Non-invasive Preimplantation Genetic Testing of Embryonic Genome in Spent Culture Medium

Lei Huang[1,2#*], Yangyun Zou[3#], Ruiqi Zhang[4#], Jin Huang[5,6,7,8#], Guangjun Yin[1,2,9], Quangui Wang[9], Yingying Xia[3], Jialin Jia[5,6,7,8], Zeyu Wu[10], Dandan Cao[9], Weiliang Song[1,2], Yaqiong Tang[1,2,9], Kai Liu[1,2,9], Xiaoran Chai[1,2], Guo-Bo Chen[11], Sijia Lu[3*], Hao Ge[1,4*], Jie Qiao[1,2,5,6,7,8,12*] and Xiaoliang S. Xie[1,2,9*]

1. Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing, 100871, China
2. Beijing Advanced Innovation Center for Genomics, Peking University, Beijing, 100871, China
3. Yikon Genomics Co., Ltd., Shanghai, 310120, China
4. Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing, 100871, China
5. State Key Laboratory of Female Fertility Promotion, Centre for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing 100191, China
6. National Clinical Research Center for Obstetrics and Gynecology, Peking University Third Hospital, Beijing 100191, China
7. Key Laboratory of Assisted Reproduction (Peking University), Ministry of Education, Beijing 100191, China
8. Beijing Key Laboratory of Reproductive Endocrinology and Assisted Reproductive Technology, Beijing 100191, China
9. Changping Laboratory, Beijing, China
10. School of Mathematical Sciences, Peking University, Beijing, 100871, China
11. Center for Reproductive Medicine, Department of Genetics and Genomic Medicine, and Clinical Research Institute, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, Zhejiang, 310024, China
12. Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, 100871, China

[#] These authors contribute equally to the work.

[*]Corresponding authors: X. S. Xie (sunneyxie@biopic.pku.edu.cn), J. Qiao (jie.qiao@263.net), L. Huang(leihuangmed@gmail.com), H. Ge (haoge@pku.edu.cn), S. Lu (lusijia@yikongenomics.com)

ABSTRACT

Traditionally, preimplantation genetic testing (PGT) for in vitro fertilization (IVF) requires invasive trophectoderm (TE) biopsy, which might be detrimental to the embryo. Recently proposed non-invasive PGT (ni-PGT) utilizing cell-free DNA from spent embryo culture medium (SCM) also faces serious challenges in accuracy, especially for monogenic diseases (niPGT-M), due to trace DNA content, maternal cell contamination, and high Allele Drop-Out (ADO) rates. In this study, an improved linear single-cell whole genome amplification method and a Bayesian linkage analysis model were used to enhance accuracy in niPGT-M. We achieved about 75% report rate across all samples and 100% accuracy in the reported samples. Meanwhile, we reconstructed the embryonic genome and calculated the risk of type II diabetes (T2D) via niPGT-P, consistent well with those from TE biopsy samples. Our approach alleviated the limitations of ni-PGT and offers a promising avenue for advancing noninvasive PGT with potential clinical applications.

Keywords: non-invasive preimplantation genetic testing, monogenic diseases, polygenic diseases, spent embryo culture medium, Bayesian linkage analysis

INTRODUCTION

Preimplantation genetic testing (PGT) is widely used in clinical *in vitro* fertilization (IVF) to screen embryos for monogenic diseases, aneuploidy, or structural rearrangements. Next-generation sequencing (NGS)-based PGT allows simultaneous analysis of chromosome copy number, linkage, and targeted haplotyping for mutations [1-8]. However, traditional PGT requires trophectoderm (TE) biopsy, which poses risks to embryo health and is prone to sampling bias due to embryo mosaicism [9]. Recently,

minimally invasive or non-invasive methods using cell-free DNA (cfDNA) from blastocoele fluid (BF) [10] or spent embryo culture medium (SCM) [11,12] have been reported as alternatives for PGT with less risk and lower price. Despite their potential, challenges like trace DNA amount, poor DNA integrality, and complicated origin led to highly variable success rates in the published references from different clinical centers [12-15]. Although the success amplification rate in SCM was higher than that in BF [13,16] using the same amplification method, it still does not meet the yield required for clinical diagnosis application.

Even worse, genetic contents amplified from SCM can be potentially affected by maternal cell contamination (MCC), as maternal DNA is often longer than embryonic DNA [17,18] thus preferentially being amplified, and by the presence of high levels of human serum albumin (HSA) protein in culture medium. Current commercial single-cell amplification kits, such as DOP-PCR, MDA, or MALBAC are not specially optimized to address these challenges. In 2017, Linear Amplification via Transposon Insertion (LIANTI) was developed [19], significantly reducing the false positive rate of Single Nucleotide Polymorphisms (SNP) detection and much lower amplification bias caused by fragment lengths. However, the yield of LIANTI amplification for SCM is still insufficient. There is an urgent need to develop an efficient and accurate amplification method to obtain embryonic DNA information from SCM and/or BF.

Numerous reports have explored on non-invasive preimplantation genetic testing for aneuploidy (niPGT-A) [11,20,21], but studies on non-invasive preimplantation genetic testing for monogenic diseases (niPGT-M) remain limited [12,13,15,16, 22-24]. In niPGT-M studies using SCM, the genotype concordance with trophectoderm biopsy results varied widely, from 21% to 88%, limiting the clinical utility of niPGT-M [12-15]. The key criterion for evaluating PGT-M is not only genotype concordance but also the misdiagnosis rate, which is very low (<0.1%) with traditional methods according to the ESHRE PGT consortium [25]. This emphasizes the need

to assess misdiagnosis risk in new niPGT-M approaches.

Accurate linkage analysis is crucial alongside the amplification method to pinpoint disease-carrying chromosomes in the embryos and minimize misdiagnosis, especially given the trace DNA content in SCM. Traditional linkage analysis typically requires at least two fully informative markers like STRs or SNPs, near the mutation site, with ADO rates below 5% [26]. Ou et al. used 4 informative SNPs [15], and Liu et al. used 10 [12]. However, since ADO rates in SCM or BF can exceed 5%, relying on a few informative sites for linkage analysis may lead to misdiagnosis. Informative SNP markers could be too far from the mutation site, increasing recombination risk. Meanwhile, maternal cell contamination can also raise the misdiagnosis rate. These challenges indicate that traditional linkage analysis methods for niPGT-M may not be suitable for clinical applications. Therefore, developing a highly accurate linkage analysis method tailored for high ADO rates and maternal cell contamination is crucial.

Furthermore, preimplantation genetic testing for polygenic disorders (PGT-P) using polygenic risk scores (PRS) holds promise in reducing the risk of polygenic diseases, such as coronary heart diseases, diabetes, and cancers at the pre-conception stage [27-29]. Non-invasive PGT-P (niPGT-P) utilizing genetic information from SCM may, help alleviate some of the ethical risk-benefit concerns by avoiding the harms of biopsy to the embryos [30]. However, the challenge lies in the low coverage and inferior genotype quality of embryo genomes due to limited genetic materials and poor quality of cell-free DNA in SCM, which hinders the availability of genome-wide genetic variants necessary for PRS evaluation. Therefore, accurate whole genome reconstruction is crucial for the successful implementation of niPGT-P.

In the present study, we improved the LIANTI method to amplify trace DNA from SCM, overcoming challenges posed by high-concentration

protein existence and amplification bias caused by DNA size difference. We introduced a Bayesian-based linkage analysis method (BASE-niPGT-M) that incorporates high ADO rates and maternal cell contamination. This method could accurately detect disease-carrying chromosomes and provide a confidence level for each diagnosis. Furthermore, using the fragmented and incomplete genomic data from SCM, we employed Pedigree-Population-based Imputation with Haploid Assumption (PPIHA) method to reconstruct the embryonic genome. After reconstruction, the high genome-wide coverage and accuracy of genetic information were obtained, enabling PRS analysis. In summary, our study demonstrates high detection accuracy for ni-PGT, offering a reliable and potentially applicable non-invasive method for preimplantation genetic testing intended for clinical use.

RESULTS

Characterization of Cell-Free DNA in Spent Embryo Culture Medium

Understanding the properties of DNA in SCM, such as fragment size and DNA quantity, is crucial for selecting appropriate amplification methods. Unlike biopsy samples, which typically contain one to three complete cell nuclei, cfDNA in SCM or BF is present in low quantities and of poor quality. Our results showed the cfDNA in the culture medium exhibits highly variable DNA fragment size distribution and distinct DNA ladder patterns (Extended data Fig. 1 and Supplementary Note S1), posing a significant challenge for DNA amplification. Based on the above evidences, an effective amplification method should minimize amplification bias caused by DNA size variation.

Analysis of the sequencing results of cfDNA revealed low genome coverage in SCM, ranging from 10% to 50% (Extended data Fig. 2), in contrast to the genome coverage of biopsy samples, which typically exceeds 95%. This deficiency is not primarily attributed to the amplification-induced allelic dropout (ADO) but rather due to the large-scale DNA information loss of paternal or maternal alleles, we called it

large fragment loss of haplotype (LFLoH), as shown in Extended data Fig. 3. This LFLoH phenomenon often results in the loss of informative SNPs both upstream and downstream adjacent to the pathogenic site, which are typically used in traditional linkage analysis. Informative SNPs are those where the SNP is heterozygous in the parent carrying the disease and homozygous in the other parent, and where the SNP remains heterozygous in the embryo (Extended data Fig. 4).

Maternal cell contamination in embryo culture medium has long been a challenge in diagnostic procedures, often leading to misdiagnosis. Using informative SNPs with homozygous but different parental genotypes at the whole-genome level, we observed MCC values ranging from -55.1% to 99.7% (Extended data Fig. 5). Negative values may result from copy number gains in segments of paternal chromosomes or losses of maternal chromosomal segments. Employing an iterative strategy within a 10 Mb range upstream and downstream of pathogenic loci in our Bayesian linkage analysis method (See below and Methods), we observed the modified MCC values ranging from 0.1% to 99.7%. Aside from the absence of negative values, with this region, 58 samples have an MCC of less than 2%, indicating minimal maternal cell contamination, 43 samples have relatively low levels of maternal cell contamination between 2% and 30%, and only 10 samples have maternal cell contamination levels between 30% and 80%, with only 6 samples exceeding 80% (Extended data Fig. 5, Supplementary Table S4).

It's important to note that the cfDNA in SCM may potentially originate from cell apoptotic processes. Using a ligation-based method (Supplementary Notes S1), we observed a clear DNA ladder in the culture medium, as shown in Extended data Fig. 1.

Whole Genome Amplification Method for SCM

Given the unique characteristics of cfDNA in SCM, a whole genome amplification method with minimal fragment size bias is essential. We compared the fragment size amplification biases of the current WGA methods for PGT-M and found that LIANTI exhibited the least bias (short:

long = 1:5), while MDA and MALBAC showed a stronger preference for amplifying longer fragments, with short: long amplification biases of 1:500 and 1:6000, respectively (Supplementary Table S1). LIANTI is a linear amplification method known for its high SNP detection accuracy [19].

However, the original LIANTI method has low efficiency in amplifying such minute quantities of DNA, failing to meet clinical requirements. To address this, we improved the LIANTI method by modifying the lysis step, adding more transposons for insertion, introducing DNA primers during first-strand DNA synthesis, and implementing exponential amplification during second-strand synthesis, as depicted in Extended data Fig. 6 (see Methods for details).

The improved LIANTI method achieved a 100% amplification success rate for clinical samples and demonstrated high SNP detection accuracy, as shown in Extended data Fig. 7 and Supplementary Table S2. We compared the sequencing results of cfDNA from BJ cell line culture medium collected at different time points using the improved LIANTI method and other scWGA methods. The improved LIANTI method demonstrated superior amplification efficiency for cfDNA compared to the other scWGA methods.

Linkage Analysis Method for Noninvasive PGT-M of Embryos

Figure 1 illustrates the complete workflow employed in this study. Briefly, data collection was divided into two stages (Figure 1A): the Pre-phasing stage and the SCM genetics information acquisition stage. In the Pre-phasing stage, family trio sequencing was performed using DNA from the father, mother, and proband or alternatively from grandparents or discarded embryos. During the SCM genetics information acquisition stage, the DNA content in SCM was amplified using the improved LIANTI method, and then was followed by sequencing and SNP calling (See Fig. 1A and Methods).

With these data in hand, we present a computational method BASE-niPGT-M (BAyesian linkage analySis mEthod for non-invasive Preimplantation Genetic Testing of Monogenic disorders) (Fig. 1B,

Extended data Fig. 8 and Supplementary Note S2). This approach addresses the challenges posed by MCC and LFLoH in SCM samples. The Bayesian model takes the sequencing data of SCM and the phased haplotypes of the parents, as input. Samples with extremely low quality are filtered out. Unlike previous approaches for high-quality sequencing data, which often predetermined the number of SNPs to be considered, our model incrementally incorporates SNPs, starting from the disease-causing mutation site. This process allows us to calculate the log-likelihood ratio of inheriting disease-carrying chromosomes versus disease-free chromosomes for each SNP subset, ultimately resulting in a log-likelihood ratio curve (Fig. 1B, Extended data Fig. 8 and Supplementary Note S2). This curve, with its distinct characteristics, enables us to classify the results into four categories: Highly Confident, Moderately Confident, Possibly confident, and Undetermined. In the first three categories, we can pinpoint the inherited chromosome.

Case Study of our niPGT-M Method

We selected one case (Family 2) to illustrate the SNP distribution of cfDNA and the results of our niPGT-M method, which consists of WGA and NGS followed by BASE-niPGT-M. This case involved an autosomal dominant genetic disease, where the husband carried a genetic variant for Marfan syndrome, a heritable connective tissue disorder. This disorder is known for its characteristic features such as elongated limbs, aortic aneurysms, and mitral valve prolapse. Genetic diagnosis of the husband revealed a mutation c.643C>T at the FBN1 gene, known to cause Marfan syndrome. Eight embryos of the couple developed to the blastocyst stage, and several trophectoderm cells were biopsied from each embryo for traditional PGT-M. All eight SCM samples were collected, and cfDNA was amplified using the improved LIANTI method. The linkage analysis results of these eight SCMs are shown in Fig. 2A. The detected SNP distribution of 2T-SCM-01 and 2T-SCM-02 appeared relatively scattered, with longer intervals between informative SNPs (Fig. 2A), and the SNPs flanking the causal mutation site indicated different haplotypes from the father (Fig. 2B),

leading to their classification as 'undetermined'. In contrast, 2T-SCM-03, 2T-SCM-04, 2T-SCM-05, 2T-SCM-06 and 2T-SCM-08 showed denser SNP distributions, and SNPs upstream and downstream of the causal mutation site indicated the same haplotype (Fig. 2B), resulting in a classification of 'Highly confident'. The detected SNPs of SCM-07 were dense, and the haplotype upstream and downstream was concordant. However, an upstream SNP located relatively close to the causal mutation site pointed to the opposite haplotype (Fig. 2B), leading to a classification of 'moderately confident' due to our conservative diagnosis strategy. All these results were consistent with those of trophectoderm biopsies, demonstrating no misclassifications.

Validation Using all Matched Biopsy/ Whole Embryo.

To further evaluate the performance of our niPGT-M method, we compared the results obtained from all SCM samples to the gold standard, which was either TE-biopsy or discarded embryo PGT-M results (Table 1). We categorized the cases into three groups: autosomal dominant disease, autosomal recessive disease, and X-linked recessive disease. For autosomal recessive diseases, we compared each allele of embryos.

In the seven cases with autosomal dominant diseases, there were 45 SCM samples. For embryos that the father carried the causal mutation, 23 out of 28 SCMs (82.1% report rate) were successfully diagnosed. For embryos that the mother carried the causal mutation, 11out of 17 SCMs (64.7% report rate) were successfully diagnosed. In the five cases with autosomal recessive disorders, 36 SCMs were analyzed. We successfully detected the paternal allele in 25 SCMs (69.4% report rate) and the maternal allele in 27 SCMs (75% report rate). For the two cases with X-linked recessive disorders, 8 out of 9 SCMs (88.9% report rate) were diagnosed. Overall, we successfully detected 94 out of 126 alleles (74.6% average report rate). The average report rate for paternal alleles was 75%, while the average report rate for maternal alleles was 74.2%.

Importantly, all detected results were concordant with the gold standard, demonstrating a 100% accuracy rate of BASE-niPGT-M (Table

1).

