Bioinformatics, 34(4), 2018, 550–557
doi: 10.1093/bioinformatics/btx637
Advance Access Publication Date: 9 October 2017
Original Paper

OXFORD

Genome analysis

findGSE: estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies

Hequan Sun[1], Jia Ding[2], Mathieu Piednoël[1] and Korbinian Schneeberger[1,*]

[1]Department of Plant Developmental Biology and [2]Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 20, 2017; revised on September 20, 2017; editorial decision on October 3, 2017; accepted on October 6, 2017

Abstract

Motivation: Analyzing *k*-mer frequencies in whole-genome sequencing data is becoming a common method for estimating genome size (GS). However, it remains uninvestigated how accurate the method is, especially if it can capture intra-species GS variation.

Results: We present *findGSE*, which fits skew normal distributions to *k*-mer frequencies to estimate GS. *findGSE* outperformed existing tools in an extensive simulation study. Estimating GSs of 89 *Arabidopsis thaliana* accessions, *findGSE* showed the highest capability in capturing GS variations. In an application with 71 female and 71 male human individuals, *findGSE* delivered an average of 3039 Mb as haploid human GS, while female genomes were on average 41 Mb larger than male genomes, in astonishing agreement with size difference of the X and Y chromosomes. Further analysis showed that human GS variations link to geographical patterns and significant differences between populations, which can be explained by variable abundances of LINE-1 retrotransposons.

Availability and implementation: *R* package of *findGSE* is freely available at https://github.com/schneebergerlab/findGSE and supported on linux and Mac systems.

Contact: schneeberger@mpipz.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome size (GS) refers to the amount of haploid nuclear DNA of an organism and is typically measured in picograms or megabases (where 1 pg is equivalent to 978 Mb) (Soltis *et al.*, 2003; Doležel *et al.*, 2003). GS estimation (GSE) is not only important for our understanding of genome evolution (Gregory, 2005), but is also required for many practical aspects in genome sequencing and assembly, including approximating the amount of sequencing data needed and evaluating the completeness of assembled genome sequences.

Methods for GSE can be classified into two broad categories, experimental or computational. Experimental techniques, like feulgen densitometry or the widely used flow cytometry, have been used for many years and were applied to tens of thousands of species leading to various GS databases (Gregory *et al.*, 2007). It is important to note that all experimental methods rely on specific genomes which are used as internal/external size standards (Bennett *et al.*, 2003; Doležel *et al.*, 2007; Hardie *et al.*, 2002). However, the goal of establishing a set of commonly used standards has remained unrealized since the importance of standards was emphasized (Doležel and Bartoš, 2005; Doležel and Greilhuber, 2010). This and other factors like differences in sample preparation, staining/dyeing strategy, and stochastic drift of instruments can result in significant differences in GSE for the same genome when analyzed in different laboratories (Doležel *et al.*, 1998).

Alternatively, GS can be estimated computationally using whole-genome sequencing data. Due to the incompleteness of most sequence assemblies, using genome assembly length as GSE is not

very accurate. Instead, GS can be inferred from sequencing reads directly by analyzing the frequencies of *k*-mers (Li and Waterman, 2003), for which there are various efficient *k*-mer counting tools, like *jellyfish* or *DSK* (Marçais and Kingsford, 2011; Rizk *et al.*, 2013). The amount and average abundance of *k*-mers can then be used to estimate GS; however, accurate estimation of these variables is not trivial. Recently mathematical models have been used to fit the histogram of distinct *k*-mer frequencies including negative binomial distributions or Poisson distributions to infer these variables and predict GS (Liu *et al.*, 2013; Vurture *et al.*, 2017). Though regularly used, there are hardly any studies that verified their predictions.

In this work, we introduce *findGSE*, a sophisticated method for *k*-mer based GSE, which relies on a mixture model by fitting *k*-mer frequencies iteratively using a skew normal distribution (Azzalini, 1985, 2005) to factor in sequencing biases as well as low-frequency *k*-mers that are usually excluded together with erroneous *k*-mers. After systematically comparing *findGSE* with state-of-the-art *k*-mer based GSE methods using simulated data, we used *findGSE* to estimate GS of 89 *Arabidopsis thaliana* accessions for which both whole-genome sequencing data and flow cytometry-based GSE were released. Estimates calculated with *findGSE* were strongly correlated to the flow cytometry-based estimates, thus demonstrating higher capability in capturing GS variation than others. Encouraged by this, we applied *findGSE* to unravel human GS variation. Using previously published sequencing data from 142 human genomes from all around the world, we estimated an average GS of 3, 039 Mb and found a significant difference of 41 Mb between male and female genomes, which is in high agreement with the size differences of the X and Y chromosome reference sequences. Besides gender we found that the highly abundant LINE-1 retrotranspons are the major contributors to present human GS variation and that this is linked to significant differences between populations.

2 Materials and methods

2.1 Algorithm for *findGSE*

The number of *k*-mers in a haploid genome with *G* bases is $G\text{-}k+1$. Assuming that each *k*-mer is sequenced on average with *C* copies (*k*-mer coverage) and *N* denotes the number of genomic *k*-mers in the reads, the relationship $N = C^*(G\text{–}k+1)$ allows to estimate GS with $G \approx N/C$ as $G \gg k$. Both *C* and *N* can be statistically inferred from a *k*-mer frequency histogram (or *k*-mer distribution in short), which summarizes how many distinct *k*-mers occur at a specific frequency within a given whole-genome sequencing data set. Figure 1A shows the shape of a typical *k*-mer distribution of a diploid genome.

