Genomics **73,** 50–55 (2001)
doi:10.1006/geno.2000.6504, available online at http://www.idealibrary.com on IDEAL®

# Complex Arrangement of Genes within a 220-kb Region of Double-Duplicated DNA on Human 2q37.1

Andreas Rump,*,[1] Grit Kasper,* Chris Hayes,† Gaiping Wen,* Heike Starke,‡ Thomas Liehr,‡,§ Rüdiger Lehmann,* Dorothee Lagemann,* and André Rosenthal*,§

*Department of Genome Analysis, Institute of Molecular Biotechnology, 07745 Jena, Germany; †Mammalian Genetics Unit, MRC, Harwell, Didcot, Oxon., OX1 3TS, United Kingdom; and ‡Institute of Human Genetics and Anthropology, §Friedrich-Schiller-University Jena, 07740 Jena, Germany

Received October 20, 2000; accepted January 8, 2001

**Gene duplication events are followed by divergence of initially identical gene copies, due to the subsequent accumulation of mutations. These mutations tend to be degenerative and may lead to either nonfunctionalization or subfunctionalization of the gene copies. Here we report the molecular characterization of a 220-kb genomic DNA fragment from human 2q37.1, in which a double duplication and a partial triplication event has taken place. As a result, this region contains four copies of alkaline phosphatase (P), four copies of the ECEL1 gene (X), two copies of a newly identified gene (N), and two copies of a cholinergic receptor subunit (R), in the order N-P-X-P-X-P-X-N-P-X-R-R. While three of the four ECEL1 copies, one copy of the phosphatase gene and one copy of the newly identified gene have lost their function, three phosphatase gene copies and the two receptor subunits are still functionally active and thus may provide an example for subfunctionalization of duplicated genes.**

## INTRODUCTION

Duplicate genes arise frequently in eukaryotic genomes, either via local events that generate tandem duplications or via large-scale events that duplicate chromosomal regions, entire chromosomes or even the whole genome. The duplication events are followed by divergence of the initially identical gene copies, due to the subsequent accumulation of mutations. With high probability, these mutations are degenerative and lead to loss of function of one gene copy (nonfunctionalization). However, it is also possible, even though with much lower probability, that both duplicates are preserved. This, however, requires that the coexistence of both duplicates is beneficial or, at least, neutral for the organism. Under the classical model for the evolution of gene duplicates, the mechanism by which members of a pair can permanently escape mutational decay is neofunctionalization, whereby one copy acquires a new beneficial function, with the other retaining the original function (Ohta, 1988; Nowak *et al.*, 1997). Since this mechanism is too unlikely to account for the observed large number of functional gene duplicates in eukaryotes, an alternative model was recently proposed (Force *et al.*, 1999; Lynch and Force, 2000). According to this model, duplicate gene preservation is achieved by the partitioning of ancestral functions, i.e., subfunctionalization of the duplicated genes. Lynch and Force (2000) define subfunctionalization “as the fixation of complementary loss-of-function alleles that result in the joint preservation of duplicate loci.” For example, a gene that is expressed originally in two tissues may diverge into two copies, each being expressed uniquely in one of the two tissues. A unique feature of the subfunctionalization model is that gene preservation is entirely a consequence of degenerative mutations, whereas beneficial mutations need not be invoked (Lynch and Force, 2000).

In the course of a large-scale sequencing project we have identified and analyzed in detail a 220-kb region on human 2q37.1, which provides an example for nonfunctionalization of genes and for maintenance of functional gene copies after a double duplication event.

## MATERIALS AND METHODS

*Fluorescence in situ hybridization (FISH).* FISH was carried out according to standard protocols. In brief, human metaphases were prepared from peripheral blood of a healthy male donor. Slides were pretreated with RNase and pepsin as described in Liehr *et al.* (1995) to diminish disturbing plasma background. One hundred twenty nanograms of the human clone CIT-HSP-D-3057K12, digoxigenated by nick translation, together with 4.0 $\mu$g of COT1 DNA was hybridized per slide. Hybridization was performed at 37°C for 3 nights in a humid chamber. After a postwashing series in 50% FA/2× SSC (3 times 5 min at 45°C) and 2× SSC (3 times 5 min at 37°C), the

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession No. AF307337.

