Title

UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy

Running Title

Modelling UMI errors improves quantification accuracy

Authors

Tom Smith [1]

Andreas Heger[1]

Ian Sudbery[2]*

1. Computational Genomics Analysis and Training Programme, MRC WIMM Centre for Computational Biology, University of Oxford, John Radcliffe Hospital/Headley Way, Oxford OX3 9DS

2. Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield, UK, S10 2TN

* Corresponding author

Keywords

UMI

Unique Molecular Identifier

Random Tag

PCR-duplicates

iCLIP

Single Cell RNA-seq

Sequencing

Abstract

Unique Molecular Identifiers (UMIs) are random oligonucleotide barcodes that are increasingly used in high-throughput sequencing experiments. Through a UMI, identical copies arising from distinct molecules can be distinguished from those arising through PCR-amplification of the same molecule. However, bioinformatic methods to leverage the information from UMIs have yet to be formalised. In particular, sequencing errors in the UMI sequence are often ignored, or else resolved in an *ad-hoc* manner. We show that errors in the UMI sequence are common and introduce network-based methods to account for these errors when identifying PCR duplicates. Using these methods, we demonstrate improved quantification accuracy both under simulated conditions and real iCLIP and single cell RNA-seq datasets. Reproducibility between iCLIP replicates and single cell RNA-seq clustering are both improved using our proposed network-based method, demonstrating the value of properly accounting for errors in UMIs. These methods are implemented in the open source *UMI-tools* software package.

Background

High throughput sequencing technologies yield vast numbers of short sequences (reads) from a pool of DNA fragments. Over the last ten years a wide variety of sequencing applications have been developed which estimate the abundance of a particular DNA fragment by the number of reads obtained in a sequencing experiment (read counting) and then compare these abundances across biological conditions. Perhaps the most widely used read counting approach is RNA-seq, which seeks to compare the number of copies of each transcript in different cell types or conditions. Prior to sequencing, a PCR amplification step is normally performed to ensure sufficient DNA for sequencing and/or enrichment for fragments with successful adapter ligation. Biases in the PCR amplification step lead to particular sequences becoming overrepresented in the final library (Aird et al. 2011). In order to prevent this bias propagating to the quantification estimates, it is common to remove reads or read pairs with the same alignment coordinates as they are assumed to arise through PCR amplification of the same molecule (Sims et al. 2014). This is appropriate where sequencing depth is low and thus the probability of two independent fragments having the same genomic coordinates are low, as with paired-end whole genome DNA-seq from a large genome. However, the probability of generating independent fragments mapping to the same genomic coordinates increases as the distribution of the alignment coordinates deviates from a random sampling across the genome and/or the sequencing depth increases. For example, in RNA-seq, highly expressed transcripts are more likely to generate multiple fragments with exactly the same genomic coordinates. The problem of PCR duplicates is more acute when greater numbers of PCR cycles are required to increase the library concentration, as in single cell RNA-seq, or when the alignment coordinates are limited to a few distinct loci, as in individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation (iCLIP). Random barcodes were initially proposed as a method to count the number of mRNA molecules in a sample (Hug and Schuler 2003), and have since been used to explicitly label PCR duplicates (McCloskey et al. 2007). More recently, random barcodes, referred to as unique molecular identifiers (UMIs), have been employed to confidently identify PCR duplicates in high-throughput sequencing experiments (König et al. 2010b; Kivioja et al. 2012; Islam et al. 2014). By incorporating a UMI into the same location in each fragment during library preparation, but prior to PCR amplification, it is possible to accurately identify true PCR duplicates as they have both identical alignment coordinates and identical UMI sequences (Figure 1a). In addition to their use in single cell RNA-seq and iCLIP (König et al. 2010b) , UMIs may be applied to almost any sequencing method where confident identification of PCR duplicates by alignment coordinates alone is not possible and/or an accurate quantification is required, including ChIP-exo (He et al. 2015), DNA-seq karyotyping (Karlsson et al. 2015), detection of rare mutations (Schmitt et al. 2012) and antibody repertoire sequencing (Vollmers et al. 2013).

Accurate quantification with UMIs is predicated on a one-to-one relationship between the number of unique UMI barcodes at a given genomic locus and the number of unique fragments that have been sequenced. However, errors within the UMI sequence including nucleotide substitutions during PCR, and nucleotide miss calling and insertions or deletions (Indels) during sequencing, create additional artefactual UMIs. Nucleotide miss-calling and substitution errors affect only the UMI sequence itself and do not affect the alignment coordinates. Hence, these errors will inflate the estimation of the number of unique molecules at a particular genomic coordinate. These errors can be identified by examining all UMIs at a single genomic coordinate. On the other hand, UMI Indels will affect the alignment position also, leading to the assignment of reads to incorrect genome coordinates. Identification of such events requires the examination of sets of UMIs at neighbouring coordinates. Recombination events, also called ‘PCR jumping’ create chimeric sequences that may change either the UMI sequence and/or alignment. Miss-calling during sequencing is by far the most prevalent error, occurring 1-2 orders of magnitude more frequently than Indels for Illumina sequencing (Marinier et al. 2015; Schirmer et al. 2015, 2016). Recombination is common when sequencing amplicons, but much rarer with the shotgun sequencing approaches where UMIs are utilised (Schloss et al. 2011; Waugh et al. 2015). We therefore focus here on improving quantification via UMIs by considering nucleotide miss-calling and substitution errors within pools of UMIs from the same genomic coordinate. Herein, we will refer to these errors as UMI errors.

UMI errors have been considered in previous analyses (Macosko et al. 2015; Bose et al. 2015; Yaari and Kleinstein 2015; Islam et al. 2014). However, their impact on quantification accuracy has not previously been demonstrated and there is no consistency in the approach taken to resolve these errors. For example, Islam *et al* (2014) removed all UMIs where the counts were below 1 % of the mean counts of all other non-zero UMIs at the genomic locus, whilst Bose *et al* (2015) merged together all UMIs within a hamming distance of two or less, with little explanation as to how this was achieved. We therefore set out to demonstrate the need to account for UMI errors, to compare different methods for resolving UMI errors and to formalise an approach for removing PCR duplicates with UMIs.

Results

We reasoned that UMI errors create groups of similar UMIs at a given genomic locus. To confirm this, we calculated the average number of bases different (edit distance) between UMIs at a given genomic locus and compared the distribution of average edit distances to a null distribution generated by randomly sampling (see methods). Using iCLIP data (Müller-McNicoll et al. 2016), we confirmed that the UMIs are more similar to one another than expected according to the null, strongly suggesting sequencing and/or PCR errors are generating artefactual UMIs (see methods; Figure 1b, see Figure S1 for other datasets). Furthermore, the enrichment of low edit distances is well correlated with the degree of PCR duplication (Figure 1c). Overall, we detected a 25-fold enrichment for positions with an average edit distance of 1, compared to our null expectation. In contrast, when we compared the UMI sequences at adjacent positions we detected an 1.1-fold (+/- standard deviation of 0.1, see materials and methods) enrichment for UMIs which may have originated from a single nucleotide deletion, suggesting UMI Indels are much less prevalent than UMI errors, as expected. We then constructed networks between UMIs at the same genomic locus where nodes represent UMIs and edges connect UMI separated by a single nucleotide difference. Whilst most of the networks contained just a single node, we observed that 3-36% of networks contained two or more nodes, of which 4-20% did not contain a single central node, and thus could not be naively resolved (Figure 1d). This indicates that the majority of networks are likely to originate from a single unique molecule prior to PCR amplification, but a minority of networks may originate from a combination of errors during PCR and sequencing or may originate from multiple unique molecules, which by chance have similar UMIs.

