Title

Genome-wide Disease Screening in Early Human Embryos with Primary Template-Directed Amplification

Yuntao Xia[1], Veronica Gonzales-Pena[1], David J Klein[1], Joe J Luquette[2], Liezl Puzon[3], Noor Siddiqui[3], Vikrant Reddy[4], Peter Park[2], Barry R Behr[4], Charles Gawad[1,5]

Department of Pediatrics, Hematology/Oncology division, Stanford University, Palo Alto, CA, 94304[1]

Department of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115[2]

Orchid Health, Palo Alto, CA, 94301[3]

Department of Obstetrics & Gynecology - Reproductive Endocrinology and Infertility, Stanford University, Sunnyvale, CA, 94087[4]

Chan Zuckerberg Biohub, San Francisco, CA, 94158[5]

Abstract

Current preimplantation genetic testing (PGT) enables the selection of embryos based on fetal aneuploidy or the presence a small number of preselected disease-associated variants. Here we present a new approach that takes advantage of the improved genome coverage and uniformity of primary template-directed amplification (PTA) to call most early embryo genetic variants accurately and reproducibly from a preimplantation biopsy. With this approach, we identified clonal and mosaic chromosomal aneuploidy, *de novo* mitochondrial variants, and variants predicted to cause mendelian and non-mendelian diseases. In addition, we utilized the genome-wide information to compute polygenic risk scores for common diseases. Although numerous computational, interpretive, and ethical challenges, this approach establishes the technical feasibility of screening for and preventing numerous debilitating inherited diseases.

Introduction

The development of high-throughput sequencing technologies has enabled the rapid acceleration of our understanding of how specific inherited genetic variants contribute to human disease[1]. In addition, the creation of polygenic risk scores has provided us with new tools to assess risk of developing multifactorial diseases[2]. Still, although we can now accurately diagnose and assess risk for numerous genetic disorders, treatment options remain limited for many diseases.

Preimplantation genetic testing (PGT) has been developed to screen embryos for aneuploidy which has significantly improved implantation and subsequent pregnancy success rates[3]. In addition, strategies have been developed for identifying known genetic variants in families by first screening the parents. Clinical labs then typically use targeted PCR-based strategies to test the embryo for those known pathogenic variants[4]. However, an accurate genome-wide method for screening embryos does not currently exist and most strategies look for either aneuploidy or small genomic changes, but not both in the same embryo. This is due to the lack of genome coverage and/or uniformity of existing whole genome amplification

(WGA) methods, which are required to produce sufficient quantities of DNA to sequence a few cells from a trophoectoderm (TE) biopsy.

We recently developed a much more accurate WGA method, primary template-direct amplification (PTA), which captures almost the entire genome of minute quantities of nucleic acid in a more accurate and uniform manner, enabling much more sensitive genome-wide variant calling[5]. Here we utilized PTA to create a comprehensive PGT strategy of TE biopsies that sensitively and precisely detects aneuploidy, small genomic variants, and heteroplasmy from the same embryo, allowing us to detect inherited and *de novo* genetic variation known to cause disease, as well as produce polygenic risk scores for common multifactorial diseases. Together, this approach established an approach for preventing many genetic diseases through more comprehensive screening of embryos prior to implantation.

Results

To assess the accuracy of our PGT strategy, we performed PTA on 8 TE biopsies harvested from 4 sibling embryos that had been donated and banked for research use. As seen in Fig. 1A, two biopsies were sampled from the same embryo to determine the consistency of variant calls. Parallel analyses were performed to call chromosome ploidy, capture genomic variants, and heteroplasmy. In addition, we were able to conduct genome-wide variant screening for pathological variants at 7.5 million known SNP locations to produce polygenic risk scores for eleven common diseases.

Coverage, uniformity, variant calling, allelic balance, and reproducibility of the PTA-based PGT

Comparable genome coverage was achieved among the biopsies, embryos, and a bulk CD34+ cord blood sample when subsampled to the same read depth (Fig.1B & S1A). The coverage initially rises rapidly with increasing number of reads, followed by a coverage saturation at 96% with 450M reads, corresponding to a mean of 14X sequencing depth (Fig.1B). The uniformity was assessed by constructing Lorenz curve and associated Gini indices (GI) for each sample. Although biopsies of 5-8 cells contained 20 times fewer cells than the ~200 cell embryos, a mean GI of 0.13 was obtained for each biopsy, which was similar to the embryos, as well as the CD34+ cord blood bulk sample (Fig.1C & S1B). In each biopsy, we called an average of 3.27 million SNVs, among which 3.14 million were shared with the corresponding embryo (Fig.S1C). This corresponds to an estimated sensitivity of 96.3% and precision of 96.2% (Fig.1D)[5].

To estimate the somatic variants in each biopsy, we employed the somatic SNV caller SCAN2 that was recently developed for highly specific variant calling from samples that had undergone PTA.[6] Using this approach, we did not see a significant difference in the number of mutations in the biopsies when compared to the corresponding embryo, suggesting that the chromosomal instability seen in early embryos[7] is not associated with the widespread acquisition of small genetic variants .

Allelic dropout due to loss of coverage or allelic imbalances, which are seen with previous WGA methods[5], will have diminished sensitivity that could result in the loss of detection of pathological variants[3]. To asses this, we created variant allele frequency (VAF) histograms, which show that somatic variants have a distribution around 0.5 as expected for acquired heterozygous variants (Fig.1F & S2B). To verify the reproducibility of PTA in PGT, amplification of the 8 biopsies was conducted in three separate batches one

week apart. As expected, we observed equivalent genomic coverage, uniformity, concordance, and allelic balance for all samples (Fig.1B-E).

Detection of fetal aneuploidy

Our first analytic assessment was to determine if we could detect aneuploidy in the embryos, which is the most common use of PGT. As seen in Fig.2, the two biopsies were concordant with the embryo in 3 of 4 cases. Interestingly, the one discordant case had one biopsy consistent with the embryo while the other showed a diploid profile. Upon closer examination of embryo S2 and the concordant biopsy, the loss of chromosome 14 was partial, suggesting one area of the embryo was mosaic for the loss of chromosome 14 while the area taken by the second biopsy was diploid, as has been previously reported (Fig.S2C)[8].

Genome-wide screening for genetic diseases

We then sought to determine the additional genetic abnormalities that could be identified using our genome-wide approach. To achieve this, we primarily focused on variants that are rare in the population and reside in transcribed regions. Variants among those regions were then annotated for predicted consequences with MutationTaster[9], and for known disease phenotypes using Clinvar[10], HGMD[11], and OMIM[12]. Importantly, the identified potential deleterious variants were concordant between both biopsies and the corresponding embryos (Fig.3A). Most of the disease-associated variants were also found in at least one additional embryo, suggesting they were inherited from one of the parents (Fig.3A). However, we also identified embryo-specific changes such as COQ8A A233T and ALOXE3 L237M, which likely represent *de novo* variants.

We then focused on variants that have been described to cause autosomal dominant diseases (Fig.3A). This included a known gene conversion event in vWF (V1279I)[13] that increases the risk of bleeding, SDHB S163P that has been shown to cause cancer predisposition[14], and APC R534Q which increases risk for clotting[15, 16]. In addition, embryo S4 has autosomal recessive alpha−actinin−3 deficiency, which is associated with increased aerobic metabolism in skeletal muscle (Fig.3A)[17]. We also reported another 13 potential autosomal recessive variants here that are either known or adjacent to pathological variants (Fig.3A). Together, these data suggest it is feasible to screen almost the entire genome of an embryo for known disease-associated variants.

Polygenic risk scores for 11 common diseases

Next, we directed our evaluation to common variants that are known to have small risk for common multifactorial diseases, and can be combined across the genome through the calculation of a polygenic risk score (PRS)[18]. With our broad genomic coverage, we successfully called more than 98% of selected SNP coordinates (7.5 million in total) from each biopsy, which is similar to the sensitivity of bulk PRS site coverage from 11 published studies (Fig.3B & S3A)[18-24]. Raw scores were then calculated and transformed into percentiles and prevalence using raw score distributions from the UK Biobank[25], and whole embryos were processed in the same manner as controls. Importantly, we again saw consistent PRS percentile

when comparing the embryo and corresponding biopsies in all cases, suggesting the feasibility of applying PRSs in PGT (Fig.3C-D & S3B-C).

Interestingly, coronary artery disease (CAD) and schizophrenia (SCZ) PRS-percentiles vary between embryos, even though they were derived from the same donors, which could be due to a combination of *de novo* variants in the embryo and the random co-segregation of risk alleles. (Fig.3C). In contrast, breast cancer and atrial fibrillation PRS-percentiles were generally low (Fig.S3A-C). We identified embryo S3 has an increased risk of type 1 diabetes (>=3X odd ratio, Fig.S3C). However, taking into account the low prevalence of type 1 diabetes in 50th percentile controls (<0.9%), the probability of developing diabetes remains low[24]. Broader GWAS analyses implied risk for 3 disease phenotypes (Table S1)[26]. Interestingly, age-related macular degeneration and venous thromboembolism exhibited levels of consistency with the results of our mendelian disease screening strategy.