The leftmost peak mostly consists of *k*-mers resulting from sequencing errors, which occur often but are at low frequencies as they are only present in one or a few reads. The second (heterozygous) peak and the third (homozygous) peak reflect genomic *k*-mers present in either one or both chromosome sets, which are shared by all reads sampled from the respective loci (except those with sequencing errors). The long tail of the distribution reflects genomic *k*-mers from repetitive elements, which occur at higher frequencies as they are shared by multiple loci. In general, at higher levels of heterozygosity, the heterozygous peak becomes more dominant and shapes the *k*-mer distribution differently (Fig. 1B).

When compared with Poisson or negative binomial distributions, skew normal distributions (Azzalini, 2005), $Y \sim SN(\xi,\ \omega^2,\ \alpha)$, can be more appropriate for fitting *k*-mer frequencies. Both sides of the distribution can be skewed independently. This is of importance if different genomic regions are represented with more reads than other regions, which has been regularly reported for Illumina sequencing, e.g. for GC rich versus GC poor regions (Ossowski *et al.*, 2008).

Given a distribution of *k*-mer frequencies, *findGSE* first fits the distribution iteratively with a skew normal distribution model; then it calculates the total number of *k*-mers (*N*) according to both of the fitted and the original counts and corrects the average *k*-mer coverage (*C*) with the skewness of the fitted curve, based on which it calculates *G* as *N/C* (Algorithm 1). Details of the algorithm are explained with an example in Supplementary Material Sections 1.1 and 1.2. Approximate *k*-mer frequency distribution based on *k*-mer sampling, which is useful in other tasks such as determining optimal *k* for genome assembly (Chikhi and Medvedev, 2014), is not recommended because such methods do not accurately estimate repetitive genomic *k*-mers (Supplementary Fig. S1).

2.2 Read simulation

For evaluation of methods (Section 3.1), we simulated 99 bp Illumina reads from the *A. thaliana* reference sequences using *pIRS* (Hu *et al.*, 2012) by tuning parameters including base coverage,

Algorithm 1. *findGSE*

Input: *k*, and a *k*-mer frequency distribution from whole-genome sequencing data
Output: size of the genome *G*
1: Initialize a vector $residual(0) \leftarrow$ raw *k*-mer counts at different frequencies.
2: Initialize an overall fitting F_o as a null vector with the same size as *residual*(0)
3: **For** *iteration* in 1: *n*
4: Find valley and (the homozygous) peak frequency f_v and f_p according to *residual*(*iteration*-1)
5: Find a set of parameters (ξ, ω, α, *s*), which minimize $\sum_{x=\ fv:\ 2^*fp}(dsnorm(x|\xi,\ \omega,\ \alpha)^*s - residual(iteration\text{-}1,\ x))^2$
6: Set $F(iteration) \leftarrow dsnorm(x|\xi,\ \omega,\ \alpha)^*s$
7: **If** *iteration*==1
8: Set $\alpha_1 \leftarrow \alpha$
9: Set $f_{v1} \leftarrow f_v$
10: **End If**
11: Set $residual(iteration) \leftarrow residual(iteration\text{-}1) - F(iteration)$
12: Update $F_o \leftarrow F_o + F(iteration)$
13: **End for**
14: Set $N \leftarrow \sum_{x=1:fv1}(x^*F_o(x)) + \sum_{x=fv1+1:end}(x^*residual(0, x))$, where *end* is length of raw *k*-mer counting
15: Calculate frequency *e* with α_1 and $\max(F_o)$ (Supplementary Material)
16: For *x* in 1 to *e*, if $x \le f_{v1}$, $H(x) \leftarrow F_o(x)$, else $H(x) \leftarrow residual(0, x)$
17: Calculate *C* as $\sum_{x=1:e}(x^*H(x))/\sum_{x=1:\ e}H(x)$
18: **Return** *G* as *N/C*
(*x*—*k*-mer frequencies; ξ—location parameter mean; ω—scale parameter sd; α—skewness parameter; *s*—density scaling factor; *dsnorm*(.)—function for computing skew normal density. Iteration number *n* can be set as 10 by default.)

