Supplemental Material for: "Targeted discovery of novel human exons by comparative genomics"

A. Siepel, M. Diekhans, B. Brejová, et al.

Supplementary Methods

Selection of Targets

Candidate genes were selected by starting with genome-wide sets of gene predictions, then removing candidates that overlapped genes (1) in the RefSeq (Pruitt et al., 2005) or Vega (Ashurst et al., 2005) sets, (2) already in the MGC, or (3) already in the MGC pipeline for full-length cloning. The ENSEMBL set (Hubbard et al., 2007) was not included among the "known genes" at this stage because it includes gene predictions that have little or no cDNA support. All predictions and known genes were represented by their UCSC Genome Browser coordinates for the May, 2004 assembly of the human genome (hg17). For cDNA-based collections, these are determined by BLAT alignments of cDNA sequences to the genome. A prediction and a known gene were considered "overlapping" if they were on the same strand and shared at least one base within their annotated coding regions. Predictions that did not contain at least one intron between coding exons were removed. Any predicted UTRs were ignored.

The initial N-SCAN candidates came from several sets of gene predictions, based on both the July 2003 (hg16) and May 2004 (hg17) human assemblies. Early candidates came from TWINSCAN (Korf et al., 2001) predictions with mouse as the informant genome, and later candidates came primarily from N-SCAN predictions with mouse (assembly mm5) as the informant genome and no EST evidence, or from N-SCAN predictions with mouse (mm5), rat (rn3), and chicken (galGal2) as informant genomes and with EST evidence (Arumugam et al., 2006). BLASTZ (Schwartz et al., 2003) pairwise alignments and MULTIZ (Blanchette et al., 2004) multiple alignments were used. Later predictions were subjected to iterative pseudogene removal (van Baren and Brent, 2006).

The EXONIPHY candidates came from a single set of predictions based on MULTIZ alignments of the human (hg16), mouse (mm3), rat (rn3) genome sequences (Siepel and Haussler, 2004). Predictions not in regions of large-scale synteny between human and mouse, as defined by the UCSC "syntenic net" (Kent et al., 2003), were discarded, because many of them reflect alignments of paralogous sequences, including processed pseudogenes aligned with homologs of their parent genes. Predictions in recent segmental duplications were also discarded, because early experiments suggested an elevated false positive rate in these regions. In addition, all candidates were manually inspected in the UCSC Genome Browser, and were discarded if they showed indications of being likely false positives, such as unusual patterns of cross-species conservation or heavily fragmented gene structures (as often occurs with pseudogenes).

The TRANSMAP predictions were obtained by starting with BLAT alignments to the mouse genome (mm6) of RefSeq mRNAs and GenBank mRNAs with CDS annotation (both from mouse), then mapping them to the human genome (hg17) via the human/mouse syntenic net (Kent et al., 2003). This approach implicitly discards candidate genes not in regions of large-scale synteny, but it allows some genes in segmental duplications to pass through. (The syntenic net does not require a one-to-one relationship between genomes.)

Targets were selected from eligible predictions by criteria that differed somewhat by prediction source, but, in all cases, candidates with little or no cDNA support—as defined by overlap in genomic coordinates with alignments of public EST or mRNA sequences—were given highest priority. Target selection occurred over a two-year period, during which the sets of known genes and cDNA alignments changed considerably.

Alignment of cDNAs to Genome Sequence

We used all EST and mRNA sequences available from GenBank as of June 1, 2007, except for those from two sources considered problematic for our analysis: sequences from the Athersys RAGE library (Harrington et al., 2001), which reflect induced gene expression; and a set of Invitrogen sequences that appear to have been modified to match the genome, in some cases based on alignments with pseudogenes (personal communication, RefSeq staff). The RSTs among the cDNAs were identified by their author and comment fields and tracked separately. All sequences were aligned to the human genome sequence (hg17) using BLAT. Each cDNA sequence with at least one high-quality alignment ($\geq$25% coverage and $\geq$95% identity) was assigned its best-matching position in the genome, plus any secondary positions having high quality alignments within 1% identity of the best match. Any cDNAs without high quality alignments were discarded. RSTs assigned multiple genomic positions (usually because of a recent genomic duplication) were excluded when determining success rates, novel exons, and NGFs, because of uncertainty about their locus of origin. Alignment gaps of no more than 12 bases were assumed to be polymorphisms or sequencing errors and were ignored (filled in) when defining the genomic coordinates for cDNAs. A direction of transcription (strand) was assigned to each aligned cDNA, if possible, based on its nucleotides at apparent splice sites. Unspliced cDNAs, and cDNAs that otherwise could not be assigned a strand, were not considered significant supporting evidence for benchmark exons (see below).

Evaluation of Hit Rates

Each RT-PCR experiment was associated with a set of gene predictions based on the PCR primer pair used in the experiment. Each primer pair was mapped to the human genome sequence (hg17) using the In-Silico PCR (isPcr) program (J. Kent, unpublished; http://hgdownload.cse.ucsc.edu/downloads.html), with `-maxSize=1000000`, `-minPerfect=18`, and `-minGood=15`. The primer pair was then associated with all gene predictions for which the genomic sequence matching each primer fell completely within a predicted exon and these exons were separated by at least one intron. The experiment defined by a primer pair is said to "test" all associated predictions. A primer pair that mapped to multiple genomic positions tested all matching predictions at all positions. A few primer pairs that could not be associated with predictions (because of clerical errors or changes in the genome sequence between assemblies) were discarded. Because the predictions tend to overlap partly but not completely, and because they have different properties (EXONIPHY predictions, for example, tending to be much shorter than N-SCAN predictions), success rates were evaluated at the level of *prediction clusters* rather than predictions. Prediction clusters correspond to the connected components of a graph in which nodes represent predictions and an edge is present between two nodes if and only if the corresponding predictions were both tested by the same primer pair.

