Supplementary Information

Data availability

Code and documentation can be found at https://github.com/wlab-bio/vdnamic. Raw sequencing data is available at SRA PRJNA1004618.

Experimental Method

All reagents used are enumerated in Table S3 and all oligonucleotides are enumerated in Table S4.

Zebrafish embryos preparation

AB-wildtype zebrafish were kept and crossed in accordance with the approved protocols and ethical guidelines of the University of Chicago Institutional Animal Care and Use Committee. Embryos were collected at 24 hours post fertilization (hpf), dechorionated with 1mg/ml pronase 5min at 28C. Dechorionated embryos were fixed in 4% paraformaldehyde in 1xPBS at 4C overnight. After dehydration in 100% methanol for 15-30min at room temperature, the embryos were stored at -80C for at least 2hrs before use. Embryos were successively rehydrated with 75%, 50% and 25% methanol in 1xPBS for 5min each and washed 4x with 1xPBST (1xPBS + 0.1%Tween-20), 5min per wash at room temperature. The embryos were then permeabilized with 0.5-1 x 10^{-4} U/ul Thermolabile proteinase K for 12min at room temperature. The Proteinase K was then inactivated at 55C for 15min. Samples were then washed 4x in PBST, 5min per wash.

RT (Reverse transcription)

After permeabilization, embryos were incubated at 4C for 1hr under slow rotation (10rpm) followed by 10min of 65C incubation in a pre-RT buffer comprising 20% formamide, 0.5U/ul Superase-In, 4.4mM DTT, 0.5 ug/ul rBSA, in 1xPBS and then cooled down to 4C immediately. After one water rinse, reverse transcription mix (1x FS buffer, 4.4mM DTT, 400uM dNTP, 32uM aminoallyl-dUTP, 0.5ug/ul rBSA, 1U/ul Superase-In, 10U/ul Superscript III, and 1uM 21.068C-8N RT-primer) was added and underwent 4C incubation for 1hr, 60C 3min, and 37C overnight under slow horizontal orbital rotation, followed by 1hr of 50C incubation. After RT, embryos were washed 3x in PBST and 1x with water, incubated in ExoI mix (1x ExoI buffer, 1.43U/ul ExoI) at 4C 1hr under slow rotation, followed by 37C 1hr to remove the RT primer and displaced cDNA. Embryos were again washed 3x in PBST and 1x with water.

Tagmentation

Transposomes were assembled according to manufacturer protocol. Briefly, oligos 22tn5.003A and 22tn5.MOS-3p were resuspended to 100uM individually in annealing buffer (40mM Tris-HCl pH8, 50mM NaCl). These were annealed at equal molar ratio at 95C 2min, followed by 1 degree-C decrements per minute for 70min. Transposome assembly mix (25uM of the annealed oligo-duplex, 1ug/ul tagmentase) was then mixed and incubated at 23C for 30min. Glycerol was then added to 50% and stored until use at -20C. Following reverse-transcription of samples transposome-glycerol stock was diluted 3:10 in tagmentase dilution buffer. This diluted solution was then added at a further 1:50 dilution to transposome reaction mix containing 5mM $MgCl_2$, 10mM Tris-HCl (pH7.5), 10% N,N-Dimethylformamide, 9% PEG8000, and 850uM ATP. This mix was added to samples, and incubated at 4C 1hr under slow rotation followed by 55C 1hr. Samples were then washed 2x in PBST and 1x in PBS.

Cross-linking and transposome denaturation

BS(PEG)5 mix was prepared in 1xPBS to 5mM concentration and added to samples for incubation at room temperature for 1hr under slow rotation. Samples were then rinsed with 1M Tris pH 8, and quenched in this buffer 30min at room temperature. 4xSSC was then added, samples were incubated at 4C for 10min under slow rotation, and then at 70C for 15min. 10% formamide/2xSSC was then added, the samples were incubated at 4C under slow rotation for 10min, and then incubated at 50C for 10min. Samples were then washed 3x in PBST, 5min each.

3' adapter ligation and circDNA annealing

After rinsing with water, 3' adapter ligation mix (500nM 22tn5.005, 500nM 22tn5.006, 1.25U/ul SplintR ligase, 1x SplintR Reaction buffer containing ATP) was added to samples, which were incubated at 4C 1hr followed by 23C overnight. After ligation, samples were washed 3x in PBST for 5min each wash, then rinsed with water, and 3' phosphates on the ligated oligos were removed with 0.5U/ul Quick CIP in 1x CutSmart buffer by incubation at 4C 1hr under slow rotation followed by 37C 1hr. Samples were then again washed 3x in 2xSSCT, 5min per wash. Circ6G1 and Circ7G1 were prepared using T4 DNA ligase and short splint oligos[a] (splint6F5 and splint7F5, respectively) and purified using a Zymo Oligo Concentrator spin column. Products were checked for size and purity via TBE-urea gel. CircDNA annealing mix (100nM Circ6G1, 100nM Circ7G1, in 1x hybridization buffer, containing 2x SSC, 10% formamide, 0.1% Tween-20) was added to

[a] An, Ran, et al. Nucleic Acids Research 45.15 (2017): e139-e139.

samples and incubated overnight at 40C under slow rotation. Samples were then washed in hybridization buffer at 40C for 30min under slow rotation, and then washed in 2xSSCT, 1xSSCT, and finally 1xPBST, at 5min per wash.

Circular DNA annealing and RCA (rolling circle amplification)

Samples were rinsed with water and RCA mix (25ng/ul T4 gene 32, 1x phi29 reaction buffer, 0.5ug/ul rBSA, 250uM dNTP, 0.2U/ul phi29 polymerase) was added, incubated at 4C 1hr under slow rotation, and then 30C overnight. In cases where fluorescence was to be observed, RCA mix was supplemented with 20uM fluorescein-12-dUTP. Samples were washed 3x in 2xSSCT.

