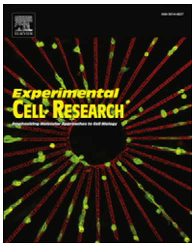
Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/yexcr



Research Article

A reference transcriptome and inferred proteome for the salamander *Notophthalmus viridescens*

Ilgar Abdullayev[a,b], *Matthew Kirkham*[a], *Åsa K. Björklund*[a,b], *András Simon*[a,*], *Rickard Sandberg*[a,b,**]

[a] *Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden*
[b] *Ludwig Institute for Cancer Research, Box 270, Stockholm, Sweden*

ARTICLE INFORMATION

Article Chronology:
Received 19 December 2012
Received in revised form
15 February 2013
Accepted 18 February 2013
Available online 28 February 2013

Keywords:
Salamanders
Transcriptomics
Regeneration
Assembly

ABSTRACT

Salamanders have a remarkable capacity to regenerate complex tissues, such as limbs and brain, and are therefore an important comparative model system for regenerative medicine. Despite these unique properties among adult vertebrates, the genomic information for amphibians in general, and salamanders in particular, is scarce. Here, we used massive parallel sequencing to reconstruct a *de novo* reference transcriptome of the red spotted newt (*Notophthalmus viridescens*) containing 118,893 transcripts with a N50 length of 2016 nts. Comparisons to other vertebrates revealed a newt transcriptome that is comparable in size and characteristics to well-annotated vertebrate transcriptomes. Identification of putative open reading frames (ORFs) enabled us to infer a comprehensive proteome, including the annotation of 19,903 newt proteins. We used the identified domain architectures (DAs) to assign ORFs phylogenetic positions, which also revealed putative salamander specific proteins. The reference transcriptome and inferred proteome of the red spotted newt will facilitate the use of systematic genomic technologies for regeneration studies in salamanders and enable evolutionary analyses of vertebrate regeneration at the molecular level.

Background

The capacity to replace lost and degenerating tissues varies between species. Aquatic salamanders in general, and newts in particular, display the widest spectrum of regeneration abilities among adult vertebrates (for a review see [1]). Although these abilities of salamanders have been known for centuries, the underlying molecular mechanisms are only now beginning to be characterized. The lack of molecular understanding of salamander regeneration hampers both the promotion of functional repair in species where it normally does not occur, as well as the posibility to address questions regarding the uneven distribution of regeneration capacities in the animal kingdom.

Several studies have significantly increased the available molecular information and tools for examining salamander regeneration during the past decade. These studies include the generation of EST databases, microarray studies, proteomic characterization, transgenic approaches, characterization of microRNAs (miRNAs)

*Correspondence to: Berzelius väg 35, 171 77 Stockholm, Sweden.
**Correspondence to: Nobels väg 3, 171 77 Stockholm, Sweden.
E-mail addresses: Andras.Simon@ki.se (A. Simon), rickard.sandberg@ki.se (R. Sandberg).

0014-4827/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.yexcr.2013.02.013

and the initiation of genome sequencing [2–10]. Despite these efforts, a comprehensive reference transcriptome for salamanders has not yet been available.

A challenge in the case of salamanders is that genome sequence data is currently sparse due to their large genome size (estimated to ~18 Gbp) and high density of repeats. However, it is now possible to reconstruct transcriptomes from high-throughput RNA-Seq data even for organisms that lack reference genome [11–13]. Sequencing of RNAs for transcriptome reconstruction has become a powerful tool for high-throughput identification of gene structures and even untranslated regions (UTRs) and non-coding RNAs that are hard to predict from comparative genomics alone.

In this work, we set out to generate a reference transcriptome for the red spotted newt, *N. viridescens*, which is one of the most commonly used model organism in regeneration research. We employed paired-end (PE) RNA-Seq and used Trinity [11] for *de novo* transcriptome assembly without a reference genome. We show that newt transcripts are of similar length and numbers as their counterparts in well-annotated vertebrate model organisms. Mining the data revealed predicted proteins ranging from evolutionarily highly conserved sequences to sequences that at present appear to be newt-specific. In addition, our study reveals novel protein domain organizations, untranslated regions, miRNA target sites, and tissue-specific transcriptome variations. This information will facilitate the discovery of molecular determinants that regulate tissue regeneration.

Material and methods

RNA isolation

RNA was isolated from 7 adult *N. viridescens*. Tissues were minced and snap-frozen before being ground under liquid nitrogen using a pestle and mortar. After thawing, the tissues were passed through a series of syringes with decreasing gauge. Total RNA was isolated using kits from Qiagen according to the instructions of the manufacturer. Briefly, tissues were placed in sample buffer, processed using shedder columns, and run through RNeasy micro kit. Sample A: Brain; B: Heart; C: Liver; D: Soft tissue; E: Upper torso and skull; F: Lower torso and tail; G: Brain (2) with mixture of intact and dopamine neuron ablated brain. The RNA quality was checked using Agilent RNA 6000 Nano Assay (Agilent Technologies), requiring RNA Integrity Number (RIN) of 8 or greater.

Library preparation

RNA-Seq library was prepared using TruSeq RNA Sample Prep Kits v2 (Illumina) according to manufacturer's recommendations combined with the dUTP method previously described [27]. Additional modifications are: first, we started with 1–5 μg of total RNA. RNA was fragmented for 1 min at 94 °C. After the first strand synthesis, the sample was purified using QIAquick PCR Purification kit (Qiagen) to remove all dNTPs and eluted into 25 μl of RNAse/DNAse free water. Different RNA Adapter Indexes were used to make indexed library. The library was amplified for 12 cycles. The resulting library was size-selected (500±25 bp) using LabChipXT (Caliper) according to manufacturer's recommendations at SciLifeLab, Stockholm. Finally, resulting libraries were PE (2 × 100 bp) sequenced using Illumina Hiseq-2000. The reported sequence read data have been deposited in Sequence Read Archive at NCBI (SRP018244)

Transcriptome assembly

Raw sequence reads were filtered for adapter containing reads and reads with low Phred scores using cutadapt (Marcel, EMBnetjournal), containing the following commands for left and right reads respectively: "cutadapt -b GTCTTCTGCTTG -b CGGCGACCACCG -O 12 –discard-trimmed –quality-base=64 -q 30 -o outfile infile" and "cutadapt -b CGGTGGTCGCCG -b CAAGCAGAAGAC -O 12 –discard-trimmed –quality-base=64 -q 30 -o outfile infile". This filtering removed 5% of the sequences (and trimmed an additional large number of reads). Transcriptome reconstructions were performed using Trinity (version 2012-04-27) on all reads from the different tissues combinations (in total 1.2 Billion pairs of sequences) requiring reported contigs to be equal or longer than 200 nucleotides. In parallel, we assembled transcripts from paired-end reads from a regenerating (dopamine neuron ablated) brain sample that was separated out due to a significant shorter insert size distribution (around 150 bp). Transcripts and inferred ORFs that were only found in the assembled transcriptome from the regenerating brain were added to yield the final newt reference transcriptome (and have component names starting with rcomp). This resulted in the reconstruction of a total of 772,128 contigs, separated in 576,061 components.

