In Press, Human Heredity 7/2011

KELVIN: a Software Package for Rigorous Measurement of Statistical Evidence in Human Genetics

Veronica J. Vieland, Yungui Huang, Sang-Cheol Seok, John Burian, Umit Catalyurek, Jeffrey O'Connell, Alberto Segre, William Valentine-Cooper

Running Title: Kelvin: Genetic Analysis Software

Corresponding Author:

Veronica J. Vieland, Ph.D.
The Research Institute at Nationwide Children's Hospital, Columbus OH
700 Children's Drive
Columbus, Ohio 43205
Veronica.Vieland@nationwidechildrens.org
Phone: (614) 355-2861
Fax: (614) 355-2728

Abstract

This paper describes the software package KELVIN, which supports the PPL framework for the measurement of statistical evidence in human (or more generally, diploid) genetic studies. In terms of scope, KELVIN supports two-point (trait-marker or marker-marker) and multipoint linkage analysis, based on either sex-averaged or sex-specific genetic maps, with an option to allow for imprinting; trait-marker linkage disequilibrium, or association analysis, in case-control data, trio data, and/or multiplex family data, with options for joint linkage and trait-marker linkage disequlibrium (LD) or conditional LD given linkage; dichotomous trait, quantitative trait and quantitative trait threshold models; and certain types of gene-gene interactions and covariate effects. Features and data (pedigree) structures can be freely mixed and matched within analyses. The statistical framework is specifically tailored to accumulate evidence in a mathematically rigorous way across multiple data sets or data subsets while allowing for multiple sources of heterogeneity, and KELVIN itself utilizes sophisticated software engineering to provide a powerful and robust platform for studying the genetics of complex disorders.

Introduction

This paper describes the statistical genetic software package KELVIN. KELVIN is a relatively comprehensive package for linkage and/or association analysis, based on familiar forms of genetic likelihoods (with some unique extensions of these likelihoods, see below). Its options include: two-point (trait-marker or marker-marker) and multipoint linkage analysis, based on either sex-averaged or sex-specific genetic maps (covering the autosomes, X, and the pseudoautosomal region of X), with an option to allow for imprinting; trait-marker linkage disequilibrium (LD), or association analysis, in case-control data, trio data, and/or multiplex family data, with options for joint linkage and LD or conditional LD given linkage; dichotomous trait (DT), quantitative trait (QT) and QT threshold (QTT) models; and certain types of gene-gene interactions and covariate effects.

The underlying algorithm is Elston-Stewart based [1], permitting analysis of fairly large pedigrees including loops; multipoint analyses [2] are done automatically by walking down each chromosome using a user-specified number of markers to perform the calculations at each position [3]. In-house implementations currently allow the use of Lander-Green [4] or MCMC algorithms for handling multipoint marker data; these options are slated for future incorporation into KELVIN itself. KELVIN accepts mixtures of different pedigree structure (cases/controls, trios, sib-pairs, nuclear families, extended pedigrees), and analysis options can be easily combined with one another. A custom graphing program, KELVIZ, facilitates visualization of KELVIN output. Sophisticated software engineering, as described below, makes possible some types of calculations that cannot be done by other programs. A rigorous protocol for semi-automated testing of all revisions to the code is employed.

KELVIN has been developed within a broader philosophical program of research focused on how to *measure statistical evidence,* and in particular, how to *accumulate evidence* as data are acquired [5, 6]. As a result, KELVIN's calculations are neither frequentist nor Bayesian nor strictly "evidentialist" [7]. Rather, they require the user to become comfortable with a novel framework, although arguably a framework that is in many ways simpler to interpret than the alternatives.

We begin, therefore, with (1) a brief historical description of KELVIN's original purpose and form. We then (2) give an overview of the full set of statistical models currently available to the user. In section (3), we describe the underlying computational architecture used to accomplish the calculations. Each of the KELVIN features described in this paper has been the subject of one or more previous peer-reviewed publications. Here we forego mathematical details and give instead a general overview, with citations to the appropriate source materials. We return to philosophical matters in the Discussion.

(1) Historical Development of KELVIN

1.1 Original Motivation Development of KELVIN was originally motivated by one simple question: When performing linkage analysis, why not get what we really want? Vieland [8] argued that what we really want to know is the probability of a clinically relevant gene at any given genomic position g based on the available data D, or in brief, P_g(linkage | D). But what we usually measure is instead P_g(D | no linkage), which is at best only indirectly related to the quantity of true interest and can under some circumstances actually be misleading (see, e.g., [7]).

In 1959, Smith [9] proposed using Bayes' theorem to directly compute what we really wanted to know, P_g(linkage | D). His proposal was of limited practical value at the time, however, for purely computational reasons. But by the 1990s computational difficulty no longer seemed a sufficient reason to forego calculating a quantity of interest. Thus, as argued in Vieland [8], all that seemed to stand between us and getting what we really wanted was to overcome our squeamishness at the use of the mathematical device of assigning prior distributions to parameters. In view of the subsequent acceptance of Bayesian methods in human genetics, it seems somewhat quaint to have had to belabor the argument at all.

In any case, that was when we began to consider a statistic we called the PPL, which stood for "posterior probability of linkage." As originally proposed, the PPL was simply an application of Bayes' theorem to genetic linkage analysis, and as such it looked quite Bayesian in spirit. Vieland [8] was careful to distinguish between accepting the PPL as a valid measure of linkage and adopting Bayesianism as a broader philosophical program of statistical investigation. However, it is only more recently that we have come to view the PPL as a truly non-Bayesian measure. We return to this topic in the Discussion.

1.2 The original PPL Vieland [8] proposed a form of the PPL based on the ordinary LOD score, calculated as was usual at the time under a given trait model. Bayes' theorem tells us that

$$PPL \triangleq P(\theta < \tfrac{1}{2} \mid D) = \frac{P(D \mid \theta < \frac{1}{2})P(\theta < \frac{1}{2})}{P(D \mid \theta < \frac{1}{2})P(\theta < \frac{1}{2}) + P(D \mid \theta = \frac{1}{2})P(\theta = \frac{1}{2})}$$

$$= \frac{\int_{\theta<\frac{1}{2}} L(\theta \mid D) f(\theta \mid \theta < \frac{1}{2})\, \pi \, d\theta}{\int_{\theta<\frac{1}{2}} L(\theta \mid D) f(\theta \mid \theta < \frac{1}{2})\, \pi \, d\theta + L(\theta = \frac{1}{2} \mid D)(1-\pi)}, \qquad Eq\ 1$$

where L(θ | D) is the likelihood for the recombination fraction θ, $f(\theta|\theta < \frac{1}{2})$ is the prior probability distribution for θ under the hypothesis of "linkage," and π = P(linkage) = P(θ < ½). Dividing each term through by L(θ = ½ | D), we can rewrite *Eq 1* as

$$PPL = \frac{\int 10^{LOD(\theta)} f(\theta \mid \theta < \frac{1}{2})\, \pi \, d\theta}{\int 10^{LOD(\theta)} f(\theta \mid \theta < \frac{1}{2})\, \pi \, d\theta + (1-\pi)}. \qquad Eq\ 2$$

This formulation makes clear that any computer program that can compute LOD scores can be used to calculate the PPL using *Eq 2*. It also establishes that the PPL is intrinsically ascertainment corrected: replacing the likelihood with the (exponentiated) LOD means that we are implicitly computing the likelihood for the marker data given the trait data [10, 11], for a form of "ascertainment assumption free" calculation [12] (see also [13, 14]).

