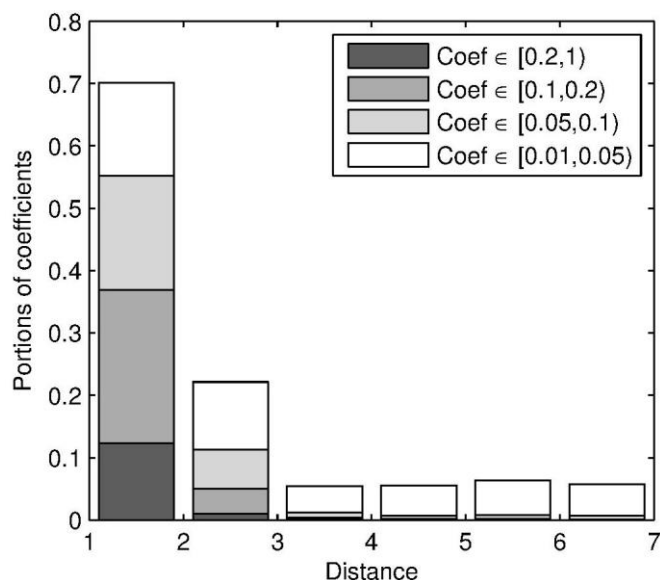


An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters

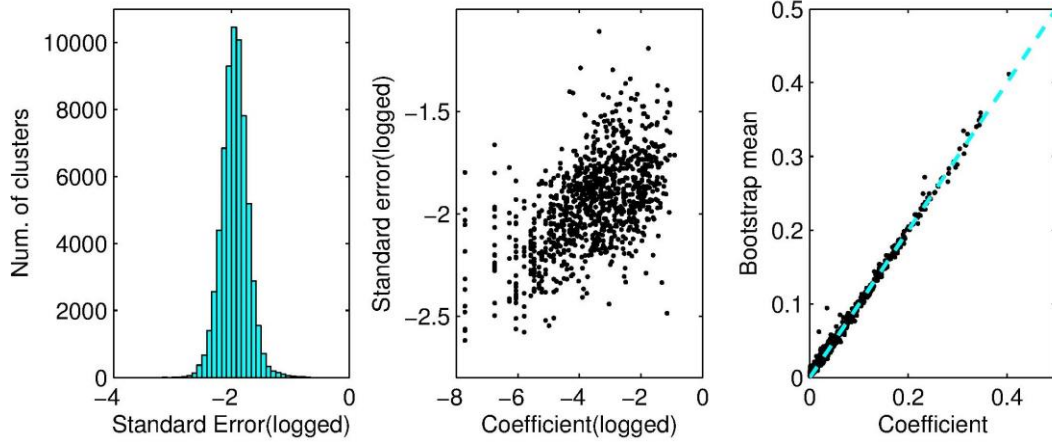
Bo Wang, Lin Wan, Anqi Wang, Lei M Li

Supplementary Materials

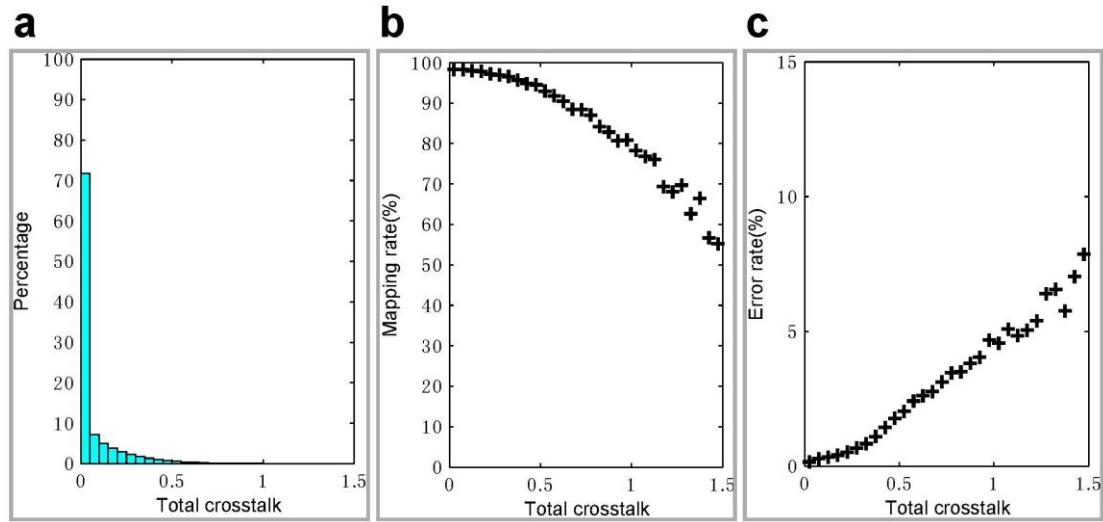
Supplementary Figures



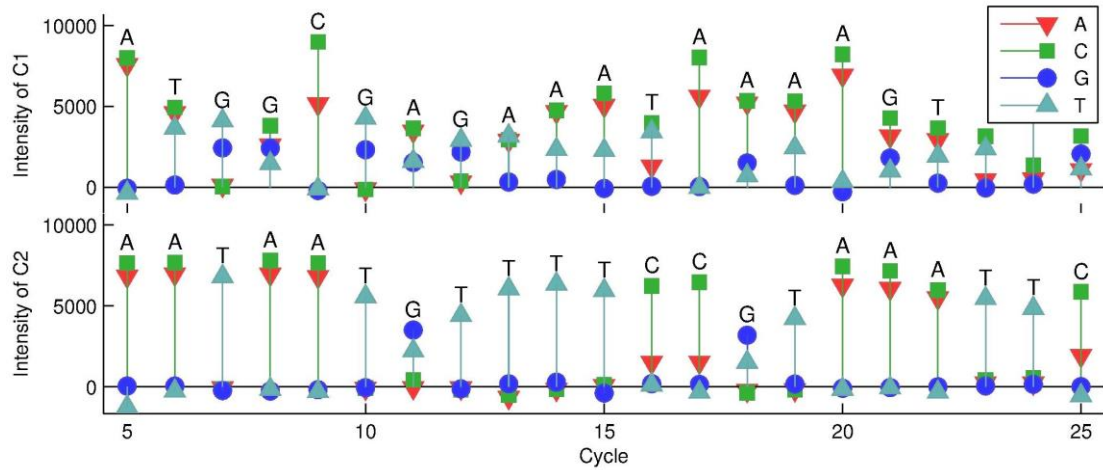
Supplementary Figure S1: Distributions of the estimated spatial crosstalk coefficients stratified by the distances between the involved cluster pairs. Each column of bars show the distribution of the coefficients. The columns are stratified by the Euclidian distances in pixels between the involved cluster pairs. For examples, the first column collects all crosstalk coefficients whose distances are in the range [1,2), and so on and so forth. More than half of the spatial crosstalk coefficients are larger than 0.05 when the cluster pairs are within 2 pixels. More than 10% of the coefficients are larger than 0.05 when the pairs fall between 2-3 pixels. The portion of the coefficients larger than 0.05 is negligible when the distances are beyond 4 pixels.



Supplementary Figure S2: The bootstrap variances and biases of the spatial crosstalk coefficient estimates. (a) The histogram of the bootstrap standard errors of the estimated coefficients on the log-10 scale. Only the non-zero coefficients are included. (b) The scatter plot of standard errors of estimates (y-axis) versus estimate values (x-axis), both on the log-10 scale. The points were randomly sampled from all non-zero coefficients. This plot indicates that standard errors of the estimates are positively correlated with the estimate values. (c) Scatter plot of the averages of bootstrap estimates of spatial crosstalk coefficients (y-axis) versus their estimates (x-axis). The points are the same as those in Panel b. The consistency indicates that the estimates are unbiased.



Supplementary Figure S3: The mapping rates and error rates stratified by the total crosstalk into each cluster. We define the total crosstalk into cluster i by summing up all c_{ij} , over all j such that $i \neq j$. (a) The histogram of the total crosstalk effects into each cluster. (b) The percentage of reads that can be mapped to the consensus reference (y-axis) stratified by the total crosstalk (x-axis) into each cluster. (c) The error rate among the mapped reads (y-axis) stratified by the total crosstalk (x-axis) into each cluster.



