Experimental design and statistical analysis of SNP data obtained in genetic association studies

Norsk Regnesentral
ANVENDT DATAFORSKNING
Norwegian Computing Center/Applied Research and Development

NOTAT/NOTE

Figure taken from The Human Genome Project
at http://www.ornl.gov/hgmis/

SAMBA/28/03
Marit Holden
Ola Haug
Oslo

December 2003

Norsk Regnesentral
ANVENDT DATAFORSKNING

NR-notat/NR Note

Tittel/Title:
Experimental design and statistical analysis of SNP data obtained in genetic association studies

Dato/Date:	December
År/Year:	2003
Notat nr: Note no:	SAMBA/28/03

Forfatter/Author:
Marit Holden, Ola Haug

Sammendrag/Abstract:
A literature search for papers describing mean-scale and large-scale SNP-experiments with focus on biomedical research has been performed. The aim has been to find the "state-of-the art" within this area to make a web page with advices to scientist planning and analysing such experiments. The results of this literature study are given in a PowerPoint presentation and on the web page http://www.nr.no/documents/samba/research_areas/SMBI/BioSNP/. The most concrete advices, including references to available software, are given for hypothesis testing and case-control studies. Short descriptions of other experimental designs and statistical methods, are also given.

Emneord/Keywords:	SNPs, genetic association studies, experimental design, case-control design, statistical analysis
Tilgjengelighet/Availability:	Open
Prosjektnr./Project no.:	220090
Satsningsfelt/Research field:	Bioinformatics
Antall sider/No. of pages:	32

Norsk Regnesentral / Norwegian Computing Center
Gaustadalléen 23, Postboks 114 Blindern, 0314 Oslo, Norway
Telefon 22 85 25 00, telefax 22 69 76 60

Contents

Experimental design and statistical analysis of SNP data 1

Appendix A – Web pages 3

Experimental design and statistical analysis of SNP data…..……... 4
Selection of individuals…………………………………………... 5
Software …………………………………...………………... 6
Selection of target SNPs……………………………..…………. 7
DNA pooling.……….…..……………………………………. 8
Data analysis…………...…………………………...……….. 9

Appendix B – PowerPoint presentation 11

Association studies…………………………….……..………. 13
Experimental design……………………………………………. 15
Statistical analysis…………………………………………… …19
A review paper…………………………………………………… 25
Software …………………………………...………………... 28
Literature…………………………..………………………. 29

Experimental design and statistical analysis of SNP data

A **literature search** for papers describing mean and large-scale SNP-experiments with focus on biomedical research has been performed for finding the “state-of-the art” within this area. Papers were searched for mainly in Nature, Nature Review Genetics, Science, Current Biology, the American Journal of Human Genetics and the reference lists in the most interesting papers. In addition several papers were found from searches in MedLine and other Internet resources. A reference list with relevant literature is given in the PowerPoint presentation "A survey of experimental designs and statistical analysis of SNP data obtained in genetic association studies" presented at a meeting during the project. This PowerPoint presentation is included in Appendix B.

Web pages with advices concerning experimental design and statistical analysis of SNP data, including **references to papers** and **software**, have been made. A copy of the web pages (http://www.nr.no/documents/samba/research_areas/SMBI/BioSNP/) is found in Appendix A. The most concrete advices are given for hypothesis testing and case-control studies (without DNA pooling). Short descriptions of other experimental designs and statistical methods, are also given.

There are two main groups of **experimental designs** for genetic associations studies. These are case-control designs and family-based designs. Here we have concentrated on case-control designs, where we have cases with disease and controls, which are unrelated to the cases, without disease. There are three main issues concerning the experimental design. These are how to select the individuals for the study, which SNPs to genotype, and whether to use DNA pooling or not. When it comes to **which SNPs to genotype** and whether to use **DNA pooling or not**, this has not been focused in this report. We have therefore assumed that the user already has defined a set of candidate genes/SNPs to be examined. This might for instance be a set of SNPs selected within or close to a candidate gene. We have also assumed that a design with DNA pooling has not been chosen. Use of DNA pooling reduces the cost of large association studies, but results in loss of information and additional experimental errors.

One potential, main problem with case-control designs is confounding due to **stratification** (subpopulations), caused by for example ethnicity, age or sex. This means that the finding of a positive association between a genetic variant and a complex disease phenotype, does not necessarily establish causality. A disease that is prevalent in one subpopulation will be associated with any alleles that are in high frequency in that subpopulation. This leads to false-positive results and loss of robustness, and thereby inconsistency and lack of reproducibility among studies. There are several ways of reducing/avoiding stratification like use of family-based designs, where the controls are related to the cases, and use of random or unlinked genetic markers. In this project we mainly consider matched case-control designs and assume that the cases and controls are chosen such that stratification is no problem. This might be done by the use of matched case-control designs, i.e. studies in which the **individuals are chosen** such that controls are matched to cases on potential confounding factors like ethnicity, age and sex.

After the SNPs are genotyped, they should be analysed using **statistical analysis methods** to find which of the examined SNPs that are associated with disease. A SNP is associated with disease if it is a causal SNP or if it is in linkage disequilibrium with a causal SNP. The input to the statistical analysis is for each SNP whether there are 0, 1 or 2 occurrences of a certain allele (individual genotyping) or allele frequencies (when DNA pooling has been used). The statistical methods are often based on hypothesis testing where H0:"The SNP is not associated with disease" is tested against H1:"The SNP is associated with disease". A few methods, which are not based on hypothesis testing are also mentioned. The hypothesis test described in detail on the web pages is the Pearson's chi-square test for case-control designs (without DNA pooling). Software for doing this standard hypothesis test for SNPs obtained from individual genotyping is available (see http://www.genetics.med.ed.ac.uk/software/). Tools for analysing genotyping data obtained from DNA pools are also included in that software.

When designing SNP experiments a decision has to be made about **how many individuals** should be included in the study. A genetic power calculator is available on the Internet (see http://statgen.iop.kcl.ac.uk/gpc/). The calculator might be used for estimating how many individuals that are needed to obtain a certain power of the test, given a significance level and a lower bound for what should be detectable.

Multiple testing is an important issue because complex diseases ask for many SNPs to be examined simultaneously, and therefore many hypotheses will be tested simultaneously. The simplest and most conservative approach for multiple-testing correction is the **Bonferroni correction,** which has been described in detail on the web pages.

Appendix A

Web pages

Web pages with advices concerning experimental design and statistical analysis of SNP data, including **references to papers** and **software**, have been made. This appendix contains a copy of those web pages (see http://www.nr.no/documents/samba/research_areas/SMBI/BioSNP/). The most concrete advices are given for hypothesis testing and case-control studies (without DNA pooling). Short descriptions of other experimental designs and statistical methods, are also given.

Experimental design and statistical analysis of SNP data

This is a description of how to design and analyse SNP experiments in genetic association studies for complex diseases.

Experimental design

There are two main groups of designs for genetic associations studies. These are case-control designs and family-based designs. Here we will concentrate on case-control designs, where we have cases with disease and controls, which are unrelated to the cases, without disease. There are three main issues concerning the experimental design. These are

- How to select the individuals for the study.
- Which SNPs to genotype.
- Weather to use DNA pooling or not.

Statistical analysis

After the SNPs are genotyped, they should be analysed using statistical analysis methods to find which of the examined SNPs that are associated with disease. A SNP is associated with disease if it is a causal SNP or if it is in linkage disequilibrium with a causal SNP.

PowerPoint presentation

The PowerPoint presentation "A survey of experimental designs and statistical analysis of SNP data obtained in genetic association studies" is found here. A reference list with relevant literature is included in this presentation.

Selection of individuals

The stratification problem

One potential, main problem with case-control designs is confounding due to stratification (subpopulations), caused by for example ethnicity, age or sex. This means that the finding of a positive association between a genetic variant and a complex disease phenotype, does not necessarily establish causality. A disease that is prevalent in one subpopulation will be associated with any alleles that are in high frequency in that subpopulation. This leads to false-positive results and loss of robustness, and thereby inconsistency and lack of reproducibility among studies.

How to reduce/avoid the stratification problem

There are several ways of reducing/avoiding stratification. These are:

- Use of family-based designs, where the controls are related to the cases. The most used such controls are parents or unaffected siblings.
- Use of random or unlinked genetic markers. These might be used to determine the extent of stratification by statistical analysis and then to adjust for population stratification.
- Study multiple case-control populations (different ethnic groups). If association is seen in each population we might conclude that there is causal association.
- Use matched case-control designs, i.e. choose the individuals such that controls are matched to cases on potential confounding factors like ethnicity, age and sex.

Here we only consider matched case-control designs and assume that the cases and controls are chosen such that stratification is no problem. There exist statistical methods to check weather stratification is a problem or not for a certain data set, see for example Pritchard & Rosenberg 1999 and Pritchard et al.2000. Software for the method described in these papers is available.

How many individuals are needed in case-control studies?

This will depend on the number of SNPs examined and the statistical method / test statistic used. See the data analysis page for more about this.

Software

There exist statistical methods to check whether stratification is a problem or not for a certain data set, see for example Pritchard & Rosenberg 1999 and Pritchard et al.2000. Software for the method described in these papers is available and is shortly described below.

structure

"The program **structure** is a free software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including microsatellites, RFLPs and SNPs."
Software is found here.

STRAT

"The program **STRAT** is a companion program to structure. This is a structured association method, for use in association mapping, enabling valid case-control studies even in the presence of population structure."
Software is found here.

Selection of target SNPs

For the time being we do not focus on this, and assume that the user has already defined a set of candidate genes/SNPs to be examined. This might for instance be SNPs selected within or close to a candidate gene. Botstein & Risch 2003 argues for a sequence-based approach where SNPs are chosen from coding regions, disrupt splice sites and from promoter regions. These are the SNPs most likely to be of functional significance and to influence directly the traits under study. A recent approach is to use haplotype tags for identification of SNPs (see Johnson et al. 2001 and Daly et al. 2001), and thereby reducing the number of SNPs to be genotyped.

DNA pooling

For the time being we do not focus on this and assume that a design with DNA pooling has not been chosen. Nevertheless, some comments concerning DNA pooling are given below.

- In a two pool design DNA material from all cases is put together into a pool, i.e. there is only one sample for the pooled cases. Similarly, there is only one sample for the pooled controls.
- In a two pool design we are genotyping two groups of individuals (affected and controls) rather than each of the individuals themselves. This reduces the cost of large association studies, but results in loss of information and additional experimental errors.
- Pools to be compared should be well-balanced, i.e. there should be equal representation of sexes, age groups, ethnicities etc. in each of the two pools.
- It is also possible to use multiple-pool designs to obtain average allele-frequency estimates from several pools and measurements. This will reduce uncertainties of obtained allele-frequency estimates.
- See Risch & Teng 1998 and Sham et al. 2002 for hypothesis tests to use for case-control and family based designs when DNA pooling has been used.
- Software for doing the standard hypothesis test for SNPs obtained from both individual genotyping and genotyping of pools is available. The additional experimental errors introduced by the DNA pooling are taken into account in the model used. A description of the model and software is given here.

Data analysis

The input to the statistical analysis is for each SNP

- whether there are 0, 1 or 2 occurrences of a certain allele (individual genotyping) or
- allele frequencies (for a DNA pool)

Here we assume that the data sets are chosen such that stratification is not a serious problem. There exist statistical methods to check weather stratification is a problem or not for a certain data set, see for example Pritchard & Rosenberg 1999.

The statistical methods are often based on hypothesis testing with alternatives

- H0: “the SNP is not associated with disease” and
- H1: “the SNP is associated with disease”.

Tests for case-control designs

The TDT test for family-based designs and its variants are described for example in Spielman et al. 1993, Risch 2000, Lazzeroni & Lange 1998, Spielman & Ewens 1998, Schaid & Rowland 1998. Different tests for DNA pooling designs (both family based and case-control designs) are found for example in Risch & Teng 1998 and Sham et al. 2002. Here we will describe the Pearson's chi-square test for case-control designs without DNA pooling. From the 2x2 contingency table

	Allele 1	Allele 2	Row totals
Cases	N_{11}	N_{12}	N_{1*}
Controls	N_{21}	N_{22}	N_{2*}
Column totals	N_{*1}	N_{*2}	N

the Pearson Chi-square statistic is computed as follows

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(N_{ij}-M_{ij})^2}{M_{ij}} \quad \text{where} \quad M_{ij}=\frac{N_{i*}}{N_{*j}}$$

Here N_{11} (N_{21}) and N_{12} (N_{22}) are the total number of occurrences of allele 1 and allele 2, respectively for the cases (controls). $N_{1*}=N_{11}+N_{12}$ ($N_{2*}=N_{21}+N_{22}$) is equal to twice the number of cases (controls).

The significance level is normally set to 5%. When testing only one hypothesis, the null hypothesis is rejected if the statistic above is larger than 3.84, i.e. the 95% percentile of the chi-square distribution with one degree of freedom. When several SNPs are examined simultaneously, several hypothesis are tested simultaneously and multiple testing is an important issue (see below).

A genetic power calculator is found here and is described in this paper. The calculator might be used for estimating how many individuals that are needed to obtain a certain power of the test (typically 80%), given a significance level and a lower bound for what should be detectable. For case-control designs choose the "Case-control for discrete traits"-entry on the before-mentioned web page.

Software for doing the standard hypothesis test described above for SNPs obtained from individual genotyping is available here. Tools for analysing genotyping data obtained from DNA pools are also included in this software. A description of the model and software is given here.

Above we have assumed that the genotype data are reported with no errors. How to deal with genotyping errors has been proposed in for example Gordon & Ott 2001, Gordon et al. 2002, and also in the paper mentioned in the paragraph above.

Multiple testing

Multiple testing is an important issue because complex diseases ask for many SNPs and therefore many hypotheses to be tested simultaneously. The simplest and most conservative approach for multiple testing correction is the Bonferroni correction. If the significance level for the entire set of n comparisons is equal to alpha, the significance level for each comparison is set equal to alpha/n. There exist several alternative, less conservative methods for correcting for multiple testing. When using such methods, less individuals are needed for obtaining the same power and significance level of the test. One such method is based on controlling the false discovery rate and is described in Sabatti et al. 2003.

