

kalis: A Modern Implementation of the Li & Stephens Model for Local Ancestry Inference in R

Louis J.M. Aslett^{1*} and Ryan R. Christ²

^{1*}Department of Mathematical Sciences, Durham University, Stockton Road, Durham, DH1 3LE, County Durham, UK.

²Department of Genetics, Yale School of Medicine, 333 Cedar Street, New Haven, 06520, CT, USA.

*Corresponding author(s). E-mail(s): louis.aslett@durham.ac.uk;
Contributing authors: ryan.christ@yale.edu;

Abstract

Background Approximating the recent phylogeny of N phased haplotypes at a set of variants along the genome is a core problem in modern population genomics and central to performing genome-wide screens for association, selection, introgression, and other signals. The Li & Stephens (LS) model provides a simple yet powerful hidden Markov model for inferring the recent ancestry at a given variant, represented as an $N \times N$ distance matrix based on posterior decodings. However, existing posterior decoding implementations for the LS model cannot scale to modern datasets with tens or hundreds of thousands of genomes.

Results We provide a high-performance engine to compute the LS model, enabling users to rapidly develop a range of variant-specific ancestral inference pipelines on top, exposed via an easy to use package, **kalis**, in the statistical programming language R. **kalis** exploits both multi-core parallelism and modern CPU vector instruction sets to enable scaling to problem sizes that would previously have been prohibitively slow to work with.

Conclusions The resulting distance matrices accessible via **kalis** enable local ancestry, selection, and association studies in modern large scale genomic datasets.

Keywords: Li & Stephens Model, R package, probabilistic haplotype model, Hidden Markov Model, genomics, high performance computation

047 Background

048
049 The hidden Markov model (HMM) of haplotype diversity proposed by Li & Stephens
050 [1] (hereinafter, the LS model) has become the basis for several probabilistic phasing,
051 ancestry inference, and demographic inference methods in modern genomics [2, 3].
052

053 The LS model provides the basis for the genome-wide ancestry inference software
054 ChromoPainter, which summarizes the ancestry of N haplotypes with an $N \times N$
055 similarity matrix [4]. This matrix is obtained by running N independent HMMs in
056 which each haplotype is modelled as a mosaic of all of the other haplotypes in the
057 sample. This ‘*all-vs-all*’ copying approach is motivated by the product of approximate
058 conditionals (PAC) likelihood originally proposed by [1] and allows ChromoPainter to
059 render a chromosome-wide estimate of the recent ancestry of the N haplotypes with
060 high resolution.
061

062 Beyond chromosome-wide summaries, the LS model can also be used for variant-
063 specific ancestry inference. The contemporary importance of this is highlighted by
064 the recent advances achieved in the RELATE software suite [2], which uses the LS
065 model internally to initialise variant-specific ancestral trees for downstream popula-
066 tion genetic analyses ranging from demography to selection inference. **kalis** focuses on
067 providing a high-performance engine to compute exclusively the LS model, enabling
068 users to rapidly develop a range of future variant-specific ancestral inference pipelines
069 on top, in the easy to use statistical programming language R [5].
070

071 At the same time, it has been recognised for over a decade [6] that the serial exe-
072 cution speed of CPUs will increase modestly, with additional performance primarily
073 coming from concurrency via multi-core architectures or the growing width of spe-
074 cialised single instruction, multiple data (SIMD) instruction sets. Whilst multi-core
075 architectures are now somewhat routinely exploited via forked processes or threading,
076 SIMD instructions remain an often overlooked source of performance gains, possibly
077 because they are harder to program. There are a cornucopia of SIMD instruction sets:
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092

on the Intel platform the genesis was in the 64-bit wide MMX instruction set [7] which allows simultaneous operation on two 32-bit, four 16-bit or eight 8-bit integers. The most recent incarnation on Intel CPUs is a suite of AVX-512 instruction sets [8], now capable of operating on 512-bits of various data types simultaneously (eg eight 64-bit floating point, or sixteen 32-bit integer values). Other CPU designs have similar SIMD technologies, such as NEON on ARM CPU [9] designs (including the Apple M1 and M2 processors, as well as Amazon Web Services Graviton range). Additionally all modern CPUs are superscalar architectures supporting instruction level parallelism, an advance that has been in the consumer Intel platform since the Pentium [10]. Judicious programming can make it easier for compilers and the deep reorder buffers of modern pipelined CPUs to exploit this more hidden form of parallelism.

In this work we provide a reformulation of the LS model and an optimised memory representation for haplotypes, which together enable us to leverage *both* multi-core and SIMD vector instruction parallelism to obtain local genetic distance matrices for problem sizes that previously appeared out of reach. This high performance implementation is programmed in C [11], with an easy to use interface provided in R [5]. We provide low-level targets of AVX2, AVX-512 and NEON instruction sets (covering the vast majority of CPUs in use today), and the whole package has an extensive suite of 162,835 unit tests.

In the Implementation section below, we start with a description of the LS model and our reformulation which makes it amenable to these high-performance CPU technologies. We also describe the technical details of the underlying low-level implementation for the interested reader. We then demonstrate the performance that can be achieved with **kalis**, including examples with 100,000 haplotypes capable of running on a single machine. To the best of our knowledge, this is the first example of running the LS model at the scale of hundreds-of-thousands of haplotypes. We also present a

real data example using **kalis** to examine the ancestry at the *LCT* gene. In the following Discussion section, we describe the user friendly R interface which enables easy use of the high performance implementation without any knowledge of the underlying CPU technologies.

Implementation

The LS model

To formalize our objective, let h be an $L \times N$ matrix of 0s and 1s encoding N phased haplotypes at L sites. Let $h_i^\ell \in \{0, 1\}$ denote the (ℓ, i) th element of h . For brevity, let h_i denote the i th haplotype (the i th column of h) and h_{-i} denote all of the haplotypes excluding the i th haplotype. The LS model proposes an HMM for $h_i|h_{-i}$ in which the hidden state at variant ℓ , $X_i^\ell \in \{1, \dots, N\} \setminus i$, is an index indicating the haplotype in h_{-i} that h_i is most closely related to (or “copies from”) at variant ℓ . We present here their proposed emission and transition kernels (see Equation A1 and Equation A2 in [1]) with a simplified parametrisation that is similar, but not identical, to that used by ChromoPainter.