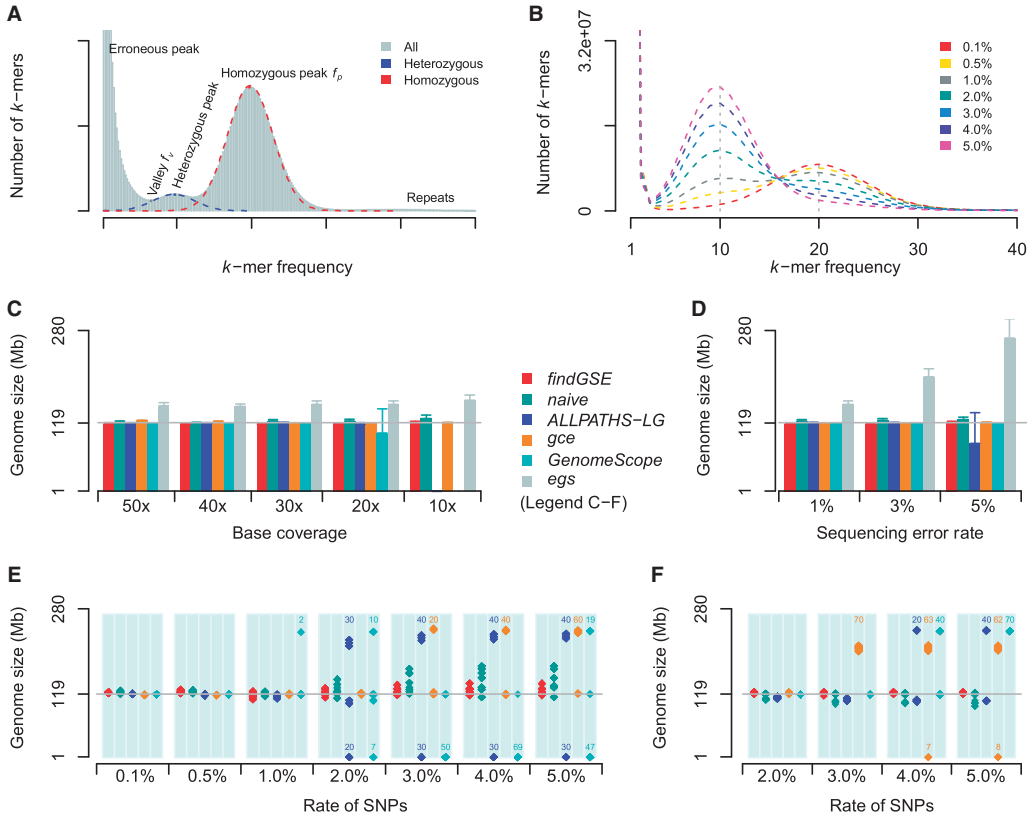

Fig. 1. Comparison of *k*-based GS estimation tools using simulated data. **(A)** Typical *k*-mer frequency histogram of a diploid genome (gray). Dashed lines describe fitted curves that reveal the amount of *k*-mers from heterozygous (left) or homozygous regions (right). **(B)** *k*-mer histograms ($k = 21$) with different levels of heterozygosity (larger rate of SNPs has a higher peak around *k*-mer frequency 10) simulated with 99 bp reads, 1% sequencing error and 30× sequencing coverage (with *in-silico A. thaliana* hybrids). **(C)** GS estimation performance of different tools with varying base coverage (99 bp reads, 1% sequencing error). **(D)** Performance with varying sequencing error (99 bp reads, 30× sequencing coverage). **(E)** Performance with different levels of heterozygosity (99 bp reads, 1% sequencing error, 30× base coverage). **(F)** Performance with different levels of heterozygosity (99 bp reads, 1% sequencing error, 100× base coverage). Note: figures in E and F gives the number of overlapping data points at the respective region. For example, at 2.0% as shown in E, *ALLPATHS-LG* has 30 (20) cases showing doubled (extremely low) GS estimations. For C–F, from left to right: *findGSE, naive, ALLPATHS-LG, gce, GenomeScope* (*egs*)

sequencing error, and rate of SNPs (to build heterozygous diploid genomes).

2.3 Pre-processing of real reads and selecting size of *k*

For evaluation of methods with real reads (Sections 3.2 and 3.3), first, adapters were trimmed with *Skewer* (Jiang *et al.*, 2014) and reads longer than 33 bp were retained. Second, reads duplicated by PCR amplification were filtered out using *FastUniq* (Xu *et al.*, 2012). Third, reads with sequence similarity to mitochondrial, chloroplast or phiX genomes were removed using *BWA* (Li and Durbin, 2009) and *SAMtools* (Li *et al.*, 2009). For the analysis of the seven Col-0-like samples, we used *k* ranging from 15 to 33 at a step of 2 in *k*-mer counting with *jellyfish*. Considering the computational requirement, we used fewer *k*s including 15, 17, 25 and 27 in the analysis of the 89 *A. thaliana* accessions, while we only used one *k* of 21 in analyzing the 142 samples of human (Section 3.4). All the selections guaranteed a *k*-mer frequency ≥10 at the homozygous peaks. If several *k*s were used, final GS was the average of all individual estimates.

3 Results

3.1 Evaluation of *k*-mer based GSE

As for any other *k*-mer analysis, the selection of *k* can be critical, however, it is not obvious what *k* is optimal (Chikhi and Medvedev, 2014). In the first simulation using the yeast reference sequence with a length of ~12 Mb (www.yeastgenome.org), we evaluated the performance of findGSE across a wide range of coverage and *k* values (Supplementary Fig. S2). The results were surprisingly stable against changes in *k*, and even simultaneous changes in sequence coverage could not affect GSE as long as coverage and *k* were reasonably selected. To minimize these already small effects, we decided to perform GSE with a range of *k* values from 21 to 33 at a step of two and use their average predictions as final estimates.

We then compared *findGSE* with five other methods using simulated sets of Illumina reads by varying base coverage, sequencing error rate, and heterozygosity level (Section 2.2). The first method used for our comparison is a naïve implementation for GSE (by this work), which avoids erroneous *k*-mers by excluding all *k*-mers occurring at frequencies smaller than frequency f_v (at the valley between the first and second peak) and uses frequency f_p (at the homozygous peak) as *C*. The second method estimates GS by dividing the total length of reads with the average base coverage that is derived from *k*-mer frequency f_p (script available at https://github.com/josephryan/estimate_genome_size.pl.git, version 0.04, referred to as *egs*). The third method is implemented in the genome assembly tool *ALLPATHS-LG* (version r52488, Gnerre *et al.*, 2011) which estimates GS using *k*-mers occurring with frequencies between f_v and $3f_p/2$ to calculate *C* while discarding both low-frequency ($<f_v$) and extremely high-frequency *k*-mers from *N*. To be mentioned,

