BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 23 2012, pages 3042–3050
doi:10.1093/bioinformatics/bts582

Sequence analysis

Advance Access publication October 7, 2012

Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes

Dennis K. Gascoigne[1], Seth W. Cheetham[1,2], Pierre B. Cattenoz[1], Michael B. Clark[1], Paulo P. Amaral[1,‡], Ryan J. Taft[1], Dagmar Wilhelm[1], Marcel E. Dinger[1,2,3,*,†] and John S. Mattick[1,*,†,§]

[1]Institute for Molecular Bioscience, The University of Queensland, St Lucia, Brisbane, Queensland 4072, Australia, [2]The University of Queensland Diamantina Institute, Princess Alexandra Hospital, Woolloongabba, Brisbane, Queensland 4102, Australia and [3]the Garvan Institute for Medical Research, 384 Victoria Street, Darlinghurst NSW 2010, Australia

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Comparing transcriptomic data with proteomic data to identify protein-coding sequences is a long-standing challenge in molecular biology, one that is exacerbated by the increasing size of high-throughput datasets. To address this challenge, and thereby to improve the quality of genome annotation and understanding of genome biology, we have developed an integrated suite of programs, called *Pinstripe*. We demonstrate its application, utility and discovery power using transcriptomic and proteomic data from publicly available datasets.

Results: To demonstrate the efficacy of Pinstripe for large-scale analysis, we applied Pinstripe's reverse peptide mapping pipeline to a transcript library including *de novo* assembled transcriptomes from the human Illumina Body Atlas (IBA2) and GENCODE v10 gene annotations, and the EBI Proteomics Identifications Database (PRIDE) peptide database. This analysis identified 736 canonical open reading frames (ORFs) supported by three or more PRIDE peptide fragments that are positioned outside any known coding DNA sequence (CDS). Because of the unfiltered nature of the PRIDE database and high probability of false discovery, we further refined this list using independent evidence for translation, including the presence of a Kozak sequence or functional domains, synonymous/non-synonymous substitution ratios and ORF length. Using this integrative approach, we observed evidence of translation from a previously unknown *let7e* primary transcript, the archetypical lncRNA *H19*, and a homolog of RD3. Reciprocally, by exclusion of transcripts with mapped peptides or significant ORFs (>80 codon), we identify 32 187 loci with RNAs longer than 2000 nt that are unlikely to encode proteins.

Availability and implementation: Pinstripe (pinstripe.matticklab.com) is freely available as source code or a Mono binary. Pinstripe is written in C# and runs under the Mono framework on Linux or Mac OS X, and both under Mono and .Net under Windows.

Contact: m.dinger@garvan.org.au or j.mattick@garvan.org.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 13, 2012; revised on August 30, 2012; accepted on September 21, 2012

1 INTRODUCTION

Although the annotation of the human genome has undergone considerable effort, with >20 000 protein-coding loci identified, the extent to which the genome is transcribed and subsequently translated remains unclear (Clark *et al*., 2011; Flicek *et al*., 2010). As well as mRNAs, long RNAs with little or no protein-coding capacity (long non-coding RNAs or lncRNAs) are abundantly expressed from the human genome with >10 000 non-coding loci currently in GENCODE (the most comprehensive annotation of the human genome). Adding to the known catalog of RNAs, transcriptomic sequencing experiments are identifying thousands of new transcripts on an ongoing basis (Trapnell *et al*., 2010).

Confidently determining which RNAs encode proteins is difficult. Traditional approaches classify the protein-coding potential of an RNA based on the presence of long canonical open reading frames (ORFs), phylogenomic evidence of codon conservation (Dinger *et al*., 2008) and/or recognized functional domains. The former approach is highly dependent on ORF length, with ∼100 codons typically used as the threshold for the identification of mRNAs (Carninci *et al*., 2005; Imanishi *et al*., 2004). However, this approach results in an unknown false negative and positive rate, as shorter proteins are known to exist, and short ORFs can occur by chance (Dinger *et al*., 2011). Phylogenomic comparisons also have limitations, notably their inability to detect orphan genes lacking homologs in other lineages (Tautz and Domazet-Lošo, 2011), and the tendency for recently evolved genes to encode smaller proteins than older genes (Lipman *et al*., 2002). Moreover, data from which evidence of ORF and codon conservation can be determined are limited, especially at short evolutionary distances. Annotation of protein-coding genes is further confounded by factors such as RNA editing,

*To whom correspondence should be addressed.
[†]These authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.
[‡]Present address: Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge CB2 1QN, UK.
[§]Present address: Garvan Institute of Medical Research, Darlinghurst, Sydney, New South Wales 2010, Australia.

genotypic variation, non-canonical start codons and ribosomal frameshifting.

Although many of the transcripts identified in *de novo* assembled transcriptomes can be annotated through comparison with known reference databases such as GENCODE (Flicek *et al.*, 2010) or RefSeq (Pruitt *et al.*, 2007), this approach does not provide a means for classifying novel transcripts and is entirely dependent on the accuracy of the reference database. Even determining the protein-coding capacity of novel transcript isoforms of known genes is not straightforward, as the insertion, deletion and extension of exons may disrupt the existing ORF or establish an alternative ORF. Identifying the consequences of such events to the encoded protein sequence on a genomic scale is challenging and not addressed by existing tools. Although the comparison of transcriptomic datasets to identify overlapping/non-overlapping features is facilitated by applications such as BEDTools (Quinlan and Hall, 2010), such applications do not consider the frame of the CDS.

The classification of novel transcripts with a putative ORF can be discerned with greater confidence using proteomic data. Mapping of proteomic sequences to transcriptomic data is typically performed using software such as BLAST (Altschul *et al.*, 1990) or BLAT (Kent *et al.*, 2002). Use of these programs with high-throughput transcriptomic datasets requires considerable pre- and post-processing. Custom nucleotide sequence dictionaries must be created from spliced genomic coordinates before BLAST comparisons, and resulting matches are reported with respect to mRNA coordinates that require transformation into genomic coordinates. After mapping, each ORF then needs to be analyzed to determine the extent of supporting proteomic data. Currently, there is no software to facilitate this process, although high-throughput proteomic–genomic comparisons in Arabidopsis (Castellana *et al.*, 2008) and proteomic–transcriptomic integration in mice (Brosch *et al.*, 2011) have been undertaken.