Validation of reconstructed transcriptome

We identified cDNAs and ESTs in GenBank derived from *N. viridescens*. The cDNAs were curated by selecting for those annotated as mRNAs ($n=82$). The raw EST sequences ($n=9746$) were highly redundant, and we used CAP3 to assemble the ESTs into EST clusters ($n=3572$), which consisted of 768 and 2804 singlets. Accuracy was estimated as the fraction of nucleotides correctly assembled using the best blastn alignments between Trinity contigs and GenBank cDNAs for *N. viridescens*. The set of cDNAs from GenBank was identified as all entries with cDNAs, excluding histones that lack polyadenylation tails. The accuracy per base was 97% for aligned cDNAs and ESTs, and for individual cDNAs the accuracy spanned from 80 to 100%. Length comparisons with other species (Fig. 1A) was based on Ensembl v67 transcripts for human ($n=176{,}981$), mouse ($n=88{,}943$), zebrafish ($n=49{,}123$) and frog ($n=22{,}878$).

Cloning and sequencing of cDNAs

N. viridescens cDNA generated for Annexin 1, Jarid2, FGF2, Sox1, Shh, ODC1, nRad and DR2 have been described previously [6,28]. cDNA was generated from whole brain or whole limb RNA using poly T primers and superscript III (Invitrogen). Primers were designed from transcriptome data except for CKM for which degenerate primers were used. PCR was performed using platinum tag (Invitrogen), and PCR fragments cloned using TA cloning kit (Invitrogen). Sequencing was performed by MWG, using Sp6 or T7 primers. The following primers were used to amplify cDNA fragments: GAPDH for- AAG AAC GTG ACC CCA CCA ACA T, rev-

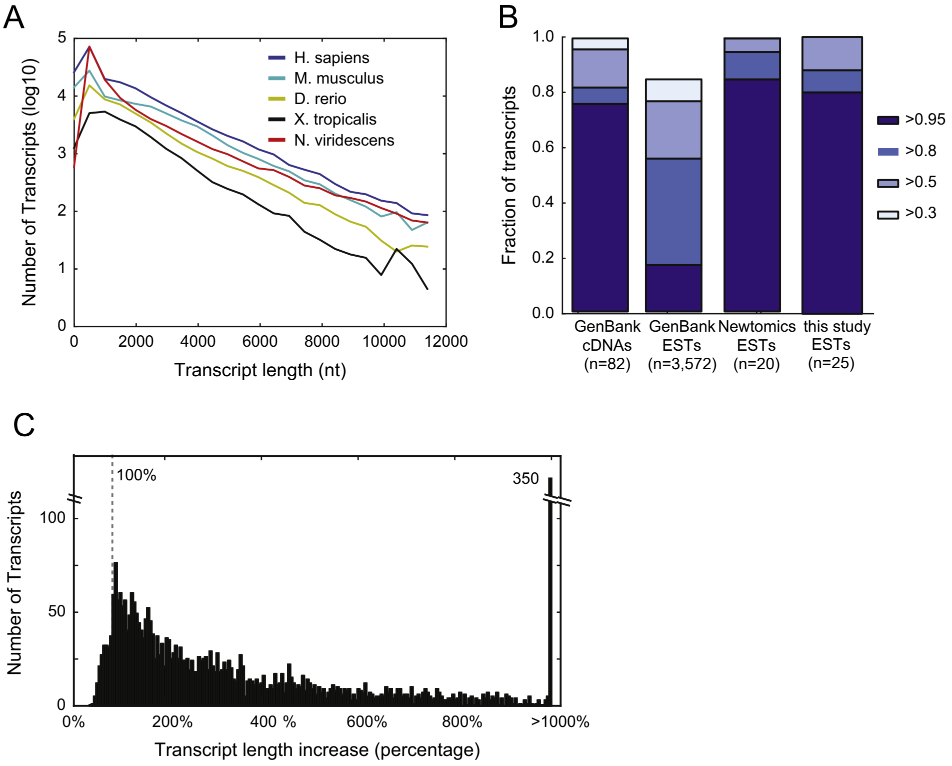


Fig. 1 – Validation of the reference transcriptome. (A) Length distributions of human, mouse, zebrafish and frog Ensembl transcripts together with reference newt transcripts. (B) Bar plot showing the fraction of newt transcripts in reference databases aligned by one or more reconstructed newt transcripts. Reference cDNAs (GenBank), ESTs (GenBank and Newtomics) and in-house sequenced EST clones were analyzed, and coloring indicates the fraction of cDNA or EST length that aligned to reconstructed newt transcript. (C) Number of reconstructed transcripts as a function of the percentage increase in transcript length, calculated as the reconstructed transcript length over matching cDNA or EST sequence length in GenBank.

CAG CAG CTG CCT TTA CCA CCT T, Actin for- ATG AAG GTT ATG CCC TGC CTC A, rev- GAG CCA CCG ATC CAA ACA GAG T, glutamine synthetase for- CTG GAA TGG TGC AGG ATG TCA C, rev- GCC ATG AGA AAG CCA AGC AAC T, BmpR2 cDNA1 for- GTG CAG TGG CAG TGG CCC AA, rev- GTT GAG GGG CGC CAC CGT TT, BmpR2 cDNA2 for- TGG GAC GTG CGT CAT CAG CG, rev- TGG GAC TCG CCT GCT GTG TG, Hes1 for- CTG GGC CAC TTG TCC GGC TG, rev- ATT GGT CGC CGA GTG CCA CG, interleukin 6 receptor for- CGC GGT TGC CAA TGG TGC AG, rev- TGT CCG GTG CCT TTG CAG GG, Jak1 for-CTC CCG CGG CAC AAG TCG AG, rev-CTG GGC GTG GTA GGC GCT TT, Mind bomb1 cDNA1 for- TGC CAT CCC AGC CTT CAG GA, rev- TCC CTG GGT GCC AAG TCC CA, mind bomb1 cDNA2 for- CCT GCA GCA GTC TCT AAA GTC GCA, rev- TCC TGA AGG CTG GGA TGG CA, Numb for- GTA GGC CGC ACC AGT GGC AA, rev- TGC TTG CTC CGT CGC TGT GG, PTRF for-ACG TCA AGA CGG TGC GGC AG, rev- CCG CCG TCT TTG ACC TGG CA, Notch 1 for-CTG GTG AAC CGC AAG CGG CA, rev-TTT AAA TGC CTC TGG GAT GT, CKM for- CTT CCT GGT STG GGT MAA YG rev- CTC AAT CAT GAG YTT CAC RC, Bcl2 for- GAG TTT GGT GGA GCC CTA AG, rev- AAG CTC CAA GGA CAG CCA GG, Casp3 for-AAT GGC GGA TAC AAA TGA CT, rev- ATC TTC GCC GTG GCT TAA CA, hif1a for- ACT ATC CGG GCT TGC TTT CC, rev- AGG TCT TAA AGA CTG CCC AG, MYOG for- TAT GCA GAC AGC CTG CCT GA, rev- CAG AGC ACC AGT TCT TGT AA, XIAP for- TTG CTA ATT TCC CTG GCA GT, rev- TCT CTC CTT GGG ATC CAT TG, ID2 for ATG AAA GCT TTT AGC CCG GT, rev GGA TAG AGT GGC TTG CTG GG.

