Compressed sensing approach for high throughput carrier screen

Yaniv Erlich, Noam Shental, Amnon Amir and Or Zuk

***Abstract*— Carrier screens are widely used in medical genetics to prevent rare genetic disorders. Current detection methods are based on serial processing which is slow and expensive. Here, we discuss a highly efficient *compressed sensing* approach for ultra-high throughput carrier screens, and highlight both similarities and unique features of our setting compared to the standard compressed sensing framework. Using simulations, we demonstrate the power of compressed carrier screens in a real scenario - finding carriers for rare genetic diseases in Ashkenazi Jews, a population that has well established wide-scale carrier screen programs. We also compare the decoding performance of two typical reconstruction approaches in compressed sensing - GPSR and Belief Propagation. Our results show that Belief Propagation confers better decoding performance in the current application.**

I. Introduction

In the past three decades extensive efforts were made to map severe genetic disorders to specific DNA sequence variations. Remarkably, large number of these disorders, such as Cystic Fibrosis and Sickle Cell Anemia, are caused by subtle sequence changes; even a single nucleotide substitution in a specific location can entirely disrupt the activity of an essential gene. However, many traits are *recessive*, meaning that the nonfunctionality of one allele copy can be compensated by the normal activity of the other copy. In those cases, the genetic disorder appears only in individuals that carry two non-functional alleles. Thus, regarding a recessive disorder, there are three groups of individuals: (a) *normal individuals* - with two functional alleles (b) *carriers* - individuals with only one functional allele (c) *affected* - individuals with two non-functional alleles. Notice that there are no phenotypic differences between a carrier and a normal individual in respect to the disease's trait. However, a breeding between two carriers may lead to devastating results as explained by Mendelian genetics (Table I): The breed of two normal individuals always gives rise to a normal offspring. The breed of a normal and a carrier has 50% chance of giving a normal offspring and 50% chance of giving a carrier, but no chance for an affected offspring. When two carriers breed, they have 25% chance of having an affected offspring for each carriage, 50% of having a carrier, and 25% of having a normal offspring (we do not consider breeds with an affected individual, since affected are usually not reproductive). This picture reveals that only families with two parental carriers are at risk for having an affected sibling, and that other breading combinations are safe.

Since there is no overt indication for being a carrier, and most carriers are born to healthy families, revealing that an offspring is affected is a shocking experience. Therefore, many countries employ wide-scale carrier screen programs,

	Offsprings		
Breed:	Normal	Carrier	Affected
Normal x Normal	100%	0%	0%
Normal x Carrier	50%	50%	0%
Carrier x Carrier	25%	50%	25%

TABLE I: Breeding outcomes as a function of parental genotype

in which individuals are genotyped for a panel of risk genes that are prevalent in the their population. The common practice is to offer the screen to the entire population regardless of their familial history, either before mate selection (premarital screens) to reduce the risk of carrier-carrier breeds, or prenatally in order to provide a reproductive choice.

The most common genotyping method is sequencing the genomic region that harbors the mutation site, and analyzing whether the DNA sequence is wild-type (WT) or carries a mutation. This approach gained popularity due to its high accuracy (sensitivity and specificity), applicability to a wide variety of genetic disorders, and technical simplicity. However, the current DNA sequencing platforms utilized in medical diagnosis provide only serial processing of a single specimen/region combination at a time. Therefore, while the genetic basis of many disorders is known, the cumbersome costs of large genotyping panels hinder applying this knowledge routinely in the clinic.

Recently, a new class of DNA sequencing methods, dubbed *next-generation sequencing technologies* (NGST) has emerged, revolutionizing molecular biology and genomics (reviewed in [1]–[3]). These sequencers process short DNA fragments in parallel and provide millions of sequence reads in a single batch, each of which corresponds to a DNA molecule within the sample. While there are several types of NGST platforms and different sets of sequencing reactions, all platforms achieve parallelization using a common concept of immobilizing the DNA fragments to a surface, so that each fragment occupies a distinct spatial position. When the sequencing reagents are applied to the surface, they generate optical signals according to the DNA sequence, which are then captured by a microscope and processed. Since the fragments are immobilized, successive signals from the same spatial location convey the DNA sequence of the corresponding fragment (Fig. 1). Using this approach millions of DNA fragments can be simultaneously sequenced to lengths of tens to hundreds of nucleotides.

