The genetic architecture of sporadic and multiple consecutive miscarriage

Laisk, Triin; Soares, Ana Luiza G.; Ferreira, Teresa; Painter, Jodie N.; Censin, Jenny C.; Laber, Samantha; Bacelis, Jonas; Chen, Chia Yen; Lepamets, Maarja; Lin, Kuang; Liu, Siyang; Millwood, Iona Y.; Ramu, Avinash; Southcombe, Jennifer; Andersen, Marianne S.; Yang, Ling; Becker, Christian M.; Børglum, Anders D.; Gordon, Scott D.; Bybjerg-Grauholm, Jonas; Helgeland, Øyvind; Hougaard, David M.; Jin, Xin; Johansson, Stefan; Juodakis, Julius; Kartsonaki, Christiana; Kukushkina, Viktorija; Lind, Penelope A.; Metspalu, Andres; Montgomery, Grant W.; Morris, Andrew P.; Mors, Ole; Mortensen, Preben B.; Njølstad, Pål R.; Nordentoft, Merete; Nyholt, Dale R.; Lippincott, Margaret; Seminara, Stephanie; Salumets, Andres; Snieder, Harold; Zondervan, Krina; Werge, Thomas; Chen, Zhengming; Conrad, Donald F.; Jacobsson, Bo; Li, Liming; Martin, Nicholas G.; Neale, Benjamin M.; Nielsen, Rasmus; Walters, Robin G.; Granne, Ingrid; Medland, Sarah E.; Mägi, Reedik; Lawlor, Deborah A.; Lindgren, Cecilia M.

Published in:
Nature Communications

DOI:
10.1038/s41467-020-19742-5

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
ARTICLE

https://doi.org/10.1038/s41467-020-19742-5  OPEN

The genetic architecture of sporadic and multiple consecutive miscarriage

Triin Laisk et al.†

Miscarriage is a common, complex trait affecting ~15% of clinically confirmed pregnancies. Here we present the results of large-scale genetic association analyses with 69,054 cases from five different ancestries for sporadic miscarriage, 750 cases of European ancestry for multiple (≥3) consecutive miscarriage, and up to 359,469 female controls. We identify one genome-wide significant association (rs146350366, minor allele frequency (MAF) 1.2%, \(P = 3.2 \times 10^{-8}\), odds ratio (OR) = 1.4) for sporadic miscarriage in our European ancestry meta-analysis and three genome-wide significant associations for multiple consecutive miscarriage (rs7859844, MAF = 6.4%, \(P = 1.3 \times 10^{-8}\), OR = 1.7; rs143445068, MAF = 0.8%, \(P = 5.2 \times 10^{-9}\), OR = 3.4; rs183453668, MAF = 0.5%, \(P = 2.8 \times 10^{-8}\), OR = 3.8). We further investigate the genetic architecture of miscarriage with biobank-scale Mendelian randomization, heritability, and genetic correlation analyses. Our results show that miscarriage etiology is partly driven by genetic variation potentially related to placental biology, and illustrate the utility of large-scale biobank data for understanding this pregnancy complication.

†A list of authors and their affiliations appears at the end of the paper.
Miscarriage is defined by the World Health Organization (WHO) as the spontaneous loss of an embryo or fetus weighing <500 g, up to 20–22 weeks of gestation. Recurrent miscarriage, which is considered to be a more severe phenotype, is currently defined as two or more miscarriages, although previous definitions also include three or more (consecutive) miscarriages. It is acknowledged that using two miscarriages as the definition of recurrent miscarriage is at least in part to facilitate research, improve shared decision making with patients and provide them psychological support, rather than because of more specific evidence that this defined a unique phenotype, leaving open the question where to draw the line for separating different miscarriage phenotypes from a biological point of view.

Miscarriage is the most common complication of pregnancy, and the majority of miscarriages, both sporadic or recurrent, happen in the first trimester. Miscarriage is associated with excessive bleeding, infection, anxiety, depression, infertility and an increased lifetime risk of cardiovascular disease. The risk of miscarriage increases with maternal age, and has been associated with a range of causes; embryo and oocyte aneuploidy, parental chromosomal abnormalities, maternal thrombophilias, obesity, and endocrine and immunological dysregulation but causal underlying factors remain largely unknown. Miscarriage has a genetic component, with most studies focusing on associations of maternal genetic variants with recurrent miscarriage. A recent systematic review illustrates the small sample sizes of these studies (vast majority <200 cases) and the heterogeneous definition of cases, and as a consequence identified largely inconsistent results.

To discover and map the maternal genetic susceptibility and underlying biology of miscarriage, we combined genome-wide association study (GWAS) results of up to 69,054 cases from different ancestries (European, Chinese, UK South-Asian, UK African, African American, Hispanic American, UK Caribbean) for sporadic miscarriage, and subsequently the results of 750 cases of European ancestry for more severe multiple consecutive miscarriage in the largest genetic study of miscarriage to date. Although there is a continuous discussion on where to draw the line for classifying miscarriages as recurrent, we aimed to capture the more severe end of the phenotypic distribution and to differentiate severe cases from sporadic miscarriage, and potentially identify any differences in the underlying genetic architecture for these two conditions, and thus defined sporadic miscarriage as 1–2 miscarriages and multiple consecutive miscarriage as having had ≥3 self-reported consecutive miscarriages, or the International Classification of Diseases (ICD-10) diagnosis code N96 for habitual abortion (Supplementary Note 1).

We identify one genome-wide significant association for sporadic miscarriage and three genome-wide significant associations for multiple consecutive miscarriage. We further investigate the genetic architecture of miscarriage with biobank-scale Mendelian randomization (MR), heritability, and genetic correlation analyses. Our results show that miscarriage etiopathogenesis is partly driven by genetic variation potentially related to placental biology and illustrate the utility of large-scale biobank data for understanding this pregnancy complication.

