DNA POOLING: A TOOL FOR LARGE-SCALE ASSOCIATION STUDIES

Pak Sham, Joel S. Bader‡, Ian Craig*, Michael O'Donovan§ and Michael Owen§*

DNA pooling is a practical way to reduce the cost of large-scale association studies to identify susceptibility loci for common diseases. Pooling allows allele frequencies in groups of individuals to be measured using far fewer PCR reactions and genotyping assays than are used when genotyping individuals. Here, we discuss recent developments in quantitative genotyping assays and in the design and analysis of pooling studies. Sophisticated pooling designs are being developed that can take account of hidden population stratification, confounders and inter-loci interactions, and that allow the analysis of haplotypes.

STUTTER BANDS
The signals that indicate the presence of DNA fragments that are one or two repeats shorter than the true allele, owing to a 'slippage' artefact that arises from the PCR reaction.

FLUORIMETRY
An assay for measuring DNA concentration in which a fluorescent dye is used that intercalates quantitatively between stacked DNA base pairs.

**P080, Institute of Psychiatry, King's College, Denmark Hill, London SE5 8AF, UK.*
‡Department of Bioinformatics, CuraGen Corporation, 555 Long Wharf Drive, New Haven, Connecticut 06511, USA.
§Department of Psychological Medicine, University of Wales College of Medicine, Heath Park, Cardiff CF14 4XN, UK.
Correspondence to P.S.
e-mail: p.sham@iop.kcl.ac.uk
doi:10.1038/nrg930

The systematic association analysis of complex disorders requires genotyping on a massive scale, to accommodate the numerous genetic markers that are required to screen genomic regions or entire genomes and to assay sufficiently large samples to achieve replicable findings[1–3]. These requirements are now motivating intensive efforts to develop efficient, high-throughput genotyping technologies[4]. For the potential power of association studies to be realized, such technologies must allow the development of assays that are rapid, robust, automated, accurate and cheap.

Until these technologies are widely available, one way to address the cost, time and labour that are involved in large-scale genotyping is to carry out analyses not on individual DNA samples, but on pools made up of DNA from many individuals. The benefits of pooled analysis are easy to appreciate. In principle, 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. The pooling of samples to increase efficiency has been used previously in non-genetic settings: for screening large populations for cases of syphilis[5], for estimating disease prevalence[6–9] and for assessing exposure to pathogens or toxins in case–control studies[10]. Pooling has also been used to maintain the anonymity of individuals when screening for HIV infection[11,12]. In genetics, pooling was first used in a case–control association study of HLA class II DR and DQ alleles in type I (insulin-dependent) diabetes mellitus[13]. Subsequently, it has been used for linkage studies in plants[14], for the homozygosity mapping of recessive diseases in inbred populations, such as the Bedouin[15–18], and for mutation detection[19].

Several groups have recently investigated methods for measuring the allele frequencies of microsatellite markers[20–24] and single-nucleotide polymorphisms (SNPs)[25–30] from pooled DNA samples. In this review, we discuss the current methodologies of DNA pooling and the issues that underlie experimental design and analysis. Although we recognize the potential value of both microsatellite and SNP markers, this review focuses only on SNP-based methodologies. This is because characterized microsatellites are less abundant in the genome than are SNPs and do not provide adequate genome-wide coverage for systematic screens for association. The estimation of microsatellite allele frequencies from pooled DNA is also complicated by the occurrence of STUTTER BANDS that can vary between markers[21,22].

DNA-pooling methodology

DNA-pool constitution. Several steps are required to construct pools that contain equal quantities of DNA from individual samples and from which robust PCR results can be obtained. In the first step, DNA concentration can be measured by ultraviolet (UV) light spectroscopy. However, this approach alone is not sufficiently accurate for measuring DNA concentration unless the

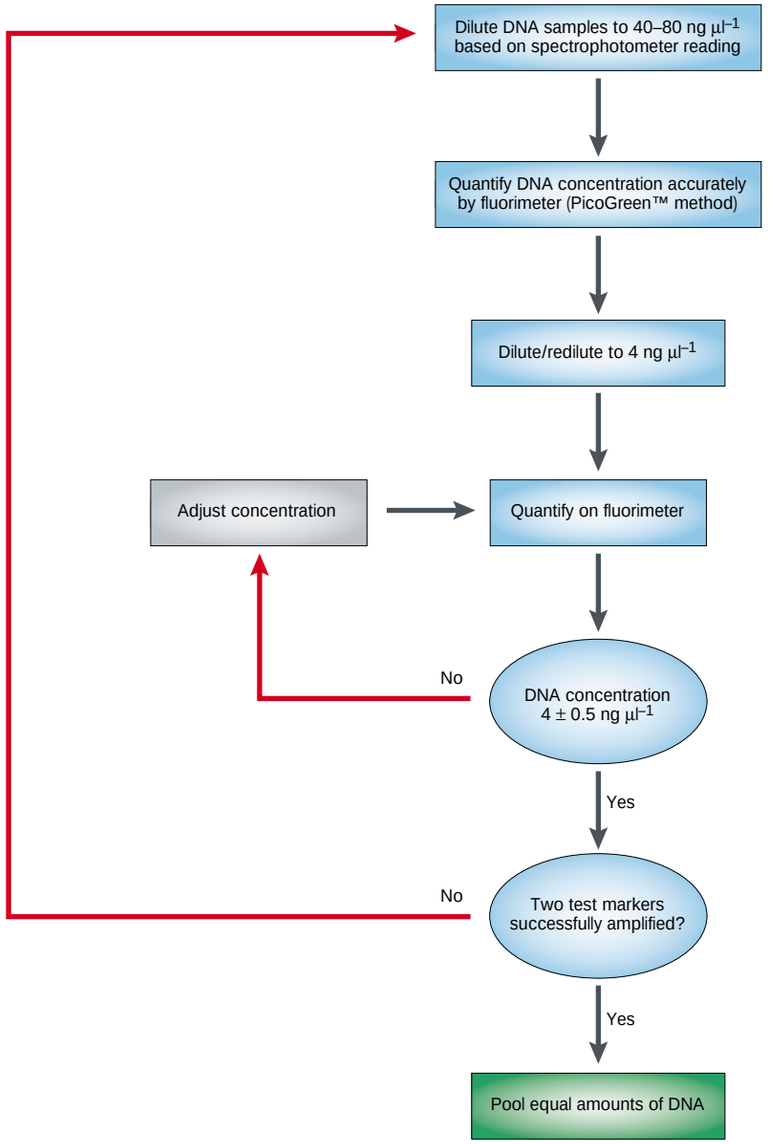


Figure 1 | **A protocol for constructing DNA pools.** The concentration of DNA in the individual samples to be pooled is first estimated by measuring ultraviolet (UV) light absorption at 260 nm (not shown). Samples are then diluted to 40–80 ng μl^{-1} and their DNA concentrations re-estimated by using fluorimetry. Samples are finally diluted to 4 ng μl^{-1}, re-quantified and adjusted if necessary. After testing each sample by PCR, equal amounts of each sample are pooled.