While the original LS model assumes that each haplotype has an equal *a priori* probability of copying from any other, following ChromoPainter, we define a left stochastic matrix of prior copying probabilities $\Pi \in \mathbb{R}^{N \times N}$ where Π_{ji} is the prior probability that haplotype j is copied by i and, by convention, $\Pi_{ii} = 0$. Here and whenever possible in **kalis**, all matrices are column-oriented such that the i th column pertains to an independent HMM where h_i is treated as the observation. There is some probability of a mis-copy at variant ℓ , μ^ℓ , which under the LS model is set proportional to

the mutation rate at ℓ . This leads to an emission kernel of the form

$$\theta_{ji}^\ell := \mathbb{P}(h_i^\ell | X_i^\ell = j) = \begin{cases} 1 - \mu^\ell & \text{if } h_i^\ell = h_j^\ell \\ \mu^\ell & \text{if } h_i^\ell \neq h_j^\ell \end{cases}. \quad (1)$$

The transition kernel between hidden states is based on the recombination rate between sites. Let m^l be the genetic distance between variant l and variant $l + 1$ in Morgans (the expected number of recombination events per meiosis). Define $N_e = 4\tilde{N}_e/N$ where \tilde{N}_e is the effective diploid population size (ie half of the haploid effective population size). Then, under the LS model the transition kernel is

$$P(X_i^\ell = k | X_i^{\ell-1} = j) = \Pi_{ki} \rho^\ell + \mathbf{1}\{k = j\} (1 - \rho^\ell), \quad (2)$$

where $\rho^\ell = 1 - \exp(-N_e m^\ell)$ and $\mathbf{1}\{\cdot\}$ is the indicator function. [1] observe that in practice the estimation of recombination rates is improved when the scaled recombination rate is raised to a power, so we adopt this approach and introduce an exponent γ . For $\gamma > 1$ the recombination map becomes more heavily peaked, whereas $\gamma < 1$ tempers the recombination map to make it more flat and smooth. Hence, in **kalis**, we set

$$\rho^\ell := 1 - \exp\left(-N_e (m^\ell)^\gamma\right), \quad (3)$$

calculated using `expm1()` to help avoid underflow.

In keeping with the nomenclature introduced by [4], we refer to h_i as the “recipient haplotype” and the remaining haplotypes, h_{-i} , as the “donor haplotypes”, in the context of the HMM where h_i is treated as the emitted observation vector. This reflects the fact that each recipient haplotype h_i is modelled as an imperfectly copied mosaic of the other observed haplotypes under the LS model. Hence, the posterior marginal probability at variant ℓ , $p_{ji}^\ell := \mathbb{P}(X_i^\ell = j | h)$, is the probability that donor j is copied

by recipient i at variant ℓ given the haplotypes h . Under the above definitions of the prior copying probabilities Π , the emission kernel (1), and the transition kernel (2), the full $N \times N$ matrix of copying probabilities at ℓ , p^ℓ , can be obtained by running the standard forward and backward recursions [12] for each column (ie for each independent HMM).

From these posterior probabilities, we calculate a local $N \times N$ distance matrix, d^ℓ . Firstly, notice that theoretically $p_{ij}^\ell > 0$, but it can be that $p_{ij}^\ell < \varepsilon$, where ε is the double precision machine epsilon ($\approx 2.22 \times 10^{-16}$, [11], pp.26). Effectively this means d_{ij}^ℓ is too large to reliably work with precisely, and so for the purposes of distance calculations we treat ε as the smallest observable posterior probability, yielding

$$d_{ji}^\ell = -\frac{\log(p_{ji}^\ell \vee \varepsilon) + \log(p_{ij}^\ell \vee \varepsilon)}{2} \quad \forall j \neq i \quad (4)$$

where \vee is the maximum binary operator. By convention $d_{ii} = 0$ for all i .

We proceed in the next Section to reformulate the forward and backward recursions so that we can more fully exploit modern high-performance CPU instruction sets, while preserving numerical precision.

Modification of the forward-backward algorithm

The N independent HMMs of the LS model have forward and backward probabilities, respectively:

$$\tilde{\alpha}_{ji}^\ell = \mathbb{P}(X_i^\ell = j, h_i^{1:\ell}), \quad \tilde{\beta}_{ji}^\ell = \mathbb{P}(h_i^{\ell+1:L} | X_i^\ell = j), \quad i \in \{1, \dots, N\},$$

where $h_i^{1:\ell}$ denotes haplotype i from variant 1 to ℓ inclusive.

Define,

$$F_i^\ell := \sum_{j=1}^N \tilde{\alpha}_{ji}^\ell \quad F_i^0 := 1 \quad (5)$$

$$G_i^\ell := \sum_{j=1}^N \tilde{\beta}_{ji}^{\ell+1} \theta_{ji}^{\ell+1} \pi_{ji} \quad G_i^L := 1 \quad (6)$$

Then the forward and backward recursions for the LS model can be written in vector notation (subscript \cdot denoting a vectorised index),

$$\tilde{\alpha}_{\cdot i}^\ell \leftarrow \theta_{\cdot i}^\ell \left((1 - \rho^{\ell-1}) \tilde{\alpha}_{\cdot i}^{\ell-1} + \rho^{\ell-1} F_i^{\ell-1} \pi_{\cdot i} \right) \quad \text{for } \ell \in \{2, \dots, L\}, \quad (7)$$

$$\tilde{\beta}_{\cdot i}^\ell \leftarrow (1 - \rho^\ell) \tilde{\beta}_{\cdot i}^{\ell+1} \theta_{\cdot i}^{\ell+1} + \rho^\ell G_i^\ell \quad \text{for } \ell \in \{1, \dots, L-1\}. \quad (8)$$

with recursions initialised with $\alpha_{\cdot i}^1 \leftarrow \theta_{\cdot i}^1 \pi_{\cdot i}$ and $\beta_{\cdot i}^L \leftarrow 1$. Note that Equation (7) corresponds to Equation A5 in [1].

