Published online 7 June 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 2 **1**
https://doi.org/10.1093/nargab/lqac042

Optocoder: computational decoding of spatially indexed bead arrays

Enes Senel[1,2], Nikolaus Rajewsky[1,2,3,4,*] and Nikos Karaiskos[1,*]

[1]Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany, [2]Humboldt-Universität zu Berlin, Institut für Biologie, 10099 Berlin, Germany, [3]DZHK (German Center for Cardiovascular Research), Partner Site Berlin, Berlin, Germany and [4]Department of Pediatric Oncology, Universitätsmedizin Charité, Berlin, Germany

Received February 04, 2022; Revised April 27, 2022; Editorial Decision May 09, 2022; Accepted May 16, 2022

ABSTRACT

Advancing technologies that quantify gene expression in space are transforming contemporary biology research. A class of spatial transcriptomics methods uses barcoded bead arrays that are optically decoded via microscopy and are later matched to sequenced data from the respective libraries. To obtain a detailed representation of the tissue in space, robust and efficient computational pipelines are required to process microscopy images and accurately basecall the bead barcodes. Optocoder is a computational framework that processes microscopy images to decode bead barcodes in space. It efficiently aligns images, detects beads, and corrects for confounding factors of the fluorescence signal, such as crosstalk and phasing. Furthermore, Optocoder employs supervised machine learning to strongly increase the number of matches between optically decoded and sequenced barcodes. We benchmark Optocoder using data from an in-house spatial transcriptomics platform, as well as from Slide-Seq(V2), and we show that it efficiently processes all datasets without modification. Optocoder is publicly available, open-source and provided as a stand-alone Python package on GitHub: https://github.com/rajewsky-lab/optocoder.

INTRODUCTION

Single-cell RNA sequencing methods (scRNA-seq) are by now well-established and of high-throughput, detecting thousands of genes at single-cell resolution (1,2). Employing scRNA-seq, researchers can readily investigate cellular heterogeneity, cell types and states, and developmental processes for a variety of tissues (3–5). One shortcoming of all scRNA-seq methods, however, is tissue dissociation that results in loss of spatial context. Spatial information is crucial to study cellular interactions in the native tissue space, to identify spatial expression patterns, and dissect tissue organisation in 3D (6–9). Such information is essential for the investigation of disease states and progression and it is anticipated that gene expression patterns in space and time will be key for the early detection and interception of complex diseases (10). In recent years, several efforts have been made to either retrieve the spatial information computationally (11–15), or to directly sequence gene expression in tissue space experimentally (16).

One way of acquiring spatially resolved transcriptomics experimentally is to use hybridisation-based methods, such as MERFISH, which achieve single-cell resolution but only for a pre-selected panel of genes (17) (although this panel can be at genome-scale). In addition to these, sequencing-based techniques that provide unbiased whole-transcriptome spatial data have become available. Methods such as the Spatial Transcriptomics (18) and the commercially available 10× Visium (18,19) use printed spatially barcoded RNA capture probes. In these techniques, however, every spot in space currently aggregates multiple cells. Seqscope is another method in which Illumina flowcells are used to amplify barcoded oligonucleotides, resulting in a higher resolution system (20). As a pioneering single-cell resolution platform, Slide-Seq (and Slide-SeqV2), was developed to spatially capture tissue gene expression (21,22).

In array-based methods, such as Slide-Seq, a tightly packed group of beads carrying DNA oligos are placed on a glass or a plate, termed *puck*. All oligos on the same bead share a random barcode sequence long enough to make this barcode unique for the bead. These barcodes and their positions on the puck are first optically decoded using subsequent rounds of hybridization to fluorescently labelled nucleotides in a microscopy setup (21,22). After this spatial registration of the beads, a tissue slice is placed on the puck. RNA is captured by the oligos on the beads, amplified, and sequenced- including, for each captured RNA molecule, the bead barcode. Thus, by matching these sequenced barcodes

*To whom correspondence should be addressed. Tel: +49 30 9406 1327; Email: nikolaos.karaiskos@mdc-berlin.de
Correspondence may also be addressed to Nikolaus Rajewsky. Email: rajewsky@mdc-berlin.de

© The Author(s) 2022. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

to the optically decoded barcodes, RNA molecules can be mapped to the spatial position of the bead which captured the respective molecules. Similar to Slide-Seq, we are currently also developing a spatial transcriptomics platform using a spatially barcoded assay. Efficient processing and analysis of the acquired datasets takes place in two fronts: in the processing of the sequencing data, for which we have developed Spacemake (23); and in the processing of the microscopy images.

Computational processing of the microscopy images to retrieve bead barcodes and their locations is challenging and requires three main steps. First, raw images are processed to correct problems such as misalignments across cycles and illumination errors, as well as to detect the beads. Next, the detected beads are processed for basecalling. Several technical issues may distort the signal, such as crosstalk caused by the overlapping laser excitation spectrum and phasing caused by inefficient reactions resulting in lagged signals. Finally, base calling quality is evaluated. Several base calling methods have been developed for sequencing data by primarily modelling the above confounding factors (24,25). While these methods provide solutions for their respective objectives, there is either no public and easy-to-use implementation, or they are not actively maintained. Hence, there is a lack of a complete pipeline that can process microscopy images from beginning-to-end in an easy-to-use, extensible and robust manner, specifically tailored for array-based spatial transcriptomics assays. In addition, array-based methods require the matching of the optically decoded barcodes to the true set obtained by high-throughput sequencing and the above methods do not make use of such information in a generalizable way.