DNA samples are all of a high purity, as contaminants can affect UV light absorbance. A quantitation step that is based on FLUORIMETRY with a DNA-specific dye (such as PicoGreen™) is therefore recommended in some protocols[21,26,28,31]. Inaccuracies can also arise from the pipetting of small volumes of viscous solutions in which DNA concentration is not homogeneous. This can be avoided by using non-viscous, dilute stock samples to construct the pools[30]. Even samples that seem to be of the same concentration can vary in their ability to be amplified by PCR, so samples need to be checked by PCR to identify those that do not yield a robust product. FIG. 1 illustrates the steps that are involved in a typical protocol for establishing the validity of individual sample input into pools. In principle, a more sophisticated approach would be to estimate the effective template concentration capable of amplification in each sample by using QUANTITATIVE REAL-TIME PCR. However, published data indicate that accurate results are routinely achievable using the simpler approach described above[24,25,30,32]. Once equi-molar amounts of each sample have been combined, the pools need to be assessed empirically before their use in large-scale studies, by comparing allele frequencies for several polymorphisms between pooled and individual samples.

Quantitative genotyping assays. SNPs confer a base-compositional difference at a polymorphic site that can be detected in an amplified PCR fragment. Various strategies have been developed to genotype SNPs, each of which has varying potential for use in the analysis of allele frequencies in DNA pools (TABLE 1). Here, we briefly consider the basis for those methods that have been applied to pooling studies and direct readers to REF. 4 for a more comprehensive discussion of SNP genotyping methodologies.

In the first general approach, a SNP can be exploited to create PCR fragments of differing size. In the simplest assay of this type, the PCR product is digested with a restriction enzyme endonuclease that cleaves a fragment from one bi-allelic SNP but not the other. Alternatively, modified nucleotides can be included in the PCR reaction that become incorporated into the PCR product in allele-specific patterns and that generate sites that are sensitive or resistant to chemical cleavage. Both methods generate PCR products of differing size that represent specific SNP alleles, which can be detected by conventional electrophoresis on gels or in capillary systems[28,33,34].

In the second approach, primers close to, or abutting, the variable SNP site are used in a primer-extension reaction. During extension, specific di-deoxyribonucleotides are incorporated that terminate the reaction in a sequence-specific manner, which results in allele-specific extension products[35–39] (BOX 1). The allele-specific extension products can be distinguished by numerous methods — for example, by conventional fluorescent tagging and electrophoretic separation. Alternatively, alleles can be detected by pyrosequencing, in which the extension is coupled to a base-specific light-emission reaction[40–42] (BOX 1). Much higher throughput can be obtained by applying extension methods to highly automated and/or highly parallel platforms. The former is best represented by mass spectroscopy. Because each nucleotide has a different mass, each allele-specific extension product can be distinguished by its mass, which can be measured in a very rapid and automated manner by matrix-assisted laser desorption ionization–time-of-flight mass spectrometry (MALDI–TOF)[43]. A parallel approach can be achieved using microarrays. In this approach, reactions that are analogous to the fluorescence method shown in BOX 1b can be carried out simultaneously for several loci[44]. To distinguish between loci, each extension primer has a unique identifier tag at its 5′-end, which causes it to bind to pre-defined complementary oligonucleotide sites

QUANTITATIVE REAL-TIME PCR
A procedure in which the PCR reaction is tracked as it progresses, by monitoring the accumulating signal that is provided by a fluorescent dye released during each PCR cycle.

Table 1 | Single nucleotide polymorphism detection technologies and their application to pooled DNA samples

Generic approach	SNP-detection method	Comment*	References
Amplification and cleavage at SNP site	Restriction enzyme digest	Technically difficult to achieve reliable cleavage	28,33
	Incorporation of UTP, followed by glycosylation and alkaline cleavage	Allele frequencies estimated to within 1–2%	32
Primer extension	Primer extension using chain termination and fluorescent tagging	Mean error of 1% in estimating differences in allele frequencies between pools	30
	Primer extension using specific base incorporation, coupled to pyrophosphate production and light emission	Allele frequencies estimated to within 5%	40–42
	Primer extension using chimeric primers that contain locus-specific, unique identifier tags; PCR products are then hybridized to oligonucleotide tag arrays	Variable allele frequency estimates, generally to within 5%	44
	Primer extension, followed by denaturing high-performance liquid chromatography	Mean error of 1% in estimating differences in allele frequencies between pools	25
	MALDI–TOF	Allele-frequency estimates deviate from real by ~3%	43
Amplification with allele-specific primers	Differentially fluorescent, tagged primers	Allele signal ratios difficult to quantify accurately; however, a procedure that used a single tagged primer estimated allele frequencies to within 1%	24,47
	Pyrophosphate-coupled bioluminometric assay	Sensitive and accurate to within 1%	80
	Real-time (kinetic) PCR coupled to quantitation of product by binding of SYBR Green 1 or TaqMan™	Allele frequencies measured to within 1–5%	26
Detection of conformational changes	Quantitative, single-strand conformation polymorphism analysis	Allele frequencies measured to within 1%	29
Hybridization of PCR products to microarrays	Fluorescent signal quantitation on Affymetrix HuSNP arrays	Used to detect reproducible differences between case and control groups; allele frequencies were not estimated	49

*The accuracy with which allele frequencies can be estimated depends on the absolute value of allele frequency. For most quantitative trait loci surveys, marker alleles with a frequency of 10–90% are considered to be the most appropriate for screening, and the estimation errors quoted are relevant to this range. MALDI–TOF, matrix-assisted laser desorption ionization–time-of-flight mass; SNP, single-nucleotide polymorphism; SYBR Green 1, a fluorescent dye that stoichiometrically binds to DNA; UTP, uridine triphosphate.

on a microarray. Because these arrays can carry probes that contain many thousands of such sites, their use allows a highly parallel approach[44].

In a third generic strategy, allele-specific hybridization is used to discriminate between local sequences at the polymorphic site. Several methods are available for this type of analysis. The two most common methods resolve alleles by either allele-specific hybridization of primers in a PCR reaction or allele-specific hybridization to primers anchored to a microarray[45–47]. In these methods, differentially fluorescence-tagged primers are used that terminate in a 3′-base that is specific to one or other of the alleles. This allows the amplification of PCR products that can be distinguished by their fluorescence signal. Improved specificity of hybridization can sometimes be obtained by the incorporation of a penultimate mismatching 3′-base in the allele-specific primers[48].

For any method of SNP detection to be applied to pooling, each step must be quantitative, from PCR through to signal detection. At present, because each SNP detection assay used depends on an initial PCR step, a common problem is the unbiased representation of allelic products that are present in a DNA pool; a problem that we return to in the next section. Here, we consider problems that are associated with the application of the three main approaches listed above in DNA pooling — cleavage, primer extension and hybridization. The main problem that affects cleavage-based methods is ensuring that the cleavage reactions are complete. Any tendency towards partial cleavage results in a systematic overestimation of the allele that corresponds to the uncut PCR product. Partial digestion is a common problem that is associated with restriction endonucleases, which makes restriction fragment length polymorphism (RFLP)-based methods poor candidates for accurate quantification, a prediction that is borne out by the very limited empirical data that are available[28]. The second method, primer extension, is the SNP genotyping method that has been used most commonly in pooling studies, and, almost without exception, the results obtained have been good. However, there are two main potential problems that are associated with its use. The first is that each base is not incorporated into the extension reaction with equal efficiency[49]. However, this can be easily allowed for when correcting for differential PCR (as discussed below). The second is that partial self-complementarity at the 3′-end of primers might result in self-annealing, which allows extension to occur independently of the target template. However, this problem can be identified readily by carrying out a control extension reaction in the absence of a template. Finally, although there are several methods

MALDI–TOF
A mass spectrometry method in which laser-vaporized PCR fragments are accelerated through a vacuum using an electric field, eventually having an impact on a detector. The time taken for the fragments to travel the distance from the plate to the detector is measured and depends on the charge-to-mass ratio of each molecule, so providing a way to distinguish between allele-specific products.

