Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

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

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

[1]Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany, [2]Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

1 Supplementary Information

1.1 Fitting *k*-mer frequencies iteratively with a skew normal distribution

To intuitively explain Algorithm 1 (Main text Section 2.1), Fig. S6 gives an example for the iterative fitting process. Initially, all raw *k*-mer counts at different frequencies (*residual*(0)) are used for the first fitting by generating a scaled density $dsnorm(x|\xi, \omega, \alpha)$ that follows the skew normal distribution (Fernández and Steel, 1998) (Https://CRAN.R-project.org/package=fGarch), which requires the optimization of four parameters including location mean ξ, scale standard deviation ω, skewness α and a density scaling factor s. Specifically, frequencies f_v and f_p (at the homozygous peak) is firstly found with *R* function *findpeaks* (library *pracma*), then the mean *k*-mer frequency of *k*-mer counts within f_v to $2*f_p$ is initialized as ξ, with the related standard deviation as ω. Meanwhile, α is initialized as 0.8 (with 1.0 meaning no skewness) and s as 100,000. The subsequent optimization of the four parameters related to skew normal distribution is performed with *R* function *optim* with a least-squares objective function, i.e., the sum of the squares of the differences between the fitted counts in the current iteration and the residual counts from the last iteration within f_v to $2*f_p$. As shown in Fig. S6A, there is a gap between the fitted (in blue) and the raw *k*-mer frequencies (in grey), which is expected mainly because the overall *k*-mer distribution is a summation of distributions of *k*-mer frequencies related to genomic regions with different copy numbers. By subtracting the fitted *k*-mer frequencies ($F(1)$) from raw frequencies (*residual*(0)), we can get a residual *k*-mer distribution (*residual*(1)) which follows a similar distribution to the original one but with a much lower peak and thus can be fitted in the same way (Fig. S6B). We continue the iteration until frequencies of the last residual *k*-mer distribution become so small that they do not show the expected distribution. Finally, we can get an overall fitting (F_o) as the summation of all individual fittings (Fig. S6C).

Overall fitting can help estimate the number of *k*-mers on the left side of the first valley point (f_{v1}), i.e., for those genomic *k*-mers which occur with the same frequencies as erroneous *k*-mers. Meanwhile, as *k*-mers occurring at frequencies larger than f_{v1} are very unlikely to be erroneous (especially for deep sequencing), we choose to use the raw original counts from reads for them. That is, by combing the number of *k*-mers occurring at frequencies lower than f_{v1} according to the fitted counts and that of *k*-mers occurring at frequencies higher than f_{v1} according to the real counts, we get the total number of *k*-mers N, as shown by the red curve in Fig. S6C.

Besides the total number of *k*-mers (N), we need the average *k*-mer coverage (C). If the *k*-mer frequencies follow a standard normal distribution, the peak frequency can be good enough to be selected as C. However, due to potential skewness in *k*-mer frequencies, C tends to shift accordingly, thus considering skewness can improve accuracy in calculating C. For this, we first select an upper bound frequency e in the *k*-mer frequencies according to Equation (1), where α_1 is the skewness of the first fitted curve, $\max(F_o)$ gives the *k*-mer frequency related to the maximum count of the overall fitting, and all the constants are empirical values. Then, we get the unique number of *k*-mers (i.e., $H(x)$) occurring at frequency x for x in $[1, e]$. Specifically, if $x \leq f_{v1}$, $H(x) = F_o(x)$, otherwise $H(x) = residual(0, x)$. Finally, we divide the total number of *k*-mers (i.e., $\sum_{x=1:e}(x*H(x))$) with the total number of unique *k*-mers (i.e., $\sum_{x=1:e}H(x)$) occurring at frequency 1 to e to estimate C, and use it to predict GS with N/C.

$$e = \begin{cases} round(1.5*\max(F_o)) & \alpha_1 < 0.93 \ (left-skew) \\ round(2.5*\max(F_o)) & \alpha_1 > 1.07 \ (right-skew) \\ round(2.0*\max(F_o)) & otherwise \ (no \ skew) \end{cases} \quad (1)$$

1.2 Calculating genome size for heterozygous genomes

findGSE targets homozygous and heterozygous genomes. For diploid heterozygous genomes, fitting the *k*-mer frequencies is slightly

different from that for homozygous ones. First, *findGSE* needs to fit the frequencies related to heterozygous *k*-mers, which requires users provide a rough estimate C_r on C as a guidance, where $C<C_r<2C$. Then, it subtracts the fitted heterozygous frequencies from the raw *k*-mer frequencies, continues to fit the remaining and calculates C in a way similar to that for a homozygous genome. For polyploid genomes, as frequencies of heterozygous and homozygous *k*-mers can overlap in a more complex manner, the current model in *findGSE* cannot guarantee accuracy in estimation.

1.3 Analysis of genome size variation in *A. thaliana*

We performed a multiple regression analysis to correlate the variability of GS estimates with the normalized read counts assigned to 45S and 5S rDNA, centromeric repeats, and transposable elements. In addition to 45S rDNA, there was also a significant association to centromeric repeat variation. Overall, the estimates of *findGSE* (R^2 0.46, *p*-value 2.0e-07) explained nearly as much of the genomic variation as flow cytometry (R^2 0.50, *p*-value 2.2e-08), which outperformed the other *k*-mer based methods (Table S3).

