More than 1,000 putative new human signalling proteins revealed by EST data mining

Jörg Schultz[1,2], Tobias Doerks[1,2], Chris P. Ponting[3], Richard R. Copley[1] & Peer Bork[1,2]

Cloning procedures aided by homology searches of EST databases have accelerated the pace of discovery of new genes[1], but EST database searching remains an involved and onerous task. More than 1.6 million human EST sequences have been deposited in public databases, making it difficult to identify ESTs that represent new genes. Compounding the problems of scale are difficulties in detection associated with a high sequencing error rate and low sequence similarity between distant homologues. We have developed a new method, coupling BLAST-based[2] searches with a domain identification protocol[3,4], that filters candidate homologues. Application of this method in a large-scale analysis of 100 signalling domain families has led to the identification of ESTs representing more than 1,000 novel human signalling genes. The 4,206 publicly available ESTs representing these genes are a valuable resource for rapid cloning of novel human signalling proteins. For example, we were able to identify ESTs of at least 106 new small GTPases, of which 6 are likely to belong to new subfamilies. In some cases, further analyses of genomic DNA led to the discovery of previously unidentified full-length protein sequences. This is exemplified by the *in silico* cloning (prediction of a gene product sequence using only genomic and EST sequence data) of a new type of GTPase with two catalytic domains.

Detecting novel members of protein families using EST databases is becoming increasingly difficult due to the rapid growth of existing sequence information. For example, if one searches for new small GTPases that are distinct from the 100 or so already known, each of the 100 must first be used to search the 1.6 million human ESTs currently known. Each of these database searches is likely to reveal an average of 500 human small GTPase ESTs that demonstrate significant sequence similarities to the query. The final task is to compare each of the ESTs found in these searches with known protein sequences to identify those human small GTPase sequences that are distinct from the initial 100 sequences. This procedure is further complicated by sequencing errors and sequence divergence of homologues. Several procedures have been developed to assist in this process. UniGene[5], for example, is a system for partitioning ESTs into single gene clusters, thereby lowering the overall size and redundancy of EST data. Programs such as ProtEST (http://circinus.ebi.ac.uk:8081/protest/) or FAST-PAN (ref. 6) assist in the analysis of ESTs corresponding to single proteins. No method, however, has yet been developed that estimates the number of new genes of homologous families that are represented in EST databases.

To determine the number of novel signalling proteins currently detectable by EST-based approaches, we chose 50 domain families that occur in extracellular proteins and 50 that are characteristic of proteins involved in intracellular signalling cascades, such as the catalytic domains of Ras-like small GTPases (Table 1). In terms of their frequency of occurrence in human proteins, these

Table 1 • Identification of new members for selected domain families

Domain	Sequenced proteins	No. ESTs	No. new ESTs	Clusters of new ESTs	No. of similar proteins
intracellular					
ARF	15	765	51	23	14
RAB	38	1,108	238	78	37
RAN	2	265	4	4	2
RAS	26	388	22	16	12
RHO	21	840	33	23	16
SAR	1	89	39	14	5
unclassified small GTPase	30	320	52	29	20
RasGEF	24	117	30	20	11
RhoGAP	48	466	196	66	26
PH	192	1,122	284	129	69
extracellular					
C1Q	13	346	55	33	14
CCP	73	730	82	52	33
COLFI	24	1,296	19	15	11
CUB	47	241	41	26	15
INB	15	136	3	2	2
KU	21	374	7	6	4
SCY	55	552	14	13	12
TNF	23	130	1	1	1
Tryp_SPc	127	2,126	96	63	45
VWA	85	810	90	48	23

The full results are available (http://smart.EMBL-Heidelberg.de/EST). For descriptions of domain families, see SMART (http://smart.EMBL-Heidelberg.de). The number of human proteins possessing a particular domain found in a non-redundant database is shown in the second column.

[1]EMBL, Heidelberg, Germany. [2]Max-Delbrück-Center, Berlin-Buch, Germany. [3]MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford, UK. Correspondence should be addressed to P.B. (e-mail: Bork@EMBL-Heidelberg.de).