Here, we present an integrated software suite, *Pinstripe*, which overcomes many of the difficulties in such tasks, providing programs (Supplementary Table S1) to analyze transcriptomes, specifically enabling: (i) ORF-aware comparison with protein-coding gene annotations identifying CDSs with in-frame overlap, (ii) mapping of proteomic data to the transcriptome (leveraging BLASTP and TBLASTN) and reporting results as genomic coordinates and (iii) annotation of ORFs supported by peptides from proteomic datasets. In addition, *Pinstripe* features a multitude of tools for analyzing transcriptomes, including intersection by feature name, identification of best representative transcripts for loci and extraction of DNA sequence or predicted ORFs from genomic coordinates. We demonstrate the utility of *Pinstripe* by analyzing transcriptomes assembled *de novo* from high-throughput RNA sequencing of 16 human tissues and mass spectrometry from the EMBL-EBI Proteomics Identifications Database (PRIDE) (Vizcaíno *et al.*, 2009), finding previously unidentified protein-coding regions and providing high-confidence annotations for thousands of long ncRNAs.

2 METHODS

Pinstripe is a multiprocessor-enabled application designed for the rapid high-throughput analysis of *de novo* or reference-guided transcriptome assemblies generated by tools such as *Tophat* (Trapnell *et al.*, 2009) and *Cufflinks* (Trapnell *et al.*, 2010) or Trinity (Grabherr *et al.*, 2011). In terms of performance, *Pinstripe* can identify ORFs, the corresponding amino acid sequence and level of support in ~5 min using a single-core 2.27 GHz Intel Xeon processor for 1 million transcripts queried against 500 000 pre-mapped peptide sequences (pepFrags). Mapped pepFrag files require <80 Mb per 1 million pepFrags. The application is programed in C#, and the compiled binary runs under Mono on Linux or MacOS X, and both Mono and .NET under Windows. Here, we demonstrate its implementation using transcriptomes assembled *de novo* from high-throughput RNA sequencing of 16 human tissues and mass spectrometry peptide identifications from public databases.

Pinstripe uses the UCSC Browser Extensible Data (BED/BED12) format to describe the chromosomal positions of input transcripts. This format is compact, storing information regarding a single transcript's exons and CDS in a single row of data. Annotation reference files for most transcript and gene collections are readily available in BED12 format from the UCSC Table Browser. For compatibility with *Pinstripe*, Sequence Alignment/Map (SAM) files can be converted to, and compared with, BED files using programs such as BEDTools. Gene Transfer Format (GTF) files can be converted to BED format using a combination of tools available from UCSC, and basic command line utilities (refer Supplementary Material).

2.1 Identification of mRNA

mRNA are most easily identified by comparing transcripts with a database of validated gene annotations such as RefSeq (Pruitt *et al.*, 2007) or GENCODE (Becker, 2011). The difficulty with this approach is that *de novo* transcriptomes, such as those assembled from high-throughput RNA sequencing experiments, can reveal both new transcripts that do not correspond to any annotation, and variants of known transcripts with additional exons, alternate splice sites or alternative UTRs, all of which may result in ORF disruption. Any comparison involving a novel transcript isoform requires consideration of the frame in which potential CDSs are translated. These problems are exacerbated for poorly annotated genomes or where a reference is not available.

To address this challenge, *Pinstripe* provides two methods for identifying mRNA in CDS-naïve transcripts. First, *Pinstripe annotate* compares each transcript with a gene reference, annotating a CDS if the three-frame conceptual translation of the query transcript yields an ORF coincident (within tolerances defined by user options) with the reference, and in the same frame.

Alternatively, *Pinstripe mapPeptides* takes peptide fragments (pepFrags) from mass spectrometry (or other proteomic data) and maps them to the transcriptome (Fig. 1). *mapPeptides* creates a BLAST (Camacho *et al.*, 2009) dictionary from the queried transcript library and maps the pepFrag amino acid sequences against the dictionary with user-definable parameters. The resultant BLAST matches are then translated from local mRNA coordinates to genomic coordinates (including those pepFrags mapping across splice junctions) and post-processed to remove redundancy (where pepFrags map to different mRNAs but resolve to the same genomic position). The original transcript library is matched to any overlapping pepFrags and potential ORFs (canonical and non-canonical) identified through *in silico* translation using *Pinstripe buildProteins*. Each transcript is reported with details of which pepFrags map to it, the amino acid sequence of the ORF and the ORF length. An example workflow is shown in Figure 2. Workflows for the main *Pinstripe* programs, *mapPeptides*, *buildProteins* and *annotate* are shown in Supplementary Figure S1.

Examples of *Pinstripe* usage to annotate ORFs from a CDS-naïve transcriptome are as follows:

(1) Map peptide fragments to a transcriptome: pstr mapPeptides [options] hg19.fa transcriptome.bed pepFrags.fa > pepMap_results.bedx.

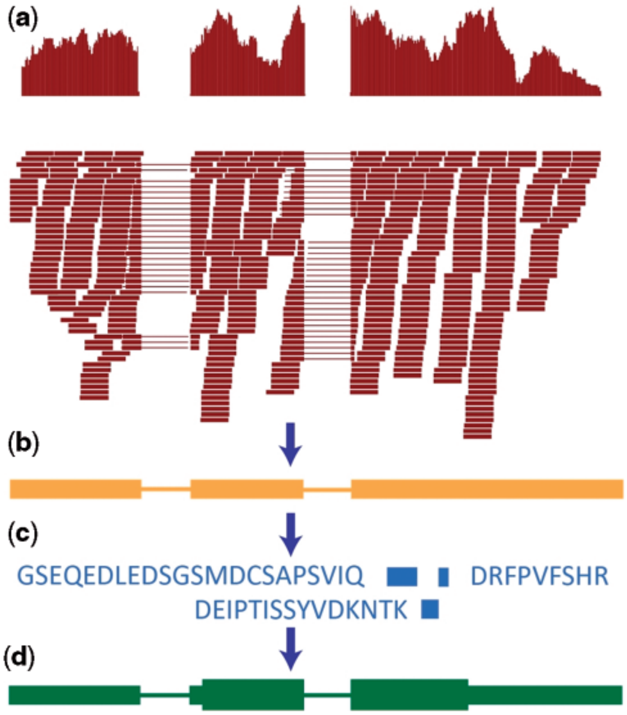

Fig. 1. Overview of the reverse peptide mapping methodology. (**a**) RNA sequencing tags from heart tissue (red) are mapped to a reference genome. (**b**) Mapped tags are assembled into a transcript. (**c**) pepFrags are mapped to the same 198-codon ORF (RD3LP shown) using *Pinstripe mapPeptides*. (**d**) Full ORF is identified and annotated using *Pinstripe buildProteins*

(2) Optionally remove pepFrags that are contained within another, or map to multiple locations: pstr deDup –exEncomp –exMisSplice –exMultiMap pepMap_results.bedx > pepMap_results_uq.bedx.