Here, we developed *Optocoder,* a computational framework to efficiently process microscopy data during the optical sequencing of the barcodes and locations of the arrays in our experimental pipeline. The framework is an open-source Python software package that inputs microscopy images, processes them, corrects confounding factors, such as crosstalk and phasing, and performs basecalling. Importantly, we developed a machine learning based basecaller that increases the number of decoded barcodes that match to sequenced ones. Furthermore, Optocoder employs several measures to control the quality of the decoded barcodes at every processing step. We demonstrate Optocoder's performance on several datasets, including in-house and publicly available ones, showing the generalizability of the pipeline to different data modalities. Optocoder is scalable, versatile, extendable and can be seamlessly integrated into existing computational pipelines.

MATERIALS AND METHODS

Optocoder consists of three distinct modules (Figure 1). The imaging module is used to align the input microscopy images and detect the beads and their respective locations (Figure 1A). Second, the barcode bases are called by correcting confounding factors, such as spectral crosstalk and phasing (Figure 1B). Finally, given that a sequencing barcode set is provided, a machine learning classifier is trained to increase the number of barcode matches between the sequencing and the optical set (Figure 1C). The output of every step is quality controlled with several metrics to create a final report of the puck, image and base calling quality (Supplementary Figures S1 and S2).

Image processing

The image processing module is used to align the microscopy images and detect the beads on the array. The input to Optocoder are the puck images acquired via microscopy for every barcode base.

Bead detection. Beads are adjoining circular-shaped objects of a given radius, and we utilise Hough Circle Transform (OpenCV, 2015, *Open Source Computer Vision Library*) to detect them from the overlay image of the last cycle (Supplementary Figure S3A, Supplementary Methods). For a given bead batch, bead sizes remain constant across experiments, so that Hough Transform requires minimal optimisation. In case a different bead batch contains larger or smaller beads, the image processing module can be readily modified through adjusting the expected bead radius parameter. Bead detection outputs the (x, y) coordinates of the beads which are subsequently used to calculate corresponding channel intensities for each cycle.

Image alignment. The experimental apparatus can physically move between cycles of optical sequencing, thus resulting in potential positional differences between cycles. To retain the bead identities during the whole sequencing process, the images need to be aligned to be able to assign correct intensities to the detected beads. To begin with, the intensities can vary across cycles and can be very low for the last ones. We therefore first create overlay images and then apply histogram matching for every cycle by using the last cycle as the reference frame. Then, we use an image registration method, Enhanced Cross Correlation Maximization (26), with a Euclidean motion model to align images from all cycles to a reference and detect warping parameters (Supplementary Figure S3B, Supplementary Methods). Finally, we evaluate the registration quality for each cycle by using the Structural Similarity Index (27) (Supplementary Methods).

Background correction. Microscopy images are affected by uneven illumination and background noise that might influence the subsequent image processing and base calling steps. To subtract this uneven background signal, we first detect the background image for every channel separately by using a morphological opening operation and then subtract it from the image (Supplementary Methods). At the end of the image processing module, Optocoder outputs a matrix containing the 2D coordinates for each bead on the puck and the average fluorescence intensity for each channel.

Basecalling

In the absence of technical noise, calling bases could be performed by calling the highest intensity channel's corresponding nucleotide. As shown in the literature for Illumina sequencing basecallers (24,25,28–30), however, there

