Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments

Atray Dixit[1,2,3]

[1] Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA 02139, USA
[2] Broad Institute of MIT and Harvard, Cambridge MA 02142 USA
[3] acdixit@broadinstitute.org

Abstract:

As part of the process of preparing scRNA-seq libraries, a diverse template is typically amplified by PCR. During amplification, spurious chimeric molecules can be formed between molecules originating in different cells. While several computational and experimental strategies have been suggested to mitigate the impact of chimeric molecules, they have not been addressed in the context of scRNA-seq experiments. We demonstrate that chimeras become increasingly problematic as samples are sequenced deeply and propose two computational solutions. The first is unsupervised and relies only on cell barcode and UMI information. The second is a supervised approach built on labeled data and a set of molecule specific features. The classifier can accurately identify most of the contaminating molecules in a deeply sequenced species mixing dataset. Code is publicly available at https://github.com/asncd/schimera.

Context Specific Definitions:

Cell barcode (CBC): A random nucleotide barcode present on the reverse transcription primer used to capture RNA that is unique for each single cell in an experiment.

Unique Molecular Identifier (UMI): also known as a Random Molecular Tag (RMT). This refers specifically to a random nucleotide barcode present on the reverse transcription primer. For a given cell (as identified by a CBC), every captured mRNA molecule should be uniquely labeled by a UMI. PCR amplification is exponential and its efficiency depends on initial concentration and various sequence features. A UMI should identify all amplicons that originated from the same transcript. Thus, UMIs can be used for normalization of transcript count and allow correction of PCR bias.

Chimera: also known as a recombinant molecule or hybrid molecule. A chimera is a PCR artifact generated when two template molecules anneal to each other rather than with a primer. This yields molecules that are comprised of a cell barcode and a mRNA from another cell.

Cell barcode UMI tag (CBCUMI): A unique pair of cell barcode and UMI. This pair can potentially be associated with more than one mRNA molecule due to random chance in which UMIs from two distinct original molecules are the same or chimera formation.

Unique Molecule (UM): A unique combination of a CBC, UMI, and an mRNA transcript.

Literature Summary:

(Meyerhans et al. 1990) Demonstrated chimeric PCR amplicons between two distinct HIV *tat* gene sequences could be **reduced from 5.4% to 2% of all molecules by increasing extension time**. Screening phage plaques using gene specific probes identified recombinant sequences.

(Odelberg et al. 1995) **Demonstrated that recombinant molecules can be produced in a single round of extension without subsequent denaturation**. They suggest the polymerase or the nascent strand can switch templates during synthesis. While template switching had been proposed in the context of RNA viruses and certain DNA polymerases, this paper demonstrated that it could occur during PCR with the widely used *Taq*. They used a set of plasmids that contained a constant middle region flanked by 20bp unique sequences. After a single round of primer extension, they performed competitive PCR (whose purpose is the same as qPCR) to demonstrate the presence of recombinant DNA. They report a 4-fold reduction in recombinant DNA when physically separating the forward and reverse strands by streptavidin beads during extension.

(Thompson 2002) Discussed a strategy for measuring heteroduplex formation during PCR of three highly related rDNA from species in the genus *Vibrio* using a capillary electrophoresis strategy. They demonstrate mixed species can constitute a large fraction of the members of the final amplified pool after 30 cycles of PCR. **Introducing a molar excess of primers or reintroducing primers at later stages of PCR mitigated the problem**.

(Ashelford et al. 2005). Suggested up to 5% of 16S rRNA sequences in public databases can be anomalies, most of which are chimeras.

(Smyth et al. 2010) By terminating PCR before the exponential phase of amplification ends, they show a reduction in chimera formation. They also emphasize the importance of polymerase processivity.

(Haas et al. 2011) Compare several chimera identification tools in the context of 16S sequences based on sequence similarity metrics including their own novel one and show a surprisingly large fraction of chimeric reads.

Results:

Computational UMI collapsing can obscure chimeras when library construction is pooled

In several single cell transcriptional profiling methods, library construction across many cells is performed in a single pool using cell-level barcoding to deconvolve transcription profiles during sequencing (Hashimshony et al. 2016; Macosko et al. 2015; Klein et al. 2015; Fan et al. 2015; Zheng et al. 2016; Jaitin et al. 2014). While some rely on in-vitro transcription to perform linear amplification, they all involve at least one step in which the library is amplified via PCR. Additionally, enrichment PCR is sometimes performed, as in (Fan et al. 2015), to selectively amplify a specific subset of molecules of interest from a complex RNA library.

