Estimation of Kinship Coefficient in Structured and Admixed Populations using Sparse Sequencing Data

Supplementary Material

Jinzhuang Dou[1,*], Baoluo Sun[1,*], Xueling Sim[2], Jason D Hughes[3], Dermot F Reilly[3], E Shyong Tai[2,4,5], Jianjun Liu[5,6], Chaolong Wang[1,4,§]

[1] Computational and Systems Biology, Genome Institute of Singapore, Singapore, Singapore
[2] Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore
[3] Genetics, Merck Sharp & Dohme Corp., Kenilworth, New Jersey, United States of America
[4] Duke-NUS Medical School, National University of Singapore, Singapore, Singapore
[5] Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[6] Human Genetics, Genome Institute of Singapore, Singapore, Singapore

[*] These authors contribute equally to this work.
[§] Correspondence: wangcl@gis.a-star.edu.sg

S1 Text. Expectations and variances of the SEEKIN estimators

A. Expectations of the SEEKIN estimators

Here we derive the expectations of our proposed kinship estimators at a locus for homogeneous samples (Equations 4 and 5) and for samples with population structure and admixture (Equations 9 and 10). We show that our estimators share the same expectations with those of genotype-based estimators, under the assumptions that the residuals of Equation 1 (defined below) are uncorrelated for any pair of individuals and that the allele frequencies are accurately estimated.

Let $\varepsilon_{im} = \tilde{G}_{im} - \mathrm{E}\left(\tilde{G}_{im} \mid G_{im}, \bar{G}_{Rm}\right)$ be the residue term for individual i at the m^{th} locus. Then the proposed kinship estimator (4) has the expression

$$\begin{aligned}
2\tilde{\phi}_{ijm} &= \frac{\left(\tilde{G}_{im} - 2\tilde{p}_m\right)\left(\tilde{G}_{jm} - 2\tilde{p}_m\right)}{2\tilde{p}_m\left(1-\tilde{p}_m\right)\left(\widehat{r_m^2}\right)^2} = \frac{\left(\tilde{G}_{im} - 2\tilde{p}_m - \varepsilon_{im} + \varepsilon_{im}\right)\left(\tilde{G}_{jm} - 2\tilde{p}_m - \varepsilon_{jm} + \varepsilon_{jm}\right)}{2\tilde{p}_m\left(1-\tilde{p}_m\right)\left(\widehat{r_m^2}\right)^2} \\
&= \frac{\left(\tilde{G}_{im} - 2\tilde{p}_m - \varepsilon_{im}\right)\left(\tilde{G}_{jm} - 2\tilde{p}_m - \varepsilon_{jm}\right) + \varepsilon_{im}\left(\tilde{G}_{jm} - 2\tilde{p}_m - \varepsilon_{jm}\right) + \varepsilon_{jm}\left(\tilde{G}_{im} - 2\tilde{p}_m - \varepsilon_{im}\right) + \varepsilon_{im}\varepsilon_{jm}}{2\tilde{p}_m\left(1-\tilde{p}_m\right)\left(\widehat{r_m^2}\right)^2} \qquad \text{(A.1)} \\
&= \frac{\left(G_{im} - 2p_m\right)\left(G_{jm} - 2p_m\right)}{2\tilde{p}_m\left(1-\tilde{p}_m\right)} + \frac{\widehat{r_m^2}\left(G_{jm} - 2p_m\right)\varepsilon_{im} + \widehat{r_m^2}\left(G_{im} - 2p_m\right)\varepsilon_{jm} + \varepsilon_{im}\varepsilon_{jm}}{2\tilde{p}_m\left(1-\tilde{p}_m\right)\left(\widehat{r_m^2}\right)^2}
\end{aligned}$$

The last equality in (A.1) holds because we can derive from Equations (1) and (2) that

$$\tilde{G}_{im} - 2\tilde{p}_m - \varepsilon_{im} = \mathrm{E}\left(\tilde{G}_{im} \mid G_{im}, \bar{G}_{Rm}\right) - 2\tilde{p}_m \approx \widehat{r_m^2}(G_{im} - 2p_m). \tag{A.2}$$

Because $\mathrm{E}\left(\varepsilon_{im}\right) = 0$, under the assumptions that $\varepsilon_{im} \perp G_{jm}$ and $\varepsilon_{im} \perp \varepsilon_{jm}$ for any $i \neq j$ and that $\tilde{p}_m \approx p_m$, we have

$$\mathrm{E}\left(\tilde{\phi}_{ijm}\right) \approx \mathrm{E}\left[\frac{\left(G_{im} - 2p_m\right)\left(G_{jm} - 2p_m\right)}{2p_m\left(1 - p_m\right)}\right] = \mathrm{E}\left(\hat{\phi}_{ijm}\right), \tag{A.3}$$

where $\hat{\phi}_{ijm}$ is the genotype-based estimator given in Equation (3) [1].

Next, we consider the self-kinship coefficient $\tilde{\phi}_{iim}$ given by Equation (5):

$$\mathrm{E}\left(\tilde{\phi}_{iim}\right) = \frac{\mathrm{E}\left(\tilde{G}_{im} - 2\tilde{p}_m\right)^2}{4\tilde{p}_m\left(1 - \tilde{p}_m\right)\widehat{r_m^2}} = \frac{\mathrm{Var}\left(\tilde{G}_{im}\right)}{4\tilde{p}_m\left(1 - \tilde{p}_m\right)\widehat{r_m^2}} = \frac{\mathrm{Var}\left(G_{im}\right)}{4\tilde{p}_m\left(1 - \tilde{p}_m\right)}. \tag{A.4}$$

The last equality in (A.4) arises from the definition of $\widehat{r_m^2} = \mathrm{Var}\left(\tilde{G}_{im}\right) / \mathrm{Var}\left(G_{im}\right)$ [2]. Under an inbreeding model with inbreeding coefficient f_i, the frequencies of genotypes 0, 1, and 2 are given by $(1 - p_m)^2 + p_m(1 - p_m) f_i$, $2p_m(1 - p_m)(1 - f_i)$, and $p_m^2 + p_m(1 - p_m) f_i$, respectively [3]. The variance of genotypes can be written as $\mathrm{Var}\left(G_{im}\right) = 2p_m(1 - p_m)(1 + f_i)$. Consequently,

$$\mathrm{E}\left(\tilde{\phi}_{iim}\right) = \frac{2p_m(1 - p_m)(1 + f_i)}{4\tilde{p}_m\left(1 - \tilde{p}_m\right)} \approx \frac{1 + f_i}{2}. \tag{A.5}$$

