LEXOGEN

Enabling complete transcriptome sequencing

Discovery and quantification of transcript variants with SQUARE™ mRNA-Seq

The majority of RNA-seq library preparation protocols quantify the expression of genes without taking into account the numerous transcript variants which contribute to this overall expression. SQUARE mRNA-seq uses a selective matrix to separate transcript variants and to enable the assembly and quantification of individual transcripts. The protocol is highly specific for full-length, polyadenylated mRNA, and transcript hypotheses are supported by directional transcription start and polyadenylation site tagging.

Introduction

The vast majority of genes are alternatively spliced and produce a variety of mature transcripts. These transcript variants often encode proteins with different structures and functions, and changes in the expression of variants from the same gene can lead to profound biological effects (reviewed in [1]). Various transcriptional events, including splicing from alternative 5′ or 3′ splice-sites, exon skipping, intron retention and the usage of different promoters or polyadenylation signals lead to a staggering number of potential transcript variants for each gene. Current next generation sequencing (NGS) technologies fail to adequately address this diversity; most RNA-seq experiments result in the cumulative quantification of all variants of a gene, without regard for the different structures and abundances of each transcript.

Lexogen's SQUARE technology enables complete transcriptome profiling by subdividing the transcriptome based on the selective amplification of full-length transcripts.

Starting with a small amount of total RNA, cDNA is generated by the reverse transcription of full-length mRNA (Fig. 1) with simultaneous tagging of transcription start sites (TSS) and end sites (TES, polyadenylation sites). Sub-populations of the total cDNA pool are then amplified using primers selective for the 5′- or 3′-most nucleotides of the transcript. As the terminal nucleotides of transcript variants are different, they are amplified in different reactions, segregating them from one another. Barcoded NGS libraries are prepared from the PCR products of each sub-population (SQUARE matrix field) individually using a proprietary protocol, and multiplexed libraries are sequenced on either the Illumina or SOLiD platform.

After demultiplexing, the reads of each individual matrix field can be used to generate transcript hypotheses and to analyze differential expression between biological samples. As each matrix field contains a single or a limited number of transcript variants, transcript models can be generated more accurately and for transcripts with low relative abundance. Additionally, the integrated, directional TSS and TES tagging provides valuable information regarding promoter usage and polyadenylation, which would require additional RACE or CAGE experiments with other mRNA-Seq protocols.