Methods to identify unique molecules

Many previous studies assume each UMI at a given genomic locus represents a different unique molecule (Collins et al. 2015; Shiroguchi et al. 2012; Soumillon et al. 2014). We refer to this method as *unique*. Islam *et al* (2014) previously identified the issue of sequencing errors and proposed removing UMIs whose counts fall below a threshold of 1% of the mean of all non-zero UMIs at the locus, a method we refer to as *percentile*.

We have developed three methods to identify the number of unique molecules at a given locus by resolving UMI networks formed by linking UMIs separated by a single edit distance (Figure 1e). In all cases, the aim is to reduce the network down to a representative UMI(s) that accounts for the network; identifying the exact sequence of the original UMI(s) is not important for the purposes of quantification. The simplest method we examined was to merge all UMIs within the network, retaining only the UMI with the highest counts. For this method, the number of networks formed at a given locus is equivalent to the estimated number of unique molecules. This is similar to the method employed by Bose et al (2015) where UMIs with an edit distance of 2 or less were considered to originate from an identical molecule. We refer to this method as *cluster*. This method is expected to underestimate the number of unique molecules, especially for complex networks. We therefore developed the *adjacency* method which attempts to correctly resolve the complex networks by using the node counts. The most abundant node and all nodes connected to it are removed from the network. If this does not account for all the nodes in the network, the next most abundant node and its neighbours are also removed. This is repeated until all nodes in the network are accounted for. In the method, the total number of steps to resolve the network(s) formed at a given locus is equivalent to the number of estimated unique molecules. This method allows a complex network to originate from more than one UMI, although UMIs with an edit distance of two will always be removed in separate steps. The excess of UMIs pairs with an edit distance of two observed in the iCLIP datasets indicate that some of these UMIs are artefactual. Reasoning that counts for UMIs generated by a single sequencing error should be higher than those generated by two errors and UMIs resulting from errors during the PCR amplification stage should have higher counts than UMIs resulting from sequencing errors, we developed a final method, *directional***.** We generated networks from the UMIs at a single locus, in which directional edges connect nodes a single edit distance apart when $n_a \geq 2n_b - 1$, where n_a and n_b are the counts of node a and node b. The entire directional network is then considered to have originated from the node with the highest counts. The ratio between the final counts for the true UMI and the erroneous UMI generated from a PCR error is dependent upon which PCR cycle the error occurrs and the relative amplification biases for the two UMIs, but should rarely be less than 2-fold. The *-1* component was included to account for strings of UMIs with low counts, each separated by a single edit distance for which the *2n* threshold alone is too conservative. This method allows UMIs separated by edit distances greater than one to be merged so long as the intermediate UMI is also observed, and with each sequential base change from the most abundant UMI, the count decreases. For this method, the number of directional networks formed is equivalent to the estimated number of unique molecules.

Comparing methods with simulated data

To compare the accuracy of the proposed methods we simulated the process of UMI amplification and sequencing for UMIs at a single locus and varied the simulation parameters (see methods). To examine the accuracy of the 5 methods, we computed two metrics: The log2-fold difference between the estimate and ground truth *Log2((estimate - truth) / truth)* and the coefficient of variance (*standard deviation / mean)* across 10,000 iterations. Increasing UMI length or sequencing depth results in a linear increase in the degree of overestimation for *unique* and *percentile* (Figure 2a, b), since increasing either parameter leads linearly increases the total amount of UMI sequence that may harbour errors. In contrast, the estimates from the network-based methods remain relatively stable, with *directional* showing the highest accuracy and lowest variance. We also simulated the effect of including a very long UMI (up to 50 bp) as there may be occasions where it is preferable to concatenate a UMI with another barcode, such as a sample barcode or cell barcode in single cell RNA-seq, leading to longer barcodes. We noted that the network-based methods showed reduced accuracy for very long barcodes (Figure S2a). Investigating further, we found this was correlated with an increase in UMIs with two errors where the single error intermediate was not observed, as detected by counting the number of networks which did not contain any of the initial UMIs prior to PCR and sequencing (Figure S2b). In order to resolve this inherent problem with very long UMIs, we modified the network-based methods so that edges joined nodes with an edit distance less than or equal to 2. This considerably decreased the number of networks without any initial UMIs and improved the accuracy of the network-based methods for very long UMI sequences (Figure S2a-b).

Increased sequencing error rate leads to an exponential overestimation for *unique* and *percentile* (Figure 2c), with a 1.3-fold overestimation observed with an error rate of 0.01, compared to less than 1.05-fold for the network based methods. Increasing the rate of errors during the PCR step had a similar impact (Figure S2c). However, this was only observed when the rate of DNA polymerase errors was simulated as greater than 0.001, considerably higher than reported error rates for even non-recombinant *Taq* DNA polymerase (Rittié and Perbal 2008; Whalen et al. 2016), confirming sequencing errors are likely to be the primary source of UMI errors. Increasing the number of PCR cycles or modifying the amplification bias had little impact on the relative accuracy of the methods (Figure S2d, e). Increasing the number of initial UMIs reduced the accuracy of the network-based methods, however even with 100 initial 8bp UMIs at a single locus, the network methods remained the most accurate (Figure S2f).

Although the network methods performed very similarly, *directional* consistently yielded more accurate and less variable estimates. For example, when the sequencing depth was increased to 400 reads, the average estimates were 19.92, 19.94 and 19.99 (truth=20) respectively for *cluster*, *adjacency* and *directional* methods, and the CVs were 0.0167, 0.0144 and 0.0099. We observed no difference between *percentile* and *unique* under most conditions tested. Increasing the number of reads sequenced per initial UMI, we were able to see an improvement in accuracy for *percentile* relative to *unique* when sequencing error rates are between 1 x 10^{-3} – 1 x 10^{-5}, however, even under this specific parameterisation, the network-based methods are more accuracy (Figure S2g).

In summary, under simulation conditions, the *directional* method outperforms all other methods, however *adjacency* and *cluster* performs equally well under simulation conditions that are expected to reflect a well-designed experiment and well-executed experiment.

