RN List

Compiled by Yurii Aulchenko

This is ”RN List (RNL)”, version 2.0

Copyright © 2004 Yurii Aulchenko and RNL members

THIS TEXT IS PROVIDED “AS IS” AND ANY EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL YURII AULCHENKO AND TEXT CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS TEXT, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Table of Contents

1 Preliminaries

1.1 What is "RNL"?

This document contains "RN List", RNL — a set of design, methodological and computational question in the filed of genetic epidemiology. These questions are coming mostly from gene-identification studies in isolated populations; the emphasise is made on disease-oriented (like DO-GRIP, see Section D.1.2 [DO-GRIP], page 19) and quantitative trait screening (like ERF, see Section D.1.3 [ERF], page 19) studies in genetically isolated populations. This is very incompelete list by now: there are very many questions that are not well-formulated at the moment: please help us to do that better!

This text was compiled from fruitful discussions with and major help and contribution from a group of people from different countries and fields (see Section 1.4 [People on the list], page 2). These are indicated at the end of the text they've contributed.

What is common about these people is that all of us are interested in genetic analysis (gene identification) in humans and / or other outbred species.

When, later in the text, "we" is used you must realize that some members may not share the view, though it is likely to be a consensus comunity view.

1.2 Aim of RNL

We aim to formulate well a set of design / statistical / computational / software questions raised by the problem of gene identification in isolated populations of humans. These questions may be the same as these rised by gene-identification problem in other curicumstances (including other studies in even other species). We hope that precise formulation of the questions may help generating answers.

Our testing framework are the data coming from isolated populations. Different testbed data (see Appendix D [Practical research in isolated populations], page 19) ara available. There are basically two major types of studies performed on isolated populations, being disease-oriented set of studies as DO-GRIP (see Section D.1.2 [DO-GRIP], page 19) and screening, QTL-oriented studies resembling ERF (see Section D.1.3 [ERF], page 19). This document aims answering the question arosen in both types.

We aim this document to be continiously updated – that is, for the problems, which are more or less solved, we will expand the "Answers" part, while new questions will appear in "Questions" part.

We are interested in answers to these questions, i. e. methodology and working, effective, easy-to-use software, because we want to apply better methods for our particular gene-identification studies, which is our primary aim. Ideally, we would like to finish every of our questions with a precise answer, including directions for the software to be used and an example of application.

1.3 Conditions of participation

Fast effective development of new methods in a team-play is one of the aims of the RNL. A team-game implies free information flow. This document is aimed to provide such a flow.

If you'd like to join the list:

1. It would be great if you inform us if you are working on a topic which is on our list (we realize that you may have worked on this topic long before this list appeared on the public). This will let coordination of efforts and time-saving.
2. Later on, it would be great if you could keep other people on the list informed, by sending short progress notes

If you join the list as a software-developer, and pick up the topic from our questions, please, bear in mind that we aim to develop free software — at least free for the members of academic community.

When sending something to the RNL, you must realise that this is an open document, which may appear on the web and many people may see it.

1.4 People on the list

Ilja Akberdin (Novosibirsk), Yurii Aulchenko (Novosibirsk, Rotterdam), Tatiana Axenovich (Novosibisrk), Harry Campbell (Edinbourgh), Cornelia van Duijn (Rotterdam), Stefano Elefante (Rotterdam), Katrin Hoffmann (Berlin), Anatoly Kiritchenko (Novosibirsk), Fan Liu (Rotterdam), Anne-Louise Leutenegger (France), Stuart Macgregor (Cardiff), Bertram Mueller (Muenchen), Luba Pardo (Rotterdam), Igor Rudan (Edinbourgh, Zagreb), Irina Zorkoltseva (Novosibirsk), Artem Zykovich (Novosibirsk), Simon Heath (France).

1.5 Structure of this document

This document is a Questions & Answers (Q&A) document. We start with very general questions, which generate ”top” question nodes. Then, the answer may be not really an answer, but a formulation of a number of sub-questions. Thus, a second layer of questions is generated. At some point, formulation of the question allows precise answer, which is either provided (closed question) or is open.

In Appendixes, we give the necssary background information:

Our view on what is genetic epi. and what are the major dimensions, limitations and data-driven classification of genetic epidemiologic studies is given in Appendix B [whatisGE], page 16.

We discuss the very fundamental question on why do we study genetically isolated populations in Appendix C [Why do we study genetically isolated populations], page 18

We show what is our paractical framework (design of contemporary studies in isolated populations) in Appendix D [Practical research in isolated populations], page 19.

There is also a glossary of terms (see Appendix F [Glossary], page 23) which may help understanding the terminology we use.

A list of software mentioned in the text with short overview is provided in Appendix H [Software], page 36.

Information for potential contributers is listed in Appendix I [Information for contributors], page 45

1.6 Navigator: suggestions on what to read

This document, which was not intended to be very lengthy (we thought of a few dozens of pages at maximum initially) became quite a long story, partly because it is aimed to a very heterogeneous auditorium. Therefore here we have some suggestions on how to read this document, as most of the readers would not need to read the whole thing. We believe that anyone should start with Appendix B [whatisGE], page 16 — the section shows our vision of the problem of gene-identificaton.

- If you are a statistical geneticist, familiar with the methods and software used for genetical research, and you are only interested in developing new methods and software, just go directly to Chapter 2 [Questions and answers], page 4. You are likely to be interested to start with see Section 2.1.4 [Open question in solved questions], page 6, and go on till the end of Chapter 2 [Questions and answers], page 4. If you feel interested, please check Appendix A [Current status], page 15, to see what is the implementation status. For latest update, contact the responsible person (see Section 1.4 [People on the list], page 2).

- If you are mainly interested in the design of studies in isolated populations, you are most likely to be interested to start reading from Appendix D [Practical research in isolated populations], page 19 and Appendix E [Guidelines for field research in human isolates], page 21. Then, you could go to Chapter 2 [Questions and answers], page 4, to check if you experience the same problems we do.

1.7 Acknowlegements

This document is supported by the joint grant from Dutch Organization for Scientific Research (NWO) and Russian Foundation for Basic Research (RFBR) (CvD, TA, YA, AK, AZ, IA).

2 Questions and answers

This chapter contains Questions and answers. First, we try to make an overview of solved questions, i.e. situations when exact computations on the pedigree data are possible (see Section 2.1 [Solved questions], page 4).

Our situation is more complex: pedigrees in question are to big to allow exact computations. One of possible solutions to this problem would be reduction of complex data down to the level when exact computations are possible; we describe this ”solution” in Section 2.3 [Reducing pedigree complexity], page 8. This solution may be not too good (see Section 2.3.2 [Advantages and disadvantages of pedigree reduction], page 9, for discussion).

Therefore later in Section 2.4 [Using whole pedigree], page 10, we consider problem of taking all pedigree data into account.

Another approach, which may be of advandate when no extensive genealogic database is present (or there are serious doubts on the value of the database) is to base your computations on genomic data only and describe the underlying (possibly, unobserved) pedigree structure by populational parameters. We discuss this approach is Section 2.5 [Using genomic data only], page 13.

How our computations may be facilitated using high-performance computing, is discussed in Section 2.6 [High-performance computing for genetic epi], page 13. In next section, Section 2.7 [Data storage ERF], page 13, we discuss some practical aspects of data storage in ERF type (see Section D.1.3 [ERF], page 19) of studies.

2.1 Solved questions

The key to all the computations on the the genetic data in general and pedigrees in particular is the likelihood. Say, if you can compute exact likelihood of your data in a second (not hours!), you may answer the whole range of questions.

The problem of estimation of parameters using genetic data (such as appearing e.g. in Complex Segregation Analysis, see Section F.4 [CSA], page 24, or variance-components <linkage> analysis, see Section F.31 [VC analysis], page 30) may be solved by numerical maximisation of the likelihood function and thus involves dosens and hundreds of evaluation of the likelihood at different parameter values. Therefore it is important that a single likelihood is evaluated fast. If maximum likelihood may be found, the hypothesis testing is about comparing maximum likelihoods under different restrictions imposed onto the model — and this is the link to model decision and gene-finding.

In IBD computations or linkage analysis, though the number of parameters to be estimated may be small (or estimation is not performed at all, as in IBD computations), this likelihood is evaluated in hundreds of points along the genome, thus, again, single likelihood must be computed fast.

There is a class of data, for which the computation of exact likelihood is possible in reasonable time. We outline these classes here and later we do not return to these in our question-list. Anything which is outlined in this chapter is not a problem.

Any algorithm computing likelihood of pedigree data (mind our definition of pedigree data, see Appendix B [whatisGE], page 16 — but here by pedigree me mean exact pedigree structure, assuming linkage equlibrium in founders) is sensitive to increasing number of people in question and to number of loci in question. There are two algorithms which do really compute exact likelihood of pedigree data fast. These are Elston-Stewart (see Section F.5 [ES algorithm], page 24) and Lander-Green (see Section F.11 [LG algorithm], page 26) algorithms. The first applies peeling (see ⟨undefined⟩ [peeling], page ⟨undefined⟩) by individual and the second applies peeling by locus.

2.1.1 Elston-Stewart algorithm

The computing time of the ES algorithm grows linearly with the number of people. However, it is sensitive to the loops (see Section F.21 [pedigree with loops], page 28): computational demands grow exponentially with the number of people required to be cut to break the loops (other possibility is to go to memory dimension — then one would need a hell of memory to process loops). Other thing is that ES algorithm computation time grows exponentially with the number of loci (say, 3-5 microsatellite loci is a maximum one can deal with currently). Thus, this algorithm can easily deal with 100s/1000s of people, but no loops (I wonder if there is any pedigree with 100s of people and no loops!) and you can do computations at 3-5 microsatellite (or 10-16 SNP — *need to check!*) loci at the time (which limits your informativity). For more formal description, see Section F.5 [ES algorithm], page 24.

2.1.2 Lander-Green algorithm

LG algorithm (see Section F.11 [LG algorithm], page 26) use peeling by locus. That means it is not sensitive to loops and computation time grows linealy with the number of loci. However, the pedigree size is crucial: computing time grows exponentially with the number of loci. More precisely, pedigree bit-size (see Section F.19 [pedigree bit-size], page 28) is a limitation. Currently, no pedigree with bit-size over 30 could be effectively analysed (depends on the particular realization, but 40 bits — a pedigree of size of 50 people or so — is definitely out of reach).

2.1.3 Overview of linkage analysis

Given one can compute pedigree likelihood in reasonable time, one is ready to perform linkage analysis. Here we give an overview of linkage analysis.

In linkage analysis, one can distinguish between extracting information about inheritance of particular genomic region in question and usage of this information to map a gene.

The inheritance of particular region is assessed by either inheritance vectors (IV) probability distribution or IBD matrix. The former is quite laborous and demanding to hardware, as the space of possible IV may be very large. For example, for a pedigree containing F founders and N phenotyped non-founders, the total number of IVs may be up to $(2 \cdot F)^{2 \cdot N}$ On the contrary, the dimension to store a primitive IBD matrix is proportional to the square of the number of phenotyped people (more exactly $(N + F) \cdot (N + F - 1)/2$). A whole range of intermediate situations may be investigated, for example, we could think of using IBD matrices containing information on sharing between all trios (three-people IBD).

If IV probability distribution is known, one may do model-based linkage analysis, both for a QT and a binary trait.

Knowlege of standard IBD matrix allow doing variance component linkage analysis for QTs and model-free linkage analysis for binary traits.

Historically, methods used to study binary and quantitative traits were different. The IBD matrix ideology and variance component analysis was developed by biometric school and extensively used in animal breeding. The inheritance vectors ideology was raised by human geneticists and was used to study inheritance of disese in human populations. Consequently, there is tendency to analyse these two types of traits differently (though both types of methods may be applied to the same data-set) and different software is more skewed to this or that side. Thus, from the historical perspective one may classify methods used for linkage analysis into methods used for binary and methods used for quantitative tarits.

Other type of classification of methods used for linkage analysis is to distinquish between so-called model-based (see Section F.15 [MB linkage analysis], page 27) and model-free ((see Section F.16 [MF linkage analysis], page 27) methods.

As here we intend overview of existing methods, we start our classification with distingushing between binary and quantitative trais and later model-based and model-free analysis.

By following the graph depicted at the figure (trait and method of choice are given in rectangles), one can reach "solution" nodes (ellipses). At these solution nodes, exact formulation of the answer is possible. We give these answers and details for the graph in Appendix G [Summary_LA], page 31.