Other statistical methods

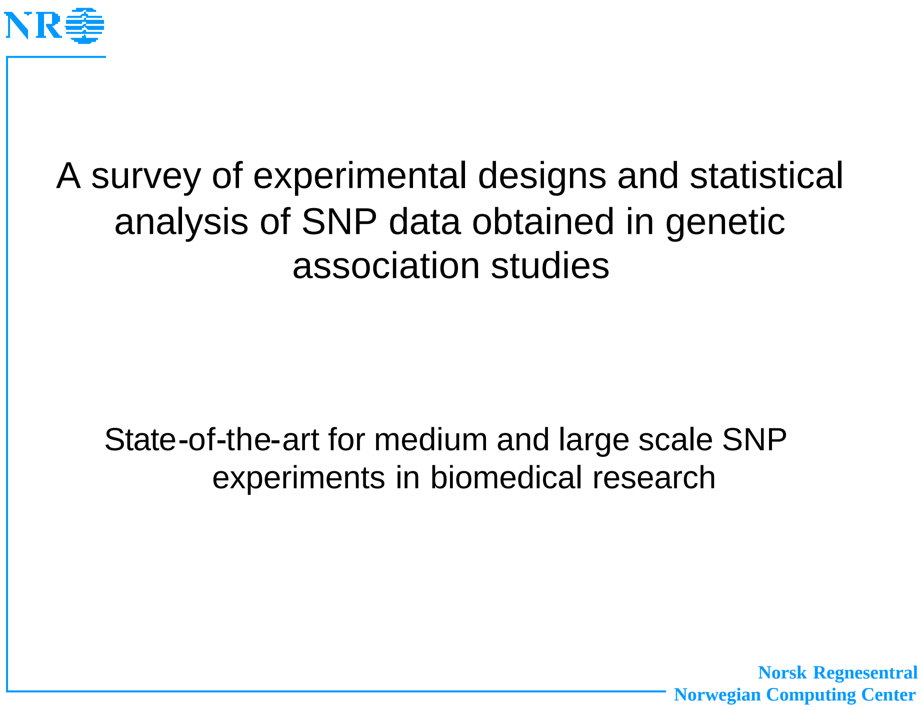
Statistical methods not based on hypothesis testing have also been proposed. Some examples of such methods are given below. Main advantages of these are that there is no need for correction for multiple testing and that the SNPs are analysed jointly rather than tested one by one.

- In Devlin & Roeder 1999 and Devlin et al. 2000 a Bayesian outlier method is described. This method also controls for population heterogeneity (stratification). Software is available on request.
- A set association method is described in Hoh et al. 2001. Software is available here.

Appendix B

PowerPoint presentation

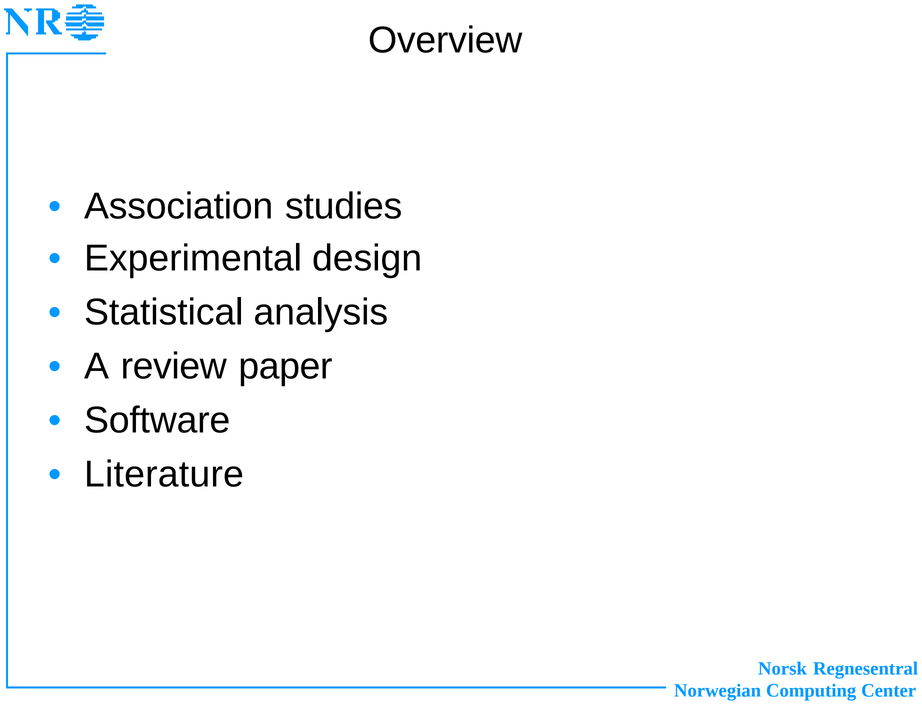
A literature search for papers describing mean- and large-scale SNP-experiments with focus on biomedical research has been performed for finding the “state-of-the art” within this area. Papers were searched for mainly in Nature, Nature Review Genetics, Science, Current Biology, the American Journal of Human Genetics and the reference lists in the most interesting papers. In addition several papers were found from searches in MedLine and other Internet resources. A reference list with relevant literature is given in the PowerPoint presentation "A survey of experimental designs and statistical analysis of SNP data obtained in genetic association studies" presented at a meeting during the project. This appendix contains a copy of that PowerPoint presentation.



A survey of experimental designs and statistical analysis of SNP data obtained in genetic association studies

State-of-the-art for medium and large scale SNP experiments in biomedical research

Norsk Regnesentral
Norwegian Computing Center

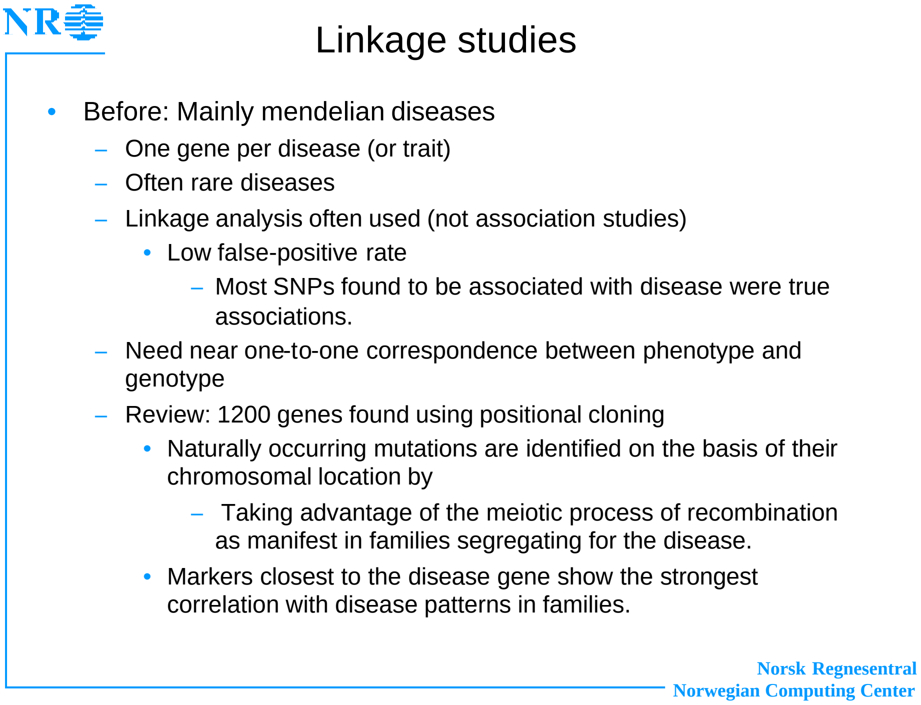


Overview

- Association studies
- Experimental design
- Statistical analysis
- A review paper
- Software
- Literature

Norsk Regnesentral
Norwegian Computing Center

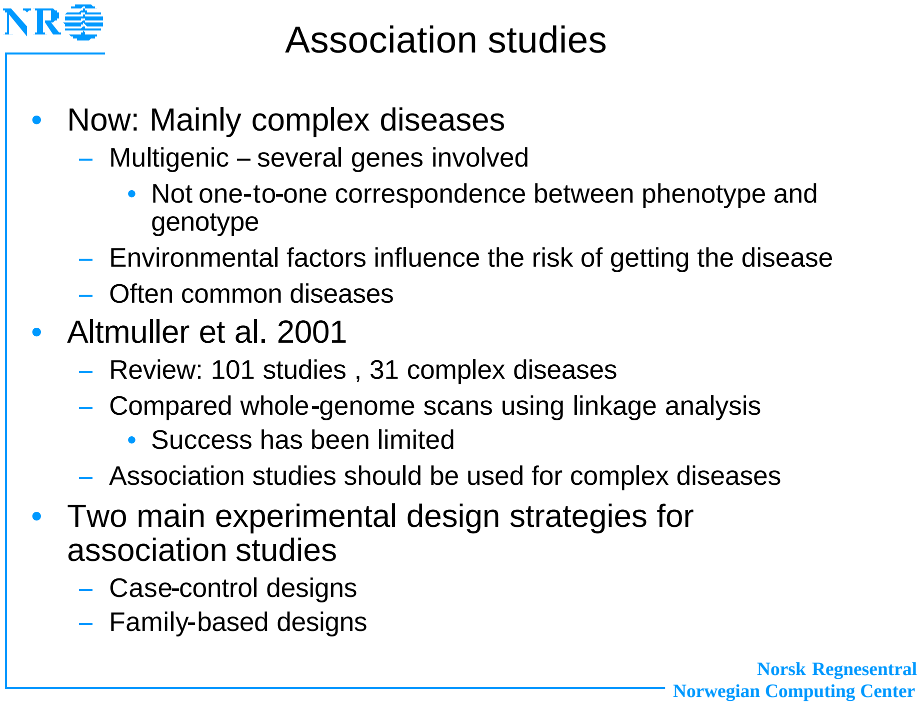
Association studies



Linkage studies

- Before: Mainly mendelian diseases
  - One gene per disease (or trait)
  - Often rare diseases
  - Linkage analysis often used (not association studies)
    - Low false-positive rate
      - Most SNPs found to be associated with disease were true associations.
  - Need near one-to-one correspondence between phenotype and genotype
  - Review: 1200 genes found using positional cloning
    - Naturally occurring mutations are identified on the basis of their chromosomal location by
      - Taking advantage of the meiotic process of recombination as manifest in families segregating for the disease.
    - Markers closest to the disease gene show the strongest correlation with disease patterns in families.

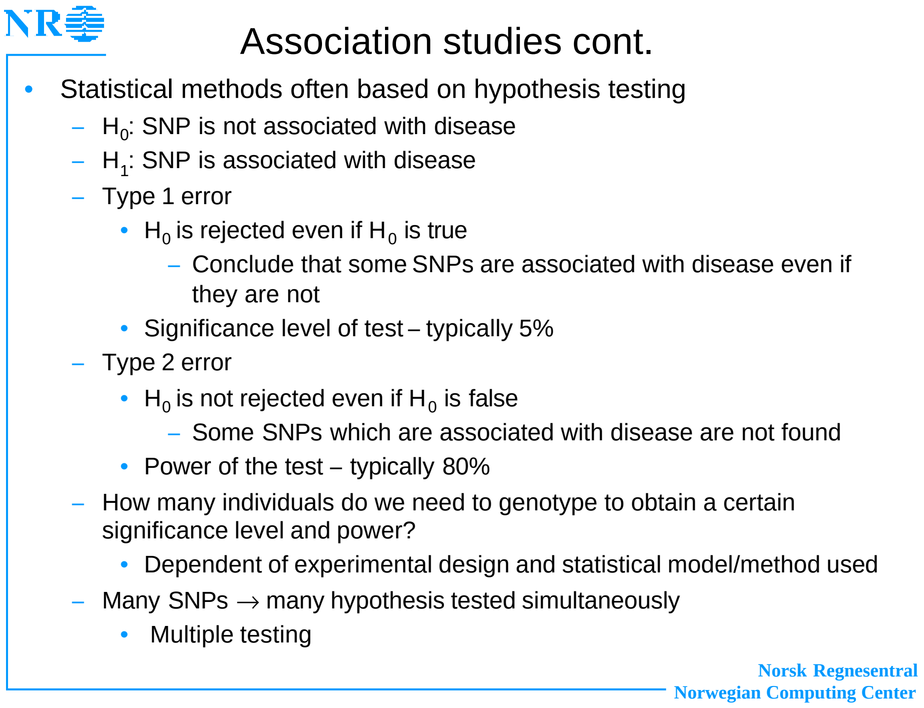
Norsk Regnesentral
Norwegian Computing Center



Association studies

- Now: Mainly complex diseases
  - Multigenic – several genes involved
    - Not one-to-one correspondence between phenotype and genotype
  - Environmental factors influence the risk of getting the disease
  - Often common diseases
- Altmuller et al. 2001
  - Review: 101 studies , 31 complex diseases
  - Compared whole-genome scans using linkage analysis
    - Success has been limited
  - Association studies should be used for complex diseases
- Two main experimental design strategies for association studies
  - Case-control designs
  - Family-based designs

Norsk Regnesentral
Norwegian Computing Center

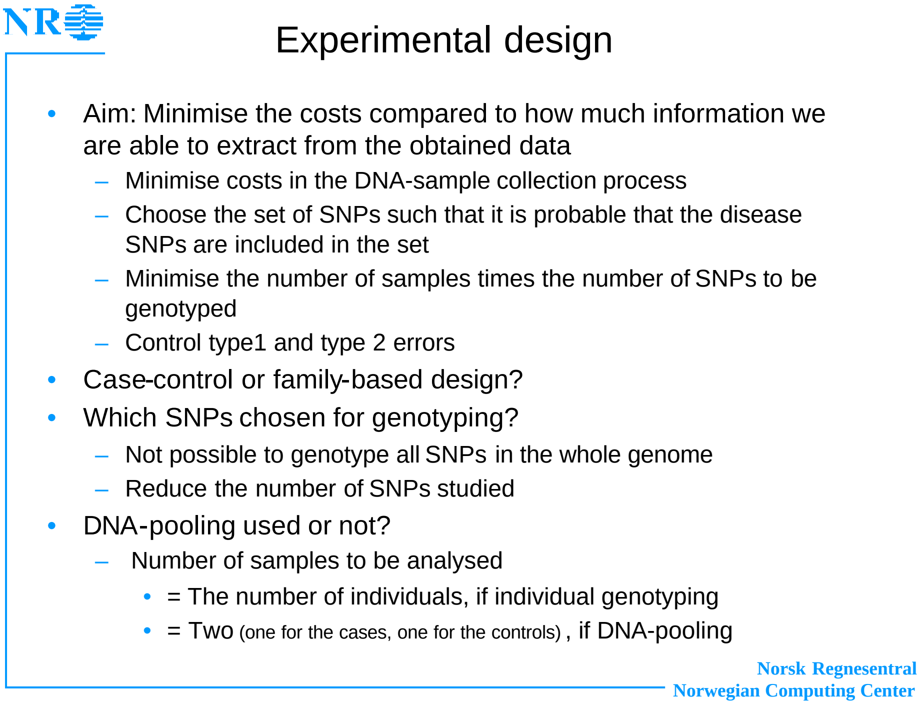


Association studies cont.

