Spliced synthetic genes as internal controls in RNA sequencing experiments

Simon A Hardwick[1,2,5], Wendy Y Chen[1,2,5], Ted Wong[1], Ira W Deveson[1,3], James Blackburn[1,2], Stacey B Andersen[4], Lars K Nielsen[4], John S Mattick[1,2] & Tim R Mercer[1,2]

RNA sequencing (RNA-seq) can be used to assemble spliced isoforms, quantify expressed genes and provide a global profile of the transcriptome. However, the size and diversity of the transcriptome, the wide dynamic range in gene expression and inherent technical biases confound RNA-seq analysis. We have developed a set of spike-in RNA standards, termed 'sequins' (sequencing spike-ins), that represent full-length spliced mRNA isoforms. Sequins have an entirely artificial sequence with no homology to natural reference genomes, but they align to gene loci encoded on an artificial *in silico* chromosome. The combination of multiple sequins across a range of concentrations emulates alternative splicing and differential gene expression, and it provides scaling factors for normalization between samples. We demonstrate the use of sequins in RNA-seq experiments to measure sample-specific biases and determine the limits of reliable transcript assembly and quantification in accompanying human RNA samples. In addition, we have designed a complementary set of sequins that represent fusion genes arising from rearrangements of the *in silico* chromosome to aid in cancer diagnosis. RNA sequins provide a qualitative and quantitative reference with which to navigate the complexity of the human transcriptome.

The human genome is pervasively transcribed into a wide array of coding and noncoding RNAs that are spliced into complex and overlapping isoforms[1,2]. This full set of expressed RNAs is collectively termed the transcriptome and conveys the genetic information required to define the cellular and developmental phenotype[3].

RNA-seq provides a global profile of the transcriptome, including high-throughput determination of transcript sequences, assembly of novel isoform structures and a quantitative measure of gene expression[4–7]. Given these advantages, RNA-seq has become a central tool with wide-ranging research and clinical applications. However, accurate resolution of gene expression is confounded by the sheer size and complexity of the transcriptome[8–10], and it is further exacerbated by technical variables during library preparation, sequencing and bioinformatic analysis[11–14].

Spike-in controls are exogenous RNA molecules that are directly added to an RNA sample of interest, allowing an objective assessment of various internal biases and accurate comparison between experimental conditions[15]. Previously, the External RNA Controls Consortium (ERCC) developed a set of spike-in standards for microarray analysis that has subsequently been co-opted for use in RNA-seq experiments[16–18]. Whilst the ERCC spike-ins can be used to assess the quantitative accuracy of RNA-seq, as a restricted set of single-exon transcripts they do not represent the complexity of eukaryotic gene expression and splicing.

We have developed a suite of synthetic RNA isoforms, termed sequins, which align to artificial gene loci encoded within an accompanying *in silico* chromosome (**Fig. 1**). The primary sequence of the *in silico* chromosome (and representative sequins) shares no homology with the genomes of known natural organisms and can be coindexed for read alignment without risk of cross-alignments contaminating analysis.

Sequins constitute internal controls that can evaluate almost all stages of the RNA-seq workflow[19], including library preparation, sequencing, split-read alignment, transcript assembly, gene expression and alternative splicing (**Supplementary Fig. 1**). Sequins are ideal for assessing downstream bioinformatic steps and optimizing parameter choice, and they act as normalization factors with which to compare multiple samples. Here we describe the design, validation, use and advantages of sequins within an RNA-seq experiment.

RESULTS

Design of the *in silico* chromosome and RNA sequins

We first designed an ~11-Mb artificial *in silico* chromosome (*chrIS_R*) sequence by proportionately sampling regions from the human genome (hg38) according to size, GC content and repeat density (see Online Methods). To remove homology to natural sequences, we performed sequence inversion and, where necessary, local shuffling to retain original nucleotide composition, repetitiveness and uniqueness in the final artificial sequence (**Supplementary Fig. 2**).

[1]Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [2]St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia. [3]School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, New South Wales, Australia. [4]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, Queensland, Australia. [5]These authors contributed equally to this work. Correspondence should be addressed to T.R.M. (t.mercer@garvan.org.au).