(3) Identify ORFs containing pepFrags, recording number of pepFrags in each ORF and the conceptual translation: pstr buildProteins [options] hg19.fa pepMap_results_uq.bedx transcriptome.bed > protein_results.bedx.

(4) Annotate ORFs using a reference: pstr annotate [options] hg19.fa Gencodev10Ref.bed transcriptome.bed > transcriptome_annotated.bed.

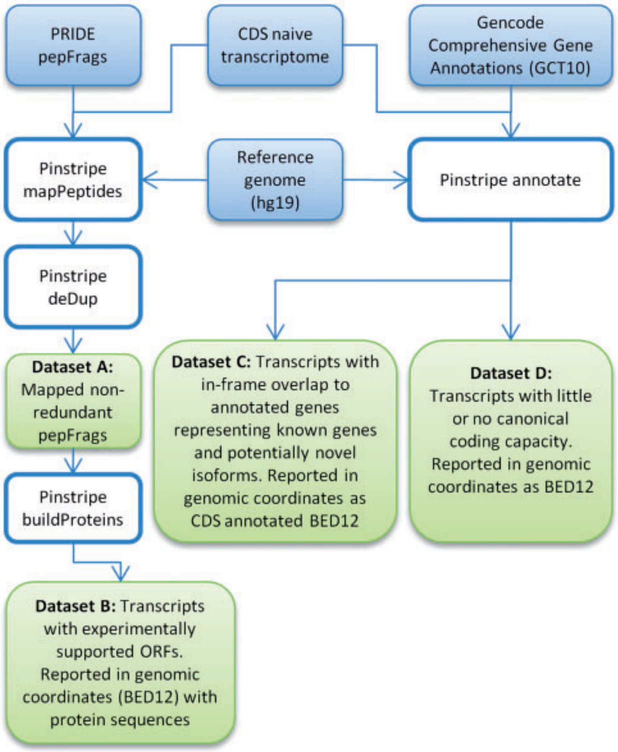
PRIDE pepFrags
CDS naive transcriptome
Gencode Comprehensive Gene Annotations (GCT10)
Pinstripe mapPeptides
Reference genome (hg19)
Pinstripe annotate
Pinstripe deDup
Dataset A: Mapped non-redundant pepFrags
Dataset C: Transcripts with in-frame overlap to annotated genes representing known genes and potentially novel isoforms. Reported in genomic coordinates as CDS annotated BED12
Dataset D: Transcripts with little or no canonical coding capacity. Reported in genomic coordinates as BED12
Pinstripe buildProteins
Dataset B: Transcripts with experimentally supported ORFs. Reported in genomic coordinates (BED12) with protein sequences

Fig. 2. Identification of mRNA and ncRNA using experimental data (*Pinstripe mapPeptides*, deDup and *buildProteins*), and by in-frame comparison to an annotation (*Pinstripe annotate*). Detailed flowcharts for these applications are included in Supplementary Fig. S1

2.2 Classification of lncRNAs

lncRNAs are defined as long mRNA-like transcripts that are not known to encode proteins, and which exhibit little or no protein-coding capacity. Because of the likelihood of long RNAs containing an ORF by chance alone, lncRNAs and mRNAs cannot be simply distinguished by the absence or presence of an ORF (Dinger *et al.*, 2008, 2011). *Pinstripe* supports classification of RNAs whose coding status is unresolved using a combination of programs, and allows the incorporation of data from other applications for assessing protein-coding potential, such as PhyloCSF (Lin *et al.*, 2011). The previously described *annotate* program not only identifies those RNAs with and without an in-frame CDS overlap to a reference but it also identifies RNAs without an ORF satisfying user-specified criteria—for example, an ORF length <80 codons and a start methionine. *Pinstripe overlap* identifies transcripts with overlapping in-frame CDSs at different tolerances, supporting identification of transcript isoforms, and *characterize* calculates the Kozak sequence strength, genomic context and probability of the ORF occurring by chance according to transcript length (refer Supplementary Text). Data from external programs are easily combined with *Pinstripe's* files using *Pinstripe join*, a BED format-specific program for identifying related data in different files, similar to a multiprocessor implementation of UNIX *sort* and UNIX *join*. Removal of mRNAs and those RNAs with uncertain coding potential reveals transcripts most likely to represent bona fide ncRNAs. An example workflow is shown in Figure 3.

3 RESULTS

To demonstrate the efficacy of *Pinstripe* for annotating transcriptomic data, we mapped non-redundant human pepFrags from the PRIDE database (Vizcaíno *et al.*, 2009) against a composite transcript library including *de novo* assembled transcriptomes from the Illumina Body Atlas of 16 human tissues (IBA2), ENCODE's GENCODE gene annotations (comprehensive) v10 (GCT10) and GenBank's database of human mRNA (Benson *et al.*, 2004). The GENCODE and GenBank datasets were included to provide coverage of transcripts from tissues not included in the IBA2, an important consideration when determining whether a pepFrag maps to a single, or multiple transcribed genomic positions.

3.1 Transcriptome assembly of the IBA2

The IBA2 is a collection of RNA sequencing experiments consisting of 2.54 billion reads across 16 human tissues: adipose, adrenal, breast, brain, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid and white blood

