A Dual Platform Approach to Transcript Discovery for the Planarian Schmidtea Mediterranea to Establish RNAseq for Stem Cell and Regeneration Biology

Martin J. Blythe[1]◑, Damian Kao[2]◑, Sunir Malla[1]◑, Joanna Rowsell[1]◑, Ray Wilson[1], Deborah Evans[2], Jamie Jowett[2], Amy Hall[2], Virginie Lemay[2], Sabrina Lam[2], A. Aziz Aboobaker[1,2]*

1 Deep Seq, Faculty of Medicine and Health Sciences, Queen's Medical Centre, University of Nottingham, Nottingham, United Kingdom, **2** Evolutionary Developmental Biology Laboratory, Centre for Genetics and Genomics, Queen's Medical Centre, University of Nottingham, Nottingham, United Kingdom

Abstract

The use of planarians as a model system is expanding and the mechanisms that control planarian regeneration are being elucidated. The planarian *Schmidtea mediterranea* in particular has become a species of choice. Currently the planarian research community has access to this whole genome sequencing project and over 70,000 expressed sequence tags. However, the establishment of massively parallel sequencing technologies has provided the opportunity to define genetic content, and in particular transcriptomes, in unprecedented detail. Here we apply this approach to the planarian model system. We have sequenced, mapped and assembled 581,365 long and 507,719,814 short reads from RNA of intact and mixed stages of the first 7 days of planarian regeneration. We used an iterative mapping approach to identify and define de novo splice sites with short reads and increase confidence in our transcript predictions. We more than double the number of transcripts currently defined by publicly available ESTs, resulting in a collection of 25,053 transcripts described by combining platforms. We also demonstrate the utility of this collection for an RNAseq approach to identify potential transcripts that are enriched in neoblast stem cells and their progeny by comparing transcriptome wide expression levels between irradiated and intact planarians. Our experiments have defined an extensive planarian transcriptome that can be used as a template for RNAseq and can also help to annotate the *S. mediterranea* genome. We anticipate that suites of other 'omic approaches will also be facilitated by building on this comprehensive data set including RNAseq across many planarian regenerative stages, scenarios, tissues and phenotypes generated by RNAi.

Citation: Blythe MJ, Kao D, Malla S, Rowsell J, Wilson R, et al. (2010) A Dual Platform Approach to Transcript Discovery for the Planarian Schmidtea Mediterranea to Establish RNAseq for Stem Cell and Regeneration Biology. PLoS ONE 5(12): e15617. doi:10.1371/journal.pone.0015617

Editor: Johannes Jaeger, Universitat Pompeu Fabra, Spain

Received September 8, 2010; **Accepted** November 13, 2010; **Published** December 14, 2010

Funding: This work was funded by project grants from the BBSRC and MRC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Aziz.Aboobaker@nottingham.ac.uk

◑ These authors contributed equally to this work.

Introduction

Understanding how we might repair and regenerate lost or damaged organs is a major goal of biomedical research. However, an integrated understanding of how stem cells can be used to replace and repair damaged or ageing tissue remains, for the most part, a distant goal [1]. The planarian model system provides a simple model in which to understand how this process is regulated in a natural context [2]. These amazing animals are able to regenerate entire animals from small starting fragments after amputations and constantly replace their somatic tissues during homeostasis as part of their normal life history [3]. This ability can be traced to a population of collectively totipotent planarian adult stem cells (pASCs), classically called neoblasts [2].

During normal tissue turnover new stem cell division progeny replace older differentiated cells. Broad positional information is maintained and re-established during regeneration by conserved signaling pathways that perform analogous roles during the embryogenesis of other animals [4,5,6,7,8,9,10,11,12,13,14,15]. Division progeny interpret these signals to ensure that they differentiate into the correct cells at the correct positions within the body. This homeostatic process is acutely responsive to nutrient conditions and planarians will grow and de-grow as they are fed and starved respectively [3]. On induction of regeneration by amputation, for example decapitation, the stem cell population undergoes a marked increase in proliferation throughout the wounded fragment(s) [16]. Stem cell progeny undergo precise differentiation to form the missing structures removed by amputation within seven days of the original injury, with the restoration of perfect scale and proportion taking around 14 days [2].

It is these two interdependent processes of ongoing homeostatic replacement and regenerative potential that make planarians a fascinating model system [17]. In recent years the inevitable application of modern molecular genetic techniques has made understanding the molecular mechanisms controlling regeneration a realistic goal. For this reason it seems timely to produce a high quality publicly accessible description of the transcribed genetic content of the model planarian *S. mediterranea*.

The development of massively parallel sequencing platforms has allowed a set of new biological questions to be asked [18]. In particular a number of recent studies have employed massively parallel sequencing to generate more complete descriptions of transcribed portions of genomes [19,20,21,22,23,24]. The approaches that can be facilitated by these large datasets include expression screens [25], large functional screens [26], interactome [27] and phylogenetic studies with large datasets of homologous genes [28].

The genome of the planarian *S. mediterranea* has been sequenced using a whole genome shotgun approach. Assembly of reads representing in excess of 11.6 fold coverage has surprisingly resulted in 43,294 super-contigs, with assembly problems caused by a number of issues. The A/T rich nature of the genome (greater that 65%) and its apparent repetitiveness (more than 40%) clearly contribute to the current poor assembly metrics (see genome.wustl.edu for information on the *S. mediterranea* genome). In addition to the genome sequence there are also 74,388 publicly available EST sequences from cDNA libraries (although the clones themselves are not publicly available). Finally, some researchers have made an effort to create a useful annotated database for the community which contains 30,930 gene predictions using the MAKER pipeline [29], attempting to incorporate data from ESTs, BLAST homology and even from presumably ongoing RNAi screens and array based studies of gene expression [30,31]. In order to facilitate a systems approach to understanding planarian generation we would like an as complete as possible description of the planarian transcriptome.

Here, we generate and describe a transcriptome from a mixed time course of regenerative stages of the asexual strain of the planarian *S. mediterranea* to use as a basis for genome wide expression analyses, such as RNAseq. We have integrated data from two next generation sequencing platforms in an attempt to increase gene discovery and transcript representation as well as avoid single platform bias. Our analysis highlights some of the current difficulties, and presents some simple solutions, with generating a dual platform transcriptome in relation to a fairly poorly assembled genome. We also show that the collection of transcripts we have defined is suitable for RNAseq approaches. We compare expression levels of genes in intact worms with those that have been irradiated to remove proliferating adult stem cells and their recent undifferentiated progeny. This defines the set of transcripts whose expression is potentially enriched in adult stem cells and their progeny.

Results and Discussion

Increased planarian gene discovery using 454 transcriptome sequencing

Currently planarian researchers are supported by the existence of a whole genome shotgun assembly and 74,388 ESTs for *Schmidtea mediterranea* [30,32]. We wished to attempt to provide a more complete description of the transcriptome to enable a broader systems approach to understanding regeneration and adult stem cells in *S. mediterranea*. In order to do this we decided to sequence transcripts from whole and regenerating worms from different stages of the first 7 days of regeneration (see Materials and Methods and Figure 1). Combining equal amounts of input material from a mix of different stages we prepared a normalized 454 cDNA library. We sequenced from this library and generated 743,464 454 reads. These reads had a mean length of 278 bp, and an overall length distribution characteristic of 454 transcriptome sequencing with Titanium chemistry on this platform (Figure 2). Assembly detected 16,967 putative "isogroups" (genes) and a total of 21,030 putative "isotigs" (isoforms). We added 74,388 publicly available ESTs for *S. mediterranea* along with our 454 data. The combined data utilized 581,365 of the 454 reads in the assembly and resulted in 17,628 putative genes and 22,698 isoforms. We refer to this dataset as Sm454EST throughout this manuscript (Table 1 and Dataset S1).

We analyzed how the number of predic ted transcripts (isogroups), predicted isoforms (isotigs), N50 length, and average contig length increased with increasing read number (Figure 3A and B) by assembling random selections of sequentially larger subsets of the total read set. These data sets suggest that even deeper sequencing with the Sm454 library would still yield significant novel transcript discovery and increases in average contig length (our 454 library is freely available on request). The length distribution of the contigs in this data set shows that most transcripts identified are over 400 bp in length (Fig 4A)

The current *S. mediterranea* genome database provides a set of 30,930 MAKER based [29,30]; here we have called this data set SmMAKER (Dataset S2). We compared Sm454EST data set with SmMAKER by BLASTN [33]. We found that 18,797 SmMAKER annotations had a cognate Sm454EST hit, representing the same transcript (Table 2). Conversely we found that this represented 13,887 Sm454EST transcripts, suggesting some level of redundancy in the SmMAKER transcripts and that some annotations may be of duplicate sequences in the genome or separate gene predictions that are actually single genes. We found that 8,811 of 22,698 Sm454EST defined transcripts had no cognate MAKER annotations, immediately highlighting the benefits of transcriptome sequencing for gene discovery, and eventual annotation. We reasoned most must represent un-annotated transcripts or be from regions not in the current genome assembly. Mapping Sm454EST transcripts to the genome by BLASTN [33] or using GMAP [34] revealed that less than 2% failed to map across at least 80% of their length. This suggests that most of the 8,829 SM454EST transcripts without an SmMAKER match are un-annotated rather than missing from the current genome assembly. We found that 12,133 MAKER predictions had no cognate hit in the Sm454EST dataset suggesting that they are not transcribed genes or, just as likely, deeper sequencing is required to identify them (see below).

Massively parallel sequencing of short reads with iterative mapping further supplements gene discovery

Short read platforms offer the possibility of ultra-deep transcriptome coverage for both transcript discovery and the analysis of expression levels with unprecedented resolution [35,36]. However, the use of short reads for the process of measuring genome wide expression levels, an approach called RNAseq, is somewhat dependent on the existence of high quality transcriptome annotations [23,37,38,39]. Using short reads to define a transcriptome poses a number of problems. Firstly, a genome sequence is still required to achieve long contig lengths despite improvements in de novo short read assemblers [40]. Secondly, short reads will be prone to certain inaccuracies when used to define a transcriptome by mapping to a genome sequence. Examples include inaccurate mapping due to sequencing errors, mapping to regions that are not transcribed but have homology to transcribed regions (for example to pseudogenes) and by missing small exons and/or introns to which isolated sort reads fail to map.

