Stem cell transcriptome profiling via massive-scale mRNA sequencing

Nicole Cloonan[1,4], Alistair R R Forrest[1,3,4], Gabriel Kolle[1,4], Brooke B A Gardiner[1], Geoffrey J Faulkner[1], Mellissa K Brown[1], Darrin F Taylor[1], Anita L Steptoe[1], Shivangi Wani[1], Graeme Bethel[1], Alan J Robertson[1], Andrew C Perkins[1], Stephen J Bruce[1], Clarence C Lee[2], Swati S Ranade[2], Heather E Peckham[2], Jonathan M Manning[2], Kevin J McKernan[2] & Sean M Grimmond[1]

We developed a massive-scale RNA sequencing protocol, short quantitative random RNA libraries or SQRL, to survey the complexity, dynamics and sequence content of transcriptomes in a near-complete fashion. This method generates directional, random-primed, linear cDNA libraries that are optimized for next-generation short-tag sequencing. We surveyed the poly(A)$^+$ transcriptomes of undifferentiated mouse embryonic stem cells (ESCs) and embryoid bodies (EBs) at an unprecedented depth (10 Gb), using the Applied Biosystems SOLiD technology. These libraries capture the genomic landscape of expression, state-specific expression, single-nucleotide polymorphisms (SNPs), the transcriptional activity of repeat elements, and both known and new alternative splicing events. We investigated the impact of transcriptional complexity on current models of key signaling pathways controlling ESC pluripotency and differentiation, highlighting how SQRL can be used to characterize transcriptome content and dynamics in a quantitative and reproducible manner, and suggesting that our understanding of transcriptional complexity is far from complete.

Over the past decade, there has been a dramatic change in our understanding of the scale of eukaryotic transcriptional complexity. Transcriptome annotation initiatives such as the expressed sequence tag (EST) projects[1,2], functional annotation of the mouse (FANTOM)[3–5] and encyclopedia of DNA elements (ENCODE)[6] have used cDNA sequencing and array-based profiling to map locus-based mRNA expression from a diverse range of mammalian tissues, cell lines and pathological states. It is now clear that rather than encoding a single transcript, loci are capable of generating, on average, six different mRNAs[6]. This is complicated by the more recent discovery that loci also frequently express noncoding transcripts[7] and that both sense and antisense gene transcripts are common[8].

Despite this extensive genome-wide cartography in mammalian transcriptomes, attempts to put the amassed mRNA complexity into the biological context of specific model systems have been largely unsuccessful primarily due to the constraints associated with current profiling technologies. Array-based profiling methods are now the most popular surveying tools with tiling[9], all-exon[10] and exon-junction[11] array strategies in use. Unfortunately, these approaches are physically constrained by issues of probe density, exon size, suitable sequence content, cross-hybridization, poor sensitivity for rarely expressed transcripts and the difficulty in identifying exon combinations. Tag-based profiling approaches such as massively parallel signature sequencing (MPSS)[12], serial analysis of gene expression (SAGE)[13], cap analysis of gene expression (CAGE)[14] and polony multiplex analysis of gene expression (PMAGE)[15] have fared better with respect to sensitivity and discrimination of some signals but suffer from the ambiguity associated with mapping of their short sequence output (17–20 nt) to complex genomes.

To create a more complete survey of transcriptome content and dynamics, we used a next-generation sequencing approach, applying the Applied Biosystems SOLiD technology to generate gigabase-scale shotgun short-sequence data from cDNA libraries. Our method, SQRL, generates short quantitative random RNA libraries, which can be used to profile transcriptome content and dynamics (**Fig. 1**).

Previous efforts to profile the dynamic transcriptome of ESC differentiation have shown that EB cultures recapitulate temporal transcriptional programs common to early stages of embryo development *in vivo*, ranging from the inner cell mass (ESC day 0), epiblast (day 3), primitive streak (day 4) and mesoderm (day 5–6)[16,17]. To test the potential of SQRL, we surveyed the locus activity and transcript-specific expression of ESC and EB transcriptomes to an unprecedented level of resolution and sensitivity by massive-scale sequencing.

RESULTS

Deep sequencing of SQRL libraries

To obtain a thorough survey of the ESC and primitive streak stage EB mRNA transcriptomes, we generated triplicate libraries from two states (ESC and EB) and sequenced them to a depth of almost 100 million high-quality mappable reads (**Supplementary Table 1**

[1]Expression Genomics Laboratory, Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St. Lucia, Queensland, 4072, Australia. [2]Applied Biosystems Inc., 500 Cummings Center, Beverly, Massachusetts 01915, USA. [3]Present address: The Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan, Queensland, 4111, Australia. [4]These authors contributed equally to this work. Correspondence should be addressed to S.M.G. (s.grimmond@imb.uq.edu.au).