An experiment was considered a "hit" if it produced a valid RST and that RST had an unambiguous mapping to the genome (see above). If the experiment did not produce a valid RST it was considered a "miss." Valid RSTs that mapped to multiple genomic positions provided ambiguous evidence about validation and were ignored. A prediction cluster was considered a "hit" if any associated experiment was a "hit," and was considered a "miss" if it had no associated "hits" and at least one "miss." At both the experiment and prediction cluster levels, hit rates were calculated as the number of hits divided by the number of hits and misses.

Definition of Benchmark Exons

Benchmark exons (BMEs) were derived from aligned cDNAs that revealed at least one canonical (GT-AG) intron and could be assigned an unambiguous direction of transcription. Each exon of a cDNA was given the genomic coordinates implied by its alignment, and was classified as initial (extreme 5′ end), terminal

(extreme 3′ end), or internal with respect to other exons of the same cDNA. cDNA exons whose flanking introns were canonical (both if internal, one if initial/terminal) were considered candidates for BMEs.

Any internal cDNA exon with canonical flanking introns defined an internal BME. The main supporting evidence for the internal BME consisted of this cDNA exon and any other internal cDNA exons sharing the same two boundaries. Initial or terminal cDNA exons that shared one boundary of the internal BME and terminated within the other boundary provided partial support for the internal BME. Similarly, an initial exon with a flanking canonical intron defined an initial BME, provided no overlapping cDNA suggested additional exons in the 5′ direction, and provided no other initial exon with the same 3′ boundary extended farther in the 5′ direction (Figure S1B). Terminal exons were defined in an exactly symmetric manner. The main supporting evidence of an initial/terminal BME consisted of itself and all other exons sharing its intron-flanking boundary. For both internal and initial/terminal BMEs, any other overlapping cDNAs—including unspliced cDNAs or cDNAs with noncanonical introns—was not considered significant supporting evidence.

Because of uncertainty in the alignment of cDNAs, two exon boundaries were considered "equal" if they were within 2bp of one another in genomic coordinates. In addition, because retained introns—which are relatively common in the cDNA databases—would otherwise create a proliferation of BMEs that span other BMEs, they were addressed by preprocessing, as follows. Any exon that spanned other (apparently spliced together) exons was split at the boundaries of the spanned exons. The new (artificial) boundaries in the set of exons produced in this way were considered initial or terminal boundaries, so these exons could provide at most partial support for existing BMEs. By these methods, a total of 457,724 BMEs were defined from all aligned public cDNAs.

Historical Exon Discovery

Novel exons and NGFs were defined for cut-off dates ranging from January 1, 2005 to June 1, 2007, in one-month intervals (Table S1). The submission dates of the GenBank entries were used to determine the date of origin of each sequence. Similarly, to measure exon discovery by cDNA sequencing over time (Figure 3), the number of BMEs with complete supporting cDNA evidence in the public database was tracked from January 1, 1990 to June 1, 2007, also in one-month intervals. A nonoverlapping subset of BMEs was used for this analysis, to help distinguish the discovery of new exons from the identification of variations on existing exons. From each set of overlapping BMEs on each date, the one with the largest number of supporting cDNAs was selected as the representative BME. The number of supporting exons per supported BME was also computed at each time point (Figure S4).

Identification of Homologs, GO Categories, and Domains

Predicted peptide sequences were generated from the NGFs, based on the reading frames of overlapping genes and gene predictions. These peptide sequences were searched against a database of vertebrate amino acid sequences using BLASTP, with an E-value threshold of 10^{-5}. Putative homologs were also identified by searching the untranslated NGF sequences against the same database using BLASTX, also with an E-value threshold of 10^{-5}. This method allowed for a direct comparison with noncoding RNAs (see below). The amino acid database was constructed from the translations of all RefSeq protein coding genes from the human, mouse, rat, cow, dog, and chicken genomes, as downloaded from the UCSC Genome Browser on November 18, 2006.

Gene Ontology (GO) categories were assigned to the RefSeq genes in the amino acid database using Entrez Gene (Maglott et al., 2007). Each NGF was then assigned the GO categories of its highest-scoring BLASTP match that had at least one GO category, and of all other matches scoring within 5% of the best match. The analysis of GO enrichments was performed with NGF clusters, rather than with individual NGFs, to avoid overcounting of especially long and/or fragmented genes. Each NGF cluster was assigned

the union of the GO categories of its constituent NGFs. In this way, 436 of 563 NGF clusters (77%), were assigned at least one GO category.

Domain matches were identified for the NGF peptides by searching with Reverse PSI-BLAST (RPS-BLAST) against the Conserved Domain Database (CDD) and the Pfam database (Marchler-Bauer et al., 2005), with an E-value threshold of 0.01. As with GO categories, the enrichment analysis was done with NGF clusters, which were assigned the domains of their constituent NGFs. 304 of 563 NGF clusters (54%) were assigned at least one domain.