[1] To whom correspondence should be addressed at MetaGen, Abteilung für Genomtechnologie, Ihnestrasse 63, 14195 Berlin, Germany. Telephone: (+)49-30-84131-683. Fax: (+)49-30-84131-674. E-mail: andreas.rump@metagen.de.

0888-7543/01 $35.00

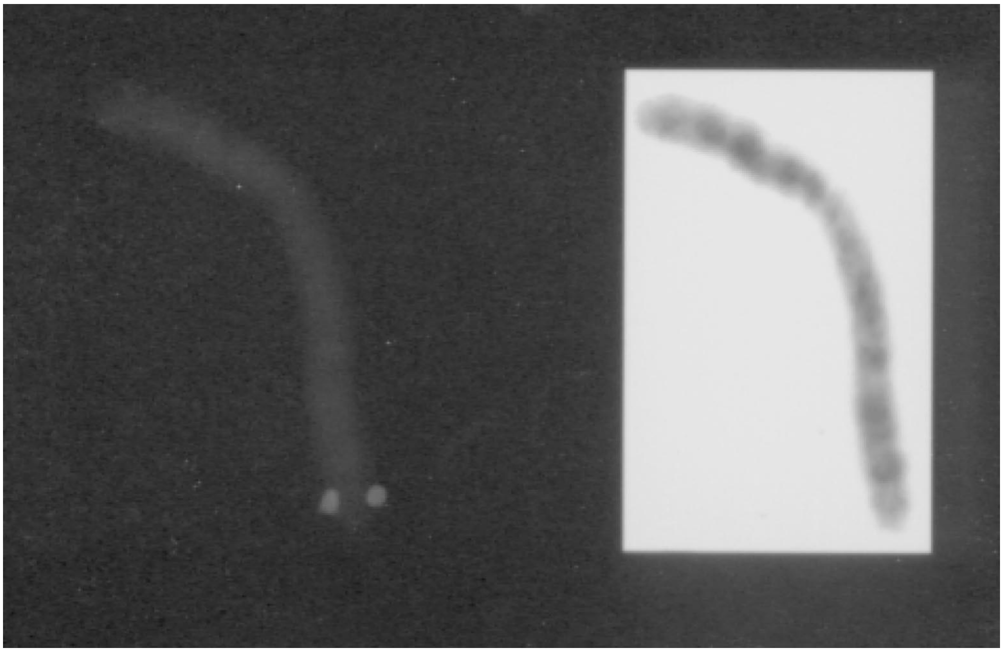

**FIG. 1.** FISH mapping of the sequenced region. The probe CIT-HSP-D-3057K12, labeled in red (**left**), has been mapped to 2q37.1. The blue DAPI counterstain (**left**) has been transformed into a GTG-like inverted DAPI-banding pattern (**right**) using the ISIS digital FISH imaging system (MetaSystems, Altlussheim, Germany). Images were taken using an XC77 CCD camera with on-chip integration (Sony).

detection of the probe was performed by antidigoxigenin–rhodamine, leading to red spots (Liehr *et al.*, 1995). Slides were evaluated on a Zeiss Axiophot fluorescence microscope. Ten metaphases were taken into account; to determine the precise subchromosomal localization, the DAPI-banding pattern was used.

*Genomic sequencing.* Nebulized fragments of the three clones P1-219H22, CIT-HSP-D-3057K12, and CIT-HSP-D-2219E1 were subcloned separately into M13mp18 vector (Yanisch-Perron *et al.*, 1985). At least 3500 plaques were selected from each clone library; M13 DNA was prepared and sequenced using dye-terminators, ThermoSequenase (Amersham), and universal M13-primer (MWG-Biotech). The gels were run on ABI 377 sequencers, and data were assembled and edited using the GAP4 program (Bonfield *et al.*, 1995). Genomic DNA sequence analysis was performed using the automated sequence annotation system RUMMAGE (Taudien *et al.*, 2000).

