A Novel Gene Family NBPF: Intricate Structure Generated by Gene Duplications During Primate Evolution

Karl Vandepoele,* Nadine Van Roy,** Katrien Staes,* Frank Speleman,† and Frans van Roy*

*Department for Molecular Biomedical Research, VIB-Ghent University, Ghent, Belgium; and †Department of Medical Genetics, Ghent University Hospital, Ghent, Belgium

Partial and complete genome duplications occurred during evolution and resulted in the creation of new genes and gene families. We identified a novel and intricate human genome family located primarily in regions of segmental duplications on human chromosome 1. We named it NBPF, for neuroblastoma breakpoint family, because one of its members is disrupted by a chromosomal translocation in a neuroblastoma patient. The NBPF genes have a repetitive structure with high intragenic and intergenic sequence similarity in both coding and noncoding regions. These similarities might expose these genomic regions to illegitimate recombination, resulting in structural variation in the NBPF genes. The encoded proteins contain a highly conserved domain of unknown function, which we have named the NBPF repeat. In silico analysis combined with the isolation of multiple full-length cDNA clones showed that several members of this gene family are abundantly expressed in a large variety of tissues and cell lines. Strikingly, no discernable orthologues could be identified in the completed genomes of fruit fly, nematode, mouse, or rat, but sequences with low homology could be isolated from the draft canine and bovine genomes. Interestingly, this gene family shows primate-specific duplications that result in species-specific arrays of NBPF homologous sequences. Overall, this novel NBPF family reflects the continuous evolution of primate genomes that resulted in large physiological differences, and its potential role in this process is discussed.

Introduction

Segmental duplications are regions of one to several hundred kilobases that exist in at least two copies in the haploid human genome and have a sequence identity >90% (Bailey et al. 2002). They comprise ~5% of the human genome and are located primarily in telomeric and pericentromeric regions. Frequent illegitimate recombination between duplicons is facilitated by their high sequence identities and plays a role in several pathologies (Emanuel and Shaikh 2001). The sequence homology between duplicons resulted in incomplete and incorrectly assembled sequences of these regions in the human genome sequencing project (Eichler, Clark, and She 2004). Some of the duplicated regions harbor genes, implying that these genes are members of gene families with high sequence identity, e.g., the morpheus and sulfotransferase 1A gene families, both located in segmental duplications on chromosome 16 (Johnson et al. 2001; Bradley and Benner 2005). Some of the duplication events and the corresponding birth of genes and gene families coincide with the recent evolution of primates. The evidence provided for strong positive selection of these genes indicates that they may have played a substantial role in the evolution of primates (Johnson et al. 2001).

Several regions on chromosome 1 consist of recently duplicated sequences (Hardas et al. 1994; Bailey et al. 2002). We identified a novel gene family located primarily on duplicated regions of chromosome 1, namely, 1p36, 1p12, and 1q21. One gene, identified by the positional cloning of a translocation breakpoint from a neuroblastoma patient (Laureys et al. 1990), was named NBPF1 for neuroblastoma breakpoint family, member 1 (Vandepoele et al., unpublished data). The NBPF genes contain numerous long-copy repetitive elements and show high intergenic and intragenic sequence identity, both in coding and noncoding regions. Because primate genomes indicate a recent expansion of the NBPF sequences resulting in species-specific genes, we speculate that NBPF genes played a role in the evolution of primates, including man.

