

# Optimization of primer design for the detection of variable genomic lesions in cancer

Ali Bashir<sup>1\*</sup>, Yu-Tsueng Liu<sup>3</sup>, Benjamin Raphael<sup>4</sup>, Dennis Carson<sup>3</sup>, Vineet Bafna<sup>2</sup>

<sup>1</sup>Bioinformatics Program, University of California, San Diego, <sup>2</sup>Department of Computer Science and Engineering, University of California, San Diego, <sup>3</sup> Moores Cancer Center, University of California, San Diego, <sup>4</sup> Department of Computer Science & Center for Computational Molecular Biology, Brown University

Associate Editor: Dr. Chris Stoeckert

## ABSTRACT

Primer approximation multiplex PCR (PAMP) is a new experimental protocol for efficiently assaying structural variation in genomes. PAMP is particularly suited to cancer genomes where the precise breakpoints of alterations such as deletions or translocations vary between patients. The design of PCR primer sets for PAMP is challenging because a large number of primer pairs are required to detect alterations in the hundreds of kilobases range that can occur in cancer. These sets of primers must achieve high coverage of the region of interest, while avoiding primer dimers and satisfying the physico-chemical constraints of good PCR primers. We describe a natural formulation of these constraints as a combinatorial optimization problem. We show that the PAMP primer design problem is NP-hard, and design algorithms based on simulated annealing and integer programming, that provide good solutions to this problem in practice.

The algorithms are applied to a test region around the known *CDKN2A* deletion, which show excellent results even in a 1:49 mixture of mutated:wild-type cells. We use these test results to help set design parameters for larger problems. We can achieve near-optimal designs for regions close to 1Mb.

## 1 INTRODUCTION

Many tumors are characterized by large-scale DNA damage. These changes include point mutations and small insertion/deletion events, but also large structural changes such as deletions, translocations, and inversions of entire chromosomal segments. Notable examples include the *TMPRSS2* fusion with *ETS* transcription factors [24], the *SMAD4/DPC4* locus which exhibits a homozygous deletion in pancreatic cancer [9], and the *CDKN2A* locus which frequently has regions mutated or deleted in many cancers [2, 22, 17]. The *CDKN2A* region is interesting in that it encodes *two* proteins, *INK4a* and *ARF*, that actively participate in major tumor suppressor networks [8]. Many such variations in tumor genomes remain undiscovered, and their characterization will be an important part of cancer genome projects.

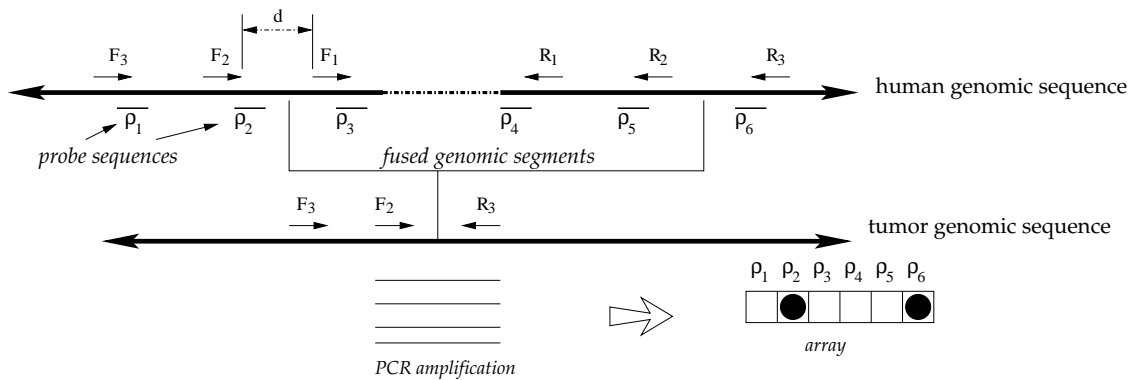
Established experimental techniques for detecting structural changes include array-CGH [19], FISH [18], and End-sequence Profiling [26]. However, array-CGH will detect only copy number changes – not structural rearrangements like inversions or translocations – and generally performs poorly if the sample is a highly

heterogeneous mix of wild type and tumor cells [19]. This presents a significant challenge when screening cells in early onset cancer patients, where the predominance of cells are actually normal, and only a small fraction contain the genomic lesion of interest. Techniques like FISH are labor intensive making them impractical for high-throughput analysis, and moreover often have poor resolution ( $> 1Mb$ ). Finally, genome sequencing techniques like End-sequence Profiling are costly for whole-genome analysis, and it is not clear how to restrict them to specific regions of the genome.

PCR provides one possible solution to this problem. The exponential growth of the reacting product allows for the amplification of weak signals. Consider two regions that are brought together by a genomic rearrangement (deletion, inversion, etc.) in a tumor. Appropriately designed primer pairs within 1 kb of the fusing breakpoints will amplify only in the presence of the mutated DNA, and can amplify even with a small population of cells. Such PCR based screening has been useful in isolating deletion mutants in *C. elegans* [10]. However, in many real tumors, further complications exist as these structural variants often do not have exact boundaries. In the *CDKN2A* region, deletion boundaries often vary over several hundred kilobases and even megabases [21]. This type of variation is even observed in deletions/translocations which result in fusion proteins; the *TMPRSS2* and *ETS* family fusion in prostate cancer not only lacks specificity in the genes it hits (*ERG* and *ETV1/4*), but also as to which exons are joined together [24]. Therefore, in order to appropriately monitor an individual's cancer progression, a test is needed that is capable of screening for, and returning accurate boundaries of, highly variable breakpoints.

To achieve this goal, Liu and Carson have recently devised a novel multiplex primer technique, *Primer Approximation Multiplex PCR (PAMP)*, that allows for the assaying of many possible lesion boundaries simultaneously [15]. A mock illustration of this experimental model can be seen in Figure 1. PAMP utilizes multiple primers whose PCR products cover a region in which breakpoints may occur. Every primer upstream of one breakpoint is in the same orientation, opposite to the primers downstream of the second breakpoint. A primer-pair can form a PCR product only if a genomic lesion places the pair in close proximity. If the primer-pairs are spatially distinct, then any lesion will cause the amplification of *exactly* one primer-pair. The resulting PCR products are easily assayed on a tiling array, identifying the breakpoints of the lesion. The result is a technique

\*Correspondence to: abashir@ucsd.edu



**Fig. 1. Schematic of PAMP design.** Forward and reverse primers approximately cover the left and right breakpoints of the fusing genomic regions. The primers are spread out so that each deletion results in a unique primer pair being amplified. The amplified product is detected by hybridization to probes on an array, the dark spots on the array correspond to amplification of primers most proximal to the breakpoint. In practice, these primer-pairs out compete all others and provide the only visible signal.

which can identify genomic lesions even in high background of normal DNA, and offers precise mappings of a genomic breakpoint (resolution of less than 1kb).