RECEIVED 25 FEBRUARY; ACCEPTED 19 MAY; PUBLISHED ONLINE 30 MAY 2008; DOI:10.1038/NMETH.1223

online). We mapped the tags to the mouse genome (release mm9, July 2007) and a database of unique exon-junction sequences (see **Supplementary Fig. 1** online for a discussion of mapping using color-space methods). We used this dual-matching strategy to ensure capture of both exonic expression and combinatorial exon usage for every locus. To confirm the directionality of the SQRL library method, we matched the tags to the opposite strand in the exon-junction database; more than 99.5% of all junction tags matched in the sense orientation (**Supplementary Table 2** online). To maximize tag-mapping accuracy and coverage, we used only high-quality sequences of at least 25 nt. We used a multi-mapping rescue strategy previously used on CAGE data to assign tags with low-level mapping ambiguity to genomic locations[18] (**Supplementary Fig. 2** and **3** online). To determine the saturation of transcriptome coverage in both the undifferentiated and differentiated states, we used start-site frequency of tags. We judged the coverage to be near-complete as tag saturation was not overcome by the inclusion of additional independent SQRL libraries for each state (**Supplementary Fig. 4** online).

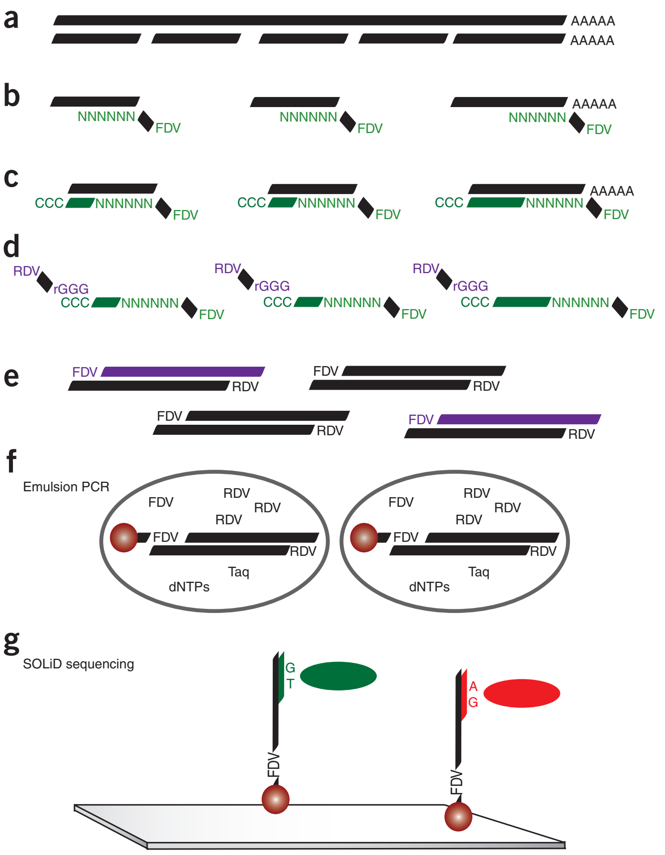

Figure 1 | Schematic of SQRL method. (**a**) Ribosomal RNA–depleted mRNA is partially fragmented. (**b**) First-strand cDNA is then reverse transcribed from a random hexamer primer tagged with a flanking sequence (FDV). (**c**) Template switching is used to add non-templated cytosines to each cDNA. (**d**) A DNA-RNA oligonucleotide template-switching primer, tagged with a second flanking sequence (RDV) is then used to incorporate a priming site for RDV into the first-strand cDNA. (**e**) The library is PCR-amplified using FDV and RDV primers. (**f**) The library is amplified and attached to beads (for subsequent sequencing) via emulsion PCR. (**g**) Beads are covalently bound to a glass slide, and short reads (25–35 nt) are generated from individual cDNA fragments using SOLiD sequencing (Applied Biosystems). Fluorescently labeled di-base probes are used to sequence the tag.

To visualize expression we prepared genome-wide, strand-specific, nucleotide-resolution plots of tag coverage for each individual library and for combined libraries corresponding to the ESC and EB states. We also generated additional plots that display the relative exon combinations observed from each locus. These resources provide a way to study the genomic landscape of gene expression alongside public genome annotations within the University of California Santa Cruz (UCSC) Genome Browser[19] as 'custom tracks' (**Fig. 2** and **Supplementary Table 3** online). Although tag coverage clearly defined the location of exons, there was considerable variation in tag coverage across exons, with many peaks and troughs in coverage observed. The cause of this variation appears to be the presence of regions of mapping ambiguity (that is, sequence to which tag matches in the exon occur in multiple regions of the genome and cannot be mapped with confidence) and insertion-deletion events between the sample and C57BL6 reference genome, which also hinders tag mapping. Additionally, in an analysis of bias in sequence content and the influence of flanking sequences on tag generation, we found a higher preference for thymine upstream of each tag and for cytosine at the random hexamer site (**Supplementary Fig. 5** online).