## RESULTS

### *Mapping of the Sequenced Region to 2q37.1*

The 220-kb sequenced region is localized on a BAC contig that consists of three partially overlapping clones. Using the central clone CIT-HSP-D-3057K12 for FISH led to the assignment of the sequenced region to human chromsome 2, band 2q37.1 (Fig. 1).

### *Identification of Three Functional Alkaline Phosphatase Genes*

Shotgun sequencing and subsequent closure of the remaining gaps resulted in a 219,708-bp contigous genomic sequence (GenBank Accession No. AF307337) that was analyzed in detail. The combined action of several exon prediction programs revealed the existence of 12 genes or pseudogenes, 8 of which are located on the forward strand, while the remaining 4 reside on the reverse strand (Fig. 2). Database searches with the predicted exons identified 3 alkaline phosphatase genes: ALPP (also known as placental alkaline phosphatase (PLAP) or Regan isozyme, cDNA Accession No. M14170), ALPPL2 (also known as PLAP-like or germ-cell alkaline phosphatase, cDNA Accession No. X55958) and ALPI (also known as intestinal alkaline phosphatase (IAP), gene Accession No. J03930). The existence of three distinct forms of human alkaline phosphatase was first shown by Lehmann (1980). Later, the corresponding phosphatase genes were mapped to human chromosome 2q37.1 (Griffin *et al.*, 1987; Martin *et al.*, 1987; Raimondi *et al.*, 1988; Wu *et al.*, 1993). As outlined in Fig. 2, the three phosphatase genes are transcribed in the forward direction and consist of 11 exons each. Since the exons of each phosphatase gene make up an open reading frame (ORF) that matches the corresponding cDNA, all 3 phosphatase genes can be considered to be functioning.

Detailed inspection of the sequence data revealed the existence of a fourth alkaline phosphatase gene (ALPPP), which, however, is highly degenerated and consists of 160 bp of genomic DNA, which shows an 86% match to ALPP exons 1 and 2, an intron not being present. This pseudogene precedes the ECEL1P3 gene on the complementary strand (Fig. 2).

### *Identification of the ECEL1 Gene and Three ECEL1 Pseudogenes*

The automated sequence annotation revealed the presence of the ECEL1 gene (Valdenaire *et al.*, 2000; cDNA Accession No. Y16187). The gene is located in reverse direction between the alkaline phosphatase

