A dual transcript-discovery approach to improve the delimitation of gene features from RNA-seq data in the chicken model

Mickael Orgeur[1,2,3], Marvin Martens[3], Stefan T. Börno[2], Bernd Timmermann[2], Delphine Duprez[3,*,#] and Sigmar Stricker[1,2,*,#].

[1] Institute for Chemistry and Biochemistry, Freie Universität Berlin, Berlin, Germany

[2] Max Planck Institute for Molecular Genetics, Berlin, Germany

[3] Sorbonne Universités, UPMC Univ. Paris 06, CNRS UMR 7622, Inserm U1156, IBPS-Developmental Biology Laboratory, 75005 Paris, France

[*] Equal contribution

[#] Corresponding authors

Delphine Duprez: delphine.duprez@upmc.fr; +33 1 4427 2753

Sigmar Stricker: sigmar.stricker@fu-berlin.de; +49 30 8387 5799

Summary statement

We combined genome-guided gene prediction and whole transcriptome assembly from RNA sequencing data to improve the chicken genome annotation. This method may be also applicable to other imperfectly annotated genomes.

Keywords

Chicken genome annotation; *Gallus gallus*; gene prediction; genome-guided transcript discovery; RNA sequencing; transcriptome reconstruction.

Abbreviations

EST, expressed sequence tags; galGal4, *Gallus gallus* genome version 4; galGal5, *Gallus gallus* genome version 5; ncRNA, non-coding RNA; ORF, open reading frame; RNA-seq, RNA sequencing; UTR, untranslated region.

Abstract

The sequence of the chicken genome, like several other draft genome sequences, is presently not fully covered. Gaps, contigs assigned with low confidence and uncharacterized chromosomes result in gene fragmentation and imprecise gene annotation. Transcript abundance estimation from RNA sequencing (RNA-seq) data relies on read quality, library complexity and expression normalization. In addition, the quality of the genome sequence used to map sequencing reads and the gene annotation that defines gene features must also be taken into account. Partially covered genome sequence causes the loss of sequencing reads from the mapping step, while an inaccurate definition of gene features induces imprecise read counts from the assignment step. Both steps can significantly bias interpretation of RNA-seq data. Here, we describe a dual transcript-discovery approach combining a genome-guided gene prediction and a *de novo* transcriptome assembly. This dual approach enabled us to increase the assignment rate of RNA-seq data by nearly 20% as compared to when using only the chicken reference annotation, contributing therefore to a more accurate estimation of transcript abundance. More generally, this strategy could be applied to any organism with partial genome sequence and/or lacking a manually-curated reference annotation in order to improve the accuracy of gene expression studies.

Introduction

Since its first release in 2004 and despite significant improvements over the last past decade, the *Gallus gallus* genome is presently incomplete and highly fragmented (Hillier et al., 2004). The chicken karyotype is composed of 38 autosomal chromosomes (1-38) and 2 additional sex chromosomes (W, Z) (Bloom et al., 1993). Out of these autosomal chromosomes, 10 are macrochromosomes (1-10), with lengths similar to those in mammals, and 28 are microchromosomes (11-38), with lengths ranging from 2 to 25 Mb (Hillier et al., 2004). Chicken microchromosomes display a high recombination rate, contain an elevated number of repetitive elements and are GC-rich, which induces significant bias and sequencing errors when using high-throughput technologies (Chen et al., 2013; Dohm et al., 2008). In addition, microchromosomes are gene dense and enriched in CpG islands, which is the result of short intronic sequences (McQueen et al., 1998; Smith et al., 2000). The fourth version of the *Gallus gallus* genome (galGal4) released in November 2011 has not fully overcome these issues. Out of the 40 chromosomes, 31 are sequenced (1-28, 32, W, Z) and contain more than 9,000 gaps, while 9 chromosomes remain missing (29-31, 33-38). The genome is also composed of about 16,000 additional contigs that are not assigned to any chromosome or assigned with low confidence. In total, the galGal4 genome sequence has a size of 1.05 Gb.

RNA sequencing (RNA-seq) data processing and results are highly dependent on the quality of the genome sequence and the associated gene annotation model. Read mapping is one of the critical steps that will further influence sample normalization, gene expression quantification and the identification of relevant genes. Gene expression profiles rely on the alignment of RNA-seq reads along the available reference genome or transcriptome, followed by their assignment to gene features. An incomplete genome sequence coupled with an inaccurate definition of gene features induce a bias in the gene expression quantification and transcript abundance estimation (Jiang and Wong, 2009; Trapnell et al., 2010). Whole

transcriptome sequencing offers valuable resources to detect novel genes and transcripts as well as to identify alternative splicing variants (Denoeud et al., 2008; Wang et al., 2008). Depending on the context, two main strategies are widely used to analyse RNA-seq data (Garber et al., 2011). One approach consists of the mapping of reads along the reference genome followed by gene prediction (Guttman et al., 2010; Trapnell et al., 2010; Yassour et al., 2009). This method can be combined with an existing reference annotation in order to detect new transcripts with respect to the provided gene annotation model (Roberts et al., 2011). The second approach aims at reconstructing the whole transcriptome independently of the reference genome (Birol et al., 2009; Grabherr et al., 2011; Robertson et al., 2010). This method is particularly suitable to study models with partial or missing genome sequence. The choice between these approaches greatly depends on the biological question and whether a reference genome is available (Conesa et al., 2016).

When analysing RNA-seq data obtained from chick embryonic limb cell cultures (so-called micromass cultures) by using the galGal4 reference genome and annotation, we observed that only 62.2% of sequencing read pairs were assigned to gene features, while 86.7% of the read pairs were mapped against the genome sequence. By comparison with the human genome, which has been nearly completely sequenced and accurately annotated, a similar analysis of RNA-seq data obtained from human blood samples depicted an assignment rate to gene features of 81.8% with a mapping rate of 92.3% (Zhao et al., 2015). We hypothesized that information was lost during the analysis of chick RNA-seq data: (i) at the mapping step, either due to low-quality sequencing reads, or due to missing genome sequence; and (ii) at the read assignment to gene features, which can be due to missing or partially annotated transcripts. To address both issues, we performed a dual transcript-discovery approach by means of genome-guided gene prediction and *de novo* transcriptome assembly. The approach described here enabled us to increase the assignment rate of RNA-seq data by nearly 20% as

compared to when using the chicken reference annotation, thus contributing to a more robust quantification of gene expression profiles.

Results

We performed RNA-seq of two independent biological replicates of chick micromass cultures infected for 5 days with empty RCAS-BP(A) replication-competent retroviral particles. 61.3 and 70.3 million of strand-specific read pairs were generated and mapped against the galGal4 version of the chicken genome by using TopHat2 (Kim et al., 2013) (Table 1). Read assignment was performed by using featureCounts (Liao et al., 2014) and a gene annotation model composed of 17,318 genes resulting from the combination of both UCSC and Ensembl reference annotations that were available at the time of analysis. Surprisingly, while 86.7% of read pairs were mapped against the chicken genome, only 62.2% of read pairs were assigned to gene features (Table 1). Therefore, 28.3% of mapped read pairs were not counted, including 93.7% of these read pairs that were not overlapping with any gene feature (Table 1). Close investigation of these unassigned read pairs highlighted genes that seemed to be absent or partially covered by the UCSC and Ensembl reference annotations (Fig. 1A,B), as well as transcripts with missing or partial exon features (Fig. 1C).

In order to improve the read assignment rate, we first performed a genome-guided transcript discovery by using Cufflinks (Trapnell et al., 2010). This approach was intended to determine more accurately exon-intron junctions, to correct or to complete existing annotated genes, and to identify unannotated gene candidates from the UCSC/Ensembl gene annotation model (Fig. 1D,E). Following this approach, 77.9% of the sequencing read pairs were assigned to gene features, corresponding to 89.8% of the read pairs that were mapped against the genome (Table 1). Therefore, the genome-guided transcript discovery enabled us to raise the read

assignment rate by 15.7% as compared to when using both UCSC and Ensembl reference annotations (Table 1). In contrast to genome-guided transcript prediction, *de novo* transcriptome reconstruction relies on overlaps between the sequencing reads to build consensus transcripts, independently of the genome sequence. We therefore applied a genome-independent strategy by using Trinity (Grabherr et al., 2011) in combination to the genome-guided approach in order to detect transcripts or regions that were not recovered from the genome sequence, such as those located within gaps or uncharacterized chromosomes (Fig. 1D,E). Reconstructed transcripts thus generated were then compared to the gene candidates obtained with the genome-guided approach in order to remove redundant sequences. Full-length transcripts or transcript regions of at least 400 bp that were not assigned to any gene candidate were extracted and grouped as an artificial chromosome. 4.0% of read pairs were found to map against this additional chromosome and 90.2% of these mapped read pairs were assigned to gene features (Table 1). By considering both transcript-discovery approaches, 90.7% of total read pairs were mapped against the galGal4 chicken genome (86.7%) and reconstructed chromosome (4.0%) (Table 1). 77.9% and 3.6% of read pairs were assigned to gene features from the genome-guided and *de novo* transcript-discovery approaches, respectively (Fig. 2A, Table 1). Therefore, 81.5% of read pairs were assigned to gene features by using this newly established gene annotation model. Given that 62.2% of sequencing read pairs were assigned to gene features by using both UCSC and Ensembl reference annotations, our transcript reconstruction model enabled us to assign 19.3% more read pairs to gene features (Fig. 2A, Table 1).

