High-Throughput SNP Genotyping by SBE/SBH[*]

Ion I. Măndoiu and Claudia Prăjescu

Computer Science & Engineering Department, University of Connecticut,
371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155, USA
{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. Our contributions include a study of multiplexing algorithms for SBE/SBH genotyping assays and preliminary experimental results showing the achievable multiplexing rates. 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 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. Determining the identity of alleles present in a DNA sample at a given set of SNP loci is called *SNP genotyping*. Despite much progress over the past decade, current 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. 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

[*] Work supported in part by NSF CAREER Award IIS-0546457 and a Faculty Large Research Grant from the University of Connecticut Research Foundation.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 742–749, 2006.

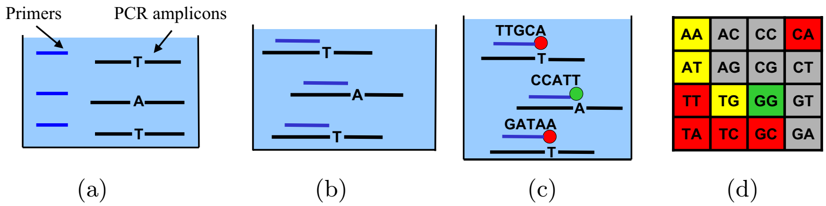

Fig. 1. SBE/SBH assay: (a) Primers complementing genomic sequence upstream of each SNP locus are mixed in solution with the genomic DNA sample. (b) Temperature is lowered allowing primers to hybridize to the genomic DNA. (c) Polymerase enzyme and dideoxynucleotides labeled with 4 different fluorescent 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). SNP genotypes are determined by analyzing the resulting hybridization pattern.

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.

Although both SBE and SBH are well-established techniques, their combination 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 [1]. 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 [2].

As the SBH multiplexing technique of [3], 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 [1], 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 [1]. We further note that PEA's effectiveness crucially depends on the ability to amplify very short genomic fragments spanning the SNP loci of interest. This limits the achievable degree of multiplexing in PCR amplification, making PCR amplification the main bottleneck for PEA assays. Full flexibility in picking PCR primers is preserved in SBE/SBH assays.

The rest of the paper is organized as follows. In Section 2 we formalize several computational problems that arise in genotyping large sets of SNPs using SBE/SBH assays. In Section 3 we propose efficient heuristics for these problems, and in Section 4 we present experimental results on both randomly generated datasets and instances extracted from the NCBI dbSNP database.

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

$$Spec_X(pe) \setminus \bigcup_{p' \in \mathcal{P}\setminus\{p\}} Spec_X(p'e) \neq \emptyset \tag{1}$$

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 $e \in E_p$,

$$Spec_X(pe) \setminus \bigcup_{p' \in \mathcal{P} \setminus \{p\}, e \in E_{p'}} Spec_X(p'e) \neq \emptyset \tag{2}$$

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 *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, *regardless of their extension.* Second, since recent studies by Naef et al. [4] 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 *decoding redundancy* parameter $r \geq 1$, and require that each SNP have at least r informative probes. 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.

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 $Spec_X(p, E) = \bigcup_{e \in E} Spec_X(pe)$.

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 $p \in \mathcal{P}$,*

$$\left| Spec_X(p) \setminus \bigcup_{p' \in \mathcal{P} \setminus \{p\}} Spec_X(p', E_{p'}) \right| \geq r \tag{3}$$

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.

For each SNP locus there are typically two different SBE primers that can be used for genotyping (one from each strand). As shown in [5] for the case of SNP genotyping using tag arrays, exploiting this degree of freedom significantly increases achievable multiplexing rates. Therefore, we next extend our Definition 1 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 $\mathcal{P} = \{P_1, \ldots, P_n\}$ 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 extension sets E_{p_i}, $i = 1, \ldots, n$.*

Primers p_i in Definition 2 are called the *representative primers* of the pools in $\mathcal{P}$, respectively.

