Quantitative Biology 2016, 4(1): 36–46
DOI 10.1007/s40484-016-0064-3

Review

Combinatorial pooled sequencing: experiment design and decoding

Chang-chang Cao and Xiao Sun*

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China
* Correspondence: xsun@seu.edu.cn

Received October 29, 2015; Revised January 3, 2016; Accepted January 22, 2016

Owing to rapid advances in the next-generation sequencing technology, the cost of DNA sequencing has been reduced by over several orders of magnitude. However, genomic sequencing of individuals at the population scale is still restricted to a few model species due to the huge challenge of constructing libraries for thousands of samples. Meanwhile, pooled sequencing provides a cost-effective alternative to sequencing individuals separately, which could vastly reduce the time and cost for DNA library preparation. Technological improvements, together with the broad range of biological research questions that require large sample sizes, mean that pooled sequencing will continue to complement the sequencing of individual genomes and become increasingly important in the foreseeable future. However, simply mixing samples together for sequencing makes it impossible to identify reads that belongs to each sample. Barcoding technology could help to solve this problem, nonetheless, currently, barcoding every sample is costly especially for large-scale samples. An alternative to barcoding is combinatorial pooled sequencing which employs pooling pattern rather than short DNA barcodes to encode each sample. In combinatorial pooled sequencing, samples are mixed into few pools according to a carefully designed pooling strategy which allows the sequencing data to be decoded to identify the reads that belongs to the sample that are unique or rare in the population. In this review, we mainly survey the experiment design and decoding procedure for the combinatorial pooled sequencing applied in rare variant and rare haplotype carriers screening, complex genome assembling and single individual haplotyping.

Keywords: combinatorial pooled sequencing; experiment design; decoding

INTRODUCTION

Over the past decade, owing to rapid advances in the next-generation sequencing technology, the cost of DNA sequencing has been reduced by over several orders of magnitude. Especially, the release of Ilumina's HiSeq X Ten System has reduced the cost of sequencing a human genome to $1000 [1]. The reduction in the costs of DNA sequencing is democratizing the extent to which individual investigators can pursue projects at a scale previously accessible only to major genome centers [2,3]. Producing tens of thousands of genomes, or so-called 'factory-scale' sequencing will revolutionize the study of population diversity and help us to understand the genetic basis of health and disease better [1].

Nevertheless, given the factory-scale, it is obvious that many research questions cannot be addressed by whole-genome sequencing of individuals despite the plunging cost for sequencing [4,5]. The main challenge exists in individually amplifying and creating sequencing libraries for thousands of samples. To efficiently use the capacity of sequencer and reduce the cost of sequencing library construction for large-scale sequencing, multiple individuals could be pooled together and sequenced, called pooled sequencing (pool-seq). Pool-seq could provide a cost-effective alternative to sequencing individuals separately, since pool-seq uses a single library for the entire sample, whereas sequencing of individuals requires a separate library to be prepared for each sample [4,6].

Pool-seq could save tremendously on sample preparations, especially for targeted sequencing projects, since the cost for target capturing is proportional to the number of samples (i.e., number of individuals without pooling

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

vs. number of pools in pool-seq) [7]. The main limitation of the naive pool-seq strategy is its inability to obtain the information for each individual sample participated in the pool. However, multiplexing using sequencing barcodes could overcome the drawback, where the DNA in each sample is cut into short fragments suitable for sequencing and ligated with a short, sample-specific DNA sequence i.e. barcode [8]. After sequencing, reads belonging to each individual could be assigned precisely based on the barcode signature.

Studies have showed that pool-seq is often more effective in SNP discovery and could provide more accurate allele frequency estimates at a lower cost than sequencing of individuals, even when taking sequencing errors into account [5,9]. Hence, pool-seq has been widely applied in researches which involve large scale samples and relies on the precise allele frequency, such as genome-wide association studies (GWAS), population genetics, reverse ecology, genome evolution studies [4].

Due to the effectiveness of pool-seq in cost and allele frequency estimation, pool-seq has been applied in a broad range of applications. The majority of these applications rely on sequencing large pools of individuals from multiple populations or generations. Although with further reductions in sequencing costs, pool-seq will remain an important tool for researches requiring adequate sample sizes. With the availability of new software packages for the analysis of pool-seq data, pool-seq will be an even more attractive research tool in the future [4].

However, original pool-seq is unable to identify reads that belongs to each sample. Although DNA barcodes make it possible to distinguish reads from different samples, barcoding every sample remains very costly currently. In 2009, Patterson *et al*. proposed the concept of the combination between combinatorics and pooled sequencing [6], defined as combinatorial pooled sequencing which allows the sequencing results to be decoded to identify the reads belonging to samples that are unique or rare among the population without barcodes. Using ideas from a branch of mathematics called combinatorics, thousands of samples are pooled and sequenced at the same time in the combinatorial pooled sequencing. In detail, samples are mixed into few pools according to a carefully designed pooling strategy where the pooling patterns instead of DNA barcodes are used to tag samples.

At present, combinatorial pooled sequencing has been utilized in many applications, such as identifying rare variants carriers and rare haplotype carriers [6,10], assembling complex genome [11], single individual haplotyping [12], sequencing of multiple viral samples [13]. Since the application and limitations of pool-seq are summarized in many articles, in this review, we particularly focus on the experiment design and decoding procedure for the combinatorial pooled sequencing applications.We will also discuss the further directions and limitation for the combinatorial pooled sequencing.