tools targeting at other tasks such as selecting optimal size of *k* and genome assembly, namely *SPAdes* (version 3.9.0, Bankevich *et al.*, 2012), *KmerGenie* (version 1.7044, Chikhi and Medvedev, 2014) and *ABySS* (version 2.0.2, Jackman *et al.*, 2017), can infer genome (assembly) size using the total number of unique *k*-mers within the reads. However, as such estimates only consider the collapsed size of repetitive regions, they can be much lower than real GSs, and thus they are not compared. Finally, we also tested two advanced methods, *gce* (version 1.0, Liu *et al.*, 2013) and *GenomeScope* (Vurture *et al.*, 2017), which determine *N* and *C* by fitting Poisson and negative binomial distributions to the *k*-mer distribution, respectively. The basis for read simulation and comparison of these tools was the reference sequence of *A. thaliana* with a length of ~119 Mb. For each of the following comparisons, we further included ten replicates, even though the variation between (*k*s and) replicates was negligibly small anyways (Supplementary Fig. S3A and B). Usages of tools are provided in Supplementary Table S1.

We started by simulating reads with a base coverage of 50× and with a sequencing error of 1% assuming an inbred, homozygous genome (Supplementary Fig. S3C). All tools showed very precise and stable estimates ranging between 118 and 123 Mb around the simulated GS of 119 Mb (Fig. 1C). The only exception was *egs*, which consistently predicted too large GS (of 149 ± 5 Mb) because of counting the total number of bases in the reads without any filtering on sequencing errors. Stepwise reducing base coverage from 50× to 10× revealed that even with 30× all tools (except egs) showed precise estimates. With base coverage of 10×, both *ALLPAHS-LG* and *GenomeScope* failed in returning any meaningful GSE, while *findGSE*, *gce* and the naïve method still estimated reasonable GS. In contrast, increasing sequencing error rates had only marginal effects on GSE (again with the exception of egs). We had to simulate unrealistically high-sequencing error rates until we observed the first effects on GSE. At an error rate of 5%, *ALLPATHS-LG* became unstable as the *k*-mer frequencies were greatly reduced while the estimates of the other tools still remained stable (Fig. 1D; Supplementary Fig. S3D).

To analyze the impact of heterozygosity, we generated *in silico A. thaliana* hybrids with SNP rates from 0.1 to 5.0% and simulated sequencing with 30× coverage including an error rate of 1%. As accurate calculation of the average *k*-mer coverage *C* relies on the correct identification of the homozygous peak, the tools need to distinguish between the homozygous and the heterozygous peaks (Fig. 1A and B). There are two different strategies for this, either by automatic identification of the homozygous peak (as implemented in *ALLPATHS-LG* and *GenomeScope*) or by prior information of an approximated average *k*-mer coverage which guides the selection of the homozygous peak (as implemented in *findGSE, gce* and the naïve method).

For low levels of heterozygosity of 0.1–1.0%, all methods predicted highly accurate estimates (Fig. 1E). However, for higher levels of heterozygosity of 2.0–5.0%, we observed strong discrepancies between the expected and predicted GS. Out of 280 predictions performed with each of the tools, we found 150 and 29 predictions of *ALLPATHS-LG* and *GenomeScope*, respectively that appeared twice as high as expected. In these cases, the automatic peak identification falsely selected the heterozygous peak as the homozygous peak, which led to wrong estimation of *C* and consequently to an inflated haploid GSE. Surprisingly even for *gce*, which requires prior information on the homozygous *k*-mer coverage we also observed 120 of such cases. Moreover, extremely low (or no) GSE were reported in additional 110 and 173 cases of *ALLPATHS-LG* and *GenomeScope*. In these cases, curve fitting around the (again falsely selected) heterozygous peak was hampered by low frequency values, and could not yield any meaning results. Increasing base coverage to 100× could efficiently avoid this complete failure in GSE by both tools; however, it did not eliminate the selection of the wrong peak and thus the doubled estimates in even more cases (Fig. 1F). In contrast, except of the problem of artificially doubling GSE, the results by *GenomeScope* showed lowest variation across all tools with high accuracy.*findGSE* was little affected by increased heterozygosity. In particular at high coverage, *findGSE* was highly accurate even in the presence of 5% heterozygosity predicting an average GS of 121 ± 1 Mb, while at lower coverage this prediction was slightly more variable (125 ± 7 Mb), but still fairly close to the simulated GS of 119 Mb.

Taken together, independent of the actual method *k*-mer based GSE can accurately predict absolute GS at high to moderate base coverage and some tools can even do it at low base coverage when applied to simulated data.

3.2 Reproducibility of *k*-mer based GSE

GSE methods need to be stable against changes in the actual experiments to ensure consistently accurate estimates even with data from different sources. To test for reproducibility of the GSE, we analyzed seven independent whole-genome sequencing data sets, all of which were published for wild type or mutant plants of the *A. thaliana* reference accession Col-0 (Becker *et al.*, 2011; Hartwig *et al.*, 2012; Jiang *et al.*, 2014; Silva-Guzman *et al.*, 2016; Zampini *et al.*, 2015). Even though small genomic differences between 'identical' *A. thaliana* lines selected from different laboratories have been reported (Zapata *et al.*, 2016), we assumed that the changes between these genomes were only of small scale and did not affect GS in any recognizable degree. As these samples were sequenced in different years, the reads were generated with different sequencing chemistry, sequencing depths and lengths (Supplementary Table S2). After adapter trimming and removal of potential PCR duplicates and reads with similarity to mitochondrial, chloroplast or phiX genomes (Section 2.3), GS were estimated with all five tools.