It is clear that harnessing next generation sequencers to carrier screens will dramatically increase its utility. The main challenge is how to fully exploit the wide capacity of the sequencers. Dedicating one sequencing batch to genotype

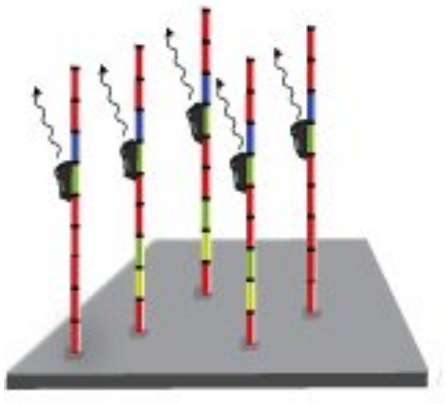


Fig. 1: Illustration of a sequencing round in a next generation platform. The rods represent immobilized DNA fragments. Optical signals (arrows) that correspond to the DNA sequences are captured by the sequencer

dozens of loci in a single individual does not realize even a small fraction of the sequencing power, and in fact, it is even less cost-effective than the serial approach. Therefore, multiplexing large number of specimens in a single batch is essential. The problem is that when specimens are simply pooled and sequenced together the sequence reads reflect only the allele frequencies of the specimens in the pools, and do not provide any information about the particular status of each specimen.

One intriguing solution for the multiplexing task is to employ combinatorial pooling before sequencing - pool the specimens according to a designated pattern that on one hand exploits the sequencing capacity, and on the other enables faithful reconstruction of the individuals' genotypes. Group testing (**GT**) and compressed sensing (**CS**) offer a rich framework for that approach, and earlier this year several groups have independently proposed multiplexing strategies based on those mathematical concepts (summarized in Table II). While these groups considered different biological scenarios, and accordingly different pooling designs and reconstruction algorithms, they reached the same conclusion - identifying rare genetic variations can be obtained by sequencing only small number of pools.

Carrier screen is also a task of finding rare genetic variations, and can be achieved by employing similar ideas. Here, we show how **CS** framework can be used for ultra-high throughout carrier screens. We focus on real problem - finding carriers of genetic diseases that are prevalent in Ashkenazi Jews. In section II, we map the problem to **CS** framework, and in section III we show its rigor by extensive simulations for two diseases that are prevalent in Ashkenazi Jews.

II. Carrier screen using compressed sensing

In **CS** [8], [9] one wishes to efficiently reconstruct an unknown vector of values $\mathbf{x} = (x_1, ..., x_N)$, assuming that $\mathbf{x}$ is *sparse*. It has been shown that $\mathbf{x}$ can be reconstructed using $k \ll N$ basic operations termed *measurements*, where a measurement is simply the output y of the dot-product of the (unknown vector) $\mathbf{x}$ with a known measurement vector $\mathbf{m}$, $y = \mathbf{m} \cdot \mathbf{x}$. By using the output of such k measurements and their corresponding $\mathbf{m}$'s, it is possible to reconstruct the original sparse vector $\mathbf{x}$.

Multiplexed carrier screen can be mapped into a **CS** setting as follows: The entry x_i in $\mathbf{x} = (x_1, ..., x_N)$ is either 0, or 1, and specifies the number of non-functional alleles of the i'th individual. We assume that individuals that carry 2 non-functional alleles are affected by a severe disorder and never participate in the screen. Since we are interested in rare genetic variations, $\mathbf{x}$ is indeed sparse as most individuals have zero copies of the non-functional allele. As for the measurements, a certain pool of individuals is chosen, and equal amounts of their DNA is mixed and then sequenced using NGST. Mathematically this simply means that an entry in the measurement vector $\mathbf{m}$ is 1 if the specific individual was part of the pool, and 0 otherwise. Our measurement y, however, is not a linear projection of $\mathbf{m} \cdot \mathbf{x}$. Instead, the sequencing results are given by the ratio of z, the number of reads corresponding to the non-functional allele, to r, the total number of reads from the specific site. Hence y in our case is an estimation of the frequency of the non-functional allele in a pool, obtained through a binomial sampling process, and given by the following equation:

$$Pr(y = \frac{z}{r}) = \binom{r}{z} p^z (1-p)^{r-z} \tag{1}$$

where

$$p = \frac{\mathbf{m} \cdot \mathbf{x}}{2 \sum_{i=1}^{N} m_i}$$

is the proportion of reads showing the non-functional allele, and for large number of sequence reads $y \cong p$.