Protein-Coding Potential

The NGFs were compared with a set of noncoding RNAs (ncRNAs) from RefSeq, which was obtained by downloading all accession numbers with the prefix "NR_" then discarding any that were annotated as expressed pseudogenes. Homologs of the ncRNAs were identified by searching them against a database of vertebrate genes with BLASTX, as described above for the NGFs. Out of 509 ncRNAs, 74 (14%) had significant BLASTX matches, while 630 of 734 (86%) NGFs had significant BLASTX matches ($P < 10^{-148}$, one-sided Fisher's exact test). To control for the shorter average length of the ncRNAs compared with the NGFs, the sequences were placed in bins of 1–199bp, 200–399bp, 400–599bp, . . . , and a one-sided Fisher's exact test was performed within each bin for which there was ample data ($\geq$10 sequences of each type). There was a significant enrichment for BLASTX matches among the NGFs in all such cases ($P < 0.01$).

The ncRNAs were searched against version 8.0 of the Rfam database (Griffiths-Jones et al., 2003) using the INFERNAL search tool (Eddy, 2002) in local mode. Based on guidelines from the INFERNAL documentation and the total length of all NGFs, bit scores of $>$20 were considered significant. 37 of 734 NGFs (5%), 32 of 269 randomly selected RefSeq protein-coding genes (12%), and 382 out of 509 ncRNAs (75%) had bit scores of $>$20.

To obtain the indel and substitution patterns (Figures 2 and S3), pairwise human-mouse alignments corresponding to the NGFs were extracted from the 17-way MULTIZ alignments in the UCSC Genome Browser. Columns with gaps in both human and mouse were removed. The NGFs were split into individual exons, and only indels completely contained within an exon were counted. Similarly, the distance between two substitutions was counted only if they were within the same exon and there was no intervening indel or unknown character. For comparison, an identical procedure was applied to nonredundant sets of protein-coding regions, UTRs, and ncRNAs from RefSeq.

Analysis of GO and Domain Enrichment

The set of human RefSeq genes in the UCSC Genome Browser (as of June 1, 2007) was used as a background set for the enrichment analyses. To mimic the procedure used for the NGF clusters, GO categories for the RefSeq genes were assigned by homology. This was accomplished by searching the translated sequences by BLASTP against a database of vertebrate amino acid sequences, exactly as for the NGFs (see above). The RefSeq genes were clustered by overlap in coding regions, and each cluster was assigned the GO categories of all constituent genes. There were 18,462 clusters, of which 17,124 (93%) were given at least one GO category. For each GO category assigned to at least five NGF clusters and five RefSeq clusters, a P-value indicating its over-representation among NGFs was computed by a one-sided Fisher's exact test. The method of Holm (1979) was used to control the family-wise error rate.

Domain enrichments were computed by a similar procedure. As for the NGFs, the domains for the RefSeq genes were identified by RPS-BLAST, and as for the GO categories, each RefSeq cluster was given the domains of its constituent genes. 15,619 out of 18,462 clusters (85%) were given at least one domain.

In Situ Hybridization to Zebrafish Embryos

We selected twenty three NGFs that could be mapped to the zebrafish genome assembly (danRer2; June, 2004) via the UCSC human/zebrafish syntenic net. At the time of selection, these NGFs were all completely novel with respect to known gene sets for both human and zebrafish. Preference was given to NGFs with very weak or no cDNA support. An exon-specific probe was synthesized for each NGF by PCR amplification with nested primer pairs. Probe synthesis was not successful for four NGFs because their exons were too small or no unique nested primers could be found. All probes were confirmed by sequencing prior to use. Zebrafish embryos at 48 and 72 hours past fertilization (hpf) were prepared by a standard protocol (http://www.zfin.org) modified for use with 96-well microtiter plates. Hybridizations were done at both 50 and 55 deg C (both near optimal for DNA-RNA hybridization) with similar results. They were done in parallel with both positive and negative controls.

Expression Levels

The "Tissues + Mixtures" sample data set for the Affymetrix GeneChip® Human Exon 1.0 ST Array was obtained from http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx. The exon-level results generated using the apt-probeset-summarize command from the Affymetrix Power Tools (APT) were used. These include probeset summaries for all tissues and tissue mixtures based on the RMA (Irizarry et al., 2003) (option -a rma-sketch) and PLIER (http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf) (option -a quant-norm,pm-gcbg,plier) algorithms. They also include DABG (detected above background) P-values (option -a dabg), indicating the probability that the intensity for each probeset $\times$ tissue could have been observed by chance based on the intensities observed for the background probes on each array. The RMA and PLIER probeset summaries produced very similar results, so we have reported only the RMA results.

Four types of exonic features were considered: novel exons, NGF clusters, RefSeq coding exons, and clusters of RefSeq genes (defined as above). Using the probeset coordinates available in the "Affy All Exon" track in the UCSC browser (hg17), a mapping was constructed between each feature of each type and the set of probesets contained within exons of that feature. At least one probeset was available for 75% of novel exons, 95% of NGF clusters, 98% of RefSeq exons, and 99.5% of RefSeq gene clusters. A P-value for each feature $\times$ tissue was calculated by combining the DABG P-values of all probesets associated with that feature using Fisher's method (Fisher, 1925) . The P-values for all three replicates per probeset $\times$ tissue were pooled in computing these combined P-values. A feature was considered to be significantly expressed above background if it had (nominal) $P < 0.01$. It was considered to be tissue-specific if it had $P < 0.01$ in the tissue in question and $P > 0.2$ in all other tissues.