Results

GWAS meta-analysis. We first performed a trans-ethnic GWAS meta-analysis for sporadic miscarriage, including genotype data for 69,054 cases and 359,469 female controls (Fig. 1, Supplementary Data 1 and 2). Association summary statistics were aggregated using trans-ethnic meta-regression implemented in the MR-MEGA software for GWAS meta-analysis. After post GWAS filtering for variants present in at least half (n = 11) of the 21 datasets, the trans-ethnic GWAS meta-analysis of 8,664,066 variants revealed a genome-wide significant locus on chromosome 7 (lead signal rs10270417, MAF = 1.7%, Pmeta = 6.0 × 10⁻⁹, Supplementary Data 3, Supplementary Fig. 2), driven by the Kadoorie Chinese-ancestry cohort (OREUR = 1.0 (0.9–1.0); ORKadoorie = 86.1 (21.1–350.3)). However, since it is known that the software used for cohort-level association testing in the China Kadoorie biobank (BOLT-LMM) can overestimate significance for rare SNPs (MAF < 1%) if the case fraction is <10% (MAFKadoorie = 0.04%, case fraction 8.9%), and the variant was absent from other Chinese-ancestry cohorts (BGI and UKBBCHI) due to low MAF, the variant was not taken forward for further analysis and interpretation. A population-specific effect cannot be ruled out but would require local replication.

We also performed a European ancestry only meta-analysis using METAL21, in 49,996 sporadic miscarriage cases and 174,109 female controls. After filtering for variants present in more than half of the 13 European ancestry cohorts (9,088,459 SNPs), we detected one genome-wide significant locus on chromosome 13 (rs146350366, MAF = 1.2%, Pmeta = 3.2 × 10⁻⁸, OR = 1.6 (1.2–1.6); Fig. 2, Supplementary Data 3, Supplementary Fig. 3).

Next, we performed a European ancestry only meta-analysis aggregating summary statistics in 750 multiple consecutive miscarriage cases and 150,215 controls from three participating cohorts (UKBB, EGCUT, ALSPAC), using Stouffer’s Z-score method implemented in METAL21, as the effect estimates from different cohorts were not directly comparable. Meta-analysis results were filtered to keep variants with an average MAF ≥ 0.5%, cohort-level MAF ≥ 0.1%, and that were present in at least two cohorts (n = 8,956,145). Four of the genome-wide significant signals (on chromosomes 2, 9, 11, and 21) were present in all three cohorts and had the same direction of effect (Fig. 3, Supplementary Data 3). Next, we applied the Firth test for significant variants to control for case-control imbalance, and to obtain uniform cohort-level association statistics and a summary effect estimate. This left us with three genome-wide significant signals: on chromosome 9 (rs7859844, MAF = 6.4%, Pmeta = 1.3 × 10⁻⁸, PFirth = 2.0 × 10⁻⁹, OR = 1.7 (1.4–2.0)), chromosome 11 (rs143445068, MAF = 0.6%, Pmeta = 5.2 × 10⁻⁹, PFirth = 1.8 × 10⁻¹⁰, OR = 3.4 (2.4–5.0)), and 21 (rs183453668, MAF = 0.5%, Pmeta = 2.8 × 10⁻⁸, PFirth = 2.5 × 10⁻⁹, OR = 3.8 (2.4–5.9)). The signal on chromosome 2 (rs138993181, MAF = 0.6%, Pmeta = 1.6 × 10⁻⁸), did not remain significant after the Firth test (PFirth = 1.7 × 10⁻⁷, OR = 3.6 (2.2–5.8)) (Supplementary Fig. 4a–d) and was not taken further for functional annotation analysis.

To clarify the potential genetic overlap between miscarriage phenotypes, we performed a cross-phenotype look-up of the associated loci. All the multiple consecutive miscarriage loci were statistically not significant in the spontaneous miscarriage analysis, and vice versa (see Supplementary Note 1). This indicates the genetic basis of sporadic and multiple consecutive miscarriage does not overlap, at least not for the reported loci, and lends further support to our phenotype definitions.

No overlap with previously proposed candidate genes. To our knowledge, no previous GWAS for recurrent or sporadic miscarriage have been carried out, but we checked the results for the 333 variants from a previous meta-analysis of published idio-pathic recurrent miscarriage candidate gene associations in our European ancestry meta-analyses for both sporadic and multiple consecutive miscarriage. None of these variants were genome-wide significant in either the sporadic or multiple consecutive miscarriage analysis, and only 14 (4.2%) and 11 (3%) were
estimates in individual cohorts and the size of the dot is proportional to effective sample size (calculated as $4/(1/N_{\text{cases}} + 1/N_{\text{controls}})$).

Supplementary Data 2. Summary effect estimate from inverse variance meta-analysis depicting 95% confidence intervals. Dashed red line indicates the line of no effect. Details of association testing in individual cohorts can be found in Supplementary Data 2. Summary effect estimate from inverse variance fixed effects meta-analysis.

![Diagram](image)

**Fig. 1** Overview of the included cohorts. Our trans-ethnic GWAS meta-analysis for sporadic miscarriage included data for 69,054 cases and 359,469 female controls, whereas our European ancestry only analysis included 49,996 sporadic miscarriage cases and 174,109 female controls. We also analyzed data for 750 multiple consecutive miscarriage cases, all of European ancestry.

![Diagram](image)

**Fig. 2** Results of the sporadic miscarriage European ancestry GWAS meta-analysis. 

(a) We detected one genome-wide significant locus on chromosome 13 (lead signal rs146350366, $n_{\text{cases}} = 49,996$, $n_{\text{controls}} = 174,109$). The 1000G EUR reference was used for plotting.

(b) The forest plots depict effect estimates in individual cohorts and the size of the dot is proportional to effective sample size (calculated as $4/(1/N_{\text{cases}} + 1/N_{\text{controls}})$). Error bars depict 95% confidence intervals. Dashed red line indicates the line of no effect. Details of association testing in individual cohorts can be found in Supplementary Data 2. Summary effect estimate from inverse variance fixed effects meta-analysis.

