

# Multistage LD studies and their computer implementation

PA-02-141 Continued development and maintenance of bioinformatics and computational biology software

Last time: József Bukszár

Today: JR. Robles and Edwin van den Oord

# Introduction

- Complex disorders such as diabetes affect the quality of life of many people and account for a substantial fraction of public health care expenditures.
- 
- It is therefore important to better understand the pathophysiology of these diseases and develop effective treatments.
- The identification of the genetic variation affecting disease susceptibility can be helpful in this process.
- No biological knowledge that relates sequence variants to diseases. Therefore necessary to screen many genetic markers for their association with the disease.
- This screening process brings along
  - 1) a considerable risk of false discoveries
  - 2) makes genetic studies expensive because many markers need to be genotyped.

# We focus on LD studies

Increasingly important. Contributing factors:

1. Disappointment with positional cloning
2. Abundance of SNPs
3. Reduction in genotyping costs

Think of candidate genes studies → fine mapping → large regions/many candidates → whole-genome LD scan

# Controlling false discoveries

If an infinite number of studies would be conducted, the FDR would be equal to the expected proportion of False Discoveries within the whole set of tests that are rejected.

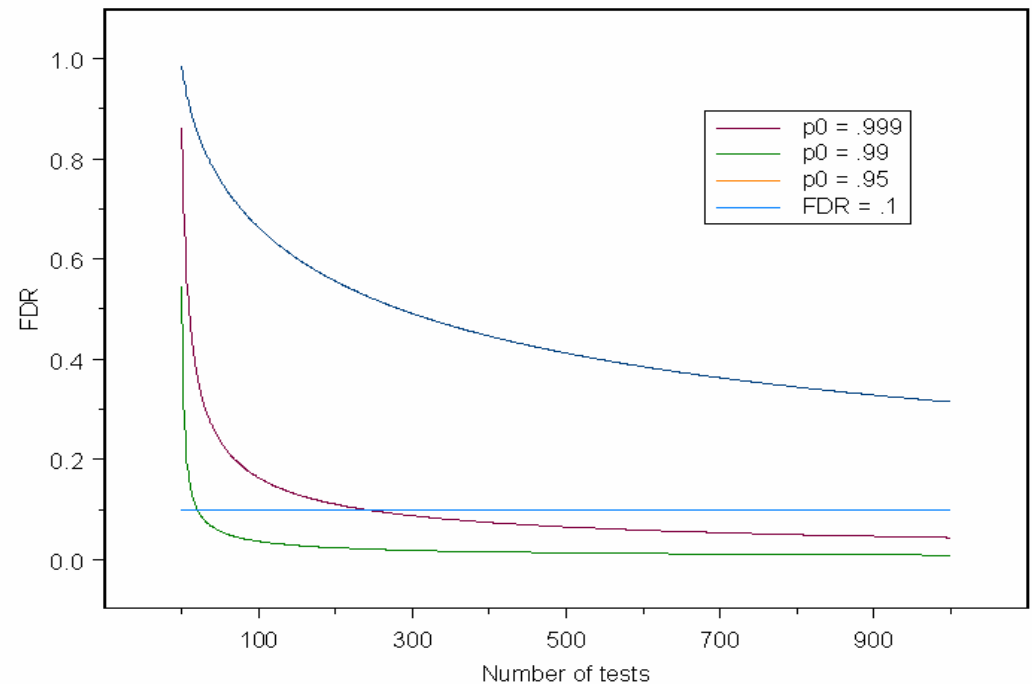
Good measure because:

1. Balances false and true discoveries

2. Does not depend on # tests

3. Works with correlated (LD) tests

FDR obtained when using Bonferroni correction



# Controlling false discoveries

$$\cancel{p_k = \alpha / m}$$

$$p_k = \frac{(1-p_0)PTD}{\frac{p_0}{FDR} - p_0}$$

The critical  $p$ -value  $p_k$  is the significance level that will control the FDR given PTD and  $p_0$ .

PTD Proportion of with True effects that are Detected (also average power)

$p_0$  is the proportion of markers with no effect (also prior probability in Bayesian terms).