RECEIVED 23 MARCH; ACCEPTED 29 JUNE; PUBLISHED ONLINE 8 AUGUST 2016; DOI:10.1038/NMETH.3958

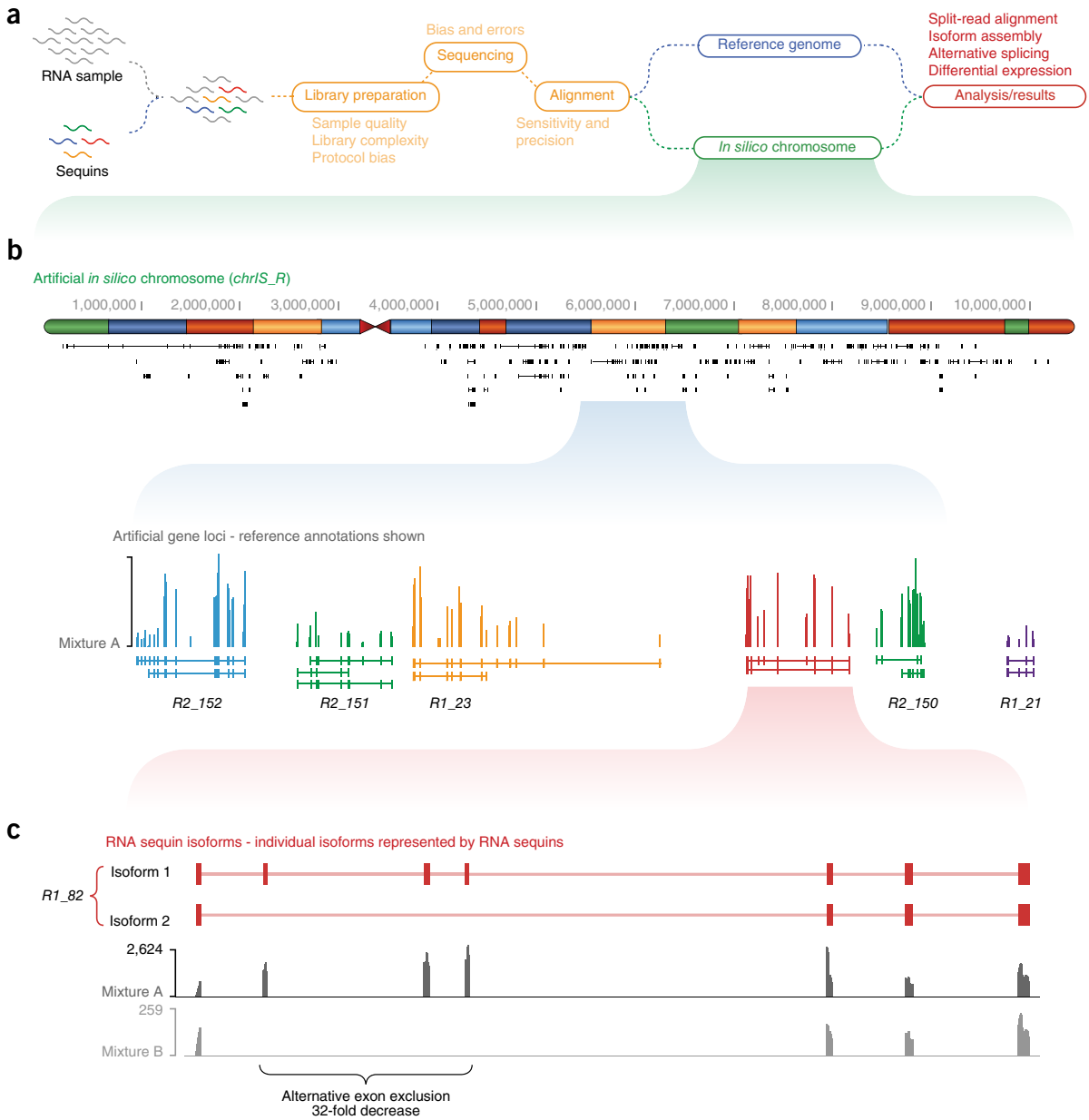

Figure 1 | Schematic overview illustrating the design and use of RNA sequins. (**a**) Schematic workflow using RNA sequins and *chrIS_R*. (**b**) 78 artificial gene loci are encoded within *chrIS_R* (top). Most genes encode multiple mRNA isoforms, each of which are represented as an RNA sequin standard. (**c**) Inset shows example of *R1_82* gene, with two alternative isoforms generated through alternative splicing. By varying the concentration of the two isoforms between two staggered mixtures (A and B), we can emulate the alternative splicing of three internal exons (RNA-seq coverage indicated by gray histograms).

To design artificial gene loci within *chrIS_R*, we representatively sampled endogenous human genes (using GENCODE annotations[20]) according to spliced transcript size, exon count and GC content (**Supplementary Fig. 3**). In total, we designed 78 artificial gene loci encoding 164 alternative isoforms comprising 869 unique exons and 754 unique introns. Genes ranged from single-exon to large multi-exon loci, with individual isoforms ranging in size from ~280 bp up to ~7 kb and comprising up to 36 exons.

The complex exon and intron architecture of spliced endogenous human genes and their relative organization into overlapping, antisense and bidirectional loci were maintained across *chrIS_R* (**Fig. 1**). Introns were populated with donor and acceptor splice sites, and because each RNA sequin represents a mature mRNA molecule (with exons spliced together and intervening introns excluded), the combination of multiple isoforms from a single gene locus can be used to emulate alternative splicing (**Fig. 1**).

To initially assess sequin alignment and assembly without biases that may be introduced during library preparation and sequencing, we simulated read libraries from RNA sequins and a matched set of endogenous protein-coding human genes (see Online Methods). The overall alignment rate of simulated reads was almost identical for sequins (99.5%) and endogenous genes

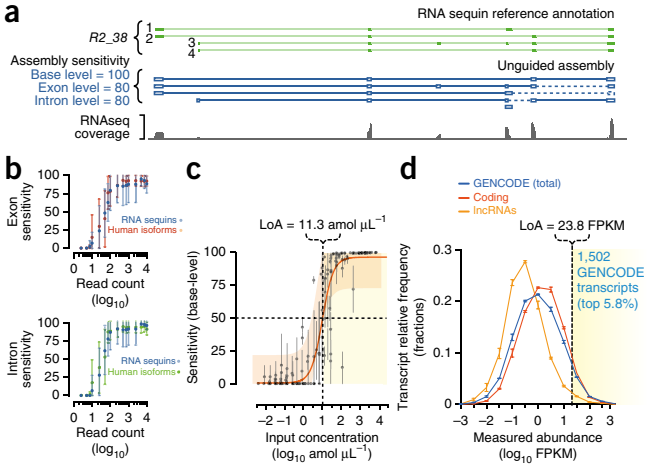

Figure 2 | Using RNA sequins to assess transcript assembly. (**a**) Assembly with Tophat2 and StringTie, unguided by reference annotations. Dashed lines indicate missing features. (**b**) Scatter plots indicate the assembly of RNA sequins and matched human isoforms with experimental reads at a range of sequence coverage levels. Assembly is measured according to the percentage of exons (top) or introns (bottom) that are assembled correctly. (**c**) Scatter plot illustrates relationship between sensitivity of assembly (base level) relative to the input concentration of RNA sequin isoforms. A sigmoidal dose–response curve is fitted to the data (R^2 = 0.858) to estimate concentration required to confidently assemble 50% of a given isoform (i.e., base-level sensitivity = 50). Orange shading indicates 90% prediction intervals; yellow shading indicates fraction sufficiently expressed above 'limit of assembly' (LoA). (**d**) Density histogram illustrates the relative fraction of endogenous human transcripts at different expression levels in the accompanying K562 sample, with the fraction sufficiently expressed above LoA indicated (shaded yellow). Frequency of protein-coding isoforms (red) and lncRNAs (orange) are indicated, as is the total of GENCODE transcripts (blue). n = 3 biological replicates, error bars indicate s.d.

(99.4%). We observed neither reads from sequins aligning to hg38 (or any other reference genome tested) nor reads from endogenous genes aligning to *chrIS_R* (**Supplementary Fig. 4a**,**c**), allowing the use of sequins in a wide range of model organisms. Synthetic and endogenous human isoforms also exhibited similar assembly profiles across a range of simulated sequence coverage levels (**Supplementary Fig. 5**).

Generation and validation of synthetic RNA sequins

RNA sequins, including terminal poly(A) tails, were synthesized and cloned into expression vectors for production by *in vitro* transcription (see Online Methods). Resultant full-length RNA molecules were purified and diluted to constitute stock inventory from which subsequent mixtures were prepared by automated liquid handling.

We initially sequenced a neat mixture of RNA sequins (i.e., without natural RNA sample) at equimolar concentrations to ensure sufficient coverage for validation of sequin designs with experimental reads. Considering the potential for experimental reads to cross-align to the reference human genome, we found that <0.0075% of reads (3,223 reads from ~43 million, which are obtained from soft clipping of poly(A) tails) aligned to both *chrIS_R* and hg38 (**Supplementary Fig. 4**). Similar results were observed for other model organism reference genomes tested and, conversely, a negligible fraction (~0.03%) of experimental reads from a large K562 RNA-seq library were found to align to *chrIS_R* (**Supplementary Fig. 4d**).

