7

The Application of Pooled DNA Sequencing in Disease Association Study

Chang-Yun Lin[1] and Tao Wang[2]
[1]*McDermott Center of Human Growth and Development and Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX,*
[2]*Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA*

1. Introduction

Hundreds of common genetic variants related to the risk of human disease, such as diabetes, hypertension, bipolar, and Crohn's disease, have been successfully discovered by Genome-wide Association Studies (GWAS) (Barret et al., 2008; Hindorff, 2009; Thomas et al., 1991; WTCCC, 2007). Current GWAS are based on the strategy of linkage disequilibrium (LD) mapping, in which a sufficient number of single nucleotide polymorphism (SNP) markers are selectively genotyped to capture the genetic variation of the whole genome. However, there are two major issues related to the results of GWAS. First, the results only explain a small fraction of the heritability of complex diseases. One of the reasons may be that many functional variants, in particular rare variants, which are not directly genotyped in GWAS, have a weak LD with SNP markers, and hence are missed by GWAS (Iyengar et al., 2004; Manolio et al., 2009). Second, the identified associations in GWAS are often inconsistent between different populations. The reason for this may be the varied LD structures between markers and underlying causal variants among populations, resulting in associations can only be observed in specific populations.

To address these issues, an ideal approach is to directly sequence all the samples in a study (Bodmer & Bonilla, 2008). However, this is not a feasible option for the traditional sequencing technology, namely Sanger sequencing, which is extremely expensive and time consumption for sequencing thousands of samples required to achieve reasonable statistical power in a typical genetic association study.

Next generation sequencing (NGS) technology, also called parallel sequencing, is a revolutionary technology for biomedical research (Shendure & Ji, 2008). The production of large numbers of low-cost reads makes NGS useful for many applications. Today there are three commonly-used next-generation sequencing systems: namely Roche's (454) GS FLX Genome Analyzer marketed by Roche Applied Sciences, Illumina's Genome Analyzer" (GA), and Applied Biosystem's SOLiD system. Several new systems have either just been introduced or will become available soon (Metzker, 2010). One of the most important applications is to identify DNA variants, in particular rare variants, responsible for human

diseases (Metzker, 2010). Now ten billion bases can be obtained routinely in a single run of NGS instrument and yields are expected to continually increase. The throughput of the smallest function unit, e.g., a single 'lane', can generate data amounting to many thousands fold coverage for a target region, which is far greater than what is needed for genotyping one individual as the individual genotype at a specific locus is expected to be accurately called at about 15-30 fold coverage. As such, it is feasible to simultaneously sequence targeted regions of multiple individuals with dramatic saving on cost and time.

To reduce the cost of large-scale association studies, one efficient approach is to sequence a large number of individuals together on a single sequence run. Two commonly-used approaches are available in disease association studies. Bar-coding ligates the DNA fragments of each sample to a short, sample-specific DNA sequence, and then sequences these DNA fragments from multiple subjects in one single sequencing run. In addition to allowing determining individual genotypes, bar-coding offers an additional advantage of reduction of sequencing variability (Craig et al. 2008). However, bar-coding at present has a limit of the multiplexing and the cost on the individual DNA amplification and sequencing template preparation could be substantial in large scale disease-association studies. Compared to bar-coding, simply pooling DNA samples is more cost-effective as it can fully make use of the high depth of sequencing and vastly reduce the efforts of sample preparation for thousands of individuals. Currently pooled DNA sequencing is particularly appealing due to its substantial cost and time-saving in large disease-association studies, i.e. pooled DNA sequencing (Shaw et al., 1998). With pooling, the sequencing throughput required per individual is much less than what is provided by a single run, and hence it is feasible to sequence multiple individuals together. For example, in a case-control study, the allele frequencies in a sample of 500 cases and 500 controls can be measured from two pooled samples, rather than from 1,000 individual samples, which represents an increase in efficiency of 500-fold.

Pooling was first used in genetic study in a case-control association study of HLA class II DR and DQ alleles in type I diabetes mellitus (Arnheim et al., 1985). Afterwards, it has been used for linkage studies in plants (Michelmore et al., 1991), for the homozygosity mapping of recessive diseases in inbred populations (Sheffield et al., 1994; Carmi et al., 1995; Nystuen et al., 1996; Scott et al., 1996), and for mutation detection (Amos et al., 2000). This strategy was also proposed for high-throughput SNP arrays (Ito et al., 2003; Shaw et al., 1998; Zeng & Lin, 2005) but it was not widely accepted as SNP array technology does not provide accurate estimates of the allele frequencies in the pooled samples. Recent next generation sequencing technology provides a high-throughput sequencing solution for examining functional variants directly. It might provide more accurate estimates of allele frequency, as shown by recent studies (Druley et al., 2009). Recently, one study adopted this strategy using 454 sequencing technology and identified associations of rare variants with insulin-dependent diabetes mellitus (Nejentsev et al., 2009). In a genome-wide analysis studies, two-stage design and DNA pooling could be used as a cost-efficient strategy to detect genetic variant regions (Chi et al., 2009; Skol et al., 2006; Wang et al., 2006; Zuo et al., 2006, 2008). In the first stage, a fraction of samples are genotyped for all SNPs and a case-control association test for each SNP is then conducted to select the most significant SNPs. In the second stage, the candidate SNPs from the first stage are further evaluated by genotyping. To reduce the cost of large-scale association studies in two-stage design, pools of DNA from many individuals have been

successfully used in the first stage of the two-stage design (Bansal et al., 2002; Boss et al., 2009; Nejentsev et al., 2009; Norton et al., 2004; Sham et al., 2002). As suggested by Out et al. (2009), the use of a pooled DNA sample for targeted regions, NGS also can be an attractive cost- effective method to identify rare variants in candidate genes.