# Reducing the genotyping burden using a 2-stage design

- Increasing although not massive interest in multistage designs:

Saito A, Kamatani N. J Hum Genet. 2002;47:360-365.

Satagopan t al. Biometrics. 2002/2004.

Van den Oord & Sullivan 2003. Hum Genet/Trends in Genetics

Satagopan JM, Elston RC. Genet Epidemiol. 2003;25:149-157.

Aplenc R, et al. Genetics. 2003;163:1215-1219.

Lowe CE, ..Clayton DG. Genes Immun. 2004;5:301-305.

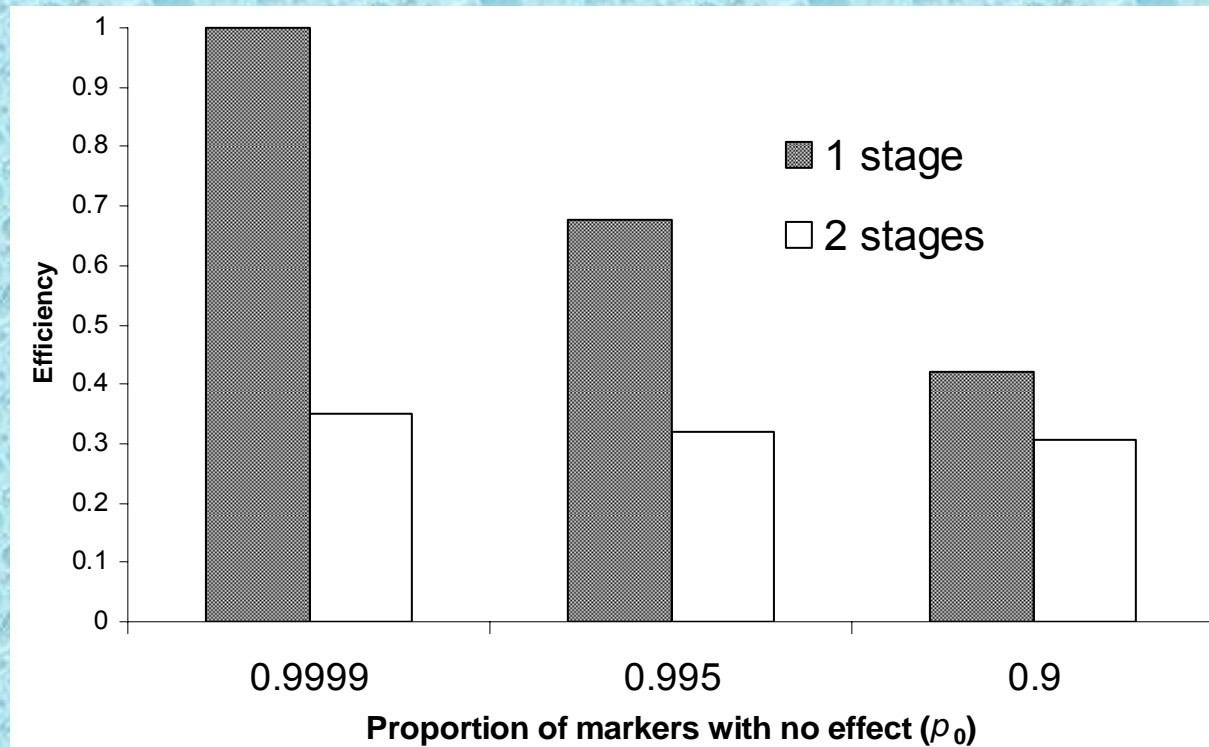
Kraft, P In preparation

- Compare population screens for breast cancer

# Our method minimizes the GB

- Genotype burden (GB) = average number of genotypes you need to do per marker
- In a 1-stage design; GB = number of individuals screened
- For a 2-stage design:  $GB = N1 + N2 \times \text{Proportion of marker select at stage 1}$
- GB does *not* depend on the number of markers that are typed. This is possible because the goals of the study (FDR and PTD) are defined in terms of proportions.
- GB pleasant measure in planning a study:
  - A) if the total budget for the study is \$10,000, 1 genotype costs 25 cents, and the GB is 500, then the total number of markers that can be typed equals  $10,000 / (.25 \times 500) = 80$ .
  - B) If 10 markers will be typed the total number of genotypes needed equals 10 times the GB.