For the PAMP technique to succeed, primers must be selected which adequately cover the entire region, such that every possible pair of deletion boundaries is represented by a corresponding pair of primers that will be amplified by PCR. Additionally, the primers must be chosen from a unique region of the genome, and not allowed to dimerize with each other. Finally, a selected primer must satisfy physico-chemical characteristics that allow it to prime the polymerase reaction. This last problem is well-studied. A number of programs, such as Primer3, select for optimal primers given a nucleotide sequence [23]. Primer Dimer (PD) formation is a common issue in multiplexing PCR reactions, and is affected by amplicon length, sequence and priming efficiency [6]. Additionally, a host of algorithms and applications are available for predicting primer-dimer interactions given a set of multiplexing primers [25, 13]. It is not hard to see that PD formation (due to cross-hybridization) is quite prevalent, using standard dimerization criteria [25]. This poses a significant challenge when the design calls for large numbers (500-1000) of primers, with  $500^2 - 1000^2$  possible dimerizations.

Some recent work addresses this problem. The general problem of optimizing primer set size under cross-hybridization constraints has previously been shown to be NP-Complete by reduction to the Multiple Choice Matching problem [16]. Thus, a number of heuristics have been developed for specific applications, such as minimizing primer set size when given a set of target objects (such as ORFs) [4, 5]. Recently, a number of papers have attempted to optimize multiplex reactions with respect to SNP genotyping. Specifically, these approaches attempt to partition primer sets into multiple multiplexing tubes and examine the trade-offs associated with various experimental design factors [20, 14]. Additionally, recent studies using bioinformatics approaches, have been able to achieve multiplexing of greater than 1000 SNPs, far exceeding previous multiplexing thresholds [27].

The PAMP technique is fundamentally different from these approaches. The design uniquely results in the amplification of a single pair out of a large set of primers (and therefore primer

pairs) due to the genomic lesion. Additionally, unlike clique or coloring based approaches for primer set partitioning [20, 14], we must simultaneously create a non-dimerizing set of primers while optimizing coverage of all possible breakpoints in a region. This *sequence* coverage criteria adds additional complexity to the optimization. Additionally, the goal is to maximize one's ability to detect a structural variant in a specific locus, no matter how variable its boundaries are within a patient population.

In this paper, we formulate the appropriate optimization problem (Section 2), show that the problem is NP-hard (Section 3) even in a restricted form. In Section 4 we describe a number of heuristics that either terminate quickly, or guarantee optimality (but not both). The algorithms are applied to a test region around the known *CDKN2A* deletion, and show excellent results even in a 1:49 mixture of mutated:wild-type cells (Section 5). These preliminary results also help set the design parameters for larger problems. We can achieve near-optimal designs for regions as large as 500kb, and describe additional improvements for larger regions<sup>1</sup>. Our results indicate that PAMP is a feasible technique for assaying lesions, up to a given size, in a heterogeneous mixture of cancer and wild-type cells.

## 2 OPTIMIZING PRIMER DESIGN

To model the problem accurately, we must establish the constraints for an appropriate set of primers. Define a *primer-design* as a set of forward primers  $F_n, \dots, F_2, F_1$  with genomic locations  $l_{F_n} < \dots < l_{F_1}$  and a set of reverse primers  $R_1, R_2, \dots, R_n$  with genomic locations  $l_{R_1} < l_{R_2} < \dots < l_{R_n}$  (Figure 1). Let  $d$  equal the maximum distance between any pair of forward primers, or any pair of reverse primers. We say that the primer design covers a genomic location  $z$  provided that there exists a pair of primers  $F_i$  and  $R_j$  such that if  $z$  is deleted from the genome then the distance between  $F_i$  and  $R_j$  is at most  $2d$ . For the protocol to be successful, the distance  $2d$  should be no greater than what can be readily amplified between a primer-pair (2kb is used as a cut-off). We define the *coverage* of the primer design to be the fraction of the genome between  $l_{F_n}$  and  $l_{R_n}$

<sup>1</sup> Additionally, though not as high-throughput, FISH could detect events for larger regions.

that is covered. Our goal is to find a primer design that maximizes coverage subject to some constraints, as described below.

First, we do not allow *primer-dimerization*: Any two primers in a single multiplex reaction should not cross-hybridize. We present an experiment below in which the presence of a single pair of dimerizing primers is sufficient to negate the amplification. This imposes a fairly stringent requirement for our specific protocol, as dimerization is fairly common, and every forward primer will be pooled with every reverse primer. With 250 forward and reverse primers, this leads to at least  $250^2$  pairs<sup>2</sup>. Second, each primer must amplify a *unique* region. We enforce this by ensuring that the primer itself is unique, and 13bp from the 3' end occur no more than expected by chance. Historically, primers have been selected only in repeat-masked genomes. However, we show that good coverage can be ensured only by allowing unique primers located within transposable elements. Third, primers in the same direction must be non-overlapping, and *at least* distance  $r$  apart where  $r$  is the length of the desired probe. Finally, the *primer-selection* must be physico-chemically appropriate, as described by melting temperature, GC content, and other parameters. These are lumped together as they are adequately addressed by primer selection programs such as Primer3 [23]. In our design, we start with some pre-processing to select unique and physico-chemically appropriate candidate primers. Next, all dimerizing pairs are identified. The problem of designing primers that obey the coverage and cross-hybridization constraints is formulated as a combinatorial optimization problem on the set of candidate primers.