- Statistical methods often based on hypothesis testing
  - H_0: SNP is not associated with disease
  - H_1: SNP is associated with disease
  - Type 1 error
    - H_0 is rejected even if H_0 is true
      - Conclude that some SNPs are associated with disease even if they are not
    - Significance level of test – typically 5%
  - Type 2 error
    - H_0 is not rejected even if H_0 is false
      - Some SNPs which are associated with disease are not found
    - Power of the test – typically 80%
  - How many individuals do we need to genotype to obtain a certain significance level and power?
    - Dependent of experimental design and statistical model/method used
  - Many SNPs → many hypothesis tested simultaneously
    - Multiple testing

Norsk Regnesentral
Norwegian Computing Center

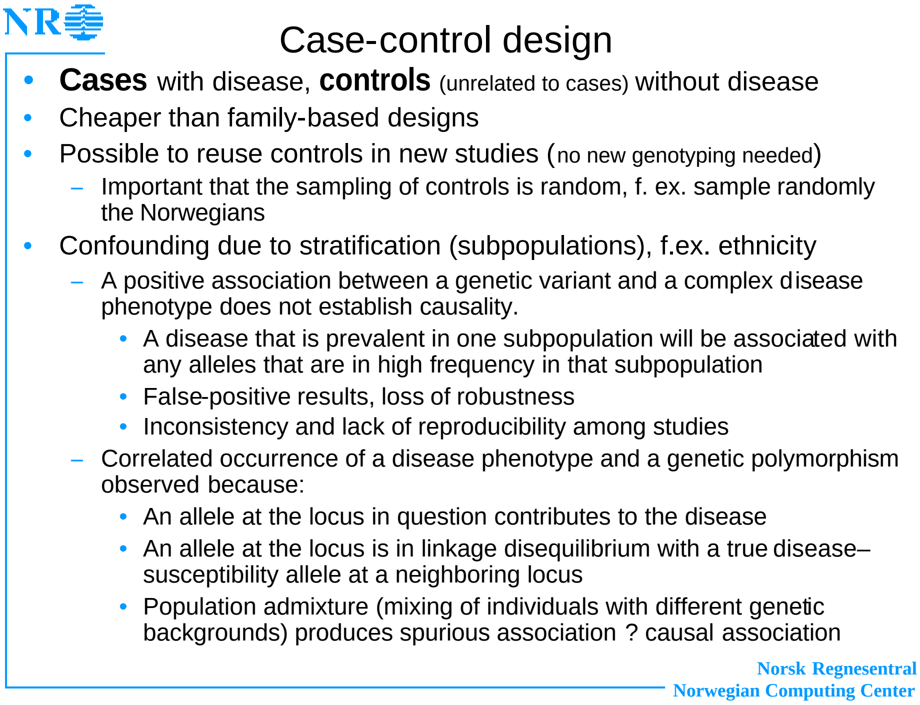
Experimental design



Experimental design

- Aim: Minimise the costs compared to how much information we are able to extract from the obtained data
  - Minimise costs in the DNA-sample collection process
  - Choose the set of SNPs such that it is probable that the disease SNPs are included in the set
  - Minimise the number of samples times the number of SNPs to be genotyped
  - Control type1 and type 2 errors
- Case-control or family-based design?
- Which SNPs chosen for genotyping?
  - Not possible to genotype all SNPs in the whole genome
  - Reduce the number of SNPs studied
- DNA-pooling used or not?
  - Number of samples to be analysed
    - = The number of individuals, if individual genotyping
    - = Two (one for the cases, one for the controls) , if DNA-pooling

Norsk Regnesentral
Norwegian Computing Center



Case-control design

- **Cases** with disease, **controls** (unrelated to cases) without disease
- Cheaper than family-based designs
- Possible to reuse controls in new studies (no new genotyping needed)
  - Important that the sampling of controls is random, f. ex. sample randomly the Norwegians
- Confounding due to stratification (subpopulations), f.ex. ethnicity
  - A positive association between a genetic variant and a complex disease phenotype does not establish causality.
    - A disease that is prevalent in one subpopulation will be associated with any alleles that are in high frequency in that subpopulation
    - False-positive results, loss of robustness
    - Inconsistency and lack of reproducibility among studies
  - Correlated occurrence of a disease phenotype and a genetic polymorphism observed because:
    - An allele at the locus in question contributes to the disease
    - An allele at the locus is in linkage disequilibrium with a true disease–susceptibility allele at a neighboring locus
    - Population admixture (mixing of individuals with different genetic backgrounds) produces spurious association ? causal association

Norsk Regnesentral
Norwegian Computing Center

NR

How to solve the stratification problem?

- Use family-based designs
  - Controls related to cases
- Use of random or unlinked genetic markers
  - Determine the extent of stratification by statistical analysis
  - Adjust for population stratification
  - Most efficient for large-scale genotyping
- Study multiple case-control populations (different ethnic groups)
  - Association seen in each population
- Data from nuclear families may be used to validate results from population-based association studies
- Matched case-control designs
  - Controls matched to cases on potential confounding factors like age, sex, ethnicity etc.

Norsk Regnesentral
Norwegian Computing Center

NR

Family-based designs

- Parents as controls
  - For each SNP
    - Test if an allele is transmitted to an affected offspring more or less often than expected by chance
  - Problem
    - Unavailability for late-onset disease
    - Some loss of power
  - Solution
    - Use sibs instead with even more loss of power
- Unaffected sibs as controls
  - Individuals in the same family are genetically related
  - Loss of power compared with a well-designed study involving unrelated controls
  - Sampling multiplex families
    - More than a single individual is affected
    - More efficient than sampling singletons

Norsk Regnesentral
Norwegian Computing Center

Which SNPs are chosen for genotyping?

- Using candidate genes or genome-wide scans
  - "The majority of publications reporting genetic studies of complex diseases investigate candidate genes and known metabolic pathways." (Peltonen et al. 2001)
- Use genome-wide random SNP approach?
  - No, many disease-causing genes would be missed
- Map-based or sequence-based approach? Yes, Botstein & Risch 2003
  - Argues for a genomic-scale sequence-based approach
    - Focus on SNPs in coding regions, disrupt splice sites and in promoter regions
      - most likely to be of functional significance and to influence directly the traits under study
  - Argues against a map based gene approach
    - Use haplotype information

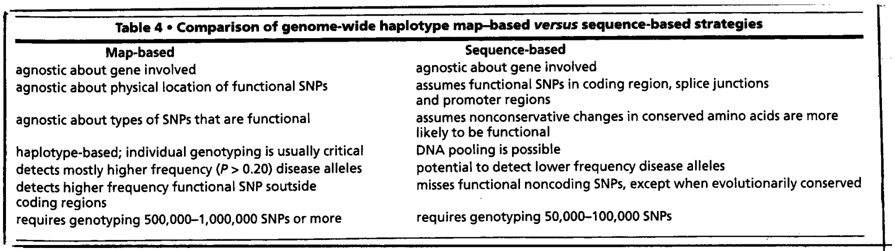


Table 4 • Comparison of genome-wide haplotype map–based *versus* sequence-based strategies

Map-based	Sequence-based
agnostic about gene involved	agnostic about gene involved
agnostic about physical location of functional SNPs	assumes functional SNPs in coding region, splice junctions and promoter regions
agnostic about types of SNPs that are functional	assumes nonconservative changes in conserved amino acids are more likely to be functional
haplotype-based; individual genotyping is usually critical	DNA pooling is possible
detects mostly higher frequency ($P > 0.20$) disease alleles	potential to detect lower frequency disease alleles
detects higher frequency functional SNP soutside coding regions	misses functional noncoding SNPs, except when evolutionarily conserved
requires genotyping 500,000–1,000,000 SNPs or more	requires genotyping 50,000–100,000 SNPs

The table is copied from the paper of D. Botstein and N. Risch, 2003 (see slides with literature overview)

Norsk Regnesentral
Norwegian Computing Center

Choosing SNPs in different populations

- Possible strategy
  - Under the assumption
    - Common genetic reason for all populations

1. In populations with high linkage disequilibrium
   - Initial detection of SNP associations
   - Coarse mapping
   - Ex.: Caucasian + Asian population
     - Two alleles equally associated with disease (+ in complete disequilibrium).
2. In populations with lower linkage disequilibrium
   - Which SNP is primary?
   - Fine mapping
   - Ex.:African:
     - The two alleles found above are not in complete disequilibrium
     - Find primary allele of these two

Norsk Regnesentral
Norwegian Computing Center

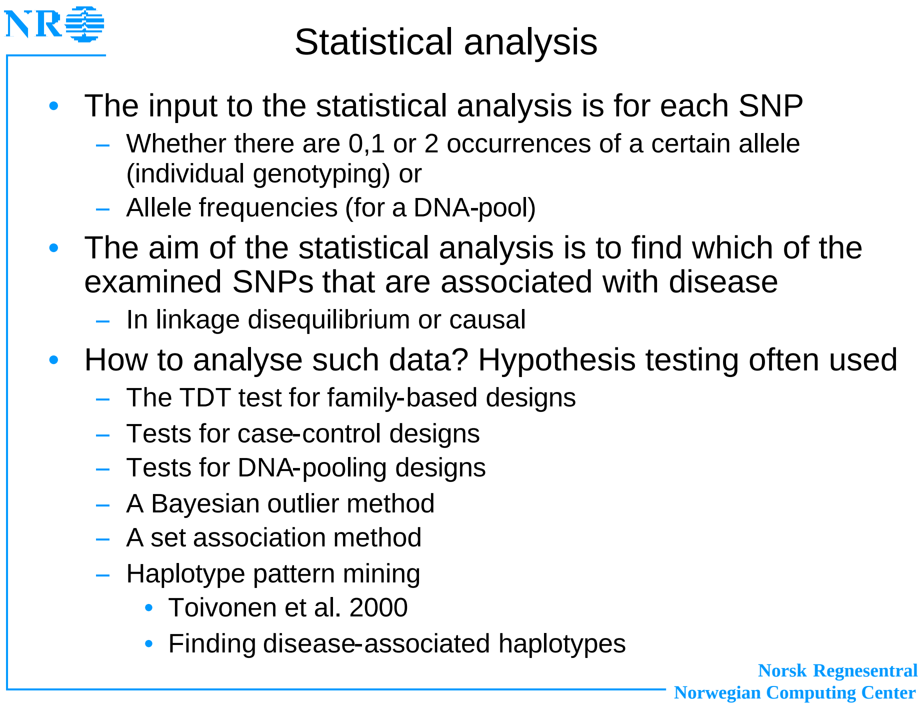
DNA-pooling

- Two pool designs
  - One sample for cases, one for controls (disease trait)
  - One pool for each of the two extremes (quantitative trait)
- Reduce the cost of large association studies, but
  - Result in loss of information
  - For quantitative traits
    - Allows examination of between-pool differences, but not within-pool differences
  - Additional experimental errors
- Efficiency of a DNA-pooling study = N_I / N_P
  - N_P is the number of individuals required to achieve same significance and power as a in a study that is based on individual genotyping with N_I individuals
  - Qualitative traits
    - Efficiency =1 in the absence of experimental error.
    - Four replicate measures are recommended for sufficient reduction in this error
  - Qualitative traits
    - Efficiency << 1

DNA-pooling cont.

- Pools to be compared should be well-balanced
  - Equal representation of sexes, age groups, ethnicities etc.
  - If risk factors for getting the disease, use four pools
    - Two pools for cases
      - One with high level exposure to risk factor, one with low
      - Similarly for controls
- Also possible to use multiple-pool designs to obtain average allele-frequency estimate from several pools and measurements
  - Reduce uncertainty of obtained allele-frequency estimate
- Two stage design most effective?
  - Find markers that show positive association in a pooling study
    - Cost saving
    - OK with some stratification and type1 error here
  - Follow up these markers by confirmatory individual genotyping
    - Full information = best accuracy
    - Avoid stratification and type1 error here

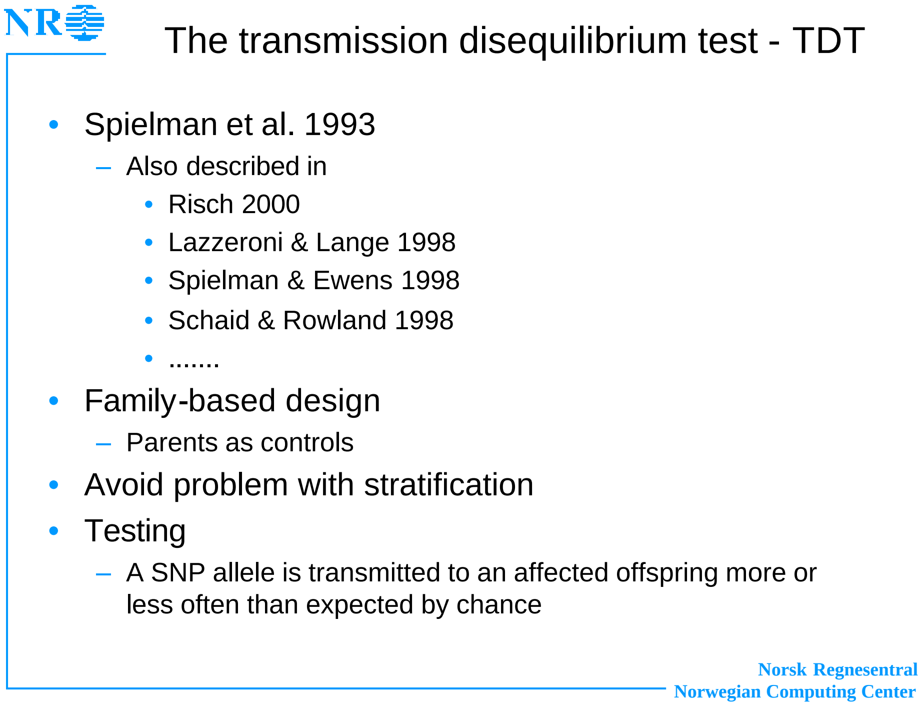
Statistical analysis



Statistical analysis

- The input to the statistical analysis is for each SNP
  - Whether there are 0,1 or 2 occurrences of a certain allele (individual genotyping) or
  - Allele frequencies (for a DNA-pool)
- The aim of the statistical analysis is to find which of the examined SNPs that are associated with disease
  - In linkage disequilibrium or causal
- How to analyse such data? Hypothesis testing often used
  - The TDT test for family-based designs
  - Tests for case-control designs
  - Tests for DNA-pooling designs
  - A Bayesian outlier method
  - A set association method
  - Haplotype pattern mining
    - Toivonen et al. 2000
    - Finding disease-associated haplotypes

Norsk Regnesentral
Norwegian Computing Center



The transmission disequilibrium test - TDT

- Spielman et al. 1993
  - Also described in
    - Risch 2000
    - Lazzeroni & Lange 1998
    - Spielman & Ewens 1998
    - Schaid & Rowland 1998
    - ....
- Family-based design
  - Parents as controls
- Avoid problem with stratification
- Testing
  - A SNP allele is transmitted to an affected offspring more or less often than expected by chance

Norsk Regnesentral
Norwegian Computing Center

TDT cont.