In the case of *S. mediterannea* the genome assembly comprises 43,294 contigs with no chromosomal structure. To deal with this problem we used an iterative mapping approach to increase the definition of splice junctions between exons and define alternative transcript sequences (see Figure 5). We generated 903,642,430 50 bp reads using two flow cells on the SOLiD 3+ sequencing platform using a library prepared from the same mixed stage RNA pools (see materials and methods). We mapped reads using the

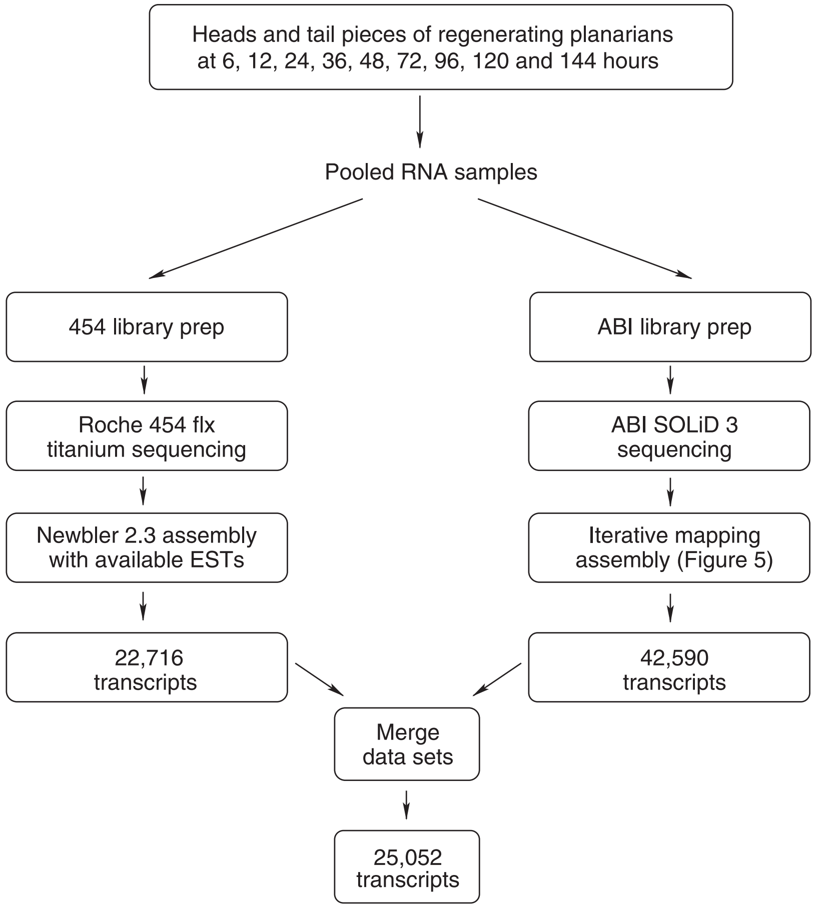
Heads and tail pieces of regenerating planarians at 6, 12, 24, 36, 48, 72, 96, 120 and 144 hours
Pooled RNA samples
454 library prep
Roche 454 flx titanium sequencing
Newbler 2.3 assembly with available ESTs
22,716 transcripts
ABI library prep
ABI SOLiD 3 sequencing
Iterative mapping assembly (Figure 5)
42,590 transcripts
Merge data sets
25,052 transcripts

Figure 1. Overview of pipeline obtaining planarian transcripts. A schematic flow chart showing the work-flow for defining the planarian transcript set using a dual platform approach. 20 size-matched worms for each time point were cut into head, middle and tail pieces. Regenerating head, tail and middle fragments for RNA preparation were collected and frozen at 6, 12, 24, 36, 48, 72, 96, 120, 144 and 168 hours of regeneration. Total RNA was prepared from these time points, pooled and libraries made for both platforms.
doi:10.1371/journal.pone.0015617.g001

existing published MAKER annotations from SmedGD as a starting point of our analysis [30]. We applied a simple filter to remove reads with low quality colour calls or low mapping scores from the resulting SAM file. We achieved high quality mapping of 507,719,814 reads, representing around 56% of the raw sequenced reads. Using the total length of the available MAKER annotations as an estimate this conservatively represents 1,060 fold coverage of the planarian transcriptome. We then used the Cufflinks program [41] to interpret the high quality mapped reads to produce a new annotated GTF (see materials and methods). Cufflinks was used to remove regions that were represented by very low read coverage and to define split-reads that crossed splice sites predicted by the SmMAKER annotations. This initial transcriptome GTF file comprised 19,429 putative multiple exon transcripts defined entirely by those genes that were in the original annotation file (i.e. in the SmMAKER dataset), and over 153,038 putative single exon transcripts. Many of these single exons reads were very short, with over 30,000 being consisting of two read lengths or less (Figure 4B).

The occurrence of mapped split-reads enables the definition of gene structure and different isoforms. To define as many split-reads as possible we remapped selected reads against the initial exon annotation defined by Cufflinks, but with exons grouped to form hypothetical “super-genes” on each strand of individual genomic contigs (see Figure 5). Of the 43,294 total contigs we were able to detect expression on only 15,897, representing 20,611 super-genes as not all contigs showed expression on both strands. A drawback of this approach is that low quality reads may erroneously map to splice junctions that are not real. Therefore we only remapped high quality reads that initially mapped giving a partial alignment string (cigar) possibly indicative of one half of a split-read. We found that this method defined many new splice junctions.

We then reassembled the new split-reads mappings in combination with the original mappings using Cufflinks. The number of multiple exon transcripts in the new assembly increased from 19,429 to 26,791 and the average length increased from 387 bp to 490 bp, reflected in the distribution of contig lengths (Figure 4C). This was a result of an increased number of multiple exon genes, additional isoform discovery, and of extension of transcripts defined in the first round of mapping. A total of 15,799 single exon transcripts were also retained to look for possible overlaps with the Sm454EST dataset in a merged assembly. We call this dataset SmABI (Dataset S3).

These data suggest that ultra-deep sequencing using short read technology clearly defines more transcribed regions than deep sequencing with 454 technology and may also define many more putative alternate splicing events. It is also likely that we failed to detect many splice events between contigs due to the fragmented nature of the planarian genome. We may have also detected many pseudogenes and duplicate regions unresolved in the current

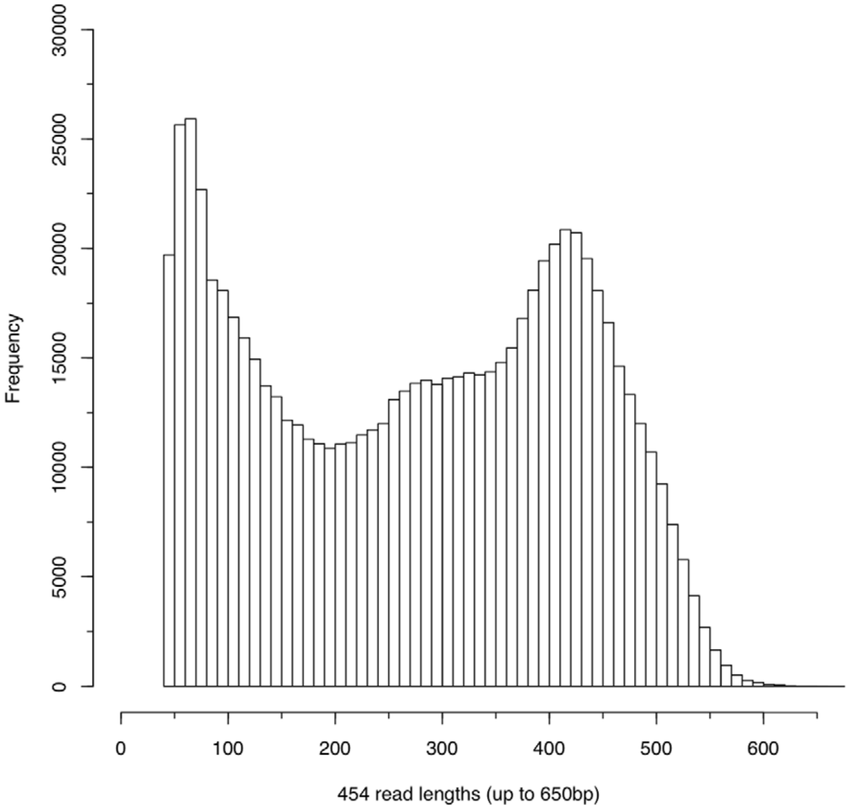

Figure 2. 454 read length distribution. A plot showing the distribution of the raw planarian 454 reads generated from one pico-titre plate run with Titanium chemistry.
doi:10.1371/journal.pone.0015617.g002

genome assembly, as short reads are able to map to accurately to these regions. Additionally, we might have detected non-coding transcripts of multiple origins and functions. We compared the SmABI dataset to the SmMAKER dataset and found that of the 42,950 putative SmABi transcripts 30,223 matched to the SmMAKER dataset, thus almost 30% of transcribed regions detected by this approach were not represented in previous annotations. When we looked for representation of the 30,920 SmMAKER annotations in the SmABi dataset we were able find similarity by BLASTN for 20,275. This is approximately 1300 more than were found in the Sm454EST dataset and maybe be the result of the deeper coverage achieved by SOLiD sequencing compared to 454 sequencing.

An overview of planarian transcriptome structure from a combined platform approach

We wished to optimize our data by combining the long read and short read data sets. We extracted the cDNA sequences of the transcripts in SmABI and merged them with the Sm454EST data sets (see materials and methods). To combine the datasets we first assembled the Sm454EST data generated by GSAssembler with CAP3 [42] to collapse very similar isoforms into single transcripts. This resulted in 21,169 contigs that represented transcripts and isoforms with differences in internal exon content (isoforms with alternate terminal exons will have assembled). We then identified all SmABI contigs that matched with this data set by screening for significant hits with BLASTN (see materials and methods). This identified 27,830 SmABI contigs that had high BLASTN homology with the CAP3 assembled Sm454EST dataset. Finally, from those SmABI contigs that did not clearly overlap as indicated by BLASTN we excluded single exon genes and kept 6,494 multiple exon genes, likely to be real transcripts as they have spliced read evidence. Single exon SmABI genes without evidence from another source were excluded as they may represent duplications and pseudogenes to which short reads are still able to map. We then assembled the 21,169 contigs derived from the Sm454EST data with 6,494 unique multiple exon ABI contigs and 27,830 ABI contigs that had BLASTN matches to the Sm454EST dataset using CAP3. This resulted in 11,308 assembled contigs (combinations of two platforms) and 19,639 singletons (only represented by one platform). We removed 5,894 singletons representing single exon genes from the SmABI dataset, as we could not ascertain whether they represented real transcripts. This combined dataset Sm454ESTABI (Dataset S4) resulted in 25,053 transcripts that were defined by long read 454EST sequencing, a combination of platforms, or spliced transcripts identified by ABi SOLiD sequencing. The distribution of contig lengths of this dataset is longer than that for either dataset alone (Figure 4D).

Table 1. List of datasets.