The alignment and assembly of RNA sequins from experimental reads were compared with a selection of human protein-coding genes with matched length and exon count expressed in the accompanying K562 library (see below). RNA sequins and endogenous genes exhibited comparable assembly across a range of matched sequencing depths (**Fig. 2** and **Supplementary Fig. 6**). Notably, even at full coverage we did not observe complete assembly of our synthetic transcriptome using TopHat2 (ref. 21) and StringTie[22] with default parameters (**Supplementary Fig. 7**). Both true- and false-positive assembly events accumulated with increasing sequencing depth, with an inverse relationship between sensitivity (Sn) and specificity (Sp) of assembly[23] (**Supplementary Fig. 7b**).

Combining RNA sequins at staggered concentrations establishes a reference ladder against which to measure gene expression. We combined sequin genes at a two-fold serial dilution, with a minimum three genes per dilution, to span an ~10^6-fold range in concentration that sufficiently represented the dynamic range of gene expression observed across the human transcriptome[24] (**Supplementary Fig. 8a**,**b**). By further combining alternative RNA sequin isoforms at different relative concentrations, we established a 32-fold ladder with which to measure alternative splicing of gene loci (**Supplementary Fig. 8a**,**b**). Using this approach, we represented a diversity of alternative splicing events (**Supplementary Fig. 9**).

Using RNA sequins to assess isoform assembly

To demonstrate the use of RNA sequins, we spiked the staggered mixture into 1 µg of total RNA harvested from K562 cells (n = 3 replicates). To enable their detailed validation, sequins were added at a high fractional concentration (10%); however, we typically recommend adding a lower fractional concentration (~1–3%) to minimize sequencing depth used. This combined sample underwent concurrent library preparation and sequencing, with reads aligned to a combined genome comprising both hg38 and *chrIS_R* (see Online Methods). Alignments to *chrIS_R* were subsampled to calibrate variable spike-in amounts between samples or replicates as required.

Comparison of reads from RNA sequins (aligning to *chrIS_R*) with reads from the accompanying K562 RNA sample (aligning to hg38) showed an indistinguishable error profile on a per-nucleotide and per-read basis, including the same distribution of Phred quality scores across the length of sequenced reads (**Supplementary Fig. 10a–d**). Additionally, endogenous genes and RNA sequins exhibited similar normalized coverage across the length of annotated transcripts (**Supplementary Fig. 10e**,**f**).

Sufficient sequence coverage is required for robust transcript assembly, with novel weakly expressed genes such as long noncoding RNAs (lncRNAs) often poorly assembled[25,26]. RNA sequins can be used to empirically determine the minimum transcript expression required for assembly of spliced isoforms. Plotting the fraction of isoforms assembled relative to their input concentration generated a sigmoidal dose–response curve (R^2 = 0.893; **Fig. 2c**) that enabled us to interpolate the minimum concentration required to achieve a threshold level of assembly. For example, a minimum concentration of 11.3 amol µL^{-1} (or 23.8 fragments per kilobase of exon model per million mapped reads (FPKM); **Fig. 2c**) was required to assemble 50% of the reference annotation

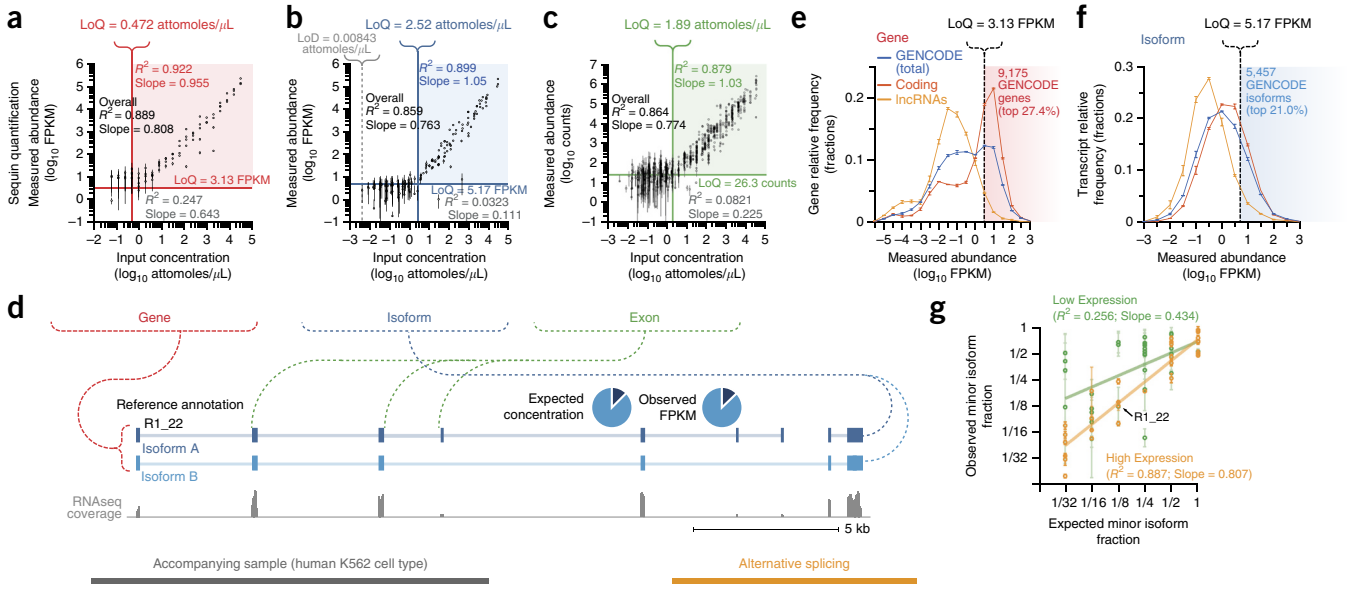

Figure 3 | Expression and alternative splicing of RNA sequins. (**a**–**c**) Scatter plots illustrate correlation between RNA sequin input concentration and observed expression at the level of genes (**a**), isoforms (**b**) and exons (**c**). Also indicated are the 'limit of detection' (LoD, gray dotted line), 'limit of quantification' (LoQ, colored lines) and the slope and correlation above the LoQ. (**d**) Genome browser view shows *R1_22* sequin reference annotations with experimental RNA-seq coverage (indicated in gray histogram below) and observed and expected major (light blue) and minor (dark blue) isoform expression. (**e**,**f**) Density histograms indicate the frequency of endogenous human genes (**e**) and isoforms (**f**) relative to expression level in the accompanying K562 transcriptome. LoQ threshold indicates the fraction of genes and isoforms sufficiently abundant to enable an accurate measurement of expression. Genes and isoforms are further divided into protein-coding (red) and lncRNA (orange) biotypes. (**g**) Scatter plot indicates correlation between observed and expected minimum isoform as a fraction of the major isoform for each sequin gene. Genes are divided into highly (yellow; $R^2 = 0.887$) and weakly (green; $R^2 = 0.256$) expressed subsets to indicate the expression dependency of isoform resolution. $n = 3$ biological replicates, error bars indicate s.d.