Materials and Methods

cDNA Isolation and Expression Constructs

cDNA was prepared from mammary gland RNA (BD Biosciences, Palo Alto, Calif.) using the Generacer kit (Invitrogen, Merelbeke, Belgium). The first cDNA strand was transcribed starting from gene-specific primer MCBU3528 (3′ untranslated region [UTR]; 5′-GCCGATCCACATCTCTCGGCTTAGTAAG-3′). Subsequently, adaptors were ligated and polymerase chain reaction (PCR) was performed with primers MCBU3673 (5′-TGAGACACGCGCGGTAC-3′) and MCBU3259 (5′-CATGCCCACGTACCACATCTATG-3′) as forward and reverse primers, respectively. A nested PCR was performed using primers MCBU3672 (5′-AGGCTGC-CGAGCCTGTGAG-3′) and MCBU3278 (5′-GCCGACCACATCTCTCGGCTTAGTAAG-3′). The resulting fragments were cloned in pGEM-T Easy (Promega, Madison, Wisc.) The cDNA was transferred to the Gateway entry vector (Invitrogen) by ligating a BeII-NciI fragment into a BamHI-NsiI-digested vector. The insert was transferred by Gateway LR cloning into pDs2-4myc (constructed by rfA Gateway cassette insertion into pCs2-MT; cloning details available upon request). Sequences were deposited in GenBank and are accessible as AY894571–AY894575 plus the accession numbers listed in Table S4 (Supplementary Material online).

Expression Analysis

In silico expression analysis was performed by Blast analysis of both the nonredundant and the expressed sequence tag (EST) databases at the National Center for Biotechnology Information. Exons were aligned with ClustalX, and clustering was performed with MEGA2 (Kumar et al. 2001). Relative gene expression levels were determined using an optimized two-step SYBR Green I reverse transcription quantitative polymerase chain reaction (RT-qPCR). Each gene was examined in at least four different expression arrays. The expression ratio was determined by the 2−ΔΔCt method.
transcriptase (RT)—PCR assay employing multiple reference genes (Vandesompele, De Paepe, and Speleman 2002). Amplification efficiencies >95% were established for all primer pairs, and the delta-Ct method was used for quantification. This method is implemented in the qBASE software we developed in-house for automated analysis of real-time PCR data (Hellemans et al., personal communication). PCR reagents were obtained from Eurogentec (Seraing, Belgium) as SYBR Green I core reagents and used according to the manufacturer’s instructions. Reactions were run on the ABI5700 instrument (Applied Biosystems, Foster City, Calif.). Gene expression levels were normalized using the geometric mean of four reference genes (UBC, HPRT1, SDHA, and GAPD) (Vandesompele et al. 2002). Relative expression levels for NBPF were measured in a panel of seven mRNAs (BD Biosciences). Primer sequences for NBPF were 5′-CAGCAGCACATTCACTATTAGAG-3′ and 5′-TCACCTGATCCCCACCGATGTC-3′.

Primate Genomic Sequence Analysis

In silico genomic analysis was performed using the NCBI build 35 version of the human genome. DNAs from chimpanzee (Pan troglodytes), gorilla (Gorilla gorilla), gibbon (Hylobates sp.), cynomolgus monkey (Macaca fascicularis), and owl monkey (Aotus trivirgatus) were obtained from the European Collection of Cell Cultures. DNAs from rhesus monkey (Macaca mulatta) and African green monkey (Cercopithecus aethiops) were obtained in a type 1 exon in a good Kozak consensus sequence by a 3′UTR of 1.5 kbp. The putative start codon is located in a type 1 exon in a good Kozak consensus sequence at a distance of 139,000. It contains a large number of charged AA (33%) and is especially rich in Glu (14.4%). The highest concentration of charged AA (24 out of 39 AA) is found in type 12 exon, which is found with variable frequency in different NBPF cDNAs. This could lead to different net charges on the proteins, which consequently might differ in function or regulation.

In NBPF1, several pairs of exon types encode a novel protein domain, which we call an NBPF repeat. This domain is partly covered by the protein family database (PFAM) domain of unknown function, DUF1220 (accession number PF06758). It is present in multiple copies in the NBPF1 protein and also once, with lower homology, in the human homologues of myomegalin (PDE4DIP) (fig. 1B). Other regions of myomegalin show low homology to NBPF exon types 2, 3, and 13. Interestingly, one of the myomegalin genes is adjacent to NBPF1 on 1p36. The NBPF repeat is always built of two exons, and the different NBPF1 exon types constituting the NBPF repeats seem to have originated from a shared set of two ancestral exons (fig. 1B). Moreover, the introns flanking either set of repeat-encoding exons also show significant homology, consolidating the hypothesis of ancestor exons.