The data produced by next-generation sequencing is different from that of SNP-chips. Next-generation re-sequencing produces large amounts of short reads. After mapping to the reference genome, an alignment of reads across the targeted regions is obtained. A schematic example of re-sequencing data in case-control study is shown in Table 1. In this example, each case and control sample consists of two pools with two individuals in each pool. The two alleles (A and a) of each individual are shown in the "Genotype" column. Each allele appears a random number of times. Although NGS have the potential to discover the entire spectrum of sequence variations in a sample of well-phenotyped individuals, NGSs also present challenges. First, the error rate of these platforms is higher than conventional sequencing methods, and many errors are not random events (Johnson & Slatkin, 2008; Chaisson et al., 2009; Lynch, 2009; Bansal et al., 2010b). These errors may be frequent enough to obscure true associations or systematic enough to generate false-positive associations. Second, the data produced by next-generation sequencing often lose linkage disequilibrium (LD) information which is lost in pooled sequencing. As the result, the powerful analytic approaches that combine multiple rare variants to examine the disease association are not directly applicable to pooled sequencing, because these approaches require individual genotypes to account for the LD between SNPs. The current single locus analysis of pooled sequencing data could be very inefficient, in particular, for rare variants.

	Pool	Individual	Genotype	Read base
Case	Pool 1	1	A	A,A,A,a,A,A,A,A,A,A,A,A
			a	a,a,a,a,a,a,a,a,a,a,a,a
		2	A	A,A,A,A,A,A,A,A,A
			a	a,a,A,a,a,a,a,a,a,a,a
	Pool 2	3	a	a,a,a,a,a,a,a,a,a,a,a,a,a,a
			a	a,a,a,a,a,a,a,a,a,a,a,a
		4	A	A,A,A,A,A,A,A,A,A,A,A,A
			a	a,a,a,a,a,a,a,a,a,a,A,a,a
Control	Pool 3	5	a	a,a,a,a,a,a,a,a,a,a,a,
			a	a,a,a,a,a,a,a,a,a,a,a,a,a,a
		6	a	a,a,a,a,a,a,A,a,a,a,a,a,a,a,a
			a	a,a,a,a,a,a,a,a,a,a,a,a,a
	Pol 4	7	A	a,A,A,A,A,A,A,A,A,A,A,A
			a	a,a,a,a,A,a,a,a,a,a,a,a,a,a
		8	a	a,a,a,a,a,a,a,a,a,a,a,a,a,a
			a	a,a,a,a,a,a,a,a,a,a,a,a,a,a,a,a

Table 1. Example of re-sequencing data in case-control. Each case and control sample consists of two pools with two individuals in each pool.

In section 2 of this chapter, we will introduce some strategies of pooling design, including PI-deconvolution, shifted-transversal design, multiplexed scheme, and overlapping pools to recover LD information. Through these well-chosen pool designs, the variant carriers can be clearly identified, which greatly enhances the pooling efficiency. In section 3, we will introduce some statistical methods for the detection of variant and case-control association study to account for high-levels of sequencing errors. A briefly summary is added in the end of this chapter.

2. Strategies of pooling design

The main idea of pooling is to sequence DNA from several individuals together on a single run. Through the observed number of re-sequencing alleles, the allele frequency can be estimated. The simplest strategy is the naïve-pooling scheme, which is also called disjoint pooling. In naïve-pooling scheme, DNA was sequenced from several individuals on a single pool and each pool includes different individuals (Table 2). It offers insight into allele frequencies, but is not able to the identity of an allele carrier.

Recently, several strategies of well-chosen pools aiming to identify variant are proposed. In these designs, each individual is tested several times in different pools. This redundancy provides a potential increase in both sensitivity and specificity. We will introduce PI-deconvolution, shifted-transversal design, multiplexed scheme, and overlapping pools.

	Individuals															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pool 1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pool 2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Pool 3	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
Pool 4	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
Pool 5	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Pool 6	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
Pool 7	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
Pool 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Table 2. Re-sequencing with naïve pooling scheme. A total of 16 individuals are divided into groups of two and pooled.

PI-deconvolution (Jin et al., 2006) The PI-deconvolution approach is a classic grid design. This strategy assigns individuals on an imaginary grid and construct pools by each row and each column. The individuals with variant then can usually be identified from the pattern of pools appearing variant. If there is a confounding among individuals, only a few candidates need to be retested. For example (Table 3), 16 individuals are arrayed on an imaginary grid and mixed in 8 pools, each containing 4 individuals (individuals 1, 2, 3, and 4 are in pool 1 and individuals 1, 5, 9, and 13 are in pool 5). If the pools 3 and 6 appeal variant, then individual 10 is the only variant carrier. If pools 2 and 7 also appear variant, we cannot distinguish whether the variant is from individuals 6 and 11 or from individuals 7 and 10. To resolve this confounding, we can add four additional pools, built along one of the grid's diagonals as indicated by the colors of the individuals. If the pink diagonal pool appears variant, individuals 6 and 11 are the variant carriers, whereas if both the orange and blue

diagonal pools appear variant, the variant is from individuals 7 and 10. The author has validated the technique in three experimental contexts: protein chips, yeast two-hybrid assay, and drug resistance screening.

