STOMP-seq: early multiplexing for high-throughput Smart-seq2 RNA-sequencing.

Matthias Eder[1,*], Nicholas Stroustrup[1,2,*]

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain

* Corresponding authors: matthias.eder@crg.eu , nicholas.stroustrup@crg.eu

Abstract

RNA-sequencing provides high-dimensional, quantitative measurements of the states of cells, tissues, organs, and whole organisms. Plate-based RNA-seq protocols allow for a wider range of experimental designs than droplet sequencing methods, but are less scalable due to the practical challenges of plate-based liquid handling. Here, we present STOMP-seq, an RNA-seq method that extends Smart-seq2 to include sample-identifying barcodes on the 5' end of each amplified transcript. These barcodes allow samples to be pooled immediately after reverse transcription, enabling a 12-fold multiplexing strategy that reduces liquid handling complexity and enzyme costs several-fold. Suitable for both manual and robotic library preparation approaches, STOMP-seq reduces protocol execution times four-fold while improving library complexity and coverage compared to Smart-seq2. Together, these advantages combine to make possible new experimental designs, in particular population-scale sequencing projects like the multi-generational study of gene-expression heritability presented here. STOMP-seq offers a "drop-in" replacement for existing protocols, removing practical barriers that limit the quality and scope of transcriptomic data.

Introduction

The measurement of gene expression using RNA Sequencing (RNA-Seq) is a powerful approach for studying the physiologic state of any living system containing mRNA. As a result of recent exponential decreases in the cost of sequencing, RNA-Seq has grown from a specialized tool used by researchers working on gene regulation into a flexible, pervasive method applied to both fundamental research as well as translational and biotechnological applications [1]. Contemporary RNA-Seq protocols fall into three loosely-defined groups: droplet-based methods like 10X or CEL-Seq [2,3], in-situ-based RNA-seq methods like DART-seq or sci-RNA-seq [4,5], and plate-based methods like MARS-Seq or Smart-Seq [6,7]. Plate-based methods are particularly useful in contexts where higher read depths are required, with applications involving complex experimental designs—for example for protocols involving pre-processing of samples using flow-cytometry—or for single-lab, exploratory studies where droplet-based methods are not cost-effective. Among plate-based approaches, SMART (Switching mechanism at the 5′ end of the RNA template) protocols are widely used due to their excellent performance, low cost, and relative ease of implementation [8–10], and a family of related protocols based on the murine leukemia RNA-dependent DNA polymerase have been developed [7,11–14]. Recent developments of the SMART protocol, for example Smart-seq3, introduce additional features like Unique Molecular Identifiers, but increase protocol complexity [11]. Consequently, Smart-seq2 remains widely used across a wide range of applications [15–20]. The flexibility of plate-based methods comes at the cost of necessitating that each sample be handled individually during library preparation. Such challenges can be overcome in several ways, e.g., by merging protocol steps [13]. Here, we explore the potential to reduce library preparation time several-fold by multiplexing biological samples early in the protocol. Multiplexing, or “early pooling” approaches have been implemented successfully for several plate based RNA-seq methods [21–25], but so far none have been developed for the widely used Smart-seq2 protocol.

In Smart-seq2, samples are multiplexed relatively late during library preparation, requiring that each sample be processed separately during reverse transcription (RT), PCR amplification, cDNA cleanup, cDNA "tagmentation", sequencing library amplification, and sequencing library purification. The greatest efficiency gains can be obtained by multiplexing after the first step, reverse transcription, and two potential methods exist for this, involving adding sample identifier barcodes either to the 3' poly-dT primer or to the 5' Template Switching Oligo (TSO). We reasoned that barcoded 5' TSOs should have many intrinsic advantages compared to barcoded 3' poly-dT primers. Any barcode added to the 3' poly-dT primer must necessarily be located before the poly-dT sequence, with the result that 30 bp of every barcoded transcript would be occupied by poly-dTs, wasting sequencing capacity. Additionally, read base quality is drastically lowered after homopolymer stretches, leading to a need for excessive trimming and a reduction in usable reads.

Therefore, to facilitate efficient and high-quality early pooling for Smart-seq2, we developed a new approach we call Stomp (Smart TSO Multiplexed 5-prime)-Seq, that introduces a unique sample barcode to the 5'-end of each transcript to enable early pooling of samples immediately following reverse transcription. Our approach enables a 12-fold multiplexing strategy for all cDNA processing steps. In addition to increasing throughput, STOMP-seq increases transcriptome coverage compared to Smart-seq2 by retaining just two informative reads per transcript while also reducing the 3' bias of the resulting cDNA libraries. Finally, our approach highlights a poorly appreciated feature of the published Smart-seq2 protocol, in which internal binding of TSOs produces substantial rates of poly-A independent reverse transcription initiation—an issue which can be identified and computationally eliminated in data generated using STOMP-seq.

Dramatically reducing processing time and enzyme costs, STOMP-seq provides a "drop-in" upgrade to Smart-seq2 that dramatically speeds up protocol times while slightly improving library quality, enabling new approaches to population-scale RNA-sequencing.

Results

Modified Template Switching Oligos (TSO) introduce 5'-end barcodes

To incorporate useful barcodes to the 5'-end of all cDNA, we first needed a method to ensure that all such barcodes would be propagated through tagmentation into the final sequencing libraries. To this end, we modified the Smart-seq2 TSOs to include a TruSeq sequencing initiation sequence usually added to fragmented cDNA by ligation in the TruSeq library preparation protocol [26]. Adding this TruSeq sequence to the 5' TSO, we provide a means for enabling comprehensive 5'-barcoding later in the protocol: after tagmentation, we can selectively amplify 5' fragments by replacing one of the Nextera indexing PCR primers with an indexing primer that is complementary to the 5' TruSeq sequence (Fig. 1a, S1a). A combination of TruSeq and Nextera primers has been proposed previously [21], however, our strategy produces libraries with the "read2" sequencing initiation sequence at the low diversity barcoded 5'-end of each fragment (Fig. S1b), avoiding problems that arise when commencing sequencing with a low sequence diversity (complexity) in "read1" [27], as produced by previous methods. To compare the performance of our new approach with Smart-seq2, we collected samples from whole *C. elegans* nematodes cultured in standard conditions [28] using either a 5'-barcoded or Smart-seq2 approach. In the resulting libraries, we were surprised to identify a substantial number of transcripts mapping to genomic regions with similarity to a fragment of the TruSeq "read2" primer sequence (Fig. S1c). In fact, 56.5% of 5'-barcoded reads and 5.7% of Smart-seq2 reads contained sequences within a 3-basepair mismatch of the TruSeq primer sequence fragment "CACGTCTGAAC". These results suggested that this TruSeq primer fragment favorably recruits Tn5 enzymes during tagmentation, resulting in the formation of short, sense-less, or even partly unsequencable library fragments by tagmenting inside the "read2" sequencing initiation sequence. Consequentially, 22.7% of paired-end 5'-barcoded reads completely lacked "read2", a phenomenon exacerbated by our strategy over Smart-seq2 due to our

inclusion of this Tn5 bias sequence on 5' TSOs. Therefore, we altered the TSO sequence to include only the last 22 bp of the full 34 bp TruSeq sequence (Fig. 1b, S1c), which reduced the prevalence of TruSeq primer complementary reads to 14.3% and paired end reads lacking "read2" to 8.1%. We thus standardize on this primer design for our STOMP-seq protocol.

Early pooling decreases library preparation time four-fold.

Introducing barcodes to 5' TSOs allows us to pool samples immediately after reverse transcription (RT) and multiplex all subsequent steps including PCR amplification, cDNA cleanup, cDNA tagmentation, sequencing library amplification, and sequencing library cleanup. For example, by pooling twelve samples in a single well after RT, we reduce the "hands-on" time spent actively handling liquids from 3.15 hours using Smart-seq2 down to 40 minutes per plate using STOMP-seq (Tab. 1). For an experiment involving twelve 96 well plates, our approach reduces "hands-off" time, mostly spent waiting for PCRs and incubations to complete, down from 7.8 hours to 3.3 hours per plate. Therefore, in aggregate our approach increases the overall protocol throughput more than 4-fold (Fig. 1c).

STOMP-seq matches or surpasses accuracy and specificity of Smart-seq2

We then assessed the quality of cDNA libraries produced by STOMP-seq. Smart-seq2 is designed to produce full-length cDNAs but often exhibits a 3' bias in read coverage [29]. As to be expected from its 5' TSO barcoding strategy, we find that STOMP-seq exhibits a much higher coverage at the 5'-end of transcripts compared to Smart-seq2 (Fig. 1d). This change in coverage bias compared to Smart-seq2 should not influence transcript counting, and like Smart-seq2 we observe reads mapping to all positions along transcripts.

We then evaluated the complexity of STOMP-seq libraries, comparing the number of unique transcripts identified to Smart-seq2 at equivalent read depths. We find that STOMP-seq detects a higher number of

unique transcripts in libraries of equivalent read depth (Fig. 1e). This improvement results from our 5' TSO strategy, which only one fragment generated in the tagmentation step and therefore samples a larger fraction of all molecules in full-length cDNA libraries to include in final fragmented sequencing libraries compared to Smart-seq2.

We then evaluated the precision of our STOMP-seq transcriptomic protocol, *i.e.*, the ability of the protocol to produce consistent count estimates across multiple technical replicates of the same biological sample. We find identical samples analyzed only with Smart-seq2 or only with STOMP-seq show equivalently high correlations: 98.9% and 98.3% respectively (Fig. 1f). We therefore conclude that technical noise contributes equivalently to the count estimates acquired using either method. Count estimates differed when comparing across sequencing methods: Smart-seq2-derived measurements were correlated 90.2% with STOMP-seq-derived measurements (Fig. 1f), an expected outcome of the differing length biases of the two approaches. While STOMP-seq retains only one fragment per full-length cDNA transcript, Smart-seq2 produces multiple fragments in proportion to the length of the source cDNA. Therefore, Smart-seq2 produces inflated read-counts for longer genes relative to STOMP-seq, explaining the observed differences in library composition between the two methods (Fig. S1d).

A set of 18 TSO barcodes with similar cDNA yields produces balanced libraries

In our early pooling strategy, each sample is processed using a TSO with a unique barcode sequence. This introduces an additional potential source of measurement error over Smart-seq2, which uses identical TSOs for all samples. To quantify any such error, we measured three aliquots from a single biological sample using 21 TSO barcode sequences differing by at least four nucleotides. As a reference, we also amplified the same aliquots using the Smart-seq2 protocol. To exclude cross-talk between barcodes from our measurements, each amplification was performed in a separate reaction using STOMP-seq. First, we considered whether barcodes might influence the cDNA yield of amplification—a

behavior that would lead to unbalanced cDNA pools in which some samples are sequenced at greater read-depth than others. We find that barcode dependent differences in cDNA yield are consistent across STOMP-seq replicates (Fig. 1g), and that the cDNA yield varied by up to ten fold across barcodes. However, by eliminating just 3 barcodes, we could obtain a set of 18 barcodes that fall within a 0.5x range around the mean cDNA yield (Fig S1e).

For STOMP-seq derived reads from these barcodes, we found that ~92% of reads amplified contain identifiable 5' TSO barcodes, with only minor differences between specific barcodes (from 87% to 95% of reads; Fig. 1h), and we chose a set of 12 barcodes with at least 89% assignable reads each.