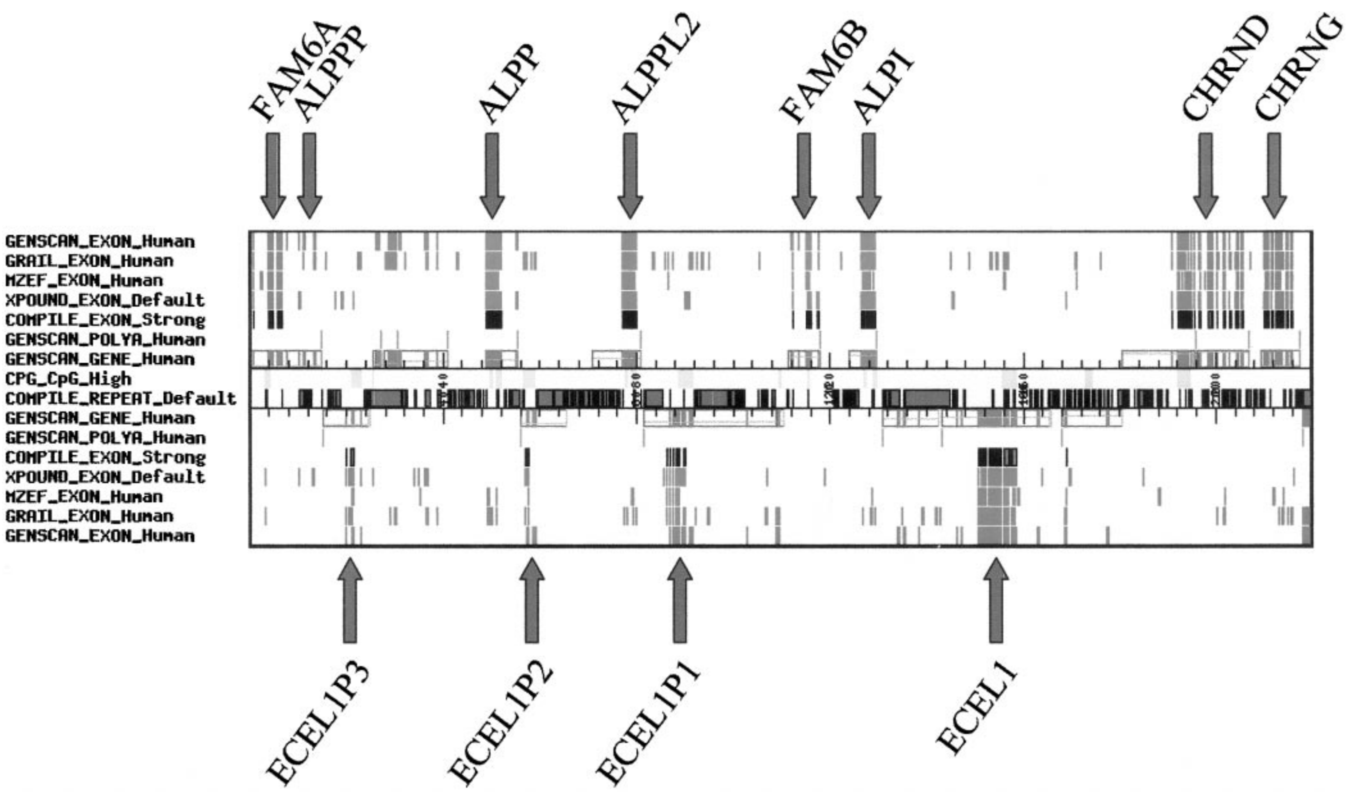

**FIG. 2.** Automated annotation of the 220-kb genomic fragment. The four phosphatase genes on the forward strand are flanked by ECEL1 pseudogenes on the reverse strand. The intestinal phosphatase gene (ALPI) is preceded by a pseudogene (FAM6B) that is derived from a newly identified gene (FAM6A), which shows similarity to the RNase type II family of proteins.

gene and the delta subunit gene of the nicotinic acetylcholine receptor (cDNA Accession No. X55019). In agreement with previous data (Valdenaire *et al.,* 2000), the ECEL1 gene consists of 18 exons and spans 8 kb of genomic sequence. What had not been known before, however, is the fact that additional ECEL1 derivatives are scattered within a region of about 100 kb in length (Fig. 2), resulting in the gene order ALPPP–ECEL1P3–ALPP–ECEL1P2–ALPP2L–ECEL1P1–ALPI–ECEL1. The ECEL1 derivatives show a decreasing match with the functional ECEL1 gene in the order ECEL1P1–ECEL1P2–ECEL1P3 (Fig. 3). The ECEL1P1 gene contains the first 11 of the original 18 exons, but the ORF is disrupted by several mutations, and some of the splice site consensi are lost. The ECEL1P2 gene is even more degenerated and has lost exons 3–18 completely. The most degenerated ECEL1 derivative is ECEL1P3, with only the 5′ portion of exon 2 retained. The nucleotide alignment of the pseudo-ECEL1 genes with the functional ECEL1 cDNA is available at http://genome.imb-jena.de/~arump/GENOMICS/align.html.

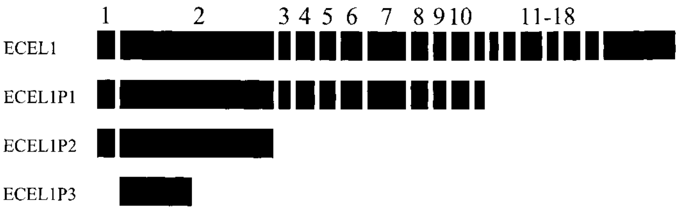

**FIG. 3.** Comparison of the functional ECEL1 gene with its three derivatives. The ECEL1 gene consists of 18 exons (black boxes). The three pseudogenes show increasing degeneration in the order ECEL1P1–ECEL1P2–ECEL1P3. In ECEL1P1 and ECEL1P2; the exon/intron structure is principally conserved, but most of the splice site consensi are lost.