We examine the abundance of PCR chimeras in scRNA-seq libraries (**Figure 1**), and which factors can contribute to their prevalence. We focus on the species mixing experiments (aka "barnyard experiments") in which human and mouse cells are mixed together before being processed (**Figure 2**). These experiments give a direct measurement of the rate at which more than one cell is incorporated into a droplet due to Poisson loading.[1]

[1] Less commonly noted, the correlation between the number of reads mapping to human and mouse for doublets can give an indication of what extent bead/droplet quality is the major factor influencing capture efficiency. (Drop-seq datasets can have an R2 greater than 0.6)

For this paper, we focus on cross-species molecules from a deeply sequenced species mixing experiment in the Drop-seq paper (**Table 1**). These molecules can come from several sources, including: 1) ambient RNA, 2) barcode collisions due to random chance, and 3) chimeric molecules formed during PCR (**Figure 1**).

	Human	Mouse
Cells	50	26
Reads/Cell	773,000	350,000
Transcripts/Cell	54,000	30,000

Table 1: Experimental Summary. Overall cDNA duplication rate (percent of molecules with more than 1 read = 79%)

Expected Chimera Rates	**Human Cell**	**Mouse Cell**
Human Gene	0.63 [n=113,000]	0.15 (n=27,230)
Mouse Gene	0.18 (n=32,014)	0.04 [n=7,200]

Table 2: Expected chimera rates assuming independence (observed numbers of events shown in parentheses, expected number of events shown in brackets).

We find that 1.7% of all unique molecules (UMs) in cells that can be unambiguously assigned to a species come from the opposite species. If, to obtain an upper bound, we assume all of these molecules to come from chimeric events, then, including presumed human-human and mouse-mouse chimeras, up to 5.2% of all molecules considered in this

experiment would be chimeric (based on expected fraction of 33% coming from cross-species products of **Table 2**).

We show that cross-species UMs increase in relative abundance as a function of sequencing depth; they are associated with a lower number of reads per molecule (**Figure 2B,C**).

Paradoxically, sequencing deeply can reduce the statistical power to detect differential expression.

Transcripts per Transcript as a Metric for PCR Chimeras in scRNA-seq

Depending on the diversity of the UMI barcodes used in library preparation, most pairs of cell barcodes and UMIs should be uniquely associated with a single gene. PCR chimeras could cause swapping between cell barcodes (**Figure 2**). However, these events are more likely to happen during later stages of PCR. Resultantly, the chimeric molecules will have a lower number of reads per molecule and are likely to have another gene containing the same cell barcode-UMI pair containing a higher number of reads.

We introduce a notion of transcript per transcript (TPT) as a measure of a chimera formation (**Figure 3**). Specifically, for transcript i in a set of molecules that share a unique pair of CBC and UMI, u, TPT is defined as:

$$TPT_{i,u} = \frac{reads_{i,u}}{\sum_j reads_{j,u}}$$

$$\text{for } \forall\ j\ in\ u$$

In the deeply sequenced Drop-seq library, we were able to increase the power to detect differential expression between human and mouse cells by filtering molecules with a TPT less than 0.02 (**Figure 3C**).

We found 10% of molecules have a TPT less than 0.01 and almost 30% have a TPT less than 0.05 (**Figure 4A**). We observed far more genes per CBC-UMI tag than we would expect based on the diversity of UMI barcodes present (**Figure 4B**).

Theseus provides an unsupervised approach for chimera filtering by performing TPT normalization and removing molecules whose TPT is below a certain threshold.

Yorimasa, a Supervised Approach

If it is possible to perform supervised labeling of cell types that should be distinct, and a subset of genes that should be mutually exclusive between the cell types, then we can perform a supervised chimera filtering approach. For example, in the species mixing experiment, cross-species molecules can be clearly identified. Without species mixing, a set of cell-type specific markers that are known to be very specific can be used instead.