TSO barcodes produce small but measurable effects on cDNA library composition

We then considered the effect of barcodes on cDNA library composition. As would be expected from 18*3 technical replicates measured from the same sample, most expression variance across our data took the form of unstructured noise, with the largest principal component (PC) derived from PC analysis explaining only 5.38% of all inter-sample variance (Fig. 2a). To identify any influence of barcode sequence, we asked whether the position of samples along each PC correlated with the specific barcodes used for amplification, and discovered only three PCs attributable to barcode biases: PC1, PC2 (4.2%), and PC4 (3.36%) (Fig S2a). These results suggest that 5' TSO barcodes exert a small but measurable effect on library composition—comparable to an additional 10-15% above technical sources of inter-sample noise shared between STORM-seq and Smart-seq2.

Barcodes differ in their rates of internal TSO priming, subtly influencing library composition

To understand the influence of 5' TSO barcodes on library composition, we first considered the barcode effects captured by inter-sample variance along PC1. The weights assigned to genes for PC1 are negatively correlated with the length of those genes' coding sequences (rho_S = -.37; Fig. S2b), suggesting

that barcodes vary in the extent to which they favor shorter transcripts over longer transcripts. Furthermore, the barcodes that favor longer transcripts (low PC1 values) also tend to produce overall higher cDNA yields (Fig. S2c), suggesting that these barcodes generate more cDNA overall in proportion to gene length. Such a phenomenon is reminiscent of the recently described "aberrant TSO priming" activity present in the Smart-seq2 protocol: previously we had found that Smart-seq2 TSOs can bind to non-terminal CCC sites within the coding sequence of specific mRNAs, catalyzing the generation of extra transcripts beyond what would normally be generated by the TSO's expected template-switching activity [30]. Since this TSO priming activity is observed in the standard Smart-seq2 TSO, we hypothesized that our incorporation of TSO barcodes might modulate this mechanism, with different barcodes producing different rates of internal TSO priming. Such differing rates of internal TSO priming would then result in some barcodes favoring longer transcripts, as the probability of internal TSO priming will increase proportionally to transcript length—the probability of a transcript containing a sequence complementary to the TSO barcode will increase in proportion to that transcript's length.

A technical feature of STOMP-seq allows us to systematically measure internal TSO priming rates and test this hypothesis. Since sequencing of every transcript's read2 is initiated by the TSO, we can detect reads derived from internal TSO priming based on their unique signature: a TSO sequence incorporated anti-sense relative to the transcript's cDNA sequence (Fig. S2d diagram). We first considered a highly-expressed gene, *vit-6,* and found peaks in read coverage distributed across the transcript (Fig. 2b). These peaks were formed almost exclusively from reads derived from internal TSO priming (Fig. 2c) rather than cDNA completion by the TSO (Fig. S2e). The rates of internal priming at these peaks vary across barcodes (Fig. 2d): The extend of internal priming at any given position varies with the similarity of the barcode sequences to the transcript's sequence downstream of a potential CCC priming site (Fig. S2f, Tab. S1). Furthermore, the total amount of internal priming across the barcodes also varies between transcripts (Fig. 2e). On average across all barcodes, TSO internal priming is responsible for roughly

3.53% of all reads in our libraries, lower than the 8.71% we find are produced by the Smart-seq2 protocol (Fig. S2g). Across barcodes, the fraction of reads derived from internal TSO priming is strongly correlated (rho_S = -.87) to the first PC of our sequenced libraries (Fig. 2f), demonstrating that differential rates of internal TSO priming generates 5.38% of inter-sample variance—the single largest influence of individual barcode sequences to library composition.

Strand Invasion is the second largest source of barcode bias.

Another potential source of barcode bias is a mechanism often described as "strand invasion" [31], which occurs when the RT enzyme switches templates from mRNA to the TSO prematurely before reaching that mRNA's 5'-end (Fig. S2d diagram)**.** Therefore, we considered that another source of barcode bias might involve some barcodes producing higher rates of strand invasion than others. If such a phenomenon were present, then barcodes with higher rates of strand invasion would generate higher proportions of truncated transcripts, which, after tagmentation, produce reads that map closer to the 3'-end of mRNAs. Looking for a bias in read mapping positions along transcripts, we find that barcodes do differ in their average mapping position (Fig. 2g). Furthermore, this variation is strongly correlated (rho_S = .61) to the second PC of our sequenced libraries (Fig. S2h). Therefore, we conclude that differential rates of strand invasion drive 4.2% of inter-sample variance—the second largest contribution of individual barcodes to library composition.

Barcode biases can be eliminated by barcode selection and permutation across samples

Our analyses of barcode biases then suggested a procedure to minimize their effects. First, we set out to measure any practical effect that they might have on common computational analyses. We asked whether barcode biases were large enough to inflate the rate of false positives in tests for statistically significant differential expression (DE). Such inflation would appear as statistically significant DE

between identical samples, which we can directly identify from our data. Comparing identical samples sequenced using different pairs of barcodes, we observe elevated rates of false-positives (Fig. S2i), with the worst performing barcode pair exhibiting false positive rates of 5.4% using an alpha expected to produce a false positive rate of .01%. Fortunately, the false positive rates were not uniform across barcodes and we chose a set of 12 barcodes in which the average false positive rate is only 1.5%, with the worst-performing pair exhibiting 3.32%.

We then sought to identify additional strategies that might completely eliminate false-positive rate inflation. We reasoned that barcode biases can effect statistical tests only when barcode sequences co-vary with the experimental covariate being considered in the statistical test. Therefore, by permuting barcode sequences across samples, such that the technical replicates of each covariate group do not share barcode sequences, we hypothesized that the influence of barcode bias on downstream statistical analysis might be completely eliminated. This proved to be correct: when identical triplet samples were compared using heterogeneous barcodes, the false positive rate of differentially expressed genes with an adjusted p-value < .01 decreased greatly (Fig. 2h), with a median false-positive rate of zero. Therefore, we conclude that by adopting a careful experimental design that randomizes TSO barcodes across experimental groups, the statistical influence of barcode bias can be essentially eliminated from STOMP-seq.

Eliminating barcode-switching errors

We then explored a final error mechanism that might be introduced by our multiplexed TSO barcoding strategy: "barcode-switching" which occurs when transcripts from one sample inadvertently acquire another sample's barcode during PCR amplification after RT. We proved the possibility for "barcode-switching" in an experiment in which PCR amplification after RT was performed without adding the standard cDNA amplification primers. Without these primers, the PCR reactions nevertheless produced

usable cDNA libraries (Fig. S2j), highlighting the potential for 5' TSO primers remaining from RT to act as primers during PCR, based on their complementary TSO sequence already appended to each cDNA. This reaction is no problem when barcode sequences of the 5' TSOs acting as primers are a perfect sequence match for the TSO sequence of the cDNA. However, it becomes problematic if barcoded TSOs with partial complementarity to the cDNA act as primers, effectively swapping the correct barcode on the cDNA that was added during RT for an incorrect barcode present during PCR amplification.

To measure the prevalence of barcode switching, we performed first strand cDNA amplification of transcripts barcoded with each barcode during RT in the presence of a surplus of 5' TSOs containing each of the other barcode sequences. Each combination of cDNA barcode and 5' TSO barcode mix was performed in a separate well, allowing us to identify barcode switching computationally in the resulting sequencing libraries as transcripts barcoded by another 5' TSO's sequence rather than the original cDNA barcode. We identified some pairs of barcodes that exhibit elevated degrees of barcode switching with rates reaching up to 1.3% of transcripts with swapped barcodes. However, in our chosen set of 12 barcodes, switching rates were less than 0.078%. This switching rate is equal to or lower than the inherent error rate of the underlying illumina sequencing technology [32], suggesting that for our set of 12 primers, the multiplexing effectively introduces no additional switching compared to Smart-seq2.

Prioritizing TSO Barcode Sequences

After systematically identifying sources of bias and error, we arrive at a final set of twelve STOMP-seq TSO barcodes suitable for swap-in replacement in existing Smart-seq2 protocols. Using these barcodes, STOMP-seq produces balanced cDNA libraries with very high rates of successful demultiplexing and low-to-undetectable rates of barcode switching. The specific barcode sequences are provided in table M2.

STOMP-seq early pooling assessment

Having validated our method on identical samples and identified a set of high performing barcodes, we then set out to demonstrate the performance of STOMP-seq to conduct differential gene expression analysis between samples with real biological differences. We collected 48 wild-type and 48 individuals sharing the *glp-1(e2141)* mutation, and processed their transcriptomes using STOMP-seq. Individuals with the *glp-1(e2141)* mutation cultivated at elevated temperatures develop into adults lacking a germline [33]—a large, organismal perturbation to the relative abundance of germline and somatic transcripts [30]. Each individual was lysed and the lysate split into two halves, with one aliquot being processed using the Smart-seq2 protocol. The other aliquot was processed using STOMP-seq in which individuals were combined after reverse transcription into eight pools of 6 wild-type and 6 *glp-1(e2141)* samples each (Fig. S3a). We reasoned that any potential problems in STOMP-seq, including barcode switching or barcode introduced biases, would be visible through a degradation of the ability to detect gene-expression differences between wild-type and *glp-1(e2141)* individuals compared to Smart-seq2. Analyzing transcriptomes collected with both Smart-seq2 and STOMP-seq, we find that the first principal component (PC) in PC analysis shows a clear separation of wild-type and *glp-1(e2141)* individuals along PC1, explaining 44.64% of all sample variance (Fig. 3a). PC2 further separates samples according to the processing protocol, explaining 9.87% of variance. This difference along PC2 is explained by the length bias of Smart-seq2, which amplifies transcripts proportionally to their length unlike STOMP-seq, which does not have this bias (Fig. S3b). Notably, the deviation of individuals of each genotype from their population-average gene expression along PC1 was preserved, highlighting the ability of STOMP-seq to identify not only large genotypic effects but also more subtle inter-individual differences between individuals.

Inter-sample variance, introduced as technical noise by a sequencing method, is important to applications seeking to measure biological variation between samples, for example in aging studies [30].

We therefore estimated the inter-sample variance of each gene across our set of identical aliquots processed with either STOMP-seq or Smart-seq2. Inter-sample variance depends on mean expression of each gene, so we estimated biases between the two methods in respect to gene-expression mean, gene-expression variance, and residual variance that uses ERCC spike-ins to remove technical noise from variance estimates. We saw no bias in gene-expression mean between STOMP-seq and Smart-seq2 (Fig. S3c,d). STOMP-seq showed a slightly elevated inter-sample variance in one out of two conditions, but this occurred for both ERCC and mRNA derived transcripts, and therefore we saw no difference in residual variance between STOMP-seq and Smart-seq2 across either condition. We therefore conclude that STOMP-seq is equivalent to Smart-seq2 in its ability to measure true biological variation between samples.

We then compared the two protocols for their ability to detect genes that are differentially expressed in *glp-1(e2141)* individuals relative to wild-type. The per-gene magnitudes of differential expression were highly correlated between the two sequencing methods (rho_S = .99) (Fig. 3b). Among differentially expressed genes, 91% were identified as significant by both methods, with 4% and 5% being identified only in STOMP-seq or Smart-seq2 respectively (adjusted p-value < .01; Fig. 3c). Since some degree of non-overlap would be expected by chance, we identified genes with strong evidence of being differentially handled by STOMP-seq and Smart-seq2 using regression to estimate a method*genotype interaction term (Fig. 3d). 328 genes, or 5.09% of all genes and 263 genes, or 4.63% of differentially regulated genes showed a statistically significant effect of method on differential expression analysis.