UEI oligos annealing and T4 DNA extension/ligation

UEI annealing mix (100nM 21.004G1/2-BC oligo mix, 100nM 21.073pt, 100nM 21.074B, 2xSSC, 5% formamide, 0.1% Tween-20) was added to samples, incubated at 4C 1hr under slow rotation, and then 50C 2hrs. After bringing to room temperature, samples were washed in UEI-hybridization buffer (2xSSC, 5% formamide, 0.1% Tween-20) 1hr at 50C, followed by washes in 2xSSCT, 1xSSCT, and then 1xPBST. After water rinse, extension/ligation mix (1x T4 ligase buffer including ATP, 1mM dNTP, 0.15U/ul T4 DNA polymerase, 20U/ul T4 DNA ligase) was added and incubated 1hr under slow ration at 4C, followed by room temperature incubation 40min. Samples were then washed 3x in PBST, followed by a water rinse.

IVT (In vitro transcription)

IVT-ligation mix was prepared by adding to final concentrations together, in order, oligo 21.075 (100nM), 21.066C3 (1uM), 1x IVT reaction buffer, 7.5mM rNTP mix, 100ng/ul T4g32, 0.5U/ul T4 RNA ligase 2, 0.25U/ul RppH, 10% T7 Enzyme Mix, 73.6ug/ul 4arm-PEG20K-Vinylsulfone, and 6.4ug/ul 3-arm Thiocure-333 (PEG reagents being thawed from -80C immediately prior to reaction). Mixes were added to individual zebrafish embryos at a total volume of 30ul. Hydrogel was allowed to form around samples for 2hrs at room temp. Reaction was then incubated at 37C 20hrs. Afterward, hydrogels were denatured via addition of 12ul denaturation solution (457.5mM KOH, 100mM EDTA, 42.5mM DTT) for 2hrs at 4C. Denaturation was stopped by addition of 12ul stop solution (600 mM Tris-HCl pH7.5, 0.4N HCl). After mixing, 30ul proteinase K mix (0.28% Tween-20, 0.09U/ul proteinase K, 8.6 mM Tris-HCl pH7.5) was added to the 54ul samples for a total of 84ul. This was incubated at 50C 1hr.

RNA isolation and cDNA synthesis

RNA was purified by addition of 1.2x RNAClean XP beads, following manufacturer protocols, and eluted into water. DNase I digestion was performed (final concentration of 0.8U/ul Superase-In, 0.1U/ul DNase I, 1x DNase I reaction buffer) at 37C for 30min. RNA was again purified via 1.2x RNAclean XP, and eluted into water. Reverse transcription was carried out in a final concentration of 500nM each of RT primers (21.077 and 21.085), 500uM dNTP, 1x FS buffer, 5mM DTT, 1U/ul Superase-In, and 10U/ul Superscript III. Primers and dNTP were added first to RNA/water-eluent and incubated at 65C 5min, after which the mixture was placed promptly on ice. The rest of the reaction mixture was then added, and samples were then incubated 1hr at 50C, followed by inactivation 15min at 70C, and kept at 4C. ExoI enzyme was then added directly to the product to final concentration of 3.3U/ul. After mixing, this was incubated at 37C for 30min, followed by heat-inactivation at 80C for 20min.

Library preparation

cDNA products (from IVT products) were then amplified in two separate PCR reactions. "cDNA-amplicons" were amplified by adding ExoI-digested product at a final 1:80 dilution into 4 separate reactions (per embryo) containing final concentrations of 300nM 21.046G1-BC primer, 300nM 21.081b primer, 1x HiFi PCR buffer, 200uM dNTP, 2mM $MgSO_4$, and 0.02U/ul Platinum Taq HiFi. This reaction was thermocycled 95C 2min, 5x(95C 30s, 56C 30s, 68C 2min), 20x(95C 30s, 68C 2min), 68C 5min, 4C. "UEI-amplicons" were amplified by adding ExoI-digested product at a final 1:40 dilution into 2 separate reactions (per embryo) containing final concentrations of 300nM 21.077-G1 primer, 300nM 21.076BB primer, 3.3uM each of 4E4.interf1 and 4E.interf2 (3'P-capped oligos to interfere with PCR recombination[7;[b]]), 5% DMSO, 1x HiFi PCR buffer, 200uM dNTP, 2mM $MgSO_4$, and 0.02U/ul Platinum Taq HiFi. This was thermocycled 95C 2min, 1x(95C 30s, 66C 30s, 68C 2min), 18x(95C 30s, 68C 2min), 68C 5min, 4C. PCR products were then purified using a 0.75x volume of Ampure XP beads, following manufacturer protocol. Products were quantified and sequenced on an Illumina NextSeq 500 instrument using 150-cycle kits (112nt read 1, 44nt read 2), including the sequencing primer sbs3b as a custom spike-in according to manufacturer protocol.

[b]Turchaninova, Maria A., et al. European journal of immunology 43.9 (2013): 2507-2515.

Sequence Analysis

Sequence analysis was performed using the pipeline previously described[7], with code updated for the larger scale of data available at https://github.com/wlab-bio/vdnamic.

Briefly, sequencing reads were demultiplexed via the barcodes depicted in Figure S1. Subsequently, for each amplicon type, sequence elements (UMI type I, UMI type II, UEIs, and cDNA inserts) were separately clustered using a 1bp difference-criterion using the EASL algorithm[7].

For UEI data sets, each UEI was assigned a UMI-pair by plurality (relevant only if a specific UEI appeared to show two different pairings of UMIs – a signature of PCR recombination). The resulting "consensus pairings" were then pruned, with each UMI required to be associated with 2 UEIs, and each association (unique UMI-UMI pair) required to be associated by at least 2 reads. The largest contiguous matrix (found via single-linkage clustering, with rarefaction depicted in Figure 1J) was retained for image inference.

For cDNA insert sequence data, reads grouped by the same UMI had a sequence-consensus generated by majority-vote. These sequences were then trimmed to eliminate sequence adapters. Those inserts retaining at least 25bp of non-artificial sequences (at least among known artificial sequences) were then counted toward the cDNA-insert UMIs (depicted as rarefaction in Figure 1I). These consensus inserts were then inputted into STAR alignment[c] using the Danio Rerio genome assembly GRCz11. Gene-assignments were performed using GTF annotations, with mappings ties (in edit-distance) between genes receiving equal weight, and priority assignment to rRNA in case of an ambiguous match with the genome.

The UMIs from genome-mapped cDNA-insert libraries (Table S1) were then matched back to the UMIs in the UEI amplicon libraries of the corresponding specimen. The gene-calls were then applied to label UMIs in the UEI-inferred image.

