SUPPLEMENTAL NOTE

Evolutionary dynamism of the primate *LRRC37* gene family

Giuliana Giannuzzi, Priscillia Siswara, Maika Malig, Tomas Marques-Bonet, NISC Comparative Sequencing Center, James C. Mullikin, Mario Ventura, Evan E. Eichler

1. *LRRC37* family organization in mammalian genomes

1.1. *LRRC37* family in the human genome

We analyzed the genomic organization of the *LRRC37* family in the human genome, identifying the regions containing a complete or partial copy of *LRRC37*. The RefSeq database (http://www.ncbi.nlm.nih.gov/refseq/) reports seven RNA belonging to the human *LRRC37* family (four mRNA and three noncoding RNA), which are distinguished in A or B type: the mRNA *LRRC37A* (NM_014834.3), *LRRC37A2* (NM_001006607.2), *LRRC37A3* (NM_199340.2), and *LRRC37B* (NM_052888.2), and the noncoding RNA *LRRC37A4* (NR_002940.2), *LRRC37A* (NR_003525.1), and *LRRC37BP1* (NR_015341.2). One additional noncoding RNA (NR_033753.2) maps on the *LRRC37F* locus. All RefSeq entries map on chromosome 17, with the exception of NR_003525, which maps on chromosome 10. To find the additional *LRRC37* loci, we mapped the exonic sequences of the four RefSeq mRNA on the human genome assembly (hg18) using the BLAT tool at the UCSC genome browser (http://genome.ucsc.edu/). We used the SMART tool (http://smart.embl-heidelberg.de/) (Schultz et al. 1998) to analyze the domain organization in proteins encoded by *LRRC37A* and *LRRC37B*, as representatives of A and B types, respectively.

1.2. *LRRC37* family in nonprimate mammalian genomes

We searched for *LRRC37* orthologs in several mammalian clades through tblastn (Altschul et al. 1990) using the human protein sequence NP_055649.4 as the query. In the platypus genome (WUGSC 5.0.1/ornAna1) we did not identify *LRRC37*, whereas in the opossum genome (Broad/monDom5) we found similarity with exons 3–9 and 13 at chr2:206 Mbp, syntenic to human chr17:40.9 Mbp (*LRRC37A4*). Since *LRRC37* is present only in mammalian genomes, it emerged in the therian ancestor.

In the dog genome (Broad CanFam3.1/canFam3) we identified one complete copy on chr9 at 10 Mbp (*LRRC37A*) having an open reading frame (ORF) and expressed sequence tag (EST) annotation supporting the expression. Dog *LRRC37A* shares exons 1–11 with human *LRRC37A*. Additional partial copies are annotated on chr9, chr14, and unassigned scaffolds. One dog predicted RefSeq mRNA sequence similar to human *LRRC37* is annotated (XM_003639254.1) and maps on chr9:10 Mbp.

Two cow RefSeq mRNA sequences similar to human *LRRC37* are annotated: XM_002696083.1 and XM_865708.3. They both map on the cow genome (Bos_taurus_UMD_3.1/bosTau6) at only one location, chr19:45,796,744–45,846,682 (minus strand), suggesting that the cow genome has one copy of *LRRC37* at chr19:45

Mbp. The syntenic human position is chr17:40,941,209–40,983,448 (NCBI36/hg18), also confirmed by the two markers CC765011 and BZ941836 (Everts-van der Wind et al. 2005) flanking the cow *LRRC37A* and uniquely mapping on human chr17 at 40.2 and 41.3 Mbp, respectively.

The NCBI Transcript RefSeq database annotates three mouse RefSeq mRNA sequences similar to human *LRRC37*: XM_137868 and XM_906537—two predicted mRNA, and NM_001033434 (GenBank AK029868.1)—a sequenced cDNA clone from mouse testis. XM_137868 and XM_906537 are 10,149 bp long, derive respectively from the reference assembly (C57BL/6J) and alternate assembly (based on Celera), and map on mouse chr11:103,312,057–103,366,163 (minus strand), spanning a genomic region of 54 kbp. NM_001033434 is 4,576 bp long and maps on chr11:103,395,891–103,476,053 (minus strand), spanning a region of 80 kbp. Therefore, the mouse genome has two tandem copies of *Lrrc37* on chromosome 11: *Lrrc37a1* mapping at 103,312,057–103,366,163 (minus strand) and *Lrrc37a2* at 103,395,891–103,476,053 (minus strand). Knockout mice for this gene are not yet available from the International Knockout Mouse Consortium and are at the embryonic stem (ES) cell phase. Five mouse strains carrying any *Lrrc37* mutation are available from the International Mouse Strain Resource but are also in the ES cell phase.

Four rat predicted RefSeq mRNA sequences similar to human *LRRC37* are annotated: XM_003750944.1 and XM_001081536.3—mapping on chr10:92.6 Mbp (minus strand, Baylor 3.4/rn4), and XM_001081550.3 and XM_002727827.3—mapping on chr10:92.7 Mbp (minus strand). Therefore, the rat genome has two tandem copies of *Lrrc37* on chromosome 10—*Lrrc37a1* and *Lrrc37a2*—mapping at 92.6 and 92.7 Mbp, respectively, both having EST sequences supporting the expression. Through global protein sequence alignment, in both mouse and rat, *Lrrc37a1* was more similar than *Lrrc37a2* to cow *LRRC37A*, suggesting rodent *Lrrc37a1* is syntenic to the ancestral mammalian single-copy gene.

Two mRNA isolated from rat prostate are annotated (M86514.1 and BC101875.1) for rat *Lrrc37a2*, corresponding to exons 1–7 and 2–14, respectively. Rombauts and colleagues cloned a 16.5 kbp androgen-regulated mRNA (GenBank:M86514.1) (De Clercq et al. 1992). They described 4.4 kDa, 38-residue proline-rich polypeptides (PRP) as the proteolytic products of a 637 kDa precursor protein encoded by the rat *Lrrc37a2* (Heyns et al. 1982; Peeters et al. 1983; Hemschoote et al. 1988). They found the PRPs are localized in the intraluminal secretion of the rat ventral prostate. The mRNA sequencing suggested the protein is encoded by the 15 kbp exon 1, with exons 2–7 belonging to the 3′ UTR. Exon 1 has a peculiar structure with a 300 bp tandem repetitive sequence that extends 11 kbp (Figure S1). Since the *LRRC37* coding sequence in other mammals comprises at least exons 1–11 and the tandem repetitive sequence gives problems in sequence assembly in the rat genome, we checked whether the premature stop codon at the end of exon 1 could derive from sequencing errors.

We reconstructed the exon-intron structure and the coding sequence of rat *Lrrc37a2* through sequence comparison of M86514.1 mRNA, rat genomic sequence, rat BAC CH230-218D21 sequence (AC111872.4), mapping to this locus, and rat EST DN934438.1. We identified a sequence error in the M86514.1 mRNA: all sequences supported the insertion of a T between the nucleotides 15017 and 15018 of M86514.

We inserted the T in the M86514.1 sequence, merged M86514.1 and BC101875.1 sequences, and extended at 5′ the exon 1 coding sequence till the putative methionine start codon. We obtained a coding sequence of 20,646 bp, corresponding to exons 1–11 of the human *LRRC37A*.

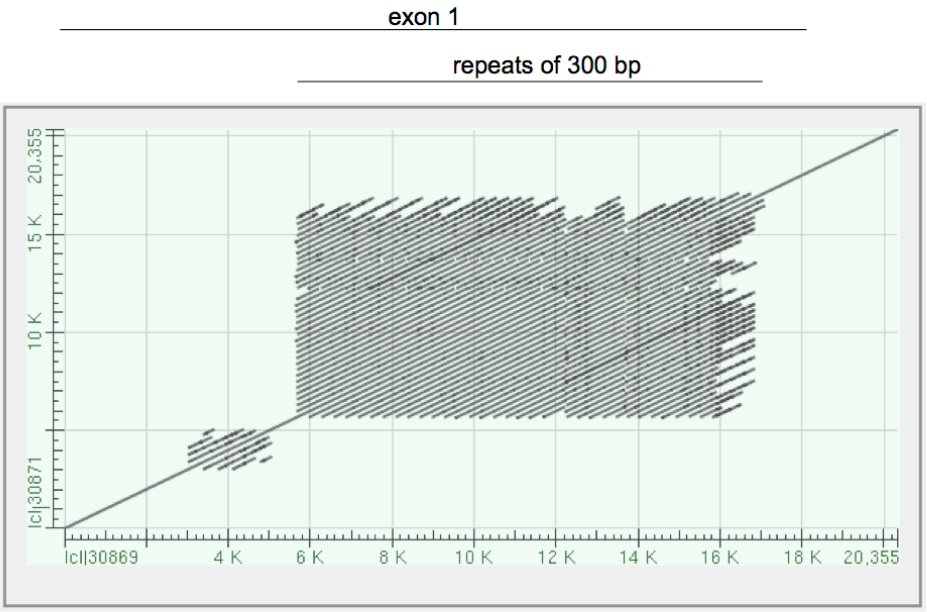


Figure S1. Blast2seq (Altschul et al. 1990) of rat *Lrrc37a2* coding sequence versus itself. The picture shows the tandem repeat of 300 bp in rat *Lrrc37a2* exon 1 sequence.

The protein products of rat *Lrrc37a2* are therefore secreted, which is consistent with our prediction of a secreted protein for human LRRC37A. Finally, the rat *Lrrc37a2* was studied several years ago and comprises 14 exons. The exon-intron structure of the gene is similar to the human one and they share exons 1–11. Our reconstructed coding sequence is 20.6 kbp long and codifies for a 766 kDa protein, which is proteolytically processed to generate polypeptides encoded by the tandem repeats through an unknown post-translational processing mechanism.