In total, these results demonstrate that STOMP-seq enables early pooling of transcripts without adversely affecting the excellent similar sensitivity, dynamic range, and noise of the widely-used Smart-seq2 protocol.

STOMP-seq enables a multi-generational study of gene-expression inheritance.

Taking advantage of the four-fold increase in throughput of STOMP-seq, we then used the protocol to perform a multi-generational study of gene expression in *C. elegans*, designed to identify heritable gene-expression states passed down from parent to progeny. Substantial work has demonstrated the inter-generation translation of phenotypes in *C. elegans* through a variety of molecular mechanisms [34–40], but the overall contribution of such mechanisms to gene expression remains unclear.

To systematically identify heritable gene-expression states, we collected the transcriptomes of 1045 siblings on the second day of adulthood while taking a subset of 24 for multi-generational analysis. For each of these 24 individuals, immediately before sample collection, we collected every progeny laid in a four-hour period. On average, 43 eggs were laid per parent, leading to a total population of. We let these cousins age until the 2nd day of adulthood and sacrificed them for transcriptomics at the same age as we had sacrificed their parents (Fig. 4a). After processing all these individuals using STORM-seq, we obtained 96 parental and 988 progeny cDNA libraries exceeding our quality thresholds.

We first investigated the overall similarity of transcriptomes between parents and progeny. Performing PCA on the parental and progeny generations separately, we find the first PC of the two generations to be correlated at rho = .71 (Fig. 4b, Fig. S4a-b), suggesting that the largest axis of inter-individual variation is shared across generations. These first PCs highlight a varying ratio of somatic to germline mRNA between individuals—previously shown to be a major source of inter-individual heterogeneity between individuals [30]. The remaining PCs showed varying degrees of correlation between the two generations (Fig. S4a-b), which we interpret as suggesting that co-variation among similar groups of genes contribute un-equally to inter-individual variation across the two generations.

To identify aspects of gene-expression variance resulting from heritable factors, we first searched for genes whose expression was more similar between siblings than between cousins. Using categorical regression, we identified 110 such genes, or about 1% of all genes measured in the experiment (Fig 4c,d; Fig. S4c,d). Remarkably, this gene set contained both *hsp-16.2* and *elo-6*, which are members of a small

group of genes whose protein abundance predicts remaining lifespan within isogenic populations of *C. elegans* [41,42]—suggesting that these genes ability to act as aging biomarkers derive in part from their transduction of heritable factors.

Though we can control for some batch effects (Statistical Methods), our experimental design does not permit us to directly separate between gene-expression patterns driven by parental influence and gene-expression patterns driven by shared environmental factors during development—a problem analogous to the environmental confounds observed in human twin studies [43]. To work around this, we hypothesized that heritable gene-expression states, unlike environmentally-driven gene-expression states, should correlate with aspects of the parental transcriptome. Parents and progeny occupy the same environment only within a four-hour window during the parents' adulthood and the progeny's late embryonic development, and so we reasoned that any progeny gene-expression states correlating with parental states would be enriched for causal parental influences on progeny gene-expression.

To identify such genes, we performed partial least squares (PLS) regression to identify aspects of the parental transcriptome associated with progeny gene-expression (Statistical Methods). This analysis revealed a set of 40 genes whose expression in progeny was significantly associated with gene expression patterns of their parent (Fig. 4e). These 40 genes were partially but not completely correlated with each other across progeny, grouping into at least three distinct co-varying, partially coupled sets(Fig. 4e,f; Fig. S4e,f). Remarkably, the expression of these genes did not appear to be directly heritable—e.g., knowing the expression of a specific gene in the parents did not predict well the expression of that same gene in the progeny (Fig. S4g-h). Rather, heritability appeared to act in *trans*, with the expression of one gene in the parent influencing the expression of a distinct gene in the child. For the 70 genes not showing strong intergenerational correlations, we conclude that either environmental factors shared among sibling groups or aspects of parental physiology not measurable using our transcriptomic method are responsible for gene-expression similarities among siblings.

We then assembled our data into an inter-generational map of heritable gene-expression variance. Using hierarchical clustering, we identify three groups of co-varying genes in the parent - X,Y,Z - that are correlated with three sets of co-varying genes in the progeny - A,B,C. At the group level, X,Y, and Z are highly but incompletely correlated (rho(X,Y) = -.94, rho(X,Z) = .68, rho(Y,Z) = -.74; Fig. 5c), as are A and B (rho(A,B) = -.83; Fig. 5b). The inter-generational correlations are lower than the within-generational correlations (rho(X,A) = .23, rho(X,B) = .24, rho(Y,A) = .3, rho(Y,B)= -.29, rho(Z,A) = -.48, rho(Z,B) = .32; Fig. 5a), as might be expected because the causal relationship between generations must be established during early embryogenesis of the progeny and then persist until adulthood. These gene-expression patterns suggest a model in which a single aspect of organismal physiology in parents involving three inter-related gene-expression groups influence a second, distinct aspect of organismal physiology in progeny that involves two anti-correlated sets of genes (Fig. 5d).

Discussion

Here, we present STOMP-seq, an RNA-sequencing protocol that combines the precision and sensitivity of Smart-seq2 with the increased throughput of early pooling. Central to our approach is a sample-identifying barcode located on the Template Switching Oligo (TSO) used to introduce this barcode to the 5'-end of each transcript. By systematically identifying and eliminating the specific molecular mechanisms reducing the fidelity of reverse transcription, PCR amplification, and tagmentation, we are able to produce a highly-multiplexed protocol with equivalent performance of Smart-seq2. STOMP-seq reduces hand-on time during sample preparation by up to 75% compared to Smart-seq2 while retaining equivalent or better full-length coverage, replicability, and accuracy.

Our methods development highlights a poorly recognized molecular activity inherent to the Smart-seq2 protocol: TSOs can act as primers for reverse transcription, generating 8% of total reads across 89% of

all unique transcripts. This behavior forms a fundamental challenge to any new sequencing protocol development involving 5' TSO barcodes, which we overcome to produce STOMP-seq.

STOMP-seq opens the door to massive plate-based experimental designs, demonstrated here by a multi-generational study of gene expression. Using STOMP-seq, we identify a set of genes whose expression in adults appears to be partially determined by measurable physiologic states in the respective parent. STOMP-seq is amenable to use by hand and by robot, minimizing liquid handling in any context. Importantly, STOMP-seq allows users to multiplex multiple samples across single sequencing barcodes allowing a larger number of samples to be loaded in a single flow cell. Pooling more samples per flow cell in turn allows users to move to larger-capacity flow cells with lower per-read costs. In summary, STOMP-seq is a drop-in replacement for Smart-seq2 that dramatically reduces the processing time and cost of high-quality, plate-based RNA-sequencing.

Acknowledgements

We thank Jochen Hecht and all members of the Dynamics of Living Systems group for discussions and encouragement throughout this project. We thank Joy Alcedo (Wayne State University) for nematode strains. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). We acknowledge support of the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa (CEX2020-001049-S, MCIN/AEI /10.13039/501100011033), the Generalitat de Catalunya through the CERCA programme and to the EMBL partnership. We are grateful to the CRG Core Technologies Programme for their support and assistance in this work. We acknowledge support from the MEIC Excelencia awards BFU2017-88615-P, PID2020-115189GB-I00, and PID2020-115439GB-I00, and support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement

No 852201). Research for this publication has been partially carried out in the Barcelona

Collaboratorium for Modelling and Predictive Biology.

References

1. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
2. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
3. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep.* **2**, 666–673 (2012).
4. Meyer, K. D. DART-seq: an antibody-free method for global m6A detection. *Nat. Methods* **16**, 1275–1280 (2019).
5. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
6. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776–779 (2014).
7. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
8. Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
9. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).
10. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
11. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
12. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1100 (2013).
13. Hahaut, V. *et al.* Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq. *Nat. Biotechnol.* **40**, 1447–1451 (2022).
14. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* **30**, 892–897 (2001).
15. Gong, W. *et al.* Dataset from RNAseq analysis of differential gene expression in germinal vesicle and metaphase I stages of granulosa cells. *Data Brief* **59**, 111409 (2025).
16. Wang, X. *et al.* High-resolution transcriptome datasets during embryogenesis of plant-parasitic nematodes. *Sci. Data* **11**, 690 (2024).
17. Zhao, H. *et al.* Spatiotemporal Proteomic and Transcriptomic Landscape of DAT+ Dopaminergic Neurons Development and Function. *iScience* **0**, (2025).
18. Blocka, J. *et al.* Single-Cell RNA Sequencing of Circulating Myeloma Cells: Clinical Implications in Times of Immunotherapies. *Blood* **144**, 1898 (2024).

19. Ikeda, H. *et al.* Protocol for high-quality single-cell RNA-seq from tissue sections with DRaqL. *STAR Protoc.* **5**, 103050 (2024).
20. Zhu, H.-R., Wei, Y.-B., Guo, J.-Q. & Liu, X.-F. Double-negative T cells with a distinct transcriptomic profile are abundant in the peripheral blood of patients with breast cancer. *Breast Cancer Res. Treat.* **209**, 103–115 (2025).
21. Li, Y. *et al.* Decode-seq: a practical approach to improve differential gene expression analysis. *Genome Biol.* **21**, 66 (2020).
22. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
23. Sholder, G. *et al.* 3'Pool-seq: an optimized cost-efficient and scalable method of whole-transcriptome gene expression profiling. *BMC Genomics* **21**, 64 (2020).
24. Alpern, D. *et al.* BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* **20**, 71 (2019).
25. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* 3236 (2014) doi:10.1101/003236.
26. TruSeq RNA Sample Preparation Guide. https://support.illumina.com/downloads/truseq_rna_sample_preparation_guide_15008136.html.
27. Mitra, A., Skrzypczak, M., Ginalski, K. & Rowicka, M. Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform. *PLOS ONE* **10**, e0120520 (2015).
28. Stiernagle, T. Maintenance of C. elegans. in *WormBook: The Online Review of C. elegans Biology [Internet]* (WormBook, 2006).
29. Guillaumet-Adkins, A. *et al.* Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* **18**, 45 (2017).
30. Eder, M. *et al.* Systematic mapping of organism-scale gene-regulatory networks in aging using population asynchrony. *Cell* **187**, 3919-3935.e19 (2024).
31. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/41/3/e44/2902349.
32. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
33. Kodoyianni, V., Maine, E. M. & Kimble, J. Molecular basis of loss-of-function mutations in the glp-1 gene of Caenorhabditis elegans. *Mol. Biol. Cell* **3**, 1199–1213 (1992).
34. Greer, E. L. *et al.* Transgenerational Epigenetic Inheritance of Longevity in C. elegans. *Nature* **479**, 365–371 (2011).
35. Ashe, A. *et al.* piRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. *Cell* **150**, 88–99 (2012).
36. Rechavi, O. *et al.* Starvation-induced transgenerational inheritance of small RNAs in C. elegans. *Cell* **158**, 277–287 (2014).
37. Houri-Ze'evi, L. *et al.* A Tunable Mechanism Determines the Duration of the Transgenerational Small RNA Inheritance in C. elegans. *Cell* **165**, 88–99 (2016).
38. Posner, R. *et al.* Neuronal Small RNAs Control Behavior Transgenerationally. *Cell* **177**, 1814-1826.e15 (2019).
39. Klosin, A., Casas, E., Hidalgo-Carcedo, C., Vavouri, T. & Lehner, B. Transgenerational transmission of environmental information in C. elegans. *Science* **356**, 320–323 (2017).
40. Deshe, N. *et al.* Inheritance of associative memories and acquired cellular changes in C. elegans. *Nat. Commun.* **14**, 4232 (2023).

