# **scientific** reports

OPEN

# TiMEG: an integrative statistical method for partially missing multi-omics data

Sarmistha Das[1,2] & Indranil Mukhopadhyay[1] ✉

Multi-omics data integration is widely used to understand the genetic architecture of disease. In multi-omics association analysis, data collected on multiple omics for the same set of individuals are immensely important for biomarker identification. But when the sample size of such data is limited, the presence of partially missing individual-level observations poses a major challenge in data integration. More often, genotype data are available for all individuals under study but gene expression and/or methylation information are missing for different subsets of those individuals. Here, we develop a statistical model TiMEG, for the identification of disease-associated biomarkers in a case–control paradigm by integrating the above-mentioned data types, especially, in presence of missing omics data. Based on a likelihood approach, TiMEG exploits the inter-relationship among multiple omics data to capture weaker signals, that remain unidentified in single-omic analysis or common imputation-based methods. Its application on a real tuberous sclerosis dataset identified functionally relevant genes in the disease pathway.

Advances in next-generation sequencing (NGS) technologies have led to harnessing robust structural and functional knowledge of the human genome. Thus, NGS provides unprecedented opportunities to understand health and disease at the present time[1–4]. But the transformation of knowledge from bench to bedside lies in mining the massively available data for genes and variants of high clinical relevance[5]. To understand the genetic architecture of disease, genome-wide association studies[6,7] and several other studies based on single-omic data such as gene expression or DNA methylation have catalogued many disease-associated loci. Nonetheless, we are yet to understand the aetiology of many complex diseases as they occur due to an intricate interplay of various genetic elements[8].

Multi-omics data integration serves as a springboard for unravelling such inherent complexity underlying the genetic architecture of disease[8,9]. Data integration through statistical and/or computational models plays a major role in the prediction of genomic and environmental perturbations underlying disease/complex traits, and transferring preclinical knowledge to clinical trials with increased speed and accuracy[10,11]. Statistical models[12–16] for data integration enhance the predictive power of gene-disease association, by incorporating prior knowledge of regulatory relation among different omics data and analysing them under one statistical framework.

With access to enormous data from several consortiums on various omics data, many data integration techniques have been developed so far. Such methods comprise of dimension reduction techniques, gene regulatory networks, feature selection techniques using supervised, unsupervised or semi-supervised learning, graph or kernel-based techniques, Bayesian, and frequentist approaches[16–23]. But more often, integration methods combine multiple omics data from large consortiums of different cohorts[15,24]. Such methods are prone to spurious prioritisation of associated genes owing to substantial cross-cell-type variation[25]. For these reasons and to reduce the stratification bias due to population diversity, increasing attempts are being made to create large scale multi-omics datasets recently by combining multiple assays from the same set of samples[26].

However, individual research groups made substantial efforts for generating data on genetic variation, gene expression, methylation, phenotype, etc. simultaneously from the same set of samples to study biomarkers. Realising its great potential, data sharing platforms/repositories store such heterogeneous data for the broader markets demanding immediate study[26–28]. But unlike large consortium data, these data might have a relatively small sample size. In addition, missing data occur across multiple omics. For example, while integrating information from different types of genomic data, typically, genotype information is available for all the individuals but gene expression and/or methylation information are often partially missing[29]. Yet gene expression and/or methylation assays are rarely repeated for generating the missing data due to various reasons such as the huge cost

[1]Human Genetics Unit, Indian Statistical Institute, Kolkata 700108, India. [2]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis 38105, USA. ✉email: indranil@isical.ac.in
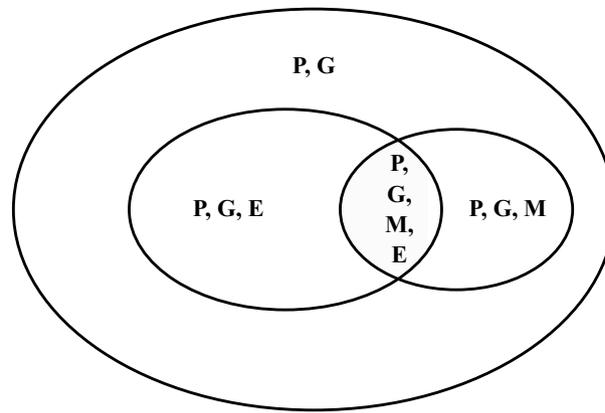
**Figure 1.** Structure of data availability for genotype (G), gene expression (E), methylation (M) and phenotype (P). Each letter indicates the presence of corresponding data.

of the assays[2], degradation of mRNA[30] and/or dearth of tissue samples, etc. So, a major challenge is to integrate multi-omics data in presence of partially missing individual-level observations[26].