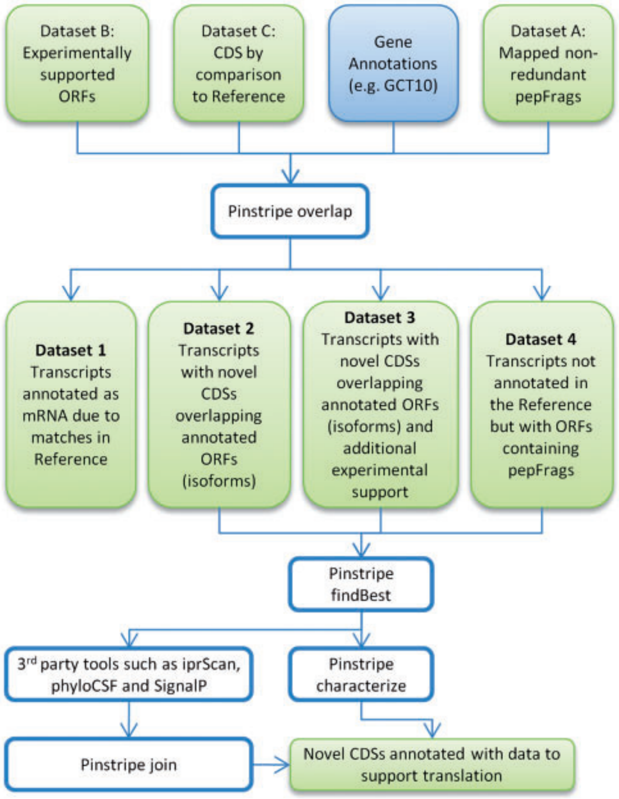
Dataset B: Experimentally supported ORFs
Dataset C: CDS by comparison to Reference
Gene Annotations (e.g. GCT10)
Dataset A: Mapped non-redundant pepFrags
Pinstripe overlap
Dataset 1 Transcripts annotated as mRNA due to matches in Reference
Dataset 2 Transcripts with novel CDSs overlapping annotated ORFs (isoforms)
Dataset 3 Transcripts with novel CDSs overlapping annotated ORFs (isoforms) and additional experimental support
Dataset 4 Transcripts not annotated in the Reference but with ORFs containing pepFrags
Pinstripe findBest
3rd party tools such as iprScan, phyloCSF and SignalP
Pinstripe characterize
Pinstripe join
Novel CDSs annotated with data to support translation

Fig. 3. Identification of different classes of novel CDSs using *Pinstripe overlap*, and annotation using characterize, *findBest* and *join*. Setting parameters to require a 100% match identifies annotated genes (Dataset 1); lower settings represent potential novel transcript isoforms of annotated genes (Datasets 2 and 3). Comparing mapped pepFrags with an annotation identifies those not included in an existing CDS. These non-CDS pepFrags can be compared with the novel isoforms to identify those with novel pepFrag support (Dataset 3) and those without (Dataset 2). The –v option of overlap reports non-overlapping CDSs (Dataset 4) representing completely novel ORFs. Additional processing with *characterize* annotates ORFs to include Kozak sequence, ORF probability and genomic context. The results of other publicly available applications such as PhyloCSF, IPRScan and Signalp can be directly appended to the annotated BED file using *Pinstripe join*

cells (see Supplementary Text). We mapped the sequence reads using *Tophat* (Trapnell *et al.*, 2009) and assembled transcriptomes for each tissue with *Cufflinks* (Trapnell *et al.*, 2010) guided by GENCODE v10, finding 59.5% of the genome was transcribed across the aggregate of tissues, with an average of 39.4% transcribed in each individual tissue, varying from 31.7% (white blood) to 48.2% (testis) (Supplementary Tables S1 and S2, Supplementary Fig. S2). Our mapping of this dataset revealed broadly similar results to those reported by Cabili *et al.* (2011) who constructed transcriptomes from the same datasets in a genome-wide study of lincRNAs (Cabili *et al.*, 2011). However, in our analysis, we report only proper pairs (paired end) or aligned reads (single end) in our results leading to a lower mapping yield averaging 77.6% across both sets (Supplementary Table S1).

Of the 28 380 RefSeq-annotated protein-coding transcripts expressed, 56.6% (16 052) were found in 14 or more of the 16 tissues, and 14.7% (4187) were present only in a single tissue, the largest fraction (42.4%; 1775) of which were found in testis.

3.2 Peptide mapping

We mapped 421 466 non-redundant pepFrags derived from the PRIDE database (Vizcaíno *et al.*, 2009) against the composite transcriptome library using *Pinstripe mapPeptides,* successfully positioning 290 104 pepFrags (no mismatch, deletion or insertions), 255 825 of which mapped to a single genomic location (allowing for up to three mismatches). A further 256 pepFrags mapped to unpositioned contigs or haplotypes but were not used in subsequent analyses.

The 131 106 pepFrags that could not be positioned demonstrate the unreliable nature of data from the PRIDE repository. Although the differing quality of individual submissions makes an aggregated assessment of the entire database impossible without reanalysis of the original spectra, even an estimate based on the number of pepFrags that can be positioned on the human genome demonstrates that PRIDE data alone provides unreliable evidence of a protein's existence. With a 31.1% chance that any entry from the database does not map to the human genome or transcriptome, even a superficial assessment requires an individual ORF to be supported by three pepFrags for a 5% False Discovery Rate (FDR) or four pepFrags to achieve 1% FDR. As a result, even ORFs with multiple matching pepFrags can only be considered to have a 'line of evidence' that an ORF may be translated, requiring additional support from other *in silico* analysis techniques such as conservation analysis, ORF probability or functional predictions, or biological validation.

For our analysis, we aimed to determine whether there was evidence of ORFs with pepFrag support and additional predictors of protein-coding potential, which remain unannotated in GENCODE v10. To identify these ORFs, we used the *Pinstripe* ORF-aware *overlap* program to identify 14 630 pepFrags with no in-frame overlap to a known GCT10 CDS from the 255 825 uniquely mapping pepFrags. We then ran *buildProteins* with only these 14 630 non-CDS coincident GCT10 pepFrags revealing a set of 12 471 unannotated ORFs. We selected the most highly supported ORF from each locus in terms of proteomic support and ORF length using *Pinstripe findBest*, (default parameters) resulting in 5561 non-overlapping CDS-annotated transcripts.

Pinstripe overlap revealed that 1130 of the 5561 CDSs contain an in-frame overlap to a GCT10 transcript, 129 of which had three or more mapping pepFrags. These ORFs arise from transcript isoforms of known genes that potentially encode novel protein variants, many of which are larger than the annotated product. For example, GIPC3 is a 312-amino acid protein encoded from chromosome 19 in which a previously unannotated splice site in the second last exon (PCR validation results in Supplementary Fig. S3) changes the reading frame of the final exon, extending the length of the ORF >400% to 1350 amino acids (Fig. 4a) and incorporating six pepFrags. Although the ORF extension partially matches the 584-amino acid Uniprot record Q8N7K9, there is no independent GENCODE-annotated transcript that would encode the ORF, and there are no histone