### *Identification of a Previously Unknown Gene*

Automated sequence annotation detected a previously unknown gene (FAM6A) at the very 5′ end of the genomic sequence, as well as a degenerated copy of this gene (FAM6B) between ALPPL2 and ALPI (Fig. 2). The gene shows similarity to the RNase type II family of proteins. Since these proteins are very large (700–1000 aa) and the newly identified gene matches only the carboxy-terminal end of these proteins, it is possible that the new gene is only partially present on the sequenced region.

### *Identification of the Acetylcholine Receptor Subunit Genes Delta and Gamma*

Analysis of the 220-kb genomic fragment revealed the presence of the two acetylcholine receptor subunits delta and gamma (cDNA Accession Nos. X55019 and NM_005199, respectively). These two genes have previously been shown to be tightly linked to each other (Shibahara *et al.,* 1985; Heidmann *et al.,* 1986) and were mapped to 2q33–q34 (Lobos *et al.,* 1989; Beeson *et al.,* 1990) and to 2q35–q37 (Pasteris *et al.,* 1993). As our data show, the delta and gamma subunit genes are

**TABLE 1**

**Exon/Intron Boundaries of CHRND**

| Exon | Size (bp) | 3′ Splice acceptor | 5′ Splice acceptor | Intron size (bp) |
|---|---|---|---|---|
| 1 | 52[a] | | ..GTGTGTGgtaag.. | 261 |
| 2 | 146 | ..cccagGCAGCTG.. | ..CTCCCTGgtgag.. | 726 |
| 3 | 45 | ..tacagAAAGAAG.. | ..AGAGCACgtaag.. | 816 |
| 4 | 110 | ..tccagGGCTGGA.. | ..AGAACAAgttga.. | 129 |
| 5 | 156 | ..cccagCAATGAC.. | ..AGTTCAGgtgtg.. | 205 |
| 6 | 110 | ..cccagTTCCCTC.. | ..TTCACAGgtgct.. | 967 |
| 7 | 201 | ..ctcagAGAACGG.. | ..GCTGACAgtgag.. | 1212 |
| 8 | 112 | ..cccagGTGGTGA.. | ..TCGGCAAgtgag.. | 78 |
| 9 | 115 | ..cacagGTTCCTG.. | ..CAAGAAGgtgag.. | 2273 |
| 10 | 205 | ..cccagCTCTTCC.. | ..ACTGCACgtggg.. | 88 |
| 11 | 119 | ..tgtagGCCGGCC.. | ..CAATGAGgtaag.. | 787 |
| 12 | 183[b] | ..cacagGAGAAAG.. | | |

[a] Size of exon refers to the ATG start codon; the 5′ UTR is not included.

[b] Size of exon refers to the TAG stop codon; the 3′ UTR is not included.

closely linked to the phosphatase genes and thus reside on 2q37.1.

Alignment of the published cDNAs with the genomic sequence reveals the exon/intron structures of the delta and gamma receptor subunits (Tables 1 and 2, respectively). In accordance with previous data (Shibahara *et al.*, 1985), the protein coding sequence of the gamma subunit is divided by 11 introns into 12 exons. Here we show that the same is true for the delta subunit (Table 1). According to our data, some of the gamma subunit exons differ in length from those published in the database (Accession No. X01715). Whereas the published cDNA codes for 520 aa, our coding region that was derived from the genomic sequence codes for 514 aa. Alignment of the published gamma subunit cDNA with our genomic sequence led to several mismatches, and at some positions the published exon structure does not fit with the occurrence of splice site consensi in our genomic sequence.

Comparison of the exon and intron sizes from the two receptor subunits delta and gamma with each other shows that the exon lengths are highly conserved, while the intron sizes differ significantly. As a consequence, the coded proteins are similar in size (517 aa for the delta subunit versus 514 aa for the gamma subunit), while the sizes of the genes differ significantly (9 kb for the delta subunit versus 6 kb for the gamma subunit).