Shifted-Transversal Design (Thierry-Mieg, 2006) This method minimizes the co-occurrence of objects and constructs pools of constant-sized intersection. They proved that it allows unambiguous decoding of noisy experimental observations. It is highly flexible and can be tailored to function robustly in a wide range of experimental settings. Let $n \geq 2$, and consider the set $\mathcal{A}_n = \{A_0, \cdots, A_{n-1}\}$ of n Boolean variables. Let σ_q be the mapping of $\{0,1\}^q$ onto itself defined by:

	Pool 5	Pool 6	Pool 7	Pool 8
Pool 1	1	2	3	4
Pool 2	5	6	7	8
Pool 3	9	10	11	12
Pool 4	13	14	15	16

Table 3. Example for the strategy of PI-deconvolution. Samples are arranged in to an array, in which each row and each column represents a pool. In this example, 16 individuals are well-chosen into 8 pools.

$$\forall\,(x_1, \cdots, x_q) \in \{0,1\}^q, \sigma_q \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} = \begin{bmatrix} x_q \\ x_1 \\ \vdots \\ x_{q-1} \end{bmatrix}.$$

For every $j \in \{0, \cdots, q\}$, let M_j be a $q \times n$ Boolean matrix, defined by its columns $C_{j,0}, \cdots, C_{j,n-1}$ as follows:

$$C_{0,0} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, and\ \forall i \in \{0, \cdots, n-1\}\ C_{j,i} = \sigma_q^{s(i,j)}(C_{0,0}),$$

where $s(i,j) = \sum_{c=0}^{\zeta} j^c \cdot \left\lfloor \frac{i}{q^c} \right\rfloor\ if\ j < q, and\ \ s(i,q) = \left\lfloor \frac{i}{q^\zeta} \right\rfloor$, the semi-bracket denotes the integer part, and ζ denotes the smallest integer γ such that $q^{\gamma+1} \geq n$. Let $L(j)$ be the set of pools of which M_j is the matrix representation. For $k \in \{1, \cdots, q+1\}$, the transversal pooling design is defined as

$$STD(n; q; k) = \bigcup_{j=0}^{k-1} L(j).$$

For example, consider the variable $\mathcal{A}_9$ and $q = 3$, we have

$$M_0 = \begin{bmatrix} 1\,0\,0\,1\,0\,0\,1\,0\,0 \\ 0\,1\,0\,0\,1\,0\,0\,1\,0 \\ 0\,0\,1\,0\,0\,1\,0\,0\,1 \end{bmatrix}, M_1 = \begin{bmatrix} 1\,0\,0\,0\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,0\,0\,0\,1 \\ 0\,0\,1\,0\,1\,0\,1\,0\,0 \end{bmatrix}, M_2 = \begin{bmatrix} 1\,0\,0\,0\,1\,0\,0\,0\,1 \\ 0\,1\,0\,0\,0\,1\,1\,0\,0 \\ 0\,0\,1\,1\,0\,0\,0\,1\,0 \end{bmatrix},$$
$$M_3 = \begin{bmatrix} 1\,1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,1\,1 \end{bmatrix}.$$

The corresponding layers of pools are the following:

Layer 0: $L(0) = \{\{A_0, A_3, A_6\}, \{A_1, A_4, A_7\}, \{A_2, A_5, A_8\}\}$,

Layer 1: $L(1) = \{\{A_0, A_5, A_7\}, \{A_1, A_3, A_8\}, \{A_2, A_4, A_6\}\}$,

Layer 2: $L(2) = \{\{A_0, A_4, A_8\}, \{A_1, A_5, A_6\}, \{A_2, A_3, A_7\}\}$,

Layer 3: $L(3) = \{\{A_0, A_1, A_2\}, \{A_3, A_4, A_5\}, \{A_6, A_7, A_8\}\}$.

The shifted-transversal design is

$$STD(9;3;2) = L(0) \cup L(1) = \{\{A_0, A_3, A_6\}, \{A_1, A_4, A_7\}, \{A_2, A_5, A_8\}, \{A_0, A_5, A_7\}, \{A_1, A_3, A_8\}, \{A_2, A_4, A_6\}\}.$$

Suppose that a single variable in $\mathcal{A}_9$ is A_8. Then pools $\{A_2, A_5, A_8\}$ and $\{A_1, A_3, A_8\}$ are positive (appealing variant), which each proves that A_8 is positive.

Multiplexed scheme (Erlich et al., 2009) In this scheme, several pooling groups are created and the individuals are assigned to pools in each group by taking use of the Chinese remainder, one of the most ancient and fundamental in number theory (Andrews 1994; Ding et al. 1996; Cormen et al. 2001). To create a w pooling groups design, the rule of pooling for group k is

$$n_k = r_k \ (mod\ x_k),$$

which brings the $r_k, r_k + x_k, r_k + 2x_k, \cdots$ individuals to the rth pools of group k, where $0 < r_k \leq x_k$ and $k = 1, \cdots, w$. According to the pooling pattern in each group, only a single, high-confidence solution would emerge for the vast majority of samples. For example (Table 4), let us create 2 groups for 20 individuals according to the following two pooling rules:

$$\begin{bmatrix} n_1 = r_1 (mod\ 5) \\ n_2 = r_2 (mod\ 8) \end{bmatrix}.$$

The total number of pools in this design is 13 (5+8). The corresponding pooling matrix is a 13×20 table and partitioned into two regions that correspond to the two pooling patterns. The staircase pattern (high- lighted in yellow) in each region is typically created by the multiplexed scheme. If pool 1 in group 1 and pool 6 in group 2 appear variant, then individual 6 can be identified as the variant carrier.

Overlapping pools (Prabhu & Pe'er, 2009) The central idea of overlapping pool design is that while sequencing DNA from several individual on a single pool, they also sequence DNA from a single individual on several pools. Individuals are assigned to pools in a manner so as to create a code: a unique set of pools for each individual. This set of pools on which an individual is sequenced defines a code word, or pool signature. If a variation is observed on the signature pools of one individual and on no other, then we identify the variant carrier. Based on the overlapping design, author proposed two algorithms for pool design: logarithmic signature designs and error-correcting designs. They showed that their designs guarantee high probability of unambiguous singleton carrier identification while maintaining the features of naïve pools in terms of sensitivity, specificity, and the ability to estimate allele frequencies.