In terms of the prior distribution for θ, we initially tried various approaches, including a β prior and an empirical Bayes approach [15]. However, in the end we settled on a simple step function, with uniform prior within steps but a higher "weight" on small values of θ; specifically, we set 95% of the prior for $\theta < ½$ to be uniform over the interval [0,..,0.05) and the remaining 5% to be uniform over [0.05,..,0.50). This prior proved to have excellent behavior under both "linkage" and "no linkage." We set $\pi = 2\%$ based on earlier calculation [16] of the probability for two random loci to be linked. This is conservative for multilocus disorders, since the prior probability of one of several genes being within linkage distance of a given genomic position is higher than 2%.

As argued in [8], one clear advantage of the PPL as a linkage statistic is simplicity of interpretation. A PPL of, say, 40% simply means that there is a 40% probability of linkage to the given marker or location, for the clinical trait under study and based on the available data. This number is readily interpreted just like any other probability – say, the probability of rain or of a successful medical intervention – and, in Smith's [9] words, it "does not need to be hedged about with qualifications, unlike a significance level." Of course this does not answer the question of whether 40% represents strong, or "strong enough," evidence for linkage. Whether we act on this evidence and in what way (say, with molecular follow up) depends on many factors, including availability of resources and alternative uses for those resources. The PPL is one piece of information available to inform such decisions, but it is not itself a decision making procedure.

By the same token, PPLs are not interpreted in terms of their associated Type 1 error rates (since there is no decision being made, there is no associated error probability). Thus the familiar paradigm of evaluating methods by assessing power for fixed-size tests is generally not applicable to statistics in the PPL framework. However, the methods development papers cited in what follows show various types of comparisons with familiar methods. The interested reader is also referred to [17-20] for some recent examples of the "power" of the PPL framework in a more general sense.

1.3 Sequential Updating It was known that including the admixture parameter α [21] in the linkage likelihood was useful for handling heterogeneity within any given data set (see, e.g., [22] for an early application; see also [23]), and the PPL could be easily reformulated in terms of the heterogeneity LOD (HLOD [24]). But in the presence of appreciable heterogeneity, the proportional representation of any given genetic form of disease can vary substantially from data set to data set as a function of multiple factors, including differences in ascertainment and clinical procedures, differences in ancestries across catchment areas, and simple sampling fluctuations. This fact undermines the use of independent replication to sort true from false findings [25], but suggests another approach to utilizing multiple data sets based on the mathematically rigorous accumulation of evidence across data sets, when the data sets can be reasonably expected to differ from one another in important ways.

We again simply use Bayes' theorem for this. We define the Bayes Ratio (BR) as

$$BR(\theta \mid D) = \int 10^{HLOD} f(\theta) f(\alpha)\, d\alpha. \qquad \text{Eq 3}$$

(The BR is a function of any parameters not being integrated out, thus in *Eq 3* it is a function of θ as indicated by the notation. In general we simply say "BR," annotating which parameters are not integrated out only as necessary to avoid ambiguity. The BR is essentially an integrated (conditional) likelihood, but the integration here is over a ratio of likelihoods as a unit, as above; see also [26].) We assume here and throughout that the prior *f*(α) is uniform (0,..,1).

Appropriately substituting BR(θ|D) into *Eq 2*, we obtain a PPL for data D based on the HLOD. (See [27] for theoretical properties of this form of the PPL.) Generalizing to *m* data sets, and allowing α to vary across data sets, the BR becomes

$$BR(\theta \mid D_1,\ldots,D_m) = \prod_{i=1}^{m} BR(\theta \mid D_i), \qquad \textit{Eq 4}$$

and the PPL is calculated as

$$PPL = \frac{\int_{\theta<\frac{1}{2}} BR(\theta \mid D_1,\ldots,D_m) f(\theta \mid \theta < \tfrac{1}{2})\, \pi \, d\theta}{\int_{\theta<\frac{1}{2}} BR(\theta \mid D_1,\ldots,D_m) f(\theta \mid \theta < \tfrac{1}{2})\, \pi \, d\theta + (1-\pi)}. \qquad \textit{Eq 5}$$

Note that the order in which the data sets are considered is irrelevant [28].

This then gives us a recipe for calculating what we really want to calculate across multiple data sets: (i) start with any program that calculates HLODs, and some numerical integration method; (ii) use *Eq 3* to calculate the marginal BR(θ) for each available data set (integrating out α); (iii) multiply BRs across data sets using *Eq 4*; (iv) use Bayes' theorem to transform the results back onto the probability scale using *Eq* 5. Note that once the BRs have been calculated for a given set of data, there is no need to reanalyze the data in order to measure the evidence considered in aggregate with another set of data. This also permits investigators to perform joint analyses without ever having to share data, since all that is required to obtain the PPL across data sets is an output file with the BR for each data set.

The process of multiplying the BR's across data sets is known as sequential updating. Here α is being treated as a nuisance parameter, integrated over separately for each data set and therefore allowed to vary freely across data sets. We have showed that this simple procedure for accumulating evidence across heterogeneous sets of genetic data has superior statistical properties compared to several alternatives [28-30]. This simple technique for accumulating evidence across data sets, while allowing nuisance parameters to vary across data sets, lies at the heart of all of the PPL's underlying methodology.

Sequential updating can be used to accumulate evidence across data collected at different sites, by different investigators, or at different time periods. It can also be used to accumulate evidence within data sets across data subsets, say, divided by ancestry or other demographic features or by clinical features. It is beneficial even when the basis for subdivision is not a perfect classifier of homogeneous subsets, and it is only mildly detrimental when data are (inadvertently) subdivided on random (genetically irrelevant) variables [31, 32]. However, in the presence of relatively homogeneous genetic effects across data sets, "pooling" families together for a single, combined analysis is always more powerful than sequentially updating across families. See [31] for practical considerations when classifying data subsets for purposes of sequential updating.

There is one other important distinguishing feature of the sequentially updated PPL: it accumulates evidence both for linkage and also against linkage as data accrue. This is because the BR will be < 1 when the data favor the denominator of the HLOD ("no linkage") over the numerator ("linkage"), as a direct benefit of utilizing integration rather than maximization to handle unknown parameters.

The combination of sequential updating and accumulation of evidence both for and against linkage gives graphs of the PPL their characteristic appearance, as was already evident in the very first application we tried [33]. As Figure 1 illustrates, PPL plots tend to show very clear separation of signal from background noise, making visual interpretation straightforward. In this application the sequentially updated PPL was also singularly successful at localizing all four of the underlying trait genes based on several heterogeneous sets of simulated data, compared with various meta-analytic approaches [34].

While the PPL and the philosophical framework in which it is embedded have both evolved considerably since the early applications, these two themes – figuring out how to compute what we really want to know, and ensuring the correct accumulation of evidence across multiple data sets – distinguish the focus of KELVIN and have influenced all facets of its development. We note also that originally we had an additional motivation in mind, namely the prospect of being able to use Bayes' theorem to include prior genomic information, e.g., regarding known sex differences in genetic maps. While this remains a topic of interest, thus far in application to typical human genetic data sets it has proved less useful than we had anticipated [35].