**PAMP design** Construct a primer-coverage graph  $G$ , over a sequence of length  $L$ , as follows: each candidate primer defines a vertex,  $u$ , with its genomic location denoted by  $l_u$ <sup>3</sup>. Add additional vertices  $l_b = 0, l_e = L$  to define the start and end of the sequence. The vertices are paired up with *primer-dimerization edges*  $E$  ( $l_b, l_e$  do not contain any primer-dimerization edges). Thus,  $(u, v) \in E$  if and only if primers  $u$  and  $v$  cross-hybridize. Each pair of nodes is also associated with a corresponding proximity cost. Consider a solution in which two forward primers  $u, v$  are adjacent. Recall that if  $|l_u - l_v| \leq d$ , then any deletion with breakpoints between  $l_u$  and  $l_v$  should lead to an amplification. Otherwise, there are at most  $|l_u - l_v| - d$  positions, where a deletion would not be marked by a PCR amplification. Based on this, each pair  $(u, v)$  is associated with a coverage-cost  $C(u, v) = \max\{0, |l_u - l_v| - d\}$ . The primer design is a *chain*  $\mathcal{P} = p_1, p_2, \dots$  of forward primers followed by reverse primers, ordered so that  $l_{p_j} < l_{p_{j+1}}$  for all  $j$  (Figure 2 top). Define the cost of the design as

$$\mathcal{C}(\mathcal{P}) = \sum_{(p_i, p_j) \in E} w_p + \sum_j w_c C(p_j, p_{j+1}) \quad (1)$$

where  $w_c, w_p$  are appropriately chosen weighting functions. To solve the PAMP design problem, we need to compute a chain of minimum cost. Note that many pairs of primers will cross-hybridize, and removing all such pairs could lead to very sparse primer datasets. This is modeled by adjusting  $w_p$ . Keeping it high may lead to a very sparse solution, while keeping it too low leads to many conflicts being allowed. While our algorithm works for a general  $w_p$ ,

preliminary results show that a single dimerizing pair can cause the entire multiplexed PCR reaction to fail (Section 5), so we describe results with  $w_p = \infty$ .

**Extensions** The model proposed here does not capture two natural extensions. The PAMP protocol does not require all forward and all reverse primers to be in a single multiplex reaction. Rather, the forward and reverse primers can be partitioned into sets, with each forward set reacted with each reverse set. This implies that dimerizations between two forward (or, two reverse) primers are allowed when they are in different sets. We model this by not adding  $(u, v)$  (which can possibly dimerize) to  $E$  if  $u$  and  $v$  are both forward (or both reverse) and occur in different sets. If we partition the forward and reverse sets into  $N$  sets each the protocol will have  $N^2$  distinct multiplex reactions. As smaller  $N$  implies smaller cost, it can be used to optimize a cost-coverage tradeoff. A second useful parameter is the total number of primers. Define *primer-density*  $\rho$  as the average number of primers every  $d$  base-pairs. Clearly  $\rho$  must be  $\geq 1$  for full coverage. We show that a modest increase in  $\rho$  can provide a significant increase in achievable coverage. The primer-density is controlled by augmenting the cost function to be

$$\mathcal{C}(\mathcal{P}) = \sum_{(p_i, p_j) \in E} w_p + \sum_j w_c C(p_j, p_{j+1}) + w_\rho \rho \quad (2)$$

Here, we select  $w_\rho = \infty$  if  $\rho$  exceeds the desired density. Otherwise,  $w_\rho = 0$ . Another point to note is that Figure 1 describes a scenario in which the left and right breakpoints are known to lie in distinct genomic segments. However, this is not critical. We can extend the protocol to a case where the left and right break-points lie in overlapping regions. Primer design considerations for more complex rearrangements (such as deletions with overlapping boundaries and translocations) are natural extensions, but are omitted for exposition purposes.

### 3 COMPLEXITY OF PAMP DESIGN

We show that the PAMP design problem is NP-hard, even in a restricted form: we consider the case  $w_p = \infty$ , so that no cross-hybridization edge is allowed in the solution. We consider an additional restriction on the problem by assuming that the right breakpoint is known exactly, and we have a single reverse primer on that side which does not conflict with any of the candidate forward primers. The decision version of the restricted problem is as follows.

**One-sided PAMP design (OPAMP):** Given a genomic region,  $G$ , of length  $L$  with a single reverse primer, a collection  $\mathcal{F}$  of candidate forward primers, with only the forward primers dimerizing, and an integer value  $D \leq L$ . *Does there exist a non-dimerizing collection  $F \subseteq \mathcal{F}$  of forward primers such that the total uncovered region is less than  $D$  with no adjacent primers*<sup>4</sup>?

Note that a polynomial time solution for the general problem implies that OPAMP is poly-time solvable. Hence, it is sufficient to prove that OPAMP is NP-hard. Supplemental 1 includes a detailed proof via reduction from Max2-SAT [7]. The problem has recently been shown to be hard to even approximate [3].

<sup>2</sup> If all primers are pooled in a single reaction this would lead to  $500^2$  pairs

<sup>3</sup> for exposition purposes, we ignore the length of the primer

<sup>4</sup> “Adjacent” primers have a spacing of less than  $r$  base pairs, where  $r > 0$  as defined in Section 2.

## 4 ALGORITHMS FOR PAMP DESIGN

Prior to optimization of the candidate set, we need to do two preliminary computations.

**Conflict Edge computation:** First, we must compute all possible primer pairs that dimerize (the set of conflict edges  $E$ ). Dimerization due to cross-hybridization is not perfectly understood, but previous studies have indicated that cross-hybridization could occur if an ungapped alignment exists with matches exceeding mismatches by at least 7, specifically in the 3' region [25]. We use this as our conflict criteria. For a genomic region of 500Kbp, there are often tens of thousands of candidate primers, each pair which must be checked for dimerization. To efficiently compute the conflict edge graph, we will employ a simple filtering technique. If the mismatches occur randomly, it can be shown that with high probability, there is a sub-alignment with 3 consecutive matches in dimerizing primers. Therefore, we construct a hash table of all 3-mers. Only primers that hash to the same location are aligned to compute  $E$ .

Additional layers of complexity regarding conflict edges are possible. Lipson proposed an extensive strategy for computing dimerization and mispriming potential, the probability of a primer pair falsely amplifying a different region of the genome [14]. In future work we intend to compare such approaches experimentally to improve the robustness of our algorithm.

**Repeat filtering:** Second, we must ensure uniqueness of the primers by filtering for repeats. Typical algorithms avoid placing primers within repeats in order to reduce the possibility of primers annealing non-uniquely [1]. However, based on the limitation in range of PCR ( $\sim 2$ kb in our experiments), a significant loss of base pair coverage would result from disallowing primers within repetitive elements. For example, in the *CDKN2A* region the optimal coverage *theoretically possible* was  $\sim 75\%$  coverage of the 500 kb flanking sequence if one disallows primers within repeats.

