Liquid drop of DNA libraries reveals total genome information

Stanislav S. Terekhov[a,b], Igor E. Eliseev[c], Leyla A. Ovchinnikova[a], Marsel R. Kabilov[d], Andrey D. Prjibelski[e], Alexey E. Tupikin[d], Ivan V. Smirnov[a,b], Alexey A. Belogurov Jr[a,b], Konstantin V. Severinov[f,g], Yakov A. Lomakin[a,1], Sidney Altman[h,i,1], and Alexander G. Gabibov[a,b,j,1]

[a]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow 117997, Russia; [b]Department of Chemistry, Lomonosov Moscow State University, Moscow 119991, Russia; [c]Nanobiotechnology Laboratory, Alferov University, St. Petersburg 194021, Russia; [d]Institute of Chemical Biology and Fundamental Medicine, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia; [e]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg 199004, Russia; [f]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow 123182, Russia; [g]Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854; [h]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520; [i]School of Life Sciences, Arizona State University, Tempe, AZ 85287; and [j]Department of Life Sciences, Higher School of Economics, Moscow 101000, Russia

Contributed by Sidney Altman, September 18, 2020 (sent for review August 13, 2020; reviewed by Adrian C. Hayday and Shuguang Zhang)

Conventional "bulk" PCR often yields inefficient and nonuniform amplification of complex templates in DNA libraries, introducing unwanted biases. Amplification of single DNA molecules encapsulated in a myriad of emulsion droplets (emulsion PCR, ePCR) allows the mitigation of this problem. Different ePCR regimes were experimentally analyzed to identify the most robust techniques for enhanced amplification of DNA libraries. A phenomenological mathematical model that forms an essential basis for optimal use of ePCR for library amplification was developed. A detailed description by high-throughput sequencing of amplified DNA-encoded libraries highlights the principal advantages of ePCR over bulk PCR. ePCR outperforms PCR, reduces gross DNA errors, and provides a more uniform distribution of the amplified sequences. The quasi single-molecule amplification achieved via ePCR represents the fundamental requirement in case of complex DNA templates being prone to diversity degeneration and provides a way to preserve the quality of DNA libraries.

quasi single-molecule amplification | uniform distribution of amplicons | diversity degeneration | emulsion PCR modeling | template mispairing

Unlike the tightly controlled replication of DNA in living cells, PCR amplification, a "workhorse" of molecular biology, balances between simplicity and accuracy. While sequence accuracy can be achieved by using high-fidelity polymerases (1–3) and specific amplification protocols (4–6), the uniform amplification of complex DNA templates remains challenging. The difficulties are associated with variable amplification rates of different templates and uncontrolled mispairings that are caused by DNA repeats or homologous sequences, commonly encountered, respectively, during genomic or antibody library preparation. Hence, robust and uniform amplification of complex DNA libraries, while of clear fundamental and practical importance for molecular biology and biotechnology, is hard to attain.

The most refined method to avoid mispairing is based on amplification of single DNA molecules encapsulated inside isolated microcompartments of water-in-oil emulsions. Compartmentalization provides critical advantages to droplet-based single-cell technologies, namely the ability to distinguish and select rare individual clones from the enormous biodiversity (7–10). Similarly, pioneering work on single-molecule PCR amplification opened the era of emulsion PCR (ePCR) (11), which was successfully applied to amplifying complex gene libraries (12), PCR on microparticles in water-in-oil emulsions (13), and directed evolution driven by in vitro compartmentalization (14). The ePCR-based approaches were extensively adopted for different types of DNA libraries, such as random DNA libraries used in aptamer selection (SELEX) (15–17), antibody libraries (18–20), T cell receptor libraries (21), and libraries linking phylogeny with function on a single-cell level (22).

Here, next-generation sequencing (NGS) was used to estimate the distribution of DNA library sequences after amplification by ePCR. Unlike qualitative estimations reported previously, direct evaluation was provided of how different ePCR protocols influence the uniformity of DNA libraries' amplification. Robust and cost-effective emulsification protocols based on magnetic stirring and vortexing were compared. ePCR not only improved the yield of amplified DNA libraries but also resulted in a more uniform distribution of amplicons with reduced amount of DNA errors produced during amplification. These effects were especially pronounced in complex libraries. Finally, a mathematical model, providing a theoretical foundation for advantages of ePCR over conventional PCR, was proposed. These advantages arise from the transition from the bulk to the quasi single-molecule amplification. Implementation of the described ePCR methodology in routine laboratory practice will provide a simple and efficient tool to counteract the degeneration of the diversity of DNA libraries.

Significance

DNA libraries are predisposed to template mispairing during conventional "bulk" PCR, leading to the loss of unique sequences. The latter is facilitated by the nonuniform distribution of templates frequently observed in DNA libraries. These effects result in a prominent reduction of the original diversity. The encapsulation of DNA repertoires in liquid droplets abolishes the effects of mispairing in DNA libraries. The fundamental advantages of emulsion PCR (ePCR) over bulk PCR are illustrated by deep sequencing and mathematical modeling, which provide the general strategy for ePCR rationalization. The quasi single-molecule ePCR reveals total genetic information by counteracting the degeneration of DNA libraries' diversity.

Author contributions: S.S.T., Y.A.L., S.A., and A.G.G. designed research; S.S.T., I.E.E., L.A.O., M.R.K., A.E.T., and Y.A.L. performed research; I.E.E. and A.D.P. contributed new reagents/analytic tools; S.S.T., I.E.E., L.A.O., I.V.S., A.A.B., K.V.S., Y.A.L., S.A., and A.G.G. analyzed data; and S.S.T., I.E.E., A.A.B., K.V.S., Y.A.L., S.A., and A.G.G. wrote the paper.

Reviewers: A.C.H., King's College London; and S.Z., Massachusetts Institute of Technology.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: yasha.l@bk.ru, sidney.altman@yale.edu, or gabibov@mx.ibch.ru.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2017138117/-/DCSupplemental.

First published October 21, 2020.

