High-Throughput SNP Genotyping by SBE/SBH

Ion I. Mândoiu and Claudia Prăjescu

CSE Department, University of Connecticut
371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155
{ion.mandoiu,claudia.prajescu}@uconn.edu

Abstract. Despite much progress over the past decade, current Single Nucleotide Polymorphism (SNP) genotyping technologies still offer an insufficient degree of multiplexing when required to handle user-selected sets of SNPs. In this paper we propose a new genotyping assay architecture combining multiplexed solution-phase single-base extension (SBE) reactions with sequencing by hybridization (SBH) using universal DNA arrays such as all \( k \)-mer arrays. In addition to PCR amplification of genomic DNA, SNP genotyping using SBE/SBH assays involves the following steps:

1. Synthesizing primers complementing the genomic sequence immediately preceding SNPs of interest;
2. Hybridizing these primers with the genomic DNA;
3. Extending each primer by a single base using polymerase enzyme and dideoxynucleotides labeled with 4 different fluorescent dyes; and finally
4. Hybridizing extended primers to a universal DNA array and determining the identity of the bases that extend each primer by hybridization pattern analysis.

Under the assumption of perfect hybridization, unambiguous genotyping of a set of SNPs requires selecting primers upstream of the SNPs such that each primer hybridizes to at least one array probe that hybridizes to no other primer that can be extended by a common base. Our contributions include a study of multiplexing algorithms for SBE/SBH genotyping assays and preliminary experimental results showing the achievable tradeoffs between the number of array probes and primer length on one hand and the number of SNPs that can be assayed simultaneously on the other. We prove that the problem of selecting a maximum size subset of SNPs that can be unambiguously genotyped in a single SBE/SBH assay is NP-hard, and propose efficient heuristics with good practical performance.

Our heuristics take into account the freedom of selecting primers from both strands of the genomic DNA as well as the presence of disjoint allele sets among genotyped SNPs. In addition, our heuristics can enforce user-specified redundancy constraints facilitating reliable genotyping in the presence of hybridization errors. Simulation results on datasets both randomly generated and extracted from the NCBI dbSNP database suggest that the SBE/SBH architecture provides a flexible and cost-effective alternative to genotyping assays currently used in the industry, enabling genotyping of up to hundreds of thousands of user-specified SNPs per assay.

1 Introduction

After the completion of the Human Genome Project has provided a blueprint of the DNA present in each human cell [15, 16], genomics research is now focusing on the study of DNA variations that occur between individuals, seeking to understand how these variations confer susceptibility to common diseases such as diabetes or cancer. The most common form of genomic variation are the so called single nucleotide polymorphisms (SNPs), i.e., the presence of different DNA nucleotides, or alleles, at certain chromosomal locations. The vast majority of SNPs are bi-allelic, i.e., only two of the four possible DNA bases are observed at the SNP locus. Since human cells contain two copies of each chromosome (with the exception of sex chromosomes in males), both SNP alleles may be present in the DNA of an individual. Determining the identity of alleles present in a DNA sample at a given set of SNP loci is called SNP genotyping.

The continuous progress in high-throughput genomic technologies has resulted in numerous SNP genotyping platforms combining a variety of allele discrimination techniques (sequencing, direct hybridization, primer extension, allele-specific PCR, ligation, and cleavage, etc.), detection mechanisms (fluorescence, mass spectrometry, etc.) and reaction formats (solution phase, solid support, bead arrays), see, e.g., [17, 19] for comprehensive reviews. However, current technologies still offer an insufficient degree of multiplexing (below 10,000 SNPs per assay) for fully-powered genome wide disease association studies that require genotyping of large sets of user-selected SNPs [7]. The highest throughput is currently achieved by work supported in part by a Faculty Large Research Grant from the University of Connecticut Research Foundation.
high-density mapping arrays produced by Affymetrix, which can simultaneously genotype a fixed set of about 250,000 manufacturer selected SNPs per array. Genotyping a comparable number of user-specified set of SNPs would require an expensive and time-consuming re-design of array probes as well as a difficult re-engineering of the primer-ligation amplification protocol.

Among technologies that allow genotyping of custom sets of SNPs one of the most successful ones is the use of DNA tag arrays [6, 11, 13, 21]. DNA tag arrays consist of a set of DNA strings called tags, designed such that each tag hybridizes strongly to its own antitag (Watson-Crick complement), but to no other antitag. The flexibility of tag arrays comes from combining solid-phase hybridization reactions with the high sensitivity of single-base extension reactions, which has also been used for SNP genotyping in combination with MALDI-TOF mass spectrometry [3]. A typical assay based on tag arrays performs SNP genotyping using the following steps [5, 13]: (1) A set of reporter probes is synthesized by ligating antitags to the 5′ end of primers complementing the genomic sequence immediately preceding the SNPs of interest. (2) Reporter probes are hybridized in solution with the genomic sample. (3) The hybridized 3′ (primer) end of reporter probes is extended by a single base in a reaction using the polymerase enzyme and dideoxynucleotides fluorescently labeled with 4 different dyes. (4) Reporter probes are separated from the template DNA and hybridized to a tag array. (5) Finally, fluorescence levels are used to determine the identity of the extending dideoxynucleotides. Commercially available tag arrays have between 2,000 and 10,000 tags [1, 2]. The number of SNPs that can be genotyped per array is typically smaller than the number of tags since some of the tags must remain unassigned due to cross-hybridization with the primers [5, 22]. Another factor limiting the wider use of tag arrays is the relatively high cost of synthesizing the reporter probes, which have a typical length of 40 nucleotides.

In the k-mer array format [9], all $4^k$ DNA probes of length $k$ are spotted or synthesized on the solid array substrate (values of $k$ of up to 10 are feasible with current high-density in-situ synthesis technologies). This format was originally proposed for performing sequencing by hybridization (SBH), which seeks to reconstruct an unknown DNA sequence based on its k-mer spectrum [25]. However, the sequence length for which unambiguous reconstruction is possible with high probability is surprisingly small [26], and, despite several suggestions for improvement, such as the use of gapped probes [12] and pooling of target sequences [14], the SBH scheme has not become practical so far.

In this paper we propose a new genotyping assay architecture combining multiplexed solution-phase single-base extension (SBE) reactions with sequencing by hybridization (SBH) using universal DNA arrays such as all k-mer arrays. SNP genotyping using SBE/SBH assays requires the following steps (see Figure 1): (1) Synthesizing primers complementing the genomic sequence immediately preceding SNPs of interest; (2) Hybridizing primers with the genomic DNA; (3) Extending each primer by a single base using polymerase enzyme and dideoxynucleotides labeled with 4 different fluorescent dyes; and finally (4) Hybridizing extended primers to a universal DNA array and determining the identity of the bases that extend each primer by hybridization pattern analysis.