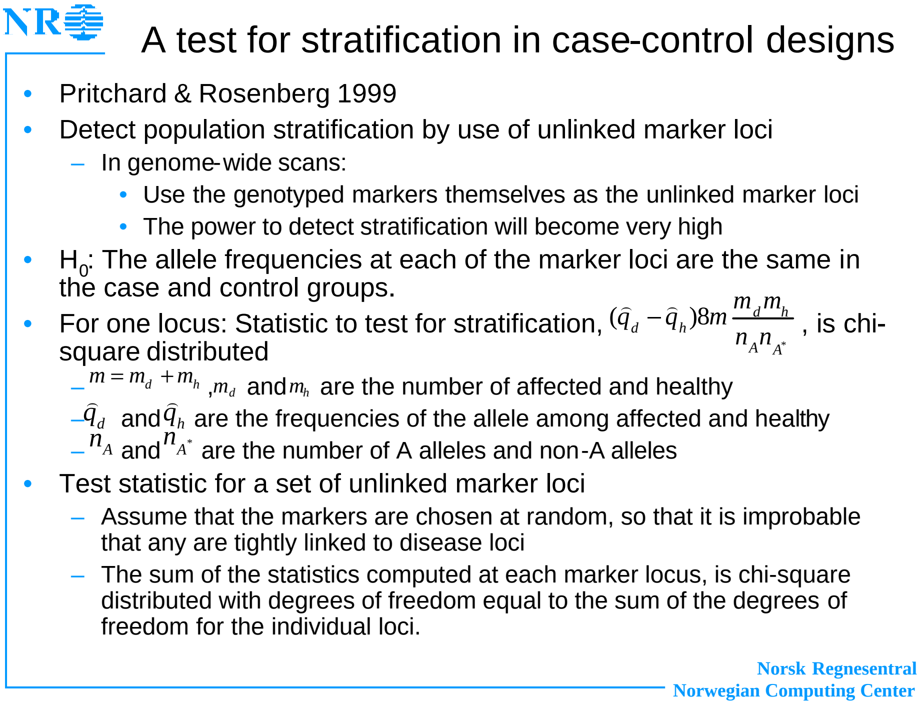
- Information used only from heterozygous parents
  - An allele transmitted by a parent to an effected child is matched to the other allele not transmitted from the same parent.
- For a biallelic locus, count the number of times allele 1 and 2 are transmitted to affected child.
  - $n_{i/j\to i}$: Number of times allele i is transmitted to affected child from a parent with one i and one j allele
  - $p_{i/j\to i}$: Probability that allele i is transmitted to affected child from a parent with one i and one j allele
  - H$_0$: $p_{1/2\to1} = p_{1/2\to2}$, H$_1$: $p_{1/2\to1} \neq p_{1/2\to2}$
  - Let $t_1 = n_{1/2\to1}$ and $t_2 = n_{1/2\to2}$
  - The test statistic $T_{MacNemar} = (t_1 - t_2)^2/(t_1 + t_2)$ follows an approximate chi-square distribution with one degree of freedom

Norsk Regnesentral
Norwegian Computing Center

Several variants/extensions of the TDT

- Lazzeroni & Lange 1998
  - Permutation extensions of the TDT to
    - Multiple alleles, multiple loci, unaffected siblings, genotypic rather than allelic associations
    - Correction for multiple tests more powerful than the standard Bonferroni correction
      - Monte Carlo approximation of p-values
      - Simultaneous TDT tests are conducted on haplotype data
- Spielman & Ewens 1998
  - Sib-TDT (S-TDT): statistic for unaffected sibs instead of parents
  - Statistic for combining S-TDT and TDT
    - Obtain “null” distribution by permuting the data (bootstrapping) for finding whether allele frequencies differ significantly
    - Software available
- Schaid & Rowland 1998
  - Parents, sibs, unrelated as controls
  - Extension of Spielman & Ewens 1998
- Teng and Risch 1999
  - TDT for sibs

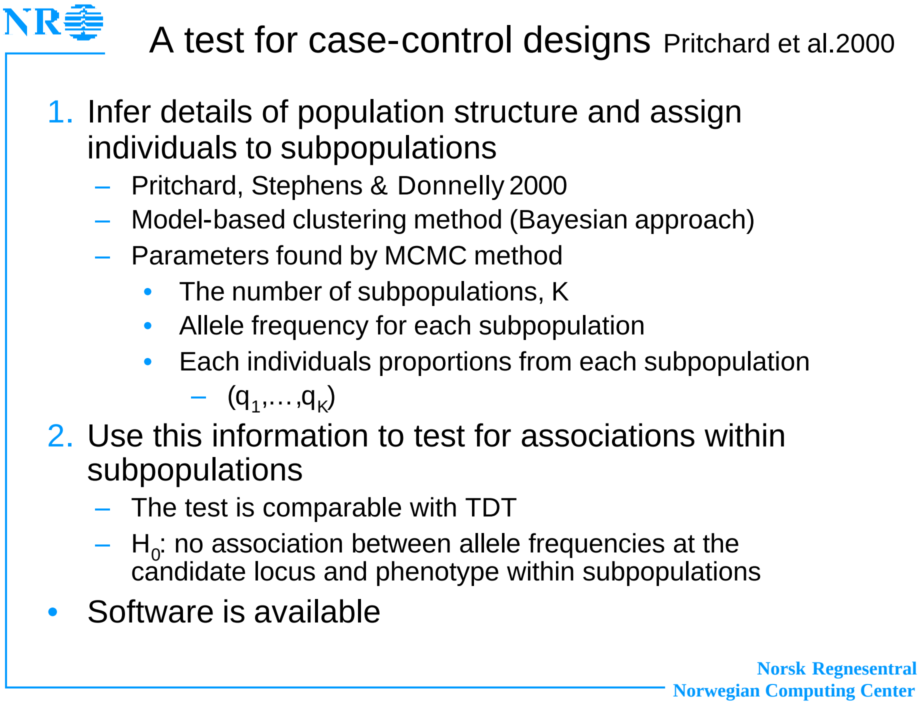
Norsk Regnesentral
Norwegian Computing Center



A test for stratification in case-control designs

- Pritchard & Rosenberg 1999
- Detect population stratification by use of unlinked marker loci
  - In genome-wide scans:
    - Use the genotyped markers themselves as the unlinked marker loci
    - The power to detect stratification will become very high
- H_0: The allele frequencies at each of the marker loci are the same in the case and control groups.
- For one locus: Statistic to test for stratification, $(\hat{q}_d - \hat{q}_h)8m\frac{m_d m_h}{n_A n_{A^*}}$, is chi-square distributed
  - $m = m_d + m_h$,m_d and m_h are the number of affected and healthy
  - $\hat{q}_d$ and $\hat{q}_h$ are the frequencies of the allele among affected and healthy
  - n_A and n_{A^*} are the number of A alleles and non-A alleles
- Test statistic for a set of unlinked marker loci
  - Assume that the markers are chosen at random, so that it is improbable that any are tightly linked to disease loci
  - The sum of the statistics computed at each marker locus, is chi-square distributed with degrees of freedom equal to the sum of the degrees of freedom for the individual loci.

Norsk Regnesentral
Norwegian Computing Center

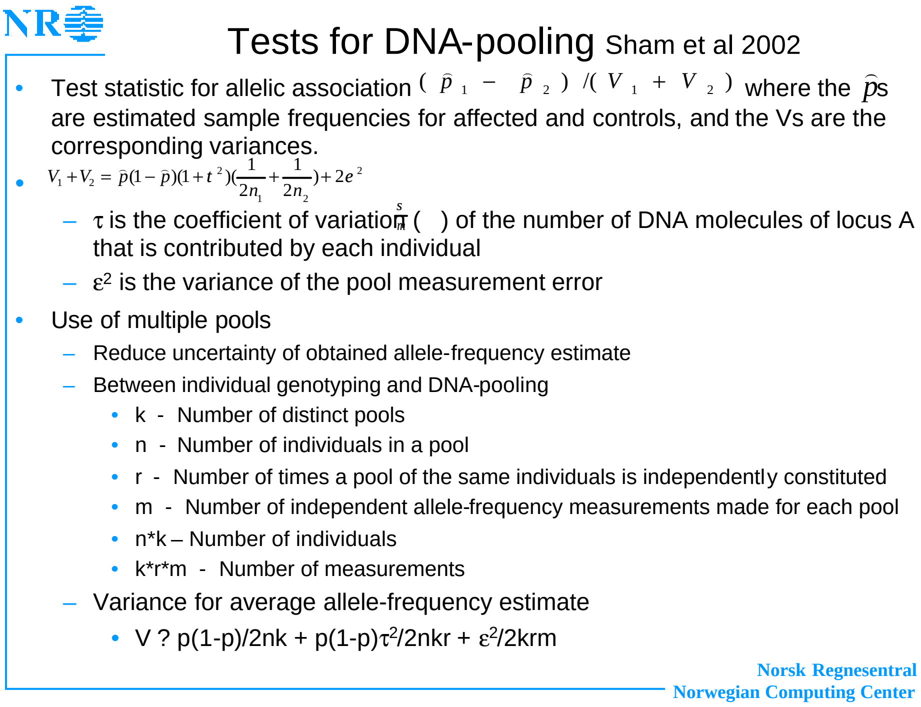


A test for case-control designs Pritchard et al.2000

1. Infer details of population structure and assign individuals to subpopulations
   - Pritchard, Stephens & Donnelly 2000
   - Model-based clustering method (Bayesian approach)
   - Parameters found by MCMC method
     - The number of subpopulations, K
     - Allele frequency for each subpopulation
     - Each individuals proportions from each subpopulation
       - $(q_1,\dots,q_K)$
2. Use this information to test for associations within subpopulations
   - The test is comparable with TDT
   - H_0: no association between allele frequencies at the candidate locus and phenotype within subpopulations

- Software is available

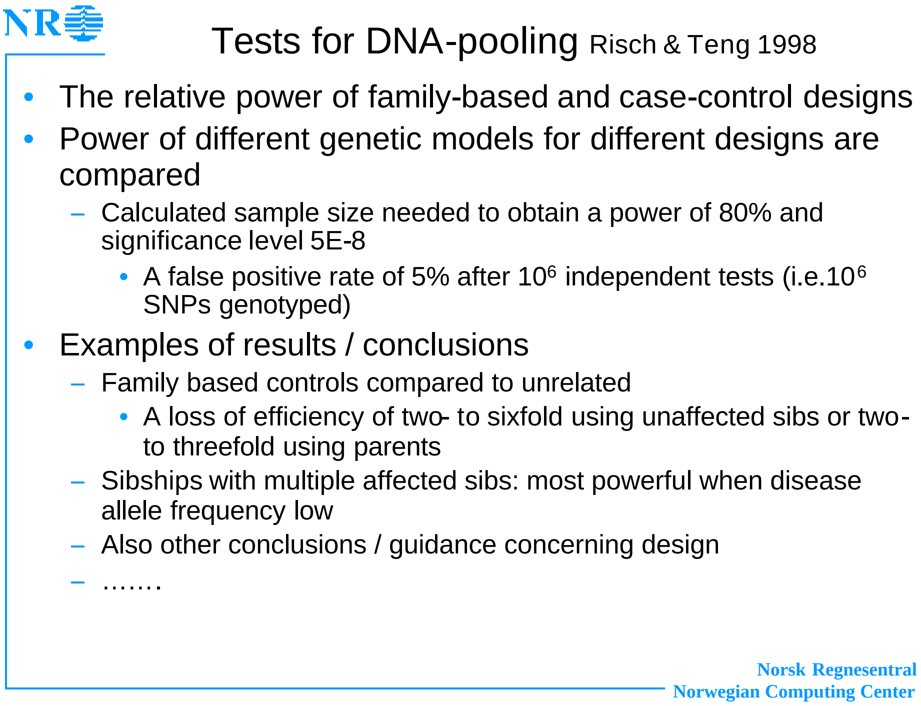
Norsk Regnesentral
Norwegian Computing Center



Tests for DNA-pooling Sham et al 2002

- Test statistic for allelic association $(\hat{p}_1 - \hat{p}_2)/(V_1 + V_2)$ where the $\hat{p}$s are estimated sample frequencies for affected and controls, and the Vs are the corresponding variances.
- $V_1 + V_2 = \hat{p}(1-\hat{p})(1+t^2)(\frac{1}{2n_1}+\frac{1}{2n_2})+2e^2$
  - τ is the coefficient of variation $\frac{s}{m}$ ( ) of the number of DNA molecules of locus A that is contributed by each individual
  - ε^2 is the variance of the pool measurement error
- Use of multiple pools
  - Reduce uncertainty of obtained allele-frequency estimate
  - Between individual genotyping and DNA-pooling
    - k - Number of distinct pools
    - n - Number of individuals in a pool
    - r - Number of times a pool of the same individuals is independently constituted
    - m - Number of independent allele-frequency measurements made for each pool
    - n*k – Number of individuals
    - k*r*m - Number of measurements
  - Variance for average allele-frequency estimate
    - V ? p(1-p)/2nk + p(1-p)τ^2/2nkr + ε^2/2krm

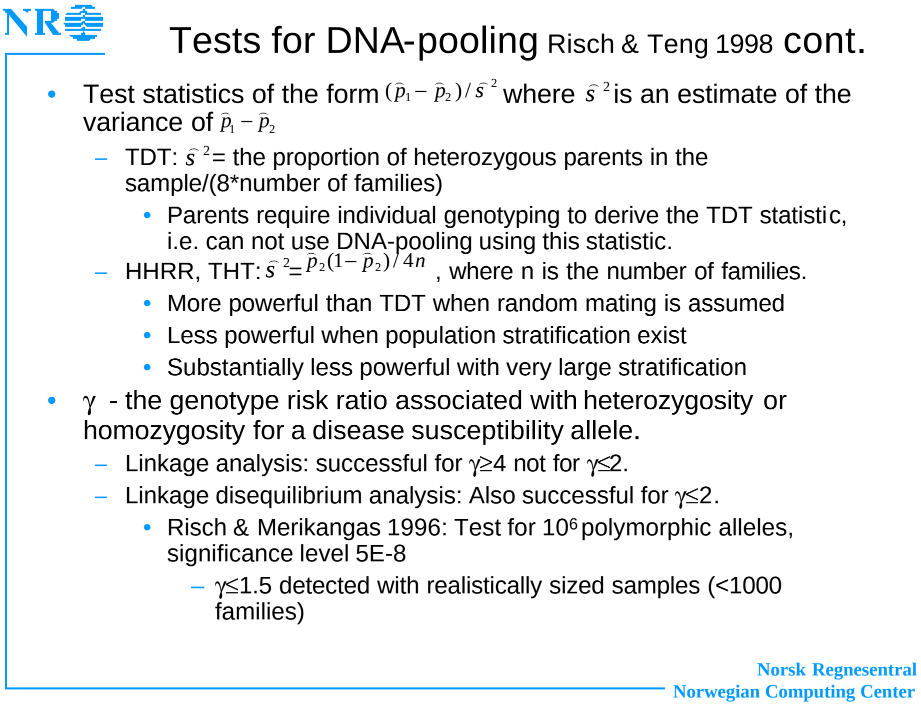
Norsk Regnesentral
Norwegian Computing Center



Tests for DNA-pooling Risch & Teng 1998

- The relative power of family-based and case-control designs
- Power of different genetic models for different designs are compared
  - Calculated sample size needed to obtain a power of 80% and significance level 5E-8
    - A false positive rate of 5% after 10^6 independent tests (i.e.10^6 SNPs genotyped)
- Examples of results / conclusions
  - Family based controls compared to unrelated
    - A loss of efficiency of two- to sixfold using unaffected sibs or two-to threefold using parents
  - Sibships with multiple affected sibs: most powerful when disease allele frequency low
  - Also other conclusions / guidance concerning design
  - …….