COMBINATORIAL POOLED SEQUENCING

As described previously, the objective of combinatorial pooled sequencing is to obtain reads specific to individuals that are unique or rare among the population such as rare variant carriers. In the combinatorial pooled sequencing, samples are mixed into few pools according to a carefully designed pooling strategy where the pooling patterns instead of DNA barcodes are used to tag each sample. And combinatorial pooled sequencing allows the sequencing results to be decoded to identify the reads that belongs to samples that are unique or rare in the population. Hence, combinatorial pooled sequencing involves two more steps than normal pool-seq: encoding and decoding (Figure 1) [6]. The encoding step refers to the design of the pooling strategy which should guarantee that pooling pattern for each sample is distinct to each other. While the decoding procedure is utilized to analyze

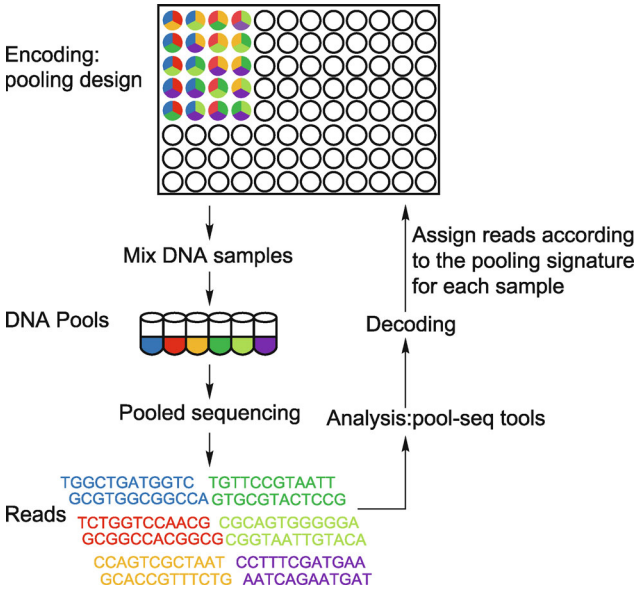

Figure 1. The combinatorial pooled sequencing involves two more steps than normal pool-seq project: encoding and decoding. Individual DNA samples are placed in multiple-wells plates. The encoding step refers to the design of the pooling strategy. Different colors denote distinct pools. Samples are mixed into pools according to the color code (i.e., pooling pattern in the combinatorial pooled sequencing) and samples with multiple colors are mixed into multiple pools. The decoding procedure is utilized to analyze the pooled sequencing results and obtain the reads belonging to samples that are unique or rare in the population, according to the pooling signature for each sample. The colors of the reads match the colors of the pools, denoting the distinct sets of reads that are produced by each pool.

the pooled sequencing results to obtain the reads that belong to each sample according to the unique pooling pattern for each sample.

Currently, many pooling strategy in the combinatorial pooled sequencing derive from the non-adaptive pooling design in the field of group testing [14], such as DNA Sudoku design [15], the shifted transversal design (STD) [16] and some random pooling designs [14]. The concept of group testing can date back to World War II, which was first proposed by Dorfman [17], for the problem of determining which blood samples contain the syphilis antigen for numerous soldiers. Besides, samples can also be mixed into multiple pools in a manner so as to create a code (a unique set of pools for each individual) and the unique code (pooling pattern) is used to tag each individual. Hence, code words from the theory of constructing error-correcting codes have been applied in creating the pooling strategy [18]. In the above pooling strategies, samples are mixed into multiple pools to guarantee unique pool signatures for each sample. Alternatively, samples could also be mixed into only one pool. Taking advantage of extra information, reads belong to distinct samples in the same pool could also be separated without barcodes [12].

In theory, decoding algorithms for non-adaptive pooling designs in the group testing field can be utilized directly to analyze the results of combinatorially-pooled sequencing as long as these algorithms match the pooling design used to construct the pools [19]. Techniques from the field of compressed sensing [20,21], designed for efficiently reconstructing a sparse signal by finding solutions to underdetermined linear systems, have also been applied in the combinatorial pooled sequencing for the decoding procedure.

In general, combinatorial pooled sequencing can utilize few pools without barcoding specimens in the pools to obtain reads belonging to samples that are unique or rare in the population. Compared with sequencing individual samples independently, combinatorial pooled sequencing could vastly reduce the efforts for sequencing library preparation as well as the cost. Besides, pools could be mixed again by employing few DNA barcodes to further reduce the sequencing cost.

In the following section, we will mainly introduce the experiment design and decoding procedure for the combinatorial pooled sequencing applications.

EXPERIMENT DESIGN AND DECODING FOR COMBINATORIAL POOL-SEQ APPLICATIONS

Rare variant carriers screening

Rare variants are responsible for a large portion of the heritability of some common complex human diseases [22,23]. Genome-wide association studies have begun to focus on the contribution of variants of low minor allele frequency (MAF 0.5%–5%), or of rare variants (MAF $< 0.5\%$) [23]. The functional and evolutionary impacts of rare variants have been reported; therefore, large-scale screening for disease-associated rare variants becomes increasingly important for investigating target biology and drug response, providing clinically actionable information [24,25].

Because of the extremely low frequency of rare variants, sample sizes must be large enough to guarantee efficient observations. As described previously, combined with multiplexing using sequencing barcodes, pool-seq could help for screening rare variant carriers. However, at present, barcode sequencing is still costly not only because of limited barcoding capacity but also due to the per sample cost for barcode preparation and ligation. To further reduce the cost of large-scale screens for rare variant carriers, several groups used techniques from the group testing theory [14,26] and compressed sensing [20,21] to construct a kind of combinatorial pooled sequencing called overlapping pool sequencing (OPS) to identify rare variant carriers efficiently. In summary, OPS allocate individual samples into a small number, but different pools and the identity of each sample is encoded within the pooling pattern rather than by its association with a particular DNA barcode. Accordingly, samples that carry variants could be identified based on their pool signatures.

An OPS design with n samples and t pools is associated with a $t \times n$ binary matrix $M = \{m_{ij}\}$, in which the rows are indexed by t pools $A_1, \ldots, A_t \subset \text{Set}\{1, \ldots, n\}$, the columns are indexed by n samples $\{1, \ldots, n\}$, and $m_{ij} = 1$ if and only if the j^{th} sample is contained in the i^{th} pool. A simple OPS design for identifying at most 1 variant carrier among 7 samples is shown in Figure 2. The most critical step in the OPS is the encoding procedure i.e. the design of pooling matrix M which should guarantee distinct pool signature for each sample. After pool-seq, pools that contain variants could be identified precisely. In the decoding procedure, variant carriers identification can be translated as sparse vector recovery on the basis of pooling matrix (M) and pool-seq results (Y). And fortunately, many algorithms from the field of group testing [14,26] and compressed sensing [20,21] are available for finding the rare variant carriers.