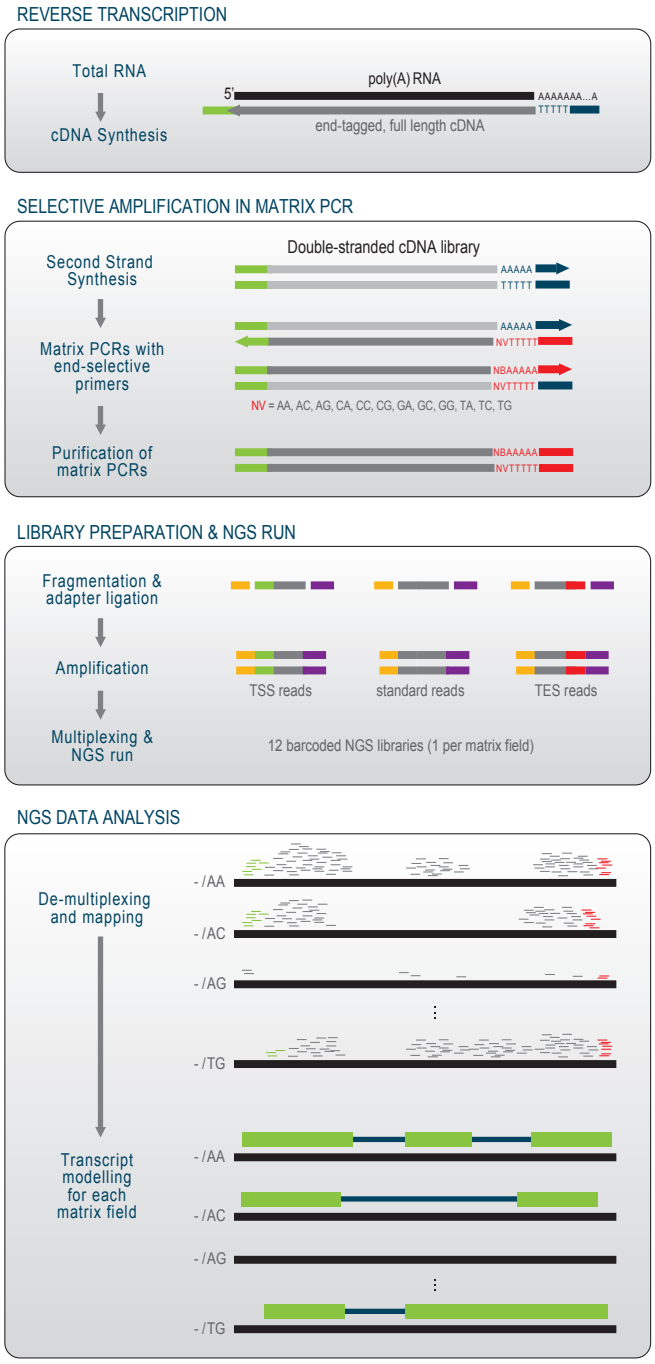

Figure 1 | Schematic of SQUARE workflow for a 3′-end selective 12-field matrix.

Table 1 | Summary of sequencing run. Reads in millions (M) or thousands (K).

	SQUARE			CONTROL		
	All Reads	TSS Reads	TES Reads	All Reads	TSS Reads	TES Reads
Demultiplexed	203.8 M (17.0 M / field)	10.1 M (842 K / field)	10.1 M (842 K / field)	30.2 M	809 K	1.3 M
Mapping	177.1 M (86.9%)	8.6 M (85.1%)	9.7 M (96.0%)	27.0 M (89.4%)	629 K (77.8%)	906 K (69.7%)
Uniquely Mapping	162.4 M (79.7%)	8.0 M (79.2%)	8.9 M (88.1%)	24.7 M (81.8%)	579 K (71.6%)	842 K (64.8%)

Experiment

We performed a 12-field 3' selective SQUARE experiment using 1200 ng total mouse liver RNA. With primers selective for the two 3'-terminal nucleotides, the entire transcriptome can be separated into a matrix with 12 fields denoted by their selective primers (-/AA, -/AC, etc.). Depending on the number of transcript variants expressed and their terminal sequences, some fields may contain multiple transcript variants. As the size of the SQUARE matrix is scalable, a higher number of fields can be employed to increase segregation power and separate a greater number of transcript variants. The 12-field 3' selective matrix was chosen here for ease of handling.

In addition to the 12 SQUARE libraries, we generated a control library with non-selective primers. Multiplexed libraries were pooled and sequenced on 7 lanes of an Illumina GAIIx flowcell with 100bp single-end reagents.

