A non-EST based method for exon-skipping prediction

Rotem Sorek*†, Ronen Shemesh†, Yuval Cohen†, Ortal Basechess†, Gil Ast* and Ron Shamir#

* Department of Human Genetics, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel

† Compugen, 72 Pinchas Rosen St., Tel Aviv 69512, Israel.

School of Computer Science, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

Correspondence should be addressed to Rotem Sorek, +972-3-7658536 (Tel); +972-3-7658555 (Fax); rotem@compugen.co.il (e-mail).

Abstract

It is estimated that between 35% and 74% of all human genes can undergo alternative splicing. Currently, the most efficient ways for large-scale detection of alternative splicing are through Expressed Sequence Tags (ESTs) or microarray analysis. As these merely sample the transcriptome, splice variants that do not appear in deeply sampled tissues have a low probability to be detected. We present a new method by which we can predict that an internal exon is skipped (namely whether it is a cassette-exon) merely based on its naked genomic sequence and on the sequence of its mouse ortholog. No other data, such as ESTs, are required for the prediction. Using our method, which was experimentally validated, we detected hundreds of novel splice variants that were not detectable using ESTs, and showed that a substantial fraction of the splice variants in the human genome could not be identified through current human EST or cDNA data.

Introduction

Alternative splicing is a mechanism allowing one gene to produce multiple, sometimes functionally distinct, protein variants (Maniatis and Tasic 2002). In recent years, numerous studies have shown that the phenomenon of alternative splicing is very prevalent in mammalian genes (Brett et al. 2000; Kan et al. 2001; Kan et al. 2002; Lander et al. 2001; Mironov et al. 1999; Modrek et al. 2001). All these studies used Expressed Sequence Tags (ESTs) or cDNAs for detection of alternative splicing. Other studies used microarrays specifically designed for detection of splice variants (Johnson et al. 2003).

Although much progress has been made in the field of computational detection of alternative splicing in recent years (reviewed in (Graveley 2001) and (Modrek and Lee 2002)), the full extent of splice variants in the human genome is far from being completely known. ESTs, which are the main source of information for alternative splicing prediction, are a problematic source of information, as they are merely a sample of the transcriptome. The detection of a particular splice variant by ESTs is possible only if its transcription level is sufficiently high in a tissue type for which an EST library has been prepared. Moreover, as most ESTs are generated from the 5' or the 3' termini of the transcript, dbEST is biased towards under-representation of splice variants involving exons that are in the middle of long transcripts (Johnson et al. 2003). In addition, ESTs are very noisy and contain numerous erroneous sequences (Sorek and Safer 2003; Sorek et al. 2003), so that some of the EST-predicted splice variants may be artifacts (Modrek and Lee 2002).

Indeed, Johnson and colleagues, who has recently investigated the extent of human alternative splicing using large-scale microarrays experiments, reported on numerous events of alternative splicing that were not represented in ESTs (Johnson et al. 2003). However, even microarray experiments are not sufficient for the identification of all splice variants, as they do not sample all combinations of possible tissues, developmental stages and conditions.

Comparative genomics has recently proven a useful approach for alternative splicing research (Modrek and Lee 2003; Nurtdinov et al. 2003; Resch et al. 2004; Sorek and Ast 2003; Sorek et al. 2003). Specifically, we have found that conserved alternatively spliced internal exons (of the "cassette-exons" type) are usually flanked by intronic sequences that are conserved between human and mouse, a feature only rarely seen in constitutively spliced exons (Sorek and Ast 2003). These conserved intronic sequences are probably involved in the regulation of alternative splicing. We have also recently found that alternative exons that are conserved between human and mouse possess characteristics, such as smaller size and divisibility by 3, which distinguish them from non-conserved alternatively spliced exons (Sorek et al. 2003). In the current study we show (and experimentally verify) how the combination of these and additional features, which distinguish alternative from constitutive exons, can be used for the accurate prediction whether an exon is an alternative cassette exon or not, even when there are no ESTs that indicate its skipping.

Results and Discussion

To identify and characterize features that distinguish between alternative and constitutive exons, we used the training exons sets from (Sorek and Ast 2003), which contained 243 alternative internal exons (cassettes) and 1753 constitutive internal exons that are conserved between human and mouse (see Methods). These sets were based on EST analysis of GenBank (release 131), where exons were defined as constitutive if there were at least 4 expressed sequences supporting them, and no EST skipping them, both in human and in mouse.

Table 1 summarizes the major classifying features that we characterized. In short, alternatively spliced exons are flanked by intronic sequences that are more conserved between human and mouse; they are shorter than constitutively spliced exons; their size tends to be a multiple of 3; and they have higher identity level when aligned to their mouse counterpart exon (Fig 1 A-E). These differences probably stem from the unique function of the alternative exons: Since these exons are cassette exons that are sometimes inserted and sometimes skipped, their size should be a multiplication of 3 so that their skipping would not alter the reading frame of the downstream exons. This constraint, which was also recently reported by (Resch et al. 2004), does not apply to constitutively spliced exons. The higher identity level between human and mouse could be explained by the fact that alternatively spliced exons frequently contain sequences that regulate their splicing (exonic splicing enhancers and silencers, reviewed in (Cartegni et al. 2002)). These regulatory sequences add another level of conservation constraint on the exon

sequence. The fact that alternatively spliced exons are smaller than constitutively spliced ones was observed before (Thanaraj and Stamm 2003), and might be related to sub-optimal recognition of smaller exons by the spliceosome (Berget 1995).

The features described above could be used to identify exons that are skipped in the human and the mouse genomes. However, each feature by itself provides only a weak classification for exons. Our goal was to find a combination of features that would detect a substantial fraction of the alternative exons, while making near-zero false positive detection errors. The features we have chosen are the following: (i) exon length, (ii) divisible/not divisible by 3, (iii) percent identity when aligned to the mouse counterpart, and (iv) conservation in the upstream and downstream intronic sequences. Each of the two "intronic conservation" features (upstream and downstream) were divided into two sub-features: (a) length of best human/mouse local alignment in the 100 intronic nucleotides nearest to the exon (where only local alignments with at least 12 consecutive perfectly matching nucleotides were considered) and (b) identity level in this local alignment.

For each of the features we defined a set of thresholds (see Methods). For example, "human/mouse exon identity" threshold can be set to 100%, at least 99%, at least 98%, and so forth. Similarly, the thresholds for "length of conserved upstream region" can be set to 100, at least 95, at least 90 and so forth. By using a threshold for each of the seven features above, one gets a *classification rule* that classifies as alternative all exons that pass all seven thresholds. Such a rule might, for example, be: "all exons that are at least 99% conserved with their mouse counterpart **and** have at least 95 conserved nucleotides upstream the exon **and** are divisible by 3 **and** …".

We enumerated all possible rules (about 100 million rules) and tested the quality of the resulting classification on our training set of 243 alternative and 1,753 constitutive exons. We sought a rule that would correctly identify a maximum number of alternative exons from the training set while making no false positive identification.