The first OPS design, presented by Erlich *et al*. [15], employed the DNA Sudoku design to mix samples. The DNA Sudoku design is constructed based on the Chinese remainder theorem and guarantees that the intersection number of each column in the pooling matrix M is equal to one, meaning that no two samples will be jointly placed in more than one pool. In theory, the number of pools in

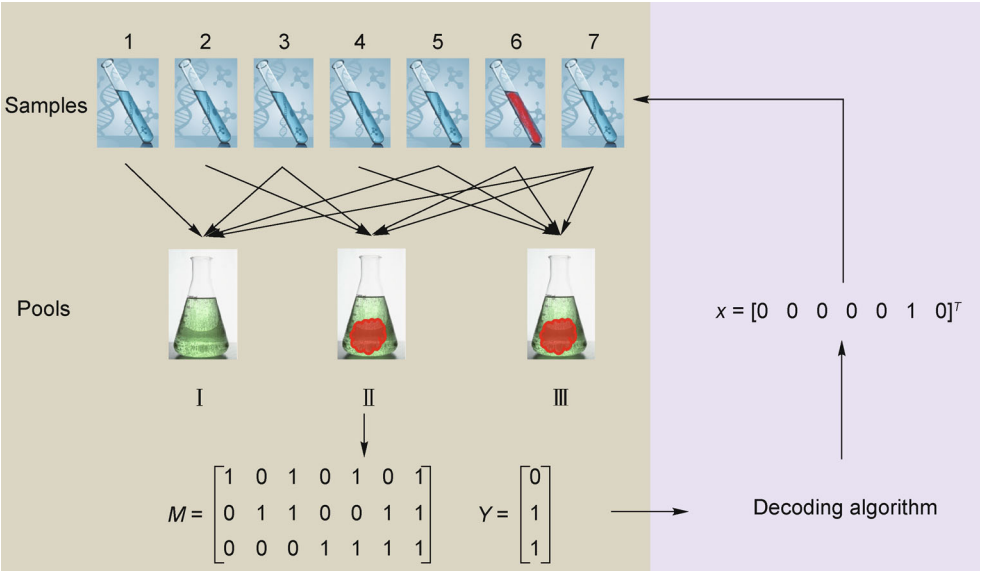

Figure 2. A simple overlapping pool sequencing design for identifying atmost 1 variant carrier among 7 samples. The left part indicates the encoding procedure and the right part indicates the decoding procedure. This design consists of three pools: {1, 3, 5, 7}, {2, 3, 6, 7} and {4, 5, 6, 7}. After pooled sequencing for these three pools, sequencing results could be represented by the vector $[0, 1, 1]^T$ (Y) where pools with the variant are denoted as 1 and pools without the variant are denoted by 0. In the decoding procedure, variant carriers identification can be translated as sparse vector recovery on the basis of pooling matrix (M) and sequencing results (Y). We can simply infer the sample vector x as $[0, 0, 0, 0, 0, 1, 0]^T$ where 1 denotes the variant carrier and 0 denotes normal individual *i.e.* the 6^{th} sample is the variant carrier.

DNA Sudoku design is a small multiple of the square root of the number of samples. As long as the pools containing variants can be identified precisely, utilizing decoding methods from the fields of non-adaptive group testing [19], samples that carry the rare variants could be inferred with very high confidence. Erlich *et al*. implemented DNA Sudoku design for screening carriers of the variants associated with cystic fibrosis among 18,000 individuals, and achieved greater than 97% accuracy with about only 2,700 pools where the frequency of the variants carriers was ~1.8%.

Prabhu and Pe'er [18] constructed an OPS design on the basis of code words from the theory of error-correcting codes. In their design, individuals are assigned to pools in a manner so as to create a code: a unique set of pools for each individual. This set of pools to which an individual is assigned defines a code word, or pool signature. Theoretically, the design requires ~log(n) pools. In the decoding procedure for discovering a variant carrier, the set of pools where the variant is observed reveals the identity of its carrier. Prabhu and Pe'er demonstrate the ability of their designs in extracting rare variant carriers using short read data from the 1,000 Genomes Pilot 3 project. With practical parameters of sequencing studies, their designs guarantee high probability of variant carrier identification. However, the design was only able to detect and identify "singletons" — a variant occurring in just one sample — with very good reliability [6].

Combinatorial pooling design in group testing field [14] could be applied in OPS strategies directly. In practice, many pooling designs are capable of detecting with high probability more variant carriers than guaranteed [16,27]. Hence, substantial simulations must be performed to figure out how many variant carriers and errors that a design can actually deal with. Incorporating weighting into existing pooling designs has also been shown that have enough power to facilitate detection of even the common alleles carriers, where different individuals donate different amounts of DNA to the pools [28]. However, it is hard to obtain exactly the same amounts of DNA as designed and differences between the actual and designed amounts may prevent the experiment from the accurate decoding [29].

Basically, decoding algorithm from group testing theory [19] can be used directly to identify these rare variant carriers based on the pooled sequencing results as long as pools that contain variants could be distinguished precisely from those that do not. In order to address these issues specific to pool-seq, Wang *et al*. [7] proposed *VIP*, a complete data analysis framework for overlapping pool sequencing designs, with novelties in variant pool and variant locus identification and variant allele frequency estimation. Meanwhile, *VIP* also improved the variant

sample decoding procedure by taking sequencing errors into consideration.