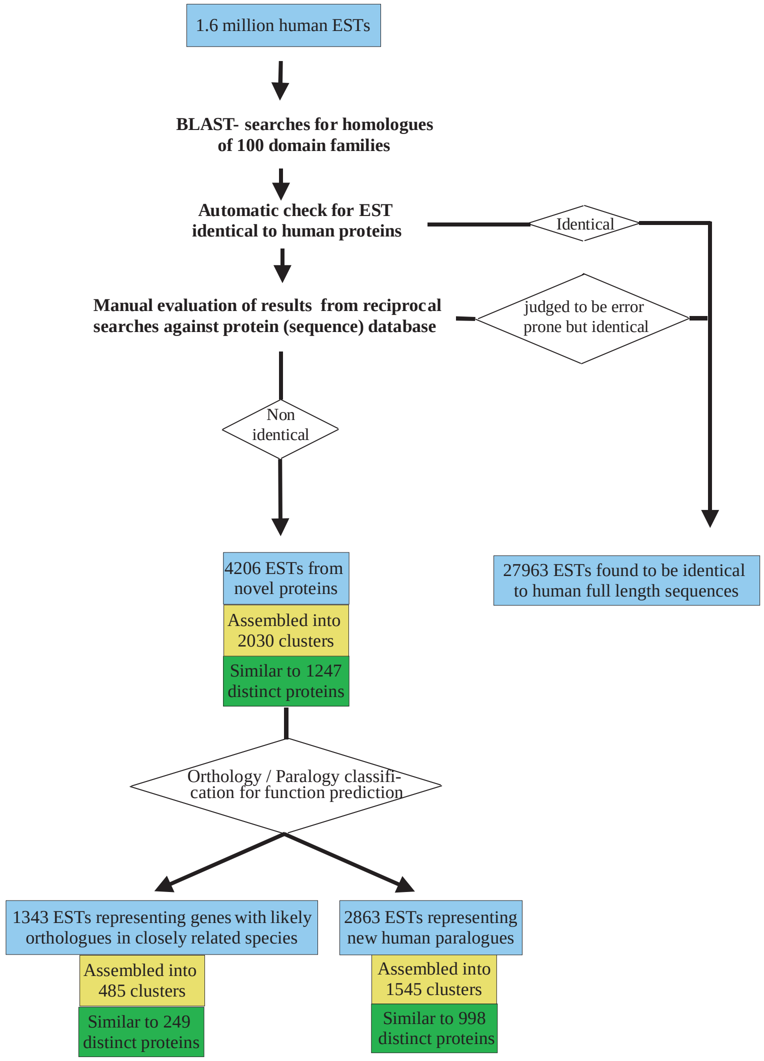
1.6 million human ESTs
BLAST- searches for homologues of 100 domain families
Automatic check for EST identical to human proteins
Identical
Manual evaluation of results from reciprocal searches against protein (sequence) database
judged to be error prone but identical
Non identical
4206 ESTs from novel proteins
Assembled into 2030 clusters
Similar to 1247 distinct proteins
27963 ESTs found to be identical to human full length sequences
Orthology / Paralogy classification for function prediction
1343 ESTs representing genes with likely orthologues in closely related species
Assembled into 485 clusters
Similar to 249 distinct proteins
2863 ESTs representing new human paralogues
Assembled into 1545 clusters
Similar to 998 distinct proteins

Fig. 1 Flow chart summarizing search protocol and results. Numbers of ESTs are shown in blue; numbers of clusters, in yellow boxes; and number of distinct full-length protein sequences most similar to the clusters, in green boxes. For the threshold criteria of each decision stage (diamonds), see text. (The results can be downloaded from ftp://booby.EMBL-Heidelberg.de/pub/data/EST.) Definition of terms: reciprocal search, a database search using a query sequence that was identified in a previous search; orthologues, genes that arose as a direct consequence of speciation; paralogues, genes that arose due to intragenomic gene duplication and not speciations; E value, the number of sequences in a given database identified in a search with a score greater than, or equal to, X expected by chance.

100 families are representative of the SMART database[3,4]. This database contains more than 400 genetically mobile and mainly signalling domains; thus our findings may reasonably be extrapolated to all 400 domain families.

The semi-automatic method we have developed couples BLAST searches with a domain identification protocol. Using the SMART system (http://smart.EMBL-Heidelberg.de), amino acid sequences of all members of the 100 domain families that were detected in *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* were collated. The sequences of the domains were used as queries for TBLASTN searches[2] against a database containing 1.6 million human ESTs derived from the EMBL EST database. Conceptual translations of all ESTs that matched a human protein with at least 95% identity using a sliding window of 30 amino acids were considered as identical to the protein. This low identity threshold was chosen to ensure that the rate of false positives (ESTs that correspond to a sequenced gene, but are not detected as such) is low, although this is at the cost of missing a few closely related genes.