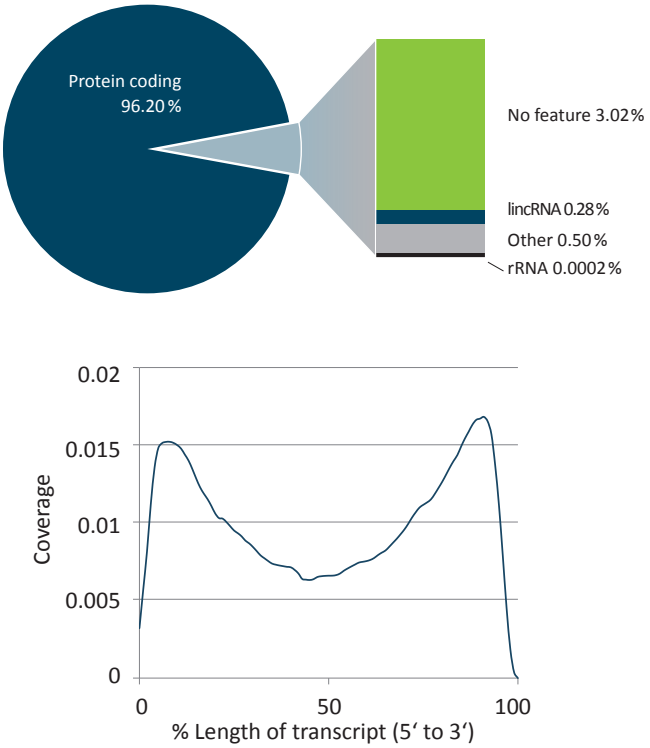

Figure 2 | SQUARE library preparation features. (A) Biotypes of detected genes. Ensembl annotations of uniquely mapped reads for all SQUARE libraries. A significant fraction map to regions without annotated features, while a small number of reads detect lincRNA (long non-coding RNA which are capped and poly-adenylated). (B) Coverage distribution. Reads cover the entire length of annotated genes with enhanced coverage at TSS and TES.

Pass-filter reads were demultiplexed, TSS and TES tag sequences were trimmed, and reads were aligned to the Ensembl65 mouse genome with Tophat[2] (Table 1). Separate alignments of TSS and TES reads permit the visualization of transcription initiation and polyadenylation sites using standard commercially available or open-source software.

Full-length, mRNA-specific library preparation

While starting with total RNA, the SQUARE workflow is highly specific for polyadenylated mRNA and does not require separate protocols for mRNA selection. The vast majority of reads (> 96%) map to annotated protein coding genes, and rRNA content is minimal (Fig. 2A). Transcript coverage is full-length and has a distinctive symmetric over-representation of 5' and 3' termini (Fig. 2B). This design feature increases TSS and TES coverage and facilitates the in-depth analysis of promoter usage and polyadenylation sites.

High matrix primer selectivity ensures efficient transcript variant segregation

The selective amplification of transcripts relied on by SQUARE requires a highly selective PCR system. Mis-hybridization during the SQUARE matrix PCR would result in high-abundant transcript variants being amplified across the matrix and not just in the correct field, obscuring variants with lower abundance.

The SQUARE PCR protocol was developed specifically for this environment and has exceptional selectivity in a complex reaction mixture with thousands of templates differing in concentration over several orders of magnitude. When TES reads were examined for evidence of mispriming events, over 95% of TES reads matched the reference genome, indicating a highly selective PCR system (Fig. 3).

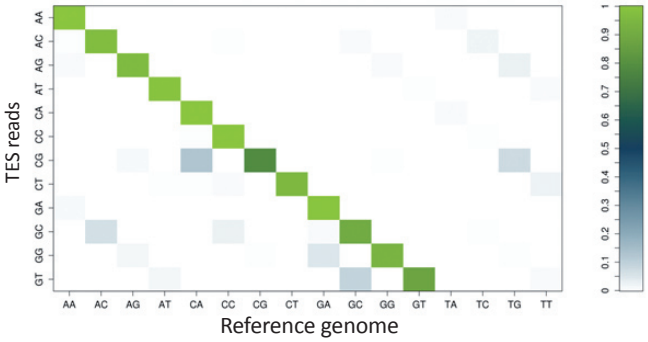

Figure 3 | Selectivity heat map. TES reads were mapped to the genome and mismatches between the expected genomic sequence and the selective nucleotides of the matrix primer used were calculated for all 12 fields. The high degree of clustering along the diagonal corresponds to correct matches and indicates that TES reads are generated almost exclusively in the expected fields.

Table 2 | SQUARE efficiently segregates transcript variants and genes.