Table S1

	GWAS results	Which Embryo
1	Psoriasis vulgaris	S3
2	**Venous thromboembolism**	All
3	**Age-related macular degeneration**	S3

Mitochondria heteroplasmic variant screening with low-pass sequencing

Further analysis of sequencing data from the same 8 biopsies and 4 embryos revealed 100% mitochondrial genome coverage with PTA with just 1 million reads, which is consistent with our previous work (Fig.4A)[5]. With 10 million reads, the mean sequencing depth of mtDNA was 342X, which is sufficient to detect heteroplasmy at approximately 1-2% frequency within the cell (Fig.4B). Meanwhile, both of sensitivity and precision reached 100%, indicating high concordance between mtDNA in embryos and biopsies using our approach (Fig.4A). As mtDNA are maternally inherited, we first confirmed high conservation of mtDNA among sibling embryos and their biopsies. All samples share the same 25 variants, including 5 in non-coding regions and 4 in rRNA regions, as well as 12 synonymous and 4 nonsynonymous coding variants (Fig.4C). None of these mtDNA variants have been reported as pathological[27].

We then looked for embryo-specific mtDNA variants where we identified a unique heteroplasmic variant, C9512T in *COIII,* in embryo S3 (Fig.4C). It is synonymous and has not been reported to associate with known diseases. Importantly, this variant was identified in both biopsies and the embryo at similar frequencies (48-50%) (Fig.4D). It could have been inherited as de novo heteroplasmy in the egg or developed at a very early stage in the embryo, resulting in a selective advantage for those mitochondria. In addition, there is another low-frequency heteroplasmic variant in embryo S1 that is present at a mean of 4% allele frequency where the two biopsies possess 0% and 8% of this variant, respectively (Fig.4D). This spatial separation of that heteroplasmic variant between biopsies suggests it occurred in a mosaic population during the initial stages of embryogenesis. The copy number and mitochondrial variant calling data with just 10M reads suggests our approach allows us to accurately detect both heteroplasmy and aneuploidy with low-pass sequencing.

Non-invasive aneuploidy assessment of spent media

It has been reported that aneuploidy can be detected noninvasively by amplifying DNA from spent media of embryos[28]. We hypothesized that using PTA we could not only detect aneuploidy, but sequence a significant proportion of the embryo genome noninvasively. We therefore modified the PTA protocol to extract DNA from spent media prior to amplification (Fig.4E). As expected, CNV calling from the two available spent media samples (embryos S1 and S3) showed aneuploid profiles consistent with the biopsies (Fig.4E & S4A). Interestingly, we saw a trend of increased DNA yield from PTA in the aneuploid embryos when compared to diploid samples, suggesting that there is increased cell death in the aneuploid cells and that DNA yield alone could be sufficient for screening embryos. Additional trials were therefore performed and compared to the clinical reports where we again found that spent media of aneuploid embryos exhibited a roughly 3-fold increase in yield (Fig.4F). Interestingly, a diploid embryo based on clinical diagnostics (Fig4F-right) exhibited high DNA yield and loss of chromosomes 22 after sequencing using our approach.

We then performed deep whole genome sequencing of the spend media where we found that genome coverage saturated at 50-70% with significant allelic drop out (Fig.S4B-C). In addition, we were unable to produce sequencable libraries from the diploid samples, which was likely the result of insufficient quantity of DNA in the media. Still, further work is needed to determine if there is sufficient DNA in the media of diploid embryos for noninvasive whole genome evaluations.

Discussion

In this study, we outline a new strategy for genome-wide PGT with PTA where we can detect CNV, small genetic variants, and heteroplasmy. Most current clinical PGT testing evaluates embryos for aneuploidy or a very small number of selected genomic variants, but not both. This is due to the inherent limitations of currently used WGA methods that introduce method-specific artifacts, limiting the accuracy of downstream analyses for a given WGA method. Previous studies have performed whole genome sequencing of embryo biopsies using multiple displacement amplification (MDA) [29, 30]. However, those studies were also hampered by the artifacts introduced by MDA, including loss of coverage and uneven coverage that hamper variant calling. Further, studies that utilize MDA typically screen an unknown number of embryos using SNP arrays or other methods prior to selecting the top candidates for whole genome sequencing, making the clinical implementation of that approach impractical.

In the present study, we took advantage of the high coverage breadth and uniformity, as well as the low error rate and high reproducibility of PTA to perform accurate genome-wide variant calling of embryo biopsies. Importantly, we did not screen embryos prior to sequencing, enabling the potential immediate translation of this approach into clinical practice. In addition, we utilized the accurate genome-wide variant data to calculate polygenic risk scores for common diseases. Finally, we provided initial feasibility data for performing whole genome sequencing of spent media with PTA where we found aneuploid embryos have significantly higher levels of DNA, making the presence of high levels of extracellular DNA a potential marker for fetal aneuploidy.

There are a number of important limitations to our study. As with all whole genome sequencing studies, variant calling from the three billion locations in the human genome is not perfect. Further work is needed to balance the tradeoff between sensitivity and precision to provide optimal clinical insights from the data. Related to that concern, even with extraordinarily accurate variant calling, the interpretation of a given variant is frequently based on imperfect, and in many cases, no empirical supporting evidence. Those concerns are further amplified when using polygenic risk scores that use multiple associated genes that each have a small change in risk. Together, these observations highlight the computational framework that needs to be developed for the responsible clinical implementation of genome-wide PGT with PTA, as well as the importance of caution when interpreting the results.

The creation and future clinical utilization of whole genome sequencing of embryos also bring up a number of important ethical concerns. First, with all the challenges in the interpretation of the results, which variants should be reported back to the family? Should future parents only receive information on variants known to cause a specific disease that arises early in life, or should reports also include adult-onset diseases or even just an increased risk of those diseases? What if parents would like additional non-disease related insights from the data, such as the probability of having specific traits? All of these ethical challenges don't take into account the quality of genomic variant annotation where there is significantly more information for those of European ancestry than all other populations, creating a disparity between people of different ancestries. Finally, there are important questions around parental consent: 1) Can parents provide consent with all the technical caveats in the interpretation? 2) If parents are more sophisticated in their understanding, can they consent to information beyond what the average person would understand? 3) What about the consent of the unborn child and any potential future consequences as a result of sequencing their whole genome? All of these challenges, and others, need careful consideration by the reproductive health community to create consensus around appropriate and ethical best practices.

In summary, we have presented a new strategy for genome-wide disease screening of embryos that is able to capture almost all the genetic variants in a sample with high precision. This approach is now technically feasible in a clinical setting, although numerous computational, interpretive, and ethical challenges remain to be addressed. Still, this approach provides a path for screening embryos for most genetic variants associated with diseases, with the potential to prevent the suffering caused by thousands of incurable genetic diseases.

Materials and Methods

Sample approval and collection

This study was approved by Stanford University IRB # 58757. Experiments were performed in accordance with protocol guidelines and regulations. The couple involved in this study had standard clinical PGT-A on each embryo, followed by a written consent to donate aneuploid embryos for research. An ethics meeting was conducted to discuss the study as part of the IRB review process.

Whole genome amplification through PTA

Embryos were first acquired retrospectively from tissue bank, followed by taking additional two biopsies from each embryo after thawing. Biopsies and embryos were transferred into a 200ul PCR-tube containing

3ul of cell buffer (Bioskryb) before PTA. For media samples, DNA in the culture media (20ul-30ul in volume) was extracted by 1X AMPure beads (Beckman) with 5 minutes incubation followed by 80% ethanol wash twice. Then 3ul of EB buffer was added to elute the DNA from the beads before PTA. Beads were left inside the tube during PTA.

PTA was performed according to Bioscryb's instruction. Specifically, 3ul of alkaline lysis buffer was added to each sample, followed by a 5-minute incubation on ice and another 5-minute incubation at room temperature to ensure DNA was denatured completely. The stop buffer was then added to neutralize the pH. After that, 3ul of exonuclease-resistant random primers were introduced, allowing them to bind to genomic DNA for 10 minutes at room temperature. At last, 8ul of reaction mix (containing phi29 polymerase and alpha-thiol terminators) was added prior to placing on thermocycler. Thermocycler program was set for 10 hours at 30^{o}C for amplification, 5 minutes at 65^{o}C for termination and infinite at 4^{o}C for storage. All reagents used in PTA were acquired from Bioskryb.

After PTA, DNA was purified using 2X AMPure XP magnetic beads (Beckman). Yields were measured using the Qubit dsDNA HS Assay Kit with a Qubit 4 fluorometer according to the manufacturer's instructions (Thermo Fisher).

Library preparation and sequencing

DNA sizes were first confirmed by running 1-2% Agarose E-Gel (Invitrogen). Then, 500ng of PTA product was used for library preparation with KAPA HyperPlus kit without fragmentation step because PTA generated sufficient quantity of DNA at optimum sizes for illumina sequencer. 2.5 µM of unique dual index adapter (Integrated DNA Technologies) was used in the ligation. 10 cycles of PCR were used in the final amplification for biopsies and embryos, and 15 cycles were used for media samples if starting DNA is less than 500 ng.

DNA concentration in the library was quantified using the Qubit 4 dsDNA HS assay as mentioned previously. Library sizes were confirmed through Agilent 4200 tapestation D1000 ScreenTape assay (Agilent Technologies). Sequencing runs were performed via either MiniSeq for QC or NovaSeq 6000 for WGS on 500ul 1.6pM DNA.