1.3. *LRRC37* family in nonhuman primate genomes

To analyze primate genomes, we used the ancestral loci of the *LRRC37* family in the human genome defined as a reciprocal best hit between human and various outgroup species, supported at least by three outgroup species, and with a size ≥1 kbp (Jiang et al. 2007). Two ancestral loci are reported, shared by human copies *LRRC37A*, *A2*, *A3*, *A4*, *A5*, *C*, and *B*: one at chr17:40,953,832–40,957,413, found in mouse, rat, and macaque, named “Anc409”, and one at chr17:41,955,779–41,956,818, found in mouse, dog, and macaque, named “Anc419”. Anc409 spans a region of 3,582 bp, including part of intron 7, exon 8, and part of intron 8; Anc419 spans a region of 1,040 bp, including exon 4, intron 4, exon 5, and part of intron 5.

We designed a probe in each ancestral locus (Anc409 and Anc419 probes) to screen primate genomic libraries. Anc409 probe includes the end of intron 7, exon 8, and the beginning of intron 8; Anc419 probe includes the end of intron 4, exon 5, and the beginning of intron 5.

For each primate species, we identified and grouped positive BAC clones containing the same gene copy through 1) end mapping on human genome assembly (hg18), 2) end mapping on orangutan (ponAbe2), macaque (rheMac2), or marmoset (calJac3) genome assemblies for CH253, CH250, and CH259 clones, respectively, and 3) PCR amplification and sequencing of a 1.2 kbp region from intron 8 and partially covering the Anc409 region, using positive BAC clones as a template.

In the PCR amplification and sequencing, we applied touchdown cycle conditions (initial denaturation 94 C 3 min; 10 cycles 94 C 30 sec, 65 C 30 sec, 72 C 2 min; 10 cycles 94 C 30 sec, 60 C 30 sec, 72 C 2 min; 30 cycles 94 C 30 sec, 55 C 30 sec, 72 C 2 min; final extension 72 C 10 min). We used the same forward primer *exon 8 F*, for all orangutan, macaque, marmoset, and lemur clones. For a reverse primer, we used *PPY intron 8 R* for CH253 clones (product size of 1.2 kbp), *MMU intron 8 R1* for CH250 clones of groups 1, 2, and 4 (product size of 1.2 kbp), *MMU intron 8 R2* for CH250 clones of group 3 (product size of 1.7 kbp), *CJA intron 8 R* for CH259 clones (product size of 1 kbp), and *LCA intron 8 R* for LBNL-2 clones (Table S7). To design primers specific for *Lemur catta*, we searched in the *Lemur catta* WGS (whole-genome shotgun) database available at the NCBI Trace Archive for the human sequence of exon 8/intron8 (chr17:41,761,522–41,762,840), using the blastn program optimized for "somewhat similar" sequences (Altschul et al. 1990). In the *Lemur catta* WGS database, we did not find any matching sequences. We then turned to the *Microcebus murinus* (gray mouse lemur) WGS database and submitted the same query. We multi-aligned the resulting entries using Clustal W (http://www.ebi.ac.uk/Tools/clustalw/) (Thompson et al. 1994), chose the same forward primer as exon 8 is very conserved, and designed a reverse primer in a region conserved among all the collected *Microcebus murinus* sequences.

PCR products were sequenced with both forward and reverse primers. Double peaks were singularly analyzed, and BAC clones presenting the same sequence, considering also the ambiguous positions, were clustered together. We multi-aligned (Clustal W) the elaborated sequences and identified BAC clones containing the same *LRRC37* copy. We mapped BAC clones representing all identified *LRRC37* copies on the respective primate metaphase chromosomes using FISH (fluorescent *in situ* hybridization) assays (Figures S2, S3, S4, 2A, and 3C).

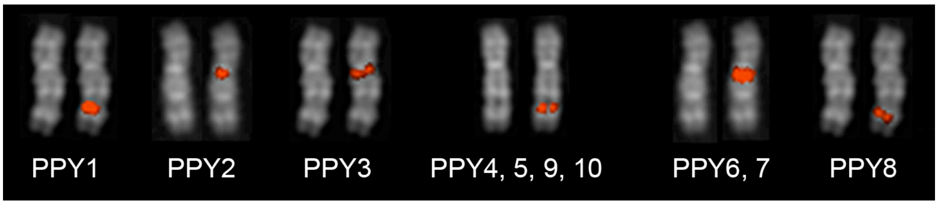


Figure S2. Chromosome location of the orangutan copies of the *LRRC37* family. FISH results of representative BAC clones for all orangutan *LRRC37* copies: CH253-10O4 (PPY1), CH253-4O2 (PPY2), CH253-149E2 (PPY3), CH253-178N18 (PPY4, 5, 9, 10), CH253-9K19 (PPY6, 7), and CH253-167E20 (PPY8) on orangutan chr17.

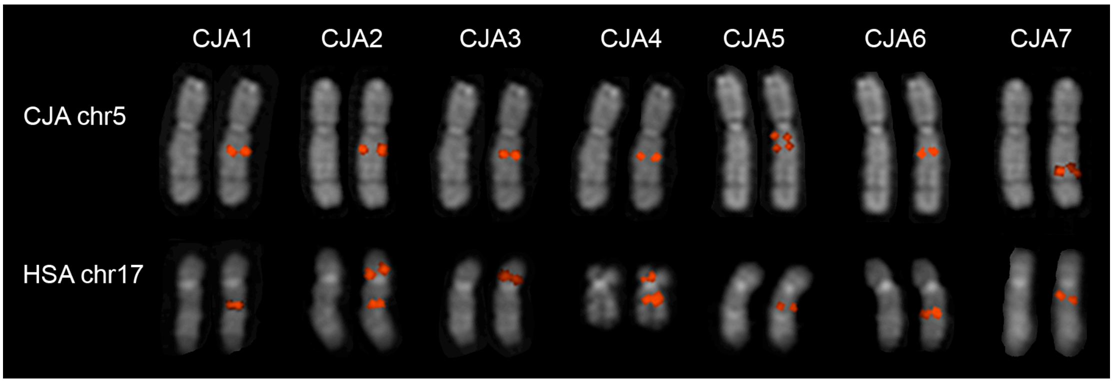


Figure S3. Chromosome location of the marmoset copies of the *LRRC37* family. FISH results of representative BAC clones for all marmoset *LRRC37* copies: CH259-1L14 (CJA1), CH259-66N11 (CJA2), CH259-10I8 (CJA3), CH259-32D5 (CJA4), CH259-152N22 (CJA5), CH259-254N24 (CJA6), and CH259-145C2 (CJA7) on marmoset chr5 (*top*) and human chr17 (*bottom*).

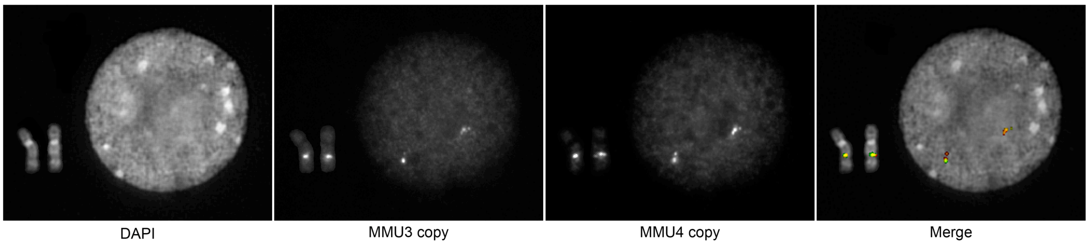


Figure S4. Co-localization of MMU3 and MMU4 *LRRC37* copies. FISH of CH250-221J22 (Fluorescein-labeled, MMU3 copy) and CH250-269O23 (Cy3-labeled, MMU4 copy) on macaque chromosomes 16 and nucleus. In the merged image, CH250-221J22 and CH250-269O23 signals are green- and red-colored, respectively.

1.3.1. *LRRC37* family in the orangutan genome

The monophyletic relationship of A and B type *LRRC37* copies in both human and orangutan is most likely explained by independent expansions of these core families in both lineages; although it is also possible that convergent evolution as a result of gene conversion could have homogenized the clusters independently. However, this seems unlikely especially since the duplicated genes are distributed in an interspersed fashion throughout chromosome 17 (Figure 1A).

Seven BAC clones of the CHORI-276 library containing a copy of *LRRC37* are being sequenced: CH276-403D18 (AC206276.2), CH276-313A19 (AC216100.2), CH276-234O9 (AC210931.1), CH276-461H24 (AC212980.2), CH276-10M16 (AC212589.2), CH276-344O6 (AC210533.4), and CH276-449H2 (AC206550.4). Through sequence comparison with intron 8 tags of each orangutan copy (PPY1–10), we identified the copies embedded in these clones. CH276-403D18 contains the exons 6–10 of the PPY1 copy; CH276-313A19 contains the exons 3–13 of the PPY5 copy; CH276-461H24 contains the PPY7 copy. The remaining clones contain partial copies without the Anc409 region, so they were not retrieved through filter screening. CH276-10M16 and CH276-344O6 contain the same *LRRC37* copy.