1.4 Calculation of repeat size in human

We considered two methods for calculating the size of a class of repeat. The first method is based on read alignment to the reference genome (hg19), we determined the total number of bases (N_{rb}) that can be aligned to the repetitive regions of interest (annotation database: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/) with 'samtools bedcov' command. Meanwhile, we inferred an average base coverage ($C_b=C*read\_length/(read\_length-k+1)$) from the *k*-mer frequency distribution (of the same read sets), where $k=21$, C is the *k*-mer frequency at the homozygous peak, and *read_length* is a constant of 100 (bp). Then, we calculated the total length of a class of repeat within the genome as N_{rb}/C_b.

The second method is based on the genomic abundance of a specific repeat. Given short read alignments against the reference sequence, the abundance was calculated as the ratio of the read counts (within the repeat), which was obtained by dividing N_{rb} with read length of 100 bp, to the overall aligned read counts. Multiplying relative abundance with GS given by *findGSE* estimated the length of each repeat type in each of the genomes.

The two methods led to the same conclusions as discussed in the main text, such as LINE-1 elements were the major contributors to human GS variation. The report is based on the second method.

Supplementary References

Azzalini A. 2005. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics* 32(2):159-88.

Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* 480(7376):245-9.

Fernández C and Steel MFJ. 1998. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 93(441):359-71.

Hartwig B, James G.V, Konrad K, Schneeberger K, Turck F. 2012. Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiology* 160(2):591-600.

Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP. 2014. Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Res*. 24(11):1821-9.

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjálmsson BJ, Korte A, Nizhynska V et al. 2013. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat Genet*. 45(8):884-90.

Mallick S, Li H, Lipson M, Mathieson L, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538:201-6.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, et al. 2013. Patterns of Population Epigenomic Diversity. *Nature* 495(7440): 193-8.

Silva-Guzman M, Addo-Quaye C, Dilkes BP. 2016. Re-evaluation of reportedly metal tolerant *Arabidopsis thaliana* accessions. *PLoS One* 11(7):e0130679.

Zampini É, Lepage É, Tremblay-Belzile S, Truche S, Brisson N. 2015. Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in *Arabidopsis* and humans. *Genome Res*. 25(5):645-54.

2 Supplementary figures

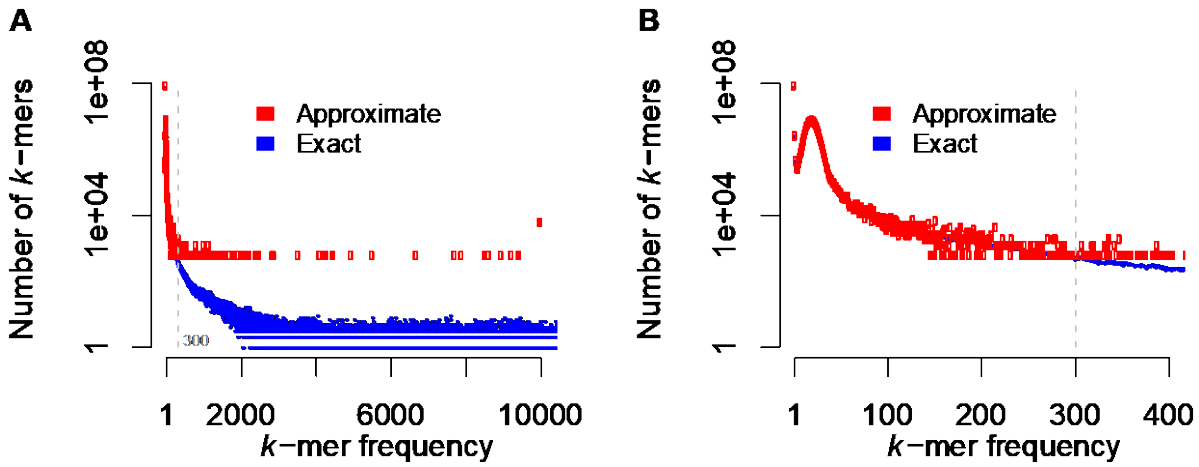

Figure S1. Comparison of exact and approximate *k*-mer frequency distributions (**Y-axis: logarithmic**).

findGSE can accept an approximate *k*-mer frequency distributions, however, exact *k*-mer distributions can lead to higher accuracy in GSE. For example, this figure gives two *k*-mer frequency distributions of the same sample 30-29_2 (one of the seven Col-0 samples to be described in the Main text Section 3.2). The approximate *k*-mer frequency that was recovered from *k*-mer counting by *KmerGenie* with sampling (rate 1/641) shows comparable performance to the exact one by *jellyfish* on capturing the counting information for the unique and less repetitive *k*-mers. Specifically, as shown in Figure S1B, before frequency 300 (vertical dashed line in gray), the exact and the approximate *k*-mer frequency distributions nearly overlap. However, they begin to separate from each other after frequency 300 (Figure S1A), where the approximate *k*-mer frequency distribution has lost accurate counting information on nearly all (more) repetitive *k*-mers. This can result in under-estimated GS, e.g., 128 Mb (approximate) *v.s.* 134 Mb (exact). Due to the above fact, an approximate *k*-mer frequency distribution should particularly not be used in intra-specific variation analysis in GS, because it is common that the variation lies in repetitive elements within the genomes (to be shown by analyses in *A. thaliana* and Human in Main text Section 3.3 & 3.4).