To the best of our knowledge the combination of the two technologies in the context of SNP genotyping has not been explored thus far. The most closely related genotyping assay is the generic Polymerase Extension Assay (PEA) recently proposed in [27]. In PEA, short amplicons containing the SNPs of interest are hybridized to an all k-mers array of primers that are subsequently extended via single-base extension reactions. Hence, in PEA the SBE reactions take place on solid support, similar to arrayed primer extension (APEX) assays which use SNP specific primers spotted on the array [28].

As in [14], the SBE/SBH assay leads to high array probe utilization since we hybridize to the array a large number of short extended primers. However, the main power of the method lies in the fact that the sequences of the labeled oligonucleotides hybridized to the array are a priori known (up to the identity of extending nucleotides). While genotyping with SBE/SBH assays uses similar general principles as the PEA assays proposed in [27], there are also significant differences. A major advantage of SBE/SBH is the much shorter length of extended primers compared to that of PCR amplicons used in PEA. A second advantage is that all probes hybridizing to an extended primer are informative in SBE/SBH assays, regardless of array probe length (in contrast, only probes hybridizing with a substring containing the SNP site are informative in PEA assays). As shown by the experimental results in Section 4 these advantages translate into an increase by orders of magnitude in multiplexing rate compared to the results reported in [27]. We further note that PEA’s effectiveness crucially depends on the ability to amplify very short (preferably 40bp or less) genomic fragments spanning the SNP loci of interest. This limits the achievable degree of multiplexing in PCR amplification [18], making PCR amplification the main bottleneck for PEA assays. Full flexibility in picking PCR primers is preserved in SBE/SBH assays.
Fig. 1. SBE/SBH assay: (a) Primers complementing genomic sequence upstream of each SNP locus are mixed in solution with the genomic DNA sample. (b) Temperature is lowered allowing primers to hybridize to the genomic DNA. (c) Polymerase enzyme and dideoxynucleotides labeled with 4 different fluorescent dyes are added to the solution, causing each primer to be extended by a nucleotide complementing the SNP allele. (d) Extended primers are hybridized to a universal DNA array (an all k-mer array for k=2 is shown) and SNP genotypes are determined by analyzing the resulting hybridization pattern. Under the assumption of perfect hybridization, unambiguous genotyping of the SNPs requires that each primer hybridizes to at least one array probe that hybridizes to no other primer that can be extended by a common base.

The rest of the paper is organized as follows. In Section 2 we formalize two problems that arise in genotyping large sets of SNPs using SBE/SBH assays: the problem of partitioning a set of SNPs into the minimum number of “decodable” subsets, i.e., subsets of SNPs that can be unambiguously genotyped using a single SBE/SBH assay, and that of finding a maximum decodable subset of a given set of SNPs. We also establish hardness results for the latter problem. In Section 3 we propose several efficient heuristics. Finally, in Section 4 we present experimental results on both randomly generated datasets and instances extracted from the NCBI dbSNP database, exploring achievable tradeoffs between the type/number of array probes and primer length on one hand and number of SNPs that can be assayed per array on the other. Our results suggest that the SBE/SBH architecture provides a flexible and cost-effective alternative to genotyping assays currently used in the industry, enabling genotyping of up to hundreds of thousands of user-selected SNPs per assay.

2 Problem Formulations and Complexity

A set of SNP loci can be unambiguously genotyped by SBE/SBH if every combination of SNP genotypes yields a different hybridization pattern (defined as the vector of dye colors observed at each array probe). To formalize the requirements of unambiguous genotyping, let us first consider a simplified SBE/SBH assay consisting of four parallel single-color SBE/SBH reactions, one for each possible SNP allele. Under this scenario, only one type of dideoxynucleotide is added to each SBE reaction, corresponding to the complement of the tested SNP allele. Therefore, a primer is extended in such a reaction if the tested allele is present at the SNP locus probed by the primer, and is left un-extended otherwise.
Let $\mathcal{P}$ be the set of primers used in a single-color SBE/SBH reaction involving dideoxynucleotide $e \in \{A,C,G,T\}$. From the resulting hybridization pattern we must be able to infer for every $p \in \mathcal{P}$ whether or not $p$ was extended by $e$. The extension of $p$ by $e$ will result in a fluorescent signal at all array probes that hybridize with $pe$. However, some of these probes can give a fluorescent signal even when $p$ is not extended by $e$, due to hybridization to other extended primers. Since in the worst case all other primers are extended, it must be the case that at least one of the probes that hybridize to $pe$ does not hybridize to any other extended primer.

Formally, let $X \subseteq \{A,C,G,T\}^*$ be the set of array probes. For every string $y \in \{A,C,G,T\}^*$, let the \textit{spectrum} of $y$ in $X$, denoted $\text{Spec}_X(y)$, be the set of probes of $X$ that hybridize with $y$. Under the assumption of perfect hybridization, $\text{Spec}_X(y)$ consists of those probes of $X$ that are Watson-Crick complements of substrings of $y$. Then, a set of primers $\mathcal{P}$ is said to be \textit{decodable} with respect to extension $e$ if and only if, for every $p \in \mathcal{P}$,

$$\text{Spec}_X(\text{pe}) \setminus \bigcup_{p' \in \mathcal{P}\setminus\{p\}} \text{Spec}_X(p'e) \neq \emptyset \quad (1)$$

Decoding constraints (1) can be directly extended to 4-color SBE/SBH experiments, in which each type of extending base is labeled by a different fluorescent dye. As before, let $\mathcal{P}$ be the set of primers, and, for each primer $p \in \mathcal{P}$, let $E_p \subseteq \{A,C,G,T\}$ be the set of possible extensions of $p$, i.e., Watson-Crick complements of corresponding SNP alleles. If we assume that any combination of dyes can be detected at an array probe location, unambiguous decoding is guaranteed if, for every $p \in \mathcal{P}$ and every extending nucleotide $e \in E_p$,

$$\text{Spec}_X(\text{pe}) \setminus \bigcup_{p' \in \mathcal{P}\setminus\{p\}, e \in E_{p'}} \text{Spec}_X(p'e) \neq \emptyset \quad (2)$$