1.3.2. *LRRC37* family in the macaque genome

According to the BAC-end anchored position in the rheMac2 assembly, macaque positive clones were arranged in four groups: 1) 12 clones spanning the region chr16:48,901,300–49,282,700, corresponding to human chr17:34,271,847–34,617,881, the *LRRC37F* locus (partial copy, exons 1–5) (Figure 1A); 2) 11 clones mapping on chr16:55,520,555–55,991,873, not mapping to a unique location in the human genome; 3) 11 clones mapping on chr16:63,321,613–63,626,197, syntenic to human chr17:63,401,640–63,854,885, the *LRRC37C* locus (Figure 1A); and 4) two clones mapping on macaque chr9, syntenic to human chr10 (Table S3). In summary, all clones mapped on macaque chr16 (syntenic to human chr17) with the exception of CH250-230P14 and CH250-287L11, which mapped on chr9.

Eight CH250 clones are being sequenced: CH250-6G22 (AC239129), CH250-14N17 (AC239243), CH250-51G5 (AC241896), CH250-197J22 (AC191449), CH250-219M3 (AC241829), CH250-221J22 (AC241249), CH250-229I23 (AC242077), and CH250-269O23 (AC143065). The combination of PCR amplification and sequencing on single BAC clones and the analysis of BAC draft sequences, along with the end-anchored position data, allowed for the identification of seven *LRRC37* copies in macaque. Clones of group 4, mapping on chr9, and clones positive only at Anc419 were PCR negative for Anc409 as both lack the annealing site for the reverse primer, designed in intron 8: in the first case because they contain a retrocopy, in the latter because they do not span the whole gene and lack this segment. Three copies, MMU1, 2, and 7 (BAC group 2), are clustered on chr16:55,520,555–55,991,873 and are annotated in the macaque assembly as well. Two copies, MMU3 and 4 (BAC group 1), derived from a macaque-specific duplication; they are collapsed in the macaque assembly in one copy at 49 Mbp. Two copies, MMU5 and 6 (BAC group 3), derived from a macaque-specific duplication; they are collapsed in the macaque assembly in one copy at 63 Mbp. Although PCR amplification and sequencing of group 3 clones resulted in fully identical sequences, their number, around twice the library redundancy, suggests they represent two separate, highly similar loci originating from a recent macaque duplication, each containing a copy of *LRRC37*. The comparison of draft sequences of CH250-197J22 and CH250-229I23 clones showed they contain the same *LRRC37* copy and belong to the same locus. Although gene conversion has homogenized macaque clustered copies 1, 2, and 7, their most likely human orthologs are *LRRC37A4*, *B*, and *A5*, respectively, suggested by end-mapping, BLAT of gene sequences on the human genome, and net alignments of human to macaque genomes (UCSC genome browser). Particularly, for MMU1, the exons 1–10 genomic region is best syntenic to the human *LRRC37A4* genomic region, whereas the exons 6–7 region is best syntenic to the human *LRRC37B* genomic region; for MMU7, most of the gene region (exons 1–5) is best syntenic to the human *LRRC37A5* genomic region.

We retrieved exon sequences of macaque *LRRC37* copies from CH250 BAC draft sequences, using BLAST sequence similarity searches (Altschul et al. 1990) and human *LRRC37A* and *LRRC37B* coding sequences as queries. MMU1 and MMU2 copies have exons from 1 to 10 plus exon 15; MMU3 copy has exons 1, 2, 3, 4, 5, 9b', 9b", 10, 12, and 13; MMU5/6 has exons 1, 2, 3, 8, 9, 10, and 11; MMU7 has exons from 1 to 8, 9b', 9b", 10, and 15. The MMU1 region has downstream exon 11 and part of exons 13 and 14. The gene organization of MMU4 could not be analyzed since no BAC sequence is available.

We found that only MMU1, MMU2, and MMU7 preserve an ORF, whereas MMU3 and MMU5/6 have disrupted reading frame. Notably, the gene modules of macaque 1, 2, and 7 copies show they all have exon 15 and lack exons 11–14, both B type features in human, suggesting human and macaque copies are both mosaic of gene segments deriving from multiple copies, thus hindering the clear reconstruction of phylogenetic relations.

1.3.3. *LRRC37* family in the lemur genome

Most lemur BAC ends contained repetitive elements and could not be uniquely mapped on the human genome. The few with nonrepetitive end sequences mapped on human chr17 to multiple locations with a similar score. PCR amplification and sequencing showed several double peaks in most electropherograms, indicating the presence of at least two copies of *LRRC37* in the same lemur BAC clone. LB2-153A17, LB2-163F7, LB2-239M16, LB2-249G20, LB2-249J24, LB2-250G20, and LB2-261P12 clones did not show double peaks and each contained one copy of *LRRC37*. LB2-163F7 and LB2-249J24, as well as LB2-249G20 and LB2-250G20, contained the same copy (LCA1 and LCA10, respectively). We identified a total of 45 different positions among the sequences of all LBNL-2 BAC clones, resulting in 12 different *LRRC37* copies in the *Lemur catta* genome (Table S5). All clones except LB2-261P12 and LB2-239M16 could be bunched in three groups respectively containing three, five, and two tandem copies (Table S5, Figure 3A). The reported order of the different groups does not reflect the real organization of the 12 copies on lemur chromosome 17 as this analysis could not provide this information. The LB2-288P4 draft sequence (AC234058.2) showed three copies of *LRRC37* organized in a tandem configuration and direct orientation, confirming PCR amplification and sequencing data for this clone (Figure 3B).

2. Phylogenetic and evolutionary analyses in primates

2.1. Phylogenetic analysis

We constructed a phylogenetic tree based on intron 8 sequence using the macaque and lemur sequences previously collected and retrieving the human and mouse ones from the hg18 and mm9 genome assemblies, respectively.

We constructed a phylogenetic tree using a 2 kbp region from intron 1. We retrieved the human sequences from the hg18 assembly, the mouse sequence from the mm9 assembly, and CJA7 sequence from the calJac3 assembly. We retrieved macaque sequences MMU1, 2, 3, 5and6, and 7, marmoset sequences CJA4, 5, and 6, and lemur sequences LCA1 and 2 from sequenced CHORI-250, CHORI-259, and LB2-288P4 BAC clones, respectively. To get the remaining sequences, we PCR-amplified (PCR Master, Roche) and sequenced using *exon 1 F* and *intron 1 R* primers (Table S7) and CHORI-253, 250, and 259 BAC clones as a template. In particular, we used CH253-10O4 (PPY1), CH253-4O2 (PPY2), CH253-149E2 (PPY3), CH253-65E19 (PPY4), CH253-97O4 (PPY5), CH253-9K19 (PPY6), CH253-31B12 (PPY7), CH253-167E20 (PPY8), CH253-122B24 (PPY9), CH253-228E23 (PPY10), CH250-269O23 (MMU4), CH259-46C7 (CJA1), CH259-66N11 (CJA2), and CH259-10I8 (CJA3) BAC clones. We completed the sequences obtained using the forward primer of orangutan copies PPY2, 3, 6, and 7 with sequences retrieved from ponAbe2 assembly: chr17:26871067–26872884 (PPY2), chr17:22896709–22898529 (PPY3), and chr17:57896782–57898602 (PPY6 and 7).

2.2. Comparative FISH analysis

We hybridized human BAC probes RP11-141H9 containing *LRRC37A2* as representative of A type copies, RP11-640N20 containing *LRRC37B* as representative of B type copies, and RP11-893O13 containing the *LRRC37F* copy as well as macaque CH250-219M3 containing MMU1, 2, and 7 copies, CH250-221J22 containing MMU3 copy as representative of MMU3 and 4 copies, and CH250-197J22 containing MMU5-6 copy on human, orangutan, and macaque metaphase chromosomes (Figure S5).

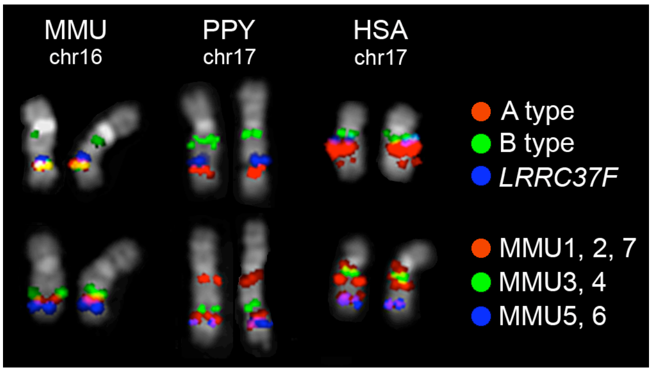


Figure S5. FISH pattern of RP11-141H9 (red), RP11-640N20 (green), and RP11-893O13 (blue) (shown on top), and CH250-219M3 (red), CH250-221J22 (green), and CH250-197J22 (blue) (bottom) on macaque, orangutan, and human chromosomes, confirming the evolutionary history of the *LRRC37* family in Catarrhini.

2.3. Evolutionary analyses

We examined the general evolutionary pressure on primate *LRRC37* genes, analyzing human, chimpanzee, orangutan, and macaque copies with complete gene structure in the ORF. We analyzed whole gene and exon 1 evolutions, the latter because exon 1 is the main difference between A and B types (Figure S6).

2.3.1. Sequence alignment and annotation of human, orangutan, chimpanzee, and macaque gene models

Human coding DNA sequences of the *LRRC37* family were retrieved from RefSeq entries: NM_014834 (*LRRC37A*), NM_001006607 (*LRRC37A2*), NM_199340 (*LRRC37A3*), and NM_052888 (*LRRC37B*). Human *LRRC37A4*, chimpanzee, and orangutan coding sequences were retrieved from the corresponding assemblies (hg18, panTro2, and ponAbe2).