Similarly, an estimated expression intensity for each feature $\times$ tissue was calculated by taking the median over all probesets assigned to the feature, of the median over the three tissue replicates, of the RMA- or PLIER-based probeset intensity summaries. The analysis of expression intensities was restricted to features significantly expressed above background so that features expressed at or near the background level did not drive the results. To summarize the variation across tissue in expression, we calculated, for each feature, the coefficient of variation (sample standard deviation/sample mean) of the estimated expression intensities for all tissues.

Supplemental Tables

Table S1: Numbers of novel exons by degree of prior support

	Jan 2005		Jun 2007	
Prior Support	No.	%	No.	%
No signif.	1983	90.6	1691	89.4
Partial	205	9.4	201	10.6
Total	2188	100.0	1892	100.0

Table S2: NGFs and NGF clusters by relationship with cDNA clusters

	NGFs		NGF Clusters	
Class	No.	%	No.	%
Completely novel[a]	245	33.4	169	30.0
5′ extension[b]	160	21.8	134	23.8
3′ extension[c]	136	18.5	102	18.1
Single internal exon[d]	87	11.9	64	11.4
Other	106	14.4	94	16.7
Total	734	100.0	563	100.0

[a] All exons novel.
[b] All novel exons 5′ of cDNA cluster.
[c] All novel exons 3′ of cDNA cluster.
[d] One novel exon that is not an extension.

Table S3: GO categories over-represented among novel gene fragments

Category	Description	N^a	n^b	$E[n]^c$	xfold^d	P^e
GO:0044421	extracellular region part	1046	63	26.6	2.4	3.1e-07
GO:0005615	extracellular space	823	54	21.0	2.6	3.6e-07
GO:0042995	cell projection	377	31	9.6	3.2	2.0e-05
GO:0030414	protease inhibitor activity	146	17	3.7	4.6	3.6e-04
GO:0051260	protein homooligomerization	45	10	1.1	8.7	4.8e-04
GO:0005576	extracellular region	1441	68	36.7	1.9	5.5e-04
GO:0031012	extracellular matrix	293	24	7.5	3.2	7.7e-04
GO:0002020	protease binding	5	5	0.1	39.3	1.2e-03
GO:0005578	proteinaceous extracellular matrix	286	23	7.3	3.2	1.7e-03
GO:0003774	motor activity	204	19	5.2	3.7	1.7e-03
GO:0001889	liver development	12	6	0.3	19.6	1.8e-03
GO:0031589	cell-substrate adhesion	87	12	2.2	5.4	3.4e-03
GO:0005518	collagen binding	24	7	0.6	11.5	4.7e-03
GO:0005903	brush border	24	7	0.6	11.5	4.7e-03
GO:0009612	response to mechanical stimulus	24	7	0.6	11.5	4.7e-03
GO:0051259	protein oligomerization	80	11	2.0	5.4	8.4e-03
GO:0005929	cilium	41	8	1.0	7.7	1.4e-02
GO:0004857	enzyme inhibitor activity	265	20	6.7	3.0	1.7e-02
GO:0044463	cell projection part	90	11	2.3	4.8	2.2e-02
GO:0022610	biological adhesion	736	38	18.7	2.0	2.8e-02
GO:0007155	cell adhesion	736	38	18.7	2.0	2.8e-02
GO:0001539	ciliary or flagellar motility	13	5	0.3	15.1	3.2e-02
GO:0007586	digestion	95	11	2.4	4.5	3.5e-02

[a] Number of genes in background set assigned to category, out of 17124 with at least one assignment.

[b] Number of NGF clusters assigned to category, out of 436 with at least one assignment.

[c] Number of NGF clusters expected to be assigned to category if categories were randomly drawn from background distribution ($N \times 436/17124$).

[d] Fold enrichment in NGF clusters ($n/N \times 17124/436$).

[e] One-sided P-value by Fisher's exact test, after adjustment for multiple comparisons. All categories with $N \geq 5$ and $P < 0.05$ are shown.

Table S4: Protein domains over-represented among novel gene fragments

Category	Description	N^a	n^b	$E[n]^c$	xfold^d	P^e	$P_{\text{adj}}{}^f$
smart00216	von Willebrand factor (vWF) type D domain	9	6	0.2	34.3	2.0e-07	**4.4e-06**
pfam00094	von Willebrand factor type D domain	9	6	0.2	34.3	2.0e-07	**4.4e-06**
COG5245	Dynein, heavy chain [Cytoskeleton]	17	5	0.3	15.1	4.9e-05	**9.9e-04**
pfam00067	Cytochrome P450	57	5	1.1	4.5	6.5e-03	1.2e-01
smart00408	Immunoglobulin C-2 Type	230	11	4.5	2.5	6.7e-03	1.2e-01
cd00931	Immunoglobulin domain cell adhesion molecule (cam) subfamily	239	11	4.7	2.4	8.8e-03	1.5e-01
smart00409	Immunoglobulin	323	13	6.3	2.1	1.3e-02	2.0e-01
cd01475	VWA_Matrilin	77	5	1.5	3.3	2.0e-02	3.0e-01
cd00096	Immunoglobulin domain family	238	9	4.6	1.9	4.8e-02	6.7e-01

[a] Number of genes in background set having domain, out of 15619 with at least one assignment.