In the following, we refine (2) to improve practical reliability of SBE/SBH assays. More precisely, we impose additional constraints on the set of probes considered to be \textit{informative} for each SNP allele. First, to enable reliable genotyping of genomic samples that contain SNP alleles at very different concentrations (as a result of uneven efficiency in the PCR amplification step or of pooling DNA from different individuals), we require that a probe that is informative for a certain SNP locus must not hybridize to primers corresponding to different SNP loci, \textit{regardless of their extension}. Second, since recent studies by Naef et al. [23] suggest that fluorescent dyes can significantly interfere with oligonucleotide hybridization on solid support, possibly destabilizing hybridization to a complementary probe on the array, in this paper we use a conservative approach and require that each probe that is informative for a certain SNP allele must hybridize to a strict substring of the corresponding primer. On the other hand, informative probes are still required not to hybridize with any other extended primer, even if such hybridizations involve fluorescently labeled nucleotides. Finally, we introduce a \textit{decoding redundancy} parameter $r \geq 1$, and require that each SNP have at least $r$ informative probes, i.e., probes that hybridize to the corresponding primer but do not hybridize to any other extended primer. Such a redundancy constraint facilitates reliable genotype calling in the presence of hybridization errors. Clearly, the larger the value of $r$, the more hybridization errors that can be tolerated. If a simple majority voting scheme is used for making allele calls, the assay can tolerate up to $\lfloor r/2 \rfloor$ hybridization errors involving the $r$ informative probes of each SNP. Furthermore, since the informative probes of a SNP are required to hybridize \textit{exclusively} with the primer corresponding to the SNP, the redundancy requirement provides a powerful mechanism for detecting and gauging the extent of hybridization errors. Indeed, each unintended hybridization at an informative probe for a bi-allelic SNP has a dye complementary to one of the SNP alleles with probability of only 1/2, and the probability that $k$ such errors pass undetected decreases exponentially in $k$.

The refined set of constraints is captured by the following definition, where, for every primer $p \in \{A,C,G,T\}^*$ and set of extensions $E \subseteq \{A,C,G,T\}$, we let

$$\text{Spec}_X(p, E) = \bigcup_{e \in E} \text{Spec}_X(pe)$$

**Definition 1.** A set of primers $\mathcal{P}$ is said to be strongly $r$-decodable with respect to extension sets $E_p$, $p \in \mathcal{P}$, if and only if, for every $p \in \mathcal{P}$,

$$\left| \text{Spec}_X(p) \setminus \bigcup_{p' \in \mathcal{P}\setminus\{p\}} \text{Spec}_X(p', E_{p'}) \right| \geq r \quad (3)$$
Note that testing whether or not a given set of primers is strongly \(r\)-decodable can be easily accomplished in time linear in the total length of the primers.

Genotyping a large set of SNPs will, in general, require more than one SBE/SBH assay. This raises the problem of partitioning a given set of SNPs into the smallest number of subsets that can each be genotyped using a single SBE/SBH assay. For each SNP locus there are typically two different primers that can be used for genotyping. As shown in [22] for the case of SNP genotyping using tag arrays, exploiting this degree of freedom significantly increases achievable multiplexing rates. Therefore, we next extend our definitions to capture this degree of freedom. Let \(P_i\) be the pool of primers that can be used to genotype the SNP at locus \(i\). Similarly to Definition 1, we have:

**Definition 2.** A set of primer pools \(P = \{P_1, \ldots, P_n\}\) is said to be strongly \(r\)-decodable if and only if there is a primer \(p_i\) in each pool \(P_i\) such that \(\{p_1, \ldots, p_n\}\) is strongly \(r\)-decodable with respect to the respective extension sets \(E_{p_i}\), \(i = 1, \ldots, n\).

Primers \(p_1, p_2, \ldots, p_n\) above are called the representative primers of pools \(P_1, P_2, \ldots, P_n\), respectively. The SNP partitioning problem can then be formulated as follows:

**Minimum Pool Partitioning Problem (MPPP):** Given primer pools \(P = \{P_1, \ldots, P_n\}\), associated extension sets \(E_{p_i}, \ p \in \cup_{i=1}^{n} P_i\), probe set \(X\), and redundancy \(r\), find a partitioning of \(P\) into the minimum number of strongly \(r\)-decodable subsets.

A natural strategy for solving MPPP, similar to the well-known greedy algorithm for the set cover problem, is to find a maximum strongly \(r\)-decodable subset of pools, remove it from \(P\), and then repeat the procedure until no more pools are left in \(P\). This greedy strategy for solving MPPP has been shown to empirically outperform other algorithms for solving the similar partitioning problem for PEA assays [27]. In the case of SBE/SBH, the optimization involved in the main step of the greedy strategy is formalized as follows:

**Maximum \(r\)-Decodable Pool Subset Problem (MDPSP):** Given primer pools \(P = \{P_1, \ldots, P_n\}\), associated extension sets \(E_{p_i}, \ p \in \cup_{i=1}^{n} P_i\), probe set \(X\), and redundancy \(r\), find a strongly \(r\)-decodable subset \(P' \subseteq P\) of maximum size. In addition, for each pool \(P_i \in P'\), find its representative primer.

Unfortunately, as shown in next theorem, MDPSP is NP-hard even for the case when the redundancy parameter is 1 and each pool has exactly one primer.

**Theorem 1.** MDPSP is NP-hard, even when restricted to instances with \(r = 1\) and \(|P| = 1\) for every \(P \in P\).

**Proof.** We will use a reduction from the maximum induced matching problem in bipartite graphs, which is defined as follows:

**Maximum Induced Matching (MIM) Problem in Bipartite Graphs:** Given a bipartite graph \(G = (U \cup V, E)\), find maximum size subsets \(U' \subseteq U, V' \subseteq V\), with \(|U'| = |V'|\) such that the subgraph of \(G\) induced by \(U' \cup V'\) is a matching.

The MIM problem in bipartite graphs is known to be NP-hard even for graphs with maximum degree 3 [20]. Let \(G = (U \cup V, E)\) be such a bipartite graph with maximum degree 3. Without loss of generality we may assume that every vertex in \(G\) has degree at least 1. We will denote by \(N(u)\) the neighborhood of vertex \(u \in U \cup V\), i.e., the set of vertices adjacent with \(u\) in \(G\).

We construct an instance of MDPSP as follows: Let \(r = 1\) and \(l = \lceil \log_2 |V| \rceil\). For every \(v \in V\) we add to \(X\) a distinct probe \(x_v \in \{A,T\}^l\); note that this can be done since \(|\{A,T\}^l| = 2^l > |V|\) by our choice of \(l\). For every \(u \in U\), with neighborhood \(N(u) = \{v_1, v_2, v_3\}\), we construct a primer \(p_u = x_{v_1}Cx_{v_2}Cx_{v_3}\) and set \(P_u = \{p_u\}\). We use a similar construction for vertices \(u \in U\) with only 1 or 2 neighbors. Note that in each case the pool \(P_u\) consists of a single primer \(p_u\) of length at most \(3l+2\). For each constructed primer \(p\), the set of possible extensions is defined as \(E_p = \{G,C\}\). Since the probes of \(X\) contain only A’s and T’s, for every primer \(p_u, \ u \in U\),

\[
Spec_X(p_u, E_{p_u}) = Spec_X(p_u) = \{x_v \in X | v \in N(u)\}
\]
Theorem 2. It is NP-hard to approximate MDPSP within a factor of $6600/6659$, even when restricted to instances with $r = 1$ and $|P| = 1$ for every $P \in \mathcal{P}$.