Simulations

All simulations were performed by taking the raw coordinates depicted in Figs 2A,E and calculating Gaussian "point-spread functions". For UMI i and UMI j at ground truth positions $\vec{x}_i$ and $\vec{x}_j$, respectively, and with N being the sum-total of all counts in the simulated data set, we assigned a raw count $n_{ij} \leftarrow \text{NegativeBinomial}(\text{mean} = \mu_{ij}, p)$ where $\mu_{ij} \leftarrow N e^{-||\vec{x}_i - \vec{x}_j||^2} / \sum_{ij} e^{-||\vec{x}_i - \vec{x}_j||^2}$ and $p \leftarrow 0.8$.

[c]https://github.com/alexdobin/STAR

Clustering

The preliminary segmentation analysis of GSE inferences (Figs 3H-I) was performed by taking UEI-associations collectively – and an equivalent total number of nearest neighbors (such that for k nearest neighbors, $k \leftarrow N_{\text{UEIs}}/m_{\text{UMIs}}$) – and calculating diffusion kernels within the GSE embedding coordinates $\{\vec{x}_i\}$: $\nu_{ij} \leftarrow e^{-||\vec{x}_i - \vec{x}_j||^2/2\sigma_{ij}^2}\sigma_{ij}^{-d}$ where $\sigma_{ij} \leftarrow n_{i\cdot}^{-1} + n_{j\cdot}^{-1}$. The symmetrized normalized Graph Laplacian matrix formed by the "pseudo-linkages" then underwent spectral clustering as previously described[7] down to a conductance threshold of 0.2, requiring a minimum segment size of 50 UMIs.

Inter-segment UEI counts then defined a new UEI-count matrix that was row-normalized. The top eigenvector – specifically, its median – provided the boundary for memberships visualized in Figs 4A-B of cephalic vs caudal.

For gene-gene UEI matrices (Fig 4I), a similar summation of categories was performed as with segments above. The symmetrized normalized Graph Laplacian was used to generate 100 eigenvectors from which proximities to each of the molecular species, rRNA, MT-rRNA, and gDNA. For a gene i relative to any one of these molecular species, here designated c, this proximity was estimated through the Gaussian kernel $e^{-||\vec{y}_i - \vec{y}_c||^2/s^2}$ with $s^2 \equiv \frac{1}{3}\left(||\vec{y}_i - \vec{y}_{\text{rRNA}}||^2 + ||\vec{y}_i - \vec{y}_{\text{MT-rRNA}}||^2 + ||\vec{y}_i - \vec{y}_{\text{gDNA}}||^2\right)$. A linear transform was then applied to affix the locations of each rRNA, MT-rRNA, and gDNA to the vertices of each ternary plot.

	Embryo 1	Embryo 2
Total cDNA UMIs (with inserts >=25bp)	8379427	8819189
Total genome mapped cDNA-inserts	5395011	5531597
rRNA	155559	222599
MT-rRNA	759518	682481
Protein-coding/mRNA	2946938	2885390
Genomic (non-CDS)	1436384	1654004
Other	96612	87123
UEI-data matched cDNA-inserts	1135105	1120439
Total in UEI-matrix (largest contig)	293352	300369
Type I UMIs in UEI-matrix (largest contig)	2118976	2252619
Type II UMIs in UEI-matrix (largest contig)	1644364	1776217
Total UEIs in UEI-matrix (largest contig)	8730431	9012629

Table S1: **UMI/UEI statistics for embryos 1 and 2**.

[d] "The Zebrafish Information Network (ZFIN)." The Zebrafish Information Network, https://zfin.org.

Cephalic genes		Caudal genes	
rfx2	neurod6b	mnx1	pmp22b
six3b	crlf1a	wnt5b	meox1
mdkb	barhl2	admp	tlcd5a
pax6a	fezf2	f2r	msgn1
npb	eomesa	hoxd12a	
sox2	dlc	fn1a	
otx1	tbr1b	hoxa11a	
stra6	dlx1a	tbx16l	
pigh	emx2	eve1	
fabp4a	dpf1	creb3l1	

Table S2: **Cephalic and caudal genes used in Fig 4F-G**. Cephalic genes were generated by performing a database search [16;[d]]. Cephalic genes were collected by filtering for "telencephelon" and selecting those genes with clear evidence of predominant expression in the head in 24hpf embryos in ISH images. Caudal genes were collected by filtering for "caudal fin" and "tail bud" and selecting those genes with clear evidence of predominant expression in the caudal region in 24hpf embryos in ISH images.

sMLE/UMAP/GSE comparisons

In Figs 2E-J, hyperparameters were chosen as follows. The top eigenvectors of the UEI matrix were applied to both sMLE (top 50, to undergo projected gradient descent) and UMAP (following common practice and to better constrain, the top 30 eigenvectors were used, with nearest neighbors set to 100). GSE used the top 50 eigenvectors ($E = 50$), along with 10 data tessellations. The total GSE eigenvectors used in the final projected gradients descent totaled 200.

GSE (Geodesic Spectral Embeddings)

GSE begins by "de-indentifying" type I and type II UMIs in our data set (Fig S3A): taking the rectangular $m_{\mathbf{I}} \times m_{\mathbf{II}}$ UEI-count matrix of $m_{\mathbf{I}}$ type I UMIs and $m_{\mathbf{II}}$ type II UMIs and converting it into a square $m \times m$ symmetric matrix, consisting of $m = m_{\mathbf{I}} + m_{\mathbf{II}}$ UMIs. UMI-UMI interactions are modeled statistically as illustrated in Fig S1B. There, the observed UEI matrix counts n_{ij} – associating UMI i with UMI j – are generated stochastically according to probabilities, w_{ij}, that go up the closer UMI i is to UMI j in the embedding space and go down the further apart they are in the embedding space. The set of all UMI/point positions which collectively best comports with this model is what we will call the optimal GSE embedding.

Evaluating embedding positions in this way, however, most of all requires a way to estimate the distance between them, given the original count data. Most commonly, this is done by taking the subspace formed by the top E eigenvectors of the data matrix and finding a straight-line Euclidean distance between two points[e][f][g] with the goal of identifying and focusing on nearest-neighbor relations. GSE does this as an initial processing step, but uses the resulting *putative* nearest neighbors only to form local tangent spaces about each point. These tangent spaces will allow us to estimate the *geodesic* distances along the d-dimensional surface (sitting in the full m-dimensional data space) we wish to represent in the final embedding.