		Individual																			
Group	Pool	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
	2	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
	3	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
	4	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
	5	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
2	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
	2	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
	3	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
	4	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
	5	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	6	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0

Table 4. The matrix is partitioned into two regions that correspond to the two pooling patterns. The staircase pattern (highlighted in yellow) is created by the multiplexed scheme.

	Individual															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pool 1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Pool 2	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Pool 3	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Pool 4	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Pool 5	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
Pool 6	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
Pool 7	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
Pool 8	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

Table 5. Sixteen distinct pool signatures are created using just eight pools in the over lapping design. If pools 1, 4, 6, and 7 appear variant, the variant carrier 10 can be identified.

3. Statistical methods for pooled DNS in GWAS

GWAS have successfully identified hundreds of variants that are associated with complex traits and pooled DNA sequencing has been considered a cost-effective approach for study rare variants in large populations. In this section, we discuss the statistical methods for the detection of variants and the case-control studies.

3.1 Detection of variants

SNPSeeker (Druley et al., 2009) This method (SNPSeeker) is an algorithm based on large deviation theory. It uses a seconder dependency error model for single-nucleotide polymorphism identification and takes into account the position in the sequencing read and the identity of the two upstream bases. This algorithm greatly improved the specificity of

SNP calling. The statistical models can be described as follows. Let $x \in \{A, C, G, T, N\}$ denote a observed base and $m \in \{A, C, G, T\}$ denote a base in the reference. The subset of nucleotides for each cycle j, sequencing run d and strand s can be defined as n i.i.d. random variables $x_{j,d,s,1}, x_{j,d,s,2}, \cdots, x_{j,d,s,n}$ and the empirical probability distribution can be written as

$$P_{j,d} = \left(\frac{As}{n}, \frac{Cs}{n}, \frac{Gs}{n}, \frac{Ts}{n}, \frac{Ns}{n}\right).$$

Under null hypothesis of no polymorphism at position i, the distribution of x is

$$Q_{j,d,s} = \sum_{m \in \{A,C,G,T\}} \Pr(x|M_i = m, j, d) * \Pr(M_i = m|s, \tau),$$

where $\Pr(x|M_i = m)$ is the probability of seeing a base x in the sequence at cycle position j on run d given that the original base at position i in the reference M_i is equal to m, and $\Pr(M_i = m|s, \tau)$ is the probability of observing nucleotide m in the reference sequence at position i, $M_i = m$ given the strand s and the true allele frequency vector τ. The cumulative p-value for each strand can be calculated by

$$\prod_d \prod_j 2^{-nD(P_{j,d,s}||Q_{j,d,s})}$$

where $D(P_{j,d}||Q_{j,d})$ is the Kullback-Leibler distance (Thomas & Joy, 1991) between $P_{j,d}$ and $Q_{j,d}$. Bonferoni-corrected is conducts for the total number of tests performed at each position in the reference sequence. The software for SNPseeker algorithm can be found at http://www.genetics.wustl.edu/rmlab/.

CRISP (Bansal, 2010a) This approach compares the distribution of allele counts across multiple pools using contingency tables and evaluates the probability of observing multiple non-reference base calls due to sequencing errors alone. The number of reads with the reference and alternate alleles at a particular position across the k pools can be modeled as a contingency table T^0 with two bases (rows) and k pools (columns) with row sums: $A = \sum_i a_i$ and $R - A = \sum_i r_i - a_i$ and column sums $r_i (1 \le i \le k)$:

	Pool 1	Pool 2	…	Pool k	Total
Reference base	$r_1 - a_1$	$r_2 - a_2$		$r_k - a_k$	$R - A$
Alternate base	a_1	a_2		a_k	A
Total	r_1	r_2		r_k	R

Under null hypothesis, the probability of the observed read can be defined as the probability of the table T^0:

$$P(T^0) = \binom{r_1}{a_1} \times \cdots \times \binom{r_k}{a_k} / \binom{R}{A}.$$

The p-value associated with the observed table T^0 is defined as the sum of all $2 \times k$ contingency tables with identical row and column sums that have equal or lower probability than the observed table:

$$p = \sum_{T \in \Gamma \ s.t. \ P(T) \le P(T^0)} P(T),$$

where Γ represents the set of all $2 \times k$ contingency tables with the same marginal sums as T^0. The p-value can be computed by Monte Carlo method:

1. Initialize $t = 0$
2. Initialize an array P of size R with $P[i] = p$ for $r_1 + \cdots + r_{p-1} \le i \le r_1 + \cdots + r_p$ $(1 \le i \le k)$
3. For $i = 1, \cdots, k$, set $a_i = 0$
4. For $i = 1, \cdots, N$, do the following:
   a. Set $P(T') = 1/\binom{R}{A}$
   b. For $a = 1, \cdots, A$, do
      i. Randomly select an integer r in the interval $[a, R]$
      ii. Set $j = P[r]$ and swap the elements $P[a]$ and $P[r]$
      iii. $P(T') = P(T') \times \frac{r_j - a_j}{a_j + 1}$, $a_j = a_j + 1$
   c. If $P(T') \le P(T^0)$: $t = t + 1$
   d. For any pool j chosen in step (2), set $a_j = 0$
5. The estimated p-value is $\frac{t+1}{N+1}$.

For example (Table 6), in contingency table (A), the p-value is 0.002 suggesting that the five base calls represent a rare SNP rather than sequencing errors; in contingency table (B), the p-value is 0.24 indicating that the presence of five alternate base calls in a single pool is likely due to sequencing error.