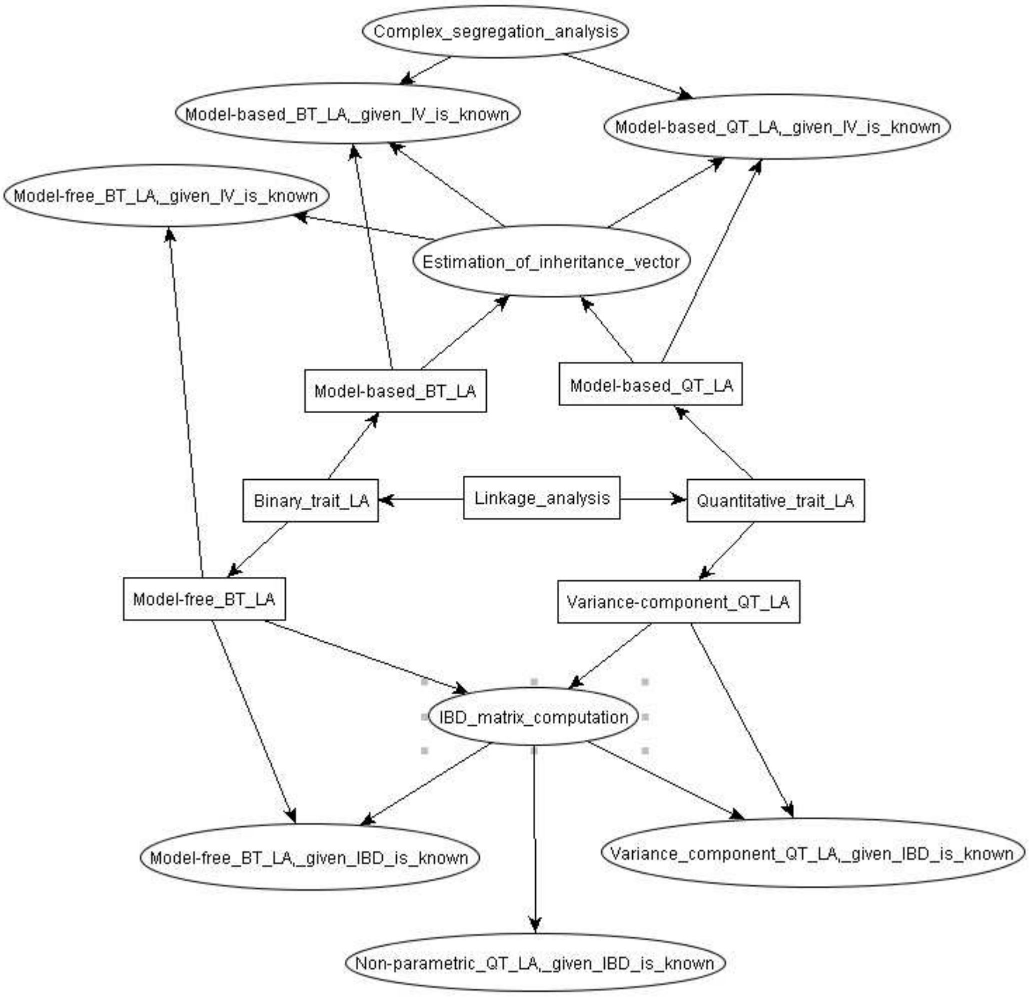
Complex_segregation_analysis
Model-based_BT_LA,_given_IV_is_known
Model-based_QT_LA,_given_IV_is_known
Model-free_BT_LA,_given_IV_is_known
Estimation_of_inheritance_vector
Model-based_BT_LA
Model-based_QT_LA
Binary_trait_LA
Linkage_analysis
Quantitative_trait_LA
Model-free_BT_LA
Variance-component_QT_LA
IBD_matrix_computation
Model-free_BT_LA,_given_IBD_is_known
Variance_component_QT_LA,_given_IBD_is_known
Non-parametric_QT_LA,_given_IBD_is_known

Figure 2.1: Summary of questions appearing in linkage analysis

2.1.4 Open question in solved questions

2.1.4.1 MOD score

Traditional LOD score is obtained in two steps: first the model of inheritance of the trait under analysis (penetrances, gene-frequencies, etc.) is estimated in complex segregation analysis or (more frequent) by eye-balling. Then, LOD score is computed under the fixed model at different genomic points. The idea of MOD score is to estimate parameters of the genetic model given the point of assumed gene location. This should be more powerful technique, though here we can run into a problem if estimating many parameters at many point along the genome (that is to say that the genome-wide-threshold must be corrected for that estimation).

The MOD score was introduced by FCD et al. in late 1980s, but we are not aware of any software realising the idea of MOD score mapping for binary traits.

Funny enough, for QTs, VC mapping with estimation of locus-specific heritability and residual heriatbility may be thought of as a realisation of MOD score.

However, for QTs, there is no complete parametric realisation of the MOD score ideology.

There seems to be some package from Newton Morton for binary traits — can anybody characterise that? — seems to be very complicated to use.

There are rumors that Konstantin Straus got a GH modification, which does MOD scores wit standard binary traits...

2.1.4.2 Can we keep inheritance vectors (IVs)?

It would be great if we could deal with IV framework exactly in the same manner as we deal with VC mapping: that is for a set of people, characterize genomic inheritance first, store the IV distribution(s) and later use for mapping of any trait. Could we store IVs (at least these making >99% of all IVs) and later use it for any trait mapping?

Yurii did something similar for pedigrees with very limited number of founders, by forcing LOKI to output the founder alleles (alleles which are unique for founders, thus there are up to 2 x No_founders alleles segregating in the pedigree) for every phenotyped person in the pedigree. That seemd to work quite well. Also, that mey work for isolated populations as effective number of founders / segregating genomes may be quite reduced there.

If we could compute the IV distribution and keep that, this may solve the problem of MB linkage analysis. This is a way more powerful then using IBD (sharing between pairs) only: it gives you information on sharing between ALL people. You can think of that as of all-people IBD: as that was shown by Elisabeth Thompson (see the grey book), even 3-people IBDs give quite some advantage over standard IBDs. This may be an option as well, if we could not keep all the IVs. However, given a dense marker map (as related to the length of connections in meioses), it might be feasible to estimate and keep the distribution of inheritance vectors.

2.1.4.3 Linkage analysis, when pedigree is not known

Assume the pedigree is not known and only some population characteristics are there (Ne, time since foundation). Meuwissen showed that in this case one can attempt estimation of IBD matrix based on marker data (that one should be quite sensitive to marker allele frequency estimates). Can we constract a method and software which estimates IV distribution based on marker data and populational parameters? Is solution different for the screening studies and sets ascertained via proband?

by Yurii Aulchenko, 13.08.2004, 07.09.2004

2.2 Characterizing complex pedigrees

Characterisation of complex pedigree is an open question. Currently, we need someone to do a practical review on pedigree characterisation. We believe that a person from livestocks genetics may help us a lot with solving this question. The starting point would be to try out the full range of PEDIG package (Section H.17 [PEDIG package], page 42).

The pedigree characteristics we could think of immediately are:

- no. people, no. phenotyped., max. no. generations, no. people in different generations
- number of nuclear families, number of re-marriages
- phenotyped sibship size distribution (by generation)
- effective number of founders (???)
- "key" people : the ones who contribute most genes to the phenotyped set — this is a kind of "research question", this information may be later used for pedigree splitting (idea from TI: compute IBD with a person excluded and included and take the difference between matrices. If difference is pronounsed, this is a key person).

- relationship coeff. distribution and distributions of number of meioses linking pairs of phenotyped relatives
- inbreeding distribution and distribution of size of inbreed loops
- number of loops by type (inbred, marrige rings, etc — see Cannings et al. 1978)

Idea is to pick up the characteristics of interest and write a software which may be used for pedigree characterization.

Realisation status: Fan has currently a program listing all the connections and their length between people. The program computes loop-specific and ancestor-specific inbreeding and kinship coefficients and some other characteristics. It will be ready for general public at the end of January 2005 (14.01.2005).

by Yurii Aulchenko, 20.08.2004, 07.09.2004

2.3 Reducing complexity: pedigree splitting for genetic mapping

Our pedigrees are too complex. As opposed to the approach discussed later in Section 2.4 [Using whole pedigree], page 10, we may try to split our pedigree into managable pieces and then analyse these using old methods (see ⟨undefined⟩ [Solved question], page ⟨undefined⟩). This would be of couse incorrect and will lead to information loss. However, given there are no methods to take all this information into acount and the fact that old methods, though less powerfull, may still work, makes ”splitting” option quite attractive.

Idea is to provide a software, which will take a large pedgree as input and split it "optimal" peaces. "Optimal" may mean

- Every piece should contain as many phenotyped individuals as possible, but the pedigree size should be less than some MAXBIT (see Section F.19 [pedigree bit-size], page 28) thus making the use of Lander-Green algorithm (see Section F.11 [LG algorithm], page 26) possible.
- No loops and thus we may apply Elston-Stewart algorithm (see Section F.5 [ES algorithm], page 24)
- Please, take into account that different splits have different power. E.g. a split breaking loops would not disturb the power under dominant or additive model much, but may destroy recessive model.
- For VC linkage analysis, a very important question is if there any split, which drops power by less then X%??? Reason: it seems that for QTL mapping very long connections between people will not increase power much, so we could get rid of these with small power drop. Say, using complete pedigree we have power of 75% to map a gene explaining 10% of variation. Could we get rid of upper generations without loosing too much power (say, it drops to 70%)? For this, a fast power-estimation algorithm is required.

ULTIMATE AIM: Assess effect of different splitting variants on power of linkage analysis; develop software for pedigree splitting.

2.3.1 Proposed answer

Developing such software includes following tasks:

- Power of different splits on QTL-mapping power, software for "most efficient" split, efficient split for ERF pedigree
  - Power of VC, assuming "common variant of moderate effect"
    - Splitting by "generation". Idea is simple — as we have only last three generations phenotyped, it might be that power of analysis using this 3 generations is not much lower then power of analysis using all pedigree. Introduce a very simple pedigree

structure which could be described by two parameters the number of sibs (S) and number of generations. One starts by generating a family with S offspring in generation 1, then every offspring marries and produce S offspring again, till generation N. Last 3 generations are phenotyped. Under specific model, estimate power of VC mapping when using 3, 4, 5, all G generations.

    - Develop software for symmetric pedigree simulation
    - Evaluate power under a range of genetic models. Later we can switch from fixed number of offspring S to some distribution to make the pedigree more reasonable.
    - Write a software implementing the splitting algorithm
    - Estimate the power effect of such a simplistic split using ERF pedigree
    - At this point, a manuscript could be written (though not really good one)
  - More complicated splitting (keep "key people" or something),
- Develop ideas (keep "key" people?) must be of large effect under large variance of distribution of offspring
- Develop software for pedigrees simulation
- Develop software for different pedigree cutting; different criteria.
- Power testing simulated pedigrees
- Power testing on ERF pedigree
- At this point, a good paper may be arranged
  - Power of QTL mapping assuming "rare variant of large effect". If there is a locus heterogeneity, linkage has some quite power to map this gene. What happens if there is locus heterogeneity?
  - Pick up a method
  - Study power

2.3.2 Advantages and disadvantages of pedigree reduction

Advantages: after pedigree split, the data could be analysed fastly with no problem by existing methods.

Disadvantages:

- Linkage equilibrium in founders will be assumed in founders of split pedigrees. We may probably relax tha assumption if we really will (I wonder how much that may contribute to the power?)
- Information on links between founders of reduced pedigrees lost. More or less the same thing as above, BUT here we mean specific connections (some founders are close, some remote...) which cannot be acounted for by introducing LD.
- Many split variants: the analysis may be sensitive to the choice of this or that variant.

Realisation status: Stuart expressed interest in testing the power under different splitting. (written on 12.09.2004)

Realisation status: Tatiana Axenovich is at the final stage with her programm, which cut all loops in an optimized way. The program and description of the algorithm will be published within coming months (14.01.2005)

by Yurii Aulchenko, 13.08.2004, 07.09.2004, 12.09.2004

2.4 Using whole pedigree

The image at Figure 2.1 summarises problems we meet in linkage analysis of pedigree data; the next section (see Appendix G [Summary_LA], page 31) formulate these.

As it follows from the image, there are several key questions in LA:

1. The question of how to compute genomic IBD matrix using all pedigree data (considered in Section 2.4.1 [Computation of IBD sharing in large pedigrees], page 10) is key to variance component, non-parametric QT linkage analysis and model-free linkage analysis of binary traits. Given IBD matrix is available, we must consider following questions (all considered in Section 2.4.3 [Gene-mapping in large pedigrees based on known IBD matrix], page 12):
   a. Variance components quantitative trait linkage analysis
   b. Non-parametric quantitative trait linkage analysis
   c. Model-free binary trait linkage analysis
2. The question of estmation of probability distribution of inheritance vectors (considered in Section 2.4.2 [Estimation of inheritance vectors distribution], page 12) is a key to model-based analysis. When this question is answered, we are ready to answer following questions (all considered in see Section 2.4.4 [Gene-mapping in large pedigrees based on known IV distribution], page 12):
   a. Model-based binary trait linkage analysis
   b. Model-based quantitative trait linkage analysis

2.4.1 Computation of IBD sharing in large pedigrees

For large pedigrees we have at Erasmus, we would like to compute multipoint IBD matrices for a set of people of interest (2500 living participants of ERF, or living participants in GRIP studies).

Given such a matrix is computed, it could be of direct use (e.g. with SOLAR (see Section H.21 [SOLAR], page 44) or ASREML (see Section H.3 [ASREML], page 37)) for QTL mapping; for binary traits mapping the solution is less clear, but we assume that some form of S_{pairs} scoring function could be used for that (this will be other question, also we can also get the total difference between the kinship-IBD matrix elements and work out it's distribution under the null by simulations).

Also, in case of reconstruction of IBD between chromosomes of the same person, this will allow us estimation of homozygousity by descent (HBD), which could be used for homozygousity mapping (we have a lot of inbreeding in some of Alzheimer cases, and in ADHD)

As the problem of computing exact IBD may be re-formulated as the problem of computing of exact likleihood on pedigree data, which has no solution nowaday, all the answers are necesserily approximate.

2.4.1.1 Proposed answer: deterministic approximation 1

Pong-Wong and collagues (2000?) proposed a deterministic approximation for computation of IBD matrix. There is no publically available software to do that.

It seems there is a certain disadvantage in this approach: basically, IBD matrix is computed for single markers and then the positions in between are approximated (not sure).

by Yurii Aulchenko, 14.01.2005; this section will be mantained by Irina Zorkoltceva

2.4.1.2 Proposed answer: deterministic approximation 2

Let we have phenotypes and genotypes for 2500 people (1:2500). At some genomic position, we want to compute the IBD matrix for these people ((1:2500) x (1:2500)). Note that direct IBD estimation is not possible: the pedigrees are too big for Lander-Green and contain too many loops fopr Elston-Stewart.

To approximate this big matrix do the following:

1. Split the pedigree in N peaces, which could be managed by Merlin (Lander-Green algorithm). Without loss of generality, let assume that peace 1 include people fro 1 to K1, peace 2 includes people (K1+1):K2, etc. till KN-1:KN. In these pedigrees, also some extra people which provide connection are included, but these are not genotyped.
2. Then, by using Merlin, we can compute exact IBD within peaces and will have N matrices: (1:K1) x (1:K1), ((K1+1):K2) x ((K1+1):K2), : (KN-1:KN) x (KN-1:KN). If we think of general (1:2500) x (1:2500) matrix, the above matrices lie on the diagonal of the general matrix. We miss off-diagonal parts, which describe IBD between more remote relatives.
3. Using above N pedigrees, reconstruct haplotypes of all people of interest. For simplicity, assume that it is known which chromosome is maternal and which one is paternal (this assumption will be relaxed later).
4. Compute the length (in meioses) of all possible connections between all chromosomes; e.g., assuming outbred pedigree, paternal chromosomes of two sibs are 2 meioses apart, the same is true for their maternal chromosomes, while the distance between paternal and maternal is infinity.
5. Use method similar to that of Meuwissen (2002) to construct gametic IBD matrix and IBD matrix (difference: we use exact number of meioses between chromosomes, not populational parameters; also we make use of allelic status, not only of IBS status)
6. replace off-diagonal elements of general IBD matrix with approximated values

Realisation status: At this point, we do not have a good splitting algorithm. So, what we do, we estimate everybody's haplotype by taking into account only first-degree relatives (cutting out the person of interest + 1st degree relatives, putting this into Merlin and extracting haplotype) and then do (4) and (5). Thus, we have program establishing all relations between chromosomes (4) (by Liu Fan) and we have (5) (by Yurii Aulchenko). Now Liu is writing a program, which will put all this together: a user provides standard linkage pedigree and data (marker description) files and a list of people with genotypes and then has set of marker IBD matrices as output.