These tangent spaces may involve the E "global" eigenvectors of the full symmetric count matrix in Fig S3A. However, global eigenvectors that describe the dominant axes of variance for the full data set may – on their own – be insufficient to differentiate the tangent spaces of neighboring points. To avoid this problem, GSE augments the global eigenvector subspace by performing several random, distinct tessellations of this subspace (Fig S3C). Each tessellation partitions the original points into $\sim \sqrt{m}$ sectors of $\sim \sqrt{m}$ points each.

The choice of $\sqrt{m}$ here is motivated by the fact that the total computational complexity of analyzing all sectors together will ultimately scale as the product of the number of tessellations and the number of points per tessellation, ie the total number of points in the data set $\sqrt{m} \cdot \sqrt{m} = m$. Each of these sectors now possesses smaller "local" count matrices generated by collapsing and summing the matrix elements belonging to all other sectors, as illustrated in Fig S3D. Each of these smaller matrices, because they include $\sim \sqrt{m}$ points and $\sim \sqrt{m}$ sectors, will be of size $\sim (2\sqrt{m}) \times (2\sqrt{m})$.

GSE then appends the E eigenvectors generated from these local matrices to the original E-dimensional global eigenvector subspace, forming a fuller subspace spanned by $2E$ eigenvectors describing each point's local neighborhood. These local neighborhoods are then used to find putative $2E$ nearest neighbors (the minimum to span the full eigenvector subspace) for each point in the data set (Fig S3E) using a standard kNN algorithm. The $2E$ nearest-neighbors for each point are then shuffled between tessellations, in order to allow each point – within each tessellation – to have a local neighborhood that extends beyond the boundaries of the sector into which it was assigned.

Although each sector now has a locally defined eigenvector subspace, this eigenvector subspace contains coordinates for the *collapsed* counts of all other sectors. These collapsed-sector coordinates can then be used to bridge the coordinates of points that have been assigned to different sectors (Fig S3F).

[e] Coifman, Ronald R., and Stéphane Lafon. App and comp harmonic analysis 21.1 (2006): 5-30.
[f] Van der Maaten, Laurens, and Geoffrey Hinton. Journal of machine learning research 9.11 (2008)
[g] McInnes, Leland, John Healy, and James Melville. arXiv preprint arXiv:1802.03426 (2018).

GSE then uses each point's $2E$ nearest neighbors from before to perform a local PCA analysis, giving each point its own d basis vectors – where d is the low-dimensionality of the intended embedding – that constitute the local tangent spaces for the high-dimensional surface (Fig S4A). For point i, we use these basis vectors to construct a tangent space projection matrix $\mathbf{M}_i$. Projecting the original counts associating different data points onto these tangent spaces then allows calculation of $d \times d$ count covariance matrices $\mathbf{\Sigma}_i$ (Fig S4B). For any vector $\vec{z}$, we can then write the re-scaled square-distance $\vec{z}^T \mathbf{\Sigma}_i^{-1} \vec{z}$ with a diffusion-distance metric $\mathbf{\Sigma}_i^{-1}$ in the neighborhood of point i.

For any *given pair* of points i and j at positions $\vec{z}_i$ and $\vec{z}_j$, respectively, GSE uses the procedure described so far to estimate a geodesic distance between them in two steps. First, the shortest connecting path between $\vec{z}_i$ and $\vec{z}_j$ is estimated by adding difference-vectors projected onto their respective tangent spaces to give the intermediate points $\vec{z}_i + \alpha \mathbf{M}_i(\vec{z}_j - \vec{z}_i)$ and $\vec{z}_j + \beta \mathbf{M}_j(\vec{z}_i - \vec{z}_j)$. The scalar values α and β are then adjusted to minimize the Euclidean distance between the resulting vector sums (Fig S4C), which uniquely specifies a piecewise linear path from $\vec{z}_i$ to $\vec{z}_j$. Second, the geodesic distance is estimated as the diffusion-distance traversed by this path, calculated using distance metrics $\mathbf{\Sigma}_i^{-1}$ and $\mathbf{\Sigma}_j^{-1}$ (Fig S4D).

GSE then collates these distances across all random tessellations (from Fig S3) and incorporates them together into a single geodesic similarity matrix that determines – for every point – the half of data points that are closer to it than the other half along the d-dimensional surface swept out by the tangent spaces calculated earlier (Fig S4E). GSE approximates this geodesic similarity matrix in a sparse matrix $\tilde{\mathbf{W}}$ by, for every point i, randomly and uniformly selecting a set of other points across the data set, estimating their geodesic distances to point i, and retaining the lowest $1/2^d$ fraction of these distances. These retained points, along with the nearest-neighbors found in the original eigenvector subspace, are inserted into the corresponding row in the form of a sparse set of Gaussian proximities.

GSE uses the geodesic similarity matrix $\tilde{\mathbf{W}}$ as part of what we call the "GSE matrix" $\tilde{\mathbf{W}}_{\text{GSE}}$: a mathematical description of how the original count matrix ought to be embedded across the d-dimensional tangent spaces used to construct $\tilde{\mathbf{W}}$. The top eigenvectors of this matrix are considered to be putative solutions to the d-dimensional embedding problem. In serving this purpose $\tilde{\mathbf{W}}_{\text{GSE}}$ will be a projection of the count data *into* the geodesic similarity matrix so that $\tilde{\mathbf{W}}_{\text{GSE}} \leftarrow \tilde{\mathbf{W}}\mathbf{N}$.

Because we consider the top eigenvectors of $\tilde{\mathbf{W}}_{\text{GSE}}$ to fit the data to the global curvature of the data set, we now apply an incremental projected gradient descent on a global objective function (Fig S4F). The objective function we use here is the Kullback-Leibler

divergence

$$D_{\mathrm{KL}} \equiv \sum_{ij} \left(\frac{n_{ij}}{n_{..}} \right) \log \left(\frac{n_{ij}/n_{..}}{w_{ij}/w_{..}} \right)$$

This is a statistical distance between the data counts n_{ij} and the proximities in the embedding space $w_{ij} \equiv e^{-||\vec{x}_i - \vec{x}_j||^2 + A_i + A_j}$, where $\vec{x}_i$ is the d-dimensional embedding coordinate of point i and where the amplitude $A_i \leftarrow \log n_{i\cdot}^{1/2}$. Here and elsewhere in this paper, subscript "$\cdot$" denotes index summation, such that $n_{i\cdot} \equiv \sum_j n_{ij}$.