	Total in Reference	Detected	Segregation		
			Absolute	Partial	None
Genes	37 532	17 051 (45 %)	2613 (15 %)	12 554 (74 %)	1884 (11 %)
Transcripts	95 084	40 112 (42 %)	12 119 (30 %)	26 367 (66 %)	1626 (4 %)
First Exons	127 788	17 064 (13 %)	6991 (41 %)	9818 (58 %)	254 (1 %)
TSS	N/A	186 786	148 264 (79 %)	38 267 (21 %)	255 (0 %)
TES	N/A	59 813	52 248 (87 %)	7 565 (13 %)	3 (0 %)
Exon-Exon Junctions	582 930	85 470 (15 %)	38 136 (45 %)	45 723 (53 %)	1 611 (2 %)

Reads derived from all twelve matrix PCRs were mapped separately to the mouse genome annotation, and the number of features detected in each was calculated.[3] Genes and transcripts were counted if the lower limit of the FPKM 95% confidence interval was greater than 0. TSS, TES and exon-exon junctions were counted if covered by 5 or more reads or junction-spanning reads. Features were counted as partly segregated if detected in 2 to 11 fields.

Between the individually mapped matrix fields, almost half of the annotated genes and transcripts were detected (Table 2). The vast majority of these were present in only a subset of matrix fields, and 15 % of the genes and 30 % of the transcripts were completely segregated and detected in only a single matrix field. Note that while in these experiments transcripts were segregating according to their 3' terminal nucleotides, start sites, first exons and splice junctions were also effectively segregated.

Examples of SQUARE data

To locate genes where SQUARE clearly segregates transcripts and detects novel variants we developed a pipeline for identifying genomic loci with segregated transcriptome features based on the pairwise Jensen-Shannon divergence between fields. Two examples identified when this method was applied to splice-junction reads are presented (Fig. 4), but this pipeline can also be applied to TES, TSS and coverage distributions.

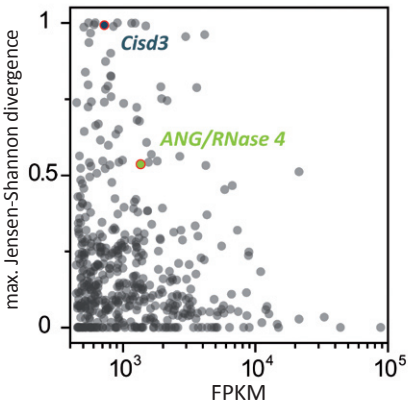

Figure 4 | Differential junction detection between matrix fields identifies transcript variant segregation. Exon-exon junction reads were tested for dissimilarities between all matrix field pairs. For each gene, the maximum Jenson-Shannon divergence was plotted against the FPKM. High divergence values indicate genes that express splice variants, which are segregated in the SQUARE matrix. Exemplary genes *Cisd3* and *ANG/RNase 4* are highlighted.

Identification of a novel variant by transcript segregation at the *Cisd3* locus

The *Cisd3* locus is annotated with three exons and a single promoter, encoding a single transcript consisting of all three exons. In the control sample, this transcript variant was assembled based on coverage distribution and junction-spanning reads (Fig. 5).
With SQUARE, an additional variant is revealed that is obscured in the control sample: By isolating the highly-expressed annotated variant in the -/AG field, a novel single-exon transcript becomes detectable in the -/AA field.