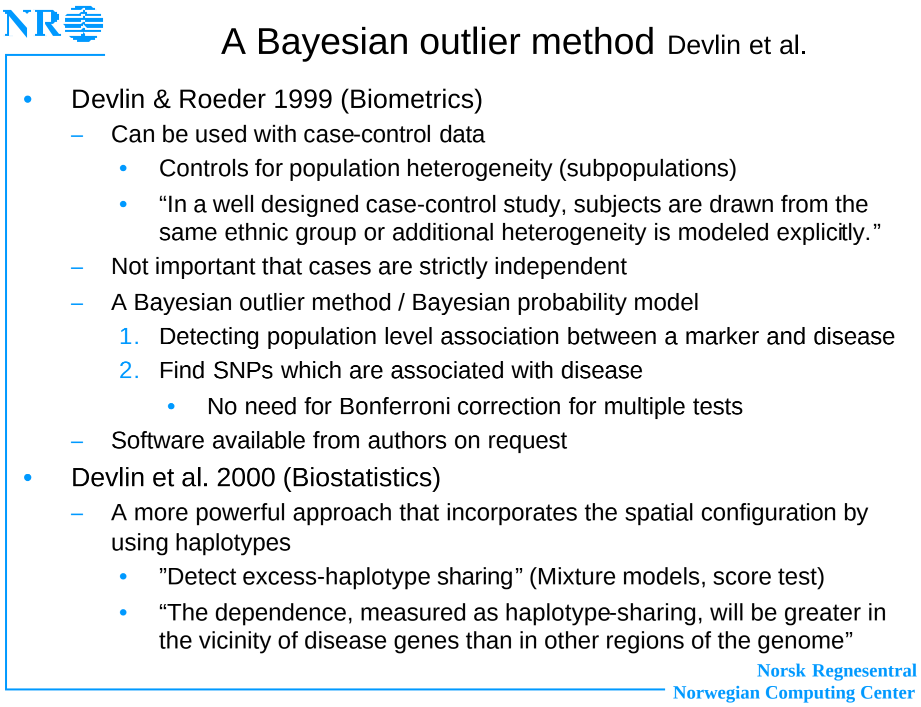
Norsk Regnesentral
Norwegian Computing Center



Tests for DNA-pooling Risch & Teng 1998 cont.

- Test statistics of the form $(\hat{p}_1 - \hat{p}_2)/\hat{s}^2$ where $\hat{s}^2$ is an estimate of the variance of $\hat{p}_1 - \hat{p}_2$
  - TDT: $\hat{s}^2$= the proportion of heterozygous parents in the sample/(8*number of families)
    - Parents require individual genotyping to derive the TDT statistic, i.e. can not use DNA-pooling using this statistic.
  - HHRR, THT: $\hat{s}^2 = \hat{p}_2(1-\hat{p}_2)/4n$, where n is the number of families.
    - More powerful than TDT when random mating is assumed
    - Less powerful when population stratification exist
    - Substantially less powerful with very large stratification
- γ - the genotype risk ratio associated with heterozygosity or homozygosity for a disease susceptibility allele.
  - Linkage analysis: successful for $\gamma \geq 4$ not for $\gamma \leq 2$.
  - Linkage disequilibrium analysis: Also successful for $\gamma \leq 2$.
    - Risch & Merikangas 1996: Test for 10^6 polymorphic alleles, significance level 5E-8
      - $\gamma \leq 1.5$ detected with realistically sized samples (<1000 families)

Norsk Regnesentral
Norwegian Computing Center



A Bayesian outlier method Devlin et al.

- Devlin & Roeder 1999 (Biometrics)
  - Can be used with case-control data
    - Controls for population heterogeneity (subpopulations)
    - “In a well designed case-control study, subjects are drawn from the same ethnic group or additional heterogeneity is modeled explicitly.”
  - Not important that cases are strictly independent
  - A Bayesian outlier method / Bayesian probability model
    1. Detecting population level association between a marker and disease
    2. Find SNPs which are associated with disease
       - No need for Bonferroni correction for multiple tests
  - Software available from authors on request
- Devlin et al. 2000 (Biostatistics)
  - A more powerful approach that incorporates the spatial configuration by using haplotypes
    - ”Detect excess-haplotype sharing” (Mixture models, score test)
    - “The dependence, measured as haplotype-sharing, will be greater in the vicinity of disease genes than in other regions of the genome”

Norsk Regnesentral
Norwegian Computing Center

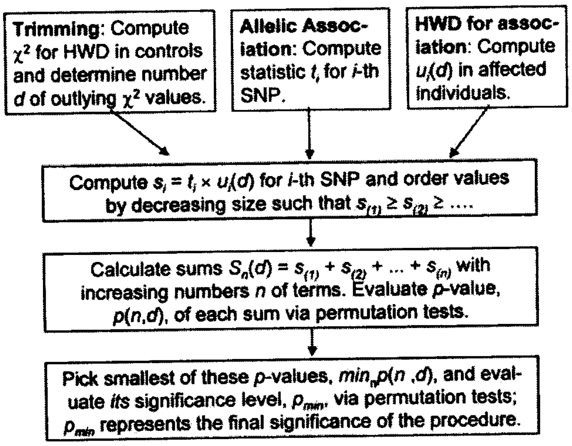
A set association method

- Hoh et al. 2001
  - Software available (Sumstat)
- Analyze markers jointly rather than testing each marker in isolation
  - Marker-by-marker approach completely ignores the multigenic nature of complex traits and does not take into account possible interactions between susceptibility genes.
  - Jorde 2000
    - “The incorporation of information from multiple loci can enhance the power and accuracy of LD mapping”
    - Several *Multilocus disequilibrium methods* described
  - Schaid & Rowland 1998
    - “Simulation result indicates that testing all alleles simultaneously is most powerful”
  - Risch 2000
    - “Many different genes, each with allelic variations, contributed to the total observed variability in a trait:
      - Application of the central limit theorem from statistics implicates a continuous normal distribution in the population for such a trait”
- Stratification not taken into account.
  - Methods described by others could be incorporated into the set association method

Norsk Regnesentral
Norwegian Computing Center

A set association method cont.

- A statistic for each marker is suitably chosen
- Application
  - 779 heart disease patients (342 cases showed restenosis, rest controls)
  - 89 SNPs in 62 candidate genes
- Ott&Hoh 2003
  - Apply the method on microarray data to find a set of differentially expressed genes



Trimming: Compute χ^2 for HWD in controls and determine number d of outlying χ^2 values.

Allelic Association: Compute statistic t_i for i-th SNP.

HWD for association: Compute $u_i(d)$ in affected individuals.

Compute $s_i = t_i \times u_i(d)$ for i-th SNP and order values by decreasing size such that $s_{(1)} \geq s_{(2)} \geq \ldots$.

Calculate sums $S_n(d) = s_{(1)} + s_{(2)} + \ldots + s_{(n)}$ with increasing numbers n of terms. Evaluate p-value, $p(n,d)$, of each sum via permutation tests.

Pick smallest of these p-values, $min_n p(n,d)$, and evaluate *its* significance level, p_{min}, via permutation tests; p_{min} represents the final significance of the procedure.

Figure 1 Flow diagram illustrating the algorithm implemented in the set-association approach.

The figure is copied from the paper of J. Ott, A. Wille and J. Hoh, 2001
(see slides with literature overview)

Norsk Regnesentral
Norwegian Computing Center

A review paper

Review by Hirschhorn et al 2002

- Review of genetic association studies and complex diseases (many references to previous studies)
  - Ignored reports on gene-gene and gene-environment interactions
  - 600 positive associations between common gene variants and disease
  - 268 genes, 133 common diseases or dichotomous traits.
- Discuss reasons for irreproducibility
  - Type1 error is an unlikely explanation
  - Stratification
  - Variable linkage disequilibrium between marker and disease allele in different populations
  - Population specific gene-gene or gene-environment interactions
  - Different sample sizes
- Suggest guidelines for performing and interpreting genetic association studies
  - Large studies and /or meta-analyses of multiple studies
  - Omission of small negative studies will bias the pooled data toward a positive result
    - Negative results also important

Norsk Regnesentral
Norwegian Computing Center

Review by Hirschhorn et al 2002 cont.

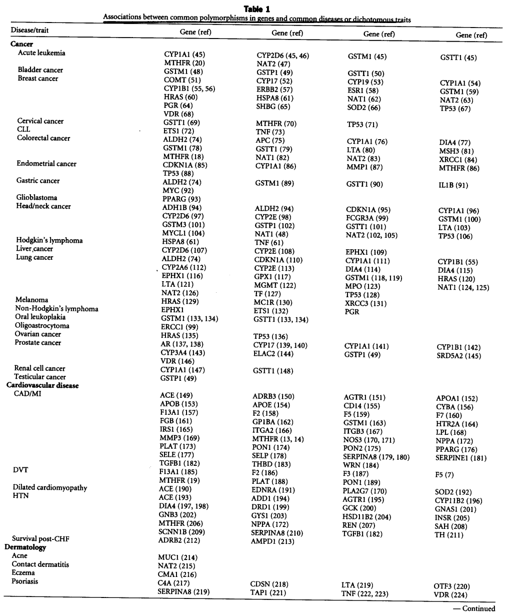


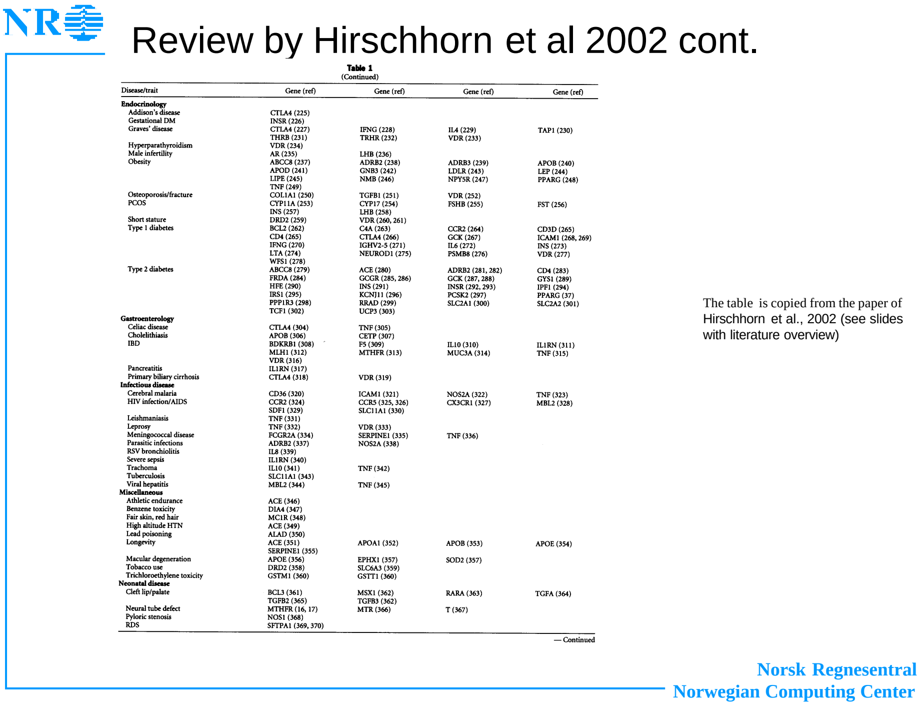
Table 1
Associations between common polymorphisms in genes and common diseases or dichotomous traits

