KAGE: Fast alignment-free graph-based genotyping of SNPs and short indels

Ivar Grytten[1,2], Knut D. Rand[1,2], Geir K. Sandve[1,2,3]

1 Department of Informatics, University of Oslo, Norway
2 Centre for Bioinformatics, University of Oslo, Norway
3 K.G. Jebsen Centre for Coeliac Disease Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway

Abstract

One of the core applications of high-throughput sequencing is the characterization of individual genetic variation. Traditionally, variants have been inferred by comparing sequenced reads to a reference genome. There has recently been an emergence of *genotyping* methods, which instead infer variants of an individual based on variation present in population-scale repositories like the 1000 Genomes Project. However, commonly used methods for genotyping are slow since they still require mapping of reads to a reference genome. Also, since traditional reference genomes do not include genetic variation, traditional genotypers suffer from reference bias and poor accuracy in variation-rich regions where reads cannot accurately be mapped.

We here present KAGE, a genotyper for SNPs and short indels that is inspired by recent developments within graph-based genome representations and alignment-free genotyping. We propose two novel ideas to improve both the speed and accuracy: we (1) use known genotypes from thousands of individuals in a Bayesian model to predict genotypes, and (2) propose a computationally efficient method for leveraging correlation between variants.

We show through experiments on experimental data that KAGE is both faster and more accurate than other alignment-free genotypers. KAGE is able to genotype a new sample (15x coverage) in less than half an hour on a consumer laptop, more than 10 times faster than the fastest existing methods, making it ideal in clinical settings or when large numbers of individuals are to be genotyped at low computational cost.

Introduction

Recent studies such as the 1000 Genomes Project [1] have produced large catalogues of known human genetic variation. We now have individual-level genotype information for thousands of individuals at known variant sites, providing unprecedented knowledge about how often genetic variants occur together at sub-population and individual levels.

Discovery and characterization of an individual's genetic variation have traditionally been done in two steps: 1) *variant calling*: discovering genetic variation in the genome, and 2) *genotyping* those called variants: determining whether each genetic variant is present in one or both of the chromosomes of the individual. Variant calling is a computationally expensive task, usually performed by first sequencing the genome of interest at high coverage, followed by mapping the sequenced reads to a reference genome and inferring variants according to where the mapped reads differ from the reference. It is now possible to skip the variant calling step, instead relying on catalogues of known genetic variation to directly genotype a given individual. For humans, the huge amount of already detected genomic variation means that a large amount (>90%) of an

individual's genetic variation can be detected this way [1]. Genotyping individuals at sites of known variation has in principle been performed already since the early 2000's based on SNP ChIPs. This has however been restricted to a limited set of fixed sites that a given chip has been designed to capture. While these ChIP-based techniques are still popular today due to their low cost, e.g. in GWAS studies [6], they are limited by ChIP architecture and thus only able to genotype about 1-2 million variants. Contrary to this, *sequence-based* genotyping techniques are, while being more expensive to perform, in theory able to genotype any genetic variant that has already been detected in the population.

Sequence-based genotyping is traditionally performed by aligning sequenced reads to a reference genome and examining how the reads support genotypes locally at each variant [4, 12]. While these methods generally have high accuracy, they are slow, mostly because read-mapping is computationally expensive. Also, since these methods require the mapping of reads to a reference genome, they have a tendency to perform poorly in regions where the reference genome is very different from the sequenced genome, such as in variant-rich regions. Furthermore, since reads are more likely to be correctly mapped when they are similar to the reference, reads supporting variant-alleles may be underrepresented among the mapped reads, resulting in a bias towards the reference alleles, a problem referred to as reference bias [9].