In order to obtain better coverage, we needed to be aggressive in our selection of primers, and selected some primers within repetitive regions. We used parameters from the Wang et. al, to help derive our filtering criteria [27]. For each repeat in our region of interest, we created a table of every 20-mer indicating its raw occurrence in the genome, as well as the occurrence of the 13bp sub-string from its 3'-end. A primer was selected if, in addition to satisfying standard primer criteria, it did not have its 3'-end occur more than was expected by chance. Additionally, the resulting sequence set is checked more rigorously for uniqueness in the region as described in the Supplemental Methods. Though the majority of repeat sequence is in far higher occurrence than expected by chance, one can sometimes find small regions that are permissible as primers.

The set of 'unique' but possibly dimerizing sets of primers forms the initial list from which a candidate-set of low cost is to be selected. Given the NP-completeness result, we focus on heuristic versions of the problem. We describe algorithms for optimal PAMP design based on *greedy selection* and *simulated-annealing* (guaranteed running time, but not optimality), and also *Integer Linear Programming* (guaranteed optimality, but not running time).

**Greedy heuristic for PAMP design:** In this simple heuristic, we attempt to greedily extend a primer set of low cost. Note that the typical value of  $w_p$  is very high (or,  $w_p = \infty$ ), which limits the total number of primers selected. Define  $P_u$  as the chain whose penultimate primer is  $u$  (the primer at  $l_e$  being the last) with cost  $\mathcal{C}(P_u)$ .  $E_u$  corresponds to the set of primers which have dimerizing edges

with  $u$ .

$$\begin{aligned} \mathcal{C}(P_u) &= \min_{v : l_v < l_u} \{\mathcal{C}(P_v + \{u\})\} \\ P_{v^*} &= \operatorname{argmin}_{v : l_v < l_u} \{\mathcal{C}(P_v + \{u\})\} \\ P_u &= P_{v^*} + \{u\} \end{aligned}$$

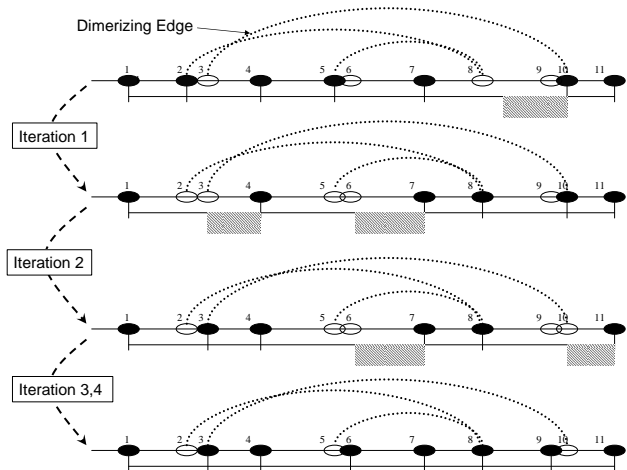
The final solution is given at the  $P_u$  with minimum cost. It is not hard to see that this will result in unevenly distributed primers, with better primer density in regions that were looked at first. In practice, the greedy heuristic is outperformed by simulated annealing (especially in large regions) and is used, along with randomized selection, only to provide initial solutions.

**Simulated annealing for PAMP design:** The simulated annealing is done over the space of all putative solutions. We start with a candidate-set  $\mathcal{P}$  of cost  $\mathcal{C}(\mathcal{P})$ , and in each step, we move to a *neighboring* solution  $\mathcal{P}'$  [12]. We will consider two solution spaces, each with its own neighborhood. In the first, we consider the case  $w_p = \infty$ . Each candidate-set  $\mathcal{P}$  induces an independent set on the primer-coverage graph, i.e. no pair of dimerizing primers is allowed. Candidate-sets  $\mathcal{P}'$  is in the neighborhood of  $\mathcal{P}$  ( $\mathcal{P}' \in N_{\mathcal{P}}$ ), if there exists a primer  $u$  such that  $\mathcal{P}' = \mathcal{P} + \{u\} - \{v : (u, v) \in E\}$ .

In the second case ( $w_p < \infty$ ), every subset  $\mathcal{P}$  subject to certain size constraints is in the solution space.  $\mathcal{P}' \in N_{\mathcal{P}}$  if  $|\mathcal{P} - \mathcal{P}'| \leq 1$ . In other words,  $\mathcal{P}'$  can be obtained from  $\mathcal{P}$  by adding or deleting a single primer. While the two approaches are similar, they do have different convergence properties. In each step  $s$ , we move from the current candidate-set  $\mathcal{P}$  to a neighboring set  $\mathcal{P}'$ . Denote the cost of this transition by  $\Delta_s = \mathcal{C}(\mathcal{P}') - \mathcal{C}(\mathcal{P})$ . From our tests, moving from one independent set to another allowed for faster convergence, therefore we use this methodology for our comparative studies.

In the simulated annealing procedure, we start with an initial solution (random, or greedy). In each transition, step  $s$  is sampled (among all possible steps) with probability proportional to  $e^{-\Delta_s/T}$ , where the temperature  $T$  is an adjustable parameter.  $T$  is decreased according to an annealing schedule. Steps that cause a large decrease in the cost are the most probable, but unfavorable steps also are possible (at higher  $T$ ) allowing for an escape from local minima. An example illustrating this point is seen in Figure 2. The speed and quality of final solution depend upon a number of factors, including the quality of initial solution, choice of neighborhood, and the setting of appropriate temperature  $T$ . We experimented with a number of strategies for optimizing the speed and quality of the solution. In practice the various annealing schedules showed little performance distance, though best results were obtained for proportional and linear schedules. For consistency, a linear schedule is used throughout the results section. Additionally, a random starting solution was given at each test, in order to compare the solutions to the naive greedy without bias.