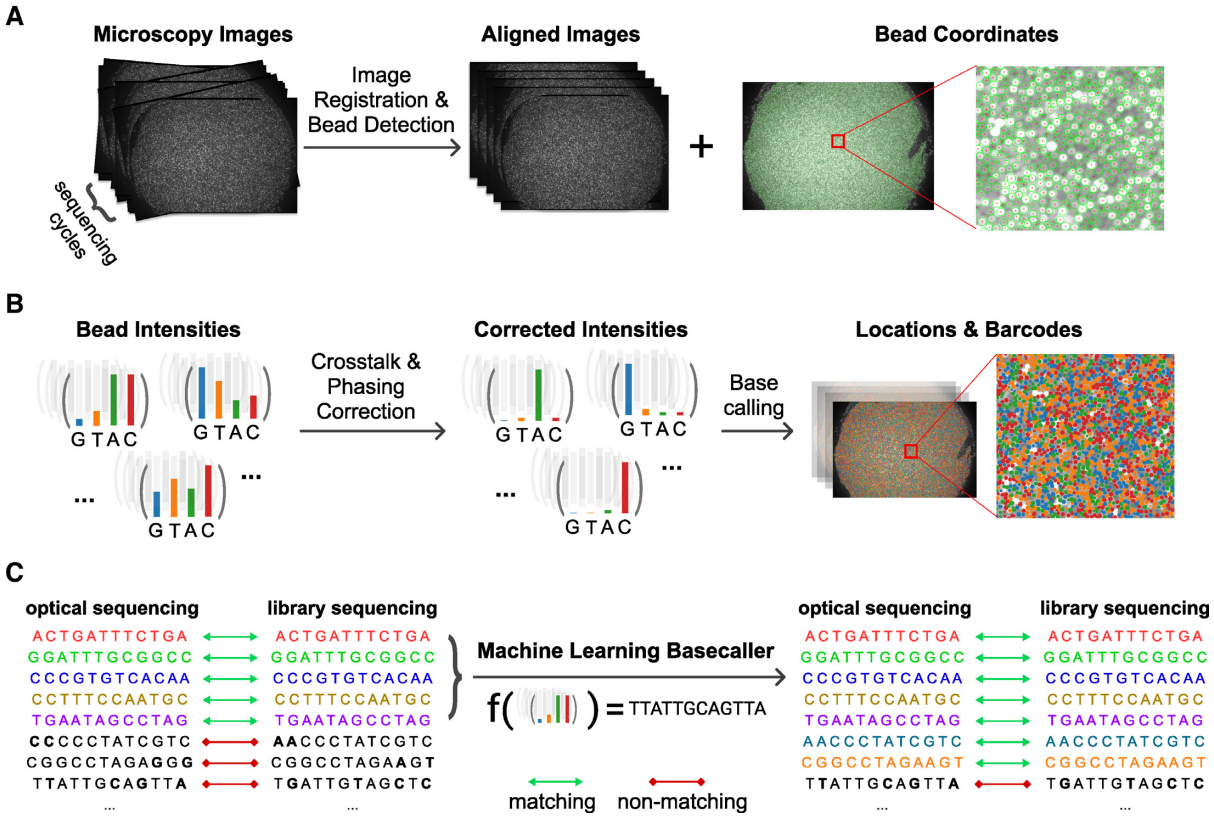

Figure 1. Schematic overview of Optocoder's modules. (**A**) Image processing is used to align microscopy images acquired across the sequencing cycles and to detect the beads and their coordinates on the array. (**B**) Crosstalk and phasing effects are corrected for high-quality basecalling. (**C**) Machine learning is employed to further correct base calling and increase matches between the optically decoded and sequenced barcodes.

exist confounding factors of the microscopy readout that need to be taken into consideration for high accuracy basecalls. Similar issues occur in the case of optical sequencing and we identified spectral crosstalk and phasing effects as the main factors that convolute the signal in our experiments.

Spectral crosstalk correction. Crosstalk refers to the correlation between the A-C and G-T channels due to overlapping emission spectra of fluorophores excited in two laser microscopy setups (Figure 2A). Optocoder utilises an estimation method (31) to detect the overlap between channels (Supplementary Methods). More specifically, the crosstalk matrix is determined by calculating the intensity overlap between every channel. First, an informative group of bead intensities is selected for every channel which is subsequently fitted with a regression model against the values in every other channel. The slope of these models represent the drift of the intensities towards the other channel (Figure 2B) and the ratio is subsequently used to correct for crosstalk, resulting in the deconvolution of the A–C and G–T channels (Figure 2C, Supplementary Methods).

Phasing and prephasing correction. Phasing and prephasing might be caused by inefficient reactions during the nucleotide incorporation process (32). A bead typically contains millions of oligos that can capture cellular molecules and missing incorporation cycles can take place for several of them. *Phasing* occurs when a nucleotide is incorporated in the next cycle instead of the current one during optical sequencing, so that the signal for that bead lags behind (Figure 3A, B). Similarly, *prephasing* occurs when multiple incorporations occur within the same cycle and the microscopy readout includes multiple nucleotides at the same time (Figure 3C). Phasing and prephasing result in convoluted signals that strongly affect basecalling quality leading to erroneous barcodes sequences. We model such effects through probabilities that correspond to the fraction of bead oligos that have phasing and prephasing for a given cycle (Supplementary Methods). Subsequently, we construct a matrix that represents the carry over signal among cycles with respect to these probabilities and use it to correct for those effects as described below.

Combined correction step. To combine the spectral crosstalk and phasing correction, we use a simplified model of the acquired signal similar to (30) as

$$B_i = C\, S_i\, P,$$

where B_i is a matrix containing the observed intensities of bead i, C is the crosstalk matrix, S_i are the true intensities of bead i, and P is the phasing matrix (Supplementary Methods). The crosstalk matrix is estimated from the first cycle and assumed to be consistent across cycles since it is a physical phenomenon of the microscopy setup and not cycle dependent. For phasing, Optocoder uses expected