To partially mitigate the risk of underflow, the forward recursion can be rearranged in terms of $\alpha_{\cdot i}^\ell := \frac{\tilde{\alpha}_{\cdot i}^\ell}{F_i^{\ell-1}}$, and the backward recursion in terms of $\beta_{\cdot i}^\ell := \frac{\tilde{\beta}_{\cdot i}^\ell}{G_i^\ell}$ (see the Appendix (Additional file 1) for details). Thus, in full for $\ell \in \{1, \dots, L\}$ we compute,

$$\alpha_{\cdot i}^1 \leftarrow \theta_{\cdot i}^1 \pi_{\cdot i} \quad \text{for } \ell = 1 \quad (9)$$

$$\alpha_{\cdot i}^\ell \leftarrow \theta_{\cdot i}^\ell \left((1 - \rho^{\ell-1}) \frac{\alpha_{\cdot i}^{\ell-1}}{\sum_j \alpha_{ji}^{\ell-1}} + \rho^{\ell-1} \pi_{\cdot i} \right) \quad \text{for } \ell > 1 \quad (10)$$

and

$$\beta_{\cdot i}^L \leftarrow 1 \quad \text{for } \ell = L \quad (11)$$

$$\beta_{\cdot i}^\ell \leftarrow (1 - \rho^\ell) \frac{\beta_{\cdot i}^{\ell+1} \theta_{\cdot i}^{\ell+1}}{\sum_j \beta_{ji}^{\ell+1} \theta_{ji}^{\ell+1} \pi_{ji}} + \rho^\ell \quad \text{for } \ell < L \quad (12)$$

Given α_i^ℓ and β_i^ℓ , the vector of posterior probabilities for recipient i , p_i^ℓ , can be calculated directly by normalising,

$$p_i^\ell = \frac{\alpha_i^\ell \odot \beta_i^\ell}{\sum_j \alpha_{ji}^\ell \odot \beta_{ji}^\ell} \quad (13)$$

where \odot denotes the Hadamard product. In the event that $\sum_j \alpha_{ji}^\ell \odot \beta_{ji}^\ell = 0$, the distance between the recipient haplotype i and all of the donor haplotypes is beyond numerical precision, so as per the earlier discussion we define $p_{ji}^\ell = \varepsilon \forall j \neq i$.

Finally, the local distances follow by taking the negative log and symmetrising. Note that if the distances are standardised for one of these columns, to account for the fact that the standard deviation will be 0, we set all of the standardised distances to 0. Please see the Appendix (Additional file 1) for a discussion on parameter values and exactly how **kalis** performs certain computations to maintain the numerical stability of the algorithm.

Core Implementation Details

The R interface described hereinbefore is a thin wrapper layer around a high-performance implementation of the core algorithm which is written in standards compliant C18 [11]. Most data structures are represented with native R types enabling user inspection and manipulation, except for the haplotype sequences themselves.

Computationally, the innermost forward and backward recursions are implemented using compiler intrinsics to exploit a variety of modern CPU instruction sets, including Streaming SIMD Extensions (SSE2 and SSE4.1), Advanced Vector Extensions (AVX, AVX2, AVX-512 and FMA) and Bit Manipulation Instructions (BMI2) on Intel platforms; as well as NEON on ARM platforms. AVX2 is supported in Intel CPUs of the Haswell generation (released Q2 of 2013) or later, AVX-512 tends to be available only in recent Intel server and workstation grade CPUs, and NEON is available for ARM

Cortex-A and Cortex-R series CPUs, as well as Apple M1/M2 and Amazon Web Services Graviton processors. Although this covers most CPUs likely to be in use today, we none-the-less provide reference implementations in pure standards compliant C which will operate on any CPU architecture with a C18 compliant compiler. During package compilation, the correct code-paths are compiled based on detection of the presence or absence of the required instruction sets, or at the direction of the user via compiler flags. See the Appendix (Additional file 1) for more details, and for guidance on how to directly check your CPU for SIMD support.

It may be worth noting at this juncture that it was an explicit design choice to target CPUs and not GPU or tensor cards initially. This is because most University high performance computing clusters have plentiful CPU resources, often with untapped power in advanced SIMD instructions sets. We believe that the problem size that can be realistically tackled in many genetics studies can be massively increased *without* needing to resort to add-on cards, though to scale beyond even this we may explore heterogeneous computing architectures in future **kalis** research.

In this section, we now describe the inner workings and design principles of the package, first covering in detail the data structures (both user facing and internal), followed by the computational implementation.

Data structures

There are three user accessible data structures utilised in the package and a low level binary haplotype representation which is not directly user accessible. The two principle data structures of interest to users are forward and backward table objects, represented as native R lists with respective S3 class names **kalisForwardTable** and **kalisBackwardTable** (detailed in Table 2 and discussed later), which are created with package functions **MakeForwardTable()** and **MakeBackwardTable()** respectively. The third user accessible data structure holds the LS model parameters, represented as

415 a native R environment with S3 class name `kalisParameters`, which can be created
416 with the package function `Parameters()`.
417

418 419 **Haplotype data** 420

421 The haplotypes are stored in an optimised binary representation which is only natively
422 accessible from within C. Note that here “optimised” is not a reference to space-
423 optimisation: it would be possible to represent the haplotypes in an even more
424 compressed manner, but we aim for streaming compute speed optimisation instead.
425
426

427 The haplotypes are loaded from disk and transformed to an in memory cache in this
428 representation via `CacheHaplotypes()`, but this function does not return any handle
429 to the loaded data. Thus the package provides the accessor function `QueryCache()`,
430 which copies genome segments from the binary representation into native R integer
431 vectors for user inspection.
432
433