41. Kong, W. *et al.* ELO-6 expression predicts longevity in isogenic populations of Caenorhabditis elegans. *Nat. Commun.* **15**, 9470 (2024).
42. Rea, S. L., Wu, D., Cypser, J. R., Vaupel, J. W. & Johnson, T. E. A stress-sensitive reporter predicts longevity in isogenic populations of Caenorhabditis elegans. *Nat. Genet.* **37**, 894–898 (2005).
43. Horwitz, A. V., Videon, T. M., Schmitz, M. F. & Davis, D. Rethinking Twins and Environments: Possible Social Sources for Assumed Genetic Influences in Twin Research. *J. Health Soc. Behav.* **44**, 111–129 (2003).
44. Baeriswyl, S. *et al.* Modulation of aging profiles in isogenic populations of Caenorhabditis elegans by bacteria causing different extrinsic mortality rates. *Biogerontology* **11**, 53–65 (2010).
45. Serra, L. *et al.* Adapting the Smart-seq2 Protocol for Robust Single Worm RNA-seq. *Bio-Protoc.* **8**, 1–20 (2018).
46. Aronesty, E. Comparison of Sequencing Utility Programs. **7**, (2013).
47. STAR: ultrafast universal RNA-seq aligner | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/29/1/15/272537.
48. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. Chen, Y., Chen, L., Lun, A. T. L., Baldoni, P. L. & Smyth, G. K. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Res.* **53**, gkaf018 (2025).
50. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

Experimental Methods

Animal handling—Transcriptomic samples in this study were obtained from the nematode model organism *C. elegans*. Two strains were used, QZ0: the wild type Bristol strain N2, and CB4037: the mutant line *glp-1(e2141)*. This mutation causes worms that are cultivated at 25 °C to develop without a germ-line and therefore to become sterile. Wild type and fertile *glp-1(e2141)* animals were cultivated at 20 °C. All lines were maintained on NGM agar plates seeded with *E. coli* NEC937 B (OP50 ΔuvrA; KanR) [44]. Populations were synchronized by egg-laying. For transcriptomics, animals were lysed according to the protocol described by Serra *et al.* [45], with the addition of ERCC spike-ins to the lysis buffer to reach a final dilution of 1:40000. Animals were individually picked into lysis buffer, 15 animals in 120 µL buffer for generating population sized samples for protocol development, and 1 animal in 8 µL of buffer for individual transcriptomics.

mRNA sequencing—To decrease sample preparation time and cost of the Smart-seq2 RNA-seq protocol, we adapted key components of the protocol to include sample identifier barcodes so that samples can be pooled and commonly processed. To obtain barcoded cDNA libraries, 4 µL of raw lysates were incubated with 4 µL of Smart-seq2 annealing mix for 3 min at 72 °C and immediately put on ice afterwards. Then, 12 µL reverse transcription mix were added that has the same composition as the Smart-seq2 reverse transcription mix but contains different barcoded STOMP-seq TSOs per sample. After running the Smart-seq2 reverse transcription protocol on a thermocycler, differently barcoded first strand cDNA samples were pooled together. 20 µL of each pool were mixed with 30 µL IS PCR amplification mix that has the same composition as the Smart-seq2 IS PCR amplification mix, except that the Smart-seq2 IS primer and the STOMP-seq IS primer are present both in a concentration of 83.3 nM. Amplified cDNA was then purified with in-house prepared SPRI paramagnetic beads that mimic AMPure XP beads (Beckman Coulter), with a bead to sample ratio of 0.8. cDNA library size distributions were

examined using a Tapestation 4150 (Agilent) and the library concentration was measured using the Quant-it DNA determination kit (Invitrogen) on a plate reader (Tecan). cDNA libraries were then tagmented using the Nextera library preparation protocol (Illumina). From tagmented sequencing libraries, barcode containing 5' end fragments are then selectively amplified with custom STOMP-seq N7 indexing primers (Tab. M1) that complement the incomplete TruSeq sequencing initiation sequence that had been added by the STOMP-seq TSO. Common Nextera S5 indexing primers are used as second part of the primer pair. To suppress the amount of non-sequenceable S5-S5 library fragments, the first gap-filling step of the Nextera library amplification PCR can be omitted. N7-S5 fragments are still being amplified since they contain one complete strand. Amplified sequencing libraries were then purified twice with in-house prepared SPRI paramagnetic beads that mimic AMPure XP beads (Beckman Coulter), with a bead to sample ratio of 0.9. Sequencing library size distributions were examined using a Tapestation 4150 (Agilent) and the library concentration was measured using the Quant-it DNA determination kit (Invitrogen) on a plate reader (Tecan). Sequencing libraries were pooled in equal masses and library pools were sequenced on Illumina sequencers NextSeq500 or NovaSeq6000, with paired end reads of length 38 bp and 50 bp respectively.

Table M1: Primers and their sequences introduced in this study.

Primer name	Primer sequence (5' – 3')
STOMP-seq barcoded TSOs	CAGACGTGTGCTCTTCCGATCT[bc_6]rGrG+G
STOMP-seq IS primer	CAGACGTGTGCTCTTCCGATCT
STOMP-seq N7XX indexing primer	CAAGCAGAAGACGGCATACGAGAT[$index_8$]GTGACTGGAGTTCAGACGTGTGCTCTTCCG
N7_TSO_fish	CAAGCAGAAGACGGCATACGAGATCCGGCAATGTGACTGGAGTTCAGACGTGTGCTCTTCC GATCTCAACGCAGAGTACATGGG

Primers used from Smart-seq2 and Nextera library preparation protocols: Smart-seq2 oligo-dT, Smart-seq2 TSO, Smart-seq2 IS primer, Nextera N7XX indexing primer, Nextera S5XX indexing primer.

Table M2: STOMP-seq TSO barcode sequences evaluated in this study. The 12 best-performing barcodes are highlighted with asterisks.

Barcode name	Barcode sequence
bc1	AACGGA
bc2*	AAGAAC
bc3	AATCCT
bc4*	ACCATT
bc5*	ACTTAG
bc6	AGAGAT
bc7	AGCTCC
bc8*	AGGCTG
bc9*	CCGCGA
bc10	CGGACT
bc11	CTAATG
bc12*	CTCTAT
bc13*	CTTGCC
bc14*	GAAGTC
bc15*	GCATCT
bc16*	GGCAAG
bc17	GGTCGC
bc18*	TCGTTC
bc19	TCTGGT
bc20*	TTCAGC
bc21	TTGGAG
bcSS2	GTACAT

Transcriptome analyses—Pooled RNA-seq reads were split by barcode using fastq-multx version 1.4.2 [46], removing the barcode sequence and GGG overhang in the process. Split reads were aligned to the *C. elegans* Wormbase reference genome (release WS265) using STAR version 2.6.0c [47]. The reference genome had been modified to include ERCC spike-in sequences. Gene counts were quantified from alignments using featureCounts version 2.0.0 [48]. The count matrix was subject to filtering for low abundant transcripts. For protocol development and evaluation, transcripts were filtered using *filterByExpr* from the R/edgeR package version 4.2.2 [49]. For the transgenerational study, a detection threshold of 5 counts in at least 75% of the samples was applied. Sample counts were normalized using normalization factors estimated from the filtered count matrix using scran version 1.32.0 [50]. Genome coverage of reads was computed using BEDTools version 2.29.1 [51].

Statistical Methods

Principal component analysis was performed using R's prcomp implementation on log counts with 0.5 pseudo counts added. Counts were centered and scaled. Hierarchical clustering of samples was performed using Spearman's correlation and the Ward algorithm. If not done by the respective algorithms themselves, p-values were adjusted for multiple testing using Benjamini-Hochberg.

The parental influence of progeny gene-expression was explored in a search for genes whose expression was more similar between siblings than between cousins. The expression of each gene among all progeny was analyzed using the log-linear regression model $\log y_i = \beta_i^E X_i^E + \beta_i^L X_i^L + \beta_i^S X_i^S + \varepsilon_i$ with y_i as the expression of gene i, two vectors β_i^E and β_i^L estimating the effect of batch effects of environmental factors and RNA-sequencing library preparation introduced by batches X_i^E and X_i^S respectively. A final categorical variable β_i^S estimating the shared influence of a parent across all of its progeny (*i.e.* among all members of each sibling group X_i^S). A shared parental of progeny's gene expression is indicated when the values of β_i^S show statistically distinct values between sibling groups X_i^S. Model parameters were estimated separately for each gene i using maximum likelihood estimation using the *glm* package in *R*.

Partial r-squared estimates comparing the influence of batch effects and parental contributions were obtained by comparing the full model to the partial models $\log y_i = \beta_{handling} X_i^H + \beta_{library} X_i^L + \varepsilon_i$ and $\log y_i = \beta_{sibling} X_i^S + \varepsilon_i$ respectively using the *rsq* package in *R*. Sibling group expressions, normalized to exclude batch effects, as plotted in Fig. 4d, were calculated from the full model's residuals as $\beta_{sibling} X_i^S + \varepsilon_i$.

Parental genes whose expression predict the expression of any gene in their progeny were identified using Partial Least Squares regression. Latent Variables were identified by decomposing the matrix of parental gene expression y into a set of 5 components, and these variables were then used to predict

the expression of a single gene's expression among progeny. Each gene in the progeny was analyzed in a separate partial least squares regression using the *pls* package in *R*. The magnitude of PLS loadings obtained from analyzing shuffled data was estimated by first assigning all individuals in a sibling group to a random parent and then re-running PLS on data permutated to match these new parent-sibling pairings. Hierarchical clustering of parental and progeny gene-expression was performed using *hclust* in *R*.

R scripts were run on R version 4.1 and RStudio version 2024.09.0.