The best rule that emerged was the following: At least 95% identity with the mouse exon counterpart; exon size is a multiple of 3; a best local alignment of at least 15 intronic nucleotides upstream the exon with at least 85% identity; and a perfect match of at least 12 consecutive intronic nucleotides downstream the exon. This combination of features identified 76 exons, or 31% of the 243 alternatively spliced exons in our training set, while none of the 1,753 constitutively spliced exons matched these features. To check the robustness of this analysis we employed 5-way cross validation. (see supplementary material for details) The average sensitivity in these five analyses was 32.3% and the average specificity was 99.72%.

The above combination of parameters can therefore be used to identify alternatively spliced exons with very high specificity, making less than 0.3% false positive calls. We note that since the ratio of constitutive to alternative exons in the genome is probably higher than in our training set, and since our training set may have some other unknown bias, the performance in genome-wide application of the rule may be somewhat lower.

To test this classifier in a genome-wide manner, as well as to discover novel splice variants in the human genome, we collected a large set of 108,983 human exons, for which a mouse counterpart could be identified (see Methods). To ensure the coherence of the analysis, we excluded our training exons from this analysis. For each of

the exons, all classifying parameters were calculated. Out of the 108,983 human exons, 952, or ~1%, were found to comply with the above-mentioned combination of parameters. Information on these 952 exons appears as Supplementary Material.

To check if these exons are indeed alternatively spliced, we searched for human expressed sequences (ESTs or cDNAs) that skip the exons but contain the two flanking exons. For 453 (48%) of the 952 candidate alternative exons there was such skipping evidence. For comparison, only 7% (7495 exons) out of our entire set of 108,983 exons had similar skipping EST evidence. This means that our classification rule indeed substantially enriches for alternatively spliced exons.

Moreover, there is evidence that EST databases can contain spurious sequences that appear as splice variants but are, in fact, artifacts caused by aberrant splicing. Such splicing artifacts are usually characterized by low EST support, although there are cases in which real, functional splice variants are supported by a single EST (Sorek et al. 2004). Indeed, only 17% of the 453 exons that were classified as 'alternative' by our rule had their exon-skipping supported by only one EST – the rest were supported by two or more. In comparison, skipping was supported by only a single EST in 46% of the total 7495 exons that showed skipping EST evidence. This suggests that our classification rule enriches for alternatively spliced exons with higher probability of being "real" relative to alternative exons merely supported by EST evidence. To calculate the classification sensitivity of the whole genome analysis while eliminating the low EST coverage factor, we took only exons that were supported by at least 10 human ESTs skipping the exon. There were 873 such exons in the entire set of 108,983 exons, and 176 in our set of 453 exons classified as alternatives. This means that the sensitivity of our analysis on the

whole genome is at least 20% (176/873). This is probably an underestimate, as we eliminated our training exons-set from the whole genome analysis.

We manually examined the remaining 499 candidate alternative exons (952-453), for which no EST/cDNA showing exon skipping event was found, by using the UCSC genome browser (April 2003). We found that for 190 additional exons (out of the 499) there was a human expressed sequence showing patterns of alternative splicing other than exon skipping (41 cases (22%) of alternative donor/acceptor; 33 cases (17%) of intron retention; 14 cases (7%) of mutually exclusive exons. More complicated types, such as double and triple exon skipping, comprise the remaining). Thus, for 643 (453 + 190; 68%) of the 952 candidate alternative exons identified by our method there was independent evidence for alternative splicing in dbEST and RefSeq.

But what about the remaining 309 candidate exons for which no EST or cDNA indicating the skipped isoform was found? These can still be rarely expressed alternatively spliced exons, or exons that are specific to a tissue, developmental stage or condition which is under-represented in dbEST, so that an EST representing their skipping isoform has not been sequenced yet. Indeed, while on average there were 32 supporting expressed sequences per exon in our general set of 108,983 exons (median 10), the support for the 309 candidate alternative exons was much smaller, averaging 14 sequences (median 7). This shows that the 309 candidate exons are supported by fewer ESTs than the average exon, in accordance with our hypothesis that under-representation in dbEST is the cause for not identifying them as alternatively spliced.

To test whether these candidate alternative exons for which no skipping ESTs were found are indeed alternative, we selected 5% of them (15 exons) for experimental

verification (Table 2). Only exons with EST support equal or less than the average (14 sequences) were selected for this verification, as such alternative splicing events are more likely to have been missed in dbEST due to low sampling and not due to a their appearance in a transient developmental state or in a rare condition. For each of these 15 exons, primers were designed from the two flanking exons. RT-PCR reactions were carried out with RNA extractions of 14 different tissue types (see Methods). For nine of these exons, a splice variant was detected in at least one of the 14 tissues tested (Figure 2). In six of the nine cases the variant represented exon skipping. Interestingly, in the other three cases the exon was alternatively spliced, but in a pattern other than exon-skipping: Two cases (genes BAZ1A and SMARCD1) of alternative acceptor site, and one case (VLDLR) of intron retention. This is consistent with our genome-wide scan, where 453/643 (70%) cases that were identified according to the classifying parameters were exon-skipping, while the remaining 30% exhibited other types of alternative splicing.

The above experimental results indicate that at least 60% (9/15) of our predictions are true (although this estimate can have a relatively large variance, due to the small size of exon set tested). Some or all of the remaining six exons might also be alternatively spliced, but in a tissue other than the ones we tested, or in an early developmental stage. We therefore believe that the actual prediction rate of this method may even be higher.

The classification rule that was chosen for the experimental verification retrieves alternatively spliced exons with a very high specificity (less than 0.3% false positive rate) but at the price of a relatively low sensitivity (20-32%). Other rules can be chosen in which sensitivity is higher, but naturally this would increase the false positive rate of the prediction. Figure 3 presents a sensitivity versus false positive rate plot (ROC curve) for

different rules selecting for increasing number of alternative exons from our test set of 243 exons. As shown in the figure, it is possible to employ a rule that would identify up to 73% of the alternative exons, but this rule would also retrieve 36% of the constitutively spliced exons (the upper limit of 73% is due to the Boolean nature of the "divisibility by 3" feature). Note, that since most of the exons in the human genome are constitutive, such a rule would have low predictability for exon skipping: Assuming, for example, that ~10%, or 20,000 out of the ~200,000 predicted exons in the human genome, are alternative, the probability that an exon identified by the 73%:36% rule would really be alternative is only 18% (0.73*20,000/[0.73*20,000 + 0.36*180,000]). This is why we preferred a rule with close to zero false positives. The curve in Fig. 3 presents a variety of alternatives, and allows the selection of a rule for a desired target specificity or sensitivity. For example, 50% sensitivity is achievable at about 1.8% false positive rate.

Our method is able to identify alternative splicing *ab initio*. Other computational approaches to detect alternative splicing were previously described, but most of them used ESTs and/or cDNAs, or information from transcripts predicted using ESTs, to predict alternative splicing (e.g., Clamp et al. (2003) and Haas et al. (2003) ; also reviewed at (Modrek and Lee 2002)). There was also an attempt to predict alternatively spliced exons using suboptimally scored exons in the gene structure prediction software GENSCAN (Burge and Karlin 1997) (see http://genes.mit.edu/GENSCANinfo.html), but as far as we know this prediction method was not tested experimentally.