We employ k measurements (pools), hence, the different measurement vectors are the rows of the measurement matrix $\mathbf{M}_{k \times N}$, and their results are the entries of $\mathbf{y}$. Although this mapping is rather straightforward, applying **CS** to the problem of detecting carriers has some unique characteristics:

'On a budget' pooling design: The central goal of **CS** is reducing k, the number of measurements, to the required minimum that enables faithful reconstruction of the input signal. This is also a goal of compressed carrier screen since sequencing each pool is rather expensive. However, the pooling procedure itself, namely mixing the specimens prior to their sequencing, has its costs and limitations. Let w be the weight of the design - the column with the maximal number of 1's in $\mathbf{M}$. The weight determines the maximal number of times a specimen is sampled in the pooling. Pooling designs with heavy weight require laborious work of technicians and liquid handling robots and consume more specimen material, which may be very limited. Moreover, pools that are composed of a large number of specimens are prone to non-uniform DNA amplification during the PCR step. Therefore, one unique feature of our compressed carrier screen is to find an 'on a budget' pooling design, such that $\mathbf{M}$

Paper	Type of pooling design	Pooling design	Number of pools	Light Design	Decoding	Remarks
[4]	Deterministic	Logarithmic signature	$O(3\lceil \log_3 N \rceil)$	No	unspecified	Designated for singletons detection
		Extended Golay Code	$24\lceil N/759 \rceil$	Yes		
[5]	Deterministic	Chinese Remainder Theorem	$\Theta(w\sqrt{N} + \frac{1}{2}w^2 \log(w))$	Yes	Minimal Discrepancy	Validated in a real experiment.
[6]	Stochastic	Dense Bernoulli matrix $p = 1/2$	$O(klogN)$	No	GPSR	Shows rare homozygous detection
		Sparse Bernoulli matrix $p = 1/\sqrt{N}$	$O(k\sqrt{N})$	Yes		
[7]	Deterministic	Chinese Remainder Theorem	$\Theta(w\sqrt{N} + \frac{1}{2}w^2 \log(w))$	Yes	Belief Propagation	

TABLE II: Summary of current **GT** and **CS** multiplexing schemes for NGST

will be sparse as possible, while also keeping the number of pools as low as possible. We refer to such pooling schemes as 'light' designs.

Apart from these "pre-sequencing" considerations, a sparse pooling design increases the reliability of sequencing results by reducing sampling noise, which is reminiscent of the dilution effect in **GT** [10], [11]. The total number of reads per pool, r, is determined by the efficiency of the sequencing batch, independently of the number of specimens in the pool. Thus, increasing the number of specimens in a pool reduces the number of reads which originate from each individual and may affect the reliability of y.

Not all pools were created equal: The traditional **CS** framework considers a measurement as a linear projection of the aggregated data points in the presence of additive white Gaussian noise. As we indicated in Eq. (1), the sequence reads reflect only relative frequencies of alleles in the pool, and not the absolute number of non-functional alleles. For example, if one carrier and 99 normal specimens are mixed in equal amounts, then the sequence reads of the non-functional allele will follow a binomial distribution with $p = 1/200$ and r total number of reads. The extent of variation in the number of non-functional allele reads determines the measurement's reliability, as larger number of reads elevates the accuracy in estimating the absolute number of carriers in a pool. However, the variance is also determined by p, the relative number of carriers in the pool, meaning that this type of noise is not additive, but depends on the pool's content. Since pools have different number of sequence reads and different number of carriers, it would be beneficial to have a reconstruction algorithm that weigh measurements according to their reliability, and a pooling design that increases the sequencing reliability.

Pooling imperfections: A third feature of this problem are pooling imperfections that introduce noise to the measurement matrix $\mathbf{M}$, as opposed to a 'standard' **CS** application in which $\mathbf{M}$ is known exactly. For example, it may happen that unequal amount of DNA are taken from each individual, or that small amounts of material from one pool contaminate the following pool to be sequenced, etc. Such problems introduce multiplicative noise into $\mathbf{M}$ which may hinder accurate reconstruction [12].