Figures

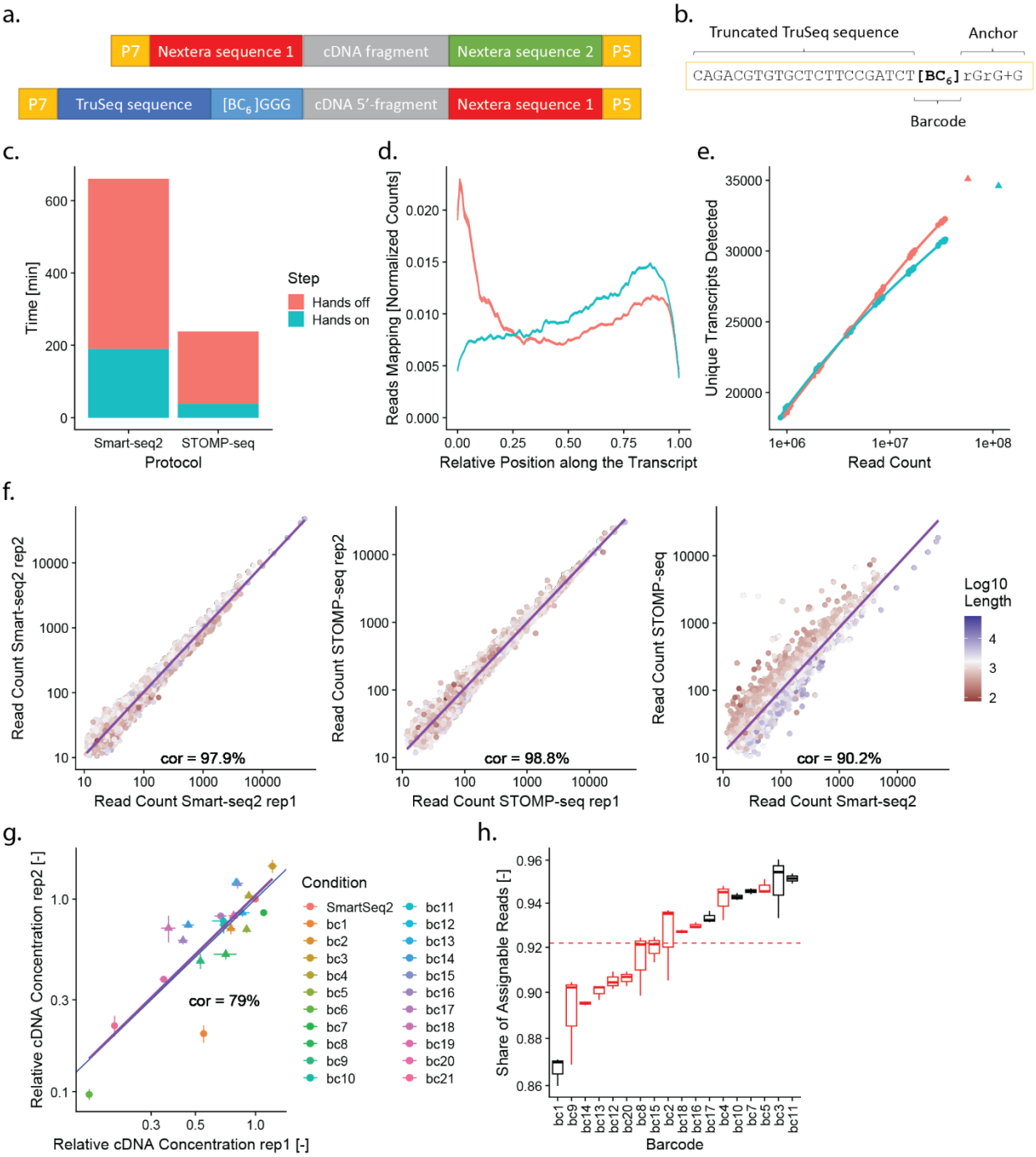


Figure 1—The STOMP-seq protocol enables a fast early-multiplexing strategy. a. Compared to Smart-seq2 (*top*), the STOMP-seq protocol generates cDNA fragments (*bottom*) with an extra 5' barcode. **b.** The STOMP-seq template-switching oligomer incorporates a truncated TruSeq read2 sequence, a six-nucleotide sample barcode, and the Smart-seq2 rGrG+G anchor containing two RNA nucleotides plus a locked nucleotide that increase affinity to first strand cDNA during template switching. **c.** The execution time of STOMP-seq and Smart-seq2 spent on active liquid handling (*blue*) and passive incubation times (*red*). **d.** Average read mapping location bias along the thousand most abundant transcripts in STOMP-seq (*red*) and Smart-seq2 (*blue*) libraries. **e.** The complexity of STOMP-seq (*red*) libraries compared to

Smart-seq2 (*blue*), measured as the number of unique transcripts detected across a range of library sizes. **f.** The abundance of genes is consistent across two replicates of Smart-seq2 (*left*) and STOMP-seq (*middle*), with linear fit (*purple*). The average abundance of genes across 48 replicates processed with either Smart-seq2 or STOMP-seq (*right*). Each gene is colored by its coding sequence length, highlighting any bias. **g.** The effect of different STOMP-seq TSO barcodes and the Smart-seq2 TSO (*colors*) on cDNA yield across three technical replicates (*error bars*) of two biological replicates (*x and y axis*), with linear fit (*purple*). **h.** The effect of STOMP-seq TSO barcodes on the fraction of reads containing an identifiable barcode, assessed in triplicate. Red dotted line indicates average across all barcodes, with the final twelve STOMP-seq barcodes chosen highlighted in red.

		Smart-seq2 - per plate		STOMP-seq - 12 plates pooled, per plate	
Step	Reaction	"Hands-on" Time [min]	"Hands-off" Time [min]	"Hands-on" Time [min]	"Hands-off "Time [min]
1	**RT Buffer Preparation**	10.0	10.0	10.0	10.0
2	**Lysis**	*Sample-specific*	*Sample-specific*	*Sample-specific*	*Sample-specific*
3	**Annealing**	10.0	5.0	10.0	5.0
4	**RT**	5.0	160.0	5.0	160.0
5	**IS PCR**	10.0	110.0	0.8	9.2
6	**cDNA clean up**	15.0	45.0	1.3	3.8
7	**cDNA concentration**	15.0	5.0	1.3	0.4
8	**cDNA dilution**	30.0	0.0	2.5	0.0
9	**Tagmentation**	10.0	10.0	0.8	0.8
10	**library PCR**	10.0	30.0	0.8	2.5
11	**library clean up**	30.0	90.0	2.5	7.5
12	**library concentration**	15.0	5.0	1.3	0.4
13	**library pooling**	30.0	0.0	2.5	0.0
	Total	***190.0***	***470.0***	***38.8***	***199.6***

Table 1—STOMP-seq protocol execution time compared to Smart-seq2, including reverse transcription (RT) cDNA amplification (IS) PCR, and sequencing library preparation (library) steps.

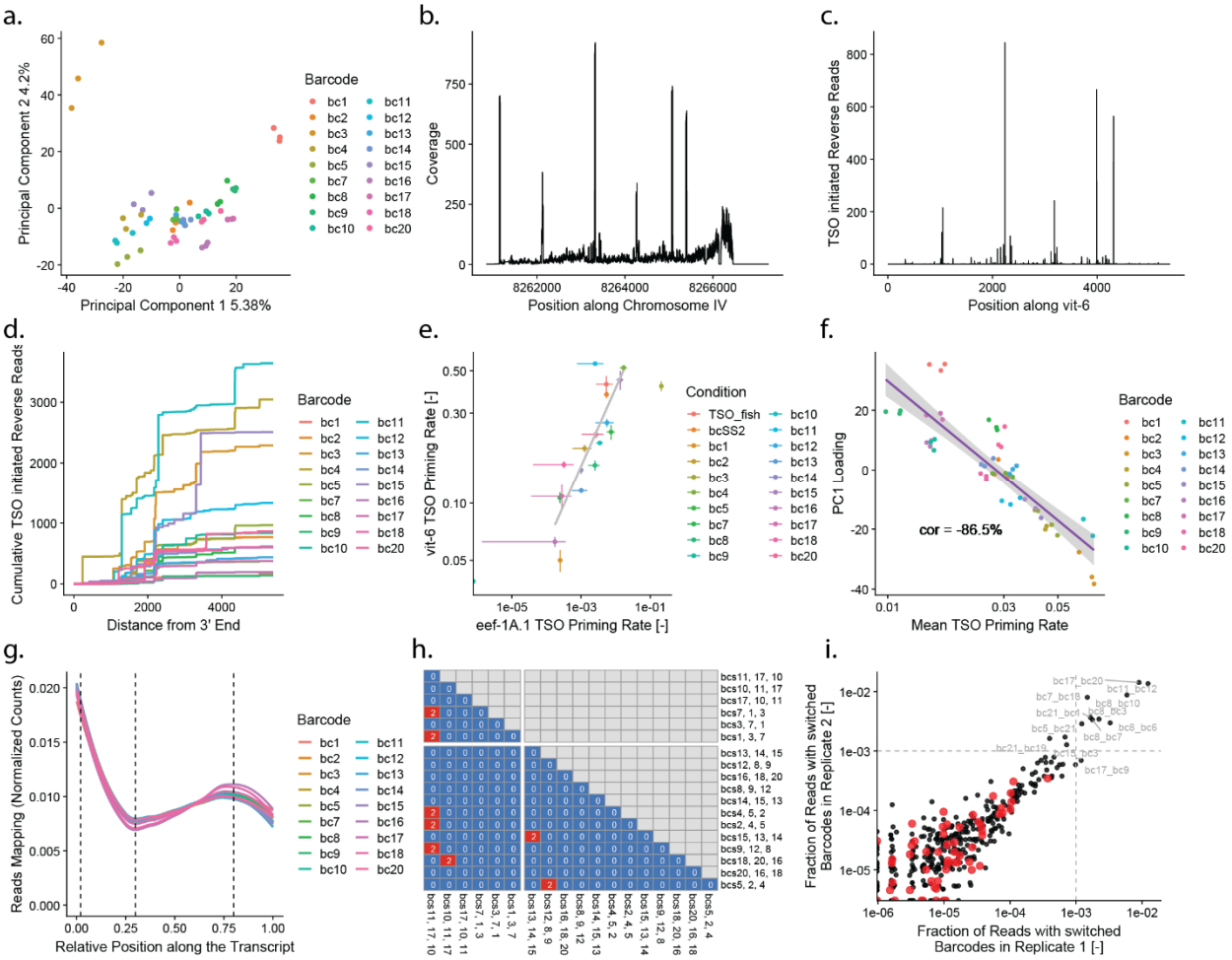


Figure 2—The effects of TSO barcode sequence on STOMP-seq cDNA libraries. a. Three replicates were performed using each of 18 different TSO barcode sequences on identical aliquots taken from the same biological sample. Principal component analysis identifies two main axes of technical noise contributing 5.4% and 2.4% of all inter-sample variation, respectively. **b.** The coverage of reads from cDNA sequences adjacent to the Smart-seq2 TSO, mapping to the highly-expressed gene *vit-6*. **c.** The coverage of reads from cDNA sequences adjacent to the Smart-seq2 TSO, resulting from aberrant TSO priming along *vit-6*. **d.** The cumulative number of aberrant TSO priming reads produced by each barcode *(colors)* along *vit-6*. **e.** The fraction of reads derived from aberrant TSO priming compared for two of the most highly-expressed genes, *eef-1A.1* and *vit-6*, measured across all TSO barcodes in triplicate technical replicates, with standard error bars. **f.** The contribution of each gene to PC1, *i.e.* the PC1 loadings, are compared to the rate of TSO priming for that gene when amplified with different TSO barcode sequences (*colors*). Linear fit in purple. **g.** The relative read coverage averaged along the 1000 most highly-expressed genes, in libraries generated using different TSO barcode sequences (*colors*). Dotted lines indicate 5', middle, and 3' positions used to calculate coverage ratios. **h.** Heatmap of false positive differentially expressed genes (p<.01) per 10,000 transcripts, identified comparing identical triplicate samples amplified using three different TSO barcode sequences each. Comparisons between the final 12 chosen STOMP-seq barcode sequences are grouped together on the bottom right. **i.** The rates of barcode swapping between each pair of barcodes are compared across two replicates. Each point shows the rate of

swapping from the correct cDNA barcode to a specific "invasive" TSO barcode. Pairs of the final 12 chosen STOMP-seq barcode sequences are highlighted in red.