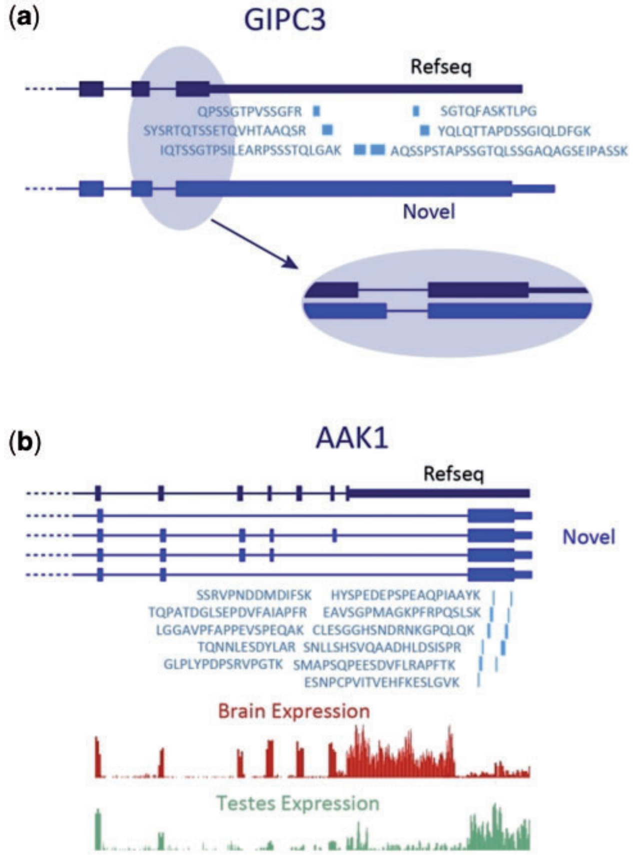

Fig. 4. Novel isoforms encoding new proteins. (**a**) A novel splice junction in the second last exon (royal blue) changes the reading frame in the last exon of *GIPC3*, extending the RefSeq ORF (dark blue) >400% through the UTR and incorporating the six pepFrags (light blue), predicting a novel protein coding region. (**b**) Several novel splicing events (royal blue) integrate a novel 3′ exon and substituting different RefSeq exons (dark blue) in *AAK1*. The novel 3′ exon is supported by 11 pepFrags (light blue). Expression data for brain (red) and testes (light green) shows different configurations are dominant in different tissues

modifications indicative of the transcription initiation necessary for an independent Q8N7K9 transcript.

Similarly, *AAK1* exhibits novel splicing variants resulting in omission of between one and five exons and a 9-kb region of the 3′-UTR (Fig. 4b). The resultant novel transcript isoform (see Supplementary Fig. S3 for PCR) substitutes up to 210 amino acids of the canonical protein's N-terminus with a completely different 307 amino acids and is supported by 11 pepFrags attributed to the 355-amino acid Uniprot protein Q6ZSR9, which, similarly to Q8N7K9, has no evidence to support its transcription except as the 3′-UTR of *AAK1*.

The reverse peptide mapping technique can also be used to provide experimental support for RefSeq genes whose status is annotated as either 'predicted', 'provisional', 'inferred' or 'unknown'. We interrogated 2213 such RefSeq genes and intersected these with the mapped PRIDE pepFrags, identifying 1100 (49.7%) ORFs having four or more pepFrags mapping to the ORF. Nineteen of these were not annotated in GENCODE v11.

3.3 Potentially novel protein-coding genes

After excluding novel isoforms of known genes, we identified 4432 canonical ORFs containing at least one uniquely mapping non-GCT10 pepFrag, of which 736 and 358 were supported by three and four or more non-GCT10 pepFrags, respectively.

The majority (81.8%) of the unannotated canonical ORFs identified (excluding isoforms) are <200 codons (Fig. 5), consistent with larger ORFs being more easily identified by both laboratory and informatic techniques. Indeed, the impact of minimum size criteria applied by *in silico* gene prediction projects such as FANTOM (Carninci *et al.*, 2005) (100 codons) and H-Invitational (Imanishi *et al.*, 2004) (80 codons) is evident in the size distributions with a difference in the distribution of ORF size where prediction algorithm cutoffs apply (Fig. 5).

As previously noted, matched pepFrags alone provide insufficient evidence for the identification of proteins. To identify which of these ORFs most likely represent bona fide protein-coding genes, and to demonstrate the *Pinstripe join* and *characterize* programs, we integrated data from other analyses including (i) protein-coding likelihood score across 29 mammal species from PhyloCSF (Lin *et al.*, 2011), (ii) the probability of the ORF occurring by chance (Dinger *et al.*, 2011), (iii) presence in the pseudogenes.org pseudogene database (Karro *et al.*, 2007), (iv) signal peptide potential (Bendtsen *et al.*, 2004), (v) similarity to known human and non-human proteins (Camacho *et al.*, 2009; Zdobnov and Apweiler, 2001) and (vi) Kozak sequence strength (Kozak, 1987). This data are available in Supplementary Table S5.

Each candidate was also annotated by its genomic context relative to GCT10-annotated genes identifying: (i) genes annotated as non-coding yet coincident with pepFrag mapping sites, (ii) coding regions within the 3′ or 5′-UTR of known protein-coding genes, (iii) genes with no overlap to known

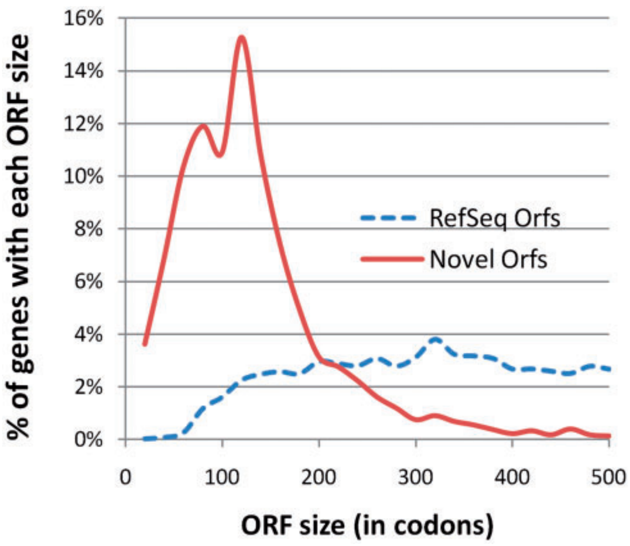

Fig. 5. Comparison of novel genes and RefSeq genes. Size distribution of ORFs in RefSeq known genes (dashed) and the novel gene candidates (solid). The distribution profile of the novel gene candidates clearly shows the 80-codon and 100-codon detection limits imposed by previous protein coding gene ORF detection methodologies

transcripts (intergenic) and (iv) genes located within the introns of other genes.