We train a random forest classifier on the species mixing experiment using the following features of each UM (**Figure 4**):

- Log_2(reads per UM); Unique molecules with low number of reads are more likely to be chimeric events.
- TPT; Low TPT is more likely to be a chimera

- Log_2(total mRNA abundance)
- Log_2(total CBC abundance)
- Log_2(Gene Length); Longer genes are expected to be more likely to form chimeras
- GC content; High GC content is also expected to be more likely to form chimeras

The classifiers AUC is 0.93. However, we note that our performance evaluation is slightly contaminated by the rate of human-human and mouse-mouse chimeras that should be present (contributing an unfair number events labeled as "false positives"). Specifically the false positive rate is defined as:

$$FPR = \frac{FP}{FP + TN}$$

There are 3,394,611 molecules that are not cross species chimeras (our labeled set of "negatives"). Of those, we estimate on the order of 120,000 molecules are actually within-species chimeras. As such, if all chimeras are detected, including the within-species chimeras, the FPR will be at least 0.034. A rough approximation of the max AUC is 0.98 (subtracting the triangle defined by an FPR of 0.034).

By setting a threshold on the out-of-bag probability estimates from the random forest classifier and setting the acceptable FPR to 0.034, 4.55% of all molecules are filtered (from before, an estimated 5.2% are chimeras).

In a more shallowly sequenced 10X genomics sample available at http://support.10xgenomics.com/single-cell/datasets/hgmm_1k that was sequenced to a depth of 69,793 reads per cell and an approximate saturation of 24.4% we still noted significant cross-species contamination that could be classified with decent accuracy (**Figure 5A, B**).

Yorimasa is an approach that allows users to specify a set of cells and a set of genes that should not be present in any of the other cells. The labels are used to train a classifier. Molecules confidently predicted as chimeras are removed.

Discussion:

The challenges of PCR chimeras in scRNA-seq include:

- Difficulty in obtaining sensitive detection of medium to lowly expressed transcripts while retaining power to detect differential expression between more highly expressed transcripts (as sequencing depth increases to detect lowly expressed transcripts, chimeras amongst highly expressed transcripts are detected)
- Artificially increased estimates of sequencing saturation if low read abundance chimeras are counted
- When performing enrichment PCR to focus on a subset of the library, there will be an increased prevalence of the chimera problem due to higher sequence similarity between amplified molecules

Potential experimental solutions that would also help with the problem include:

- Longer cell barcodes and UMIs
    - For disambiguating random barcode collisions from true chimeras
- Droplet PCR (**Figure 5C**)
    - If each input molecule is amplified individually, chimera formation is not possible
    - The library should also be more uniform in terms of reads per molecule

Pipelines that process scRNA-seq should begin to incorporate some level of chimeric correction. While certainly not perfect, we hope that our two methods, Yorimasa and Theseus, will help tackle these mythically monstrous mixed molecules in scRNA-seq.

Software Accessibility:

Examples code using Yorimasa and Theseus are at https://github.com/asncd/schimera

Acknowledgements:

The author thanks all members of the Regev Lab for helpful discussions, encouragement, and support, especially Rebecca Herbst for her help with early brainstorming, Aviv Regev for mentorship, and Carl de Boer and Christoph Muus for their insightful feedback on the manuscript. AD was supported by a NDSEG fellowship.

References:

Ashelford, K.E. et al., 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, 71(12), pp.7724–7736.

Fan, H.C., Fu, G.K. & Fodor, S.P. a., 2015. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222), pp.1258367–1258367. Available at: http://www.sciencemag.org/content/347/6222/1258367.abstract.

Haas, B.J. et al., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21(3), pp.494–504.

Hashimshony, T. et al., 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology*, 17, p.77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27121950http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4848782.

Jaitin, D. a. et al., 2014. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, 343(6172), pp.776–779. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1247651 [Accessed February 14, 2014].

Klein, A.M. et al., 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), pp.1187–1201. Available at: http://dx.doi.org/10.1016/j.cell.2015.04.044.

Macosko, E.Z. et al., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), pp.1202–1214. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26000488.

Meyerhans, A., Vartanian, J.P. & Wain-Hobson, S., 1990. DNA recombination during PCR. *Nucleic Acids Research*, 18(7), pp.1687–1691.

Odelberg, S.J. et al., 1995. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. *Nucleic Acids Research*, 23(11), pp.2049–2057.

Smyth, R.P. et al., 2010. Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, 469(1–2), pp.45–51. Available at: http://dx.doi.org/10.1016/j.gene.2010.08.009.

Thompson, J.R., 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by "reconditioning PCR." *Nucleic Acids Research*, 30(9), pp.2083–2088. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/30.9.2083.

Zheng, G.X.Y. et al., 2016. *Massively parallel digital transcriptional profiling of single cells*, Available at: http://biorxiv.org/lookup/doi/10.1101/065912.