# GB reduction



*Figure 1. Genotyping burden in 2- versus 1-stage designs.*

# Our (read JR) software: Iga972

Parameters Results

Study  
Two Stages

Goals and p0

FDR 0.150 PTD 0.715 p0 0.9660

- Transmission Disequilibrium Test
- Genotype Relative Risk
- Quantitative TDT
- Liability Model TDT
- Liability Model Relative Risk
- Quantitative Case Control
- Case-Control (genotype-based results)
- Case-Control (allele-based results)
- CC (controls from population) (genotype-based)
- CC (controls from population) (allele-based)
- Liability Model CC (genotype-based)
- Liability Model CC (allele-based)

K 0.050  
s 0.500  
lt 0.100  
ut 0.100

Exit Start

Parameters Results

Study Configuration:  
Stages=2  
False Discovery Rate (goal)=0.2  
Proportion of True Effect Detected (goal)=0.5  
Proportion of Null-effects (stage 1)=0.999

Selected Test: Liability Model TDT  
Degrees of Freedom= 1.0  
Frequency (relative) of the risk allele= 0.5  
Heritability= 0.01  
d/a (amount of dominance)= 0.0  
Proportion selected from upper tail= 0.1

Copy  
Cut  
Paste  
Help  
Labels  
About

Results

Step	FDR	PTD	Pk	Ptml	N	Genotyping Burden
1	0.96872	0.50557	0.01567	0.01616	510	1532
2	0.20000	0.98899	0.00798	0.00063	2113	102

Total Genotyping Burden=1634  
Total Sample Size=2623

Exit Start

# We need to demonstrate competitive advantage

## 1) Scientific principles

- FDR versus Bonferroni or none
- Versatility; Case-control + family based LD tests; Categorical + quantitative tests; Models for studying selection effects; 2-stage + 1 stage designs for comparison
- Optimal control; Theoretical calculations instead of simulations. For example data pooling:

# Example data pooling

Inflation Type I error.

For example, critical value chi-square test

1. Ignore: Satagopan and Elston 2003

N1, N2,, Alpha 1	No data pooling	Data pooling
500,1000	9.85	14.26
200,200	2.71	8.95

In other words, the test statistic pair  $(X_1, X_2)$  for a marker in the two stages has a bivariate normal distribution with mean vector  $(n_1\mu, n\mu)$  and covariance matrix  $\Sigma = \begin{pmatrix} n_1 & n_1 \\ n_1 & n \end{pmatrix}$ .

2. Simulation: Clayton et al. 2004

We stop after stage  $k$  if  $T_k$  fails to exceed a critical value  $c_k$ . The probability of exceeding this critical value conditional upon reaching stage  $k$  is

$$\begin{aligned} & \Pr(T_k > c_k | T_1 > c_1, \dots, T_{k-1} > c_{k-1}) \\ &= \int_{u_1} \dots \int_{u_{k-1}} \Pr(T_k > c_k | u_1, \dots, u_{k-1}) \\ & \quad \times \Pr(u_1, \dots, u_{k-1} | T_1 > c_1, \dots, T_{k-1} > c_{k-1}) du_1, \dots, du_{k-1}. \end{aligned}$$

This integral is intractable but may be approximated by simulation.

