Published in final edited form as:
*Genet Epidemiol*. 2007 ; 31(Suppl 1): S118–S123. doi:10.1002/gepi.20288.

# Multistage Designs in the Genomic Era: Providing Balance in Complex Disease Studies

**Marie-Pierre Dubé**[1], **Silke Schmidt**[2], and **Elizabeth Hauser [on behalf of on behalf of Group 14]**[2,*,†]

1*Research Centre of the Montreal Heart Institute, Department of Statistical Genetics, Montreal, Quebec and Department of Medicine, Université de Montréal, 2900, Edouard-Montpetit Blvd., Montreal, Quebec H3T 1J4, Canada*

2*Center for Human Genetics, Department of Medicine, Duke University Medical Center, Durham, North Carolina*

## Abstract

In this summary paper, we describe the contributions included in the Multistage Design group (Group 14) at the Genetic Analysis Workshop 15, which was held during November 12-14, 2006. Our group contrasted and compared different approaches to reducing complexity in a genetic study through implementation of staged designs. Most groups used the simulated dataset (problem 3), which provided ample opportunities for evaluating various staged designs. A wide range of multistage designs that targeted different aspects of complexity were explored. We categorized these approaches as reducing phenotypic complexity, model complexity, analytic complexity or genetic complexity. In general we learned that: (1) when staged designs are carefully planned and implemented, the power loss compared to a single-stage analysis can be minimized and study cost is greatly reduced; (2) a joint analysis of the results from each stage is generally more powerful than treating the second stage as a replication analysis.

### Keywords

two-stage study design; replication; joint analysis; statistical power; genetic association

## INTRODUCTION

The rapid progress of genomic technology has been accompanied by many statistical challenges. Datasets for genetic studies have grown remarkably in terms of the number of genetic locations tested and the number of phenotyped and genotyped subjects. This growth has caused both increased study costs and increased complexity of the statistical analysis. The focus of genetic studies remains the identification of DNA variants conferring increased susceptibility to disease, but these studies are also expected to tackle multiple additional challenges, such as analysis of disease-associated clinical phenotypes and environmental covariates, identification of population subgroups, and evaluation of gene-gene and gene-environment interactions. We have not yet reached the apex of this data expansion. The new mega single-nucleotide polymorphism (SNP) chips will soon be released, whole genome sequencing is making rapid progress, the P3G Consortium is facilitating harmonization of

*Correspondence to: Elizabeth R. Hauser, Center for Human Genetics, Duke University Medical Center, DUMC Box 3445, Durham, NC 27710, USA. E-mail: Elizabeth.Hauser@duke.edu
†A complete list of Group 14 primary contributing authors is given in the Acknowledgment section.

genetic data collections (http://www.p3gconsortium.org/), and the NIH is strongly encouraging efforts to make such data collections publicly available (http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/about.html).

It follows that the statistical-genetic challenges will likely increase in the near future. The primary statistical goal is usually to achieve appropriate power while controlling for multiple testing and accounting for the correlation between tests at nearby markers. The widely discussed difficulty of replicating genetic association findings has cast a long shadow over our field, and the need for validation is more pressing than ever. Reasons for lack of replication include under-powered studies with limited sample size and the population-specific nature of effects, which can themselves be genetic, environmental, or both, with the additional difficulty of discerning between causal versus indirect association findings due to population-specific linkage disequilibrium (LD) patterns [Alcais et al., 2007]. Taken together, these factors have generated many ambiguous follow-up studies and sometimes “ad hoc” staged analyses. Properly planned and implemented multistage study designs have the potential to reduce the considerable complexity of genetic-epidemiologic datasets, and thus it is not surprising that for the first time a group examining multistage designs was formed at Genetic Analysis Workshop 15 (GAW15).

The GAW15 datasets presented a variety of opportunities to tackle complexity with multistaged approaches. Problems 2 (RA data) and 3 (simulated data) included a large and diverse sample that allowed the evaluation of linkage methods, family-based and case-control association designs and their combination, using genome-wide microsatellite and SNP marker maps and a region of dense SNP genotyping on chromosome 6. Problem 1 expanded the problem to include a combination of both a dense genome wide scan on a large number of samples combined with gene expression data on the same subjects.

