Compressed Sensing-based Pooling Experiments Using Next Generation Sequencing

Noam Shental
Department of Computer Science
The Open University of Israel
Raanana, 43107, Israel
Email: shental@openu.ac.il

Amnon Amir
Department of Physics of Complex Systems
Weizmann Institute of Science
Rehovot, 76100, Israel
Email: amnon.amir@weizmann.ac.il

Or Zuk
Broad Institute of MIT and Harvard
Cambridge, 02142, MA, USA
Email: orzuk@broadinstitute.org

***Abstract*—Screening for carriers of rare variants by resequencing is important for the identification of individuals carrying disease alleles. Rapid sequencing by new technologies enables low-cost resequencing of target regions, although it is still prohibitive to test more than a few individuals. Pooling designs can enable detection of carriers of rare alleles in groups of individuals in a more economical manner. Previous pooling strategies were shown only for a relatively low number of individuals in a pool, and required the design of pooling schemes for particular cases. We propose a novel pooling design, based on a *compressed sensing* approach, which is general, simple and efficient. We model the experimental procedure and show via computer simulations our ability to recover rare allele carriers in larger groups than were possible before, especially when high coverage is obtained for each individual.**

I. Introduction

The search for carriers of rare mutations is of considerable biomedical importance and occurs in various scenarios. In recessive Mendelian diseases, individuals are affected only when they are homozygous for the rare allele, and it is thus desirable to perform carrier screens in order to identify heterozygous individuals. Novel next-generation sequencing technologies enable lower cost and higher throughput sequencing, which may facilitate the study of rare alleles. Of particular interest are sets of specific *known* Single-Nucleotide-Polymorphisms (SNPs) with low minor allele frequency which are known or suspected to be important for a certain trait. Currently identification of carriers over a pre-defined region is a feasible, yet still expensive task.

Our approach follows the lines of the field of *group testing* [1], which aims to tackle the problem of identifying individuals carrying a certain trait out of a group, by designing an efficient set of tests, i.e., pools. In a previous work, by Prabhu and Pe'er [2] , overlapping pools were used, elegantly designed based on Error-Correcting-Codes, enabling the detection of only a *single* carrier. In another work by Erlich et al. [3], a pooling strategy based on the Chinese-Reminder Theorem was employed to solve a slightly different problem. These designs offer a significant saving in resources, as they enable genotype reconstruction for N individuals, by using only $O(\log N)$ or $O(\sqrt{N})$ pools, respectively.

We present a different approach to recovering the identity of individuals carrying rare variants, based on Compressed Sensing (**CS**) (a somewhat similar approach has independently been developed by Erlich et al. [4].) We extend the scope of [2], enabling the recovery of multiple carriers, and dealing with heterozygous or homozygous rare alleles, testing larger cohorts of individuals, incorporating experimental barcoding techniques, as well as dealing with unknown sequencing read errors.

Mapping of rare allele detection into a **CS** setting is simple. We wish to reconstruct the genotypes of N individuals at a specific locus. The genotypes are represented by a vector $\mathbf{x}$ of length N, where x_i represents the genotype of the *i*'th individual. We denote the reference allele by A, and the alternative (rare) allele by B. The possible entries of x_i are 0, 1 and 2, representing a homozygous reference allele (AA), a heterozygous allele (AB) and a homozygous alternative allele (BB), respectively. Hence, x_i counts the number of (alternative) B alleles of the *i*'th individual, and since we are interested in rare minor alleles, most entries x_i are zero.

The sensing matrix M is built of k different measurements represented by the rows of M. The entry m_{ij} is set to 1 if the *j*'th individual participates in *i*'th measurement, and zero otherwise. M is a $Bernoulli(0.5)$ matrix, thus each individual participates on average in half of the pools, or may be sparser and include only $O(\sqrt{N})$ non-zero entries (in case the amount of available DNA is limited).

A measurement in our setting corresponds to sequencing the DNA of a pool of several individuals taken together, hence the measurement vector represents the individuals participating in a given pool and the output of the measurement is proportional to the total number of rare alleles in the pool. The measurement vector represents the frequency of B allele in each pool. Our experimental design is illustrated in Fig. 1.

Several noise factors occur in practice and are all incorporated into our noise model: variations in amount of DNA taken from each individual, amplification errors, sequencing read and alignment errors, and sampling noise due to insufficient coverage. The latter was found to be the major noise factor; in case the number of reads is too low to correctly estimate the fraction of rare alleles in the pool, significant noise is introduced. Another important consideration is that the sensing matrix M is also noisy, which is atypical in most classic **CS** problems, thus deserves special consideration - our results