NBPF1 is a member of a recently duplicated gene family located primarily on human chromosome 1. Yeast artificial chromosome (YAC) clones derived from chromosome 1p36 cohybridize with other regions of chromosome 1, namely, 1p12, 1q21, and 1q42 (van der Drift et al. 1994), suggesting duplication of the 1p36 sequences. This has also been described in reports on recent segmental duplications in the human genome (Bailey et al. 2002; Cheung et al. 2003). An in silico genome survey showed that NBPF1 is a member of a gene family with gene copies located in...
almost all of these recently duplicated regions of chromosome 1, but some genes (NBPF4–6, NBPF22P) also contain exons without significant homology to the NBPF1 exons (fig. 2B). The distances between the different nucleotide sequences of the 15 exon types are shown in Table S2 (Supplementary Material online). The sequences that we analyzed (NCBI build 35) still contain several gaps in the NBPF genes, preventing us from making a definitive analysis of the complete NBPF gene family. However, based on the available sequences and counting both full-length and incomplete genes as well as pseudogenes, we concluded that the NBPF gene family consists of 22 family members (fig. 2, Table S1 in Supplementary Material online). The results of the in silico analysis of NBPF sequences located on chromosome 1 were confirmed by FISH analysis of BAC RP11-4513 originating from the NBPF1 locus. Bright signals were seen at 1p36 and 1q21 and a faint signal at 1p12 (fig. 2A, inset). In silico analysis of the human genome also predicted the presence of NBPF1-related sequences in chromosome arms 3p (NBPF21P) and 5q (NBPF22P), but this could not be confirmed by FISH. Similar results were reported for BACs originating from the 1q21 region (Zhang et al. 1999; Weise et al. 2005). The absence of signal on 3p and 5q might be due to the shortness of the region homologous to the 1p/1q sequences and the lower sequence homology. Most contigs containing NBPF sequences have been significantly improved during the finishing efforts of the public genome-sequencing center (IHGSC 2004). However, some contigs still contain numerous gaps, such as contig NT_079497.2, which contains a very large number of NBPF exons. This contig was excluded from our analysis because its draft status impedes accurate analysis. The NBPF genes appear to be located on three regions of chromosome 1. The first cluster, containing NBPF1, NBPF2P, and NBPF3, is located on chromosome 1p36. The NBPF1 gene is located 5 Mbp distal to the NBPF2P and NBPF3 genes, which are separated from each other by approximately 40 kbp. The latter two very likely arose through a recent partial duplication that produced the 5’ truncated NBPF2P gene for which no cDNAs could be identified so far, but with an open reading frame (ORF) still unaffected by nonsense mutations (see below).

A second cluster of sequences similar to NBPF1 was found on the short arm of chromosome 1 close to the centromere. Here, three of the four genes identified showed almost total sequence identity (≥99.9% for exons 1–7 of NBPF4, NBPF5, and NBPF6), but they deviate from the fourth gene in this region (NBPF7, 84% nucleotide identity for the same exons). Moreover, the cDNAs from these three genes contain additional exons with no significant homology to exons present in the NBPF1 gene (indicated in fig. 2B with letters instead of numbers). One exon, classified as exon type 15, resembled type 10 and type 11 exons and was classified as a novel protein domain called the NBPF repeat. These are shown in white boxes at the bottom. (B) AA alignment of the different NBPF repeats present in the NBPF1 protein. A homologous sequence was isolated from the draft version of the bovine genome and assembled in silico (AC157123). The bottom sequence is derived from one of the human homologues of rat myomegalin (AB007923; AA 1564–1654). The vertical line denotes the exon boundaries.