TAQMAN™
A proprietary system that allows the progression of a PCR reaction to be monitored in real time.

based on hybridization principles, they generally rely on hybridization in solid phase (that is, hybridization to DNA that is anchored to filters, chips, glass slides, beads or any other solid structure) or in solution (liquid phase). A potential drawback that is likely to influence the accuracy of the massively parallel solid-phase methods, where high throughput is achieved by using both probes and targets of high complexity, is that it is difficult, perhaps impossible, to achieve 100% specificity of hybridization. This will result in some degree of background signal and cross-hybridization between probes for each allele. Although the highly parallel approach allows for multiple replicates and numerous internal controls, which minimize the effects of this complication, this approach does not allow differences in quantity within a few per cent to be distinguished at present. Higher hybridization specificity can usually be achieved in liquid phase (by using allele-specific PCR, or exonuclease assays, such as TAQMAN™), but as each allele is distinguished by oligonucleotides of slightly different sequence, there is the potential for alleles to be differentially amplified[26,50–52]. As the PCR phase is exponential, this might distort relative allele measurement to a greater extent. Fortunately, this can be overcome by using real-time detection methods that monitor amplification efficiency, as well as signal intensity, as the PCR progresses. This, coupled with corrections that are based on tests with heterozygotes (see below), allows an accurate, if expensive, assessment of allele frequencies to be made.

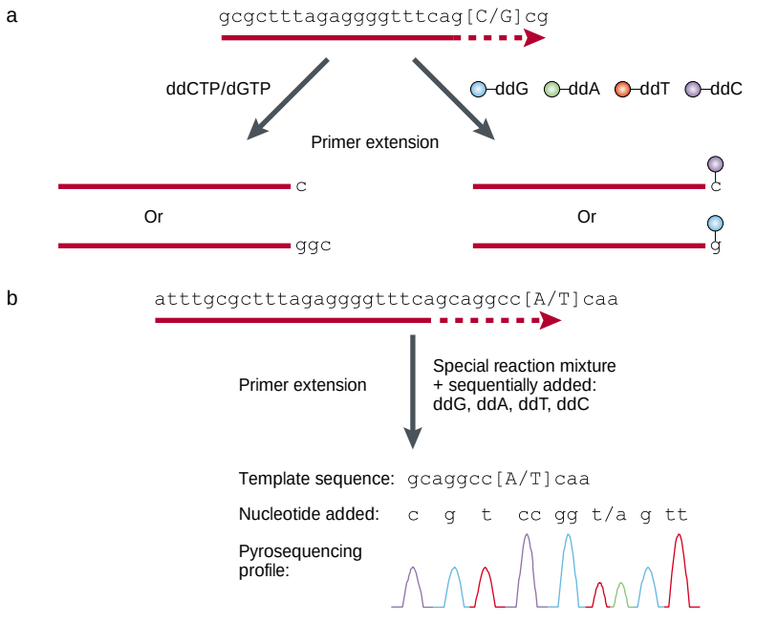
The key to accuracy in a detection method is that detection is quantitative. So, the signal strengths from both the major and the minor alleles must be in the linear range of the detector (which is easily established by generating signal–response curves), and the signal strength from the minor allele must be sufficiently above background for its frequency to be estimated reliably. Detection by autoradiography, fluorescence, photon counting, UV light absorbance and mass spectroscopy have all been used and seem to meet these criteria, and therefore the choice of detection method depends primarily on the specific method of allele discrimination and on the throughput required.

Differential amplification. As noted above, many SNPs show differential amplification during PCR, which means that one allele is more efficiently amplified than the other. Regardless of the detection method, this causes the signal that represents the more efficiently amplified allele to be higher than expected from its true frequency in a pooled sample. Other factors that can distort allele frequency estimates include allele-specific differences in the efficiency of the detection assay (such as differential efficiency of nucleotide incorporation in primer extension, differential efficiency of hybridization in hybridization-based assays) and differential detection efficiency of the allele-specific products by the detector. Examples of the latter include differences in the emission energies of different fluorescent dyes, and differences in the UV light absorbance or in the mass of extension products that are of equi-molar concentration but are of different mass and/or size. To obtain unbiased estimates of allele frequencies, the strength of allele-specific signals should be corrected by a factor that is obtained from reference samples of known allele frequencies[25]. Heterozygous individuals provide convenient reference samples because they have an equal number of copies of the two alleles. The ratio of signal strength of the allele (A) that is less well amplified to

Box 1 | Primer extension PCR

In panel **a**, a C/G polymorphism is detected by primer extension using a 20-nucleotide (nt) primer that is extended in the presence of modified nucleotides. In the first example, the primer is extended in the presence of di-deoxyribonucleotide (dd)CTP (which terminates extension) and dGTP (which allows further extension). In the presence of the G allele, the complementary ddC is added to the primer and extension terminates to produce a 21-nt product. In the presence of the C allele, a complementary dGTP is incorporated, and extension continues until ddC is incorporated, resulting in a 23-nt product. The products can be resolved by size fractionation (for example, by electrophoresis or high-phase liquid chromatography) or, for higher throughput, by mass spectroscopy. In the second example in panel **a**, the primer is extended in the presence of all four ddNTPs, each of which is tagged by a different fluorescent dye. After extension, alleles can be detected by measuring the wavelength of emitted light after laser excitation.

In pyrosequencing, shown in panel **b**, a special reaction mixture is used, which contains DNA polymerase, ATP sulphurylase, luciferase and apyrase, and which generates ATP when a nucleotide that is specific to the single-nucleotide polymorphism (SNP) allele is added to the primer. This ATP drives the luciferase-mediated conversion of luciferin to oxyluciferin, resulting in the emission of light, the amount of which is proportional to the amount of ATP released. This is detected by a charge-coupled device (CCD) camera and analysed, with the height of each peak being proportional to the number of nucleotides incorporated. dNTPs are added sequentially, and the signal obtained represents a quantitative summation of the ATP that is generated. A run of similar nucleotides in the template results in a larger signal, which is proportional to the number of nucleotides in the string. An individual that is heterozygous for a SNP will give two 'half peaks' at the SNP. Once the incorporation of a particular nucleotide stops, the next dNTP is added, until the dNTP that is required to kick off the reaction again is provided. This means that the peak heights at the SNP are constrained by the local sequence; however, software has been designed to take this into consideration (see the Online link to Pyrosequencing).



QUANTITATIVE TRAIT
A measurable trait that depends on the cumulative action of many genes and that can vary among individuals over a given range to produce a continuous distribution of phenotypes. Common examples include height, weight and blood pressure.