434 When `CacheHaplotypes()` loads haplotypes into the cache, they are interleaved
435 into a flat memory space which is organised as variant-major. That is, variant 1 of each
436 haplotype is loaded, converted to a binary 0/1 and then 32 consecutive haplotypes are
437 packed into an unsigned integer. Moreover, the initial flat memory allocation is aligned
438 on a 32-byte boundary to satisfy memory alignment requirements for some CPU vector
439 instructions¹, and after all haplotypes at a given variant are packed into consecutive
440 unsigned integers the pointer is wound forward to the next 32-byte boundary to ensure
441 the next variant starts on an SSE/AVX vector compatible memory boundary. This is
442 depicted in Figure 1.
443
444

445 Firstly, note that this orientation is natural, since the forward and backward recur-
446 sions operate variant by variant, meaning variant-major storage ensures sequential
447 memory locations are fetched during a recursion. Indeed, with the cache line size of
448 64-bytes (starting Intel Pentium IV), we essentially trigger the loading of $64 \times 8 = 512$
449

450
451
452
453
454
455
456
457
458
459 ¹Certain modern CPUs do not require specific alignment to be able to load memory to SSE/AVX registers,
460 but for maximum compatibility we honor the alignment anyway.

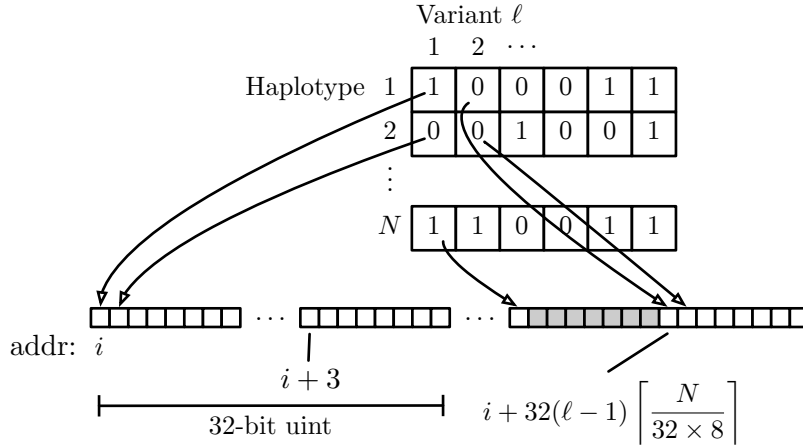


Fig. 1 Efficient binary representation of interleaved haplotypes in memory, with 32-byte boundary alignment for each variant start for SSE/AVX instructions (here $i \bmod 32 = 0$). The grey boxes indicate essentially ‘wasted’ bits which are ignored to ensure alignment for the start of the next variant.

neighbouring variants upon accessing the first variant in a recursion. This effect is even more pronounced on Apple M1/M2 whose cache line size is 128-bytes, resulting in 1024 variants being pre-fetched upon access to the first variant in a recursion.

Secondly, a possible drawback is that we must extract the individual bit into a double floating point representation in order to compute with it in the recursion. However, efficient CPU instructions can help here too: take for example the following strategy **kalis** uses on an AVX2 capable CPU. Using the PDEP instruction in BMI2, we can efficiently deposit a bit into every ninth bit of an `int` (so there are now 4 8-bit integers taking on the value of the haplotype at this variant packed in an `int`). Then, using SSE2, SSE4.1 and AVX instructions one can inflate through representations from 4 8-bit integers packed in an `int` up to 4 64-bit doubles packed in a 256-bit AVX register. As such, we are then ready to operate with this variant in parallel using AVX instructions.

During development, testing indicated the memory bandwidth and cache efficiency savings of the packed binary representation provided speed-ups thanks to these instructions efficiently enabling unpacking and spreading a haplotype variant bit for parallel use. Furthermore, such a compact representation means that more of L1/L2 cache and

<code>kalisParameters</code> object	Data type						
<code>pars</code>	Locked R environment, containing: <table> <tr> <td><code>rho</code></td><td>vector length L</td></tr> <tr> <td><code>mu</code></td><td>vector length L, or scalar</td></tr> <tr> <td><code>Pi</code></td><td>$N \times N$ matrix, or scalar</td></tr> </table>	<code>rho</code>	vector length L	<code>mu</code>	vector length L , or scalar	<code>Pi</code>	$N \times N$ matrix, or scalar
<code>rho</code>	vector length L						
<code>mu</code>	vector length L , or scalar						
<code>Pi</code>	$N \times N$ matrix, or scalar						
<code>sha256</code>	character						

Table 1 The content of the data structure representing parameter objects.

memory bus bandwidth is left available for forward and backward tables, which are the largest objects we work with in this problem.

Parameters

The `kalisParameters` object uses an environment rather than list for parameters for two reasons: (i) the parameter environment and its bindings are locked which prevents changes in parameter values between forward or backward table propagation steps, since parameters must be fixed for all steps of a given forward or backward computation; and (ii) an environment explicitly ensures the (often large) parameter vectors are not copied when associated with potentially many different tables, but will always be purely referenced.

The environment contains only two members: another environment with the actual parameter values (which is locked with `lockEnvironment()`); and a SHA-256 hash of those parameter values (details in Table 1). The purpose of the hash is to be able to efficiently determine whether the correct parameter set for a given forward or backward table has been passed when computing forward or backward recursions from an already initialised table (since it would be incorrect to propagate forward or backward using different parameter sets in different parts of the genome).

Forward/backward tables

Recall that the recipients (columns) in the forward/backward tables correspond to independent HMMs. Therefore, **kalis** enables storing only a ‘slice’ of recipients in a forward/backward table, making parallelisation across non-shared memory clusters

kalisForwardTable object		kalisBackwardTable object		Data type
alpha	$= \alpha^\ell$	beta	$= \beta^\ell$	$N \times N$ matrix
alpha.f	$= F^\ell$	beta.g	$= G^\ell$	vector length N
l	$= \ell$	l	$= \ell$	integer scalar
from_recipient		from_recipient		integer scalar
to_recipient		to_recipient		integer scalar
pars.sha256		pars.sha256		character
		beta.theta		logical scalar

Table 2 The content of the core data structures representing forward and backward table objects, together with their correspondence to mathematical quantities.

much simpler: given all haplotype data, these recipient slices can be independently propagated in a communication free manner.