**ILP and lower bounds:** We can model our problem as a binary integer linear program (ILP). Typical ILP solvers guarantee optimality, but not running time, and may not converge for large sizes. The ILP is depicted in Figure 3. For every primer in  $\mathcal{F}$ , we define a binary variable  $x_i$ . The variable  $x_i = 1$  iff the primer starting at location  $l_i$  is chosen. Clearly, for each pair of dimerizing primers  $i, j$ , and for each pair of primers such that  $l_i - l_j \leq r$ , we have the constraint



**Fig. 2. Sketch of Simulated Annealing Methodology.** Ovals represent possible primers, darkly shaded ovals represent primers in the current chain, dotted lines between two ovals represent dimerization edges, and shaded rectangles represents uncovered sequence (sequence greater than  $d$  base pairs upstream from the nearest primer). Our goal is to minimize uncovered primer space. In this case, the third iteration provides a solution with perfect coverage- which could not be reached in a single move from the initial solution. Note that iterations 1 and 2 transition through chains with higher cost than the initial solution.

$$x_i + x_j \leq 1$$

We set variable  $q_{ij} = 1$  if  $x_i = 1$ ,  $x_j = 1$  and for all other  $j < k < i$   $x_k = 0$ . In other words, primer  $j$  is the primer selected immediately prior to primer  $i$ . As  $q_{ij}$  contributes to the cost of the solution, we only need to set lower bounds on it. The constraints

$$\sum_{l_j < l_i} q_{ij} \geq x_i$$

$$q_{ij} \leq x_j$$

ensure that  $q_{ij} = 1$  only if  $x_i = 1$ , and  $x_j = 1$  for some  $j < i$ . The ‘uncovered region’ penalty  $d_i$  is constrained by

$$d_i \geq \max\{0, \sum_{l_j < l_i} (l_i - l_j - d)q_{ij}\}$$

Note, when  $l_i - l_j - d < 0$ , this term is replaced with 0. Clearly, this penalty is minimized by setting  $q_{ij} = 1$  for the primer  $j$  selected immediately prior to primer  $i$ . In that case,  $d_i$  is exactly the number of uncovered bases, which we seek to minimize.

We find empirically (see Results) that the ILP as described is intractable even for moderate regions. Therefore, we use the ILP formulation mainly to test the performance of the simulated annealing solutions on smaller regions with a sparse number of primers. For lower bounds, we also considered the Linear Programming formulation, achieved by replacing the 0, 1 constraints with

$$0 \leq x_i \leq 1, 0 \leq q_{ij} \leq 1$$

Unfortunately, the integrality gap between the ILP and the relaxed LP is quite large in practice and the bounds are not useful (data not

$$\begin{aligned} & \min \sum_i d_i \\ & s.t. \\ & x_i + x_j \leq 1 && \text{for all dimerizing primers } i, j \\ & x_i + x_j \leq 1 && i - j \leq r \\ & \sum_i x_i \leq \rho * \frac{L}{d} \\ & q_{ij} \leq x_j && \forall i, j \\ & \sum_{j < i} q_{ij} \geq x_i && \forall i \\ & d_i \geq \max \begin{cases} 0 \\ \sum_{l_j < l_i} (l_i - l_j - d)q_{ij} \end{cases} && \forall i \\ & q_{ij}, x_i \in \{0, 1\} && \forall i, j \end{aligned}$$

**Fig. 3. The Integer Linear Program for PAMP design**

Primer Set (Size)	Mat: Wild-type	Left Probe Signal?	Right Probe Signal?	Other Probe Signal?
Initial (20)	0:1	No	No	No
Initial (20)	1:49	Yes	Yes	No
Initial (20)	1:9	Yes	Yes	No
Initial (20)	1:0	Yes	Yes	No
Initial+Repeats (25)	1:49	Yes	Yes	No
Initial+Repeats (25)	1:9	Yes	Yes	No
Initial+Repeats (25)	1:0	Yes	Yes	No
Initial+Primer Dimers (28)	1:9	No	No	No
Initial+Primer Dimers (28)	1:0	No	No	No

**Fig. 4. PAMP Results.** Signal shows amplification around known breakpoints using a multiplexed set of 20 primers. The signals are obtained in a heterogeneous mix of mutant and wild-type DNA. The signal is retained when including distant primers within repeat regions, but completely disappears when a single dimerizing primer pair is used in the reaction. Note that none of the other probes shows a signal.

shown). We focus our studies on ILP solutions which can be obtained for smaller regions with a sparse number of primers. Empirical results for the simulated annealing compared to the ILP solutions can be seen in Figure 7c. In future work, we will explore the use of various cut inequalities to improve the performance of the ILP.

## 5 RESULTS

**Experimental validation** We applied our algorithm to design a set of 600 primers (300 pairs) covering a 500kb region ( $\rho = 1.2$ ) surrounding the *CDKN2A* locus. As described in the original PAMP paper, typical multiplexing reactions consist of adjacent subsets of primers from the larger, overall primer set [15]. Therefore, as a proof of concept, we used a 12 primer-pair test subset from these globally optimized primers to assay a known  $\sim 15$ kb deletion event at the *CDKN2A* locus in the Detroit 562 cell line [17]. A continuous subset of 10 forward primers and 10 reverse primers (representing 20kb of sequence coverage) were chosen from the initial set of 500, in the regions closest to the previously published breakpoints. PAMP

experiments were conducted with a varying mix of wild-type (non-deleted) and mutant samples. As shown in Figure 4, the array shows amplification of only the two probes associated with the characteristic breakpoint positions of the Detroit *CDKN2A* deletion. The signal is present even when the mutant:wild-type ratio is only 1:49. Moreover, we tested the affect of dimerizing primers with respect to suppression of a true signal. It was observed that a single pair of dimerizing primers was sufficient to destroy the signal completely, demonstrating the impact of adding dimerizing primers (Figure 4).

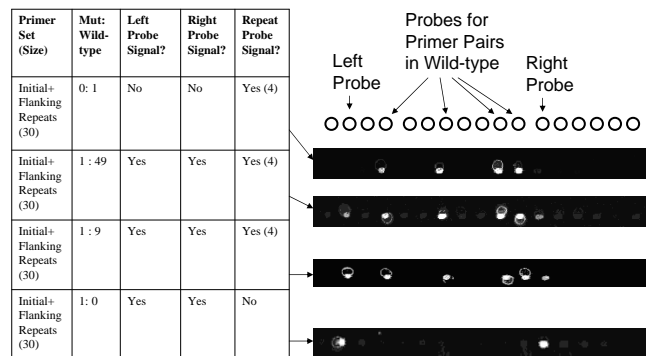
At first glance, this seems like a much simpler (and, scale-reduced) version of the optimizations we have been discussing. However, we note that the computation was done over the entire region and the 20 primers were manually picked from the 600 primer set design around the region of deletion, and the computational complexity is unchanged. Also, the experimental complexity is close to the desired experimental complexity. Recall that the experimental protocol calls for  $N$  forward, and  $N$  reverse sets for a total of  $N^2$  multiplex reactions. Our computational design was based on the most complex case ( $N = 1$ ). On the experimental side, by choosing 12 forward-reverse primers ( $N = 300/10 = 25$ ), we need a total of 625 multiplex reactions, of which exactly one would give the desired positive result. Here, we validated by only performing the single positive experiment. In other experiments, we have scaled this up to  $N \approx 10$ , sufficient for clinical settings, and are moving towards  $N = 1$  (data not shown).