PEARSON X^2-TEST
A statistical test that is used to assess whether the frequencies of individuals in different categories of one or more qualitative variables are consistent with those frequencies that are predicted under a certain hypothesis.

that of the allele (B) that is better amplified in a heterozygous individual is defined as k. Given the signal strengths H_A and H_B of a pooled sample, the corrected allele frequencies (f) for alleles A and B are $f_A = H_A/(H_A + kH_B)$ and $f_B = 1 - f_A$. In practice, there is some minor variation in k among repeated measurements of the same heterozygous individual and greater variation among different heterozygous individuals[53]. Some groups routinely estimate k from a panel of 16 (REF. 53) or 32 (REF. 54) individuals, to include some heterozygous individuals even for rare SNPs. Failure to correct for different amplification can result in biased tests of allelic association[53].

Gene duplication can also manifest itself as apparently differential amplification. Ideally, only markers with Mendelian transmission should be corrected using the above method. However, if markers with unknown properties are used, a major deviation from a ratio of 1:1 in the estimated allele frequencies from a heterozygote should alert researchers to the possibility that the target sequence is duplicated in the genome. Unknown polymorphisms in the sequence that is amplified by PCR, or in the primers, are also sources of error that cannot be allowed for by the above calculation. This problem can, however, be identified by the presence of major deviations in estimates of allele ratios between heterozygotes.

Box 2 | Efficiencies of different pooling designs

The efficiency of a study that involves pooled DNA can be quantified as the ratio of the expected χ^2-values for a true quantitative trait locus in a pooled experiment to an analogous test based on individual genotyping[78]. Below are the approximate efficiencies of four designs: first, a case–control design that compares n cases in one pool with n controls in a second pool; second, a parent–offspring trio design that compares n cases in one pool with their $2n$ parents in a second pool; third, an extreme discordant design, in which a pool consisting of high-scoring individuals is compared with a pool consisting of low-scoring individuals, in a random population sample of size N; and finally, an extreme discordant sib-pair design, in which a pool of the high-scoring sibs is compared with a pool of the low-scoring sibs, in pairs that show the largest score differences in a random population sample of $N/2$ sib-pairs.

Design	Efficiency
Case–controls	$\frac{1}{1+\tau^2+\left(2n\varepsilon^2/p(1-p)\right)}$
Parent–offspring trios	$\frac{1}{1+3\tau^2+\left(8n\varepsilon^2/p(1-p)\right)}$
Extreme individuals	$\frac{1}{1+\tau^2}\times\frac{2\phi\left(\Phi^{-1}(1-f)\right)^2}{f+f^2\left(2N\varepsilon^2/p(1-p)(1+2\tau^2)\right)}$
Extreme discordant sib-pairs	$\frac{1}{1+2\tau^2}\times\frac{2\phi\left(\Phi^{-1}(1-f)\right)^2}{f+f^2\left(4N\varepsilon^2/p(1-p)(1+2\tau^2)\right)}$

For case–controls and for parent–offspring trios, efficiency is determined primarily by experimental errors through the parameters τ and ε (REF. 47). For quantitative traits, the formulae for efficiency also involve the standard normal probability density $\phi(z)$, the cumulative normal probability $\Phi(z)$ and its functional inverse $\Phi^{-1}(z)$.
For both of the quantitative trait designs, the pooling fraction f that maximizes efficiency in the absence of experimental error is 0.27 (REFS 47,79). In the presence of experimental error, the optimal pooling fraction depends on a single collective parameter, $2N\varepsilon^2/p(1-p)(1+\tau^2)$ for extreme individuals and $4N\varepsilon^2/p(1-p)(1+2\tau^2)$ for extreme discordant sib-pairs, with an identical functional form. Abbreviating this collective parameter to κ^2 allows a universal calibration curve to be drawn.

DNA pooling studies: design and analysis

Analysis of two-pool designs. The attraction of DNA pooling is that it reduces the amount of genotyping that is required to estimate allele frequencies in a sample. The larger the sample, the greater the saving, so that the design with minimal genotyping would involve comparing just two pools, each containing DNA from numerous individuals. These two pools could be constituted from cases and controls for a disease trait, or from individuals with trait values at the two extremes of a QUANTITATIVE TRAIT[53,55]. The appropriate test for this two-pool design would be to consider the magnitude of the difference between the allele frequency estimates of the two pools in relation to its variance. However, the crucial assumption of the standard PEARSON X^2-TEST — that the variance of the difference in allele-frequency estimates is determined entirely by sampling variation — is unrealistic for pooled DNA data. This is because the variance will be inflated by experimental errors that are specific to DNA-pooling studies[53–56]. These experimental errors can potentially lead to an increase in false-positive association findings, unless they are allowed for in the statistical analysis.

For two independent pools, an appropriate test statistic for allelic association is

$$Z^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{V_1 + V_2},$$

where $\hat{p}_1$ and $\hat{p}_2$ are the estimated frequencies of the allele in the two pools, and V_1 and V_2 are its variances, respectively. The variances are determined by sampling variation and random experimental errors, due either to an unequal amount of DNA being contributed by the individuals that make up the pool (pool-formation errors), or to inaccuracies in PCR reactions and in the measurement of the allele frequencies (pool-measurement errors). Taking these sources of variation into account, the variance of the difference in allele frequency is given by

$$V_1 + V_2 = \hat{p}(1-\hat{p})(1+\tau^2)\left(\frac{1}{2n_1} + \frac{1}{2n_2}\right) + 2\varepsilon^2,$$

where τ is the coefficient of variation (that is, standard deviation divided by the mean) of the number of DNA molecules of locus A that is contributed by each individual, and ε^2 is the variance of the pool-measurement error[53,56]. The magnitudes of the two sources of experimental errors can be determined by previous experiments that involve constituting multiple DNA pools from the same individuals, as well as multiple allele-frequency measurements from the same pools[54]. Using multiple pools and multiple measurements, one recent study reported estimates of average variance

components of 1.06×10^{-4} for pool formation, and 2.93×10^{-4} and 5.55×10^{-4} for the PCR and pyrosequencing stages of allele-frequency measurement, for pool sizes ranging from 188 to 739 (REF. 54). So, for pools in this size range, errors that originate from pool formation are less important than errors in pool measurement. Values reported for ε have ranged from 0.02 to 0.04, depending on the marker (REF. 53).

Optimal two-pool designs. For two reasons, the pooling of DNA is expected to result in the loss of information that could have been obtained by individual genotyping. The first reason is that, as discussed above, pooling involves experimental errors that do not apply to individual genotyping. The second is that, for quantitative traits, pooling allows the examination of between-pool differences but not within-pool differences. Pooling studies should be designed to minimize these losses of information or, more formally, to maximize efficiency.

The efficiency of a DNA-pooling study is, by definition, inversely proportional to the sample size that is required to achieve the same significance level and power as a study that is based on individual genotyping. The efficiencies of four pooling designs are shown in BOX 2 (REFS 55,57). For QUALITATIVE TRAITS, within-pool differences are irrelevant, and efficiency becomes 1 in the absence of experimental errors. In reality, experimental error exists but can be reduced by averaging allele-frequency estimates over repeated measurements of the same pool. The standard deviation of an allele-frequency estimate from pooled DNA that results from random experimental errors is typically 0.02–0.04, so that at least four replicate measures are recommended to reduce the standard error to 0.01 (REF. 53).

For studies of quantitative traits, efficiency does not approach 1 even in the absence of experimental errors. This is due to the loss of information from within-pool phenotypic differences. This information loss can be ameliorated in a two-pool design by the optimal choice of selection criteria for the two pools. In the absence of experimental errors, the optimal selection criterion for pooling a random sample of unrelated individuals is to select individuals from the extreme tails of a quantitative trait's distribution, each comprising 27% of the sample[55–57]. Interestingly, the selection fraction of 27% has arisen previously in the optimal estimation of a correlation coefficient from frequency data[58] and in certain selection experiments[59–61]. This optimal pooling fraction is largely independent of marker frequency and of the mode of inheritance of the trait (dominant, additive or recessive). As measurement error increases, the optimal pooling fraction is decreased according to a single parameter that essentially represents the ratio of experimental errors to sampling errors (BOX 2).

A two-stage design, in which markers that show positive association in a pooling study are followed up by confirmatory individual genotyping, might represent the best trade-off between the cost savings of pooling and the full information that is provided by individual genotyping (FIG. 2). A full marker set, potentially comprising 100,000 SNPs for a genome scan, could be tested using pooled assays with a liberal *p*-value (0.01–0.001) to allow adequate power even with the information loss. Markers that show significance in the pooled assay could then be genotyped in individuals in the original population to confirm the association. A similar multi-stage design was adopted in a recent study on cognitive ability[32].

QUALITATIVE TRAIT
Those traits for which there is a sharp distinction between phenotypes — the trait is usually present or not. Often only one or a few genes are involved in the expression of qualitative traits.

More-complex pooling designs

A two-pool design provides a cheap and fast method for screening numerous SNPs for allelic association. However, for large-scale studies that involve many hundreds or thousands of individuals, it might be advantageous to limit the number of individuals that contribute to a pool and to adopt a more-complex design that involves multiple pools.

Multiple measurements, multiple pools. As mentioned above, the power of a pooling study can be improved by taking the average of allele-frequency estimates from multiple measurements or multiple pools. Such a design might involve making up multiple pools from the same individuals (to average out errors in pool constitution), and also carrying out multiple PCR reactions and making multiple allele-frequency measurements from the same pool (to average out measurement errors). In making up multiple pools from a set of individuals, either multiple replicate pools of the entire group can be made up, or the group can be divided into subsets and distinct pools made up of these subsets[54]. These two strategies are equivalent in reducing error, but the latter provides extra opportunities for examining marker–marker associations (see below). A third alternative of assembling multiple pools that are randomly constituted has been used in non-genetic settings[62].

The possible variations of such designs are characterized by: the number of distinct pools (k); the number of individuals in a pool (n); the number of times that a pool of the same individuals is independently constituted (r); and the number of independent allele-frequency measurements that is to be made for each pool (m). The total number of individuals is nk, and the total number of measurements is krm. An average allele-frequency estimate based on these multiple measurements will have the variance

$$V \approx \frac{p(1-p)}{2nk} + \frac{p(1-p)\tau^2}{2nkr} + \frac{\varepsilon^2}{2krm} \quad .$$

Note that the sampling variance is still determined by the total sample size nk, but quantification error variance is reduced by a factor of r (for any total sample size) and measurement error variance is reduced by a factor of krm. The optimal choice of the design parameters n, k, r and m depends on the relative magnitudes of the different sources of error[55]. On the basis of their estimates of the relative contributions of the different sources of experimental errors, one group of investigators has proposed a strategy of forming multiple distinct pools, each consisting of DNA from 50 individuals, as being a reasonable compromise between cost and accuracy[54].