3.4 Novel ORFs

To provide examples of potential results of *Pinstripe* analysis, we characterized selected potential proteins using bioinformatic and experimental tools. A BLAST (Camacho *et al.*, 2009) comparison of the 4431 non-GCT10 canonical ORFs (excluding isoforms) against the full UniProt (Uniprot Consortium, 2010) non-redundant database of known and predicted proteins revealed significant (E-value $< 10^{-10}$) matches for 2604 (58.8%), 601 of which had UniProt annotations of 'hypothetical' or 'predicted'. In two cases [*PRAC2*; (Olsson *et al.*, 2003) and a 5′-UTR-derived isoform of *MYC* (Choi *et al.*, 2008)], literature searches revealed candidates matching experimentally validated proteins not recorded in UniProt. PhyloCSF scored 926 (20.9%) of the ORFs as 10 or greater, a result indicating the product is considered 10 times more likely to represent a coding RNA than an ncRNA. *Signal3p* (Bendtsen *et al.*, 2004) identified 226 of the novel ORFs as having strongly predicted secretion signals, whereas the *Pinstripe specificator* program annotated 940 as specific to a single tissue (see Supplementary Text). Tissue-specific protein candidates are most frequently observed in testis (46.5%) in accordance with previously reported enrichments for novel and recently evolved genes in *Drosophila* testes (Begun *et al.*, 2007; Levine *et al.*, 2006).

Analyzing the output identified a number of novel ORFs for which translation was supported by multiple lines of evidence including RD3-like protein (RD3L; Fig. 6a), a 198-amino acid protein encoded at 14q32.33, which is conserved in tetrapods and exhibits some similarity to RD3 (27% identity, E-value $= 4 \times 10^{-15}$), a protein expressed in the eye and associated with retinal degeneration (Friedman *et al.*, 2006). Within the IBA2, *RD3L* is expressed in heart (the eye was not included in IBA2) at relatively low levels, whereas *RD3* is not observed at all. A western blot was performed using antibodies raised against RD3L revealing the bands at the expected size (Supplementary Fig. S4). High expression of the protein in the mouse eye (vitreous humor and retina) was confirmed by immunofluorescence imaging (Supplementary Fig. S4). RD3L was not detected in heart by either western blot or immunofluorescence, a result possibly attributable to the lower levels of protein expression in that tissue. Subsequent to our analysis, a predicted protein sequence for RD3L was added to Uniprot.

We also identified a transcript, which within its first intron hosts the well-studied microRNAs *let7e*, *miR-125A* and *miR-99B*. This transcript, which we term *LET7EH* (Let-7e host), contains a 283-codon ORF (Fig. 6c), which is conserved in mammals and encodes immunoglobulin-like and transmembrane domains. This ORF was revealed by the identification of a previously unrecognized 5′ exon of *NCRNA00085*, which contains a canonical start codon with a strong Kozak sequence, and is supported by H3K4me3 peaks signaling transcription initiation. The novel splice junction that encompasses the miRNAs was confirmed by PCR (Supplementary Fig. S3).

Our analysis revealed previously unannotated putative protein products larger than 600 amino acids. This includes a protein sequence with similarity (55% by BLASTP) to CROCC, a structural component of the ciliary rootlet, which contributes to centrosome cohesion before mitosis (Bahe *et al.*, 2005), that we term *CROCC2* (Fig. 6b). *CROCC2* encodes a 1480-amino acid putative protein from a large conserved ORF adjacent to *SNED1*. Orthologs of *CROCC2* are present in mammals and amphibians, with both *CROCC2* and *SNED1* remaining adjacent in mouse. Similarly, we identified another transcript expressed exclusively in testes that is predicted to encode a 638-amino acid protein with a helix–loop–helix DNA-binding domain and HMG-box domain that is conserved in mammals. The putative protein, which we term HMGDC (HMG domain containing, Fig. 6d), incorporates the predicted and unreviewed Uniprot entry *C9JSJ3_HUMAN*, adding two additional exons that form the HMG-box domain.

Our *Pinstripe* analysis highlighted possible examples of underappreciated proteome complexity. For example, we identified a transcript that overlaps with the protein-coding gene, *FIP1L1*, that introduces a previously unknown ORF that is supported by four pepFrags. The putative protein encoded by this alternate reading frame, which we term FIP1L1ARF (F1P1L1 alternate reading frame), is also predicted to contain a probable signal peptide (D-score $= 0.908$, $P = 0.99$) and a neuroendocrine domain. Although such examples of multiple proteins originating from different ORFs are rare with only a small number previously reported in mammals (Clark *et al.*, 2007; Nekrutenko *et al.*, 2005), recent studies suggest this may be a much more common phenomenon (Ingolia *et al.*, 2011).

3.5 Non-coding, coding or dual function RNA

We examined 5325 transcripts (including isoforms) from the set of RefSeq-annotated genes with no annotated CDS. Intersecting these putatively non-coding transcripts with our uniquely mapping pepFrags identified 785 (14.7%) genes with pepFrags mapping within a canonical ORF, of which 341 had multiple pepFrag support.

Although this analysis identifies those genes that may be translated, the analysis of candidates without a CDS includes, in the majority, genes whose classification as non-coding is attributable solely to a lack of evidence supporting translation. To classify genes with demonstrated or purported non-coding function, we analyzed 99 manually curated candidates from lncRNAdb (Amaral *et al.*, 2011) revealing that 10 of these non-coding genes *H19*, *HAR1A*, *MALAT1*, *TUG1*, *MEG3*, *HOTTIP*, *LOC285194*, *ST7OT*, *WT1-as* and *SOX2OT* have isoforms with uniquely mapping pepFrags contained within canonical ORFs, although only *H19* and *H19AS* had three or more pepFrags, suggesting the remainder are most likely false positives.

We selected four of the ncRNAs with pepFrag support, of which two were well supported (*H19* and an antisense transcript to *H19* and *H19AS*) and two poorly supported (*MALAT1* and *TUG1*), and created antibodies against the putative proteins. The best supported of these was *H19* with a large ORF (298 codons) matching seven uniquely mapping pepFrags (Supplementary Fig. S6). *H19* was the first long regulatory non-coding RNA described in mammals, classified as non-coding because of the absence of a detectable translated protein in mouse, along with poor ORF conservation between mouse and human compared with high overall DNA sequence identity (Brannan *et al.*, 1990).