Results

Optimization of ePCR for Routine Usage. To optimize the ePCR protocol, two of the most general emulsification strategies, vortexing (v) and magnetic stirring (m), were compared (*SI Appendix*, Table S1). Essential criteria included polydispersity of droplets, yield of the PCR product, quality of amplified DNA, and uniformity of amplification (Fig. 1). A programmable DNA microarray was used to synthesize two 169-mer DNA libraries, each containing ~1.2 × 10^4 unique oligonucleotides. These libraries encode overlapping 44-residue peptide tiles (Fig. 1*A*). Peptides have 14-residue overlaps between neighbor library members, spanning reference protein sequences of 17 human viruses for library A or 1015 selected human autoantigens for library B. Hence, these libraries contain multiple DNA sequences that partially overlap and thus are particularly prone to mispairing and product loss during conventional PCR amplification.

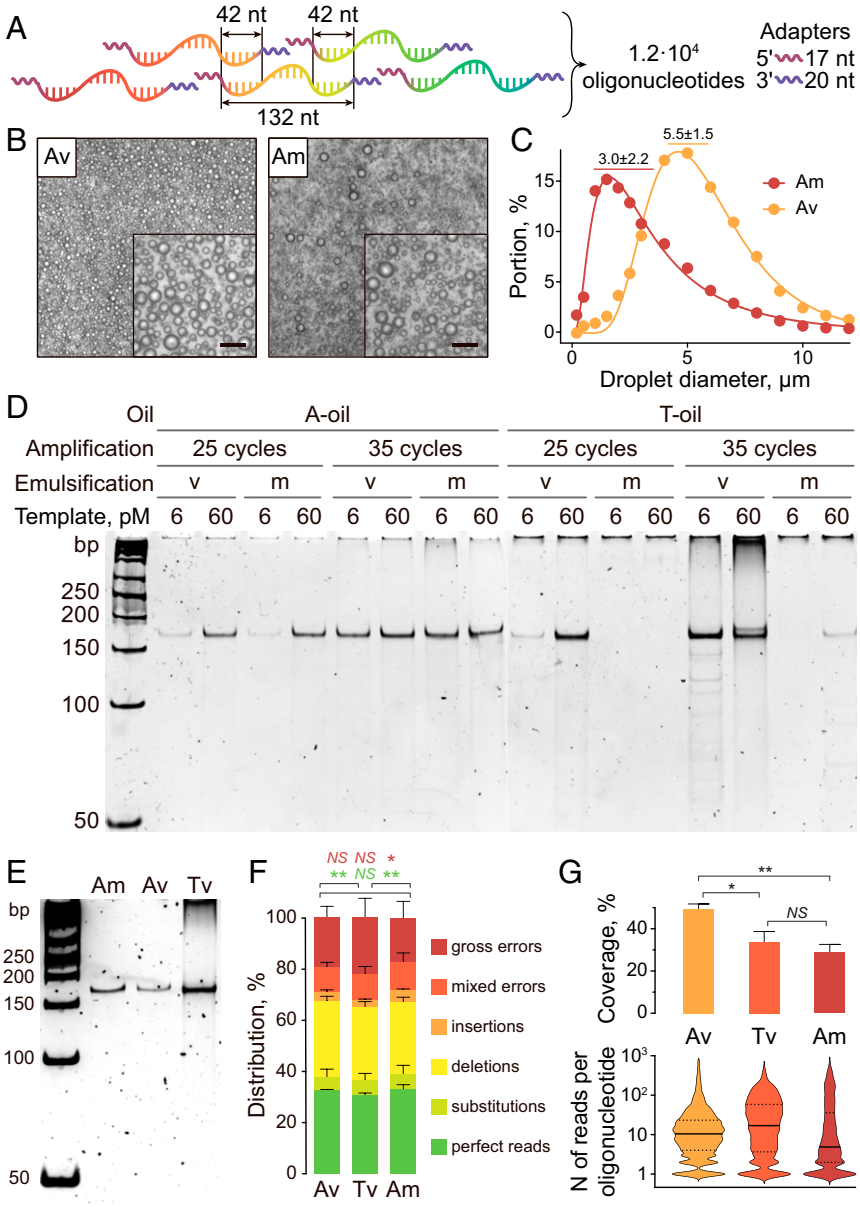

Fig. 1. The multiparametric comparison of different ePCR techniques used for amplification of a complex DNA library. (*A*) A scheme of the ssDNA library containing variable 132-nucleotide-long oligonucleotides with external overlapping 42-nucleotide-long regions flanked by adapters. (*B*) Light microscopy of emulsions produced by vortexing (Av) or magnetic stirring (Am) using Abil EM 180 emulsifier. (Scale bar, 20 μm.) (*C*) The distribution of droplet size in Am (red) and Av (orange) emulsions. Data on the size of more than 500 droplets (dots) were approximated by a lognormal distribution (line). Geometric mean and geometric SD are shown. (*D*) ePCR products obtained using different emulsifiers (Abil EM 180, A-oil and Span/Tween/Triton mix, T-oil), 25 or 35 cycles of amplification, various emulsification conditions (vortexing, v and magnetic stirring, m), and different template concentration (6 and 60 pM). (*E*) A representative electropherogram of ePCR products obtained by emulsification in A-oil with magnetic stirring (Am), A-oil with vortexing (Av), and T-oil with vortexing (Tv) after 35 cycles of amplification using 60 pM of ssDNA template. (*F*) The distributions of various types of errors observed in NGS reads of Av, Tv, and Am samples analyzed by *t* test. *$P < 0.05$; **$P < 0.01$; *NS*, not significant. Data represent the mean of three biological replicates ± 95% CI. (*G*) The distribution of NGS reads of Av, Tv, and Am samples. Coverage is the percentage of observed sequences from the initial library, estimated as the mean of three values calculated from 3.3·10^5 reads randomly sampled from NGS data, analyzed by *t* test. *$P < 0.05$; **$P < 0.01$; *NS*, not significant. Data represent the mean of three biological replicates ± 95% CI. Violin plots represent distribution of NGS reads of Av, Tv, and Am samples. Violin plots were obtained using 2.5 × 10^5 randomly sampled reads mapping to the library. Median (solid line) and interquartile range (dash line) are indicated.