(i.e., base-level Sn = 50). By comparison, we found that only 1,502 transcripts in the accompanying K562 transcriptome (comprising the top 5.8% of the 25,975 GENCODE transcripts detected) were expressed above this threshold, with lncRNAs exhibiting poorer assembly due to their lower expression (**Fig. 2d**). This indicated that the majority of novel transcripts in the K562 transcriptome will remain unassembled in a standard RNA-seq library.

Using RNA sequins to assess gene expression and splicing

RNA-seq has emerged as the primary tool in gene expression profiling. To assess the quantitative accuracy of RNA-seq, we measured the expression of sequins and endogenous human genes and isoforms using StringTie (with reference annotations supplied; see Online Methods). The measured abundance of sequins was plotted against input concentration, revealing a strong linear trend ($R^2 = 0.922$ for genes, $R^2 = 0.899$ for isoforms; **Fig. 3a**,**b**) to a lower inflection point that was determined using piecewise linear regression analysis to occur at 0.472 amol μL^{-1} for gene expression and 2.52 amol μL^{-1} for isoform expression. This inflection point constitutes the 'limit of quantification' (LoQ; ref. 27), below which the measurement of sequin expression becomes nonlinear and highly variable ($R^2 = 0.247$ for genes, $R^2 = 0.0323$ for isoforms; **Fig. 3a**,**b**).

The slope of this linear trend (0.955 for genes, 1.05 for isoforms) provided a rate with which to convert between sequin input concentration (in amol μL^{-1}) and measured abundance (in FPKM) and enabled an assessment of genes expressed in the accompanying K562 transcriptome with sequins (**Fig. 3**). This conversion indicated the LoQ to be 3.13 FPKM for genes and 5.17 FPKM for isoforms. Only 9,175 genes (comprising the top 27.4% of the 33,484 GENCODE genes detected) and 5,457 isoforms (top 21.0% of the 25,975 GENCODE isoforms detected) in the accompanying K562 transcriptome exceeded this threshold (**Fig. 3e**,**f**) and were therefore sufficiently sampled for accurate quantification within the library. Again, a smaller fraction of lncRNAs exceeded the LoQ threshold (only 6.50% and 4.73% of all lncRNA transcripts and genes detected in the K562 sample, respectively), illustrating limits on accurate gene profiling in the accompanying K562 sample.

We also considered the relative quantification of alternatively spliced isoforms by measuring minor isoform abundance as a fraction of major isoform abundance for each sequin gene (**Fig. 3g**). Whilst the overall accuracy of relative isoform quantification was poor ($R^2 = 0.509$), it was substantially better for high-abundance genes (above 30.2 amol μL^{-1}; $R^2 = 0.887$, slope = 0.807) than for low-abundance genes ($R^2 = 0.256$, slope = 0.434). At low abundances, isoform resolution (whereby isoforms are assembled and correctly quantitatively ordered) became unachievable, with no genes below 0.472 amol μL^{-1} being accurately resolved. Additional splicing complexity further exacerbated accurate quantification, so that no sequin genes with four isoforms ($n = 3$) and only one with three isoforms ($n = 12$) were correctly resolved. To further assess alternative splicing, sequin annotations were collapsed into 883 constituent exon counting bins[28] (**Supplementary Fig. 11b**). We plotted observed against expected percent spliced in (PSI; ref. 29) for each sequin exon bin (**Supplementary Fig. 11a**), again finding that the correlation was relatively weak overall ($R^2 = 0.603$) but substantially stronger for high- ($R^2 = 0.909$) than for low-abundance genes ($R^2 = 0.403$).

Finally, we investigated systematic and internal biases in the assembly and quantification of exons. Individual exon abundance was plotted against input concentration, and we observed

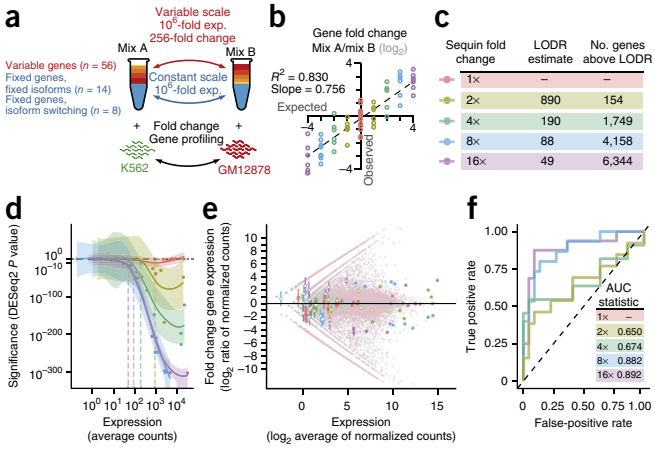

Figure 4 | Differential gene expression of sequin mixtures between samples. (**a**) Each staggered mixture of sequins comprises a variable subset (yellow, orange and red) and a constant subset (blue). Exp., expression. (**b**) Scatter plot illustrates correlation between observed and expected $\log_2$-fold change (LFC) for RNA sequin loci between mixtures A and B. (**c**) Table lists limit of detection of ratio (LODR) estimate for each sequin gene fold change (indicated by differing colors) and number of GENCODE genes in the accompanying K562 and GM12878 transcriptomes that are expressed above limit. (**d**) LODR curves illustrate the relationship between mean expression, fold change and power to detect differential gene expression. *P* value (estimated by DESeq2) is plotted relative to mean expression for each sequin gene. The false-discovery threshold is indicated (black dashed line) along with LODR estimates for each fold change (vertical colored lines). (**e**) MA plot shows relationship between observed LFC and mean expression. Sequins (colored dots) are overlaid on endogenous human genes (differential expression between K562 and GM12878 cell-types; red dots indicate significance). Empty and filled circles indicate sequins detected below and above the relevant LODR estimate, respectively. Error bars indicate s.d. (**f**) ROC (receiver operating characteristic) curves for sequin genes, showing diagnostic performance in measuring fold change.

a LoQ (at 1.89 amol μL^{-1}, or 26.3 read counts per exon) below which exon expression was not accurately measured (**Fig. 3c**). We observed superior assembly and more accurate quantitative measurement of internal exons (concentration required to achieve base-level Sn of 50 = 6.72 amol μL^{-1}; $R^2 = 0.905$) than terminal (i.e., first and last) exons (15.1 amol μL^{-1}; $R^2 = 0.825$) (**Supplementary Fig. 11c**,**d**). This artifact likely resulted from the nonuniform sequence coverage of transcript termini as a result of edge effects[13] and indicates the difficulty of achieving complete isoform structures and resolving alternative transcription initiation and termination events using RNA-seq.