Considering the fact that the number of reads covering variants could reflect the number of samples that carry variants in each pool and drawing inspiration from compressed sensing, Erlich *et al*. employ a compressed genotyping protocol to identify carriers of a severe genetic disease which occupy only a small fraction of the population [30,31]. In particularly, Erlich *et al.* presented light-weight designs based on the Chinese Reminder Theorem to pooled samples and proposed a Bayesian reconstruction approach based on belief propagation to identify the variant carriers. Similarly, Shental *et al*. put forward a novel pooling design that enables the recovery of novel or known rare alleles and their carriers in groups of individuals by utilizing techniques from compressed sensing filed [29]. In Shental's research, random pooling matrices were employed to mix samples where each pool consists of a random subset of individuals. Subsequently, the commonly used gradient projection for sparse reconstruction (GPSR) algorithm [32] in the field of compressed sensing was chosen to find the variant carriers. Cao *et al*. proposed an efficient OPS strategy on the basis of quantitative group testing [33]. In their study, random *k*-set pool designs [34] were used to mix samples, and the design parameters could be optimized according to an indicative probability. Next, using the quantitative information contained in the sequencing results, Cao *et al*. designed a heuristic Bayesian probability decoding algorithm to identify variant carriers. With the simulated pools and publicly available Illumina sequencing data, Cao's method correctly identified the carriers for 91.5%–97.9% variants with the variant frequency ranging from 0.5% to 1.5% [33].

Currently, the majority of these designed overlapping pool sequencing strategies mainly focused on screening carriers of rare variants. However, utilizing the imputation information, He *et al*. [35] and Hormozdiariy *et al*. [36] revealed that OPS could also be applied to genotype the common SNPs and obtain more accurate genotypes.

OPS has already been applied in screening individuals with rare variants associated with lung cancer [37] and bicuspid aortic valve [38]. OPS could save tremendously on sample preparations for screening rare variant carriers since the number of pools involved is much less than the number of individual samples. However, the cost for the data production cannot be neglected due to the adequate sequencing depth required for efficient observation of variants in the pools. Cao *et al*. construct the cost model for OPS which covers both library preparation and data production, and suggest that optimal OPS design should be chosen to not only minimize the cost but also guarantee correct decoding procedure [39].

Identify rare haplotype carriers

Most variants identified by GWAS so far confer relatively small increments in risk, and explain only a small proportion of heritability [23]. Studies also have begun to focus on the values of haplotypes and genome wide haplotype association (GWHA) studies have been presented to find the association between haplotypes and diseases [40]. The combination of genetic marker alleles such as SNPs on a single chromosome is called a haplotype which provides valuable information on evolutionary history and may lead to the development of more efficient strategies to identify genetic variants that increase susceptibility to diseases [41]. Rather than examining variants independent of each other, simultaneously considering the values of multiple variants within haplotypes can improve the power of detecting associations with disease [42]. Up to now, numerous rare haplotypes have been found associated with several diseases [43–46]. Therefore, identifying rare haplotype carriers is also very valuable in diseases studies.

Similarly to identifying rare variant carriers, as long as the haplotype frequencies can be estimated precisely for pool-seq, indicating that pools containing rare haplotypes can be recognized correctly, it is feasible to apply overlapping pool sequencing directly in screening for rare haplotype carriers. The only one obstacle needing to be solved is to estimate haplotype frequencies accurately from DNA pools. Fortunately, Long [47] presented *PoolHap* to estimate haplotype frequencies from pooled samples by next-generation sequencing. Based on an expectation maximization algorithm, Kessner [48] put forward *Harp* to calculate the frequencies of haplotypes from pool-seq data. However, both *PoolHap* and *Harp* required that the haplotypes for pooled samples were known for inferring the frequency which may conflict with real situation. Fortunately, several methods that take advantage of haplotype database information have showed great potential in haplotype frequency estimation [49,50]. By means of databases that contain prior haplotypes, Cao *et al*. translated the problem of estimating frequency for each haplotype into finding a sparse solution for a system of linear equations and put forward *Ehapp* to estimate the frequencies [10]. Results showed that *Ehapp* could estimate the frequencies of haplotypes with only about 3% average relative difference for pooled sequencing of the mixture of 10 haplotypes with total coverage of 50×. Comparisons on simulated data in conjunction with publicly available Illumina sequencing data indicate that *Ehapp* is state of the art for many sequencing study designs [10]. Nevertheless, *Ehapp* is sensitive to high sequencing error rate since base qualities are not taken into consideration in *Ehapp* [10].

Once a haplotype is confirmed that has association with diseases, similar to genomic variants, screening for the haplotype carriers is of great value in the practical application. On the basis of accurate estimation of haplotype frequency from pool-seq data, Cao *et al*. revealed that it is feasible to apply OPS directly to identify rare haplotype carriers cost-effectively by conducting a simulation experiment to identify 2 heterozygous carriers for an assigned haplotype among 100 simulated diploid individuals [10]. Finally, OPS with the DNA Sudoku design [15] and shifted transversal design (STD) [16] required 34 and 33 pools respectively to find 2 specified haplotype carriers precisely among 100 diploid samples. Hence, the sequencing library required for screening rare haplotype carriers is vastly reduced. Therefore, OPS showed great potentiality in reducing the sequencing library as well as the cost for screening rare haplotype carriers.

Assemble complex genomes

Although the next-generation sequencing technology allow researchers to obtain whole genome sequencing data and assemble the genome, the short reads length and substantial repetitive sequence in the genome hinder the application of the next-generation sequencing technology in assembling complex genomes such as large plant genome [51]. Alternatively, sequencing clone-by-clone has been shown feasible for large complex genomes assembly [52].

To obtain the sequence for each clone to assemble a complex genome, using DNA barcodes is unpractical due to the huge number of clones. In order to overcome the limitations imposed by using DNA barcodes when multiplexing a large number of clones, Lonardi *et al*. proposed a clone-by-clone sequencing protocol that employed combinatorial pooled sequencing to efficiently recover reads for each clone and achieve *de novo* selective genome sequencing [11,53].