Chimpanzee and orangutan assemblies, like human, have several partial duplications of *LRRC37*. We initially retrieved genomic segments containing the sequences of exons 1a, 1b, 2–8, 9, 10–14, and 15, separately, using the BLAT tool and human sequences as queries, and identified in each group exon sequences having an ORF. We kept complete copies, defined as genes having at least exon 1 and one or more exons of the LRR region (exons 2–8) that preserve an ORF. Coding sequences of MMU1, MMU2, and MMU7 copies were obtained from CH250-6G22 and CH250-219M3 draft sequences.

At the end, we recovered five complete gene sequences in human, four in chimpanzee, two in orangutan, and three in macaque; five exon 1 sequences in human, five in chimpanzee, three in orangutan, and three in macaque. According to our analysis of the orangutan *LRRC37* family, PPY2 has exon 8 sequence and PPY5 is complete. However, these loci have sequence gaps in the ponAbe2 reference, which prevented us from acquiring the complete sequences. The coding sequences were aligned using an amino acid-based alignment (Wernersson and Pedersen 2003).

2.3.2. Evolutionary rates

To determine the evolutionary history of the copies in primates, all coding DNA sequences were analyzed by maximum likelihood using PAML (Yang 1997). We used the ratio d_N/d_S (ω), which compares rates of nonsynonymous substitutions to the rate of synonymous substitutions, as a measure of evolutionary constraint. In a scenario where there is a strong functional pressure on the protein, this ratio will tend to be less than 1 since the nonsynonymous mutation would tend to be removed from the population. Alternatively, if a gene is under neutrality, the ratio should be close to 1. Finally, in cases of pervasive adaptive evolution the ratio will be higher than 1 as the new nonsynonymous substitutions acquired will be fixed more rapidly than the neutral sites. We used a codon-substitution branch model (CODEML) (Yang and Nielsen 2002) and applied a free codon-substitution model (in which every branch of the tree is allowed to have different d_N/d_S) to the accepted phylogeny for the species to estimate the evolutionary pressures at different times during the evolution of these genes.

The analysis of the whole-gene evolution of the gene family revealed that, first, all four human A type copies are highly similar resulting in ω values highly driven by lack of power (Figure S6). There are different branches with an excess of nonsynonymous substitutions but no synonymous substitutions. Second, some macaque copies might have been under neutral evolution, with values of ω of 0.98 and 1.84, suggesting they might be pseudogenes. As expected, since the LRR region is highly conserved, the analysis of exon 1 revealed an increase of evolutionary rates (14 vs. 10 values now are greater than 1), but considering we are analyzing less substitutions, it might be a result of increasing statistical noise. As observed before, most of the branches leading to the macaque copies are in values close to 1.

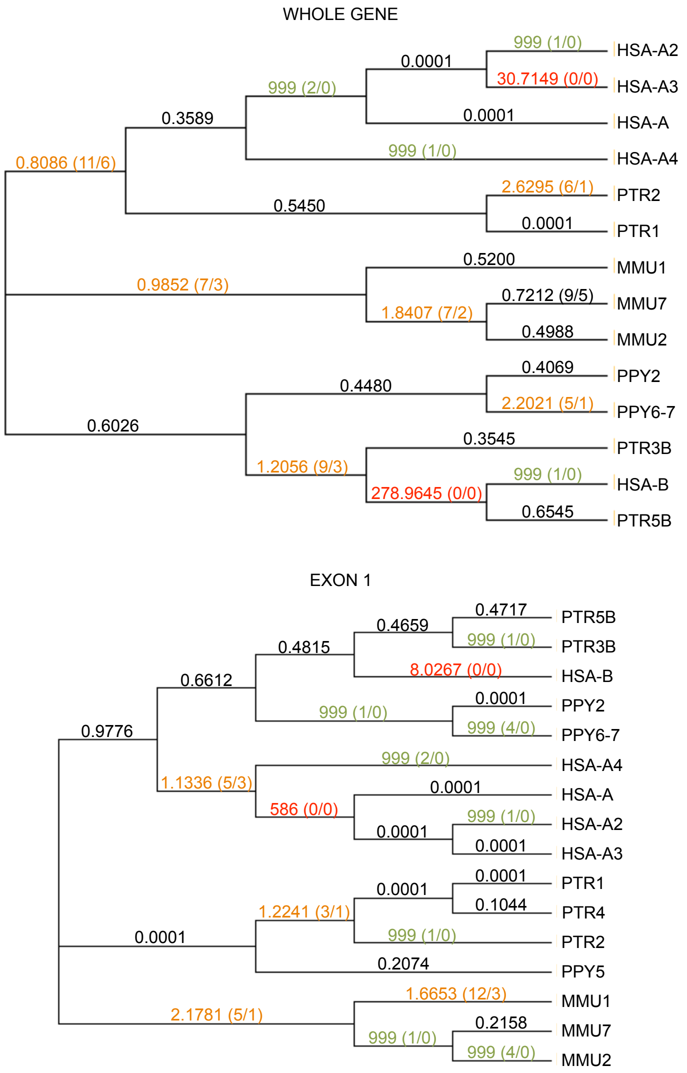


Figure S6. Evolutionary rates of *LRRC37* in primates. Evolutionary rates of the whole gene (*top*) and exon 1 (*bottom*) are shown. Ratios d_N/d_S (ω) are reported and color-coded next to branches: ω>1 (potential positive selection) in green, ω = 1 (potential neutral evolution) in orange, and ω<1 (potential negative selection) in black. Ds = 0, as indicative of branches without power to infer selection, are in red.

3. Evolutionary analysis of the ancestral locus in mammals

We analyzed the evolutionary conservation of the *LRRC37* ancestral locus in mammals. The coding sequence of human *LRRC37A4* is partly conserved in the exon-intron structure: it comprises exons 1–5 and 8 because of a premature stop codon due to the different splice acceptor site within intron 8 (see alternative splicing and copy-specific expression for human *LRRC37*). In the other mammals, the exon-intron gene structure is conserved for exons 1–11. However, the predicted exon 1 sequences are extremely variable in length unlike the short exons 2–8 (Figure 5) and determine dramatic changes in the ORF length during mammalian evolution (7668 bp in cow, 8145 bp in dog, 9555 bp in rat *Lrrc37a1*, 20,646 bp in rat *Lrrc37a2*, 10,146 bp in mouse *Lrrc37a1*, 11,178 bp in mouse *Lrrc37a2*, 5361 bp in macaque MMU1, and

5100 bp in human *LRRC37A*). The translated proteins maintain the signal peptide (except for dog and rat *Lrrc37a1*), the repetitive motifs in exon 1 sequence, the leucine-rich repeats, and the transmembrane domain (except for human *LRRC37A4*) (Figure 5). Rat *Lrrc37a1* is incomplete at 5′ because of a sequence gap in the genomic reference.

We identified the consensus sequence of the repetitive motif in rat *Lrrc37a2* exon 1 by Tandem Repeat Finder (Benson 1999). We then analyzed the repetitive structure of exon 1 in mammalian genes (option onlyintra) and the presence of the rat consensus sequence in mammalian exon 1 sequences (option onlyinter) using the Miropeats software (Parsons 1995) (Figure S7). We also searched for the number of repetitions of the rat consensus sequence in cow *LRRC37A*, dog *LRRC37A*, macaque MMU1, human *LRRC37A4*, human *LRRC37A*, rat *Lrrc37a1*, mouse *Lrrc37a1*, mouse *Lrrc37a2*, and rat *Lrrc37a2* exon 1 sequences through blastn (Altschul et al. 1990). Using both Miropeats and blastn, we found no repetitions of the rat consensus sequence in the cow, dog, mouse *Lrrc37a1*, MMU1, and human sequences; we discovered 10 and 38 repetitions in the mouse and rat *Lrrc37a2* sequences, respectively, through blastn. We identified the protein consensus motif of rodent LRRC37A2 using MEME (Bailey and Elkan 1994) (Figure S8).