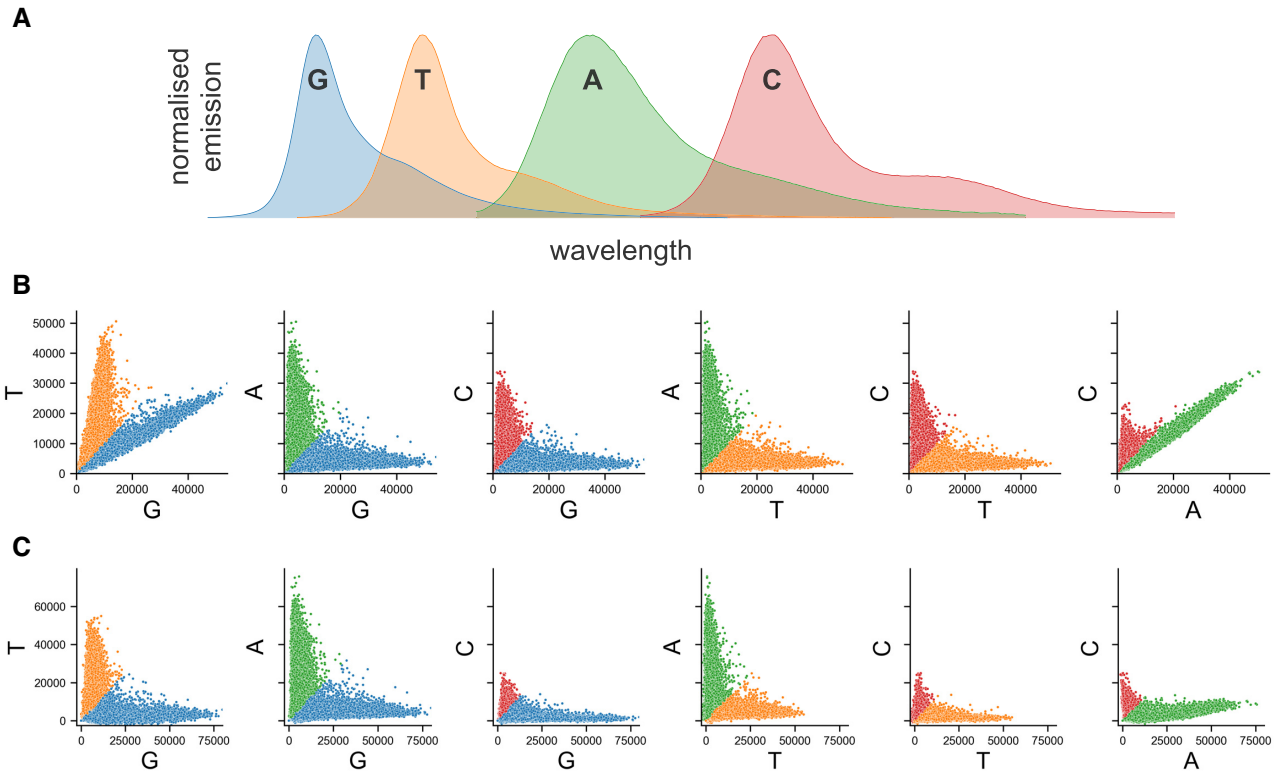

Figure 2. Optocoder efficiently ameliorates spectral crosstalk effects. (**A**) The spectra of G–T and A–C channels partially overlap in a two-laser microscopy setup. (**B**) Pairwise scatterplots of bead intensities for the puck P4 before (top) and after (bottom) crosstalk correction. Each dot represents a bead and colouring corresponds to the highest intensity of the channel pair plotted for each bead.

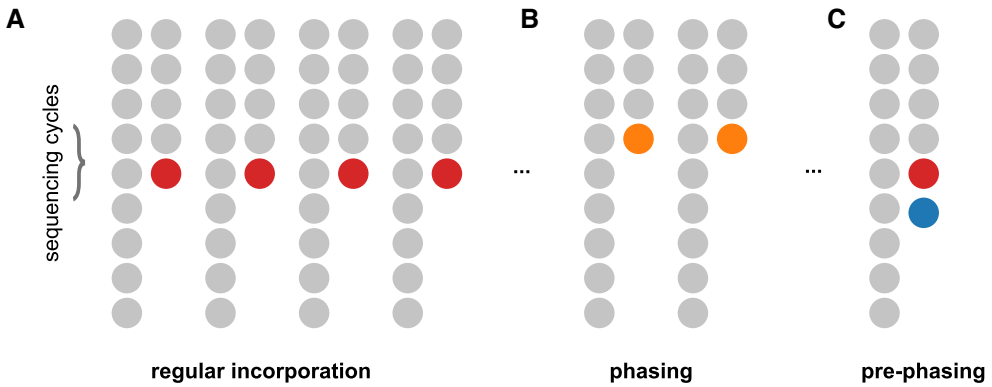

Figure 3. Optocoder efficiently corrects for phasing and prephasing effects. (**A**) In the absence of (pre)phasing effects, nucleotide incorporation takes place always in the correct cycle. (**B**) Non-incorporation of a nucleotide in the correct cycle results in phasing. (**C**) Multiple nucleotide incorporations within the same cycle result in prephasing.

phasing and prephasing probabilities chosen by the user. We have observed that for a given bead batch and experimental protocol the amount of phasing and prephasing are consistent across samples (Supplementary Figures S4 and S5). To facilitate selection of phasing parameters, we have implemented a search function in Optocoder that determines the parameter values which maximise the number of barcode matches between the optically decoded and the sequenced barcodes. As the intensity ranges of different channels vary, we apply feature scaling for every channel before basecalling (Supplementary Methods). Optocoder scales channel intensities by removing the median and scaling to the interquartile range (Robust Scaler) and also a normalised exponential function (SoftMax) is applied to each cycle's intensities for every bead before basecalling (Supplementary Methods).

Basecalling and chastity. Having corrected for spectral crosstalk, phasing and prephasing effects, we call barcode bases by selecting the nucleotide of the highest intensity for each cycle. We measure our base calling confidence by com-

puting a chastity score (33)

$$C^{pq} = \frac{I^{pq}{}_{(n)}}{I^{pq}{}_{(n)} + I^{pq}{}_{(n-1)}}$$

where $I^{pq}{}_{(n)}$ and $I^{pq}{}_{(n-1)}$ are the intensities of the channels with the highest and the second highest values for bead p in cycle q.

Machine learning

The spectral crosstalk and phasing corrections greatly improve the basecalling quality and can be readily employed via Optocoder. In array-based spatial transcriptomics methods, however, the true set of barcodes is known via high-throughput sequencing. This provides an opportunity to improve our basecalling by adding a supervised machine learning step. More specifically, we use the optically decoded barcode sequences that exactly match those stemming from the sequencing side as a training set, and we train a machine learning classifier for each sample to learn the model parameters that can use bead intensities to predict corresponding nucleotides (Figure 4). The classifier takes the background corrected intensities of all cycles after robust scaling for each bead as input features. Then, the model is trained to predict the nucleotide for every cycle.

To efficiently tackle this problem we implemented several classifiers and benchmarked their performance on a number of datasets (Supplementary Figures S6 and S7, Supplementary Methods, Supplementary Table S1). We achieved the highest performance by training Gradient Boosting classifiers for each cycle (Supplementary Figure S6), while at the same time retaining the number of false positives low (Supplementary Figure S7). Gradient Boosting is an additive model that combines weak tree models to improve model accuracy. As we input all cycle intensities to each model, models capture the effects of other cycles' intensities as well.

We begin with splitting the matching barcodes set into a randomised training (80%) and validation (20%) set. The training set is used as an input into the multi-output Gradient Boosting classifier and the validation set is used to evaluate the model's performance for hyperparameter optimization. The model with highest accuracy is then retained and used to predict nucleotide bases in the set of non-matching barcodes, which is practically the test set. We evaluate the performance in the test set by computing the number of additional matches to the sequenced barcodes.