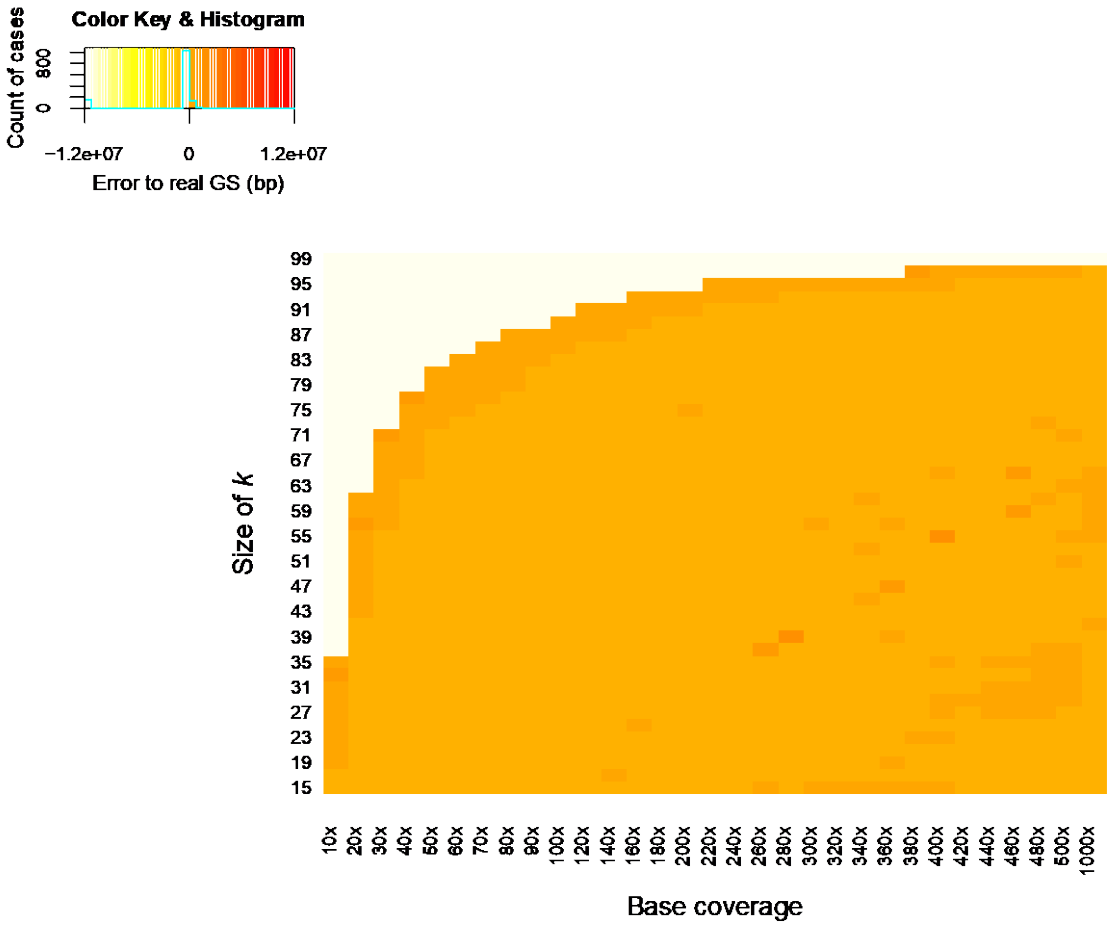

Figure S2. Performance of *findGSE* on varying size *k* and base coverage, with reads simulated from *Yeast* genome (~12Mb). *k* ranged from 15 to 99 at a step of 2, while coverage included 10x to 100x at a step of 10x, 120x to 500x at a step of 20x, and 1000x. For GSE with relatively low base coverage, smaller sizes of *k* should be preferred, while if there is sufficient coverage, a wide range of *k* can be used (though computational requirement should be considered if genome size is large).