![Diagram](image)

nominally significant ($P < 0.05$) in the respective analyses (Supplementary Data 4), which is less than expected by chance (5%). Two genome-wide linkage scans, one of 44 recurrent miscarriage cases and 44 controls and the other of 38 sibling pairs affected by idiopathic recurrent miscarriage, reported loci on 6q27, 9q33.1, Xp22.122 and 3p12.2, 9p22.1 and 11q13.413 as associated with idiopathic recurrent miscarriage but none of the multiple consecutive or sporadic miscarriage associations identified in our much larger analysis overlap with these previously reported associations.

**Heritability of miscarriage.** While previous studies have shown preliminary evidence that (recurrent) miscarriage has a genetic predisposition, the heritability of miscarriage and related phenotypes has remained unquantified. We were unable to estimate the heritability for multiple consecutive miscarriage robustly due to a relatively small number of cases. However, we estimated the traditional heritability for ‘ever having miscarried’ under a classical twin model using the QIMR twin dataset, including 1853 monozygotic (MZ) and 1177 dizygotic (DZ) complete twin pairs and 2268 individuals from incomplete pairs, and found a heritability of 29% (95% CI 20–38%) for any miscarriage (Supplementary Data 5). In parallel, we used the sporadic miscarriage European ancestry GWAS meta-analysis summary statistics and the LD Score regression (LDSC) method to calculate the SNP-based heritability for sporadic miscarriage. We found the SNP-based heritability estimate to be small, with $h^2 = 1.5\%$ (SE 0.4) on the liability scale (assuming a population prevalence 20%). Similar to other complex traits, our findings show the SNP-heritability is substantially lower than the traditional heritability, suggesting that other sources of genetic variation may have a larger contribution.
Miscarriage genetically correlated with number of children. We assessed pairwise genetic correlations ($r_g$) between sporadic miscarriage and 760 other traits (Supplementary Data 6) available from the LD Hub24 and found 82 significant (FDR < 0.05) genetic correlations with European-ancestry sporadic miscarriage analysis. For example, significant genetic correlations were observed with reproductive traits (number of children ($r_g = 0.69$, se = 0.12, FDR = 2.0 × 10$^{-6}$) and age at first birth ($r_g = -0.40$, se = 0.10, FDR = 1.1 × 10$^{-5}$)) (Supplementary Data 6). The positive genetic correlation between sporadic miscarriage and number of children is consistent with observational associations between sporadic miscarriage and greater number of live births25, and with our own observations from phenotype exploration analyses in the UKBB (Supplementary Note 1, Supplementary Fig. 1). We also observed significant correlations with traits related to smoking behavior, mental health, and general well-being (Supplementary Data 6).

Miscarriage is associated with a variety of health outcomes. We also examined phenotypic associations of sporadic and multiple consecutive miscarriage with ICD-coded disease outcomes from linked hospital episode statistics in the UKBB dataset, adjusting for number of live births and women’s age and using FDR corrected $P$-values. We focused only on outcomes with at least one observation among the cases, resulting in testing >6000 ICD codes for sporadic and >1000 ICD codes for multiple consecutive miscarriage. For sporadic miscarriage, the majority of associations were related to pregnancy, childbirth and the puerperium ($P$-values ranging between 9.9 × 10$^{-79}$ and 4.4 × 10$^{-2}$; Supplementary Data 7; Supplementary Fig. 5). Sporadic miscarriage was also positively associated with a wide variety of diagnoses, including asthma ($P = 1.6 \times 10^{-20}$, OR = 1.2 (1.19–1.3)), stillbirth ($P = 5.1 \times 10^{-5}$, OR = 74.3 (10.0–549.2)), depressive episodes ($P = 1.4 \times 10^{-7}$, OR = 1.2 (1.1–1.3)), irritable bowel syndrome ($P = 3.5 \times 10^{-9}$, OR = 1.3 (1.2–1.4)), intentional self-poisoning ($P = 9.5 \times 10^{-4}$, OR = 1.6 (1.2–2.0)) and negatively associated with endometrial cancer ($P = 9.9 \times 10^{-3}$, OR = 0.8 (0.7–1.0)). Multiple consecutive miscarriage was positively associated with tubulointerstitial nephritis ($P = 7.8 \times 10^{-5}$, OR = 5.3 (2.3–12.0)), infertility ($P = 1.9 \times 10^{-18}$, OR = 7.5 (4.8–11.7)), ectopic pregnancy ($P = 6.7 \times 10^{-17}$, OR = 25.4 (12.1–53.4)), and others (Supplementary Data 8, Supplementary Fig. 5). Due to the definitions used to extract cases with three or more consecutive miscarriages from self-reported data, this group has less children compared to the other groups (Supplementary Note 1, Supplementary Fig. 1, Supplementary Data 1) and thus represents a miscarriage subphenotype with a more severe effect on fertility, which could explain the correlation with infertility diagnosis.

Although it would be interesting to evaluate the observed correlations on a genetic level, sufficiently sized datasets were not available for the majority of these diagnoses on the LD Hub24 for the UKBB (or published previously). However, we did see some support from our genetic correlation analyses (Supplementary Data 6) for depression (‘depressive symptoms’, FDR = 0.021, $r_g = 0.32$, se = 0.1) and asthma (self-reported asthma, FDR = 0.018, $r_g = 0.24$, se = 0.07).