Relaxation of assumption that haplotypes are known for sure and knowledge which chromosome is paternal/maternal: use "sample" option of Merlin and do, say, 100 samples of conditional haplotypes. Then compute IBD between 2 chromosomes using all 100 x 100 possible combinations. We've implemented Lander-Green + parallelisation in (5), thus it works very fast and it will be feasible to do 100x100 for every pair. However, this is not done yet.

Step (1) (splitting pedigree into manageable non-overlapping parts) is not done (see also Q2).

by Yurii Aulchenko, 13.08.2004; open status, no one expressed interest trying this. Yurii did some preliminary job and even has a programm which acts on modified Meuwissen algorithm

2.4.1.3 Proposed answer: Markov Chain Monte Carlo

One can use Loki (see Section H.12 [LOKI], page 40) to compute IBD matrix. The real limitation for Loki now is the number of loops (it should be < 20-30). Thus we may think of several applications of Loki

1. Break the loops (except may be most "important" ones) in the pedigree and then apply Loki. For pedigrees like ERF (see Section D.1.3 [ERF], page 19), this will still give huge pedigree and the amount of computations may be prohibitively large.
2. Cut the pedigrees like ERF 2-3 generations over the upmost phenotyped person and apply Loki. This cut will result in almost loopless pedigree; however, many upper part connections will be lost. It may be argued this is not a problem for a screening study like ERF (see

Section D.1.3 [ERF], page 19), but definitely this is not great way to deal with disease-oriented studies like DO-GRIP (see Section D.1.2 [DO-GRIP], page 19).

3. For disease-oriented studies, pick up consecutively every pair of "small" lowest-level pedigrees, connect them and do IBD computations using Loki (these pedigrees may still be too complex and there might be a need to break the loops). Perform this operation for every pair and compose the matrix. One of the nice features of this algorithm is that it is easily parallelizable.

by Simon Heath and Yurii Aulchenko, October 2004; Stefano expressed interest in trying solution (3)

2.4.2 Estimation of inheritance vectors distribution

Inheritance vectors distribution. This splits to two questions: the first appears in see ⟨undefined⟩ [Open questions in solved questions], page ⟨undefined⟩, that is, for classical situations, can we still keep / store the IV distributions for later use? If yes, we face a problem of computation / storage of IVs for large pedigrees from genetically isolated populations. With smalle pedigrees one can use Loki to output IV. Then one can construct model-free (or even model based, but no much point...) analysis using that. Is that effective type of analysis? What is power?

For more complex pedigrees, Loki will fail. How we compute IV then?..

by Yurii Aulchenko, 07.09.2004

2.4.3 Gene-mapping in large pedigrees based on known IBD matrix

2.4.3.1 <Multivariate> QTL mapping in large pedigrees, given IBD matrix is available

QTL mapping - though we can put marker IBD matrix (when computed) in SOLAR with small efforts, we would like to make use of ASReml, as it is a way faster. We do not have a clue how to do that. Also, should (can we) we go for this option at all, as ASReml is commercial software? Then, does free software (say, DFReml), provide more or less the same functionality?

IBD -> Fudging (if singular) -> take inverse -> ASReml (easy to implement, very flexible, rather fast, but not free, no source, not truly parallel)

IBD -> Solar (very easy to implement, but slow and no source, not truly parallel)

IBD -> Fisher (not too hard to implement, but slow and possible license problem, not truly parallel)

IBD -> Compute LH with linpack / maximize with Methgi (fast, free, may be true parallel, but implementation may be a problem)

Say we have IBD. How we do multivariate QTL mapping (ASReml?)? If ASReml, how we get IBD-1 matrices (LINPACK?) and how long it takes? Will ASReml work with that big IBD-1 matrices and how long it will take? When doing multivariate QTL mapping (probably in a candidate region), what test we use and what are the properties of the test?

by Stuart Macgregor and Yurii Aulchenko, July 2004; open status, no one expressed interest to study this yet

2.4.4 Gene-mapping in large pedigrees based on known IV distribution

Model-based analyses.

by Yurii Aulchenko, Sept. 2004; open status, no one volunteered to review the topic and set the questions yet, though Tatiana Axenovich seems to be a great candidate

2.5 Using genomic data only

There are some approaches which could let compute IBD using genomic data only (and some assumptions of population history of cause). See works by S. MacPeek and M. Abney.

by Yurii Aulchenko, Sept. 2004; open status, no one volunteered to review the topic and set the questions yet

2.6 High-performance computing for genetic epi

Someone have to do a review of possible parallelisation of Elston-Stewart and Lander-Green algorithms before we go for that. These people from GHT-parallel may be good candidates (see Section H.9 [GHT parallel], page 40).

Properties of high-performance algorithms (sensitivity to hardware implementation)

Paralellisation: should we stick to fortran90 (there is a free version for academia from Itel)? MPI-F77/C/C++?

When and what hardware is best? (wait till 64-emulating processors are on the market?) - the answer depends on how changes the speed-up of differed parallel implementation of our algorithms depending on (1) number of processor (2) RAM needed for a process (3) connection speed (say having many small processes is very sensitive to connection speed)

Review:

- Primitive parallelization
  - by chromsome
  - by pedigree
  - by replica in simulation study
- Advanced optimisation
  - ES
  - LG

Status: Yurii is palnning to make a review and estimate scalability of different parallelization algorithms.

2.7 Data storage in research in genetically isolated populations, ERF-type of studies

2.7.1 Data types

Pedigree data:

Genomic data:

Phenotypic data:

Enviromental factors data:

2.7.2 Estimation of required storage space

2.7.3 Effective organisation of database

2.7.4 Effectiviness of different DBMS

2.7.5 Organisation of binary data warehouse

to be written and mantained by Anatoly Kirichenko

2.8 Later on...

Later, we are going to formalise the following questions (part is already there, but needs some editing):

- Problem of selection via genealogy in ERF
- The use of repeated measurments
- Computation of genomic and regional IBD/HBD sharing based on genomic data only, no pedigree, only some information on population history
- Utilisation of the above information in gene-mapping
- Developing new models and methods
  - Sex- and age-specific heritabilities... Sex-specific heritabilities are actually outlined in papers of Blangero: you introduce two sets of parameters for male and female + corelation. But there is no software! However, that must be reaslizable within ASReml.
  - Detecting imprinting
  - Detecting locus interaction
  - <Semi> Parametric methods for QTL mapping
- GRIP studies: doing homozygousity mapping with multiple loops present
- GRIP studies: linkage analysis with complex genealogy
- GRIP studies: detecting two-locus interaction, using model-based methods

Appendix A Current status of implementation

Topic	Program/Review	Who	Status
Utility			
	pedigree recoding and verification	TI	1
	pooling STR data	Yurii	1
	phenotipic recoding and verification	Artem	1/2
	genotypic data verification	Ira	1/2
	testing different databases for genetic epidemiologic data storage	Tolya, Ivan	1/2
	warehouse for GE data storage	Tolya	0
Section 2.2 [Characterizing complex pedigrees], page 7			
	connection description	Fan	1/2
⟨undefined⟩ [Reducing complexity], page ⟨undefined⟩			
	loop-breaker for ES usage	TI	1/2
	overlapping breaker for LG	Fan	1/2
	non-overlapping breaker for LG	Jonger	1/2
	testing power of different breaks	Stuart (?)	0
⟨undefined⟩ [IBD sharing in large pedigrees], page ⟨undefined⟩			
	Pong-Wong	Ira	0
	Loki for DO-GRIP	Stefano (?)	0

Table A.1: Current implementation status

Status 0 = not started, 1/2 = started, 1 = completed

Updated by Yurii Aulchenko, 14.01.2005

Appendix B Components of GE study

In genetic epidemiologys studies, we deal with three pure types of information, being phenotypic information (measurments on phenotypes of interest), genomic information (marker genotypes, sequence), and pedigree (or genealogic) information. A genetic epi. study may miss genomic dimension (as that is *a priori* accessed by pedigree), but missing any other is not possible.

B.1 Pedigree information

The pedigree information may be known explicitly — that is the structure of relations between the studied subjects is known in precise manner, see Section F.18 [pedigree], page 28. These might be small nuclear pedigrees (see Section F.17 [nuclear pedigree], page 27) or large extended pedigrees (see Section F.6 [extended pedigree], page 25). Ultimately, the underlying pedigree may be so complex and spread in time that it is not possible to know it or account for all these genealogic connections any more — in this case pedigree structure is assumed, though it is not known explicitly or accounted for during analysis; this structure is usually described only via some populational parameters condensing the pedigree information (effective population size, time since foundation). Some times this information is not accounted during analysis at all, though, still, it is assumed to be there.

B.2 Genomic information

Genomic information represents a wide class of data. A minimal entry, which could be still thought of as genomic data, is data from typing a single genetic polymorphism (e.g. RFL (see Section F.27 [RFLP], page 29) or STR (see Section F.29 [STRP], page 30) polymorphism). On the other end of the scale would be a complete genomic sequence (for a particular person). We must emphasise, that a sequence of human genome found in NCBI is not genetic epi. data as it applyes to our species and cannot be related to individual's trait. By sequence here we mean a sequence available for every single person in the study population, moreover, only polymorphic sites would be of interest from our viewpoint.

Two classes could be roughly distingushed in genomic information, that is a single genetic marker data — that includes also multiple unlinked markers and regional marker data – that is information on a set of linked markers for a specific region. The latter also includes genome-wide data.

The same absolute density of markers may be thought of single-marker or regional data, depending on your study design, along the lines of our definition of the concept of ”linkage” (see Section F.12 [linkage], page 26).

B.3 Phenotypic information

Phenotypic, or trait (see Section F.30 [trait], page 30), information starts with accesing disease status and comes all the stages to accesing different enviromental exposures and up to expression profiling data.

B.4 Sampling

Another information, which is crucial in a geentic epi. study (so crucial, that we assume it is always present), is the information on sampling of your subjects. Depending on sampling procedure, the analysis techniques applied later may be very different, especially when the model of inheritance is estimated (see Section F.10 [heritability analysis], page 26).

If you sample via disease status, this would correspond to sampling via proband in classical human genetics or a case/control study; you could sample via belonging to some family or population – but anyway there is a sampling schema, which is crucial to be understood. Though that may be not crucial in linkage studies, the fact that one does not know how the sample was obtained indicates a basic ignorance.

B.5 Balancing study dimentions

A practical genetic epi. study, as any study, is about balancing the dimensions to achieve most effective design for the lowest (fixed) price. As money or whatever resource is limited, one could go for one of the dimensions, but this will affect others.

Say, 100 kilo-euro is available for a study. Assume that the cost of genotyping of single SNP is 1 euro; the cost of sequencing (or doing 100,000 SNPs) is 1,000 per person. The cost of accesing height is 0.5 euro, the cost of accesing much phenomics (MRI, DEXA, expression profiling) is 2,000 euro per person. Thus, one could go for a ”phenomic flavour” – access one important SNP in an important gene and look for it's effect on many, many traits (i.e. doing phenome and one SNP for ˜ 50 people). Othervise, one could pick up a trait of interest (say, height) and try to map that along the genome (thus, accesing ˜ 100 people for height and genome-wide polymorphism).

Another option, which is statistically most welcome: take a set of candidate SNPs and a candidate phenotype(s) (related to the focus phenotype) and do as many subjects as you can (say, 1,000 to 10,000).

While balancing the costs, one should probably go for one of the dimensions (the one you are more intersted in — what you are interested in some particular gene-action or some particular trait conrol?); trying to get all dimensions at the same time will have a consequence of multiple genetic and phenotypic polymorphisms tested on few subjects, which will generate statistical ghosts and hardly any sound results.

B.6 Data-driven classification of genetic epidemiologic studies

When no data is present on genomic dimension, while there are data on nuclear or extended pedigrees, there is still a possibility to do a genetic epi study. The pedigree data indicate genetic information flow, thus genealogic clustering may be correlated with phenotipic clustering. Inheritance analysis (see Section F.1 [analysis of inheritance], page 23), including complex segregation analysis (see Section F.4 [CSA], page 24) and heritability analysis (see Section F.10 [heritability analysis], page 26) could be performed with these data. This analysis is hardly possible if all the distances between relatives are large: in this case, *a priori* probability of sharing genes does not deviate much from zero and huge number of subjects is required to detect any clustering.

When the genealogic dimension is at the top level (that is only the fact that subjects belong to the same population with some characteristics is known), this is the time for a case / control study, or, more generally a study accessing association of genetic and phenotypic polymorphisms (see ⟨undefined⟩ [association analysis], page ⟨undefined⟩).

A study with small families and single trait would generate classical linkage framework.

The DO-GRIP study (see Section D.1.2 [DO-GRIP], page 19) is generated by single / several related traits accessed in many subjects belonging to large extended pedigrees (see Section F.6 [extended pedigree], page 25), and the sampling is done according to the phenotype.

The ERF study (see Section D.1.3 [ERF], page 19) is generated by accessing multiple quantitative traits (see Section F.25 [QT], page 29) belonging to few groups of interest in a large set of people sampled from the same genealogy.

by Yurii Aulchenko, Aug. 2004

Appendix C Why do we study genetically isolated populations

From purely epidemiological point of view, human isolates are often studied everywhere for the two main reasons that have been assumed, but still need to be clearly demonstrated, I think: (i) reduced environmental variation; (ii) reduced genetic variation. I think that it would be useful to undertake two literature reviews (perhaps one of mine / your assistants may be interested in that?), where clear evidence for those two assumptions would be gathered (and, based on this, optimal ways to measure this variation and the extent of its reduction in various isolate populations could be proposed).