We have described a novel computational method for prediction of alternative splicing. A possible improvement of the method could be the addition of more classifying features. One such feature could be the comparison of the flanking intronic sequences

between the human and other genomes. For example, we were able to locate in the chicken genome 72 and 328 exons from our original alternative and constitutive training sets, respectively. Of the 72 alternatively spliced exons, 34 (47%) had conserved sequences in both their upstream and downstream introns when human and chicken genomes were compared; only 10 (3%) of the 328 constitutively spliced exons that could be found in the chicken genome had such intronic conservation (data not shown).

Currently, our classifier mainly identifies exon-skipping events in exons conserved between human and mouse. In the future, it could develop into a more general alternative splicing predictor that would identify other types of alternative splicing. The ultimate goal of such a predictor would be genome-based prediction of all splice variants, including their pattern of alternative splicing (i.e., in which tissue would the exon be inserted). This could set the foundations for understanding the absolute number of exons that are alternatively spliced and might ultimately lead to narrowing the gap between the genome and the proteome, and thereby advance towards revealing the full extent of our proteome's complexity.

Methods

Enumeration over features in training set

Training sets of alternatively spliced internal exons and constitutively spliced internal exons were taken from our previous study (Sorek and Ast 2003). For the current analysis we eliminated from our constitutive exons' set, exons for which novel evidence

for alternative splicing appeared in the newer version of GenBank, 136. This left us with 1,753 constitutive exons.

The thresholds used in the enumeration of classification rules were as follows: Exon identity thresholds were 100%, at least 99%, at least 98%, and so forth until 80%; exon lengths were below 18 bp, 23 bp, 28 bp, …, 198 bp and 1000bp; length of human/mouse local alignment of the 100 nearest upstream (or downstream) intronic nucleotides using sim4 (Florea et al. 1998) was at least 100, 95, 90, ,…, 0; minimum identity level in the locally aligned segment of the upstream (or downstream) region was 100%, 97%, 94%,…, 67%; exon divisibility by 3 had two categories, 'yes' or 'no'. Overall we enumerated more than 100 million different combinations of features.

Genome-wide retrieval of human and mouse orthologous exons

For the genome-wide compilation of human exons, human ESTs and cDNAs were obtained from NCBI GenBank version 136 (June 2003) (www.ncbi.nlm.nih.gov/dbEST) and were mapped to the human genome (April 2003 assembly) (www.ncbi.nlm.nih.gov/genome/guide/human) using the spliced alignment module described in (Sorek et al. 2002; Sorek and Safer 2003). For each expressed sequence, all mappings of internal exons on the human genome were retrieved. Only exons flanked by AG/GT or AG/GC splice sites were allowed. 185,799 human exons mapped to the human genome were thus retrieved.

To find the mouse orthologue for each human exon, we first aligned the mouse expressed sequences from GenBank version 136 to the human genome, as described in (Sorek and Ast 2003). Mouse sequences exactly spanning human exons were aligned to

the mouse genome as well, and the corresponding sequence on the mouse genome was declared as the orthologous mouse exon, if it was flanked by AG/GT or AG/GC legal splice sites.

Human exons for which no spanning mouse expressed sequence was detected were aligned directly to the mouse genome. Hits spanning the full length of the exon, that were flanked by AG/GT or AG/GC legal splice sites, were declared as the orthologous mouse exons.

Altogether, these searches retrieved 108,983 pairs of exons in the human and mouse genomes (this set does not contain the exons from our two training sets). For each such exon, all classifying parameters were calculated as follows. Conservation between exons was calculated from aligning the human and mouse exons using the global alignment program "GAP" of the GCG software package with default parameters. Conservation in the flanking intronic sequences was calculated by sim4 as described in (Sorek and Ast 2003). Sim4 detects exact matches of length 12 and extends them in both directions with a score of 1 for a match and 5 for a mismatch, stopping when extensions no longer increase the score (Florea et al. 1998). Exon size and divisibility by 3 were retrieved from the exon sequence itself.

Reverse transcription of mRNA samples

cDNA was obtained by reverse transcription of total RNA from the following human tissue samples: 1. Brain pool – a pool of brain derived RNA samples (Biochain – Normal). 2. Prostate pool – a pool of prostate derived RNA samples (Biochain –

Normal). 3. Testis pool – a pool of testis derived RNA samples (Biochain – Normal). 4. Kidney pool – a pool of kidney derived RNA samples (Biochain – Normal). 5. Thyroid pool – a pool of thyroid derived RNA samples (Biochain – Normal). 6. Assorted cell-line pool – a pool of cell-line derived RNA samples from the cell-lines: DLD, MiaPaCa, HT29, THP1, MCF7 (ATCC). 7. Cervix pool – a pool of 3 cervix derived RNA samples, mixed origin (Tumor and Normal, in-house tissue samples). 8. Uterus pool - a pool of 3 uterus derived RNA samples (Biochain – Normal), mixed origin (Tumor and Normal). 9. Ovary pool - A pool of 5 ovary derived RNA samples (Biochain – Normal), combined with two samples of mixed origin (Tumor and Normal). 10. Placenta – one sample of placenta derived RNA (Biochain – Normal). 11. Breast pool – a pool of 3 breast derived RNA samples of mixed origin (2 from tumor and one from normal - in-house tissue samples). 12. Colon and intestine – A pool of 5 colon derived RNA of mixed origin (Tumor and Normal), combined with one intestine (Normal) derived RNA sample (in-house tissue samples). 13. Pancreas – one sample of pancreas derived RNA (Biochain – Normal). 14. Liver and spleen – one sample of liver derived RNA (Biochain – Normal), one sample of spleen derived RNA (Biochain – Normal), combined with one sample of HepG2 cell line (liver tumor - ATCC) derived RNA.

RNA was incubated with a random hexamer primer mix (Invitrogen), denaturated at 70^0C for 5 minutes and transferred to 4^oC for hexamer annealing. Reverse transcription was done by Superscript II Reverse transcriptase (Invitrogen), in the presence of RNAsin™ (Promega) at 37^0C for 1 hour. Reaction was terminated by enzyme deactivation on beads (Promega).

Amplification of splicing products

For each exon tested, oligonucleotide primers were designed from its flanking exons (Supplementary Table 1). Amplification was performed for 35 cycles, consisting of 94^0C for 45 sec, annealing at a primer specific temperature (4^0C below the primer's TM) for 45 sec, and extension at 72^0C for 1 min. The cycle was ended by one stage of gap filling at 72^0C for 10 min's. The products were resolved on 2% agarose gel and confirmed by sequencing.

Acknowledgments

The authors thank Amos Tanay, Irit Gat-Viks, and Gideon Dror for fruitful discussion; and Kinneret Savitsky, Dvir Dahary and Pini Akiva for critical reading.

References

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270:** 2411-2414.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474:** 83-86.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78-94.

Cartegni, L., Chew, S.L., and Krainer, A.R. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* **3:** 285-298.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V. et al. 2003. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31:** 38-42.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967-974.

Graveley, B.R. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17:** 100-107.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. et al. 2003.

Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31:** 5654-5666.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:** 2141-2144.

Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* **11:** 889-900.

Kan, Z., States, D., and Gish, W. 2002. Selecting for functional alternative splices in ESTs. *Genome Res.* **12:** 1837-1845.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, C. Zody, M.C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Maniatis, T. and Tasic, B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236-243.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. *Genome Res.* **9:** 1288-1293.