Quality control (QC)

We have implemented several quality controls in Optocoder, which are collectively shown in an automatically generated QC sheet associated with each sample (Supplementary Figure S1 and S2). In particular, the QC sheet starts with a plot of the raw channel intensities per cycle (Supplementary Figure S1A), which facilitates experimental troubleshooting in case significant cycle deviations occur. Next, the registration accuracy score is plotted (Supplementary Figure S1B) to inspect the quality of the acquired microscopy images per cycle.

After basecalling, the overall nucleotide distribution averaged over all barcodes is plotted to measure the base content (Supplementary Figure S1C). To ensure that the barcode sequences obtained after basecalling are meaningful, Optocoder calculates two measures: string compression and Shannon entropy (Supplementary Methods). The distributions of these measures are then plotted in the QC sheet against the theoretical distributions expected for randomly uniformed sequences (Supplementary Figure S1D, E). Large deviations from either theoretical distribution would flag low confidence barcode sequences. Additionally, Optocoder plots these measures across puck space, so that areas with low confidence bead barcodes may be identified to evaluate if there are any location-specific barcode quality issues (Supplementary Figure S2C, D).

Furthermore, the distributions of chastity scores that reflect Optocoder's confidence on basecalling are plotted (Supplementary Figure S1F). Optocoder visualises the distributions of a few metrics in the array space as the spatial distribution might facilitate a better understanding of the current experiment and also better troubleshooting. Mainly, called bases and their spatial distribution (Supplementary Figure S2A), and the respective chastity scores (Supplementary Figure S2B) for each cycle are plotted and saved.

RESULTS

Performance on our data

Our array-based experimental protocol shares certain similarities with SlideSeqV2 (22) and generates microscopy images containing 4 channels for every imaging cycle. These channels display specific fluorescence profiles used to call nucleotides—four channels for the four bases G, T, A and C. After library preparation and high-throughput sequencing, the true set of bead barcodes becomes available, so that we can assess Optocoder's performance.

We benchmarked Optocoder on four different pucks that were prepared according to our protocol. Each puck contained around 70 000 beads labelled by 12 bases long barcodes and was optically sequenced under the same experimental conditions. After optical sequencing the following material was placed on the pucks: ERCC spike-ins (P1) or sections of E12 mouse brain (P2, P3 and P4). The prepared libraries were sequenced on an Illumina NextSeq 500 machine and analysed with spacemake (23) (Supplementary Methods). The 100 000 beads with the most sequencing reads were considered for matching.

Naive basecalling, i.e. without correcting for crosstalk or phasing, resulted in a low number of matches for all four pucks (Figure 5A, orange bars). Correcting for spectral crosstalk increased the number of matches by 10–20% (Figure 5A, purple bars). As expected, this increase was reflected by the corresponding chastity scores (Figure 5B). After crosstalk correction, Optocoder corrected for phasing and prephasing. This step resulted in additional matches for all four pucks. (Figure 5A, green bars) and a corresponding increase in the chastity scores (Figure 5B). We observed little to no pre-phasing in our datasets, but a strong phasing effect (Supplementary Figure S4). The combined correction step enhanced the total number of matches by 2-fold compared to naive basecalling.

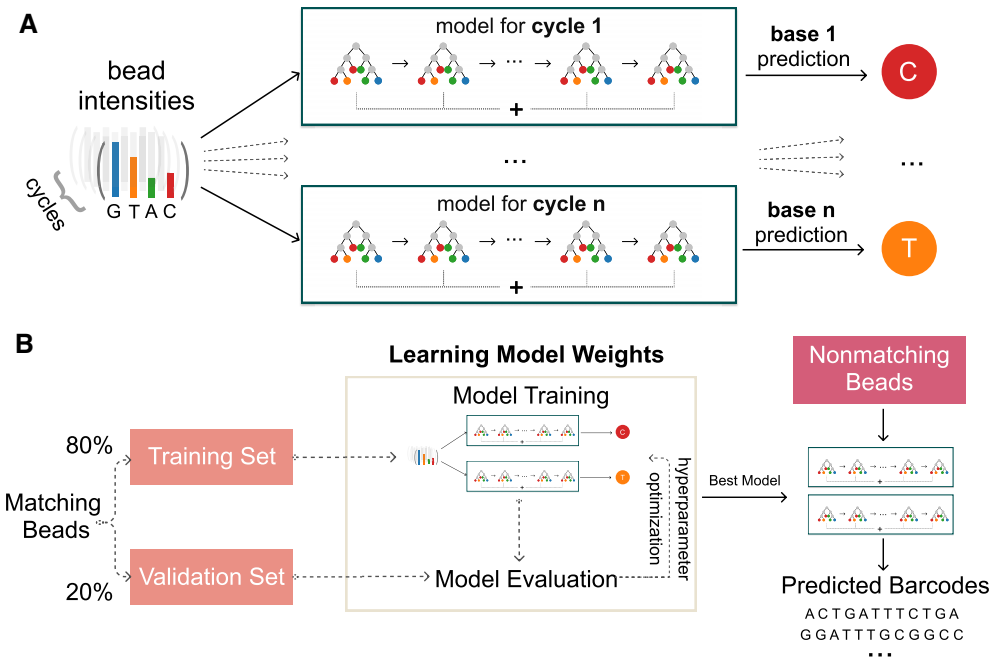

Figure 4. Supervised machine learning increases the number of matches between the optically decoded and the sequenced barcode sets. (**A**) a gradient boosting model per imaging cycle is trained to learn and predict nucleotide bases from channel intensities. (**B**) Schematic overview of the strategy employed to increase barcode matches.