Note that minimizing D_{KL} is equivalent to maximizing the log-likelihood $\mathcal{L}$ of the multinomial that uses UEI-counts as "independent trials" on the space of all possible UMI-pairings – the framing in earlier DNA microscopy work[7]. This is because $\mathcal{L} = \sum_{ij} n_{ij} \log w_{ij}/w_{..} +$ constant, which is the same as minus-the-expression for D_{KL} once constant coefficients are factored out.

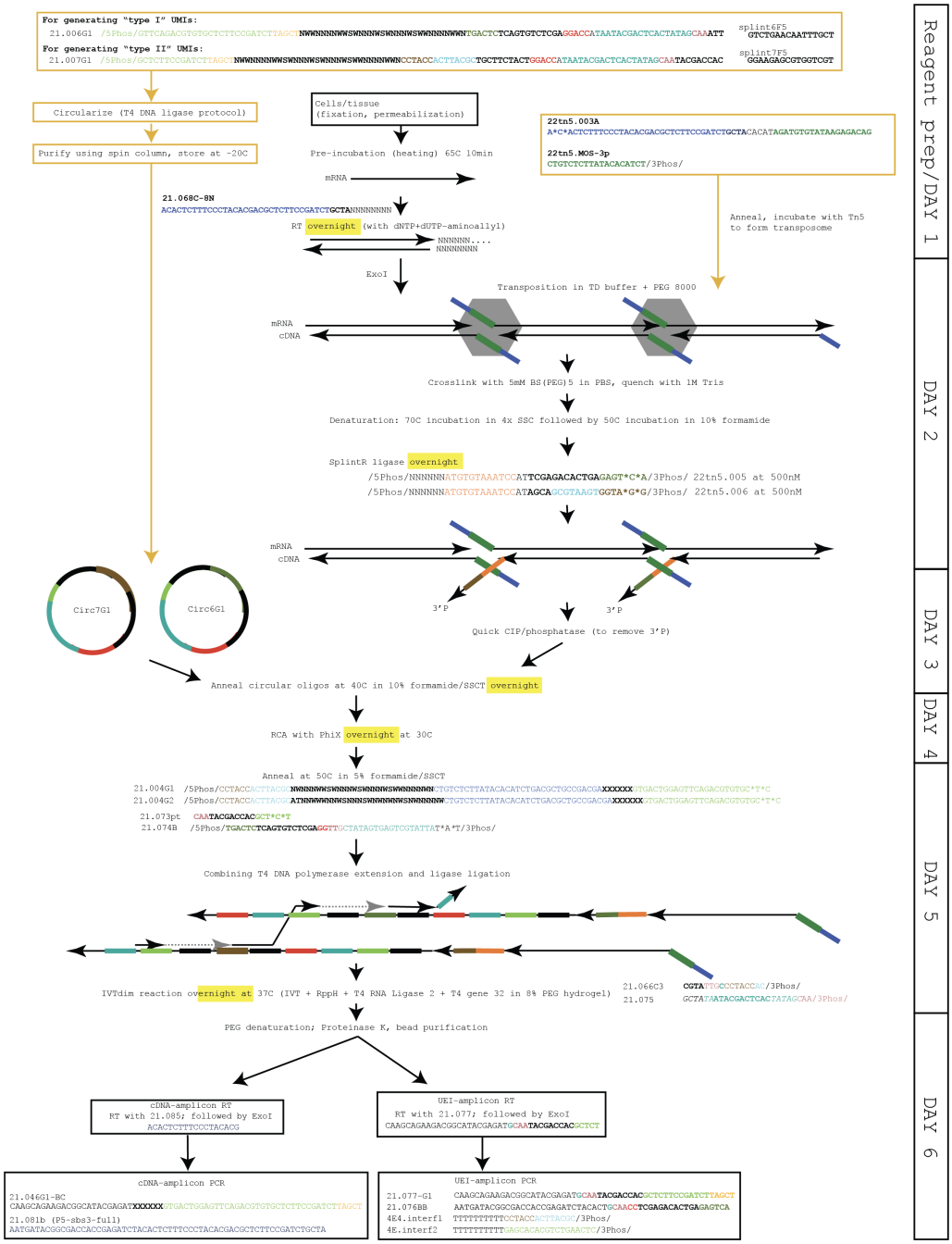

Figure S1: **Experimental flow of volumetric DNA microscopy.** "XXXXXX" corresponds to sample barcodes (arbitrary 6nt identifiers).