Computational analyses of newt ORFs

We used the Trinity utility (transcripts_to_best_scoring_ORFs.pl) to predict ORFs of at least 100 amino acids from the reconstructed contigs, resulting in the prediction of 57,035 best candidate ORFs with mean ORF length of 452 amino acids (N50=750 aa). We aligned the newt ORFs to Swiss-Prot (Release 2012_01) and TrEMBL (Release 2012_01) databases using blastp within BLAST (Release 2.2.21). We filtered alignments to only consider matches with an *E*-value below 10^{-5}. In addition, we downloaded all Ensembl v70 protein sequences for human ($n=104{,}785$), mouse ($n=50{,}877$), zebrafish ($n=42{,}157$) and frog ($n=22{,}705$). For the comparison, we identified the longest protein isoforms for human ($n=23{,}287$), mouse ($n=22{,}796$), zebrafish ($n=26{,}163$) and frog ($n=18{,}429$) and compare those with the longest newt ORF sequence ($n=29{,}316$).

Ortholog analysis

We identified orthologs to all newt ORFs using Inparanoid [16] and the set of all ORFs with unique DAs (36,406 proteins). The longest protein from each gene (Ensembl) was used for the other species, human, chicken, zebrafish and frog (*Xenopus tropicalis*). Summary of the ortholog statistics can be found in Supplementary Table 4. A set of 1-1 ortholog groups was defined only when all 5 species had 1-1 relations between all the proteins, which gave 3752 ortholog groups.

Domain annotation and ORF filtering

We identified domains using HMMER 3.0 (R.D. Finn 2011) and Pfam-A domains, requiring an E-value < 0.001, while allowing for lower E-values in repeats as previously described [29]. Of the 56,700 unique newt ORFs, 35,644 had one or more significant Pfam-A domain. To define a set of newt ORFs without including several isoforms from the same gene, the ORFs from the same Trinity component were filtered according to the following criteria: (a) when no domains were found, the longest ORF was used, (b) if the domain architectures (DAs) were identical or overlapping Pfam-A or Pfam Clan domains, the longest ORF from the longest DA was used, (c) if non-overlapping DAs were found, the longest ORF from each of those DAs were used. This rendered a set of 36,406 putative newt ORFs. Next we identified putative transposable element (TE) ORFs using TransposonPSI (http://transposonpsi.sourceforge.net) and ORFs with the PfamA domain PF02994 (L1 transposable element) or CL0418 (GIY-YIG endonu clease superfamily) in its sequence or in the sequence of its closest blast hit. This procedure identified 3282 ORFs as TE. Finally, we identified ORFs from possible contaminants, such as bacteria, viruses and parasites. All ORFs with closest blast hits to a prokaryotes or viruses were considered contaminants. Among the 1417 bacterial hits, 71% were different species of the sea-living Flavobacteria. For ORFs with closest blast hit to an invertebrate eukaryote, the expression of their transcripts in the different tissues was examined. We found that transcripts related to certain species were only found in or more abundant in the soft tissue sample that contains intestines. Thus we defined contaminant eukaryotes as sequences with > 1 fold higher expression in soft tissues and this gave a set of 4207 ORFs. Non-coding transcripts from transposable elements (TE) were identified using TransposonPSI (http://transposonpsi.sourceforge.net). We detected TEs in 6689 components (6.6%) that were removed from further analysis.

Domain analysis

To define the phylogenetic positions of DAs their presence among different phyla and classes were analyzed. DAs found in both eukaryotes and prokaryotes were defined as ancient, DAs exclusively found in amphibians as one group and DAs found in vertebrates, metazoans or other eukaryotes in that order. Next we compared the newt DAs to all the UniProt DAs considering a match also when the DA in newt was shorter but completely contained within the other DA. Thus all proteins and DAs in newt were assigned an evolutionary position, or defined as novel if no match was found. We also assigned Pfam-A domains in frog (*X. tropicalis* JGI 4.2, Ensembl v.68) proteins and repeated the same analysis as for newt ORFs. Next, we examined the candidate newt-specific proteins, defined as ORFs with a novel DA that had no match in UniProt. 118 proteins with 118 newt-specific DAs were identified (Supplementary Table 6). These DAs were also compared to the DA of the five closest blast hits, and the blast hit with the most similar DA was used to define the number of additional domains in each newt-specific DA.

Identification of ncRNAs

We identified 78,812 contigs with no ORF (less than 15 a.a.) and 14,540 of them had minimum 1 FPKM of expression. Afterwards we performed identification of non-coding RNAs according to "Annotation of Non-Coding RNA", Ensembl. We identified 1419 contigs (out of all 78,812) aligning to 314 Rfam families (database version 10.1) with less stringent criteria of $E \leq 0.01$. By using 1419 contigs, we identified 310 contigs aligning to 226 Rfam families using Infernal ("INFERence of RNA ALignment" version 1.1rc1) software for searching RNA structure and sequence similarities by seeding covariance models properly parameterized for Infernal v1.1rc1. Finally, we identified 66 micro-RNAs, 48 Small nucleolar RNAs and 33 tRNAs by following these procedures.

Analyses of 3′UTRs and microRNA target sites

We filtered all contigs for those with an ORF that ends with a stop codon and identified 17,226 non-redundant 3′UTRs. We used the ortholog mapping software Inparanoid [16] to identify 1-to-1 orthologs in human and we identified 3608 such pairs. MiRNA target prediction was performed using Miranda (version 3.3a) with strict 5′ seed pairing. The numbers of predicted miRNA target sites were normalized according to 3′UTR length to report on microRNA sites per kb of UTR.

Identification of alternatively spliced transcripts

Genome-independent transcriptome reconstructions using Trinity group transcript isoforms into subcomponent and component structures but provide no information on how the transcript relates (except the information in the splicing path graphs). In order to identify alternatively included exons resulting from exon skipping event, we compared the splice graph information pair-wise for all transcripts within the same Trinity subcomponent. We required the variation to be internal (not considering variations in first or last exons) and that the variable included portion of transcript to exceed 50 nts. This threshold was selected fairly arbitrarily to remove variations due to polymorphisms but to keep sensitivity for identification of alternative exons. Through this procedure we identified 4709 alternative exons. Likely there are alternative exons that are shorter than 50 nt and escape our identification. All identified exons are listed in Supplementary Table 3.

Expression level analyses

We aligned reads back to the reconstructed transcripts using the Trinity script (alignReads.pl) with the Bowtie aligner. Next, we used RSEM [21] to estimate expression levels as fragments per kilobase and million mappable reads (FPKMs) using the scripts (run_RSEM.pl and summarize_RSEM_fpkm.pl) available within the trinity software package. We used EdgeR [30] to identify significantly (FDR < 0.05, fold-change > 2) differentially expressed transcripts using the tissue or tissue combination data. The brain enriched transcripts used for analyses in Figs. 3b and 4a was identified as highly expressed in the brain (> 10 FPKM) and with low expression in other tissues (< 1 FPKM). Hierarchical clustering in Fig. 3a was generated using R/Bioconductor using the gplots package on newt transcripts with largest variation across samples.

Gene Ontology analyses

We used the DAVID Bioinformatics Resources 6.7 for Gene Ontology analyses of protein sets [31]. We identified significant Gene Ontology categories enriched or depleted (FDR$<$5%) among newt proteins with stringent match in human, based on the human orthologs gene names. Analyses of brain transcriptome used the UP_TISSUE categories within DAVID to find the tissues with a significant enrichment. Gene Ontology annotations in Supplementary Table 5 were inferred from the closest blast hit ($E<10^{-5}$) with annotations in Uniprot.