Whole Genome Reconstruction of Embryos

The directly detected genetic information of embryos obtained from the SCM samples is much limited due to low amount of cfDNA materials released from blastocyst cells. We therefore developed a Pedigree-Population-based Imputation with Haploid Assumption (PPIHA) to acquire the missing genetic information. This approach allows us to reconstruct the whole genome information of embryos from SCM samples.

The initial density of detected sites in the raw data was 222±127 SNPs per megabase (Mb). However, after performing PPIHA (shown in Fig. 3A), the median coverage of SNPs significantly improved from 6.4% to 93.5%, and the density of imputed sites increased to 1479±554 SNPs/Mb (Fig. 3B and 3D).

To assess the accuracy of the predicted embryo's haplotype obtained from SCM samples, we compared it with the haplotype of corresponding discarded embryos or TE biopsy samples. When the MCC of SCM was low, we observed a high similarity of haplotype configuration between SCM and the corresponding TE biopsy. Conversely, SCM samples with high MCC exhibited reduced haplotype similarity, especially in the maternal haplotypes (Fig. 3C). For a more comprehensive exploration, we obtained 43 paired samples with whole genome sequencing data from both SCMs and their corresponding discarded embryos or TE biopsy samples for comparison. To avoid bias we removed paired samples with failed quality control or aneuploid embryos for further analysis (Supplementary Table S3). Paternal or Maternal haplotype concordance at each physical site between SCM and discarded embryos/TE biopsy samples was applied to measure the haplotype accuracy constructed in SCM (Fig. 3E). We found that the concordance of maternal and paternal haplotypes in SCM versus discarded embryos/TE was around 98.6% and 99.8% respectively, for SCM samples without significant maternal cell contamination (Fig. 3E). High levels of maternal cell contamination could largely impact the haplotype construction of SCM, especially for maternal haplotypes (Fig.

3E).

Furthermore, after filtering out SNPs with Mendelian errors, we compared the constructed genome-wide genotypes from SCM samples with those of corresponding discarded embryos/TE biopsy samples. The concordance of genotypes ranged from 79.2% to 99.9%, partially impacted by maternal cell contamination in SCM (Fig. 3E). For those SCM without significant maternal cell contamination, the average genotype accuracy was approximately 96.9%. Additionally, we compared the constructed genotypes from two SCM samples with the genotypes of corresponding amniotic fluid samples, which exhibited concordance rates of 94.5% and 94.1%, respectively.

Polygenic Risk Evaluation of Type II Diabetes for Embryos

We used type II diabetes (T2D) as a complex disease model to assess the feasibility of polygenic risk evaluation based on constructed whole genome information of SCM. After applying various filters, including removing SCM samples with high maternal cell contamination, chromosomal abnormality or poor-quality, 25 SCM samples were left for polygenic risk score (PRS) calculation of T2D. We employed a PRS model of T2D from a previous study [31]. The PRS of T2D model involves a total of 114 loci, where only 28.7% of the loci can be directly detected, 52.5% being obtained through family-based genome reconstruction, and the remaining 18.8% being imputed by population-based reconstruction (Fig. 3B, Fig. 4A).

When comparing the genotypes of SNPs in the T2D PRS model between the SCM samples and corresponding discarded embryos or TE biopsy samples, we found that median 97.4% of the SNPs were identical. Among these identical SNPs, 52.5% are homozygous REF genotypes, 24.9% are homozygous ALT genotypes, and 22.6% are heterozygous genotypes (Fig. 4B).

The PRS of T2D calculated from SCM showed a high level of concordance with those calculated from the corresponding discarded embryos, TE biopsy samples, or amniotic fluid, with an R-squared value

of 0.79 (Fig. 4C). In clinical practice, PRS percentiles are typically used to evaluate the relative genetic risk of common diseases instead of raw PRS scores. To determine an individual's PRS percentile for a particular disease, a reference population comprising both patients (e.g., with T2D) and healthy individuals is needed to establish the PRS distribution, which then allows each individual to be placed within a specific percentile of the distribution. For example, if an individual's PRS is in the 90th percentile, it means he/she has a higher PRS than 90% of the reference population. Generally, the higher the PRS percentile, the greater the risk of a genetic disease. We used a reference population from the UK Biobank (https://www.ukbiobank.ac.uk/) to construct the PRS distribution for T2D and calculated each embryo's PRS percentile. We observed a correlation of T2D PRS percentiles between SCM samples and the corresponding discarded embryos, TE biopsy samples, or amniotic fluid, indicated by an R-squared value of 0.68 (Fig. 4D).

DISCUSSION

In this study, we introduced a method for non-invasive preimplantation genetic testing, addressing key challenges in the accuracy of genetic assessments using SCM. We improved a linear single-cell whole genome amplification method to successfully amplify DNA content in SCM and introduced a Bayesian linkage analysis model to improve the accuracy of ni-PGT for monogenic diseases, matching the reliability of invasive procedures. Our approach advances the potential clinical implementation of ni-PGT, reducing the risks associated with traditional PGT while maintaining high accuracy.

PGT-P holds promise for reducing the incidence of complex genetic diseases at the population level. However, its widespread use is hindered by controversies regarding ethical concerns, scientific validity, and practical issues, particularly the potential harm of the biopsy procedure to embryos if the benefits of PGT-P are not conclusive. Non-invasive PGT-P, which detects genetic materials from embryos through SCM, may

accelerate scientific research and clinical implementation by significantly reducing potential risks to embryos. We achieved comprehensive whole-genome construction of embryos, accurately assessing the risk of polygenic diseases (niPGT-P) in line with results from trophectoderm (TE) biopsy samples. Our PPIHA approach to reconstruct whole genome information of embryos from SCM samples fully considers the genetic characteristics of SCM specimens, such as LFLoH, which violates the general biological principles of diploidy, and the low amount and quality of cfDNA released from blastocysts. The high genome-wide coverage and accuracy of genetic information in SCM samples constructed through PPIHA enable non-invasive assessment of polygenic disease risk in embryos, advancing the prevention and control of common diseases to the pre-implantation stage.

Maternal cell contamination from cumulus cells or polar bodies in SCM is an issue that requires attention, which would impact the accuracy of niPGT-M and niPGT-P. Special techniques for the removal of maternal cell contamination are necessary to be developed in the future. At current phase, we created a quantitative estimation of MCC rate which plays a warning role when it is high, and ensures accurate ni-PGT results when it is not high.

In contrast to TE samples that contain alleles from two haplotypes, SCM only provides alleles from a single haplotype in many genomic regions, resulting in a high rate of allelic dropout and random segments of single haplotypes in the SCM data. To address this issue, we developed a specialized approach to first estimate the haplotype status of each SNP, specifically whether only the parental or maternal haplotype detected. Assuming the independence of haplotype status among measured SNPs would introduce an excessive number of parameters, leading to the risk of overfitting. Therefore we have adopted a cautious approach: the default assumption for the haplotype status of each SNP is set to be “both parental haplotypes are detected”, unless the SNP data strongly contradicts this default.

The accuracy of haplotype pre-phasing of the parents is important for both niPGT-M and niPGT-P. In cases where no proband is available,

sequencing data from discarded embryos must be used for pre-phasing. However, the data quality, particularly the coverage, obtained from discarded embryos is often limited, which can hinder the accuracy of haplotype phasing results for the parents. Improved precision in haplotype phasing of the parents, possibly achievable through more advanced sequencing technologies, would enhance the accuracy of niPGT-M and niPGT-P.

In the context of clinical applications, our methodology extends beyond merely estimating inherited chromosomes in the embryo; it also provides a measure of confidence in this estimation. By utilizing a Bayesian model, the log-likelihood ratio curve enables a nuanced classification of our diagnoses. This approach is crucial for aiding both medical professionals and parents in determining the most suitable course of action for the embryo. The precision of our method, combined with the refined classification outputs, significantly enhances the reliability of identifying healthy embryos. This enhancement underscores the potential translational impact of our work, promising increased report rates in pregnancies and the delivery of healthy infants.

REFERENCES

1. Yan, L. et al. Live births after simultaneous avoidance of monogenic diseases and chromosome abnormality by next-generation sequencing with linkage analyses. *Proc Natl Acad Sci U S A* . 112(52):15964-9 (2015)
2. Esteki, Z.M. et al. Concurrent whole-genome haplotyping and copy-number profiling of single cells. *Am J Hum Genet.* 96(6):894-912 (2015)
3. Xiong, L. et al. Bayesian model for accurate MARSALA (mutated allele revealed by sequencing with aneuploidy and linkage analyses). *J Assist Reprod Genet.* 36(6): 1263–1271 (2019)
4. Backenroth, D. et al. Haploseek: a 24-hour all-in-one method for preimplantation genetic diagnosis (PGD) of monogenic disease and aneuploidy. *Genet Med.* 21(6):1390-1399 (2019)
5. Masset, H. et al. Single-cell genome-wide concurrent haplotyping and copy-number profiling through genotyping-by-sequencing. *Nucleic Acids Res.* 50(11):e63 (2022)
6. Yan, L. et al. Chinese experts' consensus guideline on preimplantation genetic testing of monogenic disorders. *Hum Reprod.* 38(Supplement_2):ii3-ii13 (2023)

7. Backenroth, D. et al. SHaploseek is a sequencing-only, high-resolution method for comprehensive preimplantation genetic testing. *Sci Rep.* 13(1):18036 (2023)
8. De Coster, T. et al. Genome-wide equine preimplantation genetic testing enabled by simultaneous haplotyping and copy number detection. *Sci Rep.* 14(1):2003 (2024)
9. McCoy, R.C. Mosaicism in preimplantation human embryos: when chromosomal abnormalities are the norm. *Trends Genet.* 33(7):448-463（2017）.
10. Palini, S. et al. Genomic DNA in human blastocoele fluid. *Reprod Biomed Online* 26, 603–610 (2013).
11. Xu, J. et al. Noninvasive chromosome screening of human embryos by genome sequencing of embryo culture medium for in vitro fertilization. *Proc Natl Acad Sci* 113, 11907–11912 (2016).
12. Liu, W.Q. et al. Non-invasive preimplantation aneuploidy screening and diagnosis of beta thalassemia IVSII654 mutation using spent embryo culture medium. *Ann Med* 49, 319–328 (2017)
13. Capalbo, A. et al. Diagnostic efficacy of blastocoel fluid and spent media as sources of DNA for preimplantation genetic testing in standard clinical conditions. *Fertil Steril* 110, 870–879 (2018).
14. Cimadomo, D. et al. The dawn of the future: 30 years from the first biopsy of a human embryo. The detailed history of an ongoing revolution. *Human Reproduction Update.* 26, 453-73 (2020).
15. Ou, Z. et al. Improved non-invasive preimplantation genetic testing for beta-thalassemia using spent embryo culture medium containing blastocoelic fluid. *Frontiers in endocrinology* 12, 1941 (2022).
16. Galluzzi, L. et al. Extracellular embryo genomic DNA and its potential for genotyping applications. *Futur Sci OA* 1, FSO62 (2015).
17. Lo, Y.M.D. et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* 2, 61ra91 (2010)
18. Yu, S.C.Y. et al. Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc. Natl. Acad. Sci. USA* 111, 8583–8588 (2014)
19. Chen, C. et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* 356, 189-94 (2017).
20. Huang, L. et al. Noninvasive preimplantation genetic testing for aneuploidy in spent medium may be more reliable than trophectoderm biopsy. *Proc. Natl. Acad. Sci. USA* 116, 14105-14112 (2019)
21. Leaver, M. &Wells, D. Non-invasive preimplantation genetic testing (niPGT): the next revolution in reproductive genetics? *Human reproduction update* 26, 16-42 (2020).
22. Wu, H. et al. Medium-based noninvasive preimplantation genetic diagnosis for

human α-thalassemias-SEA. *Medicine (Baltimore)* 94, e669 (2015).
23. Zhang, Y. et al. Molecular analysis of DNA in blastocoele fluid using next-generation sequencing. *J Assist Reprod Genet* 33, 637–645 (2016).
24. Shangguan, T. et al. Detection and analysis of DNAmaterial in human blastocoel fluid. *Biomed Genet Genomics* 2, 1–5 (2017).
25. De Rycke, M. et al. ESHRE PGD Consortium data collection XIV-XV: Cycles from January 2011 to December 2012 with pregnancy follow-up to October 2013. *Hum. Reprod.* 32, 1974–1994 (2017).
26. Carvalho, F. et al. ESHRE PGT Consortium good practice recommendations for the detection of monogenic disorders. *Human Reproduction Open* 2020(3), hoaa018 (2020).
27. Treff, N.R. et al. Utility and first clinical application of screening embryos for polygenic disease risk reduction. *Frontiers in Endocrinology* 10, 845 (2019).
28. Kumar, A. et al. Whole-genome risk prediction of common diseases in human preimplantation embryos. *Nature Medicine* 28, 513–516 (2022).
29. Lennon, N.J. et al. Selection, optimization and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse US populations. *Nature Medicine* 30, 480–487 (2024).
30. Turley, P. et al. Problems with Using Polygenic Scores to Select Embryos. *New England Journal of Medicine* 385, 78–86 (2021).
31. Scott, R.A. et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* 66, 2888–2902 (2017).

METHODS

Institutional Review Board Approval.

This study was approved by the Peking University Third Hospital Ethical Review Committee (Approval no. 2020SZ-005). The patients/participants provided written informed consent to participate in this study.

Study Subjects.

Patients (n=14 couples) using ICSI and PGT-M with TE biopsy for preventing their inherited genetic diseases at Peking University Third Hospital during the period from September 2019 to April 2021 were included. All of the 14 patients had genetic testing reports.

ART Laboratory Protocols and Blastocyst Biopsy for PGT-M.

Standard protocols were used for ICSI. Embryos were cultured

individually in 25-μL microdrops of the equilibrated Vitrolife G-series media with human serum albumin (HSA) (LifeGlobal) overlain with mineral oil in incubators with 6% CO2 and 5% O_2 at 37 °C. Embryos were evaluated on day 5 or 6 for trophectoderm (TE) biopsy. A few TE cells from each hatching blastocyst were biopsied and sent to PGT lab for PGT-M testing.

Collection of Culture Medium Sample and Discarded Embryo

Immediately after the blastocysts were transferred to another dish for biopsy, 20-μL spent embryo culture media were removed from each of the residual spent medium drops. The discarded embryos and their corresponding spent embryo culture media samples were also collected for analysis. Each discarded embryo was gently moved with a pipette tip to the edge of its microdrop and then removed and transferred into a RNase-DNase–free PCR tube containing 3 μL of lysis buffer. Pipette tips were changed between collections of each sample to avoid cross-sample contamination. Media drops incubated and collected under identical conditions to those used for blastocyst culture, but never containing embryos, served as negative controls. All samples were immediately frozen after collection and stored at −80 °C until analyzed.

Spent Embryo Culture Medium Lysis and Discarded Embryo Lysis

The frozen 20-μL spent embryo culture medium samples were thawed and gently mixed, and 10.8 μL was removed to a PCR tube (Maximum Recovery, Axygen) containing 1.2 μL 10x lysis buffer (1x lysis buffer: 60 mM Tris-Ac pH 8.3, 2 mM EDTA pH 8.0, 15 mM DTT, 0.5 μM carrier ssDNA (5'-TCAGGTTTTCCTGAA-3', Thermo Fisher Scientific oligo with PAGE purification)), spun down. It was heated at 75°C for 30 min, then incubated at 75°C for 30 min. 0.5 μL 35mg/mL QIAGEN protease (dissolved in water and stored at 4°C) was added, spun down, and incubated at 55°C for 2 hrs followed by 80°C for 30 min. The resulting lysate in PCR tubes could be immediately subject to improved linear transposon-based amplification as described herein or stored in a freezer for later use.

The frozen discarded embryos were thawed and heated at 75°C for 30

min. Then the discarded embryo samples were lysed by adding 0.5 µL 5mg/mL Qiagen Protease and heating (2 h at 55 °C and 30 min at 80 °C) in 3 µL of lysis buffer.

Whole Genome Amplification

Transposon DNA (5'/Phos/CTGTCTCTTATACACATCTGAACAGAATTTAATACGACTCACTATAGGGAGATGTGTATAAGAGACAG-3', Thermo Fisher Scientific oligo with PAGE purification) was annealed by gradual cooling in annealing buffer (20 mM Tris-Ac pH 8.3, 50 mM NaCl, 2 mM EDTA pH 8.0) into a 1.5 µM self-looping structure. The 1.5 µM annealed transposon DNA was then mixed with an equal volume of ~1 µM Tn5 transposase (Lucigen, EZ-Tn5™ Transposase) and incubated at room temperature for 30 min, dimerizing into transposomes with a final concentration of ~0.25 µM. The transposomes can be stored at -20°C for a long time, and each single-cell transposonbased amplification needs ~0.5 µL transposome.