[b] Number of NGF clusters having domain, out of 304 with at least one assignment.

[c] Number of NGF clusters expected to have domain if domains were randomly drawn from background distribution ($N \times 304/15619$).

[d] Fold enrichment in NGF clusters ($n/N \times 15619/304$).

[e] Nominal one-sided P-value by Fisher's exact test. All domains with $N \geq 5$ and $P < 0.05$ are shown.

[f] P-value adjusted for multiple comparisons using Holm's correction. Values ≤ 0.05 are shown in bold.

Table S5: Summary of selected novel gene fragments

Cluster[a]	NEs[b]	Size[c]	Name[d]	Other Evidence	Functional Information
ngf51–ngf55	24	66/3766	–	Possible extension of mRNA (BC042869).	Homologous to several human axonemal dynein heavy chains.
ngf60	9	21/942	–	No prior cDNA evidence. Possible distal 3′ extension of *ELYS* transcription factor (50kb away).	Expressed in brain in zebrafish embryos (see text). Homologous to kinesin-like proteins.
ngf101–ngf103	18	52/2832	*OTOG*	5′ and 3′ extension of mRNA (AK128214).	Encodes otogelin, a glycoprotein specifically expressed in the inner ear, primarily during early development (Cohen-Salmon et al., 1997; El-Amraoui et al., 2001).
ngf132	5	90/4314	*DYNC2H1*	Recently became provisional RefSeq (NM_001080463).	Encodes cytoplasmic dynein 2 heavy chain 1.
ngf158–ngf161	41	43/7524	*MUC19*	Overlaps much shorter mRNA (AY236870). Mouse ortholog is provisional RefSeq (NM_207243).	Expressed mainly in salivary glands and may contribute to viscosity of salivary mucus (Chen et al., 2004; Culp et al., 2004).
ngf167–ngf171	24	53/2293	–	Essentially no prior cDNA evidence. 3′ end overlaps predicted RefSeq (NM_173591) by one exon.	Homologous to mammalian von Willebrand factors and mucins.
ngf191–ngf192	11	78/4471	*DNAH10*	Recently became provisional RefSeq (NM_001083900).	Encodes axonemal dynein heavy chain 10.
ngf332	5	45/2141	*SDK2*	5′ extension of RefSeq (NM_019064). Orthologous to mouse and chicken RefSeqs.	Encodes cell adhesion protein thought to guide axonal terminals to specific synapses in developing neurons.
ngf338–ngf339	14	45/4497	*DNAH17*	5′ extension of mRNA (AL832652).	Encodes axonemal dynein heavy chain 17.
ngf393–ngf396	16	70/4148	–	May bridge mRNAs FLJ37357 and LOC200383.	Homologous to several human axonemal dynein heavy chains.
ngf408–ngf409	10	45/2113	*MYO7B*	Large internal fragment of gene recently added to RefSeq.	Encodes myosin VIIB.
ngf490–ngf493	26	61/3644	–	May bridge mRNAs AK128592 and FLJ44290.	Homologous to several cytosolic dynein heavy chains.
ngf498	3	22/1029	*CNTN3*	Recently became validated RefSeq (NM_020872).	Encodes adhesive glycoprotein thought to function in neuronal outgrowth.
ngf510–ngf513	29	30/2542	–	5′ extension of predicted RefSeq (NM_153264).	Homologous to several collagens.
ngf608–ngf611	20	35/1125	*COL28A1*	Recently became provisional RefSeq (NM_001037763).	Encodes collagen type XXVIII. Expressed specifically in dorsal root ganglia and peripheral nerves. May contribute to connective tissue development (Veit et al., 2006).
ngf634–ngf638	25	40/1893	–	5′ extension of mRNA (AK025690).	Homologous to several myosin heavy chains.
ngf653–ngf657	25	103/5148	*SSPO*	Recently became provisional RefSeq (NM_198455).	Encodes glycoprotein of thrombospondin family expressed in the subcommissural organ. Thought to be involved in CNS development.
ngf698–ngf703	33	98/5101	*HMCN2*	5′ extension of mRNA AL834139. Known paralog of hemicentin-1 but not in gene catalogs.	Encodes extracellular matrix protein likely to have a role in the architecture of adhesive and flexible cell junctions (Xu et al., 2007).

[a]Ids of NGFs in cluster

[b]Number of novel exons

[c]Estimated total size, based on gene predictions, homologs, and cDNA evidence (exons/amino acids).

[d]Putative gene name, based on overlapping genes and/or orthologs.

Table S6: Overlap of novel exons by Affymetrix transfrags

Exon set	Threshold[a]	Median Coverage (%)[b]	Combined Coverage (%) [c]
Novel exons	$\geq$0.5	10.6	32.0
RefSeq CDS	$\geq$0.5	50.1	75.8
Novel exons	$>$0.0	23.8	50.4
RefSeq CDS	$>$0.0	61.1	83.2

[a]Fraction of bases in exon required to fall within annotated transfrags.

[b]Median percentage, across all 8 cell lines, of exons overlapped by transfrags. The "long RNA" ($>$200 nucleotides) data set of Kapranov et al. (2007) was used. For two cell lines (HeLa and HepG2), transfrags for nuclear RNA as well as cytosolic polyA+ RNA were included, making a total of 10 sets of transfrags.