FIG. 1.—The NBPF1 gene is composed of repetitive elements. (A) Schematic representation of the ORF of the NBPF1 gene as predicted from the human genome sequence (NT_004873.16) and present in several cDNA sequences (boxed) derived from this gene. Each rectangle represents an exon, with highly similar exons (referred to as exon types) depicted in the same color and with the same number in or above the box. The start codon is present in the type 1 exon, resulting in a truncated version of this exon type (ΔI) at the aminoterminus of the encoded protein, in contrast to internal copies. The number of amino acids (AA) encoded by each exon is shown under its respective symbol. The isolated cDNAs contain multiple copies of exon types 10, 11, 12, and 13, which are present only once in the genomic sequence. Several exon couples constitute a novel protein domain called the NBPF repeat. The Complex NBPF Gene Family in Primate Evolution 2267
FIG. 2.—Genomic overview of the NBPF gene family. (A) Ideogram of human chromosome 1 showing the localization of the NBPF genes. The in silico mapping was confirmed by FISH analysis of human lymphocytes with a BAC derived from the 1p36 locus (inset). The arrow in front of each gene indicates its orientation. The color of the gene symbol indicates whether it is full length (green: start and stop codons present in one continuous ORF in the predicted cDNA), partial (red: due to gaps in the genomic sequence), or a pseudogene (black: P). Vertical lines indicate whether genes are present in a single genomic contig of the human genome assembly (NCBI build 35). Four genes are not linked to the ideogram: NBPF19 and NBPF20 are as yet unplaced on chromosome 1 and NBPF21P and NBPF22P are located on other chromosomes. (B) The ORFs of the different NBPF genes were assembled in silico based on cDNA sequences (Table S4, Supplementary Material online). NBPF4–NBPF6 and NBPF22P contain exons with no significant homology to exons present in NBPF1. These are shown as gray boxes with letters instead of numbers. NBPF8 contains a deletion of 2 nt in the first coding exon, which results in a frameshift mutation. Due to skipping of the type 2 exon in several NBPF8 transcripts, this frameshift often does not result in a premature stop. Frameshift and nonsense mutations in pseudogenes are indicated by asterisks at the site of occurrence. Three consecutive dots indicate the presence of gaps in the current genomic assembly.
The third and largest cluster of NBPF genes is located pericentromerically on the long arm of chromosome 1. The current genomic assembly contains 11 NBPF genes distributed along 7 Mbp of genomic sequence that still contains several gaps. The NBPF8 gene contains an aberrant type 1 exon at its 5’ end (shown as type 1A in fig. 2B) resulting from the deletion of 2 bp compared to the consensus sequence of the other type 1 exons. This deletion results in a frameshift, but due to skipping or deletion of the type 2 exon in several cDNAs from this gene, the alternative ORF does not lead to a premature stop codon but results in an amino-terminal sequence that is unique to the NBPF8 protein. The 2-bp deletion occurred quite recently because a type 2 exon without nonsense or frameshift mutations is present downstream of the 1A exon. NBPF9, NBPF12, and NBPF14 are located at the boundaries of sequence contigs and therefore could be predicted only partially. The exceptionally large NBPF10 gene contains a very large number of exons of types 10, 11, 12, and 13. It is located 450 kb distal of NBPF9 and encodes a hypothetical protein with a predicted Mr of 416,000. The full-size NBPF10 transcript was not retrieved in any cDNA, possibly due to the difficulty of cloning such an exceptionally long mRNA. However, it is also possible that genomic sequences corresponding to different members of the NBPF gene family have been erroneously assembled into an artificial fusion of two or more genes. The genomic sequence of NBPF11 contains a frameshift mutation in the type 11 exon, but this gene is nevertheless expressed in a variety of tissues, and hence, we consider it a functional gene. NBPF15 and NBPF16 are present on separate genomic contigs but show very high nucleotide sequence identity (99.4%), demonstrating that these genes were duplicated very recently or subjected to gene conversion. The last two genes positioned in this 1q21 region, NBPF17P and NBPF18P, were identified as pseudogenes. Two partial genes, NBPF19 and NBPF20, are present on a genomic contig that has not been positioned on chromosome 1. In addition to the numerous NBPF sequences on chromosome 1, we also identified homologous sequences on chromosomes 3p22 (NBPF21P: NT_0022517.17) and 5q14 (NBPF22P: NT_006713.14). NBPF21P contains a nonsense mutation in its type 3 exon and was classified as a pseudogene. NBPF22P contains a type 1 and a type 13 exon, but these are the only exons with high similarity to the exons of NBPF1. NBPF22P is nevertheless included in the NBPF gene family, as several atypical exon types found in NBPF4–NBPF6 (see above) are also present in this region.