## STAGED DESIGNS IN GENETICS

The use of staged designs in genetic epidemiology is not unique to the current genomic era. Morton [1955] proposed a sequential test for linkage in his 1955 publication that was based on Wald's sequential probability ratio test. In the 1990s, Whittemore and Halpern [1997] advocated staged designs as cost-saving sampling strategies in genetic epidemiology. Multistage sampling designs entail a number of interesting statistical issues that also arise in the context of survey studies.

More recently, staged designs have been proposed with the primary aim of reducing the genotyping cost in linkage [Craddock et al., 1996; Holmans and Craddock, 1997] and association studies [Fulker et al., 1999; Satagopan and Elston, 2003; Thomas et al., 2004; Marchini et al., 2005; Van Steen et al., 2005; Rosenberg et al., 2006; Skol et al., 2006; Wang et al., 2006]. A popular approach is to divide the entire available (or anticipated) dataset into a test sample and a validation (replication) sample, in which only a subset of the original genetic tests will be performed. The subset of SNPs can be selected by single-locus association *P*-values in the test sample or by identifying SNPs that capture the distribution of neighboring SNPs via correlation. Both approaches reduce the number of tests to be performed at the second stage. An alternative use of a staged design is to identify a genomic region of interest through linkage analysis and both validate and refine this finding by case-control or family-based association analysis. Finally, two-stage strategies have also been proposed in genetic association studies to reduce the penalty due to multiple testing when modeling gene-gene interactions [Marchini et al., 2005; Ionita and Man, 2006].

## METHODS

The multistage designs employed by our group covered a wide range of applications and analyses. Most designs aimed to reduce the complexity in the first stage and then used the second stage to evaluate the reduced problem. Of primary importance is the statistical treatment of the first-stage tests as the second-stage analysis must account for the inference made at the first stage. The entire experiment must be viewed as the result of both stages for valid statistical inference. In addition, the desired reduction of complexity at the first stage typically results in a reduction of information extracted from the second stage, which impacts statistical power. The Multistage Designs group presented a variety of approaches to reducing complexity, which we have divided into four categories for clarity. However, some approaches fall into more than one category.

### PHENOTYPIC COMPLEXITY

Some contributions used a staged design to reduce the phenotypic complexity of the data. For example, Yang et al. [personal communication] worked with the gene expression data of problem 1 and aimed to identify phenotypes (traits) with high signal-to-noise ratio to carry into the next stage. Two contributions [Gray-McGuire et al., personal communication; Schmidt et al., 2007] used covariate adjustment models to identify important disease-associated covariates for testing in the next stage. This may improve the homogeneity of the phenotype [Gray-McGuire et al., personal communication] or identify a subset of cases, on the basis of both identity-by-descent sharing and covariate values, which may carry a specific disease susceptibility allele and qare analyzed in detail at the second stage [Schmidt et al., 2007].

### MODEL COMPLEXITY

In applications to real datasets, the underlying genetic model is unknown and may or may not include higher-order terms, such as gene-environment or gene-gene interactions. In a staged design approach to model building, Barhdadi and Dube [2007] selected SNPs in stage one based on the results of single-marker case-control association analysis and then tested a number of SNP-SNP interaction models by logistic regression using two different SNP entry criteria in the second stage. They realized the importance of stringency in selecting the SNPs (model terms) to be included in the second-stage interaction models. Model complexity was also reduced by applying a staged approach to a variety of data types such as expression data, family-based and case-control association data, such that results brought forward from one stage to the next helped inform the inference from the entire experiment [Yang et al., personal communication].

### ANALYTIC COMPLEXITY

One of the most novel applications of a staged design was presented by Wang et al. [2007a], who argued that the choice of the best statistic to use in any given study is highly dependent on the underlying data distribution and that a multistage design can guide the selection of the most powerful statistic for any particular sample. They compared two statistics on a portion of the dataset in the first stage, and then applied the most powerful statistic to the remaining samples at the second stage [Wang et al., 2007a]. As one might expect, the selection of the most powerful statistic in the first stage can substantially increase power compared to applying a single statistic that may or may not be a good fit for the underlying distribution of the data at hand.