The emulsion obtained by vortexing had a more uniform distribution of droplet sizes than that generated by magnetic stirring, 5.5 ± 1.5 μm versus 3.0 ± 2.2 μm, respectively (Fig. 1 *B* and *C*). The latter had a large population of small, <2-μm, droplets, which decreased the mean diameter and increased polydispersity.

Occupancy of emulsion droplets can limit the yield of amplified DNA during ePCR, since low template concentrations result in abundance of empty droplets, reducing the overall yield of ePCR products. The amplification efficiency during ePCR should depend on the average number of template molecules per droplet, defined as the λ parameter. Poisson distribution allows one to estimate the percentage of droplets with *n* template molecules for a given λ value (7). Here, two different template concentrations, 6 and 60 pM (Fig. 1*D*), corresponding, respectively, to estimated λ = 0.3 and λ = 3, were used. λ = 0.3 corresponds to 26% of filled droplets, of which 86% should contain single DNA molecules. λ = 3 results in 95% filled droplets, with only 16% of droplets having a single copy of DNA molecules. Therefore, the conditions tested allowed us to compare two radically different distributions, one with a high percentage of empty droplets, with occupied droplets containing predominantly single DNA templates, and another with almost all droplets occupied by at least one DNA template.

There are two basic compositions for ePCR based on either the Abil EM emulsifier or the Span/Tween/Triton mix (12). Mineral oil supplemented with either 3% of Abil EM 180 (A-oil) or a mixture of 4.5% Span 80/0.4% Tween 80/0.05% Triton X-100 (T-oil) was used in order to compare them in terms of emulsion stability and quality of amplified DNA library. The homogeneity of the amplified DNA was much better in the case of A-oil, whereas ePCR in T-oil resulted in the formation of high-molecular-weight by-products (Fig. 1*D*). Moreover, the outcome of A-oil ePCR was more reproducible independent of the emulsification technique. Thirty-five cycles of ePCR in A-oil were sufficient to reach saturation in the majority of the droplets (Fig. 1*D*). T-oil was less stable, displaying droplet coalescence after 25 cycles accompanied by a loss of amplified DNA quality (Fig. 1*D*).

High-throughput sequencing was performed to obtain detailed information regarding the frequency and distribution of errors that occurred during the amplification of library members. The expected outcome of quasi single-molecule amplification is reduction of template mispairing. Since mispairing leads to the formation of highly erroneous amplicons like chimeric DNA molecules and high-molecular-weight by-products, the number of errors is expected to decrease during ePCR relative to conventional bulk PCR. Regardless of the emulsification strategy, the majority of detected errors were deletions (60 to 68%), with substitutions and insertions representing, respectively, 20 to 22% and 12 to 18% (Fig. 1*F* and *SI Appendix*, Fig. S1). The probability of deletion/insertion was around 0.5% at any nucleotide position and likely reflected the initial distribution of inaccuracies of single-stranded DNA (ssDNA) oligonucleotide synthesis (*SI Appendix*, Fig. S1*A*). "Gross errors" (23)—PCR products that differed in length from the expected product by more than 3 bp—were observed in ∼20% of reads. Despite the good quality of PCR product observed after emulsification with magnetic stirrer (Am) (Fig. 1*E*), the actual distribution of reads, estimated by NGS of extracted bands of correct size, was broader in comparison with that achieved with vortexing (Av) emulsification (Fig. 1*G*). The ePCR sample that was obtained using T-oil and that displayed an intensive band of high-molecular-weight by-product demonstrated a narrower distribution of reads compared to the Am sample (Fig. 1*G*). The higher dispersion of reads in samples obtained through emulsification by magnetic stirrer resulted in overrepresentation of specific sequences. We associate this effect with more polydisperse emulsion generated by the magnetic stirring procedure. Coalescence during ePCR cycling, caused by the lower stability of T-oil, resulted in lower library coverage of Tv samples compared to Av samples (Fig. 1*G*). Hence, the appearance of clear amplicon bands on electropherograms does not correlate with the uniformity of amplification during ePCR.

ePCR Improves the Uniformity of Amplification over Conventional PCR. The sequence distribution in ePCR amplification products of two DNA libraries of different quality was compared. Both libraries were amplified by ePCR and conventional PCR (Fig. 2). The quality of library A was substantially lower than that of library B in terms of heterogeneity of sequence distribution and frameshift frequency, apparently originated from erroneous synthesis (Fig. 2 *A* and *B* and *SI Appendix*, Figs. S2 and S3). NGS analysis revealed that library B had more perfect reads and less frequent deletions and gross errors in comparison with library A. The proportion of reads with substitutions was approximately the same for ePCR libraries amplified by Am, Av, and Tv techniques and varied from 10 to 15%. The frequency of single- and double-nucleotide frameshifts was approximately the same for ePCR and bulk PCR and was most likely associated with initial errors of a chemical synthesis.

ePCR clearly outperformed bulk PCR, reducing the number of reads with gross errors by twofold (Fig. 2 *A* and *B*). Moreover, ePCR resulted in a more uniform distribution of amplified sequences in both libraries (Fig. 2 *C* and *D*). While varying template concentrations could result in different distributions of library sequences after ePCR amplification, similar results were observed at λ ≈ 0.3 or 3 (Fig. 2 *A* and *C*), indicating that ePCR amplification is reproducible irrespective of the template concentration range used. The use of bulk PCR, by contrast, resulted in substantially broader distribution of sequences in amplified DNA libraries with significantly increased frequency of overrepresented sequences (Fig. 2 *C* and *D*). Consequently, this overrepresentation resulted in lower coverage and loss of the least represented sequences. The latter effect shall be especially problematic during amplification of libraries with low initial template concentration, leading to drastic differences in gross errors, nonuniform distribution and coverage, and biases in sequence diversity after PCR amplification.