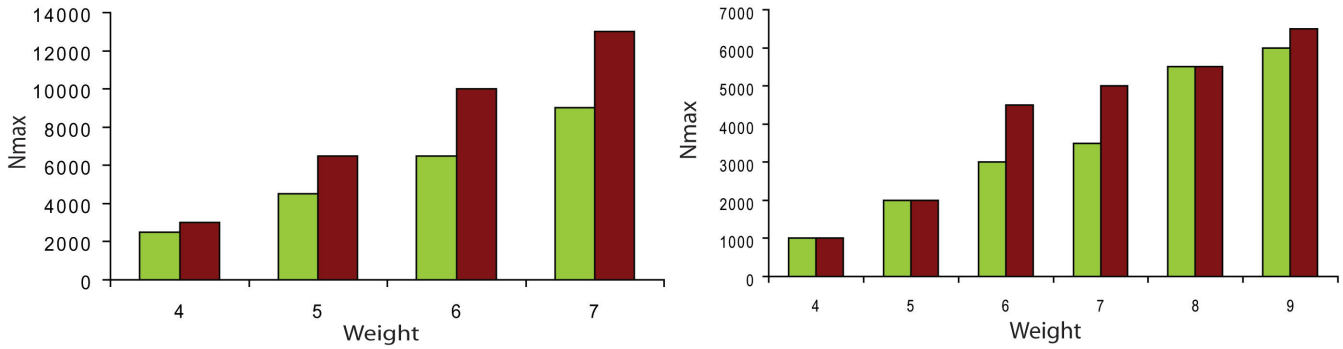
Signal domain: The last unique feature of this problem is related to the fact that in traditional **CS** the transmitted vector $\mathbf{x}$ is assumed to belong to $\mathbb{R}^N$, while in our task, $\mathbf{x} \in \{0,1\}^N$. This property implies that additional post-processing quantization step is required when one employs an off the shelf **CS** reconstruction algorithm. To the best of our knowledge systematic analysis of reconstructing binary input vectors by traditional **CS** reconstructing methods, such as convex relaxation is yet to be studied.

III. Case study - a carrier screen for prevalent genetic disorders in Ashkenazi Jews

In order to evaluate the compressed genotyping methodology, we performed simulations of carrier screens for prevalent genetic disorders in Ashkenazi Jews. This community has been a subject of extensive genetic studies, and about a dozen severe Mendelian diseases were found in relatively high prevalence (see Table III). Carrier screen programs designated for Ashkenazi Jews are widely employed; the Israeli Ministry of Health sponsors a national screening program for Tay-Sachs (TS), Cystic Fibrosis (CF), and Familial Dysautonomia (FD) [19], and similar programs exists for the communities abroad [20]. Screening the bulk population is highly apt to our compressed sensing carrier screen. First, large number of specimens increases the benefit-cost ratio of the screen, and justifies the pooling step. Second, screening large number of specimens is less prone to random fluctuations in the carrier rates. We particularly focused on the leading mutations of two genetic diseases, Tay-Sachs and

Disorder	Main clinical features	Carrier Rate(%)	Reference
Tay-Sachs	Neurodegenerative disorder. Fatal by age of 2 or 3 years	1:25	[13]
Cystic Fibrosis	Pulmonary complications. Median age of death is 30 years	1:30	[14]
Familial Dysautonomia	Severe nervous system impairment. Median age of death is 30 years	1:30	[15]
Canavan	Mental retardation	1:40	[16]
Usher Syndrome	Deaf blindness	1:40	[17]
Bloom	Early cancer onset. Sterility	1:102	[18]

TABLE III: Examples of Typical Genetic Disorders in Ashkenazi Jews



(a) Bloom Syndrome

(b) Tay-Sachs

Fig. 2: Evaluating N_{max} for GPSR (green) and BP (red)

Bloom Syndrome, which are found in the two extremes of the disorders in terms of their relative carrier rate in the Ashkenazi Jews population.