Figure 1

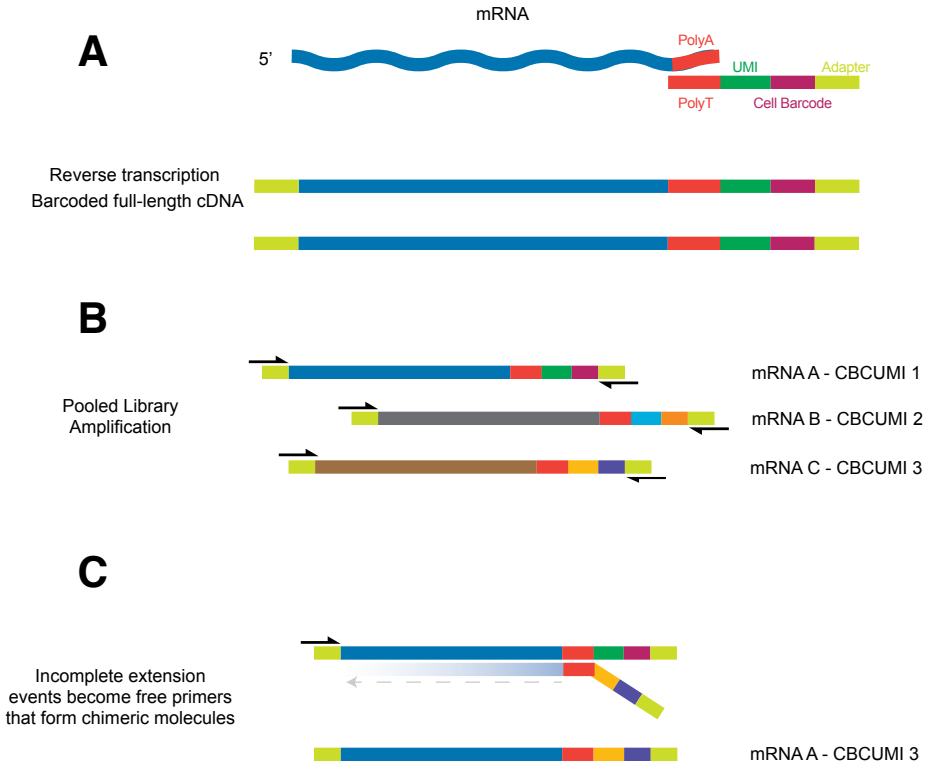
A
mRNA
5'
PolyA
UMI
Adapter
PolyT
Cell Barcode
Reverse transcription
Barcoded full-length cDNA
B
Pooled Library
Amplification
mRNA A - CBCUMI 1
mRNA B - CBCUMI 2
mRNA C - CBCUMI 3
C
Incomplete extension
events become free primers
that form chimeric molecules
mRNA A - CBCUMI 3

Figure 1: The formation of chimeric molecules during scRNA-seq

An overview of a scRNA-seq library construction protocol in which RNA capture and reverse transcription (**A**) is followed by pooled library amplification using PCR (**B**) Incomplete extension events or template switching can form chimeric molecules (**C**)