Identification of cell cycle inhibitors

We mapped newt ORFs against Swiss-Prot and TrEMBL, requiring a blastp E below $1e^{-20}$. If the newt ORF matched a cell cycle inhibitor in Swiss-Prot, it was designated as a Swiss-Prot match. Those with no match in Swiss-Prot but with one or more match ($E<10^{-20}$) in TrEMBL was designated as a TrEMBL match in Fig. 5. For other vertebrate species, we determined the presence or absence of a cell cycle in each vertebrate through requiring them to have an annotated gene name in each particular species in Swiss-Prot, TrEMBL or Ensembl databases. Note that the lack of a certain cell cycle inhibitor in any particular organism could be due to incomplete annotation in Swiss-Prot rather than absence of corresponding protein. Finally, although we identified most inhibitors in the newt, it is possible that missed inhibitors have more restricted expression patterns that precluded their identification in our samples.

Results

Generation of the newt reference transcriptome

To construct a comprehensive reference transcriptome for the newt, we purified total RNA from individual tissues and combinations of tissues. From each RNA preparation, we generated strand-specific PE sequencing libraries using the dUTP method [14]. Fragments were sequenced 100 base pairs (bp) from both ends resulting in a total of 1.2 Billion sequence reads. The mean fragment lengths of these libraries were ~300–350 bp (Supplementary Table 1), thus highly suitable for *de novo* transcript assembly. We preprocessed the sequenced reads to discard those with adapter matches and read ends with lower quality. We performed several transcriptome assemblies using Trinity [11] on each sample in parallel or combined in one assembly. The reconstructed transcriptome that was derived from a combination of tissues rendered the longest transcript contigs (Supplementary Table 2), and it was therefore chosen as the reference newt transcriptome throughout this study.

Validation of the reconstructed transcriptome

The newt reference transcriptome consists of 118,893 transcript sets (i.e. Trinity components, removing contaminant species and transposons, described below), with 48,503 transcripts of 1 kb or longer, in line with what we expected in terms of length. Through cross-species comparisons, we found that the newt reference transcriptome had a length distribution similar to human and mouse transcriptomes and that the number of transcripts exceeded the data available for *X. tropicalis* (Fig. 1A).

We assessed the completeness, accuracy and contiguity of the reconstructed transcriptome [15] through comparisons with existing nucleotide data for *N. viridescens*. Existing transcript data in GenBank consisted of 82 full-length or partial cDNAs (GenBank cDNAs) and 3572 non-redundant EST sequences (GenBank ESTs, see Methods). In addition, we manually downloaded a random subset of 20 EST sequences from a recent study (Newtomics ESTs [4]), since systematic download and computational analyses of these data were not feasible. We mapped the reference transcriptome to the cDNA and EST data sets, respectively. To determine the completeness of the reconstructed transcriptome, we analyzed the percentage of GenBank cDNAs and ESTs that had one or more significant match in the reconstructed transcriptome. We found that the newt reference transcriptome contains transcripts for all previously identified GenBank cDNAs and Newtomics ESTs, and 85% of the ESTs in GenBank (Fig. 1B). To address the length of the reconstructed transcripts, we computed the fraction of the reference transcript length (in GenBank cDNAs and ESTs) that was aligned by a reconstructed contig. For 76% of the cDNAs and 85% of Newtomics ESTs, one reconstructed transcript covers more than 95% of each sequence length (Fig. 1B). The fraction of sequences with more than 95% coverage was lower for GenBank ESTs possibly due to low quality sequences at the EST ends. To independently validate the newt transcriptome, we selected 25 transcripts for cloning and Sanger sequencing. For all new sequence data we found high agreement with the reference transcriptome (Fig. 1B), with a single assembled transcript spanning the full clone. The reconstructed transcripts were, on average, 5-fold longer than existing sequence data in GenBank (Fig. 1C), demonstrating that the reference transcriptome has significantly improved the length of the cDNAs and ESTs that were previously present in GenBank.

Using the alignments of reference transcript to GenBank cDNAs and ESTs we computed base-level identities that ranged between 97 and 99% for ESTs and cDNAs. Comparisons of cDNA alignments to closely related salamanders (*salamandridae family*) revealed DNA sequence identities of 91% to Japanese fire belly newt (*Cynops pyrrhogaster*), 89% to Iberian ribbed newt (*Pleurodeles waltl*), and a somewhat lower identity of 86% to Axolotl (*Ambystomatidae family*).

The reference transcriptome had, on average 1.34 transcript variants per gene, either due to alternative RNA splicing or polymorphic sites. To identify alternatively spliced transcripts that result from exon skipping, we compared transcript variants for internal deletions and found 4709 internal deletions with a deletion of 50 bp or longer (Supplementary Table 3), a cutoff that we selected to remove variations due to sequence polymorphism. The mean length of these putative alternative exons was 150 bp, similar to exon lengths in mammals. These results also suggest the existence of large numbers of alternatively spliced transcripts, indicating that the reference transcriptome is rich in transcript isoforms. Together, we conclude that we have generated a comprehensive reference transcriptome of the red spotted newt.

Characterization of inferred newt protein sequences

We translated each reconstructed transcripts into its longest putative ORF, requiring a length of at least one hundred amino acids (Methods). This yielded an inferred proteome of 57,035 ORFs of which 43% were complete (containing both start and stop codons), 33% partial (having either a start or stop codon) and remaining ORFs had neither start nor stop codon (Table 1). Although Trinity groups isoforms and close paralogs into components and subcomponents, we found that different ORFs with distinct domain architectures were sometimes grouped into the same component. Therefore, we compiled the longest ORF with distinct domain architecture from each component. Next we determined the completeness of the translated newt proteome through analyses of newt orthologs in other vertebrates. We used Inparanoid [16] to detect orthologs to human, chicken, zebrafish and frog (*X. tropicalis*). The number of orthologs and ortholog groups for the newt were similar to those of the other vertebrates (Supplementary Table 4). For instance, the human proteome has 17,097 proteins with homologs in the newt, similar to the numbers of human proteins with a homolog in chicken (17,121), zebrafish (16,580) and frog (16,719). Next, we identified 1-1 orthologs in all 5 species (3752 ortholog groups) in order to determine whether the newt proteins were of comparable length as their counterparts in the other species. We observed that in 85% of the ortholog groups, the newt protein was longer than the shortest representative from the other four species. In fact, the newt proteins were often longer than the mean length of orthologs from the other four species (Fig. 2A). Only human proteins had larger average length than the newt proteins. These results demonstrated that the automated translation of transcripts into protein sequences yielded a high fraction of full-length proteins.

Table 1 – Summary of inferred proteome size and annotations.

Dataset	Number	Percentage
Total number of ORFs >100aa	57,035	
Complete ORFs	24,703	43.3%
Partial ORFs	18,521	32.5%
Internal ORFs	13,811	24.2%
ORFs with unique DA *	36,703	
ORFs with UniProt match	26,373	71.9%
ORFs with Pfam-A domains	21,324	58.1%
Transposable elements	3284	9.0%
Contaminant species	4283	11.7%
Complete ORFs	14,389	39.2%
Filtered ORF set **	29,316	
ORFs with UniProt match	19,410	66.2%
ORFs with Pfam-A domains	16,663	56.8%
ORFs with Pfam-A or UniProt match	19,903	67.9%
Complete ORFs	13,263	45.2%

* Initial dataset with predicted ORFs based on unique domain architectures (DA) in each component.