The forward and backward table objects contain not only the (upto) N independent forward/backward vectors at variant ℓ , but also supporting meta-data. This includes the variant the table is currently at, the scaling constants F^ℓ (forward, Equation (5)) or G^ℓ (backward, Equation (6)), the range of recipient haplotypes represented (that is, the recipient HMMs to which the column corresponds), and a hash of the parameter values used in propagating this table.

In total, a full-size forward table for example requires $8N^2 + 8N + 1576$ bytes of memory² for storage and the small overhead of R object management. Since this grows quadratically in the number of haplotypes, most functions in the package operate on forward and backward table objects in-place, rather than via the idiomatic copy-on-write mechanism of standard R. The most important consequence of this for users is that standard assignment of a table object to another variable name only creates a reference and so an explicit copy must be made by using the `CopyTable()` utility function provided in the package.

Core SIMD code

The two most important core algorithms which are accelerated with SIMD vector instructions are the forward and backward recursions. This code is fully implemented

²Measured under R 4.2.2

599 in C, with tailored modifications accounting for all combinations of: scalar/vector μ ,
600 scalar/matrix Π , and use of the asymmetric mutation model of RELATE [2] or not
601 (ie 8 combinations); to ensure that minimal memory accesses are performed where
602 possible. So, for example, scalar μ and scalar Π parameters will be faster than any
603 other combination since these values are likely to be held in registers (or at least L1
604 cache) for the duration of the recursion.

605 Additionally, in all places where we identify SIMD instructions may be used, a
606 macro is deployed, with a header file providing all mappings from these macros to
607 a specific SIMD instruction for all supported instruction sets. Taking arguably the
608 simplest non-trivial example, all `src/ExactForward*.c` and `src/ExactBackward*.c`
609 files make use of the custom macro `KALIS_MUL_DOUBLE(X, Y)` when they need to mul-
610 tiply `KALIS_DOUBLEVEC_SIZE` double precision floating point values together. The file,
611 `src/StencilVec.h` then provides definitions for these macros under each instruction
612 set `kalis` supports (via assembly intrinsics), together with a pure C alternative. For
613 this example, we have (with ... indicating other macro definitions):

```
614 // Extract from src/StencilVec.h
615 #if defined(KALIS_ISA_AVX512)
616 #define KALIS_DOUBLEVEC_SIZE 8
617 #define KALIS_MUL_DOUBLE(X, Y) _mm512_mul_pd(X, Y)
618 ...
619 #elif defined(KALIS_ISA_AVX2)
620 #define KALIS_DOUBLEVEC_SIZE 4
621 #define KALIS_MUL_DOUBLE(X, Y) _mm256_mul_pd(X, Y)
622 ...
623 #elif defined(KALIS_ISA_NEON)
624 #define KALIS_DOUBLEVEC_SIZE 2
625 #define KALIS_MUL_DOUBLE(X, Y) vmulq_f64(X, Y)
626 ...
```

```

#elif defined(KALIS_ISA_NOASM)
#define KALIS_DOUBLEVEC_SIZE 1
#define KALIS_MUL_DOUBLE(X, Y) (X) * (Y)
...
#endif

```

The inner-most loop in these core files then includes a programmatically generated unroll to the depth specified during compilation. All this is wrapped in code which dispatches using `pthread`s to multiple threads, with automatic detection of the ability to pin to specific cores if that option is passed (important in some settings to ensure a hot L1/L2 core cache). In particular, each thread operates on a subset of columns of the forward and backward tables, ensuring spatial locality for memory accesses. Furthermore, when propagating by more than a single variant position, each column (ie each independent HMM) is propagated all the way to the target variant before proceeding to the next column, ensuing temporal locality of memory accesses.

Results

We provide a brief overview of some example performance figures, though due to the highly tuned nature of **kalis**, the exact performance you can expect will be heavily dependent on your exact computer architecture and resources.

First, it is important to note we do *not* claim to have altered the scaling properties of the LS model, only that we provide an implementation which is highly optimised within the scaling constraints inherent to the model. As such, Figure 2 demonstrates that **kalis** indeed inherits the $\mathcal{O}(N^2)$ and $\mathcal{O}(L)$ properties of the original LS model.

We turn now to the benefits **kalis** does provide.

Firstly, for some of the reasons highlighted in the previous Section, **kalis** exhibits accelerated performance when propagating the forward/backward recursions over more extended stretches of the genome. This is because every effort has been made to

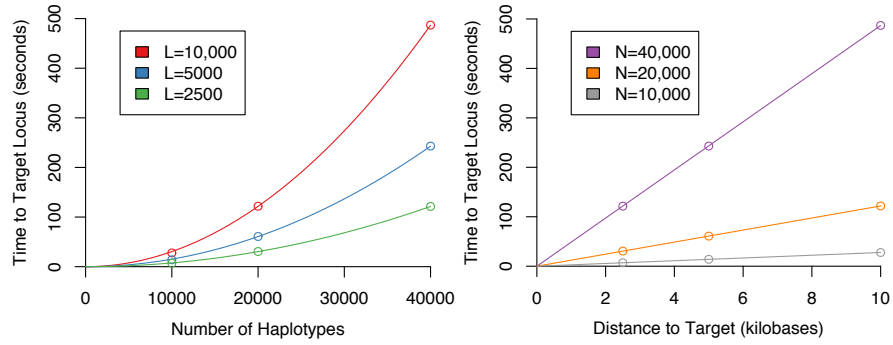


Fig. 2 **kalis** shows the expected order N^2 and order L scaling of the LS model. Computed on an Amazon Web Services **c4.8xlarge** instance (36 vCPUs, 60 GB of RAM).

be cache efficient, so that when more than a single variant step is taken, the strong cache locality design ensures that we are not memory bandwidth limited. This effect can be seen quite dramatically in Figure 3 by the rapid decrease in compute time per variant as longer stretches are propagated.