(A)	Pool 1	Pool 2	Pool 3	Pool 4
Reference base	42	40	44	50
Alternate base	0	5	0	0
(B)	Pool 1	Pool 2	Pool 3	Pool 4
Reference base	41	40	42	49
Alternate base	1	5	2	1

Table 6. Example of contingency tables for CRISP analysis.

3.2 Detect association base on case-control study

The model in case-control study can be described as follows. Let n_{ij} be the total number of chromosome segments of the region of interest in the jth pool of phenotypic group i, where $i = 1$ for case and 2 for control. Let z_{ij} be the unknown number of rare allele at a specific locus of interest. After re-sequencing, a total number of m_{ij} sequencing reads at the loci are observed and x_{ij} out of m_{ij} read report variant. We denote the random vector $Z_i = (Z_{i1}, \cdots, Z_{il_i})$ and $X_i = (X_{i1}, \cdots, X_{il_i})$. Let p_i be the frequency of the minor allele at this locus for group i. The question we are interested in the case-control study is whether this locus is associated with the disease. The hypothesis of the association can be written as:

$$H_0: p_1 = p_2 = p_0 \ versus \ H_1: p_1 \ne p_1$$

Several statistical methods can be utilized for this test.

Fisher's exact test The allele frequencies of cases and controls are calculated from the observed numbers of total reads and from the number of reads reporting the variant; and the number of variants carried by each phenotypic group is estimated by $\widehat{z_i} = \widehat{p_i} n_i = (\sum x_{ij} / \sum m_{ij}) \times n_i$, where $\widehat{p_i}$ is the estimated allele frequency for group i. The data are then summarized in a 2×2 table (Table 7) with the same row and column margins that have probabilities less than or equal to that of observed table. Although this test is simple to implement, it treats the estimated numbers of the rare variant as if they were observed without considering the uncertainty of such estimates. Thus, it may have an inflated Type I error rate. Second, the sampling scheme of Fisher's exact test is based on the hypergeometric distribution, which in principal requires both the column and row marginal totals of a 2×2 table are fixed, i.e., both the sample size (the number of cases and controls) and the number of variants and non-variants are fixed in Table I. The number of variants and non-variants are usually not fixed in a genetic case-control study. Fisher's exact test used in this way can become very conservative (Upton, 1982). Finally and most importantly, because next-generation sequencing has a relatively high rate of base-calling error and from sequence reads of pooled samples, it is difficult to distinguish true rare variants from such errors. For a rare variant whose frequency is not much higher than the error rate, the power to detect its association with a disease would be very low without adjusting for such error in the statistical method.

	Case	Controls	Total
Variant	$\widehat{z_1}$	$\widehat{z_2}$	$\hat{z}$
Non-variant	$n_1 - \widehat{z_1}$	$n_2 - \widehat{z_2}$	$n - \hat{z}$
Total	n_1	n_2	n

Table 7. A 2×2 table for Fisher's exact test.

Combined Z-test (Abraham et al., 2008) This method combines chi-square statistic and Z-statistic for testing the differences in mean allele frequencies between cases and controls. The general description of this statistic has been presented in (Sham et al., 2002; Macgregor, 2007; Kirov et al., 2009):

$$T_{comb} = \frac{\left(\bar{f}^{(2)} - \bar{f}^{(1)}\right)^2}{v_2 + v_1 + \varepsilon_2^2 + \varepsilon_1^2}$$

where $\bar{f}_i = \frac{1}{l_i} \sum_{j=1}^{l_i} f_j^{(i)}$ is the mean of the allele frequencies over K_i pool replicates, $v_i = \frac{\bar{f}_i(1-\bar{f}_i)}{2n_i}$ is the binomial sampling variance and n_i is number of controls and cases respectively $(i = 1, 2)$ and $\varepsilon_i^2 = \frac{1}{n_i(n_i-1)} \sum_{j=1}^{l_j} \left(f_j^{(i)} - \bar{f}_i\right)^2$ is the square of the standard error due to experimental error. This method considers sampling error and experimental error, which is equivalent to a simplified version of the complex regression model suggested by Macgregor (2007).

Likelihood ratio test (Kim et al., 2010) To quantify he sequencing error rate in pooled sequencing, one approach is to include a control DNA sequence in each pool, which makes it possible to obtain the empirical distribution of the sequencing error of individual pools. Let $\upsilon \in \{A, C, G, T\}$ be the variant and $\gamma \in \{A, C, G, T\}$ be the reference allele. We define $\tau_{1,ij} = \Pr(\upsilon|\tau, i, j)$ to be the false positive error rate, i.e., the probability of reporting a variant given the reference base, and $\tau_{2,ij} = \Pr(\upsilon|\upsilon, i, j)$ to be the true-positive rate, i.e., the probability of reporting a variant given

the variant based. Both the false-positive and true-positive rates can be estimated for each pool by the proportion of reads reporting the variant for the reference base and the variant base, respectively. The estimate of allele frequency with error can be calculated as

$$\theta_{ij} = \frac{z_{ij}}{n_{ij}}(1-\tau_{2,ij}) + \left(1-\frac{z_{ij}}{n_{ij}}\right)\tau_{1,ij}.$$

For next generation sequencing data, the likelihood can be computed as:

$$L(p|X) = \prod_{j=1}^{l_i}\sum_{z_{ij}=0}^{n_{ij}} Binominal(m_{ij}, x_{ij}, \theta_{ij}) Binomial(n_{ij}, z_{ij}, p_i)$$
$$= \prod_{j=1}^{l_i}\sum_{Z_{ij}=0}^{n_{ij}} \binom{m_{ij}}{x_{ij}} (\theta_{ij})^{x_{ij}}(1-\theta_{ij})^{m_{ij}-x_{ij}} \binom{n_{ij}}{z_{ij}} p_i^{z_{ij}}(1-p_i)^{n_{ij}-z_{ij}}$$

where θ_{ij} is as defined above. Then the Likelihood ratio statistic is computed as

$$LRT = -2log\frac{L(\hat{p}_0|X)}{L(\hat{p}_1,\hat{p}_2|X)}.$$

Reject H_0 if $LRT \geq \chi^2(1)$.