Maximal ePCR Yield Requires Optimization of Droplets Template Occupancy. The accumulation of high-molecular-weight by-products after ePCR at conditions of increased template loads was repeatedly documented (15, 16). A broad range of concentrations (6 to 3,000 pM) corresponding to λ = 0.3 to 150 was used to compare the yields of ePCR and bulk PCR. Bulk PCR represents a special case of ePCR when λ equals the number of molecules in a test tube ($\lambda > 10^6$). The use of a variety of template concentrations and PCR conditions resulted in a low yield of the DNA library amplified by bulk PCR (Fig. 3*A*). In contrast, ePCR resulted in efficient amplification in the 60 to 600 pM template concentrations range (λ = 3 to 30). At λ < 1, ePCR resulted in reduced yield that must have been caused by a high proportion of empty droplets. On the other hand, at λ > 30 the reduction of DNA quality and ePCR yield was detected. Thus, amplifying a DNA library by ePCR outside the optimal λ ∼ 1 to 10 range leads either to low yields at λ < 1 or to a transition from single-molecule amplification to bulk effects at λ >> 10 (Fig. 3*B*).

The Mathematical Model Illustrating the Critical Advantages of ePCR. The details of the proposed mathematical description are given in *SI Appendix*, section S2. It was assumed that droplets are identical in size, and templates are distributed randomly between them. Hence, the number of template molecules in each droplet follows the Poisson distribution:

$$\frac{C_n}{C} = \frac{\lambda^n}{n!}e^{-\lambda},$$

where C is the total number of emulsion droplets, C_n is the number of droplets with n template molecules, and λ is the average number of template molecules in a droplet.

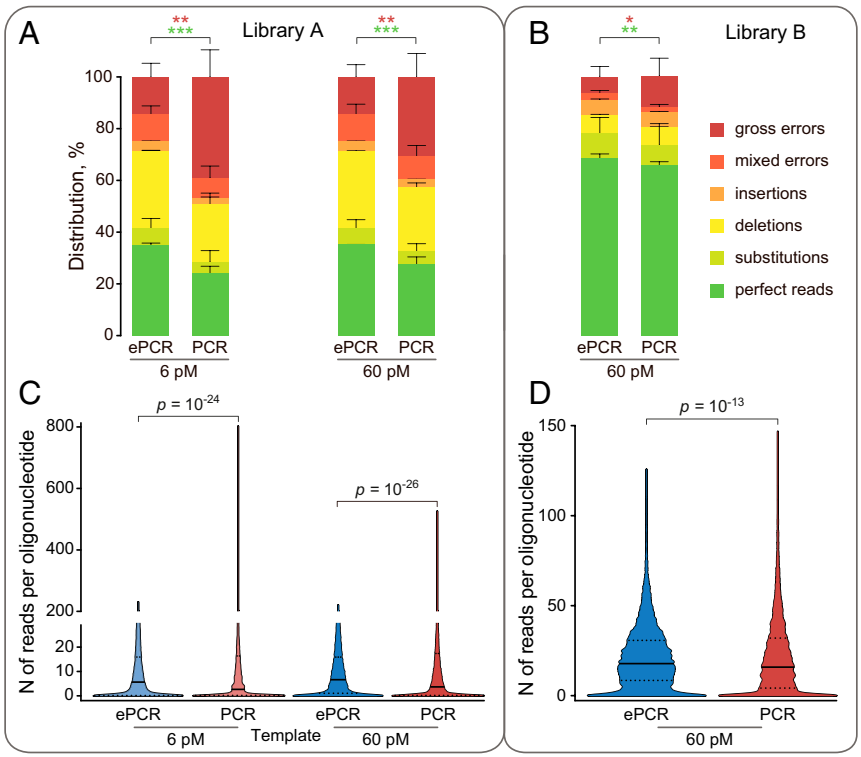

Fig. 2. ePCR results in a more uniform distribution of the amplified sequences than bulk PCR. The distribution of various types of errors observed in NGS reads of library A (*A*) or library B (*B*) amplified by ePCR and bulk PCR. All reads were divided into six groups: gross errors (burgundy), perfect reads (green), reads with substitutions only (light green), with deletions only (yellow), with insertions only (peach), and with mixed errors (orange). The distribution of errors was analyzed by *t* test; *P* values for perfect reads and gross errors are given. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. Data represent the mean of three biological replicates ± 95% CI. Violin plots represent distribution of reads for DNA library A (*C*) or library B (*D*) amplified by ePCR (blue) or bulk PCR (red). Different template concentrations (6 and 60 pM) were used for amplification of library A. A 60 pM of template was used for amplification of library B. Violin plots were obtained using 1.5×10^5 library A reads and 2.5×10^5 library B reads randomly sampled from the NGS data mapping to the library. Median (solid line) and interquartile range (dash line) are indicated. Wilcoxon rank-sum test *P* values between ePCR and bulk PCR are given.

It was assumed that the presence of two complementary template molecules in the same droplet results in PCR suppression and product loss with a constant probability denoted P_m. We speculate that the phenomenon of PCR suppression, mathematically described here by probability P_m, arises from mispairing and hybridization of complementary regions present in the DNA library (Fig. 4*A*). Therefore, the probability of appropriate template amplification in a droplet with n template molecules is estimated as $(1-P_m)^{n-1}$. The amount of ePCR product is calculated by the summation of the products of individual amplification reactions in all droplets:

$$M = \sum_{n=0}^{\infty} A_n \cdot n \cdot C_n \cdot (1-P_m)^{n-1},$$

where M is the total number of product molecules and A_n is the amplification coefficient, the number of product molecules obtained from a single template. Considering that the amplification reaction reaches a plateau after 35 cycles, and all templates are amplified with the same efficiency, we assume that $A_n = A/n$, where A corresponds to the amount of reagents in a single droplet. The summation with C_n given by the Poisson distribution results in the following expression:

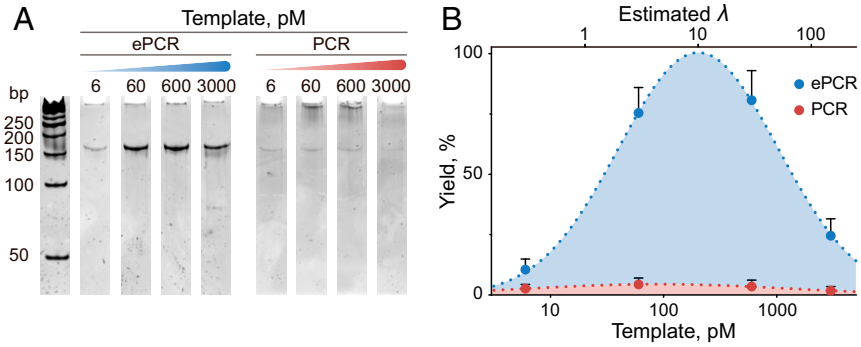

Fig. 3. The transition from single-molecule to bulk amplification of a complex DNA library by ePCR at different droplet occupancy. (*A*) Amplified DNA library obtained by ePCR and bulk PCR at different template concentrations. A representative gel obtained in one out of three independent experiments is shown. (*B*) The dependence of ePCR and bulk PCR yield on the concentration of the DNA template. The estimated average number of DNA templates encapsulated per droplet (λ) is indicated for ePCR. Data represent the mean from three independent experiments ± SD.

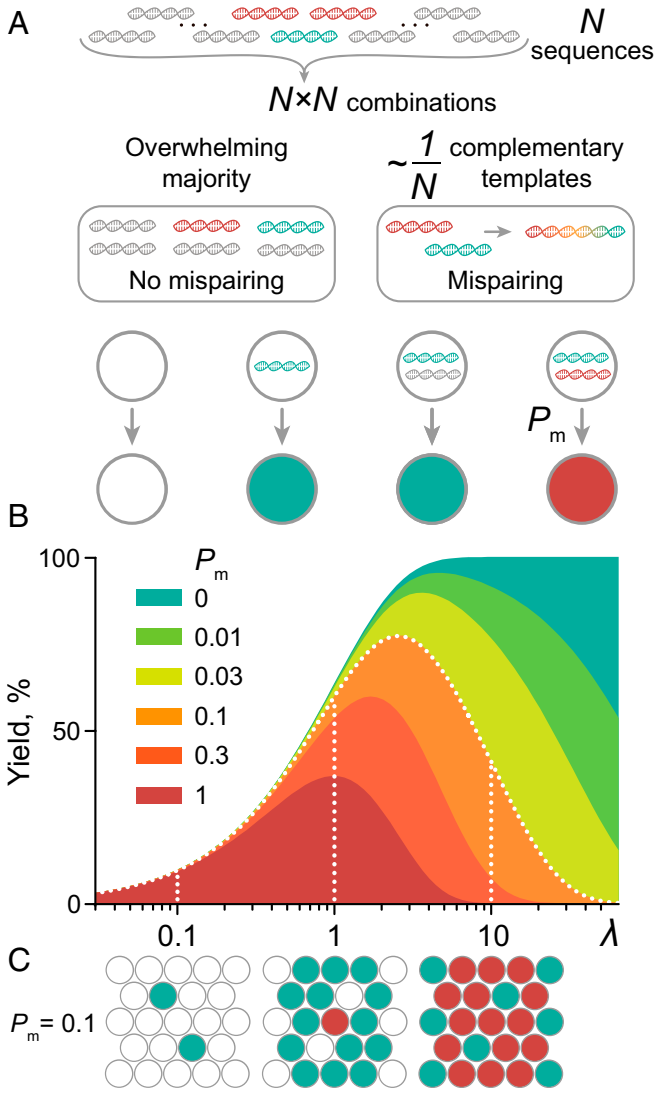

Fig. 4. The model of ePCR amplification of a complex DNA library. (*A*) A model DNA library of *N* overlapping sequences has at least ~1/*N* of complementary sequences. A representative DNA sequence, complementary DNA sequences, and noncomplementary DNA sequences are colored with aquamarine, red, and gray, respectively. PCR products are not produced in empty droplets. A single DNA molecule encapsulated in a droplet is always amplified. Multiple template molecules in a droplet are amplified (aquamarine) or face mispairing with probability $P_m \sim 1/N$, which results in the loss of PCR product (red). (*B*) The dependence of total yield of PCR product on the average number of DNA templates encapsulated in droplets (λ) for various P_m values. (*C*) Schematic illustration of the distribution of empty droplets, droplets with amplified templates, and droplets with lost PCR product for $P_m = 0.1$ at λ = 0.1, 1, and 10 from left to right, respectively.

$$Y = \frac{1}{1-P_m} \cdot \left(e^{-P_m \lambda} - e^{-\lambda}\right),$$

where Y is the yield of PCR amplification – the ratio of product molecules M to the total amount of available reagents AC. The ePCR yield demonstrates complex behavior, depending on P_m and λ (Fig. 4 *B* and *C*). When PCR suppression is absent and thus P_m equals 0, the ePCR product accumulation grows steadily with λ and asymptotically approaches 1 (a yield of 100%), meaning that every droplet with an encapsulated template gives a PCR product. Therefore, at high λ there is no difference between the yield of ePCR and bulk PCR.

On the contrary, $P_m = 1$ illustrates a very unfavorable case, when ePCR amplification can proceed exclusively in droplets with a single template molecule and every additional DNA template results in PCR suppression. In this case, the yield grows linearly with λ at small λ and then decays exponentially at large λ, according to the formula $Y = \lambda \cdot e^{-\lambda}$, reaching the maximum at $\lambda = 1$. Generally, the optimal parameter λ_{max} corresponding to the maximal yield is given by the formula

$$\lambda_{max} = \frac{1}{1-P_m} \cdot \ln \frac{1}{P_m}.$$

This expression has the limit $\lambda_{max} = 1$ at $P_m = 1$. When the suppression probability P_m decreases, the optimal λ grows very slowly under reagents' limiting conditions, making the optimal parameter λ_{max} remarkably insensitive to P_m. While it is difficult to predict the P_m value for a particular DNA library, especially if the initial distribution of DNA sequences is nonuniform, in our case it is possible to make an estimate. Assuming that a model DNA library has representativity N and contains adjacent sequences with complementary regions, we estimate that a portion of complementary templates is approximately $1/N$, and $P_m \sim 1/N$. In our model DNA libraries $P_m \sim 10^{-4}$ and, therefore, the predicted optimal λ is ~9, which correlates with experimental observations (Fig. 3 *A* and *B*).