In summary, we identified 10 complete NBPF genes, 6 partial genes (which might be completed in later releases of the human genome sequence), and 6 pseudogenes (fig. 2A, Table S1 in Supplementary Material online). As shown in figure 2B, the encoded NBPF proteins can be divided into an aminoterminal and a carboxyterminal domain, each of which can vary significantly in length, linked by a constant hinge region encoded by exons of types 7, 8, and 9. The hinge region is rich in negative charges compared to the overall charge distribution. The variable number of exons in the aminoterminal and carboxyterminal domains leads to significant structural differences, but the functional implications of these differences are unknown.

Phylogenetic Analysis of the NBPF Gene Family

To investigate the evolution of this gene family, we analyzed a genomic fragment extending from exon type 2 to exon type 4. For NBPF1, NBPF10, NBPF11, and NBPF20, two or three such regions could be identified (shown as a, b, and c in fig. 3), allowing us to investigate both gene duplications and intragenic duplications. In the case of NBPF1a, this fragment was 2,277 bp long (391 coding and 1,886 noncoding nucleotides). In total, we analyzed 22 copies of this fragment obtained from 17 NBPF genes. The resulting phylogenetic tree revealed that the genes located on chromosome 1p12 (NBPF4–7) were grouped with two pseudogenes (NBPF13P and NBPF21P) but separately from the other functional genes. Additionally, the genes in this cluster lack a LINE repeat insert, which is present in the intron between the type 7 and type 8 exons of all other NBPF genes containing these two exon types. With the exception of the NBPF3 and NBPF17P genes, whose branching showed low bootstrap values, all NBPF genes were grouped in a second major cluster. During the evolution of the ancestral gene to the functional genes located on 1p36 and 1q21, a duplication of the analyzed region occurred (middle arrow in fig. 3), after which the NBPF genes were generated through a number of duplication events. An additional duplication within the NBPF1 gene resulted in three copies (a + b vs. c) of the analyzed region.
Natural selection can be analyzed from DNA sequences using the ratio of nonsynonymous to synonymous substitution rates ($K_a/K_s$). This ratio is assumed to be equal to 1 under neutral evolution, whereas a ratio above 1 or below 1 indicates positive (adaptive) or negative (purifying) selection, respectively. To investigate the functional constraint acting on the NBPF genes, the coding sequences of the region analyzed above were used to determine the $K_a/K_s$ ratios (Table S3, Supplementary Material online). This analysis showed that negative selective pressure acted on several NBPF genes ($K_a/K_s$ significantly lower than 1), arguing against a neutralist hypothesis that the genes were duplicated without functional constraint, and for the presence of several functional NBPF paralogues in the human genome.