Genotyping a large set of SNPs will, in general, require more than one SBE/SBH assay. This rises 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, which is formulated as follows:

Minimum Pool Partitioning Problem (MPPP): *Given primer pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, associated extension sets E_p, $p \in \cup_{i=1}^{n} P_i$, probe set X, and redundancy r, find a partitioning of $\mathcal{P}$ 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 $\mathcal{P}$, and then repeat the procedure until no more pools are left in $\mathcal{P}$. This greedy strategy for solving MPPP has been shown to empirically outperform other algorithms for solving the similar partitioning problem for PEA assays [1]. 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 $\mathcal{P} = \{P_1, \ldots, P_n\}$, associated extension sets E_p, $p \in \cup_{i=1}^{n} P_i$, probe set X, and redundancy r, find a strongly r-decodable subset $\mathcal{P}' \subseteq \mathcal{P}$ of maximum size.*

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

Theorem 1 is proved by reduction from the *Maximum Induced Matching* (MIM) problem in bipartite graphs (see [6] for details). Since the reduction preserves the size of the optimal solution, it follows that any hardness of approximation result for the latter problem also holds for MDPSP. From the hardness result in [7] we get:

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 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 other two algorithms are inspired by the Min-Greedy algorithm in [7], which approximates MIM in d-regular graphs within a factor of $d - 1$. For the MIM problem, the Min-Greedy algorithm picks at each step a vertex u of minimum degree and a vertex v, which is a minimum degree neighbor of u. All the neighbors of u and v are deleted and the edge (u, v) is added to the induced matching. The algorithm stops when the graph becomes empty.

Each instance of MDPSP can be represented as a bipartite *hybridization graph* $G = ((\bigcup_{i=1}^{n} P_i) \cup X, E)$, with the left side containing all primers in the given pools and the right side containing the array probes, i.e., X. There is an edge between primer p and probe $x \in X$ iff $x \in Spec_X(p, E_p)$. As discussed in Section 2, we distinguish between the hybridizations that involve the extending nucleotides and those that do not. Thus, for every primer p, we let $N^+(p) = Spec_X(p)$ and $N^-(p) = Spec_X(p, E_p) \setminus Spec_X(p)$. Similarly, for each probe $x \in X$, we let $N^+(x) = \{p | \ x \in N^+(p)\}$ and $N^-(x) = \{p | \ x \in N^-(p)\}$.

We considered two versions of the Min-Greedy algorithm when run on the bipartite hybridization graph, depending on the side from which the minimum degree vertex is picked. In the first version, referred to as MinPrimerGreedy, we pick first a minimum degree node from the primers side, while in the second version, referred to as MinProbeGreedy, we pick first a minimum degree node from the probes side. Thus, MinPrimerGreedy greedy picks at each step a minimum degree primer p and pairs it with a minimum degree probe $x \in N^+(p)$. MinProbeGreedy greedy, selects at each step a minimum degree probe x and pairs it with a minimum degree primer p in $N^+(x)$. In both algorithms, all neighbors of p and x and their incident edges are removed from G. Also, at each step, the algorithms remove all vertices u, for which $N^+(u) = \emptyset$. These deletions ensure that the primers p selected at each step satisfy condition (3). Both algorithms stop when the graph becomes empty.

As described so far, the MinPrimerGreedy and MinProbeGreedy algorithms work when each pool contains only one primer and when the redundancy is 1. We extended the two variants to handle pools of size greater than 1 by simply removing from the graph all primers $p' \in P \setminus \{p\}$ when picking primer p from pool P. If the redundancy r is greater than 1, then whenever we pick a primer p, we also pick it's r probe neighbors from $N^+(p)$ with the smallest degrees (breaking ties arbitrarily). The primer neighbors of all these r probes will then be deleted from the graph. Moreover, the algorithm maintains the invariant that $|N^+(p)| \geq r$ for every primer p and $|N^+(x)| \geq 1$ for every probe x by removing primers/probes for which the degree decreases below these bounds. Full pseudocode and efficient implementation details for proposed algorithms are available in [6].