The standard deviations of the seven predictions for each tool were generally low, ranging from 2 Mb for the estimation of *findGSE* up to 5 Mb for the estimation of *gce* (Fig. 2A). This suggested that *k*-mer based GSE methods are very robust against changes in the actual sequencing setup, and that their estimations can be highly reproducible.

Absolute GSEs were different but all average estimates of the different tools ranged between 132 and 139 Mb, which is close to the estimated size of 135 Mb (www.arabidopsis.org). Among these methods, the estimates of *ALLPATHS-LG* were consistently lower (124 Mb on average), which mainly resulted from the removal of low and high frequency *k*-mers leading to the underestimated total number of genomic *k*-mers (Supplementary Fig. S4).

3.3 *k*-mers versus flow cytometry

Recently, the GSs of 165 diverse strains of *A. thaliana* were estimated using flow cytometry and matching whole-genome sequencing data were released (Long *et al.*, 2013; Schmitz *et al.*, 2013). This allowed us to compare flow cytometry and *k*-mer based GSE, while testing for intra-specific GS variation in *A. thaliana*. After preprocessing the sequencing reads (Section 2.3), we selected 89 accessions (Supplementary Table S3) with a base coverage larger than 19x for GSE with *findGSE*, *ALLPATHS-LG*, *GenomeScope* and *gce*.

GS predictions by *k*-mers were independent of base coverage (Fig. 2B) and the standard deviations of *findGSE*, *GenomeScope*

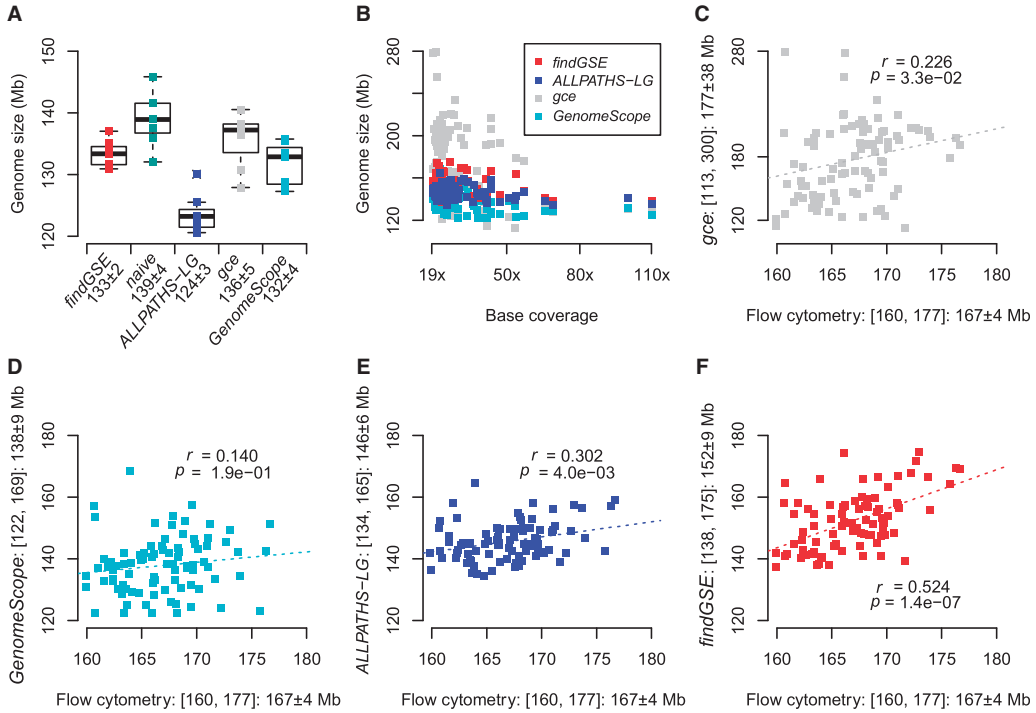

Fig. 2. Comparison of *k*-based GS estimation tools (and flow cytometry) using real sequencing data of *A. thaliana*. **(A)** GSE by different methods on seven different samples of the inbred, homozygous *A. thaliana* reference accession Col-0. **(B)** GSE by different methods on 89 *A. thaliana* accessions, with correlation to sequencing coverage. **(C)** Correlation of GSE on the 89 *A. thaliana* accessions by gce with flow cytometry estimates. **(D)** Correlation of GSE by GenomeScope with flow cytometry estimates. **(E)** Correlation of GSE by *ALLPATHS-LG* with flow cytometry estimates. **(F)** Correlation of GSE by *findGSE* with flow cytometry estimates. For C–F, regression line was drawn, and Pearson's correlation value *r* was given with respective *P* value

and *ALLPATHS-LG* ranged between 6 and 9 Mb (Fig. 2C–F). In contrast, the standard deviation of *gce* was unreasonably large (38 Mb) indicating extreme levels of noise in these estimates. The standard deviation of 4 Mb of the flow cytometry derived GSE was lower than any of the *k*-mer based methods implying that flow cytometry predictions are more stable and less noisy than the *k*-mer based predictions.