** Dataset after removing detected contaminant species and transposable elements.

Annotation of the newt proteome

Within the translated ORFs we observed high redundancy in protein sequences coming from the translation of several transcript isoforms of the same gene. We restricted the subsequent analyses to the 36,703 ORFs with a unique domain architecture (Table 1) to effectively remove isoforms that do not alter protein domains. To provide an initial annotation of the newt proteome, we mapped the protein sequences to the UniProtKB (UniProt) protein database using protein BLAST (blastp). We found that 72% (26,373) of the ORFs had a match in UniProt ($E<10^{-5}$), including matches to all sixteen N. viridescens proteins present in Swiss-Prot (Table 1). Using the annotations for the closest UniProt match to our ORFs, we could exclude 12% of the newt ORFs (with matches to bacteria, virus and metazoan parasites) as being derived from putative contaminant species. The UniProt alignments also revealed that many of the short and incomplete ORFs were mapping to transposable elements (TE), in particular to long interspersed elements (LINEs) and Gypsy elements. Using the TransposonPSI program to detect TEs we annotated 9% of the newt ORFs as being derived from TEs (Table 1), a number that is likely to be an underestimation since TE detection is difficult for short, fragmented ORFs. Studies in other salamanders found TEs to be abundant [17] and that they are re-activated in regenerating limbs of Axolotl [18]. Analyses of the remaining proteins (76%, 29,316) revealed that the newt proteins had similar length as the closest match in UniProt, whereas the ORFs mapping to contaminant species or TEs were significantly shorter than the match in UniProt (Supplementary Fig. 1). In subsequent analyses we excluded ORFs (and transcripts) derived from contaminant species or TEs. The remaining dataset contained 29,316 ORFs of which 19,903 had a Blast hit to UniProt or a PfamA domain and can be considered a high confidence set of newt proteins (Table 1). All protein annotations are provided in Supplementary Table 5 with Gene Ontology annotations inferred from closest blast hits. The length distribution of the high confidence set of newt proteins was similar to other vertebrate proteomes (Fig. 2B).

Evolutionary analyses of newt proteins and domains

We compared the amino acid identity between the red spotted newt and other salamanders and vertebrates and found highest identity to the Japanese fire belly and Iberian ribbed newt proteins (both 90%), followed by 82% amino acid identity to Axolotl (*Ambystomatidae family*) and 76% to frog (*X. tropicalis*). More distant vertebrates had identities in the range of 65–70%. Next, we compared the domain content and architecture in the newt proteome to those in the frog, the closest relative with sufficient gene annotation for a systematic comparison. A similar number of proteins in the newt and frog had one or more curated protein domain (PfamA) (Fig. 2C), but the newt had a larger number of ORFs with no curated domain. We conclude that the newt proteome is complete, although with an additional set of shorter ORFs of unknown origin.

To identify the phylogenetic positions of newt proteins we investigated their domain architecture (DA) and compared them to DAs present in proteins across different phyla and classes in UniProt. The analysis revealed that 60% of all newt proteins contained an ancient DA found in both prokaryotes and

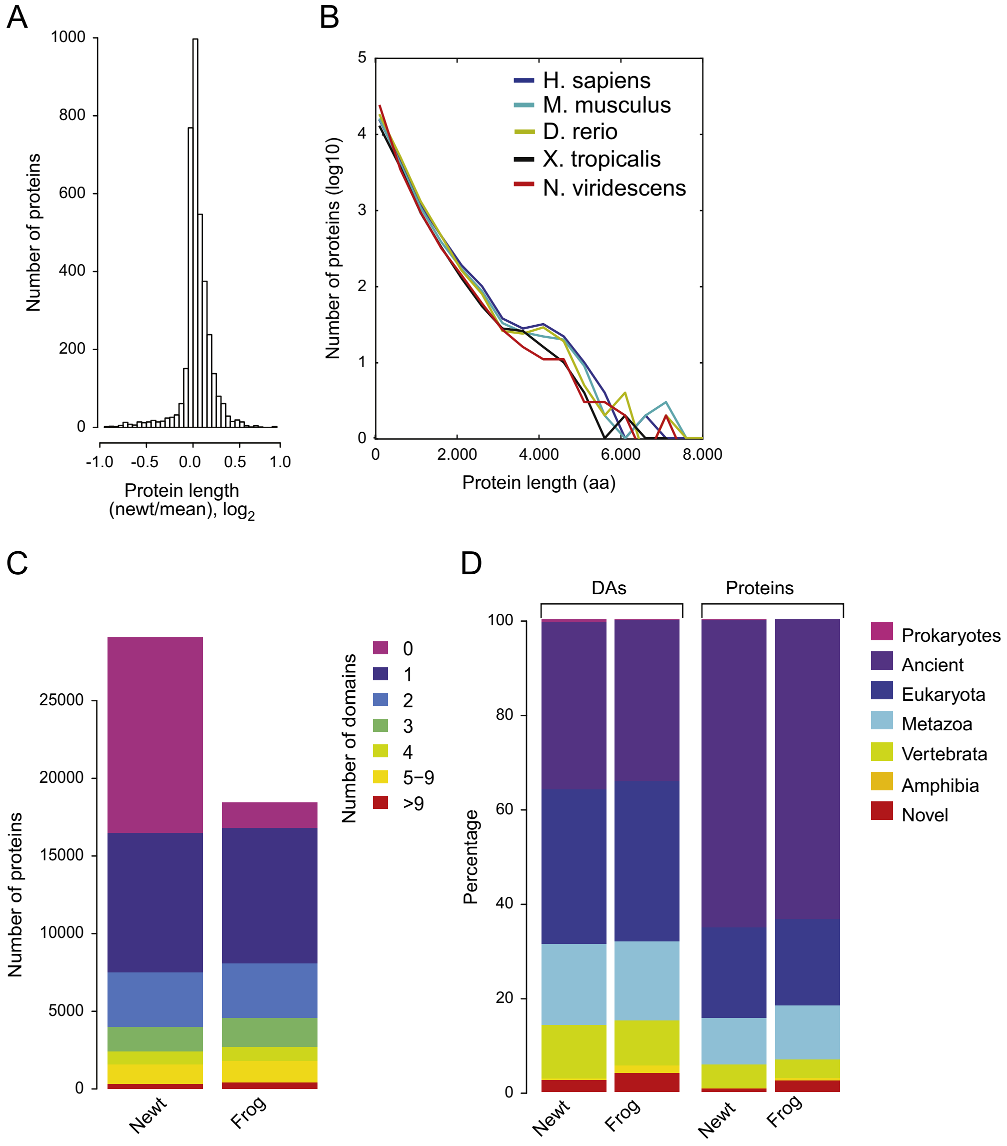


Fig. 2 – Analyses of inferred reference proteome. (A) Length distributions of the longest Ensembl protein isoform (per gene) for human, mouse, zebrafish, frog and the longest inferred newt protein isoform (per trinity component). (B) Histogram of length differences for 1-1 orthologs between newt ORFs and the mean length of proteins in ortholog groups. (C) Bar plot showing the number of newt and frog proteins with different number of PfamA domains. (D) Bar plots showing the phylogenetic positions of newt and frog proteins and DAs. Prokaryotes: only found in prokaryotes, Ancient: found in prokaryotes and eukaryotes, Eukaryota: only found in eukaryotes, Metazoa: only found in metazoans, Vertebrata: only found in vertebrates, Amphibia: only found in amphibians, Novel: only found in the newt or the frog, respectively.