For samples with population structure and admixture, the proposed kinship estimator (9) can be written as

$$\begin{aligned} 2\tilde{\phi}_{ijm} &= \frac{\left(\tilde{G}_{im} - 2\tilde{u}_{im}\right)\left(\tilde{G}_{jm} - 2\tilde{u}_{jm}\right)}{2\sqrt{\hat{p}_{im}\left(1 - \hat{p}_{im}\right)\hat{p}_{jm}\left(1 - \hat{p}_{jm}\right)}\left(\widehat{r_m^2}\right)^2} = \frac{\left(\tilde{G}_{im} - 2\tilde{p}_m - 2\widehat{r_m^2}\left(\hat{p}_{im} - \hat{p}_m\right)\right)\left(\tilde{G}_{jm} - 2\tilde{p}_m - 2\widehat{r_m^2}\left(\hat{p}_{jm} - \hat{p}_m\right)\right)}{2\sqrt{\hat{p}_{im}\left(1 - \hat{p}_{im}\right)\hat{p}_{jm}\left(1 - \hat{p}_{jm}\right)}\left(\widehat{r_m^2}\right)^2} \\ &= \frac{\left[\left(G_{im} - 2p_m\right) - 2\left(\hat{p}_{im} - \hat{p}_m\right) + \varepsilon_{im}\left(\widehat{r_m^2}\right)^{-1}\right]\left[\left(G_{jm} - 2p_m\right) - 2\left(\hat{p}_{jm} - \hat{p}_m\right) + \varepsilon_{jm}\left(\widehat{r_m^2}\right)^{-1}\right]}{2\sqrt{\hat{p}_{im}\left(1 - \hat{p}_{im}\right)\hat{p}_{jm}\left(1 - \hat{p}_{jm}\right)}}, \end{aligned} \tag{A.6}$$

where the last equality follows from (A.2). We assume accurate estimation of individual-specific allele frequencies $\hat{p}_{im} \approx p_{im}$ for $i = 1, 2, \ldots, N$. Further assuming $\varepsilon_{im} \perp G_{jm}$ and $\varepsilon_{im} \perp \varepsilon_{jm}$ for any $i \neq j$, we have

$$\mathrm{E}\left(\tilde{\phi}_{ijm}\right) \approx \mathrm{E}\left[\frac{\left(G_{im} - 2p_{im}\right)\left(G_{jm} - 2p_{jm}\right)}{4\sqrt{p_{im}\left(1 - p_{im}\right)p_{jm}\left(1 - p_{jm}\right)}}\right] = \mathrm{E}\left(\hat{\phi}_{ijm}\right), \tag{A.7}$$

where $\hat{\phi}_{ijm}$ is the genotype-based PC-Relate kinship estimator (Equation 8 with $i \neq j$) and is a consistent estimator of ϕ_{ij} [4].

Finally, we derive the expectation of the self-kinship coefficient estimator by Equation (10), which can be written as

$$\begin{aligned} 2\tilde{\phi}_{iim} &= \frac{\left(\tilde{G}_{im}-2\tilde{u}^{*}_{im}\right)^{2}}{2\hat{p}_{im}\left(1-\hat{p}_{im}\right)\widehat{r_m^2}} = \frac{\left[\tilde{G}_{im}-2\tilde{p}_{m}-2\left(\hat{p}_{im}-\hat{p}_{m}\right)\sqrt{\widehat{r_m^2}}\right]^{2}}{2\hat{p}_{im}\left(1-\hat{p}_{im}\right)\widehat{r_m^2}} \\ &= \frac{\left(\tilde{G}_{im}-2\tilde{p}_{m}\right)^{2}-4\left(\hat{p}_{im}-\hat{p}_{m}\right)\left(\tilde{G}_{im}-2\tilde{p}_{m}\right)\sqrt{\widehat{r_m^2}}+4\left(\hat{p}_{im}-\hat{p}_{m}\right)^{2}\widehat{r_m^2}}{2\hat{p}_{im}\left(1-\hat{p}_{im}\right)\widehat{r_m^2}}. \end{aligned} \tag{A.8}$$

Because $\mathrm{E}\left[\left(\tilde{G}_{im}-2\tilde{p}_{m}\right)^{2}\right]=\widehat{r_m^2}\,\mathrm{E}\left[\left(G_{im}-2p_{m}\right)^{2}\right]$ and $\mathrm{E}\left(\tilde{G}_{im}-2\tilde{p}_{m}\right)=\sqrt{\widehat{r_m^2}}\,\mathrm{E}\left(G_{im}-2p_{m}\right)=0$, we can substitute into the expectation of (A.8) and get

$$\begin{aligned} \mathrm{E}(\tilde{\phi}_{iim}) &= \mathrm{E}\left[\frac{\widehat{r_m^2}\left(G_{im}-2p_{m}\right)^{2}-4\left(\hat{p}_{im}-\hat{p}_{m}\right)\left(G_{im}-2p_{m}\right)\widehat{r_m^2}+4\left(\hat{p}_{im}-\hat{p}_{m}\right)^{2}\widehat{r_m^2}}{4\hat{p}_{im}\left(1-\hat{p}_{im}\right)\widehat{r_m^2}}\right] \\ &= \mathrm{E}\left[\frac{\left(G_{im}-2p_{m}-2\left(\hat{p}_{im}-\hat{p}_{m}\right)\right)^{2}}{4\hat{p}_{im}\left(1-\hat{p}_{im}\right)}\right] \approx \mathrm{E}\left[\frac{\left(G_{im}-2\hat{p}_{im}\right)^{2}}{4\hat{p}_{im}\left(1-\hat{p}_{im}\right)}\right] = \mathrm{E}(\hat{\phi}_{iim}) \end{aligned} \tag{A.9}$$

where $\hat{\phi}_{iim}$ is the genotype-based self-kinship estimator in PC-Relate (Equation 8 with $i=j$) and is a consistent estimator of $(1+f_i)/2$ [4].

B. Variances of the SEEKIN estimators for unrelated pairs

In this section, we derive the variances of our proposed SEEKIN estimators at a single locus for unrelated pairs in homogeneous samples (Equation 4) and in samples with population structure and admixture (Equation 9). We use the results to justify our choice of weight function when combining kinship estimates across loci under the inverse-variance weighting scheme. We do not derive the variances of kinship estimates for related pairs because the derivation is complicated without assuming independence between individuals. In practice, most pairs in a dataset are unrelated, so it is natural to choose a weight function based on unrelated pairs.