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Cancer				
Acute leukemia	CYP1A1 (45)	CYP2D6 (45, 46)	GSTM1 (45)	GSTT1 (45)
	MTHFR (20)	NAT2 (47)		
Bladder cancer	GSTM1 (48)	GSTP1 (49)	GSTT1 (50)	
Breast cancer	COMT (51)	CYP17 (52)	CYP19 (53)	CYP1A1 (54)
	CYP1B1 (55, 56)	ERBB2 (57)	ESR1 (58)	GSTM1 (59)
	HRAS (60)	HSPA8 (61)	NAT1 (62)	NAT2 (63)
	PGR (64)	SHBG (65)	SOD2 (66)	TP53 (67)
	VDR (68)			
Cervical cancer	GSTT1 (69)	MTHFR (70)	TP53 (71)	
CLL	ETS1 (72)	TNF (73)		
Colorectal cancer	ALDH2 (74)	APC (75)	CYP1A1 (76)	DIA4 (77)
	GSTM1 (78)	GSTT1 (79)	LTA (80)	MSH3 (81)
	MTHFR (18)	NAT1 (82)	NAT2 (83)	XRCC1 (84)
Endometrial cancer	CDKN1A (85)	CYP1A1 (86)	MMP1 (87)	MTHFR (86)
	TP53 (88)			
Gastric cancer	ALDH2 (74)	GSTM1 (89)	GSTT1 (90)	IL1B (91)
	MYC (92)			
Glioblastoma	PPARG (93)			
Head/neck cancer	ADH1B (94)	ALDH2 (94)	CDKN1A (95)	CYP1A1 (96)
	CYP2D6 (97)	CYP2E (98)	FCGR3A (99)	GSTM1 (100)
	GSTM3 (101)	GSTP1 (102)	GSTT1 (101)	LTA (103)
	MYCL1 (104)	NAT1 (48)	NAT2 (102, 105)	TP53 (106)
Hodgkin's lymphoma	HSPA8 (61)	TNF (61)		
Liver,cancer	CYP2D6 (107)	CYP2E (108)	EPHX1 (109)	
Lung cancer	ALDH2 (74)	CDKN1A (110)	CYP1A1 (111)	CYP1B1 (55)
	CYP2A6 (112)	CYP2E (113)	DIA4 (114)	DIA4 (115)
	EPHX1 (116)	GPX1 (117)	GSTM1 (118, 119)	HRAS (120)
	LTA (121)	MGMT (122)	MPO (123)	NAT1 (124, 125)
	NAT2 (126)	TF (127)	TP53 (128)	
Melanoma	HRAS (129)	MC1R (130)	XRCC3 (131)	
Non-Hodgkin's lymphoma	EPHX1	ETS1 (132)	PGR	
Oral leukoplakia	GSTM1 (133, 134)	GSTT1 (133, 134)		
Oligoastrocytoma	ERCC1 (99)			
Ovarian cancer	HRAS (135)	TP53 (136)		
Prostate cancer	AR (137, 138)	CYP17 (139, 140)	CYP1A1 (141)	CYP1B1 (142)
	CYP3A4 (143)	ELAC2 (144)	GSTP1 (49)	SRD5A2 (145)
	VDR (146)			
Renal cell cancer	CYP1A1 (147)	GSTT1 (148)		
Testicular cancer	GSTP1 (49)			
Cardiovascular disease				
CAD/MI	ACE (149)	ADRB3 (150)	AGTR1 (151)	APOA1 (152)
	APOB (153)	APOE (154)	CD14 (155)	CYBA (156)
	F13A1 (157)	F2 (158)	F5 (159)	F7 (160)
	FGB (161)	GP1BA (162)	GSTM1 (163)	HTR2A (164)
	IRS1 (165)	ITGA2 (166)	ITGB3 (167)	LPL (168)
	MMP3 (169)	MTHFR (13, 14)	NOS3 (170, 171)	NPPA (172)
	PLAT (173)	PON1 (174)	PON2 (175)	PPARG (176)
	SELE (177)	SELP (178)	SERPINA8 (179, 180)	SERPINE1 (181)
	TGFB1 (182)	THBD (183)	WRN (184)	
DVT	F13A1 (185)	F2 (186)	F3 (187)	F5 (7)
	MTHFR (19)	PLAT (188)	PON1 (189)	
Dilated cardiomyopathy	ACE (190)	EDNRA (191)	PLA2G7 (170)	SOD2 (192)
HTN	ACE (193)	ADD1 (194)	AGTR1 (195)	CYP11B2 (196)
	DIA4 (197, 198)	DRD1 (199)	GCK (200)	GNAS1 (201)
	GNB3 (202)	GYS1 (203)	HSD11B2 (204)	INSR (205)
	MTHFR (206)	NPPA (172)	REN (207)	SAH (208)
	SCNN1B (209)	SERPINA8 (210)	TGFB1 (182)	TH (211)
Survival post-CHF	ADRB2 (212)	AMPD1 (213)		
Dermatology				
Acne	MUC1 (214)			
Contact dermatitis	NAT2 (215)			
Eczema	CMA1 (216)			
Psoriasis	C4A (217)	CDSN (218)	LTA (219)	OTF3 (220)
	SERPINA8 (219)	TAP1 (221)	TNF (222, 223)	VDR (224)

— Continued

The table is copied from the paper of Hirschhorn et al., 2002 (see slides with literature overview)

Norsk Regnesentral
Norwegian Computing Center



Review by Hirschhorn et al 2002 cont.

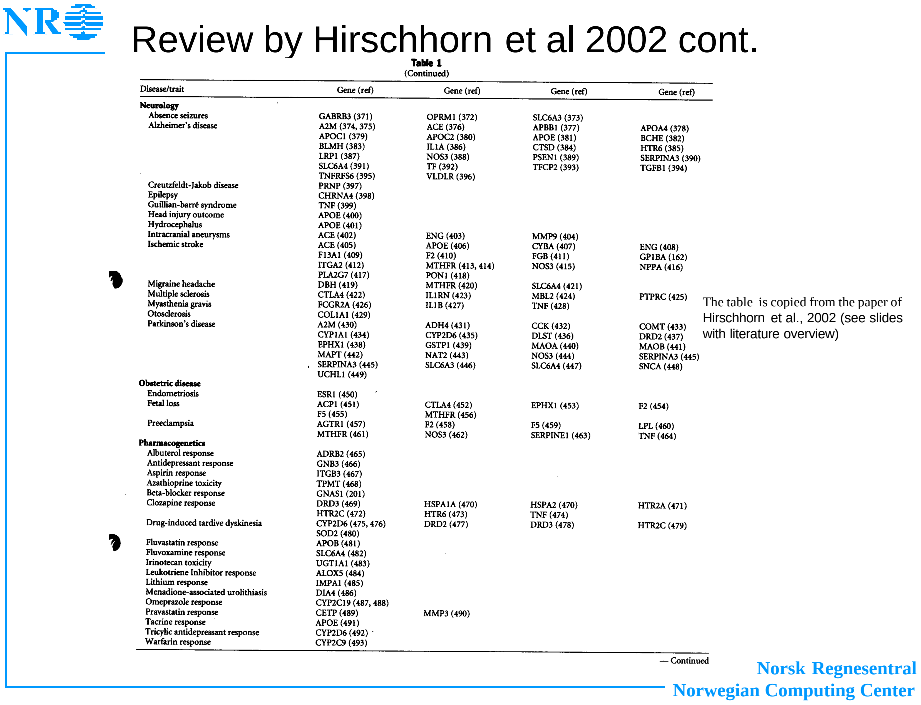
Table 1
(Continued)

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Endocrinology				
Addison's disease	CTLA4 (225)			
Gestational DM	INSR (226)			
Graves' disease	CTLA4 (227)	IFNG (228)	IL4 (229)	TAP1 (230)
	THRB (231)	TRHR (232)	VDR (233)	
Hyperparathyroidism	VDR (234)			
Male infertility	AR (235)	LHB (236)		
Obesity	ABCC8 (237)	ADRB2 (238)	ADRB3 (239)	APOB (240)
	APOD (241)	GNB3 (242)	LDLR (243)	LEP (244)
	LIPE (245)	NMB (246)	NPY5R (247)	PPARG (248)
	TNF (249)			
Osteoporosis/fracture	COL1A1 (250)	TGFB1 (251)	VDR (252)	
PCOS	CYP11A (253)	CYP17 (254)	FSHB (255)	FST (256)
	INS (257)	LHB (258)		
Short stature	DRD2 (259)	VDR (260, 261)		
Type 1 diabetes	BCL2 (262)	C4A (263)	CCR2 (264)	CD3D (265)
	CD4 (265)	CTLA4 (266)	GCK (267)	ICAM1 (268, 269)
	IFNG (270)	IGHV2-5 (271)	IL6 (272)	INS (273)
	LTA (274)	NEUROD1 (275)	PSMB8 (276)	VDR (277)
	WFS1 (278)			
Type 2 diabetes	ABCC8 (279)	ACE (280)	ADRB2 (281, 282)	CD4 (283)
	FRDA (284)	GCGR (285, 286)	GCK (287, 288)	GYS1 (289)
	HFE (290)	INS (291)	INSR (292, 293)	IPF1 (294)
	IRS1 (295)	KCNJ11 (296)	PCSK2 (297)	PPARG (37)
	PPP1R3 (298)	RRAD (299)	SLC2A1 (300)	SLC2A2 (301)
	TCF1 (302)	UCP3 (303)		
Gastroenterology				
Celiac disease	CTLA4 (304)	TNF (305)		
Cholelithiasis	APOB (306)	CETP (307)		
IBD	BDKRB1 (308)	F5 (309)	IL10 (310)	IL1RN (311)
	MLH1 (312)	MTHFR (313)	MUC3A (314)	TNF (315)
	VDR (316)			
Pancreatitis	IL1RN (317)			
Primary biliary cirrhosis	CTLA4 (318)	VDR (319)		
Infectious disease				
Cerebral malaria	CD36 (320)	ICAM1 (321)	NOS2A (322)	TNF (323)
HIV infection/AIDS	CCR2 (324)	CCR5 (325, 326)	CX3CR1 (327)	MBL2 (328)
	SDF1 (329)	SLC11A1 (330)		
Leishmaniasis	TNF (331)			
Leprosy	TNF (332)	VDR (333)		
Meningococcal disease	FCGR2A (334)	SERPINE1 (335)	TNF (336)	
Parasitic infections	ADRB2 (337)	NOS2A (338)		
RSV bronchiolitis	IL8 (339)			
Severe sepsis	IL1RN (340)			
Trachoma	IL10 (341)	TNF (342)		
Tuberculosis	SLC11A1 (343)			
Viral hepatitis	MBL2 (344)	TNF (345)		
Miscellaneous				
Athletic endurance	ACE (346)			
Benzene toxicity	DIA4 (347)			
Fair skin, red hair	MC1R (348)			
High altitude HTN	ACE (349)			
Lead poisoning	ALAD (350)			
Longevity	ACE (351)	APOA1 (352)	APOB (353)	APOE (354)
	SERPINE1 (355)			
Macular degeneration	APOE (356)	EPHX1 (357)	SOD2 (357)	
Tobacco use	DRD2 (358)	SLC6A3 (359)		
Trichloroethylene toxicity	GSTM1 (360)	GSTT1 (360)		
Neonatal disease				
Cleft lip/palate	BCL3 (361)	MSX1 (362)	RARA (363)	TGFA (364)
	TGFB2 (365)	TGFB3 (362)		
Neural tube defect	MTHFR (16, 17)	MTR (366)	T (367)	
Pyloric stenosis	NOS1 (368)			
RDS	SFTPA1 (369, 370)			

— Continued

The table is copied from the paper of Hirschhorn et al., 2002 (see slides with literature overview)

Norsk Regnesentral
Norwegian Computing Center



Review by Hirschhorn et al 2002 cont.

Table 1
(Continued)

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Neurology				
Absence seizures	GABRB3 (371)	OPRM1 (372)	SLC6A3 (373)	
Alzheimer's disease	A2M (374, 375)	ACE (376)	APBB1 (377)	APOA4 (378)
	APOC1 (379)	APOC2 (380)	APOE (381)	BCHE (382)
	BLMH (383)	IL1A (386)	CTSD (384)	HTR6 (385)
	LRP1 (387)	NOS3 (388)	PSEN1 (389)	SERPINA3 (390)
	SLC6A4 (391)	TF (392)	TFCP2 (393)	TGFB1 (394)
	TNFRFS6 (395)	VLDLR (396)		
Creutzfeldt-Jakob disease	PRNP (397)			
Epilepsy	CHRNA4 (398)			
Guillian-barré syndrome	TNF (399)			
Head injury outcome	APOE (400)			
Hydrocephalus	APOE (401)			
Intracranial aneurysms	ACE (402)	ENG (403)	MMP9 (404)	
Ischemic stroke	ACE (405)	APOE (406)	CYBA (407)	ENG (408)
	F13A1 (409)	F2 (410)	FGB (411)	GP1BA (162)
	ITGA2 (412)	MTHFR (413, 414)	NOS3 (415)	NPPA (416)
	PLA2G7 (417)	PON1 (418)		
Migraine headache	DBH (419)	MTHFR (420)	SLC6A4 (421)	
Multiple sclerosis	CTLA4 (422)	IL1RN (423)	MBL2 (424)	PTPRC (425)
Myasthenia gravis	FCGR2A (426)	IL1B (427)	TNF (428)	
Otosclerosis	COL1A1 (429)			
Parkinson's disease	A2M (430)	ADH4 (431)	CCK (432)	COMT (433)
	CYP1A1 (434)	CYP2D6 (435)	DLST (436)	DRD2 (437)
	EPHX1 (438)	GSTP1 (439)	MAOA (440)	MAOB (441)
	MAPT (442)	NAT2 (443)	NOS3 (444)	SERPINA3 (445)
	SERPINA3 (445)	SLC6A3 (446)	SLC6A4 (447)	SNCA (448)
	UCHL1 (449)			
Obstetric disease				
Endometriosis	ESR1 (450)			
Fetal loss	ACP1 (451)	CTLA4 (452)	EPHX1 (453)	F2 (454)
	F5 (455)	MTHFR (456)		
Preeclampsia	AGTR1 (457)	F2 (458)	F5 (459)	LPL (460)
	MTHFR (461)	NOS3 (462)	SERPINE1 (463)	TNF (464)
Pharmacogenetics				
Albuterol response	ADRB2 (465)			
Antidepressant response	GNB3 (466)			
Aspirin response	ITGB3 (467)			
Azathioprine toxicity	TPMT (468)			
Beta-blocker response	GNAS1 (201)			
Clozapine response	DRD3 (469)	HSPA1A (470)	HSPA2 (470)	HTR2A (471)
	HTR2C (472)	HTR6 (473)	TNF (474)	
Drug-induced tardive dyskinesia	CYP2D6 (475, 476)	DRD2 (477)	DRD3 (478)	HTR2C (479)
	SOD2 (480)			
Fluvastatin response	APOB (481)			
Fluvoxamine response	SLC6A4 (482)			
Irinotecan toxicity	UGT1A1 (483)			
Leukotriene Inhibitor response	ALOX5 (484)			
Lithium response	IMPA1 (485)			
Menadione-associated urolithiasis	DIA4 (486)			
Omeprazole response	CYP2C19 (487, 488)			
Pravastatin response	CETP (489)	MMP3 (490)		
Tacrine response	APOE (491)			
Tricylic antidepressant response	CYP2D6 (492)			
Warfarin response	CYP2C9 (493)			

— Continued

The table is copied from the paper of Hirschhorn et al., 2002 (see slides with literature overview)

Norsk Regnesentral
Norwegian Computing Center

Review by Hirschhorn et al 2002 cont.