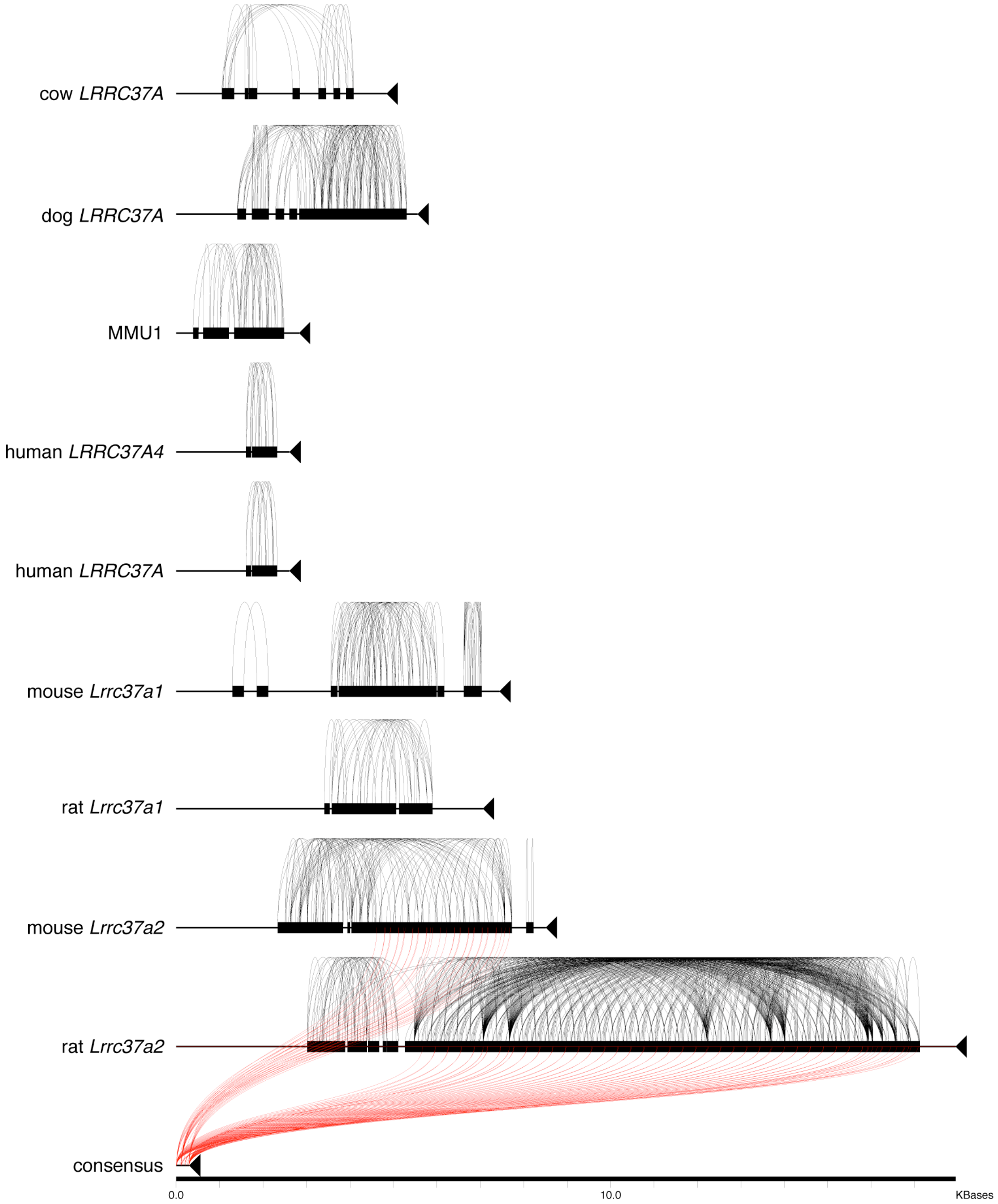


Figure S7. Miropeats output (threshold score 50) showing internal repeats (black lines) and repetitions of the rat motif (consensus) (red lines) in mammalian *LRRC37* exon 1 sequences.

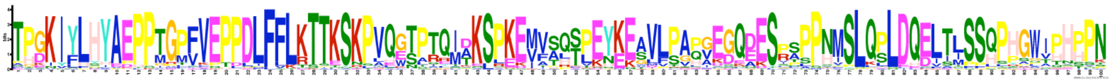


Figure S8. Protein consensus motif of the repetition identified in rodent LRRC37A2.

We then compared the ancestral copies (cow *LRRC37A*, dog *LRRC37A*, mouse *Lrrc37a1*, rat *Lrrc37a1*, macaque MMU1, and human *LRRC37A4*) as well as the duplicated coding sequences of mouse *Lrrc37a2*, rat *Lrrc37a2*, and human *LRRC37A* into the mammalian evolutionary tree, and we simulated different scenarios. Note, we could not include marmoset, orangutan, gorilla, and chimpanzee sequences because there are sequence gaps in the corresponding genomic regions and sequenced BAC clones are not anchored to a specific ancestral locus. First, we used a "one-ratio" model to test a uniform d_N/d_S across all the branches of the phylogenetic tree for exons 2 to 8 (Table S6). This model was markedly different than the neutral scenario (ω = 0.417 for exons 2–8, P-value <0.001). Second, we applied a "two-ratio" model in which we allowed the branch of interest to vary freely compared to the background (rest of branches). We interrogated the human branch (A copy) for positive selection (P-value = 0.35) and the macaque (P-value = 0.048), dog (P-value = 0.002), cow (P-value = 0.04), mouse a2 copy (P-value = 0.055), and rat a2 copy (P-value = 0.153) for negative selection with a likelihood ratio test. Results suggested that the human copy cannot be differentiated from neutral selection and that the macaque, dog, and cow copies are significantly more conserved.

4. Expression analyses

4.1. Notes for primers

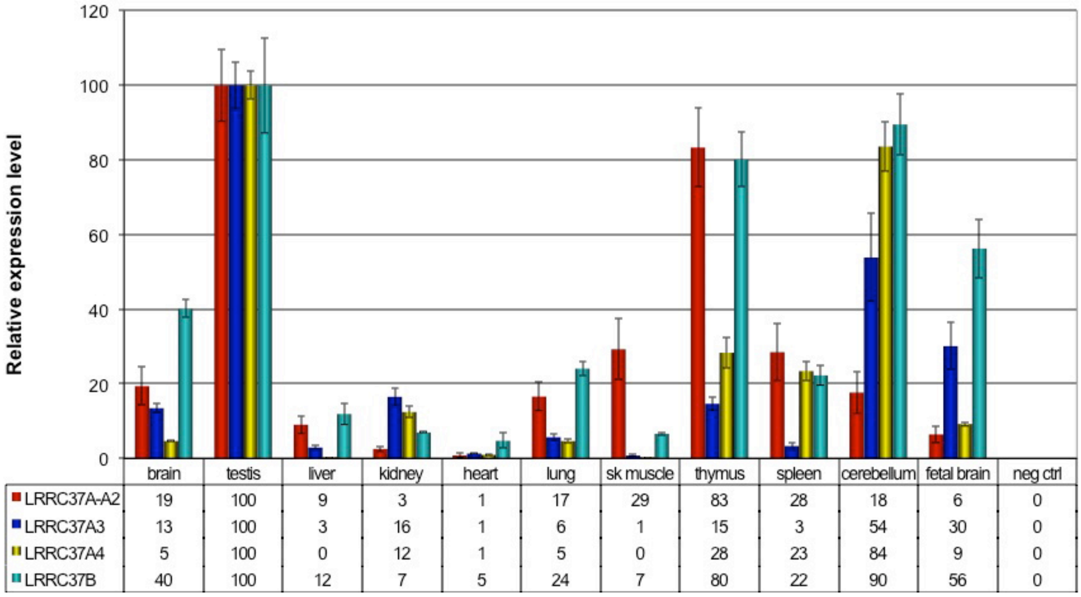
The amplification of mouse LRR regions was performed using *LRR mouse 1 F* and *R* primers for the *Lrrc37a1* copy and using *LRR mouse 2 F* and *R* primers for the *Lrrc37a2* copy. The amplification of LRR regions was performed using two different pools of degenerated primers for human and macaque samples to amplify all copies of the gene family: for human *LRR F1*, *F2*, *F3*, *R1*, *R2*, and *R3*; for macaque *LRR F2*, *F4*, *F5*, *F6*, *F7*, *R1*, and *R4*. To amplify the LRR region of the *LRRC37A4* copy, the forward primers *LRR F1*, *F2*, and *F3* and the reverse primer *LRRseqA4 R* were used (Table S7). *LRRseqA4 R* was designed in the additional *LRRC37A4* exon 9 segment, a sequence unique to *LRRC37A4* mRNA available owing to the use of a different acceptor splice site for intron 8 (see "Alternative splicing and copy-specific expression for human *LRRC37*" in Results section).

4.2. Tissue expression analyses

The tissue expression pattern of the LRR region of *LRRC37A4* showed a slightly different result than the one of the entire gene family in RT-PCR assays. Testis remained the tissue with the highest expression level; liver and skeletal muscle showed no evidence of expression; lung, thymus, spleen, cerebellum, and fetal brain showed a higher expression than brain, kidney, and heart (Figure 6B).

Interestingly, there is a clear difference between human and macaque in the expression of the *LRRC37* family. In both species the tissue with the highest expression level is testis, but whereas in macaque all the other tissues have a very low and almost even expression level, in human some tissues gain upon testis in the expression of the *LRRC37* family. Particularly, cerebellum, thymus, spleen, and fetal brain showed a significant high expression, respectively of 71%, 66%, 25%, and 22% of testis expression level (Figure 6C). If compared to heart—the tissue with the lowest expression level, the expression of the *LRRC37* family in these tissues increased by 36-, 33-, 12-, and 11-fold, respectively.

The *LRRC37* family tissue expression pattern in macaque is halfway between the mouse and human expression patterns. These data show that in the adult organism the regulation of the expression has evolved from a testis-specific expression in mouse, to a ubiquitous expression in all tissues in macaque and human. Finally, we designed four distinct quantitative RT-PCR assays to evaluate the expression of *LRRC37A/A2*, *A3*, *A4*, and *B* copies in the same panel of human tissues (Figure S9).



	brain	testis	liver	kidney	heart	lung	sk muscle	thymus	spleen	cerebellum	fetal brain	neg ctrl
🟥 LRRC37A-A2	19	100	9	3	1	17	29	83	28	18	6	0
🟦 LRRC37A3	13	100	3	16	1	6	1	15	3	54	30	0
🟨 LRRC37A4	5	100	0	12	1	5	0	28	23	84	9	0
🟩 LRRC37B	40	100	12	7	5	24	7	80	22	90	56	0

Figure S9. Human tissue quantitative expression of the *LRRC37* family members. Human tissue quantitative expression study of the *LRRC37* family copies performed using a panel of human tissue total RNA (Clontech). Values were normalized to *UBE1* expression. C_T values were elaborated by the comparative C_T method (Livak and Schmittgen 2001). The highest value of expression in each assay (testis in all the four cases) was set to 100. Error bars indicate standard error of the mean.