Starting with the culture medium lysate, a 20 µL transposition mixture was assembled in the buffer containing 2.5 mM MgCl2 and 18.75 nM transposome. The transposition reaction was carried out at 55°C for 12 min. Then 0.84 µL mixture of EDTA and ssDNA was added to each tube and was incubated at 68 °C for 30 min. After removing transposase reaction, 0.2 µL Q5 HF DNA Polymerase (New England Biolabs) was added and was heated at 73°C for 45 seconds in the presence of 2 mM MgCl2 and 200 µM dNTPs for fragments end filling and extension. 0.5 µL 4mg/mL Qiagen protease was added and the PCR tube was heated at 50 °C for 1 hour in the presence of 4 mM EDTA to inactivate DNA polymerase, followed by protease heat inactivation by 77°C for 20 min in the presence of 450 mM NaCl in a total volume of ~25 µL. DNA fragments were amplified to RNAs in a 90 µL T7 in vitro transcription reaction mixture overnight at 37 °C as described in Cheng et al's paper [19].

The next day, 10 µL 0.5M EDTA was added to each tube and RNAs were column purified (Zymo Research). 18 µL RNAs were transferred to a PCR tube containing 3.1 µL mixture of 6.5mM each dNTPs, 9.7 µM

carrier ssDNA (5'-TCAGGTTTTCCTGAA-3', Thermo Fisher Scientific oligo with PAGE purification) and 3.2 U/μL SUPERase In RNase Inhibitor (Invitrogen). The denaturation incubation was at 70 ℃ 1min, 90 ℃ 15s, and followed by ice quenching. The reverse transcription reaction for the first stand was carried out in a 30 μL SuperScript IV reserve reaction system including 0.67 mM each dNTPs, 0.6 U/μL SUPERase In RNase Inhibitor (Invitrogen) and 6 U/μL SuperScript IV Reserve Transcriptase (Invitrogen) in SuperScript IV buffer, with the primer 5'-AGATGTGTATAAGAGACAG-3', Thermo Fisher Scientific oligo with PAGE purification, and the incubation program was 25°C 1 min, 37°C 1 min, 42°C 1 min, 50°C 1 min, 55°C 15 min, 60°C 10 min, 65°C 12 min, 70°C 8 min, 75°C 5 min, 80°C 10 min. RNA was then removed by incubation at 37°C for 30 min with 10 ng/μL affinity-purified RNase A (Invitrogen) and 0.08 U/μL RNase H (New England Biolabs). Second strand synthesis was carried out in a 100 μL Q5 DNA polymerase system (New England Biolabs, 1X Q5 reaction buffer, 1X Q5 High GC enhancer, 200 μM dNTPs, 0.5 μM primer, 0.02 U/μL Q5 DNA polymerase), with the primer 5'-NNNNNNNNGGGAGATGTGTATAAGAGACAG-3', Thermo Fisher Scientific oligo with PAGE purification. Each tube was heated at 98°C for 30 s, then using 10 cycles of 10 s at 98°C, 30s at 58°C, 30s at 60°C, 30s at 65°C, 2.5 min at 70°C, then 72°C 6 min for strand extension. The resulted amplicons were column purified into 23 μL elution buffer and stored at -20°C. The DNA concentration of the product after amplification was measured using a Qubit 2.0 fluorometer (ThermoFisher Scientific) with the Qubit dsDNA HS Assay kit (Life Technologies).

Starting with the discarded embryo lysate, DNA was amplified following the LIANTI amplification method described in [19], except the DNA primer 19_1 5'-AGATGTGTATAAGAGACAG-3', Thermo Fisher Scientific oligo with PAGE purification was added in the reaction of the first strand cDNA synthesis.

Sequencing and Data Analysis.

Upon library preparation, the amplicons were subject to a conventional sonication process (Covaris S2) to the length required by

sequencing platforms. The final insert size was suitable for 2x150 bp pair-end Illumina sequencing. Library preparation was performed by NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs), following the instructions of the manufacturer and skipping the optional size selection step. For the gDNA of blood samples, all the samples were subject to a conventional sonication process (Covaris S2). Library preparation was performed by a PCR-free Library Prep Kit, NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB #E7645S/L).

The bulk samples, culture medium samples, and discarded embryo samples were sequenced on the Illumina NovaSeq 6000 System or Illumina HiSeq HiSeq 4000 platform with 2x150 bp pair-end sequencing. Sequencing lanes were shared by 24 samples with NEBNext Indexes. Each gDNA sample was sequenced to generate ~90 Gb raw data and each culture medium sample or discarded embryo was sequenced to generate ~30 Gb raw data.

Mapping and SNP Calling

Raw paired-end reads were trimmed adapters and low-quality ends by cutadapt. Clean reads were mapped to human reference genome hg19 with BWA mem. Each BAM file was marked duplicates and sorted by MarkDuplicatesSpark of GATK 4.2.0. Base quality score recalibration (BQSR) was performed using BaseRecalibrator to generate a recalibration table, followed by ApplyBQSR to adjust the base quality accordingly. For variant calling, GATK HaplotypeCaller produced a genomic variant call format (GVCF) file for each sample. GenotypeGVCFs was utilized to perform joint genotyping across all samples. SNPs were filtered using variant quality score recalibration (VQSR) with HapMap 3.3, Omni 2.5, 1000 Genome phase I, and dbSNP 151 as SNP training sets. A 99% sensitivity threshold was chosen to filter SNPs accurately.

We exclusively retain biallelic SNPs for subsequent analysis. To deduce the two parental haplotypes, we employ Plink 1.9 with the option "--mendel" to identify sites with Mendelian errors.

Parental haplotype phasing

For parental haplotype phasing, genetic information of pedigrees such as sibling embryos or other family members are needed. Likelihood-based haplotyping approach using Hidden Markov Model (HMM) and dynamic programming Viterbi algorithm were used to determine the most likely haplotype configuration of parents [32,33]. The transition probability of haplotype state changes between consecutive loci in the HMM was calculated from the recombination fraction obtained from the genetic distance map of the 1000 Genomes Project Phase 3 references (https://www.internationalgenome.org/). To address the ADO issue stemming from scWGA bias during parental haplotyping with sibling embryos, we filtered out sites that violated Mendelian rules or phased variants that did not adhere to chromosome interference theory [34]. Each chromosome of parent was then independently processed until the parental haplotypes of all chromosomes were phased.

BASE-niPGT-M: a BAyesian linkage analySis mEthod for niGPT-M

Bayesian model. We present BASE-niPGT-M, a Bayesian linkage analysis method tailored for niPGT-M. Our approach incorporates the MCC rate and haplotype status of each SNP into the likelihood function. To enhance the robustness of our data, SNPs with extremely low quality are filtered out. Then the sequencing error rate, MCC rate and haplotype status were carefully estimated. Subsequently, we calculated the likelihood of observing SNP data from the SCM samples within regions proximal to the disease-causing mutation site. The calculation is based on the single-SNP likelihood, which was attained from a binomial model using the measured allelic depth (AD) values and estimated parameters, and the recombination probability that can be obtained from the DECODE dataset or other datasets containing the recombination fractions. See Extended data Fig. 8 and Supplementary Notes S2 for more details.

Recursively estimating the MCC rate and haplotype status. MCC rate was defined as the ratio of DNA fragments originating from the maternal

chromosome. Haplotype status, representing the true parental origin of DNA at each measured SNP, is categorized into three distinct classes: 'Paternal Chromosome Only,' 'Maternal Chromosome Only,' or 'Both Parental Chromosomes.' For instance, the designation 'Paternal Chromosome Only' indicates that the DNA identified at a specific SNP exclusively originates from the father. Initial MCC rate and haplotype status were estimated, and these parameters were then iteratively and jointly calibrated until convergence (See Supplementary Notes S2 for details).

Determining the inherited chromosome. Unlike previous publications, where a fixed number of SNPs were predetermined, our methodology involves a sequential addition of SNPs. Starting from the disease-causing mutation site, we incrementally incorporate SNPs one by one, calculating the log-likelihood ratio for each subset. This iterative process yields a curve of log-likelihood ratio plotted against the physical distance of the terminal SNP.

Typically, the curve initiates at a point where the ordinate is in close proximity to zero, gradually stabilizing into a plateau with the addition of sufficient SNPs. The distinctive properties of this curve enable the determination of the inherited chromosome. Subsequently, the confidence level of our disease-carrying analysis is categorized into four classes: Highly Confident, Moderately Confident, Possibly confident, and Undetermined, based on the characteristics of the curve. See Supplementary Notes S2 for more details.

Haplotype phasing and whole genome reconstructing of SCM

In many regions of the genome from the SCM samples, genetic information from paternal and maternal origins of the embryo was not both detected. This was due to allelic dropout resulting from scWGA bias or LFLoH. The conventional diploid assumption of the genome could not be applied to haplotyping and genotype imputation of SCM samples.

Therefore, we developed a Pedigree-Population-based Imputation with Haploid Assumption (PPIHA) to reconstruct the whole genome information of SCM. Specifically, we combined pre-phased parental haplotype information and selected informative SNPs where at least one parental origin haplotype could be confidently phased, to create the haplotype scaffold of SCM (Extended Data Fig. 4). For example, for an informative locus with paternal genotype being AB (allele A comes from haplotype P1 and allele B from P2, based on the pre-phased parental haplotype), the maternal genotype being AA, and the genotype called from SCM being BB, it could infer that the B allele of this variant in SCM is derived from haplotype P2, based on the Mendelian rule (see Extended data Fig. 4 and Fig. 9). Maternal origin haplotype was not determined at this site due to ADO bias or LFLoH feature in SCM. And vice versa, for the locus with maternal, paternal and SCM genotypes being AB, AA and BB, respectively, maternal haplotype of SCM could be inferred while paternally genetic information unavailable. After examining all such kind of informative variants from the sequenced SCM data, we constructed a sparse haplotype scaffold with single parent of SCM (Fig. 3A and Extended data Fig. 9).

Paternal or maternal haplotypes in the scaffold were first corrected if continuous variants of the opposite haplotype were observed on both nearby sides. This correction was applied when the probability of a double crossover between the nearest variants with the opposite haplotype, calculated as the square of their genetic distances multiplied by 0.0001, was less than 0.005 for the maternal haplotype or 0.001 for the paternal haplotype, according to chromosome interference theory ([34], Extended Data Fig. 9). Then, with parental haplotype support and HMM strategy, paternal and maternal haplotypes of other variants were independently phased based on pre-determined haplotype scaffold of SCM (Extended Data Fig. 9). We assumed that the state (P1 or P2 for paternal haplotypes and M1 or M2 for maternal haplotypes) at variant t could be inferred by SNPs in the scaffold from t-5 to t+5. The transition probability between two SNPs was considered equivalent to their genetic distance multiplied

by 0.01. Joint probabilities of the two states of paternal or maternal haplotypes at variant t were calculated using a forward-backward algorithm based on the adjoining ten variants (from t-5 to t+5). The haplotype of variant t was then assigned to the state with the higher probability. If the joint probabilities of the two states (P1/P2 or M1/M2) for a variant were equal, phasing of that variant failed (Extended Data Fig. 9). By organizing the haplotype states of all variants according to their chromosomal positions, the pedigree-based haplotypes of SCMs were reconstructed. Subsequently, the genotypes of SCM were imputed with the genome information of the parents (Fig. 3A, Extended Data Fig. 9).

Following pedigree-based genome reconstruction, population-based genotype imputation was conducted using the 1000 Genomes haplotype reference panel with Minimac4 [35], resulting in the reconstruction of the whole genome information of the embryo from the SCM samples (Fig. 3A).

Polygenic risk score calculation

We employed a PRS model of T2D including 128 SNPs from a previous study [31]. To validate the model and determine the PRS percentile (indicating relative disease risk) in the UK Biobank (UKB) cohort, SNPs whose genotype not available or failing to be imputed in UKB cohort were excluded, leaving a total of 114 SNPs. We then directly used the effect sizes from the original study to calculate the PRS, employing the following formula:

$$\mathrm{PRS} = \sum_{i=1}^{114} \beta_i \times G_i$$

where G_i is the allelic dosage and β_i is the effect size of SNP i.

Using the T2D model and control information in UKB cohort, PRS percentile for T2D was obtained to evaluate the polygenic risk of T2D for each individual.

DATA AVAILABILITY

Access to anonymized patient data is subject to a data-sharing

agreement and protocol approval from the institutional review board committee. The study-specific analyzable dataset is, therefore, not publicly available. All summary statistics are available in Supplementary Tables.

CODE AVAILABILITY

Code for BASE-niPGT-M is available at https://github.com/Ge-lab-pku/BASE-niPGT-M

REFERENCES

32. Abecasis, G.R. et al. Merlin---rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30, 97-101 (2002).
33. Lander, E.S. and Green, P. Construction of multulocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* 84, 2363-2367 (1987).
34. Housworth, E.A. and Stahl, F.W. Crossover interference in humans. *Am J Hum Genet* 73, 188-197 (2003).
35. Howie, B. et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 44, 955-959 (2012).

ACKNOWLEDGEMENTS

We thank Long Gao, Yuhang Huan, Ye Li, Cuiping Pan, Xiaodan Shi, Wenjie Sun, Zhongwei Wang, Liying Yan, Jingyi Yang, Jiakun Zhang and Nannan Zhang for their help. This work is funded by Beijing Advanced Innovation Center for Genomics at Peking University and Changping Laboratory. H.G. is supported by National Natural Science Fundation of China (No. 11971037 and T2225001). This study was also supported by the National Key R&D Program of China (2023YFC2705604), the National Science Foundation of China (82071721 and 82371706) and the special fund of the National Clinical Key Specialty Construction Program, P. R. China (2023).

AUTHOR CONTRIBUTIONS

X.S.X., J.Q., L.H. and S.L. conceived and supervised the research. J.H. and J.J collected SCM samples and performed PGT-M experiments of biopsy samples. L.H. designed and performed the niPGT-M experiments with the help from G.Y., D.C., Y.T. and K.L.. R.Z. and H.G. performed the lineage analysis for niPGT-M with the help from L.H., Q.W., Z.W., W.S. and X.C. Y.Z. and Y.X. performed the data analysis for niPGT-P with the help from L.H. and G.C.. L.H., H.G., Y.X. and Y.Z. wrote the initial manuscript with the help from Q.W. and G.Y. All authors approved the final version of the manuscript.

COMPETING INTERESTS

L.H., H.G., R.Z. and X.S.X. are coinventors on patent application PCT/CN2023/125605 that includes the experimental discovery and BASE-niPGT-M in this manuscript. S.L., Y.Z. and Y.X. are current employees of Yikon Genomics.

FIGURE CAPTIONS:

Figure 1 niPGT-M workflow using SCM samples.

A. Haplotype pre-phasing of parents is conducted using sequencing results from the proband, grandparents, or at least a discarded embryo within the family. Fertilized embryos are cultured, and immediately after transferring the blastocysts to another dish for biopsy, SCMs are removed from each of the residual drops. A few trophectoderm cells from each hatching blastocyst were biopsied and sent to an external PGT lab for PGT-M testing, serving as the gold standard. All the SCM samples underwent scWGA and NGS, followed by SNP calling.

B. The SNP data of SCM samples, together with the phased haplotypes of the parents, are input into the BASE-niPGT-M method. After iteratively

estimating the MCC rate and haplotype states of each SNP, the log-likelihood ratio of inheriting disease-carrying chromosomes versus disease-free chromosomes was calculated. niPGT-M diagnose is performed based on the characteristics of the log-likelihood ratio curve as a function of distance to the disease-causing mutation site.

Figure 2 An illustrative case for our niPGT-M method.

A. The SNP distribution and the linkage-analysis results of cfDNA in an illustrative case. This case involved an autosomal dominant genetic disease, where the husband carried a genetic variant for Marfan syndrome, a heritable connective tissue disorder. All eight SCM samples were collected, and cfDNA was amplified using the improved LIANTI method.