In Tay-Sachs, the leading mutation is a 4 base pairs (bp) deletion in the HEXA gene (MIM: 606869.1), and the estimated carrier rate for that mutation is 2% [13]; In Bloom syndrome, the leading mutation is a 6bp deletion and 7bp insertion in RECQL3 (MIM: 604610.1), and the estimated carrier rate for that mutation is 1% [18]. The genetic alterations in these two cases are quite major, and affect several nucleotides in the sequence. Therefore, one can unambiguously determine whether a sequence read corresponds to the WT allele or to the non functional one. Accordingly, we did not introduce sequencing errors to the simulations.

For the pooling design, we used the light Chinese design, presented in [7]. In brief, this design allows the user to specify the required pooling weight w, i.e., the number of non-zero elements in $\mathbf{M}$'s columns. Hence it is capable of producing very sparse designs, conferring the 'on a budget' requirement. The number of pools in the light Chinese design are roughly $w\sqrt{N}$, and we tested $4 \leq w \leq 7$ for Bloom syndrome, and $4 \leq w \leq 9$ for Tay-Sachs, which mostly fits a single batch of Illumina GAII, the leading NGST platform.

We evaluated two reconstruction approaches: An adaption of Gradient Projection for Sparse Reconstruction (GPSR) [21] to compressed genotyping that was presented in [6], and an improvement of Belief Propagation (BP)-based solver that was presented in [7] (see Appendix for details about the improvements). GPSR and BP represent two common classes of reconstruction approaches in **CS** . The former algorithm is based on a top-down approach, a minimization of the ℓ_2 differences between the reconstructed signal and the observed results with a global cost function that evaluates the solution's sparsity by measuring its ℓ_1 norm. This approach has been widely employed in other **CS** solvers, and in machine learning for spare model selection [22]. The latter algorithm, BP, achieves sparsity by a bottom-up approach of rewarding local consistencies in the genotype assignments that inherently exclude non-sparse solutions. BP as a **CS** solving strategy has gained popularity in the past few years, and several studies demonstrated its rigor [23]–[25]. A noticeable difference between GPSR and BP in solving compressed carrier screen is that GPSR reports $\mathbf{x} \in \mathbb{R}^N$, and requires a rounding procedure to obtain the final results, whereas BP directly computes the marginals for $\mathbf{x} \in \{0,1\}^N$

We measured the success rate of the solvers by evaluating their N_{max}, the largest number of specimens that can be reconstructed with 100% accuracy in 95% of the simulations' instances. We have previously found that N_{max} tightly indicates the decoding boundary, after which a strong phase transition occurs and very low decoding accuracies are observed [6]. In order to find N_{max}, we simulated carrier screens with $1,000$ to $19,000$ specimens in steps of 500 and $1,000$ specimens for the range $1,000 - 10,000$, and $10,000 - 19,000$ correspondingly. For each number of specimens, we generated 50 independent inputs of carrier and non-carrier according to the expected frequency of each disorder. We assumed that the sequencer reports 5,000 reads for each pool.

Remarkably, our results indicate that one can obtain perfect decoding accuracy while providing a significant saving in sequencing costs. For example, accurate reconstruction was achieved even with a pool/specimen ratio as low as 6.5% for carrier screen of a very rare disorder (Bloom) and 12.5% for a more prevalent one (Tay-Sachs), and very sparse design with relatively low weights (Fig. 2). We also found that in our setting the performance of BP outperformed the performance of GPSR. In the Bloom syndrome simulations, regardless of the weight, BP decoded large number of specimens by about $20\% - 25\%$ compared to GPSR. In the Tay-Sachs simulations, BP yields the same N_{max} as for some of the weights, and achieve higher N_{max} for the others. These findings are in agreement with other studies that show a decoding advantage when using BP for sparse **CS** problems (Dror Baron - personal communication). We speculate that the performance advantage of BP, at least in our case, is attributed to its ability to find the assignment directly without any post-processing rounding step that may introduce errors. Another possibility for the advantage of

BP is its inherent mechanism that gives lower weights to less reliable pools with large number of carriers. It is thus possible that running GPSR with a modified cost function taking into account these reliabilities will produce improved results. In addition, we would like to stress the point that in other pooling settings, BP may exhibit lower performance levels, since the light Chinese design ensures that no short cycles occurs in the corresponding factor graph. With respect to the running times, GPSR has a significant advantage over over BP. While a typical BP run took couple of minutes to 1-2 hours (depend on the number of specimens), a typical GPSR run took seconds. For large scale simulations, as in this study, such differences accumulate to many CPU hours. Therefore, we recommend using GPSR for the initial performance estimations in large scale compressed carrier screen simulations.