Dataset	Number of transcripts
Sm454EST	22,716
SmABI	42,950
Sm454ESTABI	25,052
MAKER	30,930

doi:10.1371/journal.pone.0015617.t001

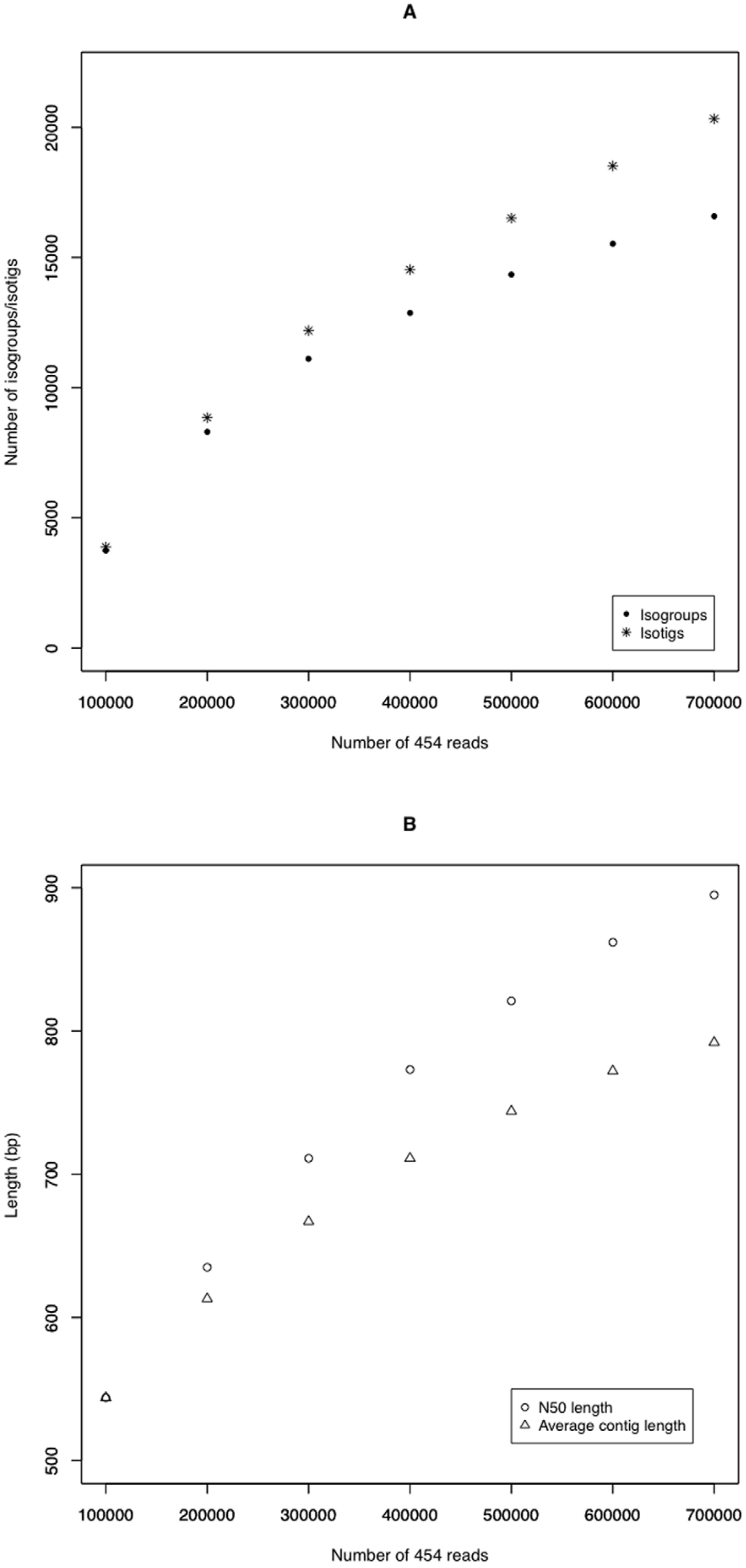
A
Number of isogroups/isotigs
0
5000
10000
15000
20000
100000
200000
300000
400000
500000
600000
700000
Number of 454 reads
● Isogroups
✳ Isotigs
B
Length (bp)
500
600
700
800
900
100000
200000
300000
400000
500000
600000
700000
Number of 454 reads
○ N50 length
△ Average contig length

Figure 3. Increasing isogroup/isotig numbers and contig length with increasing read number. A. The number of Isotigs and Isogroups continue to increase with increasing number of reads added to the assembly, however Isogroup (gene) discovery appears to be a approaching a maximum whereas Isotig discovery (transcript isoforms) continues to increase. This suggests further sequencing would be beneficial to define more alternately spliced transcripts. **B.** The lengths of contigs also increase as more reads are added to the assembly. While many contigs are not any getting longer (average contig length) further sequencing is increasing the N50 perhaps by joining some transcripts together to form single contigs. Both analyses suggest that further sequencing of the mixed stage library may generate more transcripts and continue to increase the average length of contigs.
doi:10.1371/journal.pone.0015617.g003

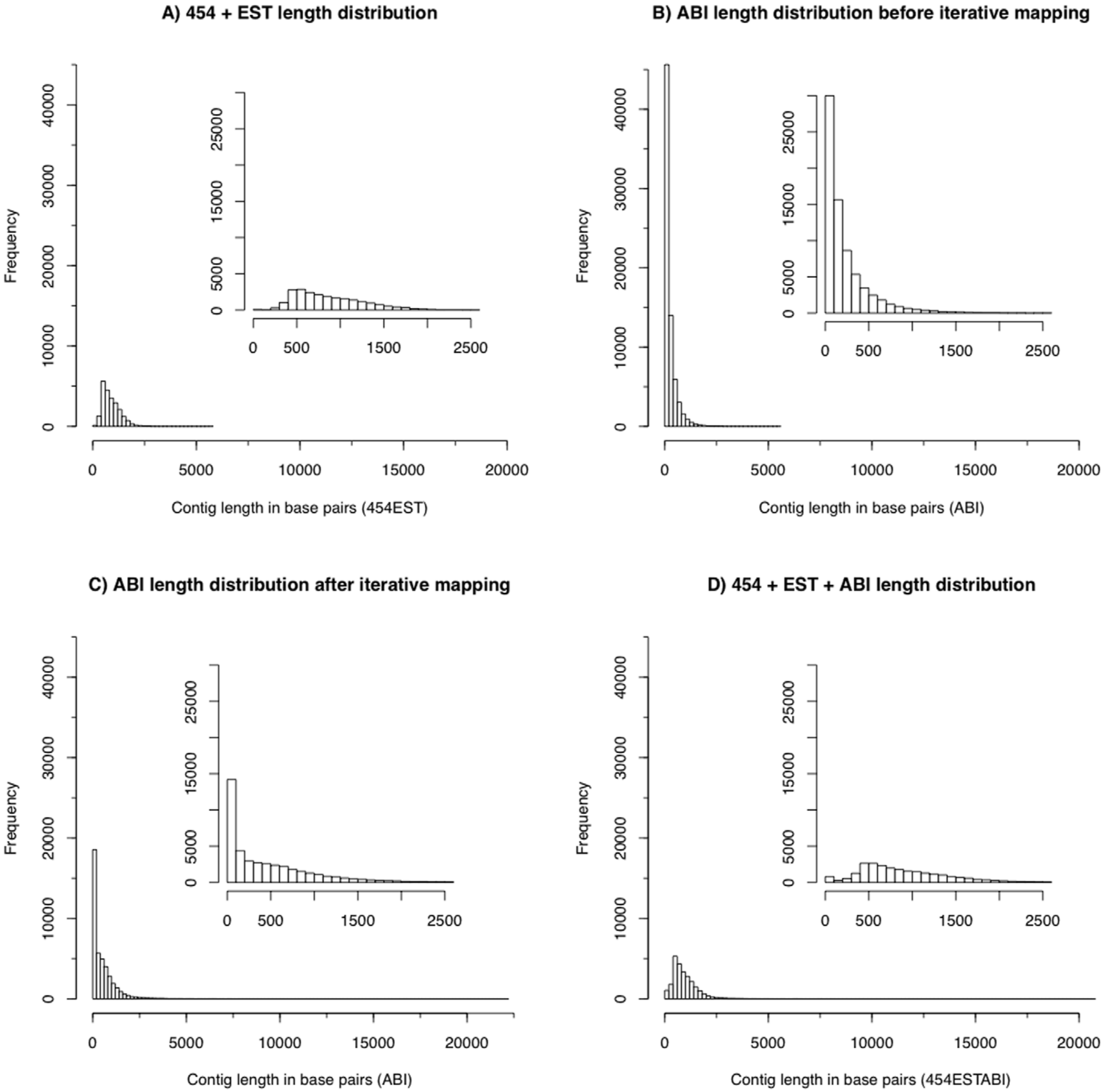

Figure 4. Contig length distributions. Contig length distribution for the following datasets (A) Sm454EST (B) SmABI before iterative mapping (C) SmABI after iterative mapping (D) Sm454ESTABI merged dataset. The subplots for each figure focus on contigs with length up to 2,500 bp. The short read approach using SOLiD clearly produces many short putative transcripts, many of these may be transcribed regions that do not encode proteins or which splice reads were not identified. Iterative mapping resulted in fewer short transcribed regions after filtering and longer contigs (B vs C) Combining the datasets led to an increase in contig lengths compared to either platform alone.
doi:10.1371/journal.pone.0015617.g004

Table 2. Comparison of transcriptomes with existing datasets.

	454/EST	ABI	454/EST/ABI	MAKER
454/EST	x	16,374	22,682	13,887
ABI	28,618	x		30,223
454/EST/ABI	21,502		x	16,726
MAKER	18,797	20,745	22,189	x

doi:10.1371/journal.pone.0015617.t002

We compared the Sm454ESTABI dataset to the SmMAKER annotation set (Table 2). We found that 22,189 MAKER annotations were now represented by 16,726 Sm454ESTABI transcripts. We interpret this as highlighting redundancy in the MAKER annotations. The combined dataset contained many more of the SmMAKER trascripts than the Sm454EST or SmABi dataset alone (Table 2), suggesting that gene discovery is increased by combining and assembling data from both platforms. The remaining approximately 8,000 SmMAKER annotations may not represent real genes or may represent transcribed regions that are expressed in life history stages we have not sampled in this work. It is also possible that some represent differences between the sexual and asexual transcriptome. We also found that 8,316 transcribed regions defined in Sm454ESTABI were absent from the SmMAKER annotations, indicating that we are likely to have identified a large number of novel planarian transcripts.