4 Experimental Results

We performed experiments with 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 [8], 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. The second type of arrays that we

Table 1. Number of SBE/SBH assays needed to cover $90-95\%$ of extracted reference SNPs using SBE 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%	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	5	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

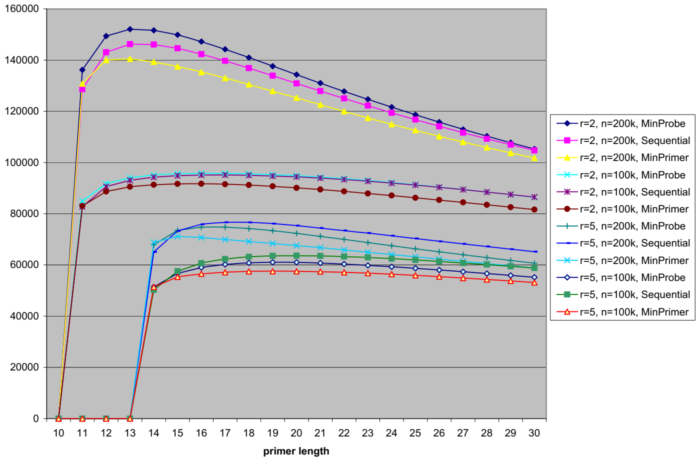

Fig. 2. Size of the strongly r-decodable pool subset computed by the three MDPSP algorithms as a function of primer length, for random instances with $n = 100 - 200k$ pools of 2 primers and all 10-mer arrays (averages over 10 test cases)

considered are all c-token arrays. 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, where the *weight* of a DNA string is defined as the number of A and

T bases plus twice the number of C and G bases. Since the weight of a c-token is either c or $c+1$, it follows that the 2-4 rule computed melting temperature of c-tokens varies in a range of 4°C.

The results of a comprehensive set of experiments comparing the three proposed MDPSP algorithms on both synthetic and genomic datasets are reported in [6]. In Table 1 we report the number of SBE/SBH assays required to cover 90%, respectively 95%, of a total of over 9 million 2-primer pools extracted from the NCBI dbSNP database build 125. We disregarded reference SNPs for which two non-degenerate SBE primers of length 20 could not be determined from the genomic sequence. The results are obtained with a simple MPPP algorithm which iteratively finds maximum r-decodable pool subsets using the sequential greedy algorithm.

Further improvements in the multiplexing rate can be achieved by optimizing the length of SBE primers (see Figure 2). Notice that constraints (3) imply a minimum length for SBE primers. Increasing the primer length beyond this minimum primer length is at first beneficial, since this increases the number of array probes that hybridize with the primer. However, if primer length increases too much, a larger number of array probes become non-specific, and the multiplexing rate starts to decline.

References

1. Sharan, R., Gramm, J., Yakhini, Z., Ben-Dor, A.: Multiplexing schemes for generic SNP genotyping assays. Journal of Computational Biology **12**(5) (2005) 514–533
2. Tonisson, N., Kurg, A., Lohmussaar, E., Metspalu, A.: Arrayed primer extension on the DNA chip - method and application. In Schena, M., ed.: Microarray Biochip Technology, Eaton Publishing (2000) 247–263
3. Hubbell, E.: Multiplex sequencing by hybridization. Journal of Computational Biology **8**(2) (2001) 141–149
4. Naef, F., Magnasco, M.: Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. In: Physical Review E. Volume 68. (2003) 11906–11910
5. Măndoiu, I., Prăjescu, C., Trincă, D.: Improved tag set design and multiplexing algorithms for universal arrays. LNCS Transactions on Computational Systems Biology **II**(LNBI 3680) (2005) 124–137
6. Măndoiu, I., Prăjescu, C.: High-throughput SNP genotyping by SBE/SBH. ACM Computing Research Repository, cs.DS/0512052 (2005)
7. Duckworth, W., Manlove, D., Zito, M.: On the approximability of the maximum induced matching problem. In: Journal of Discrete Algorithms. Volume 3. (2005) 79–91
8. Wallace, R., Shaffer, J., Murphy, R., Bonner, J., Hirose, T., Itakura, K.: Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. Nucleic Acids Res. **6**(11) (1979) 6353–6357