Implementation

To implement our methods within the framework of removing PCR duplicates from BAM alignment files, we developed a command line toolset, UMI-tools, with two commands, *extract* and *dedup*. ex*tract* takes the UMI from the read sequence contained in a FASTQ read sequence and appends it to the read identifier so it is retained in the downstream alignment. *extract* expects the UMIs to be contained at the same location in each read. Where this is not the case, e.g with sequencing techniques such as inDrop-seq (Klein et al. 2015), the user will need to extract the UMI sequence from the read sequence and append it to the read identifier. *dedup* takes an alignment BAM file, identifies reads with the same genomic coordinates as potential PCR duplicates, and removes PCR duplicates using the UMI sequence according to the method chosen. Time requirements for running *dedup* depend on number of input reads, length of UMI and level of duplication. Memory requirements depend on the number of output reads. On a desktop with a Xeon E3-1246 CPU, it takes ~220 seconds and ~100MB RAM to process a 32 million read single-end input file with 5bp UMIs to ~700,000 unique alignments. Inputs with longer UMIs may take significantly longer.

Comparing methods with iCLIP data

We next sought to examine the effect of these methods on real data, starting with the previously mentioned iCLIP data, which includes 3-6 replicates for 9 proteins (Müller-McNicoll et al. 2016). For replicate 1, the distribution of the average edit distance between UMIs present at each genomic locus showed enrichment for single edit distance relative to a null distribution from random sampling, taking into account the genome-wide distribution of UMIs (Figure 3a). For all samples, application of the *directional* method resulted in an edit-distance distribution resembling the null, whereas using the *percentile* method made little or no difference. The same was also true of other replicates of this dataset or other datasets (Figure S2). In some cases a residual enrichment of positions with an average edit distance of 2 was observed, but this was also reduced in most cases.

We reasoned that if the *directional* method removed PCR duplicates more accurately, the reproducibility between replicates should be improved. To test this we turned to a previously defined measure of iCLIP reproducibility (König et al. 2010b) . Briefly, we identified in each sample the bases with two or more tags mapping at that positions and asked what percentage had a tag present in one or more other replicates for that pull-down. We limited the analysis to the first three replicates for each protein. In each case, after de-duplication with the *directional* method, bases with two or more tags were more reproducible (Figure 3b), with the difference being very large in some cases (e.g. 21% vs 59% of bases reproducible for SRSF7 replicate 1). In contrast, the *percentile* method was little different from *unique* (Figure S3).

In order to measure reproducibility of their data, Müller-McNicoll *et al* measured the spearman's rank correlation between the numbers of significant tags in each exon across the genome. We repeated this calculation with data processed using either the *unique* or *directional* method, and compared the average spearman's correlation between each sample and other replicates of the same pull down. In all cases we see an improvement in the correlation between replicates of the same pull down when data are processed using the *directional* method (Figure 3c). As expected, the degree of improvement for a particular sample was correlated with the enrichment of positions with an average edit distance of 1 (Figure S3; R^2=0.4). Thus our method substantially improves the reproducibility of replicates in this iCLIP experiment.

Comparing methods with Single Cell RNA Seq data

To further demonstrate the utility of our network-based method, we applied it to two differentiation single cell RNA-seq data sets: the first reported use of UMIs in a single cell RNA-seq experiment seeking to describe a developmental pathway (Soumillon et al. 2014), referred to here as SCRB-seq, and a recently reported single cell RNA-seq utilising droplet-barcoding (Klein et al. 2015), referred to here as inDrop-seq . As before, network-based methods show a marked improvement in the distribution of edit distances over the *percentile* method and the *unique* method (Figure 4a). Improvements are generally less pronounced than observed with the iCLIP data, likely due to a lower maximum read depth in single cell RNA-seq. To demonstrate that this improvement in the edit distance lead to an improved accuracy in transcript abundance estimates we used the ERCC spike-ins. The naïve use of UMIs to identify PCR duplicates with the *unique* method improved the per-cell correlation between ERCC concentration and counts, compared to quantification without considering PCR duplicates (median coefficients were 0.86 and 0.89, respectively). As expected, the correlation was further improved using the *directional* method (median coefficient = 0.91; Figure 4b).

We applied hierarchical clustering to the SCRB-seq gene expression data using the *unique* method and observed the Day 0 and Day 14 cells separately relatively well (Figure 4c). However, 7 cells clustered with cells of the wrong time point, reflecting either a failure to commit to differentiation or miss-classification event due to noise in the expression estimates. With the *directional* method this was reduced to 5 cells, suggesting that failure to account for UMI errors can lead to miss-classification in single cell RNA-seq. Applying hierarchical clustering to the the inDrop-seq gene expression estimates, we observed that 44/2717 (1.6%) of cells clustered with cells from another timepoint when using the *unique* method. Biological variation in the progression of differentiation may explain Day 2, Day4 and Day 7 miss-classification events. However, 19/44 events involved undifferentiated mES cells, suggesting these miss-classification events were the result of low-accuracy quantification estimates (Figure 4d). With the application of the *directional* method, the rate of miss-classification was reduced to 0.9% and, strikingly, all the mES cells were correctly classified. These results indicate that application of the *directional* method improves the quantification estimates and can improve classification by hierarchical clustering.

Discussion

UMIs can be utilised across a broad range of sequencing techniques, however bioinformatic methods to leverage the information from UMIs have yet to be standardised. In particular, others have noted the problem of UMI errors, but the solutions applied are varied (Bose et al. 2015; Islam et al. 2014) . The *adjacency* and *directional* methods we set out here are, to our knowledge, novel approaches to remove PCR duplicates when using UMIs. Comparing these methods to previous methods with simulated data, we observed that our methods are superior at estimating the true number of unique molecules. Of the three network-based methods, *directional* was the most robust over the simulation conditions and should be preferred. We note that the performance of all network-based methods will decrease as the number of aligned reads at a genomic locus approaches the number of possible UMIs, however this is an intrinsic issue with UMIs and not one that can be solved computationally post-sequencing. For this reason, we recommend all experiments to use UMIs of at least 8 bp in length and to use longer UMIs for higher sequencing depth experiments. The simulations also indicated that very long UMIs actually decrease the accuracy of quantification when not accounting for UMI errors, since the UMIs are more likely to accumulate errors. For experiments utilising long UMIs, network-based methods therefore show an even greater performance relative to the *unique* method. The simulations provide an insight into the impact on quantification accuracy and indicate that application of an error-aware method is even more important with higher sequencing depth. This is perhaps most pertinent for single cell RNA-seq, as cost decreases continue to drive higher sequencing depths.

The analysis of iCLIP and single cell RNA-seq and data sets established that UMI errors present in all of the data sets tested and that quantification accuracy could therefore be improved by modelling these errors during the deduplication step. The frequency of UMI Indels was far less than UMI errors suggesting only minimal gains would be achieved by considering UMI Indels also. We observed an improved distribution of edit distances for all samples when using network-based methods to detect PCR duplicates, although theoretical reasoning and empirical evidence suggests that the extent of the errors depends on the quality of the sequencing base calls and the sequencing depth, as confirmed by the simulations.