### 3 Algorithms

In this section we describe three heuristic approaches to MDPSP. The first one is a naive greedy algorithm that sequentially evaluates the primers in the given pools in an arbitrary order. The algorithm picks a primer $p$ to be the representative of pool $P \in \mathcal{P}$ if $p$ together with the representatives already picked satisfy condition (3). The pseudocode of this algorithm, which we refer to as Sequential Greedy, is given in Figure 2.

The next two algorithms are inspired by the Min-Greedy algorithm in [10], which approximates MIM in $d$-regular graphs within a factor of $d - 1$. For the MIM problem, the Min-Greedy algorithm picks at
Algorithms MinPrimerGreedy and MinProbeGreedy can be implemented efficiently using a Fibonacci heap. Since each primer has bounded degree, the sorting of probe degrees requires \( O(k) \) total time. The total number of edges in the hybridization graph is \( O(N + m) \). By using a Fibonacci heap, finding a minimum degree primer (probe) can be done in \( O(\log N) \) (respectively \( O(\log m) \)) and each primer degree update can be done in amortized \( O(1) \) time. Thus, the total runtime for MinPrimerGreedy algorithm is \( O(k \log N + N + m) \), and the total runtime for MinProbeGreedy algorithm is \( O(k \log m + N + m) \).
4 Experimental Results

We considered two types of data sets:

- Randomly generated datasets containing between 1,000 to 200,000 pools with 1 or 2 primers of length between 10 and 30.
- Two-primer pools representing over 9 million reference SNPs in human chromosomes 1-22, X, and Y extracted from the NCBI dbSNP database build 125. We disregarded reference SNPs for which available flanking sequence was insufficient for determining two non-degenerate primers of desired length (due, e.g., to the presence of degenerate bases near the SNP locus).

We used two types of array probe sets. First, we used probe sets containing all \( k \)-mers, for \( k \) between 8 and 10. All \( k \)-mer arrays are well studied in the context of sequencing by hybridization. However, a major drawback of all \( k \)-mer arrays is that the \( k \)-mers have a wide range of melting temperatures, making it difficult to ensure reliable hybridization results. For short oligonucleotides, a good approximation of the melting temperature is obtained using the simple 2-4 rule of Wallace [29], according to which the melting temperature of a probe is approximately twice the number of A and T bases, plus four times the number of C and G bases. As in [4], we define the weight of a DNA string to be the number of A and T bases plus twice the number of C and G bases. For a given integer \( c \), a DNA string is called a \( c \)-token if it has a weight \( c \) or more and all its proper suffixes have weight strictly less than \( c \). Since the weight of a \( c \)-token is either \( c \) or \( c + 1 \), it follows that the 2-4 rule computed melting temperature of all \( c \)-tokens varies in a range of about 4°C. In our experiments we used probe sets consisting of all \( c \)-tokens, with \( c \) varying between 11 and 13. The considered values of \( k \) and \( c \) were picked such that the resulting number of probes is representative of current array manufacturing technologies: there are roughly 65,000 8-mers, 262,000 9-mers, 1 million 10-mers, 86,000 11-tokens, 236,000 12-tokens, and 645,000 13-tokens – the smaller probe sets can be spotted using current oligonucleotide printing robots, while the larger probe sets can be synthesized in situ using photolithographic techniques.

### 4.1 Results on Synthetic Datasets

In a first set of experiments on the randomly generated datasets we compared the three MDPSP algorithms on instances with primer length set to 20, which is the typical length used, e.g., in genotyping using tag arrays. In these experiments the set of possible extensions was considered to be \( \{A,C,T,G\} \) for all primers. Such a conservative choice gives an estimate of multiplexing rates achievable by SBE/SBH assays in more demanding genomic analyses such as microorganism identification by DNA barcoding [8], in which a primer (typically referred to as a "distinguisher" in this context) may be extended by any of the DNA bases in different microorganisms. The results of these experiments for all \( k \)-mer and all \( c \)-token probe sets are presented in Tables 1 and 2, respectively. The results show that using the flexibility of...
Input: Pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, extension sets $E_p$, $p \in \cup_{i=1}^n P_i$, probe set $X$, and redundancy $r$

Output: Strongly $r$-decodable subset of pools $\mathcal{P}' \subseteq \mathcal{P}$ and set $R$ of representative primers for the pools in $\mathcal{P}'$

Construct hybridization graph $G$

$\mathcal{P}' \leftarrow \emptyset$

$R \leftarrow \emptyset$

While $G$ is not empty do
  Find a minimum degree primer $p$, and let $P$ be the pool of $p$
  $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{P\}$
  $R \leftarrow R \cup \{p\}$
  For each $(p') \in P \setminus \{p\}$ do
    remove-primer$(p')$
  End For
  Let $|N^+(p)| = k$ and let $\{x_1, \ldots, x_k\}$ be the probes in $N^+(p)$, indexed in increasing order of their degrees
  For each $x \in \{x_1, \ldots, x_r\}$ do
    For each $(p') \in N^+(x) \cup N^-(x)$ do
      remove-primer$(p')$
    End For
    Delete vertex $x$ from $G$
  End For
  For each $x \in \{x_{r+1}, \ldots, x_k\} \cup N^-(p)$ do
    remove-probe$(x)$
  End For
End While

Fig. 5. MinPrimerGreedy greedy algorithm.

picking primers from either strand of the genomic sequence yields an improvement of up to 10% in the number of $r$-decodable pools. The MinProbeGreedy algorithm typically produces better results compared to the MinPrimerGreedy variant. On the other hand, neither Sequential Greedy nor MinProbeGreedy dominates the other algorithms for all range of instance parameters – Sequential Greedy generally gives the better results for $k$-mer experiments with high redundancy values, while MinProbeGreedy generally gives better results for $k$-mer experiments with large number of pools and low redundancy and for $c$-token experiments.

In the second set of experiments we ran the three MDPSP algorithms on datasets with the same primer length of 20, pool size of 2, and with the number of possible extensions of each primer set to 4 as in DNA-barcoding applications, and to 2 as in SNP genotyping. The results for all $k$-mer and all $c$-token probe sets are given in Tables 3 and 4. The relative performance of the algorithms is similar to that observed in the first set of experiments. As expected, taking into account the reduced number of possible extensions increases the size of computed decodable pool subsets, often by more than 5%.