We performed a simple open reading frame analysis (ORF) on all of the Sm454ESTABI transcripts to identify the longest putative protein that could be coded for by each transcript file

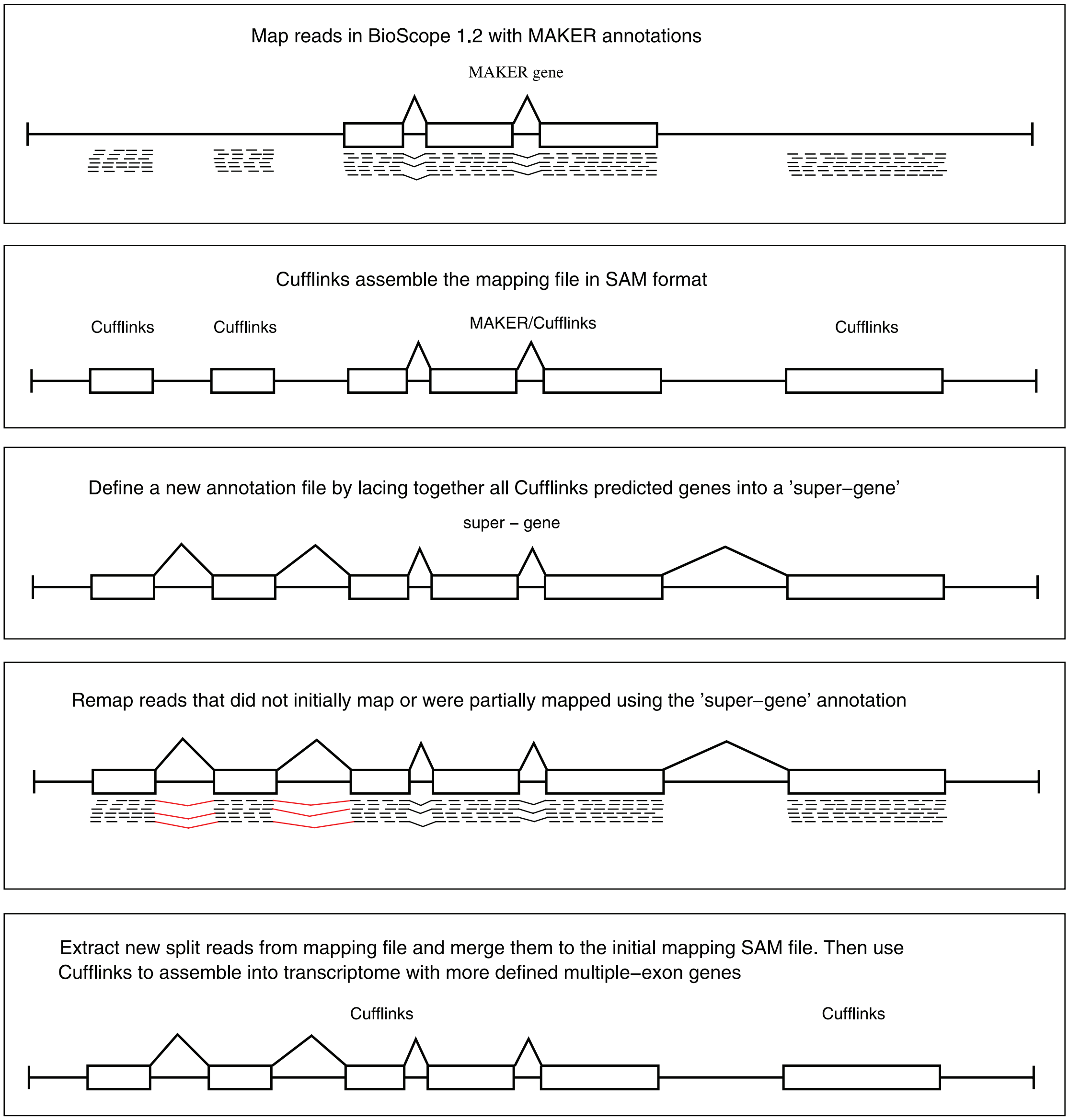
Map reads in BioScope 1.2 with MAKER annotations
MAKER gene
Cufflinks assemble the mapping file in SAM format
Cufflinks
Cufflinks
MAKER/Cufflinks
Cufflinks
Define a new annotation file by lacing together all Cufflinks predicted genes into a 'super–gene'
super – gene
Remap reads that did not initially map or were partially mapped using the 'super–gene' annotation
Extract new split reads from mapping file and merge them to the initial mapping SAM file. Then use Cufflinks to assemble into transcriptome with more defined multiple–exon genes
Cufflinks
Cufflinks

Figure 5. Defining transcript variants in more detail. A schematic flow chart showing the iterative mapping approach. New split reads found in the second round of mapping are in red.
doi:10.1371/journal.pone.0015617.g005

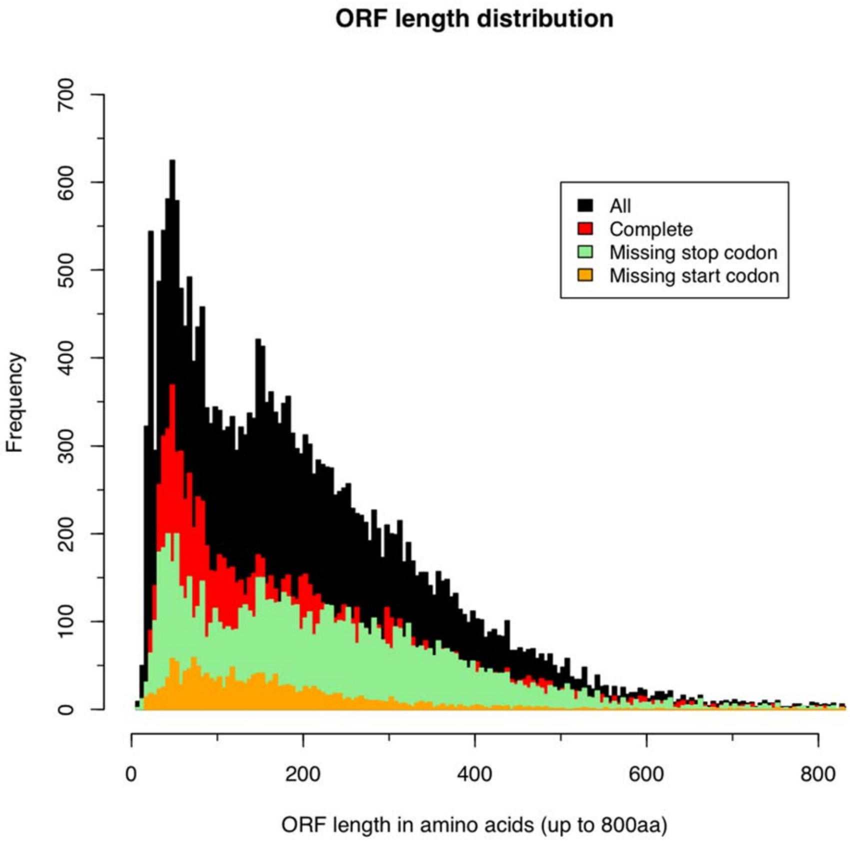

Figure 6. Open reading frame length distribution. Length distribution of the longest putative protein that could be coded for by each transcript for the Sm454ESTABI dataset (black). 11,599 open reading frames had both a putative start and stop codon (red), 9,076 missed the start (green) codon and 1,757 missed the stop codon (orange). The remaining 2620 transcripts either had neither a start or stop.
doi:10.1371/journal.pone.0015617.g006

(Dataset S5). We found 11,599 putatively full-length ORFs with a possible start and stop codon, 9,076 with the start codon missing, and 1,757 with the stop codon missing. The remaining transcripts either did not have either a start or stop codon or no clear ORF (Figure 6). We assume that the large number of short ORFs with both starts and stop codons around the 50 amino acid size are probably not real proteins but represent non-coding regions of transcripts.

Gene ontology, homology and novelty in the planarian transcriptome

In order to further validate and analyze the protein coding content of our Sm454ESTABI dataset we performed three independent analyses: 1. a cross-species BLASTX [33] analysis against the proteomes of 18 species. 2. transcript sequences were scanned for homology to Pfam protein families using Pfam-A hidden markov models v.24 [43]. 3. Gene Ontology (GO) database [44] classification analysis using the Blast2go pipeline v.2.3.5 [45] against a dataset consisting of the NCBI non-redundant invertebrate protein sequences, supplemented with sequences from the cross-species BLASTX analysis not included in the NCBI dataset. For all three analyses a lower significance threshold E-value of 1e-5 was applied for BLASTX.

For the cross-species BLASTX analysis a total of 18,742 significant blast hits were calculated for 15,000 of the 25,053 transcripts (Figure 7, Dataset S6), while the remaining 10,053 showed no significant homology. Of the matched transcripts 915 showed homology to only a single species; more than half of these (479) to S. *mansoni*. Not surprisingly the most top hits were also to the *S. mansoni* genome. The number of top hits broadly followed the accepted phylogenetic distance of planarians from the compared species.

The transcripts were translated in all 6 frames into amino acid space and scanned for homology to Pfam-A protein families. This allowed 15,691 pfam family assignments to 11,597 transcripts. The Blast2go pipeline provided 73,491 GO term assignments to 10,803 transcripts. A further 13,724 assignments to 7,365 transcripts were obtained via pfam2go GO terms for Pfam-A families. These 2 datasets were condensed to 82,872 terms for 12,133 transcripts after accounting for redundancy between datasets. The hierarchical structure of GO terms allowed transcript classification to be summarised further to first tier terms of the three main domains of the GO database (Figure 8A, B, C).

Of the 25,053 total transcripts a total of 15,610 were classified by cross-species homology, a GO assignment, or pfam family homology, while 9443 remained unclassified (see Dataset S7 for GO assignments and PFAM annotations). An analysis comparing the distribution of transcript and open reading frame lengths of assigned and unassigned transcripts shows that the average classified transcript ORF length (266aa±182) is more than double that of the unclassified (111aa±98) (Figure 9A). Some of these unclassified short ORFs are likely to belong to short partial transcripts in our dataset (Fig 9B). The large number of transcripts with ORF less than 100 aa are likely to represent short fragments of protein coding transcripts, non-coding RNAs and possibly untranslated regions.

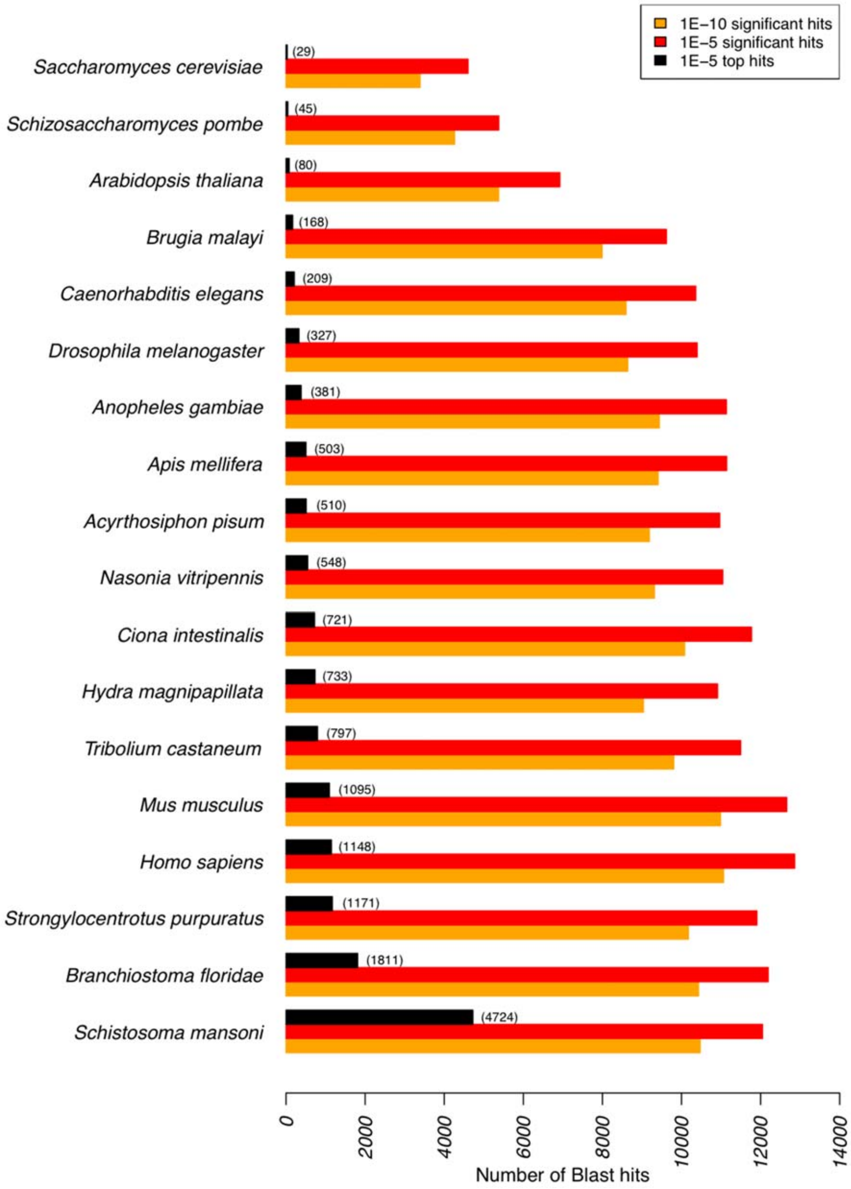

Figure 7. Cross-species BLASTX results. The BLASTX cross-species results for the Sm454ESTABI dataset. All species with a significant hit (probability < 1e-5 red bar, 1e-10 orange bar) are shown. Also shown are the number of times the top hit was to a particular species (black bar and number in brackets).