B. The curves of log-likelihood ratio and diagnosis results using BASE-niPGT-M. The diagnosis results of five SCM samples resulted in a classification of 'Highly confident', while one SCM sample was classified into 'Moderately confident' due to our conservative diagnosis strategy. The remaining two SCM samples were classified into 'Undetermined' due to either a very low density of SNPs or an opposite inference of chromosome upstream and downstream of the disease-carrying mutation site. All these results were consistent with those of trophectoderm biopsies, demonstrating no misclassifications.

Figure 3 Whole genome reconstruction of the embryo from SCM data.

A. Schematic illustration of the procedure for whole genome reconstruction. This involves four steps. Firstly, parental haplotypes were phased with sibling embryos or other family members. Secondly, using pre-phased parental haplotype information and selected informative SNPs, haplotype scaffold of SCM was created. Thirdly, pedigree-based imputation was conducted on SCM with the genome information of parents. Finally, whole genome information of SCM was reconstructed based on population-based genotype imputation.

B. Whole-genome SNP density and T2D related SNPs coverage of 10T-

SCM1. The blue triangles were SNPs in PRS model of T2D directly called from raw data; the orange triangles were SNPs imputed from pedigree information; the gray triangles represent SNPs imputed from population information.

C. Haplotypes of SCM samples with different MCC and corresponding TE biopsy samples. Different haplotypes were showed with different colors. The uncolored areas denoted regions where reconstruction failed.

D. Whole-genome coverage of SCM samples before and after our construction. See Supplementary Table S6 for the data.

E. Haplotype and genotype concordances between SCM samples and the corresponding discarded embryos or TE biopsy samples. See Supplementary Table S6 for the data.

Figure 4 T2D associated genotype imputation and polygenic risk score calculation.

A. Distribution of reconstructed genotypes of T2D associated SNPs in SCM samples. T2D related genotypes in SCM were obtained from direct detection by NGS in cfDNA from SCM (blue bar), pedigree-based genotype imputation (green bar) and population-based genotype imputation (yellow bar).

B. Genotype concordance of T2D associated SNPs in SCM versus corresponding TE or discarded embryo samples. Genotype 0/0, 1/1 and 0/1 respectively denote homozygous reference allele, homozygous alternate allele and heterozygous genotypes.

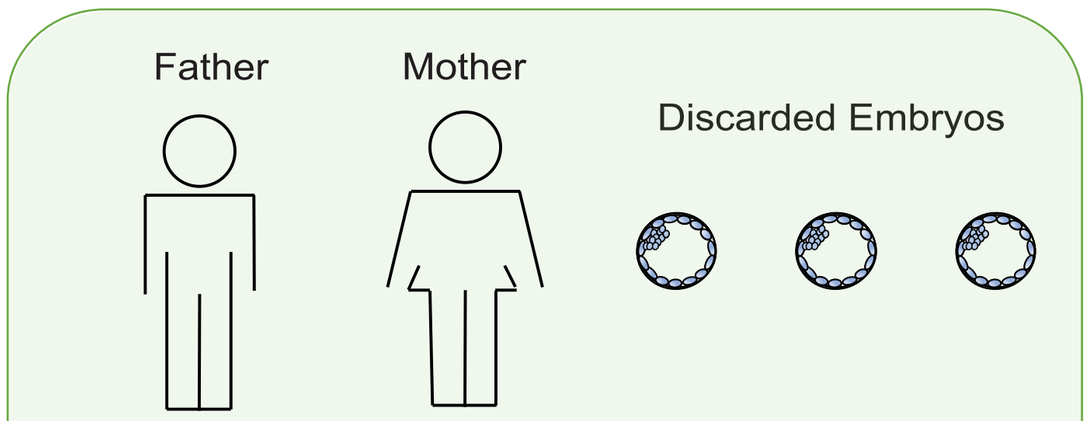
C. PRS consistency of T2D between SCM and paired TE/discarded embryo or amniotic fluid samples.

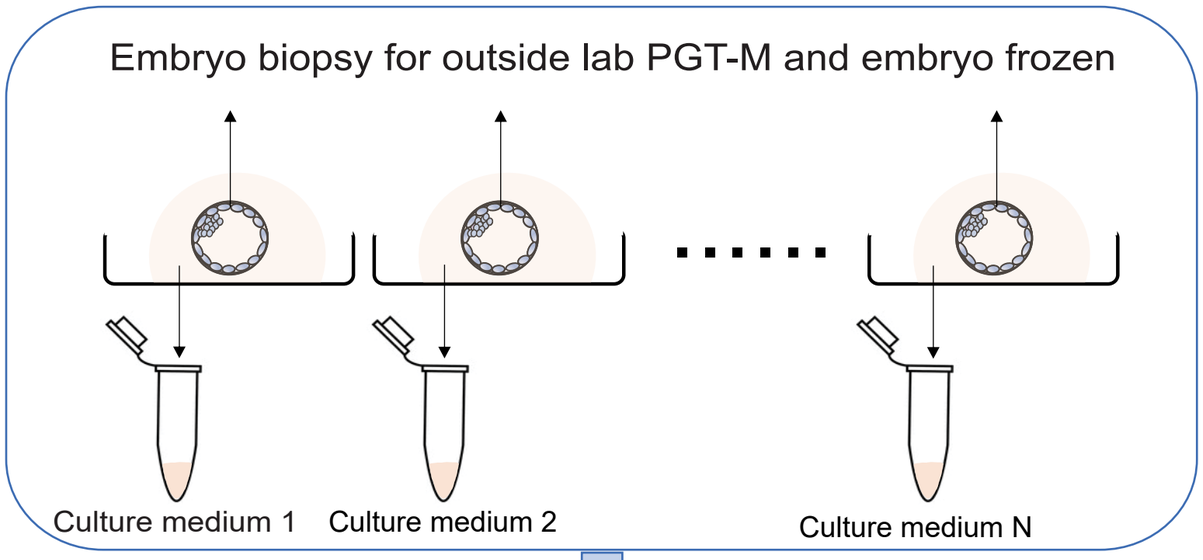
D. PRS percentile of T2D between SCM and paired TE/discarded embryo or amniotic fluid samples. 20 SCM samples matched with TE samples (green round dots), 5 SCM samples matched with discarded embryo samples (blue rhombic dots), and 2 of the 20 TE biopsied embryos had corresponding amniotic fluid samples, also be compared with the corresponding SCM samples (red triangular dots).

See Supplementary Table S7 for the data shown in this figure.

Table 1. Summarized results of BASE-niPGT-M using all matched biopsy/disgarded embryo. We categorized the results into three groups: autosomal dominant disease, autosomal recessive disease, and X-linked recessive disease. For autosomal recessive diseases, we compared each allele of the embryos. The acronyms HC, MC, PO, UN, and RO stand for Highly Confident, Moderately Confident, Possibly Confident, Undetermined, and Ruled Out, respectively. Overall, we successfully detected 95 out of 126 alleles, achieving a 75.4% average report rate. All detected results were concordant with the gold standard, demonstrating a 100% accuracy rate of BASE-niPGT-M. See Supplementary Table S5 for more detailed data.

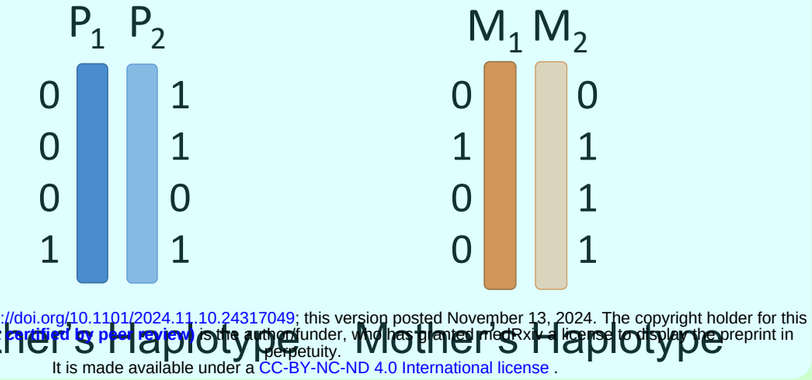
A





Amplified by improved LIANTI
+ NGS sequencing

SNP Calling and Prephasing



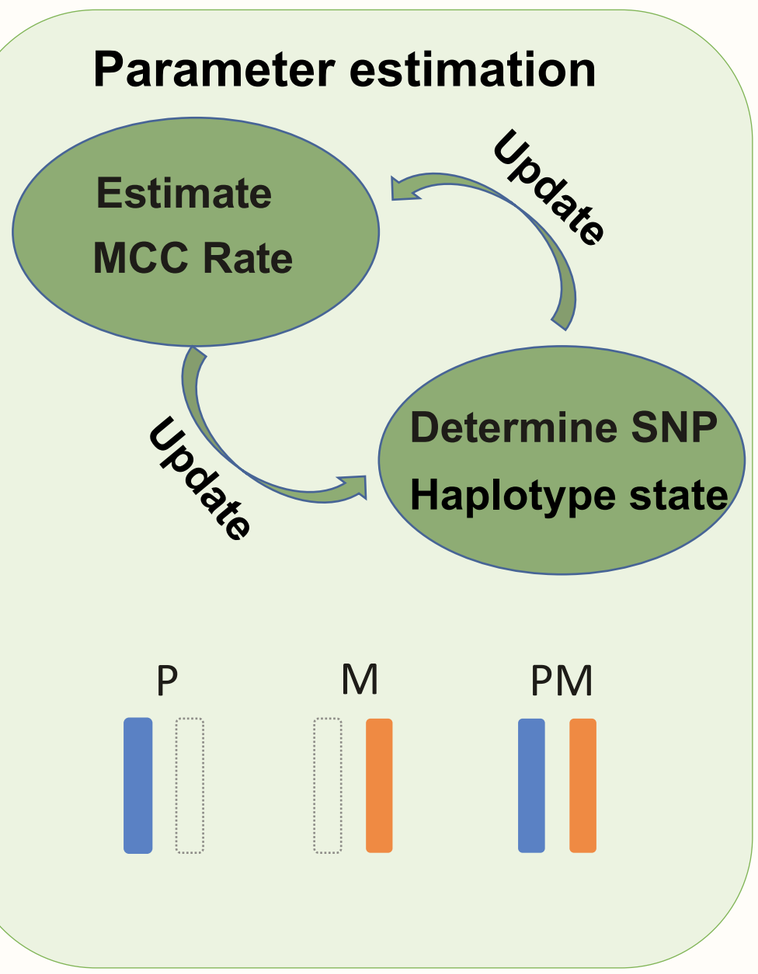
SNP Calling and Allele Depth

	Allele Depth
SNP_1	$[AD_0, AD_1]$
SNP_2	$[AD_0, AD_1]$
.....	.....

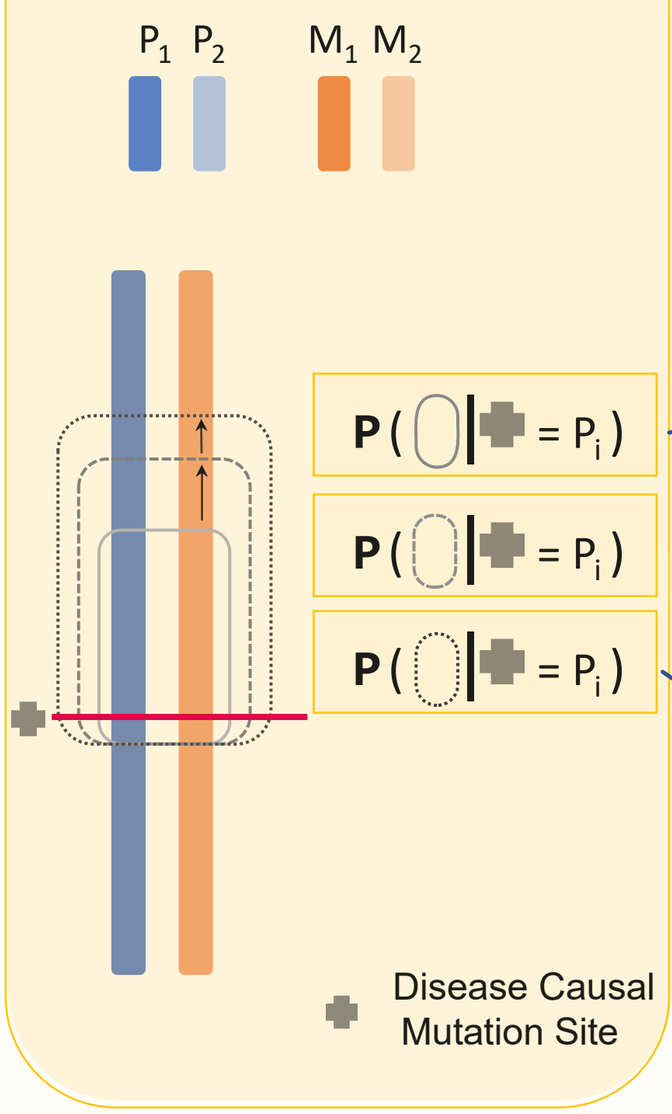
	Allele Depth
SNP_1	$[AD_0, AD_1]$
SNP_2	$[AD_0, AD_1]$
.....	.....