Discussion

DNA libraries are indispensable for the state-of-the-art technologies in genomics, biotechnology, and combinatorial chemistry. Regardless of library origin, an amplification step is generally required for downstream manipulation. In addition, DNA libraries frequently contain either rare sequences (23, 24) or sequences prone to mispairing. The uniformity of amplification of these DNA library members is extremely important to avoid their loss. This is particularly relevant to quantitative analysis of metagenomes and amplification of DNA sequences selected via phage (25, 26), ribosome (27), and messenger RNA (28) display. Although different approaches to DNA library amplification were proposed (23, 29, 30), alternative methodologies are still in demand.

The benefits of transition from bulk PCR to ePCR amplification come to the fore in droplet digital PCR, which outperforms classical DNA quantification techniques, such as qPCR (31), and is extensively used for absolute DNA quantification (32–38). Similarly, ePCR provides a reproducible DNA libraries amplification platform, outperforming conventional bulk PCR. Moreover, it enables simple amplification of complex DNA templates that could not be efficiently amplified by bulk PCR. Hence, ePCR is particularly useful and should be chosen as the primary technique of choice for the amplification of complex DNA libraries.

To date, the advantages of ePCR over conventional PCR were confirmed mainly by visualizing the heterogeneity of amplified products using electrophoresis. Analysis of DNA quality based on gel electrophoresis alone could give controversial results in terms of the uniformity of amplification of individual library members. Hence, gel electrophoresis could not be regarded as a reliable technique for estimating the quality of library amplification. Here, the NGS analysis of DNA libraries, expanded by different amplification techniques, was performed. Primarily taking into account the NGS data, the ePCR protocol was optimized to get a simple amplification pipeline that does not require any additional equipment or skills but markedly improves the resulting library quality.

Successful ePCR needs to be compliant with the following guidelines carefully considered in this paper. First, emulsion should not coalesce during the amplification cycles. The stability of emulsion in ePCR is important because it enables one to accomplish a high number of amplification cycles and to reach

saturation of amplification in every template-containing droplet. Hence, a more uniform amplification is achieved. Second, the emulsification method should satisfy the following criteria: 1) be highly productive, 2) generate a monodisperse emulsion of small (~ 5 μm) droplets, 3) be PCR-compatible, and 4) be cost-effective. Thirdly, the occupancy of droplets should be optimized in order to achieve the maximal ePCR yield. Vortex-based emulsification using Abil EM 180 surfactant with ~60 pM template concentration satisfies these criteria and thus should be a technique of choice for robust ePCR amplification.

In order to provide a numerical description of the observed transition from bulk PCR to ePCR, a simple mathematical model of ePCR was developed. The model correctly describes critical advantages that ePCR has over conventional bulk PCR when it comes to amplifying complex libraries prone to template mispairing. It also accurately predicts optimal parameters for maximum yield of the amplified library. We surmise that virtually any highly representative complex DNA library will possess, to some extent, PCR-suppressive properties. As is shown here, even a tiny probability of PCR suppression causes nonmonotonic dependence of amplification yield on the amount of templates encapsulated for ePCR. Moreover, the optimal number of encapsulated templates per emulsion droplet lies in the quasi single-molecule range (1 to 10 DNA templates) and is remarkably insensitive to suppression probability and, therefore, the composition of a particular DNA library. Hence, the fundamental advantages of ePCR over bulk PCR originate from the quasi single-molecule amplification, providing the way to preserve the quality of DNA libraries during amplification.

Materials and Methods

Oligonucleotide Reagents. Oligonucleotide library A of size 12,004 oligos was synthesized by Custom Array (http://customarrayinc.com/) and oligonucleotide library B of size 11,973 oligos was synthesized by Twist Bioscience (https://www.twistbioscience.com/). The libraries were constructed as single-stranded 169-mer oligonucleotides with the following sequence: 5′-CCCAGCCGGCCATGGCC-$(N)_{132}$-GCTAGCAGTGGTGGAGGCGG-3′, where the central 132 nucleotides contained preassigned insert sequences (N) flanked by 5′ and 3′ library adapters. GC content was set to vary between 40% and 60%. No contiguous string of identical nucleotides longer than six nucleotides was allowed. Desalted single oligos used for amplification—forward primer (5′-CCCAGCCGGCCATGGCC-3′) and reverse primer (5′-CCGCCTCCACCACTGCTAGC-3′)—were provided by Evrogen.

Emulsification. The detailed optimized ePCR protocol is provided in *SI Appendix*, section S1. Briefly, oil phases based on Abil EM emulsifier or Span/Tween/Triton mix were used. Mineral oil was supplemented with either 3% of Abil EM 180 (Evonik) emulsifier (A-oil) or a mixture of 4.5% Span 80/0.4% Tween 80/0.05% Triton X-100 (Sigma-Aldrich) emulsifiers (T-oil). The aqueous phase contained dNTP mix, buffered polymerase, bovine serum albumin (BSA) supplement, and DNA library solution. All of the reagents were kept at 4°C before emulsification. The emulsification was conducted using a vortex mixer (Heidolph 541-10000-00) in 2-mL Eppendorf tubes for v conditions or a magnetic stirrer (Heidolph MR Hei-Tec 505-30000-00) using a 5- × 15-mm polytetrafluoroethylene magnetic stir bar in 5-mL round-bottom polystyrene test tubes (Falcon) for m conditions. The emulsification was provided with a 1:6 water:oil ratio for 5 min at 1,400 rpm and at 4 °C.

ePCR. The aqueous phase of ePCR was prepared as described below: 0.3 μM of forward and reverse primers, 200 μM dATP, 200 μM dCTP, 200 μM dGTP, 200 μM dTTP, 10 mg/mL BSA, 20 unit/mL Q5 DNA polymerase (NEB), 6 to 3,000 pM of ssDNA template were added into the Q5 Buffer to adjust total volume to 75 to 300 μL. PCR amplification was performed on Thermal Cycler T100 (BioRad). All PCR procedures were carried out under the following cycling conditions: 94 °C for 2 min, 25 to 35 cycles of 94 °C for 30 s, 58 °C for 30 s, and 72 °C for 30 s In optimizing experiments, the annealing temperature ranged from 52 °C to 62 °C, the concentration of Q5 DNA polymerase ranged from 10 to 50 units/mL, and the cycles ranged from 20 to 35.