Finally, we combined all tag maps with existing genome and transcriptome annotations available in the UCSC Genome Browser and used this information to address: (i) what transcriptional units are active, (ii) what evidence supports transcriptional complexity from these loci, (iii) what transcriptional activity is present outside of currently defined annotations, and (iv) which previously unidentified nonsynonymous SNPs are expressed in these cells.

ESC and EB locus activity

We calculated locus activity by counting the number of tags mapping to exons of each UCSC Genome Browser–defined gene. Comparison of SQRL libraries revealed that the methods were highly reproducible, with Pearson correlations of 0.99 and 0.95 for technical and biological replicates, respectively (**Supplementary Fig. 6** online). We scored ESC and EB locus activity as a series of tag count thresholds (**Supplementary Table 4** online) and compared this to activity predicted by present or absent calls on a standard microarray profiling platform. Using a threshold of 50 tags per gene, we identified 11,436 genes via SQRL compared to 7,180 detected by array profiling using Illumina arrays ('present' and 'marginal' Refseqs with a precision score ≥ 0.99; **Fig. 3a**). We used quantitative reverse-transcriptase (RT)-PCR to validate 79 transcripts shown to be active by SQRL. Many of these were below the level of detection of the array platform, confirming that SQRL profiling was still accurate and quantitative for these rarely expressed genes (**Fig. 3b** and **Supplementary Table 5** online).

Given the quantitative nature of SQRL, we used tag counts to define differential expression between ESC and EB states; we used an empirical Bayes approach to define significant differential expression. We observed 1,136 differentially expressed genes with 679 genes up-regulated in ESCs and 457 up-regulated in EBs. This showed strong concordance ($P < 6.0 \times 10^{-33}$) to previously published ESC expression markers[16] (**Supplementary Fig. 7** online), with 35/50 of ESC markers, including *Spp1*, *Esrrb*, *Klf4*, *Lefty1*, *Zfp42* and *Tdh* (**Supplementary Table 6** online), also significantly up-regulated in SQRL ESC profiling. Additionally, we compared our data to publicly available MPSS data, finding good concordance (Pearson correlation = 0.65) with the MPSS SV129 ESC sample (**Supplementary Fig. 8** online).

Transcriptional complexity in ESCs and EBs

Given that short sequence tags can only truly be used to interrogate single discrete regions of a transcript, these data could not be used to define the complete exon content of every individual full-length mRNA. Instead, we developed a method that profiles complexity by the recognition of 'diagnostic' exonic and exon-junction sequences. These are sequences within individual mRNAs that are unique, against the background of all mRNA transcripts currently described in public mRNA data repositories. To create this diagnostic sequence set, we used the latest version of AceView[20], a curated, comprehensive and non-redundant representation of public mRNA and EST sequences, which is also available through the UCSC Genome Browser. The September 2007 build of AceView contains 173,008 non-redundant transcripts from 65,363 loci in the mouse; 45.7% (78,791) of these have one or more diagnostic junction sequence, and 79.9% (138,156) contain a diagnostic exonic sequence. A comparison of SQRL tags with these diagnostic sequence annotations yielded evidence for 53,056 transcripts from 31,872 ESC loci and 50,881 mRNAs from 29,606 EB loci (**Supplementary Table 7** online).

Transcriptome discovery

Total RNA contains low levels of partially spliced transcripts, and intron retention could potentially confound our results. To test this possibility, we assessed the relative exon to intron signal for each known locus. We observed mean 48.7- and 62.1-fold enrichment for exonic versus intronic tag counts for each RefSeq locus for ESC and EB states, respectively. Furthermore, the amount of intronic sequence tags does not scale with increasing abundance of transcripts (**Supplementary Fig. 9** online). Together, these results suggested that intron retention was not a serious confounder in this study.

One of the interesting outcomes of this study was the observation that 31–37% of all mappable tags in both libraries occurred outside the exons of known or predicted genes. To understand the genomic landscape of this expression, we classified all tags using a hierarchical strategy into the following categories: tags mapping to (i) known exons, (ii) predicted exons, (iii) known regions (strand-specific areas spanning the entire genomic region of a mouse gene or transcript, including intronic sequence), (iv) predicted regions (as previous), (v) conserved regions, and (vi) new sequences (**Table 1** and **Supplementary Table 8** online). Much of the non-exonic expression we observed in gene regions was pervasive and low. When we applied a threshold of 5× nucleotide coverage to these data, ~60% of all intronic signal was lost, compared to a 5% loss of exonic signal.