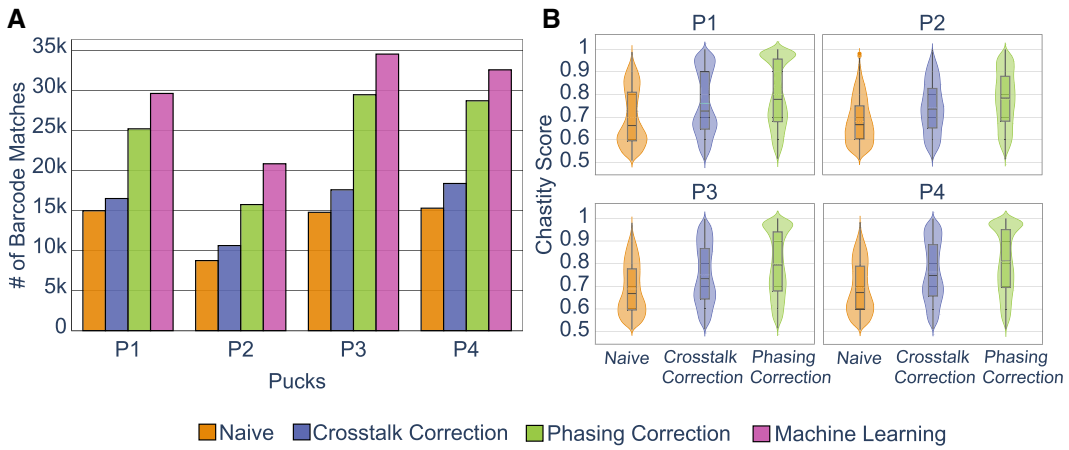

Figure 5. Optocoder exhibits high performance on own-generated data. (**A**) Optocoder efficiently corrects for crosstalk and phasing effects, and employs machine learning, resulting in a stark increase of the number of matched barcode sequences compared to naive basecalling in four different pucks. (**B**) Chastity scores consistently increase after correcting for crosstalk and phasing effects across all four pucks.

Finally, we trained machine learning models to further increase the number of matches. A model was trained for each puck separately and was used to predict the nucleotides of the non-matching barcode sequences. Optocoder's machine learning step resulted in a further 15–32% increase of matches to the combined correction matches (Figure 5A, pink bars). Interestingly, machine learning performance varied across the four pucks, with the highest increase in number of matches taking place for P2, the puck with the fewest overall matches.

We observed that all methods performed consistently, regardless of the number of top beads used (Supplementary Figure S8A, Supplementary Methods). Furthermore, we evaluated whether filtering for high quality barcodes would affect the number of matches (Supplementary Methods). For naive, crosstalk- and phasing-corrected basecallers, restricting to barcodes of very high chastity scores lowers the number of matches, however, the relative performance improvement among the methods is consistent (Supplementary Figure S9A, Supplementary Methods). All machine learning models–except for the random forests–exhibit robust performance, independent of the prediction score threshold (Supplementary Figure S10A, Supplementary Methods).

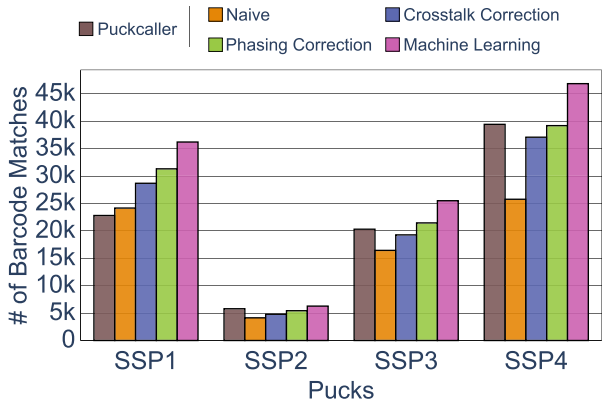

Figure 6. Optocoder efficiently processes published Slide-Seq and Slide-SeqV2 datasets. The number of matches between optically decoded and Illumina sequenced barcodes is shown for three Slide-Seq and one Slide-SeqV2 puck. In all cases Optocoder outperforms Puckcaller, the script used by the authors in the original publications.

In summary, Optocoder successfully corrected for crosstalk and phasing effects in our datasets and strongly enhanced the number of matches between the decoded and the sequenced barcodes.

Performance on external data

We primarily developed Optocoder for efficiently processing spatial transcriptomics data stemming from our in-house experimental method. Optocoder, however, is built to be versatile and adaptable. To showcase its flexibility, we used Optocoder to analyse similar microscopy datasets that are publicly available.

The initial Slide-Seq protocol (21) uses SOLiD chemistry. The optical sequencing images are generated via 20 ligations where 6 of them are constant bases. We applied Optocoder to process the microscopy images associated with three pucks SSP1, SSP2, SSP3 (Supplementary Table S2). After image processing, Optocoder detected and identified 52 754 / 48 564 / 63 643 beads, respectively. We observed little-to-no phasing and prephasing effects in the three pucks (Supplementary Figure S6). We compared the decoded barcodes against the true set of sequences that we extracted from the associated BAM file (Supplementary Methods). Matching the two barcode sets for SSP1 after crosstalk and phasing correction resulted in 31 308 exact matches which is ∼37% higher than the number of exact matches with Puckcaller, the computational pipeline that was developed and accompanied the Slide-Seq protocol. For SSP2 and SSP3, Optocoder performed similarly to Puckcaller barcodes, with −7% and +5% difference, respectively. By training the corresponding machine learning model, Optocoder resulted in 7% to 58% more matches compared to the baseline Puckcaller basecalling (Figure 6).