Benchmarking Experiments Data Analysis

Sequencing data were trimmed using trimmomatic first to remove adapter sequences and low-quality terminal bases, followed by GATK 4 best practices with genome assembly of GRCh38. In Brief, fastq files were aligned with bwa mem, and then the corresponding bam files were processed with base quality score recalibration (BQSR) and marking duplicates before loading into gatk HaplotypeCaller. Mitochondrial genome was included in the whole genome bed during variant calling so that mitochondria variants could be identified as well. The resulting vcf files were combined and genotyped with gatk combineGVCFs and GenotypeGVCFs, followed by variant quality score recalibration (VQSR). Annotation on variants was done by Annovar and HDMD. Quality metrics such as genomic coverage etc were acquired from the bam files after BQSR and MarkDuplicates using qualimap, as well as gatk AlignmentMetricsAummary and CollectWgsMetrics. Sensitivity and precision curves were generated through rtg package on the same BAM files. No regions were excluded from the analysis.

Sensitivity and precision

The sensitivity and precision were calculated using RTG package by comparing each biopsy to its corresponding embryo. Default parameters were used for data running. Sensitivity was defined as number of variants shared between embryo and biopsy over total number of variants detected in embryo. Precision was defined as number of variants shared between embryo and biopsy over total number of variants detected in that biopsy.

Variant allele frequency histogram

Allele frequency was calculated by using the number of ALT read divided by total read at that position. Each sample has roughly 3.26 million variants, which is slow in making histogram in R. We therefore randomly sampled 10000 variants and then generated the histogram.

Chromosomal copy number variation

Ginkgo was applied for CNV calling following its standard protocols. All data were aligned to reference genome hg38/GRCh38. As Ginkgo was based on hg19 in default, we created new reference files for Ginkgo using the instructions and scripts from the Ginkgo github (https://github.com/robertaboukhalil/ginkgo) and performed Ginkgo CNV calling locally. For Ginkgo analysis, samples were converted to bed files through bedtools to feed to the Ginkgo tool. A bin size of 1Mbp and independent segmentation were used as running parameters.

Polygenic risk score and GWAS analysis

Embryos and biopsies were subjected to gatk HaplotypeCaller to generate gVCF files containing all coordinates on the genome including REF bases (via BP_resolution option). Then a total of 7.5 million unique genomic coordinates were focused based on published PRSs for 11 diseases. Percent coverage in this case is defined as successfully extracted coordinates over total required coordinates for each disease. We calculated raw scores of 11 PRS and adjusted for 4 principal components of ancestry to minimize spurious ancestry associations in the resulting polygenic score. The percentile is reported based on the raw score distribution of about 100,000 UK Biobank participants. To investigate the uncertainty in the PRS percentiles due to missing coordinates, we first simulated random genotypes for missing SNPs and constructed 95% confidence intervals of PRSs for CAD and SCZ. Then the mean±errs of PRSs were converted to percentile following the same principle. The prevalence was converted using published AUC values and tools here *https://opain.github.io/GenoPred/PRS_to_Abs_tool.html*, or extrapolated from prevalence vs percentile curve if AUC is not available. Next, to evaluate the possible genotypes on a set of the most important GWAS SNPs across a variety of publications, we sorted the GWAS catalog[26] and took the 1,000 highest log P-value variants. Then, we extracted these variants from GVCFs of embryo and biopsy, followed calculating concordance metrics between embryos and biopsies for this variant subset. We then sort the filtered list based on odds ratio and present high-odds ratio examples.

Disease screening

Disease screening was performed on variants related to exonic regions, 5bp adjacent to stop/start codons, splicing regions and ncRNA regions. These variants were extracted first followed by annotation of Gnomad, Clinvar, HGMD, MutationTaster and OMIM. The criteria for analysis include: 1) Gnomad AF_popmax < 10%, 2) positive Clinvar or HGMD annotations, and 3) MutationTaster without label of “polymorphism automatic” as these variants are known to be harmless. Afterwards, ~280 variants were left and they were subjected to further analysis using content in Clinvar and HGMD. Pathological (or DM) and conflict (or DM?) were particularly focused and reviewed manually.

Code availability

Packages/software used in this study are all open-source. GATK4 from Broad institute is available at https://gatk.broadinstitute.org/hc/en-us. RTG tools from Real Time Genomics is available at https://www.realtimegenomics.com/products/rtg-tools. HGMD is provided by Stanford subscription via QIAGEN (a public version is available at http://www.hgmd.cf.ac.uk/ac/index.php). Annovar is available at https://annovar.openbioinformatics.org/en/latest. The whole genome sequencing analysis pipeline

described in "Benchmarking Experiments Data Analysis" section is available on Gawad-lab github page https://github.com/Gawad-Lab/. Other scripts are available on request.

Author Contributions

Y. Xia, V. Gonzales-Pena, and C. Gawad designed experiments. B.R. Behr, and P. Park contributed key materials, methods, and discussion. Y. Xia, D.J. Klein, J. Luquette and V. Reddy performed and analyzed experiments. L. Puzon and N. Siddiqui performed PRS and GWAS analysis. Y. Xia, C. Gawad wrote the manuscript.

Conflict of Interests

CG is a co-founder and board member of BioSkryb, which is commercializing primary template-directed amplification. NS is a founder of Orchid and BB is a Scientific Advisor to Orchid.

Acknowledgements

We would like to acknowledge Stanford center for biomedical ethics for benchside ethic consultation. C.G. is supported by Burroughs Wellcome Fund Career Award for Medical Scientists, an NIH Director's New Innovator Award (1DP2CA239145) and the Chan-Zuckerberg Biohub.

Reference

1. Ott, J., Wang, J. & Leal, S.M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* **16**, 275-284 (2015).
2. Torkamani, A., Wineinger, N.E. & Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590 (2018).
3. Vermeesch, J.R., Voet, T. & Devriendt, K. Prenatal and pre-implantation genetic diagnosis. *Nat Rev Genet* **17**, 643-656 (2016).
4. Lee, V.C.Y., Chow, J.F.C., Yeung, W.S.B. & Ho, P.C. Preimplantation genetic diagnosis for monogenic diseases. *Best Pract Res Clin Obstet Gynaecol* **44**, 68-75 (2017).
5. Gonzalez-Pena, V. et al. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc Natl Acad Sci U S A* **118** (2021).
6. Luquette, L.J. et al. Ultraspecific somatic SNV and indel detection in single neurons using primary template-directed amplification. *bioRxiv* (2021).
7. Fragouli, E. et al. The origin and impact of embryonic aneuploidy. *Hum Genet* **132**, 1001-1013 (2013).
8. McCoy, R.C. Mosaicism in Preimplantation Human Embryos: When Chromosomal Abnormalities Are the Norm. *Trends Genet* **33**, 448-463 (2017).
9. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362 (2014).
10. Landrum, M.J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014).
11. Stenson, P.D. et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665-677 (2017).
12. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**, D789-798 (2015).
13. Gupta, P.K. et al. Gene conversions are a common cause of von Willebrand disease. *Br J Haematol* **130**, 752-758 (2005).
14. Castellano, M. et al. Genetic mutation screening in an italian cohort of nonsyndromic pheochromocytoma/paraganglioma patients. *Ann N Y Acad Sci* **1073**, 156-165 (2006).
15. Liu, X., Tao, T., Zhao, L., Li, G. & Yang, L. Molecular diagnosis based on comprehensive genetic testing in 800 Chinese families with non-syndromic inherited retinal dystrophies. *Clin Exp Ophthalmol* **49**, 46-59 (2021).
16. Allikmets, R. et al. Mutation of the Stargardt disease gene (ABCR) in age-related macular degeneration. *Science* **277**, 1805-1807 (1997).
17. MacArthur, D.G. et al. Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nat Genet* **39**, 1261-1265 (2007).
18. Khera, A.V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics* **50**, 1219-1224 (2018).
19. Desikan, R.S. et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS Med* **14**, e1002258 (2017).
20. Khera, A.V. et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596 e589 (2019).
21. Schumacher, F.R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature genetics* **50**, 928-936 (2018).

22. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
23. Abraham, G. et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun* **10**, 5819 (2019).
24. Sharp, S.A. et al. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes care* **42**, 200-207 (2019).
25. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
26. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
27. Lott, M.T. et al. mtDNA variation and analysis using mitomap and mitomaster. *Current protocols in bioinformatics* **44**, 1.23. 21-21.23. 26 (2013).
28. Huang, L. et al. Noninvasive preimplantation genetic testing for aneuploidy in spent medium may be more reliable than trophectoderm biopsy. *Proceedings of the National Academy of Sciences* **116**, 14105-14112 (2019).
29. Murphy, N.M., Samarasekera, T.S., Macaskill, L., Mullen, J. & Rombauts, L.J.F. Genome sequencing of human in vitro fertilisation embryos for pathogenic variation screening. *Sci Rep* **10**, 3795 (2020).
30. Peters, B.A. et al. Detection and phasing of single base de novo mutations in biopsies from human in vitro fertilized embryos by advanced whole-genome sequencing. *Genome Res* **25**, 426-434 (2015).