In the third set of experiments we explored the degree of freedom given by the primer length. For any fixed array probe set and redundancy requirement, we need a minimum primer length to be able to satisfy constraints (3). Increasing the primer length beyond this minimum primer length is often beneficial, as it increases the number of array probes that hybridize with the primer. However, if primer length increases too much, an increasing number of these probes become non-specific, and the multiplexing rate starts to decline. Figure 7 gives the tradeoff between primer length and the size of the strongly $r$-decodable pool subsets computed by the three MDPSP algorithms for pools with 2 primers, 2 possible extensions per primer and all 10-mers, respectively all 13-tokens, as array probes. We notice that the optimal primer length increases with the redundancy parameter.

4.2 Results on dbSNP Data

To stress-test our methods, we extracted a total of over 9 million 2-primer pools corresponding to reference SNPs in human chromosomes 1-22, X, and Y in the NCBI dbSNP database build 125. We constructed a dataset for each of the 24 chromosomes by creating a 2-primer pool for each reference SNP for which dbSNP contains at least 20 non-degenerate base pairs of flanking sequence on both sides (the number of
Input: Pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, extension sets $E_p, p \in \bigcup_{i=1}^n P_i$, probe set $X$, and redundancy $r$

Output: Strongly $r$-decodable subset of pools $\mathcal{P}' \subseteq \mathcal{P}$ and set $R$ of representative primers for the pools in $\mathcal{P}'$

Construct hybridization graph $G$

$\mathcal{P}' \leftarrow \emptyset$

$R \leftarrow \emptyset$

While $G$ is not empty do

Find a minimum degree probe $x$

Find a minimum degree primer $p$ in $N^+(x)$, and let $P$ be the pool of $p$

$\mathcal{P}' \leftarrow \mathcal{P}' \cup \{P\}$

For each $p' \in P \setminus \{p\}$ do

remove-primer$(p')$

End For

Let $|N^+(p)| = k$ and let $\{x_1, \ldots, x_k\}$ be the probes in $N^+(p)$, indexed in increasing order of their degrees

For each $x \in \{x_1, \ldots, x_r\}$ do

For each $p' \in N^+(x) \cup N^-(x)$ do

remove-primer$(p')$

End For

Delete vertex $x$ from $G$

End For

For each $x \in \{x_{r+1}, \ldots, x_k\} \cup N^-(p)$ do

remove-probe$(x)$

End For

End While

Fig. 6. MinProbeGreedy greedy algorithm.

reference SNPs and extracted pools for each chromosome are given in Table 5). Since these large sets of pools must be partitioned between multiple SBE/SBH experiments, we used a simple MPPP algorithm which iteratively finds maximum $r$-decodable pool subsets using the sequential greedy algorithm.

Figures 8 and 9 give the cumulative coverage percentage for the first 50 arrays of all 10-mers, respectively all 13-tokens, on the set of pools extracted from the human chromosome 1. In these experiments we used redundancy between 1 and 5, and primer length 14 or 20. While the MDPSP size in the first few iterations of our MPPP algorithm is comparable to those reported for randomly generated datasets in Section 4.1, the number of SNPs assayed per array decreases constantly with array number – as we need to assay more and more “difficult” SNPs. Somehow surprisingly, the results also suggest using primers of different lengths in different SBE/SBH experiments: while a primer length of 14 seems to be optimal for the first few arrays, longer primers improve the degree of multiplexing when only hard to differentiate SNPs remain, especially for high redundancy.

Finally, in Table 5 we give the number of arrays (containing either all 10-mers or all 13-tokens) required to cover 90%, respectively 95% of the extracted reference SNPs, when using primers of length 20. In practical association studies a much lower SNP coverage (and hence much fewer arrays) would be required due to the high degree of linkage disequilibrium between the SNPs in the human population [24].

References

1. Affymetrix, Inc. GeneFlex Tag Array Technical Note No. 1, available online at http://www.affymetrix.com/support/technical/technotes/genflex_technote.pdf, 2001.
2. Affymetrix, Inc. Custom and application-specific genotyping with the Affymetrix GeneChip MegAllele System, available online at http://www.affymetrix.com/support/technical/other/parallele_brochure.pdf, 2005.
3. Y. Aumann, E. Manisterski, and Z. Yakhini. Designing optimally multiplexed SNP genotyping assays. In Proc. 3rd Workshop on Algorithms in Bioinformatics (WABI), pages 320–338, 2003.
4. A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini. Universal DNA tag systems: a combinatorial design scheme. Journal of Computational Biology, 7(3-4):503–519, 2000.
5. A. Ben-Dor, T. Hartman, B. Schwikowski, R. Sharan, and Z. Yakhini. Towards optimally multiplexed applications of universal DNA tag systems. In Proc. 7th Annual International Conference on Research in Computational Molecular Biology, pages 48–56, 2003.

6. S. Brenner. Methods for sorting polynucleotides using oligonucleotide tags. US Patent 5,604,097, 1997.

7. C.S. Carlson, M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D.A. Nickerson. Selecting a maximally informative set of snps for association analyses using linkage disequilibrium. American Journal of Human Genetics, 74:106–120, 2004.

8. B. DasGupta, K.M. Konwar, I.I. Mándoiu, and A.A. Shvartsman. Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications, 1(2):145–161, 2005.

9. R. Dramanac and R. Crkvenjakov. DNA sequencing by hybridization. Yugoslav patent application, 1987.

10. W. Duckworth, D.F. Manlove, and M. Zito. On the approximability of the maximum induced matching problem. In Journal of Discrete Algorithms, volume 3, pages 79–91, 2005.

11. N.P. Gerry, N.E. Witowski, J. Day, R.P. Hammer, G. Barany, and F. Barany. Universal DNA microarray method for multiplex detection of low abundance point mutations. J. Mol. Biol., 292(2):251–262, 1999.

12. S.A. Heath and F.P. Preparata. Enhanced sequence reconstruction with DNA microarray application. In Proc. 7th Annual International Conference on Computing and Combinatorics (COCOON), pages 64–74, 2005.

13. J.N. Hirschhorn, P. Sklar, K. Lindblad-Toh, Y.-M. Lim, M. Ruiz-Gutierrez, S. Bolk, B. Langhorst, S. Schaffner, E. Winchester, and E. Lander. SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. PNAS, 97(22):12164–12169, 2000.

14. E. Hubbell. Multiplex sequencing by hybridization. Journal of Computational Biology, 8(2):141–149, 2001.

15. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 431:880–921, 2001.

16. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 431:931–945, 2004.