We also performed a series of experiments to test whether primers within (or proximal to) known repeats would cause problems. The set of 20 primers was extended by adding primers flanking highly conserved repeats (such as AluSx transposable elements). Figure 4 provides a negative control by showing that repeat located primers do not destroy the *CDKN2A* deletion signal.

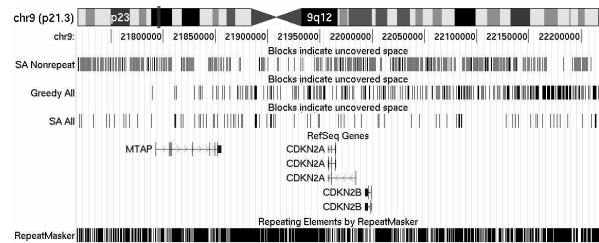
An ideal positive control would be an experiment in which repeat located primers *are* the ones being amplified. Unfortunately, such a design is not possible for the *CDKN2A* deletion in the Detroit 562 cell line, as the deletion breakpoints are not proximal to repeats that yield good primers. However, the 14kb deleted region itself contains multiple repeat elements which can be equally informative. Therefore, we conducted a series of experiments including primer pairs from repeat elements within the deleted region. In every mixed sample (wild-type + mutant), PCR products located in repeats and PCR products resulting from the deletion were detected (Figure 5). The signals of PCR products located in repeats disappeared when only the mutant sample was used, confirming that only the deleted region was being amplified by the primers located in repeats. All PAMP experiments were performed using the protocol described in Liu et al [15].

Our results show the power of the PAMP protocol in amplifying the deletion signal even in a mixed population. The negative control with dimerizing primers reveals the importance of a good design providing high coverage with non-dimerizing primers. The remainder of this manuscript describes the impact of various parameters on the performance of the simulated-annealing heuristic, with comparison to the naive greedy and ILP solutions.

**Computational modeling** Recall that if the candidate primer set places two adjacent primers at distance  $d'$ , with  $d' > d$ , then the total coverage is reduced by  $(d' - d)$  bp. We use this bound on *theoretically obtainable coverage* to compare performance. This bound



**Fig. 5. PAMP performance with repeat located primers.** Primers were chosen in a repeat located within the deleted *CDKN2A* region in the Detroit 562 cell-line. In mixed samples, a signal is obtained from the boundaries of the deleted region, as well as the repeat located primers from the wild-type sample. The repeat signal disappears in the absence of the wild-type sample.



**Fig. 6. Comparison of Missing Coverage in *CDKN2A*.** Custom tracks corresponding to missing coverage were added into the UCSC genome browser at the *CDKN2A* locus [11]. The first track indicates simulated optimized solutions with  $\rho = 1.2$ , when restricted to *nonrepeat* regions of the genome (206,250bp missed coverage). The second track indicates a greedy solution when the search space allows for primers in repeatmasked regions (92,848bp missed coverage). The third track is for the general simulated annealing solution with  $\rho = 1.2$  (17,846bp missed coverage). Less highlighting indicates better coverage.

is not tight for large regions, because as the size of the region increases, it becomes harder to find primer sub-sets with non-dimerizing pairs.

Even with the weak upper bound, early computational results are promising. On the 500kb *CDKN2A* region, we obtained non-dimerizing primer sets with greater than  $> 96\%$  coverage (with primer density  $\rho = 1.2$ ), improving upon the greedy solution ( $< 92\%$  coverage). Also, note that restricting primer selection to repeat-masked regions would have resulted in greatly reduced coverage ( $\sim 60\%$ ). Figure 6 shows the amount of coverage missed in each of these approaches. The lower coverage (across all optimizations) in the *CDKN2A* region when compared to other 500kb regions is primarily due to its high repeat content ( $> 60\%$  over 500kb), however this exemplifies the need for primers in repeat-masked genomic regions. A similar study was done for a smaller region corresponding to the *TMPRSS2:ERG* [24] fusion, achieving  $> 97\%$  coverage (data not shown).