In summary, Lonardi's protocol involves seven essential steps:

A. Obtain a clone library for the target individual;

B. Fingerprint clones and build a physical map;

C. Select a minimum tiling path (MTP) from the physical map [54,55];

D. Pool the MTP clones according to the shifted transversal design (STD) [16];

E. Sequence the DNA in each pool and filter reads with low quality;

F. Assign reads to clones (decoding);

G. Assemble reads clone-by-clone using a short-read assembler.

The novelty of Lonardi's approach is that hundreds of clones are pooled using STD and then sequenced. Pool seq can take advantage of the throughput of the current generation of sequencing instruments. Meanwhile, STD compares favorably, in terms of efficiency and robustness to errors, to the previously described combinatorial pooling designs and allows the identification of rare events among large-scale samples [16]. Since at most few MTP clones overlap for the most realistic scenario, meaning that each pool-sequenced read is specific to very few clones i.e. rare events among all these MTP clones, Lonardi's protocol allows the decoding procedure to obtain reads specific to each alone [11]. Accordingly, the whole genome assembly could be achieved by assembling reads clone-by-clone using a short-read assembler.

The most critical step in Lonardi's protocol is to assign pooled sequencing reads to each clone i.e. decoding. Initially, Lonardi *et al*. presented *HashFilter* which assign reads based on the pool signatures of all the *k-mer* strings in each reads [11]. After decoding, hundreds of millions of short reads were assigned to the clones so that the assembly can be carried out clone-by-clone. Experimental results on simulated data for the rice genome showed that the assignment was extremely accurate (99.57% of the decoded reads are assigned to the correct BAC), and the resulting BAC assemblies had very high quality (BACs were covered by contigs over about 77% of their length, on average).

Since Lonardi *et al*. pooled clones according to STD which is a Reed-Solomon based pooling design, to achieve the highest possible decoding accuracy, Duma *et al*. proposed a decoding approach that combined ideas from the fields of compressive sensing and decoding of error-correcting codes [56]. Given the result of pooled sequencing of genomic BAC clones, Duma *et al*. cast the problem as a compressed sensing problem where the unknowns are the assignments of the reads. Additionally, overlap between reads and mate pair information could also be used to further improve the accuracy of the decoding. Experimental results on synthetic data for the rice genome and real data for the barley genome showed that this decoding algorithm enables significantly higher quality assemblies than *HashFilter* [56].

Single individual haplotyping

Haplotype phase information in diploid organisms provides valuable information on human evolutionary history and may lead to the development of more efficient strategies to identify genetic variants that increase susceptibility to human diseases [41]. At present, individual haplotyping are achieved mainly by three strategies: genetic analysis, population inference and molecular haplotyping.

Genetic analysis based haplotyping requires pedigrees and can yield chromosome length haplotypes which are

highly accurate and complete. Nevertheless, genetic analysis is not always feasible to recruit the required participants for family-based studies and cannot phase positions in which all family members are heterozygous [57]. Population inference requires models of the population structure of haplotypes and relies on the availability of population-matched reference haplotype datasets which are needed to be large enough to sample rare variants. Moreover, population-based phasing methods are limited to generation of short haplotype blocks and will incorrectly phase rare combinations of variants, where those combinations are likely to be medically important [57].

Molecular haplotyping involves the direct observation of alleles on a single molecule. These molecules are often single sequence reads, ranging in size from tens of bases to thousands of bases. Whole chromosome haplotype phasing has been achieved by fluorescence activated cell sorting [58], chromosome segregation [59], microdissection based sequencing [60] and proximity-ligation sequencing using the Hi-C protocol [61]. However, these methods only phase a fraction of the heterozygous variants in an individual, and more importantly, they are technically challenging to perform or require specialized instruments [62].

Recently, a kind of single individual haplotyping methods based on combinatorial pooled sequencing have also been presented [12,63–67]. The basic principle behind these methods involves constructing pools containing DNA fragments that constitute only a few haploid copies of genomic DNA. Taking DNA fragments as samples, the encoding procedure for the combinatorial pooled sequencing used in these single individual haplotyping methods refers to mixing limited long genome fragments to a pool which should guarantee that any position of the genome is likely to be represented by haploid DNA. In the decoding step, each sample (DNA fragment) in the pool could be computationally reconstructed by identifying regions of enriched coverage after mapping pool-seq reads back to reference genome [12]. Accordingly, reconstructed fragments representing local haplotypes could be joined together to form the whole chromosome haplotypes.

Kitzman *et al*. [63], Suk *et al*. [64] and Lo *et al*. [12] construct pools that contain few haploid copies by pooling limited number of fosmids or BAC clones from the clone library for the individual. In general, given a clone library for an individual, clones are grown in separate wells on plates. Briefly, using a liquid handling robot, pools are constructed by combining clones from limited number of plates. For instance, in Lo's research [12], 5,376 clones from fourteen 384 well plates are mixed to form a pool which constitute about only 0.25 haploid copies. Hence, any position of the genome is likely to be covered by only one clone. After pool-seq and mapping reads, each clone in the pool could be computationally reconstructed by identifying regions of enriched coverage even without barcodes. However, an obvious shortcoming of these strategies is that large-insert cloning is technically challenging and not readily scalable [62].

To overcome this limitation, Peters *et al*. [65] and Kaper *et al*. [66] developed another pooling strategy by diluting sheared genome DNA fragments into plenty of pools to ensure that each pool has few haploid copies. Subsequently, DNA template in each pool is independently amplified using multiple displacement amplification (MDA) to generate sufficient input material for sequencing library preparation. Similar to pooled clone sequencing, reconstructed segments in each pool representing local haplotypes were next used for whole genome haplotyping. However, the challenge of consistent dilution of genomic DNA to each pool and non-uniform representation of the fragments introduced by MDA may hinder the application of this kind of strategy [62].

Alternative implementations of these haplotyping methods based on combinatorial pooled sequencing mainly differ in how pools are generated. The configurations for generating pools in different combinatorial pooled sequencing based haplotyping methods are summarized in Table 1. In general, the accuracy and resolution of the reconstructed haplotypes increase with length of fragments and the number of pools, and decrease with the number of haploid copies per pool as well as the number of fragments per pool.