## DISCUSSION

### *Reconstruction of the Duplication Events in 2q37.1*

We assume that the currently observed gene organization, as shown in Fig. 2 and at the bottom of Fig. 4, is the outcome of events that once started with all genes being functionally active and being present in one copy only (Fig. 4, top). Then a tandem duplication event might have taken place, with subsequent accumulation of degenerative mutations. We assume that these mutations have led to subfunctionalization of the two alkaline phosphatase gene copies and to nonfunctionalization of one ECEL1 copy and one copy of the RNase II-like gene. After this, another duplication occurred, affecting only a subfunctional phosphatase and a nonfunctional ECEL1 copy. This event was followed by a third, partial duplication, leading to the current situation.

The assumption of a third duplication event seems to be in contrast to the findings of Knoll *et al.* (1987), who reported a double duplication event. Their conclusions, however, are based on partial sequencing of separated alkine phosphatase genes, which were subcloned into Lambda vectors. Thus the whole genomic organization was not known, nor was the existence of ECEL1 pseudogenes, distributed evenly among the phosphatase genes, known.

In our reconstruction of events, the acetylcholine receptor subunits delta and gamma have not been considered. However, taking the highly conserved exon structure and the similarity at the protein level into account, it seems very likely that these two subunits also arose by a duplication event, which might have occurred independently of the events proposed in Fig. 4.

### *The ECEL1 Gene Duplicates: An Example of Nonfunctionalization*

By now, many examples of nonfunctionalization of a duplicated gene are known. However, the case of ECEL1 is particularly interesting, because the first nonfunctional copy of this gene is involved in two subsequent duplication events, leading to a cascade of degenerated products.

**TABLE 2**

**Exon/Intron Boundaries of CHRNG**

| Exon | Size (bp) | 3′ Splice acceptor | 5′ Splice acceptor | Intron size (bp) |
|---|---|---|---|---|
| 1 | 55[a] | | ..TGCCTGGgtggg.. | 189 |
| 2 | 140 | ..tgcagGGGCCCA.. | ..CTCCCTGgtaag.. | 249 |
| 3 | 45 | ..cccagAACGAGC.. | ..AGAGATGgtaag.. | 176 |
| 4 | 110 | ..tgcagCAGTGGT.. | ..AGAACAAgtgag.. | 653 |
| 5 | 156 | ..tacagCGTGGAC.. | ..TCTTCCAgtgag.. | 893 |
| 6 | 98 | ..cccagGTCCCAG.. | ..TTCACAGgtaac.. | 361 |
| 7 | 201 | ..cgcagAGAATGG.. | ..GCCAAGGgtacc.. | 192 |
| 8 | 115 | ..gatagCTGGGGG.. | ..TCAGCAAgtaag.. | 194 |
| 9 | 106 | ..gacagGTACCTG.. | ..CCGAGGGgtccg.. | 676 |
| 10 | 214 | ..tctagGTGTTCC.. | ..AAGCTAGgtgag.. | 191 |
| 11 | 131 | ..atcagAGAAAGG.. | ..TGACAATgtaag.. | 640 |
| 12 | 174[b] | ..ctcagGGGAATG.. | | |

[a] Size of exon refers to the ATG start codon; the 5′ UTR is not included.

[b] Size of exon refers to the TAG stop codon; the 3′ UTR is not included.