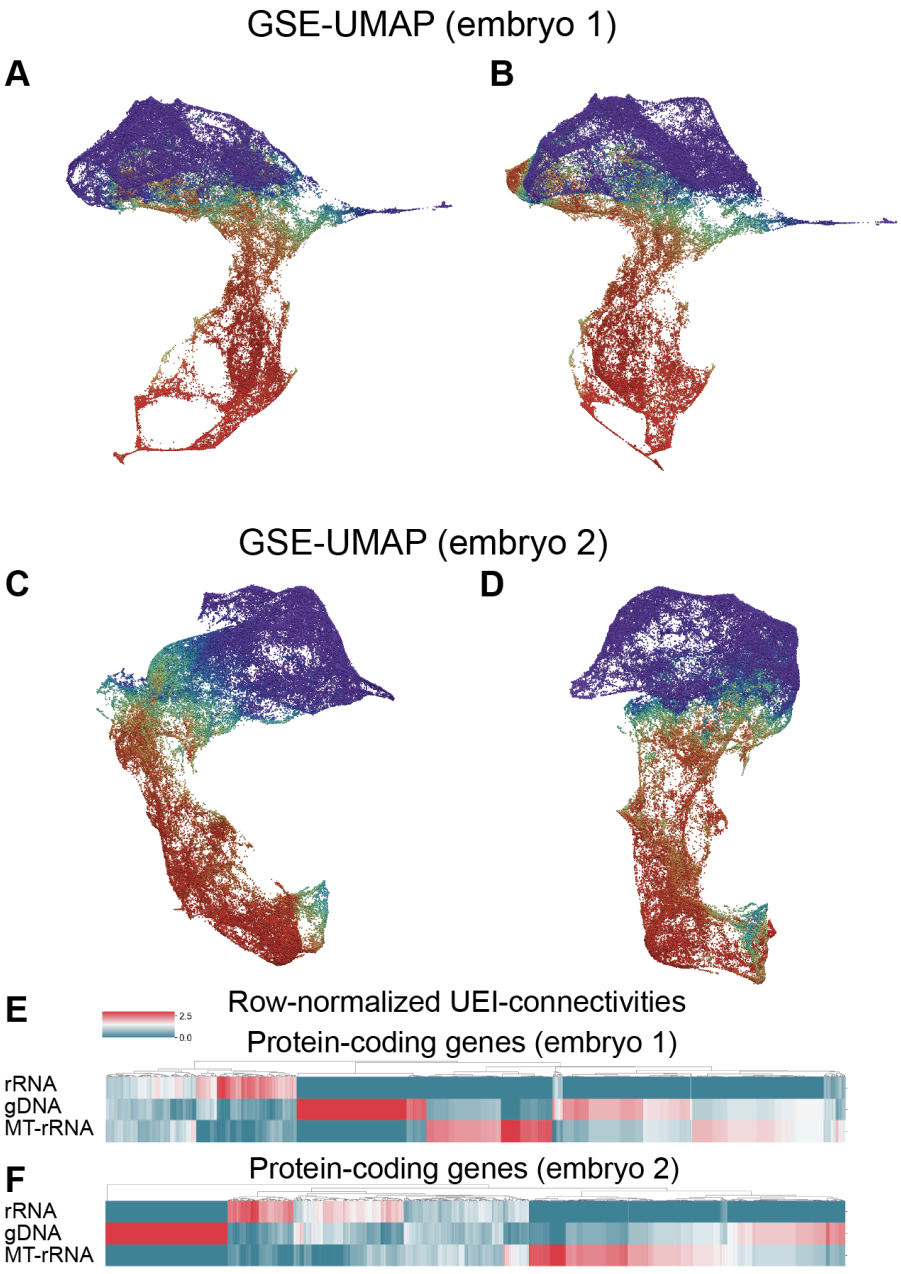

Figure S2: **Spatio-genetic visualizations.** GSE-UMAP plots from different perspective angles for embryos 1 (**A**-**B**) and 2 (**C**-**D**), with coloration as in Fig 4A-B,E. UEI connectivities in embryos 1 (**E**) and 2 (**F**) to molecular species rRNA, gDNA, and MT-rRNA, as in Fig 4H, but where UEI-counts are first normalized to the sum of each row, and are only then normalized to the mean for each protein-coding gene (columns).

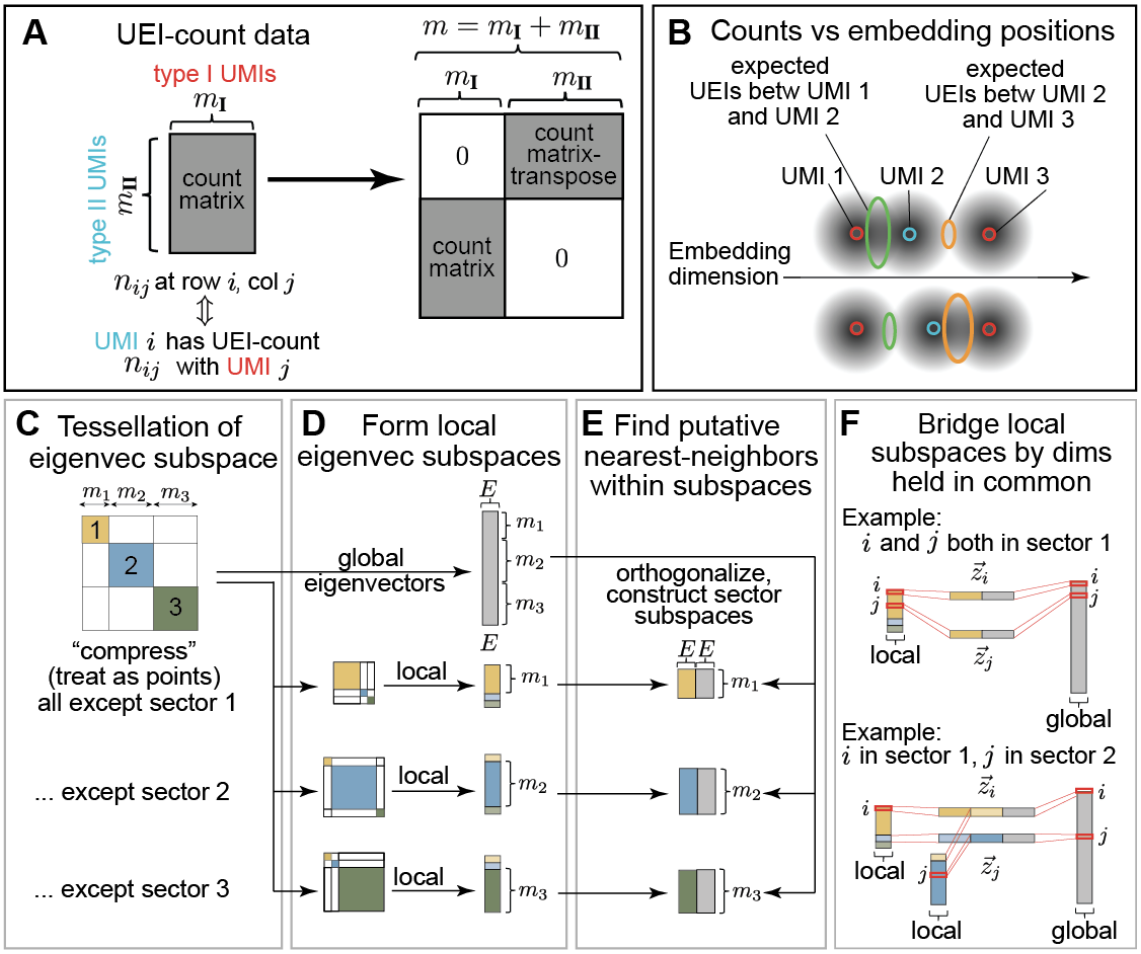

Figure S3: **Assigning global and local subspace coordinates to UEI-count matrix observables.** GSE begins by taking an arbitrary count matrix and transforming it into a block-symmetric matrix describing a bipartite graph (**A**). It asserts that the counts in this matrix describe the overlaps between diffusion "fields" of the rows and columns of this matrix in an embedding space (**B**). The GSE procedure begins by forming several random tessellations of the top eigenvectors of the $m \times m$ count matrix (**C**), with each tessellation consisting of multiple sectors (illustrated as sizes $m_1 + m_2 + m_3 = m$). The top eigenvectors of each sector are calculated by collapsing the other sectors into single rows/columns of "local" count matrices (**D**). These eigenvector subspaces, combined with eigenvectors from the "global" count matrix, are used to calculate nearest neighbors on a per-sector/per-tessellation basis (**E**). The coordinates of any pair of points across the data set can then be compared by analyzing the eigenvector subspace dimensions they share (**F**).