[c]Percentage of exons overlapped by a merged set of transfrags from all cell lines.

Supplemental Figures

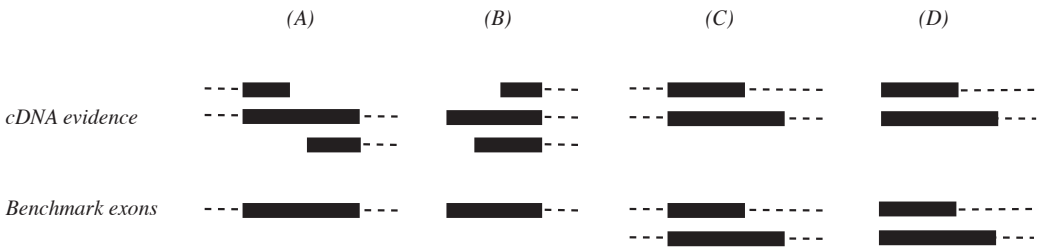

Figure S1: Examples of benchmark exons, with solid bars for exons and dashed lines for flanking introns. A benchmark exon is a best guess at the true genomic boundaries of an exon, based on all of the available cDNA evidence (including the RSTs). *(A)* Internal exons trump initial or terminal exons in defining benchmark exons, because most cDNA evidence is fragmentary. *(B)* When only initial (or only terminal) exons are present and all of them share a splice site, the longest one defines the benchmark exon. *(C)* and *(D)* Exons that overlap but do not share splice sites define separate benchmark exons.

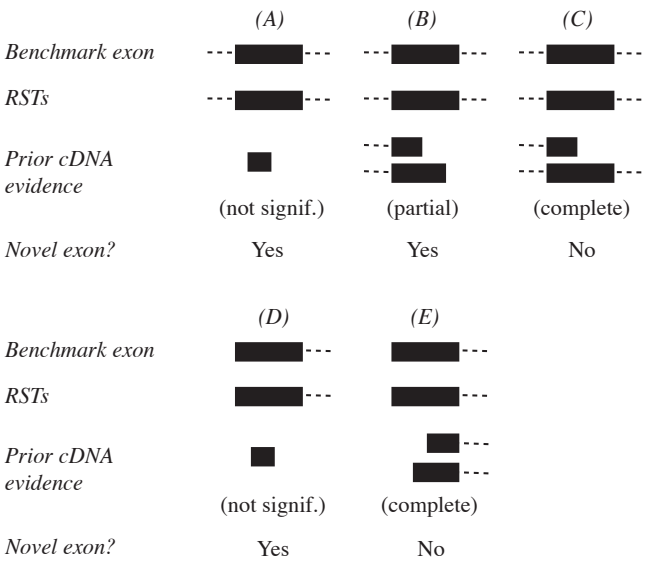

Figure S2: Definition of novel exons. The degree of support for each benchmark exon is separately evaluated based on the RSTs and prior cDNA evidence. A "novel exon" is a benchmark exon that has complete support from RSTs but incomplete support from prior cDNA evidence. Complete support is defined as evidence for both splice sites of an internal benchmark exon, or evidence for one splice site of an initial or terminal benchmark exon (cases (*C*) and (*E*)). Other levels of cDNA support include *not significant* (no splice sites supported; cases (*A*) and (*D*)) and *partial* (one splice site supported, internal exons only; case (*B*)). Note that novelty is a function of the cutoff date used to define the prior cDNA evidence (see Figure 3).

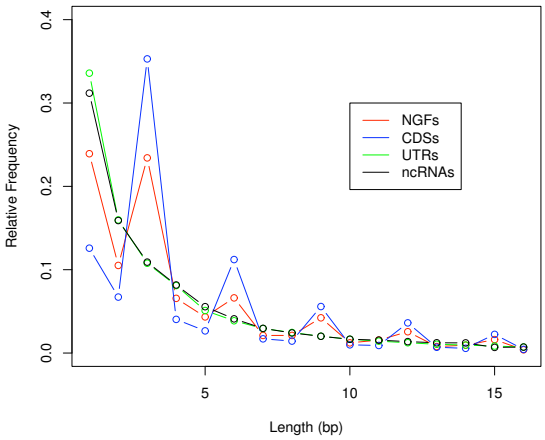

Figure S3: Distributions of indel lengths in human-mouse alignments for NGFs versus coding sequences (CDSs), UTRs, and ncRNAs from RefSeq. Like the CDSs, the NGFs show a pronounced period of three.

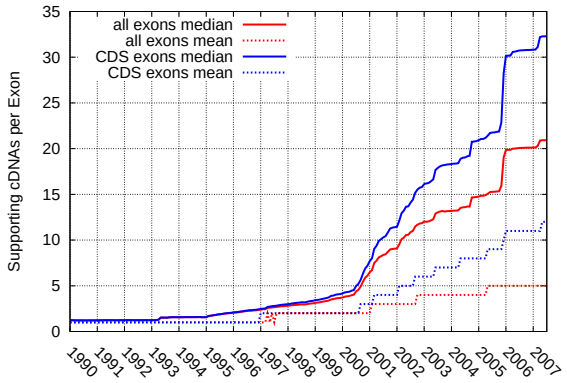

Figure S4: Mean and median numbers of supporting cDNAs per benchmark exon as a function of time, for all exons and coding exons.