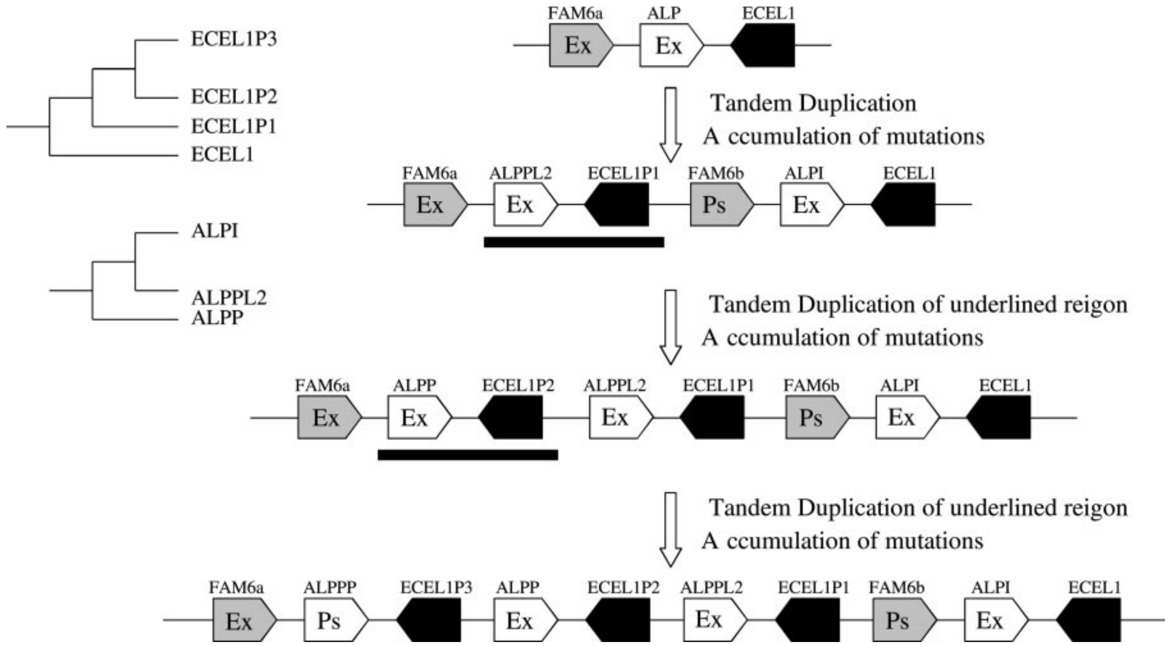

**FIG. 4.** Reconstruction of the duplication events in 2q37.1. The events started with all genes being functionally active (indicated by "Ex," for "expressed") and being present in one copy only. Due to subsequent mutations, some genes became nonfunctional (indicated by "Ps," for "pseudogene"). Gray, newly identified gene (with similarity to RNase type II); white, alkaline phosphatase gene family; black, ECEL1 gene family. The phylogenetic relationships between the ECEL1 derivatives and the ALP derivatives have been calculated with the DNA parsimony algorithm of the program PHYLIP 3.53C and are shown in unrooted phylogenetic trees.

### *The Alkaline Phosphatase Genes: An Example for Subfunctionalization*

Since the alkaline phosphatase genes on 2q37.1 are known to be expressed in different tissues (Lehmann, 1980), these genes may represent a nice example of subfunctionalization in the sense of Lynch and Force (2000). The tissue specifity can be explained by different alterations in the promoter regions of the three genes, due to degenerative mutations. This is consistent with the findings from alkaline phosphatase promoter studies (Millan, 1987; Kiledjian and Kadesch, 1990).

Although the different ALP genes do have different expression patterns, this observation alone does not prove that subfunctionalization has really occurred. To prove this, the corresponding homologous region of an anchestral organism with only one, ubiquitously expressed ALP gene must be sequenced. However, such an organism has not yet been identified. An alternative approach would be to rescue a mouse with knockout of all three ALP genes by introduction of a BAC containing only one ALP gene under the control of the promoter that drives ubiquitous expression.

## ACKNOWLEDGMENTS

This work was supported by a grant from the German "Bundesministerium für Bildung und Forschung" (BMBF, BEO-0311108), the Herbert-Quandt-Stiftung der VARTA AG, and the Madeleine-Schickedanz-Kinderkrebs Stiftung. C.H. is in receipt of a MRC Research Fellowship.

## REFERENCES

Beeson, D., Jeremiah, S., West, L. F., Povey, S., and Newsom-Davis, J. (1990). Assignment of the human nicotinic acetylcholine receptor genes: the alpha and delta subunit genes to chromosome 2 and the beta subunit gene to chromosome 17. *Ann. Hum. Genet.* **54:** 199–208.

Bonfield, J. K., Smith, K. F., and Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res.* **23:** 4992–5009.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Griffin, C. A., Smith, M., Henthorn, P. S., Harris H., Weiss, M. J., Raducha, M., and Emanuel, B. S. (1987). Human placental and intestinal alkaline phosphatase genes map to 2q34–q37. *Am. J. Hum. Genet.* **41:** 1025–1034.