We independently surveyed repeat expression as these elements are not typically profiled using array-based approaches. Repeat elements are normally excluded from array expression profiling because of cross-hybridization of mRNAs from closely related sequences. Approximately 20% of all SQRL tags could be mapped with confidence within the repeat elements and low-complexity sequences defined by Repeatmasker (**Supplementary Table 9** online). A total of 354,218 repeat elements were expressed with an average tag count of 31 tags.

We also characterized natural antisense transcription in exons of known genes in our libraries. We found that antisense expression was enriched in the 3′ exons of transcripts (**Supplementary Fig. 10** online). This is consistent with previous reports of the poly(A)$^+$ fraction of the transcriptome[21] and is due to overlapping 'tail-to-tail' gene expression on opposing strands (**Supplementary Table 10** online).

As newly identified alternative splicing generates new combinations of exon sequences, we wished to see whether it was possible to identify such events. We clustered high-quality 35-mer tags that did not map to the genome or to the junction library using a modified version of the VCAKE[22] algorithm to form a longer consensus sequence (≥50 nt) and aligned these clusters to the genome using BLAT[23]. We examined high-quality single-mapping alignments as candidates for new junctions. Using this approach we successfully identified 42 junctions excluded from our original junction set because their sequence occurred in multiple genes. Additionally, we detected 35 candidate splicing events (**Supplementary Fig. 11** and **Supplementary Table 11** online).

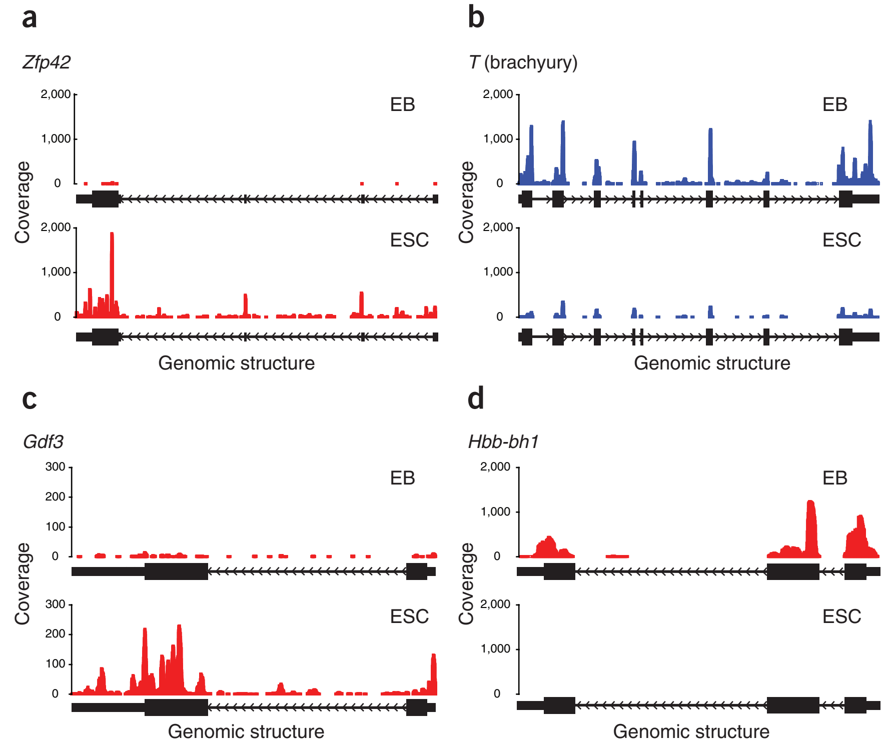

Figure 2 | Visualization of SQRL data on the UCSC genome browser. (**a**–**d**) 'Wiggle' plots showing coverage of SQRL tags for genes previously associated with the ESC state, *Zfp42* (**a**) and *Gdf3* (**c**), or differentiation-specific markers, mesoderm marker *T* (**b**) and blood cell marker *Hbb-bh1* (embryonic hemoglobin; **d**). Relative sequence coverage is displayed for all tags, showing the location of exons across the locus. Expression in the positive strand is shown in red and the negative strand in blue. The genomic structure of the gene (major Refseq transcript) is shown at the bottom of each plot.