eukaryotes and that less than 20% of the newt proteins (and 35% of the DAs) contained DAs only present among metazoans, vertebrates or amphibians (Fig. 2D). In general, newt proteins had similar phylogenetic positions as the frog (Fig. 2D) and few newt proteins (14) contained DAs that were specific to the amphibian lineage. Interestingly, we identified 118 unique ORFs with putative newt-specific DAs (i.e. DAs that did not occur in any protein in UniProt). Most of these DAs only have one or two additional domains (70% and 23%, respectively) compared to the domain content of the closest blast hit (Supplementary Table 6). This agrees well with previous findings that novel domain combinations mainly evolve through the addition of a single domain to existing DAs [19].

Identification of non-coding transcripts

To extend the annotated reference transcriptome to non-coding RNAs, we identified newt transcripts without any (even low stringency) match in Swiss-Prot and TrEMBL. These putative non-coding RNAs (ncRNAs) were aligned to Rfam using nucleotide BLAST (blastn), which resulted in 310 transcripts mapping to 226 Rfam families (requiring $E<10^{-15}$) that were subsequently analyzed using Infernal [20]. This procedure annotated a subset of the ncRNAs as miRNA precursors ($n=66$), small nucleolar RNAs (snoRNAs, $n=48$) and transfer RNAs (tRNAs, $n=33$). In addition to these sets of ncRNAs, we anticipate the existence of many long ncRNAs (e.g. lincRNAs) in our data, although the lack a newt reference genome precludes analyses of their exon-intron structures.

Brain-specific newt transcriptome

Although our selection of tissues and tissue combinations for RNA-Seq was based on obtaining good coverage across many different tissue types, it could also be used to start surveying tissue-specific regulation in gene expression. We re-aligned our sequenced reads to the reconstructed transcripts to estimate gene expression values using RSEM [21]. We found that 5145 genes with significant tissue variation in expression (FDR$<$0.00001, fold-change$>$2, min expression$>$10 FPKM) across tissues (Fig. 3A). Zooming in on expression differences in the brain compared to other tissue combinations revealed 940 ORF encoding transcripts with significantly higher expression in the brain. The human orthologs of these ORFs were significantly enriched for genes with known brain functions and had annotated brain-associated expression patterns (Fig. 3B). We also detected a large number (363) of brain-specific transcripts longer than 1 kb but with no ORF spanning more than 25% of the transcript length, suggestive of tissue-specific ncRNAs in the newt brain. These results indicate that the newt transcriptome is rich in tissue-regulated transcripts.

Untranslated regions and microRNA target sites

We next filtered for complete 3′ UTRs in the newt by only considering reconstructed transcripts with both a defined stop codon and polyadenylation signal with the last 50 nucleotides from the 3′ end (extended to allow for a partial poly(A) tail within the end of reconstructed transcripts). Through this procedure, we identified 5452 3′ UTRs with a mean length of 1.5 kb (Fig. 4A). Similarly to other species, transcripts specifically expressed in the brain utilize longer 3′ UTRs [22] (Fig. 4A), with a mean length of 2.2 kb in the newt. Genome sequence data from a closely related salamander (*Ambystoma mexcanum*) with comparable regenerative abilities identified a higher density of putative miRNA sequences in introns compared to humans. This suggested that salamander-specific biological functions, such as regeneration, might depend on an extended role of miRNA-mediated regulation [7]. In conjunction to this notion, we identified the subset of newt 3′UTRs with a direct homolog in human (3605 3′ UTRs) and compared their occurrence and density of target sites for miRNAs conserved among vertebrates. We found that newt and human 3′ UTRs had similar density of miRNA sites per kb of UTR (Fig. 4B). As the analysis only included the most evolutionary conserved miRNAs, mammalian or primate specific miRNAs were excluded, as well as effects of targeting through unknown newt-specific miRNAs.

Cell cycle regulators in the newt

Reentry to the cell cycle from the postmitotic state is a distinctive feature of newt regeneration [1] exemplified by pigmented epithelial cells of the eye and terminally differentiated myotubes. It was hypothesized that the evolution of the cell cycle inhibitor INK4A locus occurred at the expense of this plasticity [23]. Interestingly, our newt inferred proteome captured many cell cycle regulators, including a transcript with near equal homology to p15 and p16 (Fig. 5). However, we did not find a transcript corresponding to alternative reading frame (ARF) in the newt. Although our gene and protein catalogs corroborate the lack of ARF in the regenerative newt, we note that ARF is also missing in frogs.

Discussion

A prerequisite for functional analyses of salamander regeneration at the molecular level is to substantially increase the available genomic information about these animals. One of the frequently

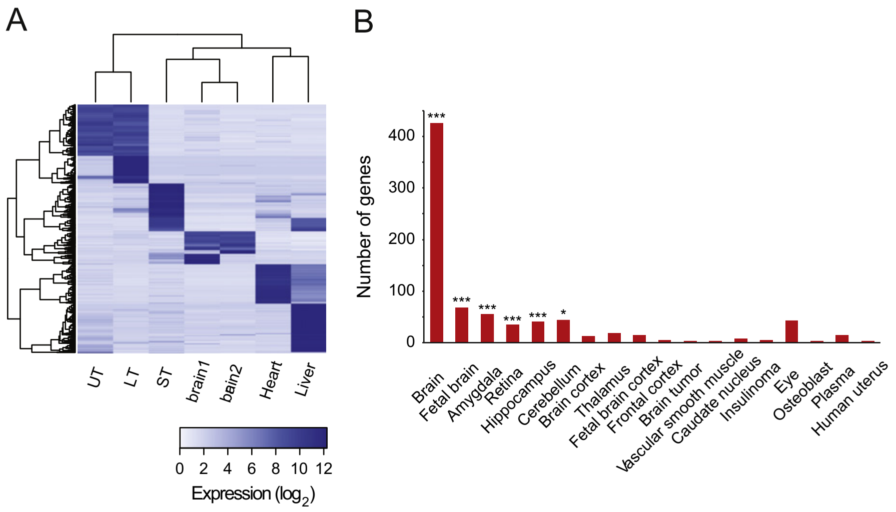


Fig. 3 – Tissue variations in the newt transcriptome. (A) Hierarchical clustering of 765 newt transcripts with largest significant variations in expression levels across samples (max expression$>$50 FPKM, lowest expression below 2 FPKM, FDR$<$$e^{-10}$). Expressions levels are shown as $\log_2$ transformed FPKM values. The sample code is UT: Upper torso and skull; LT: Lower torso and tail; ST: Soft tissue; Brain1: normal brain; Brain2: mixture of dopamine neuron ablated and normal brains. (B) Human orthologs of 622 newt transcripts with high expression in the brain ($\geq$10 FPKM) and low levels in other tissues ($\leq$1 FPKM) were significantly enriched for brain tissue categories in DAVID (*** adjusted *p*-value (*p*)$<$$e^{-10}$, ***p*$<$e-5, **p*$<$0.05).