B

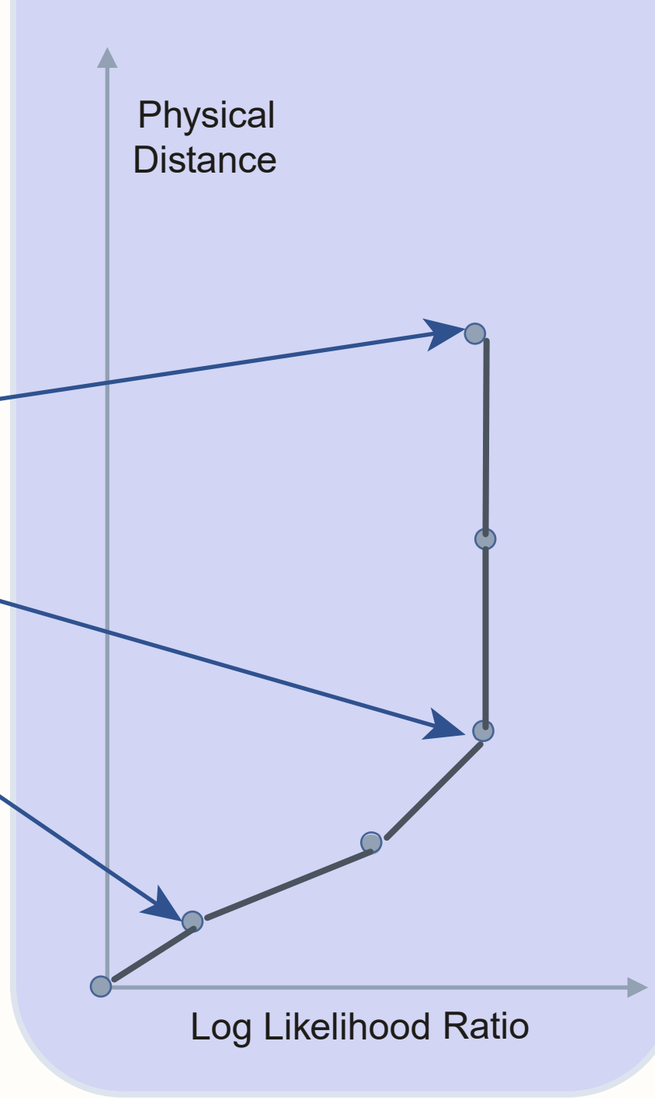
Parameter estimation



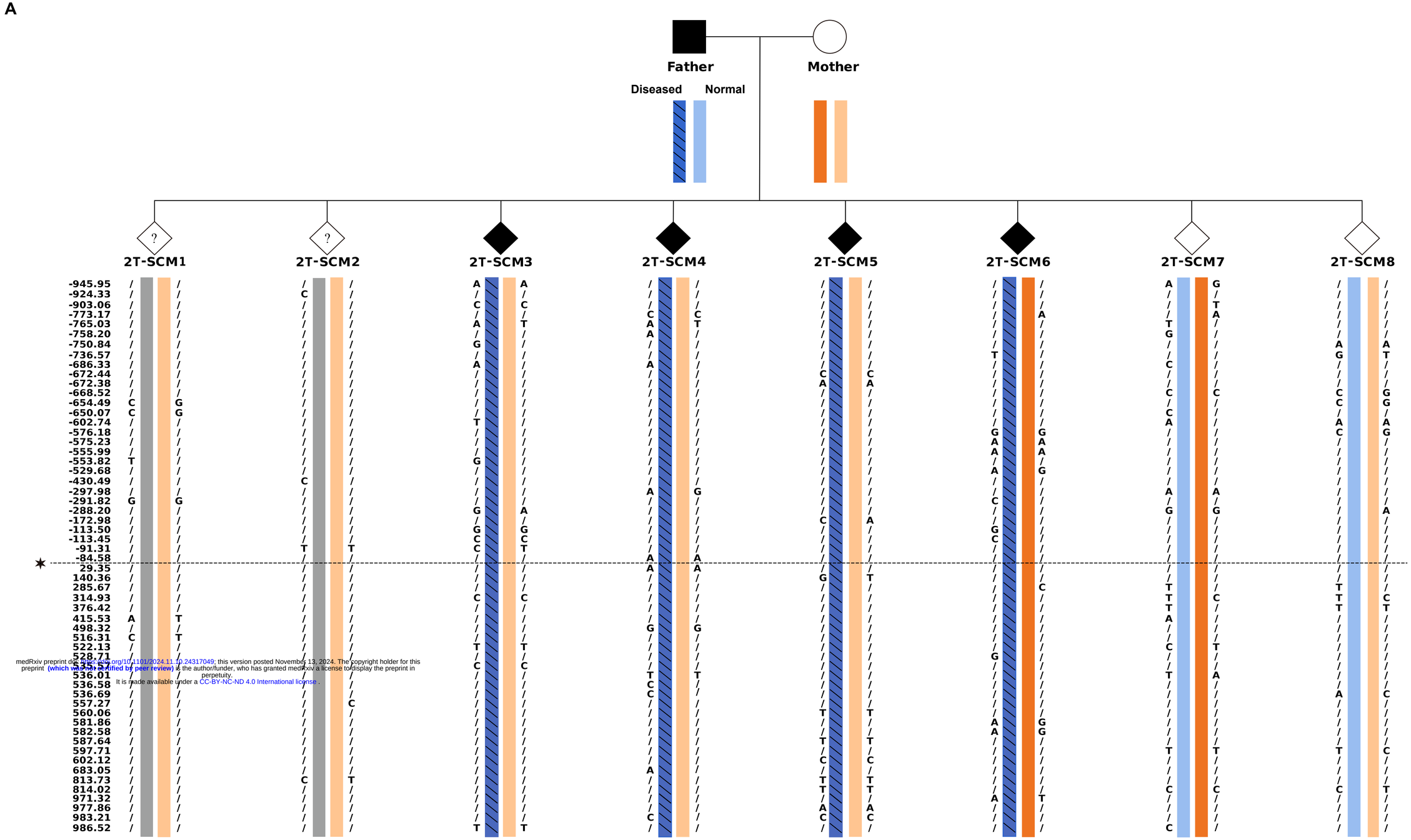
Likelihood calculation



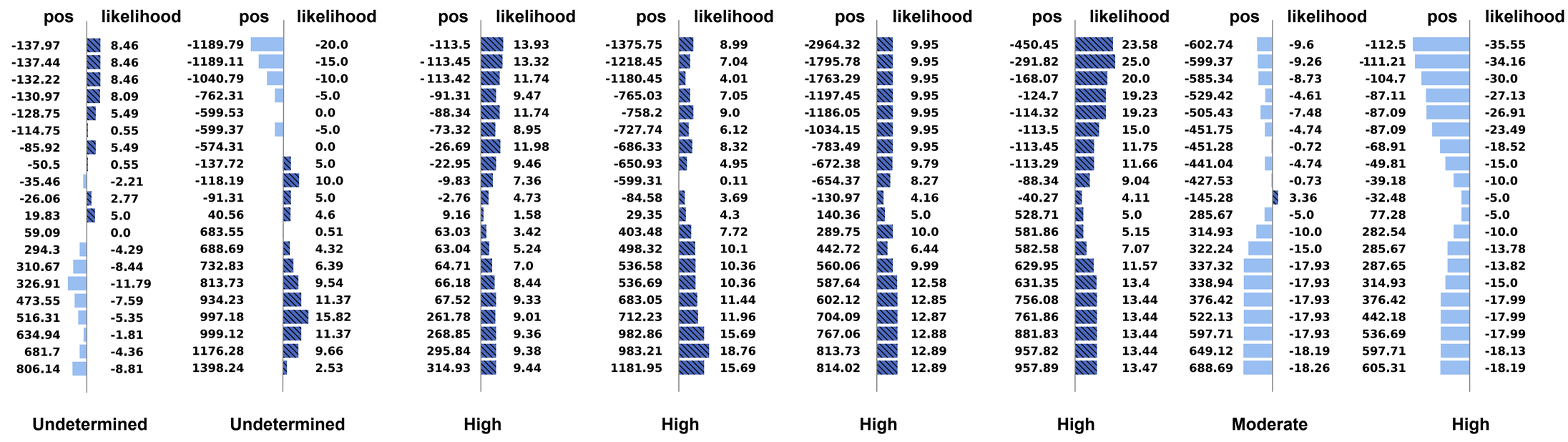
PGT-M Diagnose

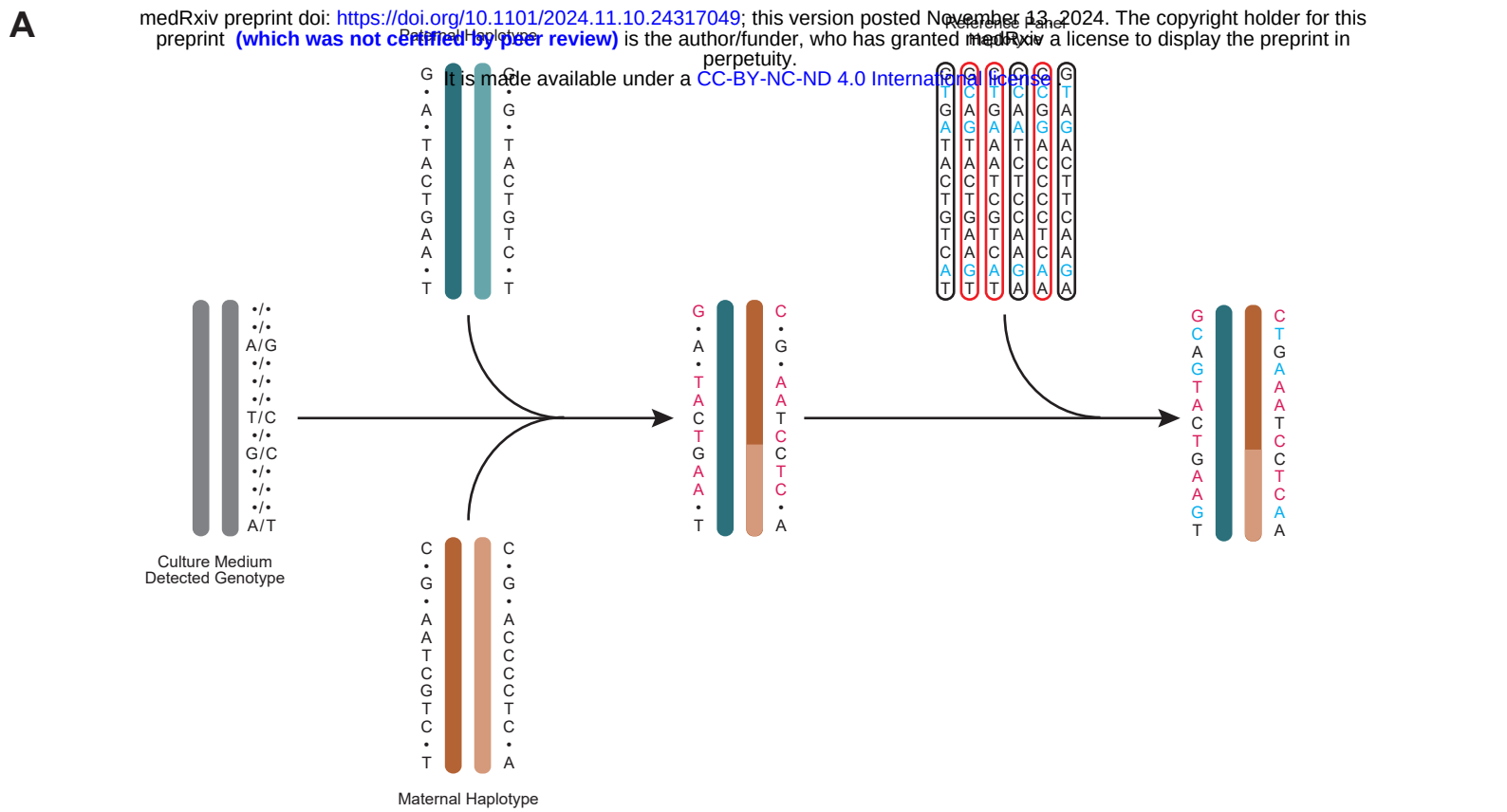


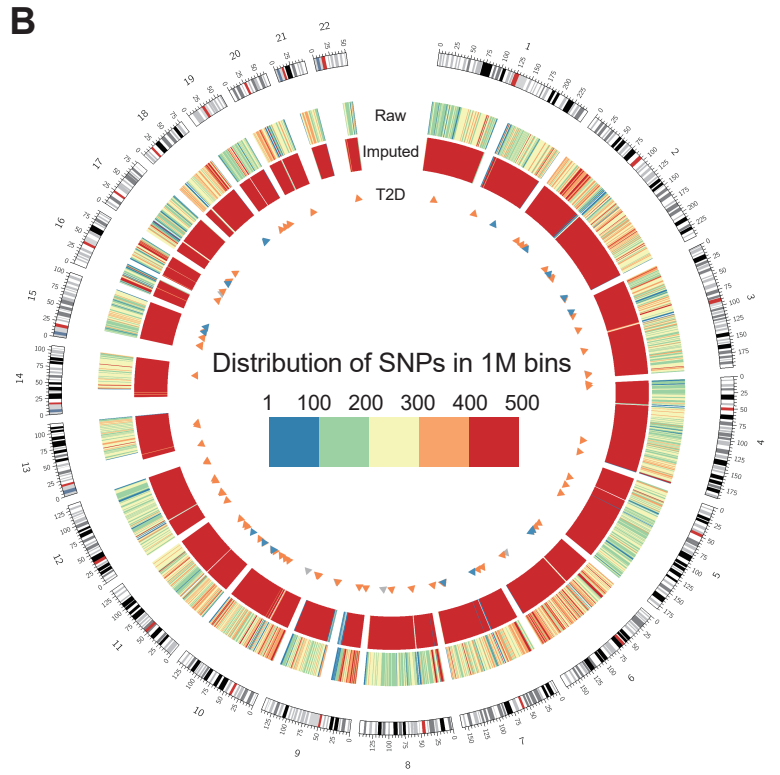
A

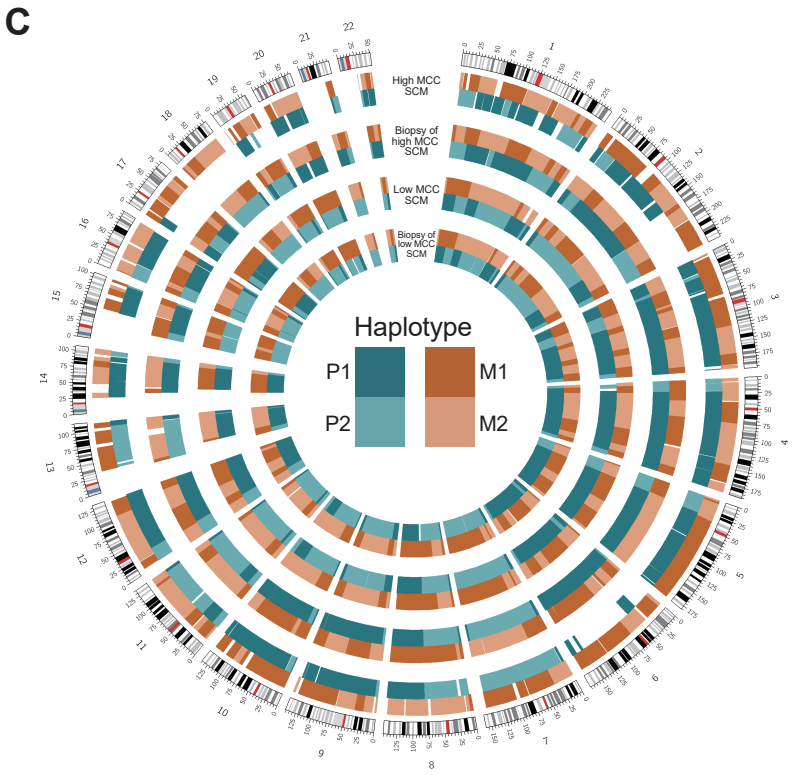


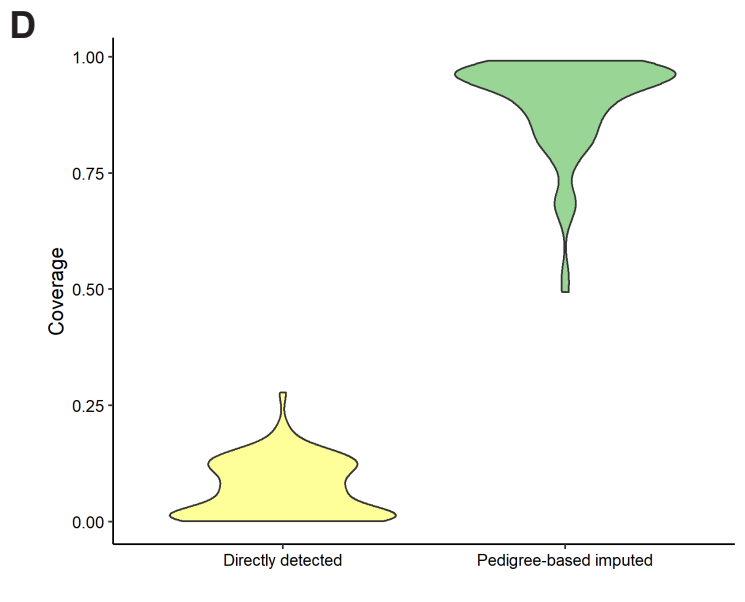
B

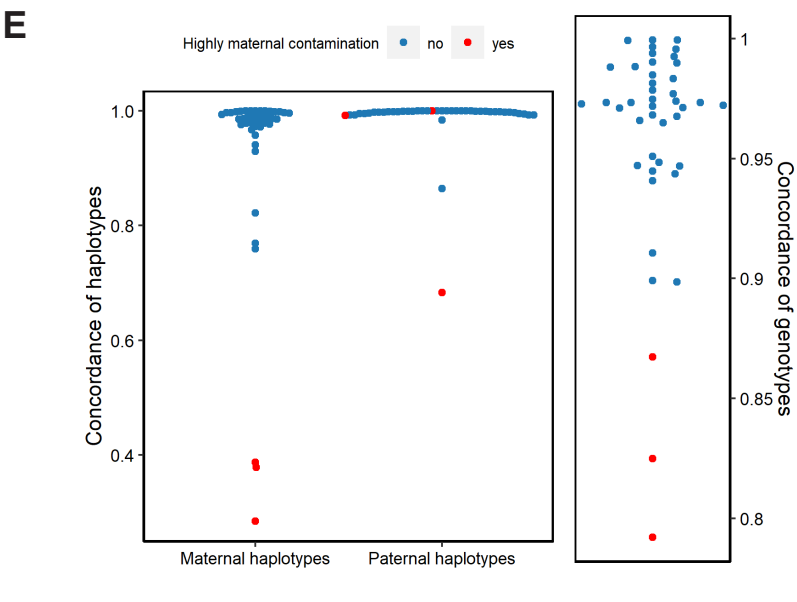


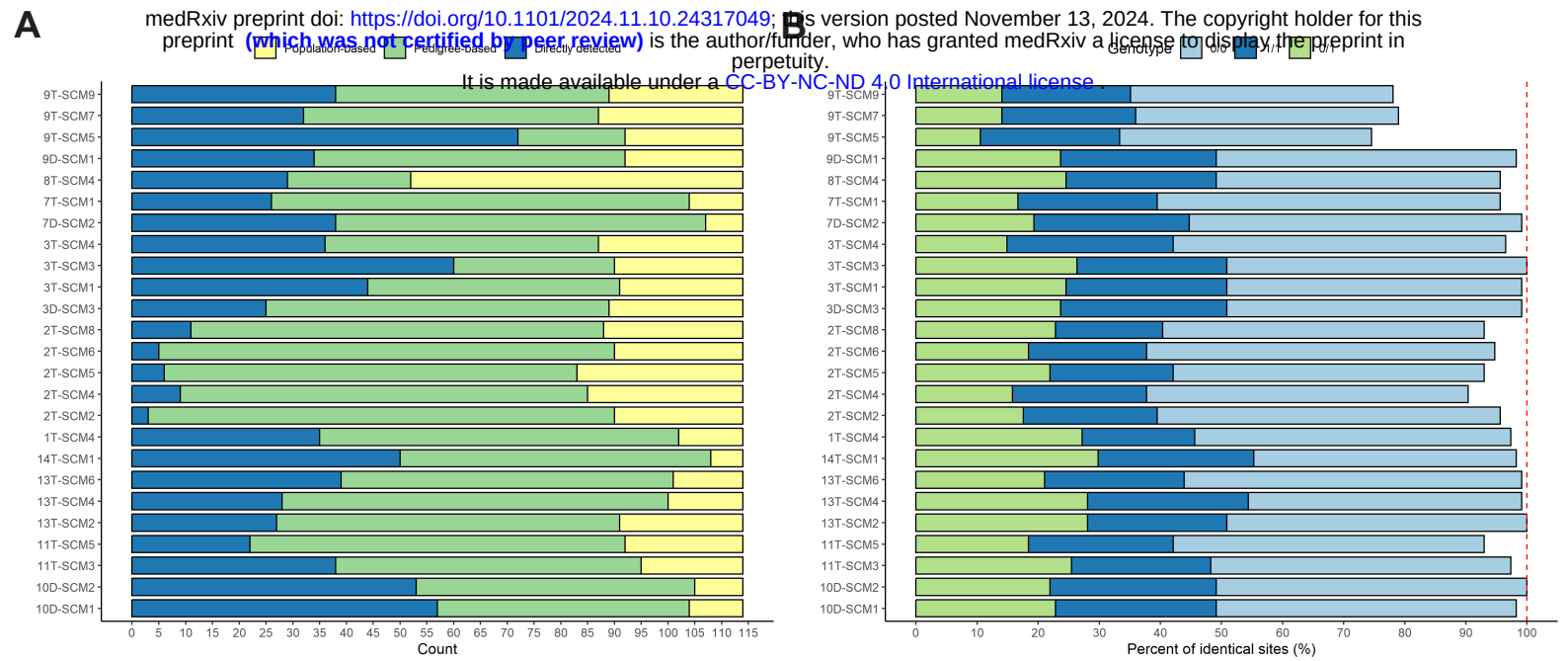


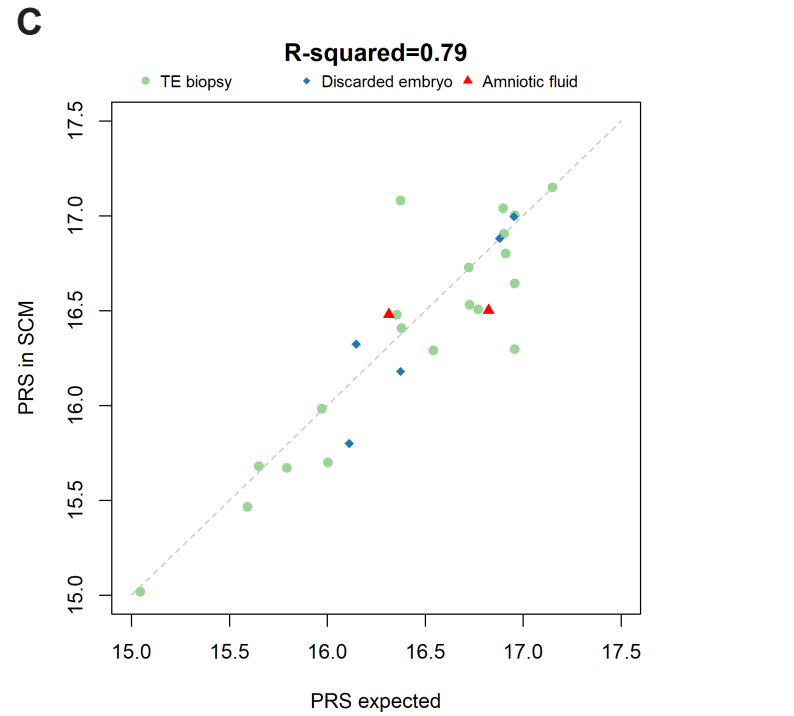


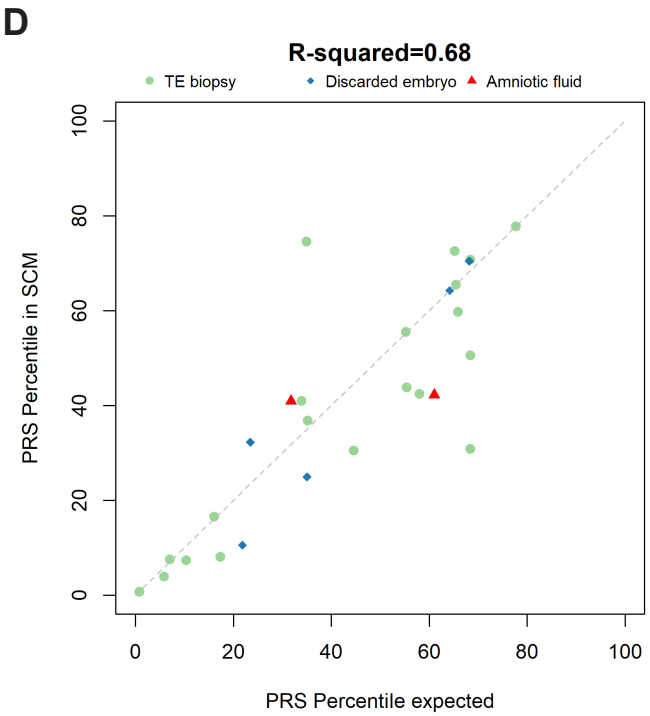












Type of Disease	Family Number	Parent	# of Samples	# of Reportable Samples	# of Correct Samples	HC	MC	PO	UN	RO	Report Rate(%)
Autosomal Dominant Disorder	1	M	6	2	2	2	0	0	0	4	33.3%
	2	P	8	6	6	5	1	0	2	0	75.0%
	3	P	8	7	7	5	0	2	0	1	87.5%
	4	M	11	9	9	3	6	0	0	2	81.8%
	5	P	4	3	3	1	2	0	0	1	75.0%
	6	P	4	4	4	2	0	2	0	0	100.0%
	7	P	4	3	3	1	0	2	1	0	75.0%
Total	\	\	45	34	34	19	9	6	3	8	75.6%
P total		\	28	23	23	14	3	6	3	2	**82.1%**
M total		\	17	11	11	5	6	0	3	6	**64.7%**
Autosomal Reccesive Disorder	8	P	5	3	3	2	1	0	0	2	60.0%
		M	5	2	2	1	1	0	1	2	40.0%
	9	P	13	9	9	5	2	2	2	2	69.2%
		M	13	10	10	6	2	2	1	2	76.9%
	10	P	5	3	3	2	1	0	1	1	60.0%
		M	5	4	4	3	0	1	0	1	80.0%
	11	P	8	7	7	2	4	1	1	0	87.5%
		M	8	7	7	4	2	1	1	0	87.5%
	12	P	5	3	3	3	0	0	1	1	60.0%
		M	5	4	4	1	1	2	0	1	80.0%
Total alleles	\	\	72	52	52	29	14	9	8	12	72.2%
P total		\	36	25	25	14	8	3	5	6	**69.4%**
M total		\	36	27	27	15	6	6	3	6	**75.0%**
X-Reccessive Disorder	13	M	6	5	5	5	0	0	1	0	83.3%
	14	M	3	3	3	3	0	0	0	0	100.0%
Total/ **M total**	\	\	9	8	8	8	0	0	1	0	**88.9%**
Total alleles			126	94	94	56	23	15	12	20	**74.6%**
P total			64	48	48	28	11	9	8	8	75.0%
M total			62	46	46	28	12	6	7	12	74.2%

ADDITIONAL INFORMATION

Extended data

Extended Data Fig. 1 DNA size distribution in Spent Embryo Culture Medium (SCM)

Size distributions of DNA molecules in six spent culture medium samples with different DNA fractions (Fragment Analyzer System). The horizontal axis represents the range of DNA fragments detected. 1bp is the lower marker, and 6000bp is the upper marker. The vertical axis represents a unit of relative fluorescence units. The first peak of insert fragments was approximately 180bp, with subsequent peaks spaced at 180bp intervals, corresponding to the length of DNA wrapped around a nucleosome. SCM 2, SCM4, SCM5 and SCM6 showed varying levels of apoptosis, suggesting different degrees of DNA degradation. Additionally, SCM3 failed to produce an effective library due to its low cfDNA content, and SCM1 did not exhibit the characteristic apoptosis signals. See Supplementary Notes S1 for more experimental details.

Extended Data Fig. 2 Genome Coverage of Spent Embryo Culture Media

Genome coverage of all SCM samples.