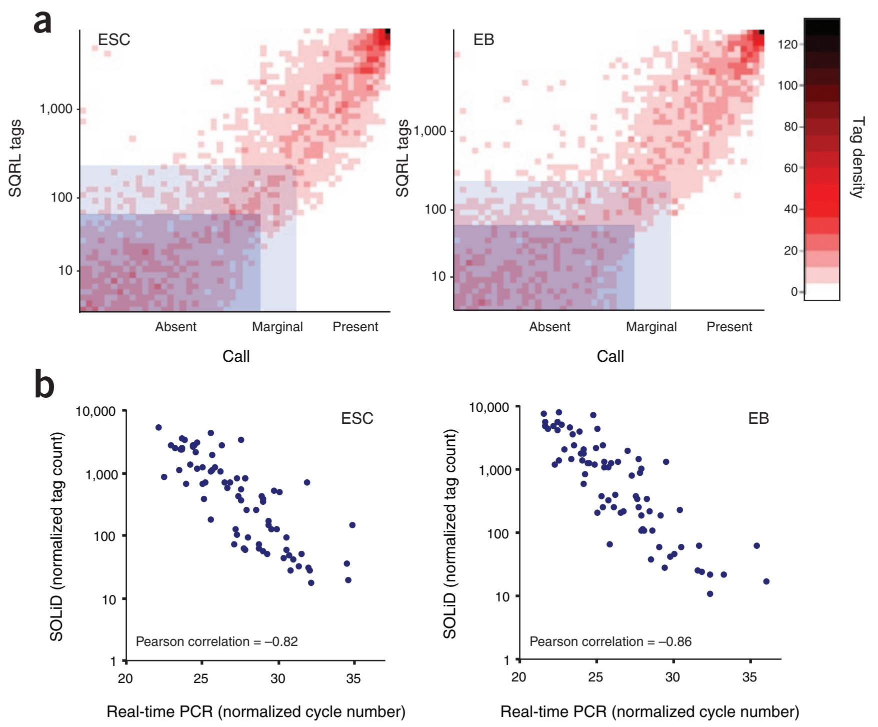

Figure 3 | Analysis of the quantitative nature and sensitivity levels of SQRL. (**a**) Comparison of SQRL tag count versus relative abundance measured by Illumina bead arrays for 17,325 Refseq genes. Absent (blue), marginal (light blue) and present (no shading) calls for Illumina and SQRL detection are indicated. (**b**) Graph comparing quantitative RT-PCR cycle number against tag abundance for 79 candidate transcripts.

Transcribed sequence variation analysis

We also used these sequence data to identify expressed polymorphisms between the reference genome (C57BL6) and the sample being studied (ESCs from a SV129 background). We found more than 2,000 candidate expressed SNPs in both ESC and EB libraries (see **Supplementary Fig. 12** online for the SNP discovery pipeline description). Of the predicted 643 SNPs found within the Refseq gene collection, 83.9% (540) of the variants had been previously reported in dbSNP (**Supplementary Table 12** online); presumably a reflection of focused SNP detection in mouse by screening mRNA and ESTs within the US National Center for Biotechnology Information (NCBI) database.

We were particularly interested in testing SQRL's ability to identify new non-synonymous SNPs between SV129 and C57BL6 in genes from RefSeq. We tested 18 candidates by PCR-amplifying their genomic region from ESC DNA, followed by Sanger sequencing. We confirmed 8/10 new non-synonymous SNPs and 8/8 UTR candidate variants, demonstrating that this approach can be used to screen for potential sequence variations (**Supplementary Table 13** online).

Functional implications of transcriptional complexity

We wished to place the previously uncharacterized transcriptional complexity in ESCs and EBs into a biological context. We specifically sought to address how it may impact current models of ESC pluripotency and pathways controlling differentiation. To do this we compiled a summary of all transcriptional output from TGFB, Wnt and FGF pathways and the previously defined pluripotency regulatory module for mouse ESCs[24] (**Supplementary Table 14** online). In ESCs we found evidence for 270 transcripts in 113 loci, which substantially increases the components known to be active in these key pathways in ESCs (**Supplementary Table 15** online). We

Table 1 | Summary of SQRL tag distribution

	ESCs combined (ES-2, ES-3 and ES-5)				EB combined (EB-1, EB-2, and EB-5)			
	Volume (all tags, Gb)	Proportion (all tags, %)	Volume with coverage ≥ 5× (Gb)	Proportion with coverage ≥ 5× (%)	Volume (all tags, Gb)	Proportion (all tags, %)	Volume with coverage ≥ 5× (Gb)	Proportion with coverage ≥ 5× (%)
Known expression								
Known exons	1.72	58.7	1.63	74.2	1.72	62.2	1.64	75.7
Predicted exons	0.13	4.5	0.12	5.6	0.18	6.4	0.17	7.8
Identified expression								
Known regions	0.42	14.3	0.17	7.9	0.32	11.4	0.12	5.6
Predicted regions	0.26	8.8	0.10	4.8	0.21	7.6	0.09	3.9
Conserved regions	0.29	10.0	0.10	4.7	0.24	8.6	0.08	3.8
Other regions	0.10	3.6	0.06	2.8	0.10	3.7	0.07	3.2
Total expression	2.93	100	2.20	100	2.77	100	2.17	100