(2) Extension of statistical models

2.1 Integration over trait model parameters The original PPL utilized a simple parameterization for a dichotomous trait (DT), based on extensive research on robust approximating single-locus models for multi-locus traits (see e.g., [23, 36-39] and the work of many others). The model was parameterized in terms of a disease allele frequency p, three penetrances f_{DD}, f_{Dd}, f_{dd} corresponding to each possible trait genotype (assuming a biallelic locus), and the heterogeneity parameter α.

Initially, following current practice at the time, we assumed that it was necessary to fix the parameters of the trait model, even if that meant fixing them at wrong values. However, theoretical developments had already made clear that this was not necessary [10-13, 40], and that maximum likelihood could be used to estimate the trait models from linkage data, by maximizing the LOD over the trait model, i.e., calculating the MOD [40]. But in the presence of heterogeneity, the trait model itself can reasonably be expected to differ across data sets. In this context, the MOD's use of maximization to handle the unknown trait model is problematic, because maximizing separately in each data set results in a statistic that tends to infinity even at unlinked loci [28].

By contrast, Bayes' theorem lends itself immediately to integrating over the nuisance parameters of the unknown trait model while maintaining the ability of sequential updating to accumulate evidence both for and against linkage. Our first application of this approach involved simple model averaging [41], based on uniform priors for each parameter with some constraints (e.g., an ordering constraint on the penetrances). This simple approach appeared to work quite well, both in simulation studies [42] and in application to real data [41, 43]. Integration over the trait parameter space is now automatic for all forms of the PPL, using essentially this same approach. Thus from this point forward, letting γ represent the parameters of the trait model, we assume that the PPL is computed based on

$$BR(\theta \mid D) = \int 10^{HLOD} f(\theta) f(\alpha) f(\gamma)\, d\gamma\, d\alpha. \qquad \textit{Eq 6}$$

2.2 Quantitative Traits KELVIN handles QT data using the QT model from LIPED [44]. This model is parameterized in terms of three normal distributions, one for each of three possible trait genotypes (under the usual two-allele model). The advantages of this parameterization include the fact that it is a direct and simple extension of the dichotomous trait model, with penetrances replaced by means, and as a result the resulting QT LOD is also inherently (approximately) ascertainment corrected [12, 13]. Moreover the assumption of normality at the genotypic level is far weaker than the assumption of normality at the population level required by some other parameterizations (see [45] for evaluation of the QT-PPL under violations of normality in comparison to other methods).

The drawback to LIPED's implementation of this QT model – and presumably the reason the model has not been more widely applied – was that the user was required to input specific values for each of the genotypic means and variances, and of course in practice these are completely unknowable prior to discovery of the underlying QT gene. Kelvin, however, integrates out the QT trait parameters, bypassing this difficulty altogether; see [45] for details.

We also extended the model to allow for both DT and QT measurements within families under the QT threshold (QTT) model. This is useful when QT measures are available in relatives but not affected individuals, e.g., in studying autoimmune thyroid disease (AITD), for which thyroid autoantibodies (TAb) can be meaningfully measured only in unaffected relatives of AITD patients (TAb levels are affected by treatment), but the patients themselves were known only to have had elevated TAb at some point in the past. In order to model a common genetic factor for AITD and TAb simultaneously, the QTT model utilizes QT values in the relatives, but assumes only that affected individuals have TAb above some threshold. The threshold itself is treated as another nuisance parameter and integrated out of the BR with the other trait parameters. See [17] and also [46] for additional applications of this model.

We note too that even the distributional assumption of normality at the genotypic level is easily modified in these models, simply by swapping out the underlying function calls to a normal density (or cumulate distribution function) with calls to some other density function.

Currently the standard KELVIN QT model utilizes calls to an underlying *t*-distribution, for reasons of numerical stability; options for using a χ^2 distribution to handle traits with floor effects are available (with integration over the single degrees of freedom parameter), as are models utilizing left and/or right truncation of the *t*-distribution. However, the truncation models are not yet well evaluated in application to real or simulated data.

2.3 Liability classes, gene-gene interactions, and imprinting KELVIN also utilizes liability classes (LCs), in much the same way that LCs are implemented in other standard linkage packages, but again with the "twist" that Kelvin integrates over penetrances (or means, variances, in the case of a QT) separately for different LCs. This is therefore a very flexible technique for allowing for covariate-dependent penetrances. Individuals can be coded into separate LCs based on sex, age-of-onset, clinical subtype, etc., allowing the data to dictate whether members of different LCs have different underlying penetrances, up to and including the case in which members of one or more LCs do not have a genetic form of disease at all (i.e., have equal penetrances regardless of genotype at a given locus).

There is, however, a significant limitation on the number of LCs KELVIN can currently handle, due to the computational burden of the numerical integration process; see also below. At present, KELVIN can in general handle no more than four LCs, with some applications permitting only three. The upper bound also depends to some extent on available hardware. In our experience, however, this is less of a limitation in practice than it might at first appear. Model integration itself tends to mitigate the effects of collapsing data into fewer, more broadly defined classes, providing a level of robustness that may be lacking from, for instance, fixed-model LODs. Allowance for larger numbers of LCs is, nevertheless, an aim of ongoing computational improvements.

2.4 Epistasis and Imprinting One application of LCs that we have found to be very effective is for modeling a particular form of two-locus gene-gene interaction. Coding individuals into LCs based on genotype at an associated SNP can be a powerful and computationally feasible approach to discovering gene-gene interactions, both in family data and case-control data, and even on a genome-wide basis [32, 47] This approach is based on a straightforward underlying epistasis model corresponding to genetic (or causal) interactions (see [48, 49] for further details).

KELVIN currently allows for one additional type of covariate-dependent penetrance (or QT mean), viz., dependence on parent-of-origin. This option is implemented through a parameterization in terms of four, rather than three, penetrances (or means), allowing the penetrance for a heterozygous individual to depend on whether the putative risk allele was inherited from the mother or the father.

2.5 Linkage Disequilibrium (LD) KELVIN's underlying parameterization of the linkage likelihood also lends itself readily to incorporation of trait-marker LD. In essence, rather than assuming equal phase probabilities for unphased genotypes, the likelihood is written in terms of unknown phase probabilities, which are then integrated out as additional nuisance parameters. In practice, this is accomplished by including the standardized LD parameter D' [50] in the HLOD's underlying likelihood, for a statistic we called the LD-LOD [51], or more explicitly, LD-HLOD. By placing a point mass over D' = 0 (no LD), we can easily formulate the posterior probability of LD (PPLD) or the PPLD|L (conditioned on linkage) [52], based on

$$BR(\theta, D') = \int 10^{LD-HLOD} f(\theta) f(D') f(\alpha) f(\gamma)\, d\gamma\, d\alpha. \qquad \textit{Eq 7}$$

By restricting the prior probability of LD to a very small constant value outside the region of tight linkage, this statistic implicitly models trait-marker allelic association due to tight linkage only. We have verified that this approach works not only for pedigrees, but also for case-control or trio data [32], and have applied the method to genome-wide association data sets [19, 53]. In principle, the PPLD and PPLD|L are immediately applicable to sequence data as well, subject to the same caveats that apply to any association-based analysis of individual sequence variants (e.g., low power for rare variants).