# ...but we have József Bukszár

The probability that a marker is rejected in the second stage is

$$\Pr(\mathbf{T} \in c, T \in c)$$

We have seen that  $T$  is approximately

$$T \approx Z_1^2 + \dots + Z_{m-1}^2 = \left( A_1 \mathbf{U} + \sqrt{n} \sqrt{v_1^2 + \frac{1}{m-1} v_m^2} \right)^2 + \dots + \left( A_{m-1} \mathbf{U} + \sqrt{n} \sqrt{v_{m-1}^2 + \frac{1}{m-1} v_m^2} \right)^2,$$

where

$$U_i = \frac{1}{\sqrt{p_i p_i + q_i q_i}} \left[ \left( \frac{q_i \sqrt{p_i} - p_i \sqrt{q_i}}{2 \sqrt{p_i p_i + q_i q_i}} \right) \frac{Y_i - q_i n}{\sqrt{q_i n}} - \left( \frac{q_i \sqrt{p_i} + p_i \sqrt{q_i}}{2 \sqrt{p_i p_i + q_i q_i}} \right) \frac{X_i - p_i n}{\sqrt{p_i n}} \right]$$

and  $\mathbf{v} = \frac{1}{\sqrt{n}} A \boldsymbol{\mu} = A \left( \frac{(p_1 - q_1) \sqrt{p q}}{\sqrt{p p_1 + q q_1}}, \dots, \frac{(p_m - q_m) \sqrt{p q}}{\sqrt{p p_m + q q_m}} \right)^T$  ← does not depend on  $n$

*Test statistic on the Stage II data only.*

$$\rightarrow T' \approx Z_1'^2 + \dots + Z_{m-1}'^2 = \left( A_1 \mathbf{U}' + \sqrt{n'} \sqrt{v_1^2 + \frac{1}{m-1} v_m^2} \right)^2 + \dots + \left( A_{m-1} \mathbf{U}' + \sqrt{n'} \sqrt{v_{m-1}^2 + \frac{1}{m-1} v_m^2} \right)^2$$

Similarly for the test statistic on the pooled data

$$T^* \approx Z_1^{*2} + \dots + Z_{m-1}^{*2}.$$

It can be seen that  $Z_i^* = \frac{\sqrt{n}}{\sqrt{n+n'}} Z_i + \frac{\sqrt{n'}}{\sqrt{n+n'}} Z_i', \quad i = 1, \dots, m.$

# We need to demonstrate competitive advantage

## 2) Software development

LGA completely coded from scratch:

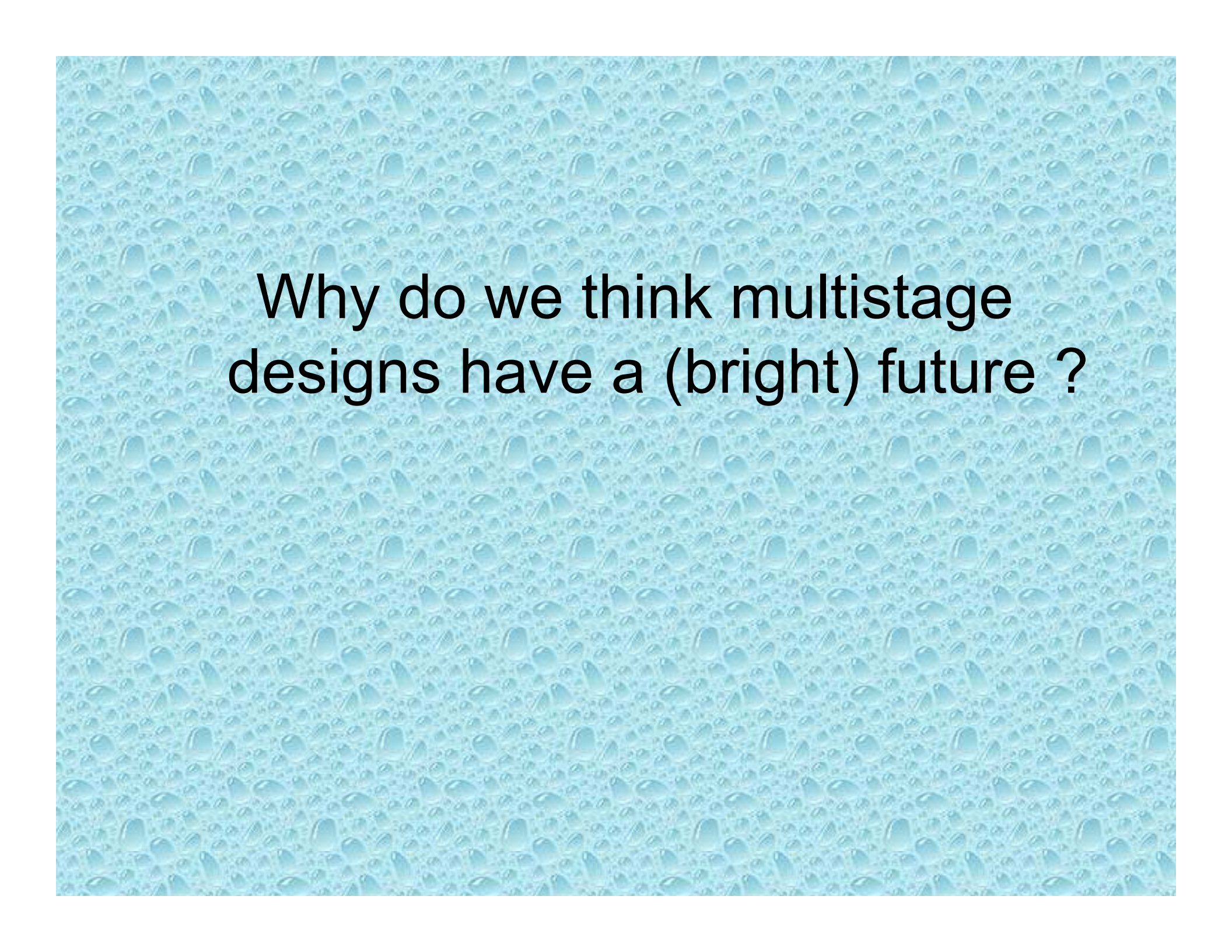
- time consuming and not necessarily easy (non-central F)
- Once you have it many advantages. For example, root (zero)-finder Power =  $f(N, \dots)$  we have Brent-Robles.