Figure 2

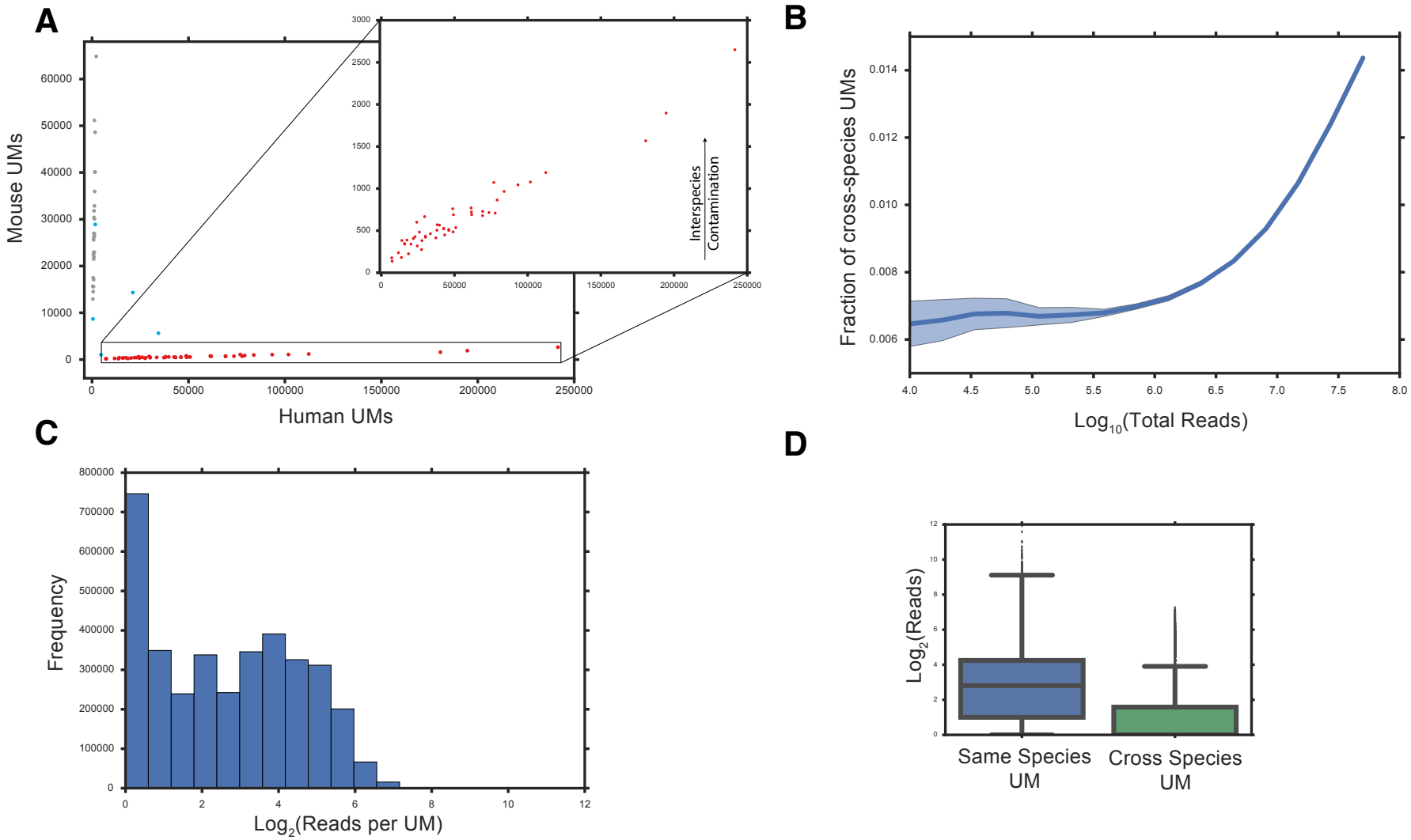
A
Mouse UMs
Human UMs
Interspecies
Contamination
B
Fraction of cross-species UMs
Log10(Total Reads)
C
Frequency
Log2(Reads per UM)
D
Log2(Reads)
Same Species UM
Cross Species UM

Figure 2: Chimeric molecules in the context of cross-species scRNA-seq

(**A**) A scatter plot of cell barcodes and their abundance of Human UMs (x-axis) and Mouse UMs (y-axis). The enlarged subpanel shows the interspecies contamination of mouse UMs is correlated with the abundance of Human UMs. (**B**) Rarefaction analysis shows the fraction of cross-species molecules observed increases with sequencing depth. Shaded error region is 1 s.d from 10 independent subsampling analyses. (**C**) Histogram of $\log_2$(reads per molecule) in this experiment (**D**) Box plots compare reads per molecule for cross species molecules and same species molecules. The box plots denote three quartiles, distribution with whiskers, and outliers as dots.