Modelling UMI errors yielded improvements in single cell RNA-seq sample clustering, demonstrating the value of considering UMI errors. Since iCLIP aims to identify specific bases bound by RNA binding proteins, datasets have a high level of PCR duplication. The effects of UMI errors are therefore particularly strong, creating the impression of reproducible cross-linking sites within a replicate but not between replicates, for example only 21% of positions with two or more tags in SRSF7 replicate 1 had any tags in replicates 2 or 3 when naive de-duplication was used, but this increased to 59% when the *directional* method was used (Figure 3b). Application of the network based methods increases the correlation between replicates in all cases, with larger differences in samples where PCR duplication was higher. From the results of the simulation and real data analyses, we recommend the use of an error-aware method to identify PCR duplicates whenever UMIs are used.

We provide our methods within the open-source UMI-tools software (https://GitHub.com/CGATOxford/UMI-tools, included here as Supplementary File 1), which can easily be integrated into existing pipelines for analysis of sequencing techniques utilising UMIs.

Methods

Simulation

To simulate the effects of errors on UMI counts, an initial number of UMIs were generated at random, with a uniform random probability of amplification [0.8-1.0] assigned to each initial UMI. To simulate a PCR cycle, each UMI was selected in turn and duplicated according to its probability of amplification. Polymerase errors were also added randomly at this stage and any resulting new UMI sequences assigned new probabilities of amplification. Following multiple PCR cycles, a defined number of UMIs were randomly sampled to model the sampling of reads during sequencing ("sequencing depth") and sequencing errors were introduced with at a given probability, with all errors (e.g A -> T, C -> G) being equally likely. The number of true UMIs within the sampled UMIs was then estimated from the final pool of UMIs using each method. To test the performance of the methods under a variety of simulation parameters, each parameter was varied in turn. The following values are the range of the parameter values tested with the value used for all other simulations in parentheses. Sequencing depth 10-400 (100), number of initial UMIs 10-100 (20), UMI length 6-16 (8), DNA polymerase error rate 1 x 10^{-3} – 1 x 10^{-7} (1 x 10^{-5}), sequencing error rate 1 x 10^{-1} -1 x 10^{-5} (1 x 10^{-3}), number of PCR cycles 4-12 (6), minimum amplification probability 0.1-1 (0.8). The maximum amplification probability was set at 1 with the probability of amplification for an each UMI drawn from a uniform distribution.

Real data

Re-analysis of the iCLIP and Single Cell RNA-seq data was performed with in-house pipelines following the methods described in the original publication with exceptions as highlighted below. Pipelines are available at https://GitHub.com/CGATOxford/UMI-tools_pipelines and as Supplementary File 2.

iCLIP

Raw sequence was obtained from the European Nucleotide Archive (accessions SRP059277 and ERR039854) (Müller-McNicoll et al. 2016; Tollervey et al. 2011). Raw sequences were processed to move the UMI sequences to the read name using 'umi_tools extract'. Sample barcodes were verified and removed, and adaptor sequence removed from the 3' end of reads using the reaper tool from the Kraken package (version 15-065) (Davis et al. 2013) with parameters: `-3p-head-to-tail 2 -3p-prefix 6/2/1`. Reads were mapped to the same genome as the original publication (mm9 for SRSF dataset, hg19 for the TDP43 dataset) using Bowtie version v1.1.2 (Langmead et al. 2009a) with the same parameters as the original publications (-v 2 -m 10 -a).

We measured the rate at which UMIs might represent Indel mutations by noting that an Indel in the UMI sequence would cause the final base of the presumed UMI to match the genomic base at position -1 relative to the mapping location of the read. Thus we examined each UMI at a particular position, and tested for the presence of a UMI that would correspond to a 1 bp deletion existed at the following base. We compared this to the situation when the UMIs at the following base were randomised, respecting the number of UMIs at the position and the genome-wide usage of each UMI. Enrichment was defined as the count at the unrandomised positions compared to the count at the randomised positions. We calculated this metric for one replicate of each pull down from SRP059277. See the *Examining_indels* notebook in the *UMI-tools_pipelines* repository (included as Supplementary File 2).

Mapped reads were deduplicated using 'umi_tools dedup' using each of the possible methods and edit_distance distribution produced using the '--output-stats' option. For the *cluster* method only the '--further-stats' option was used to output statistics on the distribution of network topology types.

Significant bases were produced by comparing tag count height at each position compared to randomised profiles (König et al. 2010a), and bases with FDR<0.05 retained.

Coverage over exons was calculated by collapsing Ensembl 67 transcripts. Where exons overlapped, they were restricted to their intersection and the number of reads mapped to significant bases counted for each exon. Exons that contained no tags in any sample were removed (König et al. 2010a). Spearman's rho between all pairwise combinations of replicates of pulldowns for the same protein were calculated and averaged for each replicate.

Reproducibility between replicates was calculated as per König *et al* (2010). Bases with a depth greater than 2 were identified in the sample in question, and then the fraction of these bases that had one or more tags in other replicates was calculated.

Single Cell RNA-seq

For both datasets, raw data was downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). For The SCRB-seq data (GSE53638) (Soumillon et al. 2014), a single Day 0 (SRR1058003) and Day 14 (SRR1058023) sample were obtained. For the inDrop data (GSE65525) (Klein et al. 2015), the mouse ES cells sample 1 (SRR1784310), mouse ES cells LIF-, 2 days (SRR1784313), mouse ES cells LIF-, 4 days (SRR1784314) and mouse ES cells LIF-, 7 days (SRR1784315) samples were obtained. FASTQ files were extracted using SRA toolkit. The sequence read filtering, preparation and alignment differed for the two data sets. In both cases, one of the paired end reads contained adapter barcodes and UMI and the other read pair contained sequence for alignment. In addition, with the inDrop data, the position of the UMI within the read varied depending on the length of the cell barcode. For this reason, for both data sets, the UMIs had to be extracted from the reads with bespoke code rather than using UMI-tools *extract*.

For SCRB-seq samples, the UMI was extracted from read 2 and appended onto the read identifier of read 1 to generate a single-end FASTQ. Reads were filtered out if any of the following conditions was not met. Phred sequence quality of all cell barcode bases >=10 and all UMI bases >=30 and cell barcode matched expected cell barcodes. A reference transcriptome was built comprising all human protein-coding genes (Ensembl v75, hg19) and the ERCC spike-ins. Since expression quantification was being performed at the gene level, overlapping transcripts from the same gene were merged so that each gene contained a single transcript covering all exons from all transcripts. Reads were aligned to the reference transcriptome using BWA Aln (Li and Durbin 2009) with the following parameters: *"-l 24 –k 2"* to set seed length to 24 bp, and mismatches allowed in the seed to 2.