Extended Data Fig. 3 Illustration of large fragment loss of haplotype (LFLoH) in one sample of SCM.

Blue and orange bars denote haplotypes respectively from paternal and maternal chromosome 1. Black region represents no genetic information observed in specific haplotypes.

Extended Data Fig. 4 Strategy for informative SNPs selection.

0 and 1 respectively denote reference and alternative allele of variant. All the informative SNPs selected are listed.

Extended Data Fig. 5 Estimated MCC rates.

The top and bottom panels illustrate the MCC rates for autosomal dominant and autosomal recessive family samples, respectively. Circles represent pathogenic loci inherited paternally, while triangles denote maternal inheritance. Hollow symbols indicate the original MCC rates, and solid symbols represent the final MCC rates. Blue symbols correspond to diagnosable samples, whereas red symbols denote undiagnosable samples. See Supplementary Table S4 for the data.

Extended Data Fig. 6 Improved protocol from LIANTI Method.

In this study, we retained the LIANTI single-cell amplification method, which involves (a) : 1) Transposon-assisted DNA fragmentation; 2) Gap-filling and extension; 3) Reverse transcription; 4) Second-strand synthesis; 5) Library construction and sequencing. To adapt this method for culture medium samples, we made the following improvements (b): 1) Adjusted the lysis system with a 10.8 μl sample volume and increased QIAGEN Protease (Qp) concentration to 35 mg/ml; 2) Doubled the transposon concentration; 3) Added a 19bp transposon ME sequence complementary primer to enhance reverse transcription efficiency; 4) Included 10 additional PCR cycles in second-strand synthesis to improve the conversion of single-stranded cDNA to double-stranded DNA.

Extended Data Fig. 7 Improved LIANTI amplification is more suitable for amplifying culture medium samples.

Clinical collection of SCM (spent culture medium) yields limited volumes, and samples are precious. In this study, we utilized the BJ cell line to simulate the embryo culture process. a) Size distributions of DNA molecules in BJ15 and BJ17 with different DNA fractions (Fragment Analyzer System). b) Genome coverage analysis of BJ15 and BJ17, which were amplified using Improved LIANTI, MALBAC, and MDA. c) SNP number analysis of BJ15 and BJ17, which were amplified using Improved LIANTI, MALBAC, and MDA. Genome coverage and SNP number

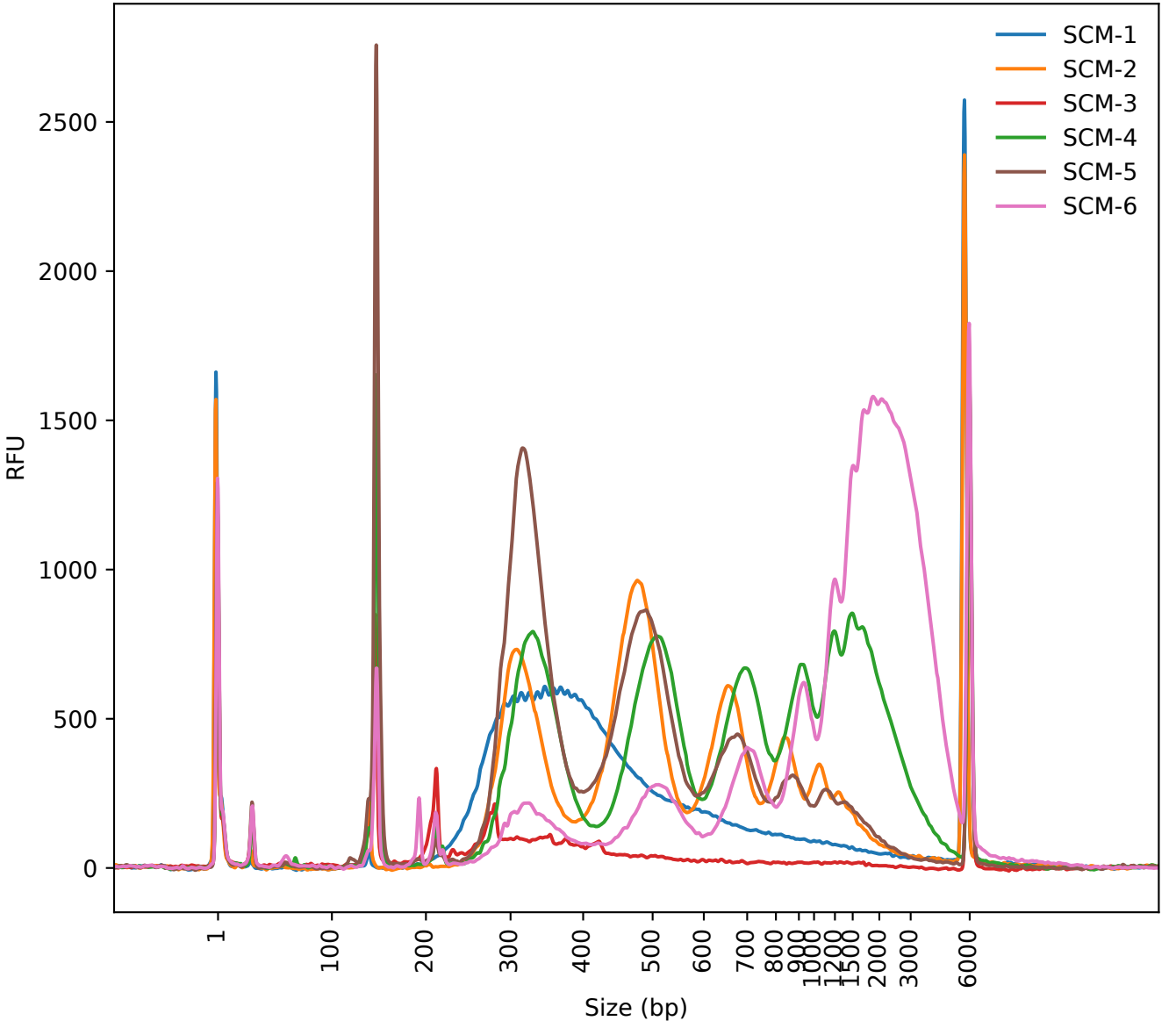
normalization were carried out with Bulk sample. The sequencing depth for bulk genomic DNA (gDNA) exceeds 30X, while for amplified products, the sequencing depth exceeds 10X. BJ15: BJ cell lines medium collected on Day5; BJ17: BJ cell lines medium collected on Day7 (see Supplementary Table S2).