IV. Conclusion

We proposed an ultra-high throughout carrier screen for rare genetic diseases. Our method harnesses the sequencing power of NGSTs by mapping the multiplexing problem to a **CS** framework, with several unique features. We demonstrated the rigor of the method in extensive simulations of carrier screens for prevalent Ashkenazi disorders, and found that a tailored Bayesian solver outperforms an off-the shelf **CS** solver.

Appendix

We introduced two improvements to the BP algorithm that was presented by us in [7]. The main limitation of the previous method was long running times for inputs with large number of specimens. In the new version, we implemented a stripping method that evaluates which pools (factor nodes) do not have carriers and stripes them off from the factor graph. Each time we remove a factor node, we fixed its connected variable nodes to be 0 (no carrier), and stripped them off also from the factor graph. Then, we updated the observed data in the remained factor nodes to fit the corrected graph.

In order to evaluate whether a factor node contains no carriers, we calculated the relative ratio of the two competing hypothesis: $\mathbf{H_0}$ - there are no carriers in the pool; and $\mathbf{H_1}$ - there is at least one carrier in the pool. Let r be the total number of reads from the pool, f - the fraction of reads that corresponds to the non-functional allele, n - the number of specimens in the pool, c - the expected carrier rate in the population, and α - the sequencing noise level, namely the probability that a sequence read from a non-functional allele will be reported as WT, and vise verse (we assume symmetric sequencing errors). $\mathbf{H_0}$ and $\mathbf{H_1}$ are given by:

$$\Pr(\mathbf{H_0}) = q_0 p_0 \tag{2}$$

$$\Pr(\mathbf{H_1}) = \sum_{k=1}^{n} q_k p_k \tag{3}$$

where p_k denotes the probability of having k carriers in the pool, and q_k denotes the likelihood of the data given that there are k carriers in the pool. p_k and q_k are Poisson probabilities given by:

$$p_k = \text{Pois}(k; nc) \tag{4}$$

$$q_k = \text{Pois}(fr; \rho r) \tag{5}$$

and:

$$Pois(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{6}$$

$$\rho = \alpha + \frac{k(1-2\alpha)}{2n} \tag{7}$$

Once the ratio $\Pr(\mathbf{H_0})/\Pr(\mathbf{H_1})$ crossed a user-determined threshold, we applied the stripping procedure to the factor node, and to its connected variables.

We then updated the observed data of the factors nodes that were connected to a stripped variable node. Let:

$$a = (1-f)r - z(1-\alpha) \tag{8}$$

$$b = fr - z\alpha \tag{9}$$

$$z = r/n \tag{10}$$

The new number of reads (r') and the new fraction of non-functional reads (f') in the factor node was set to:

$$r' = a + b \tag{11}$$

$$f' = \frac{a}{r'} \tag{12}$$

where f is the previous fraction of reads from the non-functional allele, and r is the previous number of reads in that factor node.

An additional improvement of the BP algorithm was considering two damping levels. We noticed in our previous study [7] that the convergence probability increased once the damping level crossed a threshold. Thus, instead of using a fixed damping level, we used here two damping values: 0.5 and 0.8, and if the algorithm did not converge in the lower value, it tried the higher one. We found that the stronger damping level (0.8) facilitated the convergence of heavy weight inputs ($w \geq 6$), while the weaker damping level (0.5) facilitated the convergence of light weight inputs. In is worth noting that in other studies that used BP solvers for **CS** , damping did not play any role in the convergence rate (Dror Baron - personal communication). It will be interesting in the future to study the the sensitivity of **CS** solvers to the damping level.

Acknowledgments

The authors thank Igor Carron and Nuit Blanche blog for providing useful information and communicating our results. Y.E is a Goldberg-Lindsay Fellow and ACM/IEEE Computer Society High Performance Computing PhD Fellow of the Watson School of Biological Sciences.

References

[1] M. L. Metzker, "Emerging technologies in DNA sequencing," *Genome Res.*, vol. 15, pp. 1767–1776, Dec 2005.