In an attempt to both speed up genotyping and avoid reference bias, several *alignment-free* approaches have emerged during the last few years [2,14–16]. These methods work by representing genetic variants by their surrounding kmers (sequences with length k covering each variant) and looking for support for these kmers in the sequenced reads. Since these methods do not map reads to a reference genome, they mitigate the problem of reference bias, and are usually computationally very efficient since kmer-lookup is fast compared to read mapping. However, these methods struggle when variant alleles cannot be represented by unique kmers, e.g. because a variant allele shares one or more kmers with another location in the genome. Malva [2] attempts to solve this problem by using larger kmers in cases where kmers are non-unique, but is still not able to genotype all variants using this approach. The more recent method PanGenie [3] takes Malva's approach one step further and uses known population haplotype information to infer the likelihood of genotypes for variants that do not have unique kmers, such as variants in repetitive regions of the genome. PanGenie does this by using a Hidden Markov Model (HMM) with one state for each possible pair of haplotypes from the set of known haplotypes in the population. While this approach works well for relatively few input haplotypes and variants (PanGenie was tested using haplotypes from around 10 individuals and about 7 million variants), the number of states in the HMM increases quadratically with the number of haplotypes, and the method scales poorly when more than a few haplotypes are used. This means that the more than 5000 haplotypes and 80 million variants available in the 1000 Genomes Project cannot easily be utilized by PanGenie. Furthermore, PanGenie does not use kmer-information at all for variants with non-unique kmers.

We here describe KAGE – a new genotyper for SNPs and short indels that builds on recent ideas of alignment-free genotyping from Malva and PanGenie in order to be computationally efficient. KAGE implements two novel ideas for utilizing all previously known haplotype information from repositories such as the 1000 Genomes Project in order to improve genotyping accuracy. We show that when these ideas are combined, we get a genotyper that is both more accurate and computationally efficient than existing alignment-free genotypers.

Results

We here present two new ideas inspired by recent progress in alignment-free genotyping. We show how each idea improves genotyping accuracy for alignment-free genotyping methods, and show that these two ideas can be combined into a graph-based alignment-free genotyper that is more accurate and more computationally efficient than existing alignment-free genotyping methods.