4.3. Analysis of additional total RNA panels

Two additional human tissue total RNA panels from other companies (Stratagene and Capital Biosciences) were tested for the expression of *LRRC37A* copies, which showed different results from the Clontech panel (Figure S10). Six tissues were analyzed: testis, heart, skeletal muscle, thymus, spleen, and cerebellum. The total RNA set from Stratagene showed thymus as the tissue with the highest expression, increased by 41-fold compared to heart, and there was not a significant higher expression in testis and cerebellum. The total RNA set from Capital Biosciences did not show any tissue with a significant higher expression than the others.

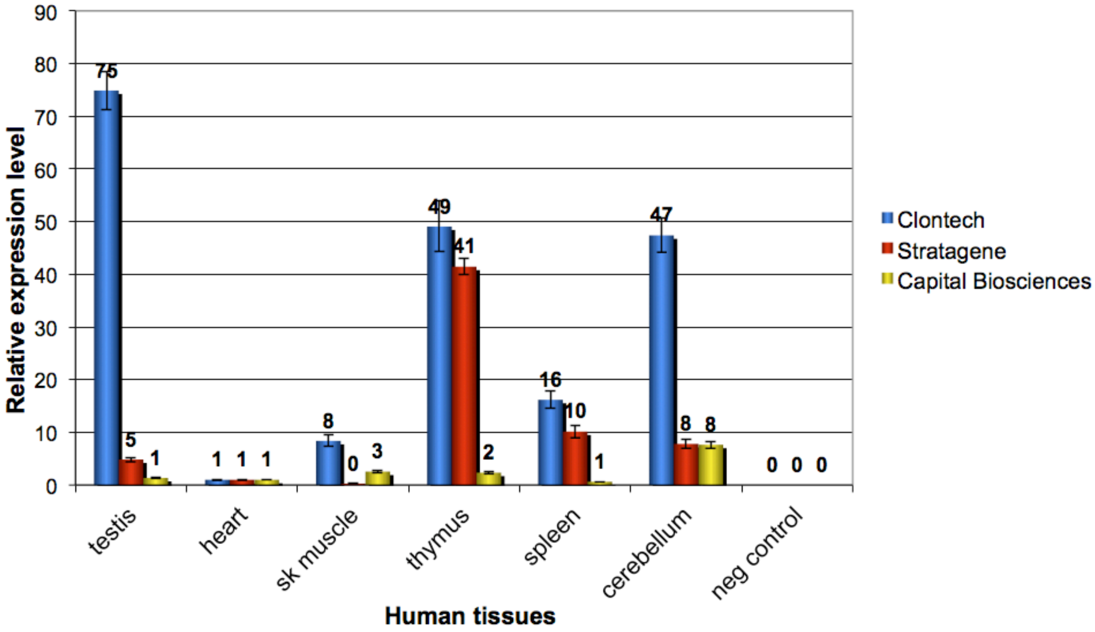


Figure S10. Comparison of quantitative *LRRC37A* expression in three different human tissue total RNA panels. Quantitative expression profiling of *LRRC37A*, *A2*, *A3*, and *A4* in three different panels of human tissue total RNA (Clontech, Stratagene, and Capital Biosciences). C_T values were analyzed by the comparative C_T method and normalized to *UBE1* expression. Expression levels are relative to heart. Error bars indicate standard error of the mean.

These tissue expression profiles did not confirm the results obtained using the Clontech panel, where a higher level of expression distinguished three tissues: testis, cerebellum, and thymus. The most unexpected result, present in both additional panels, is the low level of expression in testis. In the Stratagene panel, thymus stands out from all the other tissues, presenting the highest level of expression. In the Capital Biosciences panel all the tissues present an almost even level of expression, with an eightfold increase in cerebellum when compared to heart. The reason for this variability is not known so far and its significance might concern several aspects. For example, the number and age of donors of the different tissues is not the same; *LRRC37* is also subject to copy number variation (Alkan et al. 2009) and potential effects at the transcriptome level are not known.

4.4. RNA-seq and GeneChips expression data

We retrieved RNA-seq data regarding *LRRC37* of mouse, macaque, orangutan, gorilla, chimpanzee, bonobo, and human tissues from the publication of Brawand and colleagues (Brawand et al. 2011) (Figure S11). We normalized the expression levels among different tissues of the same species using the *GAPDH* expression levels (we could not use *UBE1* since the Ensembl gene ID is not available for all species). These data confirmed the testis-exclusive expression in the mouse and showed the emergence of a more ubiquitous expression in apes. Further, the higher expression in the cerebellum when compared to the expression level in tissues other than the testis was common among apes, suggesting it emerged in the great ape ancestor and has been conserved in the last 15 million years of hominid evolution.

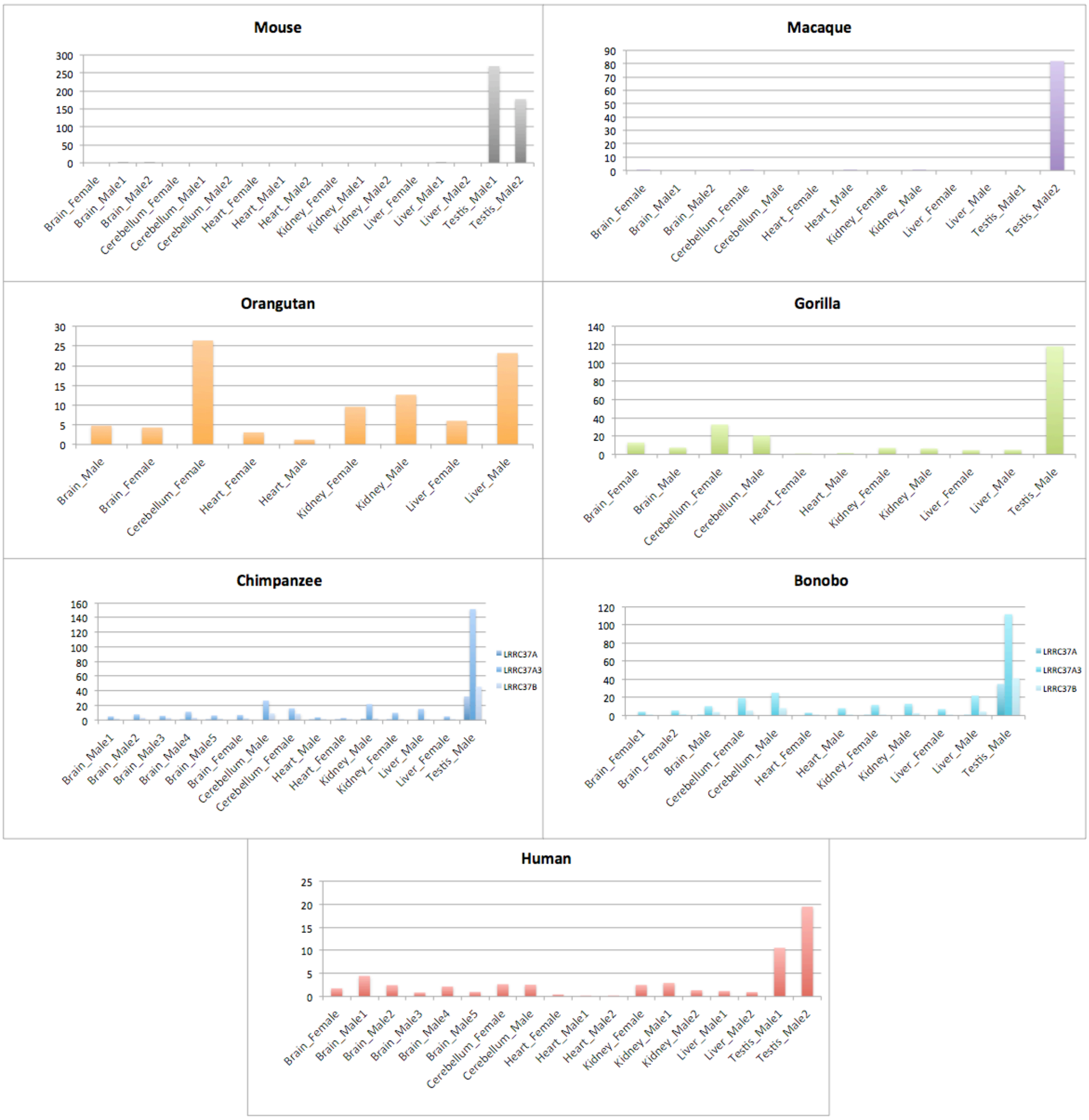


Figure S11. Tissue RNA-seq data of *LRRC37* in different mammals from the work of Brawand et al. Values were normalized according to the *GAPDH* expression.

We searched for *LRRC37A* expression data from GNF (The Genomics Institute of the Novartis Research Foundation) using Affymetrix GeneChips (Su et al. 2002). Among the human tissues analyzed, the expression was highest in fetal brain and molt-4, a T lymphoblast cell line derived from a patient with acute lymphoblastic leukemia.

4.5. Human diversity of the *LRRC37* family

RT-PCR amplification of the LRR region in macaque and human tissue total RNA showed several products on agarose gel derived from alternative splicing (Figure 6). We purified (MinElute PCR Purification Kit, Qiagen) and cloned in pGEM-T Easy Vector (Promega) the RT-PCR products of the amplification of HeLa cells, G248 lymphoblastoid cell line, human cerebellum, and human testis total RNA. Moreover, we gel-purified (MinElute Gel Extraction Kit, Qiagen) and cloned the minor products of mouse testis total RNA amplification with primers designed for *Lrrc37a2* copy.