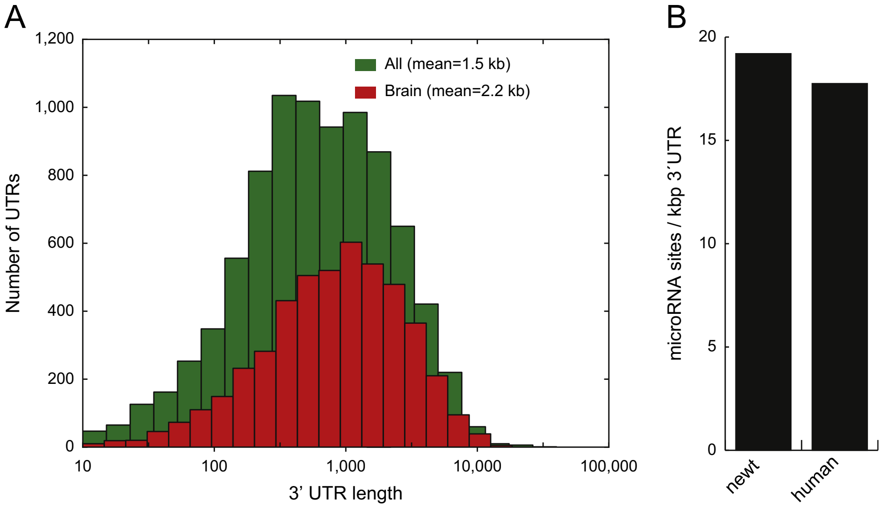
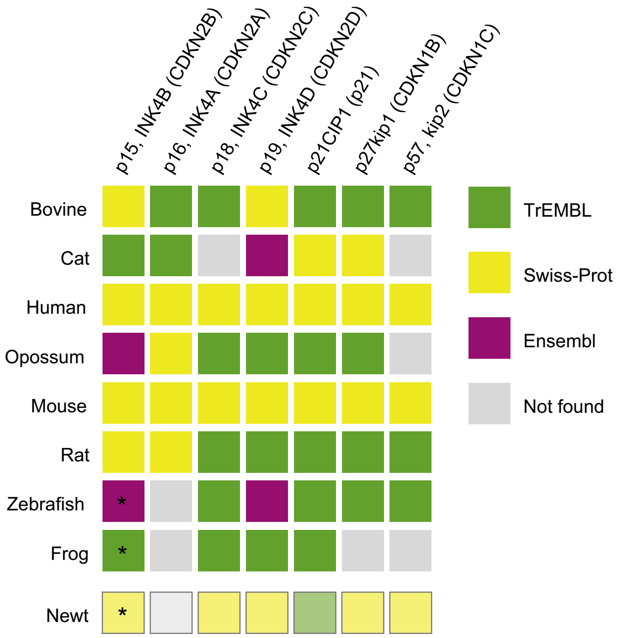


Fig. 4 – Analyses of UTRs and miRNA target sites. (A) The 3′ UTR length distribution of all (5452) and brain-specific (390) newt transcripts with apparent ORF stop codon and poly(A) signal. (B) Density of miRNA target sites (per kb) within orthologous newt and human 3′ UTRs.



	p15, INK4B (CDKN2B)	p16, INK4A (CDKN2A)	p18, INK4C (CDKN2C)	p19, INK4D (CDKN2D)	p21CIP1 (p21)	p27kip1 (CDKN1B)	p57, kip2 (CDKN1C)
Bovine	🟨	🟩	🟩	🟨	🟩	🟩	🟩
Cat	🟩	🟩	⬜	🟪	🟨	🟨	⬜
Human	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Opossum	🟪	🟨	🟩	🟩	🟩	🟩	⬜
Mouse	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Rat	🟨	🟨	🟩	🟩	🟩	🟩	🟩
Zebrafish	🟪 *	⬜	🟩	🟪	🟩	🟩	🟩
Frog	🟩 *	⬜	🟩	🟩	🟩	⬜	⬜
Newt	🟨 *	⬜	🟨	🟨	🟩	🟨	🟨

🟩	TrEMBL
🟨	Swiss-Prot
🟪	Ensembl
⬜	Not found

Fig. 5 – Cell cycle inhibitors. Listing of cell cycle inhibitors conserved amongst different vertebrate organisms, with official gene names within parenthesis. Presence of proteins in Swiss-Prot (yellow), TrEMBL (green) or Ensembl (purple) databases of other species are color-coded. Matches between the inferred newt ORFs to Swiss-Prot and TrEMBL (blastp, $E < e^{-20}$) are colored yellow and green, respectively. Asterisks indicate that the prototypic p15/p16 in the Zebrafish, Frog and the Newt only have a marginally higher similarity to p15 over p16.

used species for regeneration studies among salamanders is the red spotted newt, for which we here present a reference transcriptome and inferred proteome with comparable size and completeness as other well-annotated mammals. Indeed, the validations of the transcriptome and proteome all indicate a reference catalog of high completeness, accuracy and contiguity making this a comprehensive resource for researchers interested in regeneration biology or comparative genomics.

In this paper we examined the newt transcriptome based on a selection of questions, out of which we emphasize three that have extensively been discussed recently. First, we see that although most transcripts (65%) encode proteins that are evolutionarily ancient (found in both prokaryotes and eukaryotes), there are encoded proteins, for which orthologs only exist in some animal groups but not others. For example, 3579 proteins were conserved also in mammals, whereas only 14 proteins were specific to the amphibian lineage. Such sequences constitute only a smaller fraction of the transcripts, but may represent valuable contribution to our perception of vertebrate evolution. We also found a large number of transcripts with no known homologs in other organisms. The possibility that, in addition to species-specific gene regulation, also species-specific genes contribute to salamanders' unique regeneration capacities has been proposed (for reviews see [24,25]). Noteworthy in this context is that among the newt-specific ORFs with novel domain combinations, we found one transcript (contig 6883) which is up-regulated during regeneration of dopamine neurons in the adult newt midbrain [6]. This is in agreement with the notion that species-specific components could be involved in unique regeneration capabilities. Second, we looked at 3′UTRs in terms of the number of putative miRNA target sites. Initial characterization of genic regions of *Ambystoma mexicanum* indicated that extensive miRNA mediated gene regulation may be a distinctive feature of salamander regeneration [7]. While only functional experiments could test this hypothesis conclusively, the newt transcriptome does not provide support for this, at least based on the number of putative miRNA target sites in the 3′UTRs. Third, a search for cell cycle inhibitors resulted in the identification of many known components of the regulatory network, such as p18, p19, p21, p27, p57 and a prototypic p15/p16 but no ARF. Hence our findings support the view that ARF may be involved in maintaining postmitotic arrest in differentiated cells [26]. However, it remains to be seen to what extent cell cycle reentry by postmitotic cells is indeed required for salamander regeneration in general.

For organisms without a sequenced reference genome, *de novo* transcriptome reconstructions can provide both comprehensive transcript and protein sequence information and reveal evolutionary distances to sequences of closely related species. The newt reference transcriptome was assembled using 600 M PE reads from 5 lanes on a HiSeq 2000, but we noted that a high fraction of reference transcripts was already correctly assembled using data from a single HiSeq 2000 lane. However, multiple libraries from different tissue sources captured higher transcript diversity, indicating that multiplexing of multiple RNA-Seq in one or two Illumina lanes might provide enough data for successful reconstructions. In addition, we observed that reconstructions from PE (100 bp) RNA-Seq libraries were highly dependent upon the fragment lengths used, as an initial PE library with mean fragments around 150 bp failed to give transcriptome of expected lengths.