The genome-independent transcript assembly also enabled us to correct for gene fragmentation by gathering gene regions located on multiple chromosomes and contigs together (Fig. 1D,E). In contrast to genome-guided transcript discovery, *de novo* reconstruction of transcripts was not limited by the quality of the reference genome sequence.

By comparing transcripts generated from both reconstruction approaches, we were able to group dispersed gene features belonging to a same gene candidate together. Although 19,376 (90.8%) gene candidates were found exclusively on a single chromosome or unplaced contig, 1,971 (9.2%) gene candidates were identified as being fragmented (Fig. 2B). These fragmented gene candidates included 478 (2.2%) gene candidates that were located on multiple ordered chromosomes, 462 (2.2%) gene candidates split among multiple unplaced contigs, and 1,031 (4.8%) gene candidates with regions located on an ordered chromosome and additional unplaced contigs (Fig. 2B).

Transcript prediction and reconstruction approaches did not provide any information on gene name and function. Therefore, gene candidates identified by the dual transcript-discovery approach were then annotated by database comparison and protein domain prediction (Fig. 1E). Gene candidates were first compared to bird gene sequences, taking advantage of the recent increase of available genomic data within avian species and their high DNA sequence conservation (Dalloul et al., 2010; Huang et al., 2013; Jarvis et al., 2014; Schmid et al., 2015; Shapiro et al., 2013; Warren et al., 2010; Zhan et al., 2013; Zhang et al., 2014). Undefined gene candidates were then compared at the protein level to mouse and human databases. Finally, prediction of open reading frames (ORFs) and protein domains was performed on remaining unannotated gene candidates by using homology search against SwissProt and Pfam databases, and sequence analysis tools to identify transmembrane domains and signal peptides. Overall, the computed gene annotation model was mostly constituted of protein-coding gene candidates (16,716, 78.3%) (Fig. 2C). However, 672 (3.1%) gene candidates were only partly annotated (putative proteins having at least one protein domain detected), while 1,410 (6.6%) gene candidates remained unannotated (uncharacterized proteins with no protein domain identified but an ORF of at least 100 amino acids). Remaining gene candidates corresponded to miscellaneous genes (213, 1.0%; such as spliceosome complex

members, ribosomal RNAs and pseudogenes) and non-coding RNAs (ncRNAs; 4,418, 20.7%) for which no sufficient ORF could be predicted (Fig. 2C).

The resulting gene annotation model was composed of 21,347 unique gene candidates, encompassing 5,989 additional gene candidates as compared to the UCSC and Ensembl reference annotations associated with the galGal4 genome version. We then compared our results with the most recent version of the chicken genome (galGal5), released in December 2015, which includes 200 additional Mb, three previously missing chromosomes (30, 31, 33) and 23,400 unplaced contigs (Warren et al., 2017). Firstly, Strand-specific read pairs were mapped against the galGal5 genome version by using TopHat2 (Kim et al., 2013), and assigned to gene features by using featureCounts (Liao et al., 2014) according to a gene annotation model combining both UCSC and Ensembl annotations. This gene annotation model contained 6,280 additional genes as compared to the galGal4 UCSC/Ensembl annotations. Surprisingly, we did not observe any significant improvement of read pair mapping (+1.5%) and assignment (-0.9%) rates despite the increased genome size (Table 2). This indicated that when using galGal5, similar issues will be encountered as with galGal4. Indeed, a comparable number of reads pairs (25.5%) was not associated with any gene feature when mapped against galGal5 (Table 2). Secondly, we compared the predicted gene candidates from our annotation model to the RefSeq annotated galGal5 transcripts. We found that only 52.7% of gene candidates were covered by at least 50% of their total length by galGal5 reference genes (Table 3). In addition, 3,958 (18.5%) gene candidates were not detected at all in galGal5 reference genes (Table 3). 3,151 (79.6%) of these corresponded to gene candidates absent from galGal4 UCSC/Ensembl annotations. Lastly, we compared the gene names assigned to gene candidates with galGal5 reference genes that matched at least 50% of their length. Out of the 15,358 gene candidates that were identified in the galGal4 UCSC/Ensembl annotations, 74.1% had a concordant gene name, while 17.9% did not

significantly match any galGal5 reference gene (Table 4). Regarding the 5,989 additional gene candidates, most of these were not significantly detected among galGal5 reference genes (76.8%) or matched an undefined gene (12.7%) (Table 4). However, 223 (1.0%) gene candidates remaining partly annotated with the dual transcript-discovery approach could be successfully assigned (Table 4).

Altogether, this dual transcript-discovery approach enabled us to define an annotation model of 21,347 gene candidates that includes additional genes as compared to the reference annotation of the chicken genome. Most importantly, it enabled us to retrieve 19.3% more information from the RNA-seq data.

Discussion

The work presented here describes a dual transcript-discovery approach combining genome-guided gene prediction and *de novo* transcriptome reconstruction, which was applied to improve the assignment rate of RNA-seq data obtained from chicken samples. For the first approach, sequencing read pairs are mapped along the genome followed by a genome-dependent transcript discovery, which computes read coverage and exon-intron junctions from gapped alignments, and distance between both reads of each pair. By contrast, the second approach is carried out independently of the reference genome. Sequencing reads are *de novo* assembled by relying on their overlaps to reconstruct full-length transcripts. Genome-guided transcript discovery is more sensitive than *de novo* transcript reconstruction, but requires a reference genome along which RNA-seq reads are mapped for gene prediction (Garber et al., 2011; Roberts et al., 2011). Therefore, the choice of the latter method is obvious when no or incomplete genome sequence is available. In the case of the chicken model with its partial and fragmented genome sequence, the choice of a complementary transcript-discovery approach, combining both genome-guided and -independent methods,

appears suitable to improve RNA-seq data quantification and analysis. While the genome-guided approach contributes to correct existing annotated genes and to identify novel gene candidates, the *de novo* transcript reconstruction compensates for gene fragmentation by associating gene parts located on multiple chromosomes or contigs together; and it identifies gene regions or complete gene candidates that do not belong to the genome sequence due to the presence of gaps or uncharacterized fragments. The new annotation model is composed of 21,347 gene candidates, accounting for 5,989 additional gene candidates as compared to the UCSC and Ensembl reference annotations associated with the galGal4 genome version. 1,971 (9.2%) gene candidates have parts spread on multiple locations, while 3,340 (15.6%) gene candidates are identified among the 16,000 unplaced contigs that are not assigned to any ordered chromosome. In addition, the resulting gene annotation model increased the assignment rate of RNA-seq read pairs by 19.3% as compared to when using both galGal4 reference annotations (UCSC and Ensembl), thus contributing to a more accurate estimation of transcript abundance.

It is noteworthy to take into consideration that *de novo* assembly of short reads is prone to cause artefacts and to generate false chimeric transcripts (Yang and Smith, 2013). Such errors can be corrected for instance by comparing reconstructed transcripts with transcripts/proteins of the same organism, closely related organisms, or more accurately annotated organisms. In addition, transcriptome assemblers tend to create multiple transcript sequences per gene, which would cause reads to map at multiple locations and be subsequently ignored during read counting. Several programs have been developed in order to cluster transcript sequences into genes and to remove redundancy. TGICL (Pertea et al., 2003) and CD-HIT-EST (Fu et al., 2012), which were originally designed for clustering of expressed sequence tags (EST), can be used to create consensus gene sequences. However, since both programs perform their clustering based on all transcript sequences, paralogous genes may be erroneously merged. In

contrast, Corset (Davidson and Oshlack, 2014) identifies sequence similarity between transcripts by identifying multi-mapped reads resulting from re-mapping of reads against the reconstructed transcriptome. Although this program accurately clusters transcripts into genes, it falls short of building consensus genes from transcript sequences. To overcome these limitations, we applied a strategy that consists in a pairwise comparison of transcript sequences belonging to the same gene candidates followed by incremental concatenation of identical and unique transcript sequences to build full-length gene candidates. Very recently, a similar approach has been reported under the name of superTranscripts (Davidson et al., 2017). We observed that 99.95% of consensus gene sequences generated by superTranscripts were identical to our results. However, we note that superTranscripts tends to remove sequences specific to a unique transcript that do not overlap with any other transcript sequences although being indicated as belonging to the same gene candidates.