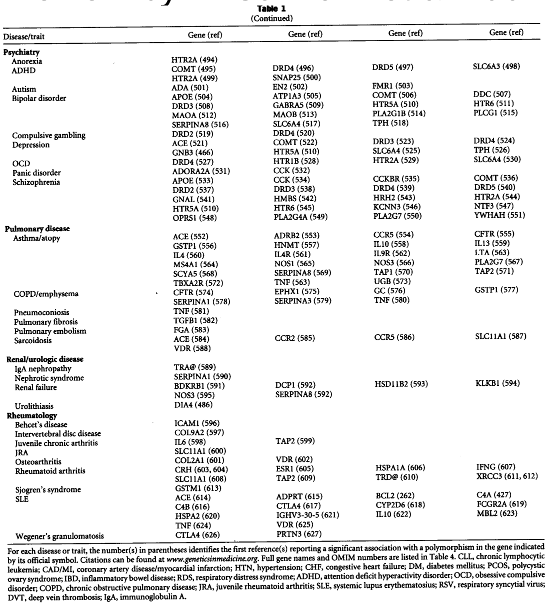


Table 1
(Continued)

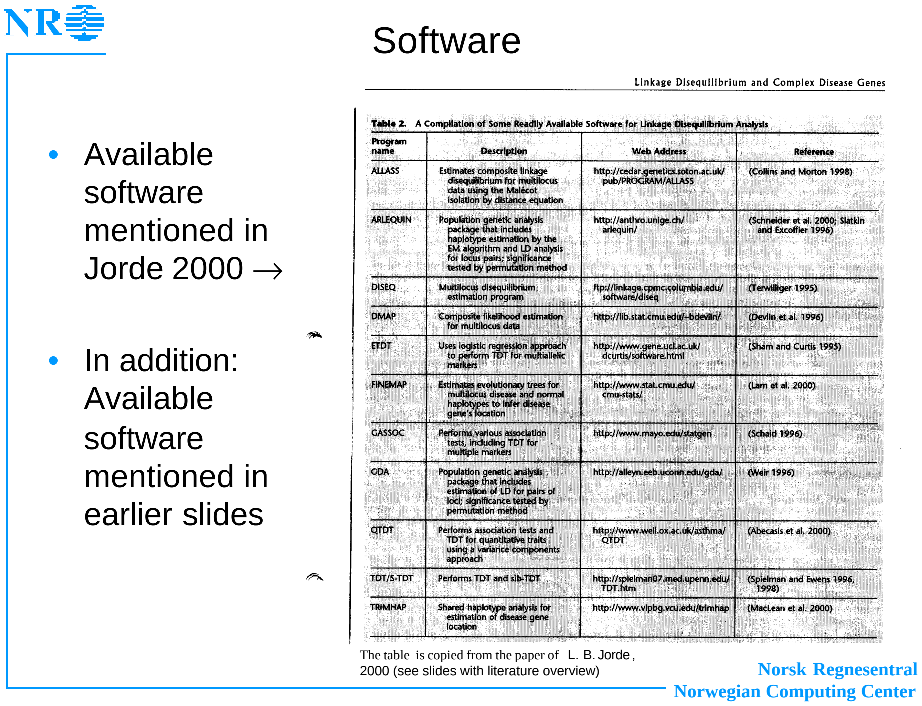
Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Psychiatry				
Anorexia	HTR2A (494)			
ADHD	COMT (495)	DRD4 (496)	DRD5 (497)	SLC6A3 (498)
	HTR2A (499)	SNAP25 (500)		
Autism	ADA (501)	EN2 (502)	FMR1 (503)	
Bipolar disorder	APOE (504)	ATP1A3 (505)	COMT (506)	DDC (507)
	DRD3 (508)	GABRA5 (509)	HTR5A (510)	HTR6 (511)
	MAOA (512)	MAOB (513)	PLA2G1B (514)	PLCG1 (515)
	SERPINA8 (516)	SLC6A4 (517)	TPH (518)	
Compulsive gambling	DRD2 (519)	DRD4 (520)		
Depression	ACE (521)	COMT (522)	DRD3 (523)	DRD4 (524)
	GNB3 (466)	HTR5A (510)	SLC6A4 (525)	TPH (526)
OCD	DRD4 (527)	HTR1B (528)	HTR2A (529)	SLC6A4 (530)
Panic disorder	ADORA2A (531)	CCK (532)		
Schizophrenia	APOE (533)	CCK (534)	CCKBR (535)	COMT (536)
	DRD2 (537)	DRD3 (538)	DRD4 (539)	DRD5 (540)
	GNAL (541)	HMBS (542)	HRH2 (543)	HTR2A (544)
	HTR5A (510)	HTR6 (545)	KCNN3 (546)	NTF3 (547)
	OPRS1 (548)	PLA2G4A (549)	PLA2G7 (550)	YWHAH (551)
Pulmonary disease				
Asthma/atopy	ACE (552)	ADRB2 (553)	CCR5 (554)	CFTR (555)
	GSTP1 (556)	HNMT (557)	IL10 (558)	IL13 (559)
	IL4 (560)	IL4R (561)	IL9R (562)	LTA (563)
	MS4A1 (564)	NOS1 (565)	NOS3 (566)	PLA2G7 (567)
	SCYA5 (568)	SERPINA8 (569)	TAP1 (570)	TAP2 (571)
	TBXA2R (572)	TNF (563)	UGB (573)	
COPD/emphysema	CFTR (574)	EPHX1 (575)	GC (576)	GSTP1 (577)
	SERPINA1 (578)	SERPINA3 (579)	TNF (580)	
Pneumoconiosis	TNF (581)			
Pulmonary fibrosis	TGFB1 (582)			
Pulmonary embolism	FGA (583)			
Sarcoidosis	ACE (584)	CCR2 (585)	CCR5 (586)	SLC11A1 (587)
	VDR (588)			
Renal/urologic disease				
IgA nephropathy	TRA@ (589)			
Nephrotic syndrome	SERPINA1 (590)			
Renal failure	BDKRB1 (591)	DCP1 (592)	HSD11B2 (593)	KLKB1 (594)
	NOS3 (595)	SERPINA8 (592)		
Urolithiasis	DIA4 (486)			
Rheumatology				
Behcet's disease	ICAM1 (596)			
Intervertebral disc disease	COL9A2 (597)			
Juvenile chronic arthritis	IL6 (598)	TAP2 (599)		
JRA	SLC11A1 (600)			
Osteoarthritis	COL2A1 (601)	VDR (602)		
Rheumatoid arthritis	CRH (603, 604)	ESR1 (605)	HSPA1A (606)	IFNG (607)
	SLC11A1 (608)	TAP2 (609)	TRD@ (610)	XRCC3 (611, 612)
Sjogren's syndrome	GSTM1 (613)			
SLE	ACE (614)	ADPRT (615)	BCL2 (262)	C4A (427)
	C4B (616)	CTLA4 (617)	CYP2D6 (618)	FCGR2A (619)
	HSPA2 (620)	IGHV3-30-5 (621)	IL10 (622)	MBL2 (623)
	TNF (624)	VDR (625)		
Wegener's granulomatosis	CTLA4 (626)	PRTN3 (627)		

For each disease or trait, the number(s) in parentheses identifies the first reference(s) reporting a significant association with a polymorphism in the gene indicated by its official symbol. Citations can be found at *www.geneticsinmedicine.org*. Full gene names and OMIM numbers are listed in Table 4. CLL, chronic lymphocytic leukemia; CAD/MI, coronary artery disease/myocardial infarction; HTN, hypertension; CHF, congestive heart failure; DM, diabetes mellitus; PCOS, polycystic ovary syndrome; IBD, inflammatory bowel disease; RDS, respiratory distress syndrome; ADHD, attention deficit hyperactivity disorder; OCD, obsessive compulsive disorder; COPD, chronic obstructive pulmonary disease; JRA, juvenile rheumatoid arthritis; SLE, systemic lupus erythematosius; RSV, respiratory syncytial virus; DVT, deep vein thrombosis; IgA, immunoglobulin A.

The table is copied from the paper of Hirschhorn et al., 2002 (see slides with literature overview)

Norsk Regnesentral
Norwegian Computing Center

Software



Software

- Available software mentioned in Jorde 2000 →
- In addition: Available software mentioned in earlier slides

Linkage Disequilibrium and Complex Disease Genes

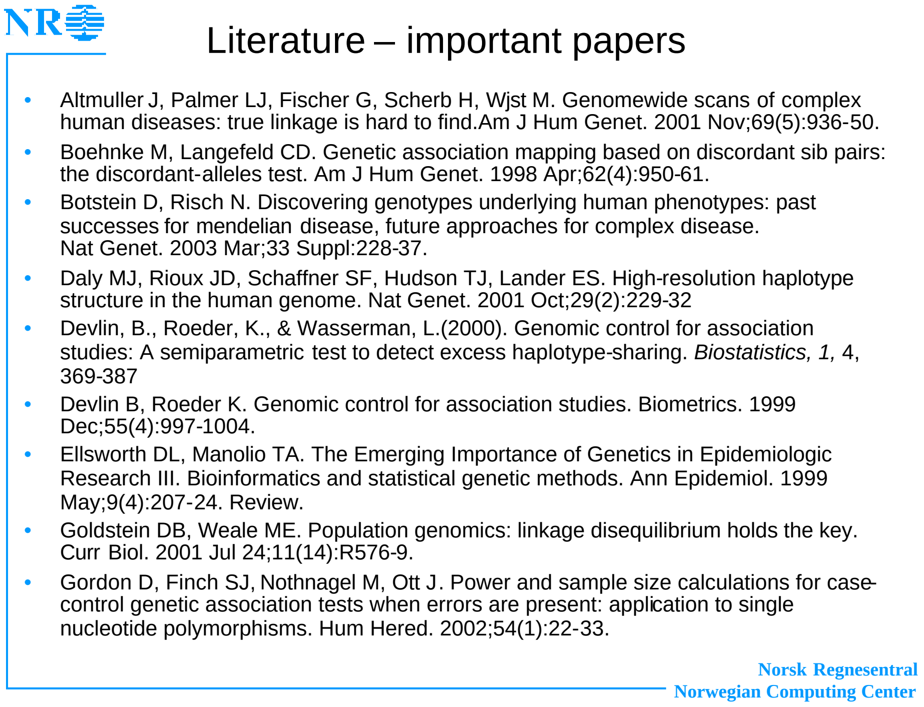
Table 2. A Compilation of Some Readily Available Software for Linkage Disequilibrium Analysis

Program name	Description	Web Address	Reference
ALLASS	Estimates composite linkage disequilibrium for multilocus data using the Malécot isolation by distance equation	http://cedar.genetics.soton.ac.uk/pub/PROGRAM/ALLASS	(Collins and Morton 1998)
ARLEQUIN	Population genetic analysis package that includes haplotype estimation by the EM algorithm and LD analysis for locus pairs; significance tested by permutation method	http://anthro.unige.ch/arlequin/	(Schneider et al. 2000; Slatkin and Excoffier 1996)
DISEQ	Multilocus disequilibrium estimation program	ftp://linkage.cpmc.columbia.edu/software/diseq	(Terwilliger 1995)
DMAP	Composite likelihood estimation for multilocus data	http://lib.stat.cmu.edu/~bdevlin/	(Devlin et al. 1996)
ETDT	Uses logistic regression approach to perform TDT for multiallelic markers	http://www.gene.ucl.ac.uk/dcurtis/software.html	(Sham and Curtis 1995)
FINEMAP	Estimates evolutionary trees for multilocus disease and normal haplotypes to infer disease gene's location	http://www.stat.cmu.edu/cmu-stats/	(Lam et al. 2000)
GASSOC	Performs various association tests, including TDT for multiple markers	http://www.mayo.edu/statgen	(Schaid 1996)
GDA	Population genetic analysis package that includes estimation of LD for pairs of loci; significance tested by permutation method	http://alleyn.eeb.uconn.edu/gda/	(Weir 1996)
QTDT	Performs association tests and TDT for quantitative traits using a variance components approach	http://www.well.ox.ac.uk/asthma/QTDT	(Abecasis et al. 2000)
TDT/S-TDT	Performs TDT and sib-TDT	http://spielman07.med.upenn.edu/TDT.htm	(Spielman and Ewens 1996, 1998)
TRIMHAP	Shared haplotype analysis for estimation of disease gene location	http://www.vipbg.vcu.edu/trimhap	(MacLean et al. 2000)

The table is copied from the paper of L. B. Jorde, 2000 (see slides with literature overview)

Norsk Regnesentral
Norwegian Computing Center

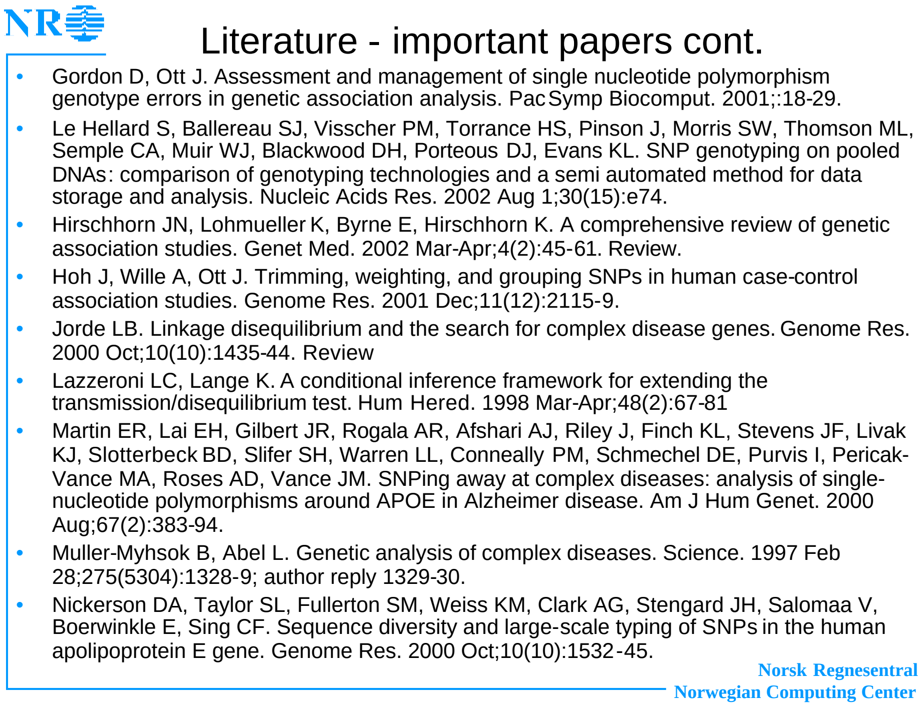
Literature



Literature – important papers

- Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide scans of complex human diseases: true linkage is hard to find.Am J Hum Genet. 2001 Nov;69(5):936-50.
- Boehnke M, Langefeld CD. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet. 1998 Apr;62(4):950-61.
- Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003 Mar;33 Suppl:228-37.
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet. 2001 Oct;29(2):229-32
- Devlin, B., Roeder, K., & Wasserman, L.(2000). Genomic control for association studies: A semiparametric test to detect excess haplotype-sharing. *Biostatistics, 1,* 4, 369-387
- Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999 Dec;55(4):997-1004.
- Ellsworth DL, Manolio TA. The Emerging Importance of Genetics in Epidemiologic Research III. Bioinformatics and statistical genetic methods. Ann Epidemiol. 1999 May;9(4):207-24. Review.
- Goldstein DB, Weale ME. Population genomics: linkage disequilibrium holds the key. Curr Biol. 2001 Jul 24;11(14):R576-9.
- Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered. 2002;54(1):22-33.