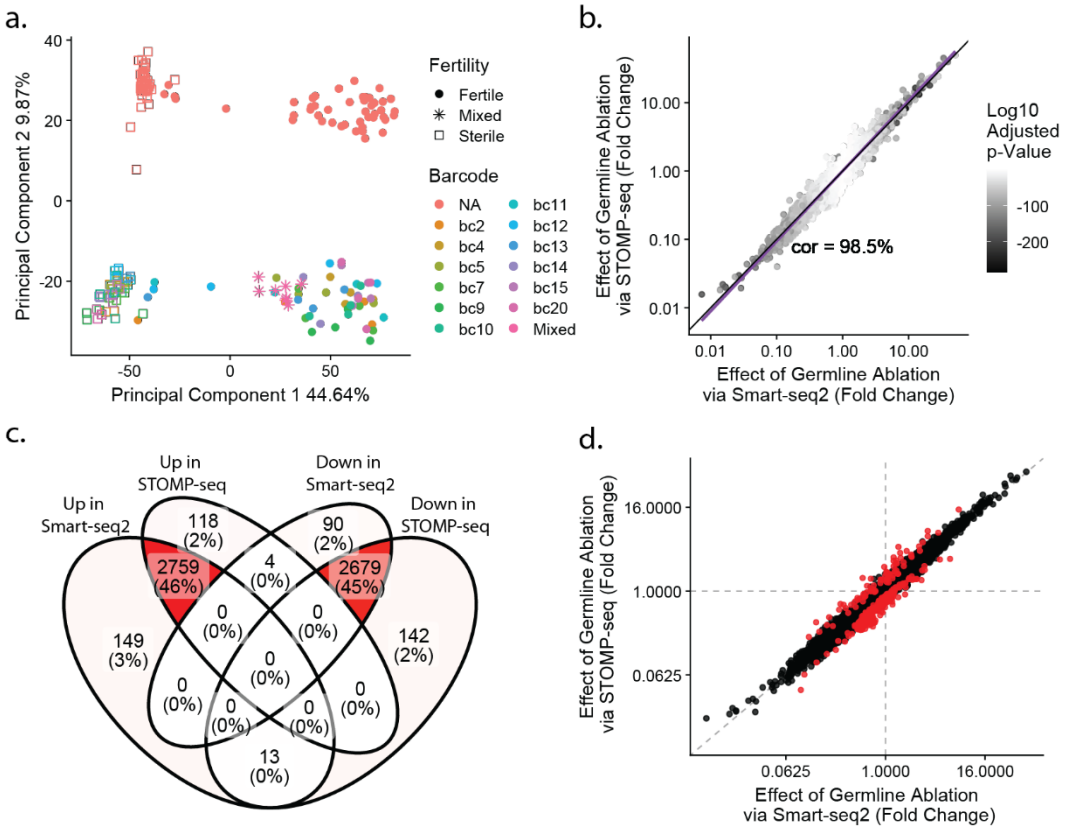


Figure 3—STOMP-seq accurately quantifies biological differences between samples. a. 48 Intact (*solid circle*) and 48 germline-ablated *glp-1(e2141)* (*open square*) individuals were individually lysed and then either processed separately using Smart-seq2 or pooled and then processed together using STOMP-seq. Samples processed with either Smart-seq2 (*red*) or STOMP-seq using different barcoded TSOs (*colors*), are compared across the first to principal components. Multiplexed samples (*star*) represent pooled intact and germline-ablated animals before demultiplexing. **b.** The magnitude of differential expression of each gene between intact and germline-ablated populations, as measured either using Smart-seq2 or STOMP-seq. Points shaded by p-value, with the linear fit in purple. **c.** The overlap in differentially-expressed genes identified by the two methods, shown as Venn diagram. **d.** The effect of germline ablation on each gene's expression, measured by Smart-Seq2 (x-axis) and STOMP-seq (y axis). Germline effects with significant differences in magnitude between Smart-Seq2 and STOMP-seq at p<.01 are highlighted red.

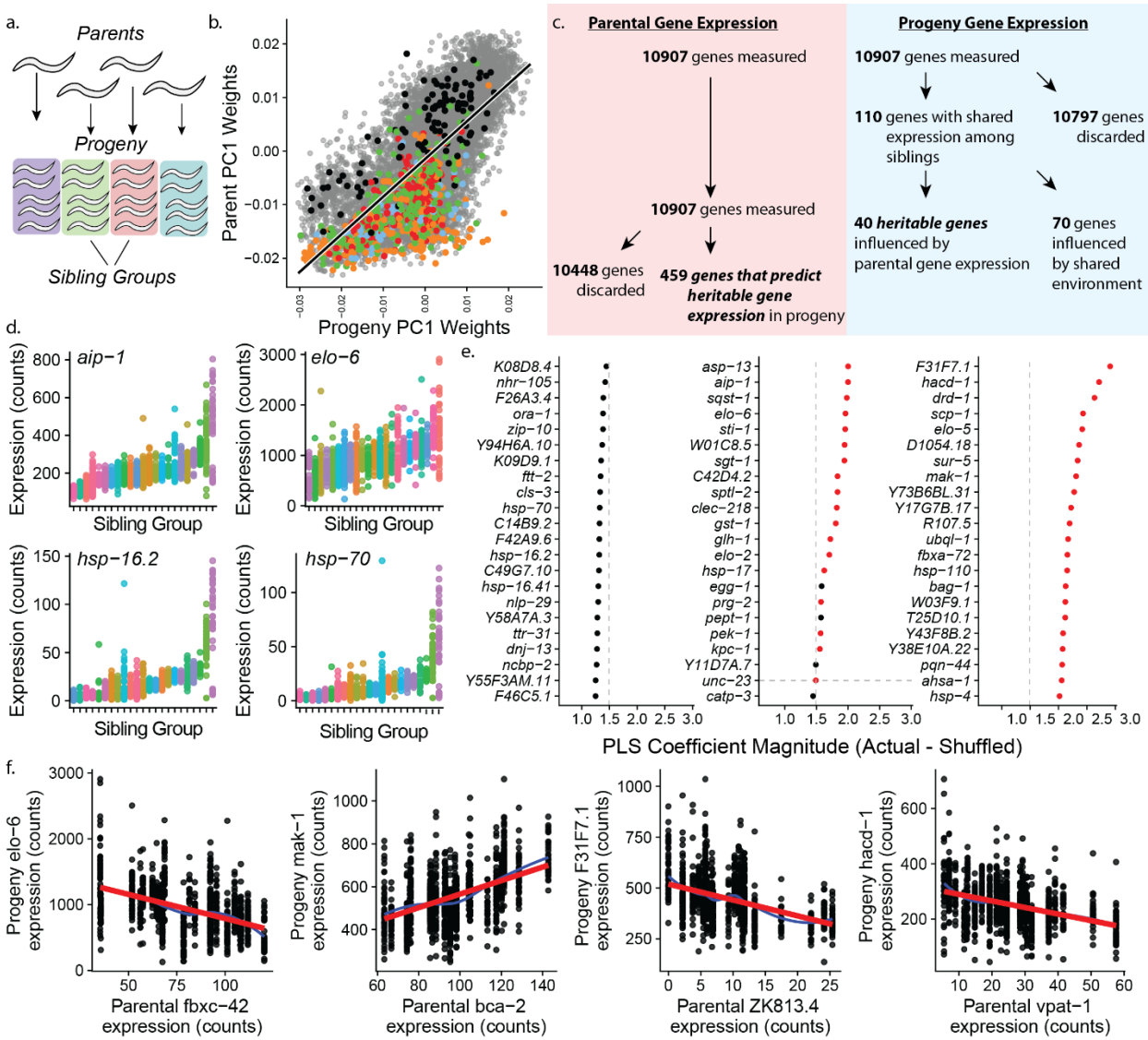


Figure 4—Using population-scale STOMP-seq to identify heritable gene-expression states. a. The experimental design, in which the transcriptomes of 96 parents and the 988 progeny of 24 of the parents were collected. **b.** The weights (loadings) of the first principal component (PC) of all progeny compared to the first PC of the parents. Genes (points) are colored by their tissue-expression pattern: germline (*black*), neurons (*red*), muscle (*orange*), intestine (*blue*), hypodermis (*green*), and multiple tissues (*gray*). **c.** A flow chart of the process for identifying heritable gene-expression states in the progeny and the parental genes that predict them. **d.** Four example genes identified as showing greater similarity in expression between siblings (*shared colors*) than between cousins (*varied colors*). Each point represents the expression in a single individual. **e.** The results of a set of partial least squares (PLS) regressions predicting each gene plotted using parental gene-expression. The relative magnitude of PLS coefficients estimated using real parents vs randomly shuffled parents, with p(actual – shuffled < .01) colored in red. Additional genes plotted in Fig. S4e. **f.** Four example genes whose expression in progeny is well-predicted by a single gene in the parents.