Figure Legends

Figure 1. Experimental workflow and data characterization. A) A total of 4 sibling embryos were collected from tissue bank. Then, 2 biopsies were sampled from each embryo, followed by a transfer into 3ul of cell buffer. Standard PTA followed by sequencing were then performed on all biopsies for characterization. In parallel, the remaining embryos were processes under the same protocol for results validation. Data analysis was present in terms of genome-wide evaluation of disease risks, mitochondrial heteroplasmy identification and non-invasive aneuploidy examine. B) A 96% genomic coverage was achieved in all biopsies at 450M reads (14X). C) Gini index, which is used to assess amplification uniformity, was consistent between embryos and corresponding biopsies. The Gini index of amplified samples is comparable to the bulk sample. D) A sensitivity and precision of 96.3% and 96.2% were observed respectively among all samples at 450M reads. E) Both heterozygous variants and homozygous variants were captured in biopsies. However, somatic variants tended to be dominantly heterozygous.

Figure 2. Copy number profiling. Aneuploidy screening was done using 1Mb bin size on 5M read samples. Consistent copy number changes are observed except for the mosaic embryo S2. However, mosaicism was able to be detected when aneuploidy cells were biopsied.

Figure 3. Genome-wide screening on disease-causing variants and polygenic risk score. A) Genome-wide screening on disease-causing variants. B) 6.6 million and 0.1 million SNPs are required respectively to calculate PRSs for CAD and SCZ. We covered an average of 98.4% SNPs for CAD and 98.6% for SCZ in each sample, ensuring a complete calculation for PRS. C) Consistent PRS percentiles were seen between embryo and corresponding biopsies. However, percentiles varied between embryos even though they were from the same donors.

Figure 3. Mitochondria heteroplasmy identification at low-pass sequencing and non-invasive attempts on PGT-A. A) Mitochondrial genome was 100% covered even at 1M reads. Using 10M-read samples for demonstration, sensitivity and precision reached 100% in both cases. B) A mean read depth of 342X was achieved at 10M reads. C) Venn plots indicated a unique heteroplasmy in embryo of S3. D) Variants in biopsies and embryos were displayed as variant minor allele frequency against location on mitochondria genome. E) We modified PTA protocol to amplify DNA from spent media of embryos. CNV measurement from spent media is consistent with CNV of embryos. F) DNA yield from each spent medium after PTA was plotted. Samples were classified into aneuploidy and diploid based on standard clinical PGT-A report on one biopsy. Aneuploidy spent media on average yielded 3X more DNA quantity. The outlier in cyan has a loss of chr19 and 22 based on our sequencing analysis. The outlier in red is S2 in figure 2. Although further validations are needed to compare traditional PGT-A with our new PGT-A approach, we have so far achieved 86% consistency with clinical PGT-A by merely measuring DNA yield.

Figure S1. Characterization on sequencing results. A) Genomic coverage of embryos and bulk CD34+ sample. B) Lorenz curve of all samples was plotted to demonstrate uniformity of amplification. C) An average of 3.27 million SNVs were called in each biopsy and 3.14 million of them were shared with the corresponding embryo.

Figure S2. Allele balance and CNV on mosaic embryo. A) Variant Allele balance is highly consistent in all samples. B) Somatic variants tend to be heterozygous. C) Illustration of CNV results on the mosaic embryo.

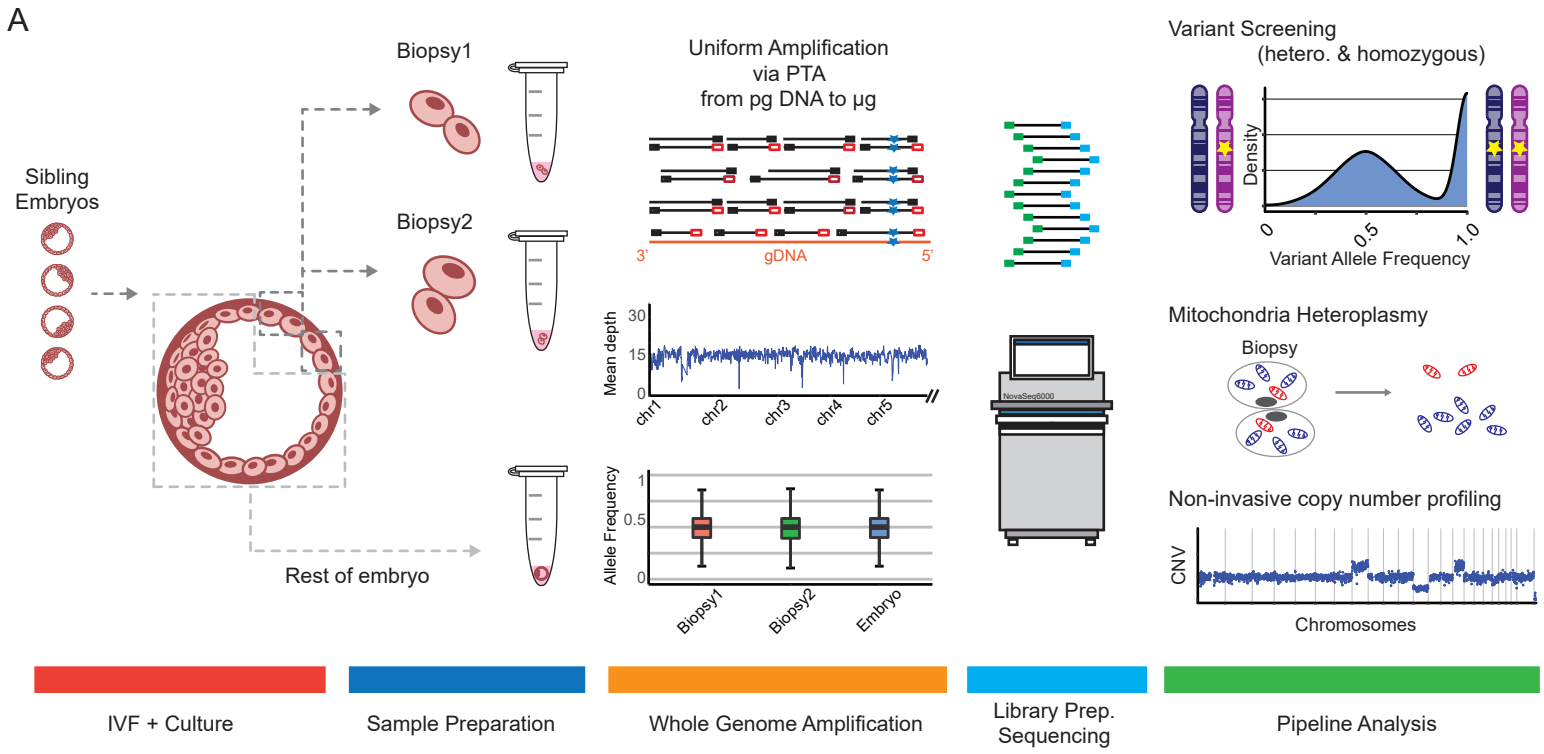
Figure S3. Polygenic risk score analysis on more diseases. A) High SNPs coverage is observed for all 9 diseases here. B) PRS percentile is plotted based on polygenic risk sore of each sample. Consistency between biopsies and corresponding embryo is observed. C) Percent prevalence or risk level is plotted based on PRS percentile.

Figure S4. Non-invasive PGT-A. A) CNV from spent media is consistent with CNV of embryo. B) After amplifying DNA from spent media, partial genomic coverage is observed even at 450M reads. C) This is due to allele dropout as majority of heterozygous variants are lost in the spent media.

Table S1. Results of GWAS analysis. Top phenotypes were recorded according to GWAS analysis on p-values and odds ratios.

Figure 1

Experiment Overview



High Genomic Coverage, Uniformity and Concodance were Achieved after Amplification