Supplementary Figure S4: The raw signals of the cluster pair in Figure 1. This figure shows the raw intensities (Y-axis) of the two clusters versus the cycle (X-axis): upper -- “cluster 1”; lower -- “cluster 2”. The two clusters are the 389752th and 389754th ones in the tile *s_4_1113*, and their locations are (1648.8, 2123.7) and (1649.8, 2124.5), respectively. The letters above the signals show the bases called by 3Dec.

Supplementary Tables

Supplementary Table S1: The estimates’ accuracies of spatial crosstalk coefficients stratified by their values. The first row shows the center and the average standard error of all non-zero spatial crosstalk coefficients in the selected block by mean and median. Next, these coefficients are stratified, and the results are shown in the following four rows. Also, we randomly selected 50,000 coefficients whose estimates are zeroes and show the results in the last row. The column denoted “Num” shows how many coefficients fall into each category. The following two columns show the averages and medians of the coefficients in each category. The last two columns are the mean and median values of the standard errors of the estimates, namely, measures of estimation accuracy. These statistics support the significance of coefficient estimates.

Strata	Num	Mean	Median	Mean of Standard Error from bootstrap	Median of Standard Errors from bootstrap
coef ≥ 0.01	51,566	0.0722	0.0459	0.0137	0.0119
$0.01 \leq \text{coef} < 0.05$	27,279	0.0258	0.0239	0.0125	0.0113
$0.05 \leq \text{coef} < 0.1$	11,795	0.0715	0.0699	0.0143	0.0126
$0.1 \leq \text{coef} < 0.2$	9,066	0.1398	0.1355	0.0158	0.0135
coef ≥ 0.2	3,426	0.2649	0.2492	0.0224	0.0170
coef = 0	50,000	0	0	5.81E-04	0

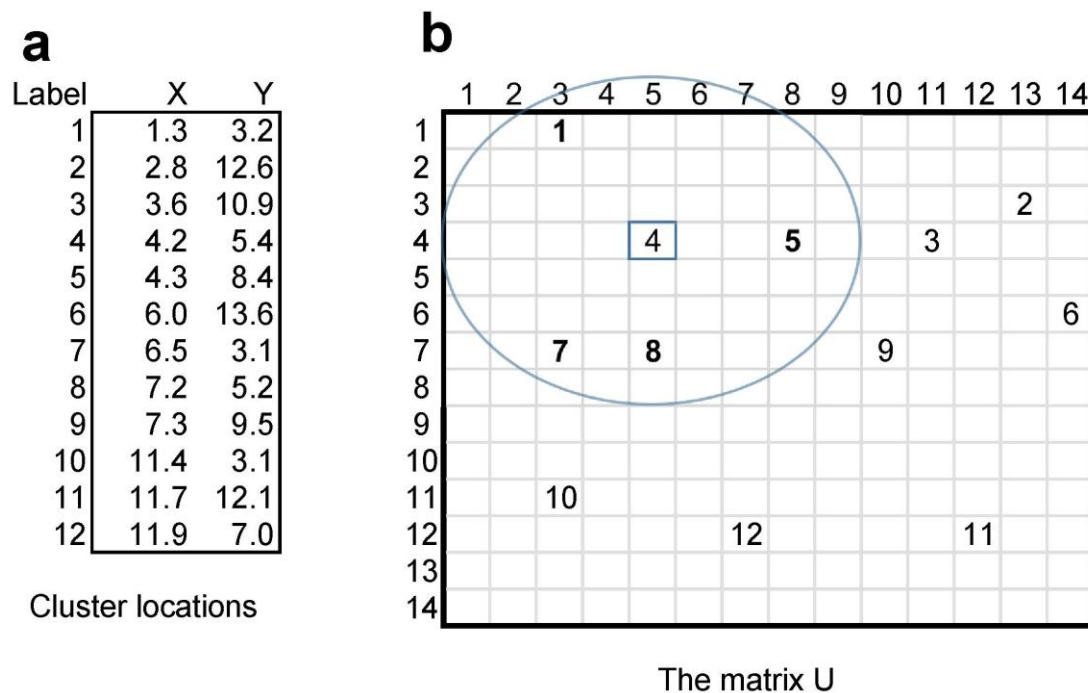
Supplementary Table S2: Deviations and standard errors of spatial crosstalk coefficients stratified by the distances and orientations of cluster pairs. We selected several two-dimensional vectors of different lengths and orientations, as shown in the first column. The column denoted by “# Cluster Pairs” represents the number of cluster pairs whose coordinate differences are equal to the vector to a precision. The next two columns are standard deviations and median absolute deviations of the coefficients in each category. The last two columns are respectively the mean and median values of the standard errors of the estimates, namely, measures of their average accuracy. The variations of the coefficients are several-fold larger than their average standard errors. This indicates that the coefficients cannot be determined by distances and orientations of the cluster pairs, demonstrating, in turn, that spatial crosstalk is cluster-specific.