17. S. Jenkins and N. Gibson. High-throughput SNP genotyping. Comparative and Functional Genomics, 3:57–66, 2002.

18. K.M. Konwar, I.I. Mándoiu, A.C. Russell, and A.A. Shvartsman. Improved algorithms for multiplex PCR primer set selection with amplification length constraints. In Y.-P. Phoebe Chen and L. Wong, editors, Proc. 3rd Asia-Pacific Bioinformatics Conference (APBC), pages 41–50, London, 2005. Imperial College Press.

19. P.Y. Kwok. Methods for genotyping single nucleotide polymorphisms. Annual Review of Genomics and Human Genetics, 2:235–258, 2001.

20. V.V. Lozin. On maximum induced matchings in bipartite graphs. In Infomation Processing Letters, volume 81, pages 7–11, 2002.

21. M.S. Morris, D.D. Shoemaker, R.W. Davis, and M.P. Mittmann. Selecting tag nucleic acids. U.S. Patent 6,458,530 B1, 2002.

22. I.I. Mándoiu, C. Prăjescu, and D. Trincá. Improved tag set design and multiplexing algorithms for universal arrays. LNCS Transactions on Computational Systems Biology, II(LNBI 3680):124–137, 2005.

23. F. Naef and M.O. Magnasco. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. In Physical Review E., volume 68, pages 11906–11910, 2003.

24. N. Patil et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science, 294:1719–1723, 2001.

25. P.A. Pevzner. Computational Molecular Biology: An Algorithmic Approach. MIT Press, 2000.

26. P.A. Pevzner and R.J. Lipshutz. Towards DNA sequencing chips. In Proc. 19th Int. Conf. on Mathematical Foundations of Computer Science, pages 143–158, 1994.

27. R. Sharan, J. Gramm, Z. Yakhini, and A. Ben-Dor. Multiplexing schemes for generic SNP genotyping assays. Journal of Computational Biology, 12(5):514–533, 2005.

28. N. Tonisson, A. Kurg, E. Lohmuskaar, and A. Metspalu. Arrayed primer extension on the DNA chip - method and application. In Mark Schena, editor, Microarray Biochip Technology, pages 247–263. Eaton Publishing, 2000.

29. R.B. Wallace, J. Shaffer, R.F. Murphy, J. Bonner, T. Hirose, and K. Itakura. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. Nucleic Acids Res., 6(11):6353–6357, 1979.
Table 1. Size of the strongly $r$-decodable pool subset computed by the three MDPSP algorithms for primer length 20 and set of possible extensions $\{A,C,T,G\}$, with redundancy $r \in \{1, 2, 5\}$ and all $k$-mer probe sets for $k \in \{8, 9, 10\}$ (averages over 10 test cases).

| # pools | Algorithm | $k=8$ | $k=9$ | $k=10$ |
|---------|-----------|-------|-------|--------|
|         |           | primer 2 primers | primer 2 primers | primer 2 primers |
| 1000    | Sequential | 1000 1000 | 1000 1000 | 1000 1000 |
|         | MinPrimer  | 1000 1000 | 1000 1000 | 1000 1000 |
|         | MinProbe   | 1000 1000 | 1000 1000 | 1000 1000 |
| 2000    | Sequential | 2000 2000 | 2000 2000 | 2000 2000 |
|         | MinPrimer  | 2000 2000 | 2000 2000 | 2000 2000 |
|         | MinProbe   | 2000 2000 | 2000 2000 | 2000 2000 |
| 10000   | Sequential | 7714 8319 | 9991 9999 | 10000 10000 |
|         | MinPrimer  | 7714 8319 | 9991 9999 | 10000 10000 |
|         | MinProbe   | 7714 8319 | 9991 9999 | 10000 10000 |
| 20000   | Sequential | 9967 11071 | 19447 19745 | 19999 20000 |
|         | MinPrimer  | 9967 11071 | 19447 19745 | 19999 20000 |
|         | MinProbe   | 9967 11071 | 19447 19745 | 19999 20000 |

| 100     | Sequential | 1000 1000 | 1000 1000 | 1000 1000 |
|         | MinPrimer  | 1000 1000 | 1000 1000 | 1000 1000 |
|         | MinProbe   | 1000 1000 | 1000 1000 | 1000 1000 |
| 200     | Sequential | 1997 2000 | 2000 2000 | 2000 2000 |
|         | MinPrimer  | 1997 2000 | 2000 2000 | 2000 2000 |
|         | MinProbe   | 1997 2000 | 2000 2000 | 2000 2000 |
| 10000   | Sequential | 6210 6901 | 9934 9999 | 10000 10000 |
|         | MinPrimer  | 6210 6901 | 9934 9999 | 10000 10000 |
|         | MinProbe   | 6210 6901 | 9934 9999 | 10000 10000 |
| 20000   | Sequential | 7463 8192 | 17948 19274 | 19992 20000 |
|         | MinPrimer  | 7463 8192 | 17948 19274 | 19992 20000 |
|         | MinProbe   | 7463 8192 | 17948 19274 | 19992 20000 |

| 100     | Sequential | 1000 1000 | 1000 1000 | 1000 1000 |
|         | MinPrimer  | 1000 1000 | 1000 1000 | 1000 1000 |
|         | MinProbe   | 1000 1000 | 1000 1000 | 1000 1000 |
| 200     | Sequential | 1997 2000 | 2000 2000 | 2000 2000 |
|         | MinPrimer  | 1997 2000 | 2000 2000 | 2000 2000 |
|         | MinProbe   | 1997 2000 | 2000 2000 | 2000 2000 |
| 10000   | Sequential | 6210 6901 | 9934 9999 | 10000 10000 |
|         | MinPrimer  | 6210 6901 | 9934 9999 | 10000 10000 |
|         | MinProbe   | 6210 6901 | 9934 9999 | 10000 10000 |
| 20000   | Sequential | 7463 8192 | 17948 19274 | 19992 20000 |
|         | MinPrimer  | 7463 8192 | 17948 19274 | 19992 20000 |
|         | MinProbe   | 7463 8192 | 17948 19274 | 19992 20000 |

| 50      | Sequential | 3745 4161 | 8674 9483 | 9972 10000 |
|         | MinPrimer  | 3745 4161 | 8674 9483 | 9972 10000 |
|         | MinProbe   | 3745 4161 | 8674 9483 | 9972 10000 |
| 200     | Sequential | 3745 4161 | 8674 9483 | 9972 10000 |
|         | MinPrimer  | 3745 4161 | 8674 9483 | 9972 10000 |
|         | MinProbe   | 3745 4161 | 8674 9483 | 9972 10000 |
Table 2. Size of the strongly r-decodable pool subset computed by the three MDPSP algorithms for primer length 20 and set of possible extensions \{A,C,T,G\}, with redundancy \(r \in \{1, 2, 5\}\) and all \(c\)-token probe sets for \(c \in \{11, 12, 13\}\) (averages over 10 test cases).