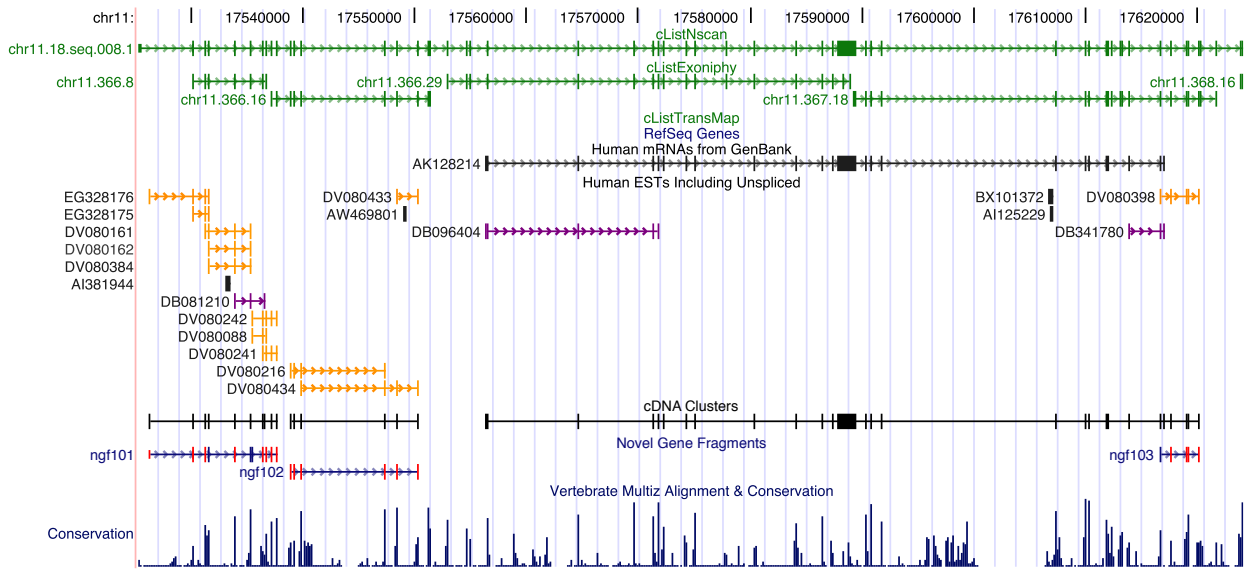

Figure S5: The region of the *OTOG* gene on chromosome 11, showing the mRNA sequence currently in GenBank (AK128214) and apparent extensions from ngf101–ngf103. Colors are as in Figure 5. This gene encodes a non-collagenous glycoprotein that is expressed only in the inner ear. Probably for this reason, it is poorly represented by ESTs.

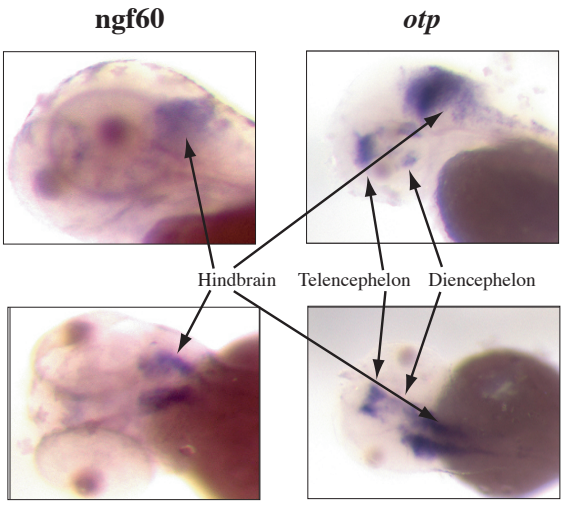

Figure S6: Whole mount *in situ* hybridization for zebrafish ortholog of ngf60 and *OTP* at 72 hours past fertilization.