Figure 3

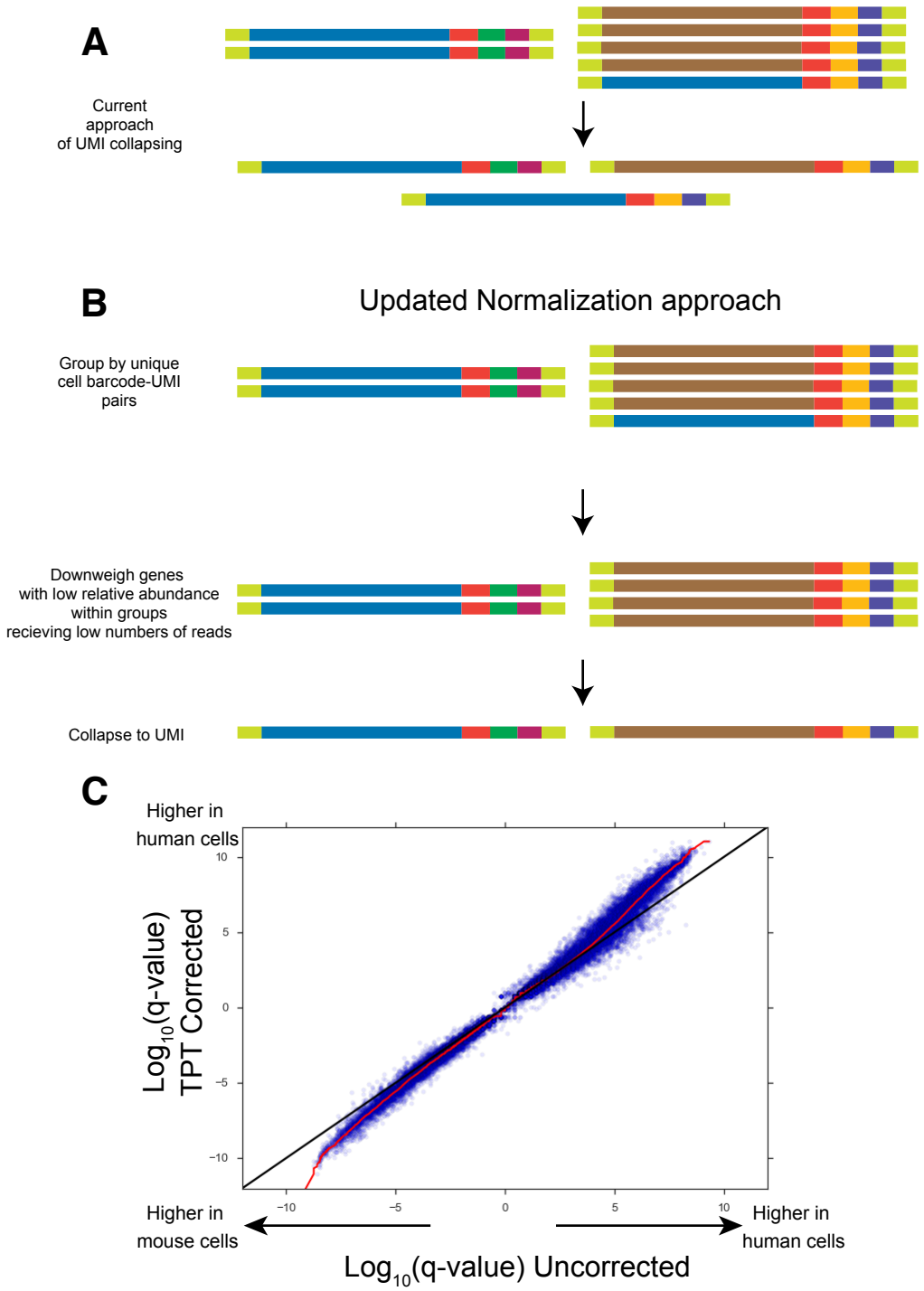
A
Current approach of UMI collapsing
B
Updated Normalization approach
Group by unique cell barcode-UMI pairs
Downweigh genes with low relative abundance within groups recieving low numbers of reads
Collapse to UMI
C
Higher in human cells
Log10(q-value) TPT Corrected
10
5
0
−5
−10
Higher in mouse cells
Higher in human cells
Log10(q-value) Uncorrected

Figure 3: A strategy to correct for chimeric molecules

(**A**) Current scRNA-seq quantification using UMIs collapses molecules into unit counts regardless of overall read abundance. An alternative approach is suggested in which molecules are first grouped by cell barcode-UMI pairs and then molecules that have a low relative abundance are filtered before collapsing (**B**). (**C**) Scatter plot comparing signed $\log_{10}$(q-values) for differential expression between human and mouse cells before and after removing molecules with TPT<0.02. Red line is smoothed lowess fit.