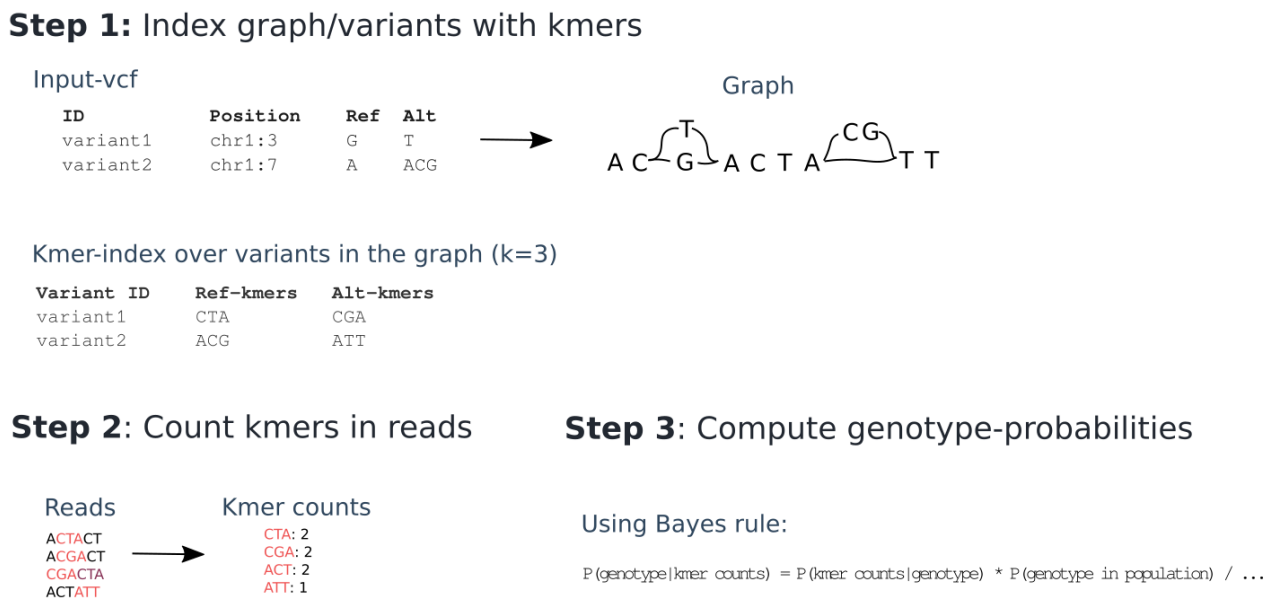

Figure 1. Overview of a typical alignment-free genotyping approach. First, kmers covering each allele of each variant is stored (step 1), e.g. by choosing kmers from a graph representation of the variants. All kmers from the reads are then collected (step 2), and genotype probabilities are computed using Bayes rule with genotype probabilities from a known population used as priors (step 3). The most likely genotype is chosen.

Idea 1: Modelling expected kmer-counts from the population

Alignment-free genotypers, like Malva and PanGenie, compute genotype likelihoods by counting the number of kmers present in the read data set that support each allele of every variant to be genotyped, as illustrated in Figure 1. A problem with this approach is that some variants may have alleles that are covered by kmers that also match elsewhere in the genome (for some or many individuals), making such variants difficult to genotype, as illustrated in Figure 2. We hypothesize that this problem may be addressed by modelling the expected match count for each individual kmer by using population data, such as data from the 1000 Genomes Project.

To test this hypothesis, we implemented a baseline genotyper that follows the scheme shown in Figure 1: each variant is represented by kmers, and the genotyper counts how many times each kmer exists in the read data sets, and computes genotype probabilities using Bayes formula (naively assuming each kmer is unique in the genome). We also created a more sophisticated version of this genotyper, where we assume that each kmer exists with a given frequency in the population, and use this information when computing the probability of observing the given number of kmers for each variant (Methods). We benchmark both these methods against Malva. As shown in Figure 3, our naive method performs similarly to Malva, whereas the approach where expected kmer counts are modelled from the population clearly performs better. Although this is an experiment performed on a single small benchmark dataset, these findings provide an indication that the idea of comparing observed kmer counts to expected kmer counts based on a population may work.

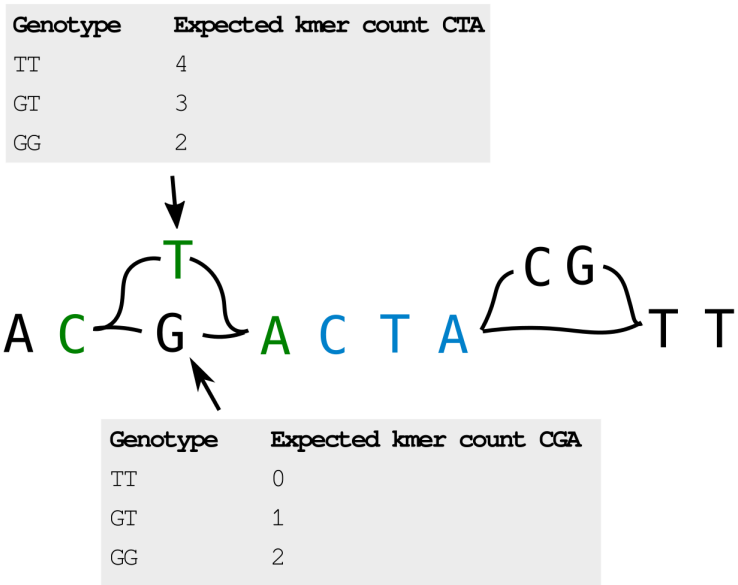

Figure 2. Graph with duplicate kmers. The SNP G/T has a kmer CTA (green) on the variant allele that also exists on the reference path (blue). If we observe each of the kmers CTA and CGA once in the read set, we might be fooled to believe that GT is the most likely genotype for this variant. However, when knowing that the kmer CTA is expected to occur at least once in the read data set, due to the duplication, we might conclude otherwise. This information can be used to adjust the probabilities used to compute the binomial probabilities of observing kmer counts given genotypes.