Previously proposed risk factors for miscarriage include smoking26, alcohol use27, and body mass index (BMI)28. To explore the possible causal effect of smoking, alcohol, and BMI on sporadic miscarriage, we used a two-sample MR approach29. Summary statistics for alcohol use (drinks per week) and smoking initiation (ever smoked regularly) were obtained from the most
recent GWAS for these traits and for BMI were obtained from Locke et al. (2015). Summary associations between each variant included in the MR analyses and sporadic miscarriage were obtained from the European-ancestry meta-analysis. We harmonized the summary datasets and used inverse variance weighted (IVW)32, weighted median (WM)33, and MR-Egger31 methods to obtain a pooled estimate of the associations of genetic variants for smoking, alcohol use and BMI with SM. Results from IVW showed evidence of a causal association between smoking and SM (OR 1.16, 95% CI 1.11; 1.22), which were consistent with results from WM (OR 1.16, 95% CI 1.09; 1.24) but not with the point estimate from MR-Egger (OR 0.95, 95% CI 0.79; 1.14), though the 95% confidence intervals overlap. The P-value for MR-Egger intercept was 0.029, suggesting that IVW result is likely to be biased by unbalanced horizontal pleiotropy. We did not observe any consistent effect of BMI or alcohol use on SM (Supplementary Data 9; Supplementary Fig. 6). It is important to note that the samples were not independent, and there was some overlap between the risk factor and sporadic miscarriage GWAS samples, especially with the alcohol analyses sample (the percentage of the sporadic miscarriage samples that were in the exposure samples were 10%, 16%, and 38% for BMI, smoking, and alcohol, respectively). This might introduce bias in the results and inflate type 1 error; MR-Egger estimates are more prone to bias due to sample overlap than IVM and WM33. This may explain some of the inconsistency that we see between the MR-Egger and other results for alcohol and possibly those for smoking. Taking these findings together, our analyses suggest that smoking may causally increase the risk of SM, but we cannot exclude the possibility of horizontal pleiotropy explaining some of this effect.

We also conducted a hypothesis generating phenome-wide MR analysis of multiple consecutive miscarriage (using a per allele genetic risk score from the GWAS significant SNPs in relation to 17,037 diseases and health related traits using the PHESANT package in UKBB (n = 168,763) (Supplementary Fig. 7), but identified no robust causal associations (Supplementary Fig. 8, Supplementary Data 10).

Gene prioritization identifies links with placental biology. In order to refine the list of candidate genes identified by eQTL and chromatin interaction mapping, we used constraint measures (pLI score and observed/expected (oe) ratio) from gnomAD v2.1.1.37. The pLI score reflects the probability that a given gene is loss-of-function (LoF) intolerant, with scores ≥0.9 reflecting extreme intolerance to protein-truncating variation, which could point to a reproductive disadvantage in heterozygous LoF cases. The oe ratio is the ratio of the oe number of LoF variants in that gene. It is a continuous measure, where low oe values indicate that the gene is under stronger selection for that kind of variation compared to a gene with higher values. The upper bound of the oe 90% CI < 0.35 is recommended for thresholding. The candidate gene list was also compared to information on mouse phenotypes based on two databases38,39, and against a recently published list of genes potentially relevant to infertility or recurrent intrauterine death.40 For the sporadic miscarriage European ancestry meta-analysis signal on chromosome 9, 53 candidate SNPs and a total of 50 candidate genes were identified by chromatin interaction data: among them protein-coding genes TLE1, TLE4, PAX1, IDN, GNAQ, RASEF, SPTA1, D1, and FRMD3 (Supplementary Data 13 and 14). Of these, TLE1, TLE4, and GNAQ have pLI scores > 0.9 (Supplementary Data 15). Hi-C data showed interactions between the association signal and TLE1 and TLE4 (Supplementary Data 15). Both TLE1 and TLE4 are repressors of the canonical WNT signaling pathway, and participate in controlling placental extravillous trophoblast motility.37 Invasion of extravillous trophoblast cells into maternal decidua is essential for proper placenta formation, and dysregulation of this process has been associated with several pregnancy complications, including miscarriage.48 Additionally, there is evidence TLE1/TLE4 may also regulate gonadotropin-releasing hormone (GnRH) expression and differentiation of GnRH neurons,49 providing a potential link with gonadotropin regulation. Further, a member of the same family of proteins (Groucho/Transducin-Like Enhancer of Split), TLE6 has been shown to be associated with early embryonic lethality,50 and is known to antagonize TLE1-mediated transcriptional repression.51 On chromosome 11, both rs143445068 and rs140847838 were highlighted as potential candidate SNPs in the associated region located in the intron of NAV2. Chromatin interaction mapping highlighted another 17 candidate genes in the locus, including DBX1, HTATIP2, E2F8, ZDHHC13, and MRGPRX2 (Supplementary Fig. 10). Both NAV2 and E2F8 show evidence of constraint, and furthermore, NAV2 is listed as one of the candidate genes potentially relevant for infertility or recurrent fetal death (Supplementary Data 15). However, there is also strong evidence to support E2F8 as a potential candidate gene in this locus, as it is required for placental development and fetal viability, and speculated to suppress the invasiveness of extravillous trophoblasts in humans. The E2F transcription factor family has also been proposed as a key regulator of placental genes differentially expressed in recurrent pregnancy loss. Finally, for the association signal on chromosome 21, no other candidate SNPs in addition to the lead signal rs183453668 were identified, and a total of 10 candidate genes were suggested by chromatin interaction data, including SIK1, U2AF1, CRYAA, HSFB2BP, and RPRP1B (Supplementary Fig. 11). Of these, U2AF1 and SIK1 exhibit intolerance to LoF (Supplementary Data 15), and SIK1 is associated with early lethality in mouse and has also been shown to play a role in trophoblast differentiation.48

We then tested for colocalization between multiple consecutive miscarriage and plasma protein levels55 and expression levels in 49 different tissues56,57 using coloc58 (Supplementary Methods). Colocalization can highlight potential mediating genes and tissues by investigating the likelihood of a shared causal variant between a disease and e.g. quantitative trait loci. There was no evidence for colocalization for any of the risk locus–gene/protein–tissue combinations, potentially reflecting that the risk is mediated through other genes than those investigated, or lack of data from the relevant tissue or time-point.
In this study, we quantify the heritability of miscarriage and identify four distinct susceptibility loci for sporadic and multiple consecutive miscarriage. We acknowledge the fact that the definition we used for multiple consecutive miscarriage differs from that currently used for recurrent miscarriage. Our rationale behind using a definition of at least three consecutive miscarriages (which was also the definition of recurrent miscarriage in much of Europe at the time of the study) was to capture the more severe end of the phenotypic spectrum, which would allow to better assess whether recurrent and sporadic events lie on the same phenotypic spectrum. Our findings that the used sporadic miscarriage definition is genetically (and phenotypically) correlated with number of children, whereas the definition used for extracting cases with three or more consecutive miscarriages from self-reported data that resulted in a group that had less children and showed correlations with infertility diagnosis confirm that the used phenotype definitions captured the contrasting ends of the phenotypic spectrum. Hopefully our study paves the way for future similar studies into the genetics of miscarriage which could provide further evidence for more informed classification of this phenotype.