Figure 4

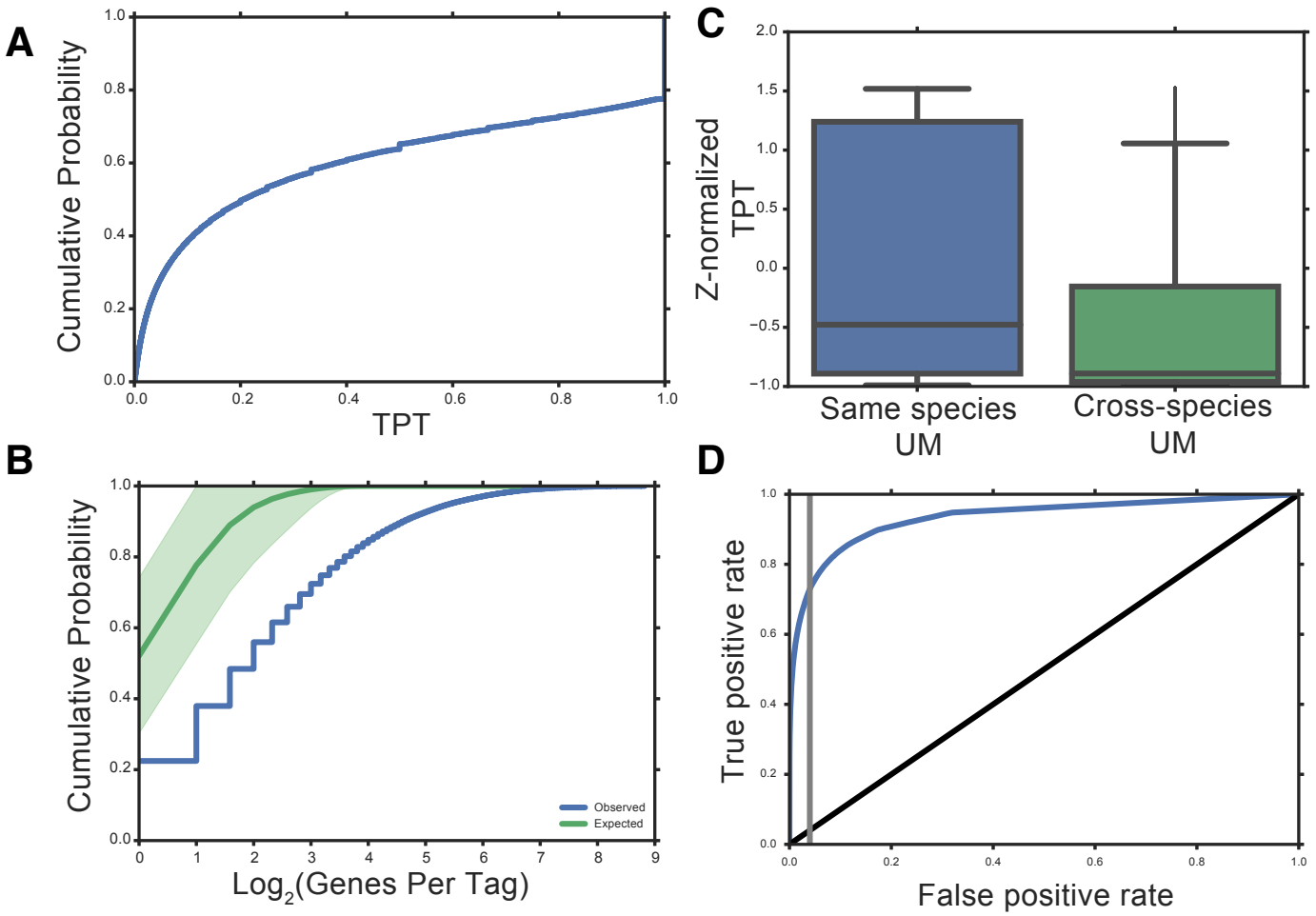
A
Cumulative Probability
TPT
B
Cumulative Probability
Log2(Genes Per Tag)
Observed
Expected
C
Z-normalized TPT
Same species UM
Cross-species UM
D
True positive rate
False positive rate

Figure 4: A supervised classifier applied to cross-species experiments

(**A**) Cumulative distribution plot of Transcripts per Transcript (TPT) in a deeply sequenced Drop-seq experiment. (**B**) Cumulative distribution of the number of genes per unique CBC-UMI tag. Blue line indicates observed distribution. Green line indicates expected distribution based on UMI length and base-pair composition on a per cell basis with 1 s.d shaded in green. (**C**) Boxplots showing how the TPT distribution (Z-normalized within each species) varies with label. (**D**) Receiver operating characteristic curves for a random forest classifier. Gray vertical line is the estimated false positive rate due to mislabeled within-species chimeras.