Few Bayesian methods[31,32] consider missing value imputation in multi-omics data integration model. But sometimes imputed data overshadow the contribution from the partially observed data for certain percentages of missing data and generally involves a huge computational cost to decide whether or not to impute[31]. Moreover, imputing the missing values might be misleading[25] as it introduces bias and uncertainty in the data[33], especially when the missing percentage is large and/or the reason for missing is unclear. To deal with the missing values it is important to understand the data source, data structure, missing mechanism, and amount of missing data[34] along with its relation with the phenotype[33]. Some network-based methods[35] considers partial multi-omics data integration using a similarity network but assume the same contribution of different omics.

In this paper, we propose a multi-omics genetic association tool, called TiMEG (Tool for integrating Methylation, gene Expression and Genotype), for the identification of disease-associated biomarkers, by integrating single nucleotide polymorphism (SNP), gene expression, and DNA methylation with partially missing omics data under case–control paradigm. Our method elucidates the effect of multiple omics on the qualitative phenotype (case-control status). It jointly dissects the information on various omics data, their inter-relationship, and the information from individuals with completely as well as partially available omics data, without imputing the missing data. Using a likelihood-based approach, TiMEG models the conditional distribution of the response variable using the missing predictor variables[36,37].

Asymptotic distribution of our test statistic under the null hypothesis of no genetic association computes $p$-values much faster compared to other computational methods like permutation-based or resampling techniques etc. Extensive simulation confirms robust performance in terms of prediction accuracy of estimation in tenfold cross-validation, controlled type I error rate, high statistical power, and consistency of the test under different missing data schemes. Application of our method on a real dataset of tuberous sclerosis (also called tuberous sclerosis complex (TSC)) patients and healthy controls (phs001357.v1.p1) identified functionally relevant genes and gene clusters belonging to the pathways that are involved in TSC pathogenesis. Even with small sample size and a substantially high percentage of missing data, TiMEG could be used for the identification of biomarkers without losing any information. This leads to capturing weaker signals that remain unidentified in larger single-omic data analysis.

## Results

**TiMEG method.** In presence of limited sample size, missing individual-level information on multiple assays poses a great loss of information. Imputation might lead to bias in such a small sample size as the percentage of missing data is large. We introduce TiMEG, a general analytical approach for the identification of biomarkers associated with a disease by integrating multiple omics data with/without missing individual-level omics information under a case–control paradigm. We integrate data from DNA sequencing, gene expression, and DNA methylation assays along with covariates and qualitative phenotype (disease status) from the same set of samples. Figure 1 gives a general structure of omics data availability. Based on Fig. 1, we design our missing data schemes (Table 1).

TiMEG relies on a likelihood-based approach to gather information on the missing data by estimating the parameters in the likelihood function containing incomplete omics data rather than imputing the missing data before the analysis. To obtain the likelihood function, we find the probability distribution of the response variable (disease status) conditional on the available omics information by (1) integrating out the missing variable from the joint likelihood function (see Methods), and (2) exploiting the interdependence among the multiple omics data.

Since an individual's gene expression level could be regulated by alteration in the DNA sequence, TiMEG considers the effect of genotype and methylation on gene expression. Similarly, it incorporates the effect of genotype on methylation and also the effect of genotype, methylation, gene expression, and covariates on disease

2

| Sample size | Data type | | | | |
|---|---|---|---|---|---|
| | Phenotype | Covariates | Genotype | Gene expression | Methylation |
| $n_1$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $n_2$ | ✓ | ✓ | ✓ | ✗ | ✓ |
| $n_3$ | ✓ | ✓ | ✓ | ✓ | ✗ |
| $n_4$ | ✓ | ✓ | ✓ | ✗ | ✗ |

**Table 1.** Missing data schemes. Notation: '✓' indicates 'available', '✗' means 'missing' data. For $n_1$ individuals, no data is missing; for $n_2$ ($n_3$) individuals only gene expression (methylation) data are missing; both gene expression and methylation data are missing for $n_4$ individuals.

status. We assume that after appropriate transformation and normalization procedures, individual-level gene expression and methylation data follows a bivariate normal distribution[13].