Modrek, B. and Lee, C. 2002. A genomic view of alternative splicing. *Nat. Genet.* **30:** 13-19.

Modrek, B. and Lee, C.J. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34:** 177-180.

Modrek, B., Resch, A., Grasso, C., and Lee, C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29:** 2850-2859.

Nurtdinov, R.N., Artamonova, II, Mironov, A.A., and Gelfand, M.S. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* **12:** 1313-1320.

Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., and Lee, C. 2004. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* **32:** 1261-1269.

Sorek, R., Ast, G., and Graur, D. 2002. Alu-containing exons are alternatively spliced. *Genome Res.* **12:** 1060-1067.

Sorek, R. and Ast, G. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13:** 1631-1637.

Sorek, R. and Safer, H.M. 2003. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* **31:** 1067-1074.

Sorek, R., Shamir, R., and Ast, G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20:** 68-71.

Thanaraj, T.A. and Stamm, S. 2003. Prediction and statistical analysis of alternatively spliced exons. *Prog. Mol. Subcell. Biol.* **31:** 1-31.

Legends to figures

Figure 1: Graphic representation of the differences between alternative and constitutive exons. For each of the following curves, constitutive exons are in squares, and alternatives are in diamond shapes. A) Length of conserved region in the nearest 100 nucleotides of the flanking upstream intron. X axis, length of conserved region (best sim4 local alignment); Y axis, percent exons with upstream conserved region greater than or equal to the value in X. Conservation was detected using local alignment with the mouse 100 counterpart intronic nucleotides. A minimum hit was 12 consecutive perfectly matching nucleotides. B) Length of conserved region in the nearest 100 nucleotides of the flanking downstream intron. Axes as in A. C) Exon size distribution. X axis, exon size; Y axis, percent exons having size lesser or equal to the size in X. D) Human-mouse exon identity. X axis, percent identity in the global alignment of the human and the mouse exons; Y axis, percent exons with identity greater or equal to the value in X. E) Human-mouse exon identity, for exons whose size is a multiple of 3. Axes as in D. Note that by combining two features we get better separation of the two exon-types.

Figure 2: Experimental validation for the existence of alternative splicing in selected predicted exons. RT-PCR for 15 exons (detailed in Table 2), for which no EST/cDNA indicating alternative splicing was found, was conducted over 14 different tissue types and cell lines (see Methods). Detected splice variants were confirmed by sequencing. For nine of these exons a splice isoform was detected in at least one of the tissues tested. Only a single tissue is shown here for each of these nine exons. Lane 1,

DNA size marker. Lane 2, exon 2 skipping in FGF11 in ovary tissue (the 344nt and 233nt products are exon inclusion and skipping, respectively). Lane 3, exon 4 skipping in EFNA5 gene in ovary tissue (exon inclusion 287nt; skipping 199nt). Lane 4, exon 8 skipping in NCOA1 gene in placenta tissue (exon inclusion 377nt; skipping 275nt). Lane 5, exon 22 skipping in PAM gene in cervix tissue (exon inclusion 323nt; skipping 215nt). Additional upper band contains a novel exon in PAM. Lane 6, exon 9 skipping in GOLGA4 gene in uterus tissue (exon inclusion 288nt; skipping 213nt). Lane 7, exon 9 skipping of NPR2 gene in placenta tissue (282nt inclusion; 207nt skipping). Lane 8, intron 8 retention in VLDLR gene in ovary tissue (wild type 324nt; intron retention 427nt). Lane 9, alternative acceptor site in exon 12 of BAZ1A in ovary tissue (wild type 351nt; alternative acceptor variant 265nt). The uppermost band represents a new exon in BAZ1A, inserted between exons 12 and 13. Lane 10, alternative acceptor site in exon 7 of SMARCD1 in uterus tissue (wild type 353nt; exon 7 extension 397nt).

Figure 3: Sensitivity vs. false positive rate in classification rules. Each square on the curve represents the performance of a single classification rule. X axis, 1-specificity, i.e., percent constitutive exons (false positives) retrieved by the rule. Y axis, sensitivity, i.e., percent alternative exons (true positives) identified by the rule. Values were computed relative to the training set. Rules that were used for this plot are provided as Supplementary Material.

Table 1: Features differentiating between alternatively spliced and constitutively spliced exons

	Alternatively spliced exons	**Constitutively spliced exons**	**P value[a]**
Average size	87	128	$p < 10^{-16}$
Percent exons whose length is a multiple of 3	73% (177/243)	37% (642/1753)	$p < 10^{-9}$
Average human-mouse exon conservation [b]	94%	89%	$p < 10^{-36}$
Percent exons with upstream intronic elements conserved in mouse [c]	92% (223/243)	45% (788/1753)	$p < 10^{-11}$
Percent exons with downstream intronic elements conserved in mouse [c]	82% (199/243)	35% (611/1753)	$p < 10^{-14}$
Percent exons with both upstream and downstream intronic elements conserved in mouse [c]	77% (188/243)	17% (292/1753)	$p < 10^{-37}$

[a] P-value was calculated using Fisher's exact test, except for the "average size" and "average human-mouse exon conservation", for which p-value was calculated using student's T test.

[b] Average percent of matching nucleotides in global alignment of the respective exons

[c] The 100 intronic nucleotides immediately upstream (or downstream) of the exon were locally aligned with the mouse 100 counterpart intronic nucleotides using sim4 (Florea et al. 1998). Conservation was defined if at least 12 consecutive perfectly matching nucleotides were found in the alignment.

Table 2: Experimental validation of predicted alternatively spliced exons

Gene	Alt Exon[a]	PCR confirmed[b]	Type of alternative confirmed[c]	Gene Description
FGF11	2	Yes	Skip	fibroblast growth factor 11
EFNA5	4	Yes	Skip	ephrin-A5
NCOA1	8	Yes	Skip	steroid nuclear receptor coactivator
PAM	22	Yes	Skip	protein associated with Myc mRNA
GOLGA4	9	Yes	Skip	golgi autoantigen, golgin subfamily a, 4
NPR2	9	Yes	Skip	natriuretic peptide receptor B/guanylate cyclase B
VLDLR	9	Yes	Int Ret [d]	very low density lipoprotein receptor
BAZ1A	12	Yes	Alt 3'ss [e]	bromodomain adjacent to zinc finger domain protein 1A
SMARCD1	7	Yes	Alt 3'ss [f]	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1
PRKCM	15	No		protein kinase C, mu
TIAM2	12	No		T-cell lymphoma invasion and metastasis 2
MDA5	4	No		melanoma differentiation associated protein-5
RNASE3L	15	No		nuclear RNase III
HAT1	7	No		histone acetyltransferase 1
DICER1	6	No		Dicer1, Dcr-1 homolog (Drosophila)

[a] Serial number of exon (out of gene's exons) identified as alternative

[b] For each predicted exons, primers were designed from its flanking exons and RT_PCR was conducted using total RNA from 14 different tissue types: cervix, uterus, ovary, placenta, breast, colon, pancreas, liver + spleen, brain, prostate, testis, kidney, thyroid, and assorted cell-lines. Products were sequenced, and alternative splicing was searched.