All other hits with *E* values less than 1.0 were recorded for further analysis (Fig. 1). This relatively lax threshold allowed detection of homologous sequences that were divergent or prone to errors. As non-homologous sequences were also recorded, however, a protocol of further validation was developed. ESTs with at least 97% identity over 100 bp were grouped into clusters and aligned. We derived a consensus sequence for each alignment, which we subsequently used to perform a reciprocal search[7] against a non-redundant protein database. Here, at least one protein containing the domain, with an *E* value less than 1.0, was required. These hits were manually analysed to ensure that the ESTs indeed represent the domain family in question and to distinguish between novel and previously known, but error-prone, sequences (Fig. 1).

This procedure resulted in 4,206 ESTs that were sufficiently dissimilar to human full-length sequences to indicate that their differences were unlikely to be due to sequencing errors. We assembled these into 2,030 separate clusters. As two or more of these clusters might correspond to different regions of the same protein, the closest known full-length protein sequence to each cluster was recorded, resulting in 1,197 distinct proteins. Whereas the number of clusters (2,030) is likely to be an overestimate of the number of novel genes represented by the ESTs, the number of most similar proteins (1,197) provides a lower bound to this number (Table 1). The clusters were further sub-divided according to whether a possible orthologue in a different species is already known. If the cluster was highly similar to non-human proteins and less similar to all known human homologues, it was classified as representing a human orthologue of a known gene. For example, we found ESTs representing likely human orthologues of 52 mouse proteins. Of the 2,030 EST clusters, 1,545 did not have a possible orthologue, thus encoding novel, uncharacterized human proteins for which precise functions remain unknown. Numbers of predicted novel proteins varied between different domain types. For five domain types (ANATO, MATH, SAND, SEA and SO), no ESTs of novel genes were found. By contrast, 152 likely new serine/threonine kinases were predicted.

Detailed results can be accessed through a World-Wide Web interface (http://smart.EMBL-Heidelberg.de/EST) that allows the selection of EST clusters according to domain families and classification into clusters that have or do not have possible orthologues. Each selected EST is hyperlinked to sequence databases and the BLAST-search output file that was used for its classification. Additionally, the function of the closest related protein is shown for each cluster.

Despite the expected availability of 90% of the human genome in spring 2000 (ref. 8), this procedure is likely to *a*

identifier (PID)	species	description	E- value
1572819	CAEEL	similar to the RAS gene family	7.9e2 1
1519671	CAEEL	similarity to ATP/GTPbi nding site	3.0e2 1
731284	YEAST	HYPOTHETICAL PROTEIN	6.2e1 7
2995366	SCHPO	conserved hypothetical protein	4.1e9
3341689	ARATH	unknown protein	0.94
5381420	ARATH	racli ke protein ARAC9	0.94
4959463	MAIZE	RACC small GTP binding protein	0.94
1350830	RAT	RABPHILIN3A	0.98