It would also be useful to attempt to demonstrate what I think is the main value of isolated populations: that they do have some extremely rare variants of large effect in common population frequencies due to combination of founder effect, subsequent genetic drift, and specific environmental conditions that may have weakened the selection against these variants.

by Igor Rudan, 14.08.2004

Appendix D Practical research in isolated populations

This appendix is dedicated to the description of the practical research we do in isolated populations. The research are grouped by geography (or, to say the same, by the research group working on that). Here, we describe Erasmus (Section D.1 [Erasmus testbed], page 19), Croatian (Section D.2 [Croatian testbed], page 20) and German (⟨undefined⟩ [German testbed], page ⟨undefined⟩) studies.

D.1 Erasmus research in isolated populations

At Erasmus, we have a number of research projects in the common framework of Genetic Research in Isolated Population (GRIP). There are basically two types of studies performed within GRIP, being disease-oriented set of studies and screening, QTL-oriented study ERF (see Section D.1.3 [ERF], page 19).

D.1.1 GRIP area

Both studies are based on the same set of closely located villages (GRIP area). From historical data, we know that this set of villages was founded by approximately 150 people in the middle of the 18th century and up until more recent decades, descendants of these founders lived in social isolation, with minimal immigration (less than 5%). From the year 1848, the population has expanded from 700 to more than 20,000 inhabitants today.

Our genealogic database currently contains >70,000 people and dates back to XVIth century. Generally, we've established that population has quite some linkage disequilibrium (LD) (Aulchenko et al, 2004) but the drift was kind of favourable — the alleles, which are rare (<1%) in general population are likely to be missing/increase frequency in our population, while common alleles keep more or less the same frequency (Pardo et al., in press)

D.1.2 Disease-oriented projects within GRIP

Disease-oriented Genetic Research in Isolated Population (DO-GRIP) is a set of disease-oriented projects aimed to identify the susceptibility genes for common disease. For DO-GRIP projects, the pedigrees are sampled via 10-200 probands with a particular disease. We also sample the first-degree relatives to established haplotypes. Normally, a genome screen with approx. 400 markers is performed in these subjects. Usually most of the affected subjects would link to a single pedigree and will share a common ancestor 10-14 generations ago. Resulting pedigree would include thousands of people; of cause only nuclear families at the last generations would have genotypes and phenotypes. Most typical papers published are Sleegers at al., Brain 2004 (on dementia), Aulchenko et al., Diabetes 2003 (type 2 diabetes), Bonifati et al., Science 2003, van Duijn et al., Am J Hum Genet 2001 (parkinson disease), Njajou et al., Nat Genet 2001, (hemochromatosis), Vaessen et al., Diabetes 2002 (type 1 diabetes).

D.1.3 Erasmus Rucphen Family (ERF) Study

Erasmus Rucphen Family (ERF) study includes 2500 relatives from the GRIP area (see Section D.1.1 [GRIP area], page 19). These were not selected on phenotype, but rather on being a part of huge genealogy. Basically, the 2500 subjects consists of 100 3-generation families which go back to 20 founding couples living in the isolate in the period 1850-1900. Moreover, all these are related to each other, which makes a huge (>10,000 people) pedigree. All 2500 living participants are screened for many (100s) of quantitative traits. We aim QTL mapping in this population. We access tarits which are risk factors and / or could serve as proxy to

1. internal / metabolic
2. ophtalmologic disorders
3. neurologic and

4. psychiatric disorders.

to be updated by Cornelia van Duijn, xx.xx.2004

D.2 Croatian research in isolated populations

to be written by Igor Rudan, xx.xx.2004

D.3 German testbed

to be written by Katrin Hoffmann, xx.xx.2004

D.4 Summary

Form above, we can see that most studies fall into QT screening () or disease-oriented frameworks. Study ... () stays a bit apart: ...

Appendix E Guidelines for field research in human isolates

I had thought of ”RQ/GRIP” a bit more, and suggest that some sort of ”Guidelines for field research in human isolates” could perhaps be agreed between several groups based on the experience that we all had. This could include:

(1) Assessing the applicability of an isolate for GE studies based on simple historic and demographic information available (low cost, pre-grant phase): studying duration of ethnic history, bottlenecks, isonymy data (surname distributions), endogamy (degree of isolation), and have some table where advantages and disadvantages for gene mapping are listed;

(2) Once modest ("seed") funding is obtained, what do we propose to invest it in: probably, demonstrating decreased genetic and environmental diversity (which will be described in detail in previous chapters); specifics could include checking and computing LD and gene diversity in a small random sample of (N?) persons using (N?) markers on the (??) chromosome(s), for which plenty of comparable data exists for other populations (and present them in previous chapter); or, genealogical reconstruction, assessment of prevalence of diseases/distribution of traits of interest, etc.

(3) Approaching and communicating with the study population: Here, there are some very important issues, we should certainly at least discuss; e.g.: Who should give ethical approval for the study, and for what (what were our experiences)? Do we recommend giving lectures to people in the villages before we start or not (and could we develop those lectures together)? Could we produce / recommend a standardised information sheet for informed consent, to be given out to people? Once the study is finished, how long can the data be used for the analyses under given consent/ethics approval? What if the results get commercialised - do we recommend that the community also should participate in sharing the money, and how, or not at all? What is the optimal way to ensure confidentiality of data collected?

There is also a slight "Catch 22": we study isolated populations because they are genetically interesting, but at the same time the education level there may be quite low. Thsu it may be hard to fulfill every single ”Sweden type” ethics requirement.

(4) Getting study started:

(a) Selecting the phenotypes to measure - suggest optimal methods of measurement and equipment for the most commonly measured phenotypes;

(b) Taking care of between-observer differences (if more than 1 observer is doing measurements), and intra-observer differences (fluctuations in values over time);

(c) Perhaps undertake repeated measurements to see which traits are more accurately repeatable;

(d) Doing heritability analyses to see which traits appear more heritable in specific population under study;

(c) Quality control procedures that need to be in place: proposed guidelines for laboratory handling of blood samples, transport of specimens and storage, etc., etc.

(5) Adjacent questionnaire on environmental/lifestyle variables:

I think that some care should be taken that, along with phenotypic, genotypic and genealogical information, we also gather some simple information for each examinee on lifestyle, marital status, nutrition, smoking, obesity, alcohol consumption, socio-economic status and quality of life, as these are all very important risks with proven effects on QT values and disease prevalence, so one should try to correct for those as much as realistically possible in further analyses;

For many of these variables, standardised questionnaires exist (by WHO, etc.), so perhaps we could develop a proposed "standard" questionnaire, which could be modified by each group respecting the specificities of settings where the studies are performed.

(6) Correlating 4 large databases:

- genotyping information
- phenotypic information
- pedigree information
- questionnaire information on environmental exposures and lifestyle

... well, everything from here forward is entering the domain of statistics, bioinformatics, power calculation, etc., etc., so I don't want to go there. We should at least mention any possibilities where things can go disasterously wrong from epidemiological point of view (e.g. choosing unhelpful population without brief and cheap checking of its history, wasting "seed" money on unhelpful analyses, all sort of related ethics issues, phetotypic inaccuracy or unreliability of measurements, and neglecting important environmental influences...

by Igor Rudan, 21.08.2004

Appendix F Glossary

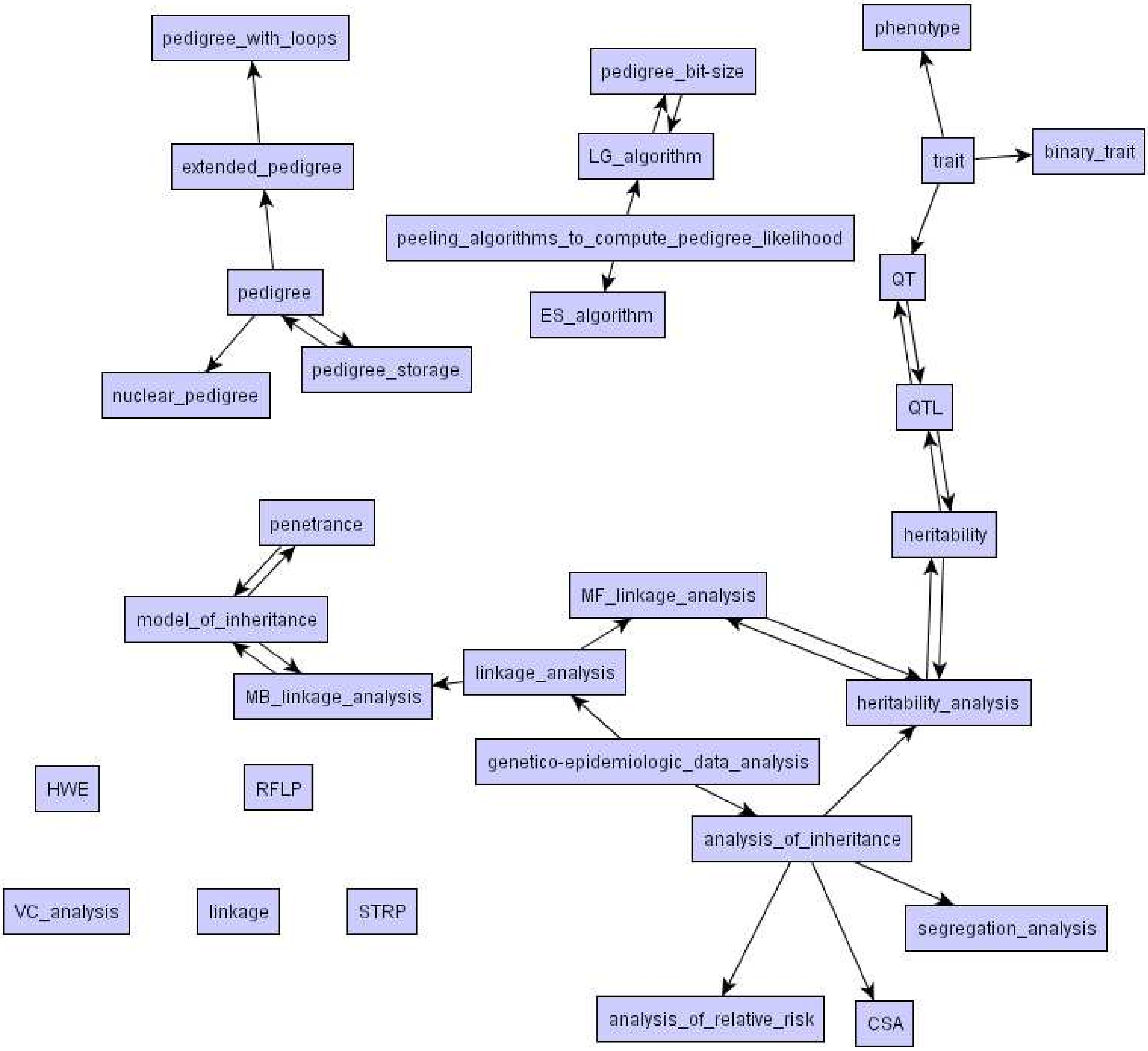

Figure F.1: Structure of Glossary

F.1 analysis of inheritance

Description: Analysis of distribution of phenotypic data (Section F.30 [trait], page 30, Section F.24 [phenotype], page 29) over the pedigree (Section F.18 [pedigree], page 28) data aiming to test if genes play essential role in the trait's control and, finally, to estimate the model of inheritance (Section F.14 [model of inheritance], page 27). As an initial step, analysis of correlation of the trait between relatives / ANOVA analysis may be performed for a quantitative trait (QT, Section F.25 [QT], page 29) and classical segregation analysis or analysis of risk for relatives of affected may be performed for a binary trait (Section F.3 [binary trait], page 24). If these show some degree of clustering between realitives, more advanced analysis may be performed, namely, heritability analysis (Section F.10 [heritability analysis], page 26) and complex segregation analysis (CSA, Section F.4 [CSA], page 24) for a QT and CSA for a binary trait.

More general terms: [Section F.7 [genetico-epidemiologic data analysis], page 25]

Share more general concept with: [Section F.13 [linkage analysis], page 26]

Sub-terms: [Section F.4 [CSA], page 24] [Section F.28 [segregation analysis], page 29] [Section F.10 [heritability analysis], page 26] [Section F.2 [analysis of relative risk], page 24]

Last updated: 11.08.2004, 20.08.2004 (Yurii)

F.2 analysis of relative risk

Description: For a set of probands, risk for particular relative of a proband (say, sib ro cousine) to develop the same disease may be estimated. When this absolute risk is divided to the populational risk, we obtain a relative risk for this particualar relative to develop the disease. Significant deviation of that from 1.0 may be thought of as an evidence of genes contributing into disease development.

More general terms: [Section F.1 [analysis of inheritance], page 23]

Share more general concept with: [Section F.4 [CSA], page 24] [Section F.28 [segregation analysis], page 29] [Section F.10 [heritability analysis], page 26]

Last updated: 20.08.2004 (Yurii)

F.3 binary trait

Description: A trait which takes two values

Example: presence or absence of a disease

More general terms: [Section F.30 [trait], page 30]

Share more general concept with: [Section F.24 [phenotype], page 29] [Section F.25 [QT], page 29]

Related terms: [⟨undefined⟩ [mendelian locus], page ⟨undefined⟩]

F.4 complex segregation analysis (CSA)

Description: Analysis of distribution of phenotypic data (Section F.30 [trait], page 30, Section F.24 [phenotype], page 29) over the pedigree (Section F.18 [pedigree], page 28) data aiming to test if genes play essential role in the trait's control and to estimate the model of inheriatnce (Section F.14 [model of inheritance], page 27) in precise parametric terms...... NO DIFF WITH ANAL OF INH

More general terms: [Section F.1 [analysis of inheritance], page 23]

Share more general concept with: [Section F.28 [segregation analysis], page 29] [Section F.10 [heritability analysis], page 26] [Section F.2 [analysis of relative risk], page 24]

Last updated: 11.08.2004

F.5 Elston-Stewart algorithm (ES algorithm)

Description: This is an algorithm to compute pedigree likelihood. This algorithm is usually contrasted with other popular algorithm for pedigree likelihood computation, Lander-Green (LG) algorithm. ES algorithm does "peeling" person-by-person, thus the computation time rise linearly with numbner of people and exponentially with number of loci. Loops present in a pedigree also increase computational time exponentially. Currently, analysis with more then 100,000 (???) potential genotypes (approx. 5 microsatellite markers) is virtually impossible. Analysis with 3-4 MS markers runs in reasonable time, depending on implementation. This algorithm may use up to XXX biallelic (SNP) markers. of considerartion. Pedigrees with size of 16-20 may Different implementations include Linkage, Vitesse, Fastlink.