doi:10.1371/journal.pone.0015617.g007

An RNAseq approach to planarians defines the set of genes specifically affected by irradiation and therefore potentially expressed in stem cells and their recent progeny

We wished to establish the use of our dataset for the whole transcriptome analysis of expression levels in different experimental conditions, broadly called RNAseq [38]. We decided to use this approach to identify the set of transcripts that are likely to be enriched in proliferative stem cells and their recent progeny. Irradiation has been used extensively as an experimental tool to remove proliferating cells in planarians [46,47]. Previous studies have utilized this to define sets of genes that are potentially expressed in pASCs and their progeny using microarrays and relatively limited gene sets [47,48]. We performed further SOLiD 3+ sequencing on intact worms (mock irradiated) and worms 7 days after lethal irradiation (see materials and methods). This timepoint was chosen as it has previously been shown to remove both pASCs, their progeny that are no yet fully differentiated as well as possibly short lived differentiated cells [47]. We performed two replicates of each condition producing 13,392,444 and 31,688,899 mapped reads for intact worms and 10,334,419 and 11,372,969 reads for irradiated worms. We discarded transcripts for which we did not detect expression in all samples or where expression was close to zero in all samples to avoid the effects of low read sampling error on expression levels (RPKM<1). We analyzed expression levels between replicates and found a high congruency among replicates (Figure 10A and B).

We then pooled the results of each replicate experiment to compare in tact worms with 7 days irradiated worms (Figure 10C).

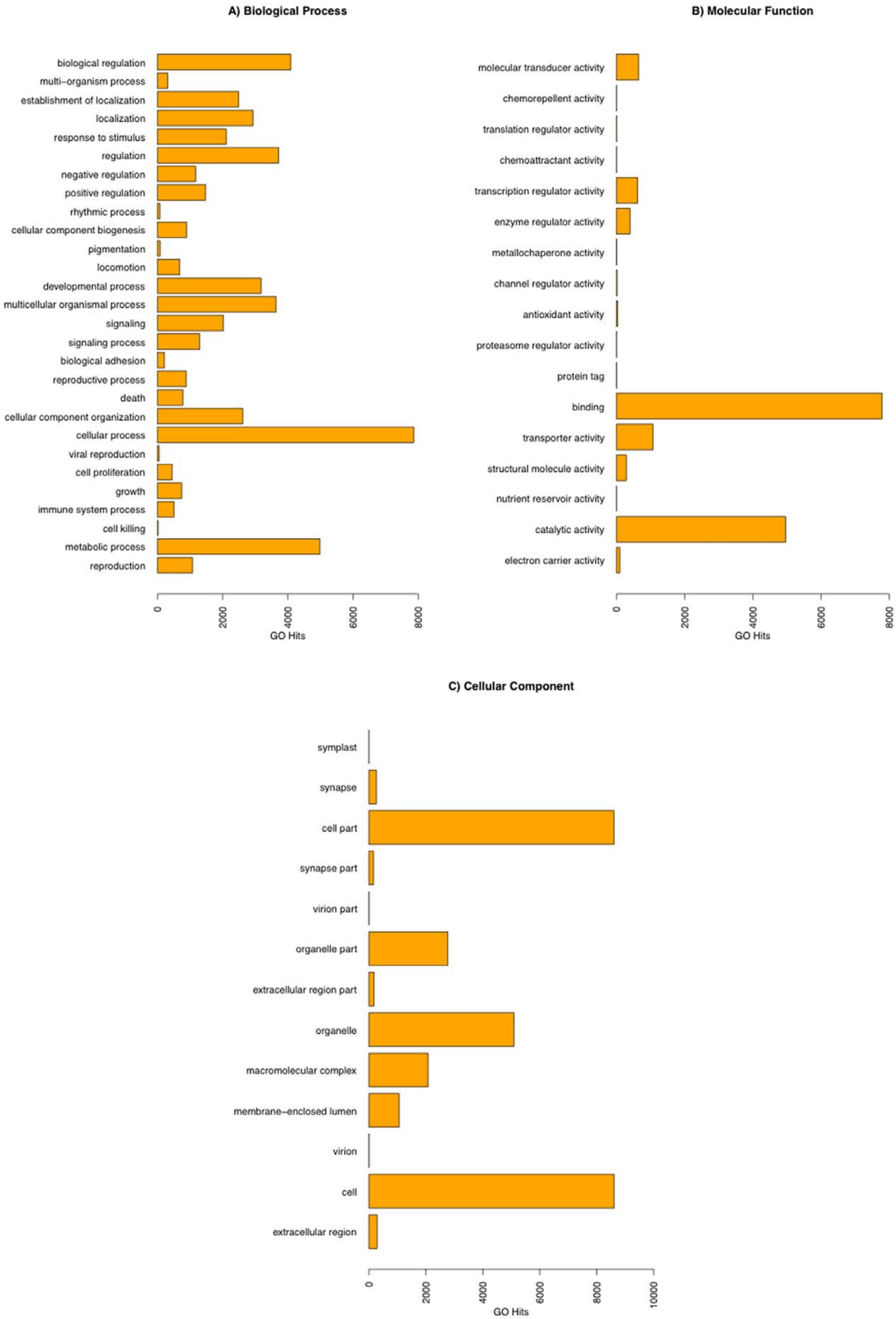
A) Biological Process
biological regulation
multi−organism process
establishment of localization
localization
response to stimulus
regulation
negative regulation
positive regulation
rhythmic process
cellular component biogenesis
pigmentation
locomotion
developmental process
multicellular organismal process
signaling
signaling process
biological adhesion
reproductive process
death
cellular component organization
cellular process
viral reproduction
cell proliferation
growth
immune system process
cell killing
metabolic process
reproduction
0
2000
4000
6000
8000
GO Hits
B) Molecular Function
molecular transducer activity
chemorepellent activity
translation regulator activity
chemoattractant activity
transcription regulator activity
enzyme regulator activity
metallochaperone activity
channel regulator activity
antioxidant activity
proteasome regulator activity
protein tag
binding
transporter activity
structural molecule activity
nutrient reservoir activity
catalytic activity
electron carrier activity
0
2000
4000
6000
8000
GO Hits
C) Cellular Component
symplast
synapse
cell part
synapse part
virion part
organelle part
extracellular region part
organelle
macromolecular complex
membrane−enclosed lumen
virion
cell
extracellular region
0
2000
4000
6000
8000
10000
GO Hits

Figure 8. Gene ontology hits. The number of gene ontology (GO) hits for the second tier categories: (A) Biological process (B) Molecular function (C) Cellular component. More detailed GO term assignment is available as an additional file.
doi:10.1371/journal.pone.0015617.g008

We found a total of 2,304 transcripts that had differential expression (p-value less than 0.01, see Dataset S8 for FASTAs of this file, Dataset 9 containing information of P value, fold change and annotation of differentially regulated transcripts) and a decreased expression in irradiated samples of more than 2 fold, thus defining a definitive set of genes down regulated by irradiation treatment and potentially enriched in pASCs and there recent progeny (these are available in FASTA format as Dataset 8 and with fold down regulation, p-value and annotations as Dataset 9). The highest fold difference observed for any gene between irradiated and intact samples was a 459.64 fold decrease.Our findings are somewhat comparable to those of previous investigations using microarrays, though we have investigated many more genes [47,48]. These transcripts are likely to fall into four categories: transcripts expressed in proliferative stem cells; transcripts expressed in recent stem cell progeny; transcripts that require the presence of proliferative stem cells for their expression but are expressed in differentiated cells, and finally, transcripts downregulated by irradiation irrespective of cell type expression profile. To validate our RNAseq results we looked at the response of genes that have been experimentally confirmed as having reduced expression after irradiation, as they are expressed in stem cells or their progeny, although some of them are also expressed in differentiated cell types (Table 3). We looked at 18 genes from different planarian species that have a role in pASC biology and/or their irradiation sensitive progeny, and have corroborating in situ hybridization data showing loss of expression after irradiation [47,49,50,51,52,53,54,55,56,57,58,59,60]. We found that in each case expression was significantly reduced ($p<0.01$, using the test of Baggerley et al [61] in irradiated samples and in 17/18 by at least two fold (Table 3). This suggests that our results are representative of the transcript set whose expression is affected by irradiation and are therefore candidates for potential roles in pASC biology.

Conclusions

We have successfully combined two different next generation sequencing platforms to produce a more complete view of the transcriptome of the planarian *S. mediterranea*. This dataset will help better define the genetic context of this model organism and allow the development of new reagents sets for 'omic approaches to understanding planarian biology. Of immediate significance is the ability to use our dataset combined with whole transcriptome sequencing to look at broadly and changes in expression. Here we have shown that this is possible by describing the set of transcripts that are down regulated by irradiation to remove proliferative stem cells and their recent progeny.

The transcriptome dataset present here defines many putative *S. mediterranea* proteins. Simple homology and ontological comparison of these proteins reveal an evolutionary pattern in agreement with our current view of the animal tree of life. *S. mediterranea* has added significance as a model system, beyond its regenerative capacity, due to its phylogetic position within the lophotrochozoan protostome group, an under represented group with respect to functional molecular studies.

Our dual platform approach to defining the transcriptome has a number of advantages over using just either a short read or long read platform alone. While a long read approach produces long reads that we can be confident represent real transcripts the ultra deep coverage afforded by short reads allows increased transcript discovery. Combining data from both platforms also likely alleviates platform biases specific to each workflow. As more and more animals genomes are sequenced using Next Generation sequencing platforms we are likely to have to work with genomes that although somewhat complete remain unassembled in many independent contigs. For these animals a dual platform transcriptome approach will be an affordable way to experimentally annotate these genomes with their cognate transcribed regions.