b

```
1572819         1 MSDDETLADVRIVLIGD                 GCGKTSLVMSLLEDEWVDA                VPRRLDRVLIPAD
                          DVRI+L+G+    Intron 1       GKTSL++SL+ +E+ +     Intron 2    VP R + + IPAD
1869775     25323      MRRDVRILLLGE  [25256:25141]  QVGKTSLILSLVGEEFPEE  [25081:25006] VPPRAEEITIPAD

1572819        51 VTPENVTTSIVDLS                KEEDENWIVSEIRQANVICVVYSVTDESTVDG               IQT
                  VTPE V T IVD S    Intron 3         +   S + QANV+CVVY V++E+T++    Intron 4    I+T
1869775     24966 VTPEKVPTHIVDYS  [24923:23982] SSPVTHTS SLFPQANVVCVVYDVSEEATIEK [23886:23371] IRT

1572819       101 KWLPLIRQSFGEYH                 TPVILVGNKSD-GTANNTDKILPIMEANTEVETCVE
                  KW+PL+     +      Intron 5      P+ILVGNKSD  + ++ + +LPIM    E+ETCVE   Intron 6
1869775     23361 KWIPLVNGGTTQGP  [23317:23245]  VPIILVGNKSDLRSGSSMEAVLPIMSQFPEIETCVE [23135:23038]

1572819       151 CSARTMKNVSEIFYYAQKAVIYPTRPLYDADTKQ                 LTDRARKALIRVFKICDRDNDGYLSD
                  CSA+ ++N+SE+FYYAQKAV++PT PLYD + KQ    Intron 7     L     +AL R+F++ D+D D  LSD
1869775     23040 CSAKNLRNISELFYYAQKAVLHPTAPLYDPEAKQ  [22935:22819]  LRPACAQALTRIFRLSDQDLDQALSD

1572819       211 TELNDFQ                 KLCFGIPLTSTALEDVKRAVSDGCPDGVANDSLMLA
                   ELN FQ    Intron 8       K CFG PL   ALEDVKV      GV  D L L     Intron 9
1869775     22740 EELNAFQ  [22719:22602]  KSCFGHPLAPQALEDVKTVVCRNVAGGVREDRLTLD  [22492:22411]

1572819       254 FLYLHLLFIERGRHETTWAVLRKFGYETSLKLSEDYLYP                 ITIPVGCSTELSPEGVQFVSALF
                  FL+L+ LFI+RGRHETTW +LR+FGY  +L+L+ DYL P    Intron 10    I +P GCSTEL+  G QFV  +F
1869775     22411 FLFLNTLFIQRGRHETTWTILRRFGYSDALELTADYLSP  [22289:21797]  IHVPPGCSTELNHLGYQFVQRVF

1572819       318 EKYDE                 DKDGCLSPSELQNLFSVCPVPVITKDNILALETNQRGWLTYNGYMAYW
                  EK+D+    Intron 11    D+DG LSP ELQ+LFSV P      +    + T   G L  +GY+  W   Intron 12
1869775     21726 EKHDQ  [21711:21634]  DRDGALSPVELQSLFSVFPAAPWGPELPRTVRTEA GRLPLHGYLCQW [21490:21410]

1572819       371 MTTLINLTQTFEQLAYLGFPVGRSGPGRAGNTLDSIR               TRERKKDLENHGTDRKVFQCLVV
                  + T +++   +  L YLG+P              +I    Intron 13   TRE++ D E   T R V+ C VV
1869775     21412 LVTYLDVRSCLGHLGYLGYPTLCEQ-----DQAHAIT [21311:21240] TREKRLDQEKGQTQRSVLLCKVV

1572819       433 GAKDAGKTVFMQSLAGRGMA                 DVAQIGRRHSP-FVINRVRVKEESKYLL
                  GA+  GK+ F+Q+  GRG+     Intron 14    D     R   P + I+ V+V  + KYL+    Intron 15
1869775     21168 GARGVGKSAFLQAFLGRGLG  [21108:20999]  DT----REQPPGYAIDTVQVNGQEKYLI       [20926:20791]

1572819       481 LREVDVLSPQDAL--GSGETSADVVAFLYDISNPDSFAFCATVYQ                KYFYRTKTPCVMIATKVEREEVDQRW
                  L EV      D L   S + + DV  +++D S+P SFA CA+VY+   Intron 16     ++   +TPC+++++K +  E
1869775     21004 LCEVGT --- DGLLATSLDATCDVACLMFDGSDPKSFAHCASVYK [20667:20566] HHYMDGQTPCLFVSSKADLPEGVAVS

1572819       549 EVPPEEFCRQFELPKPIKFSTGNIGQSSSPIFEQLAMMAVYPHLRRVFYLNDSNLLSKITFGAAIVALAGFLVLKNL
                     P EFCR+  LP P+ FS     + S+ IF QLA MA +P +     L  S++    +      A    L +  +
1869775     20487 GPSPAEFCRKHRLPAPVPFSCAGPAEPSTTIFTQLATMAAFPWVPSSAALGTSSVCHPSSGQRWCQAWNWGLAVSVI
```