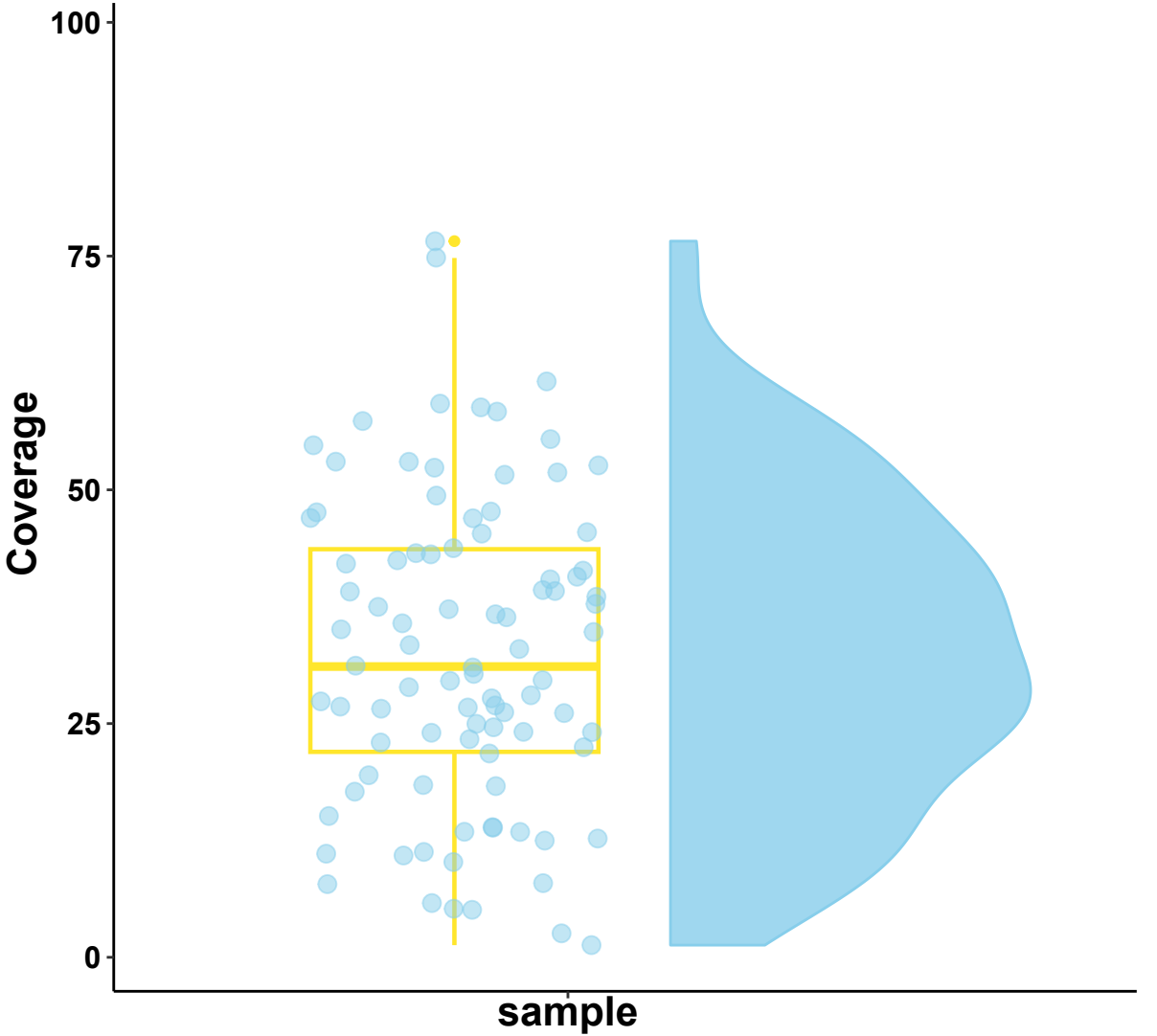
Extended Data Fig. 8 The workflow of BASE-niPGT-M.

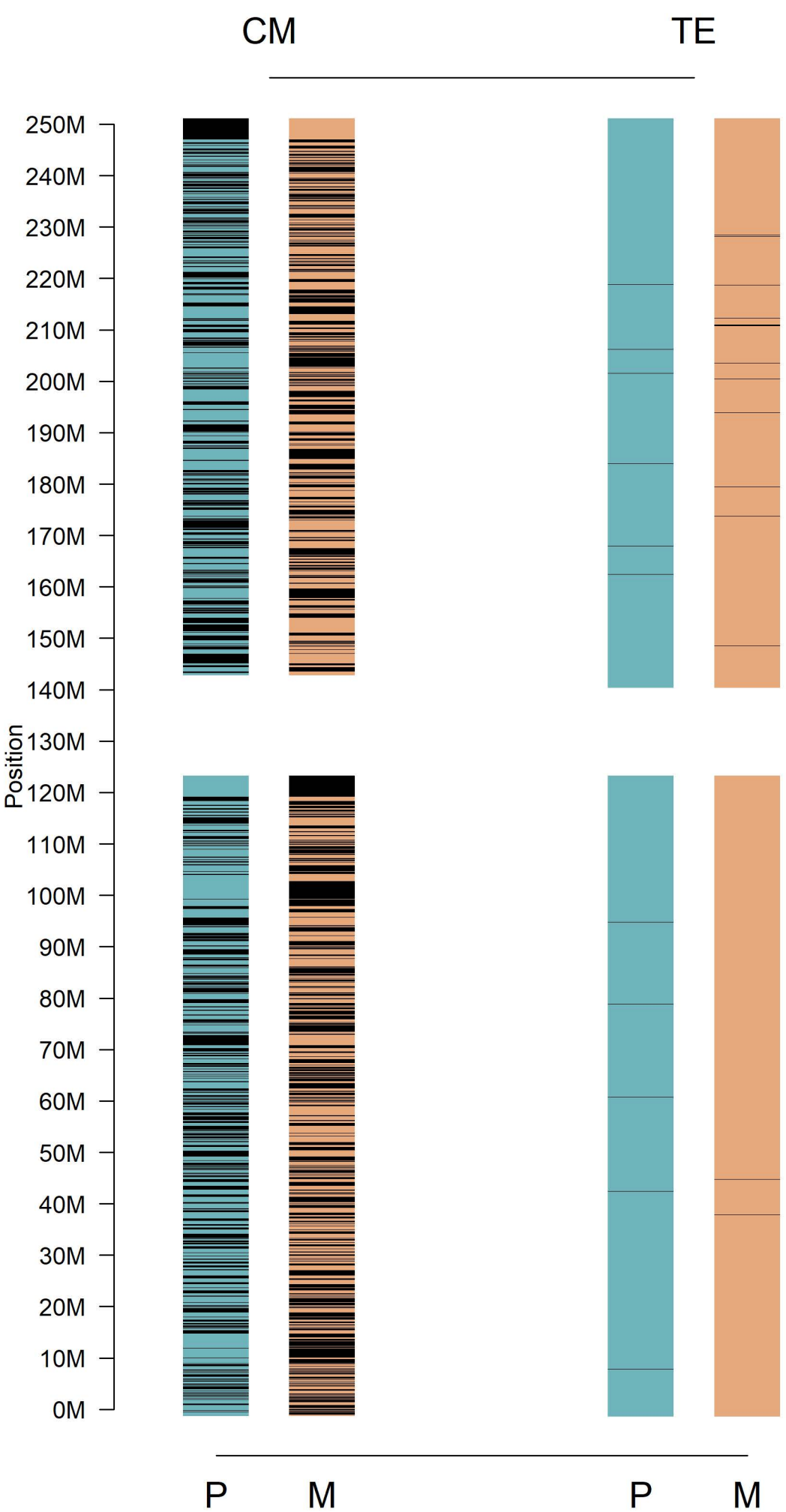
The Bayesian model uses the sequencing data of SCM and the phased haplotypes of the parents as input. By iteratively estimating the MCC rate and haplotype states of each SNP, it calculates the log-likelihood ratio of inheriting disease-carrying chromosomes versus disease-free chromosomes as a function of distance to the disease-causing mutation site. This curve, with its distinct characteristics, allows classification of the results into four categories: Highly Confident, Moderately Confident, Possibly confident, and Undetermined.

Extended Data Fig. 9 An example illustrating the workflow of pedigree-based whole genome reconstruction.

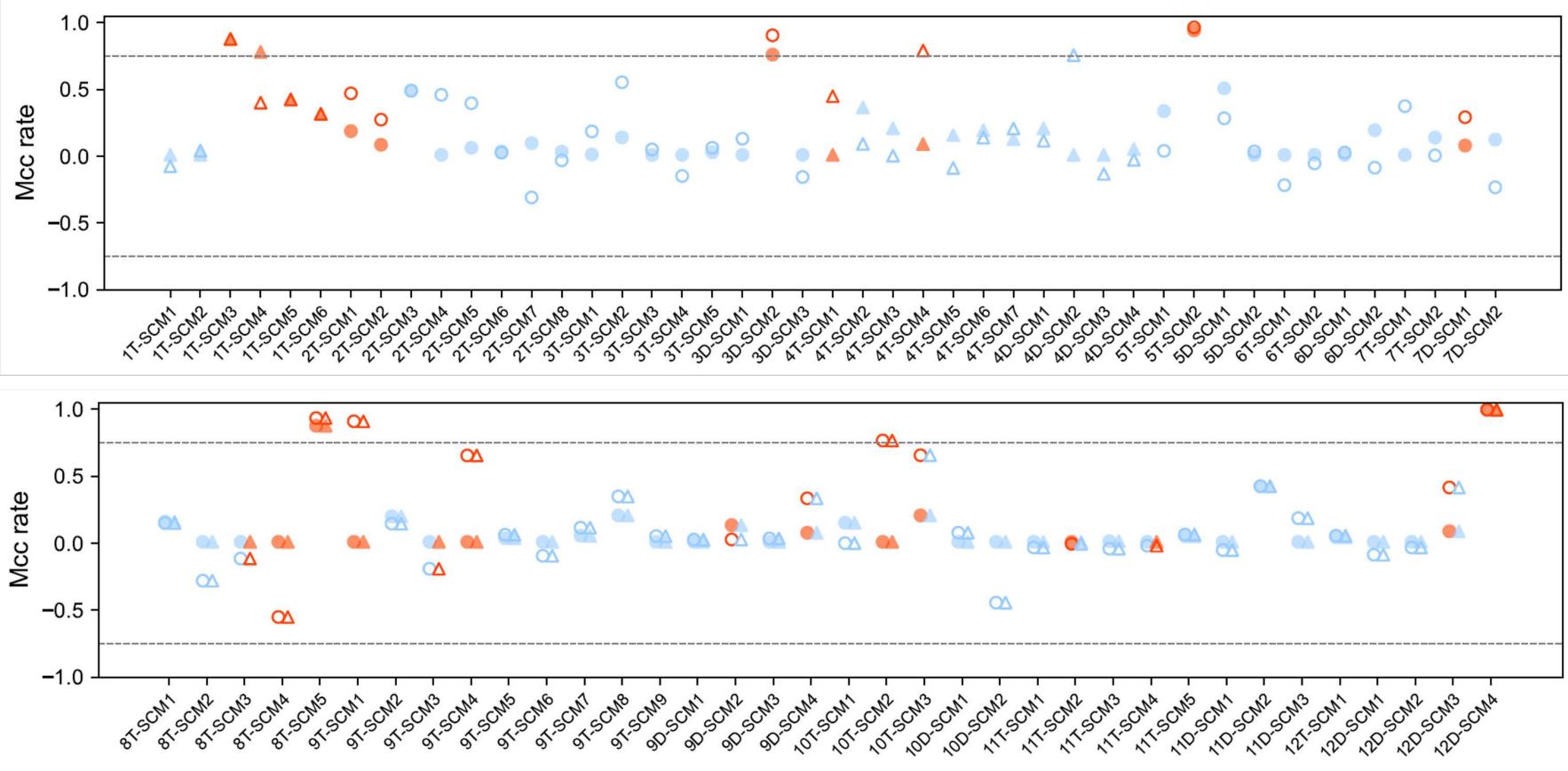
In the initial called genotype of the SCM, red numbers indicate the informative alleles, which can identify the inherited allele from one parent. Black regions represent the LFLoH pattern, characterized by consecutive SNPs missing the same haplotype. Blue letters denote the imputed haplotypes on non-informative SNPs, determined using the HMM method with genetic distance as the transition probability. Green letters indicate corrected haplotypes from the raw phasing results by the HMM method, with incorrect ones shown in red. Green numbers show the imputed allele or genotype of the genome.


SCM-1
SCM-2
SCM-3
SCM-4
SCM-5
SCM-6
RFU
2500
2000
1500
1000
500
0
1
100
200
300
400
500
600
700
800
900
1000
1200
1500
2000
3000
6000
Size (bp)

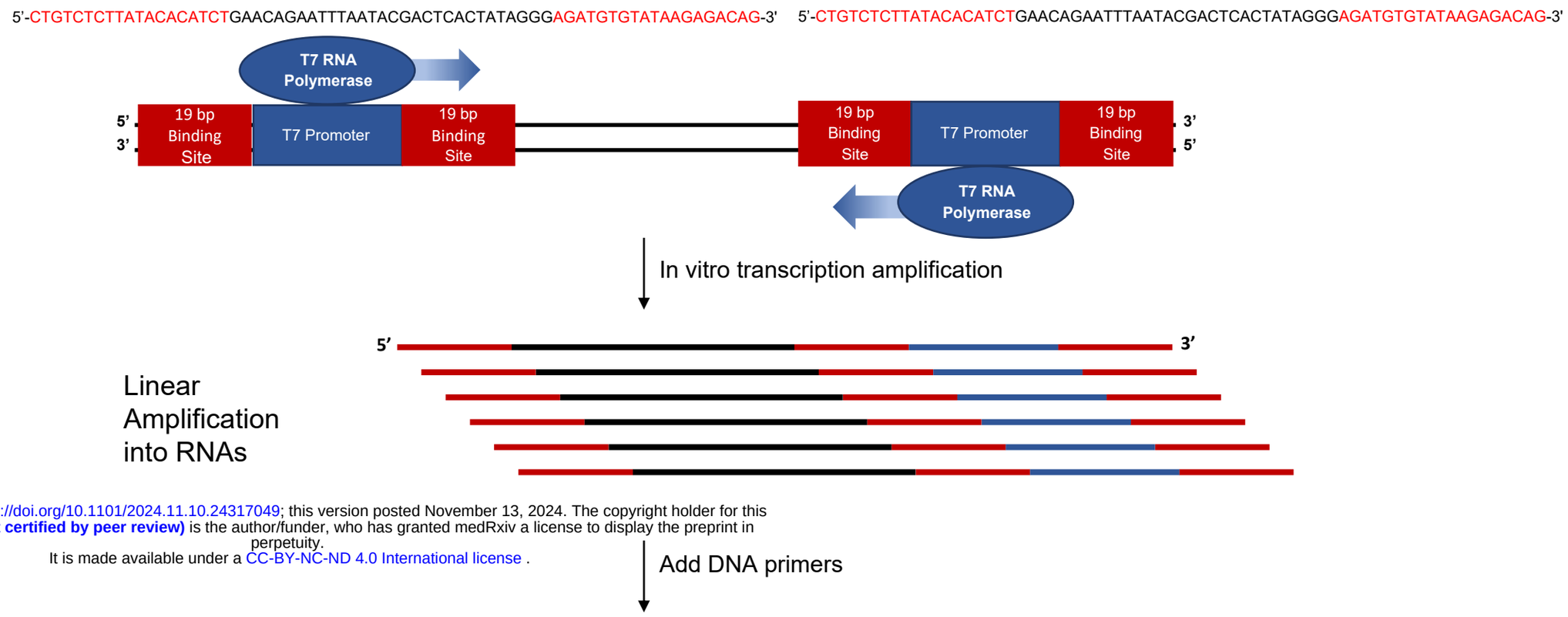

100
75
50
25
0
Coverage
sample

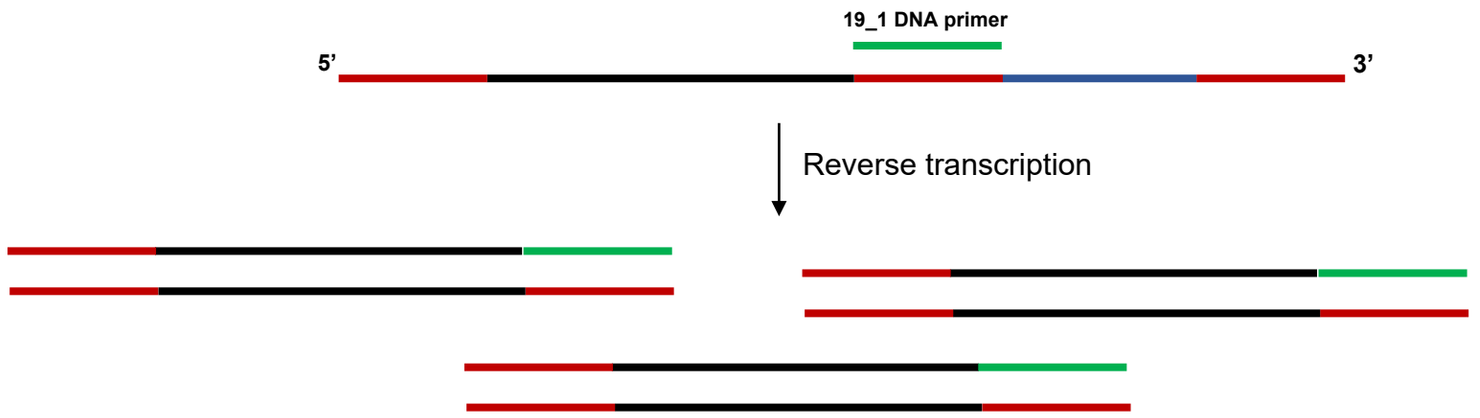


Paternal genotype	Maternal genotype	SCM genotype	Key allele	SNP type
0/1	0/0	0/1	1	Paternal informative SNP
0/1	0/0	1/1	1	Paternal informative SNP
0/1	1/1	0/1	0	Paternal informative SNP
0/1	1/1	0/0	0	Paternal informative SNP
0/0	0/1	0/1	1	Maternal informative SNP
0/0	0/1	1/1	1	Maternal informative SNP
1/1	0/1	0/1	0	Maternal informative SNP
1/1	0/1	0/0	0	Maternal informative SNP