To address some limitations in these haplotyping methods based on combinatorial pooled sequencing, Amini *et al*. [67] reported a distinct method for direct haplotyping, termed contiguity-preserving transposition

Table 1. Configuration for generating pools in different pool-seq based haplotyping methods.

Refs.	# of fragments per pool	Length of fragments (kbp)	# of haploid copies per pool	# of pools
Lo *et al*. [12]	5,000	140	0.25	24
Kitzman *et al*. [63]	5,000	37	0.06	115
Suk *et al*. [64]	5,000	40	0.07	288
Peters *et al*. [65]	5,000 –10,000[a]	60	0.15	384
Kaper *et al*. [66]	16,377[b]	13.8	0.075	192

[a, b] Estimated value based on the length of fragments and the number of haploid copies in the pool.

sequencing (CPT-seq). Combinatorial 96-plex indexing at both the transposition and PCR stage enables the construction of 9,216 'virtual pools'. CPT-seq employed large effective number of virtual compartments per physical compartment and could avoid the amplification biases associated with MDA [62].

CONCLUSION AND DISCUSSION

Pool-seq could take advantage of the capacity of sequencer and save tremendously on the cost of sequencing library construction for large-scale sequencing, especially for targeted sequencing projects. Utilizing DNA barcodes, each sample in the pool is ligated with a short and sample-specific DNA sequence, the information for each individual sample participating in the pool could be recovered from the pooled sequencing results. Furthermore, pool-seq is often more effective in SNP discovery and could provide more accurate allele frequency estimates at a lower cost than sequencing of individuals even when taking sequencing errors into account, making pool-seq appealing in various genetics researches such as GWAS, population genetics, reverse ecology, and genome evolution. With the tremendous increase in the throughput for the massive parallel sequencing instruments and the availability of advanced new algorithms, pool-seq will become more attractive in the foreseeable future.

However, original pool-seq is unable to identify reads that belongs to each sample. Barcoding remains very costly for distinguishing reads from different individuals especially for large sample size. Fortunately, combinatorial pooled sequencing has been proposed on the basis of the combination between the pooled sequencing and a branch of mathematics called combinatorics. In summary, combinatorial pooled sequencing utilizes pooling patterns rather than DNA barcodes to tag each sample and allows the sequencing results to be decoded to identify the reads belonging to the sample that are unique or rare in the population.

In general, combinatorial pooled sequencing involves two more steps than normal pool-seq: encoding and decoding. The encoding step refers to the design of the pooling strategy. Many pooling designs from the fields of group testing and compressed sensing could be applied in the combinatorial pooled sequencing. Accordingly, the corresponding decoding algorithms can be utilized directly to analysis the pooled sequencing results. Since the combinatorial pooled sequencing allows the using of few pools to obtain the reads that belong to each individual for large-scale studies, the efforts for sequencing library preparation could be vastly saved.

In this paper, we mainly survey the encoding and decoding procedure for the combinatorial pooled sequencing applied in several experiments. The combinatorial pooled sequencing could employ much fewer pools than the samples to identify rare variant and haplotype carriers among large-scale samples. Due to the fact that few clones in the library have overlaps, the combinatorial pooled sequencing has been utilized to obtain the reads specified to each clone which allows achieving the draft of complex genomes by assembling reads clone-by-clone. By constructing pools that contain a few haploid copies of genomic DNA, combinatorial pooled sequencing also showed great potential in single individual haplotyping.

Obviously, combinatorial pooling sequencing involves pooling large numbers of specimens which makes liquid handling a challenge. Using liquid-handling robot to execute experiments automatically could help to vastly decrease the labor for pooling [68].

Current pool-seq with short sequencing reads makes it a challenge for linkage disequilibrium estimation and haplotype phasing [4]. Only variants that can be observed within a read or pair of reads could be phased accurately [69,70]. However, this could be improved by advances in sequencing technologies, such as Nanopore [71] and PacBio [72] sequencing platforms, which allow haplotype sequencing for DNA fragments of up to 10 kb. Pooled RNA sequencing has exhibited accuracy comparable with pooled DNA sequencing in estimating allele frequency [73]. However, variation in expression level between individuals should be assessed and accounted for. Hill *et al*. also presented MMAPPR, a mutation mapping analysis pipeline for pooled RNA-seq [74]. We anticipate that these applications will benefit from pool-seq as well as the combinatorial pooled sequencing.

With the availability of new dedicated software tools facilitating the analysis of pool-seq data and the anticipated continued interest in pool-seq, we expect that combinatorial pooled sequencing will be a more attractive tool in the future.

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China (No. 2012CB316501) and the National Natural Science Foundation of China (No. 61472078) and the Scientific Research Foundation of Graduate School of Southeast University.

COMPLIANCE WITH ETHICS GUDELINES

The authors Chang-chang Cao and Xiao Sun declare no competing financial interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014) Ten years of next-generation sequencing technology. Trends Genet., 30, 418–426