Idea 2: Using a single variant to adjust the prior

While Malva uses the allele frequency from the population as priors for genotype likelihoods, PanGenie uses a more sophisticated approach where genotypes are inferred using a Hidden Markov Model (HMM), meaning that the predicted genotype of one variant provides information that can support the determination of the genotypes of other variants. While this approach may allow PanGenie to predict genotypes with higher accuracy than Malva, this method does not scale to handle a large number of known haplotypes, since the HMM must represent every possible pair of haplotypes as a separate state, making the number of HMM states grow quadratically with the number of haplotypes. We hypothesize that a simpler approach may be better for two reasons: First, a simple approach may be able to utilize all the haplotype information available (e.g. the 2548 individuals available from the 1000 Genomes project, rather than only a handful that the PanGenie approach would typically employ), and second, such an approach may be computationally faster than an HMM-approach that requires inference on hidden states. To test this hypothesis, we implement a way of efficiently computing prior probabilities for each genotype of each variant given genotype probabilities of other variants (Methods). We let every variant only depend on a single other variant, which leads to a very efficient way of genotyping all variants, and allows us to use as many underlying haplotypes as we would like. Finding variants that can be used to help predict other variants is done prior to genotyping, and a fast pre-computed lookup table is then used when genotyping. We test this method against PanGenie, where PanGenie is run with a varying number of known haplotypes as input. In order to test PanGenie with as many haplotypes as possible from the 1000 Genomes Projects, we perform this experiment on a small test data set (5 million base pairs of chromosome 1).

The result of this experiment is shown in Figure 4, where it can be seen that the accuracy of both our method (KAGE) and PanGenie increase with the number of known haplotypes. This experiment is performed on a small test dataset, so as to allow the inclusion of a larger number of individuals in Pangienie, and is not meant as a benchmarking of the methods per se. These findings provide an indication that the use of a single variant as prior still allows us to leverage information from a large number of individuals.

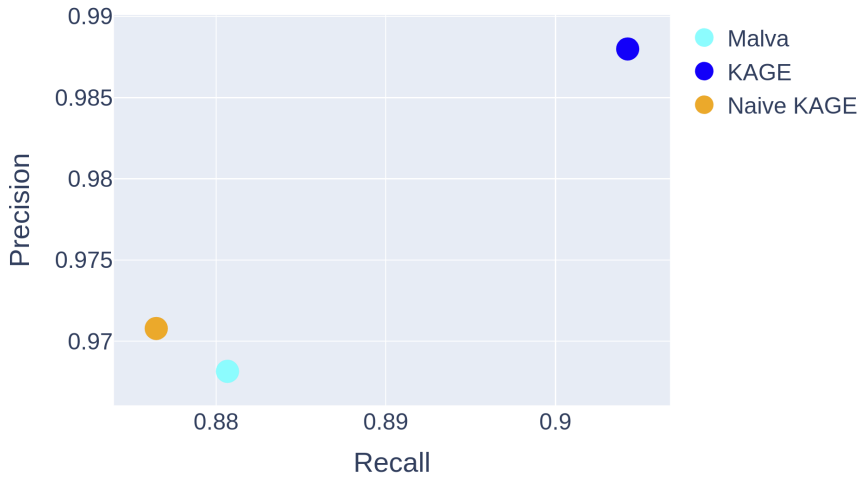

Figure 3. Genotype accuracy increases when expected kmer counts are modelled from the population. Our prototype alignment-free genotyper performs a bit worse than Malva, as expected. However, when we compute expected kmer counts from the population, using 1000 Genomes data, and use these counts to compute the probability of observing the given kmers in our data, the genotyping accuracy greatly improves.

A new graph-based alignment-free genotyper

We combine the two ideas discussed above into a new alignment-free genotyper, KAGE, and compare this genotyper in terms of running time, memory usage and accuracy against existing alignment-free genotypers as well as the most commonly used alignment-based genotypers. We follow the guidelines from "Best practices for benchmarking germline small-variant calls in human genomes" [7] and compare the output of each genotyper against a truth data set (similarly to how PanGenie and Malva assess their methods).