More general terms: [Section F.22 [peeling algorithms to compute pedigree likelihood], page 28]

Share more general concept with: [Section F.11 [LG algorithm], page 26]

Last updated: 07.08.2004

F.6 extended pedigree

Description: a pedigree involving more then one nuclear pedigree

More general terms: [Section F.18 [pedigree], page 28]

Share more general concept with: [Section F.17 [nuclear pedigree], page 27]

Sub-terms: [Section F.21 [pedigree with loops], page 28]

Last updated: 09.08.2004, 13.08.2004

F.7 genetico-epidemiologic data analysis

Description: Analysis of the data consisting of pedigree information, genomic data and phenotipic data with the aim to find genes, polimorphism in which is related to the trait manifistation (gene-finding study) or accessing the effect and public health importance of a polymorphism already known to affect the trait.

Sub-terms: [Section F.1 [analysis of inheritance], page 23] [Section F.13 [linkage analysis], page 26] [⟨undefined⟩ [association analysis], page ⟨undefined⟩]

Last updated: 11.08.2004

F.8 Hardy-Weinberg equilibrium (HWE)

Description: A state of diploid population, when frequencies of genotypes may be described by the products of allelic frequencies, that is pairing of alleles is random. May be violated because of multiple reasons (assortative mating, selection, mutation, etc.), but in practice this violation is vary hard to detect, which let to the common practice of using HWE as a quality control tool in genetic studies.

Related terms: [⟨undefined⟩ [population], page ⟨undefined⟩]

Last updated: 08.08.2004

F.9 heritability

Description: Generally, the proportion of a trait's variation, which is due to the genes (broad-sense heritability). This definition comes from the concept of variance components (VC) analysis: that is, the variance of a traits in a population may be partitioned in genetic (σ_g^2) and environmental (σ_e^2) components (total trait variance $\sigma_T^2 = \sigma_g^2 + \sigma_e^2$). Then heritability is σ_g^2/σ_T^2 . However, This broad-sense hetitability is hard to estimate as it is partly confounded by enviromental factors, also it is not straightforward computationally as it requires estimation of many terms (additive, dominance and epistatic components of the genetic variance). Usually, a "narrow-sense heritability" is estimated, that is the ratio of additive genetic variance (σ_a^2) to the total variance: $h^2 = \sigma_a^2/\sigma_T^2$. Additive genetic variance applyes to the part of genetic variance, which may be described by additive effects of alleles.

Example: in humans, heritability of height is high (approx. 0.8)

Related terms: [Section F.10 [heritability analysis], page 26] [Section F.26 [QTL], page 29]

F.10 heritability analysis

Description: Analysis of genetic clustering of a trait aimed to estimate heritability (Section F.9 [heritability], page 25)

More general terms: [Section F.1 [analysis of inheritance], page 23]

Share more general concept with: [Section F.4 [CSA], page 24] [Section F.28 [segregation analysis], page 29] [Section F.2 [analysis of relative risk], page 24]

Related terms: [Section F.9 [heritability], page 25] [Section F.16 [MF linkage analysis], page 27]

F.11 Lander-Green algorithm (LG algorithm)

Description: This is an algorithm to compute pedigree likelihood. This algorithm is usually contrasted with other popular algorithm for pedigree likelihood computation, Elston-Stewart (ES) algorithm. LG algorithm does ”peeling” locus-by-locus, thus the computation time rise linearly with numbner of loci and exponentially with number of people (more precise, the computation time is restricted by pedigree bit-size). Currently, pedigrees with bit-size over 40 are out of considerartion. Pedigrees with size of 16-20 may be analysed withing reasonable time, depennidg on the implementation. Different implementations include gene-hunter (gh), allegro, merlin, gene-hunter-plus (ghp), gene-hunter-twolocus (ght)

More general terms: [Section F.22 [peeling algorithms to compute pedigree likelihood], page 28]

Share more general concept with: [Section F.5 [ES algorithm], page 24]

Related terms: [Section F.19 [pedigree bit-size], page 28]

Last updated: 07.08.2004

F.12 linkage

Description: For two genes, a phenomenon of non-independent transmission is called linkage. This is a consequence of proximal phisical location of two genes. Closer the genes are phisically, close the linkage is. Absolute linkage refers to perfect correlation in segregation, while no linkage refers to independent segregation. Generally, any genes located on the same (not too long) chromosome, may be called linked genes. In this document, however, this term will be used in relative framework: that is we define, for a study population, the genes to be linked, if the rule of random assortment is violated. Thus, for a set of nuclear families or small extended pedigrees, that will be the same as classical definition of linkage; for a population data, that would coincide with the definition of linkage disequilibrium.

Last updated: 09.08.2004

F.13 linkage analysis

Description: Analysis of pedigree, phenotypic and genomic data aimed to map a gene associated with trait variation on the genetic map

More general terms: [Section F.7 [genetico-epidemiologic data analysis], page 25]

Share more general concept with: [Section F.1 [analysis of inheritance], page 23]

Sub-terms: [Section F.15 [MB linkage analysis], page 27] [Section F.16 [MF linkage analysis], page 27]

Last updated: 11.08.2004

F.14 MODEL OF INHERITANCE

Description: For particular trait in a population, model of inheritance is a set of parameters specifying populational distribution of genotypic frequencies, mode of transmission of genes from one generation to the other and parameters describing the effect of different genotypes onto the trait value. For a binary trait, this model is specified by at least eight parameters, being frequency of the disease allele (populational distribution), three transmission probabilities (mode of transmission) and the penetrances (describe effect of genotypes on tarits) of three genotypes made of disease and normal alleles. For a quantitaive tarits, this model is specified by at least nine parameters, with penetrances replaced by the expectation of the trait value for the three genotypes and one extra parameter is introduced to describe residual (enviromental) variance. When popualtion distriburion of genotypes is described by the only parameter, assumption of Hardy-Weinberg equilibrium (HWE) is to be used.

Related terms: [Section F.23 [penetrance], page 29] [⟨undefined⟩ [transmission probability], page ⟨undefined⟩] [Section F.15 [MB linkage analysis], page 27]

Last updated: 08.08.2004

F.15 MODEL-BASED LINKAGE ANALYSIS (MB LINKAGE ANALYSIS)

Description: Analysis of linkage, based on precise formulation of the model of inheritance of the trait to be genetically mapped. This analysis is also called "parametric linkage analysis". It is frequently contrasted with model-free (MF) linkage analysis

More general terms: [Section F.13 [linkage analysis], page 26]

Share more general concept with: [Section F.16 [MF linkage analysis], page 27]

Related terms: [Section F.14 [model of inheritance], page 27]

Last updated: 08.08.2004

F.16 MODEL-FREE LINKAGE ANALYSIS (MF LINKAGE ANALYSIS)

Description: In contrast to model-based (MB) linkage analysis, this set of methods for genetic mapping does not assume a precise model of the inheritance of the trait being mapped. MF linkage analysis applies to binary taits. General idea of these methods is that, if the gene controlling the trait variation is localised at some region, similarities at this region, as accessed by similarities in marker genotypes, should lead to similarities in trait value.

More general terms: [Section F.13 [linkage analysis], page 26]

Share more general concept with: [Section F.15 [MB linkage analysis], page 27]

Related terms: [Section F.10 [heritability analysis], page 26]

Last updated: 08.08.2004

F.17 NUCLEAR PEDIGREE

Description: a "minimal" pedigree still having genetic transmission events: two parents and set of their offspring

More general terms: [Section F.18 [pedigree], page 28]

Share more general concept with: [Section F.6 [extended pedigree], page 25]

Last updated: 09.08.2004

F.18 PEDIGREE

Description: Technically speaking, pedigree is a graph representing genetic relation between subjects. A human pedigree may be represented as an oriented graph, with nodes presenting people and arcs showing the direction of the flow of genetic material. Thus, any node has at most two incoming arcs (from maternal and paternal nodes) and some limited number of outgoing arcs. If parents for some indifidual are not known (and thus there are no incoming arcs), this one is termed ”founder” individual. Such a graph may be presented in a table format by a three-column table, coding the person and its mother and father. For hystorical and technical reasons, pedigrees are usually stored as 5-column tables. Technically, thi minimal pedigree unit is a person with missing parents (a singleton), however, the minimal pedigree, which makes sense in term of genetic transmission is a nuclear pedigree

Sub-terms: [Section F.17 [nuclear pedigree], page 27] [Section F.6 [extended pedigree], page 25]

Related terms: [Section F.20 [pedigree storage], page 28]

Last updated: 08.08.2004

F.19 PEDIGREE BIT-SIZE

Description: Bit-size of a pedigree is defined as twice the number of founders minus the number of non-founders, $BS = 2 \cdot N_{non-foundes} - N_{founders}$. Bit-size limits implementation of Lander-Green (LG) algorithm.

Related terms: [⟨undefined⟩ [inheritance vector], page ⟨undefined⟩] [Section F.11 [LG algorithm], page 26]

Last updated: 08.08.2004

F.20 PEDIGREE STORAGE

Description: In human genetics, pedigree information is usually stored as five-column table. This table contain pedigree identification number (ID), personal ID, father’s ID, mother’s ID and sex. For any person, either both parents must be known and presented in pedigree (i.e. are found in personla ID), or both are not presented (coded as ”0”). Sex is usually coded as ”1” for male and ”2” for female.

Related terms: [Section F.18 [pedigree], page 28]

Last updated: 08.08.2004

F.21 PEDIGREE WITH LOOPS

Description: A pedigree (see Section F.18 [pedigree], page 28) could be represented as an oriented graph. A pedigree is told to contain loops if there are loops in this graph. There is no way to have cycles (loops where all arcs are in the same clock- or anticlock-wise direction) as this would imply that a child is it’s own ancestor.

Example: A marriage between sibs makes a short inbred loop; a marriage bitween two brothers and two sisters makes a loop

More general terms: [Section F.6 [extended pedigree], page 25]

Last updated: 13.08.2004

F.22 PEELING ALGORITHMS TO COMPUTE PEDIGREE LIKELIHOOD

Description: Algorithms to reduce computational time for pedigree likelihood by factorization

Sub-terms: [Section F.11 [LG algorithm], page 26] [Section F.5 [ES algorithm], page 24]

Last updated: 11.08.2004

F.23 PENETRANCE

Description: When talking of a disease, the penetrance of a genotype is defined as probability of being affected given this genotype. May be also defined as a frequency of the disease in people with this genotype. More formally, penetrance of a genotype with the respect to some trait value is conditional probability of this value given the genotype.

Related terms: [Section F.14 [model of inheritance], page 27]

Last updated: 08.08.2004

F.24 PHENOTYPE

Description: A particular manifistation of a trait

Example: black hair, heighth over 1 meter 80 cm.

More general terms: [Section F.30 [trait], page 30]

Share more general concept with: [Section F.25 [QT], page 29] [Section F.3 [binary trait], page 24]

Last updated: 09.08.2004

F.25 QUANTITATIVE TRAIT (QT)

Description: A trait which can be measured on continious scale

Example: Height, weight, cholesterol level

More general terms: [Section F.30 [trait], page 30]

Share more general concept with: [Section F.24 [phenotype], page 29] [Section F.3 [binary trait], page 24]

Related terms: [Section F.26 [QTL], page 29]

F.26 QUANTITATIVE TRAIT LOCUS (QTL)

Description: A locus, variation in which controls the variation of a quantitative trait (QT, Section F.25 [QT], page 29)

Example: Height, weight, cholesterol level

Related terms: [Section F.25 [QT], page 29] [⟨undefined⟩ [locus], page ⟨undefined⟩] [Section F.9 [heritability], page 25]

F.27 Restriction Fragment Length Polymorphism (RFLP)

Description: A bialleleic genetic polymorphism which is accessed by using a context-dependent restrictase. These polymorpjisms are usually SNPs.

F.28 SEGREGATION ANALYSIS

Description: Classical segregation analysis refers to analysis of distribution of affected / unaffected in the progeny of parents with different phenotypes. E.g. for a recessive disease one expects that if both parents are affected, all the progeny will be affected, too. If the proportion of affected offspring does not depend on parents' phenotypes, it is most likely that the genes contribute little into the trait's control.

More general terms: [Section F.1 [analysis of inheritance], page 23]

Share more general concept with: [Section F.4 [CSA], page 24] [Section F.10 [heritability analysis], page 26] [Section F.2 [analysis of relative risk], page 24]

Last updated: 20.08.2004 (Yurii) — should ask TI to extend

F.29 Short Tandem Repeat Polymorphism (STRP)

Description: Ususally a multi-allelic genetic polymorphism which is accessed by PCR / gel-elevtrophoresis or other technique

F.30 trait

Description: We define this term quite technically: in a study population, a trait is any feature which can be used to distinguish between individuals. There are three points to stress: (1) a trait is somthing polymorphic in the study population (e.g. being literally brainless is not a trait for a population of living humans); (2) this definition is population specific, i.e. a "trait" in some population may be not a trait in other (e.g. wooden/natural leg is a trait for a population of XVIIIth century pirats, but does not apply to contemporary population of Netherlands (3) ANY characteristic could ve used, thus, things, usually called ”enviromental exposure” are also traits, that is smoking, having particular social class, and even being in a wrong place in a wrong time.