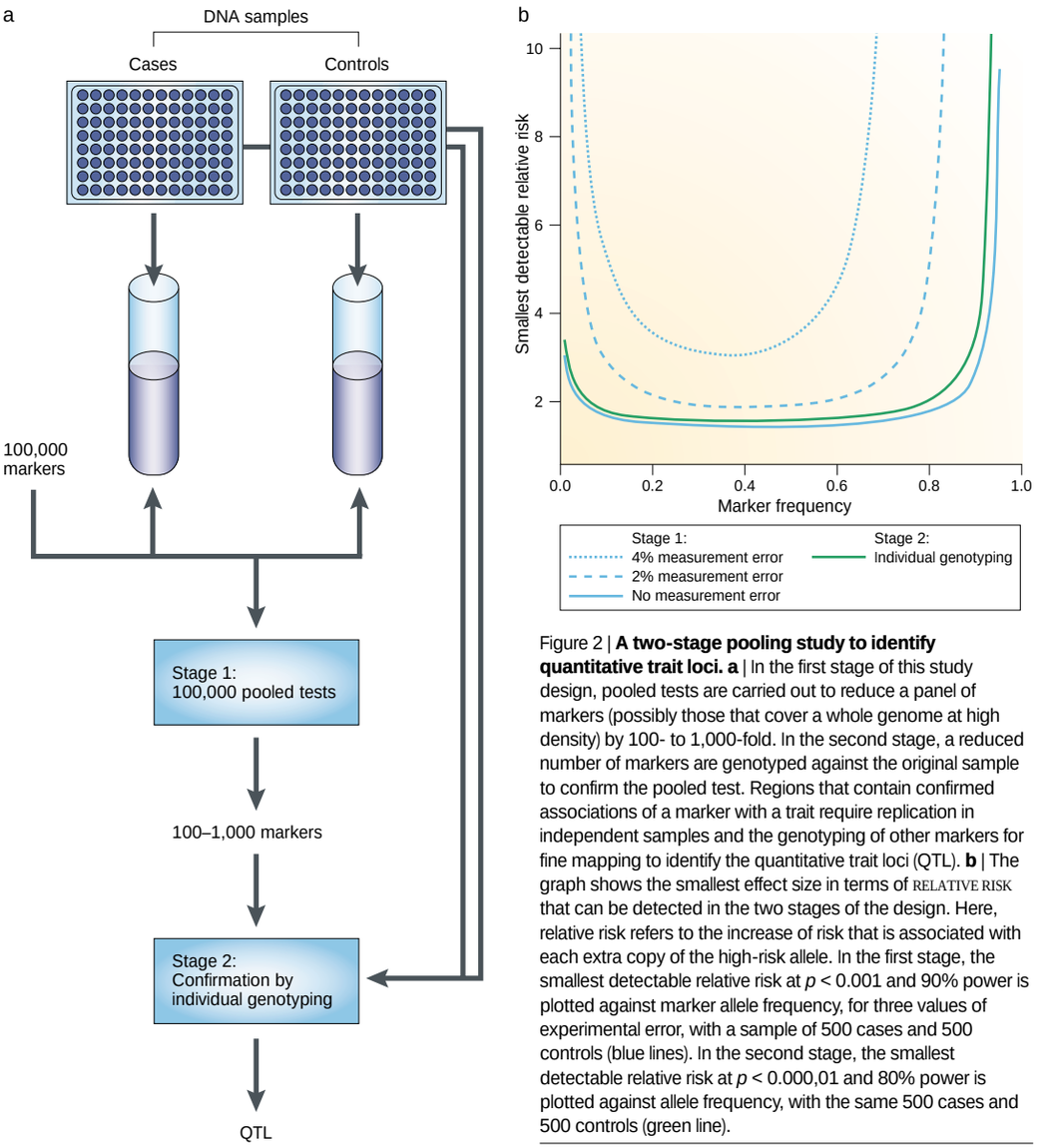


Figure 2 | **A two-stage pooling study to identify quantitative trait loci. a** | In the first stage of this study design, pooled tests are carried out to reduce a panel of markers (possibly those that cover a whole genome at high density) by 100- to 1,000-fold. In the second stage, a reduced number of markers are genotyped against the original sample to confirm the pooled test. Regions that contain confirmed associations of a marker with a trait require replication in independent samples and the genotyping of other markers for fine mapping to identify the quantitative trait loci (QTL). **b** | The graph shows the smallest effect size in terms of RELATIVE RISK that can be detected in the two stages of the design. Here, relative risk refers to the increase of risk that is associated with each extra copy of the high-risk allele. In the first stage, the smallest detectable relative risk at $p < 0.001$ and 90% power is plotted against marker allele frequency, for three values of experimental error, with a sample of 500 cases and 500 controls (blue lines). In the second stage, the smallest detectable relative risk at $p < 0.000,01$ and 80% power is plotted against allele frequency, with the same 500 cases and 500 controls (green line).

Population stratification and family studies. DNA-pooling studies, like other association study designs, are subject to false-positive and false-negative findings as a result of hidden POPULATION STRATIFICATION[63]. Two main methods have been proposed to eliminate this problem: the use of background markers and the use of related controls. In the first approach, numerous unlinked markers are used to detect and adjust for hidden population stratification. In the 'genomic control' variants of this method, the markers are used to estimate the average factor by which the association test statistics are inflated. This factor is then used to adjust association test statistics[63,64]. In the second, more-sophisticated, 'structured association' approach, the marker data are used to model the population substructure; the resulting model is then used to adjust the test statistics[65–67]. The genomic control method can be applied to pooled DNA studies because it involves simply considering the case–control differences across numerous unlinked markers. The structured association method cannot be applied to a study that compares two DNA pools, although its use is theoretically possible in a study that compares two sets of multiple distinct pools.

Another popular safeguard against hidden population stratification is to examine parent–offspring or sibship data for intra-familial association[68,69]. In the parent–offspring design, pooled DNA from affected offspring can be compared with pooled DNA from their parents[70]. Similarly, in a discordant sib-pair design, pooled DNA from the affected sib can be compared with pooled DNA from the unaffected sib. In family designs that use pooled DNA, robustness against population stratification requires that the ratio of individuals from each family that contributes to the two pools is constant across families[71]. So, in samples that consist of an affected offspring

RELATIVE RISK
The ratio of the risk of developing a disease in individuals who have been exposed to a risk factor to that in individuals who have not been exposed to the risk factor.

POPULATION STRATIFICATION
The presence of multiple population subgroups that show limited inter-breeding. When such subgroups differ both in allele frequency and in disease prevalence, this can lead to erroneous results in association studies.

and both parents, it is robust to make up a control pool that contains either both parents from all families, or one parent from all families, but not both parents from some families and only one parent from other families.

The analysis of pooled DNA data from family-based designs is complicated by the fact that the sampling variances of allele-frequency estimates, and of between-pool allele-frequency differences, need to take account of the fact that individuals in the same family are genetically related and therefore do not constitute independent observations. If non-independence is ignored, then sampling variances will be underestimated, resulting in an increase in false-positive associations. The sampling variances of between-pool differences in allele-frequency estimates of several family-based designs — involving sibships with a variable number of affected and unaffected individuals, with or without parents and with or without unrelated controls — have been derived[71]. These results allow the construction of valid statistical tests for allele-frequency differences in family-based pooling studies.

Haplotype analysis. Both theoretical models[72,73] and empirical data[74] indicate that the power of association analysis to detect a causative allele might be increased by the use of multi-locus HAPLOTYPES as compared with single marker alleles. The fact that DNA-pooling studies allow allele frequencies, but not haplotype frequencies, to be estimated directly might, therefore, be considered to be a disadvantage of pooling. However, as SNP maps become more comprehensive and association studies more systematic, it will become increasingly feasible to conduct analyses with SNPs that are not merely markers but might be the actual causative alleles. When the causative alleles are included in the set of markers tested, and when the number of causative alleles is less than the number of haplotypes, single-locus association tests have been shown to be more powerful than multi-locus haplotype analyses[75,76].

Nevertheless, haplotype analysis of pooled DNA data is in fact possible in some situations. Indeed, a method of estimating LINKAGE DISEQUILIBRIUM (LD) from pooled DNA samples has been proposed, but this requires the construction of many pools, each containing DNA from only a few (<10) individuals[77]. Haplotype analysis from pooled DNA data is also possible in regions of strong LD, where only a small proportion of the theoretically possible allele combinations is actually present as haplotypes in the population[54]. This method requires the identification of those haplotypes that actually exist in the population, by genotyping several individuals. The frequency of an allele can be written as the sum of the frequencies of all the existing haplotypes that contain the allele. Each SNP therefore gives rise to a linear equation that relates an allele frequency to the haplotype frequencies; for s SNPs there are s such simultaneous linear equations. If the number of existing haplotypes is h, and if $h \leq s$, then it might be possible to 'solve' certain subsets of the simultaneous linear equations to express the h haplotype frequencies in terms of allele frequencies. In addition to providing haplotype-frequency estimates and, therefore, the possibility of testing haplotype-frequency differences between two groups from pooled data, this method also provides a criterion for the selection of SNPs for pooling studies.

Confounders and gene–environment interactions. In association studies, the standard methods for dealing with potential confounding variables are matching and statistical adjustment. The principle of matching is clearly applicable to association studies using pooled DNA. For example, the individuals who comprise two pools that are to be compared should have equal representations of sexes, age groups and ethnicities. Adequate matching by socio-demographic variables should reduce the risk of spurious association. When a disorder has known risk factors, it might be desirable to design multiple pools that differ in the level of exposure to the risk factors[10]. For example, two pools of cases could be made up, one with a high level and the other with a low level of exposure to the risk factor, and similarly for the controls. Such a design might increase the power to detect an allelic association with disease, and allow the examination of possible interactions between the known factors and putative genetic risk variants.

Conclusion

This review has highlighted the potential increase in efficiency that DNA pooling has to offer for systematic association studies. However, it has also drawn attention to several methodological issues that must be attended to for DNA pooling to work effectively. Clearly, laboratory procedures must be optimized to obtain reliable results with minimal biases and errors. Further reduction of errors will depend on experimental design and statistical analysis. At present, the most effective use of DNA pooling might be in a two-stage design in which markers that show putative association are followed up by individual genotyping. In this way, pooling can be used as an efficient and sensitive method of screening numerous markers to identify a subset for more detailed studies.