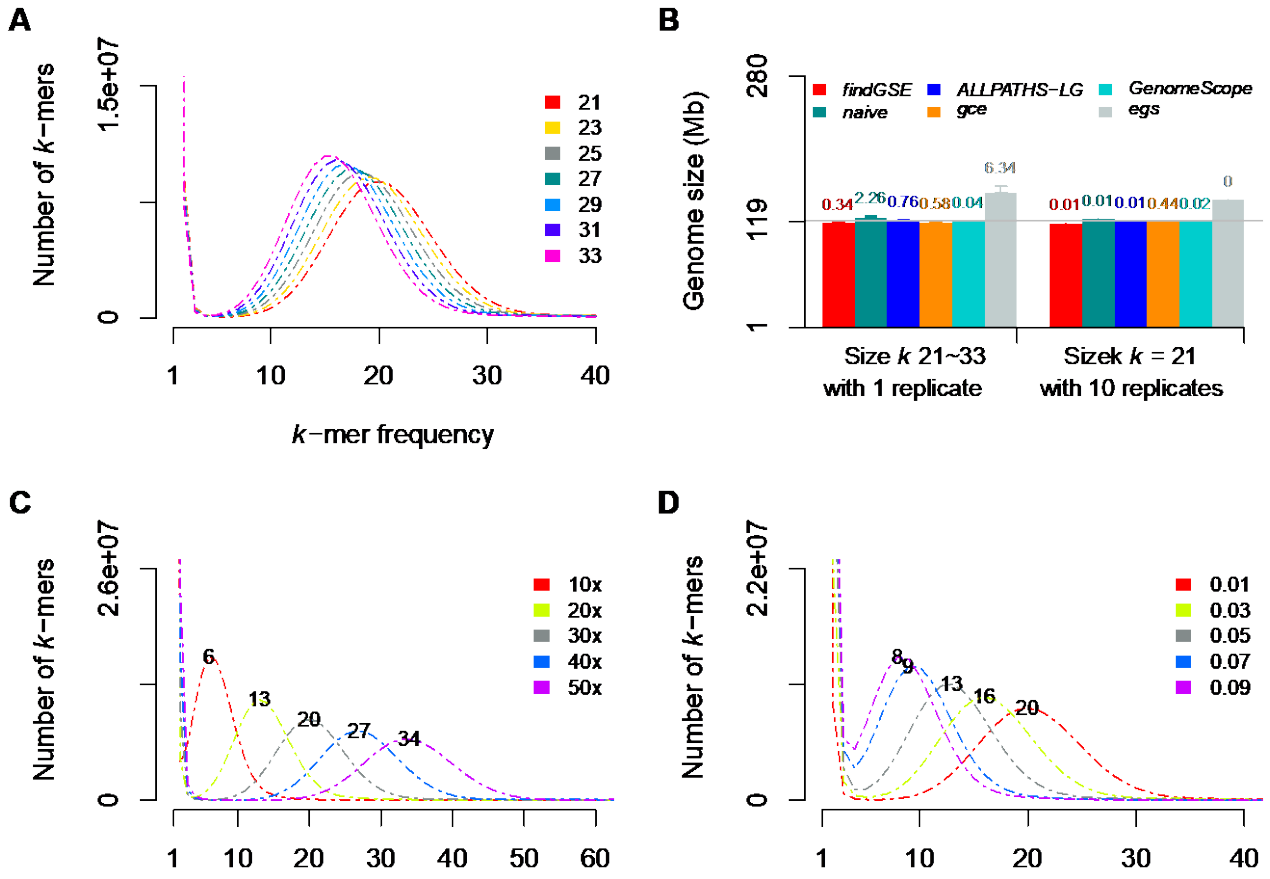

Figure S3. Factors influencing *k*-mer distribution and thus GSE. A) Increasing size of *k* (21 - 33) decreases *k*-mer frequency, with 99 bp reads, 1% sequencing error and 30x base coverage. **B)** Little difference was observed from varying *k*s and replicates. The standard deviations were labeled above bars. **C)** Decreasing base coverage (50x - 10x) decreases *k*-mer frequency, with 99 bp reads and 1% sequencing error. Numbers labeled on curves are frequencies at the peaks. **D)** Increasing sequencing error rate (1% - 9%) decreases *k*-mer frequency, with 99 bp reads and 30x base coverage (expected). The average base coverage derived based on the peak frequency would become smaller than the expectation. For example, for error rate 0.01, the estimated average base coverage was 20*99/(99-21+1)~=25 < 30, thus resulting overestimation in genome size (e.g., by method *egs*). Note: reads were simulated from the *A. thaliana* reference genome; *k*-mer_coverage = base_coverage * (read_length – *k* + 1) / read_length, when all reads are with the same length.

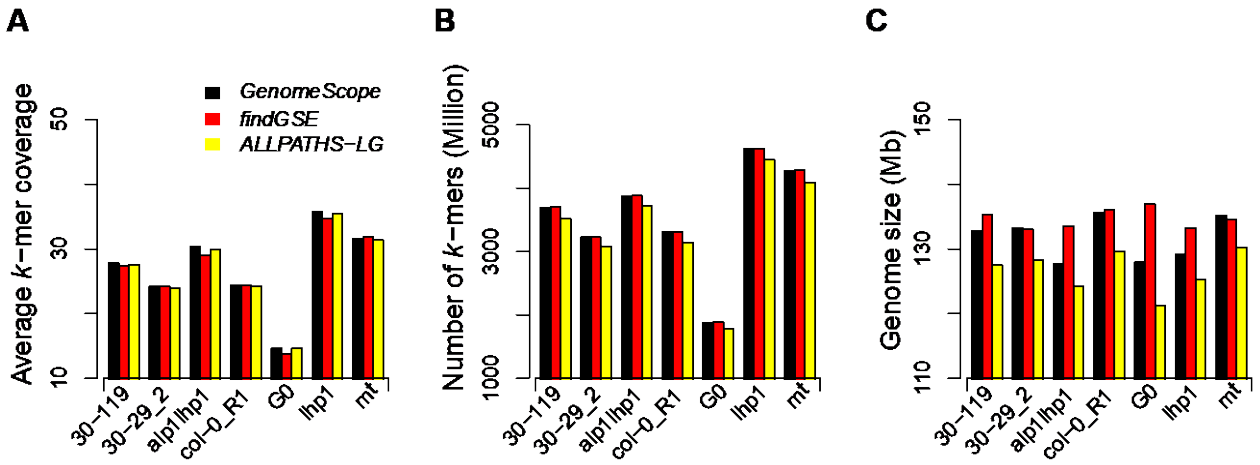

Figure S4. Comparison of *findGSE*, *ALLPATHS-LG* and *GenomeScope* on key parameters influencing genome size estimation with seven Col-0 samples (*k*=21). A) For average *k*-mer coverage, *findGSE* is -0.92~0.45 larger than *ALLPATHS-LG*, while *findGSE* is -1.32~0.24 larger than *GenomeScope*. **B)** For total number of *k*-mers, *findGSE* is 103.9~195.4 M larger than *ALLPATHS-LG*, while *findGSE* is -4.4~11.1 M larger than *GenomeScope*. **C)** For genome size estimate, *findGSE* is 4.3~15.7 Mb larger than *ALLPATHS-LG*, while *findGSE* is -0.7~8.9 Mb larger than *GenomeScope*. As seen, the main difference in estimates of *findGSE* and *ALLPATHS-LG* was due to that *ALLPATHS-LG* underestimated total number of *k*-mers, while that of *findGSE* and *GenomeScope* was mainly due to that *GenomeScope* derived larger estimates on the average *k*-mer coverage.