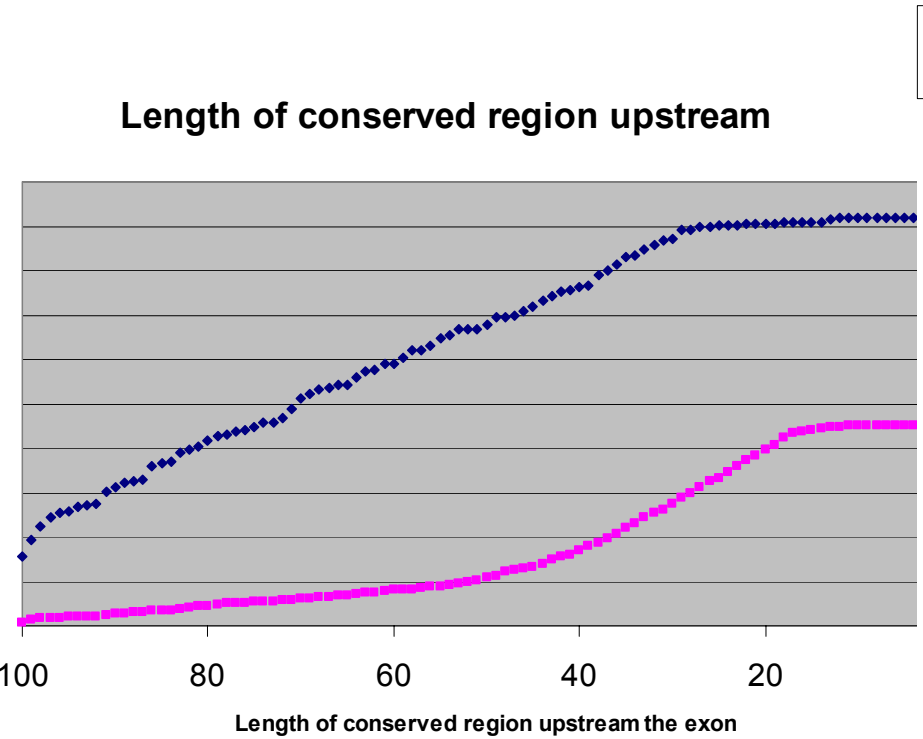
[c] Type of alternative splicing: Skip, exon-skipping; Alt 3'ss, alternative 3' splice site (acceptor); Int Ret., intron retention.

[d] Retention of intron 8 (size 103 nucleotides) was detected in VLDLR.

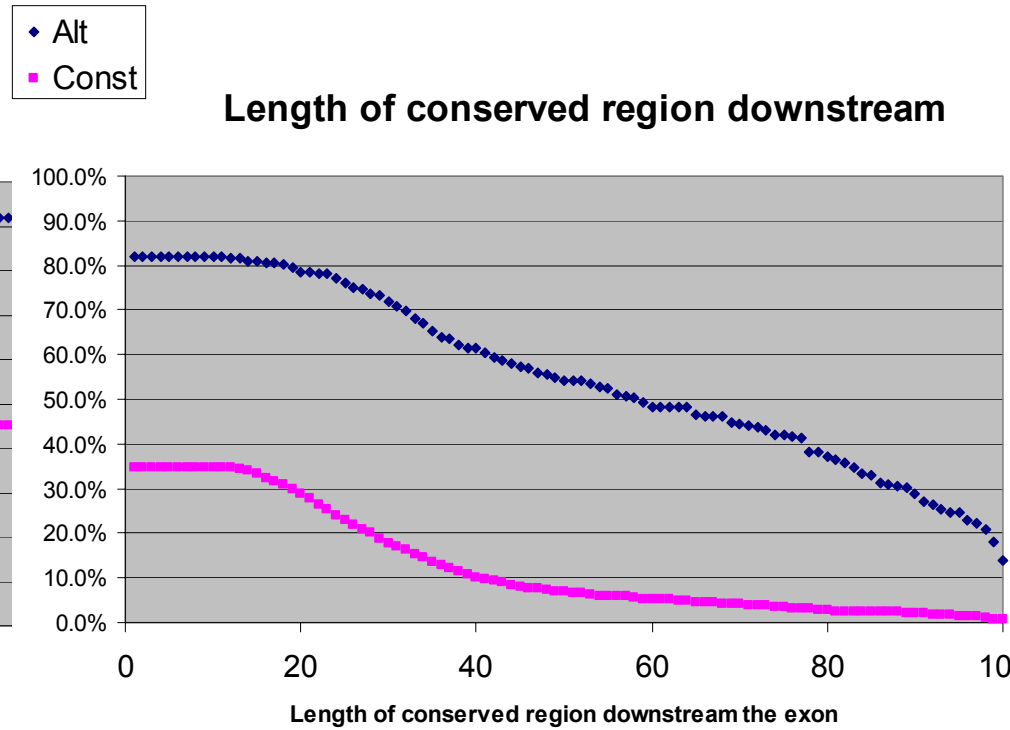
[e] Deletion of 86 nucleotides was detected on the 3' end of exon 12 7 of BAZ1A.

[f] Extension of 44 nucleotides was detected on the 3' end of exon 12 of SMARCD1.