| \(c\) | Algorithm | \# pools | 1 primer 2 primers | 1 primer 2 primers | 1 primer 2 primers |
|------|------------|----------|---------------------|---------------------|---------------------|
| 1    | Sequential | 1000     | 991 1000            | 999 1000            | 1000 1000           |
|      | MinPrimer  | 992 999  | 999 1000            | 1000 1000           |                     |
|      | MinProbe   | 993 1000 | 999 1000            | 1000 1000           |                     |
| 2    | Sequential | 2000     | 1881 1982           | 1986 2000           | 1999 2000           |
|      | MinPrimer  | 1890 1959| 1987 1998           | 1999 2000           |                     |
|      | MinProbe   | 1906 1994| 1988 2000           | 1999 2000           |                     |
| 3    | Sequential | 10000    | 5556 6401           | 8005 8782           | 9472 9801           |
|      | MinPrimer  | 5556 6401| 8005 8782           | 9472 9801           |                     |
|      | MinProbe   | 6385 7972| 8436 9688           | 9550 9980           |                     |

| \(c\) | Algorithm | \# pools | 1 primer 2 primers | 1 primer 2 primers | 1 primer 2 primers |
|------|------------|----------|---------------------|---------------------|---------------------|
| 1    | Sequential | 2000     | 16221 18510         | 32728 39552         | 61351 76037         |
|      | MinPrimer  | 14967 17278| 30762 36618         | 57530 70048         |                     |
|      | MinProbe   | 20574 24329| 40580 49300         | 72230 91488         |                     |
| 2    | Sequential | 10000    | 1711 1905           | 1940 1995           | 1995 2000           |
|      | MinPrimer  | 1697 1815| 1942 1981           | 1995 2000           |                     |
|      | MinProbe   | 1766 1948| 1951 1997           | 1996 2000           |                     |
| 3    | Sequential | 20000    | 13708 16042         | 26407 32202         | 45064 56877         |
|      | MinPrimer  | 14967 17278| 30762 36618         | 57530 70048         |                     |
|      | MinProbe   | 20574 24329| 40580 49300         | 72230 91488         |                     |
| 4    | Sequential | 1000     | 995 999             | 999 1000            | 1000 1000           |
|      | MinPrimer  | 995 999  | 999 1000            | 1000 1000           |                     |
|      | MinProbe   | 995 999  | 999 1000            | 1000 1000           |                     |
| 5    | Sequential | 2000     | 1711 1905           | 1940 1995           | 1995 2000           |
|      | MinPrimer  | 1697 1815| 1942 1981           | 1995 2000           |                     |
|      | MinProbe   | 1766 1948| 1951 1997           | 1996 2000           |                     |
Table 3. Size of the strongly $r$-decodable pool subset computed by the three MDPSP algorithms for primer length 20 and 2 primers per pool, with number of possible extensions $|E_p| \in \{2, 4\}$, redundancy $r \in \{1, 2, 5\}$ and all k-mer probe sets for $k \in \{8, 9, 10\}$ (averages over 10 test cases).

| # SNPs | Algorithm | $|E_p| = 4$ | $|E_p| = 2$ | $|E_p| = 4$ | $|E_p| = 2$ | $|E_p| = 4$ | $|E_p| = 2$ |
|---|---|---|---|---|---|---|---|
| 1000 | Sequential | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| | MinPrimer | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| | MinProbe | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| 2000 | Sequential | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 |
| | MinPrimer | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 |
| | MinProbe | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 |
| 10000 | Sequential | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| | MinPrimer | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| | MinProbe | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| 20000 | Sequential | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| | MinPrimer | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| | MinProbe | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| 100000 | Sequential | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 |
| | MinPrimer | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 |
| | MinProbe | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 |
| 200000 | Sequential | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 |
| | MinPrimer | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 |
| | MinProbe | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 |
Table 4. Size of the strongly $r$-decodable pool subset computed by the three MDPSP algorithms for primer length 20 and 2 primers per pool, with number of possible extensions $|E_p| \in \{2, 4\}$, redundancy $r \in \{1, 2, 5\}$ and all $c$-token probe sets for $c \in \{11, 12, 13\}$ (averages over 10 test cases).

| $c$ | SNPs | Algorithm | $|E_p| = 4$ | $|E_p| = 2$ | $|E_p| = 4$ | $|E_p| = 2$ | $|E_p| = 4$ | $|E_p| = 2$ |
|-----|------|-----------|------------|------------|------------|------------|------------|------------|
| 1   | 1000 | Sequential | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
|     |     | MinPrimer | 999  | 999  | 1000 | 1000 | 1000 | 1000 |
|     |     | MinProbe | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
|     | 2000 | Sequential | 1982 | 1990 | 2000 | 2000 | 2000 | 2000 |
|     |     | MinPrimer | 1959 | 1968 | 1998 | 1998 | 2000 | 2000 |
|     |     | MinProbe | 1994 | 1998 | 2000 | 2000 | 2000 | 2000 |
| 1   | 10000 | Sequential | 6955 | 7124 | 9218 | 9412 | 9927 | 9953 |
|     |     | MinPrimer | 6401 | 6776 | 8782 | 9034 | 9801 | 9866 |
|     |     | MinProbe | 7972 | 8280 | 9688 | 9782 | 9980 | 9990 |
| 1   | 20000 | Sequential | 9743 | 10358 | 15919 | 15843 | 18934 | 19194 |
|     |     | MinPrimer | 8798 | 9489 | 14080 | 14797 | 18204 | 18573 |
|     |     | MinProbe | 11548 | 12187 | 17094 | 17599 | 19613 | 19746 |
| 1   | 100000 | Sequential | 16042 | 17216 | 20000 | 20000 | 20000 | 20000 |
|     |     | MinPrimer | 14736 | 15817 | 36618 | 39500 | 51540 | 55031 |
|     |     | MinProbe | 17278 | 18483 | 70048 | 75470 | 70048 | 75470 |