Differential test (Wang et al., 2010) Following the approaches of Liddell (1976) and Barry & Choongrak (1900), Wang et al. (2010) defined a test statistic by

$$T_{X_1,X_0} = \frac{\sum_j \frac{w_{1j}}{\sum w_{1j}} X_{ij}/m_{ij} - \sum_j \frac{w_{2j}}{\sum w_{2j}} X_{2j}/m_{2j}}{\sqrt{\hat{V}_X}}$$

where $\hat{V}_X = \sum_i \frac{\hat{p}_i}{\sum\left[\frac{m_{ij}n_{ij}}{m_{ij}+n_{ij}}\right]}$, $\hat{p}_i = \sum_j \left(\frac{w_{ij}}{\sum w_{ij}}\right) X_{ij}/m_{ij}$for $i = 1, 2$, and the weight w_{ij} is inversely proportional to the variance of X_{ij}/m_{ij}. Under null hypothesis, $\hat{p}_i = \hat{p}_0 = \sum_i \sum_j \left(\frac{w_{ij}}{\sum w_{ij}}\right) X_{ij}/m_{ij}$. The p-value can be calculated as

$$P\left(T_{X_1,X_2} \geq T_{x_1,x_2}\middle|H_0\right)$$
$$= \sum_{X_{11}=0}^{m_{11}} \cdots \sum_{X_{1l_1}=0}^{m_{1l_1}} \sum_{X_{21}=0}^{m_{21}} \cdots \sum_{X_{2l_2}=0}^{m_{2l_2}} \Pr\left(X_{11},\cdots,X_{1l_1}\middle|\hat{p}_1\right) \Pr\left(X_{21},\cdots,X_{2l_2}\middle|\hat{p}_2\right)$$
$$\times I(|T_{X_1,X_2}| \geq |T_{x_1,x_2}|)$$

where

$$Pr\left(X_{i1},\cdots,X_{il_i}\middle|p_i\right) = \prod_{j=1}^{l_i}\sum_{Z_{ij}=0}^{n_{ij}} \binom{m_{ij}}{x_{ij}} \left(\frac{z_{ij}}{n_{ij}}\right)^{x_{ij}} \left(1-\frac{z_{ij}}{n_{ij}}\right)^{m_{ij}-x_{ij}} \binom{n_{ij}}{z_{ij}} p_i^{z_{ij}}(1-p_{ij})^{n_{ij}-z_{ij}}$$

The two-side p-value for testing (1) is defined by the probability of observing an absolute value of the statistic T_{X_0,X_1} that is equal to or larger than the absolute value of the observed T_{X_1,X_0} under the null hypothesis.

If the sequencing error rates in pooled sequencing are considered, then

$$Pr(X_{i1},\cdots,X_{il_i}|p_i)=\prod_{j=1}^{l_i}\sum_{Z_{ij}=0}^{n_{ij}}\binom{m_{ij}}{x_{ij}}(\theta_{ij})^{x_{ij}}(1-\theta_{ij})^{m_{ij}-x_{ij}}\binom{n_{ij}}{z_{ij}}p_i^{z_{ij}}(1-p_i)^{n_{ij}-z_{ij}}$$

The parametric bootstrap (PB) procedure can be used to determine the p-value (Krishnamoorthy & Thomoson, 2004) by the following steps:

1. Estimating the sequencing error rate from the control sequence. $\hat{\tau}_{1,ij}=m_{v|\gamma}/(m_{v|\gamma}+m_{nv|\gamma})$, in which $m_{v|\gamma}$ and $m_{nv|\gamma}$ are the number of reads that report a variant or a non-variant for a reference base in the control sequence.
2. Estimate the allele frequency under null hypotheses by a weighted-average estimate across single pools with weight proportional to $n_{ij}m_{ij}/(n_{ij}+m_{ij})$ and the allele frequency of a single pool is estimated by $\hat{p}_{ij}=\sum_{z_{ij}=0}^{n_{ij}}\pi_{ij}z_{ij}/n_{ij}$ where

$$\pi_{ij}=\frac{\binom{m_{ij}}{x_{ij}}\left[\frac{z_{ij}}{n_{ij}}\hat{\tau}_{2,ij}+\frac{n_{ij}-z_{ij}}{n_{ij}}\hat{\tau}_{1,ij}\right]^{x_{ij}}\left[\frac{z_{ij}}{n_{ij}}(1-\hat{\tau}_{2,ij})+\frac{n_{ij}-z_{ij}}{n_{ij}}(1-\hat{\tau}_{1,ij})\right]^{m_{ij}-x_{ij}}}{\sum_{z_{ij}=0}^{n_{ij}}\binom{m_{ij}}{x_{ij}}\left[\frac{z_{ij}}{n_{ij}}\hat{\tau}_{2,ij}+\frac{n_{ij}-z_{ij}}{n_{ij}}\hat{\tau}_{1,ij}\right]^{x_{ij}}\left[\frac{z_{ij}}{n_{ij}}(1-\hat{\tau}_{2,ij})+\frac{n_{ij}-z_{ij}}{n_{ij}}(1-\hat{\tau}_{1,ij})\right]^{m_{ij}-x_{ij}}}$$