We first derive the variance of our SEEKIN estimator for unrelated pairs in a homogeneous population (Equation 4). According to (A.1) with the assumption that $\tilde{p}_m = p_m$, we have

$$\begin{aligned}
\operatorname{Var}\left(2\tilde{\phi}_{ijm}\right) &= \operatorname{Var}\left[\frac{\left(G_{im}-2p_m\right)\left(G_{jm}-2p_m\right)}{2p_m\left(1-p_m\right)}+\frac{\left(G_{jm}-2p_m\right)\varepsilon_{im}}{2p_m\left(1-p_m\right)\widehat{r_m^2}}+\frac{\left(G_{im}-2p_m\right)\varepsilon_{jm}}{2p_m\left(1-p_m\right)\widehat{r_m^2}}+\frac{\varepsilon_{im}\varepsilon_{jm}}{2p_m\left(1-p_m\right)\left(\widehat{r_m^2}\right)^2}\right] \\
&= \operatorname{Var}\left[\left(\frac{G_{im}-2p_m}{\sqrt{2p_m\left(1-p_m\right)}}+\frac{\varepsilon_{im}}{\sqrt{2p_m\left(1-p_m\right)}\widehat{r_m^2}}\right)\left(\frac{G_{jm}-2p_m}{\sqrt{2p_m\left(1-p_m\right)}}+\frac{\varepsilon_{jm}}{\sqrt{2p_m\left(1-p_m\right)}\widehat{r_m^2}}\right)\right] \qquad \text{(B.1)} \\
&= \operatorname{Var}\left(\frac{G_{im}-2p_m}{\sqrt{2p_m\left(1-p_m\right)}}+\frac{\varepsilon_{im}}{\sqrt{2p_m\left(1-p_m\right)}\widehat{r_m^2}}\right)\operatorname{Var}\left(\frac{G_{jm}-2p_m}{\sqrt{2p_m\left(1-p_m\right)}}+\frac{\varepsilon_{jm}}{\sqrt{2p_m\left(1-p_m\right)}\widehat{r_m^2}}\right) \\
&= \left(\frac{\operatorname{Var}\left(G_{im}\right)}{2p_m\left(1-p_m\right)}+\frac{\operatorname{Var}\left(\varepsilon_{im}\right)}{2p_m\left(1-p_m\right)\left(\widehat{r_m^2}\right)^2}\right)\left(\frac{\operatorname{Var}\left(G_{jm}\right)}{2p_m\left(1-p_m\right)}+\frac{\operatorname{Var}\left(\varepsilon_{jm}\right)}{2p_m\left(1-p_m\right)\left(\widehat{r_m^2}\right)^2}\right),
\end{aligned}$$

in which, the last equity holds because $G_{im} \perp \varepsilon_{im}$ under the linear model in Equation (1). The second last equity in (B.1) holds because $\mathrm{E}\left(G_{im}-2p_m\right)=\mathrm{E}\left(\varepsilon_{im}\right)=0$ and $G_{jm} \perp G_{im}, \varepsilon_{im}$ when individuals *i* and *j* are unrelated.

According to (A.2) and because $G_{im} \perp \varepsilon_{im}$, we have

$$\operatorname{Var}\left(\tilde{G}_{im}-2\tilde{p}_m\right)=\left(\widehat{r_m^2}\right)^2 \operatorname{Var}\left(G_{im}-2p_m\right)+\operatorname{Var}\left(\varepsilon_{im}\right). \qquad \text{(B.2)}$$

Because $\widehat{r_m^2}=\operatorname{Var}\left(\tilde{G}_{im}\right)\big/\operatorname{Var}\left(G_{im}\right)$, $\operatorname{Var}\left(\tilde{G}_{im}-2\tilde{p}_m\right)=\operatorname{Var}\left(\tilde{G}_{im}\right)$, and $\operatorname{Var}\left(G_{im}-2p_m\right)=\operatorname{Var}\left(G_{im}\right)$, we have

$$\operatorname{Var}\left(\varepsilon_{im}\right)=\widehat{r_m^2}\left(1-\widehat{r_m^2}\right)\operatorname{Var}\left(G_{im}\right). \qquad \text{(B.3)}$$

Substituting (B.3) into (B.1), the variance of our SEEKIN estimate of kinship between unrelated individuals in a homogeneous population can be written as

$$\operatorname{Var}\left(2\tilde{\phi}_{ijm}\right)=\left(\frac{\operatorname{Var}\left(G_{im}\right)}{2p_m\left(1-p_m\right)\widehat{r_m^2}}\right)\left(\frac{\operatorname{Var}\left(G_{jm}\right)}{2p_m\left(1-p_m\right)\widehat{r_m^2}}\right)=\left(1+f_i\right)\left(1+f_j\right)\left(\widehat{r_m^2}\right)^{-2}, \qquad \text{(B.4)}$$

in which, f_i and f_j are inbreeding coefficients for individuals *i* and *j*, respectively. Under Hardy-Weinberg Equilibrium (HWE), we have $f_i = f_j = 0$ and $\operatorname{Var}\left(2\tilde{\phi}_{ijm}\right)=\left(\widehat{r_m^2}\right)^{-2}$.

The derivation for the variance of our SEEKIN estimator for unrelated pairs in a sample with population structure and admixture (Equation 9) is analogous, except that we assume that individual-specific allele frequencies are accurately estimated: $\hat{p}_{im}=p_{im}$ for $i=1,2,...,N$, and that $\operatorname{Var}\left(G_{im}\right)=\operatorname{Var}\left(G_{im}-2p_{im}\right)=2p_{im}(1-p_{im})(1+f_i)$ under the inbreeding model. Briefly, according to (A.6) and (B.3), we can derive that

$$
\begin{aligned}
\operatorname{Var}\left(2\tilde{\phi}_{ijm}\right) &= \operatorname{Var}\left[\left(\frac{\left(G_{im}-2p_{im}\right)}{\sqrt{2p_{im}\left(1-p_{im}\right)}}+\frac{\varepsilon_{im}}{\widehat{r_m^2}\sqrt{2p_{im}\left(1-p_{im}\right)}}\right)\left(\frac{\left(G_{jm}-2p_{jm}\right)}{\sqrt{2p_{jm}\left(1-p_{jm}\right)}}+\frac{\varepsilon_{jm}}{\widehat{r_m^2}\sqrt{2p_{jm}\left(1-p_{jm}\right)}}\right)\right] \\
&= \operatorname{Var}\left(\frac{G_{im}-2p_{im}}{\sqrt{2p_{im}\left(1-p_{im}\right)}}+\frac{\varepsilon_{im}}{\widehat{r_m^2}\sqrt{2p_{im}\left(1-p_{im}\right)}}\right)\operatorname{Var}\left(\frac{G_{jm}-2p_{jm}}{\sqrt{2p_{jm}\left(1-p_{jm}\right)}}+\frac{\varepsilon_{jm}}{\widehat{r_m^2}\sqrt{2p_{jm}\left(1-p_{jm}\right)}}\right) \\
&= \left(\frac{\operatorname{Var}\left(G_{im}\right)}{2p_{im}\left(1-p_{im}\right)}+\frac{\operatorname{Var}\left(\varepsilon_{im}\right)}{\left(\widehat{r_m^2}\right)^2 2p_{im}\left(1-p_{im}\right)}\right)\left(\frac{\operatorname{Var}\left(G_{jm}\right)}{2p_{jm}\left(1-p_{jm}\right)}+\frac{\operatorname{Var}\left(\varepsilon_{jm}\right)}{\left(\widehat{r_m^2}\right)^2 2p_{jm}\left(1-p_{jm}\right)}\right) \\
&= \left(\frac{\operatorname{Var}\left(G_{im}\right)}{2p_{im}\left(1-p_{im}\right)\widehat{r_m^2}}\right)\left(\frac{\operatorname{Var}\left(G_{jm}\right)}{2p_{jm}\left(1-p_{jm}\right)\widehat{r_m^2}}\right)=\left(1+f_i\right)\left(1+f_j\right)\left(\widehat{r_m^2}\right)^{-2}.
\end{aligned} \tag{B.5}
$$