In the following experiments, we run PanGenie with 64 input haplotypes (from 32 individuals), since more haplotypes than that make PanGenie too slow and memory demanding when run on whole-genome data. We run our own method with all of the 5096 haplotypes available from the 1000 Genomes Project. When listing run times, we consider only the time used to genotype a given sample, not the time used to index the reference genomes, graphs, or variants, since such indices can be created once and then be re-used for genotyping multiple samples. All experiments presented here consist of genotyping variants from the 1000 Genomes Project using short reads as input. Accuracy is presented by recall and precision when predicted genotypes are compared against a truth dataset. Further details about these experiments can be found in the Methods section, and a Snakemake pipeline for reproducing all the results can be found at `https://github.com/ivarg/genotyping-benchmarking`.

In the experiments, we compare the genotypers listed in Table 1, which includes the alignment-based genotyper Graphtyper as well as all alignment-free genotypers we believe are relevant or

commonly used. For comparison, we also run the commonly used variant caller GATK [12], which does not only genotype a specified set of variants, but tries to detect any variation from the reference. We include GATK so that the results of the genotypers can be contrasted to the alternative approach of performing full variant calling.

We genotype the individual HG002 from GIAB [17], which is an individual commonly used to benchmark variant callers and genotypers. We run the genotypers listed in Table 1 using short Illumina reads (also from GIAB) and measure the run-time, memory usage and accuracy of all the methods. Table 2 shows the results from this experiment using reads with 15x coverage, and Supplementary Table 1 shows the results from 30x coverage, which are very similar to the results from the 15x experiment. These results show that KAGE is about as accurate as or more accurate than all the other alignment-free methods and faster than any of the methods. GATK comes out as clearly more accurate than all other methods, but is also as expected slower than the alignment-free methods. KAGE is close to the alignment-based method Graphtyper in terms of accuracy while being more than 10 times faster.

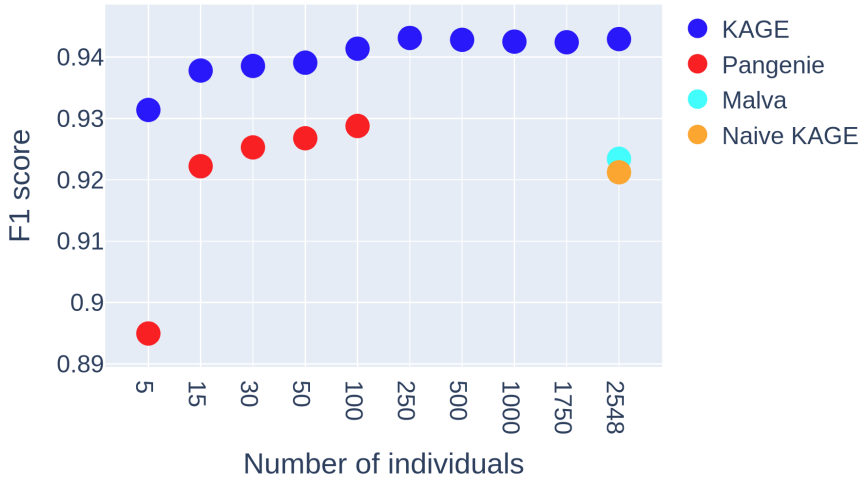

Figure 4. Accuracy of Pangenie and KAGE as a function of the number of individuals included in the model on a small test dataset. Recall and precision (shown as F1 score) are computed for the predicted genotypes by the various methods on a small test dataset (genome size 5 million bp) with reads simulated from HG002, in order to test whether the accuracy of PanGenie and KAGE increases with a larger number of individuals in the model. The F1 score of Malva and "our naive approach" are shown as a single dot, since both of these only use the whole population genotype frequency when estimating genotype probabilities.

Genotyper	Alignment-free/alignment-based	Uses prior haplotype-information
Malva	Alignment-free	No
Bayestyper	Alignment-free	No
PanGenie	Alignment-free	**Yes**
Graphtyper	Alignment-based	No
GATK	Alignment-based	No

Table 1. Overview of genotypers.