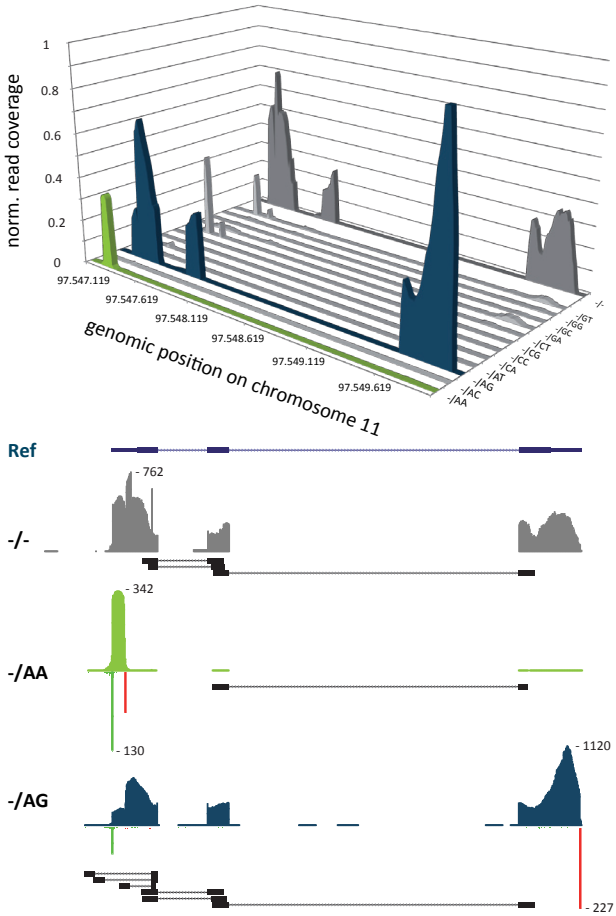

Figure 5 | Transcript segregation at the *Cisd3* locus. (**A**) Coverages at the *Cisd3* gene locus are shown for the non-segregated sample (-/-) and for all 12 SQUARE fields. SQUARE fields were down-sampled so that the total number of reads for each SQUARE field was 1/12th that of the control. Spikes in coverage visible in the -/GT and -/GA fields correspond to repeat elements and lack corresponding TSS and TES. (**B**) UCSC genome browser visualization of the control library (-/-) and the two matrix fields -/AA and -/AG that show transcript segregation. For SQUARE samples, the total coverage is shown with TSS (green) and TES (red) read coverages projected downwards to denote mapped transcript start and end-sites. The RefSeq annotated exon-intron structure is shown as reference (Ref).

Segregation of mRNAs coding for different proteins and discovery of multiple novel transcripts at the *ANG/RNase 5 - RNase 4* locus

This locus is annotated with four exons and two promoters. Depending on the promoter, transcription starts at exon 1 or exon 2, and the respective 5' UTR is spliced to either exon 3 or exon 4 which contain open reading frames encoding RNase 5/angiogenin and RNase 4, respectively (Fig. 6). In liver tissue, transcription occurs primarily from the downstream promoter P2.[4]

In the non-segregated control library, coverage is distributed across all four transcript variants. Transcription mainly starts at the P2 promoter, and exon 2 is then spliced to exon 3 or exon 4.

These two major transcript variants are segregated by SQUARE. While some fields, such as -/CT, contain reads arising from multiple variants, the P2-exon 3 variant (ANG) and the P2-exon 4 variant (RNase 4) are separated in the fields -/GG and -/GT, respectively. This enables field-wise FPKM calculations and differential expression analysis based on transcript variants and not on the less informative gene level.

In contrast to other fields, the RNase 4 variant observed in the -/CT library terminate predominantly at a novel TES in exon 4 upstream of the normal polyadenylation site, resulting in a shorter 3' UTR. Additionally, SQUARE segregation in field -/CT reveals a region immediately downstream of the annotated exon 2 with high read coverage, as well as a not yet annotated TES. Together with the TSS mapping to P2, the coverage and TES in -/CT provide solid evidence for a novel, short variant transcribed from the P2 promoter and terminating within the intron upstream of exon 3.

In field -/GT, transcript segregation reveals a novel exon between exons 2 (containing the P2 promoter) and 3. Junction-spanning reads, TES, and TSS support a novel transcript variant consisting of exon 2, the newly-discovered exon and exon 4.

In addition, novel junction reads spanning exon 3 and exon 4 suggest the existence of a variant transcribed from P2 that consists of both exons 3 and 4, potentially encoding an ANG/RNase 5 – RNase 4 fusion protein.

Conclusions

SQUARE technology accomplishes a transcriptome-wide segregation of transcript variants, allowing the detailed analysis of transcripts from the same genomic locus in a scalable number of sub-pools. Combined with directional TSS and TES tagging, this segregation enables the detection and assembly of novel transcripts normally obscured by the major variant. The segregation of transcript variants allows differential expression analysis to be carried out on individual transcripts and not simply on a gene-by-gene basis, providing a deeper understanding of the transcriptome.