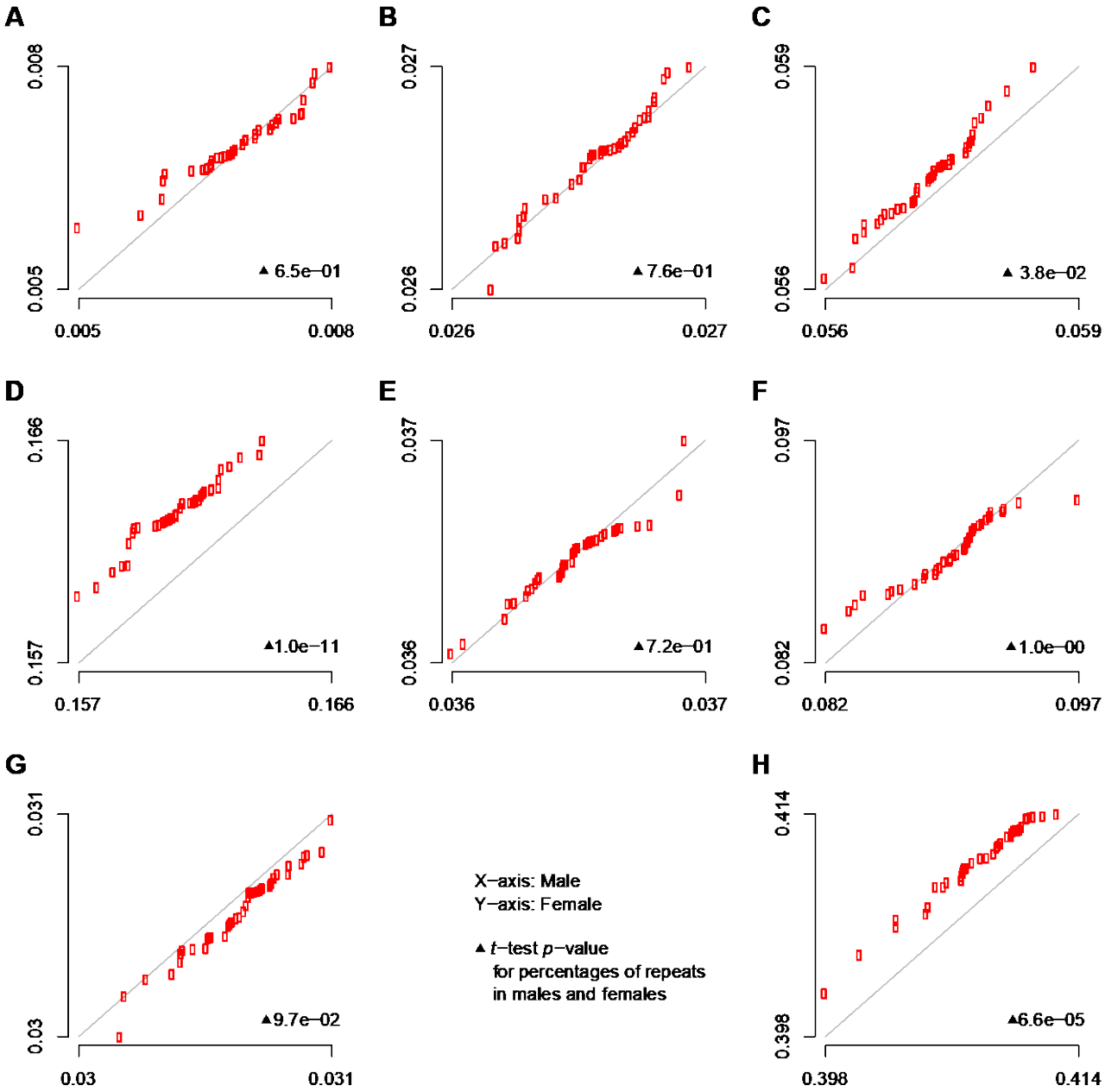

Figure S5. Q-Q plot of percentages of repeats in human female (40) and male (40) genomes. A) Centromeric repeats. **B)** ERV1. **C)** ERVL, larger in female than male. **D)** LINE-1, larger in female than male. **E)** LINE-2. **F)** Alu. **G)** MIR, slightly larger in male than female. **H)** Overall, percentages of repeats in female were larger than male, mainly due to ERVL and LINE-1 elements.