### GENETIC COMPLEXITY

The most popular use of staged designs was to perform SNP selection procedures to reduce the number of SNPs tested in the second stage. The reduction was achieved by a number of

different methods. Some contributions selected SNPs by evaluating the correlation or LD relationships between them, or by using a SNP clustering algorithm [Li, 2007;Aschard et al., 2007; Darabi et al., personal communication; Yang et al., personal communication]. Darabi et al. [personal communication] reported that localization may be very poor if too few SNPs were selected for the second stage. Other contributions selected SNPs by single-marker association tests, or multipoint linkage tests, using statistics that were independent between the first (SNP selection) and the second (SNP-phenotype association) stages [Aschard et al., 2007;Barhdadi and Dube, 2007;Rohlfs et al., 2007;Schmidt et al., 2007;Wang et al., 2007b]. Some groups used tests that were not independent between stages [Wang et al., 2007a].

We discussed in some detail the statistical options for combining tests performed at the two stages. It is known that a joint analysis is more powerful than treating the second stage as a stand-alone replication sample [Skol et al., 2006]. Li [2007] reported on the use of a joint analysis assuming heterogeneity as detailed in his paper. He found that a joint analysis of his two-stage design was very close in power to that of a single-stage analysis including all samples and all markers, consistent with the previous report [Skol et al., 2006]. Wang et al. [2007a] compared the efficiency of two methods to combine *P*-values, one based on Fisher's test and the other based on the weighted inverse normal distribution. *P*-values were generated from two different statistics, the nonpara-metric Wilcoxon test and Hotelling's $T^2$ statistic. They found that the difference in power was primarily due to the choice of statistic rather than the choice of method for combining the *P*-values.

The use of multistage procedures usually requires that the dataset be divided in some way. The samples can either be divided into a testing set and a validation set, with or without joint analysis of the respective statistics [Li, 2007;Wang et al., 2007a], or alternatively, the same dataset can be examined from different angles at the two stages. For example, Wang et al. [2007b] selected SNPs by comparing allele frequencies in affected parents versus unaffected parents in stage one and then tested a subset of SNPs by family-based association tests in stage two. Rohlfs et al. [2007] relied on the use of Lange's [2003] two-stage method that uses a conditional means model in the first stage followed by a family-based association test in the second stage. Aschard et al. [2007] selected regions of SNPs based on a case-control analysis and followed up the selected markers in stage two with a novel multi-marker family-based association test. Schmidt et al. [2007] used ordered subset analysis [Hauser et al., 2004] to implement a twofold selection approach in stage one. They selected a subset of SNPs within candidate regions identified by linkage analysis, and also selected a subset of cases for further analysis on the basis of both identity-by-descent sharing and covariate information. Unrelated cases (one per family) were then compared to unrelated controls in a second-stage logistic regression analysis.

## RESULTS

A summary of the published contributions is included in Table I. Despite the diversity of analysis approaches, we were able to identify some common themes across studies. In general, we observed that very liberal selection criteria in the first stage were likely to increase the false positive rate in the second stage [Aschard et al., 2007;Barhdadi and Dube, 2007;Li, 2007; Gray-McGuire et al., personal communication]. On the other hand, selection criteria that were too stringent in the first stage could lead to severely decreased power to detect true-positive SNPs in the second stage [Schmidt et al., 2007]. Clearly, type I and type II error rates can and should be balanced through appropriate selection criteria, and whenever possible, this should be planned prior to performing the data analyses and not be driven by interim results.

We found interesting uses of multistage designs in particular for the selection of a best model, a best test statistic, and for phenotype clarification. Further exploration of these aspects are worth additional study as they may be even more useful than the two-stage designs focused on

SNP reduction and may allow complexity reduction beyond the well-recognized genetic and genomic complexity induced by LD.

The majority of the papers discussed by the Multistage Designs group used the simulated dataset (problem 3). In general, we felt that the simulations provided a wide range of genetic effects to evaluate, however, some care was required in order to make useful comparisons among the different approaches. For example, the genetic effect simulated on chromosome 6 was overwhelmingly large and analyses concentrating only on this genomic region might not have detected variation in performance of different approaches. In contrast, the genetic effect simulated on chromosome 9 was very weak and difficult to detect with any method. The only method that was even modestly successful in detecting this locus was the joint method implemented by Li [2007]. The number of simulated replicates also proved to be somewhat problematic. Our conclusions about estimated type I error rates for any of the methods were hampered by the availability of only 100 replicates. On the other hand, 100 replicates provided a substantial computational challenge for groups like Barhdadi and Dube [2007], whose goal was to comprehensively evaluate gene-gene interactions. In general, the simulated dataset was very interesting and complex and should be useful for other evaluations of multistage designs in the future.