Approaches combining genome-dependent and -independent gene prediction have already been proposed before and reported to better recover the transcriptome of a given organism (Davidson et al., 2017; Jain et al., 2013; Visser et al., 2015). However, the approach presented here also includes a method to assign a putative name or function to the gene candidates resulting from gene prediction, which helps with the identification of relevant target genes in downstream analysis. The recent genome sequencing of the zebra finch (Warren et al., 2010), the turkey (Dalloul et al., 2010), the pigeon (Shapiro et al., 2013), the falcon (Zhan et al., 2013), the duck (Huang et al., 2013), and a wide range of additional avian species (Jarvis et al., 2014; Zhang et al., 2014) have provided extensive insights into evolutionary and adaptive traits within birds. DNA conservation of protein-coding genes among avian species considerably facilitated the annotation of the 21,347 gene candidates identified by the dual transcript-discovery approach. By combining DNA sequence comparison against avian genes with protein sequence comparison against mammal species

and protein domain prediction, 14,847 (69.6%) gene candidates could be assigned and 672 (3.1%) putative protein-coding gene candidates could be identified. The 5,828 (27.3%) remaining gene candidates were divided between uncharacterized proteins and ncRNAs depending on the length of the predicted ORF. However, gene candidates encoding uncharacterized proteins could be also potentially non-coding since none of the protein domains investigated was detected within their putative ORF. On the other hand, ncRNAs remain challenging to annotate according to a recent study comparing an extensive repertoire of long multi-exonic ncRNAs across 11 tetrapods separated by up to 370 million years (Necsulea et al., 2014). Besides their overall weak conservation as compared to protein-coding sequences, long ncRNAs display high tissue specificity and rapidly diverge through evolution, which renders their annotation difficult by comparing with other species.

Since the first draft released in 2004, considerable efforts have been made to improve the *Gallus gallus* reference genome and its annotation (Hillier et al., 2004; Kuo et al., 2017; Schmid et al., 2015; Thomas et al., 2014; Warren et al., 2017). In December 2015, the fifth version of the chicken genome (galGal5) was released (Warren et al., 2017). As compared to the fourth version, this release is 200 Mb longer and includes three additional chromosomes (30, 31, 33) but remains highly fragmented. Indeed, this fifth version is still composed of 15,400 unassigned contigs and 8,000 contigs assigned with low confidence, accounting for about 17% of the total genome size. While we found that some gene candidates still remain missing or partly annotated in this new release, our gene prediction is consistent with other comparisons identifying novel genes absent from galGal4 reference annotation but present in galGal5 reference annotation or other birds (Bornelöv et al., 2017; Hron et al., 2015; Lovell et al., 2014; Warren et al., 2017). Improvement of the chicken genome is an on-going project and a new version should be released within the next few years. It is reasonable to believe that continuing efforts will contribute to elucidate the full sequence of the chicken genome in

a near future. Until then, applying the dual transcript-discovery approach described here prior to the analysis of RNA-seq data *per se* enhances the sensitivity of gene expression profiles. This is particularly relevant considering that genes and splicing variants are specifically expressed in certain cell types or tissues, at different developmental stages and conditions within a single organism. For instance, we used the gene annotation model presented here as guide in a recent study, where we aimed at identifying genes that were regulated upon overexpression of connective tissue-associated transcription factors in chick micromass cultures (Orgeur et al., in preparation). More broadly, this approach could be also employed to analyse RNA-seq data of other organisms lacking manually-curated, high-quality reference annotation.

Materials and methods

A complete description of tools, command lines, parameters and database links used for this study is provided as Supplementary Methods. The gene annotation model and Python scripts are accessible via SourceForge: https://dualtranscriptdiscovery.sourceforge.io/.

Chick embryos

Fertilized chick eggs were obtained from VALO BioMedia (Lohmann Selected Leghorn strain, Osterholz-Scharmbeck, Germany). Chick embryos were staged according to the number of days *in ovo* at 37.5°C.

Chick micromass cultures

Two independent biological replicates of micromass cultures were prepared from limb buds of E4.5 chick embryos, infected with RCAS-BP(A) retroviruses carrying no recombinant protein and cultivated for 5 days as described previously (Solursh et al., 1978; Ibrahim et al., 2013). Briefly, ectoderm was dissociated by using a Dispase solution (Gibco) at 3 mg/mL and limb mesenchyme was digested by using a solution composed of 0.1% Collagenase type

Ia (Sigma-Aldrich), 0.1% Trypsin (Gibco) and 5% FBS (Biochrom) in DPBS (Gibco). Prior to seeding, mesenchymal cells were mixed with retroviruses (1:1) and maintained in culture for 5 days at 37°C in DMEM/Ham's F-12 (1:1) medium (Biochrom) supplemented with 10% FBS, 0.2% chicken serum (Sigma-Aldrich), 1% L-glutamine (Lonza) and 1% penicillin/streptomycin (Lonza).

RNA sequencing

For both replicates, RNA extracts were obtained by harvesting 6 micromass cultures with RLT buffer (Qiagen). Total RNAs were purified by using the RNeasy mini kit (Qiagen) in combination to a DNase I (Qiagen) treatment to prevent genomic DNA contamination. RNA libraries were prepared by using the TruSeq Stranded mRNA Library Preparation kit (Illumina), which enables to preserve the RNA strand orientation. Strand-specific 50-bp paired-end reads were generated by using a HiSeq 2500 sequencer (Illumina) with a mean insert size of 150 bp.

Genome-guided transcript discovery

RNA-seq data obtained from both biological replicates of micromass cultures were processed independently. Strand-specific read pairs were mapped against the chicken genome galGal4 (Hillier et al., 2004) by using TopHat2 v0.14 (Kim et al., 2013) (parameters: -r 150; -N 3; --read-edit-dist 3; --library-type fr-firststrand; -i 50; -G). UCSC (galGal4) and Ensembl (release 75) annotations were downloaded from Illumina iGenomes (http://support.illumina.com/sequencing/sequencing_software/igenome.html) and compared by using Cuffcompare from the Cufflinks suite v2.1.1 (Trapnell et al., 2010). Identical genes were retrieved only once and merged with the unique genes from each annotation. In case of discordant genes, the gene annotation with the best coverage was selected. The resulting gene annotation model composed of 17,318 genes was used as input for TopHat2 mapping.

Transcript discovery was performed for each replicate by using Cufflinks v2.1.1 (Trapnell et al., 2010) (parameters: -b; -u; -library-type, fr-firststrand; -g) and the combined gene annotation model as guide. Resulting annotations were merged into a single model by using the Cufflinks tool Cuffmerge v2.1.1 (Trapnell et al., 2010).

De novo transcript discovery

A second transcript-discovery approach was led independently of the genome sequence. Low-quality RNA-seq reads from each replicate of micromass cultures were first filtered out by using the FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). Reads with a median quality value lower than 28 were discarded. Filtered read pairs were then trimmed by using Trimmomatic v0.32 (Bolger et al., 2014) (parameters: ILLUMINACLIP TruSeq3 paired-end for HiSeq, seedMismatches 2, palindromeClipThreshold 30, simpleClipThreshold 10; LEADING 5; TRAILING 5; MINLEN 36). Complete read pairs were then assembled by using Trinity r20140717 (Grabherr et al., 2011) (default parameters except for the strand-specific library orientation set at RF).

Gene fragmentation correction

Contigs resulting from the *de novo* assembly were compared to the gene candidate sequences obtained by the first approach by using BLASTN from BLAST+ v2.2.31+ (Camacho et al., 2009) (parameters: -strand plus; -dust no; -soft_masking no). Contigs were assigned to a given gene candidate if they matched at least 40 bp that were not covered by a previous hit with a percentage of identities higher than 90%. Assigned contigs that were not fully covered by a given gene candidate were further processed to extract continuous uncovered regions of at least 400 bp. Remaining contigs were mapped against the galGal4 genome by using BLASTN (parameters: -perc_identity 90; -dust no; -soft_masking no). Contigs were assigned to a given gene candidate if they were located between two gene features, potentially

corresponding to an exon missed by Cufflinks, or in the vicinity of a first or last exon, potentially corresponding to a missing 5'- or 3'-untranslated region (UTR), respectively. Remaining unmapped contigs were retrieved as they could correspond to non-defined genomic regions. Unmapped, unassigned and non-covered contigs or regions of at least 400 bp were further processed to remove redundant sequences between multiple isoforms. This step was necessary to prevent read pairs to be mapped on multiple gene features and to be consequently discarded during fragment counting. Isoforms belonging to the same gene candidates defined by Trinity were compared to the longest isoforms by using BLASTN (parameters: -perc_identity 90; -strand plus; -dust no; -soft_masking no; -ungapped). Sequence alignments were then examined to build consensus gene sequences by merging identical sequences between two isoforms and by adding sequences unique to each isoform. Pairwise sequence comparison was performed until all isoforms of the same gene candidates were processed and concatenated. Resulting contig sequences were gathered together as an artificial chromosome and separated to each other by 250 bp of nucleotides N, corresponding to the total length of read pairs (50 bp for each read and 150 bp as insert size).