Supplementary References

1. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.
2. Hu YJ, Li Y, Auer PL, Lin DY (2015) Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations. Proc Natl Acad Sci U S A 112: 1019-1024.
3. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet 10: 639-650.
4. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, et al. (2012) Estimating kinship in admixed populations. Am J Hum Genet 91: 122-138.

Supplementary Tables

S1 Table. Performance of relationship classification based on homogeneous kinship estimators in ~0.15X sequencing data of 254 Chinese.

Call set	Method	3rd degree		2nd degree		PO/FS	
		Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Bcftools	lcMLkin	0.009	1.000*	0.950*	1.000*	1.000*	1.000*
	GCTA	0.027*	0.045	0.000	0.000	--	0.000
	KING	0.001	0.864	--	0.000	--	0.000
BEAGLE	SEEKIN	0.941*	0.727*	0.946*	0.972*	1.000*	0.986*
	GCTA	0.029	0.045	0.022	0.083	1.000*	0.082
	KING	0.056	0.091	0.020	0.083	1.000*	0.014
BEAGLE+1KG3	SEEKIN	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
	GCTA	0.800	0.727	0.800	0.889	1.000*	0.945
	KING	0.824	0.636	0.846	0.917	1.000*	0.959

Precision is defined as the proportion of correct classification among all pairs of a relationship type inferred from the sequence-based kinship estimates. Sensitivity is defined as the proportion of correct classification among pairs of a relationship type inferred from the gold standard kinship estimates.

* Highest values of precision or sensitivity in each call set and each relationship type.

S2 Table. Performance of homogeneous kinship estimators in ~0.75X sequencing data of 254 Chinese.

Call set	Method	Unrelated (31,925 pairs)		3rd degree (22 pairs)		2nd degree (36 pairs)		PO/FS (146 pairs)		Self-kinship (254 individuals)	
		RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
Bcftools	lcMLkin	0.054	0.054	0.034	0.034	0.015*	0.012*	0.014*	-0.012*	--	--
	GCTA	0.005*	-0.004*	0.033	-0.033	0.065	-0.065	0.124	-0.124	0.048	-0.045
	KING	0.051	0.051	0.019*	-0.016*	0.090	-0.089	0.224	-0.223	--	--
BEAGLE	SEEKIN	0.006	-0.004	0.009*	-0.007*	0.013*	-0.011*	0.018*	-0.014*	0.031	-0.003*
	GCTA	0.005*	-0.003*	0.020	-0.019	0.036	-0.035	0.065	-0.064	0.019	0.015
	KING	0.014	-0.011	0.028	-0.027	0.038	-0.038	0.070	-0.070	--	--
BEAGLE +1KG3	SEEKIN	0.004*	-0.004	0.004*	-0.001*	0.005*	-0.004*	0.007*	-0.005*	0.018	-0.011*
	GCTA	0.004*	-0.003*	0.009	-0.008	0.014	-0.013	0.022	-0.022	0.015*	-0.014
	KING	0.005	0.004	0.005	-0.004	0.006	-0.005	0.014	-0.013	--	--

RMSE is the root mean squared error and BIAS is defined as the mean difference to the array-based estimates from PC-Relate for each type of relatedness. Negative values of BIAS suggest underestimation for results based on sparse sequencing data and vice versa.

* Smallest magnitude of RMSE or BIAS in each call set and each type of relatedness

S3 Table. Performance of relationship classification based on homogeneous kinship estimators in ~0.75X sequencing data of 254 Chinese.

Call set	Method	3rd degree		2nd degree		PO/FS	
		Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Bcftools	lcMLkin	0.001	0.773*	0.844*	1.000*	1.000*	1.000*
	GCTA	0.000	0.000	0.000	0.000	--	0.000
	KING	0.000	0.636	--	0.000	--	0.000
BEAGLE	SEEKIN	0.950*	0.864*	0.972*	0.972*	1.000*	0.993*
	GCTA	0.348	0.364	0.438	0.583	1.000*	0.815
	KING	0.158	0.136	0.357	0.556	1.000*	0.747
BEAGLE+1KG3	SEEKIN	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
	GCTA	0.950	0.864	1.000*	0.972	1.000*	1.000*
	KING	1.000*	0.955	1.000*	1.000	1.000*	1.000*

Precision is defined as the proportion of correct classification among all pairs of a relationship type inferred from the sequence-based kinship estimates. Sensitivity is defined as the proportion of correct classification among pairs of a relationship type inferred from the gold standard kinship estimates.

* Highest values of precision or sensitivity in each call set and each relationship type.

S4 Table. Performance of heterogeneous kinship estimators in ~0.75X sequencing data of 762 Chinese and Malays.

Call set	Method	Unrelated (289,205 pairs)		3rd degree (148 pairs)		2nd degree (147 pairs)		PO/FS (437 pairs)		Self-kinship (762 individuals)	
		RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
BEAGLE	SEEKIN	0.005	-0.002	0.008*	0.004*	0.010*	0.006*	0.021*	0.015*	0.045	0.022
	PC-Relate	0.003*	0.000*	0.013	-0.012	0.025	-0.024	0.048	-0.047	0.035	0.026
	REAP	0.003*	-0.001	0.015	-0.014	0.028	-0.028	0.052	-0.051	0.030*	0.017*
	RelateAdmix	0.003*	0.002	0.016	-0.015	0.030	-0.029	0.057	-0.056	--	--
BEAGLE +1KG3	SEEKIN	0.003	-0.001	0.003*	0.001*	0.004*	0.001*	0.007*	0.003*	0.018	0.002*
	PC-Relate	0.002*	0.000*	0.005	-0.004	0.010	-0.009	0.020	-0.020	0.017*	-0.014
	REAP	0.002*	-0.001	0.008	-0.008	0.014	-0.014	0.024	-0.023	0.018	-0.017
	RelateAdmix	0.002*	0.001	0.005	-0.005	0.009	-0.009	0.016	-0.016	--	--

RMSE is the root mean squared error and BIAS is defined as the mean difference to the array-based estimates from PC-Relate for each type of relatedness. Negative values of BIAS suggest underestimation for results based on sparse sequencing data and vice versa.