Secondly, the hard-coded loop unrolling functionality which can be controlled at compile time by the user can be seen to be beneficial in Figure 3. Clearly excessive loop unrolling is harmful, with depth 32 unrolls actually being substantially slower than no unrolling. However, unrolling to depth 8 does give a clear improvement. The best choice of unrolls will be both problem and architecture dependent, so we recommend testing different unroll levels on the target problem before performing long compute runs.

Figure 3 also illustrates that the hand-designed use of low-level vector SIMD instructions is not superfluous, with substantial speed-up afforded by their use (the difference between dashed and solid lines of the same colour).

Finally, Figure 4 shows that in certain very large problem settings **kalis**' ability to pin threads can make a substantial difference. In this setting, AVX2 showed the greatest benefit from eliminating context switching, ensuring that the cache is not invalidated by threads migrating between cores. The lack of substantial difference between AVX2 and AVX-512 here once thread pinning is employed calls for some

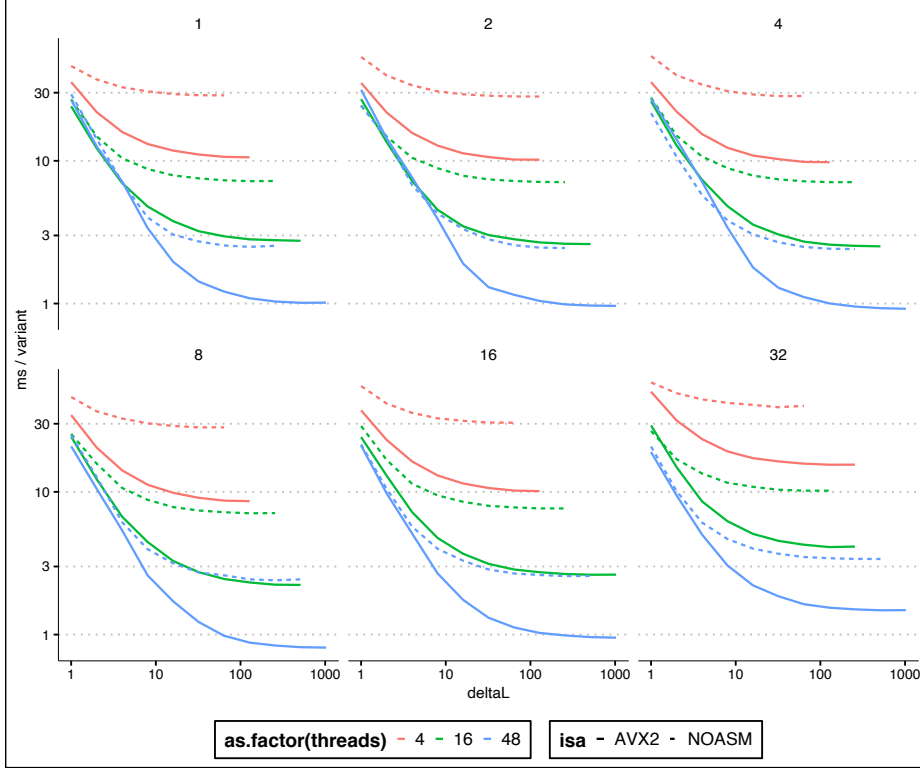


Fig. 3 Log-log plot of milliseconds per variant performance (y -axis) of the forward algorithm on 10,000 haplotypes, against the number of variants propagated (x -axis). Each panel is a different loop unrolling depth (panel title gives loop unrolling level). Line colour denotes number of CPU threads, whilst a dashed line indicates vanilla C and a solid line indicates hand-coded AVX2 instructions. In total, using AVX2, 48 threads, and loop unrolls to depth 8, it takes less than 10 seconds to propagate a 10000×10000 forward table over 10,000 variants.

investigation, though this may be a result of thermal/power throttling which is known to occur especially for AVX-512 heavy code [13].

These performance examples again highlight the importance of pilot benchmark runs with different configurations of instruction set and unroll settings before embarking on long compute runs to ensure the greatest compute efficiency is achieved for a given problem and compute architecture.

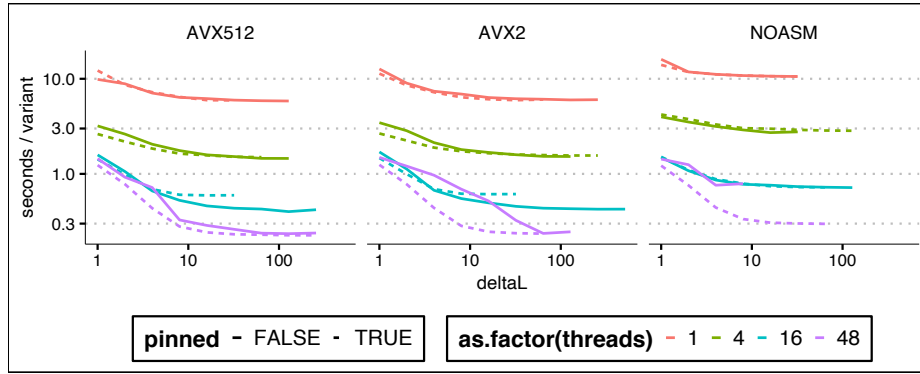


Fig. 4 Log-log plot of seconds per variant performance (y -axis) of the forward algorithm on 100,000 haplotypes, against the number of variants propagated (x -axis). Each panel is a different instruction set (AVX-512/AVX2/none). Line colour denotes number of CPU threads, whilst a dashed line indicates pinned threads and a solid line indicates no thread pinning. In total, using AVX-512, 48 threads, and pinned threads, it takes less approximately 38 minutes to propagate a 100000×100000 forward table over 10,000 variants.