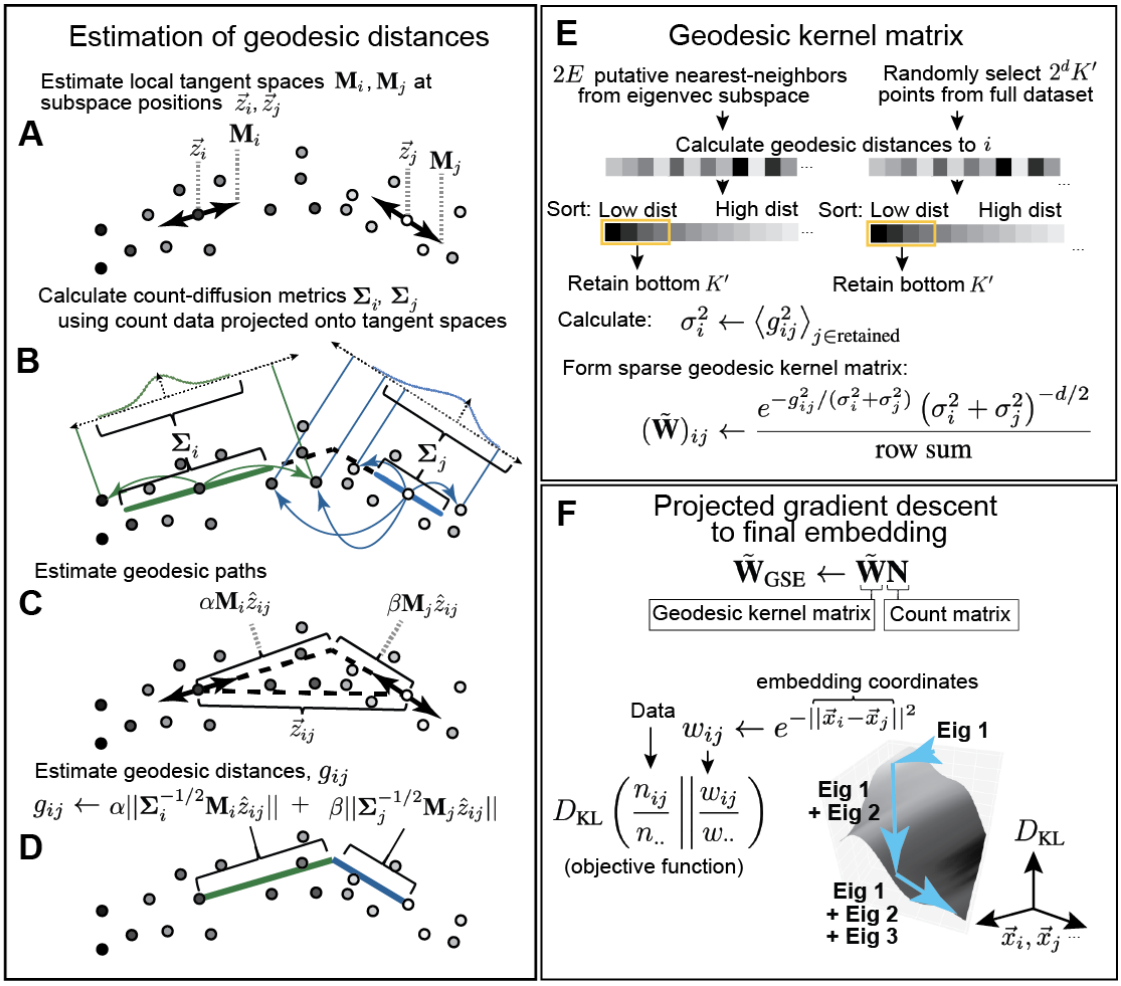

Figure S4: **GSE's numerical procedure to estimate the geodesic distance between any two points** begins by constructing local tangent spaces within the eigenvector subspaces in Fig S3 to construct local tangent spaces (**A**). Projecting each point's counts onto its own local tangent space then allows the calculation of count covariance matrices, labeled as "count-diffusion metrics" (**B**). Then, taking any two points in the data set, a shortest piecewise-linear path is constructed using knowledge of tangent spaces alone (**C**). This path is then inputted into a rescaled distance that applies the count-diffusion metrics from earlier (**D**). This geodesic estimate can then be applied both to a point's already established $2E$ nearest neighbors (from Fig S3) and to a random selection of $2^dK'$ other points – where here we set $K' \leftarrow 10d$ – across the data set. Both sets of distances are sorted independently, and the lowest K' distances from each set are retained and placed in a Gaussian kernel matrix (**E**). This kernel matrix is then used to generate a "geodesic kernel matrix", $\tilde{\mathbf{W}}_{\text{GSE}}$, whose eigenvectors are then used to construct the solution to the embedding problem (**F**).