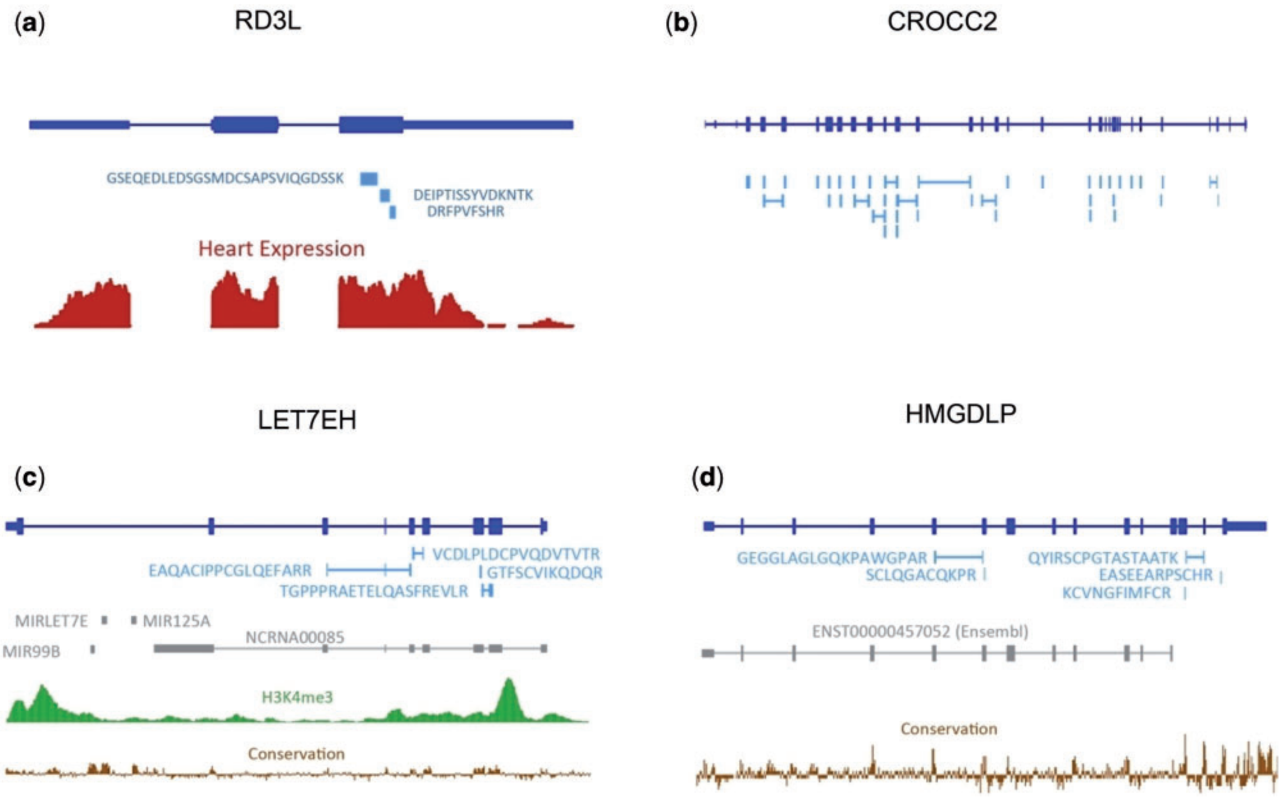

Fig. 6. Novel protein coding genes. (**a**) *RD3L* encodes a 198-amino acid protein with some similarity to RD3. (**b**) *CROCC2* encodes a 1480-amino acid protein and is supported by 40 pepFrags. Orthologs are identifiable in mammals and amphibians. (**c**) *LET7EH* encodes a 283-amino acid protein and is the primary transcript for miR-LET7E, miR-125A and miR-99B. It is conserved in mammals and contains a immunoglobulin-like and transmembrane domain. The start of transcription is supported by H3K4me3 modifications indicative of initiation. (**d**) *HMGDC* encodes a 638-amino acid testis-specific protein with helix–loop–helix DNA binding and HMG-box domains. The protein is conserved in mammals and extends a predicted Ensembl gene

The RNA is thought to be bifunctional, acting as both a miRNA and lncRNA (Smits *et al.*, 2008).

Western blots revealed no bands at the expected size for *MALAT1*, *TUG1* or *H19AS*, but the predicted ~26-kDa product for *H19* was identified in fetal liver, K562 cells and testes (Supplementary Fig. S5a) and appears to localize to the cytoplasm of T47D breast cancer cells (Supplementary Fig. S5b). As there is only limited evidence for *H19* action as a regulatory lncRNA (Runge *et al.*, 2000; Zhao *et al.*, 2010), its translation suggests *H19* is dual functional but possibly as a miRNA and mRNA, rather than lncRNA and miRNA. Although we did not successfully identify any of the other three candidates, *H19AS* has subsequently been identified by others as encoding the protein HOTS (Onyango and Feinberg, 2011).

Although the lack of confirmation for *MALAT1* and *TUG1* is unsurprising given the lack of quality controls on the PRIDE peptide database, and only single pepFrag support, the protein-coding evidence of H19 and HOTS supports the hypothesis that the binary delineation of long RNA transcripts as either ncRNA or mRNA is an oversimplification, and that genes may exhibit a range of protein-coding and non-coding functionality as either small ncRNAs, long ncRNAs, mRNAs or a combination thereof (Dinger *et al.*, 2011). On the other hand, the vast majority of spliced transcripts from the IBA2 show no evidence of translation after integrating current transcriptomic and proteomic datasets. A total of 44 843 loci were composed only of transcripts that do not contain pepFrags, compared with 34 803 loci with at least one transcript containing one or more pepFrags. This result is consistent with recent datasets produced by GENCODE (Becker, 2011).

4 DISCUSSION

Differentiating between ncRNA and mRNA is far from trivial, with traditional strategies relying on characteristics or features of known mRNAs such as codon conservation, ORF length or homology. Such strategies are effective at identifying mRNA that are well conserved, contain large canonical ORFs, are similar to known mRNA or encode known protein domains, but less reliable for detecting small poorly conserved proteins with specific non-redundant function. As a supplement to such approaches, *Pinstripe* provides a suite of tools designed to integrate and aggregate large sets of proteomic and transcriptomic experimental data, providing not only new tools for the identification of RNA potentially encoding proteins but also for integrating data from *in silico* genomic comparisons by existing applications such as PhyloCSF, SignalP and IPRScan.

There is growing evidence that large numbers of RNAs without well-conserved long canonical ORFs are actually translated. For example, the Humanin genes encode short ORFs (24–30 amino acids) and are found only in humans (Hashimoto *et al.*, 2001); translation of upstream ORFs is a demonstrated regulatory mechanism (Baek *et al.*, 2009; Wen *et al.*, 2009), and non-canonical start sites are common in mouse embryonic stem cells (Ingolia *et al.*, 2011). As a result, algorithms dependent on ORF length, canonical start codons, conservation or synonymous/non-synonymous substitution rates will not always be sufficient to identify and differentiate between coding and non-coding genes. This is especially true in the case of short proteins, which are challenging to identify as ORFs <100 codons regularly occur by chance (Dinger *et al.*, 2008). As other methods including conservation analysis and dinucleotide content are only predictive and have shortcomings, the integration of experimental evidence supporting translation is often required for confident annotation of novel proteins.