2. Metzker, M. L. (2010) Sequencing technologies— the next generation. Nat. Rev. Genet., 11, 31–46
3. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. Nat. Biotechnol., 26, 1135–1145
4. Schlötterer, C., Tobler, R., Kofler, R.and Nolte, V. (2014) Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. Nat. Rev. Genet., 15, 749–763
5. Futschik, A. and Schlötterer, C. (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics, 186, 207–218
6. Patterson, N. and Gabriel, S. (2009) Combinatorics and next-generation sequencing. Nat. Biotechnol., 27, 826–827
7. Wang, W., Yin, X., Soo Pyon, Y., Hayes, M.and Li, J. (2013) Rare variant discovery and calling by sequencing pooled samples with overlaps. Bioinformatics, 29, 29–38
8. Smith, A. M., Heisler, L. E., St Onge, R. P., Farias-Hesson, E., Wallace, I. M., Bodeau, J., Harris, A. N., Perry, K. M., Giaever, G., Pourmand, N., *et al.* (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res., 38, e142
9. Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C. and Estoup, A. (2013) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol. Ecol., 22, 3766–3779
10. Cao, C.-C. and Sun, X. (2015) Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplotype carriers by overlapping pool sequencing. Bioinformatics, 31, 515–522
11. Lonardi, S., Duma, D., Alpert, M., Cordero, F., Beccuti, M., Bhat, P. R., Wu, Y., Ciardo, G., Alsaihati, B., Ma, Y., *et al.* (2013) Combinatorial pooling enables selective sequencing of the barley gene space. PLoS Comput. Biol., 9, e1003010
12. Lo, C., Liu, R., Lee, J., Robasky, K., Byrne, S., Lucchesi, C., Aach, J., Church, G., Bafna, V. and Zhang, K. (2013) On the design of clone-based haplotyping. Genome Biol., 14, R100
13. Skums, P., Artyomenko, A., Glebova, O., Ramachandran, S., Mandoiu, I., Campo, D. S., Dimitrova, Z., Zelikovsky, A. and Khudyakov, Y. (2015) Computational framework for next-generation sequencing of heterogeneous viral populations using combinatorial pooling. Bioinformatics, 31, 682–690
14. Ngo, H., and Du, D. (2000) A survey on combinatorial group testing algorithms with applications to DNA library screening. Discrete mathematical problems with medical applications, 55, 171–182.
15. Erlich, Y., Chang, K., Gordon, A., Ronen, R., Navon, O., Rooks, M. and Hannon, G. J. (2009) DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res., 19, 1243–1253
16. Thierry-Mieg, N. (2006) A new pooling strategy for high-throughput screening: the Shifted Transversal Design. BMC Bioinformatics, 7, 28
17. Dorfman, R. (1943) The detection of defective members of large populations. Ann. Math. Stat., 14, 436–440.
18. Prabhu, S. and Pe'er, I. (2009) Overlapping pools for high-throughput targeted resequencing. Genome Res., 19, 1254–1261
19. Chen, H.-B. and Hwang, F. K. (2008) A survey on nonadaptive group testing algorithms through the angle of decoding. J. Comb. Optim., 15, 49–59.
20. Candes, E., Romberg, J. and Tao, T. (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math., 59, 1207–1223.
21. Donoho, D. (2006) Compressed sensing. IEEE Trans. Inf. Theory, 52, 1289–1306.
22. Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet., 40, 695–701
23. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., *et al.* (2009) Finding the missing heritability of complex diseases. Nature, 461, 747–753
24. Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S. A., Fraser, D., *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science, 337, 100–104
25. Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., *et al.*, (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science, 337, 64–69
26. Du, D. and Hwang, F. (2000) Combinatorial group testing and its applications, 2nd ed. Singapore: World Scientific
27. Thierry-Mieg, N. and Bailly, G. (2008) Interpool: interpreting smart-pooling results. Bioinformatics, 24, 696–703
28. Golan, D., Erlich, Y. and Rosset, S. (2012) Weighted pooling—practical and cost-effective techniques for pooled high-throughput sequencing. Bioinformatics, 28, i197–i206
29. Shental, N., Amir, A. and Zuk, O. (2010) Identification of rare alleles and their carriers using compressed se(que)nsing. Nucleic Acids Res., 38, e179
30. Erlich, Y., Gordon, A., Brand, M., Hannon, G. J. and Mitra, P. P. (2010) Compressed Genotyping. IEEE Trans. Inf. Theory, 56, 706–723
31. Erlich, Y., Shental, N., Amir, A. and Zuk, O. (2009) Compressed sensing approach for high throughput carrier screen. In Communication, Control, and Computing, 2009 Allerton 2009 47th Annual Allerton Conference
32. Figueiredo, M. A., Nowak, R. D., and Wright, S. J. (2007) Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. Selected Topics in Signal Processing, 1, 586–597
33. Cao, C.-C., Li, C. and Sun, X. (2014) Quantitative group testing-based overlapping pool sequencing to identify rare variant carriers. BMC Bioinformatics, 15, 195
34. Hwang, F. (2000) Random *k*-set pool designs with distinct columns. Probab. Engrg. Inform. Sci., 14, 49–56.
35. He, D., Zaitlen, N., Pasaniuc, B., Eskin, E. and Halperin, E. (2011) Genotyping common and rare variation using overlapping pool sequencing. BMC Bioinformatics, 12, S2
36. Hormozdiariy, F., Wang, Z., Yangy, W. -Y. and Eskiny, E. (2012) Efficient genotyping of individuals using overlapping pool sequencing and imputation. In Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference. 1023–1027.
37. Zuzarte, P. C., Denroche, R. E., Fehringer, G., Katzov-Eckert, H., Hung, R. J. and McPherson, J. D. (2014) A two-dimensional pooling strategy for rare variant detection on next-generation sequencing platforms. PLoS One, 9, e93455
38. Bonachea, E. M., Zender, G., White, P., Corsmeier, D., Newsom, D., Fitzgerald-Butt, S., Garg, V. and McBride, K. L. (2014) Use of a

targeted, combinatorial next-generation sequencing approach for the study of bicuspid aortic valve. BMC Med. Genomics, 7, 56