As genome-independent transcriptome reconstructions are becoming more refined, it will be important to develop approaches that automatically catalog transcriptome isoforms that results from alternative promoter usage or RNA processing. In this study, we predicted a large number of putative skipped exons through comparisons of transcriptome isoforms, and similar procedures could be used to catalog other variations. Ultimately, it would be informative to compare isoform quantifications based on genome-guided and genome-independent transcriptome reconstructions.

In addition to the above-presented analyses, the data can be mined in many additional ways, and is freely downloadable from our server (http://sandberg.cmb.ki.se/redspottednewt) in various formats. Currently, we are also working out ways how to amalgamate our databases with the Newtomics platform [4]. The reference transcriptome presented here provides a new basis for functional studies of salamander-specific processes at the molecular level and allows addressing key aspects of vertebrate regeneration using novel approaches.

Conflict of interest

The author(s) declare that they have no conflict of interests.

Author contributions

Ilgar Abdullayev created the RNA-Seq libraries, reconstructed transcriptome, analyzed transcriptome data, and contributed manuscript text. Matthew Kirkham extracted RNA, performed additional cloning and sequencing of newt transcripts and contributed manuscript text. Åsa Björklund performed computational analyses of proteins and domains and contributed manuscript text. András Simon conceived the study, participated in study design and wrote the manuscript. Rickard Sandberg conceived the study, participated in study design, assisted in computational analyses and wrote the manuscript with input from other authors.

Acknowledgments

We thank Manfred Grabherr for comments on the manuscript, Anna Johnsson and Sten Linnarsson for HiSeq2000 sequencing. András Simon was supported by grants from the European Research Council, Swedish Research Council, Cancerfonden and AFA Insurances. Rickard Sandberg was supported by European Research Council (Starting grant 243066), Swedish Research Council (2008–4562), Foundation for Strategic Research (FFL4) and Åke Wiberg Foundation (756194131). The reported sequence read data have been deposited in Sequence Read Archive at NCBI (SRP018244).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.yexcr.2013.02.013.

REFERENCES

[1] J.P. Brockes, A. Kumar, Comparative aspects of animal regeneration, Annu. Rev. Cell Dev. Biol. 24 (2008) 525–549.

[2] K. Nakamura, et al., miRNAs in newt lens regeneration: specific control of proliferation and evidence for miRNA networking, PLoS One 5 (2010) e12058.

[3] M.M. Casco-Robles, et al., Expressing exogenous genes in newts by transgenesis, Nat. Protocols 6 (2011) 600–608.

[4] M. Bruckskotten, M. Looso, R. Reinhardt, T. Braun, T. Borchardt, Newt-omics: a comprehensive repository for omics data from the newt *Notophthalmus viridescens*, Nucleic Acids Res. 40 (2012) D895–900.

[5] N. Rao, et al., Proteomic analysis of blastema formation in regenerating axolotl limbs, BMC Biol. 7 (2009) 83.

[6] D.A. Berg, et al., Efficient regeneration by activation of neurogenesis in homeostatically quiescent regions of the adult vertebrate brain, Development 137 (2010) 4127–4134.

[7] J.J. Smith, et al., Genic regions of a large salamander genome contain long introns and novel genes, BMC Genomics 10 (2009) 19.

[8] S. Putta, et al., From biomedicine to natural history research: EST resources for ambystomatid salamanders, BMC Genomics 5 (2004) 54.

[9] L. Sobkow, H.-H. Epperlein, S. Herklotz, W.L. Straube, E.M. Tanaka, A germline GFP transgenic axolotl and its use to track cell fate: dual origin of the fin mesenchyme during development and the fate of blood cells during regeneration, Dev. Biol. 290 (2006) 386–397.

[10] T. Sehm, C. Sachse, C. Frenzel, K. Echeverri, miR-196 is an essential early-stage regulator of tail regeneration, upstream of key spinal cord patterning events, Dev. Biol. 334 (2009) 468–480.

[11] M.G. Grabherr, et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. (2011).

[12] M.H. Schulz, D.R. Zerbino, M. Vingron, E. Birney, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, Bioinformatics 28 (2012) 1086–1092.

[13] G. Robertson, et al., De novo assembly and analysis of RNA-seq data, Nat. Methods 7 (2010) 909–912.

[14] D. Parkhomchuk, et al., Transcriptome analysis by strand-specific sequencing of complementary DNA, Nucleic Acids Res. 37 (2009) e123.

[15] J.A. Martin, Z. Wang, Next-generation transcriptome assembly, Nat. Rev. Genet. 12 (2011) 671–682.

[16] G. Ostlund, et al., InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, Nucleic Acids Res. 38 (2010) D196–203.

[17] C.C. Sun, et al., LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders, Genome Biol. Evol. 4 (2011) 168–183.

[18] W.W. Zhu, et al., Retrotransposon long interspersed nucleotide element-1 (LINE-1) is activated during salamander limb regeneration, Dev. Growth Differ. 54 (2012) 673–685.

[19] A.D. Moore, A.K. Björklund, D. Ekman, E. Bornberg-Bauer, A. Elofsson, Arrangements in the modular evolution of proteins, Trends Biochem. Sci. 33 (2008) 444–451.

[20] E.P. Nawrocki, D.L. Kolbe, S.R. Eddy, Infernal 1.0: inference of RNA alignments, Bioinformatics 25 (2009) 1335–1337.

[21] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinf. 12 (2011) 323.

[22] D. Ramsköld, E.T. Wang, C.B. Burge, R. Sandberg, An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data, PLoS Comput. Biol. 5 (2009) e1000598.

[23] K.V. Pajcini, S.Y. Corbel, J. Sage, J.H. Pomerantz, H.M. Blau, Transient inactivation of Rb and ARF yields regenerative cells from postmitotic mammalian muscle, Cell Stem Cell 7 (2010) 198–213.

[24] A.A. Garza-Garcia, P.C. Driscoll, J.P. Brockes, Evidence for the local evolution of mechanisms underlying limb regeneration in salamanders, Integr. Comp. Biol. 50 (2010) 528–535.

[25] A. Simon, E.M. Tanaka, Limb regeneration, WIREs Dev. Biol. (2012).

[26] H.M. Blau, J.H. Pomerantz, Reevolutionary regenerative medicine, JAMA 305 (2011) 87–88.

[27] J.Z. Levin, et al., Comprehensive comparative analysis of strand-specific RNA sequencing methods, Nat. Methods 7 (2010) 709–715.

[28] D.A. Berg, M. Kirkham, H. Wang, J. Frisén, A. Simon, Dopamine controls neurogenesis in the adult salamander midbrain in homeostasis and during regeneration of dopamine neurons, Cell Stem Cell 8 (2011) 426–433.

[29] A.K. Björklund, D. Ekman, A. Elofsson, Expansion of protein domain repeats, PLoS Comput. Biol. 2 (2006) e114.

[30] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010) 139–140.

[31] D.W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. protocols 4 (2008) 44–57.