Using RNA sequins to assess differential expression between samples

The formulation of alternative RNA sequin mixtures provides constant scaling factors for intersample normalization[30] and variable ladders to measure differences in gene expression and splicing between samples (**Fig. 4a**). We prepared a new mixture (mixture B) in which the concentration of 56 genes and 126 isoforms is varied to emulate differential gene expression and splicing in comparison with the previous mixture (mixture A). RNA sequin fold changes between mixtures A and B are described by 9 subgroups at the gene level and 19 subgroups at the isoform level (**Supplementary Fig. 8c**).

To demonstrate the use of RNA sequins in gene profiling, we spiked mixture B into 1 µg of total RNA harvested from GM12878 cells ($n = 3$ replicates). For a quantitative comparison to qRT-PCR, we selected a subset of sequins that represented a range of different expected fold changes across mixtures, finding that RNA-seq and qRT-PCR assays yielded a similar correlation between expected and observed fold change ($R^2 = 0.988$ for RNA-seq versus 0.929 for qRT-PCR; **Supplementary Fig. 12**).

We employed DESeq2 (ref. 31) to measure fold change in RNA sequins between mixtures, finding a robust linear relationship ($R^2 = 0.830$, slope = 0.756; **Fig. 4b**) between observed and expected fold change. DESeq2 correctly called all negative-control genes ($n = 22$; adjusted *P* values > 0.1), with the exception of two false positives that underwent extreme isoform switching in circumstances where the alternative isoforms have substantially different lengths despite constant overall gene expression. This supported similar findings from simulated data sets[14,32]. DESeq2 missed 20 (of 56) genes whose expression differed between mixtures, all of which had very low input concentrations, demonstrating the challenge in detecting expression changes of low-abundance transcripts.

We plotted limit of detection of ratio (LODR) curves[16] for five sequin gene subsets to define the fold change and mean expression limits at which differential gene expression can be measured with confidence (see Online Methods; **Fig. 4c**,**d**). When applied to differences in gene expression between the accompanying K562 and GM12878 cell lines, this indicated the fraction of endogenous human genes with sufficient expression to surpass these limits (**Fig. 4e**). As expected, there was a concomitant increase in diagnostic performance with increasing fold change, as demonstrated by receiver operating characteristic (ROC) analysis (**Fig. 4f**).

To assess dynamic alternative splicing across the two mixtures, we plotted fold change between mixtures at both the isoform (**Supplementary Fig. 13a**) and exon (**Supplementary Fig. 13b**) levels. We used DEXSeq (ref. 28) to assess differential exon usage; of the 883 exon bins, 212 were expected to have constant expression (true negatives), while the remaining 671 were expected to vary across a 512-fold scale (true positives). DEXSeq correctly identified 164 true negatives (77.4% of total) but only 305 true positives (45.5% of total). This limited ability to detect differential exon usage at significance was largely due to insufficient read counts for low-abundance isoforms that were exacerbated for terminal exons, and it illustrates the current limitations in measuring dynamic alternative splicing between samples.

Representing fusion genes with sequins

Chromosomal rearrangements that juxtapose two previously separated loci to form a fusion gene are a major cause of cancer[33,34]. RNA-seq can be used to diagnose a range of fusion genes, including the detection of novel or complex rearrangements. However, accurate diagnosis requires sufficient alignments across the intron junction traversing the chromosomal breakpoint, which can be further complicated by the presence of homologous or repetitive DNA sequences[35].

We modified our RNA sequins to represent a set of 24 fusion genes that resulted from 12 balanced inversions of *chrIS_R* (**Fig. 5**). Fusion sequins were designed to encompass a range of different genes with varying size, exon number and splicing complexity, and are accompanied by their respective unmodified parent gene versions. To validate fusion sequin designs, we showed that

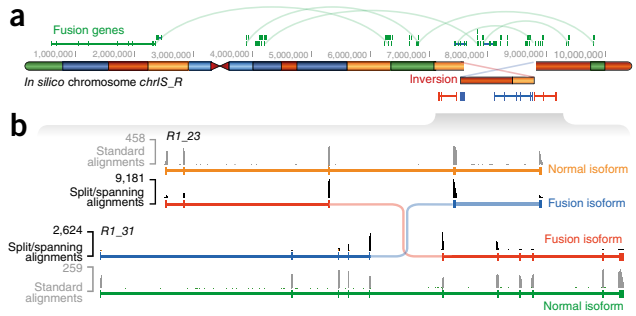

Figure 5 | Overview of fusion RNA sequins. (**a**) Synthetic rearrangements of *chrIS_R* can generate fusion genes that are represented by RNA sequins. (**b**) Example of heterozygous rearrangement resulting in two reciprocal fusion genes generated from the *R1_23* and *R1_31* loci. RNA-seq coverage shows normal alignments (gray histogram) and split–spanning alignments (black histogram) that traverse the fusion junction.

sequins exhibit comparable detection sensitivity to a set of endogenous oncogenic human fusions[36] across a range of simulated sequencing depths (**Supplementary Fig. 14a**).

Fusion sequins were manufactured by *in vitro* transcription and initially sequenced as a neat equimolar mixture, as described above. STAR-Fusion[37] identified all fusion events with sufficient coverage. By preparing sequins as a staggered mixture, we could assess the quantitative accuracy of fusion detection, and we found that split reads (reads spanning the fusion breakpoint) correlated much more closely with input concentration ($R^2 = 0.929$) than did spanning pairs (read pairs mapping to different genes; $R^2 = 0.479$; **Supplementary Fig. 14b**).

Fusion genes are often expressed at subclonal frequencies that can have prognostic value[34]. Accordingly, we combined fusion sequin isoforms at a 64-fold range in abundance relative to their normal gene counterparts and combined corresponding fusion and normal pairs to encompass a dynamic 128-fold range in expression (**Supplementary Fig. 14c**). This mixture was then pooled with the remaining unaffected RNA sequins and spiked at high fractional concentration into K562 total RNA before library preparation and sequencing (see Online Methods). Comparing observed to expected abundance for fusion sequins showed a strong linear relationship ($R^2 = 0.907$) and indicated the limit of sensitivity (7.56 amol μL^{-1}) with which fusion genes could be detected in the accompanying K562 library (**Supplementary Fig. 14d**). We also observed a strong linear relationship in fold change between fusion and normal parent gene expression ($R^2 = 0.853$; **Supplementary Fig. 14e**). Whilst we detected 20 of 24 fusion genes on *chrIS_R* (Sn = 83.3), STAR-Fusion also identified 21 false-positive fusion events (Sp = 48.8; **Supplementary Fig. 14f**). By requiring all fusion junctions to be supported by >1 split read and >1 spanning pair, and by requiring both breakpoints to be aligned with a reference exon splice site, we filtered the 942 putative K562 fusions identified down to just 12 remaining candidates, 7 of which had gene partners with shared homologous or repetitive sequence[35], and 2 of which were the previously validated *BCR–ABL1* (ref. 38) and *NUP214–XKR3* (ref. 39) fusions.