Heidmann, O., Buonanno, A., Geoffroy, B., Robert, B., Guenet, J.-L., Merlie, J. P., and Changeux, J.-P. (1986). Chromosomal localization of muscle nicotinic acetylcholine receptor genes in the mouse. *Science* **234:** 866–868.

Kiledjian, M., and Kadesch, T. (1990). Analysis of the human liver/bone/kidney alkaline phosphatase promoter *in vivo* and *in vitro. Nucleic Acids Res.* **18:** 957–961.

Knoll, B. J., Rothblum, K. N., and Longley, M. (1987). Two gene duplication events in the evolution of the human heat-stable alkaline phosphatases. *Gene* **60:** 267–276.

Lehmann, F. G. (1980). Human alkaline phosphatases. Evidence of three isoenzymes (placental, intestinal and liver–bone–kidney–type) by lectin-binding affinity and immunological specificity. *Biochim. Biophys. Acta* **6:** 41–59.

Liehr, T., Thoma, K., Gehring, C., Ekici, A, Bathke, K. D., Grehl, H., and Rautenstrauss, B. (1995). Direct preparation of uncultured EDTA-treated or heparinized blood for interphase FISH analysis. *Appl. Cytogenet.* **21:** 185–188.

Lobos, E. A., Rudnick, C. H., Watson, M. S., and Isenberg, K. E. (1989).Linkage disequilibrium study of RFLPs detected at the human muscle nicotinic acetylcholine receptor subunit genes. *Am. J. Hum. Genet.* **44:** 522–533.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459–473.

Martin, D., Tucker, D. F., Gorman, P., Sheer, D., Spurr, N. K., and Trowsdale, J. (1987). The human placental alkaline phosphatase gene and related sequences map to chromosome 2 band q37. *Ann. Hum. Genet.* **51:** 145–152.

Millan, J. L. (1987). Promoter structure of the human intestinal alkaline phosphatase gene. *Nucleic Acids Res.* **15:** 10599.

Nowak, M. A., Boerlijst, M. C., Cooke, J., and Smith, J. M. (1997). Evolution of genetic redundancy. *Nature* **388:** 167–171.

Ohta, T. (1987). Simulating evolution by gene duplication. *Genetics* **115:** 207–213.

Pasteris, N. G., Trask, B. J., Sheldon, S., and Gorski, J. L. (1993). Discordant phenotype of two overlapping deletions involving the PAX3 gene in chromosome 2q35. *Hum. Mol. Genet.* **2:** 953–959.

Raimondi, E., Talarico, D., Moro, L., Rutter, W. J., Della Valle, G., and De Carli, L. (1988). Regional mapping of the human placental alkaline phosphatase gene (ALPP) to 2q37 by in situ hybridization. *Cytogenet. Cell. Genet.* **47:** 98–99.

Shibahara, S., Kubo, T., Perski, H. J., Takahashi, H., Noda, M., and Numa, S. (1985). Cloning and sequence analysis of human genomic DNA encoding gamma subunit precursor of muscle acetylcholine receptor. *Eur. J. Biochem.* **146:** 15–22.

Taudien, S., Rump, A., Platzer, M., Drescher, B., Schattevoy, R., Gloeckner, G., Dette, M., Baumgart, C., Weber, J., Menzel, U., and Rosenthal, Rummage A. (2000). A high-throughput sequence annotation system. *Trends Genet.* **16:** 519–527.

Valdenaire, O., Rohrbacher, E., Langeveld, A., Schweizer, A., and Meijers, C. (2000). Organization and chromosomal localization of the human ECEL1 (XCE) gene encoding a zinc metallopeptidase involved in the nervous control of respiration. *Biochemistry* **346:** 611–706.

Wu, B. L., Milunsky, A., Wyandt, H., Hoth, C., Baldwin, C., and Skare, J. (1993). In situ hybridization applied to Waardenburg syndrome. *Cytogenet. Cell. Genet.* **63:** 29–32.

Yanisch-Perron, C., Vieira, J., and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: Nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33:** 103–119.