For inDrop samples, the cell barcode and UMI were extracted from read 1 and read 2 was written out to a single end FASTQ file with the cell barcode incorporated into the file name and the UMI appended to the read identifier. Only reads containing the adapter sequence (allowing 2 mismatches) were retained. For each sample, only reads containing one of the *n* most abundant cell barcodes were retained, where *n* was the number of cells in a given sample. The resulting single end reads were filtered using trimmomatic v0.32 (Bolger et al. 2014) with the following options: "*LEADING:28 SLIDINGWINDOW:4:20 MINLEN:19*" to remove bases with Phred quality scores below 28 from the 5' end, scan the reads in 4 bp sliding windows and trim when average quality score falls below 20, and retain all reads at least 19bp in length following trimming. Our alignment procedure is a deviation to the method used by Klein *et al* (2015) which involved alignment of reads to a reference transcriptome containing all transcripts (e.g not collapsed into one gene model), reporting up to 200 alignments per read, and dealing with multi-mapping alignments in a downstream step. As this method was not compatible with our de-duplication method we took a simpler approach. A reference transcriptome was built comprising all mouse protein-coding genes (Ensembl v78, mm10) plus ERCC spike-ins. Since expression quantification was being performed at the gene level, overlapping transcripts from the same gene were merged so that each gene contained a single transcript covering all exons from all transcripts. Reads were aligned to the reference transcriptome with Bowtie v1.1.2(Langmead et al. 2009b) with the following options: "*-n1 -l 15 -M 1 --best --strata*" to allow one mismatch, set seed length to 15 bp and report only one alignment where multiple "best" alignments were found. The seed length and mismatch parameters were the same as the Klein *et al* (2015) alignment method.

Following alignment, de-duplication was performed with UMI-tools dedup with *unique*, *percentile* and *directional* used in turn. Both data sets were generated with sequencing methods which generate reads with different alignment coordinates from the same initial DNA fragment (SCRB-seq, CEL-Seq). De-duplication was therefore performed with the "*--per-contig*" option so that the UMI and the contig (in this case, gene) rather than the exact alignment coordinates were used to identify duplicate reads. The "--stats-output" and "--further-stats" options were used to generate summary statistics for the alignment files pre and post de-duplication. Gene expression was quantified by counting the number of remaining reads per gene following de-duplication

Exploratory gene expression analysis

PCA was performed in R (R Core Team 2015) using the *prcomp* function. Hierarchical clustering was performed in R using the *hclust* function and heatmaps generated using the *heatmap.2* function from the gplots package. Clustering was performed using 1 - spearman's correlation coefficient as the distance measure and "ward.D2" as the clustering method. Since many genes show very low expression in the SCRB-seq data, the top 100 most highly expressed genes were selected for clustering.

Data access

UMI-tools is available from pypi (package: umi_tools) and conda (channel: https://conda.anaconda.org/toms, package: umi_tools) or GitHub (https://GitHub.com/CGATOxford/UMI-tools). Analyses conducted in this manuscript used version 0.2.6 - archived on Zenodo as https://doi.org/10.5281/Zenodo.165403, and in Supplementary File 1. Analyses were performed using automated python pipelines. iCLIP specific analyses were completed using the iCLIPlib python library (manuscript in preparation). Figures were created by python pipelines or in Jupyter notebooks using the ggplot2 package (Wickham 2009) unless otherwise noted. All pipelines, notebooks and other code, along with configuration files used are available from the GitHub repository (https://GitHub.com/CGATOxford/UMI-tools_pipelines), archived on Zenodo as https://doi.org/10.5281/zenodo.215974 and in Supplementary File 2.

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. T.S and A.H are supported by an MRC programme grant [G1000902]. We thank David Sims and members of CGAT for reviewing the manuscript. We thank the GitHub community for helping us to improve the UMI-Tools codebase.

Author contribution: T.S and I.S. conceived the study. I.S. implemented the first iteration of UMI-Tools. T.S. implemented further methods and improved the code base. A.H implemented performance improvements to UMI-Tools and advised on software development. T.S. performed the simulation and single cell RNA-seq analyses. I.S. performed the iCLIP analysis. T.S and I.S wrote the original draft of the manuscript. T.S, I.S and A.H edited the final draft of the manuscript. I.S. supervised the study.

Disclosure Declaration

The authors declare that we have no competing interests

References

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**: R18.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bose S, Wan Z, Carr A, Rizvi AH, Vieira G, Pe'er D, Sims P a. 2015. Scalable microfluidics for single cell RNA printing and sequencing. *Genome Biol* **16**: 120.

Collins JE, Wali N, Sealy IM, Morris JA, White RJ, Leonard SR, Jackson DK, Jones MC, Smerdon NC, Zamora J, et al. 2015. High-throughput and quantitative genome-wide messenger RNA sequencing for molecular phenotyping. *BMC Genomics* **16**: 578.

Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. 2013. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**: 41–49.

He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* **33**: 395–401.

Hug H, Schuler R. 2003. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J Theor Biol* **221**: 615–624.

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**: 163–6.

Karlsson K, Sahlin E, Iwarsson E, Westgren M, Nordenskjöld M, Linnarsson S. 2015. Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations. *Genomics* **105**: 150–8.

Kivioja T, Väharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. 2012. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**: 72–4.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**: 1187–1201.

König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010a. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909–915.

König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010b. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909–15.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009a. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009b. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**: 1202–1214.

Marinier E, Brown DG, McConkey BJ. 2015. Pollux: platform independent error correction of single and mixed genomes. *BMC Bioinformatics* **16**: 10.

McCloskey ML, St??ger R, Hansen RS, Laird CD. 2007. Encoding PCR products with batch-stamps and barcodes. *Biochem Genet* **45**: 761–767.

Müller-McNicoll M, Botti V, de Jesus Domingues AM, Brandl H, Schwich OD, Steiner MC, Curk T, Poser I, Zarnack K, Neugebauer KM. 2016. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev* **30**: 553–66.

R Core Team. 2015. R: A Language and Environment for Statistical Computing.

Rittié L, Perbal B. 2008. Enzymes used in molecular biology: A useful guide. *J Cell Commun Signal* **2**: 25–45.

Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**: 125.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* **43**: 1–16.

Schloss PD, Gevers D, Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing Artifacts on 16s rRNA-based studies. *PLoS One* **6**.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci* **109**: 14508–14513.

Shiroguchi K, Jia TZ, Sims PA, Xie XS. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci U S A* **109**: 1347–52.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**: 121–32.

Soumillon M, Cacchiarelli D, Semrau S. 2014. Characterization of directed differentiation by high-throughput single-cell RNA-seq. *BioRxiv*.

Tollervey JR, Curk T, Rogelj B, Briese M, Cereda M, Kayikci M, Hortobágyi T, Nishimura AL, Župunski V, Patani R, et al. 2011. Characterising the RNA targets and position-dependent splicing regulation by TDP-43; implications for neurodegenerative diseases. *Nat Neurosci* **14**: 452–8.

Vollmers C, Sit R V, Weinstein JA, Dekker CL, Quake SR. 2013. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* **110**: 13463–8.

Waugh C, Cromer D, Grimm A, Chopra A, Mallal S, Davenport M, Mak J. 2015. A general method to eliminate laboratory induced recombinants during massive, parallel sequencing of cDNA library. *Virol J* **12**: 55.

Whalen S, Truty RM, Pollard KS. 2016. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**: 488–96.

Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysiss*. Springer-Verlag, New York.

Yaari G, Kleinstein SH. 2015. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* **7**: 121.