Breaking the Emulsion. The emulsified PCR was pooled in a 1.7-mL microcentrifuge tube and centrifuged at 16,000 × *g* for 5 min at room temperature (RT). The upper (oil) phase was removed. To remove the surfactants mineral oil was added to each tube, firmly vortexed for 2 min, and centrifuged at 16,000 × *g* for 2 min at RT and the upper (solvent) phase was removed. Water-saturated diethyl ether was added to each tube, firmly vortexed for 2 min, and centrifuged at 16,000 × *g* for 2 min at RT and the upper (solvent) phase was removed. This procedure was repeated three times. For A-oil an additional washing step with ethyl acetate was performed. The lower phase was centrifuged at 16,000 × *g* for 10 min at 4 °C and the supernatant was transferred into the new tube. For the complete removal of solvent from the broken emulsion the tubes were incubated for 10 to 15 min with open caps at RT. The purified water phase was subjected to agarose gel and the band of the correct size was cleaned via PCR Clean-Up System (Evrogen).

NGS Library Preparation and Library Sequencing. The PCR product of the correct size was extracted from polyacrylamide gel electrophoresis. About 5 ng of the PCR product was ligated with adapters from NEBNext Multiplex Oligos (NEB) using NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). For quality control of DNA libraries, the High Sensitivity DNA Kit on Bioanalyzer 2100 (Agilent) was used. DNA libraries were quantified by KAPA Library Quantification Kit (Roche) on CFX96 Touch (Bio-Rad), pooled in equimolar amounts, and sequenced on Miseq using 2 × 300 bp paired-ends sequencing kit (Illumina) in the Siberian Branch of the Russian Academy of Sciences Genomics Core Facility (ICBFM SB RAS, Novosibirsk, Russia). All double-stranded DNA library amplification procedures with subsequent NGS library preparation were run in triplicate and showed high reproducibility.

Data Analysis. The script for sequence data analysis was written in Python and used the biopython package (39), seqtk toolkit (https://github.com/lh3/seqtk), and htslib library (40) with pysam wrapper (https://github.com/pysam-developers/pysam) for reading FASTA, FASTQ, and SAM files.

First, the 5′ library adapter (5′-CCCAGCCGGCCATGGCC-3′) with the linked 3′ library adapter (5′-GCTAGCAGTGGTGGAGGCGG-3′) and their reverse complementary sequences were identified in Illumina paired end reads with cutadapt (41) using a mild error rate threshold of 0.2 with no insertions or deletions allowed. In each Illumina read, the 5′ library adapter was mandatory and anchored, that is, it had to appear exactly at the start of the read. Otherwise, the read was discarded. Adapters were trimmed with cutadapt and the resulting insert sequences were kept for further processing.

For each read pair, the inserts derived from forward and reverse reads were merged into a single consensus sequence. For the positions in which paired reads had different nucleotides, the one with the higher quality was chosen. If only one read of the pair had adapters, it was used without merging. When the percentage of mismatches between paired reads was higher than 10%, the reads were considered unreliable and discarded. The reads having insert length 132 ± 3 nucleotides were used for mapping and error analysis.

The high-quality reads obtained via adapter trimming and sequence merging were mapped to the library of designed oligos with the minimap2 program (42). These read–olig pairs were used as inputs to obtain pairwise alignments by the Smith–Waterman algorithm implemented in the SSW library (43). The alignments were used to calculate the number of reads with different error types, namely having only insertions, only deletions or only mismatches, or reads with mixed errors. Since the main goal of the analysis was to compare different amplification techniques, the frequency of Illumina errors was considered to be identical in all samples. The alignments without insertions and deletions were used to calculate the library coverage statistics and the distribution of reads.

Data Availability. The Python script used for NGS data analysis and all data supporting the findings of this study are available from the corresponding authors upon request. All study data are included in the paper and *SI Appendix*.

ACKNOWLEDGMENTS. We thank Mr. Gregory Leleytner for his contribution to the computational design of the DNA libraries. We also thank St. Petersburg University for providing computational resources. This study was supported by Russian Science Foundation Grant 17-74-30019; Y.A.L. received personal financial support from The Russian Foundation for Basic Research Grant HP 17-04-01233 A; and Grant 18-29-08054 (to S.S.T. and I.V.S.).