Materials and Methods

Animal culture and preparation of regenerative stages

A clonal line of the asexual strain of *Schmidtea mediterranea*, AAANOTBIOL01, was used for all experiments. Animals were reared at 20°C in tap water filtered through activated charcoal and buffered with 0.5 ml/L 1 M $NaHCO_3$. Planarians were fed veal liver and starved for at least one week prior to experiments or amputation. All worms used were 7–8 mm in length.

Regenerating fragments for RNA preparation were collected and frozen at 6, 12, 24, 36, 48, and 72 hours, 96, 120 and 144 hours of anterior and posterior regeneration. 20 worms for each time point were cut in head, middle and tail fragments. Whole intact animals unfed for 7 days were also included as a separate sample.

Preparation of RNA for sequencing

Planarian pieces were frozen at –80°C overnight in 300 μl Trizol (Invitrogen, Cat. No. 155966) and then RNA prepared using the manufacturers protocol. Samples were resuspended in nuclease-free water and treated with 2 U of Turbo DNase (Ambion, Cat. No. AM2238) in 1 x DNase buffer for 30 mins at 37°C. Samples were extracted with phenol, pH 4.3: chloroform (50:50; Sigma, Cat. No. P4682-100 ML) and alcohol precipitated overnight at –80°C. Total RNA was pelleted by centrifugation at 12,000×g for 15 mins at 4°C, washed with 70% ethanol and air dried for 5 mins. RNA was resuspended in nuclease-free water.

Preparation of 454 transcriptome libraries and sequencing

Concentration of total RNA for samples from each timepoint was measured with Qubit flourometer RNA assay kit (Invitrogen, Cat. No. Q32852). Equal amounts of total RNA from each stage was pooled together into a single sample. Ribominus eukaryotic kit (Invitrogen, Cat. No. A10837-08) was used according to manufacturer's instructions for depletion of ribosomal RNA. 1 μg of ribo-depleted RNA was used to make double stranded cDNA (dscDNA) with SMARTer PCR cDNA synthesis kit (Clontech, Cat. No. 634925) and normalisation of amplified dscDNA was carried out with the TRIMMER kit (Evrogen, Cat. No. NK001), according to the manufacturers instructions. 5 μg of dscDNA in 100 μl of TE was sonicated with a Covaris S2 in frequency sweeping mode at a water bath temperature of 5°C (fragmentation parameters: Duty cycle –5%, Intensity – 3, cycles/burst – 200 and Time – 120 secs). Following size fractionation with conventional gel electrophoresis, DNA fragments 400–900 bp were excised and purified with QIAquick gel extraction columns (Qiagen, Cat. No. 28760). Finally, gel purified dscDNA was used to create a fragment library as described in the Roche GS FLX titanium general library preparation guide. Sequence was performed on Roche 454 sequencing machine according to the manufacturers

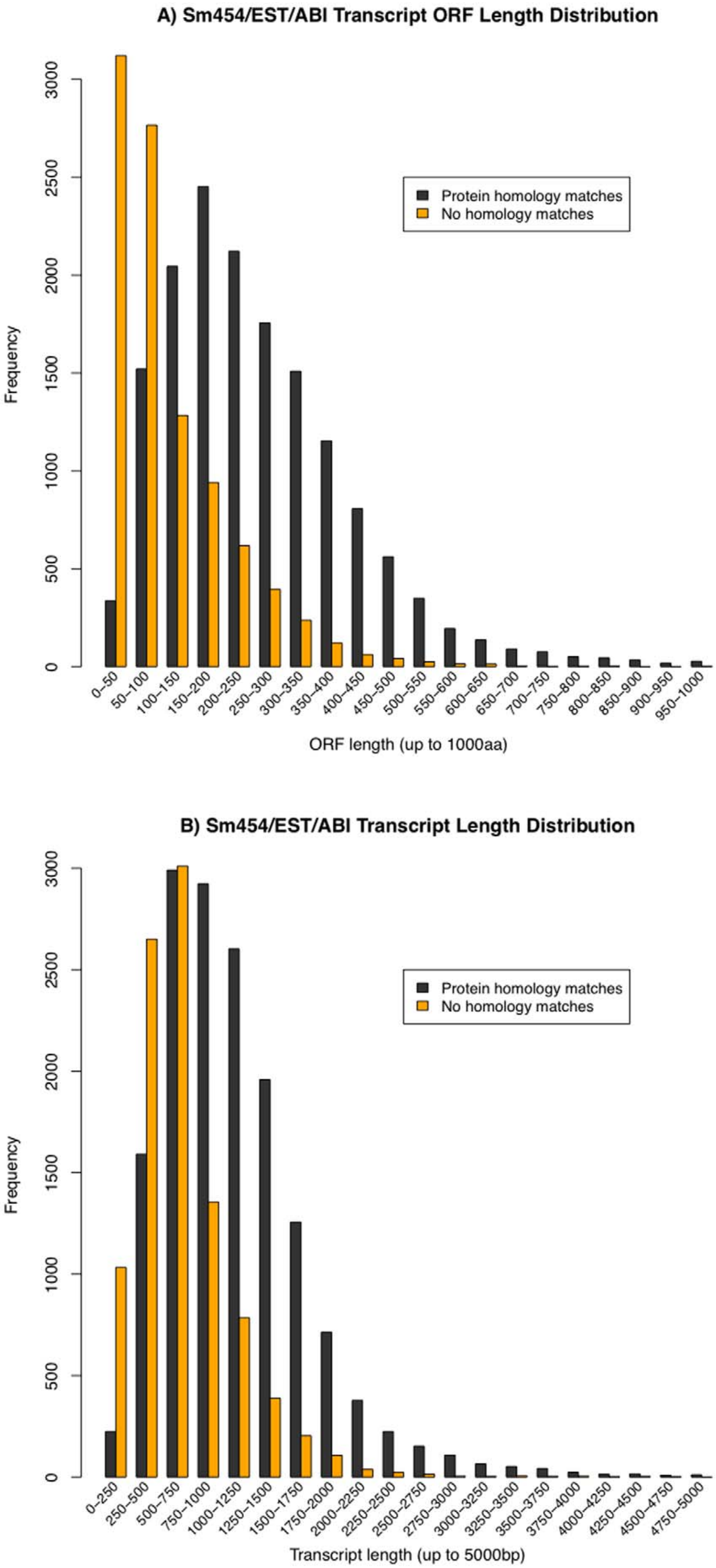
A) Sm454/EST/ABI Transcript ORF Length Distribution
Frequency
0
500
1000
1500
2000
2500
3000
■ Protein homology matches
■ No homology matches
0–50
50–100
100–150
150–200
200–250
250–300
300–350
350–400
400–450
450–500
500–550
550–600
600–650
650–700
700–750
750–800
800–850
850–900
900–950
950–1000
ORF length (up to 1000aa)
B) Sm454/EST/ABI Transcript Length Distribution
Frequency
0
500
1000
1500
2000
2500
3000
■ Protein homology matches
■ No homology matches
0–250
250–500
500–750
750–1000
1000–1250
1250–1500
1500–1750
1750–2000
2000–2250
2250–2500
2500–2750
2750–3000
3000–3250
3250–3500
3500–3750
3750–4000
4000–4250
4250–4500
4500–4750
4750–5000
Transcript length (up to 5000bp)

Figure 9. Distribution of ORF and transcript lengths with and without homology based annotation. Frequency plots of open reading frame length (A) and total transcript length separated by whether a transcript had homology (black) to another species by BLAST, PFAM and/or GO assignment or did not have any homology (orange) while shorter coding regions and transcript lengths are less likely to have homology this analysis suggests that many short transcripts and of the short ORFs (in the 0–50 aa range) are likely to be non-coding transcripts of the untranslated regions of non-coding transcripts.

doi:10.1371/journal.pone.0015617.g009

instructions. All data is available at the NCBI short read archive under study number SRP002478.

Preparation and sequencing of SOLiD 3 whole transcriptome libraries

Pooled total RNA from regenerative stages and in tact worms was enriched for mRNA enriched using the Poly A Purist Kit (Ambion, Cat. No. AM1919) followed by further depletion of ribosomal RNA using the Ribominus Eukaryotic kit (Invitrogen, Cat. No. A10837-08). Solid whole transcriptome libraries were made as outlined in the Solid Whole transcriptome kit protocol (Applied Biosystems, Cat. No. 4425680). The Quant-it HS dsDNA assay kit (Invitrogen, Cat. No. Q32851) was used to measure the concentration of libraries. Sequencing was performed on a SOLiD 3 ABi sequencer according to the manufacturers instructions to generate 50 bp reads in colour space. All data is available at the NCBI short read archive under study number SRP002478.

Assembly of 454 reads

Roche's de novo GS Assembler (Newbler 2.3) was used to assemble reads from the 454 platform. Newbler produces the assembly in FASTA format along with several metric files. The Newbler algorithm allows discovery of potential isoforms which are contained in the output as 'isotigs'. The publicly available ESTs for *S. meidterranea* were also added to the assembly and the resulting sequences are available in the Sm454EST.fa FASTA file.

Mapping and assembly of SOLiD 3 reads

The ABi transcriptome dataset was generated with an iterative mapping approach. Reads from the ABi SOLiD 3 platform was mapped using the Whole-Transcriptome Pipeline (WTP) within the BioScope 1.0 software package. The WTP of this mapping pipeline does not produce de novo split reads that span possible introns. It requires gene structure definitions in the form of a GTF formatted file to find possible split reads. We used MAKER gene annotations from SmedGD as the input gene definitions for WTP. The mapping produced a SAM formatted file containing mapping coordinates, mapping quality, and information on split reads. The Cufflinks program [41] was used define transcribed regions using the SAM mapping file as the input. We used the following parameters optimized to give correct coverage of a defined set of planarian genes that have been experimentally cloned and are present in the NCBI database. On this basis Cufflinks was run with the following parameters: Q 100 -I 100000 -F 0 -f 0.3 -j 0.1. Cufflinks based assembly was performed using only high quality uniquely mapped reads. The assembly produces a GTF gene annotation file of transcribed regions with all potentially transcribed regions, we required that these be represented by at least 2 uniquely mapped high quality reads.

Cufflinks uses split-reads to lace together exons into a genes and to define isoforms. In order to define as many possible split-reads as possible, a new gene structure definition file was created using the assembled GTF file. This file contains all assembled genes on a genomic contig laced together as one 'super-gene'. One super-gene was created for each strand of the genomic contig. This new pseudogene definition file was then used for remapping using WTP with only high quality reads that had previously only failed to map or only mapped along part of their length.

From the new SAM mapping file, all new split-reads were extracted and added to the initial SAM file. A re-assembly was done using Cufflinks on this new SAM file with more split-reads resulting in a dataset with significantly more multiple-exon genes. Putative transcripts were then extracted as cDNA sequences with introns removed. This data is available as the SmABi.fa FASTA file.