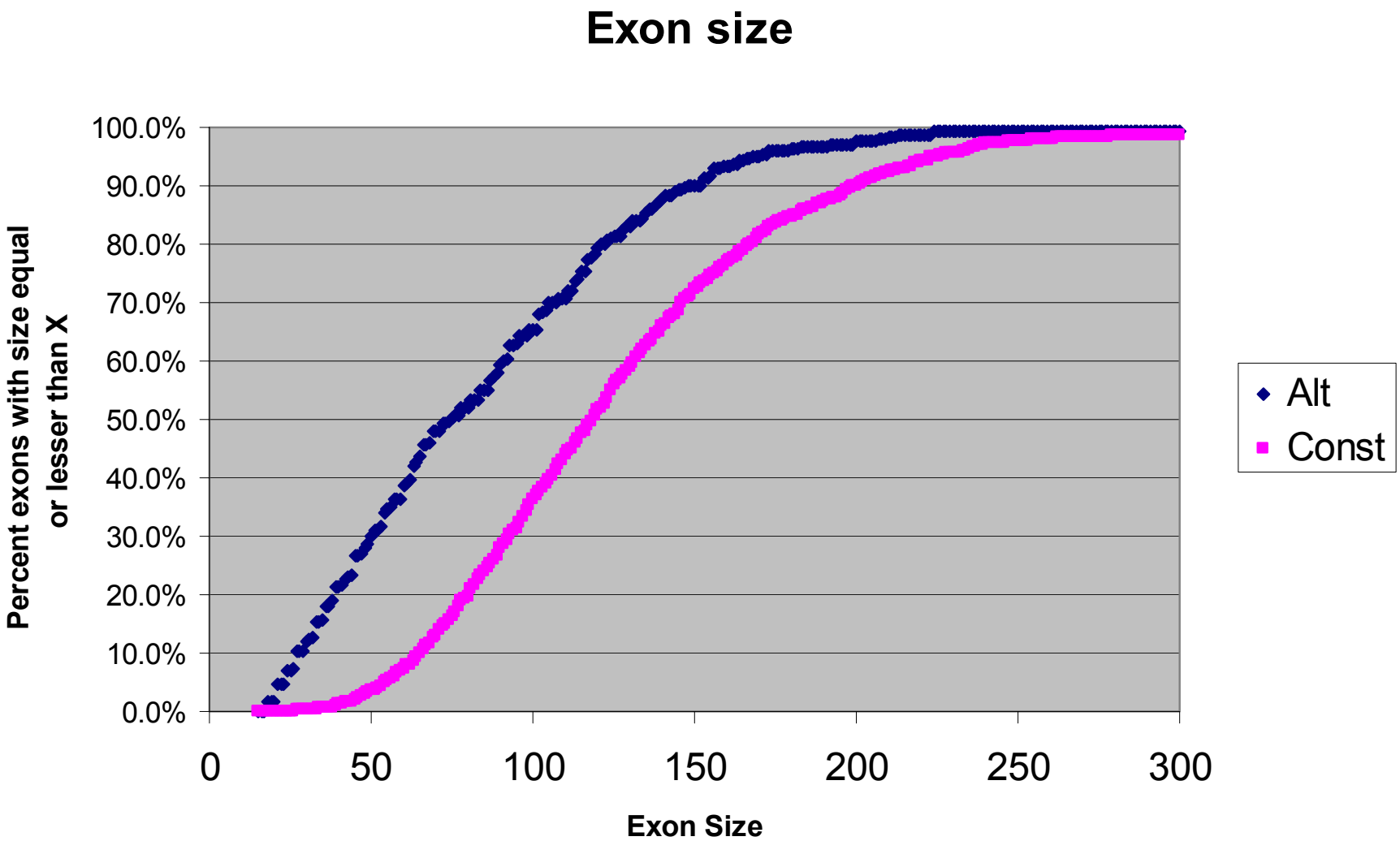
Sorek Figure 1a



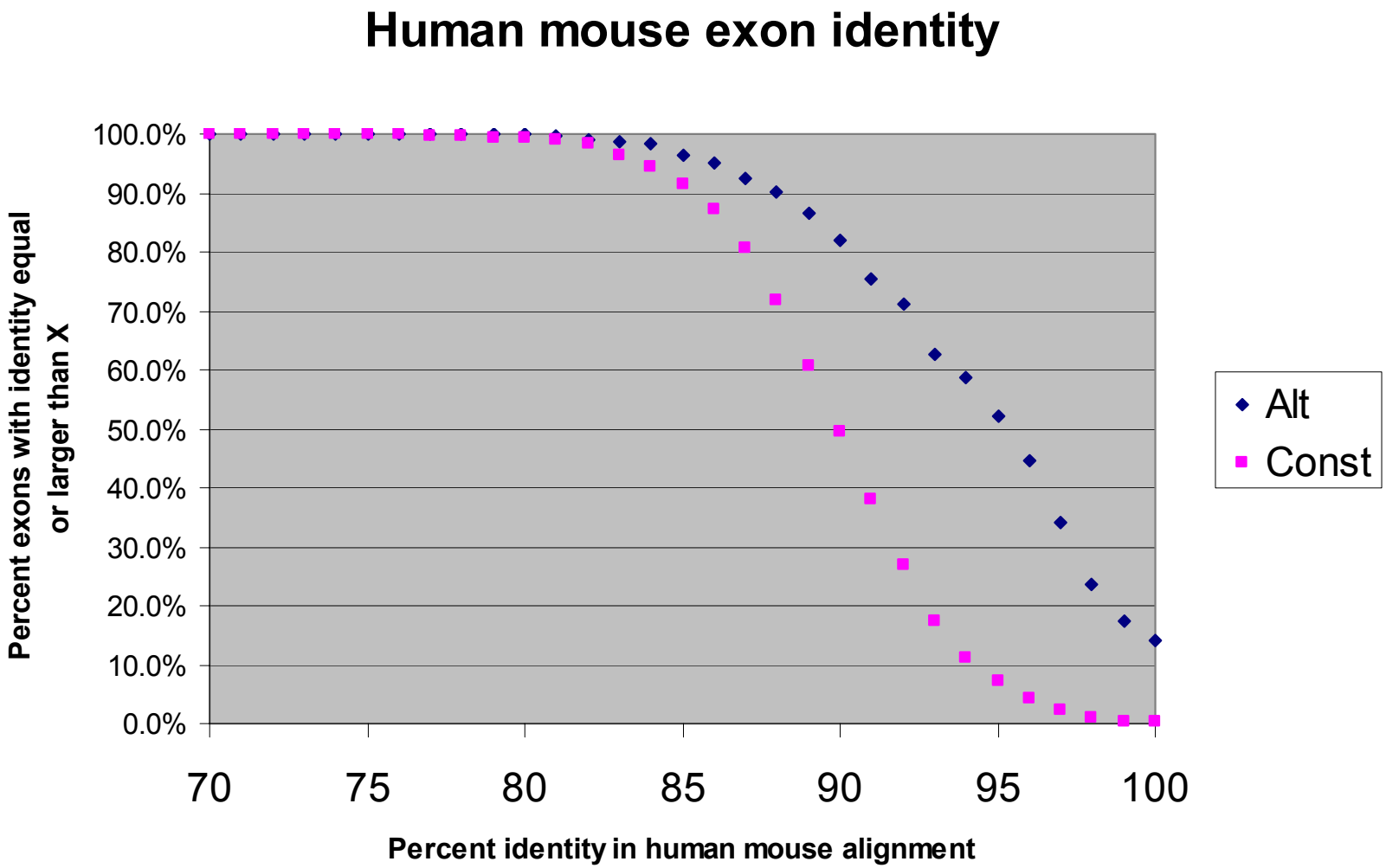
Sorek Figure 1b



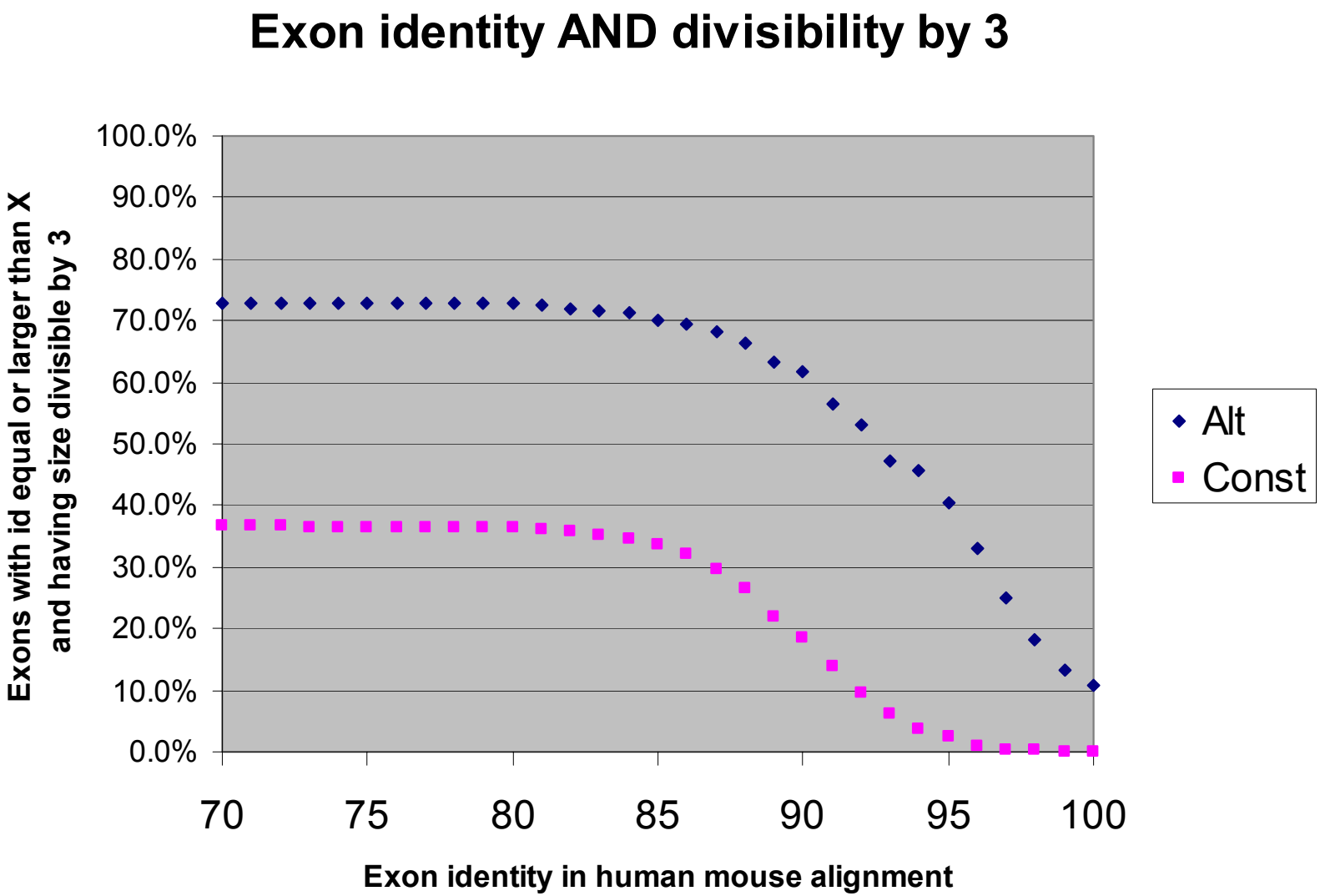
Sorek Figure 1c



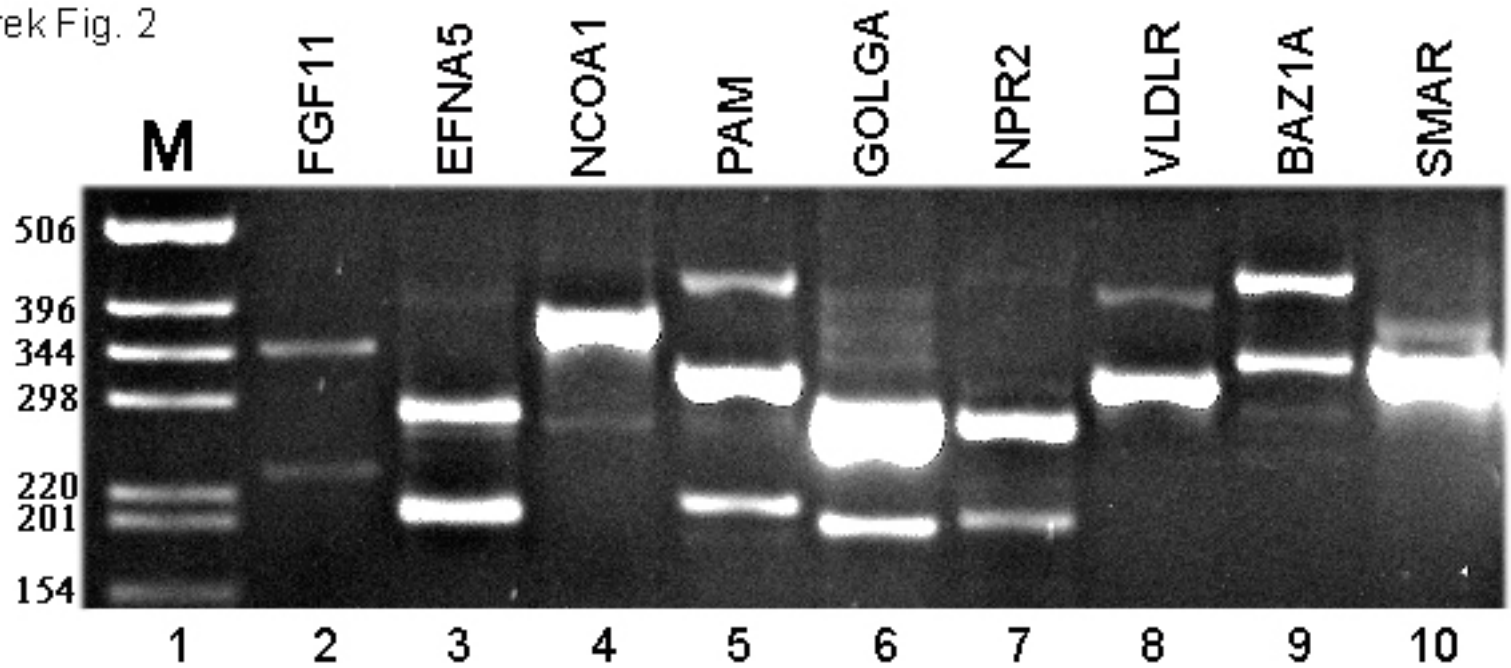
Sorek Figure 1d



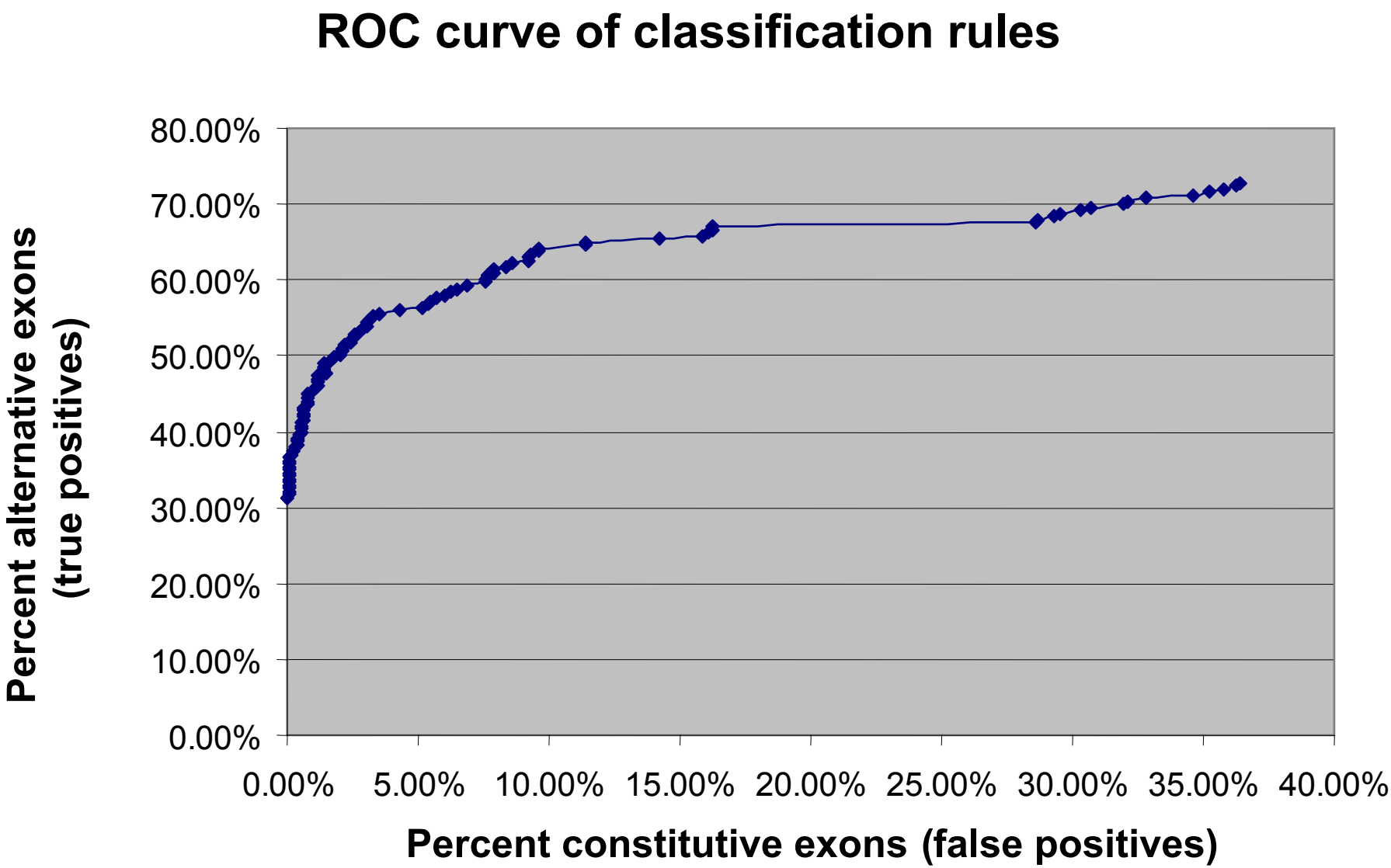
Sorek Figure 1e



Sorek Fig. 2


M
FGF11
EFNA5
NCOA1
PAM
GOLGA
NPR2
VLDLR
BAZ1A
SMAR
506
396
344
298
220
201
154
1
2
3
4
5
6
7
8
9
10

Sorek Figure 3



Supplementary Table 1: Primers used for validation of alternative exons.

Gene and direction	Primer sequences	TM
FGF11 Forward	5' – CCAAGGTGCGACTGTGCGG – 3'	68^0C
FGF11 Reverse	5' – GGTAGAGAGCAGAGGCGTACAGGACG – 3'	66^0C
EFNA5 Forward	5' – ACCGGCCTCACTCTCCAAATGG – 3'	65^0C
EFNA5 Reverse	5' – TGGCTCGGCTGACTCATGTACGG – 3'	67^0C
NCOA1 Forward	5' – AGGCAACACGACGAAATAGCCATACC – 3'	66^0C
NCOA1 Reverse	5' – TCTGGCATAAGATGGTTCTCTGCCC – 3'	65^0C
PAM Forward	5' – TGTCCCAGTGCCCGGG – 3'	61^0C
PAM Reverse	5' – GGTGAAATCCACAGCTGACTTGG – 3'	62^0C
GOLGA4Forward	5' – TCAAGAGAACCTACTTAAGCGTTGTAAGG – 3'	61^0C
GOLGA4Reverse	5' – TGAGCAATTTCTTCTTCTTTCATTTCC – 3'	61^0C