To evaluate *k*-mer based predictions, we correlated them with the flow cytometry GS estimates, assuming that a high positive correlation implies high accuracy in capturing real variation in GS. The estimates of *findGSE* showed by far the strongest correlation (Pearson's $r = 0.52$), while the correlations of *gce* (Pearson's $r = 0.23$), *GenomeScope* (Pearson's $r = 0.14$) and *ALLPATHS-LG* (Pearson's $r = 0.30$) appeared to be rather weak (Fig. 2C–F). Though this does not allow on any conclusion of the absolute GS, it reveals a remarkable similarity within the results of *findGSE* and flow cytometry.

Lower variability within the flow cytometry-based predictions as compared with the *k*-mer based predictions does not imply that their absolute GS predictions are better as well. As *k*-mer based predications are independent of any internal standard they have the potential to predict absolute GS independent of any prior assumptions. *findGSE*, *ALLPATHS-LG* and *GenomeScope* predicted an average GS of 152, 146 and 138 Mb across all 89 *A. thaliana* accessions, while *gce* showed unexpectedly large estimates with an average of 177 Mb (and an unreasonable range of 113–300 Mb). In contrast, the flow cytometry based predictions were 167 Mb on average. Overall, we think that the slightly lower *k*-mer based GS predictions of ~150 Mb are more accurate, as flow cytometry-based GS predictions are known for slightly overestimating GS. For example, different flow cytometry based GSE of the reference accessions *A. thaliana* Col-0 include estimates of ~157 and even ~201 Mb (Bennett *et al.* 2003; Schmuths *et al.* 2004), while the reference sequence consortium predicted a GS of only 135 Mb (www.arabidopsis.org).

When compared with the other *k*-mer based tools, the estimates of *findGSE* showed the strongest correlation with the relative variation in the amount of 45 S rDNA, which has recently been identified as the major contributor to GS variation in *A. thaliana* (Long *et al.*, 2013; Rabanal *et al.*, 2017) (Supplementary Table S4; Supplementary Material Section 1.3). This could indicate that the *k*-mer based predictions given by *findGSE* captured more of the true variation in GS (than other methods).

3.4 Estimating and explaining human GS variation

Several recent projects have analyzed human genomes at population scale and released whole-genome sequencing data (Mallick *et al.*, 2016; The 1000 Genomes Project Consortium, 2015). We analyzed the short read data (with base coverage ≥30×) of 142 human individuals (71 male and 71 female) from seven continent-level populations (Mallick *et al.*, 2016) (Supplementary Table S5) and estimated their GS with *findGSE* (Section 2.3).

Haploid human GS was estimated to be 3039 Mb on average, while female genomes were 3059 Mb and male genomes 3018 Mb on average (Fig. 3A). As it is difficult to differentiate chromosomes on *k*-mer level, the haploid GS of a female individual is the sum of the average sizes of each pair of autosomes plus the average size of the two X chromosomes, while the haploid GS of a male individual is the sum of the average sizes of each pair of autosomes plus the average size of the X and Y chromosomes. According to the reference assembly (version GRCh38.p9; The Genome Reference Consortium), female haploid GS would therefore be expected around 3043 Mb and male haploid GS around 2, 994 Mb, implying

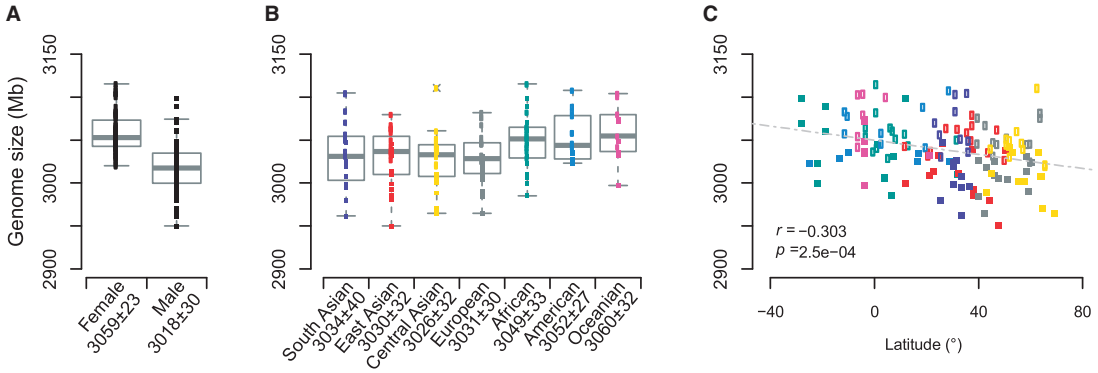

Fig. 3. Human GS variation. **(A)** GSE distribution at the gender level, with 71 female and 71 male genomes. **(B)** GS distributions between populations (males in squares, females in circles). **(C)** GS distribution along latitude (coloring scheme and symbols as in B). Regression line was drawn (dashed gray line), and Pearson's correlation value *r* given with significance value *P*