Example: For any human population: height, eye colour, presence of hypertension

Sub-terms: [Section F.24 [phenotype], page 29] [Section F.25 [QT], page 29] [Section F.3 [binary trait], page 24]

Last updated: 7.08.2004, 9.08.2004

F.31 variance component analysis (VC analysis)

Description: See Section F.10 [heritability analysis], page 26

Appendix G Summary_LA

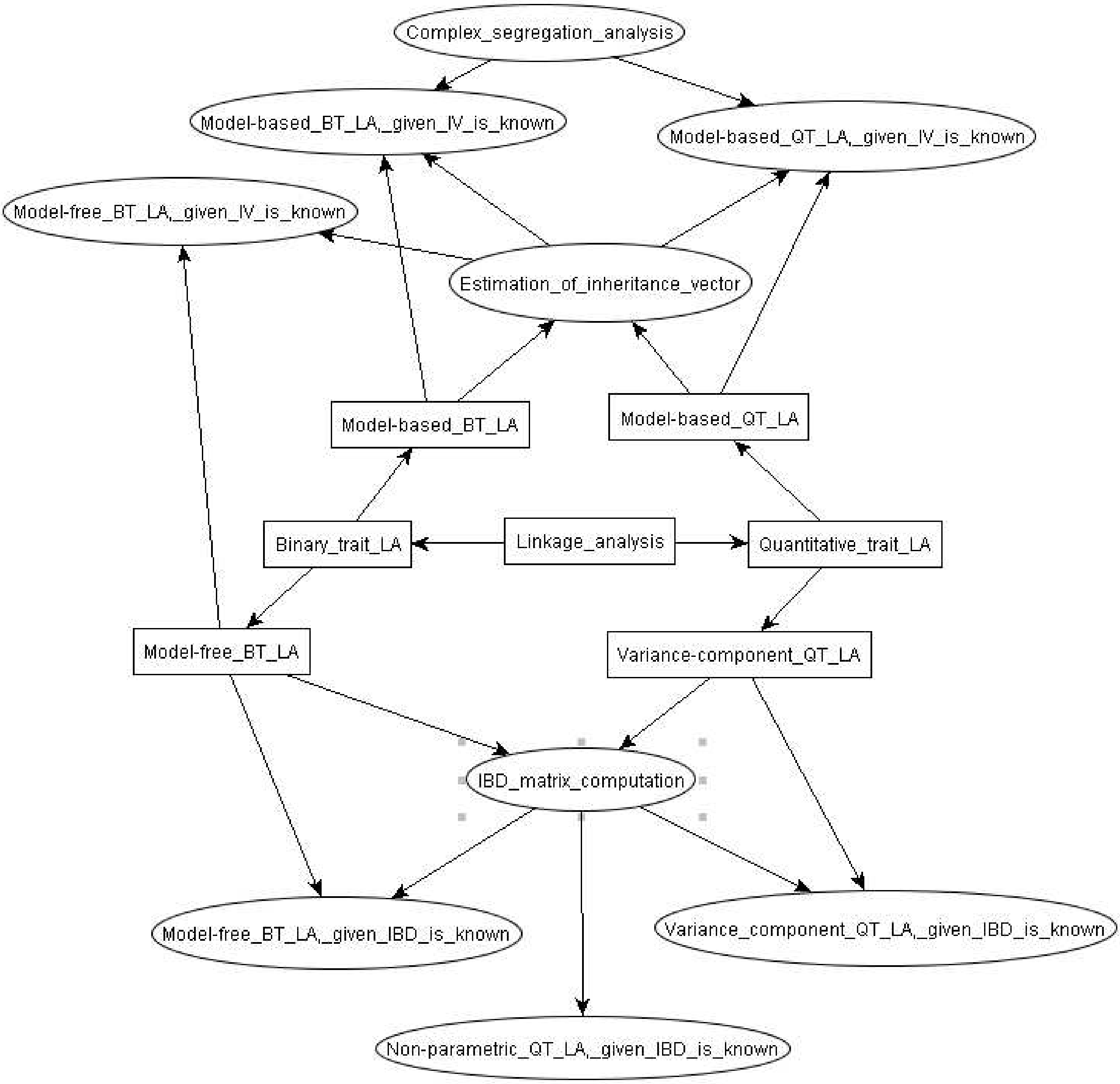
Complex_segregation_analysis
Model-based_BT_LA,_given_IV_is_known
Model-based_QT_LA,_given_IV_is_known
Model-free_BT_LA,_given_IV_is_known
Estimation_of_inheritance_vector
Model-based_BT_LA
Model-based_QT_LA
Binary_trait_LA
Linkage_analysis
Quantitative_trait_LA
Model-free_BT_LA
Variance-component_QT_LA
IBD_matrix_computation
Model-free_BT_LA,_given_IBD_is_known
Variance_component_QT_LA,_given_IBD_is_known
Non-parametric_QT_LA,_given_IBD_is_known

Figure G.1: Structure of Summary_LA

G.1 Binary trait LA

Author: Yurii Aulchenko

Description: You can analyse binary traits using either model-free linkage analysis (see Section F.16 [MF linkage analysis], page 27) or model-based linkage analysis (see Section F.15 [MB linkage analysis], page 27).

More general terms: [Section G.5 [Linkage analysis], page 32]

Share more general concept with: [Section G.14 [Quantitative trait LA], page 35]

Sub-terms: [Section G.12 [Model-free BT LA], page 34] [Section G.7 [Model-based BT LA], page 33]

Last updated: 16.08.2004

G.2 Complex segregation analysis

Author: Yurii Aulchenko

Description: Estimation of model of inheritance of the trait in question. Can be done with many user-rude programs

Sub-terms: [Section G.7 [Model-based BT LA], page 33] [Section G.9 [Model-based QT LA], page 33]

Last updated: 16.08.2004

G.3 Estimation of inheritance vector

Author: Yurii Aulchenko

Description: Estimation of IV is usually done ”in house” by GENE-HUNTER family of programs. One can Summary_LA.xml sump that though on the HDD.

More general terms: [Section G.7 [Model-based BT LA], page 33] [Section G.9 [Model-based QT LA], page 33]

Share more general concept with: [Section G.7 [Model-based BT LA], page 33] [Section G.9 [Model-based QT LA], page 33]

Sub-terms: [Section G.7 [Model-based BT LA], page 33] [Section G.12 [Model-free BT LA], page 34] [Section G.9 [Model-based QT LA], page 33]

Last updated: 16.08.2004

G.4 IBD matrix computation

Author: Yurii Aulchenko

Description: For LG framework, use GENE-HUNTER family to compute IBD. For ES framework, one may use SOLAR (see Section H.21 [SOLAR], page 44)

More general terms: [Section G.12 [Model-free BT LA], page 34] [Section G.17 [Variance-component QT LA], page 35]

Share more general concept with: [Section G.12 [Model-free BT LA], page 34] [Section G.12 [Model-free BT LA], page 34] [Section G.16 [Variance component QT LA], page 35]

Sub-terms: [Section G.12 [Model-free BT LA], page 34] [Section G.16 [Variance component QT LA], page 35] [Section G.13 [Non-parametric QT LA], page 35]

Last updated: 16.08.2004

G.5 Linkage analysis

Author: Yurii Aulchenko

Description: The answer to the question ”how do I perform linkage analysis” depends on the type of data you have.

Sub-terms: [Section G.1 [Binary trait LA], page 31] [Section G.14 [Quantitative trait LA], page 35]

Last updated: 16.08.2004

G.6 Model-based BT LA

Author: Yurii Aulchenko

Description: Model-based BT LA may be performed in two steps. At first step, probability distribution of inheritance vectors is egenrated. At second step, based on this information and some model, LA is performed

More general terms: [Section G.1 [Binary trait LA], page 31]

Share more general concept with: [Section G.12 [Model-free BT LA], page 34]

Sub-terms: [Section G.3 [Estimation of inheritance vector], page 32] [Section G.7 [Model-based BT LA], page 33]

Last updated: 16.08.2004

G.7 Model-based BT LA, given IV is known

Author: Yurii Aulchenko

Description: Given the model of inheritance is known, as assessed by CSA, and IV is estimable, one can use GENE-HUNTER (see Section H.5 [GH], page 38) when Lander-Green algorithm is applicable and LINKAGE (see ⟨undefined⟩ [LINKAGE], page ⟨undefined⟩) packages to perform MB BT LA. However, these allow implementing only simplistic major-genic models.

More general terms: [Section G.2 [Complex segregation analysis], page 32] [Section G.3 [Estimation of inheritance vector], page 32] [Section G.7 [Model-based BT LA], page 33]

Share more general concept with: [Section G.9 [Model-based QT LA], page 33] [Section G.12 [Model-free BT LA], page 34] [Section G.3 [Estimation of inheritance vector], page 32]

Sub-terms: [⟨undefined⟩ [Monogenic models for MB BT LA], page ⟨undefined⟩] [⟨undefined⟩ [Polygenic models for MB BT LA], page ⟨undefined⟩] [⟨undefined⟩ [Regressive models for MB BT LA], page ⟨undefined⟩] [⟨undefined⟩ [Mixed models for MB BT LA], page ⟨undefined⟩]

Last updated: 16.08.2004

G.8 Model-based QT LA

Author: Yurii Aulchenko

Description: Model-based quantitative trait linkage analysis may be performed in a number of frameworks

More general terms: [Section G.14 [Quantitative trait LA], page 35]

Share more general concept with: [Section G.17 [Variance-component QT LA], page 35]

Sub-terms: [Section G.3 [Estimation of inheritance vector], page 32] [Section G.9 [Model-based QT LA], page 33]

Last updated: 16.08.2004

G.9 Model-based QT LA, given IV is known

Author: Yurii Aulchenko

Description: Given the model of inheritance is known, as assessed by CSA, and IV is estimable, one can use GENE-HUNTER (see Section H.5 [GH], page 38) when Lander-Green algorithm is applicable and LINKAGE (see ⟨undefined⟩ [LINKAGE], page ⟨undefined⟩) packages to perform MB QT LA. However, these allow implementing only simplistic major-genic models.

More general terms: [Section G.2 [Complex segregation analysis], page 32] [Section G.3 [Estimation of inheritance vector], page 32] [Section G.9 [Model-based QT LA], page 33]

Share more general concept with: [Section G.7 [Model-based BT LA], page 33] [Section G.12 [Model-free BT LA], page 34] [Section G.3 [Estimation of inheritance vector], page 32]

Sub-terms: [⟨undefined⟩ [Monogenic MB QT LA], page ⟨undefined⟩] [⟨undefined⟩ [Regressive models for MB QT LA], page ⟨undefined⟩] [⟨undefined⟩ [Mixed models for MB QT LA], page ⟨undefined⟩]

Last updated: 16.08.2004

G.10 Model-free BT LA

Author: Yurii Aulchenko

Description: Model-free binary trait linkage analysis requires computation of *a priori* IBD matrix (kinship matrix) and genomic point-specific IBD matrix, as assessed with the help of marker data. If this information is available, different forms of model-free BT LA may be performed

More general terms: [Section G.1 [Binary trait LA], page 31]

Share more general concept with: [Section G.7 [Model-based BT LA], page 33]

Sub-terms: [Section G.4 [IBD matrix computation], page 32] [Section G.12 [Model-free BT LA], page 34] [Section G.12 [Model-free BT LA], page 34]

Last updated: 16.08.2004

G.11 Model-free BT LA, given IBD is known

Author: Yurii Aulchenko

Description: If your pedigree is OK for Lander-Green algorithm, use GENE-HUNTER (see Section H.5 [GH], page 38), GENE_HUNTER-PLUS (see Section H.6 [GHP], page 38) or ALLEGRO (see Section H.1 [ALLEGRO], page 36) to compute S_{pairs} function. We are not aware of software solution for Elston-Stewart framework.

More general terms: [Section G.4 [IBD matrix computation], page 32] [Section G.12 [Model-free BT LA], page 34]

Share more general concept with: [Section G.16 [Variance component QT LA], page 35] [Section G.13 [Non-parametric QT LA], page 35] [Section G.4 [IBD matrix computation], page 32] [Section G.12 [Model-free BT LA], page 34]

Last updated: 16.08.2004

G.12 Model-free BT LA, given IV is known

Author: Yurii Aulchenko

Description: If your pedigree is OK for Lander-Green algorithm, use GENE-HUNTER (see Section H.5 [GH], page 38), GENE_HUNTER-PLUS (see Section H.6 [GHP], page 38) or ALLEGRO (see Section H.1 [ALLEGRO], page 36) to compute S_{all} function. We are not aware of software solution for Elston-Stewart framework.

More general terms: [Section G.3 [Estimation of inheritance vector], page 32] [Section G.12 [Model-free BT LA], page 34]

Share more general concept with: [Section G.7 [Model-based BT LA], page 33] [Section G.9 [Model-based QT LA], page 33] [Section G.4 [IBD matrix computation], page 32] [Section G.12 [Model-free BT LA], page 34]

Last updated: 16.08.2004

G.13 Non-parametric QT LA, given IBD is known

Author: Yurii Aulchenko

Description: For sibship... ranks...

More general terms: [Section G.4 [IBD matrix computation], page 32]

Share more general concept with: [Section G.12 [Model-free BT LA], page 34] [Section G.16 [Variance component QT LA], page 35]

Last updated: 16.08.2004

G.14 Quantitative trait LA

Author: Yurii Aulchenko

Description: You can analyse quantitative traits using either variance-components linkage analysis (see ⟨undefined⟩ [VC linkage analysis], page ⟨undefined⟩) or model-based linkage analysis (see Section F.15 [MB linkage analysis], page 27).

More general terms: [Section G.5 [Linkage analysis], page 32]

Share more general concept with: [Section G.1 [Binary trait LA], page 31]

Sub-terms: [Section G.17 [Variance-component QT LA], page 35] [Section G.9 [Model-based QT LA], page 33]

Last updated: 16.08.2004

G.15 test

Author: Yurii Aulchenko

Description: test

Last updated: 16.08.2004

G.16 Variance component QT LA, given IBD is known

Author: Yurii Aulchenko

Description: Use SOLAR (see Section H.21 [SOLAR], page 44) or ASREML (see Section H.3 [ASREML], page 37) to do VC QT LA.