We found heritability of 29% (95% CI 20–38%) for any miscarriage and a considerably smaller SNP-based heritability for sporadic miscarriage ($h^2 = 1.5\%$ (SE 0.4)). Our study design is limited to interrogate maternal contribution to the genetic architecture of the trait, and it is likely that paternal and fetal contributions are responsible for a proportion of the heritability gap. Overall, it might be expected that genetic factors increasing susceptibility to miscarriage are under negative selection due to their impact on reproductive fitness and hence these will be rare. Up to two-thirds of miscarriages are unrecognized and/or undiagnosed, particularly early miscarriages, and thus ‘control’ women will include some misclassified as not having experienced a miscarriage. This would be expected to attenuate results towards the null and means larger numbers may be
required to accurately quantify SNP-heritability and identify genome-wide significant SNPs; this is likely to affect sporadic miscarriage more so than recurrent miscarriage.

Our results confirm a partly genetic basis to both sporadic and multiple consecutive miscarriage, that does not seem to overlap. Mapping of potential candidate genes at associated loci identified several genes (FGF9, TLE1, TLE4, E2F8, SIK1) with a plausible biological role in placental biology and miscarriage etiopathogenesis. However, the involvement of other transcripts at these loci cannot be ruled out and further functional studies are needed. Similarly to other traits, our larger GWAS study fails to replicate findings from previous smaller candidate gene studies13,15,25, underlining the importance for future larger collaborations to jointly dissect the genetic background of miscarriage.

The MR analyses suggest that smoking may causally increase the risk of sporadic miscarriage, as has also been suggested by epidemiological studies;26 however, this needs to be validated in independent datasets and the effect of horizontal pleiotropy cannot be ruled out. Finally, our analysis of health outcomes associated with miscarriage confirms previous observations and identifies several novel ones. For some of these diagnoses, including irritable bowel syndrome, asthma, endometrial cancer, self-harm, and ectopic pregnancy, similar epidemiologic associations have been reported previously61–65, warranting further investigation into underlying mechanisms.

In conclusion, our results show that miscarriage etiopathogenesis is partly driven by genetic variation potentially related to placental biology, and illustrate the utility of large-scale biobank data for understanding this pregnancy complication.

Methods

Study cohorts. Descriptive statistics of the cohorts included in the sporadic and multiple consecutive miscarriage GWAS meta-analyses are presented in Supplementary Data 1 and Supplementary Methods. Our analysis included 69,054 sporadic miscarriage cases, 750 multiple consecutive miscarriage cases and up to 359,469 controls. All individuals gave informed consent at enrollment and recruitment was monitored by relevant institutional ethics review boards.

Case definitions. Depending on the type of data available in individual cohorts, miscarriage cases were identified as follows:

- Sporadic miscarriage: 1 or 2 self-reported miscarriages, or ICD-10 codes O02.1 and O03 on 1 or 2 separate time-points (at least 90 days between episodes). As a result, 26,044 sporadic miscarriage cases were identified from cohorts using only self-report data, 12,031 from cohorts using only electronic health record data, and 41,779 from cohorts where both self-reported and EHR data were available.

- Multiple consecutive miscarriage: (i) five or more self-reported miscarriages, one live birth, no pregnancy terminations, (ii) three or more self-reported miscarriages, no live births, no pregnancy terminations, or (iii) three or more consecutive miscarriages. The first two criteria were used to ensure the consecutive nature of the miscarriages; and (iv) ICD-10 diagnosis code N96.

For a more detailed explanation why these definitions were chosen, please see Supplementary Note 1, Supplementary Fig. 1 and Supplementary Table 1.

Association analyses. Details on how association analyses were carried out on the cohort level can be found in Supplementary Data 2. Cohort-level association analyses had been performed using genotype data imputed to suitable reference panels and adjusted for year of birth. Where available and appropriate, additional cohort-specific covariates, such as principal components or genotyping array, were used to correct for potential within-cohort stratification.

Meta-analysis. Central QC was conducted using EasyQC66. During central QC, allele frequencies and alignment were compared against available reference datasets (HapMap Reference Consor77, HapMap Genotyper68) to determine potential strand issues or large allele frequency deviations from the reference population. Monomorphic markers, and also markers with strand mismatch, poor imputation quality (INFO score < 0.4) or an arbitrary minor allele count cutoff ≤5 were excluded from each study prior to the meta-analysis. The results from individual cohorts were meta-analyzed in parallel by two different analysts. All genome-wide significant variants that passed the applied filters (see below) are listed in Supplementary Data 3.