	Indels recall	Indels precision	Indels F1	SNPs recall	SNPs precision	SNPs F1	Runtime	Memory usage
Us	0.579	0.892	0.702	0.93	0.976	0.952	21 min	24 GB
Pangenie	0.561	0.923	0.698	0.906	0.979	0.941	3.8 hours	113 GB
Bayestyper	0.477	0.99	0.644	0.814	0.995	0.895	7.2 hours	31 GB
Malva	0.32	0.879	0.469	0.943	0.923	0.881	16.9 hours	50 GB
Graphtyper	0.552	0.942	0.696	0.889	0.991	0.938	4.3 hours	23 GB
Gatk	0.864	0.95	0.905	0.959	0.981	0.97	7.3 hours	70 GB

Table 2. Results from genotyping HG002.

Methods

Genotyping

The following describes the model and assumptions used in KAGE. We here assume all variants are biallelic variants, each with two *alleles*, a *reference allele* and an *alternative allele*. An individual can have the following three genotypes at a variant: Having the reference allele in both chromosomes (we refer to this as having the genotype 0/0), having the alternative allele in both chromosomes (1/1), or having the reference allele in one chromosome and the alternative allele in the other (0/1). Note that we are not concerned with the phasing of genotypes, so 1/0 and 0/1 are considered as being the same genotype.

Assuming a variant of interest i, we define $P(G_i)$ as the probability of having the genotype G at this variant, where G is either 0/0, 0/1 or 1/1. We let $K_i = (K_{ir}, K_{ia})$ be the observed kmer counts on the reference and variant allele of variant i. Every variant of interest i has a preselected helper variant h with kmer counts K_h. In order to find the most likely genotype for variant i, we want to calculate the posterior probabilities $P(G_i|K_i, K_h)$:

$$P(G_i|K_i, K_h) = P(K_i|G_i) \sum_{G_h G} P(G_h)P(G_i|G_h)P(K_h|G_h) \tag{1}$$

In the above formula, $P(K_i|G_i)$ is the likelihood of observing the kmer counts K_i on the variant of interest given a genotype G_i. The summation corresponds to a prior probability of the genotype G_i given the observed counts on the helper variant and the population structure, with the individual probabilities corresponding to the following:

- $P(G_h)$ is the probability (over the population) of genotype G_h at the helper variant.
- $P(G_i|G_h)$ is the probability (over the population) of having genotype G_i on the variant of interest given genotype G_h at the helper variant
- $P(K_h|G_h)$ is the likelihood of observing the kmer counts K_h given genotype G_h at the helper variant, similarly to $P(K_i|G_i)$.

The probabilities $P(G_h)$ and $P(G_i|G_h)$ are based on the observed genotypes on the specific variants i and h in the population, as well as priors (pseudo counts) that are based on all the variant-helper pairs across the genome. The full Bayesian model is detailed in the supplementary material.

The probability distributions for $P(K_i|G_i)$ and $P(K_h|G_h)$ are estimated by marginalizing over Poisson distributions for the unknown number of occurrences of the kmers in the genome (where occurrence uncertainty follows from variant uncertainty). If we knew the number of times the kmer of a variant occurs elsewhere in the genome (D_ir, or D_ia), we can calculate the probability of observing the observed number of kmers in the sample as:

$$P(K_{ir}|D_{ir}, G_i) = Pois(K_{ir}; \lambda(H_r(G_i) + D_{ir} + \epsilon)) \tag{2}$$

where ϵ is the probability of a kmer not belonging to the allele being erroneously observed and H counts the number of copies of each allele for a genotype, i.e: $H_r(0/0)) = H_a(1/1) = 2$, $H_r(0/1) = H_a(0/1) = 1$ and $H_r(1/1) = H_a(0/0) = 0$. With uncertainty in the number of duplicates we get:

$$P(K_i|G_i) = \sum_{(} D_i)P(K_i|D_i, G_i)P(D_i). \tag{3}$$

The corresponding probabilities $P(K_{ia}|G_i)$, $P(K_{hr}|G_r)$ and $P(K_{ha}|G_a)$ are calculated using the same setup. We refer to the Supplementary material for more details.