3. Calculating the test statistic

$$T_{X_1,X_0}=\frac{\sum_j\frac{\frac{w_{1j}}{\sum w_{1j}}X_{ij}}{m_{ij}}-\sum_j\frac{\frac{w_{2j}}{\sum w_{2j}}X_{2j}}{m_{2j}}}{\sqrt{\hat{V}_X}}$$

in which

$$w_{ij}=\frac{n_{ij}\,m_{ij}}{(\hat{\tau}_{2,ij}-\hat{\tau}_{1,ij})n_{ij}\,\hat{p}_0+\hat{\tau}_{1,ij}n_{ij}+(\hat{\tau}_{2,ij}-\hat{\tau}_{1,ij})^2m_{ij}\hat{p}_0}$$

and

$$\hat{V}_x=\sum_i\frac{1}{\sum(m_{ij}n_{ij})/[m_{ij}\hat{p}_0(\tau_{2,ij}-\tau_{1,ij})^2+n_{ij}(\tau_{2,ij}\hat{p}_0-\tau_{1,ij}\hat{p}_0+\tau_{1,ij})]}$$

4. Sampling $\tilde{X}_i=(\tilde{x}_{i1},\cdots)$ from $\Pr(X_{i1},\cdots,X_{il_i}|\hat{p}_0,\hat{\tau}_{i,ij},\hat{\tau}_{2,ij})$ and calculating the statistic $T_{\tilde{X}_1,\tilde{X}_2}$.
5. Replicating (4) r times and estimating the p-value by the proportion of $|T_{\tilde{X}_1,\tilde{X}_2}|\geq|T_{x_1,x_2}|$

4. Summary

Pooled DNA has been widely used as a cost-effective strategy for genome wise association studies, which have successfully identified hundreds of variants that are associated with complex traits. Pooling scheme may provide less information comparing to the sequencing

for each individual. However, though well-chosen pool designs as introduced in Section 2, there is still high chance to identify the variant carrier. Recently, next-generation sequencing technologies have made it feasible to sequence several human genomes entirely. SNPSeeker and CRISP are efficient statistical methods to detect the variant from the short read generated by NGS platforms. For case-control analysis, Fisher's exact test is common used but has been proved to be inappropriate. Several test methods such as Combined Z test, Likelihood ratio test, and differential test can be conducted for the association analysis.

5. References

Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, Dowzell K, Cichon S, Hillmer AM, O'Donovan MC, Williams J, Owen MJ, Kirov G. (2008). A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. BMC Med Genomics. 29, 1-44.

Arnheim N, Strange C, Erlich H. (1985). Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. 82, 6970-6974.

Amos CI, Frazier ML, Wang W. (2000). DNA pooling in mutation detection with reference to sequence analysis. Am J Hum Genet. 66, 1689-1692.

Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A. (2002). Association testing by DNA pooling: an effective initial screen. Proc Natl Acad Sci USA. 99, 16871-16874.

Bansal V. (2010a). A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics. 26, 318-324.

Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA. (2010b). Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res. 20, 537-545.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 40, 955-962.

Barry ES, Choongrak k. (1900). Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions . Journal of the American Statistical Association. 85, 146-155.

Bodmer W, Bonilla C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 40, 695-701.

Boss Y, Bacot F, Montpetit A, Rung J, Qu HQ, Engert JC, Polychronakos C, Hudson TJ, Froguel P, Sladek R, Desrosiers M. (2009). Identification of susceptibility genes for

complex diseases using pooling-based genome-wide association scans. Hum Genet. 125, 305-318.

Carmi R, Rokhlina T, Kwitek-Black AE, Elbedour K, Nishimura D, Stone EM, Sheffield VC. (1995). Use of a DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15. Hum Mol Genet. 4, 9-13.

Chaisson MJ, Brinza D, Pevzner PA. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 19, 336-346.

Chi A, Schymick JC, Restagno G, Scholz SW, Lombardo F, Lai SL, Mora G, Fung HC, Britton A, Arepalli S, Gibbs JR, Nalls M, Berger S, Kwee LC, Oddone EZ, Ding J, Crews C, Rafferty I, Washecka N, Hernandez D, Ferrucci L, Bandinelli S, Guralnik J, Macciardi F, Torri F, Lupoli S, Chanock SJ, Thomas G, Hunter DJ, Gieger C, Wichmann HE, Calvo A, Mutani R, Battistini S, Giannini F, Caponnetto C, Mancardi GL, La Bella V, Valentino F, Monsurr MR, Tedeschi G, Marinou K, Sabatelli M, Conte A, Mandrioli J, Sola P, Salvi F, Bartolomei I, Siciliano G, Carlesi C, Orrell RW, Talbot K, Simmons Z, Connor J, Pioro EP, Dunkley T, Stephan DA, Kasperaviciute D, Fisher EM, Jabonka S, Sendtner M, Beck M, Bruijn L, Rothstein J, Schmidt S, Singleton A, Hardy J, Traynor BJ. (2009). A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. Hum Mol Genet. 18, 1524-1532.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ. (2008). Identification of genetic variants using barcoded multiplexed sequencing. Nat Methods. 5, 887-893.

Druley TE, Vallania FL, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS, Fay JC, Mitra RD. (2009). Quantification of rare allelic variants from pooled genomic DNA. Nat Methods. 6, 263-265.

Erlich Y, Chang K, Gordon A, Ronen R, Navon O, Rooks M, Hannon GJ. (2009). DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 19, 1243-1253.

Hindorff LA, J. H. (2009). A Catalog of Published Genome-Wide Association Studies. National Human Genome Research Institute.

Ito T, Chiku S, Inoue E, Tomita M, Morisaki T, Morisaki H, Kamatani N. (2003). Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. Am J Hum Genet. 72, 384-398.

Iyengar SK, Song D, Klein BE, Klein R, Schick JH, Humphrey J, Millard C, Liptak R, Russo K, Jun G, Lee KE, Fijal B, Elston RC. (2004). Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration. Am J Hum Genet. 74, 20-39.