In addition to above, we employed Optocoder to analyse Slide-SeqV2 microscopy datasets. This is a sequence-by-synthesis generated microscopy data and the cellular barcodes consist of 14 nucleotides. Optocoder readily processed the microscopy images and decoded 63 643 barcodes for the puck SSP4 (Supplementary Table S2). Little-to-no phasing and prephasing effects were also observed for that puck too (Supplementary Figure S6). Comparing the decoded barcodes against the true set of sequenced barcodes that we obtained from the associated BAM file (Supplementary Methods) resulted in 39 188 exact matches, similar to what the authors acquired for the same dataset. Furthermore, training the machine learning model starkly enhanced the number of matches by ∼17%, resulting in a total of 46 329 exact matches. Finally, similar to the in-house pucks, all methods performed independently of the top number of barcodes used (Supplementary Figure S8B, Supplementary Methods), and we further evaluated the effect of barcode filtering through chastity scores (Supplementary Figures S9B and 10B, Supplementary Methods).

Taken together, the above demonstrates that Optocoder can reliably analyse different types of datasets, such as microscopy data of different chemistry, and achieve higher performance than existing methods.

DISCUSSION

Spatial transcriptomics methods such as Slide-Seq use spatially barcoded bead arrays that are optically sequenced. We anticipate an increase in both the development of similar methods and the utilisation of these methods in various research labs for biological insights. In this study, we present Optocoder, a computational pipeline that efficiently processes microscopy images for optical sequencing of bead barcodes. Optocoder is an easy-to-use, open-source Python package and provides a complete pipeline that processes raw microscopy images to assign bead barcodes in space. Optocoder provides functions to align images, detect beads, correct crosstalk and phasing issues and finally call the bases. Furthermore, we implemented a machine learning pipeline to increase the number of barcode matches between the optically decoded and library sequencing barcodes. We have implemented and compared four different models that are trained separately for each sample and we show that the machine learning approach substantially increases the number of matches.

We initially developed Optocoder for our in-house spatial transcriptomics platform and we evaluated Optocoder on four different samples. Additionally, we have tested Optocoder performance on Slide-Seq and Slide-SeqV2 samples and demonstrated that Optocoder efficiently processes different datasets and experimental setups with minimal modifications. In particular, we showed that correcting for crosstalk and phasing effects improved basecalling quality for our datasets, whereas for Slide-Seq datasets Optocoder performed similarly to what was originally reported. Employing Optocoder's machine learning module, however, resulted in a stark increase of barcode matches for both in-house and Slide-Seq datasets.

One drawback of the current crosstalk and phasing correction pipeline is that it implements a linear model and also the correction parameters are not tailored to beads. Beads with unique phasing properties would therefore not be efficiently processed with this approach. As a future development, implementing a more complex model that would allow for bead specific correction parameters might be beneficial.

Improved performance with the machine learning models indicates that nonlinear interactions are not fully captured by the crosstalk and phasing correction model. For machine learning basecalling, one model for each sample is trained to provide a sample-specific basecalling tool without the need for a general training set. However, this approach requires samples that already have a high number of matches before machine learning, so that a model can be trained accurately. For example, relatively poor machine learning performance for SSP2 might be explained by the small size of the initial matching set. While our current model provides a general pipeline that can be used for any new dataset and platform, an additional general model that can be trained commonly and used for different samples might be beneficial. Investigating what is learned by the machine learning models can prove useful to analyse and troubleshoot the experimental reasons for basecalling errors.

Finally, machine learning tools have been previously used for basecalling from raw signals in different platforms such as Illumina and Nanopore to improve basecalling quality (29,34,35). In principle, the machine learning approach described here utilises matched sequences and can be potentially extended to such platforms to further improve the basecalling quality by using already called reads in specific contexts, such as genome mappability.

DATA AVAILABILITY

We have deposited the microscopy images for P1-P4 on Zenodo under the DOI 10.5281/zenodo.5850813, and the corresponding Illumina sequencing data on GEO under the accession number GSE193472. The Slide-Seq microscopy images and sequencing data were downloaded from the Single Cell Portal. The Slide-SeqV2 microscopy images were provided by Evan Murray and Evan Macosko and the sequencing data were downloaded from the Single Cell Portal.

Optocoder is publicly available, open-source and provided as a stand-alone software package on GitHub: https://github.com/rajewsky-lab/optocoder.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We are indebted to our colleagues S. Abbiati, J. Alles, S. Ayoub, A. Boltengagen, A. Diag, S. Ehrig, M. Jens, J. Licha, G. Macino, M. Schott, T.R. Sztanka-Toth, S. Tagliaferro and A. Woehler for the adaptation and further development of the Slide-Seq protocol, as well as the whole Rajewsky Lab for discussions. We would also like to thank Evan Murray and Evan Macosko for kindly providing the Slide-SeqV2 microscopy images.

Author contributions: N.K. conceived and together with N.R. and E.S. designed the initial version of the pipeline. E.S. implemented and developed the pipeline. N.K. conceived and together with E.S. designed the machine learning module, which was then implemented by E.S. E.S. performed all computational and data analyses apart from the processing and analysis of Illumina sequenced libraries that was performed by N.K. N.K. and N.R. supervised the study. All authors wrote the manuscript.

FUNDING

N.K. was supported by the DFG [KA 5006/1-1, RA 838/5-1]; E.S. was supported by the DFG [KA 5006/1-1]; all authors were supported by MDC Berlin.

Conflict of interest statement. This work is part of a larger patent application in which the authors are among the inventors. The patent application was submitted through the Technology Transfer Office of the Max-Delbrück Center (MDC), with the MDC being the patent applicant.