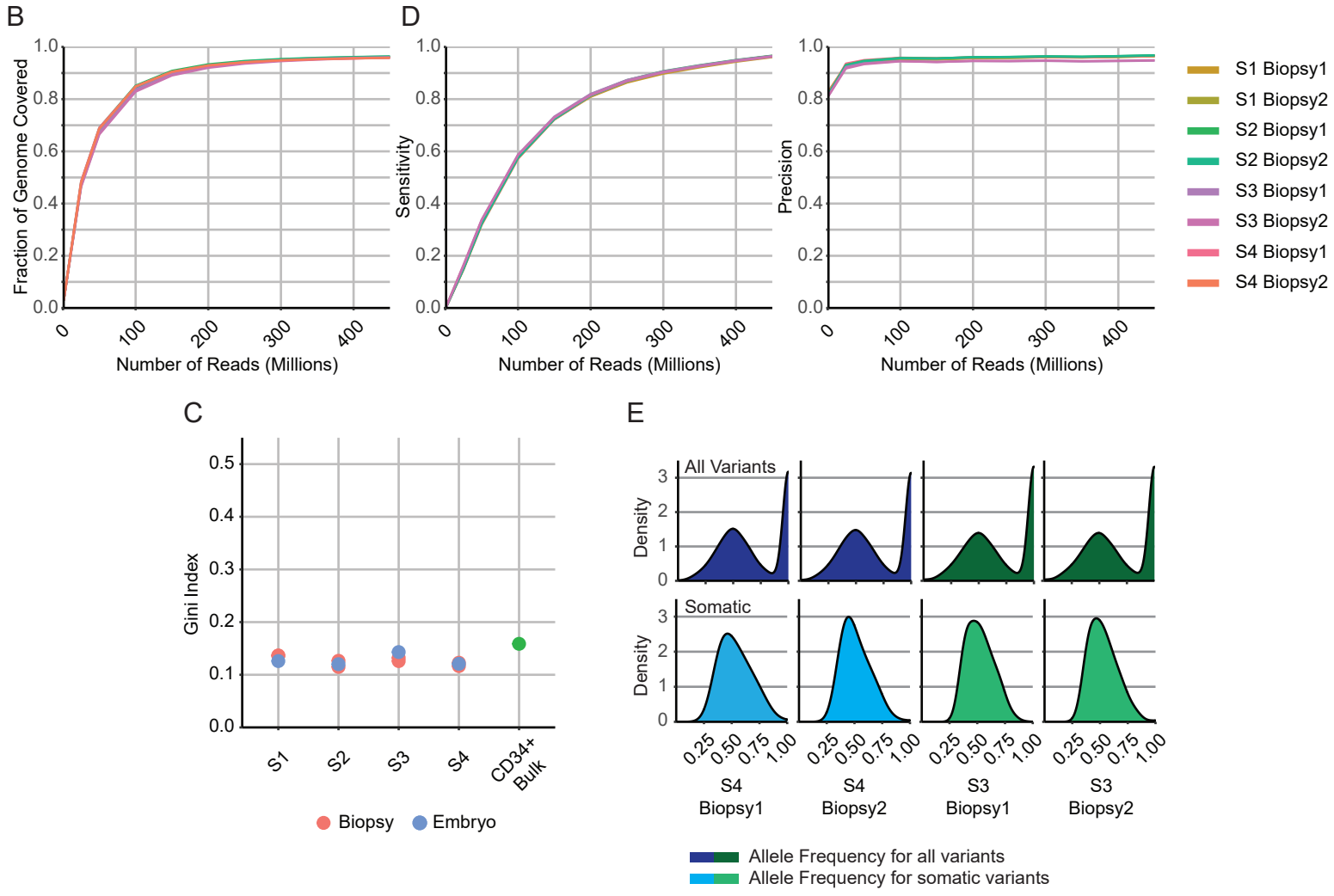


Figure 2

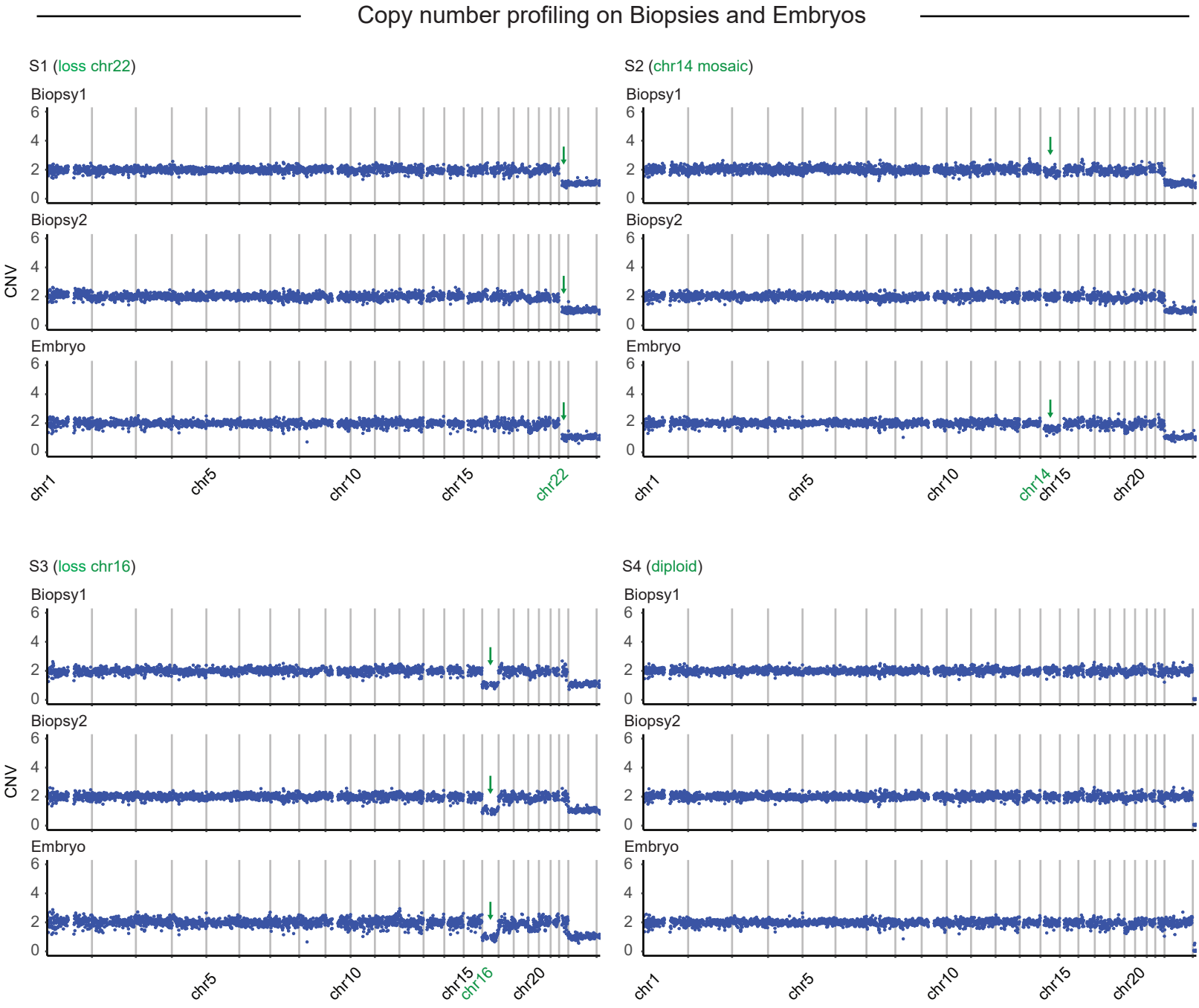
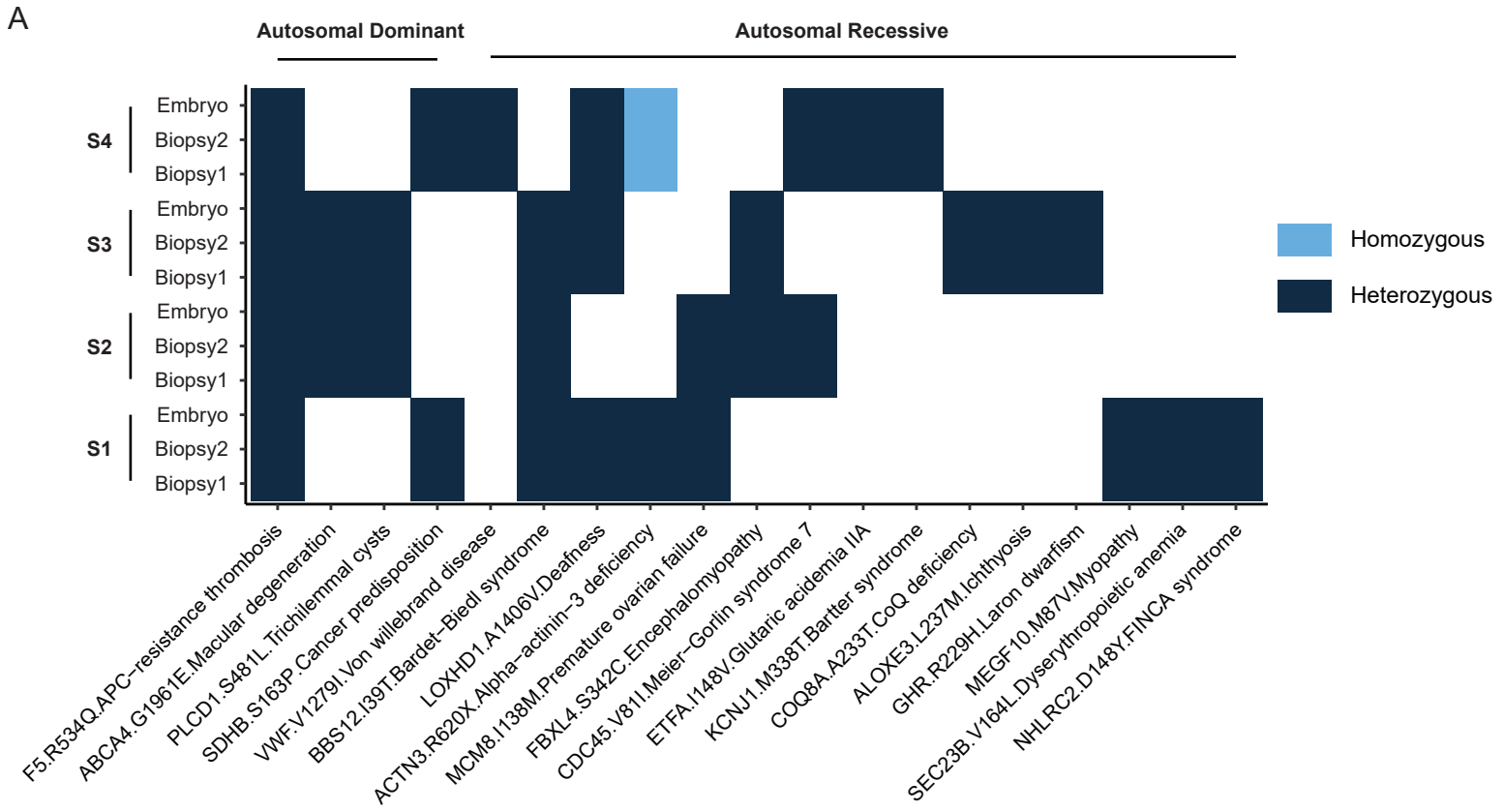