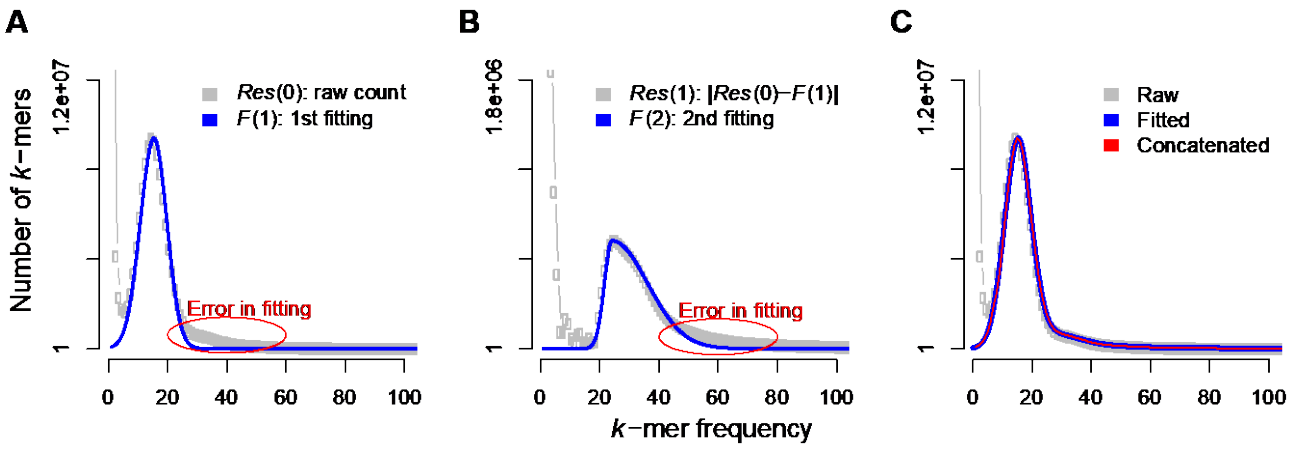

Figure S6. Example for iterative fitting of a *k*-mer distribution. A) First fitting with all initial raw *k*-mer frequencies. **B)** Second fitting with residual *k*-mer frequencies from the first fitting. **C)** Final fitting. *Res*: short for '*residual*'.

3 Supplementary tables

Table S1. Usage of tools (Example data: 99 bp paired-end reads, *21mer_counting.histo* with a homozygous peak at 25)

Tool	Command(s)	File with keyword to extract genome size
egs	*perl estimate_genome_size.pl --kmer=21 --peak=25 --fastq=R1.fq --fastq=R2.fq > egs.log*	In egs.log, after "Estimated genome size"
gce	*gce -f 21mer_counting.histo -M 1024 -D 8 -m 1 >gce.table 2>gce.error* *gce -f 21mer_counting.histo -M 1024 -D 8 -m 1 -H 1 -c 25 >gce.table 2>gce.error* **note: the first and the second are for homozygous and heterozygous respectively*	In gce.error, after "Final estimation table"
ALLPATHS-LG	*FastqToFastbQualb FASTQ=R1.fq.gz OUT_HEAD=R1* *FastqToFastbQualb FASTQ=R2.fq.gz OUT_HEAD=R2* *MergePairedFastbs HEAD1_IN=R1 HEAD2_IN=R2 HEAD_OUT=mergedR12* *FindErrors K=21 HEAD_IN=mergedR12 HEAD_OUT=out PLOIDY_FILE=ploidy VERBOSITY=2 _MM_OUT=out2 > allpaths-lg.log* **note: ploidy is a file containing a single number, e.g. 2, if one wants to get heterozygosity estimation as well.*	In allpaths-lg.log, after the first "Genome size estimate"
GenomeScope	*Rscript genomescope.R 21mer_counting.histo 21 99 out_folder > genomescope.log* **note: automatic detection of heterozygosity*	In genomescope.log, after "len:"
findGSE (naive)	*findGSE(histo="21mer_counting.histo", sizek=21, outdir="out_folder")* *findGSE(histo="21mer_counting.histo", sizek=21, exp_hom=26, outdir="out_folder")* **note: the first and the second are for homozygous and heterozygous respectively*	In v1.94*.txt file, after "sizek_cor2"
SPAdes	*spades.py -o spades -k 21 -1 R1.fq.gz -2 R2.fq.gz > spades.log*	In spades.log, after "Estimated genome size (ignoring repeats)"
ABySS	*abyss-pe name=abyss k=21 in='R1.fq.gz R2.fq.gz' > abyss.log*	In abyss.log, after "The reconstruction is"
KmerGenie	*kmergenie reads.file -l 21 -k 21 -o kmergenie* **note: reads.file contains two lines, namely R1.fq.gz and R2.fq.gz*	In kmergenie_report.html, after "Predicted assembly size"

Table S2. List of samples of *Arabidopsis thaliana* reference accession Col-0.

Sample name	NCBI SRA	Instrument	Read length (bp)	Base coverage	
				Expected	Observed
30-119 (Becker et al., 2011)	ERR420412	Illumina Genome Analyzer II	101	41x	35x
30-29_2 (Becker et al., 2011)	ERR420403			34x	29x
alp1_lhp1 (Hartwig et al., 2012)	(1001genomes)	Illumina Genome Analyzer GAIIx	95	51x	39x
Col-0_R1 (Zampini et al., 2015)	SRR1810274	Illumina HiSeq 2000	100	39x	29x
G0 (Jiang et al., 2014)	SRR1581142		75	27x	18x
lhp1 (Hartwig et al., 2012)	(1001genomes)	Illumina Genome Analyzer GAIIx	95	53x	45x
metal_tolerant (Silva-Guzman et al., 2016)	SRR2032872	Illumina HiSeq 2000	151	45x	35x

Table S3 List of 89 accessions of *Arabidopsis thaliana* (Long et al., 2013; Schmitz et al., 2013), and *k* value used in GSE.