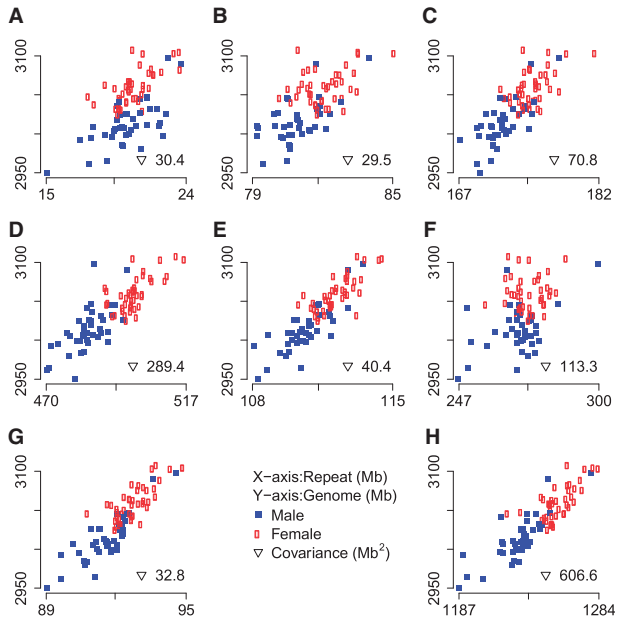

Fig. 4. Covariance analysis of human repeat and GS variations. **(A)** Centromeric repeats (of 41 female and 40 male genomes). **(B)** ERV1 elements. **(C)** ERVL elements. **(D)** LINE-1 elements. LINE-1 elements were the major contributors to human GS variation. **(E)** LINE-2 elements. **(F)** Alu elements. **(G)** MIR elements. **(H)** All repeats in A–G

a 49 Mb size difference between the two sexes. Intriguingly, this is in nearly perfect agreement with estimates by *findGSE* and the estimated 41 Mb size difference between female and male genomes (*t*-test, *P*-value 4.5e-15) implies that GS variations as small as ∼1% of the total GS can be reliably identified.

At the population level, Asians (with Siberians) and Europeans showed smaller GS than Africans, Americans and Oceanians (Fig. 3B; Supplementary Table S6). These differences could also be observed as a significant correlation between GS and latitude, where GS decreased from South to North (Fig. 3C). Though it would be tempting to speculate on environmental factors shaping GS, it would require additional analyses to support such a claim as these patterns could simply arise from population structure without any underlying adaptive pressure.

Absolute GSE of 142 individuals ranged from 2950 to 3115 Mb implying up to 6% size difference in human GS. Even though parts of this variation might come from inaccurate predictions, we speculated that we also captured some of real underlying variation in human GS. To investigate the underlying reasons for these differences, we estimated the percentage of abundant repeats within 41 female and 40 male genomes using the read counts in short read alignments against the reference sequence. rDNA which was the major source for GS variation in *A. thaliana*, only makes up a minor fraction (<0.1%) of human genomes and was not further considered. There were no differences in the relative abundance of repeats between male and female (Supplementary Fig. S5), except for the non-LTR retrotransposons LINE-1 and the endogenous retrovirus L (ERVL) elements where females showed significantly larger ratios probably because these are enriched in the X chromosome (Bailey *et al.*, 2000).

Multiplying relative abundance with GS given by *findGSE* estimates the length of each repeat type in each of the genomes (Supplementary Material Section 1.4). The abundance of the repeat across the genomes revealed the contribution of each of the repeat types to human GS variation (co-variance values in Fig. 4A–G). The LINE-1 retrotransposable elements, which have the highest genome occupancy in the human genome (International Human Genome Sequencing Consortium, 2001), showed by far the largest covariance among all repeats accounting for 29% of the GS variation in human. A combination of all analyzed repeats even accounts for 62% of the variation implying that repetitive elements and in particular the LINE-1 elements are the main source for variation in human GS (Fig. 4H).

4 Conclusion

Counting *k*-mer frequencies within whole-genome sequencing data, which are continuously generated for many organisms anyhow, enables a cheap and elegant way to estimate GS. We have developed an advanced method, *findGSE*, for GS estimation using iterative fitting of *k*-mer frequencies with a skew normal distribution model. A case study on a global collection of 142 human genomes revealed a surprisingly large variation in GS, which follows geographical distributions and is predominately caused by retro-transposons.*k*-mer based GSE using simulated reads recovered absolute GS nearly perfectly while replication with and without changes in error rates and sequence coverage did not affect the predictions of most methods tested here. High levels of heterozygosity challenged some of the tools, which try to automatically deal with heterozygosity, but worked well for *findGSE*. The genomes sizes estimated with seven different sequencing data sets of the same *A. thaliana* background were also very stable.*k*-mer based estimates derived from the

sequencing data of 89 *A. thaliana* accessions were much lower than the flow cytometry based estimates. Though there is no direct evidence, the fact that flow cytometry has to rely on internal standards, which themselves are not accurately measured, suggests that *k*-mer based predictions have the potential to be more accurate in estimating absolute GSs even though individual measurements might be noisier. Nevertheless, there were still notable differences between *k*-mer based estimates by different tools, which indicates that some models might be more accurate than others; however, it is not very clear which one. To improve the reliability of *k*-mer based GSE, several obvious advances could be implemented, like statistically analyzing the precision of GS estimates using confidence intervals, which so far has not been considered by any of the models.

The genomes analyzed in this work are diploid with no or low levels of heterozygosity. For polyploid genomes, which feature a much more complex *k*-mer frequency pattern, a mixture model like implemented in *GenomeScope*, but with a flexible number of distributions that is able to scale with the ploidy, might be more powerful to fit the *k*-mers frequencies. Likewise, within the raw reads of the 89 accessions of *A. thaliana* we found up to 36% of organellar DNA, which can easily confound GS if not filtered out, thus data generation using sterile and nuclei enriched samples could also improve accuracy in GSE.*k*-mer based GSE is still in its infancy, but its application on data, which is usually developed for different purposes anyhow, makes it a cost-effective and widely applicable method. The lack of gold standard genomes with precise GS estimates, however, impedes a final conclusion on how accurate the absolute GSE of *k*-mer based methods are and in particular if they are more accurate than the GS estimates based on flow cytometry. However, independence of any internal standards but direct measurement of the length of DNA, like any of the *k*-mer based methods do, seems to be a promising way for accurate models in the future. In the meantime, careful analysis of GS of the size standard genomes used for flow cytometry with different *k*-mer methods might help to refine their predicted GS and reduce the discrepancies between the methods in future.