Figure 3

Genome-wide screening on disease-associated variants



Polygenic risk scores on diseases

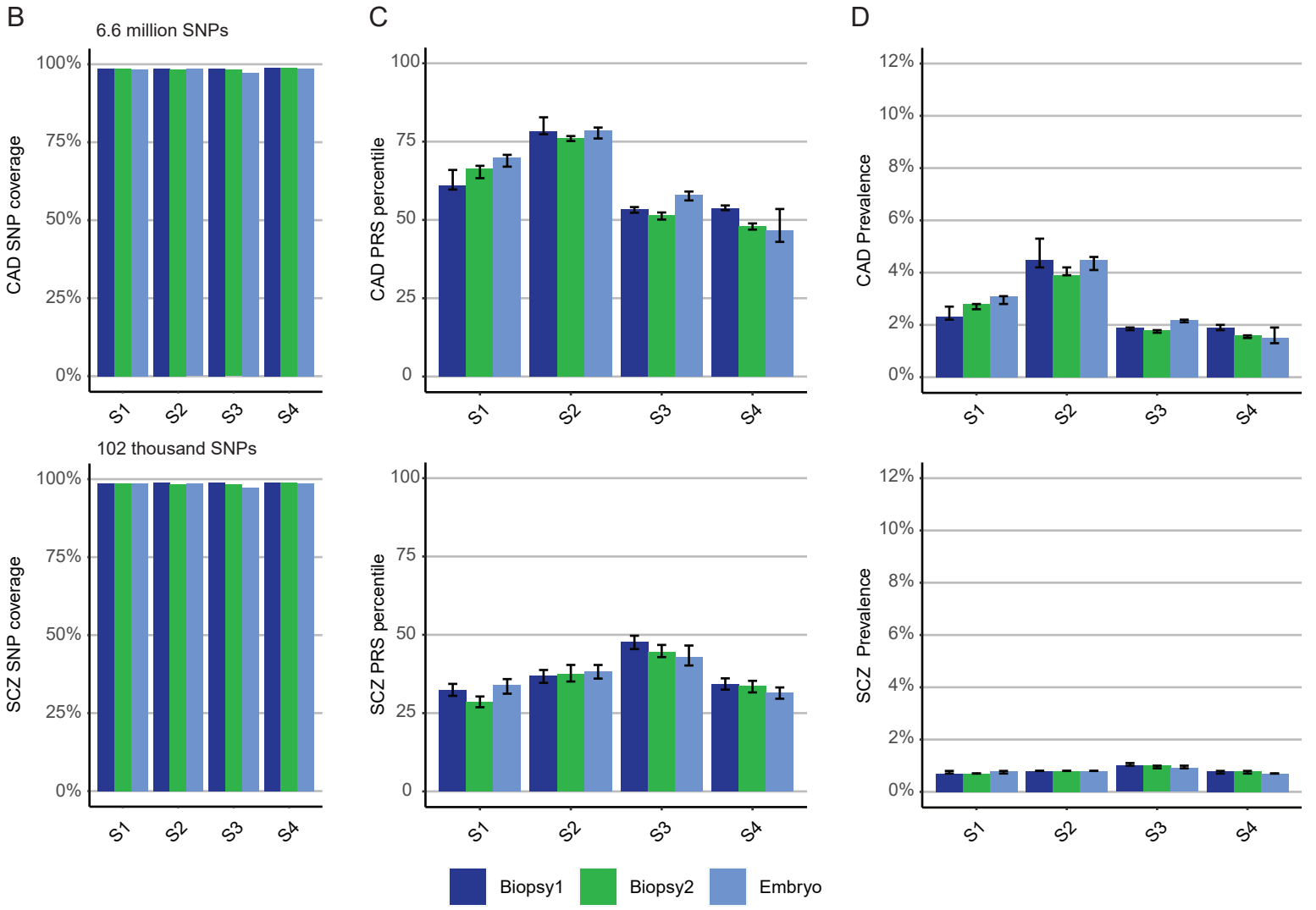
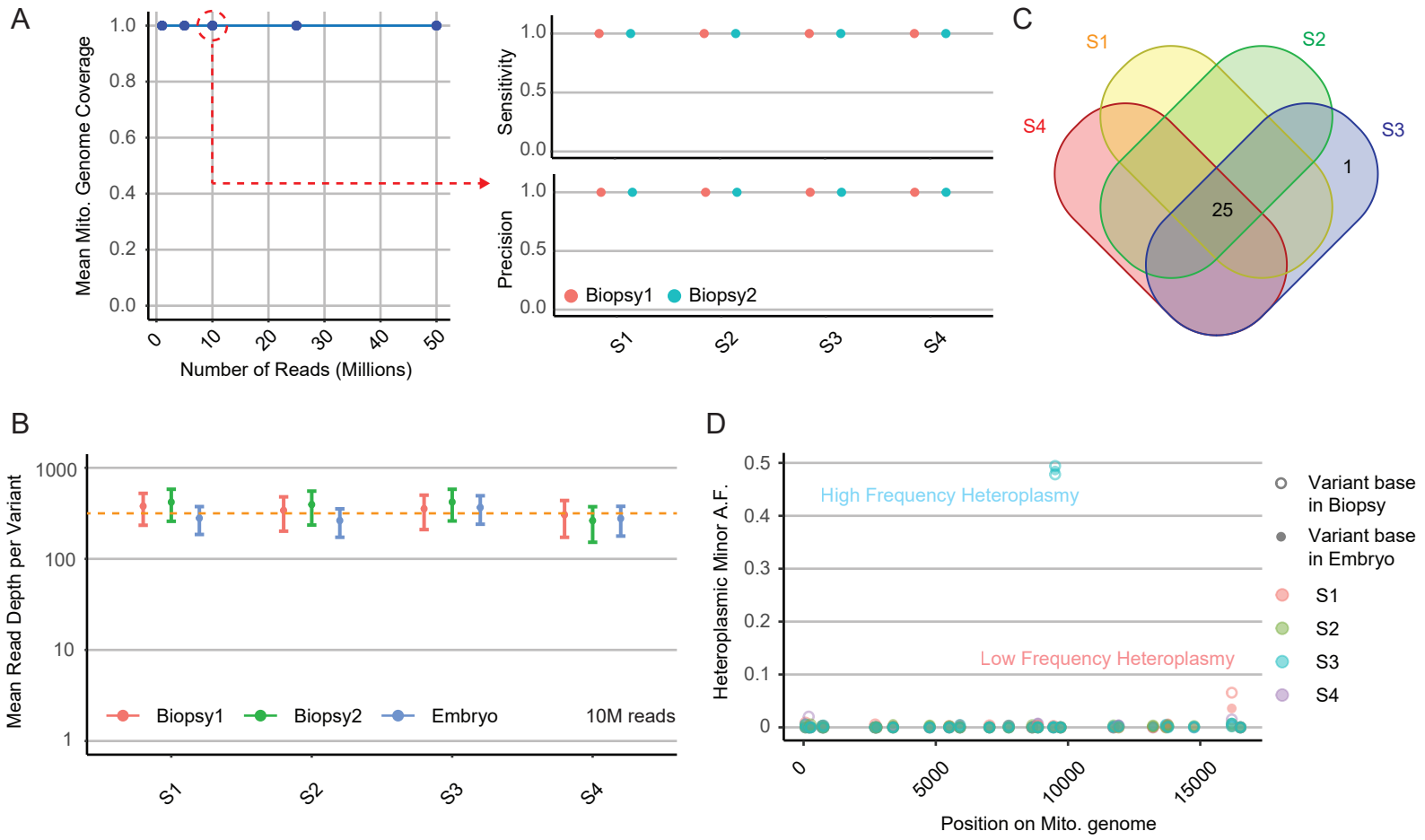