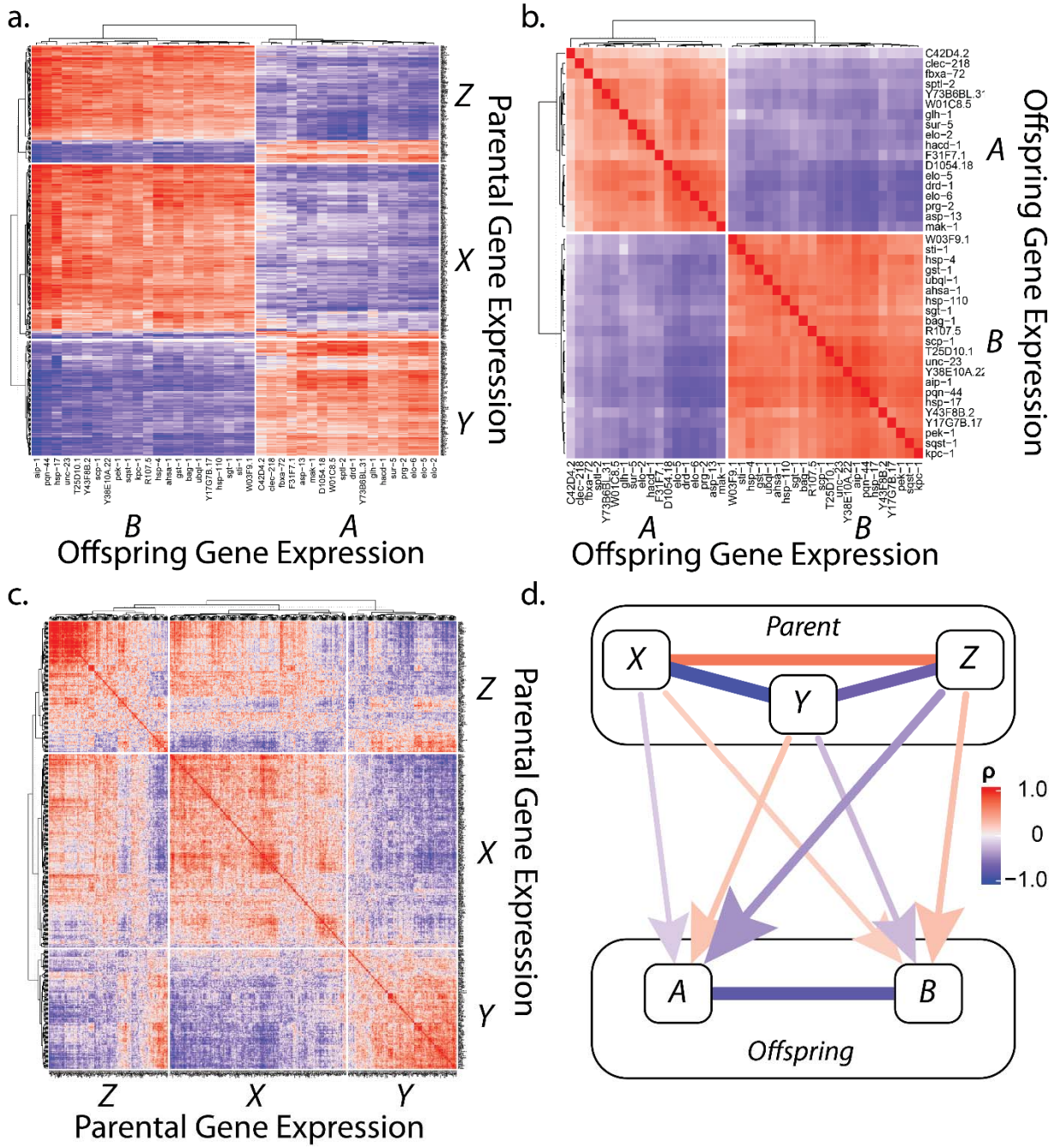
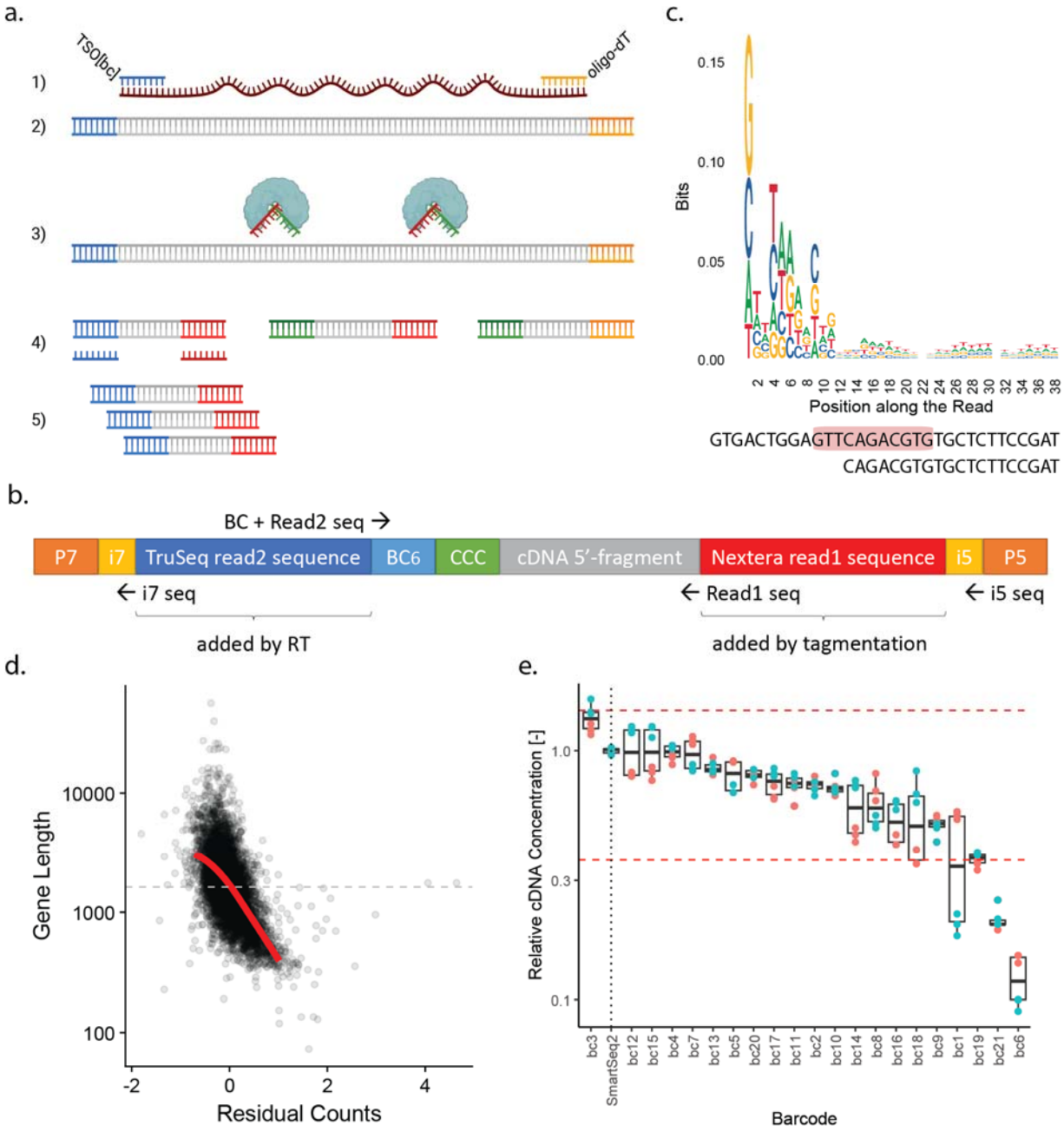


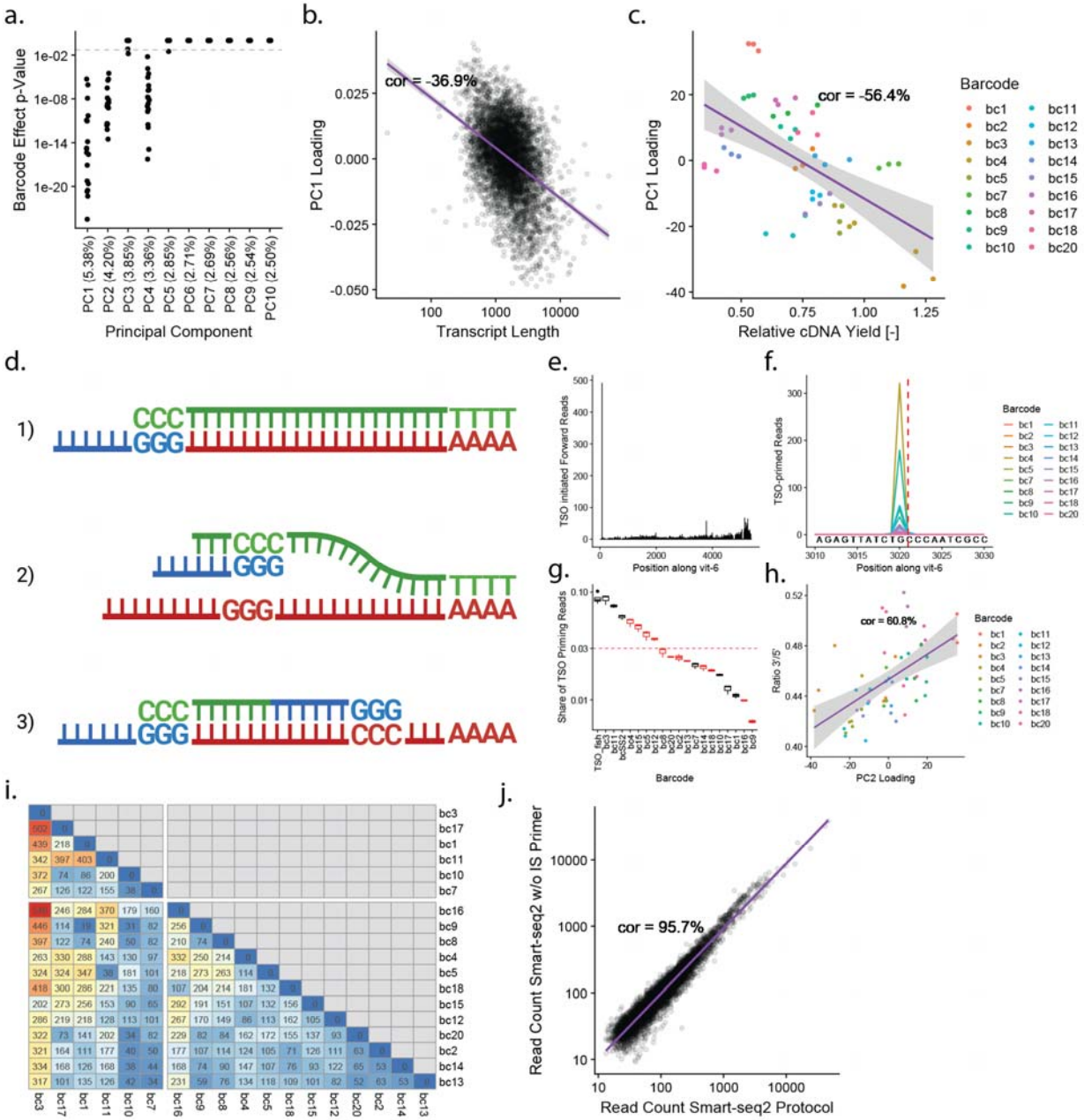
Figure 5—An inter-generational map of gene-expression variation in *C. elegans*. a. Hierarchical clustering of heritable genes reveals two groups of genes (A, B) whose expression in progeny correlates with the expression of three distinct groups of genes (X,Y,Z) in their parents. Colors indicate Pearson correlation between -1 (*blue*), 0 (*white*) and 1 (*red*). **b.** Within the progeny population, hierarchical clustering reveals strong correlations between the genes in groups A and B. **c.** Within the parental population, hierarchical clustering identifies a more complex correlation pattern between genes in groups X, Y, and Z. **d.** The full causal model suggested by these data, in which three sets of genes in parents drive heritable gene-expression states in two sets of genes in offspring. Line color and width are proportional to the correlation strength estimated using gene-expression data.

Supplementary Figures



Supplementary Figure 1—The STOMP-seq protocol enables a fast early-multiplexing strategy. a. Overview of the STOMP-seq protocol with 1) cDNA generation and 2) amplification from poly-adenylated mRNA, 3) tagmentation with Tn5, loaded with Nextera primer sequences, and 4) cDNA 5'-fragment fishing and 5) amplification using the TruSeq sequence added through the TSO. **b.** Detailed composition of the resulting STOMP-seq sequencing fragment with sequences of interest highlighted. **c.** Deviation from randomness at the tagmentation site of tn5 (top). Full TruSeq read2 sequence; sequence fragment with similarity to the Tn5 bias highlighted in red (middle). Truncated TruSeq read2 sequence with suppressed Tn5 bias used in STOMP-seq (bottom). **d.** Gene-length plotted against the residual difference in gene-expression between the two methods. **e.** cDNA yield of assessed barcoded TSOs compared to the cDNA yield achieved with Smart-seq2. Normalized technical triplicates of two biological

replicates (*red, blue*). Black pointed line highlights the Smart-seq2 control. Red dotted lines mark the +-0.5x interval of the barcoded samples' mean.