All distributions were calculated for unfiltered tag coverage and for regions of ≥5× coverage per nucleotide. The known exon set was derived from a combination of Refseq, Mammalian Gene Collection open reading frames, the UCSC Known Genes dataset, mouse mRNAs and mouse ESTs as annotated in the UCSC tracks. Known regions of transcription were defined from the above tracks, using the genomic start and end coordinates of each transcript (and therefore including intronic regions). Overlapping regions were clustered as above. Predicted exons and regions of transcription were derived from non-mouse Refseqs, non-mouse ESTs and predictions from Aceview, Geneid and Genescan. Conserved regions of the genome were derived from the phastCons 30-way vertebrate conservation track.

performed isoform-specific RT-PCR on gene variants from these pathways with differing amounts of SQRL tag support. From 14 loci, we detected 34/37 variants tested in ESCs. These results (**Supplementary Table 5b**) confirmed not only the presence of these variants but also the dynamic expression of at least 8 of these transcripts between ESCs and EBs. The presence of alternative isoforms and previously unidentified transcription, even from small networks or pathways (**Fig. 4**), suggests that current models of regulatory networks are far from complete.

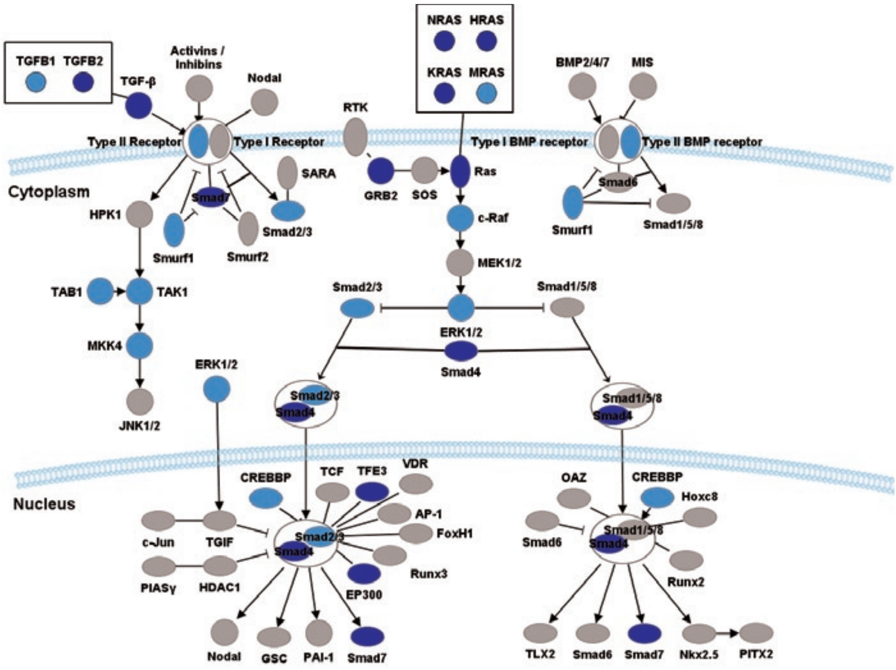

Figure 4 | Potential transcript variants for genes encoding components of the TGFB pathway. Genes that have evidence for expression of one transcript from Aceview are colored light blue, and those which have evidence for more than one transcript are shown in dark blue. Each transcript must have evidence for at least 2 tags that match across a diagnostic junction (**Supplementary Table 7**). The signaling pathways depicted in this figure and tabulated in **Supplementary Table 14** were derived from Ingenuity Pathways Analysis (Ingenuity Systems).

DISCUSSION

There is little dispute that array profiling methods have been a revolutionary tool in the post–genomic sequencing era, providing the means to survey gene expression and correlate gene activity with biological processes and pathological states. However, the extensive transcriptional complexity in many organisms means that transcriptomic content and dynamics cannot be effectively surveyed with current array technologies. Shotgun sequencing provides an effective alternative method for interrogating the transcriptome due to its open-ended scalability and the superior discriminatory power of sequence content over hybridization signals. The ability of SQRL profiling to detect expression events below the level of detection of standard microarrays and the ability to survey repeat elements are testament to the power of sequence-based profiling approaches. Our work on this mammalian transcriptome, using the Applied Biosystems SOLiD technology, is complementary to recent studies that have used Illumina-based sequencing to analyze the transcriptome in Arabidopsis and in yeast[25, 26] and highlights the importance of much larger scales of sequencing and the maintenance of tag strandedness in transcriptome analyses.

A major motivation behind the development of SQRL was to create a sensitive method to determine which loci are transcriptionally active and to determine which of all the known examples of alternative splicing events defined for the mouse are actually expressed in mouse ESC-EB differentiation. The SQRL analytical approach is now best suited to studying mammalian systems in which massive-scale shotgun profiling capability, completed genome sequences and well-annotated estimates of transcriptional complexity (from previous EST and full-length cDNA sequencing) provide a framework in which to characterize expression. Also, although known splicing events can be readily recognized with this approach, our ability to identify transcriptional complexity is limited. We demonstrated that it is possible to identify splice junctions by *de novo* assembly of new SQRL tags via clustering methods. However, the data obtained via this approach are not likely to represent all new 'events', as their detection is dependent on high levels of expression. We anticipate that longer tags or mate-pair sequencing strategies could improve the ability to link new splicing events to full-length variant mRNAs.