DISCUSSION

RNA sequins constitute a novel analytical tool for understanding the biases and limitations of RNA-seq, improving our understanding of the transcriptome and forming a standardized reference against which library preparation methods[40], sequencing platforms[11,41] and bioinformatic software[12,26,42,43] can be benchmarked and compared. To facilitate the use of RNA sequins for research purposes, aliquots of RNA sequins mixtures can be requested from http://www.sequin.xyz, along with the accompanying *in silico* chromosome sequence, associated gene annotations and supporting software.

Understanding the biases and limitations of RNA-seq is a critical prerequisite to its adoption as a diagnostic tool in a clinical context[44,45]. The routine use of internal controls will be expected not only to improve the accuracy and reliability of RNA-seq-based diagnostic assays but also to provide internal reference scales against which to anchor diagnostic guidelines and normalize large patient sample numbers.

Within this study, we have designed sequins to emulate the human transcriptome. However, sequins can be designed to investigate virtually any desired feature of genome biology. In an accompanying manuscript, Deveson *et al.*[46] develop a set of DNA sequins that represent genetic variation for use in whole-genome or exome-sequencing experiments. The synthetic representation of the genome and transcriptome with sequins provides researchers with a set of faithful internal controls that can be readily applied to improve the performance and extend the application of almost any next-generation sequencing assay.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing libraries have also been deposited to GEO with the accession code GSE77072.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to thank the following funding sources: Australian National Health and Medical Research Council (NHMRC) Australia Fellowship (1062470 to T.R.M. and 1062606 to W.Y.C.). S.A.H. and I.W.D. are supported by Australian Postgraduate Award scholarships. The contents of the published material are solely the responsibility of the administering institution, a participating institution or individual authors and do not reflect the views of NHMRC. The authors would also like to thank D. Thomson and M. Smith (Garvan Institute of Medical Research) for helpful discussions during manuscript preparation.

AUTHOR CONTRIBUTIONS

T.R.M. and J.S.M. conceived the project, designed sequins and *in silico* chromosome, and conceived experiments. W.Y.C. and S.B.A. performed experimental work. J.B. performed qRT-PCR validation. L.K.N. contributed supervision and manuscript preparation. S.A.H., T.W. and T.R.M. performed bioinformatic analyses. S.A.H., I.W.D. and T.R.M. prepared the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
2. Kapranov, P., Willingham, A.T. & Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* **8**, 413–423 (2007).

3. Kratz, A. & Carninci, P. The devil in the details of RNA-seq. *Nat. Biotechnol.* **32**, 882–884 (2014).
4. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
5. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
6. Wilhelm, B.T. & Landry, J.-R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249–257 (2009).
7. Martin, J.A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671–682 (2011).
8. Mercer, T.R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9**, 989–1009 (2014).
9. Vijay, N., Poelstra, J.W., Künstner, A. & Wolf, J.B.W. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. *Mol. Ecol.* **22**, 620–634 (2013).
10. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
11. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32**, 915–925 (2014).
12. Li, S. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32**, 888–895 (2014).
13. Lahens, N.F. *et al.* IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15**, R86 (2014).
14. Rehrauer, H., Opitz, L., Tan, G., Sieverling, L. & Schlapbach, R. Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching. *BMC Bioinformatics* **14**, 370 (2013).
15. Chen, K. *et al.* The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Mol. Cell Biol.* **36**, 662–667 (2015).
16. Munro, S.A. *et al.* Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat. Commun.* **5**, 5125 (2014).
17. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
18. Baker, S.C. *et al.* The External RNA Controls Consortium: a progress report. *Nat. Methods* **2**, 731–734 (2005).
19. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
20. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
21. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
22. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
23. Burset, M. & Guigó, R. Evaluation of gene structure prediction programs. *Genomics* **34**, 353–367 (1996).
24. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
25. Clark, M.B. *et al.* Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat. Methods* **12**, 339–342 (2015).
26. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
27. Armbruster, D.A. & Pry, T. Limit of blank, limit of detection and limit of quantitation. *Clin. Biochem. Rev.* **29**, S49–S52 (2008).
28. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
29. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
30. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
31. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
32. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
33. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).
34. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
35. Stransky, N., Cerami, E., Schalm, S., Kim, J.L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat. Commun.* **5**, 4846 (2014).
36. Tembe, W.D. *et al.* Open-access synthetic spike-in mRNA-seq data for cancer gene fusions. *BMC Genomics* **15**, 824 (2014).
37. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
38. Naumann, S., Reutzel, D., Speicher, M. & Decker, H.-J. Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leuk. Res.* **25**, 313–322 (2001).
39. Maher, C.A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 12353–12358 (2009).
40. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, 419 (2014).
41. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
42. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95 (2013).
43. Engström, P.G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).
44. Van Keuren-Jensen, K., Keats, J.J. & Craig, D.W. Bringing RNA-seq closer to the clinic. *Nat. Biotechnol.* **32**, 884–885 (2014).
45. Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D. & Craig, D.W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
46. Deveson, I.W. *et al.* Representing genetic variation with synthetic DNA standards. *Nat. Methods* http://dx.doi.org/10.1038/nmeth.3957 (2016).

ONLINE METHODS

Cell lines. K562 and GM12878 cells were sourced from ATCC and Coriell Institute. Cells were not independently verified or tested for mycoplasma. Cells were cultured according to Coriell Institute's growth protocols and standards. Briefly, K562 and GM12878 were cultured in RPMI 1640 medium (Gibco) supplemented with 10% FBS at 37 °C under 5% CO_2.

Total RNA preparation. Total RNA was extracted from K562 and GM12878 using TRIzol (Invitrogen) according to the manufacturer's protocol. DNase treatment was subsequently performed on each sample with TURBO DNase (Life Technologies) followed by a cleanup with the RNA Clean and Concentrator-25 Kit (Zymo Research). Total RNA was run on an Agilent 2100 Bioanalyzer to assess RNA quality and both the NanoDrop (Thermo Scientific) and Qubit 2.0 Fluorometer (Life Technologies) were used to determine RNA concentration. Only RNA with an RNA integrity number (RIN) > 8.0 was used for library preparation.

Design of *in silico* chromosome. An *in silico* chromosome (*chrIS_R*) sequence was designed to incorporate intergenic regions, artificial gene loci and repeat elements. Sequences were originally downloaded from the reference human genome (hg38).