Multiple NBPF Genes Are Expressed in a Variety of Tissues

When comparing NBPF cDNA sequences to the genomic sequences, we were unable to identify a single cDNA with complete identity to the genomic sequence. This was in some cases due to the presence of multiple copies of exons in the cDNAs but only single copies in the genomic sequence (see fig. 1A for NBPF1), in addition to single-nucleotide differences. Although the high error rate in EST sequences makes it difficult to annotate cDNAs to the corresponding genes, we were able to do so for several transcribed sequences (Table S4, Supplementary Material online). We found a large number of cDNAs for one group of genes, namely, NBPF1, NBPF3, NBPF8–NBPF11, NBPF15, and NBPF16, showing that numerous members of this novel gene family are effectively expressed. The expression of these cDNAs was fairly ubiquitous as they were derived from a variety of tissues and cell types, including embryonic stem cells, fetal and adult tissues, and normal and cancerous tissues. The ubiquitous expression pattern seen in this in silico analysis was confirmed by RT-PCR experiments using nested family-specific primers (data not shown).

For another group of genes (NBPF4–7, NBPF12, NBPF14, NBPF17P, NBPF18P, and NBPF22P), a very small number (<5) of cDNAs was retrieved, showing that these genes can also be transcribed, although less frequently than the former group. No cDNAs could be identified for the remaining genes (NBPF2P, NBPF13P, NBPF19, and NBPF21P) in this analysis.

The relative abundance of NBPF transcripts in seven human tissues was analyzed by real-time quantitative PCR. The results for an amplicon located in the central region of the NBPF-coding sequence are shown in figure 4. The low Ct values obtained for this amplicon and for others located elsewhere in the NBPF transcripts (data not shown) indicate that the NBPF gene family is abundantly expressed in the tested tissues; the highest expression levels were observed in normal breast and liver.

The NBPF Gene Family Expanded During Primate Evolution

Segmental duplications occurred quite recently, and it is presumed that genes residing in these regions played a role in the evolution of primates. However, an extensive search for orthologous NBPF genes in genomes of model organisms, such as nematodes, flies, and several mammals (including mouse and rat), was unsuccessful. Unfinished genomes from cow and dog, however, contain partial sequences with low homology to the human NBPF sequences (fig. 1B).

To analyze the recent evolution of this gene family, we amplified a genomic fragment of 1.2 kbp, containing exon type 6, from seven primate species: chimpanzee (P. troglodytes, Ptr), gorilla (G. gorilla, Ggo), gibbon (Hylobates sp., Hyl), cynomolgus monkey (M. fascicularis, Mfa), rhesus monkey (M. mulatta, Mmu), African green monkey (C. aethiops, Cae), and owl monkey (A. trivirgatus, Atr). To maximize the chance of obtaining specific products, the primers were designed on regions that are completely conserved in the functional human NBPF genes. The amplified fragments were cloned, and 10 inserts were fully sequenced for each species. As long as the corresponding genomes are not fully sequenced, short evolutionary distances can hamper correct analysis because it is impossible to determine whether two highly similar sequences are siblings (with single-nucleotide changes due to PCR errors and/or allelic polymorphisms) or are derived from different genes. To overcome this problem and to reduce the complexity of the phylogenetic tree, we grouped sequences with Jukes-Cantor distances smaller than 0.010. The human sequences were obtained from the genomic sequence available to us, but those of the nonfunctional genes and NBPF4–7 were omitted, as all of them clustered apart from all other sequences (fig. 3).

The sequences from the owl monkey (Atr) were used as out-group sequences to build a rooted tree (fig. 5). With the exception of sequence Atr7, all sequences from the owl monkey were derived from pseudogenes and contain a nonsense mutation in the beginning of the amplified exon. The NBPF genes appear to have been duplicated early in the evolution of the Catarrhini, as there were at least two NBPF sequences before the split of the Old World monkeys (Cercopithecidae) from the apes (Hominidae + Hylobatidae) 25 MYA. These are shown in figure 5 as one large and one small cluster, both with sequences from most of the tested Catarrhini. Interestingly, both clusters replicate the known relationships between these primate species.
The smaller cluster contains only one human gene (NBPF3), together with one sequence from chimpanzee (Ptr4). No sequences from gorilla are present in this cluster, possibly due to the limited number of sequences we determined. But it is also possible that this gene is no longer present in the gorilla genome or has mutated too extensively to allow amplification with the primers used. The sequences retrieved from Old World monkeys showed that the orthologous genes have also been duplicated separately in African green monkey and macaques, leading to species-specific NBPF genes orthologous to the human NBPF3 gene.