By virtue of the underlying implementation, KELVIN's approach to LD analyses is uniquely flexible. Virtually all features available to the PPL are also available to the PPLD, e.g., the PPLD can readily be run for DT, QT, or QTT models, combined with LCs for purposes of gene-gene interaction modeling, etc.. The PPLD can also be seamlessly applied to mixtures of case-control, trio, affected sib pair, and extended pedigree data; and for QT data unrelated individuals can be analyzed on a "case-only" basis (no separate controls). Finally, in situations in which the PPL has been calculated based on previous data sets, it can be used as the “prior” information for linkage in the joint PPLD for linkage and association. This provides a rigorous method for utilizing linkage information to inform association analyses. See [32] for further details, and [18, 19] for illustrative applications.

One drawback to the PPLD as currently implemented is that there is no ready mechanism for covariate adjustment for population structure. Ancestrally homogeneous data subsets, as determined by standard methods, can be analyzed separately, and sequential updating can then be used to accumulate evidence across subsets if desired. While the PPLD appears to be no more sensitive to issues such as Hardy-Weinberg violation than other methods (see [32]), sensitivity to subtle population structure effects remains an ongoing area of investigation.

(3) KELVIN's Computational Underpinnings

3.1 Polynomial Likelihoods KELVIN is essentially a re-engineered version of Vittesse [54, 55], maintaining all of the essential efficiencies and functionality of Vittesse while adding some additional features of its own [56]. The first substantial distinction between the programs lies in KELVIN's use of polynomial representations of underlying likelihoods.

The motivation for this approach has to do with the computational demands of numerical integration over the trait parameter space. Our original implementation was easy to program but quite crude: we specified a fixed set of values for each parameter, leading to a fixed grid of values across all parameters; the LOD was then computed by brute force at each point in the grid and the results averaged across the grid. By "brute force," we mean that a linkage program (such as Vittesse [54] or Genehunter [57]) was run one time for each point in the grid. For a DT multipoint analysis this involved writing 33,000 input files and running the linkage program 33,000 times for each calculating position on the genome. But almost all of this file writing and computation is in principle unnecessary, because the underlying pedigree peeling and ancillary operations are identical across the 33,000 runs; all that changes are the numerical values of the trait parameters.

By contrast, KELVIN takes a single input file, peels through each pedigree once (per position), and stores the resulting likelihood as a polynomial in the trait parameters, now represented as algebraic variables rather than numbers (Figure 2) ([58, 59]; see also [60]). This polynomial can then be evaluated as many times as is necessary to traverse the parameter grid. Many terms will also recur repeatedly within and across polynomials. Upon first evaluation, KELVIN stores the values for these terms in hash tables for subsequent look-up, achieving additional savings in compute time. Since the peeling step is relatively computationally intensive, whereas the evaluation of a polynomial function can be done extremely fast, this approach results in a dramatically faster calculation of the PPL. Depending upon pedigree structures and particular marker configurations, the speed up can be on the order of 1,200 fold using this approach [58].

The entire polynomial can also be stored as an optimized, compiled dynamic library for reuse in further analyses. There are situations in which this results in substantial additional time savings. Finally, we can embed the compiled Elston-Stewart based polynomial for the trait data in Lander-Green [4] or MCMC calculations covering the marker data, for a very efficient approach to integrating over trait parameters when utilizing these other algorithms.

3.2 Fast and accurate numerical integration A second major innovation embedded in KELVIN's architecture has to do with the underlying framework for numerical integration. The brute-force approach of averaging over a predefined fixed grid of parameter values is not just inefficient, but also potentially inaccurate. E.g., for bounded parameters this method overweights edges (e.g., $\theta = 0$, $f_{dd} = 0$, etc.). But we were reluctant to turn to MCMC, both because it was difficult to guarantee accuracy in any given application, and because each future extension to the statistical models could necessitate substantial additional work on samplers. We wanted something guaranteed to be accurate, fast, and readily deployed for newly implemented models.

KELVIN therefore uses a modified version of DCUHRE [61], a form of adaptive or dynamic quadrature, combining deterministic selection of points within regions with dynamic decisions regarding progressive subdivision of regions, based on error estimates on the integral. The algorithm is theoretically guaranteed to be accurate up to 13-15 dimensions, and even the more complex models in Kelvin generally stay well within this limit. As adapted and implemented in KELVIN, this approach is both highly accurate and extremely efficient, being on average 31,000 times faster than our original approach [62] (Figure 3).

3.3 Additional software engineering features KELVIN takes advantage of commonly available distributed computing resource management systems in order to efficiently distribute computation across commodity computing hardware, utilizing other standard programming (e.g., multi-threading) and hardware (e.g., solid state drives) tools to gain further efficiencies (Figure 4). We are currently exploring the possibility of distributing calculations to the cloud.

KELVIN makes use of a multi-agent system (MAS) to decompose and distribute the computational workload across available resources while utilizing DCUHRE's dynamic integration capabilities. This is particularly handy for applications involving "mixing and matching," e.g., it allows for analysis of a single set of pedigrees that have been genotyped at different markers, or that have differences in inter-marker distances and marker allele frequencies. Such calculations would otherwise be impossible in conjunction with DCUHRE's dynamic integration algorithm. Use of a MAS also allow for calculations based on a mixture of algorithms, e.g., Lander-Green [4] for smaller pedigrees and Elston-Stewart [1] for larger ones, within a single data set. While at present the MAS is only implemented in-house, we are currently working on incorporating this functionality into the distributed version of KELVIN.

KELVIN has some other embedded computational efficiencies beyond those already present in Vittesse [54]. For instance, the code automatically identifies repetitions of identical data structures (for instance, all cases with a given marker genotype in a case-control data set), computing likelihoods only once for each such structure and multiplying an appropriate number of times to obtain the likelihood for the set of such structures. This can result in substantial savings particularly for case-control or trio data, where only a small number of distinct observations are possible within data sets.

3.4 Running KELVIN KELVIN uses common input file formats; details are in the documentation, which is downloadable from http://kelvin.mathmed.org. Prebuilt binaries are also available at http://kelvin.mathmed.org, for a variety of platforms including Windows/cygwin, Macintosh, and Linux environments. (Users are asked to create an account and provide an email address before downloading the software, to facilitate notification regarding bugs and program updates. Documentation can be downloaded without an account.) System requirements depend heavily on the data: case-control and trio LD analyses are feasible on almost any machine, as are linkage analysis of nuclear families to small extended pedigrees based on a small number of markers per position (for instance using microsatellites). Calculations on larger pedigrees, particularly multi-generation pedigrees with loops and/or substantial missing data in top generations, may require more memory than most desktop machines will have, as do multipoints using several markers at a time. (We assume that when using SNPs for linkage analysis, the user would select a subset of SNPs up front, culling the map for strong marker-to-marker LD.) We are currently supporting a limited set of KELVIN functions through our website (kelvin.mathmed.org), permitting users to run jobs using our large Linux cluster. The web version is under ongoing development, with a goal of eventually making MAS, Lander-Green [4] and MCMC options available to outside users. Also planned is a downloadable tool for estimating the required compute time and resources for particular data. In the mean time, users interested in knowing whether specific data sets would be amenable to KELVIN analysis are welcome to email us at kelvin@nationwidechildrens.org for assistance.

The user selects options using a simple syntax in the configuration file. (The web version provides a simple interface for setting up the file.) While most combinations of options are compatible with one another, KELVIN will notify the user if incompatible choices are made (for instance, "multipoint" and "LD," since multipoint trait-marker LD analysis is not currently supported). The website includes a KELVIN manual, which provides tutorial examples illustrating how to set up the various types of runs.