Norsk Regnesentral
Norwegian Computing Center



Literature - important papers cont.

- Gordon D, Ott J. Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput. 2001;:18-29.
- Le Hellard S, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple CA, Muir WJ, Blackwood DH, Porteous DJ, Evans KL. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. Nucleic Acids Res. 2002 Aug 1;30(15):e74.
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002 Mar-Apr;4(2):45-61. Review.
- Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res. 2001 Dec;11(12):2115-9.
- Jorde LB. Linkage disequilibrium and the search for complex disease genes. Genome Res. 2000 Oct;10(10):1435-44. Review
- Lazzeroni LC, Lange K. A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered. 1998 Mar-Apr;48(2):67-81
- Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am J Hum Genet. 2000 Aug;67(2):383-94.
- Muller-Myhsok B, Abel L. Genetic analysis of complex diseases. Science. 1997 Feb 28;275(5304):1328-9; author reply 1329-30.
- Nickerson DA, Taylor SL, Fullerton SM, Weiss KM, Clark AG, Stengard JH, Salomaa V, Boerwinkle E, Sing CF. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. Genome Res. 2000 Oct;10(10):1532-45.

Norsk Regnesentral
Norwegian Computing Center

Literature - important papers cont.

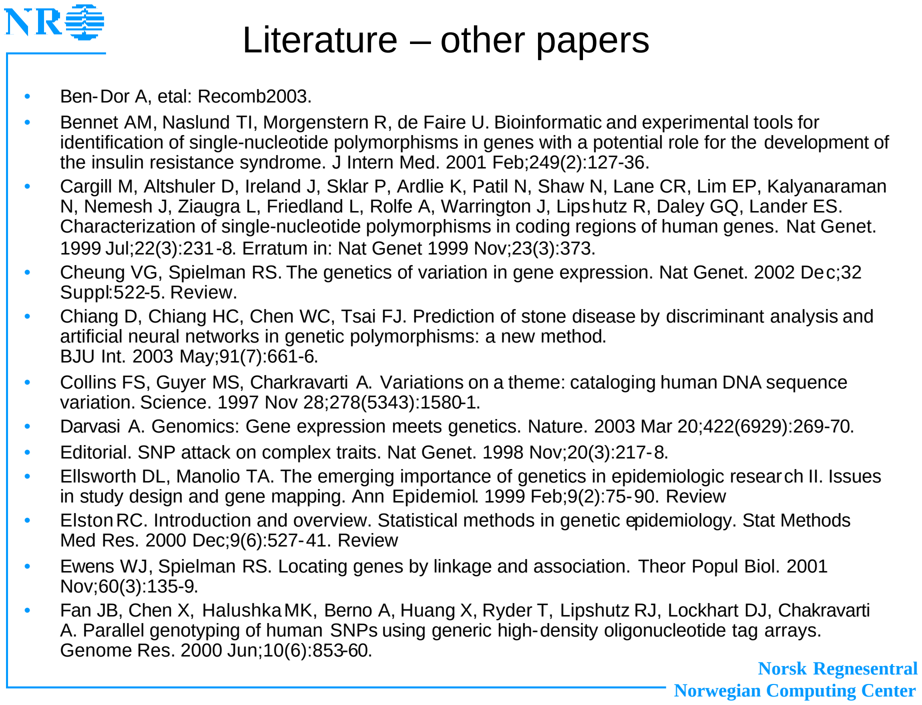
- Ott Jurg, Hoh Josephine. Set association analysis of SNP case-control and microarray data. Journal of Computational Biology 2003; 10(3-4): 569-574.
- Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. Science. 2001 Feb 16;291(5507):1224-9.
- Pfeiffer RM, Gail MH. Sample size calculations for population- and family-based case-control association studies on marker genotypes. Genet Epidemiol. 2003 Sep;25(2):136-48.
- Pociot F, McDermott MF. Genetics of type 1 diabetes mellitus. Genes Immun. 2002 Aug;3(5):235-49.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000 Jul;67(1):170-81.
- Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999 Jul;65(1):220-8
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun;155(2):945-59.
- Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003 Jan;19(1):149-50.
- Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, Kulbokas EJ, O'Leary S, Winchester E, Dewar K, Green T, Stone V, Chow C, Cohen A, Langelier D, Lapointe G, Gaudet D, Faith J, Branco N, Bull SB, McLeod RS, Griffiths AM, Bitton A, Greenberg GR, Lander ES, Siminovitch KA, Hudson TJ. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet. 2001 Oct;29(2):223-8
- Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000 Jun 15;405(6788):847-56. Review.

Norsk Regnesentral
Norwegian Computing Center

Literature - important papers cont.

- 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 Dec;8(12):1273-88. Review.
- Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. Genetics. 2003 Jun;164(2):829-33.
- Schaid DJ, Rowland C. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet. 1998 Nov;63(5):1492-506
- Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA Pooling: a tool for large-scale association studies. Nat Rev Genet. 2002 Nov;3(11):862-71. Review
- Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet. 1998 Feb;62(2):450-8
- 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. 1999 Mar;9(3):234-41.
- Toivonen HT, Onkamo P, Vasko K, Ollikainen V, Sevon P, Mannila H, Herr M, Kere J. Data mining applied to linkage disequilibrium mapping. Am J Hum Genet. 2000 Jul;67(1):133-45.
- Tsalenko A, Ben-Dor A, Cox N, Yakhini Z. Methods for analysis and visualization of SNP genotype data for complex diseases. Pac Symp Biocomput. 2003;:548-61.
- Zhu X, Bouzekri N, Southam L, Cooper RS, Adeyemo A, McKenzie CA, Luke A, Chen G, Elston RC, Ward R. Linkage and association analysis of angiotensin I-converting enzyme (ACE)-gene polymorphisms with ACE concentration and blood pressure. Am J Hum Genet. 2001 May;68(5):1139-48.

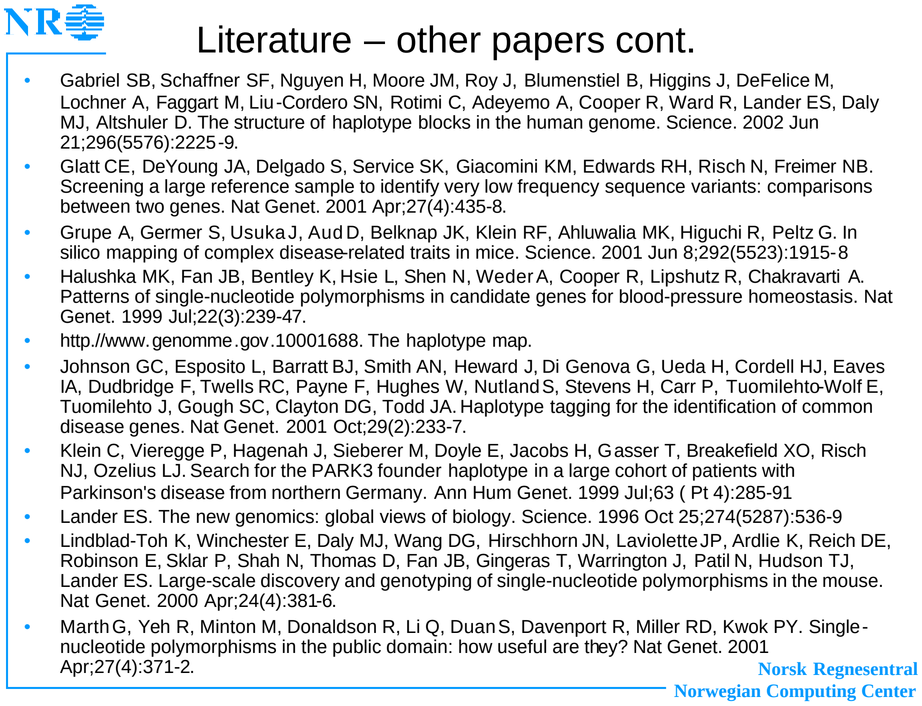
Norsk Regnesentral
Norwegian Computing Center



Literature – other papers

- Ben-Dor A, etal: Recomb2003.
- Bennet AM, Naslund TI, Morgenstern R, de Faire U. Bioinformatic and experimental tools for identification of single-nucleotide polymorphisms in genes with a potential role for the development of the insulin resistance syndrome. J Intern Med. 2001 Feb;249(2):127-36.
- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999 Jul;22(3):231-8. Erratum in: Nat Genet 1999 Nov;23(3):373.
- Cheung VG, Spielman RS. The genetics of variation in gene expression. Nat Genet. 2002 Dec;32 Suppl:522-5. Review.
- Chiang D, Chiang HC, Chen WC, Tsai FJ. Prediction of stone disease by discriminant analysis and artificial neural networks in genetic polymorphisms: a new method. BJU Int. 2003 May;91(7):661-6.
- Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. Science. 1997 Nov 28;278(5343):1580-1.
- Darvasi A. Genomics: Gene expression meets genetics. Nature. 2003 Mar 20;422(6929):269-70.
- Editorial. SNP attack on complex traits. Nat Genet. 1998 Nov;20(3):217-8.
- Ellsworth DL, Manolio TA. The emerging importance of genetics in epidemiologic research II. Issues in study design and gene mapping. Ann Epidemiol. 1999 Feb;9(2):75-90. Review
- Elston RC. Introduction and overview. Statistical methods in genetic epidemiology. Stat Methods Med Res. 2000 Dec;9(6):527-41. Review
- Ewens WJ, Spielman RS. Locating genes by linkage and association. Theor Popul Biol. 2001 Nov;60(3):135-9.
- Fan JB, Chen X, Halushka MK, Berno A, Huang X, Ryder T, Lipshutz RJ, Lockhart DJ, Chakravarti A. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. Genome Res. 2000 Jun;10(6):853-60.

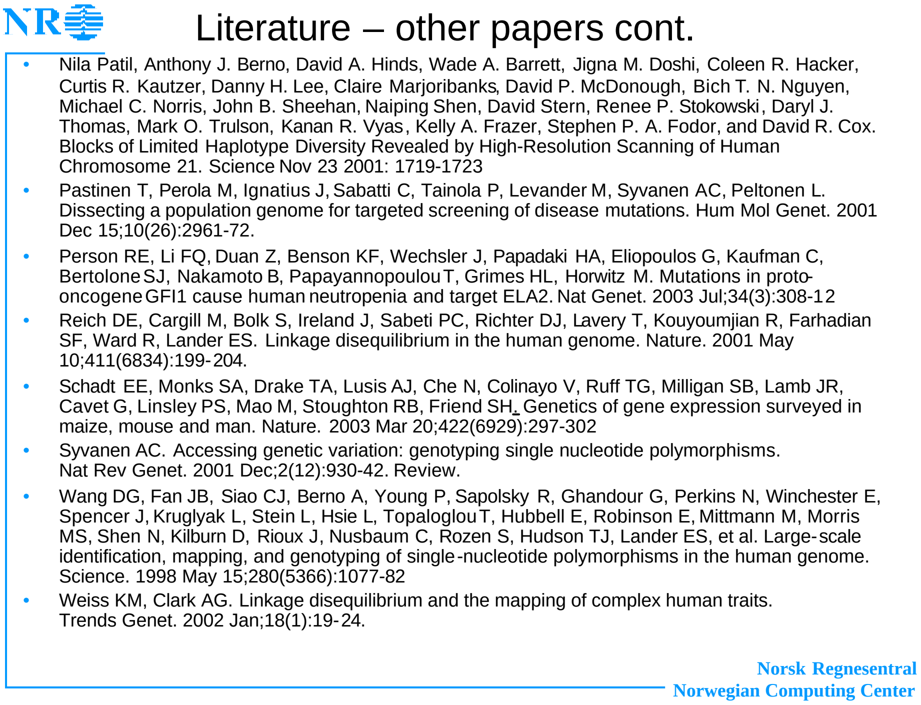
Norsk Regnesentral
Norwegian Computing Center



Literature – other papers cont.

- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296(5576):2225-9.
- Glatt CE, DeYoung JA, Delgado S, Service SK, Giacomini KM, Edwards RH, Risch N, Freimer NB. Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. Nat Genet. 2001 Apr;27(4):435-8.
- Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G. In silico mapping of complex disease-related traits in mice. Science. 2001 Jun 8;292(5523):1915-8
- Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet. 1999 Jul;22(3):239-47.
- http.//www.genomme.gov.10001688. The haplotype map.
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. Haplotype tagging for the identification of common disease genes. Nat Genet. 2001 Oct;29(2):233-7.
- Klein C, Vieregge P, Hagenah J, Sieberer M, Doyle E, Jacobs H, Gasser T, Breakefield XO, Risch NJ, Ozelius LJ. Search for the PARK3 founder haplotype in a large cohort of patients with Parkinson's disease from northern Germany. Ann Hum Genet. 1999 Jul;63 (Pt 4):285-91
- Lander ES. The new genomics: global views of biology. Science. 1996 Oct 25;274(5287):536-9
- Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan JB, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat Genet. 2000 Apr;24(4):381-6.
- Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, Davenport R, Miller RD, Kwok PY. Single-nucleotide polymorphisms in the public domain: how useful are they? Nat Genet. 2001 Apr;27(4):371-2.

Norsk Regnesentral
Norwegian Computing Center



Literature – other papers cont.

- Nila Patil, Anthony J. Berno, David A. Hinds, Wade A. Barrett, Jigna M. Doshi, Coleen R. Hacker, Curtis R. Kautzer, Danny H. Lee, Claire Marjoribanks, David P. McDonough, Bich T. N. Nguyen, Michael C. Norris, John B. Sheehan, Naiping Shen, David Stern, Renee P. Stokowski, Daryl J. Thomas, Mark O. Trulson, Kanan R. Vyas, Kelly A. Frazer, Stephen P. A. Fodor, and David R. Cox. Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. Science Nov 23 2001: 1719-1723
- Pastinen T, Perola M, Ignatius J, Sabatti C, Tainola P, Levander M, Syvanen AC, Peltonen L. Dissecting a population genome for targeted screening of disease mutations. Hum Mol Genet. 2001 Dec 15;10(26):2961-72.
- Person RE, Li FQ, Duan Z, Benson KF, Wechsler J, Papadaki HA, Eliopoulos G, Kaufman C, Bertolone SJ, Nakamoto B, Papayannopoulou T, Grimes HL, Horwitz M. Mutations in proto-oncogene GFI1 cause human neutropenia and target ELA2. Nat Genet. 2003 Jul;34(3):308-12
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibrium in the human genome. Nature. 2001 May 10;411(6834):199-204.
- Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003 Mar 20;422(6929):297-302
- Syvanen AC. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet. 2001 Dec;2(12):930-42. Review.
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998 May 15;280(5366):1077-82
- Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. Trends Genet. 2002 Jan;18(1):19-24.

Norsk Regnesentral
Norwegian Computing Center