For the trans-ethnic meta-analysis, we used the MR-MEGA software19, adjusting for the first two principal components. After the meta-analysis, we applied an additional filter for markers present in at least half of the cohorts, to rule out spurious associations. This resulted in 8,664,066 variants, and a genome-wide significant association on chromosome 7 (rs10720417). Indels were not considered due to their lower quality. A closer inspection of the effect sizes for the observed association in individual cohorts revealed the association was mainly driven by one of the Chinese-ancestry cohorts (Supplementary Fig. 2c) where the MAF was 0.04%. It is known that BOLT-LMM, used for analysis in the Kadoorie cohort, can overestimate significance for rare SNPs (MAF < 1%) if the case fraction is <10%; therefore we performed additional analyses. The Kadoorie samples have been collected from 10 different region centers, therefore we checked for batch effects. Adding batch ID as a covariate did not have a significant impact on the association statistics (original P-value 5.0 × 10−10, after adding batch ID as covariate P = 2.2 × 10−15). To check possible confounding effect from samples being collected from 10 different region centers, we performed separate analyses for each region center followed by meta-analysis. As the SNP is very rare, SNPTST failed to converge in five research datasets; however, after meta-analysis detected a similar effect direction in the remaining datasets, although with significant differences in effect magnitude across different region centers (Pmeta = 4.8 × 10−4; Preg = 5.2 × 10−5) and a considerably larger P-value compared to the BOLT-LMM results. Although the two methods (SNPTST and BOLT-LMM) are not directly comparable, given the absence of this variant in other Chinese ancestry cohorts (BGI and UKBBCHI), the rs10720417 signal was not taken further for functional annotation.

European-ancestry only sporadic miscarriage meta-analysis was carried out with METAL13 using inverse variance fixed effects meta-analysis and single genomic correction. Multiple consecutive miscarriage meta-analysis was conducted using METAL21 and Stouffer’s (P-value-based effective sample size weighted) method and single genomic correction. After the analysis, sporadic miscarriage European ancestry meta-analysis results were additionally filtered to exclude markers not present in at least half of the cohorts (n = 7). From the multiple consecutive miscarriage meta-analysis results we excluded variants that were not present in at least two cohorts, had an average MAF of <0.5%, and a MAF of <0.1% in any of the three cohorts. From the genome-wide significant signals we selected those with the same direction of effect in all three cohorts for further consideration. Indels were not considered due to their lower quality. The quantile–quantile plots, Manhattan plots, and locus zoom plots of the meta-analyses are shown in Figs. 2 and 3 in Supplementary Figs. 3 and 4.

Look-up of variants previously associated with recurrent miscarriage. We conducted a lookup in our summary statistics of all variants included in a recent, extensive and systematic review/meta-analysis of published genetic association studies in idiopathic recurrent spontaneous abortion15. The results of this lookup are given in Supplementary Data 4.

Heritability analysis. The sporadic miscarriage GWAS European-ancestry meta-analysis summary statistics and LDSC method36 were used for heritability estimation. The linkage disequilibrium (LD) estimators, European ancestry samples in the 1000 Genomes projects were used as a reference. Heritability estimates were converted to the liability scale using a population prevalence of 0.2 for sporadic miscarriage. Using the UKBB SNP-Heritability Browser (https://nealelab.github.io/UKBB_ldsc/b2_browser.html), we also did a look-up for different versions of the miscarriage phenotype using the imputed data from the UKBB dataset and observed similar heritability estimates for ‘Ever had stillbirth, spontaneous miscarriage or termination’ (HF = 0.04; s.e. = 0.007; population prevalence 31.7%) and ‘Number of spontaneous miscarriages’ (HF = 0.03; s.e. = 0.01).

Data from 1853 complete female MZ and 1177 DZ twin pairs and 2268 women from incomplete or opposite sex twin pairs (mean year of birth 1954, range 1893–1989) from the QIMR dataset were used to estimate heritability under a classical twin model, using a multifactorial threshold model in which discrete traits are assumed to reflect an underlying normal distribution of liability (or predisposition). Liability, which represents the sum of all the multifactorial effects, is known to reflect the combined effects of a large number of environmental factors each of small effect69. All data analyses were conducted using maximum-likelihood analyses of raw data within R67. Corrections for year
of birth were included with the model, such that the trait value for individual $i$ from family $f$ was parameterized as $y_i = \beta_0 + \beta_1 f_i + \epsilon_i$. The phenotypic data, which were collected at age 30, was parameterized as $y_i = \sigma_1^2 + \epsilon_i$ or $\epsilon_i = \sigma_2^2 + \epsilon_i$, where, $\sigma_1^2$ represents additive genetic effect ($A$); $\sigma_2^2$ represents non-additive genetic effects ($D$); $\epsilon_i$ represents shared environmental effects (C) and $\epsilon_i$ represents non-shared or unique environmental effects (E). The covariance terms were parameterized as $\text{Cov}_{y_{ij}} = \sigma_1^2 + \sigma_2^2$ or $\sigma_1^2 + \epsilon_i$ and $\text{Cov}_{y_{ij}} = 0.5\sigma_1^2 + 0.25\sigma_2^2$ or $0.5\sigma_1^2 + \epsilon_i$.

The significance of variance components was tested by comparing the fit (minus twice the log-likelihood) of the full model which included the effect to that of a nested model in which the effect had been dropped from the model. The difference in log-likelihoods follows an asymptotic chi-square distribution with the degrees of freedom equal to the difference in estimated parameters between the two models. The results of these analyses are summarized in Supplementary Data 5.

**Genetic correlation analyses.** The LDSC method was implemented in LD-Hub (http://lsmc.broadinstitute.org) to test for genetic correlations between sporic miscarriage traits. For sporadic miscarriage, we used the European-ancestry only GWAS meta-analysis summary statistics and data available within the LD Hub resource. False discovery rate (FDR) multiple testing correction at level 10% was used to account for multiple testing. Results of the analysis are presented in Supplementary Data 6.