More general terms: [Section G.4 [IBD matrix computation], page 32] [Section G.17 [Variance-component QT LA], page 35]

Share more general concept with: [Section G.12 [Model-free BT LA], page 34] [Section G.13 [Non-parametric QT LA], page 35] [Section G.4 [IBD matrix computation], page 32]

Last updated: 16.08.2004

G.17 Variance-component QT LA

Author: Yurii Aulchenko

Description:

More general terms: [Section G.14 [Quantitative trait LA], page 35]

Share more general concept with: [Section G.9 [Model-based QT LA], page 33]

Sub-terms: [Section G.4 [IBD matrix computation], page 32] [Section G.16 [Variance component QT LA], page 35]

Last updated: 16.08.2004

Appendix H Software

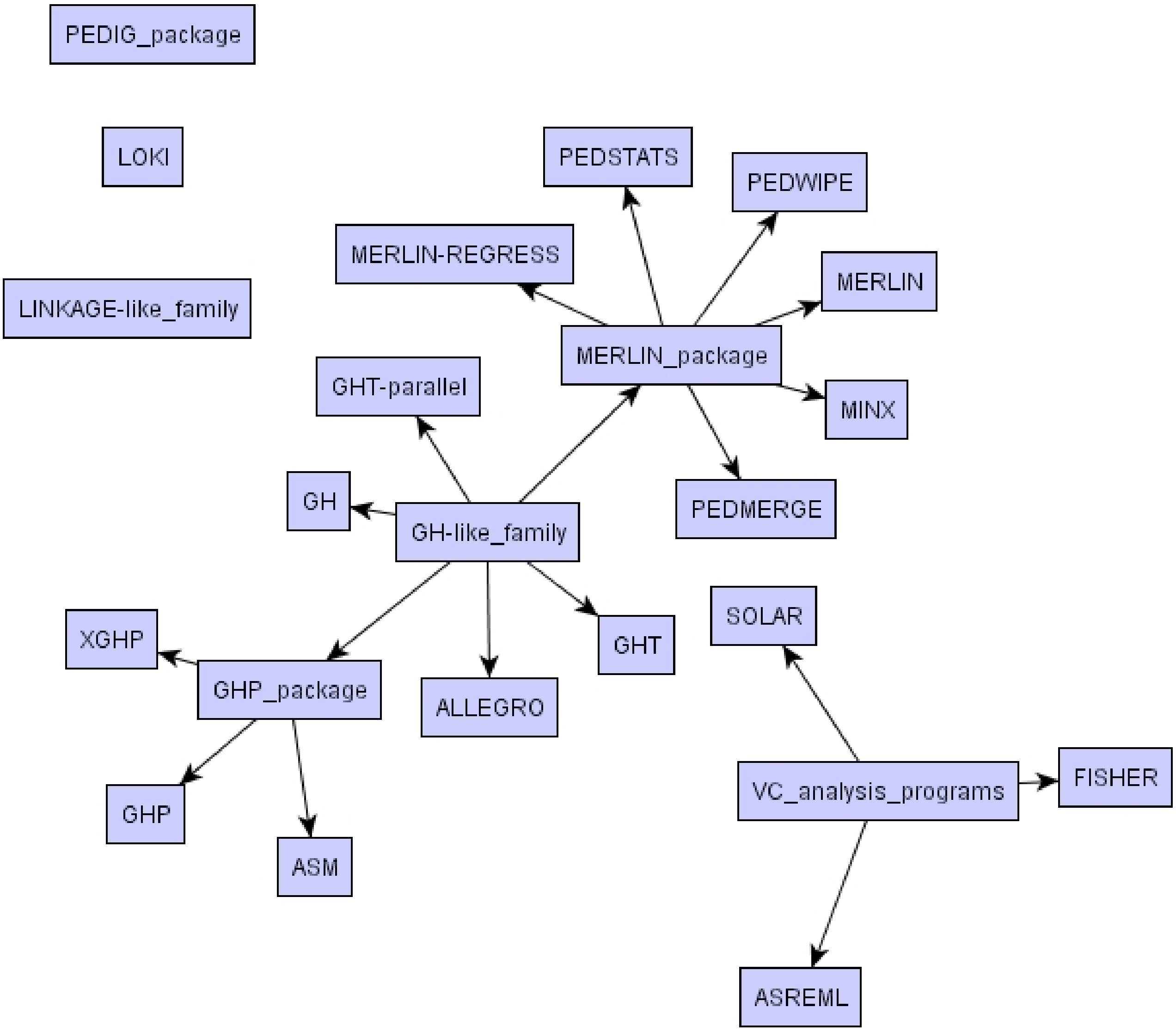

Figure H.1: Structure of Software

H.1 ALLEGRO

Latest release: 1.1d (May 2000)

Author: Daniel F. Gudbjartsson and Kristjan Jonasson, (C) DeCode genetics

Computer language: GNU C

OS: Compiles for Linux, Windows (using Cygwin)

Description: A faster implementation of GENEHUNTER (see Section H.5 [GH], page 38). Has functionality of GH plus two simulation options and recombination estimation; also MF linkage analysis options are more extended in ALLEGRO.

Positive: MF linkage analysis with exponential / linear models, different weighting schemas. Quite fast as compared to any GH-like programs, except MERLIN (see Section H.14 [MERLIN package], page 41) However, MERLIN cannot do model-based linkage analysis and weighted MF linkage analysis.

Negative: It may be a bit tricky to get the program — you have to receive license agreement form from DeCode and sign that and send it back. It is free for academic use, but not for commercial (does one know the price?)

Described in: Gudbjartsson DF, Jonasson K, Frigge M, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nature Genetics 25:12-13

More general terms: [Section H.10 [GH-like family], page 40]

Share more general concept with: [Section H.5 [GH], page 38] [Section H.8 [GHT], page 39] [Section H.7 [GHP package], page 39] [Section H.14 [MERLIN package], page 41]

Last updated: 11.08.2004, 18.08.2004 (Yurii)

H.2 ASM

Author: (C) 1997, 1998 Michael Frigge

Computer language: C

Description: A program from GHP package, which, based on the intermediate output of GHP (Section H.6 [GHP], page 38) does linear / exponential modelling and weighted model-free linkage analysis. See notes at Section H.7 [GHP package], page 39.

External links: `http://galton.uchicago.edu/genehunterplus/`

More general terms: [Section H.7 [GHP package], page 39]

Share more general concept with: [Section H.6 [GHP], page 38] [Section H.23 [XGHP], page 44]

Last updated: 17.08.2004, 18.08.2004 (Yurii)

H.3 ASREML

Latest release: 1.10 (15 May 2003)

Author: Arthur Gilmour (owner: VSN's International)

Description: A program for fast linear mixed modelling, using restricted maximum likleihood. Suited for very fast variance components analysis in large pedigrees, also for multivariate analysis.

Positive: VERY flexible, allows formulation of many, many varieties of models

Negative: Commercial software which costs you ~450 UK pounds. For this money you get a personal license (lifetime, you can request several keys to set up the program at work and at home). May be too flexible. You must be expert to use that in any non-standard form.

External links: `http://www.vsn-intl.com/ASReml/index.htm`

More general terms: [Section H.22 [VC analysis programs], page 44]

Share more general concept with: [Section H.21 [SOLAR], page 44] [Section H.4 [FISHER], page 37]

Last updated: 11.08.2004, 18.08.2004 (Yurii)

H.4 FISHER

Latest release: 2.1

Author: (C) 1985, 1987, 1988 Kenneth Lange

Computer language: F77

OS: compiles with GNU F77: Linux, Cygwin

Description: A set of F77 programs for maximum likelihood variance components analysis on general pedigrees

Negative: You must be understand something about Fortran to make real use of it. Does not utilise marker data.

External links: `http://hpcio.cit.nih.gov/lserver/FISHER.html`

Described in: Lange, Weeks, Boehnke, Genet Epid, 5, 471-471 (1988)

More general terms: [Section H.22 [VC analysis programs], page 44]

Share more general concept with: [Section H.21 [SOLAR], page 44] [Section H.3 [ASREML], page 37]

Last updated: 11.08.2004, 18.08.2004 (Yurii)

H.5 GENEHUNTER (GH)

Latest release: 2.1_r5 beta

Author: 1995-2003 Leonid Kruglyak

Computer language: GNU C

OS: binaries available for Linux, Mac OS X, SunOS; could be compiled with MingW/Cygwin for Windows after minor modifications (constants definition)

Description: An original and famous program for linkage analysis, using Lander-Green (LG) algorithm. Does model free and model based linkage analysis, VC linkage analysis, TDT, "best" haplotypes estimation.

Positive: The oldest (and thus probably least prone to bugs GH-like program (the family was named after it). Quite user-frieandy in terms of interface.

Negative: As compared to GHP (see Section H.7 [GHP package], page 39), ALLEGRO (see Section H.1 [ALLEGRO], page 36), or MERLIN (see Section H.14 [MERLIN package], page 41) no exponential or linear models (as suggested by Kong and Cox) is implemented. No weighting schema implemented for MF linkage analysis (thus restricting MF analysis to set of pedigrees of approximately same size, e.g. sib-pairs).

External links: `http://www.fhcrc.org/labs/kruglyak/Downloads/`

Described in: L. Kruglyak, M.J. Daly, M.P. Reeve-Daly, and E.S. Lander. "Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach". American Journal of Human Genetics 58:1347-1363 (June 1996). For versions 1.2 and above, please also cite: L. Kruglyak and E.S. Lander. "Faster Multipoint Linkage Analysis Using Fourier Transforms". Journal of Computational Biology 5:1-7 (1998).

More general terms: [Section H.10 [GH-like family], page 40]

Share more general concept with: [Section H.8 [GHT], page 39] [Section H.1 [ALLEGRO], page 36] [Section H.7 [GHP package], page 39] [Section H.14 [MERLIN package], page 41]

Last updated: 16.08.2004

H.6 GENEHUNTER-PLUS (GHP)

Author: (C) 1997, 1998 Michael Frigge

Computer language: C

Description: Implementation of GH 1.3, which produce some intermediate files lates used by ASM to do linear / exponential modelling and weighting for MF linkage analysis. See notes at Section H.7 [GHP package], page 39.

External links: `http://galton.uchicago.edu/genehunterplus/`

More general terms: [Section H.7 [GHP package], page 39]

Share more general concept with: [Section H.23 [XGHP], page 44] [Section H.2 [ASM], page 37]

Last updated: 17.08.2004, 18.08.2004 (Yurii)

H.7 GENEHUNTER-PLUS package (GHP package)

Latest release: 1.2 (based on GH 1.2)

Author: (C) 1997, 1998 Michael Frigge

Computer language: C

Description: Implementation of GH 1.3, which introduces linear and exponential models and weighting for MF linkage analysis.

Positive: One of the few packages which allows weighting for pedigree size. As ALLEGRO (see Section H.1 [ALLEGRO], page 36) and MERLIN (see Section H.14 [MERLIN package], page 41), allows for exponential and linear models. X-chromosome analysis is possible.

Negative: Consider using this if you will to do weighting, exp / lin model in MF linkage analysis or X-chromosome analysis, as it is based on old GH 1.3 and thus the program is quite slow. It is not too user friendly, as to perform exponential / linear modeling and / or weighted analysis one has to call an extra program (ASM) from the suite.

External links: `http://galton.uchicago.edu/genehunterplus/`

Described in: A. Kong and N.J. Cox. "Allele-Sharing Models: LOD Scores and Accurate Linkage Tests". American Journal of Human Genetics 61:1179-1188 (November, 1997).

More general terms: [Section H.10 [GH-like family], page 40]

Share more general concept with: [Section H.5 [GH], page 38] [Section H.8 [GHT], page 39] [Section H.1 [ALLEGRO], page 36] [Section H.14 [MERLIN package], page 41]

Sub-terms: [Section H.6 [GHP], page 38] [Section H.23 [XGHP], page 44] [Section H.2 [ASM], page 37]

Last updated: 17.08.2004

H.8 GENEHUNTER-TWOLOCUS (GHT)

Author: (C) 1999 Konstantin Strauch

Computer language: C

Description: Implementation of GH 1.3, which makes possible MB linkage analysis with imprinting and two-locus

Positive: The only program implementing imprinting and twolocus parametric models

Negative: Is a bit too slow, as based on old GH 1.3. MAXBIT 15 is kind of maximum you can analyse with a PC, using two-locus model. Available only on request from author (positive: this happens fast!).

Described in: K. Strauch, R. Fimmers, T. Kurz, K.A. Deichmann, T.F. Wienker, and M.P. Baur. "Parametric and Nonparametric Multipoint Linkage Analysis with Imprinting and Two-Locus-Trait Models: Application to Mite Sensitization". American Journal of Human Genetics 66:1945-1957 (June 2000).

More general terms: [Section H.10 [GH-like family], page 40]

Share more general concept with: [Section H.5 [GH], page 38] [Section H.1 [ALLEGRO], page 36] [Section H.7 [GHP package], page 39] [Section H.14 [MERLIN package], page 41]

Last updated: 17.08.2004, 18.08.2004 (Yurii)

H.9 GENEHUNTER-TWOLOCUS-PARALLEL (GHT PARALLEL)

Author: (C) 2003 (?) Konstantin Strauch and (?)

Computer language: C

Description: Highly optimized and parallelised re-implementation of GHT (see Section H.8 [GHT], page 39) GH 1.3,

Positive: The only program implementing imprinting and twolocus parametric models

Negative: MAXBIT 15 is kind of maximum you can analyse on a single PC, using two-locus model. Available only on request from author (positive: this happens fast!).

Described in: Johannes Dietter, Alexander Spiegel, Dieter an Mey, Hans-Joachim Pflug, Hussam Al-Kateb, Katrin Hoffmann, Thomas F Wienker and Konstantin Strauch Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia. 2004. Eur. J. Hum. Genet. (in press)

Last updated: 17.08.2004, 18.08.2004 (Yurii)

H.10 GH-like family

Description: Programs implementing Lander-Green algorithm (see Section F.11 [LG algorithm], page 26) for multipoin model-based (see Section F.15 [MB linkage analysis], page 27) and / or MF (see Section F.16 [MF linkage analysis], page 27) linkage analysis. Usual limitations of LG on pedigree size applies.

Sub-terms: [Section H.5 [GH], page 38] [Section H.8 [GHT], page 39] [⟨undefined⟩ [GHT-parallel], page ⟨undefined⟩] [Section H.1 [ALLEGRO], page 36] [Section H.7 [GHP package], page 39] [Section H.14 [MERLIN package], page 41]

Last updated: 18.08.2004 (Yurii)

H.11 LINKAGE-like family

Description: Programs for VC analysis (see Section F.31 [VC analysis], page 30): heritability estimation, including locus-specific.

Sub-terms: [⟨undefined⟩ [FASTLINK], page ⟨undefined⟩] [⟨undefined⟩ [VITESSE], page ⟨undefined⟩]

Last updated: 18.08.2004 (Yurii)

H.12 LOKI

Latest release: 2.4.5 (March 2003)

Author: Simon Heath

Computer language: GNU C

OS: Compiles well with GNU C on Linux

Description: A program for Monte-Carlo Markov-Chain (MCMC) segregation and linkage analysis. Can be also used to compute *a priori* or genomic IBD.

External links: `http://www.stat.washington.edu/thompson/Genepi/Loki.shtml`

Described in: Heath S. "Markov Chain Monte Carlo Segregation and Linkage Analysis for Oligogenic Models." American Journal of Human Genetics 61 (1997): 748-760.