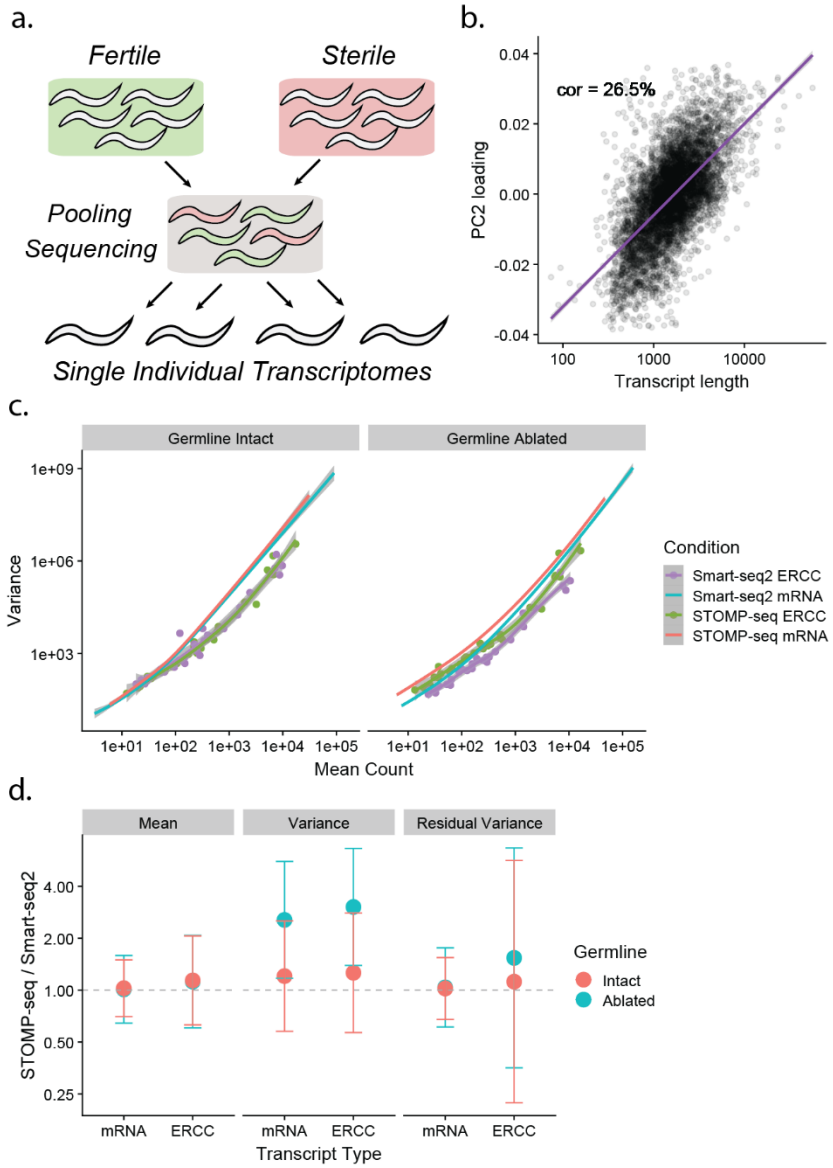


Supplementary Figure 2—The effects of TSO barcode sequence on STOMP-seq cDNA libraries. a. The adjusted p-value of the barcode effect on the PC loadings plotted for the first 10 PCs from PCA. **b.** Transcript PC1 loadings plotted against the length of the respective transcripts. Linear model fit in purple. **c.** Sample PC1 loadings plotted against the respective obtained cDNA yield of barcoded samples. Linear model fit in purple. **d.** Scheme of potential TSO interactions with the mRNA. 1) Completion of the first strand cDNA through template switching, 2) strand invasion, and 3) priming of the reverse transcription. **e.** Mapping of read2 positions (TSO initiated sequencing) generated by Smart-seq2 on the forward strand of *vit-6*. **f.** Visualization of extend of TSO priming at position 3020 (red dotted line) of *vit-6* colored by barcode. **g.** The global share of reads generated by TSO priming, separated by triplicates of

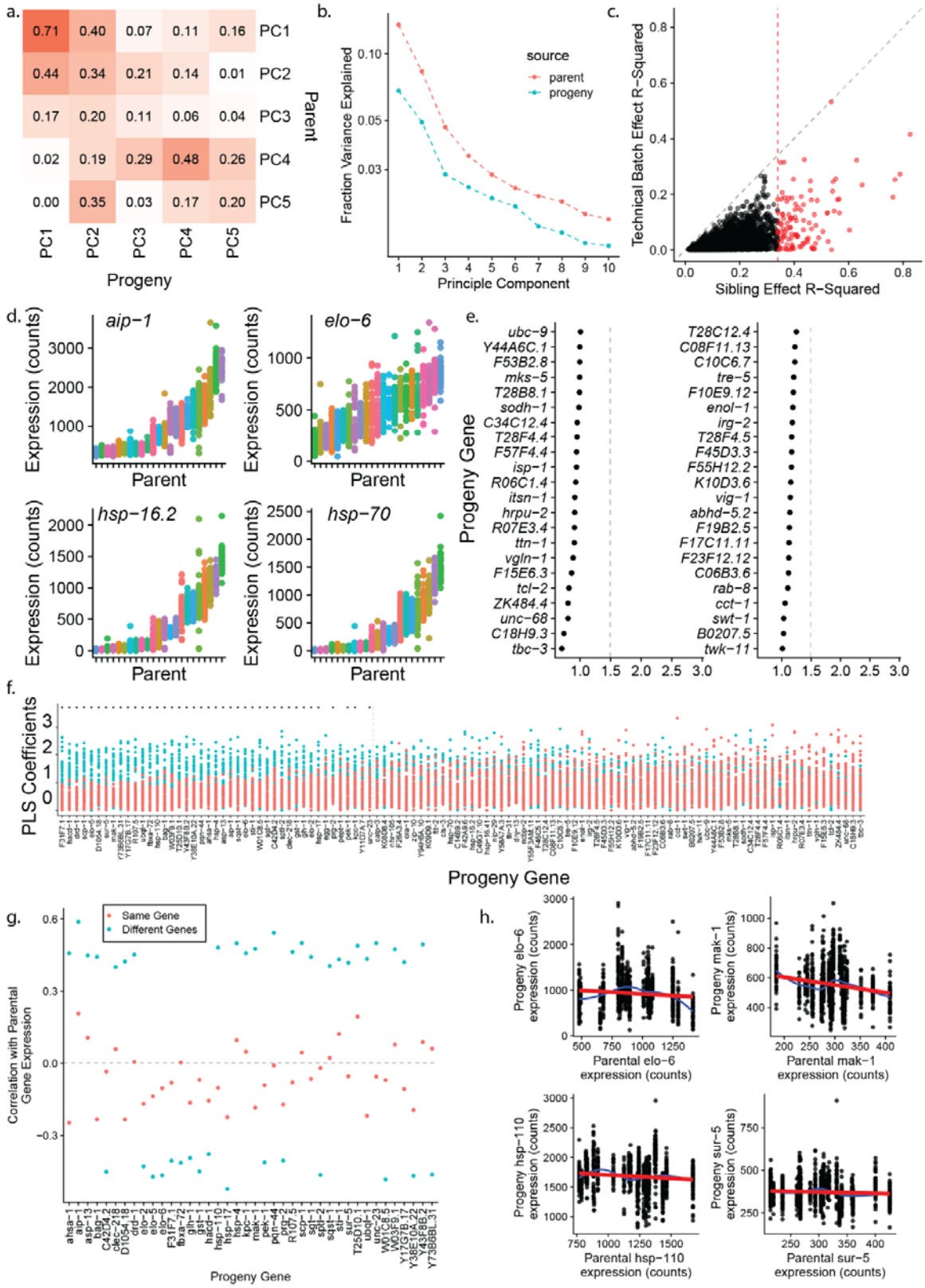
each barcode and with the mean share across the barcodes (*red dotted line*). Two strategies, TSO_fish and bcSS2, try to assess the distribution of TSO priming among transcripts processed with Smart-seq2. The final 12 chosen STOMP-seq barcode sequences are highlighted in red. **h.** The ratio of 3' to 5'-end coverage plotted against the sample loadings of PC2 for barcoded samples in triplicates. Linear model fit in purple. **i.** Heatmap of false positive differentially expressed genes (p<.01) per 10,000 transcript, identified using triplicates of identical samples amplified using one barcode each. Comparisons between the final 12 chosen STOMP-seq barcode sequences are grouped together on the bottom right. **j.** Correlation of cDNA profiles generated with or without the addition of IS primers for first-strand cDNA amplification. Linear model fit in purple.

Barcode	Normalized coverage reverse read	Genome Sequence	Barcode Sequence
bc4	272.73	CGCTAA	ACCATT
bc10	153.17	CGCTAA	CGGACT
bc12	57.29	CGCTAA	CTCTAT
bc9	54.95	CGCTAA	CCGCGA
bc8	49.36	CGCTAA	AGGCTG
bc11	36.00	CGCTAA	CTAATG
bc15	17.74	CGCTAA	GCATCT
bc13	12.89	CGCTAA	CTTGCC
bc17	5.65	CGCTAA	GGTCGC
bc3	4.91	CGCTAA	AATCCT
bc5	3.38	CGCTAA	ACTTAG
bc2	3.28	CGCTAA	AAGAAC
bc1	2.46	CGCTAA	AACGGA
bc7	1.17	CGCTAA	AGCTCC
bc14	0.58	CGCTAA	GAAGTC
bc20	0.00	CGCTAA	TTCAGC

Supplementary Table 1—The influence of TSO barcode sequence on TSO mispriming. The number of reads derived from TSOs acting as PCR primers mapping to position 3020 of *vit-6* (Fig. S2f) for each barcode.



Supplementary Figure 3—STOMP-seq accurately quantifies biological differences between samples. a. Experimental layout of the early pooling assessment using barcoded samples from fertile and sterile *glp-1(e2141)* individuals. **b.** The PC2 loading of each transcript plotted against its length. Linear model fit in purple. **c.** The variance in transcript counts plotted against the transcript's mean count for fertile (*left*) and sterile (*right*) animal samples. Samples are separated by protocol and by technical and biological variance. For visibility, only spline fits are displayed for biological variance. **d.** Summary statistics for the relative abundance of transcripts *(mean, left)*, the inter-sample variance of transcripts *(variance, center)*, and the residual variance *(residual variance, right)*, estimated using parametric methods as described in Eder *et al.* 2024. Statistics shown are the average across all mRNAs and ERCC transcripts measured in intact (*red*) and germline-ablated populations (*blue*), with error bars indicating one standard deviation.



Supplementary Figure 4—Using population-scale STOMP-seq to identify heritable gene-expression states. a. The correlation between the weights (loadings) of the first five principal components (PCs) calculated for all progeny (x axis) vs all parents (y axis). **b.** The fraction of variance explained by each PC

in the respective populations. **c.** The fraction of gene-expression variance explained by sibling groups for each gene, compared to the fraction explained by shared technical groups (worm handling batch + sequencing batch). The cutoff of sibling effects at $R^2 > 0.4$ is highlighted in red. **d.** Four example genes identified as showing greater similarity in expression between siblings (*shared colors*) than between cousins (*varied colors*). Each point represents the expression in a single individual. The same as in panel 4d, but here plotted using raw counts. **e.** The remaining PLS coefficients computed from real vs shuffled data, continued from Fig. 4e. **f.** The PLS coefficients estimated using real parental data (*blue*) compared to shuffled parental data *(red), i.e.* the underlying data used to calculate Fig. 4e and Fig. S4e. Genes with statistically significantly higher coefficients in real data are decorated with an asterisk. **g.** For the set of 40 genes identified as having heritable expression levels, the Pearson correlation in expression was calculated between each gene *(x-axis)* and the most predictive gene in the parental transcriptome *(blue)*. Separately, the correlation between the same gene *(x-axis)* in both progeny and parent *(red)* was calculated. **h.** Four example genes whose expression in progeny is well-predicted by a single gene in the parents, but whose expression in progeny is not well-predicted by the same gene's expression in the parent.