# **Supplementary methods**

# **SNP Calling**

## **Maximum log-likelihood estimation of allele frequency**

We improved the method of allele frequency calculation of a recent study by using log-likelihood to handle ultra-low coverage and very large sample size without suffering from precision loss under the current computer architecture (1). First, for one chromosome site, we randomly selected one read for each of the samples and used *samtools mpileup -Q20* to pile up the bases with a base quality score larger than 20. The log-likelihood of the allele frequency of a candidate variant site is defined as:

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  | (3) |

where *N* is the population size, *p* denotes the allele frequencies at the site and is the allele frequency of allele *b* and. is the sampled base of the *i*thsample at the site. is the sequencing error rate converted from the PHRED-scale quality of the selected read with at the site.

***EM optimization***

The allele frequency of a variant candidate site is estimated by maximizing ln(L(p)) (**Eq. 2**) through Expectation-Maximization (EM) optimization, which is the same with (1). First, the allele frequency *p* is initialized by:

|  |  |
| --- | --- |
|  | (4) |

In the expectation step, we calculate the posterior probability of the expected allele frequency at the candidate variant site, given by the observed nucleotide bases:

|  |  |
| --- | --- |
|  | (5) |

In the maximization step, the allele frequency is updated by maximizing the likelihoods.

|  |  |
| --- | --- |
|  | (6) |

The algorithm will be terminated if the change in ln(L(p)) is less than 0.001 or the iteration number has reached 2000.

## **Log-likelihood ratio test**

The aim of the log-likelihood ratio test is to remove the potential noises (such as very low allele frequencies) resulting from the maximum log-likelihood estimation of allele frequency. Instead of using a downstream way to test the hypothesis, we used the estimated allele frequency obtained from maximum log-likelihood estimation as the alternate hypothesis all the time.

The estimated allele frequency was considered as the alternate hypothesis, which was formulated as:

|  |  |
| --- | --- |
|  | (7) |
|  | (8) |

where denotes the allele frequency *p* derived from maximum log-likelihood estimation, which is used as alternate hypothesis in the log-likelihood ratio test. is the estimated allele frequency of base *b, .* is encoded as 0 or 1, meaning whether the allele *b* is regarded as a candidate variant.Anddenotes the initial allelic type of the candidate site (1 for mono-allelic site, 2 for bi-allelic site, 3 for tri-allelic site, 4 for tetra-allelic site). The candidate sites with > 1 will be used for log-likelihood ratio test. Then, the alternate hypothesis will test against the null hypothesis in a prioritized order using the following strategies.

***1 VS X***

First of all, we generate all possible mono-allelic models as the null hypothesis. The *i*th null hypothesis can be denoted as:

|  |  |
| --- | --- |
|  | (9) |
|  | (10) |

 is normalized to 1 by:

|  |  |
| --- | --- |
|  | (11) |

where means the allele frequency for the *i*th null hypothesis. In our case, the number of null hypothesis equals .

Then, the estimator of the log-likelihood ratio test is shown as:

|  |  |
| --- | --- |
|  | (12) |

where is formulated in **Eq. 2**.

We let the estimator follow a chi-square distribution. The degree of freedom is determined by how many different parameters are between the null and alternate hypothesis. Thus, in this case, the degree of freedom is .

This test would be applied to all possible null hypotheses. Then, we select the that produces the maximum p value. If this p value is larger than the threshold of 1e-6, the process is terminated here and the candidate site is regarded as a mono-allelic site. If the p value is smaller than the threshold, and , then the process is also terminated here and the candidate site is regarded as a bi-allelic site. Otherwise, it is considered there should be at least 2 variants and the process will proceed to the next procedure.

***2 VS X***

If the maximum p value of 1 vs X is smaller than the threshold and , we will continue to generate all possible bi-allelic models as the null hypothesis and test them against the alternate hypothesis .

We generated the null hypothesis of the allele frequency based on the following rules:

|  |  |
| --- | --- |
|  | (13) |
|  | (14) |
|  | (15) |

 is normalized to 1 using **Eq. 11**. Then, log-likelihood ratio test is applied to test all possible bi-allelic models against the alternate hypothesis using **Eq. 12**.

If the maximum p value is larger than the threshold of 1e-6, the process is terminated here, and the candidate site is regarded as a bi-allelic site. If this p value is smaller than the threshold, and , the process is also terminated here and the candidate site is regarded as a tri-allelic site. Otherwise, it is considered there should be at least 3 variants and the process will proceed to the next procedure.

***3 VS X***

We generated all possible tri-allelic models based on the following rules:

|  |  |
| --- | --- |
|  | (16) |
|  | (17) |
|  | (18) |

is normalized to 1 using **Eq. 11**. Then, log-likelihood ratio test is applied to test all possible tri-allelic models against the alternate hypothesis based on **Eq. 12**.

If the maximum p value is larger than the threshold of 1e-6, the process is terminated here, and the candidate site is regarded as a tri-allelic site. Otherwise, it will be considered as the tetra-allelic site, indicated by the alternate hypothesis.

# **Principal Component Analysis (PCA)**

In our study we used Principal Component Analysis in two analyses. 1) We identified potential biases using all 17,844 embryo datasets at the dbSNP known SNPs with MAF≥0.05 (Additional file 3: Figure S4) and concluded that the two most significant principal components represent different whole genome amplification kits, and sequencing platforms. 2) So, to remove the biases, we performed PCA on imputed and filtered SNPs in the 1,744 born embryos with gestational age, and used the eight most significant principle components as covariates to be considered in subsequent GWAS analyses.

The genotype probabilities calculated by STITCH (2) were used for calculating the covariance matrix as:

|  |  |
| --- | --- |
|  | (19) |
|  | (20) |

where denotes the imputed genotype probability for genotype *k*, with *k*=0 means genotype 0/0, *k*=1 means genotype 0/1 or 1/0, and *k*=2 means genotype 1/1. is the minor allele frequency of the *m*th SNP and *M* is the total number of SNPs. is the covariance matrix.

With over ten thousand samples and a hundred million SNPs, it was infeasible to get the PCA done in a reasonable time on CPU. Instead, we calculated the covariance matrix using CuPy (3) on GPU. The decomposition of the covariance matrix was calculated by the SciPy package (4).

# **Supplementary references**

1. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, et al. Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. Cell. 2018;175(2):347-59.e14.

2. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nat Genet. 2016;48(8):965-9.

3. Nishino R, Loomis SHC, editors. CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations. Proceedings of Workshop on Machine Learning Systems (LearningSys) in the Thirty-first Annual Conference on Neural Information Processing Systems (NIPS); 2017.

4. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for Python. 2001.