Fig. 5.—Neighbor-joining tree of the genomic regions surrounding type 6 exons as cloned from several primates. The sequences of the Catarrhini are grouped into two clusters, both of which resemble the currently accepted relationship between primates. Only bootstrap values higher than 70 are shown. Sequences with Jukes-Cantor distances smaller than 0.01 are grouped and shown as cones with multiple sequence identifiers. NBPF, database sequences from the human genome; Ptr, Pan troglodytes; Ggo, Gorilla gorilla; Hyl, Hylobates sp.; Mfa, Macaca fascicularis; Mmu, Macaca mulatta; Cae, Cercopithecus aethiops; and Atr, Aotus trivirgatus.
The large cluster has essentially the same build as the smaller cluster and is subdivided into two subtrees, one for the Cercopithecidae and the other combining the Hylobatidae and the Hominidae. Gene duplication after the split of the Macaca from the other Old World monkeys led to two clades and the Hominidae. Gene duplication after the split of the smaller cluster and is subdivided into two subtrees, one for the Cercopithecidae and the other combining the Hylobatidae and the Hominidae. Gene duplication after the split of the Macaca from the other Old World monkeys led to two clades and the Hominidae. Gene duplication after the split of the smaller cluster and is subdivided into two subtrees, one for the Cercopithecidae and the other combining the Hylobatidae and the Hominidae. Gene duplication after the split of the Macaca from the other Old World monkeys led to two clades and the Hominidae.

The NBPF Proteins Are Located in the Cytoplasm

To investigate the subcellular localization of the protein products of this novel gene family, we transfected several human cell lines with an expression plasmid encoding GFP- or myc-tagged NBPF1. The overexpressed proteins had a cytoplasmic reticular staining pattern (fig. 6). NBPF constructs encoding other family members showed similar results (data not shown). Several stainings with markers for subcellular compartments did not result in significant overlay.

Discussion

Primate evolution coincided with the birth of several genes and gene families, and it is tempting to speculate that a causal relationship exists between evolution and the birth of these genes. Functional analysis of these genes is necessary to further unravel the recent evolution of human beings and the remarkable differences between them and their closest relatives. We identified a novel gene family, located on segmental duplications of human chromosome 1, that has no discernable orthologues in rodent genomes and observed the effects of striking species-specific gene duplication events that occurred during primate evolution. The human NBPF genes have a repetitive structure with high intergenic and intragenic sequence conservation, both in coding and noncoding regions. Analysis of a large number of NBPF cDNA clones showed that some of them contain more exons than found in the genomic sequence available to us. Two possible explanations for this phenomenon can be offered. First, regions of segmental duplications are difficult targets for the human genome sequencing project, where so-called “muted” gaps can occur due to misassembled homologous sequences (Eichler, Clark, and She 2004) resulting in deleted sequences. Second, the NBPF gene family may be polymorphic in the human population. As both coding and noncoding regions show high sequence conservation and are sometimes repeated within a gene, it is plausible that homologous recombination between different alleles can easily occur at shifted locations, leading either to deletion or duplication of genomic sequences. Studies of the genomes of different individuals have shown frequent genomic polymorphisms in the regions of the NBPF genes (Iafrate et al. 2004; Sebat et al. 2004). The presence of structural polymorphisms in the NBPF1 gene was recently shown in a fine-scale analysis of genomic variation (Tuzun et al. 2005), strengthening the second explanation. The functional implications of these structural polymorphisms are potentially of great importance as it has been shown that duplication of the CCL3L1 gene confers partial resistance to human immunodeficiency virus infection (Gonzalez et al. 2005), linking segmental duplications to pathology. The functional implications of the structural polymorphism in the NBPF1 gene remain obscure at this moment. In addition, it should be taken into account that our description of the NBPF gene family is based on the sequences of the human genome project and that gene number and structure may differ between individuals.