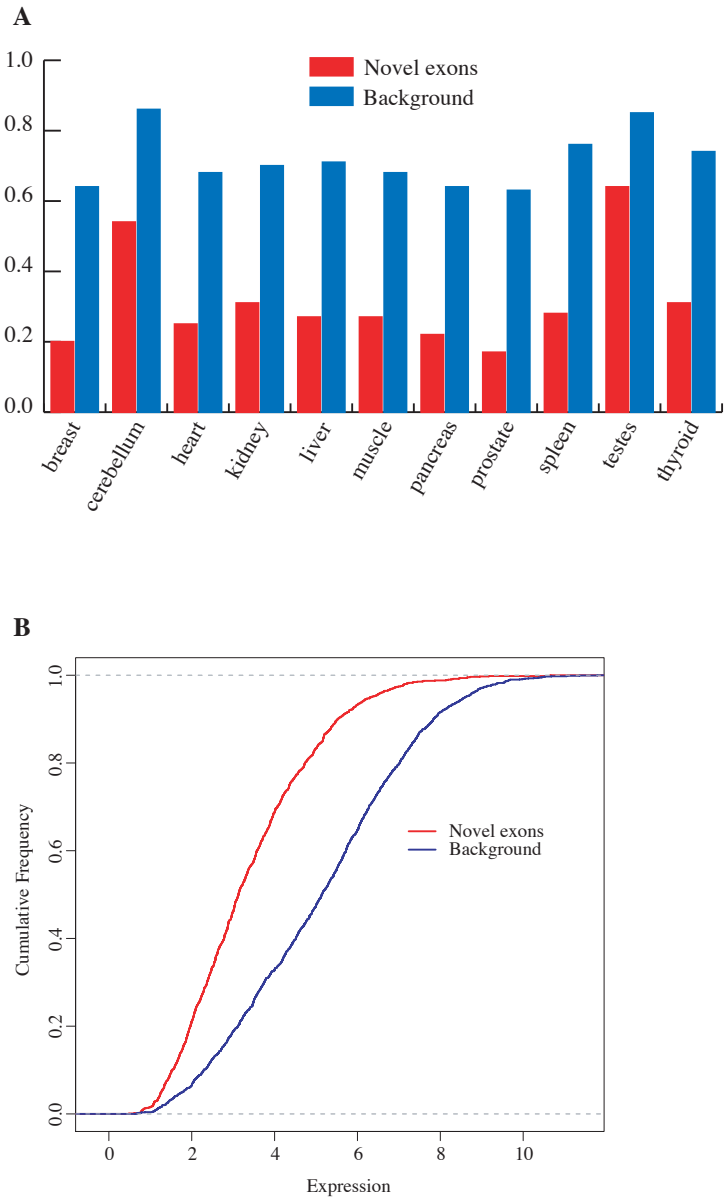

Figure S7: (A) Fractions of novel exons (red) and coding exons from RefSeq (blue) that show significant expression above background (DABG $P < 0.01$) in various tissues, according to the Affymetrix Human Exon Array. (B) Cumulative distribution of expression levels among exons showing significant expression above background. The expression levels are based on RMA probeset summaries. Shown are the data for a mixture of heart, testes, and cerebellum RNA, but the plots for all tissues and tissue mixtures were similar.

References

Arumugam, M., Wei, C., Brown, R. H., and Brent, M. R., 2006. Pairagon+N-SCAN_EST: a model-based gene annotation pipeline. *Genome Biol*, **7 Suppl 1**:1–10.

Ashurst, J. L., Chen, C.-K., Gilbert, J. G. R., Jekosch, K., Keenan, S., Meidl, P., Searle, S. M., Stalker, J., Storey, R., Trevanion, S., *et al.*, 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res*, **33**(Database issue):459–465.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.*, 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, **14**(4):708–715.

Chen, Y., Zhao, Y. H., Kalaslavadi, T. B., Hamati, E., Nehrke, K., Le, A. D., Ann, D. K., and Wu, R., 2004. Genome-wide search and identification of a novel gel-forming mucin MUC19/Muc19 in glandular tissues. *Am J Respir Cell Mol Biol*, **30**(2):155–165.

Cohen-Salmon, M., El-Amraoui, A., Leibovici, M., and Petit, C., 1997. Otogelin: a glycoprotein specific to the acellular membranes of the inner ear. *Proc Natl Acad Sci U S A*, **94**(26):14450–14455.

Culp, D. J., Latchney, L. R., Fallon, M. A., Denny, P. A., Denny, P. C., Couwenhoven, R. I., and Chuang, S., 2004. The gene encoding mouse Muc19: cDNA, genomic organization and relationship to Smgc. *Physiol Genomics*, **19**(3):303–318.

Eddy, S. R., 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**:18.

El-Amraoui, A., Cohen-Salmon, M., Petit, C., and Simmler, M. C., 2001. Spatiotemporal expression of otogelin in the developing and adult mouse inner ear. *Hear Res*, **158**(1-2):151–159.

Fisher, R. A., 1925. *Statistical methods for research workers*. Oliver & Loyd, London, 13th edition.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R., 2003. Rfam: an RNA family database. *Nucleic Acids Res*, **31**(1):439–441.

Harrington, J. J., Sherf, B., Rundlett, S., Jackson, P. D., Perry, R., Cain, S., Leventhal, C., Thornton, M., Ramachandran, R., Whittington, J., *et al.*, 2001. Creation of genome-wide protein expression libraries using random activation of gene expression. *Nat Biotech*, **19**:440–445. 10.1038/88107.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**:65–70.

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., *et al.*, 2007. Ensembl 2007. *Nucleic Acids Res*, **35**(Database issue):610–617.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P., 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, **31**(4):e15.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D., 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA*, **100**:11484–11489.

Korf, I., Flicek, P., Duan, D., and Brent, M. R., 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**:S140–S148.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T., 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, **35**(Database issue):26–31.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., *et al.*, 2005. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*, **33**(Database issue):192–196.

Pruitt, K. D., Tatusova, T., and Maglott, D. R., 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**(Database issue):501–504.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W., 2003. Human-mouse alignments with BLASTZ. *Genome Res*, **13**:103–107.

Siepel, A. and Haussler, D., 2004. Computational identification of evolutionarily conserved exons. In *Proc. 8th Int'l Conf. on Research in Computational Molecular Biology*, pages 177–186.

van Baren, M. J. and Brent, M. R., 2006. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res*, **16**(5):678–685.

Veit, G., Kobbe, B., Keene, D. R., Paulsson, M., Koch, M., and Wagener, R., 2006. Collagen XXVIII, a novel von Willebrand factor A domain-containing protein with many imperfections in the collagenous domain. *J Biol Chem*, **281**(6):3494–3504.

Xu, X., Dong, C., and Vogel, B. E., 2007. Hemicentins assemble on diverse epithelia in the mouse. *J Histochem Cytochem*, **55**(2):119–126.