One of the exciting outcomes of these types of studies is the opportunity to robustly profile some of the more recently identified expression events, such as noncoding transcript expression, strand-specific expression and expression from ultraconserved regions, providing the means to correlate their transcriptional activity with biological phenomena. More interesting still is the ability to survey transcribed repeat elements, which clearly make up a large proportion of the transcriptome content and to which array-based approaches have been 'blind'.

This study has shown that it is possible to use SQRL sequencing to screen for expressed SNPs. The redundancy of transcripts per cell means that expressed SNPs are frequently sampled by SQRL sequencing, providing high confidence in SNP calls. The ability to study sequence variation concurrent with expression provides opportunities to study allele-specific expression, mutation status and RNA editing events on a genome-wide scale. This approach would certainly prove informative for studies in which information about genotype, copy-number variation or mutation burden in expressed genes is required.

Finally, although this study has focused on the poly(A)$^+$ fraction of the transcriptome, it could be equally applied to other fractions such as the nonadenylated or nuclear fractions of RNA. Future efforts are required to fully classify antisense and noncoding RNA expression[21] in these fractions.

METHODS

ESC and EB culture. We maintained W9.5 ESCs (P18) and differentiated EBs in 1% methylcellulose as previously described[27].

Library generation. SQRL creates random-primed double-stranded linear cDNA libraries without any adaptor ligation steps (**Supplementary Methods** online). We used fragmented rRNA-depleted poly(A)$^+$ mRNA as it closely matched the starting templates for other genomic analyses to which this study was compared. The pre-processing of mRNA maximized informative sequence content and ensured efficient generation of small cDNAs. We used a 3′ adaptor–tagged random hexamer primer to initiate first-strand cDNA synthesis and a reverse transcriptase capable of promoting efficient addition of non-templated cytosines to the termini of full-length first-strand cDNAs. Next we added a 5′ template switching' adaptor-tagged oligonucleotide[28]. Although this approach has been frequently used to identify capped 5′ ends of mRNAs[29], it is also possible to perform template switching on noncapped RNAs in standard conditions[30]. We observed no enrichment of 5′ end–derived tags in any of the SQRL libraries (**Supplementary Fig. 13** online), suggesting that fragmented RNA is captured just as efficiently as 5′-capped material. The incorporation of two different terminal adaptor tags into each cDNA fragment ensured that the strand of origin of the starting mRNA could be ascertained. Finally, we used limited rounds of PCR to amplify a pool of the SQRL cDNAs using primers complementary to the 5′ and 3′ adaptor sequences (**Fig. 1**).

SOLiD sequencing. SQRL libraries were sequenced using the Applied Biosystems SOLiD sequencing. We drove 500 pg of SQRL library onto 1-μm-diameter beads using emulsion PCR. We sequenced ~120,000,000 beads using 'sequencing by ligation' chemistry[31] on a SOLiD sequencer (Applied Biosystems). Approximately 60% of beads deposited generated high-quality sequence tags 25–35 nt in length. Wiggle plots and BED tracks displaying this data are available in **Supplementary Table 3**.

Illumina BeadChip gene expression profiling. We isolated total RNA from ESCs and EB cells as described above and hybridized it to Illumina Sentrix Mouse 6 v1.1 BeadChip arrays.

Quantitative real-time PCR. We designed 79 primer sets for real-time PCR for 70 genes. We designed PCR primers with a T_m of 58–60 °C and a length of 100–101 bp (**Supplementary Table 16** online). Using Superscript III and standard protocols (Invitrogen) we produced cDNA. We performed real-time PCR using Sybr Green (Applied Biosystems) on an ABI 7900HT using the following conditions: 50 °C for 2 min, 95 °C for 10 min and 40 cycles of 95 °C for 15 s, 60 °C for 1 min.

Statistical analysis. We compared SQRL and Illumina array profiling by comparing Illumina probe intensities versus SQRL tag counts for a non-redundant Refseq transcript set ($n = 19{,}005$). We mapped raw tag counts to Refseqs and normalized the gene signal relative to the length of the transcript. We mapped Illumina probes to the Refseq set using Vmatch. We compared SQRL and Illumina based on ranked data using the 'rank' method in R. When correlating library replicates, we performed quantile normalization using the Limma package in R on length-normalized SQRL tag counts with a floor set to 10. The Pearson correlation coefficient was calculated on the $\log_2$-transformed data.

