Human specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70** (in the press).
25. Keller, M. P., Seifried, B. A. & Chance, P. F. Molecular evolution of the *CMT1A-REP* region: a human- and chimpanzee-specific repeat. *Mol. Biol. Evol.* **16**, 1019–1026 (1999).
26. Courseaux, A. & Nahon, J. L. Birth of two chimeric genes in the Hominidae lineage. *Science* **291**, 1293–1297 (2001).
27. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
28. Emanuel, B. S. & Shaikh, T. H. Segmental duplications: an 'expanding' role in genomic instability and disease. *Nature Rev. Genet.* **2**, 791–800 (2001).
29. Osborne, L. R. *et al.* A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nature Genet.* **29**, 321–325 (2001).
30. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279–286 (2001).
31. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
32. Turleau, C., De Grouchy, J. & Klein, M. Chromosomal phylogeny of man and the anthropomorphic primates (*Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus*). Attempt at reconstitution of the karyotype of the common ancestor. *Ann. Genet.* **15**, 225–240 (1972).
33. Dutrillaux, B. Chromosomal evolution in primates: tentative phylogeny from *Microcebus murinus* (Prosimian) to man. *Hum. Genet.* **48**, 251–314 (1979).
34. Nickerson, E. & Nelson, D. L. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* **50**, 368–372 (1998).
35. Valero, M. C., De Luis, O., Cruces, J. & Perez Jurado, L. A. Fine-scale comparative mapping of the human 7q11.23 region and the orthologous region on mouse chromosome 5G: the low-copy repeats that flank the Williams–Beuren syndrome deletion arose at breakpoint sites of an evolutionary inversion(s). *Genomics* **69**, 1–13 (2000).
36. Dehal, P. *et al.* Human chromosome 19 and related regions in mouse: conservative and lineage specific evolution. *Science* **293**, 104–111 (2001).
37. Stankiewicz, P., Park, S. S., Inoue, K. & Lupski, J. R. The evolutionary chromosome translocation 4;19 in *Gorilla gorilla* is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal *CMT1A-REP*. *Genome Res.* **11**, 1205–1210 (2001).
38. Tunnacliffe, A. *et al.* Duplicated KOX zinc finger gene clusters flank the centromere of human chromosome 10: evidence for a pericentric inversion during primate evolution. *Nucleic Acids Res.* **21**, 1409–1417 (1993).
39. Maresco, D. L., Chang, E., Theil, K. S., Francke, U. & Anderson, C. L. The three genes of the human *FCGR1* gene family encoding FcγRI flank the centromere of chromosome 1 at 1p12 and 1q21. *Cytogenet. Cell Genet.* **73**, 157–163 (1996).
40. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
41. Iyer, G. *et al.* Identification of a testis-expressed creatine transporter gene at 16p11.2 and confirmation of the X-linked locus to Xq28. *Genomics* **34**, 143–146 (1996).
42. Walsh, J. B. How often do duplicated genes evolve new functions? *Genetics* **139**, 421–428 (1995).
43. Patthy, L. Genome evolution and the evolution of exon-shuffling — a review. *Gene* **238**, 103–114 (1999).
44. Nei, M., Zhang, J. & Yokoyama, S. Color vision of ancestral organisms of higher primates. *Mol. Biol. Evol.* **14**, 611–618 (1997).
45. Yokoyama, S., Starmer, W. T. & Yokoyama, R. Paralogous origin of the red- and green-sensitive visual pigment genes in vertebrates. *Mol. Biol. Evol.* **10**, 527–538 (1993).
46. Rosenberg, H. F. & Dyer, K. D. Eosinophil cationic protein and eosinophil-derived neurotoxin. Evolution of novel function in a primate ribonuclease gene family. *J. Biol. Chem.* **270**, 21539–21544 (1995).

47. O'Neill, R. J., O'Neill, M. J. & Graves, J. A. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**, 68–72 (1998).
48. Petrov, D. A., Schutzman, J. L., Hartl, D. L. & Lozovskaya, E. R. Diverse transposable elements are mobilized in hybrid dysgenesis in *Drosophila virilis*. *Proc. Natl Acad. Sci. USA* **92**, 8050–8054 (1995).
49. Archidiacono, N. *et al.* Evolution of chromosome Y in primates. *Chromosoma* **107**, 241–246 (1998).
50. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
51. Zhang, J., Rosenberg, H. F. & Nei, M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA* **95**, 3708–3713 (1998).
52. Duda, T. F. & Palumbi, S. R. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc. Natl Acad. Sci. USA* **96**, 6820–6823 (1999).
53. Civetta, A. & Singh, R. S. Sex-related genes, directional sexual selection, and speciation. *Mol. Biol. Evol.* **15**, 901–909 (1998).
54. Hutter, H. *et al.* Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science* **287**, 989–994 (2000).
55. McClelland, M. *et al.* Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica serovars*, *typhimurium*, *typhi* and *paratyphi*. *Nucleic Acids Res.* **28**, 4974–4986 (2000).
56. King, M. & Wilson, A. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
57. Wilson, A. C., Bush, G. L., Case, S. M. & King, M. C. Social structuring of mammalian populations and rate of chromosomal evolution. *Proc. Natl Acad. Sci. USA* **72**, 5061–5065 (1975).
58. McConkey, E. H. & Varki, A. A primate genome project deserves high priority. *Science* **289**, 1295–1296 (2000).