Figure Legends

Figure 1. Modelling errors in UMIs

A. Schematic representation of how UMIs are used to count unique molecules. Fragmented DNA is labelled with a random UMI sequence (short oligonucleotide; represented as coloured blocks). Following PCR amplification, sequencing and bioinformatics steps, the sequence read alignment coordinates and UMI sequences are used to identify sequence reads originating from the same initial DNA fragment (PCR duplicates) and so count the unique molecules. **B**. Average edit distances (rounder to integers) between UMIs with the same alignment coordinates. Genomic positions with a single UMI are not shown. Null = Null expectation from random sampling of UMIs, taking into account the genome-wide distribution of UMIs. **C.** Correlation between duplication rate and enrichment of positions with an average edit distance of 1 for iCLIP data. **D**. Topologies of networks formed by joining reads with the same genomic coordinates and UMIs a single edit distance apart. Single hub = One node connected to all other nodes. Complex = No node connected to all other nodes. **E**. Methods for estimating unique molecules from UMI sequences and counts at a single locus. Where the method uses the UMI counts, these are shown. Red bases are inferred to be sequencing errors, blue bases inferred to be PCR errors. The inferred number of unique molecules for each method is shown in parentheses.

Figure 2. Comparison of methods with simulated data

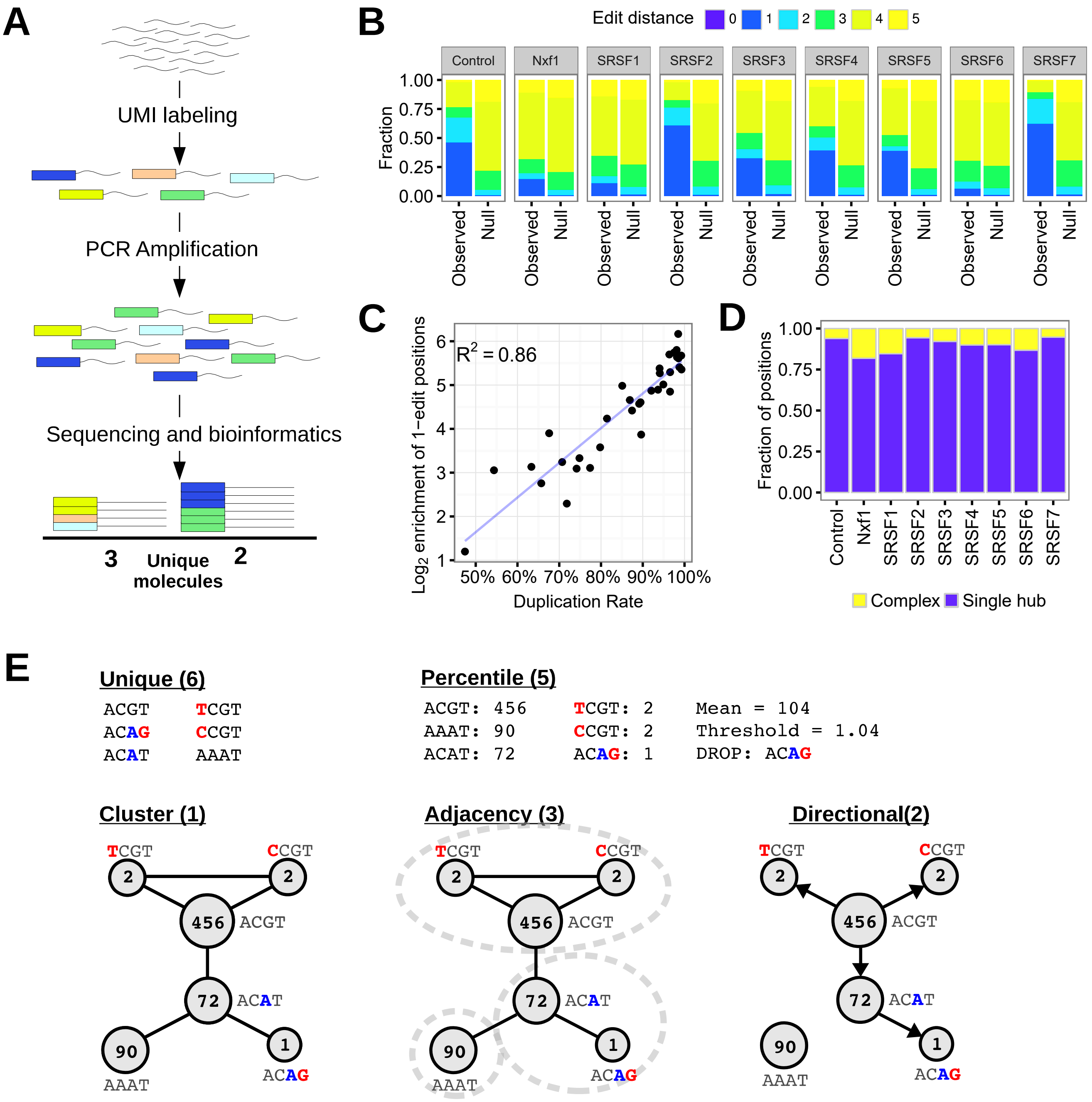
In each panel, all but one of the simulation parameters are held constant, with the remaining parameter varied as shown on the x-axis. **A.** UMI length. **B.** Sequencing depth. **C.** Sequencing error rate. Left plot shows the accuracy of quantification, presented as the log2-transformed normalised difference between the estimate and ground truth. Right plot shows the coefficient of variation (standard deviation / mean). The dashed red line represents the value used for this parameter in all other simulations. The dashed grey line represents perfect accuracy. The *unique* and *percentile* methods give identical results with the parameters shown here and are hence overplotted.