NPR2 Forward	5' – CATGTTTGGTGTTTCCAGCTTCC – 3'	62^0C
NPR2 Reverse	5' – CGGGTCAGCTCAATGCGC – 3'	62^0C
VLDLR Forward	5' – TGAGCCCCTGAAAGAGTGTCATATAAACG – 3'	66^0C
VLDLR Reverse	5' – TCTAAGCCAATCTTCCTGATGTCTCTTCG – 3'	66^0C
BAZ1A Forward	5' – TGCTCTGATGGTTTTGGAGTTCC – 3'	61^0C
BAZ1A Reverse	5' – CGTTTTTGATATCTATACTTTGCATTTGC – 3'	60^0C
SMARCD1Forward	5' – CAGCCTTGTCCAAATATGATGCC – 3'	61^0C
SMARCD1Reverse	5' – AAACTCCCGCTCGTGAGGG – 3'	61^0C
DICER1 Forward	5' – AACTCATTCAGATCTCAAGGTTGGG – 3'	61^0C
DICER1 Reverse	5' – CCAGGTCAGTTGCAGTTTCAGC – 3'	61^0C
HATB Forward	5' – AGGCTTCAGACCTTTTTGATGTGG – 3'	62^0C

HATB Reverse	5’ – CTTCCGCTGTAATATCAAGAACTGTAGG – 3’	61^0C
PRKCM Forward	5’ – AAGTACTGGGTTCTGGACAGTTTGG – 3’	61^0C
PRKCM Reverse	5’ – CTGGTTTGAGGTCACAGTGAACG – 3’	61^0C
RNASE3L Forward	5’ – CGGAGAATTTTTGTGTGAAAGGG – 3’	61^0C
RNASE3L Reverse	5’ – CCAGCTCCTCCCACTGAAGC – 3’	61^0C
TIAM2 Forward	5’ – AACGACAGTCAGGCCAACGG – 3’	62^0C
TIAM2 Reverse	5’ – CCAGAAACACCTTCTGAAACTCAAGC – 3’	62^0C
MDA5 Forward	5’ – AAATCTGGAGAAGGAGGTCTGGG – 3’	61^0C
MDA5 Reverse	5’ – CCACTCTGGTTTTTCCACTCCC – 3’	61^0C