To calculate differential expression of SQRL tag data we analyzed the normalized gene signals (tags per Refseq transcript, length-normalized) for each library using the Limma package in R. After Quantile normalization, we used Limma to fit a linear model to the $\log_2$-transformed data using an empirical Bayes method[32] to moderate standard errors. Differentially expressed genes were defined as those with a B statistic > zero. A hyper-geometric test with Bonferroni correction was used to calculate the probability of gene list overlap.

Accession numbers. NCBI Short Read Archive v0.6: SRA000306 (short tag data); Gene Expression Omnibus (GEO): GSE10518 (microarray).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

A.R.R.F. and S.M.G. are funded by the Australian National Health and Medical Research Council. A.R.R.F. is a C.J. Martin fellow (428261), and S.M.G. is a senior research fellow. N.C. is a University of Queensland postdoctoral fellow. G.J.F. and M.K.B. are supported by Australian Postgraduate awards. We acknowledge the Australian Research Council Centre for Functional and Applied Genomics Array Facility for expression profiling, and R. Nutter, G. Weightman and L. Stubberfield for support of this initiative.

AUTHOR CONTRIBUTIONS

N.C. created and integrated the sequence mapping and visualization pipeline, performed SOLiD sequencing bioinformatics, SNP analysis and splicing studies. A.R.R.F. conceived and pioneered the SQRL library strategy, performed preliminary genomic analysis, and developed the initial visualization methods. G.K. led the array-SQRL analyses, RT-PCR, pathway analysis and contributed to SNP analysis. B.B.A.G., M.K.B., G.K. and N.C. contributed to method design. A.L.S., G.K., S.J.B. and A.C.P. contributed to sample generation. G.K., A.R.R.F. and B.B.A.G. constructed libraries. C.C.L., S.S.R., B.B.A.G., G.B. and K.J.M. contributed to library sequencing. N.C., G.K., G.J.F., A.R.R.F., S.M.G., D.F.T., H.E.P. and J.M.M. contributed to data analysis. G.K., A.J.R., S.W., N.C. and A.L.S. contributed to experimental validation. S.M.G. supervised the work and prepared the manuscript with N.C., G.K. and A.R.R.F.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at
http://npg.nature.com/reprintsandpermissions/

1. Boguski, M.S. & Schuler, G.D. Establishing a human transcript map. *Nat. Genet.* **10**, 369–371 (1995).
2. Lee, Y. *et al.* The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.* **33**, D71–D74 (2005).
3. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
4. Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
5. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
6. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
7. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
8. Engstrom, P.G. *et al.* Complex loci in human and mouse genomes. *PLoS Genet.* **2**, 564–577 (2006).
9. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
10. Gardina, P.J. *et al.* Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* **7**, 325 (2006).
11. Clark, T.A., Sugnet, C.W. & Ares, M. Genome-wide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**, 907–910 (2002).

12. Reinartz, J. *et al.* Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief. Funct. Genomic. Proteomic.* **1**, 95–104 (2002).
13. Velculescu, V.E. *et al.* Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
14. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
15. Kim, J.B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
16. Bruce, S.J. *et al.* Dynamic transcription programs during ES cell differentiation towards mesoderm in serum versus serum-free (BMP4) culture. *BMC Genomics* **8**, 365 (2007).
17. Hirst, C.E. *et al.* Transcriptional profiling of mouse and human ES cells identifies SLAIN1, a novel stem cell gene. *Dev. Biol.* **293**, 90–103 (2006).
18. Faulkner, G.J. *et al.* A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**, 281–288 (2008).
19. Karolchik, D. *et al.* The UCSC genome browser database: 2008 update. *Nucleic Acids Res.* **36** (Suppl. 1), D773–779 (2008).
20. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **7** (Suppl. 1), S12 (2006).
21. Kiyosawa, H. *et al.* Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res.* **15**, 463–474 (2005).
22. Jeck, W.R. *et al.* Extending assembly of short DNA sequences to handle error. *Bioinformatics* **23**, 2942–2944 (2007).
23. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
24. Zhou, Q. *et al.* A gene regulatory network in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **104**, 16438–16443 (2007).
25. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).
26. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* published online, doi:10.1126/science.1158441 (1 May 2008).
27. Bruce, S.J. *et al.* In vitro differentiation of murine embryonic stem cells toward a renal lineage. *Differentiation* **75**, 337–349 (2007).
28. Matz, M. *et al.* Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.* **27**, 1558–1560 (1999).
29. Ohtake, H. *et al.* Determination of the capped site sequence of mRNA based on the detection of Cap-dependent nucleotide addition using an anchor ligation method. *DNA Res.* **11**, 305–309 (2004).
30. Schmidt, W.M. & Mueller, M.W. CapSelect: a highly sensitive method for 5′ CAP–dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res.* **27**, e31 (1999).
31. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
32. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).