## CONCLUSIONS

There are many ways to implement a multistage study design. Ideally, a carefully planned and accurately executed staged design can help manage whichever dimension of complexity is considered most challenging in any given study, i.e., phenotypic complexity, model complexity, analytic complexity, or genetic complexity. The GAW15 contributions to the Multistage Designs group evaluated multiple approaches to reduce these various forms of complexity. It was clear from most studies that an analysis of all SNPs in all available samples provides maximum statistical power in the absence of cost constraints. However, given limited resources, a staged design can provide substantial cost savings at little loss of power. Consistent with previous studies [Skol et al., 2006; Wang et al., 2006], we confirmed that a joint statistical analysis that combines results from the two experimental stages is virtually always more powerful than treating the second-stage sample as an independent replication study. This statement was true, whether the complexity reduction was in the genetic or analytic dimension. However, it is important to carefully consider corrections for multiple comparisons in the multistage setting.

The possible two-stage study designs are quite diverse, as are the reasons for using these designs. However, the ability to balance costs and information gain is an important feature of all multistage designs. Financial considerations were the original and are still often the main reason for considering two-stage study designs. Given rapidly decreasing genotyping costs, some have argued that two-qstage study designs are no longer necessary. However, costs may not be the only factor in considering a two-stage study design. As data from genome-wide association studies are becoming publicly available, these datasets may be viewed as first stage data, making multistage designs an important tool in the new era of genetic studies. We believe that the time is ripe for an additional method development for multistage designs, including stages related to phenotypic and analytic complexity.

## ACKNOWLEDGMENTS

Group 14 primary contributing members include Hatef Darabi, Jing Li, Amina Barhdadi, Marie-Pierre Dube, Mike Schmidt, Silke Schmidt, Xuexia Wang, Quiying Sha, Zhaogong Zhang, Tao Wang, Hugues Aschard, Mickael Guedj, Rori Rohlfs, Amy Anderson, Chelsea Taylor, Lucia Mirea, Radoslav Nickolov, Valentin Milanov, Hsin-Chou Yang, Yeunjoo Song, and Ritwik Sinha. This work was supported by NIH grants MH59528 and EY015216 (ERH and SS),

the Neurosciences Education and Research Foundation (ERH), and the Fonds de la recherche en santé du Québec (MPD).

Contract grant sponsor: NIH; Contract grant numbers: MH59528, EY015216; Contract grant sponsor: Neurosciences Education and Research Foundation; Contract grant sponsor: Fonds de la recherche en santé du Québec.

## REFERENCES

Alcais A, Alter A, Antoni G, Orlova M, Nguyen VT, Singh M, Vanderborght PR, Katoch K, Mira MT, Vu HT, Ngyuen TH, Nguyen NB, Moraes M, Mehra N, Schurr E, Abel L. Stepwise replication identifies a low-producing lymphotoxin-alpha allele as a major risk factor for early-onset leprosy. Nat Genet 2007;39:517–522. [PubMed: 17353895]

Aschard H, Guedj M, Demenais F. A two step multiple-marker strategy for genome-wide association studies. BMC Proceedings 2007;1(Suppl 1):S134. [PubMed: 18466477]

Barhdadi A, Dube MP. Two-stage strategies to detect gene-gene interactions in case control data. BMC Proceedings 2007;1(Suppl 1):S135. [PubMed: 18466478]

Craddock N, Daniels J, Holmans P, Williams N, Owen MJ. Increasing the efficiency of genomic searches for linkage in complex disorders by DNA pooling of affected sib-pairs. Mol Psychiatry 1996;1:59–64. [PubMed: 9118316]

Fulker DW, Cherny SS, Sham PC, Hewitt JK. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 1999;64:259–267. [PubMed: 9915965]

Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M. Ordered subset analysis in genetic linkage mapping of complex traits. Genet Epidemiol 2004;27:53–63. [PubMed: 15185403]

Holmans P, Craddock N. Efficient strategies for genome scanning using maximum-likelihood affected-sib-pair analysis. Am J Hum Genet 1997;60:657–666. [PubMed: 9042927]

Ionita I, Man M. Optimal two-stage strategy for detecting interacting genes in complex diseases. BMC Genet 2006;7:39. [PubMed: 16776843]

Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. Using the noninformative families in family-based association tests: a powerful new testing strategy. Am J Hum Genet 2003;73:801–811. [PubMed: 14502464]

Li J. Marker selection for whole genome association studies with two stage designs using dense SNPs. BMC Proceedings 2007;1(Suppl 1):S136. [PubMed: 18466479]

Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37:413–417. [PubMed: 15793588]

Morton NE. Sequential tests for the detection of linkage. Am J Hum Genet 1955;7:277–318. [PubMed: 13258560]

Rohlfs RV, Taylor C, Mirea L, Bull SB, Corey M, Anderson AD. One-stage design is empirically more powerful than two-stage design for family-based genome-wide association studies. BMC Proceedings 2007;1(Suppl 1):S137. [PubMed: 18466480]

Rosenberg PS, Che A, Chen BE. Multiple hypothesis testing strategies for genetic case-control association studies. Stat Med 2006;25:3134–3149. [PubMed: 16252274]

Satagopan JM, Elston RC. Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 2003;25:149–157. [PubMed: 12916023]

Schmidt M, Qin X, Martin ER, Hauser ER, Schmidt S. Two-stage study designs for analyzing disease-associated covariates: linkage thresholds and case selection strategies. BMC Proceedings 2007;1 (Suppl 1):S138. [PubMed: 18466481]

Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 2006;38:209–213. [PubMed: 16415888]

Thomas D, Xie R, Gebregziabher M. Two-stage sampling designs for gene association studies. Genet Epidemiol 2004;27:401–414. [PubMed: 15543639]

Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C. Genomic screening and replication using the same data set in family-based association testing. Nat Genet 2005;37:683–691. [PubMed: 15937480]

Wang H, Thomas DC, Péer I, Stram DO. Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 2006;30:356–368. [PubMed: 16607626]

Wang T, Lu Q, Torres-Caban M, Elston RC. Two-stage analysis strategy for identifying the IgM quantitative trait locus. BMC Proceedings 2007a;1(Suppl 1):S139. [PubMed: 18466482]

Wang X, Zhang Z, Zhang S, Sha Q. Genome-wide association tests by two-stage approaches with unified analysis of families and unrelated individuals. BMC Proceedings 2007b;1(Suppl 1):S140. [PubMed: 18466484]

Whittemore AS, Halpern J. Multi-stage sampling in genetic epidemiology. Stat Med 1997;16:153–167. [PubMed: 9004389]

**TABLE I**

Summary of referenced papers by GAW15 group 14: Multistage Designs

| Contribution | GAW dataset | Stage 1 test statistic | Selection | Stage 2 test statistic | Results |
|---|---|---|---|---|---|
| Aschard et al. | Problem 3 | Local score statistic using all case-control data | SNPs in candidate regions | Multimarker family-based association | Stringency in first stage is best. Family-based test controls effect of population stratification |
| Barhdadi et al. | Problem 3 | Single-marker case-control association test | SNPs that pass a *P*-value threshold | SNP-SNP interaction coefficient in logistic regression | Stringency in first stage is best |
| Li | Problem 3 | Case-control test on a fraction of dataset, using regions with 300K and 10K SNP density | SNPs are selected based on *P*-value and further clustered according to D′ | Joint statistic on selected SNPs | Can reduce number of tests by clustering SNPs. Joint analysis is almost as powerful as single-stage analysis |
| Rohlfs et al. | Problems 3, 1 and other sources | Conditional means model | 10 SNPs | Family-based association | Single stage analysis is best. Family-based test controls effect of population stratification |
| Schmidt et al. | Problem 3 | Linkage analysis on all families and ordered subset analysis for covariates | Cases (one per family) and SNPs are selected for stage 2 analysis | Case-control association test | Covariates can increase power in the presence of genetic heterogeneity |
| Wang T et al. | Problem 3 | Different statistics are compared on a fraction of the case-control dataset | Most powerful statistic is chosen | Chosen statistic is applied to the remaining samples and joint statistic is used | Combination of *P*-values from both stages is most powerful |
| Wang X et al. | Problem 3 | Case-control association test in parental generation, with or without controls | SNPs | Selected SNPs are tested for family-based association | Including controls in stage 1 is more powerful |

**GAW, Genetic Analysis Workshop; SNPs single-nucleotide polymorphisms.**