Comparison of datasets with Smed GD MAKER annotations

Comparison of Sm454EST, SmABI, Sm454ESTABI, and SmMAKER was done using BLASTN with the e-value threshold set at 1E-10. All three dataset fasta files were concatenated into one file and an all versus all blast was done. The blast results were then parsed, categorized, and filtered for only blast matches that had 95% or better identity.

Merging of datasets from 454 and SOLiD

The 454 dataset was used as a scaffold for the ABi dataset. The 454 assembly from Newbler contains possible isoforms. The isoforms with only terminal differences at the 3′ and 5′ ends were collapsed by running the dataset through CAP3 assembler [42]. The ABi dataset is was then blasted (BLASTN) against this collapsed 454 dataset. A list of blast matches and a list of non-matches are generated by filtering the blast results (BLASTN with the e-value threshold set at 1E-10, and filtered for only blast matches that had 95% or better identity across 50 nucleotides or more). Out of the list with non-matches, any ABi transcripts that had multiple exon structures were maintained. Finally, ABi transcripts with BLAST match(es) and ABi transcripts that made multiple exons with no BLAST matches were assembled with the 454 dataset with CAP3 assembler. This results in a list of assembled contigs and a list of singlets. From the list of singlets, any transcript that was only represented single exon ABi SOLiD data was removed. The assembled contigs and filtered singlets list is then concatenated together as the merged dataset resulting in 25,0053 putative transcripts. These are available in the Sm454ESTABI.fa.txt FASTA file (Dataset S4).

Open Reading Frame analysis

We used CLC Genomics Workbench (www.clcbio.com) for open reading frame analysis. ORFs were allowed to start and end outside the given sequence. Putiative peptides are available in the file Smed454ESTABI.orf.fa.txt FASTA file (Dataset S5)

Gene ontology and PFAM analysis

The complete six-frame amino acid translations of each transcript were tested for homology to Pfam-A protein families using the hmmscan algorithm applied to the Hidden Markov Model dataset (Pfam-A.hmm v.24) [43]. An e-value threshold of 1E-5 was applied to determine family homology of each transcript. Gene Ontology (GO) terms where assigned to Pfam matched transcripts via the pfam2go lookup table (http://www.geneontology.org/external2go/pfam2go). This procedure was implemented using a Perl script.

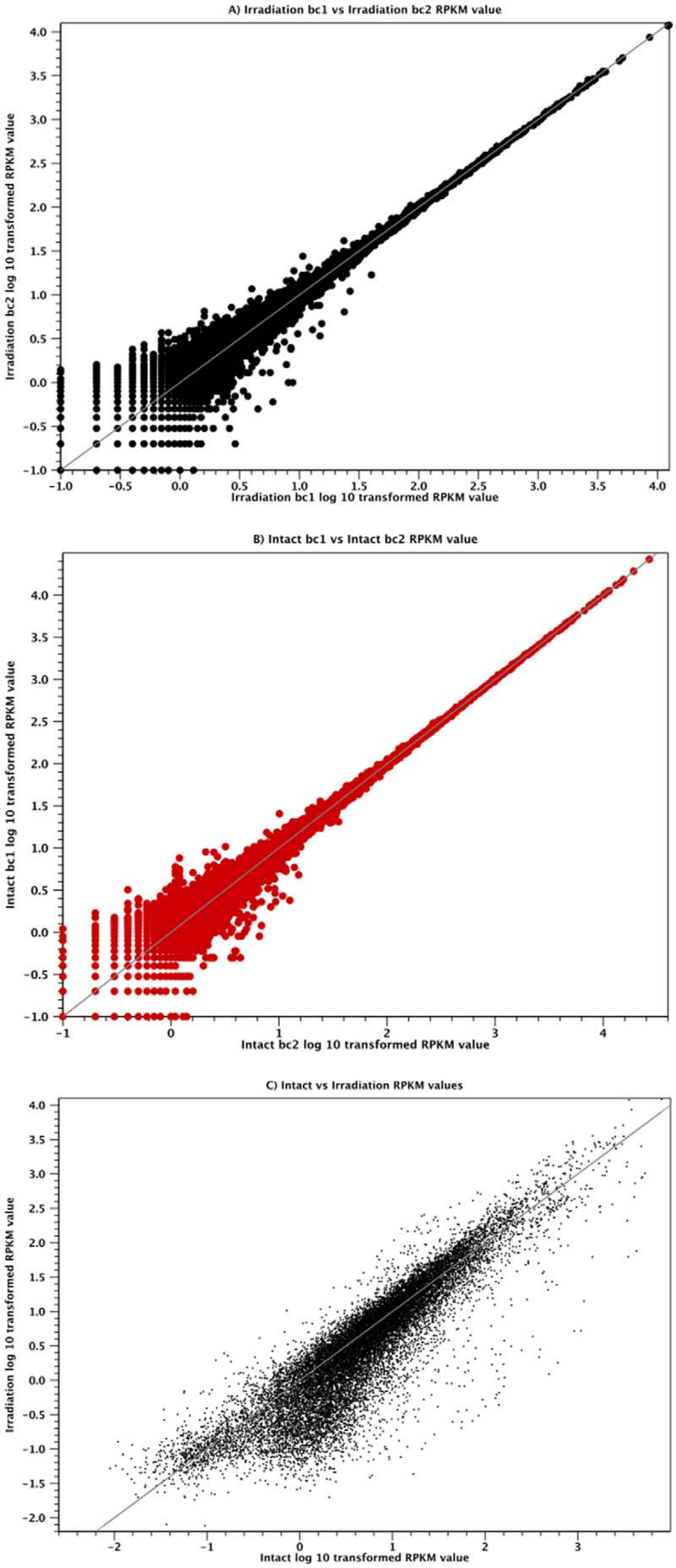
A) Irradiation bc1 vs Irradiation bc2 RPKM value
Irradiation bc2 log 10 transformed RPKM value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Irradiation bc1 log 10 transformed RPKM value
B) Intact bc1 vs Intact bc2 RPKM value
Intact bc1 log 10 transformed RPKM value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
2
3
4
Intact bc2 log 10 transformed RPKM value
C) Intact vs Irradiation RPKM values
Irradiation log 10 transformed RPKM value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3
Intact log 10 transformed RPKM value

Figure 10. RNA seq expression values of Whole vs Irradiated expression. Plots show Log 10 transformed RPKM values for transcripts detected above an RPKM of >1 (Log 10 RPKM>0) in at least one sample. We compared the two replicates of RNAseq of irradiated worms (A) and intact (mock irradiated) to each other demonstrating a high level of congruency with differences in expression values only apparent at low expression levels. We then compared the replicates of irradiated worms against the replicates of intact worms (C) and are able to detect many differentially expressed transcripts, many of which decrease after irradiation. These represent a group of genes potentially expressed in planarian stem cell and their recent progeny.
doi:10.1371/journal.pone.0015617.g010

GO terms were also assigned to transcripts using the Blast2go pipeline version 2.3.5 [45]. Transcript sequences were compared initially to a BLAST database comprising the NCBI invertebrate non-redundant (nr) protein dataset supplemented with the proteome sequences used in the cross-species homology analysis. Sequence redundancy was removed according to unique GenBank Identification (GI) code. A lower e-value threshold of 1e-5 was used to determine significant homology matches and results were recorded in XML format. GO terms where then assigned to each transcript through automated queries to the Blast2go MySQL term tables. The resulting Blast2go annotations where then added to those defined by Pfam2go with duplicate GO term assignments removed. The classifications for individual transcripts are presented as features in Dataset S6. Determining the total number of transcripts classified according to first tier terms of the three main domains of the GO database was achieved using a Perl script querying the Blast2go MySQL term tables.

Analysis of homology with other animals

The number of protein sequences defined for each species used in the cross-species homology analysis were as follows: *Strongylocentrotus purpuratus*: 42420, *Homo sapiens*: 37391, *Mus musculus*: 34962, *Arabidopsis thaliana*: 33049, *Branchiostoma floridae*: 28623, *Caenorhabditis elegans*: 23894, *Drosophila melanogaster*: 17702, *Hydra magnipapillata*: 17398, *Ciona intestinalis*: 13842, *Schistosoma mansoni*: 12856, *Anopheles gambiae*: 12659, *Brugia malayi*: 11470, *Acyrthosiphon pisum*: 10466, *Tribolium castaneum*: 9833, *Nasonia vitripennis*: 9254, *Apis mellifera*: 9244, *Schizosaccharomyces pombe*: 5000, *Saccharomyces cerevisiae*: 4027. Proteome sequences were sourced from NCBI except those of *Schistosoma mansoni*; sourced from GeneDB [62]. Transcript sequences where compared to the blast database of the proteins using BLASTX. A minimum e-value threshold of 1e-5 was applied and results were recorded in XML format. Default values for all other parameters where used. Blast results where processed using scripts implementing BioPerl modules. The protein sequence conferring the closest match to each transcript was determined by E-value, then by bit score. These matches are also presented as features in Dataset S6.

RNAseq analysis of irradiated versus whole worms

Two batches of 10 whole worms were starved for one week prior to gamma irradiation, exactly as previously described. Two batches of matched controls were mock irradiated. RNA from all four batches was prepared as described above and used to make four separate SOLiD 3 transcriptome libraries (see above). These were then sequenced as four separate barcodes on one flow cell to generate 50 bp color space reads. Reads were mapped to the Sm454ESTABI transcriptome with Bioscope version 1.0 WTO pipeline. The number of unique tags per transcript for each sample were counted and converted into Reads per KB per million (RPKM) values. These are values normalized by number of reads and gene length.

We used the statistical package for gene expression analysis within the CLC BIO Genomics Workbench (www.clcbio.com) to

Table 3. RPKM values of known genes in whole vs. irradiated worms.

Gene/EST and species	Accession	Reference	Sm454ESTABI Dataset ID	Mean Intact RPKM	Mean Irradiated RPKM	P-value	Fold-Change
Smedwi-1	DQ186985	[52]	9082	275.7	19.3	2.17E-121	−14.28
Smedwi-2	DQ186986	[52,58]	20642	134.7	0.65	1.19E-77	− 207.23
Smed-bruli	DQ344977	[55]	2960	34.3	19.4	5.29E-03	−1.77
Smed-Histone-H2B	DN290330	[55]	17226	50.45	15.1	1.34E-8	−3.34
Smed-nanos-like	EF035555	[57]	24540	12.6	0.6	1.71E-07	−21
Dj-PCNA	EU856391	[59]	22122	46.85	2.35	9.25E-24	−19.94
Spol-Tudor	FJ655915	[50]	16133	49.25	20.65	1.36E-06	−2.38
Smed-HDAC	NP_004955	[47]	22368	190.25	94.9	3.65E-15	−2
Smed-NB.21.11.e	FG311845	[47]	17431	155.9	0.9	3.64E-89	−173.22
Smed-AGAT-1	AY967547	[47]	17877	95.75	0.8	3.2E-45	−118.75
Smed-p53	AY068713	[53]	19247	15.0	0.8	1.64E-8	−18.25
Dj-rrm2	AY067447	[48]	13945	70.2	8.55	1.34E-26	−8.21
smed-SMB	GU562964	[56]	19507	122	36.15	3.05E-23	−3.36
Smed-INX-11	Q851133	[54]	18619	50.2	13.4	1.34E-11	−3.75
Dj-Pumilo	AJ639658	[48,49]	19575	72.25	28.55	6.22E-10	−2.53
Dj-EST, similar to H2Az)	BP186772	[48]	15734	47.55	1.4	5.39E-26	−33.96
DjEST, similar to TAF1-beta	BP187634	[48]	16295	584.95	207.15	1.42E-82	−2.82
Smed-CHD4	GU980571	[60]	4553	22.15	2.2	3.77E-10	−10.07

doi:10.1371/journal.pone.0015617.t003

analyze gene expression levels between replicates and between intact and irradiated worms ($p > 0.01$, using the test of Baggerey et al [61]). We did not consider genes that had very low expression levels across all 4 samples (RPKM<1, Log10 of RPKM<0).