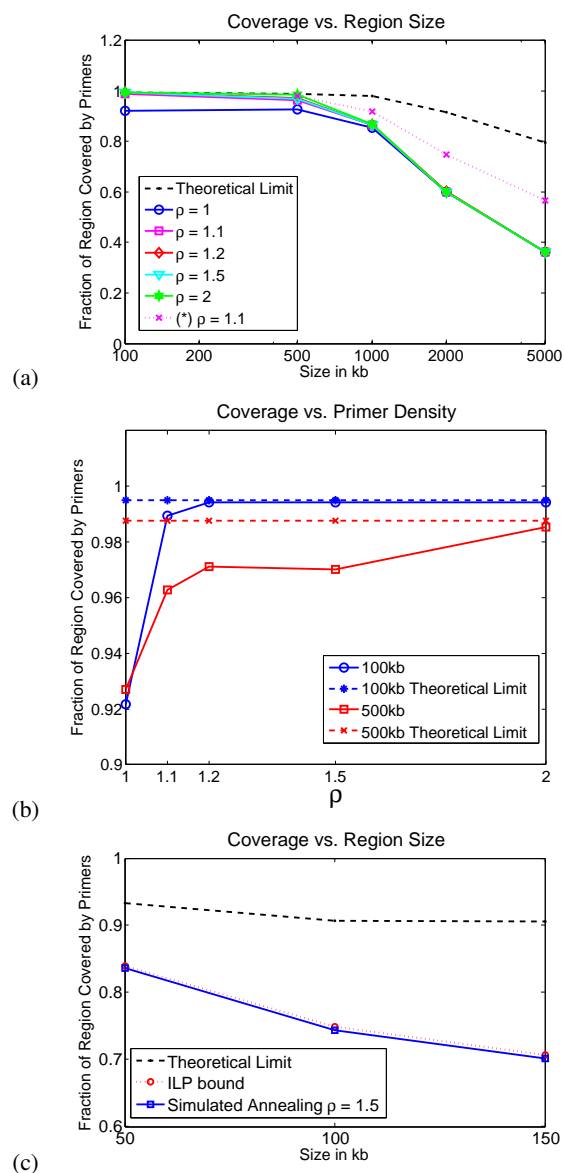
We consider the factors that would impact the quality of the final solution. The size of the region is an important consideration, as discussed earlier. Also, different genomic regions have fairly different compositional characteristics, which will influence performance. Finally, the performance is also influenced by algorithm specific parameters such as the primer-density  $\rho$ . To examine these issues, we selected a number of regions at random from the genome, with size varying from 100Kb to 5Mb, with  $\geq 5$  replicates for each size. Figures 7a,b show the performance as a function of size and primer-density  $\rho$ , additional parameters for primer selection and optimization can be found in the Supplemental Methods. As can be seen, for low sizes ( $\leq 500$ kb), and higher primer-density, the designed primers are very close to theoretical optimum. The coverage diverges from the theoretical optimum over large regions, primarily due to extensive primer dimerization, which not only restricts the overall number of primers, but also greatly limits choices in primer sparse regions of the genome. The primer-density provides a cost-coverage trade-off. For mid-sized regions, a small increase in primer-density greatly improves coverage (Figure 7a). Specifically, a significant improvement is observed between  $\rho = 1$  and  $\rho = 1.2$ , suggesting that  $\rho = 1.2$  provides a good cost-coverage tradeoff. This is made more apparent by comparison to the unrestricted greedy approach which outperforms  $\rho = 1$  in certain cases. Additionally, the simulated annealing solution consistently improves upon the greedy heuristic (even when no restriction is placed on the greedy heuristic for  $\rho$ ) (Figure 7a,b).

For larger-sized regions ( $> 1Mb$ ) there is a significant reduction in coverage. To improve coverage further, we exploit the fact that the PAMP protocol employs multiple ‘tubes’ of multiplexing in practice. The forward and reverse sets are themselves partitioned into  $N$  sets each ( $N \leq 10$ ). Each multiplex reaction consists of a forward and a reverse partition. Thus, each forward primer is only multiplexed with the forward primers in its own partition, and can dimerize with all other forward primers. This relaxation on dimerization constraints allows us to get improved coverage (See Figure 7a,  $p = 1.1$ , *partitioned*). An improvement similar to the unpartitioned sets was observed as  $p$  was increased from 1.1 to 2 (data not shown). In future work, we will include the optimization of the number of rounds as an explicit part of the primer design.

As mentioned, it was observed that extensive primer-dimerization makes it difficult to obtain high coverage, and our results may well be close to the true optima, but the weak bounds on optimal coverage makes it difficult to test this directly. However, in small, sparse regions<sup>5</sup> informative ILP bounds could be obtained. It should be noted that even when the simulated annealing diverged significantly from the theoretical lower bound, it almost perfectly approximated the observed ILP solution (Figure 7c). Additionally, the revised lower bound placed by the ILP (since true convergence could not always be observed) greatly reduced the gap between the theoretical and observed coverage.

**Convergence and running time** The performance of the simulated-annealing algorithm depends upon a number of factors, including the annealing schedule, choice of neighborhood, and so on. We restrict discussion to the length of the schedule. We experimented with a linearly decreasing temperature  $T$ , using a number of runs. In

<sup>5</sup>  $\frac{1}{10}$ th as dense as the normal sequence, corresponding to 88, 179, 262 primers for the 50, 100, 150 kb regions respectively



**Fig. 7.** (a) Coverage versus region size. Each datapoint is the mean over 5 randomly chosen genomic regions of a fixed size. For large regions, the coverage is improved by allowing dimerization to be possible between 2 forward (or, 2 reverse) primers ( $\rho = 1.1$  partitioned), if the two primers are never together in a single multiplex experiment (i.e. they belong to different multiplexing ‘sets’). (b) Improvement in coverage with increasing primer-density  $\rho$ . There is a distinct improvement as  $\rho$  goes from 1 to 1.2, after which the improvement is less pronounced. (c) The best ILP and simulated annealing solutions are virtually indistinguishable at equivalent  $\rho$ . Each data represents a single iteration.

each run, the annealing was set to be  $10\times$  slower than the previous run. We stop when little (less than 1%) or no improvement is recorded over the previous run. The number of iterations is plotted as a function of region size and primer-density in Supplemental 2. Interestingly, the number of iterations peaks around 1Mb, decreasing again. Once again, primer-dimers in large regions severely restrict the search space, leading to fast convergence, but low coverage. The

number of iterations also increase with increasing  $\rho$  due to added flexibility in selection.

## 6 DISCUSSION

We have shown a method to design appropriate sets of primers for PAMP that cover a region without dimerizing, map uniquely in the genome, and possess the requisite physico-chemical characteristics for the PCR reaction. Using this design in multiplex PCR, allows us to detect most deletions within a given region. However, the real advantage of this method is that protocols can be designed for any structural variation that brings two disparate genomic regions together. Therefore, deletions, inversions, translocations, and transpositions can all be assayed with PAMP and appropriate primer designs.

A critical part of our study is the formulation of the problem as a combinatorial optimization problem with goal of improving coverage, while satisfying a collection of constraints. This is not simply an academic question. A diagnostic test will fail in patients for whom the deletion boundaries lie within uncovered regions. The greater the uncovered region, higher the failure probability. In this respect, bringing the coverage up from  $< 90\%$  to over  $97\%$  is very significant. We also provide an ILP formulation, which guarantees true optimality, but did not converge in our formulation. We are exploring a number of alternative approaches, using different cut inequalities to speed up ILP convergence. Even if we guarantee optimality, the ubiquitous dimerization will keep coverage low for large regions. For these regions, we have alternative formulations with more complex multiplexing scenarios to improve coverage. Future work will address the optimization for those problems.