Real-data example: recent selection for lactase persistence

LCT is a gene on chromosome 2 that encodes lactase, the enzyme responsible for the breakdown and digestion of lactose, the sugar commonly found in milk. Ancestral humans had a regulatory ‘switch’ on chromosome 2 that stops lactase production after infancy when children would be weaned off breast milk. Mutations that disrupt this switch allow lactase production to persist into adulthood, conferring a lifelong ability to extract energy from milk [14]. Such mutations have arisen independently at least twice in human history, in Europe and in East Africa, and are among the strongest examples of recent positive natural selection in humans [15, 16]. These mutations have been shown to spread across standard human population boundaries. Using another implementation of the LS model, [17] estimated that a European haplotype conferring lactase persistence became prevalent within the West African Fula population due to natural selection sometime over the past two thousand years.

Here we run **kalis** on 5008 haplotypes from the 1000 Genomes Phase 3 release to revisit the haplotype structure around *LCT*; the haplotypes are sampled from 26 sub-populations all over the world [18]. Figure 5 shows a clustered version of a

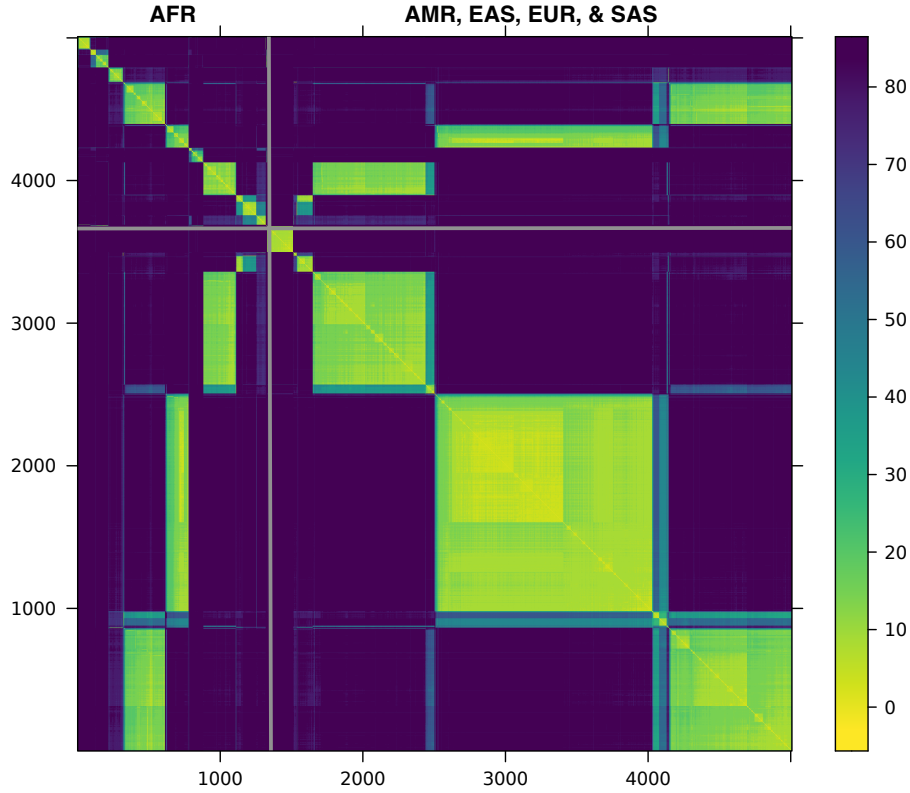


Fig. 5 Distance matrix among 5008 haplotypes calculated at **rs4988235**, upstream of *LCT*. African haplotypes are clustered in the upper left corner and separated by grey lines from non-African haplotypes from the Americas (AMR), East Asia (EAS), Europe (EUR), and SAS (South Asia). The scale on the right maps the colours to distances.

distance matrix, calculated as in Equation (4), at a variant in the regulatory region of *LCT* (**rs4988235**). To see if we could observe a pattern of gene-flow into or out of Africa similar to what was observed by [17], we use average pairwise linkage [19] to cluster the African haplotypes separately from the non-African haplotypes. In Figure 5, distances between African haplotypes are shown in the upper left corner; non-African haplotypes, in the lower right corner.

Rather than 26 clusters reflecting the 26 sampled human populations, we see that there are three very distinct lactase haplotypes that are common both within and

875 outside Africa. This suggests that these three haplotypes, under strong positive selec-
876 tion pressure, recently spread across population boundaries and presumably confer
877 lactase persistence. However, these three haplotypes are not the only structure we see:
878 in the upper left corner of the African (AFR) block we see some haplotypes that are
880 only found inside Africa; and in the non-African block, a haplotype that is only found
881 outside Africa. We can also see some sub-structure within the clear haplotype blocks.
882
883
884

885 Discussion

886
887 In the User Guide (Additional file 2), we introduce the package from a user perspective,
888 from package installation right through to decoding a single variant position in R using
889 **kalis**.
890
891

892
893 There are many avenues for future research in developing **kalis**. On the model side,
894 for example, allowing for different recombination rates between sub-populations as
895 done in fastPHASE [20] would be a natural extension.
896
897

898
899 On the computational side, ARM scalable vector extensions [21] represent an inter-
900 esting new approach to SIMD instruction sets, where the width of instructions need
901 not be hard coded prior to compilation. At present it is not widely available, but as this
902 rolls out, it would be natural to extend **kalis** to enable targeting this new instruction
903 set.
904
905

906
907 An important utility extension is expanding the file formats that **kalis** can natively
908 read via `CacheHaplotypes()`, to enable simpler and more streamlined software
909 pipelines when bioinformaticians incorporate **kalis** into their workflows.
910
911

912
913 Finally, a future avenue of potential development is extension of **kalis** to support
914 GPU or tensor cards. Note that it was an explicit design choice to initially target CPU
915 SIMD extensions, since the vast majority of University high performance computing
916 clusters have a huge amount of untapped compute power in this form, but often
917 much more limited availability of specialist extension cards. Therefore, by pushing
918
919
920

performance as extensively as possible via CPU only means, we provide the greatest
potential impact for end users. This does not preclude future versions adding support
for add-on compute cards.

Conclusion

kalis provides a R interface to a highly optimized C implementation of the LS model
that enables local ancestry, selection, and associations studies in modern large genomic
datasets.