Functional annotation

Gene candidate sequences retrieved from both transcript-discovery approaches were then compared to existing databases for gene name assignment. First, gene candidates were compared to the NCBI RefSeq transcript database by using BLASTN (parameters: -strand plus; -dust no; -soft_masking no). Comparison was limited to Aves (birds) sequences (taxid 8782). Gene candidates with a percentage of identities higher than 90% for chicken genes or 75% for bird genes, and bidirectionaly covered on at least 50% of their length were assigned to the corresponding hits. Gene candidates matching several discordant gene names, such as chimeric and fused gene candidates, were manually investigated and corrected. Non-annotated gene candidate sequences were then compared to the NCBI human (taxid 9606)

and mouse (taxid 10090) non-redundant protein database by using BLASTX from BLAST+ v2.2.31+ (Camacho et al., 2009) (parameters: -strand, plus; -seg, no). Gene candidates with a percentage of homology of at least 30% and covered by at least 50% of their length were filtered. Matching protein accession numbers were converted into gene accession numbers by using the Hyperlink Management System (Imanishi and Nakaoka, 2009). ORF prediction was finally performed on remaining gene candidates by using TransDecoder v2.1.0 (Haas et al., 2013) (strand specificity parameter: -S). ORFs of at least 100 amino acids were annotated by using Trinotate v3.0.1 (https://trinotate.github.io/). Functional annotation was based on the following protein predictions: (i) BLASTX and BLASTP homology search against the SwissProt database (Bairoch et al., 2004); (ii) protein domain prediction against the Pfam database (Punta et al., 2012) by using HMMER v3.1b2 (Finn et al., 2011); (iii) signal peptide prediction by using SignalP v4.1 (Petersen et al., 2011); and (iv) transmembrane domain prediction by using tmHMM v2.0c (Krogh et al., 2001). Resulting functional annotation was divided into three categories: (i) putative proteins, for which at least one protein domain could be identified; (ii) uncharacterized proteins, corresponding to ORFs for which no protein domain could be identified; and (iii) ncRNAs, corresponding to genes with an ORF shorter than 100 amino acids.

Comparison with galGal5

UCSC (galGal5) and Ensembl (release 89) reference annotations associated with the galGal5 genome version were downloaded from the UCSC browser and merged by using the Cufflinks tool Cuffmerge v2.1.1 (Trapnell et al., 2010). RNA-seq strand-specific read pairs were mapped against the chicken genome galGal5 (Warren et al., 2017) by using TopHat2 v0.14 (Kim et al., 2013) (parameters: -r 150; -N 3; --read-edit-dist 3; --library-type fr-firststrand; -i 50; -G) and the merged reference annotations as guide. Sequences of annotated galGal5 transcripts were retrieved from the RefSeq database

(ftp://ftp.ncbi.nih.gov/genomes/Gallus_gallus/RNA/) and compared to the predicted gene candidates by using BLASTN (parameters: -perc_identity 90; -strand plus; -dust no; -soft_masking no). On one hand, the total length coverage of predicted gene candidates was assessed by identifying all regions matching with galGal5 gene sequences. On the other hand, gene name assignment between predicted gene candidates and annotated galGal5 genes was compared by retrieving only the hits that matched at least 50% of their length.

Fragment counting

Strand-specific read pairs mapped against the chicken genome and the artificial chromosome generated from the *de novo* transcript discovery were first split by strand by using SAMtools v1.2 (Li et al., 2009) according to their FLAG field (strand plus: -f 128 -F 16, -f 80; strand minus: -f 144, -f 64 -F 16). Fragments (both reads of a pair) mapped on gene features were counted by using featureCounts v1.4.6-p3 (Liao et al., 2014) (parameters: -p; -s 2; --ignoreDup; -B; -R). Chimeric fragments aligned on different chromosomes were taken into consideration to overcome the gene fragmentation due to the location of gene parts on multiple chromosome contigs.

Acknowledgements

We thank Georgeta Leonte (Freie Universität, Berlin, Germany) for helping preparing and collecting the samples. We also thank Stéphane Descorps-Declère and Marc Monot (Institut Pasteur, Paris, France) for critical reading of the manuscript, as well as Roman Eremchenko for fruitful discussion. We are grateful to the Sequencing Core Facility of the Max Planck Institute for Molecular Genetics for processing the RNA-seq. We are thankful to Peter Hansen and Peter N. Robinson (Charité Universitätsmedizin, BCRT, Berlin, Germany), as well as Marius van den Beek and Christophe Antoniewski (Institut de Biologie Paris-Seine, ARTbio, Paris, France) for providing access to Galaxy web servers.

Competing interests

The authors declare that no competing interests exist.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG; grant GK1631), the Université Franco-Allemande (UFA/DFH; grants CDFA-06-11 and CT-24-16), the Association Française contre les Myopathies (AFM; grants 16826 and 18626), the Fondation pour la Recherche Médicale (FRM; grant DEQ20140329500), the INSERM and the CNRS. MO was part of the MyoGrad International Research Training Group for Myology and received financial support from the FRM (grant FDT20150532272).

Data availability

RNA-seq data have been deposited on the Gene Expression Omnibus (GEO) database under the SuperSeries accession number GSE100517. Both samples that have been used for this study are available under the SubSeries GSE100516 via the accession numbers GSM2685833 and GSM2685834. Gene annotation model in GTF format associated with the galGal4 version of the chicken genome and sequence of the artificial chromosome in FASTA format are available at: https://dualtranscriptdiscovery.sourceforge.io/.

Authors' contributions

MO, DD and SS designed and conceived the study. MO performed the experiments and collected the samples. MO and MM analysed the data. STB and BT performed and supervised the RNA-seq procedure. MO, DD and SS wrote the manuscript, with comments and approval from all authors.

Supplementary information

Supplementary Methods. Supplementary methods providing command lines and parameters as well as a detailed description of the dual transcript-discovery approach.

References

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. (2004). Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* **5**, 39-55.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., et al. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872-2877.

Bloom, S., Delany, M. and Muscarella, D. (1993). Constant and variable features of avian chromosomes. In *Manipulation of the Avian Genome.* (ed. R. J. Etches and A. M. Verrinder Gibbins), pp. 39-59. CRC Press, Inc.

Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Bornelöv, S., Seroussi, E., Yosefi, S., Pendavis, K., Burgess, S. C., Grabherr, M., Friedman-Einat, M. and Andersson, L. (2017). Correspondence on Lovell et al.: identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biol.* **18**, 112.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y. and Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* **8**, e62856.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey

of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13.

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg, L. A., Bouffard, P., Burt, D. W., Crasta, O., Crooijmans, R. P. M. A., et al. (2010). Multi-platform next-generation sequencing of the domestic Turkey (*Meleagris gallopavo*): Genome assembly and analysis. *PLoS Biol.* **8**, e1000475.

Davidson, N. M. and Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* **15**, 410.

Davidson, N. M., Hawkins, A. D. K. and Oshlack, A. (2017). SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biol.* **18**, 148.

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., et al. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175.

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105.

Finn, R. D., Clements, J. and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152.

Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8**, 469-477.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I.,

Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503-510.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494-1512.

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., Bork, P., Burt, D. W., Groenen, M. A. M., Delany, M. E., et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716.

Hron, T., Pajer, P., Pačes, J., Bartůněk, P. and Elleder, D. (2015). Hidden genes in birds. *Genome Biol.* **16**, 164.

Huang, Y., Li, Y., Burt, D. W., Chen, H., Zhang, Y., Qian, W., Kim, H., Gan, S., Zhao, Y., Li, J., et al. (2013). The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.* **45**, 776-783.

Ibrahim, D. M., Hansen, P., Rödelsperger, C., Stiege, A. C., Doelken, S. C., Horn, D., Jäger, M., Janetzki, C., Krawitz, P., Leschik, G., et al. (2013). Distinct global shifts in genomic binding profiles of limb malformation-associated HOXD13 mutations. *Genome Res.* **23**, 2091-2102.