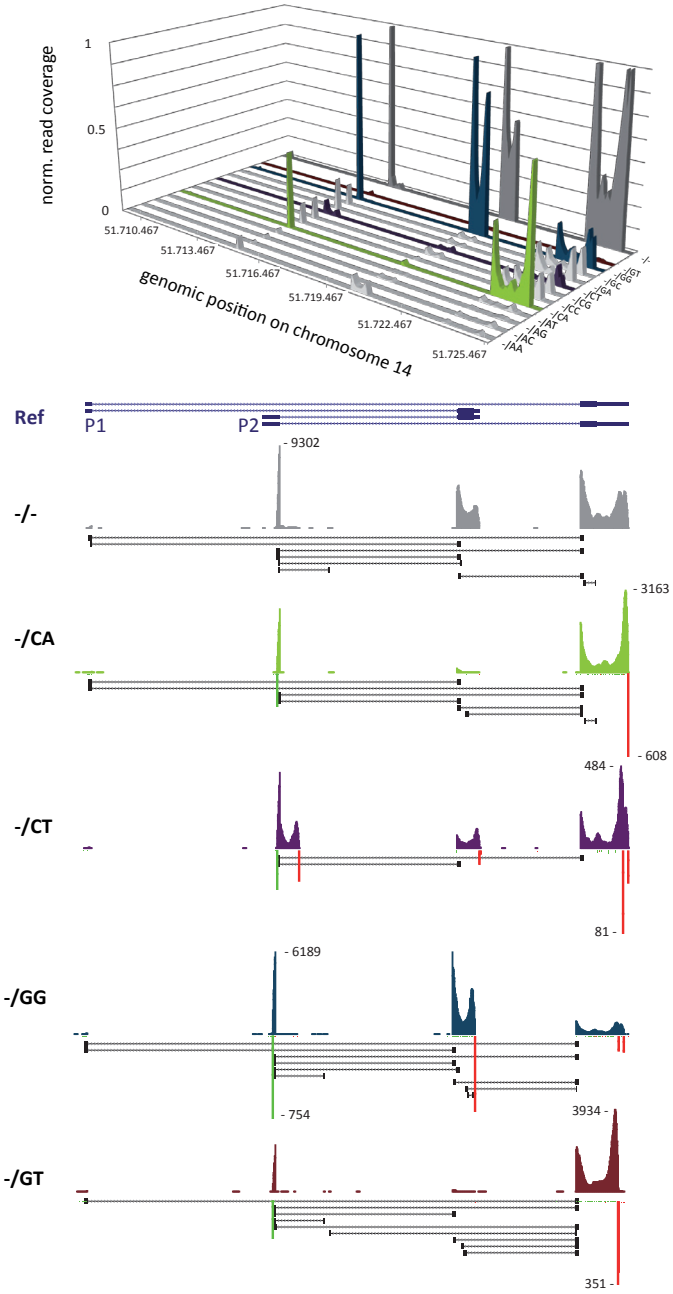

Figure 6 | Transcript segregation at the *ANG/RNase5* & *RNase4 locus*. (**A**) Coverages are shown for the non-segregated sample (-/-) and for all 12 SQUARE samples. SQUARE fields were down-sampled, so that the total number of reads for each SQUARE field was 1/12th that of the control. (**B**) UCSC genome browser visualization of non-segregated sample (-/-) and matrix fields of interest. The RefSeq database (Ref) contains 4 transcript variants. P1 and P2 refer to the two promoters located in exon 1 and exon 2, respectively; exon 3 contains the CDS of ANG/RNase 5, exon 4 the CDS of RNase 4.

[1] Faustino, N. A. & Cooper, T. A. (2003). Pre-mRNA splicing and human disease. Genes Dev. 17, 419-437
[2] Trapnell, C. et al. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111
[3] Trapnell C. et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28: 511-515
[4] Dyer, K. D. & Rosenberg, H. F. (2005). The mouse RNase 4 and RNase 5/ang 1 locus utilizes dual promoters for tissue-specific expression. Nucleic Acids Res. 33, 1077-1086

For more information visit www.lexogen.com.
Correspondence should be addressed to info@lexogen.com.