KELVIN produces 1-2 standard output files, depending on the analysis being run. These include the PPL (and/or PPLD) for each position, organized by chromosome, and the BRs at each position. The BRs serve as the input to a utility program that performs sequential updating and returns final PPLs (and/or PPLDs) (Figure 5). A separate option outputs the MOD score [40] (the LOD maximized over trait parameters) and the maximizing values of those parameters (maximum likelihood estimators, or MLEs) [10, 11, 13]; fixed-model LODs can also be calculated. The MLEs should be interpreted with some care, however, as DCUHRE is optimized for integration rather than maximization. In general, MLEs reported by Kelvin will be close approximations to the true MLEs, but in particular cases of interest this should be independently verified.

A separate graphing program, KELVIZ, is distributed as a freestanding application with the KELVIN code. KELVIZ allows the user to read in KELVIN output on a per-chromosome or genome-wide basis, with the ability to overlap output from multiple files for visual comparison of results across different data sets or subsets. Various options for annotating peaks and peak boundaries are available, and graphs are readily configurable for output under several available formats (.png, .eps, .pdf, .svg, .tif). Figure 6 illustrates some of these features.

Discussion

Two themes have characterized the development of KELVIN as a piece of software. The first is our attitude towards software engineering. Human geneticists have a tendency to prefer statistical software that can be written by statisticians, is extremely easy to use, and runs very quickly. But state-of-the-art code does not always fit this prescription: such code can be expensive to develop, requiring computational techniques beyond the competence of most statisticians; it can be difficult to run, for instance, requiring a certain level of familiarity with command line execution in Linux and access to hardware beyond a desktop machine; and it may run very slowly. But collection of clinical and molecular data in human genetics generally entails considerable investment of time and resources, and surely we want our statistical methods to be as robust and accurate as possible in extracting all available information from the data. Thus it seems strange to place a premium on simplicity and ease at the very final, data analytic, step of such studies. We have, therefore, repeatedly allowed our statistical models to outrun our computational capacity, relying on hardware upgrades and software engineering to render the required calculations feasible in real time. This is no different than what is done in the laboratory, where we embrace difficult molecular technologies that promise new types of data, working to bring down costs and logistical obstacles over time.

We have also placed an emphasis on a unified framework, adding new features within the same program. This requires long-term coordination on code development by a team of professional programmers. By contrast, a more common model is to assign development of a novel statistical approach to an individual trainee who programs the approach de novo. This results not just in inefficient programs, but also in a proliferation of programs each of which is quite limited in scope (e.g., programs that can handle imprinting only nuclear families but not extended pedigrees, or perform association analyses in case-control data or trios but not both, etc.). By contrast, virtually all KELVIN options are available in conjunction with one another and in seamless application to multiple data structures. We have also placed a premium on extensive, ongoing testing of the code in order to ensure continued accuracy.

The second theme is the philosophical framework in which KELVIN is embedded. Throughout development of the code there has been an emphasis on maintaining the scale of KELVIN's outcome measures. To a great extent we feel that we have consistently achieved this objective, so that the PPL and PPLD remain directly interpretable regardless of number of parameters in the model, regardless of the type of model, and without requiring reference to a null sampling distribution in order to "correct" the scale.

But scaling of evidence measures remains an ongoing concern, and since the original PPL proposal we have developed a certain agnosticism with regard to the particular scale of the PPL. The original argument [8] rested on the premise that what we really want to know is in fact the *probability* of linkage, and indeed, probabilities have the advantage of being on a familiar and readily interpreted scale. But if we rephrase the objective slightly, and say instead that what we are interested in is the *evidence* for linkage, this opens up the possibility that the probability scale would not necessarily be optimal. Thus at present we view the choice of the probability scale for the statistics reported by KELVIN to be somewhat arbitrary, albeit still convenient in terms of familiarity and interpretation. Possibly, future statistics in this "PPL" framework might no longer be probabilities at all, which is one reason we do not consider the PPL framework to be particularly Bayesian. In any case, KELVIN development has always progressed in parallel with a separate research program on the measurement of evidence, which is itself an ongoing area of active investigation [5, 6, 63, 64].

Acknowledgments

This work was funded in part by NIH grant MH086117. We thank Linda Brzustowicz for her great patience and perseverance as a perennial β tester, and for the many suggestions she has made that have helped turn KELVIN into a better and more useful program.

References

1. Elston, R.C. and J. Stewart, *A general model for the genetic analysis of pedigree data.* Hum Hered, 1971. **21**(6): p. 523-42.
2. Logue, M.W. and V.J. Vieland, *A new method for computing the multipoint posterior probability of linkage.* Hum Hered, 2004. **57**: p. 90-99.
3. George, A.W., et al., *Calculation of multipoint likelihoods using flanking marker data: a simulation study.* BMC Genet, 2005. **6 Suppl 1**: p. S44.
4. Lander, E.S. and P. Green, *Construction of multilocus genetic linkage maps in humans.* Proc Natl Acad Sci U S A, 1987. **84**(8): p. 2363-7.
5. Vieland, V.J., *Thermometers: something for statistical geneticists to think about.* Hum Hered, 2006. **61**(3): p. 144-56 PMID: 16770079.
6. Vieland, V.J., *Where's the Evidence?* Human Heredity, 2011. **71**(1): p. 59-66.
7. Vieland, V.J. and S.E. Hodge, *Review of Statistical Evidence: A Likelihood Paradigm.* Am J Hum Genet, 1998. **63**: p. 283-289.
8. Vieland, V.J., *Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage.* Am J Hum Genet, 1998. **63**(4): p. 947-54 PMID: 9758634.
9. Smith, C.A.B., *Some comments on the statistical methods used in linkage investigations.* Am J Hum Genet, 1959. **11**: p. 289-304.
10. Greenberg, D.A., *Inferring mode of inheritance by comparison of lod scores.* Am J Med Genet, 1989. **34**(4): p. 480-6.
11. Elston, R.C., *Man bites dog? The validity of maximizing lod scores to determine mode of inheritance.* Am J Med Genet, 1989. **34**(4): p. 487-8.
12. Ewens, W.J. and N.C. Shute, *A resolution of the ascertainment sampling problem. I. Theory.* Theor Popul Biol, 1986. **30**(3): p. 388-412.
13. Vieland, V.J. and S.E. Hodge, *The problem of ascertainment for linkage analysis.* Am J Hum Genet, 1996. **58**(5): p. 1072-84 PMID: 8651268.
14. Vieland, V.J. and S.E. Hodge, *Ascertainment bias in linkage analysis: Comments on Ginsburg et al.* Genet Epi, 2005. **28**: p. 283-285 PMID: 15578735.
15. Wang, K., *Ph.D. Thesis, Univ of iowa.* 1999.
16. Elston, R.C. and K. Lange, *The prior probability of autosomal linkage.* Ann Hum Genet, 1975. **38**(3): p. 341-50.
17. Vieland, V.J., et al., *A Multilocus Model of the Genetic Architecture of Autoimmune Thyroid Disorder, with Clinical Implications.* American journal of human genetics, 2008. **82**(6): p. 1349-1356 PMID: 18485327.
18. Wratten, N.S., et al., *Identification of a schizophrenia-associated functional noncoding variant in NOS1AP.* Am J Psychiatry, 2009. **166**(4): p. 434-41.
19. Vieland, V.J., et al., *Novel method for combined linkage and genome-wide association analysis finds evidence of distinct genetic architecture for two subtypes of autism.* J Neurodev Disord, 2011. **3**(2): p. 113-123.