1. C. Ralec, E. Henry, M. Lemor, T. Killelea, G. Henneke, Calcium-driven DNA synthesis by a high-fidelity DNA polymerase. *Nucleic Acids Res.* **45**, 12425–12440 (2017).
2. M. A. Quail *et al*., Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* **9**, 10–11 (2011).
3. S. Filges, E. Yamada, A. Ståhlberg, T. E. Godfrey, Impact of polymerase fidelity on background error rates in next-generation sequencing with unique molecular identifiers/barcodes. *Sci. Rep.* **9**, 3503 (2019).
4. J. Quick *et al*., Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).
5. E. E. Wrenbeck *et al*., Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* **13**, 928–930 (2016).
6. I. W. Deveson *et al*., Chiral DNA sequences as commutable controls for clinical genomics. *Nat. Commun.* **10**, 1342 (2019).
7. S. S. Terekhov *et al*., Microfluidic droplet platform for ultrahigh-throughput single-cell screening of biodiversity. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2550–2555 (2017).
8. S. S. Terekhov *et al*., Ultrahigh-throughput functional profiling of microbiota communities. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9551–9556 (2018).
9. S. S. Terekhov *et al*., Deep functional profiling facilitates the evaluation of the antibacterial potential of the antibiotic amicoumacin. *Antibiotics (Basel)* **9**, 157 (2020).
10. S. S. Terekhov *et al*., A kinase bioscavenger provides antibiotic resistance by extremely tight substrate binding. *Sci .Adv.* **6**, eaaz9861 (2020).
11. F. J. Ghadessy, J. L. Ong, P. Holliger, Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4552–4557 (2001).
12. R. Williams *et al*., Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* **3**, 545–550 (2006).
13. F. Diehl *et al*., BEAMing: Single-molecule PCR on microparticles in water-in-oil emulsions. *Nat. Methods* **3**, 551–559 (2006).
14. O. J. Miller *et al*., Directed evolution by in vitro compartmentalization. *Nat. Methods* **3**, 561–570 (2006).
15. K. Shao *et al*., Emulsion PCR: A high efficient way of PCR amplification of random DNA libraries in aptamer selection. *PLoS One* **6**, e24910 (2011).
16. M. Witt *et al*., Comparing two conventional methods of emulsion PCR and optimizing of Tegosoft-based emulsion PCR. *Eng. Life Sci.* **17**, 953–958 (2017).
17. R. Yufa *et al*., Emulsion PCR significantly improves nonequilibrium capillary electrophoresis of equilibrium mixtures-based aptamer selection: Allowing for efficient and rapid selection of aptamer to unmodified ABH2 protein. *Anal. Chem.* **87**, 1411–1419 (2015).
18. T. Sumida, H. Yanagawa, N. Doi, In vitro selection of fab fragments by mRNA display and gene-linking emulsion PCR. *J. Nucleic Acids* **2012**, 371379 (2012).
19. A. S. Adler *et al*., Rare, high-affinity anti-pathogen antibodies from human repertoires, discovered using microfluidics and molecular genomics. *MAbs* **9**, 1282–1296 (2017).
20. B. Wang *et al*., Functional interrogation and mining of natively paired human V_H:V_L antibody repertoires. *Nat. Biotechnol.* **36**, 152–155 (2018).
21. M. A. Turchaninova *et al*., Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43**, 2507–2515 (2013).
22. S. J. Spencer *et al*., Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME J.* **10**, 427–436 (2016).
23. A. Pinto, S. X. Chen, D. Y. Zhang, Simultaneous and stoichiometric purification of hundreds of oligonucleotides. *Nat. Commun.* **9**, 2467 (2018).
24. R. A. Hughes, A. D. Ellington, Synthetic DNA synthesis and assembly: Putting the synthetic in synthetic biology. *Cold Spring Harb. Perspect. Biol.* **9**, a023812 (2017).
25. D. Mohan *et al*., Publisher correction: PhIP-seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. *Nat. Protoc.* **14**, 2596 (2019).
26. Y. A. Lomakin *et al*., Probing surface membrane receptors using engineered bacteriophage bioconjugates. *Bioconjug. Chem.* **30**, 1500–1506 (2019).
27. I. Zimmermann *et al*., Generation of synthetic nanobodies against delicate proteins. *Nat. Protoc.* **15**, 1707–1741 (2020).
28. Matilda S. Newton, Yari Cabezas-Perusse, Cher Ling Tong, Burckhard Seelig, *In vitro* selection of peptides and proteins-advantages of mRNA display. *ACS Synth. Biol.* **9**, 181–190 (2020).
29. J. C. Klein *et al*., Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2016).
30. S. Kosuri, G. M. Church, Large-scale de novo DNA synthesis: Technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
31. S. C. Taylor, G. Laperriere, H. Germain, Droplet digital PCR versus qPCR for gene expression analysis with low abundant targets: From variable nonsense to publication quality data. *Sci. Rep.* **7**, 2409 (2017).
32. K. R. Sreejith, C. H. Ooi, J. Jin, D. V. Dao, N.-T. Nguyen, Digital polymerase chain reaction technology–Recent advances and future perspectives. *Lab Chip* **18**, 3717–3732 (2018).
33. M. Postel, A. Roosen, P. Laurent-Puig, V. Taly, S.-F. Wang-Renault, Droplet-based digital PCR and next generation sequencing for monitoring circulating tumor DNA: A cancer diagnostic perspective. *Expert Rev. Mol. Diagn.* **18**, 7–17 (2018).
34. I. Giovannelli *et al*., Utility of droplet digital PCR for the quantitative detection of polyomavirus JC in clinical samples. *J. Clin. Virol.* **82**, 70–75 (2016).
35. M. Massanella, S. Gianella, S. M. Lada, D. D. Richman, M. C. Strain, Quantification of total and 2-LTR (long terminal repeat) HIV DNA, HIV RNA and herpesvirus DNA in PBMCs. *Bio Protoc.* **5**, e1492 (2015).
36. D. Pekin *et al*., Quantitative and sensitive detection of rare mutations using droplet-based microfluidics. *Lab Chip* **11**, 2156–2166 (2011).
37. A. B. Košir, B. Spilsberg, A. Holst-Jensen, J. Žel, D. Dobnik, Development and inter-laboratory assessment of droplet digital PCR assays for multiplex quantification of 15 genetically modified soybean lines. *Sci. Rep.* **7**, 8601 (2017).
38. S. A. Byrnes *et al*., Simple polydisperse droplet emulsion polymerase chain reaction with statistical volumetric correction compared with microfluidic droplet digital polymerase chain reaction. *Anal. Chem.* **90**, 9374–9380 (2018).
39. P. J. A. Cock *et al*., Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
40. H. Li *et al*.; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
42. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
43. M. Zhao, W.-P. Lee, E. P. Garrison, G. T. Marth, SSW library: An SIMD smith-waterman C/C++ library for use in genomic applications. *PLoS One* **8**, e82138 (2013).