**Associated phenotypes analysis.** The UKBB has extensive phenotype data for 500K individuals, of which 273,465 are women. After applying the same QC criteria that were applied to the subset of individuals used in the genetic association analysis, extensive phenotype data was available for 220,804 women. Among these there were 39,411 sporadic miscarriage cases and 133,545 controls, and 458 multiple consecutive miscarriage cases and 133,545 controls (defined using the criteria applied to define cases and controls for the association analysis). We explored differences in the prevalence of diseases between sporadic miscarriage cases and controls and between multiple consecutive miscarriage cases and controls. Diseases were identified from the UKBB linked Hospital Episode Statistics which provide ICD10 diagnosis codes (UKBB data fields 41202 and 41204). First, for each hospital diagnosis observed among the cases (defined by an ICD10 code; n = 6840 for sporadic and n = 1323 for multiple consecutive miscarriage, respectively; excluding those used to define the cases), we tested the difference in the proportion of cases with the diagnosis with that among the controls (using a two-sample test for difference in proportions when the number of successes is >500). If there was statistical evidence of a difference between cases and controls, and between multiple consecutive miscarriage cases and controls. Diseases of the respiratory system were included with the model, such that the trait value for individual $i$ from family $f$ was parameterized as $y_i = \beta_0 + \beta_1 f_i + \epsilon_i$. The phenotypic data, which were collected at age 30, was parameterized as $y_i = \sigma_1^2 + \epsilon_i$ or $\epsilon_i = \sigma_2^2 + \epsilon_i$, where, $\sigma_1^2$ represents additive genetic effect ($A$); $\sigma_2^2$ represents non-additive genetic effects ($D$); $\epsilon_i$ represents shared environmental effects (C) and $\epsilon_i$ represents non-shared or unique environmental effects (E). The covariance terms were parameterized as $\text{Cov}_{y_{ij}} = \sigma_1^2 + \sigma_2^2$ or $\sigma_1^2 + \epsilon_i$ and $\text{Cov}_{y_{ij}} = 0.5\sigma_1^2 + 0.25\sigma_2^2$ or $0.5\sigma_1^2 + \epsilon_i$.

The significance of variance components was tested by comparing the fit (minus twice the log-likelihood) of the full model which included the effect to that of a nested model in which the effect had been dropped from the model. The difference in log-likelihoods follows an asymptotic chi-square distribution with the degrees of freedom equal to the difference in estimated parameters between the two models. The results of these analyses are summarized in Supplementary Data 5.

**Functional annotation.** The FUMA platform designed for prioritization, annotation and interpretation of GWAS results was used for functional annotation of association signals from the GWAS meta-analyses (Supplementary Methods).

To narrow down potential candidate genes, we first extracted SNPs in high LD ($r^2 > 0.8$) with the lead signals and overlapped those with chromatin data (ChIPseq for histone modifications and DHS for chromatin accessibility, both indicative of regulatory elements) using HaploReg. We then used Hi-C chromatin interaction datasets to visualize topologically associated domains (TADs) in the region and Capture Hi-C data for various tissues to further explore interactions within the TAD domain. Data was visualized using the 3D Genome Browser (http://3dgenome.org) using the datasets available via the browser11. TADs are relatively conserved across different tissue types and define the boundaries for potential genomic interactions. In the main text (Fig. 4), visual representations are shown for single tissue to show representative signals at a given locus and to illustrate chromatin architecture (promoter-enhancer interactions) at the locus, to infer possible target genes. The visual representations do not infer/confirm potential tissue-specific interactions for the associations.

For sporadic miscarriage, we used the summary statistics of our EUR-ancestry meta-analysis. A total of five candidate SNPs were identified ($r^2 > 0.6$ with rs164350366 in the associated locus on chr13, all of them intergenic (Supplementary Data 14), among them protein-coding genefi and certain disease genes (Table 1), including those involved in the immune mechanism, D50-D89), circulatory (“Diseases of the circulatory system”, I00-I99), congenital (“Congenital malformations, deformations and chromosomal abnormalities”, Q00-Q99), digestive (“Diseases of the digestive system”, K00-K93), ears (“Diseases of the ear and mastoid process”, H60-H65), endocrine (“Endocrine, nutritional and metabolic diseases”, E00-E90), musculoskeletal (“External causes of morbidity and mortality”, V01-V98), eyes (“Diseases of the eye and adnexa”, H00-H05), genitourinary (“Diseases of the genitourinary system”, N00-N99), infectious (“Certain infectious and parasitic diseases”, A00-B99), injury (“Injury, poisoning and certain other consequences of external causes”, S00-T98), musculoskeletal (“Diseases of the musculoskeletal system and connective tissue”, M00-M99), neoplasms (“Neoplasms”, C00-D48), nervous (“Diseases of the nervous system”, G00-G99), other (“Factors influencing health status and contact with health services”, U04-Z99), perinatal (“Certain conditions originating in the perinatal period”, P00-P96), pregnancy (“Pregnancy, childbirth and the puerperium”, O00-O99), psychiatric (“Mental and behavioral disorders”, F00-F99), respiratory (“Diseases of the respiratory system”, J00-J99), skin (“Diseases of the skin and subcutaneous tissue”, L00-L99), symptoms (“Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified”, R00-R99). To rule out the confounding effect of woman’s age and parity, we then conducted multivariate logistic regression adjusted for age and number of children for any diseases for which there was statistical evidence of a difference between cases and controls, and applied FDR 5% correction for the $P$-values (Supplementary Data 7 and 8).

**MR analyses.** We used a two-sample MR approach to explore the possible causal effect of smoking, alcohol, and BMI on sporadic miscarriage. Summary statistics for alcohol use (drinks per week) and smoking initiation (ever smoked regularly) were obtained from the most recent GWAS for these traits28 and for BMI were obtained from Locke et al. (2015)29. Summary associations between each variant and sporadic miscarriage were obtained from the European-ancestry meta-analysis filtered for variants present in at least half of the cohorts. We harmonized the summary datasets and used IVW32, WM33, and MR-Egger34 methods to obtain a pooled estimate of the associations of genetic variants for smoking, alcohol, and BMI with sporadic miscarriage. Including only variants that reached genome-wide significance in the exposure GWAS in our MR analyses, 369/378 variants associated with smoking, 96/99 variants associated with alcohol use, and 97/97 variants associated with BMI were present in our European-ancestry sporadic miscarriage GWAS. The results of these analyses are presented in Supplementary Data 9. Using a hypothesis-free (hypothesis-free) MR approach, we conducted multiple multiple testing correction (using a per allele genetic accuracy score from the four GWAS significant SNPs; rs7859844, rs14345068, rs138993181, rs185435668) to relation to 17,037 outcomes using the PHESANT35 package in (Supplementary Fig. 7). The analysis was adjusted for year of birth and the top 10 PCs. Overall, the MR-PheWAS did not show any evidence of causal effects (Supplementary Fig. 8). Only three outcomes reached Bonferroni-corrected levels of statistical significance ($P < 2.3 \times 10^{-8}$), including one outcome related to alcoholism and one related to post-traumatic stress disorder. However, both of these were single items from instruments that included 11 items (alcohol use questions) and 21 items (post-traumatic/stress event questionnaire), respectively, with none of the other items reaching suggestive thresholds of statistical significance. The third outcome to show association below this $P$-value threshold was a job code (scenery designer or costume designer) that is one of which lies in 42-item employment history category (MR analyses did not suggest effects on any other jobs in this list).