Intergenic regions. Hg38 was initially binned into 1-Mb windows. Windows were then ranked according to gene density and repeat density and systematically sampled to ensure proportional representation. Intergenic regions between annotated genes (GENCODE v19 basic annotation[20]) were extracted and inverted to generate artificial sequence.

Artificial genes. Human isoforms were ranked according to primary transcript length, mature transcript length and exon count and systematically sampled to ensure proportional representation. Additional alternative isoforms to selected isoforms were also retrieved from original annotations. The exon and intron sequences of selected isoforms were inverted separately to generate an artificial sequence that maintains exon and intron architecture. Splice sites and polypyrimidine tracts were maintained in introns to ensure directionality and local sequence content.

Repeat DNA. Repetitive elements were retrieved from RepeatMasker (downloaded from UCSC Genome Browser; http://www.genome.ucsc.edu/) and ranked according to length. Repeat elements were then systematically sampled to ensure proportional representation.

Assembly. Intergenic regions, artificial genes and repeat DNA were assembled into a single contiguous chromosome sequence. To ensure an artificial sequence, ~75 nt windows were matched to the nucleotide collection database (nr/nt) using BLASTN[47] (word size = 28; expected threshold = 0.01; 1,−2 mismatch score and linear gap costs). Any sequences with significant matches to natural nucleotide sequences (E-value < 0.01) were subjected to local shuffling, nucleotide substitution and/or manual curation to abolish matches. Regions in which significant homology could not be abolished—for example, low-complexity or homopolymeric sequences—were removed.

The final ~11 Mb *chrIS_R* sequence (FASTA format) and synthetic transcriptome annotations (GTF format) are available for download at http://www.sequin.xyz.

Generation of synthetic RNA sequins. The sequences of individual RNA sequin isoforms, along with their molar concentrations in mixtures, are available for download at http://www.sequin.xyz. An SP6 promoter sequence was added to the 5′ end of the sequence to allow for transcription, and a poly(A) tail and EcoRI restriction enzyme site were added to the 3′ end of the sequence to allow for linearization of the plasmid. The sequence was synthesized and inserted into a pMA vector by GeneArt (Life Technologies). Competent *E. coli* cells (Bioline) were thawed on ice and transformed with 2 µL of diluted NEBuilder HiFi DNA Assembly product per the manufacturer's suggested protocols. Transformed cells were plated on prewarmed 100 µg/mL ampicillin plates and incubated at 37 °C overnight (18 h). One colony from each plate was used to inoculate 5 mL Luria–Bertani (LB) broth containing 100 µg/mL ampicillin. Inoculated tubes were incubated overnight on a shaker at 37 °C. Plasmids were isolated using the Zyppy Plasmid Miniprep Kit (Zymo Research). The sequences of the purified plasmids were validated with Sanger sequencing. For RNA synthesis, each plasmid was linearized with EcoRI-HF (New England BioLabs), followed by Proteinase K treatment. The linearized plasmid was cleaned up using the Zymo ChIP DCC columns (Zymo Research). An *in vitro* transcription reaction was performed to synthesize the RNA transcripts. Full-length RNA transcripts were synthesized using the MEGAscript SP6 kit (Life Technologies) according to the manufacturer's protocol. RNA was purified using an RNA Clean & Concentrator-25 column (Zymo Research) using the manufacturer's >200 nt protocol. Purified RNA transcripts were verified on an Agilent 2100 Bioanalyzer with the RNA Nano kit (Agilent Technologies).

Preparation of RNA sequins mixtures. The concentrations of all synthetic transcripts were measured on a Qubit 2.0 Fluorometer (Life Technologies). RNA sequin isoforms were initially combined at equimolar concentrations (82.5 amol/µL) to ensure sufficient coverage for validation of sequin designs with experimental reads. Staggered mixtures were manufactured in two formulations, Mixtures A & B, each containing the full complement of 164 transcripts. The transcripts in each mixture were pooled using an ep*Motion* 5070 epBlue software program (Eppendorf) to make the final mixtures.

Spiking of RNA libraries with sequins. 1 µg of total RNA was used in each library preparation. Sequins were added at 10% of the total RNA concentration of K562 and GM12878 before library preparation. The RNA mixture was depleted for rRNA using Ribo-Zero Magnetic Kit (Human/Mouse/Rat) (Epicentre). Ribodepleted RNA was used to prepare libraries using KAPA Stranded RNA-Seq Library Preparation Kit for Illumina platforms (KAPA Biosystems) according to the manufacturer's protocol. Prepared libraries were quantified using the HS dsDNA Qubit Assay on a Qubit 2.0 Fluorometer (Life Technologies) and verified on an Agilent 2100 Bioanalyzer (Agilent Technologies) before samples were pooled for sequencing.

RNA sequencing. Samples were sequenced on an Illumina HiSeq 2500 instrument with 125 bp paired-end reads. All standard RNA sequins mixtures (neat and spiked-in) were sequenced on a single lane, while all fusion sequins mixtures (neat and spiked-in) were sequenced on a separate lane. Standard Illumina adaptor sequences were trimmed from reads using CutAdapt[48] (v1.8.1).

doi:10.1038/nmeth.3958

Generation of simulated libraries. Simulated libraries were generated for all RNA sequins ($n = 164$) and a randomly chosen subset of 78 human genes (comprising 170 isoforms) localized to chr21 (from GENCODE) using Sherman (http://www.bioinformatics.babraham.ac.uk/projects/sherman/) (v0.1.7; default parameters were used, with bisulfite conversion disabled). Simulated reads were aligned using TopHat2 (v2.1.0; ref. 21) to a combined genome comprising *chrIS_R* and hg38, with the option–max-intron-length 550000 (without this option, TopHat fails to assemble the *R1_92_1* isoform, whose fifth intron spans ~545 kb). The genome index was compiled using Bowtie2 (v2.1.0; ref. 49) (with default parameters) after concatenating both FASTA files together. Full-length transcripts were assembled using StringTie[22] (v1.1.2) with default parameters. BAM alignment files were sequentially subsampled using SAMtools[50] in order to compare transcriptome assembly at different simulated sequencing depths. We used Cuffcompare[51] (v2.2.1) to measure the sensitivity (Sn) of assembly at the base, exon, intron and gene levels.

Alignment, assembly and quantification of RNA sequins. Trimmed reads were aligned to a combined genome index comprising both *chrIS_R* and hg38. Reads were aligned using TopHat2 (v2.1.0) with the same parameters described above with the additional option–library-type fr-firststrand. Alignment was performed both with and without a reference GTF file comprising GENCODE and synthetic gene annotations. Alignments to *chrIS_R* were subsampled using SAMtools to achieve coverage equivalent to the accompanying human genome. Output BAM files were assembled into full-length transcripts using StringTie (v1.1.2) with the following option: –f 0.01 (if the default setting of 0.1 is used, low-abundance RNA sequin isoforms may be missed). We calculated quantitative expression values (including FPKM) for genes and isoforms using StringTie's –G, –B and –A options, again supplying our combined GTF file. Transcripts were visualized in IGV[52], using our combined FASTA file as the reference genome.