Imanishi, T. and Nakaoka, H. (2009). Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Res.* **37**, W17-22.

Jain, P., Krishnan, N. M. and Panda, B. (2013). Augmenting transcriptome assembly by combining de novo and genome-guided tools. *PeerJ* **1**, e133.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320-1331.

Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25**, 1026-1032.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567-580.

Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L. and Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Liao, Y., Smyth, G. K. and Shi, W. (2014). featureCounts: an efficient general purpose

program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930.

Lovell, P. V, Wirthlin, M., Wilhelm, L., Minx, P., Lazar, N. H., Carbone, L., Warren, W. C. and Mello, C. V. (2014). Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol.* **15**, 565.

McQueen, H. A., Siriaco, G. and Bird, A. P. (1998). Chicken microchromosomes are hyperacetylated, early replicating, and gene rich. *Genome Res.* **8**, 621-630.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F. and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., et al. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-652.

Petersen, T. N., Brunak, S., von Heijne, G. and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785-786.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* **40**, D290-301.

Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-seq. *Bioinformatics* **27**, 2325-2329.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909-912.

Schmid, M., Smith, J., Burt, D. W., Aken, B. L., Antin, P. B., Archibald, A. L., Ashwell, C., Blackshear, P. J., Boschiero, C., Brown, C. T., et al. (2015). Third report on chicken genes and chromosomes 2015. *Cytogenet. Genome Res.* **145**, 78-179.

Shapiro, M. D., Kronenberg, Z., Li, C., Domyan, E. T., Pan, H., Campbell, M., Tan, H., Huff, C. D., Hu, H., Vickrey, A. I., et al. (2013). Genomic diversity and evolution of the head crest in the rock pigeon. *Science* **339**, 1063-1067.

Smith, J., Bruley, C. K., Paton, I. R., Dunn, I., Jones, C. T., Windsor, D., Morrice, D. R., Law, A. S., Masabanda, J., Sazanov, A., et al. (2000). Differences in gene density on chicken macrochromosomes and microchromosomes. *Anim. Genet.* **31**, 96-103.

Solursh, M., Ahrens, P. B. and Reiter, R. S. (1978). A tissue culture analysis of the steps in limb chondrogenesis. *In Vitro* **14**, 51-61.

Thomas, S., Underwood, J. G., Tseng, E., Holloway, A. K., on behalf of the Bench To Basinet CvDC Informatics Subcommittee. (2014). Long-read sequencing of chicken transcripts and identification of new transcript isoforms. *PLoS One* **9**, e94650.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511-515.

Visser, E. A., Wegrzyn, J. L., Steenkmap, E. T., Myburg, A. A. and Naidoo, S. (2015). Combined de novo and genome guided assembly and annotation of the *Pinus patula* juvenile shoot transcriptome. *BMC Genomics* **16**, 1057.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A. J., Fairley, S., et al. (2010). The genome of a songbird. *Nature* **464**, 757-762.

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., Markovic, C., Bouk, N., Pruitt, K. D., Thibaud-Nissen, F., et al. (2017). A new chicken genome assembly provides insight into avian genome structure. *G3 (Bethesda)* **7**, 109-117.

Yang, Y. and Smith, S. A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* **14**, 328.

Yassour, M., Kaplan, T., Fraser, H. B., Levin, J. Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., et al. (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 3264-3269.

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M. G., Ni, P., Hu, L., Liu, Y., Hou, H., et al. (2013). Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat. Genet.* **45**, 563-566.

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., et al. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311-1320.

Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D. and Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics* **16**, 675.

Tables

Table 1. RNA-seq read pair assignment.

Sample	Read pairs	RCAS-BP(A) genome	Chicken reference genome (galGal4)				*De novo* assembly (Trinity)		Total gain of read assignment	
		Mapped pairs	Mapped pairs	Assigned pairs [UCSC/Ensembl]	Assigned pairs [Cufflinks]	Gain of assigned pairs	Mapped pairs	Assigned pairs		
Rep1	61.3 M	1.7 M	53.1 M	38.0 M	47.6 M	**+9.6 M**	2.4 M	**2.2 M**	**+11.8 M**	
			Mapped pairs with no gene feature	14.2 M	4.6 M					
Rep2	70.3 M	2.1 M	61.0 M	43.9 M	55.0 M	**+11.1 M**	2.9 M	**2.6 M**	**+13.7 M**	
			Mapped pairs with no gene feature	16.0 M	5.0 M					
Average (Rep1/2)		**2.9%**	**86.7%**	**62.2%**	**77.9%**	**+15.7%**	**4.0%**	**3.6%**	**+19.3%**	total pairs
			Assigned mapped pairs	71.7%	89.8%			90.2%		total mapped pairs
			Unassigned mapped pairs	28.3%	10.2%			9.8%		
			Mapped pairs with no gene feature	26.5%	8.4%					

Abbreviation: M, million of read pairs.

Table 2. RNA-seq read pair assignment against galGal5.

Sample	Read pairs	Reference genome (galGal5)		Reference annotations (UCSC/Ensembl)			
		Mapped reads	As compared to galGal4		Assigned pairs	As compared to galGal4	
Rep1	61.3 M	53.9 M	**+0.8 M**		37.3 M	**-0.6 M**	
				Mapped pairs with no gene feature	13.9 M	-0.3 M	
Rep2	70.3 M	62.2 M	**+1.2 M**		43.4 M	**-0.5 M**	
				Mapped pairs with no gene feature	15.6 M	-0.4 M	
Average (Rep1/2)		**88.2%**	**+1.5%**		**61.3%**	**-0.9%**	total pairs
				Assigned mapped pairs	69.5%	-2.2%	total mapped pairs
				Unassigned mapped pairs	30.5%	+2.2%	
				Mapped pairs with no gene feature	25.5%	-1.0%	

Abbreviation: M, million of read pairs.

Table 3. Length coverage of gene candidates as compared to galGal5 reference genes.

Length coverage	Number of gene candidates	Cumulative number	Cumulative percentage
100%	3,620	3,620	17.0%
≥ 75% and < 100%	4,822	8,442	39.5%
≥ 50% and < 75%	2,801	11,243	52.7%
≥ 25% and 50%	3,282	14,525	68.0%
> 0% and < 25%	2,864	17,389	81.5%
0%	3,958	21,347	100%

Table 4. Comparison of galGal4 gene candidates to galGal5 reference genes.

Gene candidates	Number	Percentage
galGal4 reference genes	15,358	
- concordant assignment	11,384	74.1%
- concordant and undefined assignments [a]	368	2.4%
- partly annotated with assignment [b]	41	0.3%
- assigned with different gene symbol	126	0.8%
- undefined assignment [c]	441	2.9%
- discordant assignment [d]	244	1.6%
- without assignment	2,754	17.9%
galGal4 additional genes	5,989	
- concordant assignment	376	6.3%
- concordant and undefined assignment [a]	29	0.5%
- partly annotated with assignment [b]	182	3.0%
- assigned with different gene symbol	28	0.5%
- undefined assignment [c]	760	12.7%
- discordant assignment [d]	16	0.3%
- without assignment	4,598	76.8%

[a] Gene candidates matching a correct gene and one or several undefined genes (LOC, ORF).

[b] Gene candidates resulting from ORF and protein domain prediction.

[c] Gene candidates matching one or several undefined genes (LOC, ORF).

[d] Includes highly repeated genes such as those encoding histone proteins and myosin heavy chains.