Benchmarking genotypers

We compared KAGE against the following other tools: GATK HaplotypeCaller version 4.2.3.0, Graphtyper version 2.7.1 (using the genotype subcommand), Malva version 1.3.1, PanGenie (commit ID da87f55cdccf31cd2c0008fd4848e33ba42021fc) and Bayestyper version 1.5. We also tried including Platypus in the comparison, but were not able to get the tool to work. Both Bayestyper and Malva use kmer-counts from kmc [5]. We ran kmc version 3.1.1 and included singleton kmers (using the flag -ci1) for Bayestyper and Malva, as this seemed to give the best results. For Bayestyper and Malva, we used the suggested k specified by the tool documentation, which was 51 for Bayestyper and 31 (short kmers) and 45 (long kmers) for Malva. We ran KAGE and PanGenie with the recommended k=31.

GATK and Graphtyper were run on reads mapped to GRCh38 using BWA-MEM [8] version 0.7.17-r1188. A Snakemake [10] pipeline for running all the experiments and a conda environment with all the necessary tools can be found at `https://github.com/ivargr/genotyping-benchmarking`. Reads were simulated using Graph Read Simulator (version 0.0.2) with substitution, deletion and insertion rate all set to 0.001. When using experimental data all reads were obtained from the GIAB repository and downsampled to the given coverage used in the experiment. Exact data URLs can be found in the config file of the Snakemake pipeline.

Variants genotyped in the experiments are from the 1000 Genomes Project v2a. Only variants with allele frequency >0.1% were used. We used Hap.py version 0.312 [7] to compute accuracy (precision and recall) against a truth dataset.

Software implementation

KAGE was implemented using Python 3, and some run-time critical code (kmer lookup) were written in Cython. All code developed as part of this method has been modularized into four Python packages, each available through the PyPi package repository: kage-genotyper (main tool for genotyping), graph_kmer_index (building and using the kmer indexes necessary for the genotyping), kmer-mapper (counting kmers) and obgraph (building and working with the sequence graphs). KAGE with all dependencies can be installed directly by installing only the kage-genotyper package.

Discussion

We find that our genotyper generally has higher accuracy than the other methods. PanGenie seems to perform similarly on indels, but performs slightly worse on SNPs. Malva performs worse than all the other methods, which we believe makes sense considering that Malva does not have a good way of dealing with variants represented by non-unique kmers. While PanGenie has decent accuracy, it requires more than 10 times the run time and 5 times the memory of KAGE on the experiments we have run. KAGE is able to genotype a full sample (15x coverage) with less than 30 GBs of RAM in less than 30 minutes using 16 compute cores, which means that it would be possible to genotype a sample quickly and easily on a standard laptop, such as in

a clinical setting. The short run time also means that KAGE could be valuable if thousands of samples are to be genotyped in reasonable time, which could become a relevant situation as more than a million genomes are expected to be sequenced in the coming years [13].

We confirm that performing full variant calling with a method such as GATK yields considerably higher accuracy than only genotyping a preset of specified variants. One should therefore be aware of this speed/accuracy tradeoff: genotyping a preset of known variants can be very efficient and give decent accuracy now that a lot of genomic variation is known. However, when the aim is to accurately discover as many variants in a sample as possible and runtime is not an issue, a variant caller like GATK or DeepVariant [11] would be the best option. We note that GATK has higher accuracy in our experiments than what has been found in the previous benchmarks presented in the Graphtyper [4] and Malva [2] papers, where Malva and Graphtyper are both shown to have as good as and in some cases better genotyping accuracy than GATK. We believe this might be because GATK has been significantly improved in the last few years since these papers were published [12].

We are surprised to see that KAGE and PanGenie, which are completely alignment-free, are able to achieve very close accuracy to Graphtyper, which first maps and aligns all reads using BWA-MEM and then locally realigns all reads to a sequence graph. We believe the reason must be that KAGE and PanGenie use more sophisticated ways of computing the genotype likelihoods, by using population information. We speculate that PanGenie should be able to achieve even better accuracy by using more known haplotypes as input, but this would lead to very long run-time and high memory usage due to the quadratic complexity of the Hidden Markov Model used.