Previous studies of fast-evolving gene families have shown that they are often involved in sexual reproduction or in the regulation of immune responses (Cheung et al. 2003). The frequent occurrence of chromosome 1 breakpoints in male infertility (Bache et al. 2004), specifically 1q21 breaks, could indicate that one or more members of the NBPF family play a role in male reproduction. Additional evidence is provided by the fact that many cDNAs of this gene family are derived from testis. Moreover, several proteins involved in male reproduction have highly charged regions (Lahn and Page 2000), which is also true for the NBPF proteins and particularly in type 12 exons.

Recent reports showed that NBPF sequences are overexpressed in sarcoma (Meza-Zepeda et al. 2002) and non–small-cell lung cancer (NSCLC) (Petrozziello et al. 2004). This hints at a cancer-related function for one or more members of this gene family. The COAS1 transcript (Meza-Zepeda et al. 2002) was identified by hybridizing a YAC containing a genomic fragment frequently amplified in sarcomas to a cDNA library. Subsequently, the COAS1 transcript was shown to be highly expressed in tumors with genomic amplification of the 1q21 region. The analysis made by Meza-Zepeda et al. (2002) resembles ours in analyzing all NBPF transcripts as hybridization experiments.
cannot discriminate between different NBPF paralogues. From the analysis of an NBPF cDNA sequence (KIAA1245), the authors concluded that its 3′ UTR is built of repetitive elements, but according to us these repetitive elements are coding exon types 10, 11, 12, and 13. This discrepancy is due to an alternative splicing event in this transcript that skips part of the type 8 exon and induces a frameshift that results in a premature stop. We confirmed the frameshift variation in this clone by sequencing and detected similar results in a premature stop. We confirmed the frameshift due to an alternative splicing event in this transcript that shows low homology to parts of the ancestral sequences. The myomegalin genes (Beckers et al. 2001; Johnson et al. 2001; Paulding, Ruvolo, and Haber 2003). However, the human myomegalin gene has an orthologue in lower mammals, we may conclude that the NBPF genes either originated from the primate-specific (Beckers et al. 2001; Johnson et al. 2001; Paulding, Ruvolo, and Haber 2003). However, the human myomegalin gene has an orthologue in lower mammals, we may conclude that the NBPF genes either originated from the primate-specific evolution. In this respect, the above-mentioned presumptive role of NBPF genes in sexual reproduction may have played a role in the creation of species barriers. Functional analysis of the encoded NBPF proteins is necessary to determine the impact these novel genes had on recent primate evolution and on the diversity of the human population.

Supplementary Material

Tables describing the chromosomal position of the NBPF genes (Table S1), the distances between exons belonging to the same exon types (Table S2), the Ks/Ka values for human NBPF genes (Table S3), and the annotated NBPF cDNA clones (Table S4) are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jo Vandepoele, Geert Berx, and Geneviève Laureys for helpful discussions, Amin Bredan for editorial assistance, and Miek De Roover for sequence analysis. K.V. is supported by the Instituut voor de Aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen. N.V.R. is a postdoctoral researcher with the Fund for Scientific Research, Flanders (FWO). This work was supported by the FWO, the Geconcerteerde Onderzoeksacties of Ghent University, Fortis Verzekeringen (Belgium), and Interuniversity Attraction Poles Program—Belgian Science Policy.

Literature Cited


Naruya Saitou, Associate Editor

Accepted July 22, 2005