Exon-level analysis of RNA sequins. Sequin isoform structures were first collapsed into 883 distinct exon counting bins (i.e., genomic intervals corresponding to unique exons, or parts thereof; **Supplementary Fig. 11b**). Percent spliced in (PSI) index was calculated for all exon bins by modifying the code provided in Schafer *et al.*[53] When comparing terminal (i.e., first and last) exons with internal exons, we excluded single-exon transcripts ($n = 8$) and ambiguous exon bins (i.e., bins that were terminal in one isoform but internal in another) from the analysis. Base-level assembly of terminal and internal exons was calculated using the Coverage utility from BEDTools[54] (v2.25.0). We used DEXSeq[28] to test for differential exon usage (i.e., alternative splicing) between Mixtures A & B, using an adjusted *P* value threshold of 0.1 to assign statistical significance.

Testing the potential for sequins to cross-align to other genomes. Simulated and experimental neat sequins libraries were aligned to hg38, as well as five separate nonhuman eukaryotic genomes (*Mus musculus* (mm10), *Gallus gallus* (gg4), *Danio rerio* (dr10), *Drosophila melanogaster* (dm6) and *Caenorhabditis elegans* (ce10)). Genome sequences were downloaded from UCSC Genome Browser, and alignment was performed using TopHat2 with parameters described above.

To determine potential for cross-alignment between sequins and hg38, we aligned neat sequins library to *chrIS_R* or hg38. Reads that aligned to both sequences within mismatch and alignment score (default unless specified) were considered cross-aligners. The same process was performed for reads from the K562 RNA-seq library to determine potential for human-derived reads to cross-align.

Comparison of RNA sequins with endogenous human genes. To compare performance of RNA sequins to endogenous human genes using experimental reads, we selected the top ten most highly expressed GENCODE protein-coding transcripts detected in the accompanying K562 library. For each highly expressed gene, we selected the RNA sequin that had the most comparable transcript size and exon count. Alignments associated with either the relevant endogenous gene or matched sequin were extracted using BEDTools, and gene–sequin pairs were normalized to correct for differences in alignment coverage. Finally, BAM files from each gene–sequin pair were sequentially subsampled and assembled into full-length transcripts using StringTie, and sensitivity of assembly was measured using Cuffcompare.

Assessment of quality and coverage distribution of sequenced reads. We used FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (v0.10.1) to compare the per-base quality, per-sequence quality and length distribution of reads emanating from sequins and endogenous genes in our spiked-in K562 library. To compare read coverage distribution of sequins with endogenous transcripts, we selected the top 164 expressed protein-coding GENCODE transcripts (from K562 library) and calculated normalized transcript coverage using the CollectRnaSeqMetrics utility from Picard Tools (http://broadinstitute.github.io/picard/) (v2.0.1).

Differential gene expression analysis. We used HTSeq-Count[55] (v0.6.0) to count reads mapping to annotated genes and DESeq2 (v1.10.0; ref. 31) to test for differential expression between Mixtures A & B. We used an adjusted *P* value threshold of 0.1 to assign statistical significance. We generated LODR curves, MA plots and ROC curves for five sequins gene subsets by modifying the code from the ercdashboard[16] Bioconductor package. For these analyses, we grouped together sequin genes that had expected $\log_2$-fold changes (LFC) of the same magnitude, regardless of direction (i.e., genes with expected LFC of 4 and −4 were grouped together, 3 with −3, 2 with −2 and 1 with −1).

Validation of RNA sequin abundance using qRT-PCR. First-strand cDNA synthesis was performed with 600 ng of RNA sequin Mixtures A and B using SuperScript II Reverse Transcriptase (Life Technologies) and random primers (Random Primer 6, New England BioLabs) with the standard protocol.

A 1:5 dilution of the resulting cDNA was used to perform qRT-PCR reactions, using Power SYBR Green PCR Master Mix (Life Technologies) and 2 μM primers (Integrated DNA Technologies) designed for ten RNA sequin genes. Primer sequences are available for download at http://www.sequin.xyz. These ten sequins were selected so as to encompass a range of different expected fold changes.

doi:10.1038/nmeth.3958

qRT-PCRs were run in triplicate on an Applied Biosystems 7900HT Fast Real-Time PCR System—including no template controls—using standard conditions and an additional dissociation step. Cycle threshold (CT) values were normalized to the lowest-concentration RNA sequin for each mixture. Mean average ΔCT values were compared between mixtures to determine fold change for each RNA sequin.

Comparison of fusion sequins with endogenous human gene fusions. To compare the detection sensitivity of fusion sequins with endogenous human gene fusions, we used a set of nine synthetic transcripts that corresponded to previously reported oncogenic gene fusions for comparison[36]. We simulated reads from these nine fusion transcripts, as well as a size-matched set of nine fusion sequins, using Sherman (with the parameters described above). We then sequentially subsampled FASTQ files using the Reformat utility from BBMap (http://sourceforge.net/projects/bbmap/) (v35.50) and compared the sensitivity of fusion detection (see below) at a range of different simulated read coverages.

Detection of gene fusions. We aligned simulated and experimental reads to our combined genome with STAR[37] (v2.4.2a) using the parameters suggested by Stransky *et al.*[35] for optimal fusion detection. We then supplied the resultant chimeric SAM files to STAR-Fusion (https://github.com/STAR-Fusion/STAR-Fusion) (v0.5.4), along with our combined GTF file, to compile a list of fusion candidates. Putative fusion events were filtered on the basis of minimal split-read and spanning-fragment support, alignment of breakpoints with reference exon splice sites, and shared homology or repetitive sequence between fusion partners. To determine parent gene expression, we measured the number of reads spanning the normal intron immediately downstream of the location of the fusion breakpoint.

Statistical analyses. We generated plots and performed regression analyses using GraphPad Prism (GraphPad Software, v6.0f) and R (R Foundation for Statistical Computing, v3.1.3). We used piecewise linear regression to calculate the limit of quantification at the gene, isoform and exon levels (defined as the point which separates the two segmental regression lines fitting the data as closely as possible while minimizing the total residual sum of squares).

Data availability. Requests for RNA sequins and associated data files (including *in silico* chromosome, sequin isoform sequences, annotations, sequencing libraries and primer sequences) and supporting software can be made at http://www.sequin.xyz.

47. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
48. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
49. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
52. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
53. Schafer, S. *et al.* Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). *Curr. Protoc. Hum. Genet.* **87**, 11.16.11–11.16.14 (2015).
54. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
55. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

doi:10.1038/nmeth.3958