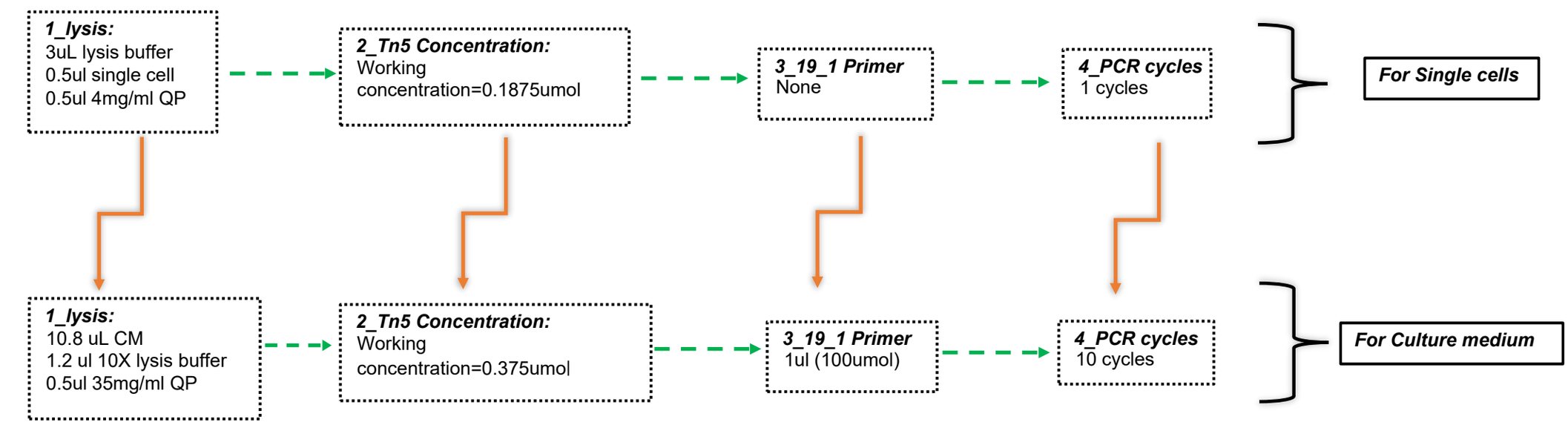

Mcc rate
1.0
0.5
0.0
−0.5
−1.0
1T-SCM1
1T-SCM2
1T-SCM3
1T-SCM4
1T-SCM5
1T-SCM6
2T-SCM1
2T-SCM2
2T-SCM3
2T-SCM4
2T-SCM5
2T-SCM6
2T-SCM7
2T-SCM8
3T-SCM1
3T-SCM2
3T-SCM3
3T-SCM4
3T-SCM5
3D-SCM1
3D-SCM2
3D-SCM3
4T-SCM1
4T-SCM2
4T-SCM3
4T-SCM4
4T-SCM5
4T-SCM6
4T-SCM7
4D-SCM1
4D-SCM2
4D-SCM3
4D-SCM4
5T-SCM1
5T-SCM2
5D-SCM1
5D-SCM2
6T-SCM1
6T-SCM2
6D-SCM1
6D-SCM2
7T-SCM1
7T-SCM2
7D-SCM1
7D-SCM2
Mcc rate
1.0
0.5
0.0
−0.5
−1.0
8T-SCM1
8T-SCM2
8T-SCM3
8T-SCM4
8T-SCM5
9T-SCM1
9T-SCM2
9T-SCM3
9T-SCM4
9T-SCM5
9T-SCM6
9T-SCM7
9T-SCM8
9T-SCM9
9D-SCM1
9D-SCM2
9D-SCM3
9D-SCM4
10T-SCM1
10T-SCM2
10T-SCM3
10D-SCM1
10D-SCM2
11T-SCM1
11T-SCM2
11T-SCM3
11T-SCM4
11D-SCM1
11D-SCM2
11D-SCM3
12T-SCM1
12D-SCM1
12D-SCM2
12D-SCM3
12D-SCM4

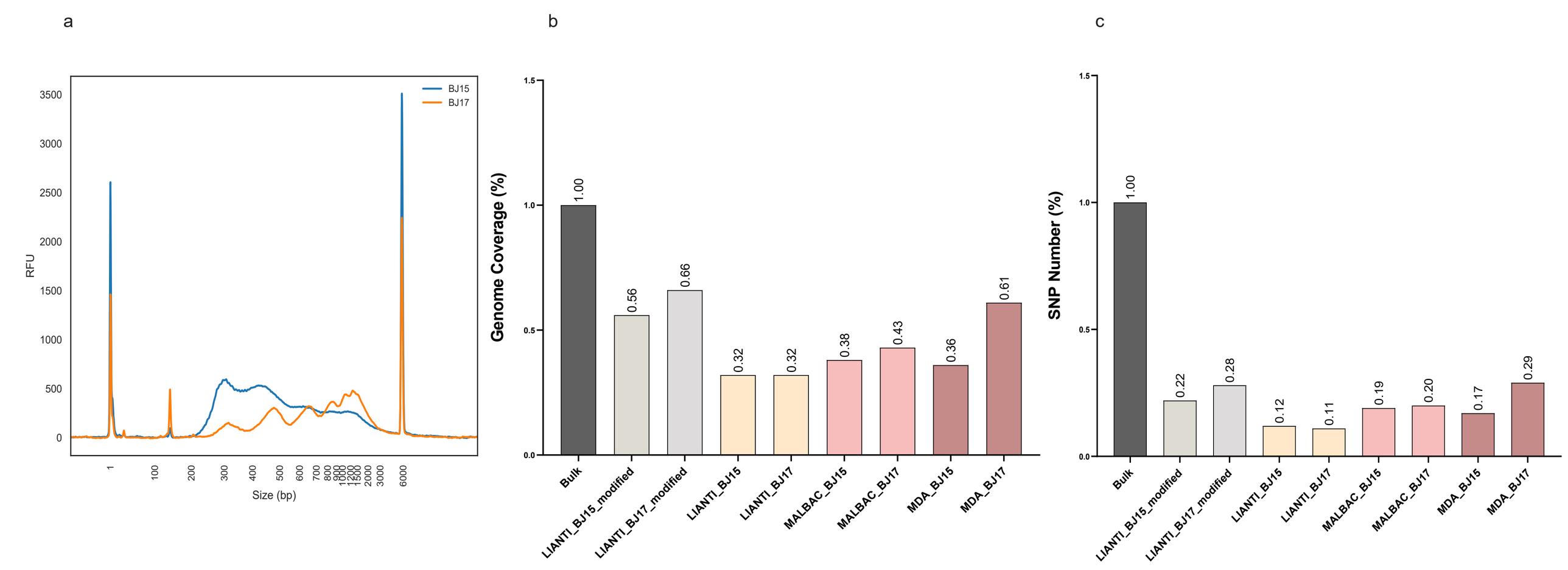
a





b

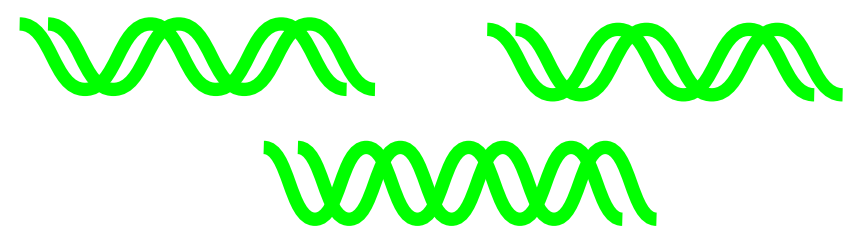



a
BJ15
BJ17
RFU
3500
3000
2500
2000
1500
1000
500
0
1
100
200
300
400
500
600
700
800
900
1000
1100
1200
1500
2000
3000
6000
Size (bp)
b
Genome Coverage (%)
1.5
1.0
0.5
0.0
1.00
0.56
0.66
0.32
0.32
0.38
0.43
0.36
0.61
Bulk
LIANTI_BJ15_modified
LIANTI_BJ17_modified
LIANTI_BJ15
LIANTI_BJ17
MALBAC_BJ15
MALBAC_BJ17
MDA_BJ15
MDA_BJ17
c
SNP Number (%)
1.5
1.0
0.5
0.0
1.00
0.22
0.28
0.12
0.11
0.19
0.20
0.17
0.29
Bulk
LIANTI_BJ15_modified
LIANTI_BJ17_modified
LIANTI_BJ15
LIANTI_BJ17
MALBAC_BJ15
MALBAC_BJ17
MDA_BJ15
MDA_BJ17

BASEni-PGT-M

DENIFITIONS
10MB (Million Base pairs) upstream and downstream of the disease-causing mutation site in the biological sample is defined as the adjacent region of the disease-causing mutation site.
Maternal Cell Contamination rate (MCC rate) is defined as the proportion of maternal DNA in the culture medium of an embryo compared to the total amount of DNA.
Large Fragment Loss of Haplotypes (LFLoH) is defined as the phenomenon that long and contiguous DNA fragments from one or both parents do not enter the culture medium, and the hyplotype status of each SNP is defined to represent whether only DNA from one parent is present in the culture medium.

Amplify the DNA molecules from spent embryo culture medium, sequencing and performing mapping and SNP calling, filtering the SNPs of extremely low quality.



INPUT

Initially estimate the SER, the relative density of SNPs, the MCC rate and the hyplotype status of each SNP.

Recursively estimate and update the MCC rate and the haplotype status of each SNP, using a conservative strategy with the single-SNP likelihood.

Iterately calculate the likelihood of observing the SNP data within regions adjacent to the disease-causing mutation site, combining the likelihood at each single SNP and the recombination probabilities.

$$\mathrm{P}(AD_{n:end}|F_n=i,M_n=j)=L_{ij}^{(n)}\sum_{k,l=1}^{2}\mathrm{P}(AD_{n+1:end}|F_{n+1}=k,M_{n+1}=l)\cdot P(F_{n+1}=k,M_{n+1}=l|F_n=i,M_n=j)$$

AD is the measured Allele Depth data of SNPs.

INPUT

Conduct haplotype pre-phasing, for the blood samples of both parents or solely the disease-carrying parent.

Incorporate the maternal cell contamination (MCC) rate and haplotype status into the likelihood of the four possible inheritance scenarios at each single SNP representing whether the genetic chains are inherited from the paternal or maternal alleles of the parents.

$$\hat{L}_{ij}^{(n)}=\frac{AD_n!}{AD_{0,n}!\cdot AD_{1,n}!}\left[\hat{p}_{ij}^{(n)}\widehat{SER}+\left(1-\hat{p}_{ij}^{(n)}\right)\left(1-\widehat{SER}\right)\right]^{AD_{0,n}}\cdot\left[\hat{p}_{ij}^{(n)}\left(1-\widehat{SER}\right)+\left(1-\hat{p}_{ij}^{(n)}\right)\widehat{SER}\right]^{AD_{1,n}}$$

$\widehat{SER}$ is the estimated sequencing error rate, and $\hat{p}_{ij}^{(n)}$ is the theoretical probability that an ALT gene is obtained by randomly sampling one read at n-th SNP.

Recombination probability derived from database

$$P(F_{n+1}=k,M_{n+1}=l|F_n=i,M_n=j)$$

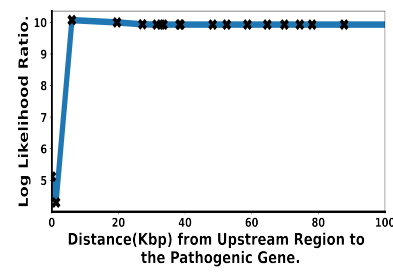
The linkage state of the n-th SNP:
F_n=1(P1), F_n=2(P2), M_n=1 (M1), M_n=2(M2).

Plot the log-likelihood-ratio(LLR) curves as a function of the physical distance to the disease-causing mutation site, using the SNPs within this region.

$$\mathrm{LLR}=\log\left(\frac{\sum_{j=1}^{2}\mathbb{P}\left(\mathrm{AD}_{0:N}\middle|\mathrm{F}_0=1,\mathrm{M}_0=j\right)}{\sum_{j=1}^{2}\mathbb{P}\left(\mathrm{AD}_{0:N}\middle|\mathrm{F}_0=2,\mathrm{M}_0=j\right)}\right)$$

$$\mathrm{LLR}=\log\left(\frac{\sum_{i=1}^{2}\mathbb{P}\left(\mathrm{AD}_{0:N}\middle|\mathrm{F}_0=i,\mathrm{M}_0=1\right)}{\sum_{i=1}^{2}\mathbb{P}\left(\mathrm{AD}_{0:N}\middle|\mathrm{F}_0=i,\mathrm{M}_0=2\right)}\right)$$

One example:



Determine whether the embryo carries the disease-causing chromosome, and providing a level of confidence.

OUTPUT

DIAGNOSE: P1 OR P2? M1 OR M2?

- UNDETERMINED
- POSSIBLY confident
- MODERATELY confident
- HIGHLY confident

Phasing informative SNPs of SCM							Building haplotype of SCM			Reconstructing genotype of SCM		
Paternal haplotype		Maternal haplotype		Initial genotype of SCM	SNP type	Paternal Haplotype of SCM	Maternal Haplotype of SCM	Genetic distance	Paternal Haplotype of SCM	Maternal Haplotype of SCM	Paternal allele of SCM	Maternal allele of SCM	Imputed genotype of SCM
P1	P2	M1	M2										
1	0	0	0	1/1	Haploid	P1		0.12	P1	M1	1	0	1|0
1	1	0	1	0/0	Haploid		M1	0.48	P1	M1	1	0	1|0
0	0	1	0	1/1	Haploid		M1	0.61	P1	M1	0	1	0|1
0	1	1	1	0/0	Haploid	P1		0.85	P1	M1	0	1	0|1
1	1	0	1	1/1	Uncertain			1.02	P1	M1	1	0	1|0
0	1	1	1	0/0	Haploid	P1		1.16	P1	M1	0	1	0|1
1	0	1	1	0/0	Haploid	P2		1.22	P1	M1	1	1	1|1
0	1	1	1	0/0	Haploid	P1		1.34	P1	M1	0	1	0|1
1	1	0	1	0/1	Diploid		M1	1.98	P1	M1	1	0	1|0
1	0	0	0	0/1	Diploid	P1		2.39	P1	M1	1	0	1|0
0	0	1	0	1/1	Haploid		M1	2.64	P1	M1	0	1	0|1
1	1	1	0	0/0	Haploid		M2	2.73	P1	M1	1	1	1|1
0	0	1	0	1/1	Haploid		M1	2.85	P1	M1	0	1	0|1
0	0	1	0	1/1	Haploid			2.96	P1	M1	0	1	0|1
1	0	0	0	0/1	Diploid	P1		3.15	P1	M1	1	0	1|0
0	0	1	0	0/1	Diploid		M1	3.21	P1	M1	0	1	0|1
0	0	0	1	0/1	Diploid		M2	4.37	P1	M2	0	1	0|1
1	0	0	0	0/1	Diploid	P1		4.93	P1	M2	1	0	1|0
0	1	1	0	0/1	Diploid			5.30	?	M2	?	0	./.
1	0	1	1	0/1	Diploid	P2		5.67	P2	M2	0	1	0|1
0	1	0	0	0/1	Diploid	P2		5.69	P2	M2	1	0	1|0
0	0	0	1	1/1	Haploid		M2	5.74	P2	M2	0	1	0|1
0	1	0	0	1/1	Haploid	P2		5.88	P2	M2	1	0	1|0

Supplementary Information

Supplementary Notes

Supplementary Tables