* Smallest magnitude of RMSE or BIAS in each call set and each type of relatedness.

S5 Table. Performance of relationship classification based on heterogeneous kinship estimators in ~0.15X sequencing data of 762 Chinese and Malays.

Call set	Method	3rd degree		2nd degree		PO/FS	
		Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
BEAGLE	SEEKIN	0.934*	0.858*	0.916*	0.959*	1.000*	0.986*
	PC-Relate	0.389	0.392	0.161	0.388	1.000*	0.325
	REAP	0.302	0.331	0.090	0.238	1.000*	0.195
	RelateAdmix	0.335	0.351	0.113	0.306	1.000*	0.188
BEAGLE+1KG3	SEEKIN	0.966*	0.946*	0.954*	0.993*	1.000*	1.000*
	PC-Relate	0.867	0.791	0.866	0.878	1.000*	0.954
	REAP	0.758	0.635	0.796	0.796	1.000*	0.931
	RelateAdmix	0.836	0.723	0.900	0.857	1.000*	0.968

Precision is defined as the proportion of correct classification among all pairs of a relationship type inferred from the sequence-based kinship estimates. Sensitivity is defined as the proportion of correct classification among pairs of a relationship type inferred from the gold standard kinship estimates.

[*] Highest values of precision or sensitivity in each call set and each relationship type.

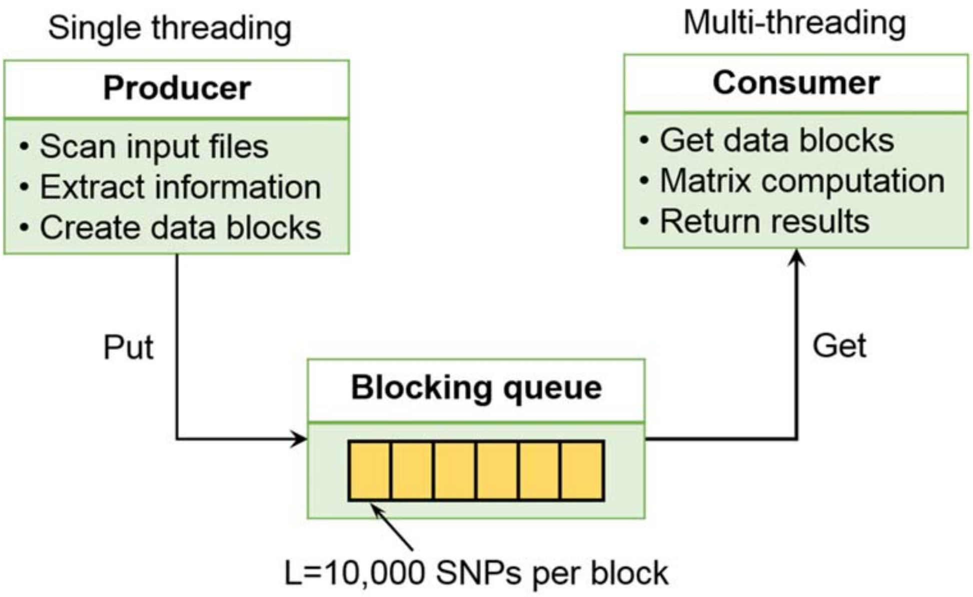
S6 Table. Performance of relationship classification based on heterogeneous kinship estimators in ~0.75X sequencing data of 762 Chinese and Malays.

Call set	Method	3rd degree		2nd degree		PO/FS	
		Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
BEAGLE	SEEKIN	0.964*	0.905*	0.935*	0.980*	1.000*	1.000*
	PC-Relate	0.836	0.723	0.882	0.864	1.000*	0.961
	REAP	0.778	0.615	0.823	0.823	1.000*	0.941
	RelateAdmix	0.736	0.622	0.814	0.776	1.000*	0.941
BEAGLE+1KG3	SEEKIN	0.980*	0.980*	1.000*	0.993*	1.000*	1.000*
	PC-Relate	0.959	0.953	1.000*	0.966	1.000*	0.998
	REAP	0.948	0.858	0.986	0.952	1.000*	0.995
	RelateAdmix	0.965	0.919	1.000*	0.966	1.000*	1.000*