Location	NCBI SRA	*k* in analysis	Location	NCBI SRA	*k* in analysis
SWE	SRR519503	15-17	SWE	SRR519546	15-17
..	SRR519509	15-17	..	SRR519644	15-
..	SRR519512	15-17	..	SRR519506	15-17
..	SRR519538	15-17	..	SRR519598	15-17
..	SRR519536	15-17	..	SRR519599	15-
..	SRR519661	15-17	..	SRR519581	15-17
..	SRR519539	15-17	..	SRR519713	15-17
..	SRR519663	15-17	..	SRR519593	15-17
..	SRR519695	15-17	..	SRR519562	15-17
..	SRR519579	15-17	..	SRR519600	25-27
..	SRR519551	15-17	..	SRR519493	15-17
..	SRR519557	15-	..	SRR519568	15-17
..	SRR519515	15-17	..	SRR519665	15-17
..	SRR519483	15-17	..	SRR519584	15-
..	SRR519563	15-17	..	SRR519596	15-17
..	SRR519654	15-17	..	SRR519561	15-17
..	SRR519507	15-17	..	SRR519602	15-17
..	SRR519659	15-17	..	SRR519601	25-27
..	SRR519666	15-17	..	SRR519500	25-27
..	SRR519529	15-17	..	SRR519636	25-27
..	SRR519574	15-17	RUS	SRR519699	25-27
..	SRR519585	15-	SWE	SRR519629	15-
..	SRR519582	15-17	FRA	SRR492358	25-27
..	SRR519526	15-17	USA	SRR492298	25-27
..	SRR519505	15-17	USA	SRR492366	25-27
..	SRR519650	15-	USA	SRR492418	25-27
..	SRR519542	15-17	TJK	SRR492375	25-27
..	SRR519559	15-17	RUS	SRR492324	25-27
..	SRR519594	15-17	GER	SRR492266	25-27
..	SRR519573	15-17	CZE	SRR492223	25-27
..	SRR519664	15-17	FRA	SRR492200	25-27
..	SRR519681	15-17	USA	SRR492409	25-27
..	SRR519649	15-17	ESP	SRR492388	25-27
..	SRR519560	15-17	USA	SRR492300	25-27
..	SRR519549	15-	GER	SRR492326	25-27
..	SRR519687	25-27	BEL	SRR492205	25-27
..	SRR519620	15-17	GER	SRR492271	25-27
..	SRR519527	15-17	USA	SRR492301	25-27
..	SRR519545	15-17	ESP	SRR492389	25-27
..	SRR519497	15-17	GER	SRR492272	25-27
..	SRR519537	15-17	ESP	SRR492316	25-27
..	SRR519569	15-17	FIN	SRR492384	25-27
..	SRR519583	15-17	IN	SRR492292	25-27
..	SRR519488	15-17	ITA	SRR492322	25-27
..	SRR519595	15-17			

Table S4. Multiple regression analysis with GS and copy number of repeats.

Method	Feature-specific statistics							Overall statistics		
	Feature	DF	Sum of squares	Mean square	*F*	*p*-value	R^2	Total R^2	Adjusted R^2	*p*-value
Flow cytometry	**45S rDNA**	**1**	**334.0**	**334.0**	**45.7**	**6.9e-09**	**0.386**	**0.499**	**0.465**	**2.2e-08**
	5S rDNA	1	2.8	2.8	0.4	5.4e-01	0.008			
	Centromeric	**1**	**76.3**	**76.3**	**10.4**	**2.0e-03**	**0.062**			
	TE	1	16.6	16.6	2.3	1.4e-01	0.044			
	Error	59	431.6	7.3		N.A.				
findGSE	**45S rDNA**	**1**	**1340.3**	**1340.3**	**32.7**	**3.7e-07**	**0.313**	**0.456**	**0.422**	**2.0e-07**
	5S rDNA	1	78.3	78.3	1.9	1.7e-01	0.030			
	Centromeric	**1**	**613.4**	**613.4**	**15.0**	**2.7e-04**	**0.124**			
	TE	1	14.5	14.5	0.4	5.5e-01	-0.007			
	Error	59	2413.2	40.9		N.A.				
allpaths-lg	**45S rDNA**	**1**	**273.2**	**273.2**	**9.6**	**3.0e-03**	**0.132**	**0.255**	**0.204**	**1.4e-03**
	5S rDNA	1	3.3	3.3	0.1	7.3e-01	0.002			
	Centromeric	**1**	**298.5**	**298.5**	**10.5**	**2.0e-03**	**0.119**			
	TE	1	0.4	0.4	0.0	9.1e-01	0.002			
	Error	59	1681.7	28.5		N.A.				
gce	45S rDNA	1	2635.0	2635.0	2.5	1.2e-01	0.037	0.116	0.056	1.2e-01
	5S rDNA	1	2343.0	2343.0	2.2	1.4e-01	0.041			
	Centromeric	1	1510.0	1509.8	1.4	2.4e-01	0.025			
	TE	1	1740.0	1740.4	1.6	2.1e-01	0.013			
	Error	59	62744.0	1063.5		N.A.				
Genome-Scope	45S rDNA	1	2.7	2.7	0.0	8.6e-01	0.001	0.065	0.002	4.0e-01
	5S rDNA	1	63.7	63.7	0.8	3.8e-01	0.011			
	Centromeric	1	241.3	241.3	2.9	9.3e-02	0.051			
	TE	1	31.4	31.4	0.4	5.4e-01	0.001			
	Error	59	4880.0	82.7		N.A.				

*Alpha level: 0.05.

Table S5. List of 142 human samples (EBI:PRJEB9586; Mallick et al., 2016) used in analysis.