Figures

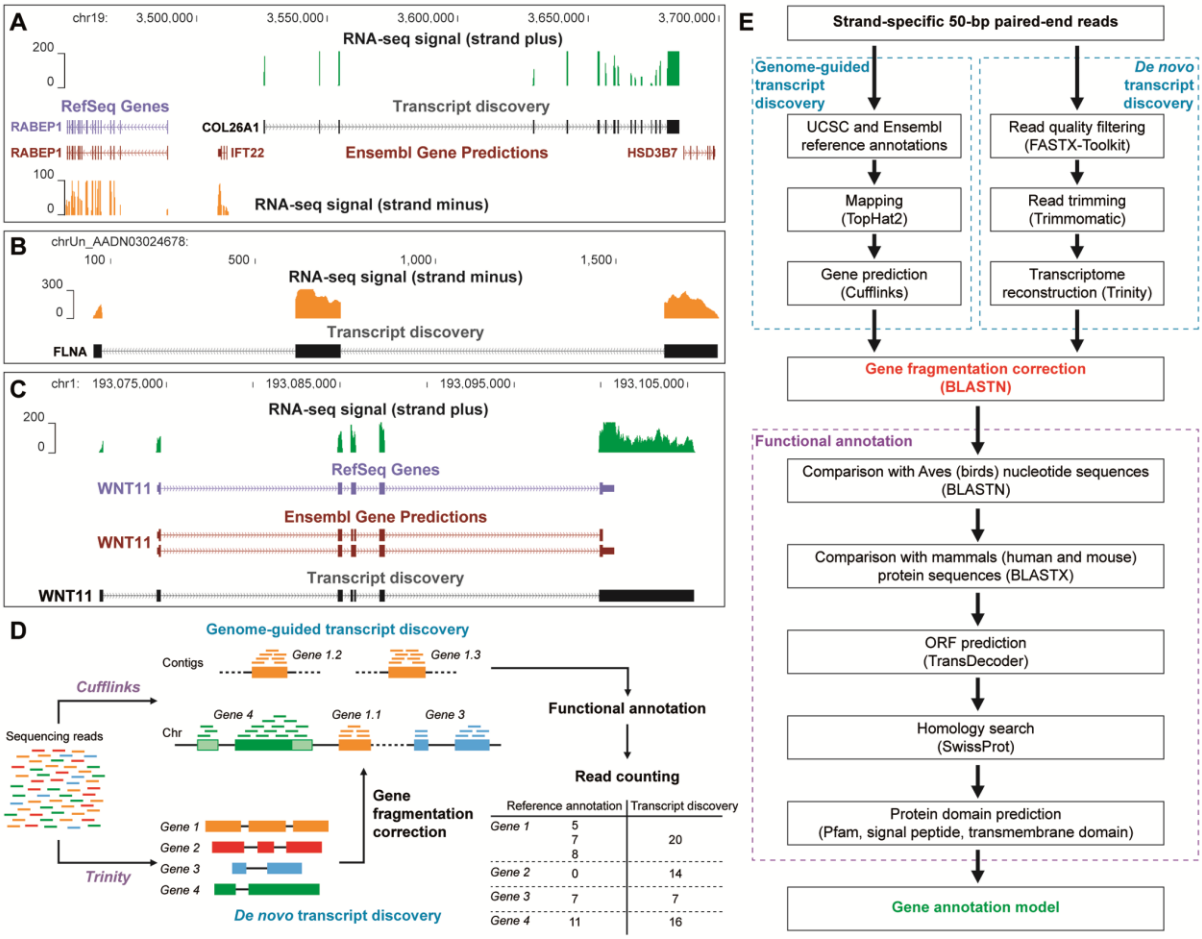

Fig. 1. Dual transcript-discovery approach. (**A**) Region surrounding the genes *RABEP1* and *HSD3B7* on chromosome 19. RNA-seq signal on strand plus (green), which does not overlap any gene from UCSC and Ensembl reference annotations, corresponds to the gene *COL26A1*. (**B**) RNA-seq signal (orange) on strand minus of an uncharacterized contig delimitating 3 exons of the gene *FLNA*. (**C**) Region of the gene *WNT11* on chromosome 1. As visible from the RNA-seq signal on strand plus (green), both UCSC and Ensembl reference annotations lack an exon of the 5'-UTR and display a shorter 3'-UTR. (**D**) The dual transcript-discovery approach combined a genome-guided gene prediction with a *de novo* transcriptome reconstruction. This dual approach enabled us to correct for gene fragmentation

(orange), to identify missing gene candidates (red) and to adjust existing annotated genes (green), thus improving the assignment rate of RNA-seq read pairs. (**E**) Workflow to design the comprehensive gene annotation model.

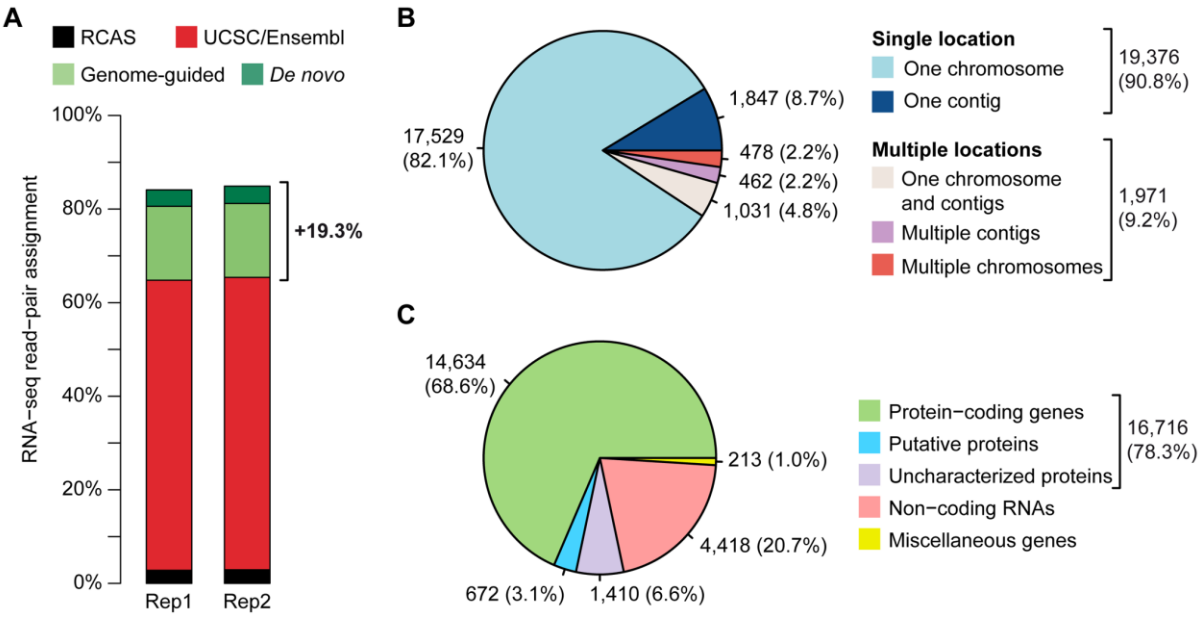

Fig. 2. Characteristics of the new gene annotation model. (**A**) The dual transcript-discovery approach combining genome-guided gene prediction (light green) and de novo transcriptome reconstruction (dark green) raised the read-pair assignment rate by 19.3% as compared to when using the UCSC and Ensembl reference annotations (red). The proportion of read pairs coming from the RCAS-BP(A) replication competent retroviruses is depicted in black. (**B**) Proportion of gene locations on chromosomes and contigs of the chicken reference genome galGal4. 9.2% of identified gene candidates are fragmented due to their location on multiple chromosomes and contigs. (**C**) Proportion of annotated gene biotypes. Most of the annotated gene candidates potentially encode proteins (78.3%). Putative proteins correspond to gene candidates for which at least one protein domain could be detected (3.1%). Uncharacterized proteins are gene candidates with an ORF of at least 100 amino acids without protein domain identified (6.6%). Gene candidates with no sufficient predicted ORF (less than 100 amino acids) are classified as non-coding RNAs (20.7%). Gene candidates encoding spliceosome complex members and ribosomal RNAs, as well as pseudogenes are classified as miscellaneous genes (1.0%).

A dual transcript-discovery approach to improve the delimitation of gene features from RNA-seq data in the chicken model: Supplementary Methods

1) Requirements:

- TopHat2 v0.14
- Cufflinks v2.1.1
- FASTX-toolkit v0.0.13
- Trimmomatic v0.32
- Trinity r20140717
- BLAST+ v2.2.31+
- BEDtools v2.24.0
- TransDecoder v2.1.0
- HMMER v3.1b2
- SignalP v4.1
- tmHMM v2.0c
- Trinotate v3.0.1
- Python v2.7 (scripts are available at: https://dualtranscriptdiscovery.sourceforge.io/)

2) Datasets:

Strand-specific paired-end reads (length of 50 bp, mean insert size of 150 bp) were generated by using a HiSeq 2500 sequencer (Illumina). Datasets used for this study are available at:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2685833
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2685834
The reference sequence of the chicken galGal4 genome can be downloaded from the UCSC browser via the following link: http://hgdownload.soe.ucsc.edu/goldenPath/galGal4/bigZips/galGal4.fa.gz.
Length of each chromosome and contig associated with the chicken gaGal4 genome are available at: http://hgdownload.soe.ucsc.edu/goldenPath/galGal4/bigZips/galGal4.chrom.sizes.
Ensembl and UCSC gene annotation models related to galGal4 are accessible via Illumina iGenomes: https://support.illumina.com/sequencing/sequencing_software/igenome.html.