Last updated: 18.08.2004 (Yurii)

H.13 MERLIN

Latest release: Version: 0.10.2

Author: (C) 2000-2003 Goncalo Abecasis

Computer language: GNU C++

OS: binaries available for Linux, Sun OS, Windows, MacOS X G5

Description: A program for multipoint linkage analysis using Lander-Green (LG) algorithm. Usual limitations apply. MERLIN uses sparse trees to represent gene flow in pedigrees. The program carries out single-point and multipoint analyses of pedigree data, including IBD and kinship calculations, nonparametric and variance component linkage analyses, error detection and information content mapping. For multipoint analyses in dense maps, Merlin allows the user to impose constraints on the number of recombinants between consecutive markers (thus increase the speed). Merlin estimates haplotypes by finding the most likely path of gene flow or by sampling paths of gene flow at all markers jointly. It can also list all possible nonrecombinant haplotypes within short regions. Finally, Merlin provides swap-file support for handling very large numbers of markers as well as gene-dropping simulations for estimating empirical significance levels.

Positive: Probably the fastest implementation, deal with pedigrees of size ˜ 30 bits. Except standard IBD, estimates identity coefficients (extended IBD). Can sample haplotypes from postrior distribution

Negative: Does not attempt model-based linkage analysis. There is no any weighting schema implemented in MF linkage analysis

External links: http://www.sph.umich.edu/csg/abecasis/Merlin/

More general terms: [Section H.14 [MERLIN package], page 41]

Share more general concept with: [Section H.15 [MERLIN-REGRESS], page 42] [Section H.16 [MINX], page 42] [Section H.19 [PEDSTATS], page 43] [Section H.18 [PEDMERGE], page 43] [Section H.20 [PEDWIPE], page 43]

Last updated: 07.08.2004, 18.08.2004 (Yurii)

H.14 MERLIN package

Latest release: Version: 0.10.2

Author: (C) 1999-2004 Goncalo Abecasis

Computer language: GNU C++

OS: binaries available for Linux, Sun OS, Windows, MacOS X G5

Description: A suite of programs for pedigree analysis

External links: http://www.sph.umich.edu/csg/abecasis/Merlin/

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

More general terms: [Section H.10 [GH-like family], page 40]

Share more general concept with: [Section H.5 [GH], page 38] [Section H.8 [GHT], page 39] [Section H.1 [ALLEGRO], page 36] [Section H.7 [GHP package], page 39]

Sub-terms: [Section H.13 [MERLIN], page 41] [Section H.15 [MERLIN-REGRESS], page 42] [Section H.16 [MINX], page 42] [Section H.19 [PEDSTATS], page 43] [Section H.18 [PEDMERGE], page 43] [Section H.20 [PEDWIPE], page 43]

Last updated: 18.08.2004 (Yurii)

H.15 MERLIN-REGRESS

Latest release: Version: 0.10.2

Author: (C) 2000-2003 Goncalo Abecasis

Computer language: GNU C++

OS: binaries available for Linux, Sun OS, Windows, MacOS X G5

Description: A regression-based replacment for VC analysis by MERLIN. Also estimates informativity / ELOD.

Positive: Applicable to selected samples and robust to minor deviations from normality

Negative: Requires knwolege of mean, var. and covar. of trait distribution in population (rarely known), may be slow (?)

External links: `http://www.sph.umich.edu/csg/abecasis/Merlin/`

Described in: Sham PC, Purcell S, Cherny SS and Abecasis GR Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet (2002) 71:238-53

More general terms: [Section H.14 [MERLIN package], page 41]

Share more general concept with: [Section H.13 [MERLIN], page 41] [Section H.16 [MINX], page 42] [Section H.19 [PEDSTATS], page 43] [Section H.18 [PEDMERGE], page 43] [Section H.20 [PEDWIPE], page 43]

Last updated: 18.08.2004 (Yurii)

H.16 MINX

Latest release: Version: 0.10.2

Author: (C) 2000-2003 Goncalo Abecasis

Computer language: GNU C++

OS: binaries available for Linux, Sun OS, Windows, MacOS X G5

Description: MERLIN fo X-chromosome analysis

External links: `http://www.sph.umich.edu/csg/abecasis/Merlin/`

More general terms: [Section H.14 [MERLIN package], page 41]

Share more general concept with: [Section H.13 [MERLIN], page 41] [Section H.15 [MERLIN-REGRESS], page 42] [Section H.19 [PEDSTATS], page 43] [Section H.18 [PEDMERGE], page 43] [Section H.20 [PEDWIPE], page 43]

Last updated: 18.08.2004 (Yurii)

H.17 PEDIG package

Author: Didier Biochard

Description: This software, specifically developed for the analysis of large pedigrees (tens of thousands of subjects), is a set of independent programs written in F77, to calculate probabilities of gene origin, relationship and inbreeding coefficient, and to characterize the quality of pedigree information. Also pedigree manipulation.

Positive: Really fast and useful.

Negative: Pedigree encoding is a bit specific: people must be numbered by conseqутive integers, from 1 to N; also this is not a standard linkage file as it must include three extra fields

External links: `http://dga.jouy.inra.fr/sgqa/diffusions/pedig/pedigE.htm`

Described in: Boichard D., Maignel L., et Verrier E. (1997). The value of using probabilities of gene origin to measure genetic variability in a population. Genet. Sel. Evol 29, 5-23. ALSO Maignel L., Boichard D., Verrier E. (1996). Genetic variability of French dairy breeds estimated from pedigree information. Interbull meeting, Veldhoven, Pays Bas, 23-24 Juin 1996, Interbull Bull., 14, 49-54

Sub-terms: [⟨undefined⟩ [NGEN], page ⟨undefined⟩] [⟨undefined⟩ [MEUW], page ⟨undefined⟩] [⟨undefined⟩ [VANRAD], page ⟨undefined⟩] [⟨undefined⟩ [PAR], page ⟨undefined⟩] [⟨undefined⟩ [PAR2], page ⟨undefined⟩] [⟨undefined⟩ [PAR3], page ⟨undefined⟩] [⟨undefined⟩ [PARENTE], page ⟨undefined⟩] [⟨undefined⟩ [PROB_ORIG], page ⟨undefined⟩] [⟨undefined⟩ [SEGREG], page ⟨undefined⟩] [⟨undefined⟩ [ETR], page ⟨undefined⟩] [⟨undefined⟩ [INTGEN], page ⟨undefined⟩]

Last updated: 16.08.2004, 18.08.2004 (Yurii)

H.18 PEDMERGE

Author: (C) 1999 Goncalo Abecasis

Computer language: GNU C++

OS: binaries available for Linux, Sun OS, Windows, MacOS X G5

Description: Attempts to merge paired pedigree and data files

External links: http://www.sph.umich.edu/csg/abecasis/Merlin/

More general terms: [Section H.14 [MERLIN package], page 41]

Share more general concept with: [Section H.13 [MERLIN], page 41] [Section H.15 [MERLIN-REGRESS], page 42] [Section H.16 [MINX], page 42] [Section H.19 [PEDSTATS], page 43] [Section H.20 [PEDWIPE], page 43]

Last updated: 18.08.2004 (Yurii)

H.19 PEDSTATS

Latest release: Version: 0.4.6

Author: (C) 1999-2004 Goncalo Abecasis (1999-2004), (C) 2002-2004 Jan Wigginton

Computer language: GNU C++

OS: binaries available for Linux, Sun OS, Windows, MacOS X G5

Description: Provide pedigree description statistics, also some description of genotypes. PDF options are especially interesting.

External links: http://www.sph.umich.edu/csg/abecasis/Merlin/

More general terms: [Section H.14 [MERLIN package], page 41]

Share more general concept with: [Section H.13 [MERLIN], page 41] [Section H.15 [MERLIN-REGRESS], page 42] [Section H.16 [MINX], page 42] [Section H.18 [PEDMERGE], page 43] [Section H.20 [PEDWIPE], page 43]

Last updated: 18.08.2004 (Yurii)

H.20 PEDWIPE

Author: (C) 2002 Goncalo Abecasis

Computer language: GNU C++

OS: binaries available for Linux, Sun OS, Windows, MacOS X G5

Description: Updates genotypes from MERLIN pedigree file by deleting possible genotypic errors as listed in merlin.err (the later is produced when using –error option of MERLIN).

External links: http://www.sph.umich.edu/csg/abecasis/Merlin/

More general terms: [Section H.14 [MERLIN package], page 41]

Share more general concept with: [Section H.13 [MERLIN], page 41] [Section H.15 [MERLIN-REGRESS], page 42] [Section H.16 [MINX], page 42] [Section H.19 [PEDSTATS], page 43] [Section H.18 [PEDMERGE], page 43]

Last updated: 18.08.2004 (Yurii)

H.21 Sequential Oligogenic Linkage Analysis Routines (SOLAR)

Latest release: 2.1.2 (Feb. 21, 2004)

Author: John Blangero, Kenneth Lange, Tom Dyer, Laura Almasy, Harald G?ring, Jeff Williams, and Charles Peterson SOLAR is (C) 1995-2004, Southwest Foundation for Biomedical Research.

Description: A sute of programs for variance components linkage analysis, using maximum likelihood. Allows heritability estimation (see Section F.9 [heritability], page 25) and variance components linkage analysis with covariates. Bivariate analysis implemented. Computational core: FISHER (see Section H.4 [FISHER], page 37)

Positive: Positive: quite user-friendly, does not require too sofisticated files, well-documented.

Negative: Linkage analysis with really large pedigrees wityh multiple non-genotyped members takes ages because of IBD estimation step. Also, ”multipoint” IBD is not really multipoint, as single-marker IBD computations are done and then between-marker points are regressed. Binary traits analysis does not seem to work too well.

External links: `http://www.sfbr.org/solar/`

Described in: Almasy L, Blangero J (1998) Multipoint quantitative trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198-1211. ALSO MORE REFs FOR BIVARIATE, LIABILITY THRESHOLD, ETC – SEE LINKs

More general terms: [Section H.22 [VC analysis programs], page 44]

Share more general concept with: [Section H.4 [FISHER], page 37] [Section H.3 [ASREML], page 37]

Last updated: 11.08.2004, 18.08.2004 (Yurii)

H.22 VC analysis programs

Description: Programs for VC analysis (see Section F.31 [VC analysis], page 30): heritability estimation, including locus-specific.

Sub-terms: [Section H.21 [SOLAR], page 44] [Section H.4 [FISHER], page 37] [Section H.3 [ASREML], page 37]

Last updated: 18.08.2004 (Yurii)

H.23 X-GENEHUNTER-PLUS (XGHP)

Author: (C) 1997, 1998 Michael Frigge

Computer language: C

Description: Implementation of GH 1.3, which allows for GH-like X-chromosome analysis, also produce some intermediate files lates used by ASM to do linear / exponential modelling and weighting for MF linkage analysis. See notes at Section H.7 [GHP package], page 39.

External links: `http://galton.uchicago.edu/genehunterplus/`

More general terms: [Section H.7 [GHP package], page 39]

Share more general concept with: [Section H.6 [GHP], page 38] [Section H.2 [ASM], page 37]

Last updated: 17.08.2004, 18.08.2004 (Yurii)

Appendix I Information for contributors

The text is a texinfo (http://texinfo.org/) file.

If you woukd like to contribute some text, please ask Yurii (yurii@bionet.nsc.ru) to send the source code or send the text in plain ASCII.

The Glossary and Software appendices are generated from the same form of XML source. If you would like to contribute, please provide the topics using the following form:

```
<term>
        <name></name>
        <author></author>
        <release></release>
        <clang></clang>
        <OS></OS>
        <meaning></meaning>
        <positive></positive>
        <negative></negative>
        <child></child>
        <relterm></relterm>
        <link></link>
        <manuscript></manuscript>
        <updated></updated>
</term>
```

Any order is accepted. There may be more then one <child></child> pairs and also <relterm></relterm>. Other fields must be unique. Explanation:

```
<term>
        <name>Name to appear here</name>
        <author>For a software: put author here</author>
        <release>
                For a software: put version and date of release here
        </release>
        <meaning>
                Describe the term / software here
        </meaning>
        <positive>
                Usually for software, list positive features here
        </positive>
        <negative>
                Usually for software, list negative features here
        </negative>
        <child>children-node 1 comes here</child>
        <child>children-node 2 comes here</child>
        <child>children-node ... comes here</child>
        <relterm>
                put related terms / software here (do not put
                parents and sibs! --- these are automatively
                deduced from <child> description; also do not
                put any relations which could be deduced from
                ''children'' description)
        </relterm>
        <link>@uref{Put external web link here}</link>
        <clang>For software, put the computer language here</clang>
        <OS>For a software, put operation ssytem here</OS>
        <manuscript>Put references here</manuscript>
        <updated>Put information of the update date, also your name</updated>
</term>
```

For example, entry GENE-HUNTER-PLUS package (Section H.7 [GHP package], page 39) looks internally like this:

```
<term>
        <name>
                GENEHUNTER-PLUS package
        </name>
        <author>
                (C) 1997, 1998 Michael Frigge
        </author>
        <release>
                1.2 (based on GH 1.2)
        </release>
        <abbreviation>GHP package</abbreviation>
        <meaning>
                Implementation of GH 1.3, which introduces linear
                and exponential models and weighting for MF linkage analysis.
        </meaning>
        <positive>
                One of the few packages
                which allows weighting for pedigree size. As
                ALLEGRO (@pxref{ALLEGRO}) and MERLIN
                (@pxref{MERLIN package}), allows for exponential and
                linear models. X-chromosome analysis is possible.
        </positive>
        <negative>
                Consider using this if you will to do weighting,
                exp / lin model in MF linkage analysis or X-chromosome analysis, as
                it is based on old GH 1.3 and thus the program
                is quite slow.
                It is not too user friendly, as to
                perform exponential / linear modeling and / or
                weighted analysis one has to call an extra program
                (ASM) from the suite.
        </negative>
        <example>
        </example>
        <manuscript>
                A. Kong and N.J. Cox. "Allele-Sharing Models: LOD
                Scores and Accurate Linkage Tests". American Journal
                of Human Genetics 61:1179-1188 (November, 1997).
        </manuscript>
        <link>
                @uref{http://galton.uchicago.edu/genehunterplus/}
        </link>
        <clang>C</clang>
        <child>GHP</child>
        <child>XGHP</child>
```

```
          <child>ASM</child>
          <manuscript>
          </manuscript>
          <updated>
                  17.08.2004
          </updated>
  </term>
```

Index