Cross validation (supplementary material for Sorek et al)

Five-way cross validation: The data were partitioned to five equal size sets; each set in turn was put aside as the test set, and the remaining 80% of the data were used as training set to determine the best rule, that was then applied on the test set. The following table describes the 5 rules obtained in each of the cross validation sets. For comparison we include also the rule obtained on the complete training set.

Rule	Ex ID	Upst Len	Upst ID	Downst Len	Downst ID	Div 3
1	95	15	85	12	>0	Yes
2	95	15	82	12	>0	Yes
3	95	15	85	12	>0	Yes
4	94	15	85	12	>0	Yes
5	95	15	85	12	>0	Yes
Full set	95	15	85	12	>0	Yes

Cross validation sensitivity and false positive rate

Rule	Training set		Test set		Sensitivity	False Positive Rate
	Alternative	Constitutive	Alternative	Constitutive		
1	65/195	0/1403	11/48	0/350	33.3%	0%
2	67/194	0/1402	17/49	2/351	34.5%	0.569%
3	56/194	0/1402	20/49	0/350	28.8%	0%
4	64/194	0/1402	20/49	3/351	32.9%	0.854%
5	62/195	0/1402	14/48	0/351	31.8%	0%
Full set	76/243	0/1753			31.3%	0%

Average sensitivity: 32.26%
Average specificity: 0.28%