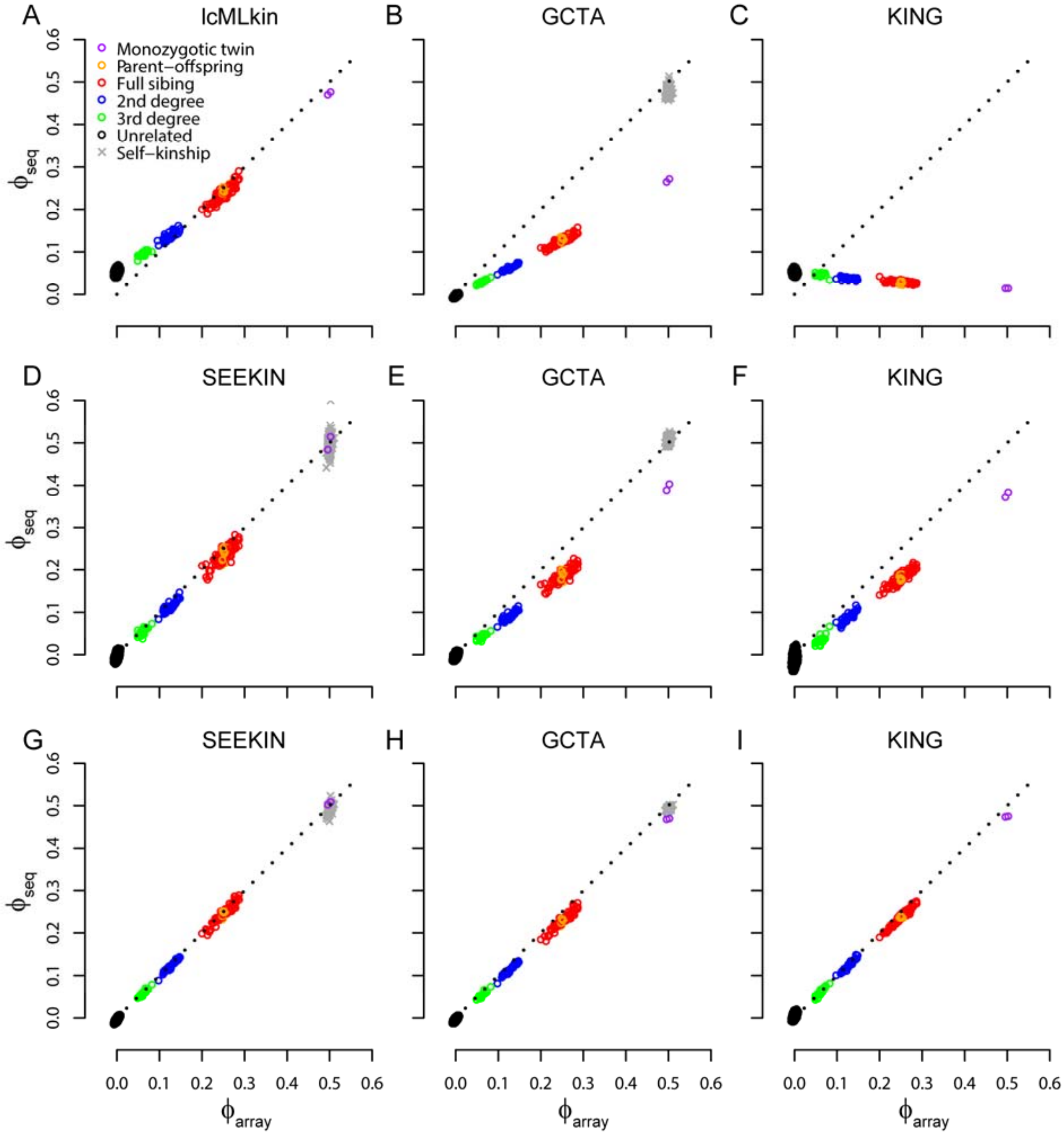
Precision is defined as the proportion of correct classification among all pairs of a relationship type inferred from the sequence-based kinship estimates. Sensitivity is defined as the proportion of correct classification among pairs of a relationship type inferred from the gold standard kinship estimates.

[*] Highest values of precision or sensitivity in each call set and each relationship type.