Fig. 2 *In silico* cloning of a novel small GTPase. ***a***, Result of a reciprocal database search with a candidate EST (GenBank nucleotide identifier [NID]: 564419) that encodes a putative small GTPase. It identifies the closest full-length homologue (K08F11.5, a hypothetical *C. elegans* protein), but also shows similarity with other GTPases. CAEEL, *C. elegans*; YEAST, *S. cerevisiae*; SCHPO, *Schizosaccharomyces pombe*; ARATH, *A. thaliana*; MAIZE, *Zea mays*; RAT, *Rattus norvegicus*. ***b***, *In silico* cloning of a new human small GTPase using EST 564419 (validated in *a*) and its closest hit. The EST is identical to a region of chromosome 16 (the NID of the respective contig is 1869775), which is not annotated as coding. Using *C. elegans* hypothetical protein sequence K08F11.5 (GenBank protein ID: 1572819), identified in (*a*) as being the sequence most similar to EST 564419, the full-length sequence and gene structure of this protein can be predicted. An alignment is shown of the *C. elegans* protein with the predicted human protein. The positions of 16 introns are indicated as predicted by GeneWise[11], which uses the alignment information in combination with statistical parameters of intron distribution. The region identical to the EST is shown in bold. Both proteins contain an amino-terminal small GTPase domain (blue), followed by two EF hands, typical calcium-binding domains (green). The carboxy termini of the proteins also show significant similarity to small GTPases (a second PSI-BLAST iteration revealed RAD_HUMAN (PID:1710005) with an *E* value of 0.0004; shown in red). This domain structure is found in likely orthologues in *S. pombe* (PID:2995366) and *S. cerevisiae* (YAE8_YEAST, PID:731284, a protein involved in a secretory pathway[12]).

remain useful due to long-standing problems in accurate gene annotation. As only about 44%, on average, of coding positions are represented in ESTs (as measured in a set of 9,400 distinct human mRNAs; ref. 9), *in silico* full-length cloning using only ESTs (ref. 10) will be unsuccessful in most cases. If an EST is identical to a region of human genomic DNA, however, further analysis can result in the identification of the complete gene structure. An example of this is the identification of the full-length sequence of a small GTPase (Fig. 2). The gene structure was predicted by first identifying the genomic region identical in part to the EST, and then aligning its conceptual translation against the protein sequence of its closest full-length homologue. Here (Fig. 2*b*) the conceptual translation of a region of human genomic sequence has been aligned with its presumed *C. elegans* orthologue. The gene of this new GTPase was not reported in the human genomic sequence.

Resources such as UniGene[5] address problems associated with the high redundancy of information in EST databases. To quantify the loss of information in the UniGene clustering procedure, we compared the 4,206 ESTs representing new human signalling proteins with the UniGene set (release August 1999). An EST was classified as present in UniGene if a cluster matched the EST with an identity of at least 97% in a sliding window of 100 bp. More than 20% of these ESTs are absent from this release. Thus, it appears advisable to mine the original EST data rather than the UniGene set to detect novel genes.

The 100 domain families used here are representative of all 400 domain families curated in SMART, in terms of numbers of human proteins in which they occur. Although a linear extrapolation is influenced by abundance of sequenced ESTs for each domain family as well as by the degree of EST mining done previously, we estimate that roughly 4,000 novel signalling proteins may be indentified in current public EST databases using the data inherent in SMART. The number of human small GTPases in current protein databases could be almost doubled using this procedure (Table 1). This expansion illustrates both the complexity of signalling and a major challenge facing the modern biologist: deciphering the *in vivo* role of a profusion of proteins by *in silico* techniques.

Note added in proof: Presented data were calculated using sequence databases of December 1999.

Acknowledgements

J.S., T.D. and P.B. are supported by the DFG and by the EC (grant 01KW9602/6) as well as by the BMBF grants MEDSEQ and TARGID.

Received 27 September 1999; accepted 17 March 2000.

1. Pandey, A. & Lewitter, F. Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem. Sci.* **24**, 276–280 (1999).
2. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
3. Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA* **95**, 5857–5864 (1998).
4. Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28**, 231–234 (2000).
5. Schuler, G.D. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75**, 694–698 (1997).
6. Retief, J.D., Lynch, K.R. & Pearson, W.R. Panning for genes—a visual strategy for identifying novel gene orthologs and paralogs. *Genome Res.* **9**, 373–382 (1999).
7. Bork, P. & Gibson, T.J. Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184 (1996).
8. Wadman, M. Human Genome Project aims to finish 'working draft' next year. *Nature* **398**, 177 (1999).
9. Sunyaev, S. *et al*. Individual variation in protein coding sequences of the human genome. *Adv. Protein Chem.* (in press).
10. Prigent, C., Gill, R., Trower, M. & Sanseau, P. *In silico* cloning of a new protein kinase, Aik2, related to Drosophila Aurora using the new tool: EST Blast. *In Silico Biol.* **1**, 11 (1998).
11. Birney, E. & Durbin, R. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Ismb* **5**, 56–64 (1997).
12. Wolff, A.M., Petersen, J.G.L., Nilsson-Tillgren, T. & Din, N. The open reading frame YAL048c affects the secretion of proteinase A in *S. cerevisiae*. *Yeast* **15**, 427–434 (1999).