3) Genome-guided transcript discovery:

3.1) Read mapping: TopHat2

With a reference gene annotation model

```
$ tophat -r 150 -N 3 --read-edit-dist 3 --library-type fr-firststrand -i 50 \
    -G genes.gtf genome reads_R1.fq.gz reads_R2.fq.gz
```

Without a reference gene annotation model

```
$ tophat -r 150 -N 3 --read-edit-dist 3 --library-type fr-firststrand -i 50 \
    genome reads_R1.fq.gz reads_R2.fq.gz
```

3.2) Gene prediction: Cufflinks

With a reference gene annotation model

```
$ cufflinks -b genome.fa -u -library-type fr-firststrand -g genes.gtf \
    accepted_hits.bam
```

Without a reference gene annotation model

```
$ cufflinks -b genome.fa -u -library-type fr-firststrand accepted_hits.bam
```

3.3) Merge gene annotation models: Cuffmerge

Create a file listing the name of the gene annotation models generated for each replicate

```
./transcripts_Rep1.gtf
./transcripts_Rep2.gtf
```

Create a single gene annotation model

```
$ cuffmerge list_models.txt
```

4) *De novo* transcript discovery:

4.1) Merge reads from both replicates

```
$ cat reads_R1_Rep1.fq reads_R1_Rep2.fq > reads_R1.fq
$ cat reads_R2_Rep1.fq reads_R2_Rep2.fq > reads_R2.fq
```

4.2) Filter reads by quality: FASTX-Toolkit

```
$ fastq_quality_filter -q 28 -p 50 -i reads_R1.fq -o reads_R1.filtered.fq
$ fastq_quality_filter -q 28 -p 50 -i reads_R2.fq -o reads_R2.filtered.fq
```

4.3) Trim reads by quality: Trimmomatic

```
$ java -jar trimmomatic-0.32.jar PE reads_R1.filtered.fq reads_R2.filtered. \
     reads_R1.trimmed.fq reads_R1.unpaired.fq \
     reads_R2.trimmed.fq reads_R2.unpaired.fq \
     ILLUMINACLIP:TruSeq3-PE:2:30:10 LEADING:5 TRAILING:5 MINLEN:36
```

4.4) *De novo* assembly: Trinity

```
$ $TRINITY_HOME/Trinity.pl --seqType fq --JM 10G --SS_lib_type RF \
     --left reads_R1.trimmed.fq --right reads_R2.trimmed.fq
```

5) Gene fragmentation correction

5.1) Retrieve transcript sequences

Upload the gene annotation model "merged.gtf" resulting from the genome-guided transcript discovery on the UCSC browser. Transcript sequences can be retrieved by using the Table Browser tool.

5.2) Create a BLAST database: BLAST+

```
$ makeblastdb -in transcripts.fa -dbtype nucl
```

5.3) Compare Trinity contigs to transcripts: BLAST+

```
$ blastn -query contigs.fa -db transcripts.fa -perc_identity 90 \
     -strand plus -dust no -soft_masking no -outfmt "7 std qlen slen sstrand" \
     -out contigs_vs_transcripts.blastn
```

5.4) Convert transcript IDs into gene IDs in the BLAST output file: Python script

```
$ python convert_tids_into_gids.py \
     merged.gtf contigs_vs_transcripts.blastn contigs_vs_genes.blastn
```

5.5) Assign Trinity contigs to genes: Python script

The minimum number of overlapping base pairs not covered from previous hits can be fixed by adjusting the parameter "t_aln_length=40".

```
$ python assign_contigs_to_genes.py \
     contigs_vs_genes.blastn 40 \
     assigned_contigs.txt
```

5.6) Extract unassigned Trinity contigs: Python script

```
$ python extract_unassigned_contigs.py \
     contigs.fa assigned_contigs.txt unassigned_contigs.fa
```

5.7) Create a BLAST database: BLAST+

```
$ makeblastdb –in genome.fa –dbtype nucl
```

5.8) Compare unassigned Trinity contigs to genome: BLAST+

```
$ blastn –query unassigned_contigs.fa –db genome.fa –perc_identity 90 \
     –dust no –soft_masking no –outfmt "7 std qlen slen sstrand" \
     –out unassigned_contigs_vs_genome.blastn
```

5.9) Filter hits based on cumulative alignment length: Python script

The minimum number of overlapping base pairs not covered from previous hits can be fixed by adjusting the parameter "t_aln_length=40". The minimum percentage of cumulative alignment length can be fixed by adjusting the parameter "p_cumul_length=50".

```
$ python parse_blast_hits_genome.py \
     unassigned_contigs_vs_genome.blastn 40 50 \
     unassigned_contigs_vs_genome.blastn.txt
```

5.10) Extract genome coordinates from filtered hits: Python script

```
$ python extract_genome_coordinates.py \
     unassigned_contigs_vs_genome.blastn.txt \
     unassigned_contigs_vs_genome.blastn.bed
```

5.11) Sort genome coordinates: sort

```
$ sort –k1,1 –k2,2n unassigned_contigs_vs_genome.blastn.bed \
     > unassigned_contigs_vs_genome.blastn.sort.bed
```

5.12) Extract gene coordinates: Python script

```
$ python extract_gene_coordinates.py \
     merged.gtf genes.bed
```

5.13) Extend gene boundaries by 1000 bp: BEDtools

```
$ bedtools slop –i genes.bed –g galGal4.chrom.sizes –b 1000 \
     > genes.extended.bed
```

5.14) Sort gene coordinates: sort

```
$ sort –k1,1 –k2,2n genes.extended.bed > genes.extended.sort.bed
```

5.15) Compare unassigned Trinity contigs to genes: BEDtools

```
$ bedtools intersect –a unassigned_contigs_vs_genome.blastn.sort.bed \
     –b genes.extended.bed –wo –s > unassigned_contigs_vs_genes.bed
```

5.16) Assign Trinity contigs to genes: Python script

```
$ python assign_unassigned_contigs_to_genes.py \
     unassigned_contigs_vs_genes.bed \
     unassigned_contigs.txt
```

5.17) Extract unmapped Trinity contigs: Python script

```
$ python extract_unmapped_contigs.py \
     contigs.fa assigned_contigs.txt unassigned_contigs.txt \
     unmapped_contigs.txt
```

5.18) Select regions of assigned, unassigned and unmapped Trinity contigs: Python script

The minimum number of continuous base pairs not covered can be fixed by adjusting the parameter "min_length=400".

```
$ python select_contig_regions.py \
     contigs.fa assigned_contigs.txt \
     unassigned_contigs.txt unmapped_contigs.txt 400 \
     contig_regions.fa
```

6) Remove redundant sequences from selected Trinity contigs

6.1) Extract Trinity contigs with multiple isoforms: Python script

```
$ python extract_gene_isoforms.py \
     contig_regions.fa \
     contigs_with_single_isoform.fa \
     contigs_with_multiple_isoforms.fa
```

6.2) Extract longest isoforms: Python script

```
$ python extract_longest_isoforms.py \
     contigs_with_multiple_isoforms.fa \
     0_contigs_longest_isoforms.fa \
     0_contigs_longest_isoforms.index
```

6.3) Create a BLAST database: BLAST+

```
$ makeblastdb -in 0_contigs_longest_isoforms.fa -dbtype nucl
```

6.4) Compare Trinity contigs with multiple isoforms to longest isoforms: BLAST+

```
$ blastn -query contigs_with_multiple_isoforms.fa \
     -db 0_contigs_longest_isoforms.fa \
     -perc_identity 90 -strand plus -dust no -soft_masking no -ungapped \
     -outfmt "7 std qlen slen sstrand" \
     -out 0_contigs_with_multiple_isoforms_vs_longest_isoforms.blastn
```

6.5) Parse BLAST hits to build contig scaffolds: Python script

```
$ python build_scaffolds.py \
     0_contigs_longest_isoforms.index \
     0_contigs_with_multiple_isoforms_vs_longest_isoforms.blastn \
     0_contigs_longest_isoforms.scaffolds
```

6.6) Build contig sequences based on scaffolds: Python script

```
$ python build_consensus_sequences.py \
     contigs_with_multiple_isoforms.fa \
     0_contigs_longest_isoforms.scaffolds \
     0_contigs_longest_isoforms.fa 0_contigs_longest_isoforms.index \
     1_contigs_longest_isoforms.fa 1_contigs_longest_isoforms.index
```

6.7) Incremental Trinity contig concatenation: Python script

The four previous steps are executed to include sequences that were not processed before until all contig sequences are concatenated.

```
$ python concatenate_gene_sequences.py \
     contigs_with_multiple_isoforms.fa \
     contigs_with_multiple_isoforms_vs_longest_isoforms.blastn \
     contigs_longest_isoforms
```

7) Functional annotation: birds

7.1) Download the latest release of the NCBI nucleotide sequence database

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz
```

7.2) Retrieve taxonomy data to build the custom database

Download taxonomy IDs related to the class Aves (birds, taxid 8782): https://www.ncbi.nlm.nih.gov/taxonomy/?term=txid8782[Subtree].

Download taxid mapping for nucleotide sequence records: ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/nucl_gb.accession2taxid.gz.

Download taxonomy information for the database: ftp://ftp.ncbi.nlm.nih.gov/blast/db/taxdb.tar.gz.

7.3) Create taxid map of Birds accession IDs: Python script