TABLE S3

Reagents	Manufacturer	Catalog number
Pronase	Sigma-Aldrich	10165921001
Water	Invitrogen	10977023
PBS (10x, pH7.4)	Invitrogen	AM9624
Paraformaldehyde (16%)	Thermo Scientific	28906
Methanol	Sigma-Aldrich	34860-100ML-R
Tween-20	Sigma-Aldrich	P9416-100ML
Thermolabile Proteinase K (0.12U/ul)	NEB	P8111S
Formamide	Sigma-Aldrich	47671-250ML-F
rBSA (20 ug/ul)	NEB	B9200S
Superase-In (20 U/µL)	Invitrogen	AM2696
dNTP	Thermo Scientific	R0181
Aminoallyl-dUTP (50 mM)	Thermo Scientific	R1101
SuperScript III Reverse Transcriptase (200 U/µL)	Invitrogen	18080085
ExoI (20U/ul)	NEB	M0293L
Tris-HCl (1M, pH8.0)	Fisher Scientific	BP1758-500
NaCl (5M)	Invitrogen	AM9760G
Tagmentase (2ug/ul)	Diagenode	C01070010-20
Glycerol	Sigma-Aldrich	G5516-100ML
Tagmentase dilution buffer	Diagenode	C01070011
Tris-HCl (1M, pH7.5)	Fisher Scientific	BP1757-500
MgCl2 (1M)	Invitrogen	AM9530G
N,N-Dimethylformamide	Sigma-Aldrich	D4551-250ML
PEG8000 (50%)	NEB	B0216L
ATP (10mM)	NEB	B0216L
BS(PEG)5	Thermo Scientific	A35396
Tris (1M, pH8.0)	Invitrogen	AM9855G
SSC (20x)	Invitrogen	AM9770
SplintR ligase (25U/ul)	NEB	M0375S
Quick CIP (5U/ul)	NEB	M0525S
Zymo Oligo Concentrator	Zymo Research	D4060
TBE-urea Gels (15%)	Invitrogen	EC68855BOX
T4 gene 32 (10ug/ul)	NEB	M0300S
Phi29 polymerase (10U/ul)	NEB	M0269L
Fluorescein-12-dUTP (1mM)	Thermo Scientific	R0101
T4 DNA polymerase (3U/ul)	NEB	M0203L
T4 DNA ligase (400U/ul)	NEB	M0202L
4arm-PEG20K-Vinylsulfone	Sigma-Aldrich	JKA7025-1G
3-arm Thiocure-333	Bruno Bock	345352-19-4
T4 RNA ligase 2 (10U/ul)	NEB	M0239L
RppH (5U/ul)	NEB	M0356S
rNTP (25mM each)	NEB	N0466L
MEGAscript T7 Transcription Kit (T7 enzyme mix)	Invitrogen	AM1334
KOH (1M)	Honeywell	319376-500ML
EDTA (0.5M)	Sigma-Aldrich	03690-100ML
DTT (1M)	Thermo Scientific	P2325
HCl (1N)	Sigma-Aldrich	H9892-100ML
Proteinase K (0.8U/ul)	NEB	P8107S
RNAClean XP beads	Beckman Coulter	A63987
DNase I (2U/ul)	NEB	M0303S
Platinum Taq HiFi (5U/µL)	Invitrogen	11304029
Ampure XP beads	Beckman Coulter	A63881
Nextseq 500/550 Mid-Output v2.5 Kit (150 cycles)	Illumina	20024904
Nextseq 500/550 High Output Kit v2.5 (150 Cycles)	Illumina	20024907

TABLE S4	
Circular DNA	
21.006G1	/5Phos/GTTCAGACGTGTGCTCTTCCGATCTTAGCTNWWNNNNWWSWNNNWSWNNNWSWWNNNNWWNTGACTCTCAGTGTCTCGAGGACCATAATACGACTCACTATAGCAAATT
21.007G1	/5Phos/GCTCTTCCGATCTTAGCTNWWNNNNWWSWNNNWSWNNNWSWWNNNNWWNCCTACCACTTACGCTGCTTCTACTGGACCATAATACGACTCACTATAGCAATACGACCAC
splint6F5	GTCTGAACAATTTGCT
splint7F5	GGAAGACGCGTGGTCGT
In situ RT	
21.068C-8N	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTANNNNNNNN
Transposome	
22nt5.003A	A*C*ACTCTTTCCCTACACGACGCTCTTCCGATCTGCTACACATAGATGTGTATAAGAGACAG
22nt5.MOS-3p	CTGTCTCTTATACACATCT/3Phos/
3' Ligation	
22nt5.005	/5Phos/NNNNNNATGTGTAAATCCATTCGAGACACTGAGAGTC*A/3Phos/
22nt5.006	/5Phos/NNNNNNATGTGTAAATCCATAGCAGCGTTAAGTGGTA*G*G/3Phos/
UEI formation	
21.004G1-BC	/5Phos/CCTACCACTTACGCNWNNNWWSWNNNWSWWNNNNWNCTGTCTCTTATACACATCTGACGCTGCCGACGAXXXXXXGTGACTGGAGTTCAGACGTGTGC*T*C
21.004G2-BC	/5Phos/CCTACCACTTACGCATNWWWNNNWSNNNSWNNWNWSNWNNNNWCTGTCTCTTATACACATCTGACGCTGCCGACGAXXXXXXXGTGACTGGAGTTCAGACGTGTGC*T*C
21.073pt	CAATACGACCACGCT*C*T
21.074B	/5Phos/TGACTCTCAGTGTCTCGAGGTTGCTATAGTGAGTCGTATTAT*A*T/3Phos/
IVT dimerization	
21.066C3	CGTATTGCCCTACCAC/3Phos/
21.075	GCTATAATACGACTCACTATAGC AA/3Phos/
RT-PCR	
UEI	
21.077	CAAGCAGAAGACGGCATACGAGATGCAATACGACCACGCTCT
21.077-G1	CAAGCAGAAGACGGCATACGAGATGCAATACGACCACGCTCTTCCGATCTTAGCT
21.076BB	AATGATACGGCGACCACCGAGATCTACACTGCAACCTCGAGACACTGAGAGTCA
4E4.interf1	TTTTTTTTTTCCTACCACTTACGC/3Phos/
4E.interf2	TTTTTTTTTTGAGCACACGTCTGAACTC/3Phos/
IVTdim	
21.078	TCCAGTAGAAGCAGCGT
21.079B	CAAGCAGAAGACGGCATACGAGATTCCAGTAGAAGCAGCGTAAGTGGTAGG
21.076BB	AATGATACGGCGACCACCGAGATCTACACTCGAACCTCGAGACACTGAGAGTCA
cDNA	
21.085	ACACTCTTTCCCTACACG
21.081b	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTA
21.046G1-BC	CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGCT
Custom sequencing primer	
sbs3B	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG