Family-based genome-wide association studies

In the last 2 years, the effort to identify genes affecting common diseases and complex traits has been accelerated through the use of genome-wide association studies (GWAS). The availability of existing large collections of linkage data paved the way for the use of family-based GWAS. Although most published GWAS used population-based designs, family-based designs have played an important role, particularly in replication stages. Family-based designs offer advantages in terms of quality control, the robustness to population stratification and the ability to perform genetic analyses that cannot be achieved using a sample of unrelated individuals, such as testing for the effect of imprinted genes on phenotypes, testing whether a genetic variant is inherited or *de novo* and combined linkage and association analysis.

KEYWORDS: association, family, genome-wide association study, GWAS, SNPs, TDT, Transmission Disequilibrium Test

Beben Benyamin[†], Peter M Visscher & Allan F McRae

[†]*Author for correspondence:*
Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research,
300 Herston Road, Brisbane,
QLD 4029, Australia
Tel.: +61 733 620 169;
Fax: +61 733 610 101;
bebenB@qimr.edu.au

Over the last few decades, a huge effort has been invested into the identification and characterization of genes influencing human diseases and phenotypes. While success stories have been reported for Mendelian traits, the identification of genes underlying complex diseases has been slow and difficult [1]. However, the development of large-scale genotyping platforms, which was precipitated by the completion of the Human Genome Project [2,3], the availability of SNPs in the public databases [2] and the completion of the first and second stages of the International Haplotype Map (HapMap) Project [3,4], allowed the extension of genotype–phenotype association studies from the realm of a small number of candidate genes to that of an entire genome.

Unlike candidate gene association studies, which test for the association between phenotypes and variant(s) in biologically selected gene(s), genome-wide association studies (GWAS) test hundreds of thousands or millions of SNPs covering the entire genome without reference to any particular gene(s). The rapid explosion in the number of GWAS has been accompanied by the increase on the number of reported associations between genetic variants and common complex diseases [101] and has led to the implication of novel pathways in the development of various diseases.

Published GWAS have mostly used samples of unrelated individuals as, for a given genotyping budget, this is in general the most powerful study design [101]. However, much effort has been put into collecting family-based samples, both in resources such as twin registries that collect both genetic and phenotypic information on a large number of phenotypes in twins and their families [5] and from clinicians studying disease phenotypes. These collections formed the basis of the many linkage studies performed in the pre-GWAS era and remain valuable resources in the study of many complex traits, including GWAS.

This review will provide a brief overview of the methodology used in family-based association studies followed by an examination of the advantages and disadvantages of family-based design compared to population-based designs. Finally, we will review the published GWAS that use family-based data, which were identified from a catalog of published GWAS [101]. The summary of the published family-based GWAS was not intended to be comprehensive, but it is hoped to capture some of the interesting roles of family data in published GWAS.

Analysis of family-based association data

There are several possible family-based designs, ranging from simple cases of parent–offspring trios to large multigenerational pedigrees. Different methods have been developed to perform an association analysis on these family-based data [6,7]. While most methods use the transmission of allele within informative families (families with at least one heterozygous parent) to assess the evidence for genetic association [6], a number of methods were developed to analyze all available data (e.g., Abecasis's 'total' association test [8]). Here, we will provide

10.2217/14622416.10.2.181 © 2009 Future Medicine Ltd ISSN 1462-2416

a brief overview of a simple family-based association design in order to demonstrate the sources of information on genetic association within these designs and which of these components are protected against population stratification. A more detailed discussion of the methodology for family-based association tests in more general pedigrees is beyond the scope of this review and can be found elsewhere (e.g., [7]).

To illustrate the use of allelic transmissions within families, consider a simple family-based design for detecting a genetic association to a disease, the parent–offspring trio design, which consists of families with an affected offspring and both parents. The genotype–phenotype association in this design is tested using the Transmission Disequilibrium Test (TDT) [9]. This test was originally designed to detect linkage in the presence of association [9], but since it requires the presence of both linkage and association in order to be significant, it is now typically used as a test for association [1]. The requirement for the presence of both linkage and association is one of the biggest advantages of this association test as the presence of linkage makes the result robust against population stratification and admixture, both of which can cause potential false-positive associations [10].

In the TDT, an association between a marker and disease is tested by comparing the number of transmitted with nontransmitted alleles from heterozygous parents to the affected offspring. Any deviation from the 1:1 ratio expected from Mendel's laws suggests an association between the allele and disease. To illustrate this, consider the family in Figure 1A. In this trio, the offspring can have two possible genotypes depending on the allele inherited from its mother, AA and Aa. Under the null hypothesis, both genotypes are expected to occur with 50% probability. Similarly, in Figure 1B the offspring can have three possible genotypes, AA, Aa and aa, with probabilities of 25%, 50% and 25%, respectively. If in a collection of such families, one allele, say the A allele, is preferentially transmitted to the offspring and thus creating a significant deviation from the expected genotype probabilities, then the locus is said to be linked to the disease of interest. Note that in the family represented in Figure 1C, the offspring always has genotype AA so this family does not contribute to the test. It follows that only families with at least one heterozygous parent can contribute to the test for association. In other words, there is a requirement for potential genotype variation within a family and thus this test is called a 'within-family' test for association.

Extensions to the methodology for the analysis of family-based association analysis allow the analysis of general pedigree structures [8,11], quantitative traits [12,13] and multiple phenotypes [14]. Tests of association using family-based samples have also been extended to include additional association information from across families [8]. As its name suggests, a within-family association test only uses information from the allelic transmissions within families. Additional information on genetic association is available through a comparison of allele frequencies across families and testing for its association with the trait of interest [15]. With carefully constructed test statistics, it is possible to separate the total association data into independent within-family and between-family components [15]. Ethnicity or population substructure does not vary within families, but they can vary between families. Thus, while the total (combined between- and within-family) association is more powerful, only the within-family component is robust to the effects of population stratification. Furthermore, the independence of the between- and within-family tests for association makes it possible to explicitly test for the effects of population stratification by comparing the estimated allelic effect on the trait of interest from the two tests [15].

The genome-wide approach to association studies provide a good coverage of the genome, but at the same time it creates a multiple testing problem due to hundreds of thousands of statistical tests performed. The question on what is the association test p-value to be declared significant may be answered by a Bonferroni correction, false-discovery rate or other methods [16]. For example, the Wellcome Trust Case Control Consortium declared that association test p-value of 5×10^{-7} to be significant [17].

Advantages & disadvantages of family-based design in genome-wide association studies

While it is generally accepted that association analysis using unrelated individuals is more powerful than using related individuals [18,19], there are several advantages that family-based designs have to offer. Primary among the advantages of family-based association studies is the robustness of the design to the effects of population stratification or structure as discussed above. It is well known that population-based genetic association analyses are subject to spurious associations caused by factors such as ethnicity, admixture and population stratification [20].

With dense genome-wide SNP data, it becomes possible to detect and remove individuals with admixed ancestries (e.g., through the use of multidimensional scaling as in [17]). Such methods can even detect subtle population differences between countries in Europe [21,22]. In order to achieve the large sample sizes necessary to have sufficient power for a GWAS, it is often necessary to accept a small amount of subtle stratification. However, even small amounts of stratification can lead to false positives, and care must be taken in the analysis to avoid this when not using within-family association tests. An example of the effect of subtle population stratification in an association study is provided by Campbell *et al.* who found an association between a SNP in the lactase (*LCT*) gene and height in a European American population (a mixture of populations derived from different parts of Europe) [23]. Later they discovered that the apparent association between the *LCT* gene SNP and height was due to the fact that the *LCT* gene SNP and height were correlated with grandparental ancestry along an approximately northwestern–southwestern axis in Europe [23].

Family-based designs offer a more thorough genotype quality control mechanism, especially with respect to the detection of Mendelian errors. Genotyping errors can be detected by noting inconsistencies between a parent and his/her offspring's genotype, providing a direct estimate of genotyping error rate. As SNPs generally have only two alleles, the proportion of genotyping errors detected is in the range of 25% [24] for a parent–offspring pair, but the detectable proportion of genotyping errors is increased with additional relatives and the examination of genotypes that would force unlikely recombination events within the family. Apart from the removal of incorrect genotypes, the use of Mendelian inconsistencies also allows the detection, and possible resolution, of sample mix-ups.

A further advantage of family-based design is the possibility of genotyping a subset of individuals within families, but including the phenotypes and imputed genotype probabilities of the ungenotyped relatives in the total association analysis. Chen and Abecasis have shown that for the same number of genotyped individuals, the total association test in related individuals that includes ungenotyped relatives is much more powerful than an association test using unrelated individuals [25]. This approach is particularly advantageous in a study where, the genotyping budget is limited to genotype, only a subset of available samples and marker data from previous linkage analysis is available across individuals that were not genotyped for genome-wide association [25].

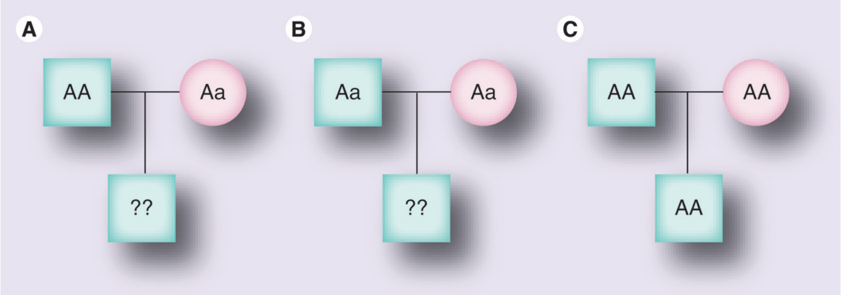

Figure 1. Example of a simple family-based study design demonstrating the source of within- and between-family information in family-based association studies. In families **(A)** and **(B)** the genotype of the offspring is not completely determined by the parental genotypes. Across a collection of such families, potential variation in offspring genotype is used for the within-family association test. The genotype of the offspring in family **(C)** is completely determined by the parental genotypes and so does not contribute to the within-family test, but it can still contribute to between-family and total tests of association.

Family-based designs offer a variety of genetic analyses that cannot be performed using a sample of unrelated individuals. By using family-based designs, we can test for the effect of imprinted genes on phenotypes [26]. Some studies have shown that the parental origin of genotypes has an effect on the phenotypic expression of complex traits (e.g., [27]), although these parent-of-origin effects may only affect a small proportion of genes [28]. We can also examine whether a particular allele is inherited or *de novo* [29,30]. This is of particular interest when examining the effect of copy-number variants where *de novo* variants appear to occur with a greater frequency. For example, Sebat *et al.* have demonstrated that *de novo* variants were sugnificantly associated with autism [30]. Another use of family-based data is the possibility to perform combined linkage and association analysis. This type of analysis tests whether a linkage between disease locus (generated from linkage analysis) and disease can be explained by association of candidate SNP. Combined linkage and association analysis can be useful for fine mapping [31] and for testing locus heterogeneity in the population [32].

As with any other design, family-based association designs also have their disadvantages. Compared to population-based design, which uses a sample of unrelated individuals, a notable disadvantage of family-based design is that it has less power per genotype. Theoretical and simulation studies have shown that designs based on affected and unaffected sibs have less power than designs using unrelated individuals as controls [18,19]. In the absence of population stratification, the loss of power by using the TDT (within-family test) method can be substantial [33]. Chen and Abecasis noted that the power loss in family-based design is due to the fact that the same marker was used for both testing the association and to guard against population stratification [25]. This loss in power can be substantial when considering only the within-family test. However, if the total association is being considered, the loss of power due to the relatedness of individuals in family-based designs is small [34]. Also, the power can be increased by including sibships with multiple affected sibs [18,19] and ungenotyped relatives [35].

Other disadvantages of family-based designs include their sensitivity of the results of their analyses to genotyping error [36,37], although this is somewhat circumvented due to the increased genotype error checking afforded by the related individuals. The analysis of family-based association studies is also computationally more demanding compared to that of a sample of unrelated individuals, thus requiring specialized software. Finally, family-based designs that require parental information, such as the parent–offspring trio design may not be practical for late-onset diseases. However, this can be overcome by using sib design rather than parent–offspring design [38].

Statistical packages for the analysis of family-based association data

There are several publicly available statistical software packages that can be used to perform family-based association analysis (Table 1). While some were developed specifically for family-based designs (e.g., family-based association testing [FBAT] [39], GHOST [25], PBAT [40], quantitative tansmission disequilibrium tests [QTDT] [8] and TRANSMIT [41]), others were developed for general association analyses, but provide support for some family-based designs (e.g., PLINK [42], UNPHASED [43] and whole-genome association pipeline [WASP]). Also, several statistical genetics packages that are primarily aimed toward other genetics analyses (e.g., linkage, linkage disequilibrium or haplotype block analysis) can also be used to perform family-based association analysis, including FAMHAP [44], HAPLOVIEW [45], multipoint engine for rapid likelihood inference (MERLIN) [46] and SIB-PAIR [47].

Published genome-wide association studies using family data

Prior to August 2008, there were 173 published GWAS [101] with at least 36 of them (including the 17 published GWAS from the Framingham Heart Study [48]) using a family-based study design either during the initial screening or the replication stage. A list of published GWAS that used family data (plus a summary of 17 GWAS from the Framingham Heart Study series [48]) is presented in Table 2.

It can be seen from Table 2 that only one published GWAS used a family-based design for both the initial screening and replication samples [49]. The Framingham Heart Study series, which used family data in the initial screening stage, did not feature a replication stage in any of its publications. Most studies used a combination of family-based and case–control designs, with the use of population-based design for the initial screening sample and family-based design as the replication sample being favored (e.g., [50–52]). This approach is attractive as the higher power of population-based samples compared to the family-based samples is particularly advantageous during the screening of large numbers of SNPs due to the rigors of multiple testing. Using a family-based design for the replication stage removes the potential for significant associations to be caused by population stratification and thus any replicated genotype–phenotype associations are more likely to be genuine.

Another point to be taken from the published family-based association studies is that not all use the within-family test (TDT test). For example, Scuteri *et al.* used the total association (using all observed/estimated genotypes) for testing the association between SNPs and obesity related traits [49]. While this test uses more data and thus is more powerful, it is not protected against the effects of population stratification. As false-positives caused by population stratification are a primary concern, they then applied genomic control to adjust for the effects of population stratification.

The series of papers published from the Framingham Heart Study [48] form the largest family-based association study published in terms of the number of phenotypes analyzed [53–69]. A total of 17 phenotype groups were examined, ranging from obesity to cancers, but mostly related to cardiovascular diseases, with each published in a separate paper. By genotyping 1345 individuals from 310 families using 100K Affymetrix GeneChips®, these studies demonstrate the use of large phenotypic and

Table 1. List of publicly available software that support family-based association analyses.

Study	Software	Analysis options	Website	Ref.
Abecasis *et al.*	MERLIN	Total association analysis (not robust to population stratification)	www.sph.umich.edu/csg/abecasis/merlin/	[25,46]
Abecasis *et al.*	QTDT	Family-based association analysis, including total- and within-family association Combined association and linkage analyses	www.sph.umich.edu/csg/abecasis/QTDT/index.html	[8]
Barrett *et al.*	HAPLOVIEW	Single SNP and haplotype association tests (including TDT) Linkage disequilibrium and haplotype block analyses Visualization and plotting of PLINK GWAS results	www.broad.mit.edu/node/443	[45]
Becker *et al.*	FAMHAP	Association analysis of nuclear family data Test for imprinting in nuclear family data GWAS and genotype imputation Haplotype association analysis	http://famhap.meb.uni-bonn.de/	[44]
Chen *et al.*	GHOST	Designed for GWAS Association analysis of family data Can handle large pedigree and infer missing genotypes	www.sph.umich.edu/csg/chen/ghost/	[25]
Clayton *et al.*	TRANSMIT	TDT analysis Marker haplotypes based on several closely linked markers	www-gene.cimr.cam.ac.uk/clayton/software/ (no longer maintained)	[41]
Dudbridge	UNPHASED	Analysis of nuclear families and unrelated subjects, and combinations of the two Analysis of discrete or quantitative traits Maximum likelihood treatment of missing genotype data and uncertain haplotypes Global association tests and tests of individual haplotypes Conditioning tests that allow for previous associations of linked loci Inclusion of information from additional tag markers Support for nongenetic covariates including parent-of-origin Permutation tests allowing for multiple testing	www.mrc-bsu.cam.ac.uk/personal/frank/software/unphased/	[43]
Duffy	SIB-PAIR	Various basic genetic analyses Allelic association with a binary or quantitative trait Combined association and linkage analyses Imputation of genotypes	www.qimr.edu.au/davidD/sib-pair.html	[102]
Laird *et al.*	FBAT	A variety of family-based association analyses Association tests on sex-linked X-chromosome markers	www.biostat.harvard.edu/~fbat/default.html	[39]
Lange *et al.*	MENDEL	Haplotypes estimation Allelic association using TDT or gamete competition model Association analysis on quantitative traits Combined association and linkage analyses	www.genetics.ucla.edu/software/mendel	[73]
Lange *et al.*	PBAT	A variety of family-based association analyses, including for univariate and multivariate data, gene/covariate interaction and time to onset/survival data	www.biostat.harvard.edu/~clange/default.htm	[40]
Purcell *et al.*	PLINK	Designed for GWAS Summary statistics for data quality control Various association test, including family-based association test (TDT, sibships test) Multimarker/haplotypic tests Joint SNPs and CNVs association tests Epistasis, gene–environment analyses	http://pngu.mgh.harvard.edu/~purcell/plink/	[42]
Vanderbilt University	WASP	Designed for GWAS Summary statistics for data quality control Association analyses (TDT and case–control) GUI data plotter	http://chgr.mc.vanderbilt.edu/wasp/	–

CNV: Copy number variant; GUI: Graphical user interface; GWAS: Genome-wide association studies; TDT: Transmission disequilibrium test.

genetic collections previously used in genetic linkage studies are still valuable resources in the era of GWAS. Since these papers were aimed as an initial resource for future replication studies or meta-analysis, there was no attempt to replicate the findings in independent sample(s).

Table 2. List of published family-based genome-wide association studies.

Study	Disease/trait	Initial sample	Replication sample	Statistical methods for family data	Software	Ref.
Barrett *et al.*	Crohn's disease	3230 cases, 4829 controls	2325 cases, 1809 controls 1339 affected trios	TDT in trios of replication sample	Not specified	[76]
Cupples *et al.*	987 phenotypes related to cardiovascular diseases	1345 individuals from 310 families (Framingham Heart Study)	Not available	Total (GEE) and within family associations	FBAT and R	[48]
Duerr *et al.*	Inflammatory bowel disease	547 cases, 548 controls	401 cases, 433 controls, 883 families, 1119 affected offspring	TDT in replication sample	FBAT and HAPLOVIEW	[80]
Florez *et al.*	Type 2 diabetes and six quantitative traits	1087 family members	1465 unrelated individuals; 2175 cases and 2412 controls	Total (GEE) and within family associations	FBAT	[87]
Franke *et al.*	Irritable bowel syndrome	393 cases, 399 controls	2920 cases, 1961 controls, 1248 trios	TDT in replication sample	HAPLOVIEW	[81]
Graham *et al.*	Systemic lupus erythematosus	431 cases, 2155 controls	740 trios	TDT in replication sample	PLINK	[50]
Hafler *et al.*	Multiple sclerosis	931 trios, 2431 controls	609 trios, 2322 cases, 2987 controls	TDT	PLINK, UNPHASED and WASP	[82]
Hakonarson *et al.*	Type 1 diabetes	561 cases, 1143 controls, 467 trios	1333 individuals in 549 families; 390 trios	TDT in trios	HAPLOVIEW	[84]
Hakonarson *et al.*	Type 1 diabetes	563 cases, 1146 controls, 483 trios	549 families, 1092 individuals from 364 trios	TDT in trios	FBAT	[85]
Herbert *et al.*	Obesity	694 offspring	3489 cases, 6392 controls, 361 trios	Two-stage procedures in PBAT using the same sample: use parental genotype to select SNPs and genetic model that best predict offspring' phenotype; use FBAT to test the association between selected SNPs and phenotype	PBAT	[83]
Hinney *et al.*	Early onset extreme obesity	487 young cases, 442 controls	2269 individuals in 644 families	Family-based association testing (FBAT additive) in replication sample	Not specified	[51]
Kirov *et al.*	Schizophrenia	574 cases, 605 controls, 1148 parents of cases	Not available	TDT in DNA pooling experiment, followed by TDT in individual genotypes	Not specified	[70]
Libioulle *et al.*	Crohn's disease	547 cases, 928 controls	1266 cases, 559 controls, 428 trios	TDT in trios of replication sample	Not specified	[77]
Moffatt *et al.*	Childhood asthma, *ORMDL3* expression	994 cases, 1243 controls (including TDT analysis in 378 children and 405 parents)	2320 cases, 3301 controls	TDT in a subset of individuals in initial sample	TRANSMIT	[75]
Poduslo *et al.*	Alzheimer's disease	29 siblings from two affected families, 60 unrelated controls	199 patients, 85 spouses	Allelic and haplotype associations	HAPLOVIEW and HELIXTREE	[74]
Raelson *et al.*	Crohn's disease	382 trios	521 trios, 750 cases, 828 controls	For trios with affected parents, spousal chromosomes were used as controls. For child-affected trios, parental nontransmitted chromosomes were used as controls	Not specified	[78]

GEE: Generalized estimating equation; GLIMM: Generalized linear mixed models; TDT: Transmission disequilibrium test. Table 2 is modified from [101].

Table 2. List of published family-based genome-wide association studies.

Study	Disease/trait	Initial sample	Replication sample	Statistical methods for family data	Software	Ref.
Rioux *et al.*	Crohn's disease	946 cases, 977 controls	530 trios, 353 cases, 207 controls	Not specified	PLINK and WHAP	[79]
Scuteri *et al.*	Obesity-related traits: (BMI, hip circumference, body weight)	4741 individuals from large families	3205 individuals from families	Total association (use all observed/estimated genotypes) and use genomic control to adjust for population stratification	Not specified	[49]
Todd *et al.*	Type 1 diabetes	1963 cases, 2938 controls	4000 cases, 5000 controls, 2997 trios	TDT in replication sample	R or STATA	[86]
Wallace *et al.*	Biochemical markers for cardiovascular disease	1955 unrelated hypertensive individuals	2033 individuals in 519 families; 1461 twins (1/pair selected randomly)	GLMM	WINBUGS	[52]

GEE: Generalized estimating equation; GLIMM: Generalized linear mixed models; TDT: Transmission disequilibrium test. Table 2 is modified from [101].

Family-based designs have not only been used for studies where all individuals are genotyped, but also in DNA pooling experiments. Kirov *et al.* used parent–offspring trio design in DNA pooling experiments to identify genetic variants affecting schizophrenia [70]. Although the difference between parents and offspring is roughly half of that of cases and controls, resulting in reduced power, this approach provided a protection against false positives due to potential population stratification. While no SNP in that particular study reached genome-wide significance, the use of pooling is an attractive option for studies with a limited budget and its use with family-based data warrants further investigation.

Future perspective

A summary of published GWAS shows that population-based design is currently favored as a design of choice for identifying genes for common complex diseases [101]. However, family-based designs are regularly used in replication studies and this is likely to become more prevalent in the future. Published GWAS have revealed that, for most diseases, the identified genetic variants explain only small proportion of genetic variance. Larger sample sizes will be needed in order to identify the remaining genetic variants. As the sample sizes of a population-based study are increased, so is the probability of false positives resulting from the effect of population stratification [71]. Genomic control, which uses the genotypes throughout the genome to determine an appropriate correction for population stratification, has been suggested as a solution for eliminating potential false positives in population-based studies [71]. However, the use of genomic control is only appropriate in situations where a large number of markers have been genotyped across all individuals. This is not the case in replication studies where typically only a few markers are examined. Thus, genomic control only serves as a complement method to a family-based design for correcting the association between genotypes and phenotypes for the effect of population stratification [72]. Therefore, for any GWAS based on population samples, it would be preferable to have a family-based design for the replication samples to ensure the replicated associations are genuine. Family-based designs can also be successfully used during the primary stages of GWAS, as shown by the Framingham Heart Study. As the price of large-scale genotyping continues to decrease, it is likely that many of the other large resources collected for linkage studies of complex disease will become targets for GWAS.

Acknowledgments

We thank David Duffy of the Queensland Institute of Medical Research for his suggestion related to the software used for family-based association analysis.

Financial & competing interests disclosure

AFM and PMV are funded by the National Health and Medical Research Council, Australia. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Family-based genome-wide association studies

- Family-based genome-wide association studies (GWAS) are aimed to identify genetic variants associated with complex diseases/phenotypes using samples of related individuals genotyped with a very large number of genetic markers (e.g., SNPs).

Advantages of family-based compared to population-based designs

- Can make use of the existing family-based (linkage) data collected in pre-GWAS era.
- Robust to the possible effects of population stratification.
- The ability to perform a thorough genotyping quality control (e.g., Mendelian inheritance error and sample mix-ups).
- Provide a platform to perform additional genetic analyses, such as testing for the effect of imprinted genes on phenotypes, testing whether an allele is inherited or *de novo* and performing combined linkage and association analyses.

Disadvantages of family-based compared to population-based designs

- Less power.
- Sensitive to genotyping error.
- Computationally more demanding and requires specialized software.
- May not be practical for late-onset diseases.

Published family-based genome-wide association studies

- Family-based designs are frequently used during the replication stage.

Available software for analysis

- A variety of software for analyzing family-based GWAS data is available.

Bibliography

Papers of special note have been highlighted as:

▪ of interest

▪▪ of considerable interest

1 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108 (2005).

▪ **Good introduction to the genome-wide association studies.**

2 Sachidanandam R, Weissman D, Schmidt SC *et al.*: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933 (2001).

3 International HapMap Consortium: A haplotype map of the human genome. *Nature* 437, 1299–1320 (2005).

4 Frazer KA, Ballinger DG, Cox DR *et al.*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861 (2007).

5 Boomsma D, Busjahn A, Peltonen L: Classical twin studies and beyond. *Nature* 3, 872–882 (2002).

6 Ewens WJ, Li M, Spielman RS: A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. *PLoS Genet.* 4(9), E1000180 (2008).

7 Laird NM, Lange C: Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* 7, 385–394 (2006).

▪▪ **Comprehensive review on statistical methods for family-based genome-wide association studies.**

8 Abecasis GR, Cardon LR, Cookson WO: A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* 66, 279–292 (2000).

9 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.* 52, 506–516 (1993).

▪ **First study to demonstrate the application of the transmission disequilibrium (TDT) for gene mapping.**

10 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat. Genet.* 36(5) 512–517 (2004).

11 Lange C, DeMeo DL, Laird NM: Power and design considerations for a general class of family-based association tests: quantitative traits. *Am. J. Hum. Genet.* 71, 1330–1341 (2002).

12 Allison DB: Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* 60, 676–690 (1997).

13 Rabinowitz D: A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* 47, 342–350 (1997).

14 Lange C, Silverman EK, Xu X, Weiss ST, Laird NM: A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 4, 195–206 (2003).

15 Fulker DW, Cherny SS, Sham PC, Hewitt JK: Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* 64, 259–267 (1999).

16 Ziegler A, Konig IR, Thompson JR: Biostatistical aspects of genome-wide association studies. *Biom. J.* 50, 8–28 (2008).

17 Consortium WTCC: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678 (2007).

18 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.* 8, 1273–1288 (1998).

19 Teng J, Risch N: The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* 9, 234–241 (1999).

20 Lander ES, Schork NJ: Genetic dissection of complex traits. *Science* 265, 2037–2048 (1994).

21 Lao O, Lu TT, Nothnagel M *et al.*: Correlation between genetic and geographic structure in Europe. *Curr. Biol.* 18, 1241–1248 (2008).

22 Tian C, Plenge RM, Ransom M *et al.*: Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet.* 4(1), E4 (2008).

23 Campbell CD, Ogburn EL, Lunetta KL *et al.*: Demonstrating stratification in a European American population. *Nat. Genet.* 37, 868–872 (2005).

▪▪ **Example of the effects of population stratification on association studies.**

24 Gordon D, Heath SC, Ott J: True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum. Hered.* 49, 65–70 (1999).

25 Chen WM, Abecasis GR: Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* 81, 913–926 (2007).

26 Becker T, Baur MP, Knapp M: Detection of parent-of-origin effects in nuclear families using haplotype analysis. *Hum. Hered.* 62, 64–76 (2006).

27 Hawi Z, Segurado R, Conroy J *et al.*: Preferential transmission of paternal alleles at risk genes in attention-deficit/hyperactivity disorder. *Am. J. Hum. Genet.* 77, 958–965 (2005).

28 Wilkinson LS, Davies W, Isles AR: Genomic imprinting effects on brain development and function. *Nat. Rev. Neurosci.* 8, 832–843 (2007).

29 Mefford HC, Sharp AJ, Baker C *et al.*: Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* 359(16), 1685–1699 (2008).

30 Sebat J, Lakshmi B, Malhotra D *et al.*: Strong association of *de novo* copy number mutations with autism. *Science* 316, 445–449 (2007).

31 Das SK, Hasstedt SJ, Zhang Z, Elbein SC: Linkage and association mapping of a chromosome 1q21-q24 Type 2 diabetes susceptibility locus in northern European Caucasians. *Diabetes* 53, 492–499 (2004).

32 Gordon S, Visscher PM: Residual linkage: why do linkage peaks not disappear after an association study? *Hum. Genet.* 121, 77–82 (2007).

33 Hernandez-Sanchez J, Haley CS, Visscher PM: Power of QTL detection using association tests with family controls. *Eur. J. Hum. Genet.* 11, 819–827 (2003).

34 Visscher PM, Andrew T, Nyholt DR: Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *Eur. J. Hum. Genet.* 16, 387–390 (2008).

35 Visscher PM, Duffy DL: The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. *Genet. Epidemiol.* 30, 30–36 (2006).

36 Gordon D, Heath SC, Liu X, Ott J: A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am. J. Hum. Genet.* 69, 371–380 (2001).

37 Gordon D, Ott J: Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac. Symp. Biocomput.* 18–29 (2001).

38 Spielman RS, Ewens WJ: A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* 62, 450–458 (1998).

39 Laird NM, Horvath S, Xu X: Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* 19(Suppl. 1), S36–S42 (2000).

40 Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM: PBAT: tools for family-based association studies. *Am. J. Hum. Genet.* 74, 367–369 (2004).

41 Clayton D: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am. J. Hum. Genet.* 65, 1170–1177 (1999).

42 Purcell S, Neale B, Todd-Brown K *et al.*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).

43 Dudbridge F: Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.* 66, 87–98 (2008).

44 Becker T, Knapp M: Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet. Epidemiol.* 27, 21–32 (2004).

45 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265 (2005).

46 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30, 97–101 (2002).

47 Duffy DL: Sib-pair: a program for non-parametric linkage/association analysis. *Am. J. Hum. Genet.* 61, 1140 (1997).

48 Cupples LA, Arruda HT, Benjamin EJ *et al.*: The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Med. Genet.* 8(Suppl. 1), S1 (2007).

▪ **Introductory paper to the series of 17 papers on family-based genome-wide analysis studies from the Framingham Heart Study.**

49 Scuteri A, Sanna S, Chen WM *et al.*: Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLoS Genet.* 3, E115 (2007).

▪ **Example of family-based genome-wide anaylsis studies that uses the total association test and corrects for the effects of population stratification using genomic control.**

50 Graham RR, Cotsapas C, Davies L *et al.*: Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat. Genet.* (2008) (Epub ahead of print).

51 Hinney A, Nguyen TT, Scherag A *et al.*: Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (*FTO*) variants. *PLoS ONE* 2, E1361 (2007).

52 Wallace C, Newhouse SJ, Braund P *et al.*: Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am. J. Hum. Genet.* 82, 139–149 (2008).

53 Benjamin EJ, Dupuis J, Larson MG *et al.*: Genome-wide association with select biomarker traits in the Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S11 (2007).

54 Fox CS, Heard-Costa N, Cupples LA, Dupuis J, Vasan RS, Atwood LD: Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. *BMC Med. Genet.* 8(Suppl. 1), S18 (2007).

55 Gottlieb DJ, O'Connor GT, Wilk JB: Genome-wide association of sleep and circadian phenotypes. *BMC Med. Genet.* 8(Suppl. 1), S9 (2007).

56 Hwang SJ, Yang Q, Meigs JB, Pearce EN, Fox CS: A genome-wide association for kidney function and endocrine-related traits in the NHLBI's Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S10 (2007).

57 Kathiresan S, Manning AK, Demissie S *et al.*: A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S17 (2007).

58 Kiel DP, Demissie S, Dupuis J, Lunetta KL, Murabito JM, Karasik D: Genome-wide association with bone mass and geometry in the Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S14 (2007).

59 Larson MG, Atwood LD, Benjamin EJ *et al.*: Framingham Heart Study 100K project: genome-wide associations for cardiovascular disease outcomes. *BMC Med. Genet.* 8(Suppl. 1), S5 (2007).

60 Levy D, Larson MG, Benjamin EJ *et al.*: Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Med. Genet.* 8(Suppl. 1), S3 (2007).

61 Lunetta KL, D'Agostino RB Sr, Karasik D *et al.*: Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study. *BMC Med. Genet.* 8(Suppl. 1), S13 (2007).

62 Meigs JB, Manning AK, Fox CS *et al.*: Genome-wide association with diabetes-related traits in the Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S16 (2007).

63 Murabito JM, Rosenberg CL, Finger D *et al.*: A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S6 (2007).

64 Newton-Cheh C, Guo CY, Wang TJ, O'Donnell CJ, Levy D, Larson MG: Genome-wide association study of electrocardiographic and heart rate variability traits: the Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S7 (2007).

65 O'Donnell CJ, Cupples LA, D'Agostino RB *et al.*: Genome-wide association study for subclinical atherosclerosis in major arterial territories in the NHLBI's Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S4 (2007).

66 Seshadri S, DeStefano AL, Au R *et al.*: Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham Study. *BMC Med. Genet.* 8(Suppl. 1), S15 (2007).

67 Vasan RS, Larson MG, Aragam J *et al.*: Genome-wide association of echocardiographic dimensions, brachial artery endothelial function and treadmill exercise responses in the Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S2 (2007).

68 Wilk JB, Walter RE, Laramie JM, Gottlieb DJ, O'Connor GT: Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC Med. Genet.* 8(Suppl. 1), S8 (2007).

69 Yang Q, Kathiresan S, Lin JP, Tofler GH, O'Donnell CJ: Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. *BMC Med. Genet.* 8(Suppl. 1), S12 (2007).

70 Kirov G, Zaharieva I, Georgieva L *et al.*: A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Mol. Psychiatry* (2008) (Epub ahead of print).

71 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 55, 997–1004 (1999).

72 Bacanu SA, Devlin B, Roeder K: The power of genomic control. *Am. J. Hum. Genet.* 66, 1933–1944 (2000).

73 Lange K, Sinsheimer JS, Sobel E: Association testing with Mendel. *Genet. Epidemiol.* 29, 36–50 (2005).

74 Poduslo SE, Huang R, Huang J, Smith S: Genome screen of late-onset alzheimer's extended pedigrees identifies *TRPC4AP* by haplotype analysis. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* (2008) (Epub ahead of print).

75 Moffatt MF, Kabesch M, Liang L *et al.*: Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* 448, 470–473 (2007).

76 Barrett JC, Hansoul S, Nicolae DL *et al.*: Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 40, 955–962 (2008).

77 Libioulle C, Louis E, Hansoul S *et al.*: Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of *PTGER4*. *PLoS Genet.* 3(4), E58 (2007).

78 Raelson JV, Little RD, Ruether A *et al.*: Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc. Natl Acad. Sci. USA* 104, 14747–14752 (2007).

79 Rioux JD, Xavier RJ, Taylor KD *et al.*: Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* 39, 596–604 (2007).

80 Duerr RH, Taylor KD, Brant SR *et al.*: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314(5804), 1461–1463 (2006).

81 Franke A, Hampe J, Rosenstiel P *et al.*: Systematic association mapping identifies *NELL1* as a novel IBD disease gene. *PLoS ONE* 2(1), E691 (2007).

82 Hafler DA, Compston A, Sawcer S *et al.*: Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J Med.* 357, 851–862 (2007).

83 Herbert A, Gerry NP, McQueen MB *et al.*: A common genetic variant is associated with adult and childhood obesity. *Science* 312, 279–283 (2006).

84 Hakonarson H, Grant SF, Bradfield JP *et al.*: A genome-wide association study identifies KIAA0350 as a Type 1 diabetes gene. *Nature* 448, 591–594 (2007).

85 Hakonarson H, Qu HQ, Bradfield JP *et al.*: A novel susceptibility locus for Type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* 57, 1143–1146 (2008).

86 Todd JA, Walker NM, Cooper JD *et al.*: Robust associations of four new chromosome regions from genome-wide analyses of Type 1 diabetes. *Nat. Genet.* 39, 857–864 (2007).

87 Florez JC, Manning AK, Dupuis J *et al.*: A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets. *Diabetes* 56, 3063–3074 (2007).

Websites

101 Hindorff L, HA Junkins, TA Manolio, A Catalog of Published Genome-Wide Association Studies. National Human Genome Research Institute (2008). www.genome.gov/26525384.

102 Duffy DL: SIB-PAIR 1.00b: A program for elementary genetical analyses (2008). www2.qimr.edu.au/davidD/sib-pair.html