In general, competitive advantage LGA;

- Reparability; Complete control over the code.
- Evolvability; Genetic Algorithm for optimization task Object oriented program;
- User-friendly; Graphical user-interface; Cross-platform Stand-alone: no libraries needed; Well documented

# The future of multistage designs????

- Genotyping costs are decreasing fast. Most extreme example:
  - Affymetrix The GeneChip® Mapping 100K genotype: 100,000 SNPs with a single primer.
  - In prospect of their new chip with 500k SNPs, Affymetrix already dropped the price of their 100k chip set to \$600.
  - Illumina will bring a chip on the marker with 100k SNPs from HapMap.
- One could argue that the genotyping costs will eventually become so low that it may no longer be worth while to use multistage designs that
  - 1) are more laborious
  - 2) require somewhat larger sample sizes

The background of the slide is a light blue surface covered with numerous small, clear water droplets of varying sizes, creating a textured, dewy appearance.

Why do we think multistage designs have a (bright) future ?

# 1. Sample size increase may not be such a serious disadvantage.

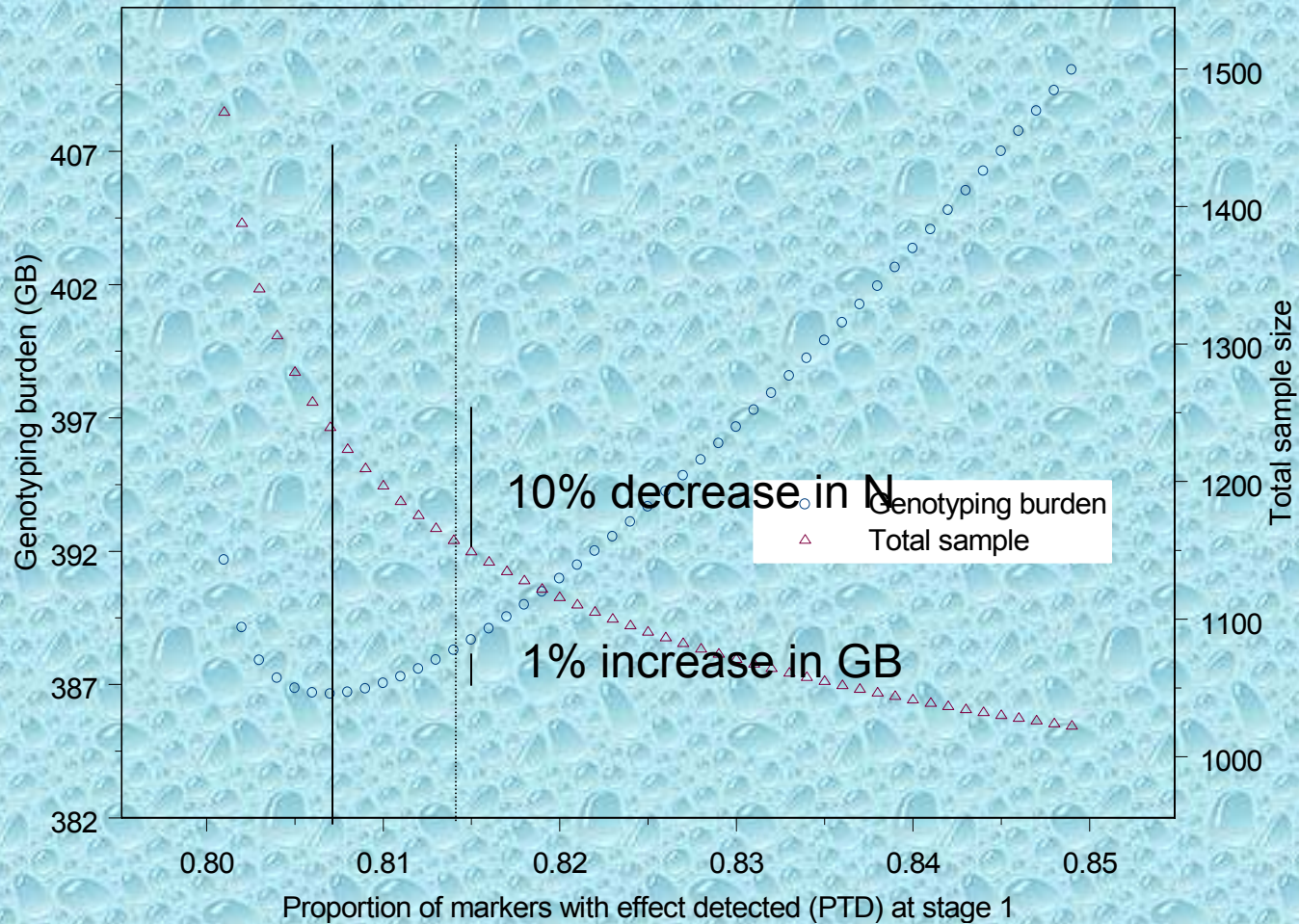
a) costs of a study  $\neq$  phenotyping cost + genotyping costs.

Because sampling and phenotyping costs are onetime. Once the sample has been collected it can be used for many genotyping projects. Better to collect a sample that forms a fruitful basis for further projects and reserve a part of the budget for genotyping in the initial project.