Figure 5

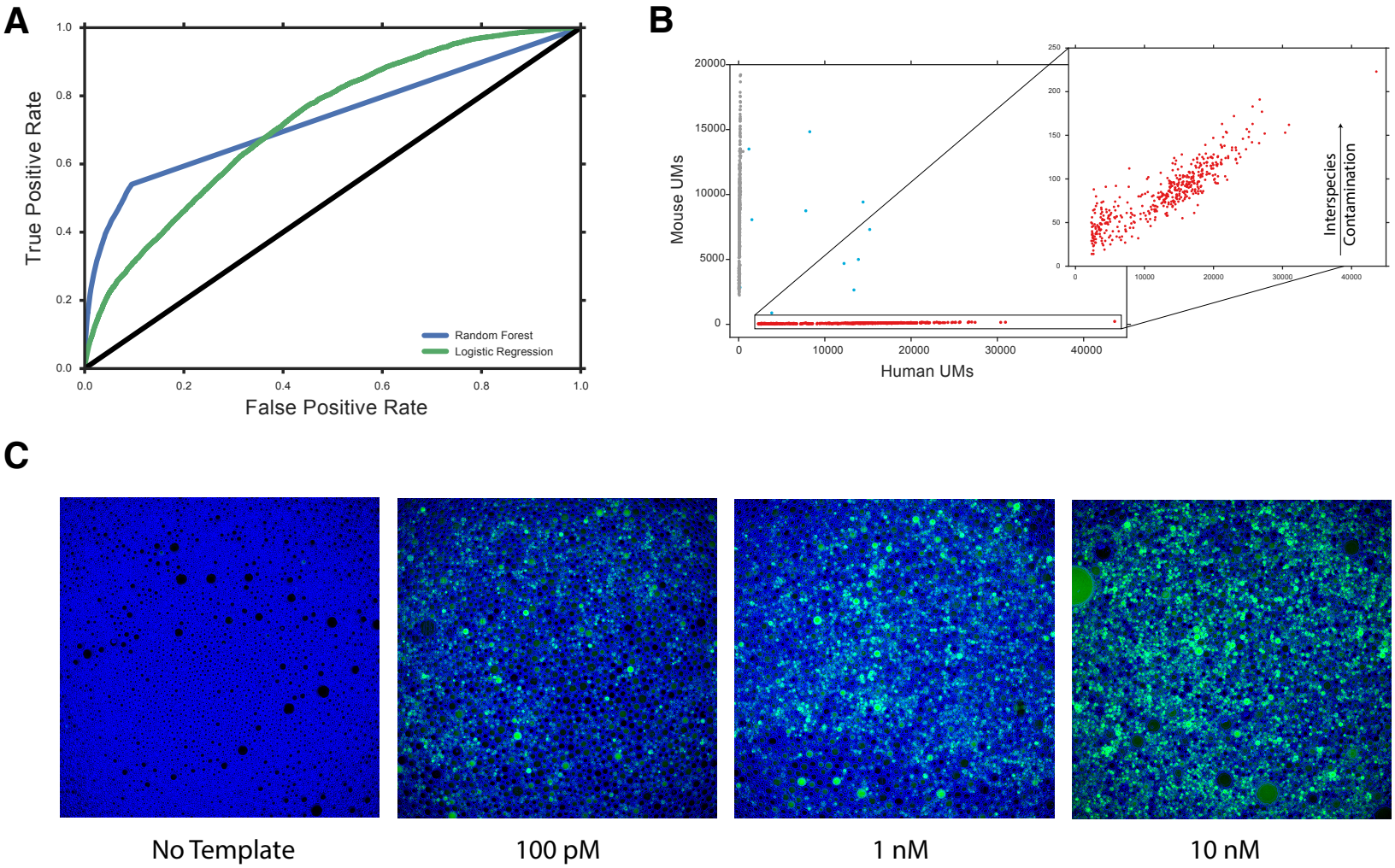
A
True Positive Rate
False Positive Rate
Random Forest
Logistic Regression
B
Mouse UMs
Human UMs
Interspecies Contamination
C
No Template
100 pM
1 nM
10 nM

Figure 5: Future Applications, 10X and droplet PCR

(**A**) Receiver operating characteristic curves for a logistic regression classifier compared to a random forest model. (**B**) A scatter plot of cell barcodes and their abundance of Human UMs (x-axis) and Mouse UMs (y-axis). The enlarged subpanel shows the interspecies contamination of mouse UMs is correlated with the abundance of Human UMs. (**C**) Fluorescent imaging of vortex emulsion PCR droplets containing increasing amounts of a diverse template. SYBR green was added to the PCR mix. Blue outlines were obtained from the bright field image of the droplets.