Acknowledgements
This work is supported by the National Institutes of Health, the Department of Energy and the Charles B. Wang Foundation.

Online links

DATABASES
The following terms in this article are linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/
ABCD1 | *CMT1A* | ECP | NPIP | *SLC6A8*
OMIM: http://www.ncbi.nlm.nih.gov/Omim/
Charcot–Marie–Tooth neuropathy type 1A | Williams–Beuren syndrome
UniGene: http://www.ncbi.nlm.nih.gov/UniGene/Hs.135840
Access to this interactive links box is free online.

OPINION

No post-genetics era in human disease research

James Gusella and Marcy MacDonald

In the 1980s, linkage emerged as a route to discovering genetic defects, spurring the rise of genomics and making gene-based approaches available to previously phenotype-orientated researchers. In the post-genomics era, genetics is fundamental to understanding disease at all stages of the pathogenic process.

The traditional approach to studying human disease has been to describe the disease phenotype in as much clinical, histological and biochemical detail as possible, and then to try to work back to the mechanism. We refer to this strategy of identifying events that lie earlier in a pathogenic cascade, based on the nature of later phenotypes, as the phenotypic approach (FIG. 1). In a genetic disease, the goal of this approach would be to work back along the cascade until the causative gene and its mechanism are identified by hypothesis-driven molecular and biochemical phenotyping. This approach has been successful in disorders in which the nature of the clinical phenotype has indicated obvious candidate genes or pathways. However, for disorders in which the phenotypic approach does not lead to the identification of the disease gene, a genetic approach offers a powerful option. For the past two decades, genetics has been used, most successfully in monogenic disorders, to bypass the biochemical steps in the pathogenic cascade and jump directly to the genetic defect — the starting point of the disease process. Usually, this has required an indirect approach — positional cloning — in which the genetic defect is first assigned to a chromosome by genetic linkage studies of carefully phenotyped families, and is then identified on the basis of its chromosomal location. The identification of a disease gene offers a compelling alternative to the purely phenotypic approach for studying the pathogenic process. By using detailed, quantitative clinical phenotyping to formulate genetic criteria, the investigator then attempts to work forwards from the defective gene by defining sequential molecular phenotypes that are indicative of stepwise events in the cascade that leads to the final disease state. We refer to this strategy of relying on genotype to discover new, disease-relevant molecular phenotypes as the genotypic approach to disease mechanism (FIG. 1). Whereas the value of the genotypic approach might seem obvious to geneticists, it has been less obvious to the many non-geneticists that are involved in human disease research, for whom phenotype alone has long been the driving force. Although not intended as a comprehensive overview of the topic, this article draws from the experience in Huntington disease (HD) to provide examples of the phenotypic and genotypic approaches to disease, and how genetic analysis can be integral to interpreting and reconciling the findings in each strategy.

Huntington disease

HD is a late-onset neurodegenerative disorder that is inherited in an autosomal-dominant fashion[1]. The typical carrier of the *HD* gene lives for four decades with no evident clinical or physiological abnormality and then suffers the onset of a subtle movement disorder. The symptoms progress inexorably over the next 10 or 20 years to profound CHOREA, complete debilitation and finally death. Behavioural and cognitive changes also occur, with psychiatric symptoms sometimes preceding the onset of abnormal movements. Neuropathological phenotyping has revealed a hallmark gradient of progressive loss of STRIATAL NEURONS that begins in the tail of the CAUDATE NUCLEUS[2,3]. The neuronal cell type that is especially vulnerable is the most abundant striatal neuron — the medium-sized spiny GABAERGIC PROJECTION NEURON. Other striatal neuronal populations are relatively spared. Neuronal cell loss also occurs elsewhere in the BASAL GANGLIA and in the CEREBRAL CORTEX, the shrinkage of which leads to an overall reduction of brain weight of 30–40% at the final stages of the disease. About 15 years after onset, the HD patient usually succumbs to aspiration pneumonia, heart disease or another complication that results from their physical devastation.

The phenotypic approach

The purely phenotypic approach to disease mechanisms relies on hypothesis-driven research in which the investigator tries to place the observed disease manifestations in the framework of known biology. Experiments that test these mechanistic hypotheses have the potential to add a new layer to our understanding of known biological pathways in both the normal and disease state. To understand fully the disease process, one would like to work back, step by step, to earlier events in the disease process until the crucial causative factor(s) is found (FIG. 1). The number of steps, which is *a priori* unknown, and the overall time