Other technologies have been developed for assaying structural changes in tumor genomes, such as BAC End Sequence Profiling [26] and array CGH [19]. However, the ability of array CGH to detect alterations is impeded by the presence of normal cells or other genomic heterogeneity in the tumor sample. End Sequence Profiling (ESP) can potentially overcome genomic heterogeneity with deep sequencing, but at great expense. Moreover, it is not clear how to restrict ESP to specific regions of the genome. In contrast, the selective amplification of the structurally modified region allows PAMP to detect even weak signals in a heterogeneous population. PAMP could become the technique of choice for probing of specific variants in cancer patients, although the *de novo* discovery of such variants will still rely on array CGH, ESP, or other techniques.

## ACKNOWLEDGEMENTS

This work is supported in part by UCSD NanoTumor Center of Excellence for Cancer Nanotechnology (CA119335) from the National Cancer Institute. B.J.R. is supported by a Career Award at the Scientific Interface (CASI) from the Burroughs-Wellcome Fund. A.B. is supported by a Graduate Research Fellowship from the National Science Foundation.

## REFERENCES

[1]Andreson, R., Reppo, E., Kaplinski, L., and Remm, M. (2006). GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics*, **7**, 172. Comparative Study.

[2]Castellano, M., Pollock, P. M., Walters, M. K., Sparrow, L. E., Down, L. M., Gabrielli, B. G., Parsons, P. G., and Hayward, N. K. (1997). Cdkn2a/p16 is inactivated in most melanoma cell lines. *Cancer Res*, **57**(21), 4868–4875.

[3]Chuzhoy, J. (2007). Hardness of PAMP. Personal Communication.

[4]Doi, K. and Imai, H. (1997). Greedy Algorithms for Finding a Small Set of Primers Satisfying Cover and Length Resolution Conditions in PCR Experiments. *Genome Inform Ser Workshop Genome Inform*, **8**, 43–52.

[5]Doi, K. and Imai, H. (1999). A Greedy Algorithm for Minimizing the Number of Primers in Multiple PCR Experiments. *Genome Inform Ser Workshop Genome Inform*, **10**, 73–82.

[6]Fan, J.-B., Chee, M., and Gunderson, K. (2006). Highly parallel genomic assays. *Nat Rev Genet*, **7**(8), 632–644.

[7]Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness*. W.H. Freeman and Company.

[8]Gil, J. and Peters, G. (2006). Regulation of the INK4b-ARF-INK4a tumour suppressor locus: all for one or one for all. *Nat Rev Mol Cell Biol*, **7**(9), 667–677.

[9]Hahn, S. A., Schutte, M., Hoque, A. T., Moskaluk, C. A., da Costa, L. T., Rozenblum, E., Weinstein, C. L., Fischer, A., Yeo, C. J., Hruban, R. H., and Kern, S. E. (1996). DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science*, **271**(5247), 350–353.

[10]Jansen, G., Hazendonk, E., Thijssen, K., and Plasterk, R. (1997). Reverse genetics by chemical mutagenesis in *Caenorhabditis elegans*. *Nat Genet*, **17**, 119–121.

[11]Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, **12**(6), 996–1006.

[12]Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, **220**(4598), 671–680.

[13]Kmpke, T., Kieninger, M., and Mecklenburg, M. (2001). Efficient primer design algorithms. *Bioinformatics*, **17**(3), 214–225.

[14]Lipson, D. (2002). *Optimization Problems in Design of Oligonucleotides for Hybridization based Methods*. Master's thesis, Technion - Israel Institute of Technology.

[15]Liu, Y.-T. and Carson, D. A. (2007). A novel approach for determining cancer genomic breakpoints in the presence of normal DNA. *PLoS ONE*, **2**, e380.

[16]Nicodme, P. and Steyaert, J. M. (1997). Selecting optimal oligonucleotide primers for multiplex PCR. *Proc Int Conf Intell Syst Mol Biol*, **5**, 210–213.

[17]Nobori, T., Miura, K., Wu, D. J., Lois, A., Takabayashi, K., and Carson, D. A. (1994). Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. *Nature*, **368**(6473), 753–756.

[18]Perry, A., Nobori, T., Ru, N., Anderl, K., Borell, T. J., Mohapatra, G., Feuerstein, B. G., Jenkins, R. B., and Carson, D. A. (1997). Detection of p16 gene deletions in gliomas: a comparison of fluorescence in situ hybridization (FISH) versus quantitative PCR. *J Neuropathol Exp Neurol*, **56**(9), 999–1008.

[19]Pinkel, D. and Albertson, D. G. (2005). Array Comparative Genomic Hybridization and its applications in cancer. *Nat Genet*, **37** Suppl, S11–S17.

[20]Rachlin, J., Ding, C., Cantor, C., and Kasif, S. (2005). Computational tradeoffs in multiplex PCR assay design for SNP genotyping. *BMC Genomics*, **6**, 102.

[21]Raschke, S., Balz, V., Efferth, T., Schulz, W. A., and Florl, A. R. (2005). Homozygous deletions of CDKN2A caused by alternative mechanisms in various human cancer cell lines. *Genes Chromosomes Cancer*, **42**(1), 58–67.

[22]Rocco, J. W. and Sidransky, D. (2001). p16(MTS-1/CDKN2/INK4a) in cancer progression. *Exp Cell Res*, **264**(1), 42–55.

[23]Rozen, S. and Skaletsky, H. (2000). Primer3 on the www for general users and for biologist programmers. *Methods Mol Biol*, **132**, 365–386.

[24]Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., V., S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**(5748), 644–648.

[25]Vallone, P. M. and Butler, J. M. (2004). Autodimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques*, **37**(2), 226–231.

[26]Volik, S., Raphael, B. J., Huang, G., Stratton, M. R., Bignel, G., Murnane, J., Brebner, J. H., Bajsarowicz, K., Paris, P. L., Tao, Q., Kowbel, D., Lapuk, A., Shagin, D. A., Shagina, I. A., Gray, J. W., Cheng, J.-F., de Jong, P. J., Pevzner, P. A., and Collins, C. (2006). Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res*, **16**(3), 394–404.

[27]Wang, H.-Y., Luo, M., Tereshchenko, I. V., Frikker, D. M., Cui, X., Li, J. Y., Hu, G., Chu, Y., Azaro, M. A., Lin, Y., Shen, L., Yang, Q., Kambouris, M. E., Gao, R., Shih, W., and Li, H. (2005). A genotyping system capable of simultaneously analyzing  $> 1000$  single nucleotide polymorphisms in a haploid genome. *Genome Res*, **15**(2), 276–283.