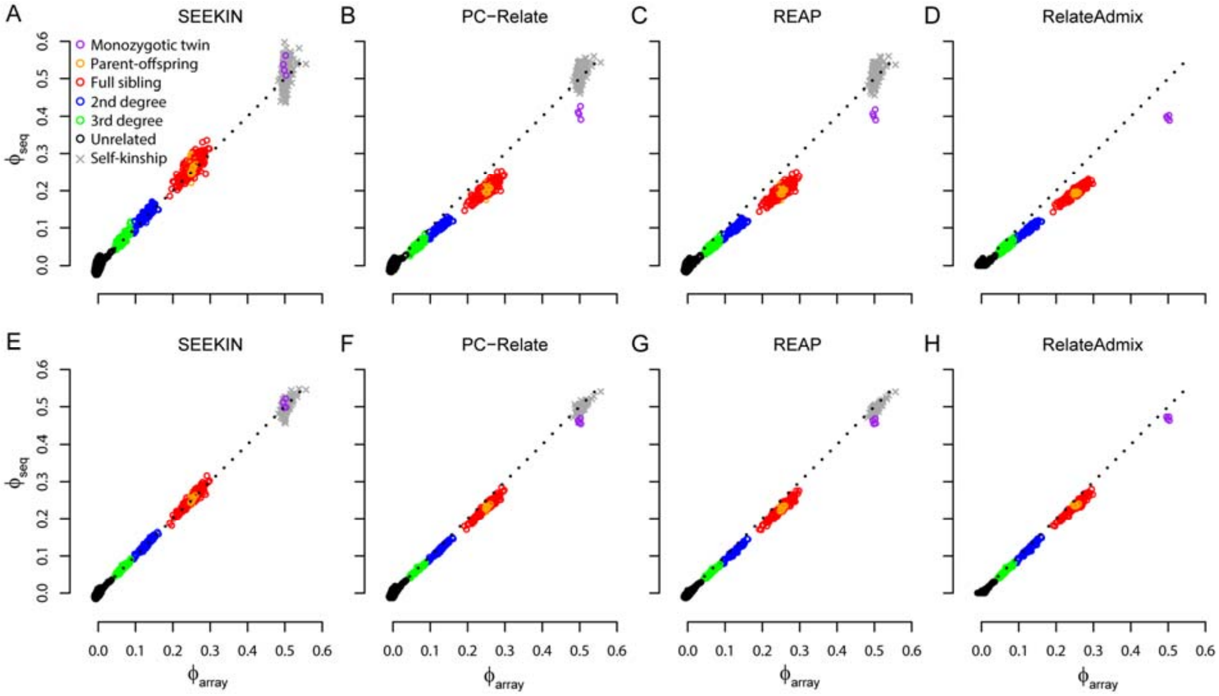
Supplementary Figures



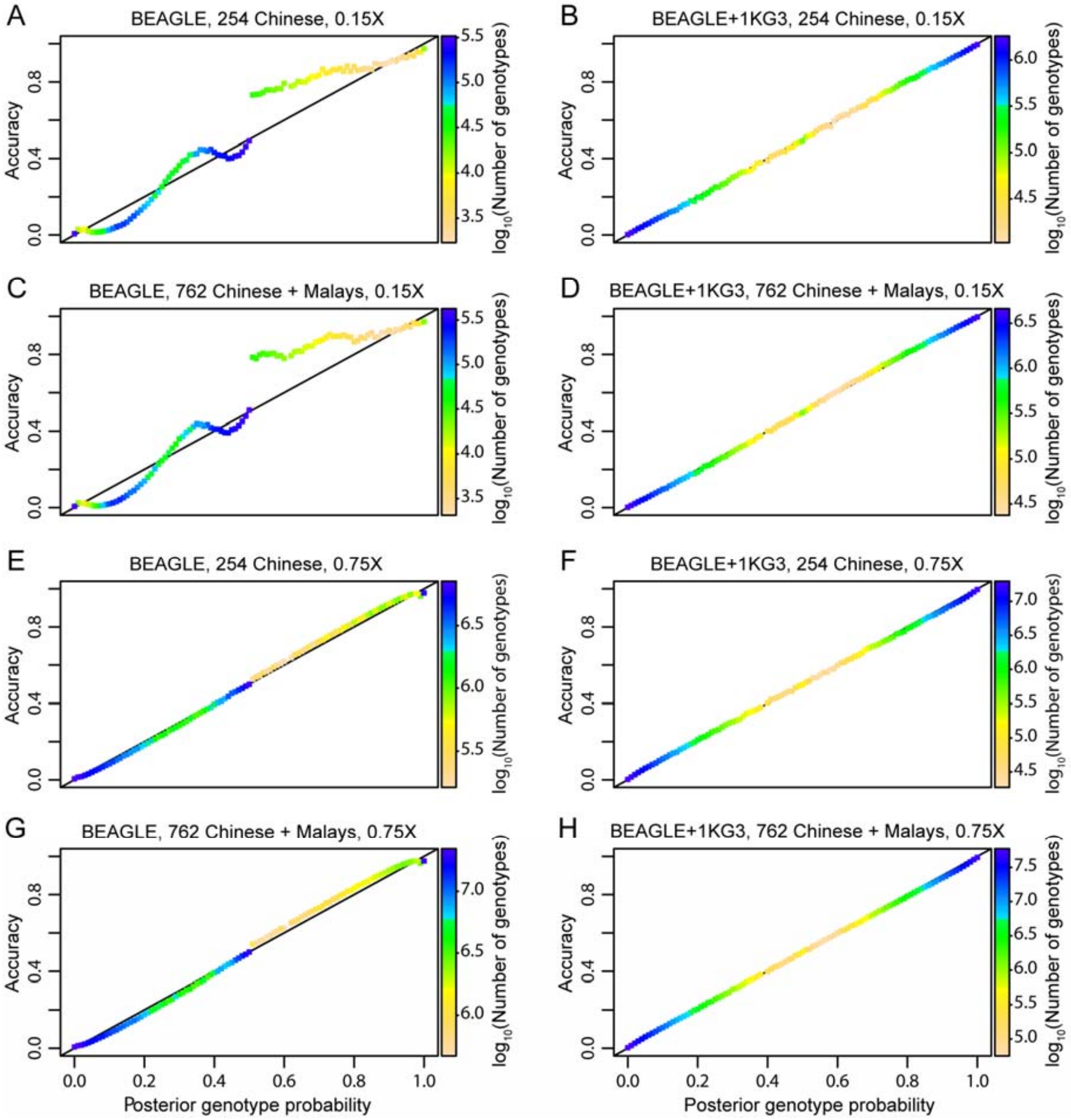
S1 Fig. Illustration of the "single producer/consumer" design in the SEEKIN software. A single-threading "producer" job scans the input files, extracts required information for each SNP, and packs into a data block for every *L* SNPs. These data blocks are stored in the buffer, labeled as the blocking queue. Concurrently, a "consumer" job takes the data blocks one by one, performs multi-threading computation, and returns results. The results from different blocks are automatically combined after all blocks are analyzed. The "producer" and the "consumer" are synchronized through the blocking queue; the "producer" will become inactive if the blocking queue is full, and the "consumer" will become inactive if the blocking queue is empty. The best performance is achieved when production and consumption are balanced (i.e., the blocking queue is neither full nor empty).



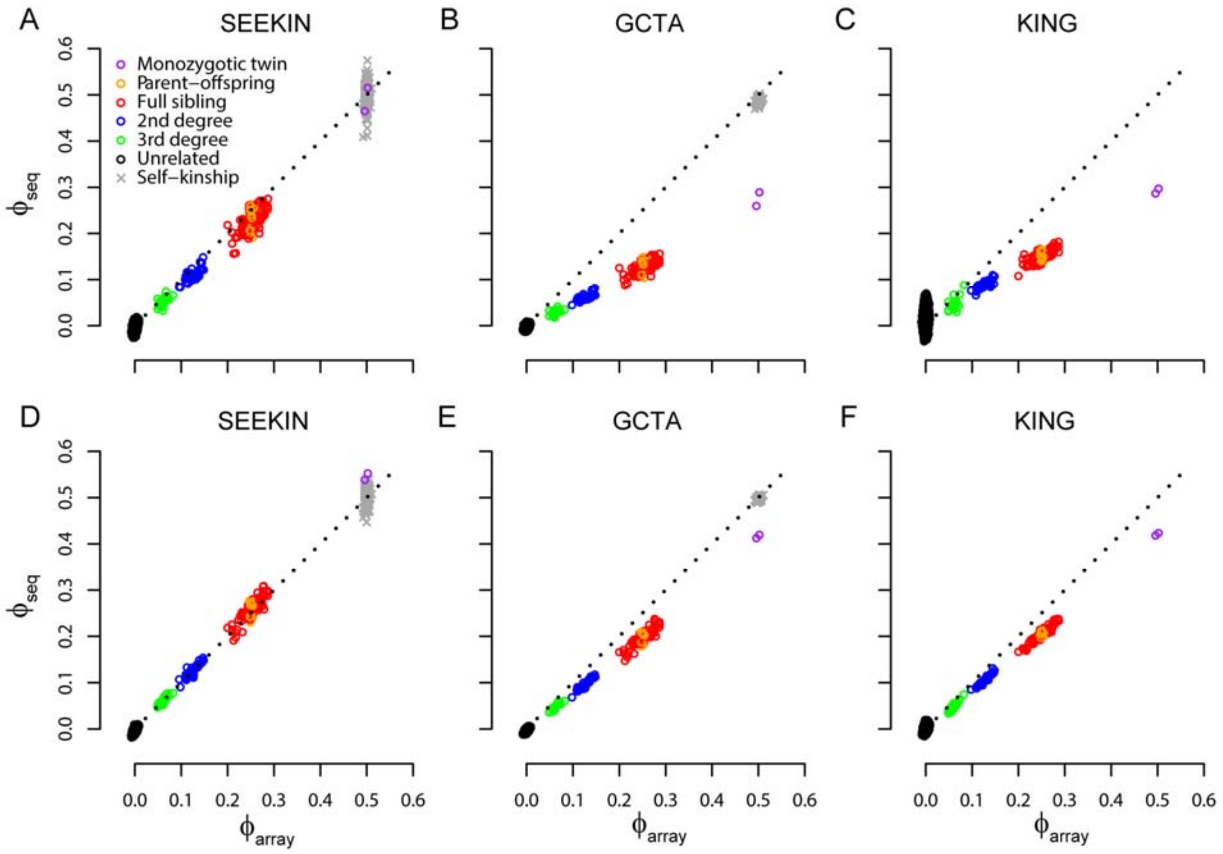
S2 Fig. Performance of homogeneous kinship estimators in ~0.75X sequencing data of 254 Chinese. In each panel, we compared sequence-based estimates (ϕ_{seq}, y-axis) with the array-based estimates from PC-Relate (ϕ_{array}, x-axis). Colored circles represent kinship coefficients between two individuals and different types of relatedness were determined in Figure 2. Grey crosses represent self-kinship coefficients. We evaluated lcMLkin (A), GCTA (B, E, H), KING (C, F, I), and SEEKIN (D, G) using the bcftools call set (A-C), the BEAGLE call set (D-F), and the BEAGLE+1KG3 call set (G-I). Note that KING does not estimate self-kinship coefficients.