Figure 4

Heteroplasmy Screening on MitoDNA at Low-pass Sequencing



Aneuploidy Identification from Spent Mmedia

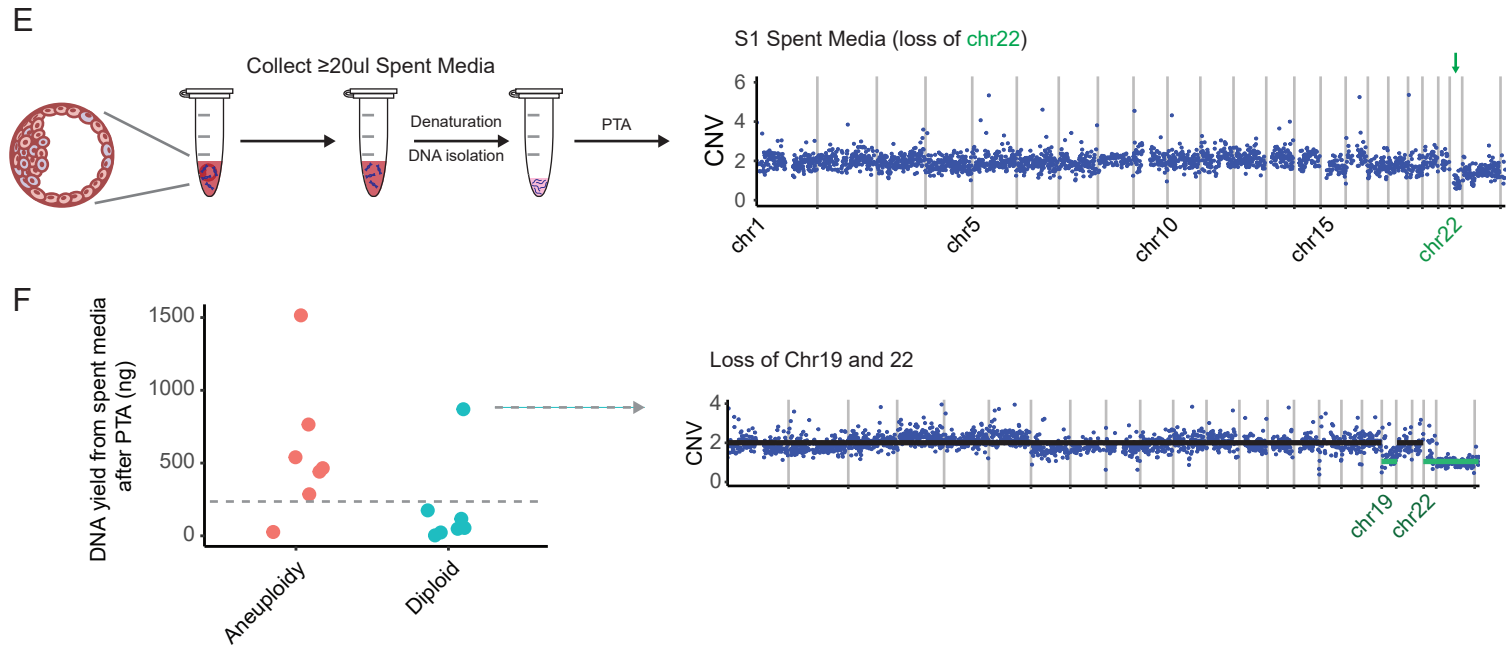
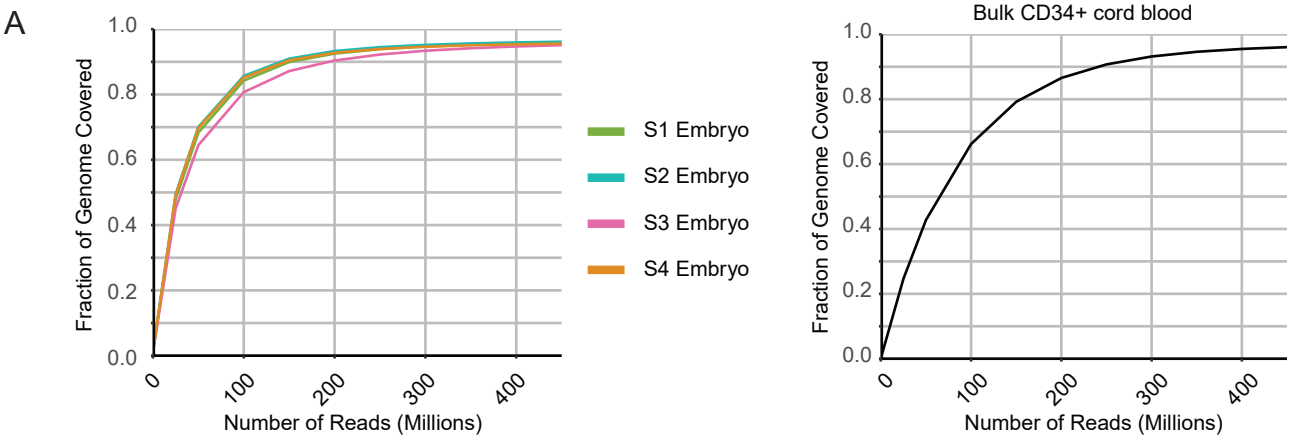
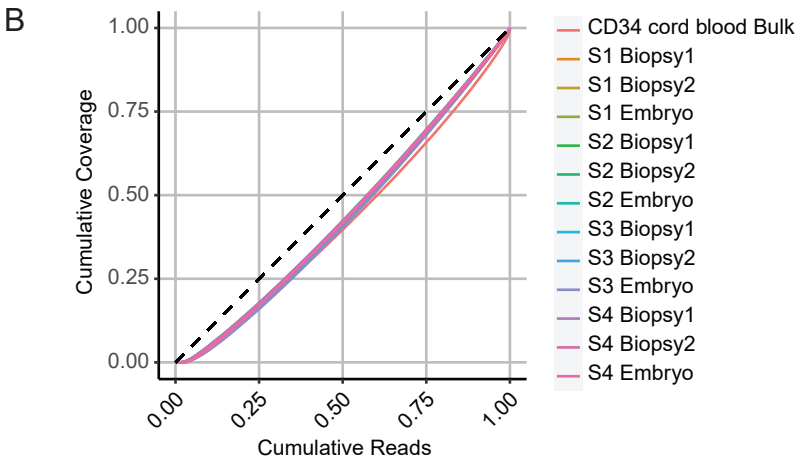


Figure S1





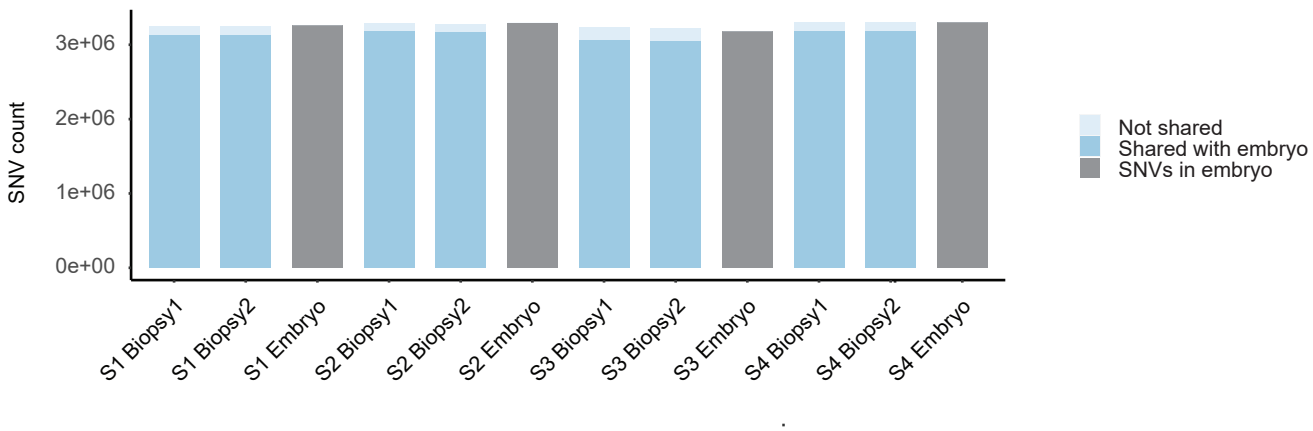
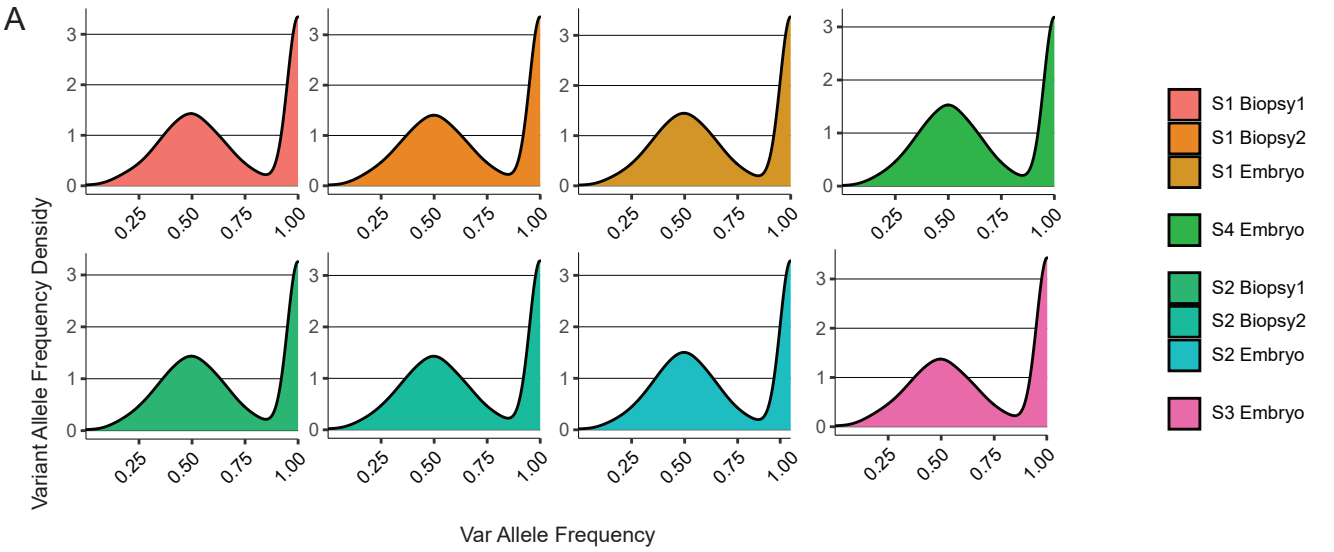
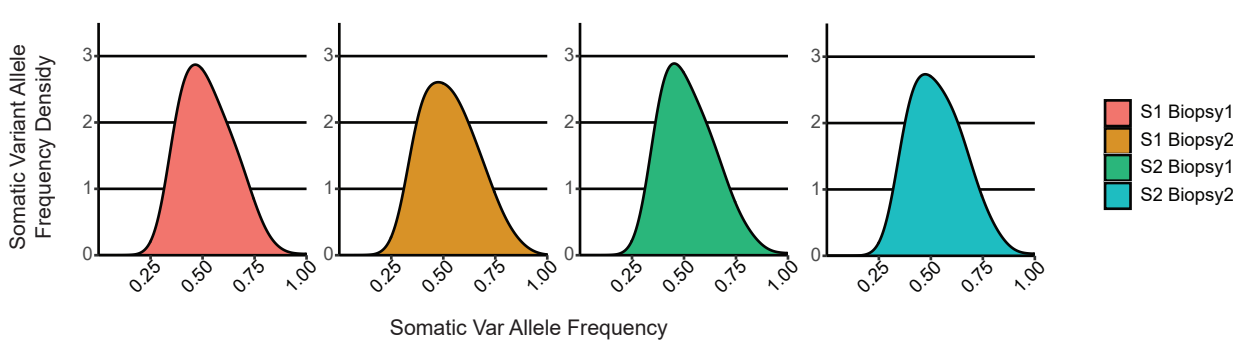


