Molecular Psychiatry (2009) 14, 796–803

www.nature.com/mp

ORIGINAL ARTICLE

A genome-wide association study in 574 schizophrenia trios using DNA pooling

G Kirov[1], I Zaharieva[2], L Georgieva[1], V Moskvina[1], I Nikolov[1], S Cichon[3], A Hillmer[3], D Toncheva[2], MJ Owen[1] and MC O'Donovan[1]

[1]*Department of Psychological Medicine, Cardiff University, Henry Wellcome Building, Heath Park, Cardiff, UK;* [2]*Department of Medical Genetics, University Hospital 'Maichin Dom', Sofia, Bulgaria and* [3]*Department of Genomics, Life and Brain Center, University of Bonn, Bonn, Germany*

The cost of genome-wide association (GWA) studies can be prohibitively high when large samples are genotyped. We conducted a GWA study on schizophrenia (SZ) and to reduce the cost, we used DNA pooling. We used a parent–offspring trios design to avoid the potential problems of population stratification. We constructed pools from 605 unaffected controls, 574 SZ patients and a third pool from all the parents of the patients. We hybridized each pool eight times on Illumina HumanHap550 arrays. We estimated the allele frequencies of each pool from the averaged intensities of the arrays. The significance level of results in the trios sample was estimated on the basis of the allele frequencies in cases and non-transmitted pseudocontrols, taking into account the technical variability of the data. We selected the highest ranked SNPs for individual genotyping, after excluding poorly performing SNPs and those that showed a trend in the opposite direction in the control pool. We genotyped 63 SNPs in 574 trios and analysed the results with the transmission disequilibrium test. Forty of those were significant at $P<0.05$, with the best result at $P=1.2\times10^{-6}$ for rs11064768. This SNP is within the gene CCDC60, a coiled-coil domain gene. The third best SNP ($P=0.00016$) is rs893703, within *RBP1*, a candidate gene for schizophrenia.
Molecular Psychiatry (2009) **14,** 796–803; doi:10.1038/mp.2008.33; published online 11 March 2008

Keywords: schizophrenia; pooled DNA; pooling; Illumina; genome-wide association

Introduction

Schizophrenia (SZ) has a strong genetic component, as shown by heritability estimates of ~80%.[1] However, there is no clear mode of transmission and most cases appear to be sporadic, suggesting a complex pattern of inheritance. Like other disorders of complex inheritance, it is now believed that many genes of small effect operate in the aetiology of this disorder. Epidemiological and molecular genetic studies suggest that genetic risk is mainly attributable to multiple alleles each with a small to moderate effect on liability. Several promising candidate susceptibility genes have been reported, but most of the genetic risk has not yet been attributed to specific genes.[2]

Recent technological advances have enabled researchers to perform genome-wide association (GWA) studies using hundreds of thousands of single-nucleotide polymorphisms (SNPs) that can capture the majority of common variation in the human genome. Several such studies have already been published for disorders such as, diabetes, ischaemic heart disease, hypertension and others (for example see ref[3]). These studies have produced many definitive findings, and in many cases, implicated genes that were not expected to be involved (for example, in ischaemic heart disease). The odds ratios found were modest even for the top loci (1.18–5.49 for heterozygote odds ratios and 1.48–18.52 for homozygote odds ratios in the largest study on seven different common disorders[3]) and required thousands of cases and controls to be used in the initial or the replication stages, to reach results that are genome-wide significant (that is, corrected for the number of SNPs tested). It is expected that any future susceptibility factors discovered in complex diseases will have even lower odds ratios, unlikely to be >2.0.

The need to genotype very large samples translates into very high costs that are beyond the budgets of most research groups in most countries. Although the sensitivity and specificity of pooled DNA analyses are imperfect, previous work in our own and other laboratories suggests that such analysis at the level of single locus[4] and highly parallel chip-based methods[5–8] can offer an economic alternative to individual genotyping, although clearly, not all loci with evidence for association in the samples so analysed will be detected. Several array-based

Correspondence: Dr G Kirov, Department Psychological Medicine, Cardiff University, School of Medicine, Heath Park, Henry Wellcome Building, Cardiff CF14 4XN, UK.
E-mail: kirov@cardiff.ac.uk
Received 24 October 2007; revised 1 February 2008; accepted 8 February 2008; published online 11 March 2008

DNA-pooling studies were published in the last few months, that identified new, or replicated known illness/trait loci, thus providing proof of principle that DNA pooling provides a valid alternative to these very expensive studies, at a hugely reduced cost.[9–12]

Here, we apply the pooling principle to a large sample of SZ parent–offspring trios and controls recruited in Bulgaria, comprising a total of over 2000 individuals. Although we previously demonstrated that the Affymetrix system could yield reliable pooled genotypes,[5] unpublished data since that time suggested that the Illumina system performed rather better than the two 250 K Affymetrix arrays available at the time when the current work started. Therefore, in this study, we chose the Illumina HumanHap550 array. This array interrogates ~550 000 SNPs and provides excellent coverage of known common variation in the human genome. Ninety per cent of all Phase I+II HapMap loci with a minor allele frequency of f $\geqslant$0.05 are covered by at least one SNP in high linkage disequilibrium on the HumanHap550 BeadChip for the Utah residents with ancestry from northern and western Europe (CEU) population (http://www.illumina.com/downloads/HUMANHAP550_DataSheet.pdf).

Materials and methods

Samples

DNA pools were constructed from three sources. Two pools were constructed from a nuclear-family-based association sample, one pool consisted of DNA from 574 unrelated SZ patients and the other of DNA from all their 1148 parents. The trios were recruited in Bulgaria between 1999 and 2004 by a team organized and trained by GK. All probands satisfied DSM-IV criteria for schizophrenia,[13] and DNA was available from both parents. Diagnoses were made on the basis of a semi-structured interview with the SCAN instrument[14] and inspection of hospital discharge summaries. The third pool was constructed from DNA samples of 605 healthy controls from the same regions in Bulgaria as the trios. The controls were recruited in several settings: random people applying for driving licences, non-psychiatric attendees at a GP surgery and hospital staff. No matching for age was implemented.

Pool construction

DNA pools were constructed by taking equimolar amounts of DNA from each individual. DNA was initially serially diluted to 5–15 ng mcl^{-1} and then measured in duplicate with PicoGreen DNA quantitation reagent (Molecular Probes, Eugene, OR, USA) on a Labsystems Ascent Fluoroscan (Life Sciences International, Basingstoke, UK). The final pool concentrations were ~8–9 ng mcl^{-1} and then concentrated to 50 ng mcl^{-1} with Microcon YM-100 Centrifugal Filter Units (Millipore Corporation, Billerica, MA USA), as required for hybridization with Illumina arrays.

Genotyping on microarrays

We used Illumina HumanHap550 arrays. Genotyping was performed at the laboratory in Bonn, using the manufacturer's protocols. We used eight replicate arrays for each pool.

Statistical analysis

Approximation of allele *A* frequencies for each replicate, *i*, was produced on the basis of the raw data as follows: f_alleleA = Xraw/(Xraw + Yraw), averaged over the number of replicates in each pool. We firstly examined the array/array correlations produced by the replicates and excluded arrays that appeared to be outliers (see Results). *P*-values were estimated using the following statistic combining experimental and sampling errors, which is based on previous work on DNA pooling:[4,8]

$$T_{\mathrm{comb}} = \frac{(\bar{f}^{(\mathrm{c})} - \bar{f}^{(\mathrm{p})})^2}{v + \varepsilon_{\mathrm{c}}^2 + \varepsilon_{\mathrm{p}}^2} \qquad (1)$$

where

$$\bar{f}^{(\mathrm{c})} = (1/n_{\mathrm{c}}) \sum_{i=1}^{n_{\mathrm{c}}} f_i^{(\mathrm{c})}$$

is the mean of the allele frequencies over n_{c} children pool replicates,

$$\bar{f}^{(\mathrm{p})} = (1/n_{\mathrm{p}}) \sum_{i=1}^{n_{\mathrm{p}}} f_i^{(\mathrm{p})}$$

is the mean of the allele frequencies over n_p parents pool replicates,

$$\varepsilon_{\mathrm{c}}^2 = (1/(n_{\mathrm{c}} - 1)) \sum_{i=1}^{n_{\mathrm{c}}} (f_i^{(\mathrm{c})} - \bar{f}^{(\mathrm{c})})^2 \text{ and}$$
$$\varepsilon_{\mathrm{p}}^2 = (1/(n_{\mathrm{p}} - 1)) \sum_{i=1}^{n_{\mathrm{p}}} (f_i^{(\mathrm{p})} - \bar{f}^{(\mathrm{p})})^2$$

are the variances due to experimental error in the children and parents pools, respectively, $v = (\bar{f}^{(\mathrm{p})}(1-\bar{f}^{(\mathrm{p})}))/(4N)$ is the variance estimated due to the sampling error and *N* is the number of trios. Estimation of *v* is based on the estimated proportion of heterozygous parents in the sample $\hat{\mathrm{h}} = 2\bar{f}^{(\mathrm{p})}(1-\bar{f}^{(\mathrm{p})})$, which requires the assumption of Hardy–Weinberg equilibrium in parents.[15]

We examined if there was a systematic inflation of the test statistic using the genomic control lambda, as defined by Devlin and Roeder.[16] When standard χ^2-statistic was used (directly comparing transmitted and non-transmitted allele frequencies, assuming Hardy–Weinberg equilibrium in parents), the genomic control, λ, was equal to 2.22, (indicating a systematic inflation), however, when our preferred statistic T_{comb} was used (which is corrected for experimental error), the genomic control λ was equal to 0.938, indicating that further adjustment for systematic variation between pools was redundant.

Data filtering

We filtered the data as follows: Firstly, we wanted to exclude SNPs whose allele frequencies were poorly predicted by pooled analysis. To do this, we obtained

population allele frequencies from the HapMap database (http://www.hapmap.org) and estimated the *k*-correction coefficients for individual SNPs by using the CEU frequencies as explained in detail in our previous work.[5] Briefly, *k* is a coefficient that is used to correct data in pooling experiments, due to differential amplification of the two alleles of a SNP. In a heterozygous individual, there should be an equal strength of signal obtained from the two alleles, however this is rarely the case, due to different dye intensities and amplification of different nucleotides.

$$k = h_A/h_B \quad (2)$$

where h_A and h_B are the measurements representing alleles A and B in heterozygous individuals (for example, signal intensities). To estimate *k*, instead of measuring the intensities of alleles in heterozygous individuals obtained from individual genotyping (clearly a problematic task when dealing with large numbers of SNPs), we used a method by which *k* can be approximated from the known allele frequencies in the CEU population.[5] We assume that the CEU frequency approximates to the true frequency in our control sample and that differences between the two are the result of a bias in estimating the frequency of the alleles in the pooling experiment. Of course this is only an approximation because there may be ethnic-related differences in allele frequency and also, since the CEU population is small, there is an appreciable effect of sampling variance. Under this assumption, we obtained the correction coefficients (*k*) for the SNPs on the array required to convert the control pools to match the CEU frequencies, according to the following formula:[5]

$$k = \frac{H_A}{H_B} \cdot \frac{f_B}{f_A} \quad (3)$$

The mean *k* was 0.45 for the whole sample, (s.d. = 0.62), indicating the presence of a systematic under-representation of one of the alleles, possibly due to differential intensity signals from the two dyes in this particular experiment. This bias is evident on Figure 1, which demonstrates the preferential under-estimation by pooling of allele A (Figure 1).

We have previously shown that the use of SNPs with extreme values of *k* results in high error rates[17] and, therefore, we excluded SNPs with extreme *k* values. We filtered out the SNPs with the worst 5% of *k*-values (2.5% in each direction). This translated to retaining SNPs with values of $0.15 < k < 1.8$. This also excluded non-polymorphic SNPs in the CEU sample (which would produce *k*-values of 0 or infinity) and SNPs with no frequency data in the HapMap at that time.

Ignoring the effect of sampling variance, in the absence of pooled genotyping errors, we would expect that the parental allele frequencies would be intermediate between cases and controls, since parents are genetically closer to the cases than are unrelated controls. Thus, to enrich for findings in the pooled data that are not attributable to pooled genotyping error, we excluded SNPs where the control allele frequencies were in the opposite direction of the trend found in the trios, as those were most likely false-positive findings (for example, if a SNP had a frequency of 0.20 in cases, and a frequency of 0.25 in parents, then it was excluded if the frequency in controls was < 0.25.

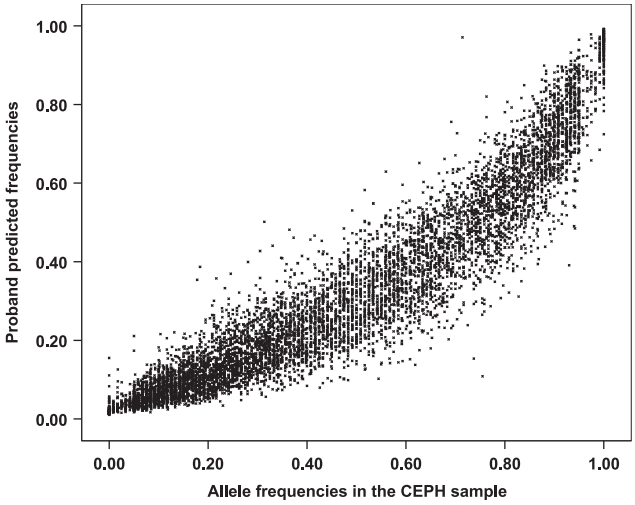

Figure 1 Correlation between predicted allele frequencies and population frequencies obtained from the CEU sample, Pearson's correlation: $r = 0.938$.

Individual genotyping

We decided to follow up only the top-ranked SNPs, choosing a cut-off criterion of $P < 0.001$ according to formula (1). There were 763 SNPs that reached that level of significance, after excluding SNPs on the X-chromosome. Of those, 146 were in the vicinity of each other (we only typed the single best SNP from a cluster of significant ones), 123 SNPs had either poor *k*-values, or had no HapMap frequency data at the time and 380 SNPs showed an effect in the opposite direction between controls and trios (some SNPs were excluded on the basis of more than one of these criteria and a small number of the lowest-ranked SNPs were not included as we had to restrict the individual genotyping to three Sequenom panels). There were 133 SNPs that satisfied our filtering and were selected for primer design. Follow-up genotyping was performed in the SZ trios comprising the pooled, proband and parent samples. Individual genotyping was performed with the Sequenom MassARRAY using iPlex chemistry (Sequenom, San Diego, CA, USA, http://www.sequenom.com), according to the recommendations of the manufacturer. Three panels of SNPs markers were designed using Sequenom Assay Design 3.1 software. Of the 133 SNPs we presented for design, 42 were not included by the software into the three panels we wanted to construct. Of the SNPs that were initially included, 28 were dropped from the analysis at various stages of genotyping and data cleaning, as they produced unreliable genotypes.

All assays were first optimized in 30 reference CEU trios from the HapMap database. Genotypes were called in duplicate by two independent raters (one blind to sample identity). Genotypes of CEU samples were compared to those available on the HapMap database, to provide a measure of genotyping accuracy. The comparison was performed using a computer programme named 'CephCheck', which allows an automated comparison of genotypes, thus minimizing manual intervention (available upon request from its author Dobril Ivanov: ivanovdk@cardiff.ac.uk). Genotyping assays were only considered suitable for analysis if our genotypes were identical to those in the HapMap database.

Statistical analysis

Individual genotype data were analysed with the transmission disequilibrium test,[18] which considers the preferential transmission of alleles from heterozygous parents.

Results

The Illumina arrays delivered highly reproducible results from array replicates. Figure 2 shows a picture of the correlation between predicted allele frequencies obtained from the same pool hybridized on two separate arrays. The correlations between the predicted allele frequencies from one pool hybridized on eight arrays varied between $r=0.992$ and $r=0.998$ (Figure 2). For association analysis, we excluded arrays that gave correlations with several other arrays below an arbitrary chosen cutoff point of $r<0.996$, as these appeared the only outliers. This left all eight arrays from controls, seven from probands and six from parents. Based upon these arrays, averaged allele frequencies in parents and probands were, as expected, very highly correlated ($r=0.999$, Figure 3).

Correlation with CEU population frequencies

Population frequencies for the SNPs on the Illumina arrays were obtained from the HapMap database (www.hapmap.org) and, where appropriate, reversed according to the DNA strand used. Figure 1 shows the correlations between the predicted allele frequencies and the CEU population frequencies, $r=0.938$. It is clear from the figure that a large number of SNPs had severely distorted frequencies and that there was a systematic underestimation of the frequency of allele A on the arrays. As detailed in the Methods section, we filtered out SNPs with extreme values of k ($k>1.8$ or <0.15) under the assumption that the allele frequencies were predicted poorly for such SNPs. After exclusions, the samples were analysed for 433 680 SNPs. In the present study, we did not apply k to the association statistic, as pilot unpublished data, and previous work[5] did not reveal a superior outcome when using SNP arrays was corrected with k.

Of the 63 SNPs we successfully genotyped individually, 40 showed significant transmission to probands at $P<0.05$, 15 of those at $P<0.005$ and 4 of those at <0.0005 (Table 1). All markers were in Hardy–Weinberg equilibrium (cutoff $P>0.001$). The strongest result was obtained for rs11064768: $P=1.2\times10^{-6}$. This SNP is on chromosome 12, within the gene CCDC60, a coiled-coil domain gene. The result does not reach genome-wide significance, which has been estimated at 1.85×10^{-7} taking into account the linkage disequilibrium between SNPs, which reduce the effective number of independent tests.[19] Although calculated for the Affymetrix 500 K array, this level is likely to be of a similar magnitude in the Illumina HumanHap550 array and probably even slightly more conservative, as more SNPs on the

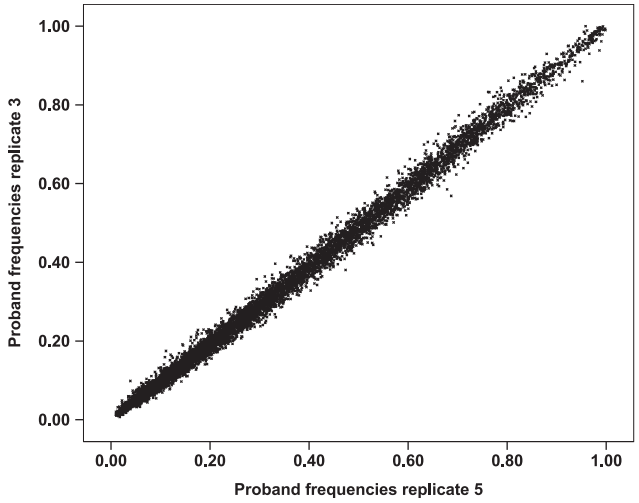

Figure 2 Correlations between predicted allele frequencies from the same pool (probands) replicated on two arrays. About 27 000 single-nucleotide polymorphisms are shown, $r=0.996$.

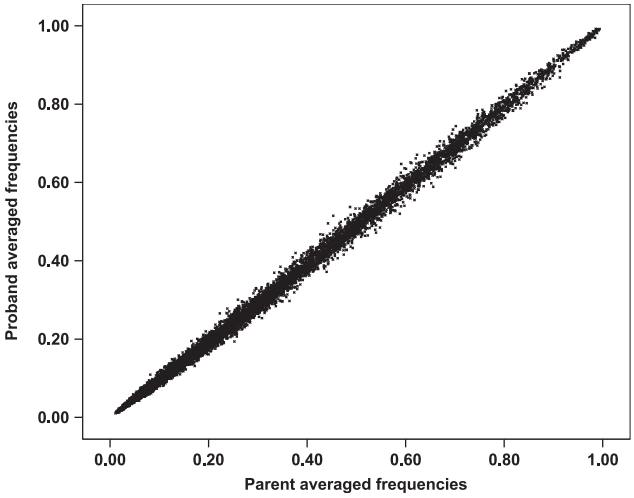

Figure 3 Correlations between predicted averaged allele frequencies in proband and parents' pool, $r=0.999$. Presented are data on ~27 000 single-nucleotide polymorphisms.

Table 1 Individual genotyping results

RS number	*Gene name*[a]	*Chromo-some*	*Position in bp (built 36, March, 2006)*	*SNP type*	*Allele A*	*Pool frequencies*				*Individual genotyping frequencies*		*Difference in pools*	*Difference individual genotyping*	p-*Comb test*	p-*TDT*	*T:NT*
						Parents	*Children*	*Controls*	*CEU*	*Parents*	*Children*					
rs11064768	*CCDC60*	12	118302892	A/G	A	0.851	0.781	0.855	0.905	0.936	0.911	−0.070	−0.025	$3.3x10^{-7}$	0.0000012	33:86
rs11782269		8	8527561	G/T	G	0.320	0.254	0.321	0.670	0.694	0.654	−0.066	−0.040	0.0003	0.000084	186:270
rs893703	*RBP1*	3	140733339	A/G	A	0.685	0.621	0.685	0.845	0.881	0.854	−0.064	−0.027	0.000004	0.00016	86:143
rs2288039	*CIRH1A*	16	67745613	C/T	C	0.432	0.385	0.435	0.642	0.752	0.721	−0.047	−0.031	0.0007	0.00043	163:233
rs9883916	*STAG1*	3	137691478	C/T	C	0.460	0.404	0.479	0.750	0.789	0.762	−0.056	−0.027	0.0005	0.001	143:204
rs6078931	*SPTLC3*	20	13053853	C/T	C	0.119	0.087	0.121	0.150	0.151	0.127	−0.032	−0.024	0.0004	0.002	111:162
rs980616		9	108238336	G/T	G	0.391	0.305	0.408	0.675	0.795	0.768	−0.086	−0.027	0.0001	0.002	148:206
rs4767235		12	113122292	A/G	A	0.664	0.713	0.627	0.792	0.863	0.886	0.048	0.023	0.00002	0.002	151:102
rs2659504	*PPP3CA*	4	102314520	C/T	C	0.750	0.711	0.786	0.922	0.922	0.907	−0.039	−0.015	0.00007	0.0023	56:94
rs16874040	*CLIC5*	6	46031767	A/C	A	0.978	0.950	0.982	0.992	0.992	0.985	−0.028	−0.006	4.8×10^{-10}	0.0027	2:14
rs1478684		11	27293921	G/T	G	0.458	0.397	0.477	0.800	0.750	0.720	−0.061	−0.030	0.0003	0.0029	168:227
rs2029099	*BC035112*	2	14314367	G/T	G	0.786	0.813	0.784	0.883	0.900	0.920	0.027	0.020	0.0006	0.0042	118:78
rs10461669		5	27657934	A/G	A	0.371	0.322	0.392	0.602	0.671	0.640	−0.049	−0.031	0.0008	0.0044	193:253
rs9457631		6	159772770	C/T	C	0.802	0.767	0.817	0.875	0.917	0.900	−0.035	−0.017	0.0007	0.0046	63:99
rs11144978	*KIAA0367*	9	78415148	A/C	A	0.310	0.243	0.335	0.733	0.748	0.723	−0.068	−0.025	0.00008	0.0056	163:217
rs424970	*PGM5*	9	70303655	C/T	C	0.463	0.411	0.469	0.758	0.730	0.704	−0.053	−0.026	0.0007	0.0096	183:236
rs3759700		14	58632889	G/T	G	0.605	0.645	0.593	0.750	0.770	0.790	0.040	0.020	0.0006	0.01	216:166
rs11203820		8	16884525	G/T	G	0.546	0.608	0.522	0.483	0.493	0.522	0.062	0.029	0.0002	0.01	289:231
rs1510881	*SMARCA5*	4	144658532	A/G	A	0.474	0.416	0.502	0.775	0.781	0.759	−0.058	−0.022	0.0007	0.011	143:189
rs9548798	*LHFP*	13	38995853	A/C	A	0.672	0.717	0.665	0.758	0.811	0.832	0.045	0.021	0.0008	0.012	193:147
rs6027861		20	58859975	C/T	C	0.664	0.610	0.674	0.833	0.851	0.832	−0.054	−0.019	0.0001	0.014	119:160
rs176512	*SYPL1*	7	105548396	A/G	A	0.838	0.801	0.851	0.898	0.905	0.892	−0.036	−0.014	0.0006	0.014	74:107
rs10518356	*BC054887*	1	71336573	C/T	C	0.650	0.694	0.609	0.917	0.925	0.938	0.044	0.013	0.0009	0.014	91:61
rs12224013	*TTC17*	11	43471286	G/T	G	0.921	0.886	0.934	0.946	0.976	0.968	−0.035	−0.008	0.000001	0.015	16:33
rs8045220		16	54323846	G/T	G	0.095	0.072	0.102	0.183	0.175	0.157	−0.023	−0.018	0.0005	0.018	123:163
rs9345837		6	67242433	A/C	A	0.476	0.412	0.481	0.692	0.736	0.713	−0.065	−0.023	0.0003	0.018	199:249
rs12455939	*KIAA0802*	18	8709122	C/T	C	0.349	0.298	0.383	0.695	0.697	0.674	−0.050	−0.023	0.00007	0.021	193:241
rs13084692	*DCBLD2*	3	100092139	C/T	C	0.888	0.847	0.901	0.900	0.970	0.960	−0.041	−0.010	0.00001	0.021	19:36
rs6965651	*DPP6*	7	153767488	G/T	G	0.784	0.732	0.799	0.904	0.920	0.900	−0.053	−0.020	0.00004	0.027	67:95
rs2395174	*HLA-DRA*	6	32512856	G/T	G	0.755	0.802	0.747	0.767	0.796	0.815	0.047	0.019	0.0003	0.030	191:151
rs12034664	*GRRP1*	1	26354176	C/T	C	0.608	0.538	0.623	0.850	0.850	0.830	−0.070	−0.020	0.00002	0.030	121:157
rs10509722	*HPSE2*	10	100268007	C/T	C	0.797	0.828	0.762	0.871	0.948	0.958	0.031	0.010	0.0004	0.035	66:44
rs4761874	*GALNT6*	12	50040346	C/T	C	0.182	0.133	0.193	0.164	0.246	0.228	−0.049	−0.018	0.00006	0.037	174:215
rs2985662	*AK057351*	13	21347588	A/C	A	0.566	0.610	0.531	0.658	0.650	0.680	0.043	0.030	0.0006	0.038	258:213
rs5752019	*RUTBC2*	22	23646350	A/G	A	0.521	0.461	0.532	0.703	0.756	0.736	−0.060	−0.020	0.0004	0.038	175:216
rs1188568	*OR11G2*	14	19726485	A/G	A	0.553	0.502	0.554	0.692	0.721	0.702	−0.051	−0.019	0.0001	0.041	192:234
rs7172362		15	96992329	C/T	C	0.929	0.906	0.940	0.942	0.980	0.973	−0.023	−0.006	0.0001	0.047	15:28
rs930767		2	7220450	G/T	G	0.475	0.528	0.463	0.442	0.503	0.524	0.053	0.021	0.0008	0.047	293:247
rs546464		9	115434335	C/T	C	0.865	0.896	0.865	0.917	0.960	0.968	0.031	0.008	0.00003	0.049	51:33
rs946442	*FLJ00377*	1	54434328	A/G	A	0.932	0.955	0.929	0.949	0.986	0.991	0.024	0.006	0.0008	0.049	18:8
rs2007451		1	30571886	A/G	A	0.362	0.301	0.366	0.658	0.647	0.628	−0.061	−0.019	0.00009	0.054	218:260
rs12455836	*TWSG1*	18	9347497	A/C	A	0.876	0.829	0.881	0.941	0.957	0.948	−0.047	−0.008	0.000003	0.055	35:53
rs10484735	*TCBA1*	6	124453652	A/G	A	0.613	0.567	0.613	0.758	0.814	0.799	−0.046	−0.015	0.0004	0.067	146:179
rs17035181	*PDGFC*	4	157897961	G/T	G	0.639	0.569	0.680	0.808	0.857	0.844	−0.070	−0.013	0.00006	0.08	114:142
rs6926332	*PTPRK*	6	128527832	A/C	A	0.611	0.544	0.634	0.825	0.880	0.870	−0.068	−0.010	0.00009	0.094	99:124
rs1996794	*SBF2*	11	9779172	A/C	A	0.447	0.377	0.467	0.758	0.728	0.713	−0.070	−0.016	0.0002	0.11	199:232

Table 1 Continued

RS number	*Gene name*[a]	*Chromo-some*	*Position in bp (built 36, March, 2006)*	*SNP type*	*Allele A*	*Pool frequencies*			*Individual genotyping frequencies*			*Difference in pools*	*Difference individual genotyping*	p-*Comb test*	p-*TDT*	*T:NT*
						Parents	*Children*	*Controls*	*CEU*	*Parents*	*Children*					
rs6950779	*COBL*	7	51114044	G/T	G	0.350	0.281	0.349	0.575	0.720	0.710	−0.069	−0.010	0.00008	0.11	193:225
rs7930681	*HBG2*	11	5560551	C/T	C	0.317	0.268	0.320	0.635	0.519	0.501	−0.049	−0.018	0.0005	0.12	254:290
rs7122479	*ME3*	11	85842795	G/T	G	0.587	0.640	0.575	0.763	0.800	0.810	0.054	0.010	0.00006	0.14	169:143
rs2372441		2	36234908	C/T	C	0.871	0.835	0.878	0.900	0.955	0.949	−0.036	−0.006	0.00001	0.17	39:52
rs629310		6	153234604	G/T	G	0.617	0.681	0.598	0.661	0.691	0.704	0.064	0.013	0.0002	0.2	237:210
rs768214		18	3332256	A/C	A	0.609	0.555	0.626	0.733	0.800	0.790	−0.054	−0.010	0.0005	0.22	153:175
rs16934812	*TMTC1*	12	29763585	G/T	G	0.813	0.775	0.832	0.808	0.933	0.927	−0.039	−0.006	0.0003	0.22	61:75
rs17692695		10	29209389	A/G	A	0.855	0.821	0.870	0.922	0.937	0.931	−0.034	−0.006	0.00007	0.24	57:70
rs6657332	*RYR2*	1	235897518	G/T	G	0.296	0.338	0.297	0.433	0.412	0.424	0.042	0.012	0.0006	0.25	256:231
rs17231292	*SSBP2*	5	80811791	A/G	A	0.765	0.720	0.776	0.858	0.887	0.880	−0.044	−0.007	0.0004	0.29	97:112
rs1463535		3	28649285	A/G	A	0.300	0.256	0.318	0.492	0.450	0.440	−0.044	−0.010	0.0007	0.34	239:260
rs6926853	*TIAM2*	6	155603867	C/T	C	0.878	0.907	0.863	0.924	0.972	0.975	0.028	0.003	0.0002	0.37	34:27
rs512089		9	23864047	G/T	G	0.394	0.317	0.403	0.742	0.760	0.750	−0.077	−0.010	0.00009	0.37	176:193
rs532210		11	51249087	C/T	C	0.629	0.582	0.624	0.573	0.664	0.656	−0.047	−0.008	0.0007	0.42	220:237
rs13406291	*ARHGAP15*	2	143818289	A/C	A	0.759	0.802	0.752	0.808	0.822	0.829	0.043	0.007	0.0007	0.42	163:149
rs4833722		4	122604933	G/T	G	0.437	0.372	0.459	0.707	0.723	0.717	−0.065	−0.006	0.0004	0.53	217:230
rs320203		9	103983047	A/C	A	0.198	0.241	0.195	0.183	0.136	0.136	0.043	0.001	0.0006	0.89	126:124

[a]Gene names are given for intragenic single-nucleotide polymorphisms (SNPs) and SNPs within 10 kb from a gene. Allele A for each SNP is specified (column 7, 'Allele A') and its predicted frequencies in pools (parents, children and controls) and individual genotyping frequencies (CEU, parents and children) are presented (columns 8–13). Predicted allele frequency differences and real individual genotyping differences are given (column 14, 'Difference in pools', column 15 'Difference individual genotyping'), significance level (*p*-combined test value) is calculated according to formula (1) (see Materials and methods section) (column 16, '*p*-comb test'). *P*-value from transmission disequilibrium test (TDT) is given (column 17, '*p*-TDT') and transmission/non-transmission (T/NT) counts from heterozygous parents for allele A (column 18, 'T:NT'). All markers were in Hardy–Weinberg equilibrium (cutoff $P > 0.001$). The *P*-value from TDT is likely to differ slightly from that produced by the difference between the true parental and offspring frequencies.

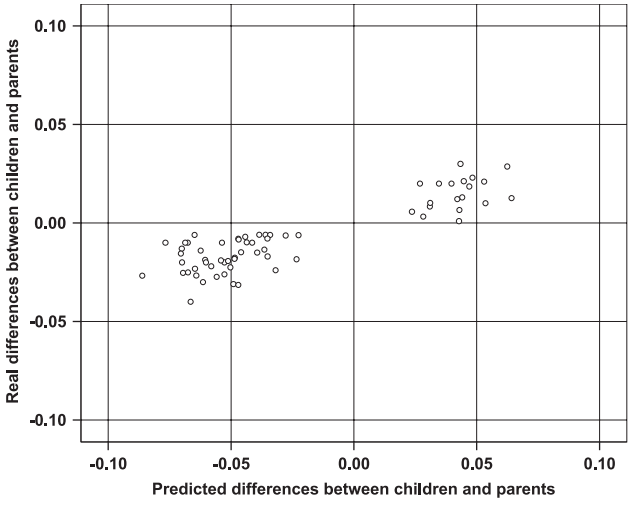

Figure 4 Correlation between predicted and real differences in allele frequencies between SZ proband and their parents, $r = 0.88$.

Affymetrix 500 K array are in high linkage disequilibrium and are, therefore, redundant, thus reducing the number of effective tests, compared with Illumina HumanHap550.

The differences in allele frequencies between offspring and their parents in the 63 SNPs were predicted extremely well by pooled genotyping (Figure 4). The Pearson's correlation is very high at $r = 0.88$, $P = 10^{-21}$. However, the magnitude of the differences was consistently overestimated, reflecting the fact that we had targeted the biggest predicted differences.

Discussion

We present results from the first parent–offspring GWA study in SZ and one of the very few on any phenotype. We genotyped 574 complete and unrelated parent–offspring trios, where the proband suffers with SZ. To reduce the cost of the study, we performed a two-stage analysis, first, genotyping DNA pools on eight Illumina HumanHap550 arrays each and then individually genotyping some of the most promising SNPs.

The Illumina arrays performed remarkably well, compared to other techniques we have tried over the years.[4] The single array/array correlations of predicted allele frequencies varied between $r = 0.992$ and 0.998, and we were able to select only those that were at $r \geqslant 0.996$. The vast majority of SNPs predicted the correct population frequencies with only modest distortion (as reflected by the 433 680 SNPs that had correction coefficients k ranging from 0.15 to 1.8, although the frequency of allele A was systematically overestimated in this particular experiment (Figure 1). (This bias should not affect substantially the sensitivity and specificity of the results in this study, as there is differential amplification in any pooling experiment, but we are not able to estimate the exact size of this effect).

The SNPs were ranked according to the significance level estimated by formula (1). In addition to the sampling error, this statistic takes into account the variability due to technical error. To reduce the number of false-positive results, we performed several filtering procedures. We excluded SNPs with severely distorted allele frequencies (using the correction coefficient k as our criterion, $1.8 > k > 0.15$) and those SNPs, which showed a trend in the opposite direction in the controls sample (to reduce the number of results that, although significant in the trios sample, were unlikely to survive replication in an independent sample). To further reduce the cost of the project, we genotyped only SNPs that we could fit into three panels of Sequenom genotyping. We obtained high-quality results on 63 SNPs. These SNPs gave identical genotypes to the CEU trios from the HapMap database and were in Hardy–Weinberg equilibrium (at $P > 0.001$).

A study using parent–offspring trios is more difficult to perform with DNA pooling than a case–control study, as the differences between parents and offspring are roughly half of the differences that could be expected to be encountered in a similar-sized case–control study, and could easily approach the resolution limit of the pooling approach. We reasoned that this drawback would be offset by the elimination of false-positive results due to population stratification and by the use of an additional sample of controls, where we wanted to observe a similar direction of effect. Our method produced highly significant results in detecting the correct direction and magnitude of the differences in allele frequencies between SZ offspring and their parents: $r = 0.88$, $P = 10^{-21}$ (Figure 4).

Our method provided significant results: 68% of the genotyped SNPs were significant at $P < 0.05$ level, with four results < 0.0005 level. The best result in the current study is for rs11064768, a SNP on chromosome 12, within the gene CCDC60, a coiled-coil domain gene. The $P = 1.2 \times 10^{-6}$ does not reach genome-wide significance, which we estimate to be at least 1.85×10^{-7} (see Materials and methods), but is better than the expected value for the smallest P-value out of 500 000 independent tests assuming the null hypothesis ($P = 2 \times 10^{-6}$). This indicates that this might indeed be the best P-value that this sample could achieve, if every SNP was typed individually. The most interesting finding is, however, for rs893703, an SNP on chromosome 3, within the gene RBP1, a cellular retinol-binding protein which inhibits PI3K/Akt signalling.[20] The genes in this pathway have been implicated in SZ pathogenesis.[21]

These results demonstrate that DNA pooling can be used successfully, especially in phenotypes where large genetic effects from single SNPs are likely to exist, as such differences between two samples appear easy to detect with these methods. The lack of genome-wide significant findings in this study could be due to several factors. First, due to cost constraints and problematic SNPs, we genotyped only

about half of the top-ranked SNPs (63 of 133). Second, we might have filtered out some true significant results that could, by chance, also produce poor *k* or have an opposite direction in the control sample. It is also possible that some *k*-estimates were quite poor, because we used the small CEU HapMap set of 60 parents for a US population, which could introduce more variance. Third, it is possible that the best SNPs were not identified by the pooling as highly significant. This could be due for example by technical artefacts or violation of our assumption of Hardy–Weinberg equilibrium in parents, for some SNPs. Another possibility is that our sample of 574 trios does not have the power to detect results with genome-wide significance, that is, no SNP on the HumanHap550 array might reach this level of significance when typed individually in a sample of this size. Current thinking in GWA studies is that several thousand cases and controls are required to detect signals in complex disorders with genome-wide significance. Several confirmed associations in diabetes escaped initial detection when 2000 cases and 3000 controls were tested in the largest study so far[3] but were confirmed after genotyping much larger numbers. Therefore, several of our best results need to be tested for replication in other samples, and conversely, the best results of other GWA studies in SZ need to be examined against our pooling data as a replication attempt.

We expect that the results from several GWA studies would need to undergo a meta-analysis, to identify the true susceptibility variants. At a fraction of the cost (using 24 instead of >2000 Illumina arrays), our DNA-pooling methodology appears to provide great value for money as a first-pass analysis in GWA studies. This methodology could enable researchers to obtain data from many existing collections of cases and controls, at an affordable cost.

Acknowledgments

This work was funded by the International Centre for Genetic Engineering and Biotechnology, Trieste, grant to the Department of Medical Genetics, Sofia (CRP/BUL04-01) and a Schizophrenia programme grant from the MRC to the Department of Psychological Medicine, Cardiff University (ref G9309834). The recruitment of trios was funded by the Janssen Research Foundation, Beerse, Belgium.

References

1 Cardno AG, Gottesman II. Twin studies of schizophrenia: from bow-and-arrow concordances to star war Mx and functional genomics. *Am J Med Genet* 2000; **97**: 12–17.
2 Kirov G, O'Donovan MC, Owen MJ. Finding schizophrenia genes. *J Clin Invest* 2005; **115**: 1440–1448.
3 The Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
4 Sham PC, Bader JS, Craig I, O'Donovan M, Owen M. DNA pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002; **2**: 862–871.
5 Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'Donovan MC. Pooled DNA genotyping on affymetrix SNP genotyping arrays. *BMC Genomics* 2006; **7**: 27.
6 Wilkening S, Chen B, Wirtenberger M, Burwinkel B, Försti A, Hemminki K *et al*. Allelotyping of pooled DNA with 250 K SNP microarrays. *BMC Genomics* 2007; **8**: 77.
7 Docherty SJ, Butcher LM, Schalkwyk LC, Plomin R. Applicability of DNA pools on 500 K SNP microarrays for cost-effective initial screens in genomewide association studies. *BMC Genomics* 2007; **8**: 214.
8 MacGregor S. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *Eur J Hum Genet* 2007; **15**: 501–504.
9 Shifman S, Bhomra A, Smiley S, Wray NR, James MR, Martin NG *et al*. A whole genome association study of neuroticism using DNA pooling. *Mol Psychiat* 2008; **13**: 302–312.
10 Melquist S, Craig DW, Huentelman MJ, Crook R, Pearson JV, Baker M *et al*. Identification of a novel risk locus for progressive supranuclear palsy by a pooled genomewide scan of 500, 288 single-nucleotide polymorphisms. *Am J Hum Genet* 2007; **80**: 769–778.
11 Steer S, Abkevich V, Gutin A, Cordell HJ, Gendall KL, Merriman ME *et al*. Genomic DNA pooling for whole-genome association scans in complex disease: empirical demonstration of efficacy in rheumatoid arthritis. *Genes Immun* 2007; **8**: 57–68.
12 Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C *et al*. A genomewide association study of skin pigmentation in a South Asian population. *Am J Hum Genet* 2007; **81**: 1119–1132.
13 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Press: Washington, DC, 1994.
14 Wing JK, Babor T, Brugha J, Cooper JE, Giel R, Jablensky A *et al*. Schedules for clinical assessment in neuropsychiatry. *Arch Gen Psychiat* 1990; **47**: 137–144.
15 Risch N, Teng J. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res* 1998; **8**: 1273–1288.
16 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
17 Moskvina V, Norton N, Williams N, Holmans P, Owen M, O'Donovan M. Streamlined analysis of pooled genotype data in SNP-based association studies. *Genet Epidemiol* 2005; **28**: 273–282.
18 Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–516.
19 Moskvina V, Schmidt KM. On multiple testing correction in genome-wide association studies. *Genet Epidemiol* (in press).
20 Farias EF, Marzan C, Mira-y-Lopez R. Cellular retinol-binding protein-I inhibits PI3K/Akt signalling through a retinoic acid receptor-dependent mechanism that regulates p85-p110 heterodimerization. *Oncogene* 2005; **24**: 1598–1606.
21 Kalkman HO. The role of the phosphatidylinositide 3-kinase-protein kinase B pathway in schizophrenia. *Pharmacol Ther* 2006; **110**: 117–134.