**Features of TiMEG.** TiMEG is a statistical tool for the identification of disease-associated combinations of multiple omics. For example, a significant combination provides a gene along with a cis-methylation and a cis-genotype (collectively called a 'trio' throughout the paper). In this article, we illustrate a detailed pipeline for the identification of significant trios, if at least one component of the trio is associated with the disease.

Other advantages of TiMEG include the ability:

- to theoretically handle an unrestricted number of missing transcriptomic and epigenetic data as elucidated in the Methods section and confirmed using simulations,
- to capture weaker signals that remain unidentified in single-omic analysis by incorporating more information from the inter-relation among multiple omics data in the likelihood framework,
- of robust performance in terms of prediction accuracy of estimation in tenfold cross-validation, controlled type I error rate, and high statistical power,
- of efficient incorporation of correlated omics information to decipher significant signals resulting in reduced false-negative rate, and
- prompt calculation of *p*-values using the asymptotic distribution as opposed to computation-intensive permutation-based resampling methods.

Since alteration in gene expression regulation is expected to alter phenotype more than any change in the DNA sequence, the performance of TiMEG reveals that the effect (in terms of statistical power) of a certain percentage of missing gene expression data is more than the same percentage of missing methylation data (see Simulations). Missing both omics information for a subset of individuals will lead to much more loss of information and therefore statistical power than missing gene expression or methylation data on any subset of equivalent size. Thus, TiMEG agrees with the biologically accepted notion. Regardless of the sample size and/or percentage of missing omics data, wet-lab researchers will be able to promptly identify significant biomarkers from their data by calculating *p*-values using this tool.

**Performance of TiMEG.** *Simulations.* We perform extensive simulations to study the performance of TiMEG for varying percentages of missing omics data under different missing data schemes. As more often gene expression and methylation data are missing for a subset of genotyped individuals, we assume that genotype, phenotype, and covariate data are available for the entire sample of size $n$ (say). As shown in Table 1, for these $n$ individuals, there could arise four different scenarios (1) none of the other two omics data is missing for a certain subset of size $n_1$ (say), (2) only gene expression data are missing for another subset of size $n_2$ (say), (3) only methylation data are missing for the third subset of size $n_3$ (say), and (4) both gene expression and methylation data are missing for the remaining subset $n_4$ (say). Here, we consider two covariates, age and gender and simulate them from $N(40, 6)$ and $Bin(1, 0.5)$ respectively. First, we simulate data under scheme 1 i.e when there is no missing observation.

For generating genotype data, we assume an SNP having two alleles $A$ and $a$ with $A$ as a minor allele. Considering di-allelic loci, we simulate genotype data from Bernoulli distribution assuming Hardy-Weinberg Equilibrium (HWE) for controls with minor allele frequency (MAF) 0.2 for associated SNP. We generate genotypes for cases using additive model for relative risk[38] based on disease prevalence = 0.1, genotypes ($A$, $Aa$, and $aa$), and relative risk = 1.2.