Recombinant constructs were transformed in TOP10 Chemically Competent E. Coli cells (Invitrogen). We randomly picked up several colonies and extracted plasmid DNA using the Mini-prep protocol. Inserts were fully sequenced using *sp6_plasmid* and *t7_plasmid* primers (Table S7) and analyzed in three aspects: 1) the paralogous locus that triggered the transcription, using the BLAT tool (UCSC genome browser) and annotating the locus with the highest score (in some cases, there was the ambiguity among two or three loci); 2) the exon content; and 3) whether the sequence was in an ORF or not.

The total number of sequenced and positive clones was 39, 40, 86, 42, and 21 for HeLa cells, G248 cell line, human cerebellum, human testis, and mouse testis, respectively. The results of all the analyses are reported and summarized in Figure 7.

Since HeLa cells are aneuploidy, we addressed their endogenous genomic organization of the *LRRC37* family through a FISH assay with BAC clones RP11-141H9, RP11-640N20, and RP11-893O13 on HeLa cells chromosome metaphases. We found that HeLa cells have four chromosomes 17 with the same *LRRC37* organization observed for the human lymphoblastoid cell line (Figure S12).

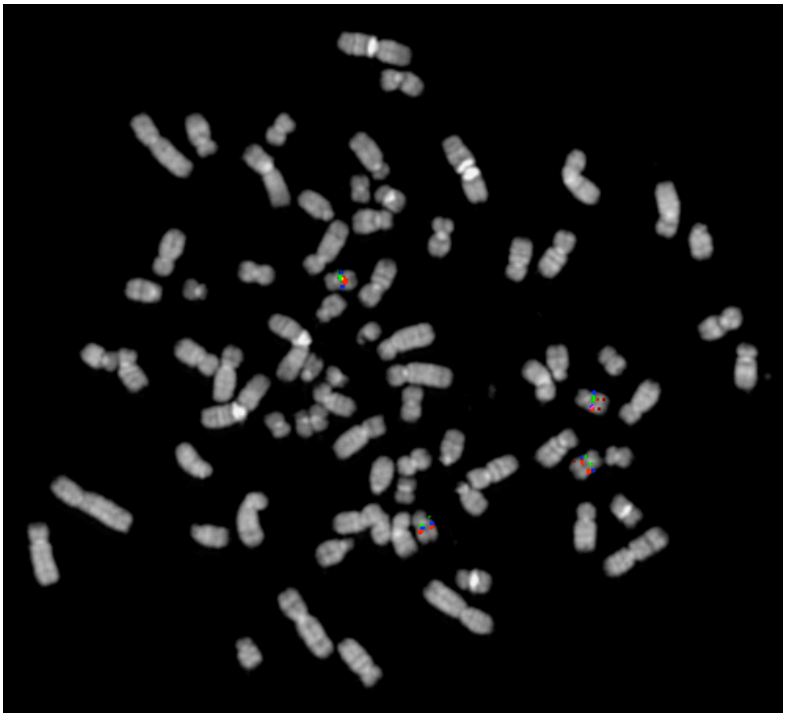


Figure S12. FISH results on HeLa cells metaphase chromosomes of BAC probes RP11-141H9 (red), RP11-640N20 (green), and RP11-893O13 (blue).

5. Subcellular localization of the LRRC37A protein

We amplified the full-length coding sequence of the *LRRC37A2* paralog from HeLa cells cDNA and cloned it in mammalian expression vectors to generate a recombinant LRRC37A protein bearing a FLAG-tag either at the amino terminus or at the carboxyl terminus. We transiently transfected the two recombinant plasmids to HeLa cells in two parallel experiments and determined the expression of the recombinant protein by immunofluorescence and Western blot analyses. HeLa cells seeded the same day and under the same conditions were left nontransfected and represented the negative control. We used the antibody binding the FLAG epitope and a polyclonal antibody developed to recognize human LRRC37 proteins.

In immunofluorescence, HeLa cells transfected with the N-terminal tagged construct did not show any strong labeling with the anti-FLAG antibody, while HeLa cells transfected with the C-terminal tagged construct showed a staining of the Golgi apparatus, vesicle boundaries, and plasma membrane (Figure 8A). Moreover, some transfected cells presented a change in their shape and a plasma membrane deformation with the formation of filopodia-like protrusions.

In Western blot, anti-FLAG antibody did not detect any band in the lysate of cells expressing the N-terminal tagged protein (first line in Figure 8B), while it revealed multiple bands in the lysate of cells expressing the C-terminal tagged protein: two close upmost bands with a molecular mass higher than 260 kDa, one band around 100 kDa, two close bands around 70 kDa, and the most abundant band around 20 kDa (second line). The same cell lysates were also probed with anti-LRRC37 antibody. This antibody revealed the same bands in both lysates, i.e. two close high molecular weight bands, corresponding to the two upmost bands detected by anti-FLAG antibody in cells expressing carboxyl terminal tagged protein. Anti-FLAG antibody did not detect any band in both conditioned media, while anti-LRRC37 antibodies revealed two close bands of high molecular weight at the same height of the two upmost bands present in cell lysates. The two bands in lysates and the ones in conditioned media appeared different in the relative abundance: in lysates the higher band was more abundant than the lower one, whereas in conditioned media the lower one was much more abundant (Figure 8C). The double band might be due to a phosphorylated or glycosylated form of the protein.

The most likely reason for the lack of recognition of N-terminal tagged LRRC37A by anti-FLAG antibody in both immunofluorescence and Western blot assays is the removal of the FLAG-tag from the recombinant protein as part of the signal peptide. Thereby, the LRRC37A protein tagged at the N-terminus cannot be traced with anti-FLAG antibody as it loses the tag during the posttranslational processing (Figure S13). The two upmost bands in both cell lysates corresponded to the complete recombinant protein, as both antibodies in the lysate of cells expressing the C-terminal tagged protein recognized them.

The FLAG-tag weighs 1 kDa and the molecular weight of the recombinant FLAG-tagged LRRC37A protein is 190 kDa. Since the molecular weight estimated according to its running behavior on an SDS-PAGE gel is higher than 260 kDa, the protein presented a very low electrophoretic mobility. This behavior might be due to some physical and chemical characteristics of the protein itself. Moreover, the running delay of LRRC37A was also observed in the GST-fusion LRRC37A whole protein and protein fragments, revealing that the portion that confers this characteristic is the moiety encoded by exon 1 (see “GST-fusion LRRC37A full protein and protein fragments”).

The multiple bands at lower molecular weights revealed by the anti-FLAG antibody in the lysate of cells transiently transfected and expressing the C-terminal tagged

LRRC37A indicate that the recombinant protein might be proteolytically processed or is degraded. None of these bands are recognized by anti-LRRC37 antibody.

C-terminal FLAG-tagged LRRC37A recombinant protein revealed in the conditioned media by anti-LRRC37 antibody is no longer detected by anti-FLAG antibody. The most abundant band detected by anti-FLAG antibody in the lysate of cells expressing the protein tagged at the C-terminus is the one with the lowest molecular weight, around 20 kDa. This protein should include the N-terminal juxta-membrane region, the transmembrane domain, and the C-terminal tail. These two observations suggest that the C-terminal FLAG-tagged LRRC37A recombinant protein transiently expressed in HeLa cells might be delivered to the plasma membrane, proteolytically processed in the extracellular juxta-membrane domain, and secreted outside of the cell (Figure S13).