39. Cao, C.-C., Li, C., Huang, Z., Ma, X. and Sun, X. (2013) Identifying rare variants with optimal depth of coverage and cost-effective overlapping pool sequencing. Genet. Epidemiol., 37, 820–830
40. Trégouët, D.-A., König, I. R., Erdmann, J., Munteanu, A., Braund, P. S., Hall, A. S., Grosshennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M., *et al.* (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat. Genet., 41, 283–285
41. Niu, T. (2004) Algorithms for inferring haplotypes. Genet. Epidemiol., 27, 334–347
42. Iliadis, A., Anastassiou, D. and Wang, X. (2012) Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled DNA data. BMC Genet., 13, 94
43. Chang, Y.-C., Chang, L.-Y., Chang, T.-J., Jiang, Y.-D., Lee, K.-C., Kuo, S.-S., Lee, W.-J. and Chuang, L.-M. (2010) The associations of LPIN1 gene expression in adipose tissue with metabolic phenotypes in the Chinese population. Obesity (Silver Spring), 18, 7–12
44. Jin, H., Stewart, T. L., Hof, R. V., Reid, D. M., Aspden, R. M. and Ralston, S. (2009) A rare haplotype in the upstream regulatory region of COL1A1 is associated with reduced bone quality and hip fracture. J. Bone Miner. Res., 24, 448–454
45. Lambert, J. C., Grenier-Boley, B., Harold, D., Zelenika, D., Chouraki, V., Kamatani, Y., Sleegers, K., Ikram, M. A., Hiltunen, M., Reitz, C., *et al.* (2013) Genome-wide haplotype association study identifies the *FRMD4A* gene as a risk locus for Alzheimer's disease. Mol. Psychiatry, 18, 461–470
46. Martin, R. J. L., McKnight, A. J., Patterson, C. C., Sadlier, D. M., Maxwell, A. P. and Group, T. W. U. G. S., and the Warren 3/UK GoKinD Study Group. (2010) A rare haplotype of the vitamin D receptor gene is protective against diabetic nephropathy. Nephrol. Dial. Transplant., 25, 497–503
47. Long, Q., Jeffares, D. C., Zhang, Q., Ye, K., Nizhynska, V., Ning, Z., Tyler-Smith, C. and Nordborg, M. (2011) PoolHap: inferring haplotype frequencies from pooled samples by next generation sequencing. PLoS One, 6, e15292
48. Kessner, D., Turner, T. L. and Novembre, J. (2013) Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. Mol. Biol. Evol., 30, 1145–1158
49. Pirinen, M. (2009) Estimating population haplotype frequencies from pooled SNP data using incomplete database information. Bioinformatics, 25, 3296–3302
50. Gasbarra, D., Kulathinal, S., Pirinen, M. and Sillanpää, M. J. (2011) Estimating haplotype frequencies by combining data from large DNA pools with database information. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 8, 36–44
51. Treangen, T. J. and Salzberg, S. L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet., 13, 36–46
52. Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., *et al.* (2012) The oyster genome reveals stress adaptation and complexity of shell formation. Nature, 490, 49–54
53. Lonardi, S., Duma, D., Alpert, M., Cordero, F., Beccuti, M., Bhat, P. R., Wu, Y., Ciardo, G., Alsaihati, B. and Ma, Y. (2011) Barcoding-free BAC pooling enables combinatorial selective sequencing of the barley gene space. arXiv:1112.4438.
54. Engler, F. W., Hatfield, J., Nelson, W.and Soderlund, C. A. (2003) Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res., 13, 2152–2163
55. Bozdag, S., Close, T. J. and Lonardi, S. (2008) Computing the minimal tiling path from a physical map by integer linear programming. In Algorithms in Bioinformatics. 148–161. Berlin: Springer Berlin Heidelberg
56. Duma, D., Wootters, M., Gilbert, A. C., Ngo, H. Q., Rudra, A., Alpert, M., Close, T. J., Ciardo, G. and Lonardi, S. (2013) Accurate decoding of pooled sequenced data using compressed sensing. In Algorithms in Bioinformatics.70–84. Berlin: Springer Berlin Heidelberg
57. Glusman, G., Cox, H. C.and Roach, J. C. (2014) Whole-genome haplotyping approaches and genomic medicine. Genome Med., 6, 73
58. Yang, H., Chen, X. and Wong, W. H. (2011) Completely phased genome sequencing through chromosome sorting. Proc. Natl. Acad. Sci. USA, 108, 12–17
59. Fan, H. C., Wang, J., Potanina, A. and Quake, S. R. (2011) Whole-genome molecular haplotyping of single cells. Nat. Biotechnol., 29, 51–57
60. Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K. and Song, Q. (2010) Direct determination of molecular haplotypes by chromosome microdissection. Nat. Methods, 7, 299–301
61. Selvaraj, S., R Dixon, J., Bansal, V.and Ren, B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol., 31, 1111–1118
62. Snyder, M. W., Adey, A., Kitzman, J. O. and Shendure, J. (2015) Haplotype-resolved genome sequencing: experimental methods and applications. Nat. Rev. Genet., 16, 344–358
63. Kitzman, J. O., Mackenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R., Eichler, E. E., *et al.* (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat. Biotechnol., 29, 59–63
64. Suk, E.-K., McEwen, G. K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D. T., McLaughlin, S., Peckham, H., *et al.* (2011) A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res., 21, 1672–1685
65. Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., *et al.* (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature, 487, 190–195
66. Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H.-Y., Kruglyak, S., Ronaghi, M., Eberle, M. A., *et al.* (2013) Whole-genome haplotyping by dilution, amplification, and sequencing. Proc. Natl. Acad. Sci. USA, 110, 5552–5557
67. Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., *et al.* (2014) Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat. Genet., 46, 1343–1349
68. Zielinski, D., Gordon, A., Zaks, B. L. and Erlich, Y. (2014) iPipet: sample handling using a tablet. Nat. Methods, 11, 784–785
69. Cradic, K. W., Murphy, S. J., Drucker, T. M., Sikkink, R. A., Eberhardt, N. L., Neuhauser, C., Vasmatzis, G. and Grebe, S. K. (2014) A simple method for gene phasing using mate pair sequencing. BMC Med. Genet., 15, 19
70. Feder, A. F., Petrov, D. A. and Bergland, A. O. (2012) LDx: estimation of linkage disequilibrium from high-throughput pooled resequencing data. PLoS One, 7, e48588

71. Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol., 4, 265–270
72. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. Science, 323, 133–138
73. Konczal, M., Koteja, P., Stuglik, M. T., Radwan, J. and Babik, W. (2014) Accuracy of allele frequency estimation using pooled RNA-Seq. Mol. Ecol. Resour., 14, 381–392
74. Hill, J. T., Demarest, B. L., Bisgrove, B. W., Gorsi, B., Su, Y. -C., and Yost, H. J. (2013) MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. Genome Res., 23, 687–697.