**Other methods.** Use of the mouse phenotype ontology (https://phenomecentral.org) and the Mouse Genome Informatics database38 (http://www.informatics.jax.org/phenotypes.shtml) is mentioned. In brief, the $P$-value reflects the probability that a given gene is LoF intolerant and pLI scores ≥0.9 reflect extreme intolerance to protein-truncating variation, which could point to a reproductive disadvantage in heterozygous LoF cases. The oe ratio is the ratio of the oe of LoF variants in that gene. It is a continuous measure, where lower oe values indicate the gene is under stronger selection for that kind of variation compared to another. For the upper bound of genefi = 0.35 is recommended for thresholding. In mouse phenotype data, we first and foremost looked at potentially early (embryonic/prenatal/preweaning) lethal phenotypes. We
also compared our candidate genes against a recently published carefully curated list of genes potentially relevant to infertility or recurrent intra-uterine death, which were selected based on scores of genetic constraint, murine phenotypic information, and the cell 'essentialism', i.e., genes that are essential for human cell lines.

Using this approach, in each locus the most likely candidate genes were considered to be those that either show evidence of evolutionary constraint and are therefore relevant for reproductive fitness or are associated with an early lethal phenotype in mice (Supplementary Data 15). Previous publications showing evidence the gene is linked with either placental function, embryo viability, phenotype in mice (Supplementary Data 15). Previous publications showing evidence the gene is linked with either placental function, embryo viability, phenotype in mice (Supplementary Data 15).

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The GWAS meta-analysis summary statistics that support the findings of this study are available for download from http://www.geenivaramu.eu/tools/mifs_sumstats.zip. The analyses in this manuscript included data from UK Biobank, http://www.ukbiobank.ac.uk/, under applications 17805, 11867, and 16729; Estonian Biobank, https://www.geenivaramu.eu/en; A1SPAC (http://www.bristol.ac.uk/a1spac/); China Kadoorie Biobank (http://www.kdb.org.hk). All QC and GWAS analyses were carried out with standard tools and pipelines. The analyses in this paper also use data from the 3D Genome Browser, http://promoter.bx.psu.edu/hc-c/; GTEx, https://gtexportal.org/home/; International Mouse Phenotype Consortium, https://www.mousephenotype.org; Mouse Genome Informatics database; http://www.informatics.jax.org/phenotypes.shtml; GWAS atlas, https://atlas.cglab.nl/

**Code availability**

Cohort-level analyses were carried out with SNPESTX v2.5.0, BOLT-LMM v2.3.2, EPACTS 3.3, plink v1.9, RAREMETALWORKER,adget2dat. Before central meta-analysis, data quality control was conducted using Easy QC software (v17.6), central meta-analysis was conducted using the MR-MEGA (0.1.5;https://genomics.ut.ee/en/tools/mr-mega) and METAL (version released on 2011-03-25;http://genome.sph.umich.edu/wiki/METAL_Documentation) software. Follow-up analyses were in part carried out using FUMA (1.3.1;http://fuma.cglab.nl/), using data from ANNOVAR (17-07-2017) and GWAS catalog (e91_r2018-02-06). Gene-based testing was carried out with MAGMA 1.06 implemented in FUMA. Functional annotations of variants were obtained from HaploReg 4.1. SNP heritability and genetic correlations were calculated using LDSC software (1.0.0) and LDSC software implemented in LD-Hub (http://ldsc.broadinstitute.org/), respectively. Colonization analyses were performed using R versions 3.4.3 and 3.5.1 (https://www.r-project.org/), the colco R package and the coloc.abf() function and LocusCompareR (http://www.locuscompare.com/) was used for visualization. MR-PheWAS was conducted using the PHEASANT R package. Fig 1 was created using the maps package in R. All other analyses were carried out in the R environment.

Received: 16 September 2019; Accepted: 23 October 2020; Published online: 25 November 2020

**References**


17. Cohain, J. S., Buxbaum, R. E. & Mankuta, D. Spontaneous first trimester miscarriage rates per woman among parous women with 1 or more pregnancies of 24 weeks or more. BMC Pregnancy Childbirth 17, 437 (2017).


62. W.M., A.P.M., P.B.M., P.R.N., D.R.N., M.L., S.S., A.S., K.Z., I.G. and D.A.L. have received support from Roche Diagnostics and Medtronic Ltd for biomarker research unrelated to that presented here. The other authors have no competing interests.
Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark. Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. Department of Epidemiology & Biostatistics, Peking University Health Science Centre, Peking University, Beijing, China. Department of Integrative Biology, University of California Berkeley, Berkeley, CA, USA. Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. Bristol National Institute of Health Research Biomedical Research Centre, Bristol, UK. Program in Medical and Population Genetics, Broad Institute, Boston, MA, USA. These authors contributed equally: Triin Laisk, Ana Luiza G. Soares, Teresa Ferreira, Jodie N. Painter. These authors jointly supervised this work: Ingrid Granne, Sarah E. Medland, Reedik Mägi, Deborah A. Lawlor, Cecilia M. Lindgren.  

✉email: triin.laisk@ut.ee; cecilia.lindgren@bdi.ox.ac.uk