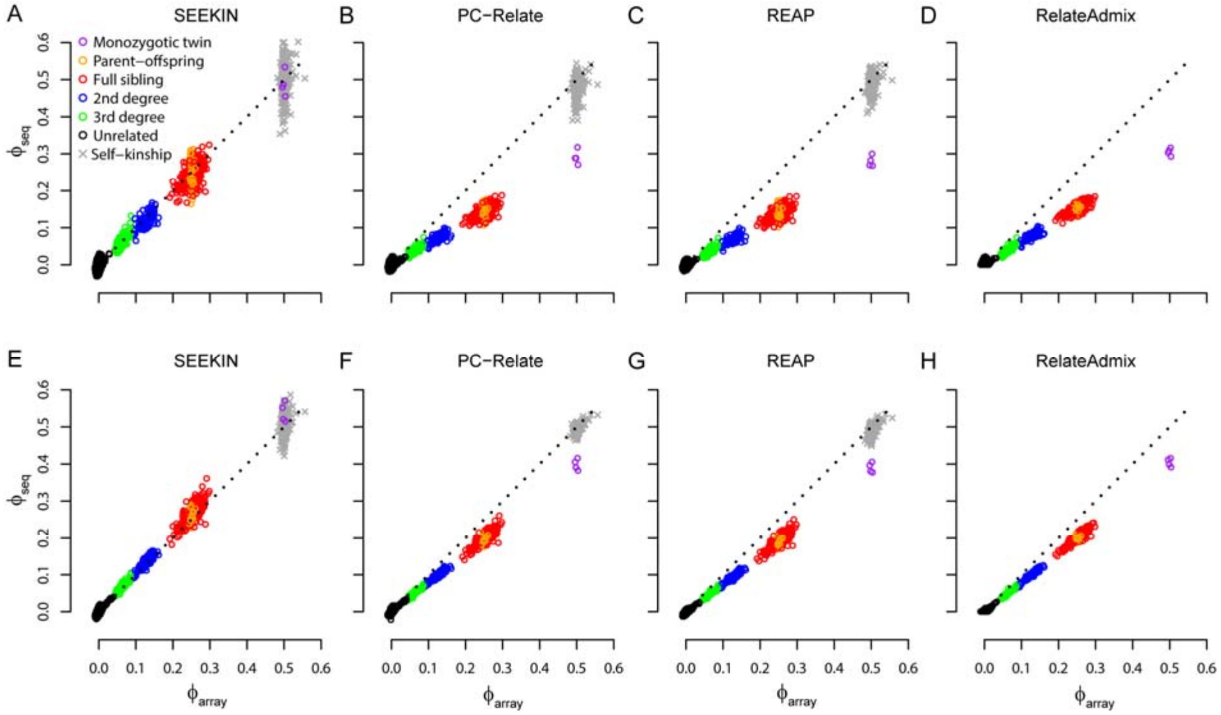
S3 Fig. Performance of heterogeneous kinship estimators in ~0.75X sequencing data of 762 Chinese and Malays. In each panel, we compared sequence-based estimates (ϕ_{seq}, y-axis) with the array-based estimates from PC-Relate (ϕ_{array}, x-axis). Colored circles represent kinship coefficients between two individuals and different types of relatedness were determined in Figure 2. Grey crosses represent self-kinship coefficients. We evaluated SEEKIN (A, E), PC-Relate (B, F), REAP (C, G), and RelateAdmix (D, H) using the BEAGLE call set (A-D), and the BEAGLE+1KG3 call set (E-H). We only included SNPs overlapping with the SGVP dataset in the analyses, because we used the SGVP dataset as the reference panel to estimate individual-specific allele frequencies for SEEKIN, REAP and RelateAdmix.



S4 Fig. Calibration of posterior genotype probabilities from BEAGLE for different sequencing datasets. For each dataset, we binned the genotype probabilities into 100 bins spaced by 0.01 from 0 to 1 (x-axis). For each bin, we calculated the proportion of correct genotypes by comparing to the array genotypes (y-axis). The number of genotypes in each bin is color-coded according to the logarithmic scale in the color bar. When the genotype probabilities are well calibrated, we expect all data points on the diagonal. (A) BEAGLE call set for 254 Chinese at 0.15X. (B) BEAGLE+1KG3 call set for 254 Chinese at 0.15X. (C) BEAGLE call set for 762 Chinese and Malays at 0.15X. (D) BEAGLE+1KG3 call set for 762 Chinese and Malays at 0.15X. (E) BEAGLE call set for 254 Chinese at 0.75X. (F) BEAGLE+1KG3 call set for 254 Chinese at 0.75X. (G) BEAGLE call set for 762 Chinese and Malays at 0.75X. (H) BEAGLE+1KG3 call set for 762 Chinese and Malays at 0.75X.



S5 Fig. Performance of homogeneous kinship estimators in ~0.15X sequencing data of 254 Chinese using SNPs with r^2>0.3. In each panel, we compared sequence-based estimates (ϕ_{seq}, y-axis) with the array-based estimates from PC-Relate (ϕ_{array}, x-axis). Colored circles represent kinship coefficients between two individuals and different types of relatedness were determined in Figure 2. Grey crosses represent self-kinship coefficients. We evaluated SEEKIN (A, D), GCTA (B, E), and KING (C, F) using the BEAGLE call set (A-C), and the BEAGLE+1KG3 call set (D-F). Note that KING does not estimate self-kinship coefficients.



S6 Fig. Performance of heterogeneous kinship estimators in ~0.75X sequencing data of 762 Chinese and Malays using SNPs with r^2>0.3. In each panel, we compared sequence-based estimates (ϕ_{seq}, y-axis) with the array-based estimates from PC-Relate (ϕ_{array}, x-axis). Colored circles represent kinship coefficients between two individuals and different types of relatedness were determined in Figure 2. Grey crosses represent self-kinship coefficients. We evaluated SEEKIN (A, E), PC-Relate (B, F), REAP (C, G), and RelateAdmix (D, H) using the BEAGLE call set (A-D), and the BEAGLE+1KG3 call set (E-H). We only included SNPs overlapping with the SGVP dataset in the analyses, because we used the SGVP dataset as the reference panel to estimate individual-specific allele frequencies for SEEKIN, REAP and RelateAdmix.