Pinstripe is designed to allow users to integrate any available experimental data in the form of gene annotations, peptide identifications or external analyses using applications such as PhyloCSF (Lin *et al.*, 2011) or SignalP (Bendtsen *et al.*, 2004). Although *Pinstripe* is entirely dependent on the integrity of the pepFrags, annotations and transcriptomes used for analysis, users can adjust the thresholds for identification to improve the probability of correct classification. The appropriate thresholds will largely depend on the quality of the data, notably the accuracy of pepFrag identifications. Most commonly, the process of correctly identifying pepFrags from spectra relies on the statistically robust concordance of the queried peptide fragment with predicted spectra from a target database. Although using pepFrags that have already been sequenced dramatically reduces the computational task, the fidelity of the original identification is lost. This can be addressed by pre-screening of spectra selecting only those with a set level of confidence, using peptides from quality-controlled sources such as the PeptideAtlas (Deutsch *et al.*, 2008), or at the protein identification stage by requiring an increased number of coincident pepFrags, or other supporting information such as conservation or functional prediction to identify coding genes.

Our analysis of the IBA2 data demonstrates that application of *Pinstripe's* programs accurately maps pepFrags and integrates them with transcripts to identify supported ORFs and identifies RNA with little or no protein-coding capacity. Although not all ORFs containing mapped pepFrags represent genuine protein-coding genes, additional lines of evidence including signal peptide potential, positive PhyloCSF scores and multiple pepFrags identify potential protein-coding genes, including the validated protein RD3LP, variants of GIPC3 and AAK1 and a protein translated from the putative ncRNA *H19*. *Pinstripe* allows researchers to identify transcripts with little or no protein-coding potential, and which have no pepFrags mapping to putative ORFs, regardless of size. This is especially important, as knowledge of whether a given transcript is coding or non-coding will guide the design of any subsequent experimental work to characterize its function.

Pinstripe is freely available for download at pinstripe.matticklab.com.

Funding: This work was supported by a National Health and Medical Research Council (NHMRC) Australia Fellowship (631668; J.S.M.), an NHMRC Career Development Award (631542; M.E.D.), and a Queensland Government Department of Employment, Economic Development and Innovation Smart Futures Fellowship (M.E.D.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Amaral,P.P. *et al.* (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.

Baek,I.C. *et al.* (2009) A novel mutation in Hr causes abnormal hair follicle morphogenesis in hairpoor mouse, an animal model for Marie Unna Hereditary Hypotrichosis. *Mamm. Genome*, **20**, 350–358.

Bahe,S. *et al.* (2005) Rootletin forms centriole-associated filaments and functions in centrosome cohesion. *J. Cell Biol.*, **171**, 27–33.

Becker,P.B. (ed.) (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.

Begun,D.J. *et al.* (2007) Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics*, **176**, 1131–1137.

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides: signalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Benson,D.A. *et al.* (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.

Brannan,C.I. *et al.* (1990) The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, **10**, 28–36.

Brosch,M. *et al.* (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res.*, **21**, 756–767.

Cabili,M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Carninci,P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.

Castellana,N.E. *et al.* (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl Acad. Sci. USA*, **105**, 21034–21038.

Choi,H. *et al.* (2008) mrtl-A translation/localization regulatory protein encoded within the human c-myc locus and distributed throughout the endoplasmic and nucleoplasmic reticular network. *J. Cell. Biochem.*, **105**, 1092–1108.

Clark,M.B. *et al.* (2007) Mammalian gene PEG10 expresses two reading frames by high efficiency -1 frameshifting in embryonic-associated tissues. *J. Biol. Chem.*, **282**, 37359–37369.

Clark,M.B. *et al.* (2011) The reality of pervasive transcription. *PLoS Biol.*, **9**, e1000625; discussion e1001102.

Deutsch,E.W. *et al.* (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.*, **9**, 429–434.

Dinger,M.E. *et al.* (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.*, **4**, e1000176.

Dinger,M.E. *et al.* (2011) The evolution of RNAs with multiple functions. *Biochimie*, **93**, 2013–2018.

Flicek,P. *et al.* (2010) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Friedman,J.S. *et al.* (2006) Premature truncation of a novel protein, RD3, exhibiting subnuclear localization is associated with retinal degeneration. *Am. J. Hum. Genet.*, **79**, 1059–1070.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Hashimoto,Y. *et al.* (2001) Mechanisms of neuroprotection by a novel rescue factor humanin from Swedish mutant amyloid precursor protein. *Biochem. Biophys. Res. Commun.*, **283**, 460–468.

Imanishi,T. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.

Ingolia,N.T. *et al*. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.

Karro,J.E. *et al*. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res*., **35**, D55–D60.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Kozak,M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res*., **15**, 8125–8148.

Levine,M.T. *et al*. (2006) Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA*, **103**, 9935–9939.

Lin,M.F. *et al*. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.

Lipman,D.J. *et al*. (2002) The relationship of protein conservation and sequence length. *BMC Evol. Biol*., **2**, 20.

Nekrutenko,A. *et al*. (2005) Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. *PLoS Genet*., **1**, e18.

Olsson,P. *et al*. (2003) PRAC2: a new gene expressed in human prostate and prostate cancer. *Prostate*, **56**, 123–130.

Onyango,P. and Feinberg,A.P. (2011) A nucleolar protein, H19 opposite tumor suppressor (HOTS), is a tumor growth inhibitor encoded by a human imprinted H19 antisense transcript. *Proc. Natl Acad. Sci*., **108**, 16759–16764.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Runge,S. *et al*. (2000) H19 RNA binds four molecules of insulin-like growth factor II mRNA-binding protein. *J. Biol. Chem*., **275**, 29562–29569.

Smits,G. *et al*. (2008) Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. *Nat. Genet*., **40**, 971–976.

Tautz,D. and Domazet-Lošo,T. (2011) The evolutionary origin of orphan genes. *Nat. Rev. Genet*., **12**, 692–702.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*., **28**, 511–515.

Uniprot Consortium (2010) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res*., 38, D142–D148.

Vizcaíno,J.A. *et al*. (2009) A guide to the proteomics identifications database proteomics data repository. *Proteomics*, **9**, 4276–4283.

Wen,Y. *et al*. (2009) Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat. Genet*., **41**, 228–233.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.

Zhao,J. *et al*. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–953.