Availability and requirements

Project name: **kalis**

Project home page: <https://kalis.louisaslett.com/>

Operating system(s): Linux, MacOS, Windows

Programming language: R, C

Other requirements: R ($\geq 3.5.0$)

License: GPL (≥ 3)

Any restrictions to use by non-academics: None beyond GPL (≥ 3).

List of abbreviations

LS model = Li & Stephens model

HMM = hidden Markov model

SIMD = single instruction, multiple data

Supplementary information.

Declarations

Ethics approval and consent to participate. Not applicable.

967 **Consent for publication.** Not applicable.
968
969 **Availability of data and materials.** The package source code repository is at
970 <https://github.com/louisaslett/kalis>. All scripts for reproducing the results of this
971 paper are available in this repository <https://github.com/louisaslett/kalis-bmc>. The
972 two external dependencies are: 1000 Genomes data which are available for download
973 from <https://www.internationalgenome.org/>; and the msprime simulator, which may
974 be downloaded from <https://tskit.dev/software/msprime.html>.
975
976 **Competing interests.** The authors declare that they have no competing interests.
977
978 **Funding.** This project was supported by the NHGRI Centers for Common Disease
979 Genomics grant (UM1-HG008853), active from 2015-2022.
980
981 **Authors' contributions.** LA architected and wrote the C-core. LA and RC collab-
982 orated on the R interface. RC conducted the real-world lactase persistence example.
983
984 LA and RC wrote and approved the final manuscript.
985
986 **Acknowledgements.** Both authors would like to acknowledge Professor Ira Hall,
987 Professor Chris Holmes, and Dr Chris Spencer for their discussions and advice on this
988 project.
989
990

997 References

- 998
999
1000 [1] Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recom-
1001 bination hotspots using single-nucleotide polymorphism data. *Genetics* **165**,
1002 2213–2233 (2003). URL <http://www.genetics.org/content/165/4/2213>.
1003
1004
1005 [2] Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide
1006 genealogy estimation for thousands of samples. *Nature Genetics* **51**, 1321–1329
1007 (2019).
1008
1009
1010
1011
1012

[3] Song, Y. S. Na Li and Matthew Stephens on Modeling Linkage Disequilibrium.	1013
<i>Genetics</i> 203 , 1005–1006 (2016). URL http://www.genetics.org/content/203/3/	1014
1005.	1015
	1016
	1017
	1018
[4] Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population	1019
structure using dense haplotype data. <i>PLoS genetics</i> 8 , e1002453 (2012).	1020
	1021
	1022
[5] R Core Team. <i>R: A Language and Environment for Statistical Computing</i> . R	1023
Foundation for Statistical Computing, Vienna, Austria (2023). URL https://	1024
www.R-project.org/ .	1025
	1026
	1027
	1028
[6] Sutter, H. The free lunch is over: A fundamental turn toward concurrency in	1029
software. <i>Dr. Dobbs's Journal</i> 30 , 202–210 (2005).	1030
	1031
	1032
[7] Peleg, A. & Weiser, U. MMX technology extension to the Intel architecture.	1033
<i>IEEE Micro</i> 16 , 42–50 (1996).	1034
	1035
	1036
[8] Intel Corporation. Intel Architecture Instruction Set Extensions and Future	1037
Features. Tech. Rep. 319433-046 (2022).	1038
	1039
	1040
[9] ARM. NEON Programmer's Guide. Tech. Rep. DEN0018A ID071613 (2013).	1041
	1042
	1043
[10] Alpert, D. & Avnon, D. Architecture of the Pentium microprocessor. <i>IEEE Micro</i>	1044
13 , 11–21 (1993).	1045
	1046
	1047
[11] ISO. <i>ISO/IEC 9899:2018 Information technology — Programming languages —</i>	1048
<i>C</i> Fourth edn (BSI, 2018). URL https://www.iso.org/standard/74528.html .	1049
	1050
	1051
[12] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in	1052
speech recognition. <i>Proceedings of the IEEE</i> 77 , 257–286 (1989).	1053
	1054
	1055
	1056
	1057
	1058

1059 [13] Schöne, R., Ilsche, T., Bielert, M., Gocht, A. & Hackenberg, D. IEEE (ed.)
1060 *Energy efficiency features of the Intel Skylake-SP processor and their impact on*
1061 *performance.* (ed.IEEE) *2019 International Conference on High Performance*
1062 *Computing & Simulation (HPCS)*, 399–406 (2019).
1063
1064
1065
1066 [14] Ingram, C. J., Mulcare, C. A., Itan, Y., Thomas, M. G. & Swallow, D. M. Lactose
1067 digestion and the evolutionary genetics of lactase persistence. *Human genetics*
1068 **124**, 579–591 (2009).
1069
1070
1071
1072 [15] Ranciaro, A. *et al.* Genetic origins of lactase persistence and the spread of
1073 pastoralism in Africa. *The American Journal of Human Genetics* **94**, 496–510
1074 (2014).
1075
1076
1077
1078 [16] Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the
1079 lactase gene. *The American Journal of Human Genetics* **74**, 1111–1120 (2004).
1080
1081
1082 [17] Busby, G. *et al.* Inferring adaptive gene-flow in recent African history. *BioRxiv*
1083 205252 (2017).
1084
1085
1086 [18] Consortium, . G. P. *et al.* A global reference for human genetic variation. *Nature*
1087 **526**, 68 (2015).
1088
1089
1090 [19] Sokal, R. R. A statistical method for evaluating systematic relationships. *Univ.*
1091 *Kansas, Sci. Bull.* **38**, 1409–1438 (1958).
1092
1093
1094 [20] Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale pop-
1095 ulation genotype data: applications to inferring missing genotypes and haplotypic
1096 phase. *The American Journal of Human Genetics* **78**, 629–644 (2006).
1097
1098
1099 [21] Stephens, N. *et al.* The ARM Scalable Vector Extension. *IEEE Micro* **37**, 26–39
1100 (2017).
1101
1102
1103
1104