b) Data pooling mitigates effect on total sample. Rule of thumb: gain is half the stage 1 sample size.

c) We can design studies with suboptimal GB

# Suboptimal GB: GB-Sample size trade-off



## 2. Genotyping is still a bottleneck and may remain so foreseeable future

- a) Extremely cheap Very-high-throughput technologies may not be suitable for all genotyping projects. For example, animals, specific regions in detail, need non-standard set of SNPs because of population/genetic (e.g. population/selection) effects
- c) Even current chips limited. Most recent estimates (Ke et al. e-published) suggest that 100– 300k htSNPs may be required to cover only the high LD regions making up about 50% of the genome. For near-complete coverage 1,000k or more selected htSNPs may be required.
- c) Multistage designs remain advantageous even with chips

	FDR	PTD	$p_{(k)}$	Required sample	GB
1-stage design chip used for whole sample					
Stage 1	0.1	0.8	.000010	3083	3083
Chip stage 1, 20 times more expensive genotyping in stage 2					
Stage 1	0.983	0.817	.004725	1622	1534
Stage 2	0.1	0.979	0.00188	2156	207
Total				3778	1,829
Fixed genotype costs					
Stage 1	0.9903	0.815	0.08358	852	
Stage 2	0.1	0.981	0.00011	3459	
Total				4311	

- 2-stage design is still merely 60% of the costs of the 1-stage design
- In absolute terms the savings might be substantial. For example, costs 1-stage design using a 100k chip costing \$600 equals approximately 1,8 million (= \$600 × 3083 subjects). With a 2-stage design this would be 60% × 1.8 = 1.1 million, which is a saving of \$700,000.