REFERENCES

1. Aldridge,S. and Teichmann,S.A. (2020) Single cell transcriptomics comes of age. *Nat. Commun.*, **11**, 4307.
2. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
3. Trapnell,C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.*, **25**, 1491–1498.
4. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.
5. Birnbaum,K.D. (2018) Power in numbers: single-cell RNA-Seq strategies to dissect complex tissues. *Annu. Rev. Genet.*, **52**, 203–221.
6. Saunders,A., Macosko,E.Z., Wysoker,A., Goldman,M., Krienen,F.M., de Rivera,H., Bien,E., Baum,M., Bortolin,L., Wang,S. *et al.* (2018) Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, **174**, 1015–1030.
7. Wang,X., Allen,W.E., Wright,M.A., Sylwestrak,E.L., Samusik,N., Vesuna,S., Evans,K., Liu,C., Ramakrishnan,C., Liu,J. *et al.* (2018) Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, **361**, eaat5691.
8. Sun,S., Zhu,J. and Zhou,X. (2020) Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods*, **17**, 193–200.
9. Moor,A.E., Harnik,Y., Ben-Moshe,S., Massasa,E.E., Rozenberg,M., Eilam,R., Bahar Halpern,K. and Itzkovitz,S. (2018) Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell*, **175**, 1156–1167.
10. Rajewsky,N., Almouzni,G., Gorski,S.A., Aerts,S., Amit,I., Bertero,M.G., Bock,C., Bredenoord,A.L., Cavalli,G., Chiocca,S. *et al.* (2021) Publisher correction: lifetime and improving european healthcare through cell-based interceptive medicine. *Nature*, **592**, E8.
11. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
12. Karaiskos,N., Wahle,P., Alles,J., Boltengagen,A., Ayoub,S., Kipar,C., Kocks,C., Rajewsky,N. and Zinzen,R.P. (2017) The *Drosophila* embryo at single-cell transcriptome resolution. *Science*, **358**, 194–199.
13. Nitzan,M., Karaiskos,N., Friedman,N. and Rajewsky,N. (2019) Gene expression cartography. *Nature*, **576**, 132–137.
14. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
15. Moriel,N., Senel,E., Friedman,N., Rajewsky,N., Karaiskos,N. and Nitzan,M. (2021) NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nat. Protoc.*, **16**, 4177–4200.
16. Rao,A., Barkley,D., França,G.S. and Yanai,I. (2021) Exploring tissue architecture using spatial transcriptomics. *Nature*, **596**, 211–220.
17. Xia,C., Fan,J., Emanuel,G., Hao,J. and Zhuang,X. (2019) Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 19490–19499.

18. Ståhl,P.L., Salmén,F., Vickovic,S., Lundmark,A., Navarro,J.F., Magnusson,J., Giacomello,S., Asp,M., Westholm,J.O., Huss,M. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82.
19. Rao,N., Clark,S. and Habern,O. (2020) Bridging genomics and tissue pathology. *Genetic Eng. Biotechnol. News*, **40**, 50–51.
20. Cho,C.-S., Xi,J., Si,Y., Park,S.-R., Hsu,J.-E., Kim,M., Jun,G., Kang,H.M. and Lee,J.H. (2021) Microscopic examination of spatial transcriptome using seq-scope. *Cell*, **184**, 3559–3572.
21. Rodriques,S.G., Stickels,R.R., Goeva,A., Martin,C.A., Murray,E., Vanderburg,C.R., Welch,J., Chen,L.M., Chen,F. and Macosko,E.Z. (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, **363**, 1463–1467.
22. Stickels,R.R., Murray,E., Kumar,P., Li,J., Marshall,J.L., Di Bella,D.J., Arlotta,P., Macosko,E.Z. and Chen,F. (2021) Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.*, **39**, 313–319.
23. Sztanka-Toth,T.R., Jens,M., Karaiskos,N. and Rajewsky,N. (2021) Spacemake: processing and analysis of large-scale spatial transcriptomics data. bioRxiv doi: https://www.biorxiv.org/content/10.1101/2021.11.07.467598v1,08 November 2021, preprint: not peer reviewed.
24. Rougemont,J., Amzallag,A., Iseli,C., Farinelli,L., Xenarios,I. and Naef,F. (2008) Probabilistic base calling of solexa sequencing data. *BMC Bioinf.*, **9**, 431.
25. Erlich,Y., Mitra,P.P., delaBastide,M., McCombie,W.R. and Hannon,G.J. (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat. Methods*, **5**, 679–682.
26. Evangelidis,G.D. and Psarakis,E.Z. (2008) Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30**, 1858–1865.
27. Wang,Z., Bovik,A.C., Sheikh,H.R. and Simoncelli,E.P. (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, **13**, 600–612.
28. Kao,W.-C., Stevens,K. and Song,Y.S. (2009) BayesCall: a model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res.*, **19**, 1884–1895.
29. Kircher,M., Stenzel,U. and Kelso,J. (2009) Improved base calling for the illumina genome analyzer using machine learning strategies. *Genome Biol.*, **10**, R83.
30. Massingham,T. and Goldman,N. (2012) All your base: a fast and accurate probabilistic approach to base calling. *Genome Biol.*, **13**, R13.
31. Li,L. and Speed,T.P. (1999) An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis*, **20**, 1433–1442.
32. Fuller,C.W., Middendorf,L.R., Benner,S.A., Church,G.M., Harris,T., Huang,X., Jovanovich,S.B., Nelson,J.R., Schloss,J.A., Schwartz,D.C. *et al.* (2009) The challenges of sequencing by synthesis. *Nat. Biotechnol.*, **27**, 1013–1023.
33. Whiteford,N., Skelly,T., Curtis,C., Ritchie,M.E., Löhr,A., Zaranek,A.W., Abnizova,I. and Brown,C. (2009) Swift: primary data analysis for the illumina solexa sequencing platform. *Bioinformatics*, **25**, 2194–2199.
34. Cacho,A., Smirnova,E., Huzurbazar,S. and Cui,X. (2016) A comparison of Base-calling algorithms for illumina sequencing technology. *Brief. Bioinformatics*, **17**, 786–795.
35. Boža,V., Brejová,B. and Vinař,T. (2017) DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS One*, **12**, e0178751.