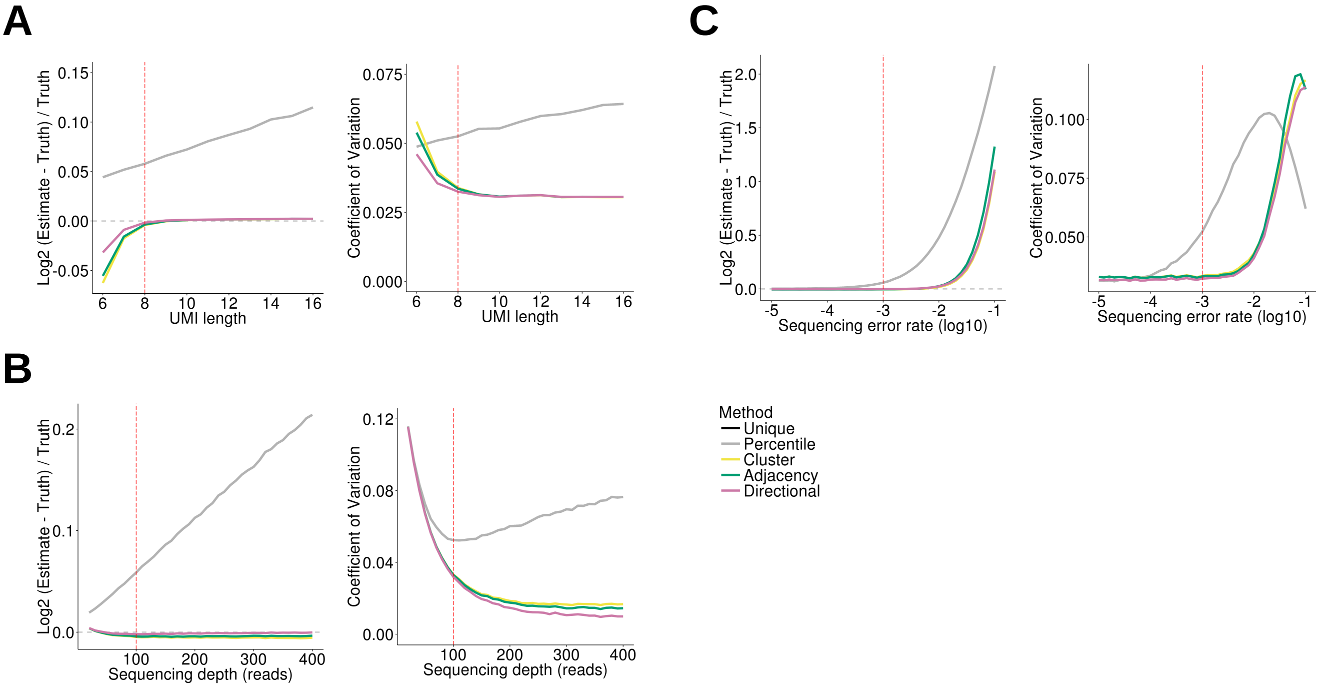
Figure 3. UMI-Tools improves reproducibility between iCLIP replicates

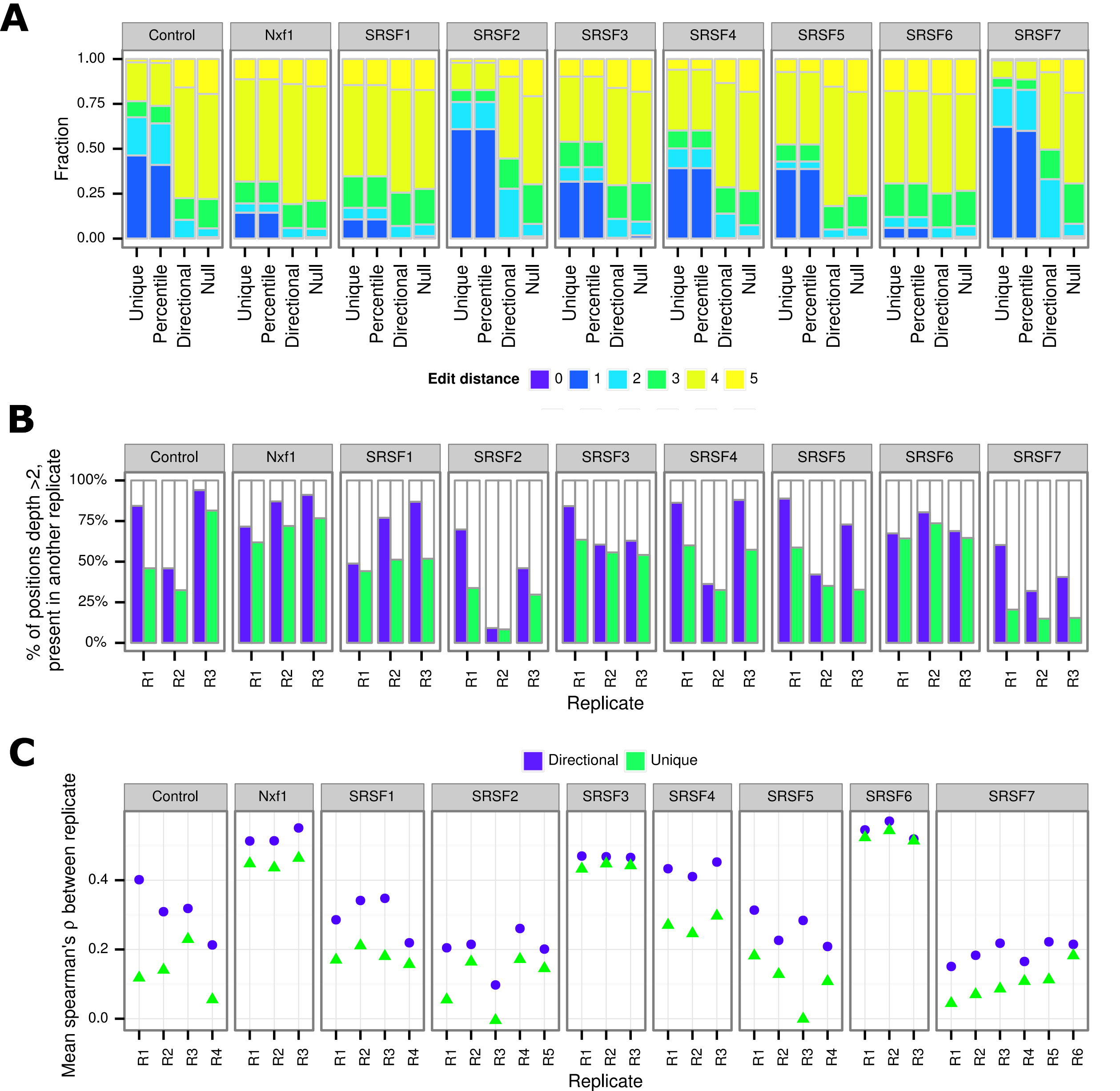
A. Average edit distances between UMIs with the same alignment coordinates. Genomic positions with a single UMI are not shown. Null = Null expectation from random sampling of UMIs, taking into account the genome-wide distribution of UMIs. Only the first replicate of the dataset is shown for each pull down **B.** iCLIP reproducibility as represented by the percentage of positions with >2 tags also cross-linked in at least one of 2 other replicates. **C.** Spearman's rank correlation between the numbers of significant tags in each exon