Coordinate Difference Vector	# Cluster Pairs	Standard Deviation	Median Absolute Deviation	Mean of Standard Error from bootstrap	Median of Standard Errors from bootstrap
(0,1.5)	346	0.1055	0.0785	0.0162	0.0147
(-1.5,0)	389	0.1015	0.0713	0.0169	0.0164
(1,-1.5)	997	0.0823	0.0603	0.0119	0.0117
(2,0)	1,746	0.0702	0.0239	0.0092	0.0091
(1.5,-1.5)	1,939	0.0606	0.0197	0.0082	0.0080
(3,0)	2,124	0.0190	0	0.0019	0

Supplementary Note

The algorithm of finding clusters' neighbors.

1. The location file of clusters in a tile include their x- and y-coordinates. We label these clusters according to their orders in the location file (the label for the first cluster is 1, the label for the second cluster is 2, etc.). One such example is illustrated in Panel (a) of Supplementary Figure S5.
2. We generate a matrix U to approximately represent the cluster locations in the tile, as illustrated in Panel (b). The initial values of its entries are set to zero. Then we allocate each cluster to an entry in U. That is, for each cluster, we find the non-zero entry denoted by (i,j) in the matrix such that the Euclidian distance between (i,j) and the x- and y-coordinates of this cluster is the smallest; then we set the entry value as the cluster label. For example, Cluster 1 goes to the (1,3)-th entry in U. Cluster 2 goes to the (3,13)-th entry in U. We do this for cluster 3, 4, 5 etc. successively.
3. Next we identify the neighbors of each cluster. That is, we scan all the entries in U within a certain distance to the entry of the current cluster, and collect the non-zero ones. They include all neighbor clusters to the current cluster. Thus we calculate the Euclidian distances from the clusters indexed by these non-zero entries to the current cluster. Those clusters whose distance are within the given threshold (4 for 3Dec, and 15 when drawing Figure 2) are selected as the neighbors to the current one. The nearest neighbor is then the one whose distance is the shortest.



Supplementary Figure S5: Illustration of the algorithm for finding cluster's

neighbors. (a) The locations of 12 clusters and their indices. Each row represents a cluster, the first column represents the cluster label and the last two columns represents the X- and Y- coordinates of each cluster. (b) The matrix U representing the cluster locations. The vertical axis (row) and horizontal axis (column) respectively correspond to the x- and y-coordinate. Entry (i,j) in U equals 0 if no cluster is mapped to the entry of Row i, Column j, otherwise it equals the label of the cluster that is mapped to the cell, see Supplementary Note 2. Boldface numbers in the blue ellipse illustrate the neighbors of Cluster 4.

The error rates under different criteria

All the three metrics of error rates (See Table 1) are calculated based on the simple edit distance, namely

$$\text{error rate} = \frac{\text{total edit distances}}{\text{total bases in the mapped reads}}$$

But they used different sets of mapped reads for the calculations. We introduced two more metrics because the traditional one (error rate at a given mapping criterion) is insufficient to measure the accuracy of base callers. For example, we consider a small dataset containing 5 reads of 50 cycles. Suppose that the base caller A identified 0, 0, 2, 4 and 16 errors in the 5 reads, respectively; and the base caller B identified 0, 0, 1, 2 and 8 errors in the 5 reads, respectively. Obviously B is more accurate because its results contain only 11 errors while A's results contain 22 errors. Assume the mapping rule is that reads with less than 10 errors are defined to be successfully mapped back to the reference. Then the error rate under this mapping criterion for the base caller A is $(2+4)/(50*4)=3\%$, because the last read cannot be mapped under the criterion and thus is discarded. Meanwhile, the same error rate for the base caller B equals $(1+2+8)/(50*5)=4.4\%$, which is larger than the former one. In our opinion, this comparison is unfair. Therefore, we introduced the second and third metric along with the first one.