Run	Population	Gender	Run	Population	Gender	Run	Population	Gender
ERR1019043	African	male	**ERR1025628**	East-Asian	female	**ERR1419103**	European	female
ERR1019044	African	female	**ERR1019056**	East-Asian	male	**ERR1019082**	European	male
ERR1019077	African	female	**ERR1025627**	East-Asian	male	ERR1019062	European	male
ERR1019078	African	male	**ERR1025610**	East-Asian	male	ERR1346534	European	male
ERR1025600	African	female	**ERR1025598**	East-Asian	male	ERR1347661	European	female
ERR1025601	African	male	**ERR1019038**	East-Asian	male	ERR1347666	European	male
ERR1025602	African	female	ERR1025646	East-Asian	male	ERR1347674	European	male
ERR1025612	African	male	ERR1025647	East-Asian	female	ERR1347691	European	female
ERR1025621	African	male	ERR1347688	East-Asian	male	ERR1419128	European	female
ERR1025640	African	male	ERR1347700	East-Asian	female	ERR1419195	European	female
ERR1025641	African	female	ERR1347709	East-Asian	male	**ERR1019058**	South-Asian	male
ERR1025657	African	female	ERR1395570	East-Asian	male	**ERR1019059**	South-Asian	female
ERR1025658	African	male	ERR1419125	East-Asian	female	**ERR1025649**	South-Asian	female
ERR1025622	African	female	ERR1419178	East-Asian	female	**ERR1025664**	South-Asian	female
ERR1019076	African	male	**ERR1019034**	C-Asian,Siberia	male	**ERR1025634**	South-Asian	female
ERR1025613	African	male	**ERR1025644**	C-Asian,Siberia	male	**ERR1019064**	South-Asian	male
ERR1347660	African	female	**ERR1025645**	C-Asian,Siberia	female	**ERR1025663**	South-Asian	male
ERR1347662	African	female	ERR1347657	C-Asian,Siberia	female	**ERR1025626**	South-Asian	male
ERR1347677	African	female	ERR1347658	C-Asian,Siberia	female	**ERR1019051**	South-Asian	male
ERR1347678	African	female	ERR1347663	C-Asian,Siberia	female	**ERR1025625**	South-Asian	male
ERR1347714	African	male	ERR1347672	C-Asian,Siberia	female	**ERR1025616**	South-Asian	male
ERR1347723	African	male	ERR1347679	C-Asian,Siberia	female	**ERR1025648**	South-Asian	male
ERR1347732	African	male	ERR1347682	C-Asian,Siberia	male	**ERR1019035**	South-Asian	male
ERR1019075	African	male	ERR1347689	C-Asian,Siberia	male	**ERR1019052**	South-Asian	female
ERR1419099	African	female	ERR1347690	C-Asian,Siberia	female	**ERR1019065**	South-Asian	female
ERR1419105	African	female	ERR1347698	C-Asian,Siberia	male	**ERR1019081**	South-Asian	male
ERR1019070	American	female	ERR1347699	C-Asian,Siberia	male	ERR1347676	South-Asian	female
ERR1019071	American	female	ERR1347705	C-Asian,Siberia	male	ERR1395564	South-Asian	female
ERR1025603	American	male	ERR1347706	C-Asian,Siberia	male	ERR1419142	South-Asian	female
ERR1025604	American	female	ERR1347707	C-Asian,Siberia	female	**ERR1019040**	Oceanian	male
ERR1025636	American	female	ERR1347708	C-Asian,Siberia	female	**ERR1019048**	Oceanian	female
ERR1025642	American	female	ERR1347715	C-Asian,Siberia	male	**ERR1019049**	Oceanian	female
ERR1025643	American	female	ERR1347724	C-Asian,Siberia	female	**ERR1019080**	Oceanian	female
ERR1025635	American	male	ERR1347725	C-Asian,Siberia	male	ERR1347685	Oceanian	male
ERR1019066	American	female	ERR1347733	C-Asian,Siberia	female	ERR1347701	Oceanian	male
ERR1019067	American	female	ERR1395586	C-Asian,Siberia	male	ERR1347702	Oceanian	male
ERR1347686	American	male	**ERR1019046**	European	female	ERR1347711	Oceanian	female
ERR1347738	American	male	**ERR1019053**	European	male	ERR1347728	Oceanian	male
ERR1395589	American	male	**ERR1025614**	European	male	ERR1347736	Oceanian	male
ERR1395547	American	male	**ERR1025620**	European	male	ERR1395557	Oceanian	female
ERR1019057	East-Asian	female	**ERR1025630**	European	male	ERR1395580	Oceanian	male
ERR1025618	East-Asian	female	**ERR1025659**	European	female	ERR1419089	Oceanian	female
ERR1025617	East-Asian	male	**ERR1019045**	European	male			
ERR1019039	East-Asian	male	**ERR1025629**	European	female			
ERR1019055	East-Asian	female	**ERR1019069**	European	female			
ERR1019072	East-Asian	female	**ERR1025650**	European	male			
ERR1019074	East-Asian	female	**ERR1025651**	European	female			
ERR1025637	East-Asian	female	**ERR1419102**	European	female			
ERR1025638	East-Asian	male	**ERR1019068**	European	male			
ERR1019061	East-Asian	female	**ERR1025631**	European	female			

Note: a subset of 81 samples with run ids in bold were used in repeat analysis.

Table S6. *p*-values for pair-wise *t*-tests in average genome sizes of seven populations.

Population	South-Asian	East-Asian	Central-Asian Siberia	European	**African**	**American**	**Oceanian**
South-Asian	-	0.7168	0.5207	0.7649	0.1821	0.1234	0.0511
East-Asian	-	-	0.7287	0.9296	**0.0403**	**0.0281**	**0.0119**
Central-Asian Siberia	-	-	-	0.6574	**0.0203**	**0.0148**	**0.0066**
European	-	-	-	-	**0.0419**	**0.0295**	**0.0126**
African	-	-	-	-	-	0.7323	0.3355
American	-	-	-	-	-	-	0.5261
Oceanian	-	-	-	-	-	-	-

*Alpha level: 0.05.