[2] K. R. Chi, "The year of sequencing," *Nat. Methods*, vol. 5, pp. 11–14, Jan 2008.

[3] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nat. Biotechnol.*, vol. 26, pp. 1135–1145, Oct 2008.
[4] S. Prabhu and I. Pe'er, "Overlapping pools for high-throughput targeted resequencing," *Genome Res.*, vol. 19, pp. 1254–1261, Jul 2009.
[5] Y. Erlich, K. Chang, A. Gordon, R. Ronen, O. Navon, M. Rooks, and G. J. Hannon, "DNA Sudoku–harnessing high-throughput sequencing for multiplexed specimen analysis," *Genome Res.*, vol. 19, pp. 1243–1253, Jul 2009.
[6] N. Shental, A. Amir, and O. Zuk, "Rare-Allele Detection Using Compressed Se(que)nsing," 2009, 0909.0400v1.
[7] Y. Erlich, A. Gordon, M. Brand, G. Hannon, and P. Mitra, "Compressed Genotyping," 2009, 0909.3691v1.
[8] E. Candes, "Compressive sampling," in *Int. Congress of Mathematics*, (Madrid, Spain), pp. 1433–1452, 2006.
[9] D. Donoho, "Compressed sensing," *IEEE Transaction on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
[10] G. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," Jul 2009, 0907.1061.
[11] K. R. and W. P., "Multiplexed experiment design for high-throughput screening."
[12] M. A. Herman and T. Strohmer, "General deviants: An analysis of perturbations in compressed sensing," *CoRR*, vol. abs/0907.2955, 2009.
[13] N. Risch, "Molecular epidemiology of Tay-Sachs disease," *Adv. Genet.*, vol. 44, pp. 233–252, 2001.
[14] S. Orgad, S. Neumann, R. Loewenthal, I. Netanelov-Shapira, and E. Gazit, "Prevalence of cystic fibrosis mutations in Israeli Jews," *Genet. Test.*, vol. 5, pp. 47–52, 2001.
[15] S. L. Anderson, R. Coli, I. W. Daly, E. A. Kichula, M. J. Rork, S. A. Volpi, J. Ekstein, and B. Y. Rubin, "Familial dysautonomia is caused by mutations of the IKAP gene," *Am. J. Hum. Genet.*, vol. 68, pp. 753–758, Mar 2001.
[16] O. N. Elpeleg, Y. Anikster, V. Barash, D. Branski, and A. Shaag, "The frequency of the C854 mutation in the aspartoacylase gene in Ashkenazi Jews in Israel," *Am. J. Hum. Genet.*, vol. 55, pp. 287–288, Aug 1994.
[17] T. Ben-Yosef, S. L. Ness, A. C. Madeo, A. Bar-Lev, J. H. Wolfman, Z. M. Ahmed, R. J. Desnick, J. P. Willner, K. B. Avraham, H. Ostrer, C. Oddoux, A. J. Griffith, and T. B. Friedman, "A mutation of PCDH15 among Ashkenazi Jews with the type 1 Usher syndrome," *N. Engl. J. Med.*, vol. 348, pp. 1664–1670, Apr 2003.
[18] C. Oddoux, C. M. Clayton, H. R. Nelson, and H. Ostrer, "Prevalence of Bloom syndrome heterozygotes among Ashkenazi Jews," *Am. J. Hum. Genet.*, vol. 64, pp. 1241–1243, Apr 1999.
[19] G. Rosner, S. Rosner, and A. Orr-Urtreger, "Genetic testing in Israel: an overview," *Annu Rev Genomics Hum Genet*, vol. 10, pp. 175–192, 2009.
[20] J. Ekstein and H. Katzenstein, "The Dor Yeshorim story: community-based carrier screening for Tay-Sachs disease," *Adv. Genet.*, vol. 44, pp. 297–310, 2001.
[21] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 586–597, 2007.
[22] J. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, pp. 1030–1051, March 2006.
[23] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," Dec 2008, 0812.4627.
[24] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *Signal Processing, IEEE Transactions on*, vol. 56, pp. 2346–2356, June 2008.
[25] M. A. Sheikh, S. Sarvotham, O. Milenkovic, and R. G. Baraniuk, "Dna array decoding from nonlinear measurements by belief propagation," in *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on*, pp. 215–219, Aug. 2007.