Next, we generate methylation and gene expression values using Eqs. 2 and 3 (see Methods) assuming the values of the parameters as $\alpha_0 = 1.3, \alpha_g = 2.4, \gamma_0 = 1.9, \gamma_g = 0.6, \gamma_m = 2.3$. Methylation-gene expression pair follows a bivariate normal distribution with variances $\sigma_1^2 = \sigma_2^2 = 1$. We assume that the means of the bivariate distributions for cases and controls differ by 0.3. Now, based on covariates, genotype, gene expression, and methylation, we simulate the phenotype of each individual using Bernoulli distribution from Eq. 1 (see Methods) with parameters $\beta_0 = 1, \boldsymbol{\beta_x}' = 0.01\mathbf{1}', \beta_g = 0.1, \beta_m = 0.2, \beta_e = 0.3$. We generate data for $n$ cases and $n$ controls where the sample size $n$ is taken as 100, 150 and 200. For the sake of power comparison, we take equal sample sizes for cases and controls. However, our method works for unequal sample sizes for cases and controls as well.
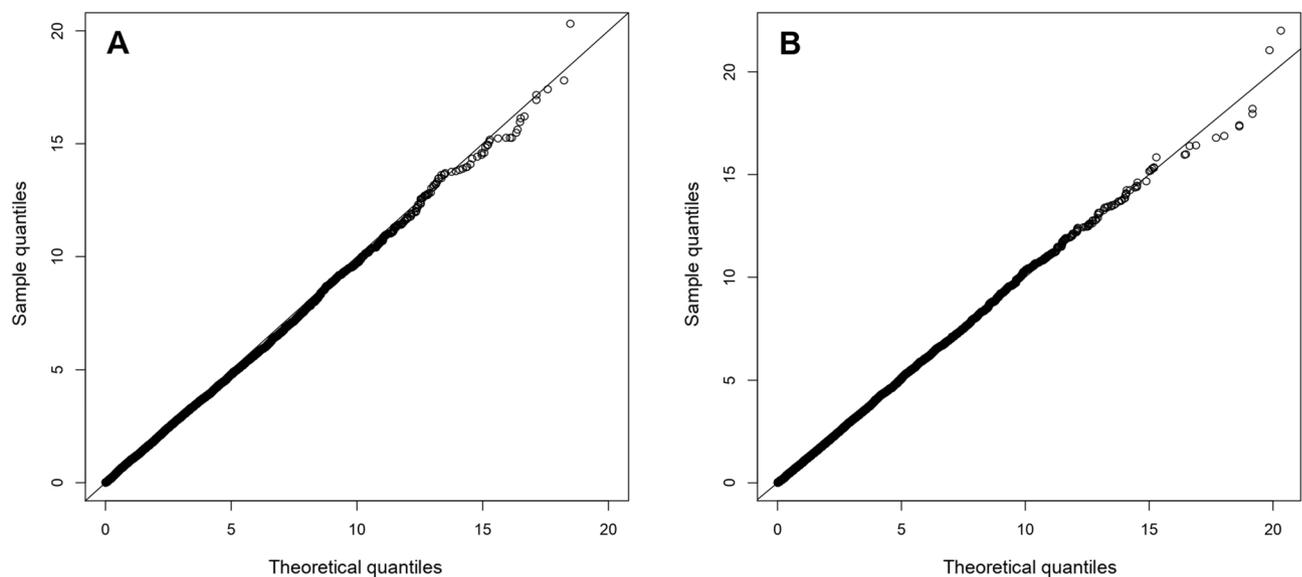
**Figure 2.** QQ-plot with sample size 200 based on the performance of simulated data. (**A**): QQ-plot with no missing data, (**B**): QQ-plot with 10% both gene expression and methylation missing, 10% only methylation and 20% only gene expression missing.