20. Pagnamenta, A.T., et al., *Rare familial 16q21 microdeletions under a linkage peak implicate cadherin 8 (CDH8) in susceptibility to autism and learning disability.* J Med Genet, 2011. **48**(1): p. 48-54.
21. Smith, C.A.B., *Testing for heterogeneity of recombination fraction values in human genetics.* Ann Hum Genet, 1963. **27**: p. 175-182.
22. Hodge, S.E., et al., *The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity.* Am J Hum Genet, 1983. **35**(6): p. 1139-55.
23. Vieland, V.J. and M. Logue, *HLODs, trait models, and ascertainment: implications of admixture for parameter estimation and linkage detection.* Human Heredity., 2002. **53**(1): p. 23-35 PMID: 11901268.
24. Ott, J., *Linkage analysis and family classification under heterogeneity.* Ann Hum Genet, 1983. **47 (Pt 4)**: p. 311-20.
25. Vieland, V.J., *The replication requirement.* Nat Genet, 2001. **29**(3): p. 244-5 PMID: 11687787.
26. Hodge, S.E. and V.J. Vieland, *Expected monotonicity - A desirable property for evidence measures?* Hum Hered, 2010. **70**(3): p. 151-66.
27. Wang, K., J. Huang, and V.J. Vieland, *The consistency of the posterior probability of linkage.* Ann Hum Genet, 2000. **64**(Pt 6): p. 533-53.
28. Vieland, V.J., K. Wang, and J. Huang, *Power to detect linkage based on multiple sets of data in the presence of locus heterogeneity: comparative evaluation of model-based linkage methods for affected sib pair data.* Hum Hered, 2001. **51**(4): p. 199-208.
29. Huang, J. and V.J. Vieland, *Comparison of 'model-free' and 'model-based' linkage statistics in the presence of locus heterogeneity: single data set and multiple data set applications.* Human Heredity., 2001. **51**(4): p. 217-25 PMID: 11287743.
30. Bartlett, C.W., R. Goedken, and V.J. Vieland, *Effects of updating linkage evidence across subsets of data: reanalysis of the autism genetic resource exchange data set.* Am J Hum Genet, 2005. **76**(4): p. 688-95 PMID: 15729670.
31. Govil, M. and V.J. Vieland, *Practical considerations for dividing data into subsets prior to PPL analysis.* Hum Hered, 2008. **66**: p. 223-237 PMID: 18612207.
32. Huang, Y. and V.J. Vieland, *Association statistics under the PPL framework.* Genet Epidemiol, 2010. **34**(8): p. 835-45.
33. Wang, K., V. Vieland, and J. Huang, *A Bayesian approach to replication of linkage findings.* Genet Epidemiol, 1999. **17**(Suppl 1): p. S749-54.
34. Greenberg, D.A., *Summary of analyses of problem 2 simulated data for GAW 11.* Genet Epidemiol, 1999. **17 Suppl 1**: p. S449-59.
35. Logue, M.W. and V.J. Vieland, *The incorporation of prior genomic information does not necessarily improve the performance of Bayesian linkage methods: an example involving sex-specific recombination and the two-point PPL.* Hum Hered, 2005. **60**: p. 196-205.
36. Greenberg, D.A. and S.E. Hodge, *Linkage analysis under "random" and "genetic" reduced penetrance.* Genetic Epidemiology., 1989. **6**(1): p. 259-64.
37. Vieland, V.J., S.E. Hodge, and D.A. Greenberg, *Adequacy of single-locus approximations for linkage analyses of oligogenic traits.* Genetic Epidemiology., 1992. **9**(1): p. 45-59 PMID: 1634106.

38. Vieland, V.J., D.A. Greenberg, and S.E. Hodge, *Adequacy of single-locus approximations for linkage analyses of oligogenic traits: extension to multigenerational pedigree structures.* Human Heredity., 1993. **43**(6): p. 329-36 PMID: 8288263.
39. Slager, S.L. and V.J. Vieland, *Investigating the numerical effects of ascertainment bias in linkage analysis: development of methods and preliminary results.* Genet Epidemiol, 1997. **14**(6): p. 1119-24 PMID: 9433634.
40. Clerget-Darpoux, F., C. Bonaiti-Pellie, and J. Hochez, *Effects of misspecifying genetic parameters in lod score analysis.* Biometrics, 1986. **42**(2): p. 393-9.
41. Logue, M., et al., *Bayesian analysis of a previously published genome screen for panic disorder reveals new and compelling evidence for linkage to chromsome 7.* Am J Med Genet B (Neuropsychiat Genet), 2003. **121B**: p. 95-99.
42. Logue, M.W., *Complications of an unknown genetic model in the presence of heterogeneity for linkage analysis [Ph. D. Dissertation]*, 2001, THE UNIVERSITY OF IOWA: Iowa City. p. 124.
43. Logue, M., et al., *A Posterior Probability of Linkage (PPL) Based Re-Analysis of Schizophrenia Data Yields Evidence of Linkage to Chromosomes 1 and 17.* Human Heredity, 2006. **62**(1): p. 47-54.
44. Ott, J., *A computer program for linkage analysis of general human pedigrees.* Am J Hum Genet, 1976. **28**(5): p. 528-9.
45. Bartlett, C.W. and V.J. Vieland, *Accumulating quantitative trait linkage evidence across multiple datasets using the posterior probability of linkage.* Genet Epi, 2006. **31**: p. 91-102 PMID: 17123305.
46. Bartlett, C.W. and V.J. Vieland, *Two novel quantitative trait linkage analysis statistics based on the posterior probability of linkage: application to the COGA families.* BMC Genet, 2005. **6 Suppl 1**: p. S121 PMID: 16451579.
47. Huang, Y., et al., *Exploiting gene x gene interaction in linkage analysis.* BMC Proceedings, 2007. **1(Suppl 1)**: p. S64(1-5) PMID: 18466565.
48. Vieland, V.J. and J. Huang, *Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pair data.* Am J Hum Genet, 2003. **73**(2): p. 223-32. PMID: 12830401.
49. Vieland, V.J. and J. Huang, *Reply to Cordell and Farrall.* Am J Hum Genet, 2003. **73**: p. 1471-1473.
50. Lewontin, R.C., *The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.* Genetics, 1964. **49**(1): p. 49-67.
51. Slager, S.L., J. Huang, and V.J. Vieland, *Power comparisons between the TDT and two likelihood-based methods.* Genet Epidemiol, 2001. **20**(2): p. 192-209.
52. Yang, X., et al., *The posterior probability of linkage allowing for linkage disequilibrium and a new estimate of disequilibrium between a trait and a marker.* Hum Hered, 2005. **59**: p. 210-9 PMID: 16015031.
53. Flax, J.F., et al., *Combined linkage and linkage disequilibrium analysis of a motor speech phenotype within families ascertained for autism risk loci.* J Neurodev Disord, 2010. **2**(4): p. 210-223.
54. O'Connell, J. and D. Weeks, *The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance.* Nature Genetics, 1995. **11**: p. 402-408.