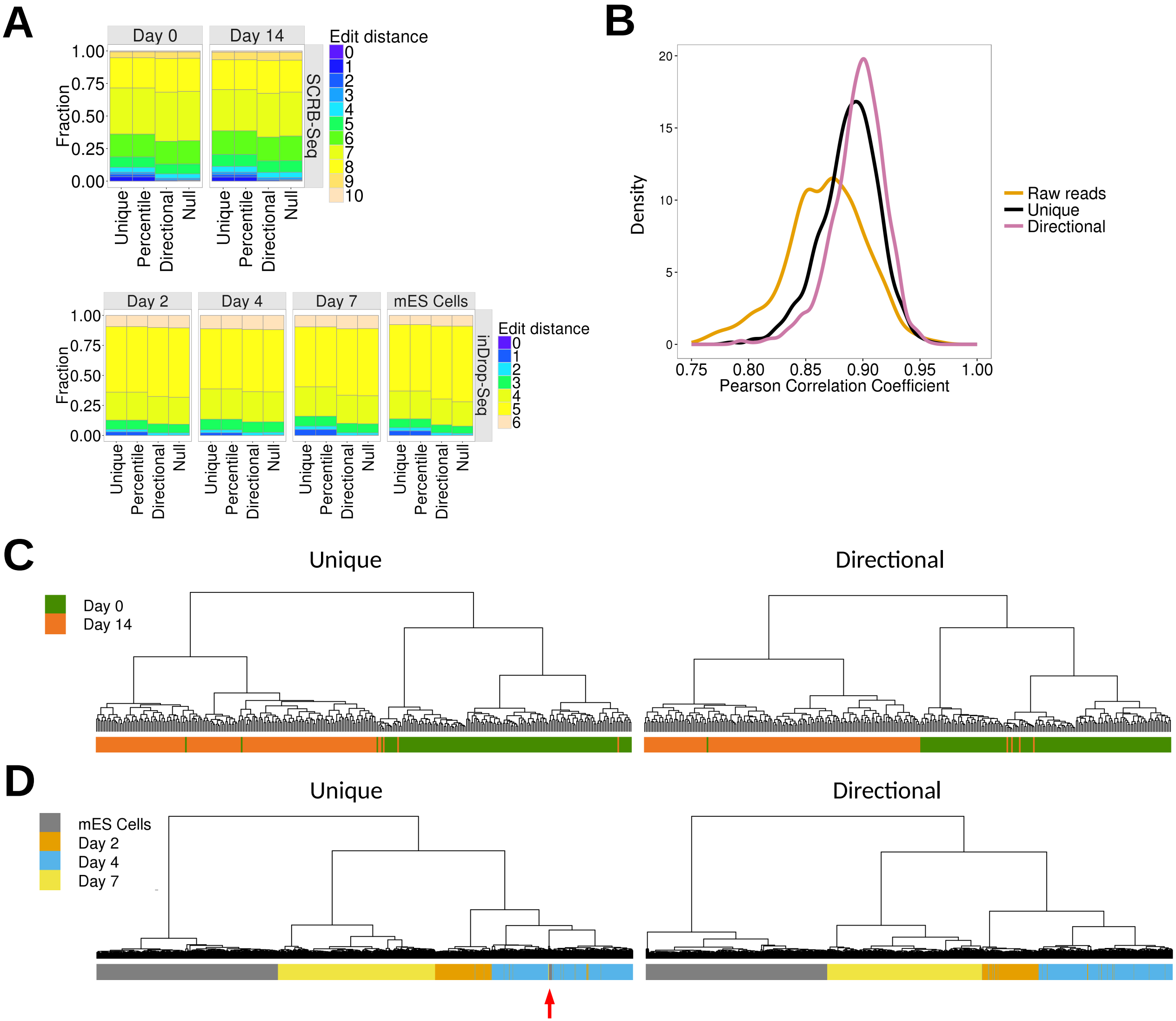
Figure 4. UMI tools improves accuracy and *cluster*ing in Single Cell RNA-seq

A. Average edit distances between UMIs with the same alignment coordinates following removal of PCR duplicates using the methods indicated on the x-axis. Genomic positions with a single UMI are not shown. Null: Null expectation from random sampling of UMIs, taking into account the genome-wide distribution of UMIs. Top = SCRB-seq. Bottom = inDrop-seq. **B**. Distribution of pearson correlation coefficients between log ERCC concentration and log counts for raw reads (UMIs ignored) and *unique* and *directional* methods. C & D. Hierarchical clustering based on the gene expression estimates obtained using *unique* and *directional* Colour bars represent differentiation stage. **C**. SCRB-seq. **D.** inDrop-DSeq. Red arrow indicates mES Cells clustering with Day 4 cells.


A
UMI labeling
PCR Amplification
Sequencing and bioinformatics
3
Unique molecules
2
B
Edit distance 0 1 2 3 4 5
Control
Nxf1
SRSF1
SRSF2
SRSF3
SRSF4
SRSF5
SRSF6
SRSF7
Fraction
1.00
0.75
0.50
0.25
0.00
Observed
Null
C
$R^2 = 0.86$
Log$_2$ enrichment of 1−edit positions
6
5
4
3
2
1
50% 60% 70% 80% 90% 100%
Duplication Rate
D
Fraction of positions
1.00
0.75
0.50
0.25
0.00
Control
Nxf1
SRSF1
SRSF2
SRSF3
SRSF4
SRSF5
SRSF6
SRSF7
Complex
Single hub
E
Unique (6)
ACGT
TCGT
ACAG
CCGT
ACAT
AAAT
Percentile (5)
ACGT: 456
TCGT: 2
Mean = 104
AAAT: 90
CCGT: 2
Threshold = 1.04
ACAT: 72
ACAG: 1
DROP: ACAG
Cluster (1)
Adjacency (3)
Directional(2)
TCGT
CCGT
2
2
456
ACGT
72
ACAT
90
AAAT
1
ACAG


A
Log2 (Estimate - Truth) / Truth
0.15
0.10
0.05
0.00
-0.05
6
8
10
12
14
16
UMI length
Coefficient of Variation
0.075
0.050
0.025
0.000
6
8
10
12
14
16
UMI length
B
Log2 (Estimate - Truth) / Truth
0.2
0.1
0.0
100
200
300
400
Sequencing depth (reads)
Coefficient of Variation
0.12
0.08
0.04
0.00
100
200
300
400
Sequencing depth (reads)
C
Log2 (Estimate - Truth) / Truth
2.0
1.5
1.0
0.5
0.0
-5
-4
-3
-2
-1
Sequencing error rate (log10)
Coefficient of Variation
0.100
0.075
0.050
-5
-4
-3
-2
-1
Sequencing error rate (log10)
Method
Unique
Percentile
Cluster
Adjacency
Directional


A
Control
Nxf1
SRSF1
SRSF2
SRSF3
SRSF4
SRSF5
SRSF6
SRSF7
Fraction
1.00
0.75
0.50
0.25
0.00
Unique
Percentile
Directional
Null
Edit distance
0
1
2
3
4
5
B
% of positions depth >2, present in another replicate
100%
75%
50%
25%
0%
R1
R2
R3
Replicate
C
Directional
Unique
Mean spearman's ρ between replicate
0.4
0.2
0.0
R1
R2
R3
R4
R5
R6
Replicate


A
Day 0
Day 14
SCRB-Seq
1.00
0.75
0.50
0.25
0.00
Fraction
Unique
Percentile
Directional
Null
Edit distance
0
1
2
3
4
5
6
7
8
9
10
Day 2
Day 4
Day 7
mES Cells
inDrop-Seq
Edit distance
0
1
2
3
4
5
6
B
Density
20
15
10
5
0
0.75
0.80
0.85
0.90
0.95
1.00
Pearson Correlation Coefficient
Raw reads
Unique
Directional
C
Unique
Directional
Day 0
Day 14
D
Unique
Directional
mES Cells
Day 2
Day 4
Day 7

CSH PRESS GENOME RESEARCH

UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy

Tom Sean Smith, Andreas Heger and Ian Sudbery

Genome Res. published online January 18, 2017
Access the most recent version at doi:10.1101/gr.209601.116