While Malva's strength is supposed to be relatively low run-time, Malva is unfortunately not optimized to run on multiple compute cores/threads. We believe Malva would be able to be close to our genotyper in run time if it was able to utilize multiple compute cores in the genotyping step. We observe that Bayestyper differs from the other methods by having a lower recall and a higher precision, meaning that Bayestyper is likely more conservative when genotyping. We were not able to tune Bayestyper to provide higher recall at the cost of lower precision.

There are a few limitations of KAGE. First, KAGE is for now only able to genotype SNPs and short indels, and it is important to note that our benchmarks only compare the tools on SNPs and short indels, while several of the methods have been developed also for structural variation. We believe, however, that the ideas that KAGE are based on can be generalized to structure variant calling with short reads, which would be an interesting opportunity for further work. Second, KAGE relies on a relatively good database of known variation from a population, meaning it would not work on species where such databases are not yet available. This does, however, also mean that we expect KAGE to potentially perform even better in the future when more genetic variation data from more individuals will be available. Finally, KAGE and the other kmer-based methods are not able to use information from paired-end reads. We do, however, not think this is a major problem since single-end sequencing is very common and cheaper to perform than paired-end sequencing.

Conclusions

We have presented KAGE, a new genotyper for SNPs and short indels. KAGE is alignment-free and inspired by ideas from previous alignment-free genotypers such as Malva and PanGenie, but is able to get more than 10x speedup over other methods and improved accuracy by using simple yet powerful ideas for utilizing prior population information from databases such as the 1000 Genomes Project. We believe KAGE will be a relevant method in the coming years when millions of genomes will be sequenced and fast and accurate genotyping will be even more important than before.

References

1. 1000 Genomes Project Consortium and others. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

2. L. Denti, M. Previtali, G. Bernardini, A. Schönhuth, and P. Bonizzoni. MALVA: genotyping by Mapping-free ALlele detection of known VAriants. *Iscience*, 18:20–27, 2019.

3. J. Ebler, W. E. Clarke, T. Rausch, P. A. Audano, T. Houwaart, J. Korbel, E. E. Eichler, M. C. Zody, A. T. Dilthey, and T. Marschall. Pangenome-based genome inference. *bioRxiv*, 2020.

4. H. P. Eggertsson, H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, K. E. Hjorleifsson, A. Jonasdottir, A. Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, 49(11):1654–1660, 2017.

5. M. Kokot, M. Długosz, and S. Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017.

6. A. Korte and A. Farlow. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9(1):1–9, 2013.

7. P. Krusche, L. Trigg, P. C. Boutros, C. E. Mason, M. Francisco, B. L. Moore, M. Gonzalez-Porta, M. A. Eberle, Z. Tezak, S. Lababidi, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, 37(5):555–560, 2019.

8. H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013.

9. R. Martiniano, E. Garrison, E. R. Jones, A. Manica, and R. Durbin. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology*, 21(1):1–18, 2020.

10. F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, et al. Sustainable data analysis with Snakemake. *F1000Research*, 10, 2021.

11. R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, 2018.

12. R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178, 2017.

13. G. Saunders, M. Baudis, R. Becker, S. Beltran, C. Béroud, E. Birney, C. Brooksbank, S. Brunak, M. Van den Bulcke, R. Drysdale, et al. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nature Reviews Genetics*, 20(11):693–701, 2019.

14. A. Shajii, D. Yorukoglu, Y. William Yu, and B. Berger. Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics*, 32(17):i538–i544, 2016.

15. J. A. Sibbesen, L. Maretty, and A. Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, 50(7):1054–1059, 2018.

16. C. Sun and P. Medvedev. Toward fast and accurate SNP genotyping from whole genome sequencing data for bedside diagnostics. *Bioinformatics*, 35(3):415–420, 2019.

17. J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3(1):1–26, 2016.