55. O'Connell, J., *Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm.* Human Heredity, 2001. **51**(4): p. 226-240.
56. Huang, Y., et al. *KELVIN: a 2nd generation distributed multiprocessor linkage and linkage disequilibrium analysis program*. in *Am Soc Hum Genet*. 2006. New Orleans, LA.
57. Kruglyak, L., et al., *Parametric and nonparametric linkage analysis: a unified multipoint approach.* Am J Hum Genet, 1996. **58**(6): p. 1347-63.
58. Wang, H., et al., *Fast computation of human genetic linkage.* Proceedings of IEEE 7th Symposium on Bioinformatics and Bioengineering, 2007. **BIBE 2007**: p. 857-863.
59. Wang, H., et al., *Rapid computation of large numbers of LOD scores in linkage analysis through polynomial expression of genetic likelihodos.* Proceedings of IEEE Workshop on High-Throughput Data Analysis for Proteomics and Genomics, 2007. **IEEE 2007**: p. 197-204.
60. Kramer, R.W., D.E. Weeks, and D.M. Chiarulli, *An incremental algorithm for efficient multipoint linkage analysis.* Hum Hered, 1995. **45**(6): p. 323-36.
61. Bernsten, J., T. Espelid, and A. Genz, *An adaptive multidimensional integration routine for a vector of integrals.* ACM Trans Math Softw, 1991. **17**: p. 452-456.
62. Seok, S.-C., M. Evans, and V.J. Vieland, *Fast and accurate calculation of a computationally intensive statistic for mapping disease genes.* J Comput Biol, 2009(16(5)): p. 659-676.
63. Hodge, S.E. and V.J. Vieland, *Expected Monotonicity - A Desirable Property for Evidence Measures?* Hum Hered, 2010(70): p. 151-166.
64. Vieland, V.J. and Hodge, S.E., *Measurement of Evidence and Evidence of Measurement (Invited Commentary).* Stat Applications Genetics and Molec Biology, (In Press).

Figure 1 The first application of the PPL: The original PPL sequentially updated across four highly heterogeneous simulated datasets. The x-axis shows all six simulated chromosomes; the y-axis shows the PPL, on the probability scale. The locations of all four underlying loci are clearly and accurately revealed.

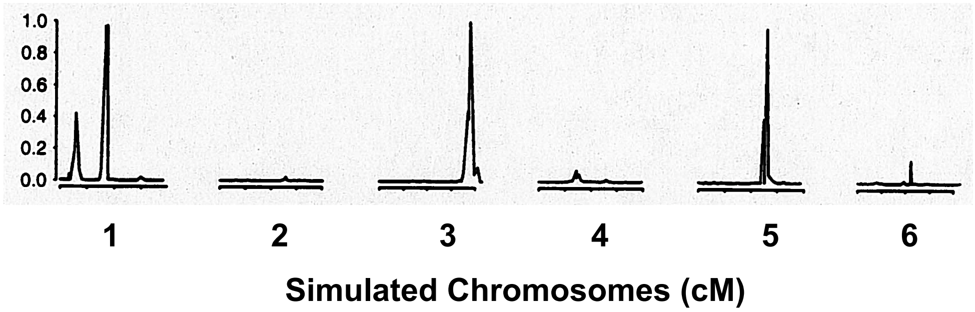
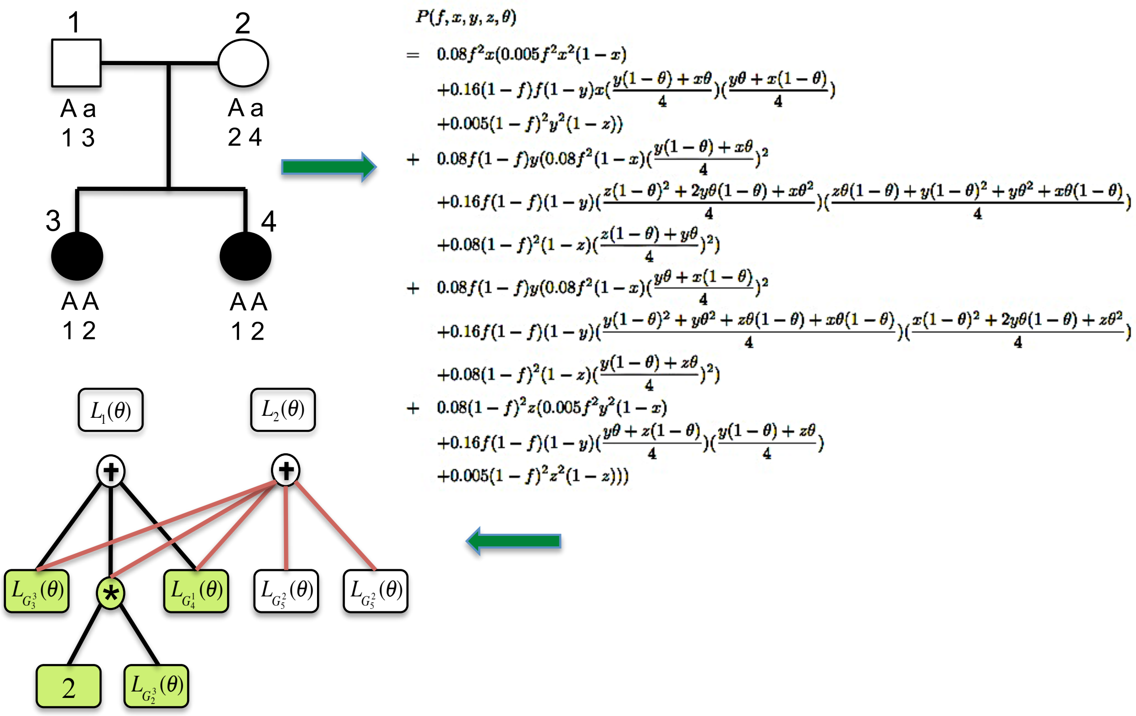


Figure 2 Polynomial representation of pedigree likelihoods: The likelihood for a pedigree at a given genomic position is stored internally as a polynomial in the parameters of the model for rapid evaluation over large numbers of parameter values, exploiting redundancies in evaluated polynomial subterms using hash tables for further computational savings.



$$
\begin{aligned}
& P(f,x,y,z,\theta) \\
= \ & 0.08f^2x(0.005f^2x^2(1-x) \\
& +0.16(1-f)f(1-y)x(\frac{y(1-\theta)+x\theta}{4})(\frac{y\theta+x(1-\theta)}{4}) \\
& +0.005(1-f)^2y^2(1-z)) \\
+ \ & 0.08f(1-f)y(0.08f^2(1-x)(\frac{y(1-\theta)+x\theta}{4})^2 \\
& +0.16f(1-f)(1-y)(\frac{z(1-\theta)^2+2y\theta(1-\theta)+x\theta^2}{4})(\frac{z\theta(1-\theta)+y(1-\theta)^2+y\theta^2+x\theta(1-\theta)}{4}) \\
& +0.08(1-f)^2(1-z)(\frac{z(1-\theta)+y\theta}{4})^2) \\
+ \ & 0.08f(1-f)y(0.08f^2(1-x)(\frac{y\theta+x(1-\theta)}{4})^2 \\
& +0.16f(1-f)(1-y)(\frac{y(1-\theta)^2+y\theta^2+z\theta(1-\theta)+x\theta(1-\theta)}{4})(\frac{x(1-\theta)^2+2y\theta(1-\theta)+z\theta^2}{4}) \\
& +0.08(1-f)^2(1-z)(\frac{y(1-\theta)+z\theta}{4})^2) \\
+ \ & 0.08(1-f)^2z(0.005f^2y^2(1-x) \\
& +0.16f(1-f)(1-y)(\frac{y\theta+z(1-\theta)}{4})(\frac{y(1-\theta)+z\theta}{4}) \\
& +0.005(1-f)^2z^2(1-z)))
\end{aligned}
$$

Figure 3 Fast and accurate non-stochastic numerical integration: Schematic representation of (a) the original parameter grid for numerical integration over the trait space (points are shown for θ and D' only for purposes of illustration), (b) the grid as generated by DCUHRE, and (c) DCUHRE integrals compared to highly accurate (extremely dense grid) numerical integration, across a large range of BR values.