Figure S2





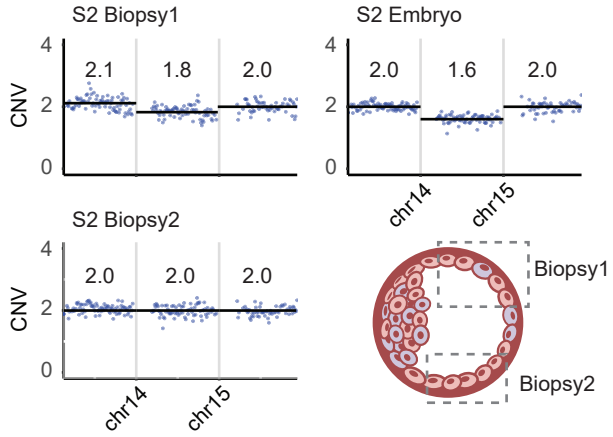


Figure S3

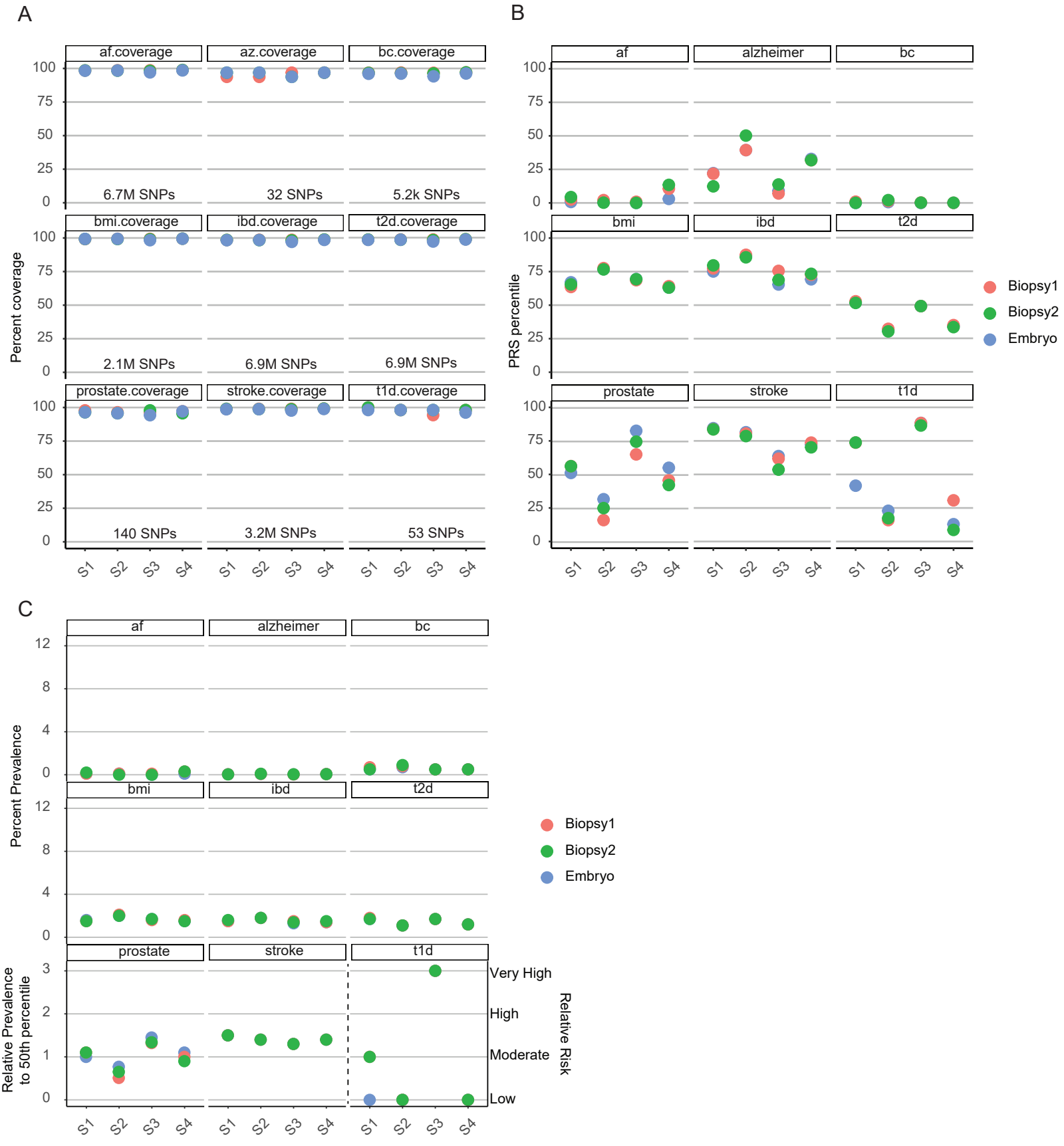
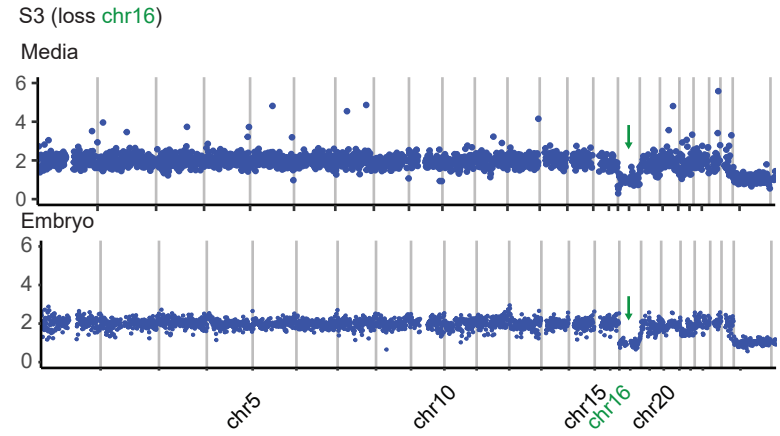
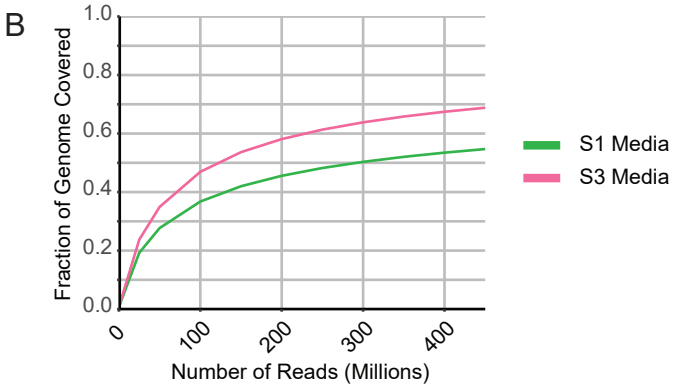


Figure S4

A



B



C