| | SS | | | | | | | SS | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | 100 | | 150 | | 200 | | | 100 | | 150 | | 200 | |
| *missing %* | TiMEG | CC | TiMEG | CC | TiMEG | CC | *missing %* | TiMEG | CC | TiMEG | CC | TiMEG | CC |
| (0,0,0) | 0.0534 | 0.0534 | 0.0528 | 0.0528 | 0.0533 | 0.0533 | (10,0,10) | 0.0860 | 0.0559 | 0.0847 | 0.0544 | 0.0708 | 0.0529 |
| (0,0,10) | 0.0836 | 0.0527 | 0.0805 | 0.0549 | 0.0757 | 0.0505 | (0,10,10) | 0.0889 | 0.0536 | 0.0519 | 0.0540 | 0.0469 | 0.0519 |
| (0,0,20) | 0.0828 | 0.0590 | 0.0760 | 0.0531 | 0.0758 | 0.0519 | (10,10,0) | 0.0848 | 0.0557 | 0.0795 | 0.0533 | 0.0750 | 0.0520 |
| (0,0,40) | 0.0758 | 0.0586 | 0.0803 | 0.0527 | 0.0776 | 0.0565 | (20,0,20) | 0.0826 | 0.056 | 0.0810 | 0.0585 | 0.0793 | 0.0536 |
| (0,0,60) | 0.0826 | 0.0632 | 0.0789 | 0.0628 | 0.0741 | 0.0551 | (0,20,20) | 0.0798 | 0.0537 | 0.0812 | 0.0514 | 0.0476 | 0.0558 |
| (0,0,80) | 0.0869 | 0.0805 | 0.0848 | 0.0689 | 0.0761 | 0.0625 | (20,20,0) | 0.0871 | 0.0542 | 0.0814 | 0.0552 | 0.0834 | 0.0516 |
| (0,10,0) | 0.0552 | 0.0540 | 0.0508 | 0.0509 | 0.0497 | 0.0530 | (40,0,40) | 0.0884 | 0.0783 | 0.0773 | 0.0652 | 0.0794 | 0.0617 |
| (0,20,0) | 0.0538 | 0.0607 | 0.0497 | 0.0564 | 0.0520 | 0.0519 | (0,40,40) | 0.0546 | 0.0621 | 0.0507 | 0.0594 | 0.0455 | 0.0550 |
| (0,40,0) | 0.0552 | 0.0602 | 0.0533 | 0.0580 | 0.0516 | 0.0492 | (10,10,10) | 0.0551 | 0.0565 | 0.0551 | 0.0549 | 0.0566 | 0.0526 |
| (0,60,0) | 0.0558 | 0.0649 | 0.0537 | 0.0594 | 0.0507 | 0.0578 | (10,20,10) | 0.0792 | 0.0531 | 0.0527 | 0.0569 | 0.0530 | 0.0494 |
| (0,80,0) | 0.0641 | 0.0837 | 0.0570 | 0.0637 | 0.0507 | 0.0620 | (10,10,20) | 0.0556 | 0.0577 | 0.0512 | 0.0547 | 0.0502 | 0.0529 |
| (10,0,0) | 0.0869 | 0.0533 | 0.0755 | 0.0531 | 0.0747 | 0.0495 | (20,10,10) | 0.0536 | 0.0551 | 0.0574 | 0.0554 | 0.0495 | 0.0510 |
| (20,0,0) | 0.0866 | 0.0595 | 0.0796 | 0.0532 | 0.0760 | 0.0552 | (20,10,20) | 0.0556 | 0.0544 | 0.0511 | 0.0590 | 0.0573 | 0.0513 |
| (40,0,0) | 0.0813 | 0.0598 | 0.0783 | 0.0570 | 0.0796 | 0.0556 | (20,20,10) | 0.0523 | 0.0588 | 0.0497 | 0.0548 | 0.0501 | 0.0571 |
| (60,0,0) | 0.0836 | 0.0624 | 0.0871 | 0.0564 | 0.0766 | 0.0593 | (10,20,20) | 0.0523 | 0.0567 | 0.0515 | 0.0556 | 0.0467 | 0.0556 |
| (80,0,0) | 0.0982 | 0.0849 | 0.0854 | 0.0679 | 0.0836 | 0.0624 | (20,20,20) | 0.0538 | 0.0600 | 0.0510 | 0.0581 | 0.0518 | 0.0532 |

**Table 2.** Type I error rate under different combination of sample sizes and varying percentages of missing methylation and/or gene expression values based on 10,000 simulations. *missing*% $(m_1, m_2, m_3) \equiv (m_1\%$ both missing, $m_2\%$ only methylation missing, $m_3\%$ only gene expression missing); *SS*: sample size for case (or control).

For the other three schemes, we generate complete data as above and remove some omics information to introduce missingness. For the second scheme where only gene expression is missing, we remove varying percentages (10%, 20%, 40%, 60%, 80%) of gene expression values. Similarly, we remove varying percentages of methylation values for the third scheme. For both omics missing scheme, we remove gene expression and methylation values for different combinations of missing percentages (Tables 2 and 3).