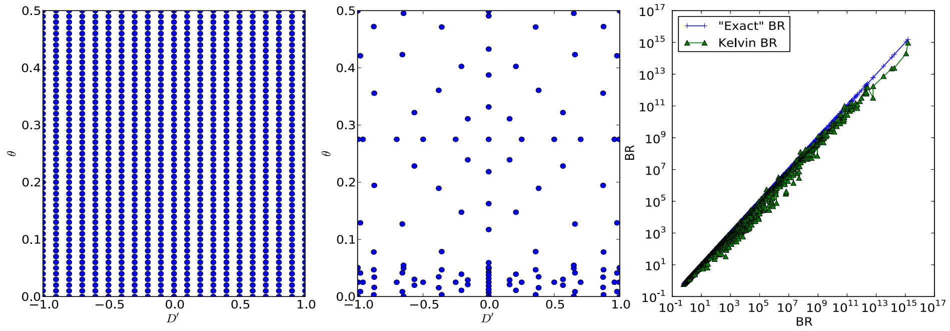


Figure 4 KELVIN's underlying software engineering: KELVIN utilizes commodity software for distributing calculations across a Linux cluster while exploiting hardware and software capabilities for further computational economies, including OpenMP for multithread processing, solid state drives (SSDs) for adjuvant memory, storage of polynomials as compiled code for rapid and repeated reuse, and a fast and accurate dynamic numerical integration algorithm (DCUHRE).

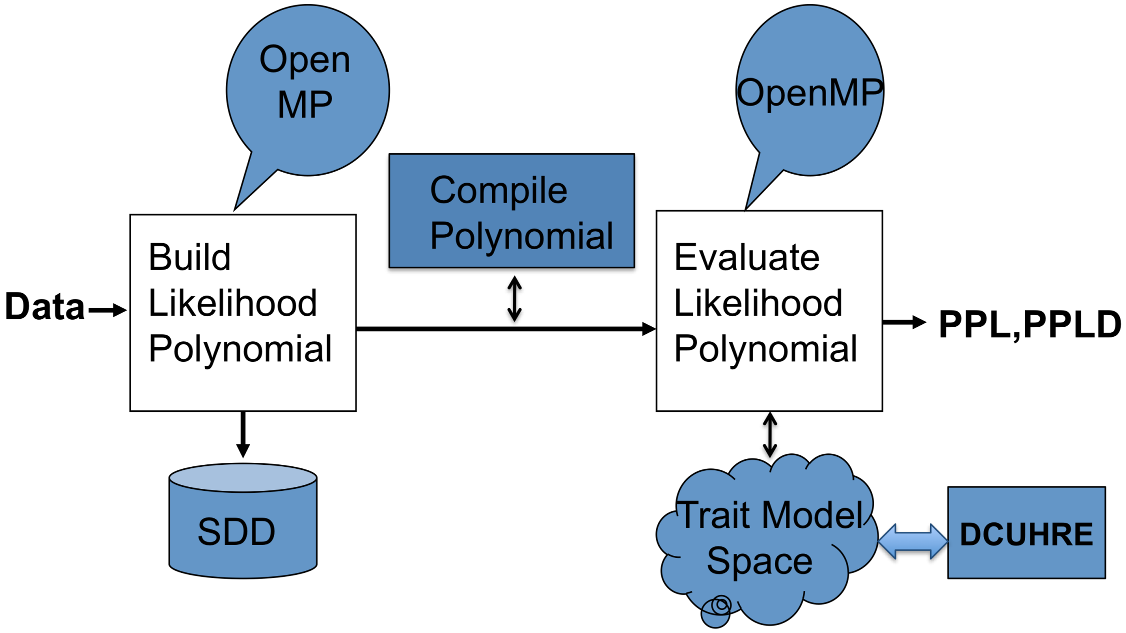

Open MP
OpenMP
Compile Polynomial
Data
Build Likelihood Polynomial
Evaluate Likelihood Polynomial
PPL,PPLD
SDD
Trait Model Space
DCUHRE

Figure 5 Sequential updating: KELVIN uses sequential updating to accumulate evidence across multiple sets of data or data subsets, via multiplication of the subset-specific BRs at each genomic location.

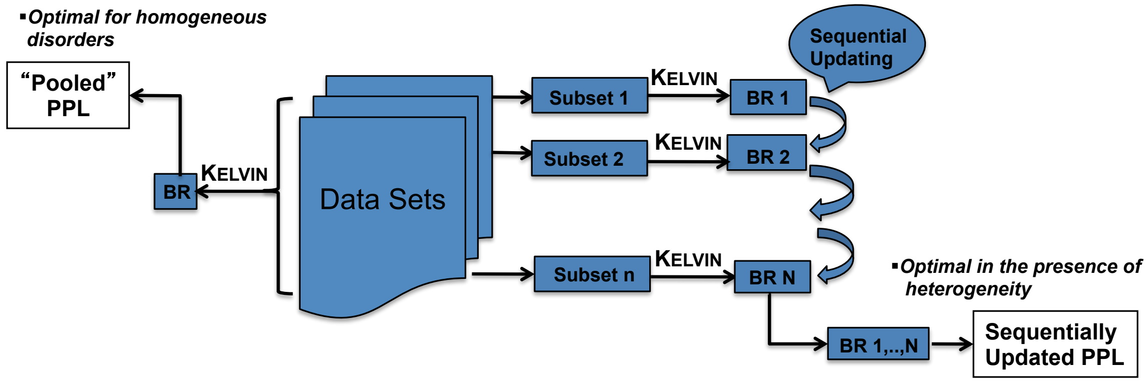

▪Optimal for homogeneous disorders
"Pooled" PPL
BR
KELVIN
Data Sets
Subset 1
Subset 2
Subset n
KELVIN
KELVIN
KELVIN
BR 1
BR 2
BR N
Sequential Updating
▪Optimal in the presence of heterogeneity
BR 1,..,N
Sequentially Updated PPL

Figure 6 Using Kelviz to explore and graph Kelvin output: Kelviz is specifically designed for plotting Kelvin output. It has a large range of options for custom graph construction. Shown here are just a few of these: (a) a genome-wide plot of the PPL; (b) a view comparing genome-wide PPLs across two different clinical subgroups; (c) a zoom in to a particular chromosome, displaying both clinical groups' results overlaid, with an option to annotate features of the graph using information in the input file or other edited content; (d) a zoom in to a portion of a chromosome, overlaying two PPL plots with PPLD results from a genome-wide association data set. Data are taken from [19].