## 2. Genotyping is still a bottleneck and may remain so (Cont.)

- d) Common Disease/Common Variant (CD/CV) hypothesis might capture only % of the causal variants
- Solution:
  - 1) Weak version (Botstein & Risch, 2003)
  - 2) Strong version we need to sequence to find and test all the rare variants

### 3. Other advantages multistage designs

- a) Multistage designs also save DNA

Reducing the amount of genotyping with 50-70% also implies the same amount of savings in DNA. This may be important when DNA is limited (cheek swabs)

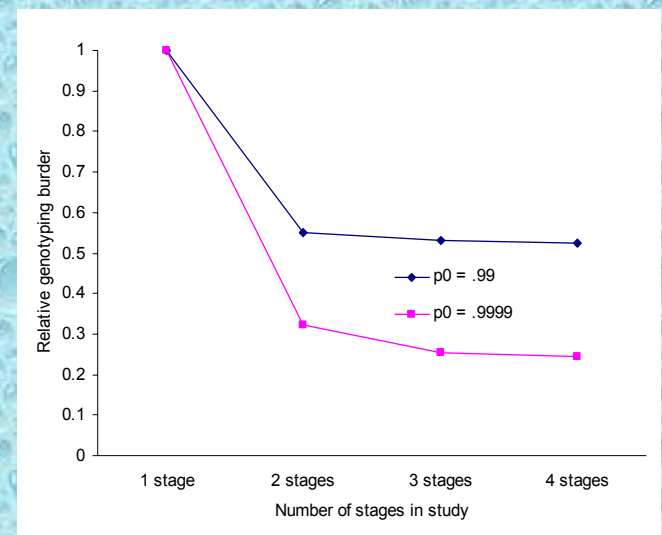
- b) Possibility of adjustment the study

$$\text{FDR stage 1} = f(\text{PTD}, p_0) = p_0 \text{ stage 2}$$

- c) Research groups make lack the N for a complete study. Also think about, GxG/GxE (smaller effect sizes, higher  $p_0$ ). Now replication as the 2<sup>nd</sup> stage. Provide guidelines.

# Concrete extensions

- **Mitigating the sample size increase in multistage designs**
  - Allowing for a suboptimal GB
  - Data pooling
- **Variable genotyping costs**
  - very-high-throughput genotyping chips special case
- **Sample size constraints**
  - Fixed sample size in one of the stages
  - Optimal solutions with fixed total sample size (+different selection criteria)
- **Practical constraints**
  - N is plate size times constant
  - Chip size fixed (e.g. Illumina method greatest efficiency by typing sets of 1536 markers).
- **Test extensions**
  - GxG, GxE
  - Siblings
  - Haplotypes/trend test
- **Design extensions**
  - Different samples across stages
  - Three stage design



# Extensions initially requiring further scientific work

- Learn how to feed back stage 1 results into the program to update the stage 2 design or provide guidelines for future replication studies
- Determine characteristics of optimal designs that are shared across many tests and those that are highly test dependent. Use this knowledge to create a program designing optimal studies for non-standard tests where the test dependent characteristics are user-specified.

Step	FDR	PTD	Pk	N	Genotyping Burden
Transmission Disequilibrium Test					
1	0.92882	0.73019	0.09624	246	740
2	0.10000	0.95866	0.00816	916	281
Case-Control (allele-based results)					
1	0.94062	0.75196	0.12031	322	322
2	0.10000	0.93090	0.00653	1140	144
Quantitative Case Control					
1	0.93165	0.72767	0.10018	638	638
2	0.10000	0.96198	0.00784	2248	239



# JR: Computer implementation