We specifically identified transcripts representing genes that have been experimentally confirmed in published literature as being expressed in neoblast stem cells or their progeny and compared their RPKMs between whole and irradiated worms to evaluate our dataset (Table 3).

Supporting Information

Dataset S1 TheSm454EST dataset in FASTA format is an assembly fof 454 reads generated in this work and publicly available ESTs.
(TXT)

Dataset S2 The SmMAKER dataset in FASTA format.
(TXT)

Dataset S3 The SmABI datset in FASTA format.
(TXT)

Dataset S4 Sm454ESTABI dataset iN FASTA format.
(TXT)

Dataset S5 The longest predicted open reading frames of the Sm454ESTABI dataset in FASTA format with positional coordinates and strand information along the coordinates
(TXT)

Dataset S6 The data file detailing the cross species BLAST results top hits.
(TXT)

Dataset S7 The data file detailing all of the GO and PFAM annotations of the Sm454ESTABI dataset.
(TXT)

Dataset S8 The data file containing the FASTA files of transcripts differentially regulated after irradiation.
(TXT)

Dataset S9 Data file containing P-values, fold change, BLAST and GO information for transcripts differentially regulated after irradiation.
(TXT)

Acknowledgments

We thank members of the Aboobaker lab for encouragement and discussion.

Author Contributions

Conceived and designed the experiments: AAA. Performed the experiments: MJB DK SM JR RW DE AH JJ VL SL AAA. Analyzed the data: MJB DK JR AAA. Contributed reagents/materials/analysis tools: AAA. Wrote the paper: AAA MJB DK JR.

References

1. Brockes JP, Kumar A (2008) Comparative aspects of animal regeneration. Annu Rev Cell Dev Biol 24: 525–549.
2. Reddien PW, Sanchez Alvarado A (2004) Fundamentals of planarian regeneration. Annu Rev Cell Dev Biol 20: 725–757.
3. Salo E (2006) The power of regeneration and the stem-cell kingdom: freshwater planarians (Platyhelminthes). Bioessays 28: 546–559.
4. Yazawa S, Umesono Y, Hayashi T, Tarui H, Agata K (2009) Planarian Hedgehog/Patched establishes anterior-posterior polarity by regulating Wnt signaling. Proc Natl Acad Sci U S A 106: 22329–22334.
5. Rink JC, Gurley KA, Elliott SA, Sanchez Alvarado A (2009) Planarian Hh signaling regulates regeneration polarity and links Hh pathway evolution to cilia. Science 326: 1406–1410.
6. Petersen CP, Reddien PW (2009) A wound-induced Wnt expression program controls planarian regeneration polarity. Proc Natl Acad Sci U S A 106: 17061–17066.
7. Adell T, Salo E, Boutros M, Bartscherer K (2009) Smed-Evi/Wntless is required for beta-catenin-dependent and -independent processes during planarian regeneration. Development 136: 905–910.
8. Petersen CP, Reddien PW (2008) Smed-betacatenin-1 is required for anteroposterior blastema polarity in planarian regeneration. Science 319: 327–330.
9. Iglesias M, Gomez-Skarmeta JL, Salo E, Adell T (2008) Silencing of Smed-betacatenin1 generates radial-like hypercephalized planarians. Development 135: 1215–1221.
10. Gurley KA, Rink JC, Sanchez Alvarado A (2008) Beta-catenin defines head versus tail identity during planarian regeneration and homeostasis. Science 319: 323–327.
11. Reddien PW, Bermange AL, Kicza AM, Sanchez Alvarado A (2007) BMP signaling regulates the dorsal planarian midline and is needed for asymmetric regeneration. Development 134: 4043–4051.
12. Molina MD, Salo E, Cebria F (2007) The BMP pathway is essential for re-specification and maintenance of the dorsoventral axis in regenerating and intact planarians. Dev Biol 311: 79–94.
13. Kobayashi C, Saito Y, Ogawa K, Agata K (2007) Wnt signaling is required for antero-posterior patterning of the planarian brain. Dev Biol 306: 714–724.
14. Felix DA, Aboobaker AA (2010) The TALE class homeobox gene Smed-prep defines the anterior compartment for head regeneration. PLoS Genet 6: e1000915.
15. Adell T, Cebria F, Salo E (2010) Gradients in planarian regeneration and homeostasis. Cold Spring Harb Perspect Biol 2: a000505.
16. Wenemoser D, Reddien PW (2010) Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. Dev Biol 344: 979–991.
17. Newmark PA, Sanchez Alvarado A (2002) Not your father's planarian: a classic model enters the era of functional genomics. Nat Rev Genet 3: 210–219.
18. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–141.
19. Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, et al. (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11: 400.
20. Cantacessi C, Mitreva M, Jex AR, Young ND, Campbell BE, et al. (2010) Massively parallel sequencing and analysis of the Necator americanus transcriptome. PLoS Negl Trop Dis 4: e684.
21. Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, et al. (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Mol Ecol 19 (Suppl 1): 197–211.
22. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11: 180.
23. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6: 377–382.
24. Fullwood MJ, Wei CL, Liu ET, Ruan Y (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 19: 521–532.
25. Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, et al. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol 8: R145.
26. Boutros M, Ahringer J (2008) The art and design of genetic screens: RNA interference. Nat Rev Genet 9: 554–566.
27. Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, et al. (2008) A protein domain-based interactome network for C. elegans early embryogenesis. Cell 134: 534–545.
28. Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19: 706–712.
29. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18: 188–196.
30. Robb SM, Ross E, Sanchez Alvarado A (2008) SmedGD: the Schmidtea mediterranea genome database. Nucleic Acids Res 36: D599–606.
31. Reddien PW, Bermange AL, Murfitt KJ, Jennings JR, Sanchez Alvarado A (2005) Identification of genes needed for regeneration, stem cell function, and tissue homeostasis by systematic gene perturbation in planaria. Dev Cell 8: 635–649.
32. Zayas RM, Hernandez A, Habermann B, Wang Y, Stary JM, et al. (2005) The planarian Schmidtea mediterranea as a model for epigenetic germ cell specification: analysis of ESTs from the hermaphroditic strain. Proc Natl Acad Sci U S A 102: 18491–18496.

33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
34. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.
35. Vivancos AP, Guell M, Dohm JC, Serrano L, Himmelbauer H (2010) Strand-specific deep sequencing of the transcriptome. Genome Res 20: 989–999.
36. Mollet IG, Ben-Dov C, Felicio-Silva D, Grosso AR, Eleuterio P, et al. (2010) Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. Nucleic Acids Res 38: 4740–4754.
37. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509–1517.
38. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.
39. Kircher M, Kelso J (2010) High-throughput DNA sequencing–concepts and limitations. Bioessays 32: 524–536.
40. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, et al. (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25: 2872–2877.
41. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515.
42. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868–877.
43. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–222.
44. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, et al. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. Nucleic Acids Res 37: D396–403.
45. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.
46. Chandebois R (1965) [Changes in the regenerative capacity of irradiated planarians (Dugesia subtentaculata) as a function of the x-ray dose]. C R Acad Sci Hebd Seances Acad Sci D 260: 4834–4837.
47. Eisenhoffer GT, Kang H, Sanchez Alvarado A (2008) Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian Schmidtea mediterranea. Cell Stem Cell 3: 327–339.
48. Rossi L, Salvetti A, Marincola FM, Lena A, Deri P, et al. (2007) Deciphering the molecular machinery of stem cells: a look at the neoblast gene expression profile. Genome Biol 8: R62.
49. Salvetti A, Rossi L, Lena A, Batistoni R, Deri P, et al. (2005) DjPum, a homologue of Drosophila Pumilio, is essential to planarian stem cell maintenance. Development 132: 1863–1874.
50. Solana J, Lasko P, Romero R (2009) Spoltud-1 is a chromatoid body component required for planarian long-term stem cell self-renewal. Dev Biol 328: 410–421.
51. Salvetti A, Rossi L, Deri P, Batistoni R (2000) An MCM2-related gene is expressed in proliferating cells of intact and regenerating planarians. Dev Dyn 218: 603–614.
52. Reddien PW, Oviedo NJ, Jennings JR, Jenkin JC, Sanchez Alvarado A (2005) SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. Science 310: 1327–1330.
53. Pearson BJ, Sanchez Alvarado A (2010) A planarian p53 homolog regulates proliferation and self-renewal in adult stem cell lineages. Development 137: 213–221.
54. Oviedo NJ, Levin M (2007) smedinx-11 is a planarian stem cell gap junction gene required for regeneration and homeostasis. Development 134: 3121–3131.
55. Guo T, Peters AH, Newmark PA (2006) A Bruno-like gene is required for stem cell maintenance in planarians. Dev Cell 11: 159–169.
56. Fernandez-Taboada E, Moritz S, Zeuschner D, Stehling M, Scholer HR, et al. (2010) Smed-SmB, a member of the LSm protein superfamily, is essential for chromatoid body organization and planarian stem cell proliferation. Development 137: 1055–1065.
57. Wang Y, Zayas RM, Guo T, Newmark PA (2007) nanos function is essential for development and regeneration of planarian germ cells. Proc Natl Acad Sci USA 104: 5901–5906.
58. Palakodeti D, Smielewska M, Lu YC, Yeo GW, Graveley BR (2008) The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. RNA 14: 1174–1186.
59. Orii H, Sakurai T, Watanabe K (2005) Distribution of the stem cells (neoblasts) in the planarian Dugesia japonica. Dev Genes Evol 215: 143–157.
60. Scimone ML, Meisel J, Reddien PW (2010) The Mi-2-like Smed-CHD4 gene is required for stem cell differentiation in the planarian Schmidtea mediterranea. Development 137: 1231–1241.
61. Baggerly KA, Deng L, Morris JS, Aldaz CM (2003) Differential expression in SAGE: accounting for normal between-library variation. Bioinformatics 19: 1477–1483.
62. Hertz-Fowler C, Peacock CS, Wood V, Aslett M, Kerhornou A, et al. (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. Nucleic Acids Res 32: D339–343.