Jin F, Hazbun T, Michaud GA, Salcius M, Predki PF, Fields S, Huang J. (2006). A pooling-deconvolution strategy for biological network elucidation. Nat Methods. 3, 183-189.

Johnson PL, Slatkin M. (2008). Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol. 25, 199-206.

Kim SY, Li Y, Guo Y, Li R, Holmkvist J, Hansen T, Pedersen O, Wang J, Nielsen R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. Genet Epidemiol. 34, 479-491.

Kirov G, Zaharieva I, Georgieva L, Moskvina V, Nikolov I, Cichon S, Hillmer A, Toncheva D, Owen MJ, O'Donovan MC. (2009). A genome-wide association study in 574 schizophrenia trios using DNA pooling. Mol Psychiatry. 14, 796-803.

Krishnamoorthy K, Thomson J. (2004). A more powerful test for comparing two Poisson means. J Stat Plan Inference. 119, 23-35.

Liddell, D. (1976). Practical Tests of 2 × 2 Contingency Tables . Journal of the Royal Statistical Society. 25, 295-304.

Lynch M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics. 182, 295-301.

Macgregor S. (2007). Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. Eur J Hum Genet. 15, 501-504.

Maher, B. (2008). Peasonal genomes: the case of the missing heritability. 456, 18-21.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. (2009). Finding the missing heritability of complex diseases. Nature. 461, 747-753.

Metzker ML. (2010). Sequencing technologies - the next generation. Nature reviews. 11, 31-46.

Michelmore RW, Paran I, Kesseli RV. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A. 88, 9828-9832.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 324, 387-389.

Norton N, Williams NM, O'Donovan MC, Owen MJ. (2004). DNA pooling as a tool for large-scale association studies in complex traits. Ann Med. 36, 146–152.

Nystuen A, Benke PJ, Merren J, Stone EM, Sheffield VC. (1996). A cerebellar ataxia locus identified by DNA pooling to search for linkage disequilibrium in an isolated population from the Cayman Islands. Hum Mol Genet. 5, 525-531.

Prabhu S, Pe'er I. (2009). Overlapping pools for high-throughput targeted resequencing. Genome Res. 19, 1254-1261.

Scott DA, Carmi R, Elbedour K, Yosefsberg S, Stone EM, Sheffield VC. (1996). An autosomal recessive non-syndromic- hearing-loss locus identified by DNA pooling using two inbred Bedouin kindreds. Am J Hum Genet. 59, 385-391.

Sham P, Bader JS, Craig I, O'Donovan M, Owen M. (2002). DNA Pooling: a tool for large-scale association studies. Nat Rev Genet. 3, 862-871.

Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. (1998). Allele frequency distributions in pooled DNA samples: applica- tions to mapping complex disease genes. . Genome Res. 8, 111-123.

Sheffield VC, Carmi R, Kwitek-Black A, Rokhlina T, Nishimura D, Duyk GM, Elbedour K, Sunden SL, Stone EM. (1994). Identification of a Bardet–Biedl syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping. Hum. Mol. Genet. . Hum. Mol. Genet. 3, 1331-1335.

Shendure J, Ji H. (2008). Next-generation DNA sequencing. Nat Biotechnol. 26, 1135-1145.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 38, 209–213.

Thierry-Mieg N. (2006). A new pooling strategy for high-throughput screening: the Shifted Transversal Design. BMC Bioinformatics. 19, 7-28.

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, Brinton LA, Sigurdson A, Doody M, Bhatti P, Alexander BH, Buring J, Lee IM, Vatten LJ, Hveem K, Kumle M, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Chanock SJ, Hunter DJ. (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet. 41, 579-584.

Thomas MC, Joy AT. (1991). Elements of Information Theory. Wily Interscience.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2 x 2 comparative trial. J. R. Statist. Soc. A, 145, 86-105.

Wang H, Thomas DC, Pe'er I, Stram DO. (2006). Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol. 30, 356-368.

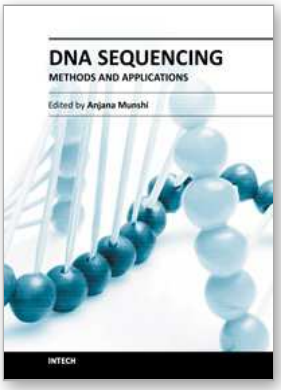
Wang T, Lin CY, Rohan TE, Ye K. (2010). Resequencing of pooled DNA for detecting disease associations with rare variants. Genet Epidemiol. 34, 492-501.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447, 661-678.

Zeng D, Lin DY. (2005). Estimating haplotype-disease associations with pooled genotype data. . Genet Epidemiol , 28, 70-82.

Zuo Y, Zou G, Zhao H. (2006). Two-stage designs in case-control association analysis. Genetics. 173, 1747-1760.

Zuo Y, Zou G, Wang J, Zhao H, Liang H. (2008). Optimal two-stage design for case-control association analysis incorporating genotyp- ing errors. Ann Hum Genet. 72, 375-387.



DNA Sequencing - Methods and Applications

Edited by Dr. Anjana Munshi

ISBN 978-953-51-0564-0
Hard cover, 174 pages
Publisher InTech
Published online 20, April, 2012
Published in print edition April, 2012

This book illustrates methods of DNA sequencing and its application in plant, animal and medical sciences. It has two distinct sections. The one includes 2 chapters devoted to the DNA sequencing methods and the second includes 6 chapters focusing on various applications of this technology. The content of the articles presented in the book is guided by the knowledge and experience of the contributing authors. This book is intended to serve as an important resource and review to the researchers in the field of DNA sequencing.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821