More sophisticated pooling designs that involve families or multiple, moderately sized pools can account for population stratification, confounders and interactions, and can provide greater control of pooling-specific experimental errors. However, if pooling is used as a screen to be followed by individual genotyping, it might be argued that some of these safeguards can be built in at the follow-up stage. Similarly, pooled DNA can be used for haplotype analysis provided that haplotype-diversity data on the population have been obtained before the pooling study and that an appropriate set of SNPs has been selected for genotyping. It remains to be established whether this sophisticated approach of marker selection and haplotype analysis, or a simple approach of maximizing the number of SNPs examined so that these have a greater chance of including causative SNPs, is a more cost-effective design for the systematic screening of human genomes for disease association.

HAPLOTYPE
The allelic configuration of two or more alleles on a single chromosome of a given individual.

LINKAGE DISEQUILIBRIUM
This occurs when the frequency of a particular haplotype for two or more loci deviates significantly from that expected from the product of the observed allelic frequencies at each locus.

1. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2002).
2. Cardon, L. R. & Bell, J. I. Association study designs for complex disease. *Nature Rev. Genet.* **2**, 91–99 (2001).
3. Tabor, H. K., Risch, N. J. & Myers, R. M. Candidate-gene approaches for studying complex traits: practical considerations. *Nature Rev. Genet.* **3**, 1–7 (2002).
4. Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet.* **2**, 930–942 (2001).
   This review provides a good introduction to SNP-genotyping methods.
5. Dorfman, R. The detection of defective members of large populations. *Ann. Math. Stat.* **14**, 436–440 (1943).
6. Thompson, K. H. Estimation of the proportion of vectors in a natural population of insects. *Biometrics* **18**, 568–578 (1962).
7. Sobel, M. & Elashoff, R. M. Group testing with a new goal, estimation. *Biometrics* **62**, 181–193 (1975).
8. Tu, X. M., Litvak, E. & Pagano, M. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika* **82**, 287–297 (1995).
9. Brookmeyer, R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics* **55**, 608–612 (1999).
10. Weinberg, C. R. & Umbach, D. M. Using pooled exposure assessment to improve efficiency in case–control studies. *Biometrics* **55**, 718–726 (1999).
11. Gastwirth, J. L. & Hammick, P. A. Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of AIDS antibodies in blood. *J. Stat. Planning Inference* **22**, 15–27 (1989).
12. Gastwirth, J. L. & Johnson, W. Screening with cost effective quality control: potential application to HIV and drug testing. *J. Am. Stat. Assoc.* **89**, 972–981 (1994).
13. Arnheim, N., Strange, C. & Erlich, H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of *HLA* class II loci. *Proc. Natl Acad. Sci. USA* **82**, 6970–6974 (1985).
14. Michelmore, R. W., Paran, I. & Kesseli, R. V. 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. USA* **88**, 9828–9832 (1991).
15. Sheffield, V. C. *et al.* Identification of a Bardet–Biedl syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping. *Hum. Mol. Genet.* **3**, 1331–1335 (1994).
16. Carmi, R. *et al.* Use of a DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15. *Hum. Mol. Genet.* **4**, 9–13 (1995).
    An example of the successful application of pooling.
17. Nystuen, A., Benke, P. J., Merren, J., Stone, E. M. & Sheffield, V. C. 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 (1996).
18. Scott, D. A. *et al.* An autosomal recessive non-syndromic-hearing-loss locus identified by DNA pooling using two inbred Bedouin kindreds. *Am. J. Hum. Genet.* **59**, 385–391 (1996).
19. Amos, C. I., Frazier, M. L. & Wang, W. DNA pooling in mutation detection with reference to sequence analysis. *Am. J. Hum. Genet.* **66**, 1689–1692 (2000).
20. Pacek, P., Sajantila, A. & Syvanen, A. C. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. *PCR Methods Appl.* **2**, 313–317 (1993).
21. Barcellos, L. F. *et al.* Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* **61**, 737–747 (1997).
22. Daniels, J. *et al.* A simple method for analysing microsatellite allele image patterns generated from DNA pools and its applications to allelic association studies. *Am. J. Hum. Genet.* **62**, 1189–1197 (1998).
23. Shaw, S. H., Carrasquillo, M. M., Kashuk, C., Puffenberger, E. G. & Chakravarti, A. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* **8**, 111–123 (1998).
24. Kirov, G., Stephens, M., Williams, N., O'Donovan, M. & Owen, M. Automated genotyping of single-nucleotide polymorphisms by extension of fluorescently labelled primers: analysis of individual and pooled DNA samples. *Balkan J. Med. Genet.* **3**, 23–28 (2000).
25. Hoogendoorn, B. *et al.* Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum. Genet.* **107**, 488–493 (2000).
26. Germer, S., Holland, M. J. & Higuchi, R. High-throughput SNP allele frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.* **10**, 258–266 (2000).
27. Ross, P., Hall, L. & Haff, L. A. Quantitative approach to single-nucleotide polymorphism analysis using MALDI–TOF mass spectrometry. *Biotechniques* **29**, 620–626, 628–629 (2000).
28. Breen, G., Harold, D., Ralston, S., Shaw, D. & St Clair, D. Determining SNP allele frequencies in DNA pools. *Biotechniques* **28**, 464–470 (2000).
29. Sasaki, T. *et al.* Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. *Am. J. Hum. Genet.* **68**, 214–218 (2001).
30. Norton, N. *et al.* Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum. Genet.* **110**, 471–478 (2002).
31. Plomin, R. *et al.* A genome-wide scan of 1847 DNA markers for allelic associations with general cognitive ability: a five-stage design using DNA pooling. *Behav. Genet.* **31**, 497–509 (2002).
    This study illustrates the use of pooling as an efficient screening tool in a multi-stage design.
32. Curran, S. *et al.* Validation of single nucleotide polymorphism (SNP) quantification in pooled DNA samples using SNaPIT™ technology, a glycosylase-mediated polymorphism detection method. *Biotechniques* (in the press).
33. Craig, I. W. & McClay, J. in *Behavioral Genetics in the Post-genomics Era* (eds Plomin, R., DeFries, J., Craig, I. & McGuffin, P) 19–40 (APA Books, Washington, DC, 2002).
    This book reviews genotyping methods for microsatellite and SNP markers, with comments on pooling strategy.
34. Vaughan, P. & McCarthy, T. V. A novel process for mutation detection using uracil DNA-glycosylase. *Nucleic Acids Res.* **26**, 810–815 (1998).
35. Syvanen, A. C., Aalto-Setala, K., Kontula, K. & Soderlund, H. A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8**, 684–692 (1990).
36. Syvanen, A. C. From gels to chips: 'minisequencing' primer extension for analysis of point mutations and single nucleotide polymorphisms. *Hum. Mutat.* **13**, 1–10 (1999).
37. Tully, G., Sullivan, K. M., Nixon, P., Stones, R. E. & Gill, P. Rapid detection of mitochondrial sequence polymorphisms using mutiplex solid phase fluorescent minisequencing. *Genomics* **34**, 107–113 (1996).
38. Pastinen, T. *et al.* A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res.* **10**, 1031–1042 (2000).
39. Braun, A., Little, D. P. & Koster, H. Detecting *CFTR* gene mutations by using primer oligo base extension and mass spectrometry. *Clin. Chem.* **43**, 1151–1158 (1997).
40. Nordfors, I. *et al.* Large-scale genotyping of single nucleotide polymorphisms by pyrosequencing and validation against the 5′ nuclease (TaqMan) assay. *Hum. Mutat.* **19**, 395–401 (2000).
41. Gruber, J. D., Colligan, P. B. & Wolford, J. K. Estimation of single nucleotide polymorphism allele frequency in DNA pools by using pyrosequencing. *Hum. Genet.* **110**, 395–401 (2002).
42. Wasson, J., Skolnick, G., Love-Gregory, L. & Permutt, M. A. Assessing allele frequencies of single nucleotide polymorphisms in DNA pools by pyrosequencing technology. *Biotechniques* **32**, 1144–1152 (2002).
43. Werner, M. *et al*. Large scale determination of SNP allele frequencies in DNA pools using MALDI–TOF mass spectroscopy. *Hum. Mutat.* **20**, 57–64 (2002).
44. Fan, J. B. *et al.* Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.* **10**, 853–860 (2000).
45. Hacia, J. G. *et al.* Strategies for mutation analysis of the large multi-exon *ATM* gene using high-density oligonucleotide arrays. *Genome Res.* **8**, 1245–1258 (1998).
46. Germer, S. & Higuchi, R. Single tube genotyping without oligonucleotide probes. *Genome Res.* **9**, 72–78 (1999).
47. McClay, J., Sugden, K., Koch, H. G., Higuchi, S. & Craig, I. W. High-throughput single-nucleotide polymorphism genotyping by fluorescent competitive allele-specific polymerase chain reaction (SNiPTag). *Anal. Biochem.* **301**, 200–206 (2002).
48. Livak, K. J. Allelic discrimination using fluorogenmeic probes and the 5′ nuclease assay. *Genet. Anal.* **14**, 143–149 (1999).
49. Uhl, G., Liu, Q.-R., Walther, W., Hess, J. & Naiman, D. Polysubstance abuse — vulnerability genes: genome scans for association, using 1,004 subjects and 1,494 single nucleotide polymorphisms. *Am. J. Hum. Genet.* **69**, 1290–1300 (2001).
50. Holland, P. M., Abramson, R. D., Watson, R. & Gelfland, D. H. Detection of specific polymerase chain reaction product by utilizing the 5′ to 3′ exonuclease activity of *Thermus aquaticus* polymerase. *Proc. Natl Acad. Sci. USA* **88**, 7276–7280 (1991).
51. Higuchi, R. G., Dolligenger, P. S., Walsh, P. S. & Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology* **10**, 413–417 (1992).
52. Luedeck, H. & Blascyk, R. Fluorotyping of HLA-C: differential detection on amplicons by sequence-specific priming and fluorogenic probing. *Tissue Antigens* **50**, 627–638 (1997).
53. Le Hellard, S. *et al.* SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res.* (in the press).
    This paper describes the correction of differential amplification and assesses the accuracy of allele-frequency estimation in pooled samples.
54. Barratt, B. J. *et al.* Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.* (in the press).
    This paper considers the sources of errors in the estimation of allele frequency in pooled samples and proposes the use of multiple pools, each containing DNA from a small number of individuals.
55. Bader, J. S., Bansal, A. & Sham, P. C. Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. *GeneScreen* **1**, 143–150 (2001).
    A mathematical description of the optimal pooling study designs for analysing quantitative phenotypes.
56. Jawaid, A., Bader, J. S., Purcell, S., Cherny, S. S. & Sham, P. Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur. J. Hum. Genet.* (in the press).
57. Bader, J. S. & Sham, P. C. Family-based association tests for quantitative traits using pooled DNA. *Eur. J. Hum. Genet.* (in the press).
58. Mosteller, F. On some useful 'inefficient statistics'. *Ann. Math. Stat.* **17**, 377–408 (1946).
59. Hill, W. G. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics* **27**, 293–311 (1971).
60. Kimura, M. & Crow, J. F. Effect of overall phenotypic selection on genetic change at individual loci. *Proc. Natl Acad. Sci. USA* **75**, 6168–6171 (1978).
61. Ollivier, L., Messer, L. A., Rothschild, M. F. & Legault, C. The use of selection experiments for detecting quantitative trait loci. *Genet. Res.* **69**, 227–232 (1997).
62. Hammick, P. A. & Gastwirth, J. L. Group testing for sensitive characteristics: extension to higher prevalence levels. *Int. Stat. Rev.* **62**, 319–331 (1994).
63. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
64. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
65. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
66. Satten, G. A., Flanders, W. D. & Yang, Q. Accounting for unmeasured population substructure in case–control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* **68**, 466–477 (2001).
67. Zhang, S. & Zhao, H. Quantitative similarity-based association tests using population samples. *Am. J. Hum. Genet.* **69**, 601–614 (2001).
68. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
69. Curtis, D. Use of siblings as controls in case–control association studies. *Ann. Hum. Genet.* **61**, 319–333 (1997).
70. Kirov, G., Williams, N., Sham, P., Craddock, N. & Owen, M. J. Pooled genotyping of microsatellite markers in parent–offspring trios. *Genome Res.* **10**, 105–115 (2000).
71. Risch, N. & Teng, J. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. *Genome Res.* **8**, 1273–1288 (1998).
    A key paper that discusses the design of pooling studies for family-based association studies.
72. Akey, J., Jin, L. & Xiong, M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur. J. Hum. Genet.* **9**, 291–300 (2001).
73. Zollner, S. & von Haessler, A. A coalescent approach to study linkage disequilibrium between single nucleotide polymorphisms. *Am. J. Hum. Genet.* **66**, 615–628 (2000).