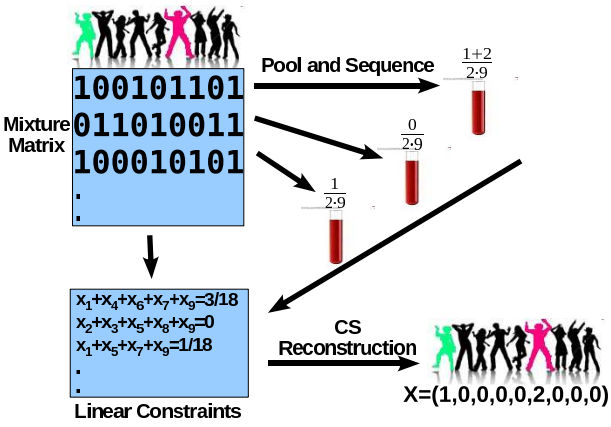

Fig. 1: Schematic description of the CS based procedure. Shown is a case of 9 people, out of which one is a heterozygotic carrier of the rare SNP (marked green), and another one who is a homozygous alternative allele carrier (marked red.) Each sample is randomly assigned to a pool with probability 0.5, as described by the sensing matrix. For example, individuals $1, 4, 6, 7$ and 9 are assigned to the first pool. The DNA of the individuals participating in each pool is mixed, and the fraction of rare alleles in each pool is measured. For example, the first pool contains the two carriers, hence the frequency of the $B's$ is $1+2$ out of the 2×9 alleles. The sensing matrix and the resulting frequencies are incorporated into an underdetermined set of linear constraints, from which the original rare SNP carriers are reconstructed.

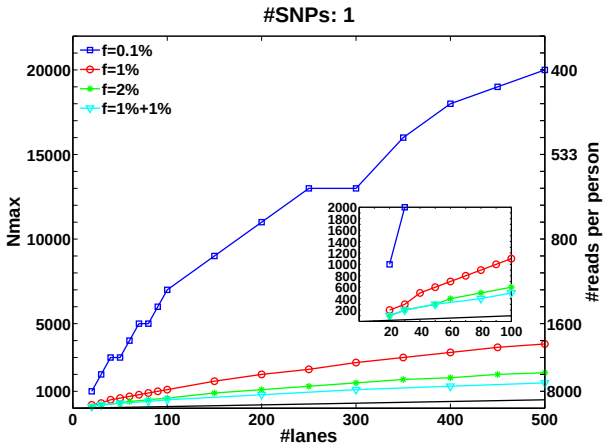

Fig. 2: The maximal number of individuals N_{max} for which our procedure has achieved accurate reconstruction, as a function of the number of lanes used, for various minor allele frequencies f. Smaller frequencies correspond to sparser genotype vectors, thus yield larger improvement in performance. The $f = 1\% + 1\%$ case corresponds to 1% AB and 1% of BB alleles. Each lane contained 4 million mappable reads. The right vertical axis represents the average number of reads obtained for a single individual in a particular lane.

show that the **CS** reconstruction is almost always robust to these types of errors. Introducing the relevant noise factors results in the following optimization problem:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\{0,1,2\}^N}{argmin} \|\mathbf{x}\|_1 \quad \text{s.t. } ||\frac{1}{2}M\mathbf{x} - \frac{1}{r}\mathbf{z}||_2 < \epsilon \qquad (1)$$

where $\mathbf{z}$ provides the measurement vector, with z_i representing the number of reads providing the alternative allele in the $i'th$ pool (out of a total of r reads). Hence we seek a solution x with the lowest L_1 norm, while keeping the L_2 error term smaller than a desired level ϵ. Problem (1) is similar to a standard **CS** problem with some modifications (such as the error in the matrix M, the requirement for an integer solution and the specific details of the noise model) - it was solved using the GPSR algorithm [5], followed by post-processing steps to account for the discrete nature of the problem.

II. Example

In order to study the performance of our approach, we have simulated data according to our pooling strategy, then reconstructed the genotype vector (without knowing the true genotypes), and compared the true and reconstructed genotypes. This process was repeated many times and performance was recorded. As a measure of performance we have chosen N_{max}, the maximal number of individuals for which our approach was successful (i.e. reconstructing the entire genotype vector without errors), in at least 95% of the simulations. A naive approach, not using pooling, will have N_{max} equal to the number of lanes used - our strategy achieves a many-fold improvement over this performance, as is seen in fig. 2.

References

[1] D. Du and F. Hwang, *Combinatorial group testing and its applications*. World Scientific Pub., 2000.

[2] S. Prabhu and I. Pe'er, "Overlapping pools for high-throughput targeted resequencing," *Genome Research*, vol. doi:10.1101/gr.088559.108, 2009.

[3] Y. Erlich, K. Chang, A. Gordon, R. Ronen, O. Navon, M. Rooks, and G. Hannon, "Dna Sudoku-harnessing high-throughput sequencing for multiplexed specimen analysis," *Genome Research*, vol. doi:10.1101/gr.092957.109, 2009.

[4] Y. Erlich, A. Gordon, M. Brand, G. Hannon, and P. Mitra, "Compressed Genotyping," *Arxiv preprint Quantitative Biology/0909.3691*, 2009.

[5] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.