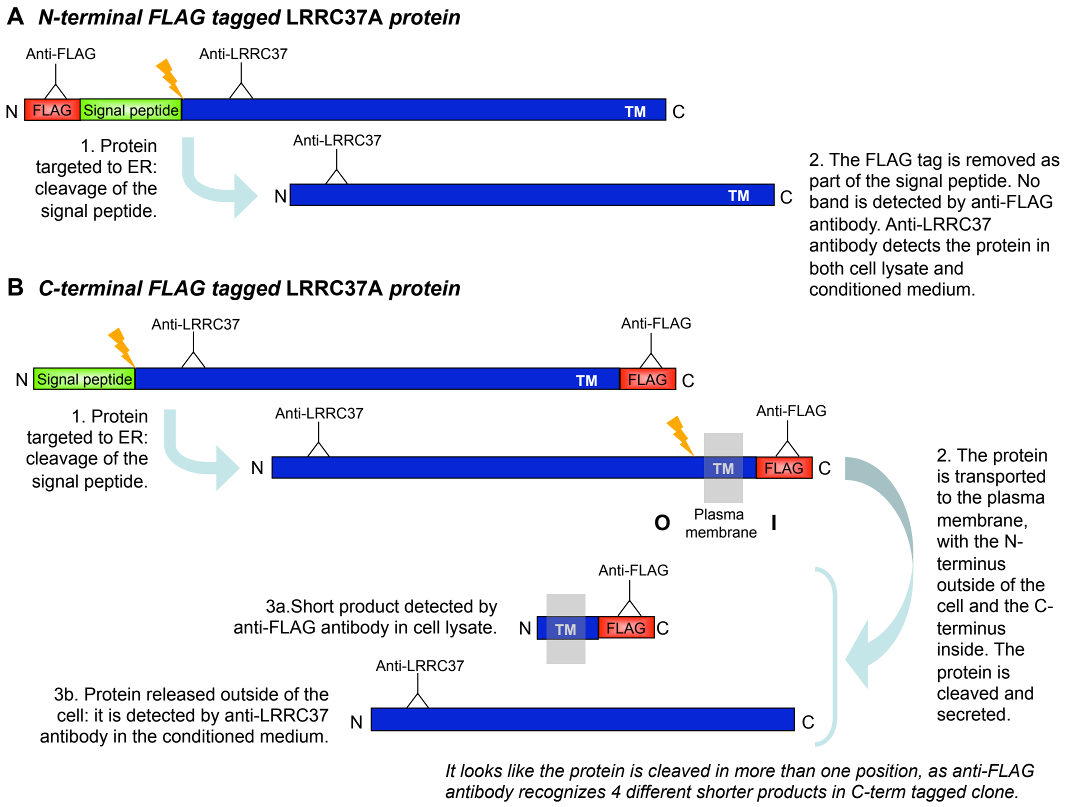

A *N-terminal FLAG tagged* **LRRC37A** *protein*
Anti-FLAG
Anti-LRRC37
N
FLAG
Signal peptide
TM
C
1. Protein targeted to ER: cleavage of the signal peptide.
Anti-LRRC37
N
TM
C
2. The FLAG tag is removed as part of the signal peptide. No band is detected by anti-FLAG antibody. Anti-LRRC37 antibody detects the protein in both cell lysate and conditioned medium.
B *C-terminal FLAG tagged* **LRRC37A** *protein*
Anti-LRRC37
Anti-FLAG
N
Signal peptide
TM
FLAG
C
1. Protein targeted to ER: cleavage of the signal peptide.
Anti-LRRC37
Anti-FLAG
N
TM
FLAG
C
O
Plasma membrane
I
2. The protein is transported to the plasma membrane, with the N-terminus outside of the cell and the C-terminus inside. The protein is cleaved and secreted.
Anti-FLAG
3a.Short product detected by anti-FLAG antibody in cell lysate.
N
TM
FLAG
C
Anti-LRRC37
3b. Protein released outside of the cell: it is detected by anti-LRRC37 antibody in the conditioned medium.
N
C
It looks like the protein is cleaved in more than one position, as anti-FLAG antibody recognizes 4 different shorter products in C-term tagged clone.

Figure S13. Outline of a possible explanation of immunofluorescence and Western blot results.

5.1. Time-course experiment

To further investigate the formation of the filopodia-like protrusions, we performed a time-course experiment, transiently transfecting HeLa cells with the C-terminal FLAG-tagged construct and analyzing the cells by immunofluorescence at different times after the beginning of transfection. The transfection was stopped after 3, 6, 12, 24, and 48 hours. Time zero (0 h) represented the nontransfected cells. HeLa cells were incubated with the anti-FLAG antibody. In Figure 9 are reported pictures depicting the most frequent aspect shown by transfected HeLa cells, at that certain interval of time.

HeLa cells incubated for 3 hours showed a Golgi staining. After 6 hours some transfected cells presented, besides the Golgi staining, a weak initial staining of vesicles and plasma membrane. After 12 and 24 hours the transfected cells showed a strong Golgi, vesicles, and plasma membrane staining. Transfected cells deformed their plasma membrane and made filopodia-like protrusions. After 24 hours many plasma membrane remnants strongly labeled by the anti-FLAG antibody began to appear. Most of these remnants were anucleated. After 48 hours the sample showed mostly membrane remnants, most of them anucleated.

6. GST-fusion LRRC37A full protein and protein fragments

We cloned the full-length coding sequence and the coding sequence of LRRC37A protein portions (exon 1, exons 1–7, and exons 2–7) fused with the Glutathione S-Transferase (GST) at the N terminus in pGEX-4T-1 vector (GE Healthcare) (primers listed in Table S7). We expressed the recombinant constructs in BL21 bacterial cells (Invitrogen) after IPTG-induction. We lysed the bacterial cells (BugBuster Protein Extraction Reagent, Novagena) and analyzed the solubilized and unsolubilized fractions through SDS-PAGE and Coomassie blue staining.

The molecular weight of the GST protein is 26 kDa. The predicted molecular weight of LRRC37A fragments and whole protein (second column), and the predicted molecular weight of the corresponding fusion proteins with GST (third column), are:

– exon 1	95 kDa	121 kDa
– exons 1–7	112 kDa	138 kDa
– exons 2–7	17 kDa	43 kDa
– whole protein	188 kDa	214 kDa

The GST-fusion LRRC37A full-length protein was not detected in either the detergent-soluble fraction or in the insoluble one, likely because the resulting fusion protein was too large and toxic for the bacterial cells (Figure S14). The GST-exon 1 fusion protein was detected in the soluble fraction (Figure S14). Conversely, both GST-exons 1–7 and GST-exons 2–7 (belonging to the LRR region) proteins were detected in the insoluble fraction (Figure S14). The GST-exon 1 and GST-exons 1–7 fusion proteins showed a low yield, likely for the same reason as in the case of the full-length protein. However, the fusion protein made up of the GST and exons 2–7 showed a high yield in the insoluble fraction, and its position in the gel is consistent with the calculated molecular weight. The GST-exon 1 and GST-exons 1–7 fusion proteins present a position in the gel much higher than the one expected according to their molecular weight. This running delay might be due to characteristics of the protein fragment encoded by exon 1.

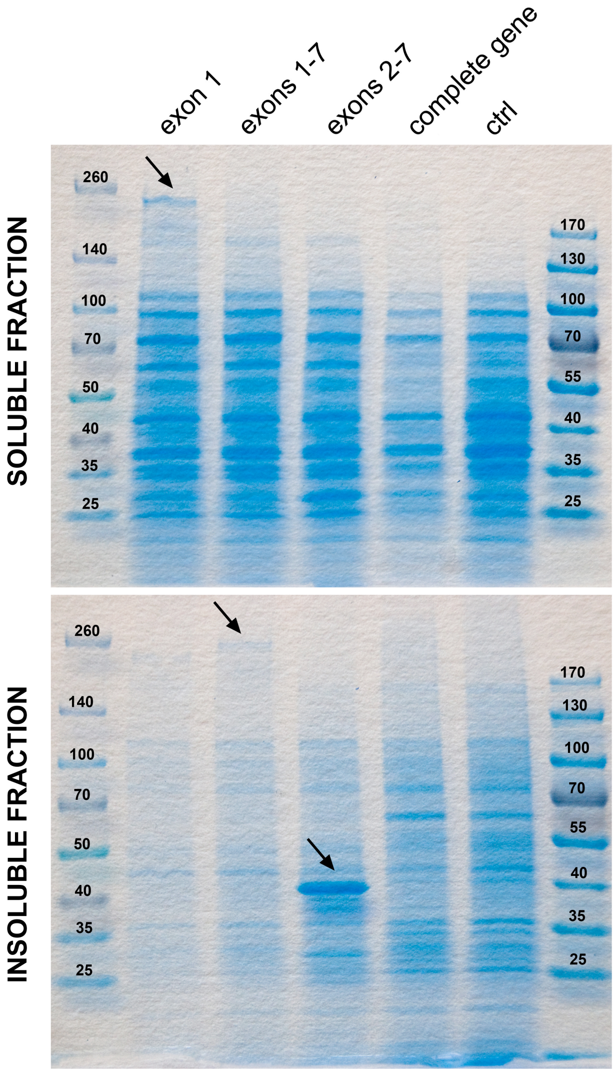


Figure S14. Coomassie stained SDS-PAGE gels of soluble and insoluble fractions of BL21 cell lysates. Arrows point the fusion products.

7. *LRRC37* family and pancreatic cancer

We searched the International Cancer Genome Consortium database (http://dcc.icgc.org) for mutation and expression data in cancers concerning the *LRRC37* family. *LRRC37A*, *LRRC37A2*, *LRRC37A3*, and *LRRC37A4* showed a copy number alteration in 14/67 (21%), 11/67 (16%), 10/67 (15%), and 4/67 (6%), respectively, of donors with pancreatic cancer (QCMG, AU) and a higher expression level in pancreatic endocrine tumors. The copy number alteration was either *LRRC37* specific or involved segments including *LRRC37* and one adjacent gene.

References

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**(10): 1061-1067.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol* **215**(3): 403-410.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* **2**: 28-36.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**(2): 573-580.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**(7369): 343-348.

De Clercq N, Hemschoote K, Devos A, Peeters B, Heyns W, Rombauts W. 1992. The 4.4-kilodalton proline-rich polypeptides of the rat ventral prostate are the proteolytic products of a 637-kilodalton protein displaying highly repetitive sequences and encoded in a single exon. *The Journal of biological chemistry* **267**(14): 9884-9894.

Everts-van der Wind A, Larkin DM, Green CA, Elliott JS, Olmstead CA, Chiu R, Schein JE, Marra MA, Womack JE, Lewin HA. 2005. A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**(51): 18526-18531.

Hemschoote K, Peeters B, Dirckx L, Claessens F, De Clercq N, Heyns W, Winderickx J, Bannwarth W, Rombauts W. 1988. A single 12.5-kilobase androgen-regulated mRNA encoding multiple proline-rich polypeptides in the ventral prostate of the rat. *The Journal of biological chemistry* **263**(35): 19159-19165.

Heyns W, Bossyns D, Peeters B, Rombauts W. 1982. Study of a proline-rich polypeptide bound to the prostatic binding protein of rat ventral prostate. *The Journal of biological chemistry* **257**(13): 7407-7413.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**(11): 1361-1368.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**(4): 402-408.

Parsons JD. 1995. Miropeats: graphical DNA sequence comparisons. *Computer applications in the biosciences : CABIOS* **11**(6): 615-619.

Peeters B, Heyns W, Bossyns D, Rombauts W. 1983. Proline-rich polypeptides bound to rat prostatic binding protein. The primary structure of the two main components, proline-rich polypeptides IV and V. *The Journal of biological chemistry* **258**(23): 14206-14211.

Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **95**(11): 5857-5864.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**(7): 4465-4470.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**(22): 4673-4680.

Wernersson R, Pedersen AG. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res* **31**(13): 3537-3539.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**(5): 555-556.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**(6): 908-917.