```
$ python create_taxid_map.py \
     Birds_taxids.txt nucl_gb.accession2taxid \
     Birds_taxid_map.txt
```

7.4) Parse Birds sequences from NCBI nt database: Python script

```
$ python parse_db_seqids.py \
     Birds_taxid_map.txt nt \
     nt_Birds
```

7.5) Create a BLAST database: BLAST+

```
$ makeblastdb –in nt_Birds –dbtype nucl \
     –taxid_map Birds_taxid_map.txt –parse_seqids –hash_index
```

7.6) Merge all transcripts and contigs

```
$ cat transcripts.fa contigs_with_single_isoform.fa \
     contigs_longest_isoforms.fa > all_gene_candidates.fa
```

7.7) Compare transcripts and contigs with Birds nucleotide sequences: BLAST+

```
$ blastn –query all_gene_candidates.fa –db nt_Birds \
     –perc_identity 75 –strand plus –dust no –soft_masking no \
     –outfmt "7 std qlen slen sstrand sallseqid salltitles staxids sscinames" \
     –out all_gene_candidates_vs_Birds.blastn
```

7.8) Parse BLAST hits for Birds gene assignment: Python script

The minimum percentages of identities for chicken genes and for other bird genes can be fixed by adjusting the parameters "ggal_pcid=90" and "other_pcid=75", respectively. The minimum percentages of matching cumulative length for the query and for the subject can be fixed by adjusting the parameters "q_clen=50" and "s_clen=50", respectively.

```
$ python parse_blast_hits_birds.py \
     all_gene_candidates_vs_Birds.blastn 90 75 50 50 \
     all_gene_candidates_vs_Birds.hits.txt
```

7.9) Retrieve gene information from the NCBI RefSeq database:

Download RefSeq gene report: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz.

7.10) Convert Birds nucleotide accession IDs into gene symbols: Python script

```
$ python get_gene_symbols.py \
     Birds_taxids.txt gene2accession nucl \
     all_gene_candidates_vs_Birds.hits.txt \
     all_gene_candidates_vs_Birds.assignment.txt
```

8) Functional annotation: human and mouse

8.1) Download the latest release of the NCBI protein sequence database

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz
```

8.2) Retrieve taxonomy data to build the custom database

Create a file listing the taxonomy IDs related to *Homo sapiens* (9606) and *Mus musculus* (10090) species.

Download taxid mapping for protein sequence records:
ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz

8.3) Create taxid map of human/mouse accession IDs: Python script

```
$ python create_taxid_map.py \
     Mammals_taxids.txt prot.accession2taxid \
     Mammals_taxid_map.txt
```

8.4) Parse human/mouse sequences from NCBI nr database: Python script

```
$ python parse_db_seqids.py \
     Mammals_taxid_map.txt nr \
     nr_Mammals
```

8.5) Create a BLAST database: BLAST+

```
$ makeblastdb –in nr_Mammals –dbtype prot \
     –taxid_map Mammals_taxid_map.txt –parse_seqids –hash_index
```

8.6) Extract unannotated transcripts and contigs from comparison with Birds: Python script

```
$ python extract_nonannotated_genes.py \
     all_gene_candidates_vs_Birds.assignment.txt \
     all_gene_candidates.fa \
     Birds_nonannotated_gene_candidates.fa
```

8.7) Compare transcripts and contigs with human/mouse protein sequences: BLAST+

```
$ blastx –query Birds_nonannotated_gene_candidates.fa \
     –db nr_Mammals –strand plus –seg no \
     –outfmt "7 std qlen slen qframe sallseqid salltitles staxids sscinames" \
     –out Birds_nonannotated_gene_candidates_vs_Mammals.blastx
```

8.8) Parse BLAST hits for Mammals gene assignment: Python script

The minimum percentage of identities for human/mouse proteins can be fixed by adjusting the parameters "mam_pcid=30". The minimum percentage of matching cumulative length for the query can be fixed by adjusting the parameters "q_clen=50".

```
$ python parse_blast_hits_mammals.py \
     Birds_nonannotated_gene_candidates_vs_Mammals.blastx 30 50 \
     Birds_nonannotated_gene_candidates_vs_Mammals.hits.txt
```

8.9) Convert Mammals protein accession IDs into gene symbols: Python script

```
$ python get_gene_symbols.py \
     Mammals_taxids.txt gene2accession prot \
     Birds_nonannotated_gene_candidates_vs_Mammals.hits.txt \
     Birds_nonannotated_gene_candidates_vs_Mammals.assignment.txt
```

9) Functional annotation: ORF and protein domain prediction

Remaining unassigned transcripts and contigs were annotated according to the Trinotate pipeline. Procedure including tools and database links is described at: http://trinotate.github.io/.

9.1) Extract unannotated transcripts and contigs from comparison with Mammals: Python script

```
$ python extract_nonannotated_genes.py \
     Birds_nonannotated_gene_candidates_vs_Mammals.assignment.txt \
     Birds_nonannotated_gene_candidates.fa \
     Mammals_nonannotated_gene_candidates.fa
```

9.2) ORF prediction: TransDecoder

```
$ TransDecoder.LongOrfs -t Mammals_nonannotated_gene_candidates.fa -S
```

9.3) Create the BLAST UniProt database: BLAST+

```
$ makeblastdb -in uniprot_sprot.pep -dbtype prot
```

9.4) Compare transcripts and contigs to UniProt database: BLAST+

```
$ blastx -query Mammals_nonannotated_gene_candidates.fa \
     -db uniprot_sprot.pep -strand plus -seg no \
     -max_target_seqs 1 -outfmt 6 \
     -out Mammals_nonannotated_gene_candidates_vs_UniProt.blastx
```

9.5) Compare predicted ORFs to UniProt database: BLAST+

```
$ blastp -query TransDecoder_predicted_ORFs.pep \
     -db uniprot_sprot.pep -seg no -max_target_seqs 1 -outfmt 6 \
     -out TransDecoder_predicted_ORFs_vs_UniProt.blastp
```

9.6) Create the HMMER Pfam database: HMMER

```
$ hmmpress Pfam-A.hmm
```

9.7) Pfam protein domain prediction: HMMER

```
$ hmmscan --domtblout TransDecoder_predicted_ORFs_vs_Pfam.out \
     Pfam-A.hmm TransDecoder_predicted_ORFs.pep \
     > TransDecoder_predicted_ORFs_vs_Pfam.log
```

9.8) Signal peptide prediction: SignalP

```
$ signalp -f short -n TransDecoder_predicted_ORFs_vs_SignalP.out \
     TransDecoder_predicted_ORFs.pep \
     > TransDecoder_predicted_ORFs_vs_SignalP.log
```

9.9) Transmembrane domain prediction: tmHMM

```
$ tmhmm --short TransDecoder_predicted_ORFs.pep \
     TransDecoder_predicted_ORFs_vs_tmHMM.out
```

9.10) Create gene-to-transcript mapping file: Python script

```
$ python create_gene_transcript_map.py \
     Mammals_nonannotated_gene_candidates.fa \
     Mammals_nonannotated_gene_candidates_transcript.map
```

9.11) Functional annotation and analysis: Trinotate

```
# Initialize database
$ Trinotate Trinotate.sqlite init \
     --gene_trans_map Mammals_nonannotated_gene_candidates_transcript.map \
     --transcript_fasta Mammals_nonannotated_gene_candidates.fa \
     --transdecoder_pep TransDecoder_predicted_ORFs.pep
# Load BLASTX transcript hits
$ Trinotate Trinotate.sqlite \
     LOAD_swissprot_blastx
     Mammals_nonannotated_gene_candidates_vs_UniProt.blastx
# Load BLASTP protein hits
$ Trinotate Trinotate.sqlite \
     LOAD_swissprot_blastp TransDecoder_predicted_ORFs_vs_UniProt.blastp
# Load Pfam protein domain prediction
$ Trinotate Trinotate.sqlite \
     LOAD_pfam TransDecoder_predicted_ORFs_vs_Pfam.out
# Load SignalP signal peptide prediction
$ Trinotate Trinotate.sqlite \
     LOAD_signalp TransDecoder_predicted_ORFs_vs_SignalP.out
# Load tmHMM transmembrane domain prediction
$ Trinotate Trinotate.sqlite \
     LOAD_tmhmm TransDecoder_predicted_ORFs_vs_tmHMM.out
# Export Trinotate annotation report
$ Trinotate Trinotate.sqlite report > Trinotate_report.xls
```