Acknowledgements

The authors would like to thank Fernando Rabanal, Quan Long and Magnus Nordborg (Gregor Mendel Institute, Vienna, Austria) for providing the flow cytometry-based GS estimate data of the *A. thaliana* accessions.

Funding

This work was supported by the Max Planck Society postdoctoral fellowship.

Conflict of Interest: none declared.

References

Azzalini,A. (1985) A class of distributions which include the normal ones. *Scand. J. Stat.*, **12**, 171–178.

Azzalini,A. (2005) The skew-normal distribution and related multivariate families. *Scand. J. Stat.*, **32**, 159–188.

Bailey,J.A. *et al.* (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis. *Proc. Natl. Acad. Sci. USA*, **97**, 6634–6639.

Bankevich,A. *et al.* (2012) *SPAdes*: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Becker,C. *et al.* (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, **480**, 245–249.

Bennett,M.D. *et al.* (2003) Comparisons with *Caenorhabditis* (~100Mb) and *Drosophila* (~175Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157Mb and thus ~25% larger than the *Arabidopsis* genome initiative estimate of ~125Mb. *Ann. Botany*, **91**, 547–557.

Chikhi,R. and Medvedev,P. (2014) Informed and automated *k*-mer size selection for genome assembly. *Bioinformatics*, **30**, 31–37.

Doležel,J. and Bartoš,J. (2005) Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Bot.*, **95**, 99–110.

Doležel,J. *et al.* (2003) Nuclear DNA content and genome size of trout and human. *Cytometry*, **51**, 127–128.

Doležel,J. and Greilhuber,J. (2010) Nuclear genome size: are we getting closer? *Cytometry Part A*, **77**, 635–642.

Doležel,J. *et al.* (1998) Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann. Bot.*, **82**, 17–26.

Doležel,J. *et al.* (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.*, **2**, 2233–2244.

Gnerre,S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA*, **108**, 1513–1518.

Gregory,T.R. (2005) Synergy between sequence and size in large-scale genomics. *Nat. Rev. Genet.*, **6**, 699–708.

Gregory,T.R. *et al.* (2007) Eukaryotic genome size databases. *Nucleic Acids Res.*, **35**, D332–D338.

Hardie,D.C. *et al.* (2002) From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis denstometry. *J. Histochem. Cytochem.*, **50**, 735–749.

Hartwig,B. *et al.* (2012) Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol.*, **160**, 591–600.

Hu,X. *et al.* (2012) *pIRS*: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, **28**, 1533–1535.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Jackman,S. *et al.* (2017) *ABySS*2.0: resource-efficient assembly of large genomes using Bloom filter. *Genome Res.*, **27**, 768–777.

Jiang,C. *et al.* (2014) Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Res.*, **24**, 1821–1829.

Jiang,H. *et al.* (2014) *Skewer*: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**, 1–12.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and *SAMtools*. *Bioinformatics*, **25**, 2078–2079.

Liu,B.H. *et al.* 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome project. *arXiv.org* arXiv: 1308.2012.

Li,X. and Waterman,M.S. (2003) Estimating the repeat structure and length of DNA sequences using *l*-tuples. *Genome Res.*, **13**, 1966–1922.

Long,Q. *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.*, **45**, 884–890.

Mallick,S. *et al.* (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, **538**, 201–206.

Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, **27**, 764–770.

Ossowski,S. *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.*, **18**, 2024–2033.

Rabanal,F.A. *et al.* (2017) Unstable inheritance of 45S rRNA genes in *Arabidopsis thaliana*. *G3*, **7**, 1201–1209.

Rizk,G. *et al.* (2013) *DSK*: *k*-mer counting with very low memory usage. *Bioinformatics*, **29**, 652–653.

Schmitz,R.J. *et al.* (2013) Patterns of population epigenomic diversity. *Nature*, **495**, 193–198.

Schmuths,H. *et al.* (2004) Genome size variation among accessions of *Arabidopsis thaliana*. *Ann. Bot.*, **93**, 317–321.

Silva-Guzman,M. *et al.* (2016) Re-evaluation of reportedly metal tolerant *Arabidopsis thaliana* accessions. *PLoS One*, **11**, e0130679.

Soltis,D.E. *et al.* (2003) Evolution of genome size in the angiosperms. *Am. J. Bot.*, **90**, 1596–1603.

The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Vurture,G. *et al*. (2017) *GenomeScope*: Fast reference-free genome profiling from short reads. *Bioinformatics*, **33**, 2202–2204.

Xu,H. *et al*. (2012) *FastUniq*: A fast de novo duplicates removal tool for paired short reads. *PLoS One*, **7**, e52249.

Zampini,É. *et al*. (2015) Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in *Arabidopsis* and humans. *Genome Res*., **25**, 645–654.

Zapata,L. *et al*. (2016) Chromosomal-level assembly of *Arabidopsis thaliana* L*er* reveals the extent of translocation and inversion polymorphisms. *Proc. Natl. Acad. Sci. USA*, **113**, E4052–E4060.