74. Martin, E. R. *et al.* SNPing away at complex disease: analysis of single-nucleotide polymorphisms around APOE in Alzheimer's disease. *Am. J. Hum. Genet.* **67**, 383–394 (2000).
75. Long, A. D. & Langley, C. H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
76. Bader, J. S. The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics* **2**, 11–24 (2001).
77. Pfeiffer, R. M., Rutter, J. L., Gail, M. H., Struewing, J. & Gastwirth, J. L. Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet. Epidemiol.* **22**, 94–102 (2002).
78. Cohen, J. *Statistical Power Analysis for the Behavioural Sciences* 2nd edn (Academic, New York, 1988).
79. Haff, L. A. & Smirnov, I. P. Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI–TOF mass spectrometry. *Genome Res.* **7**, 378–388 (1997).
80. Zhou, G.-H. *et al.* Quantitative detection of single nucleotide polymorphisms for a pooled DNA sample by a bioluminometric assay coupled with modified primer extension reactions (BAMBER). *Nucleic Acids Res.* **29**, E93 (2001).

Acknowledgements

P.S. was supported by grants from the UK Medical Research Council, the Wellcome Trust and the National Eye Institute.

Online links

DATABASES

The following terms in this article are linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
HLA class II DR | HLA class II DQ
OMIM: http://www.ncbi.nlm.nih.gov/Omim
type I (insulin-dependent) diabetes mellitus

FURTHER INFORMATION

Joel Bader's lab: http://www.curagen.com
Pak Sham's lab: http://statgen.iop.kcl.ac.uk
Pyrosequencing: http://www.pyrosequencing.com
Access to this interactive links box is free online.