| 1   | 1000 | Sequential | 998  | 998  | 1000 | 1000 | 1000 | 1000 |
|     |     | MinPrimer | 996  | 999  | 999  | 1000 | 1000 | 1000 |
|     |     | MinProbe | 998  | 999  | 1000 | 1000 | 1000 | 1000 |
|     | 2000 | Sequential | 1905 | 1931 | 1995 | 1998 | 2000 | 2000 |
|     |     | MinPrimer | 1815 | 1852 | 1981 | 1986 | 2000 | 2000 |
|     |     | MinProbe | 1948 | 1962 | 1997 | 1999 | 2000 | 2000 |
| 1   | 10000 | Sequential | 10569 | 11151 | 16139 | 17409 | 18939 | 20000 |
|     |     | MinPrimer | 9893 | 10352 | 20811 | 22486 | 39839 | 42814 |
|     |     | MinProbe | 11174 | 11894 | 24966 | 26563 | 50811 | 54858 |
|     | 20000 | Sequential | 9839 | 10352 | 20811 | 22486 | 39839 | 42814 |
|     |     | MinPrimer | 9071 | 9819 | 19192 | 20864 | 36597 | 39542 |
|     |     | MinProbe | 12695 | 13562 | 26341 | 28190 | 48131 | 51125 |
| 2   | 100000 | Sequential | 11174 | 11894 | 24966 | 26563 | 50811 | 54858 |
|     |     | MinPrimer | 10418 | 11212 | 23155 | 25122 | 47357 | 51396 |
|     |     | MinProbe | 14541 | 15467 | 31714 | 34015 | 63112 | 67567 |
| 2   | 200000 | Sequential | 11174 | 11894 | 24966 | 26563 | 50811 | 54858 |
|     |     | MinPrimer | 10418 | 11212 | 23155 | 25122 | 47357 | 51396 |
|     |     | MinProbe | 14541 | 15467 | 31714 | 34015 | 63112 | 67567 |

| 3   | 1000 | Sequential | 906  | 932  | 992  | 996  | 1000 | 1000 |
|     |     | MinPrimer | 837  | 868  | 971  | 981  | 999  | 999  |
|     |     | MinProbe | 905  | 928  | 990  | 994  | 1000 | 1000 |
|     | 2000 | Sequential | 1433 | 1497 | 1870 | 1896 | 1991 | 1995 |
|     |     | MinPrimer | 1284 | 1350 | 1753 | 1800 | 1960 | 1974 |
|     |     | MinProbe | 1437 | 1511 | 1856 | 1885 | 1986 | 1990 |
| 3   | 10000 | Sequential | 2713 | 2944 | 4988 | 5344 | 6762 | 8000 |
|     |     | MinPrimer | 2467 | 2608 | 4495 | 4925 | 6976 | 7324 |
|     |     | MinProbe | 2875 | 3081 | 5118 | 5436 | 7651 | 7988 |
| 3   | 100000 | Sequential | 2713 | 2944 | 4988 | 5344 | 6762 | 8000 |
|     |     | MinPrimer | 2467 | 2608 | 4495 | 4925 | 6976 | 7324 |
|     |     | MinProbe | 2875 | 3081 | 5118 | 5436 | 7651 | 7988 |
| 3   | 200000 | Sequential | 2713 | 2944 | 4988 | 5344 | 6762 | 8000 |
|     |     | MinPrimer | 2467 | 2608 | 4495 | 4925 | 6976 | 7324 |
|     |     | MinProbe | 2875 | 3081 | 5118 | 5436 | 7651 | 7988 |
Fig. 7. Size of the strongly $r$-decodable pool subset computed by the three MDPSP algorithms as a function of primer length, for pools with 2 primers, 2 possible extensions per primer, and array probes consisting of all $4^{10}$ 10-mers (a), respectively all 645,376 13-tokens (b) (averages over 10 test cases).
Fig. 8. Cumulative coverage rates for the first 50 10-mers arrays used to decode the SNPs in Chromosome 1 with primer length 14 or 20 and redundancy $r \in \{1, 2, 5\}$. 
Fig. 9. Cumulative coverage rates for the first 50 13-tokens arrays used to decode the SNPs in Chromosome 1 with primer length 14 or 20 and redundancy $r \in \{1, 2, 5\}$. 
Table 5. Number of arrays needed to cover 90 – 95% of the reference SNPs that have unambiguous primers of length 20.

| Chr | ID   | Ref. SNPs | Extracted Pools | # 10-mer arrays | # 13-token arrays |
|-----|------|-----------|-----------------|-----------------|------------------|
|     |      |           |                 | r=1 r=2 r=5     | r=1 r=2 r=5      |
|     |      |           |                 | 90% 95% 90% 95% | 90% 95% 90% 95% |
| 1   | 786058 | 736850    | 5 7 8 11 15 24 | 10 14 17 23     | 39 56            |
| 2   | 758368 | 704415    | 5 6 7 9 14 18  | 9 12 14 18      | 32 42            |
| 3   | 647918 | 587531    | 5 6 7 8 13 16   | 8 10 12 15      | 26 35            |
| 4   | 690063 | 646534    | 5 6 7 9 14 17   | 8 10 12 15      | 26 34            |
| 5   | 590891 | 550794    | 5 6 6 8 12 16   | 7 10 12 15      | 26 34            |
| 6   | 791255 | 742894    | 10 20 14 29 30 54 | 15 29 23 38     | 49 73            |
| 7   | 666932 | 629089    | 6 9 8 12 16 25  | 10 15 16 22     | 36 48            |
| 8   | 488654 | 456856    | 4 5 5 7 10 12   | 7 8 10 13       | 22 29            |
| 9   | 465325 | 441627    | 4 6 6 8 11 17   | 7 10 11 16      | 26 36            |
| 10  | 512165 | 480614    | 4 6 6 8 11 16   | 8 10 12 16      | 27 38            |
| 11  | 505641 | 476379    | 4 6 6 8 11 15   | 8 10 12 15      | 26 35            |
| 12  | 474310 | 443988    | 4 6 6 8 11 18   | 7 10 11 15      | 25 36            |
| 13  | 371187 | 347921    | 3 4 5 6 9 11    | 5 7 8 10        | 16 22            |
| 14  | 292173 | 271130    | 3 4 4 5 7 10    | 5 7 8 10        | 16 23            |
| 15  | 277543 | 258094    | 3 4 4 5 7 11    | 5 7 8 10        | 17 24            |
| 16  | 306530 | 288652    | 4 6 5 9 9 18    | 7 10 11 15      | 25 35            |
| 17  | 269887 | 249563    | 3 4 4 8 9 18    | 7 10 11 15      | 25 37            |
| 18  | 268582 | 250594    | 3 3 4 5 7 9     | 4 6 6 8         | 14 18            |
| 19  | 212057 | 199221    | 4 6 5 9 11 21   | 8 11 12 17      | 29 43            |
| 20  | 292248 | 262567    | 3 4 4 5 7 11    | 6 8 9 12        | 20 27            |
| 21  | 148798 | 138825    | 2 3 3 3 5 6     | 3 4 5 6         | 10 13            |
| 22  | 175939 | 164632    | 3 4 3 6 6 13    | 6 8 9 12        | 21 29            |
| X   | 380246 | 362778    | 4 6 6 8 10 15   | 6 9 9 13        | 19 26            |
| Y   | 50725  | 49372     | 2 2 2 2 3 3     | 2 2 2 3         | 4 5              |