To estimate the parameters in the model, we maximise the likelihood function using a numerical optimisation technique (see Methods). We construct a test statistic using the above estimates to test whether a trio is associated with the phenotype. Here, we use the likelihood ratio test for testing the null hypothesis ($H_0$) of no effect of genotype, gene expression, and methylation on affection status. The asymptotic distribution of this test statistic follows a $\chi^2$ distribution with 3 degrees of freedom under $H_0$. Figure 2 illustrates the QQ plot of sample

4

| missing % | SS 100 TiMEG | CC | 150 TiMEG | CC | 200 TiMEG | CC | missing % | SS 100 TiMEG | CC | 150 TiMEG | CC | 200 TiMEG | CC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (0,0,0) | 0.697 | 0.697 | 0.878 | 0.878 | 0.950 | 0.950 | (10,0,10) | 0.564 | 0.564 | 0.748 | 0.745 | 0.907 | 0.881 |
| (0,0,10) | 0.640 | 0.636 | 0.800 | 0.810 | 0.933 | 0.908 | (0,10,10) | 0.638 | 0.631 | 0.831 | 0.815 | 0.923 | 0.919 |
| (0,0,20) | 0.579 | 0.570 | 0.760 | 0.770 | 0.917 | 0.885 | (10,10,0) | 0.616 | 0.613 | 0.857 | 0.827 | 0.931 | 0.909 |
| (0,0,40) | 0.552 | 0.423 | 0.730 | 0.617 | 0.866 | 0.745 | (20,0,20) | 0.483 | 0.435 | 0.701 | 0.637 | 0.834 | 0.767 |
| (0,0,60) | 0.485 | 0.288 | 0.676 | 0.416 | 0.829 | 0.573 | (0,20,20) | 0.611 | 0.590 | 0.794 | 0.789 | 0.901 | 0.879 |
| (0,0,80) | 0.445 | 0.147 | 0.634 | 0.218 | 0.795 | 0.278 | (20,20,0) | 0.601 | 0.563 | 0.796 | 0.736 | 0.897 | 0.868 |
| (0,10,0) | 0.605 | 0.618 | 0.842 | 0.822 | 0.937 | 0.935 | (40,0,40) | 0.327 | 0.135 | 0.488 | 0.214 | 0.642 | 0.290 |
| (0,20,0) | 0.646 | 0.537 | 0.830 | 0.767 | 0.933 | 0.888 | (0,40,40) | 0.514 | 0.422 | 0.745 | 0.613 | 0.879 | 0.757 |
| (0,40,0) | 0.647 | 0.441 | 0.817 | 0.610 | 0.936 | 0.778 | (10,10,10) | 0.567 | 0.567 | 0.762 | 0.752 | 0.889 | 0.877 |
| (0,60,0) | 0.619 | 0.281 | 0.818 | 0.451 | 0.931 | 0.566 | (10,20,10) | 0.598 | 0.609 | 0.753 | 0.739 | 0.907 | 0.888 |
| (0,80,0) | 0.596 | 0.145 | 0.802 | 0.215 | 0.916 | 0.289 | (10,10,20) | 0.543 | 0.482 | 0.724 | 0.697 | 0.867 | 0.817 |
| (10,0,0) | 0.612 | 0.604 | 0.804 | 0.795 | 0.911 | 0.920 | (20,10,10) | 0.524 | 0.482 | 0.755 | 0.690 | 0.858 | 0.843 |
| (20,0,0) | 0.539 | 0.533 | 0.778 | 0.763 | 0.882 | 0.885 | (20,10,20) | 0.484 | 0.441 | 0.710 | 0.607 | 0.797 | 0.780 |
| (40,0,0) | 0.425 | 0.411 | 0.641 | 0.603 | 0.793 | 0.769 | (20,20,10) | 0.518 | 0.494 | 0.740 | 0.687 | 0.871 | 0.821 |
| (60,0,0) | 0.362 | 0.285 | 0.514 | 0.436 | 0.663 | 0.557 | (10,20,20) | 0.528 | 0.513 | 0.740 | 0.690 | 0.886 | 0.838 |
| (80,0,0) | 0.228 | 0.129 | 0.366 | 0.208 | 0.488 | 0.304 | (20,20,20) | 0.497 | 0.426 | 0.695 | 0.590 | 0.831 | 0.762 |

**Table 3.** Power under different combination of sample sizes and varying percentages of missing methylation and/or gene expression values based on 1000 simulations. $missing\%$ $(m_1, m_2, m_3) \equiv (m_1\%$ both missing, $m_2\%$ only methylation missing, $m_3\%$ only gene expression missing); $SS$: sample size for case (or control).

quantiles from the empirical distribution of the test statistic under $H_0$ to theoretical quantiles of $\chi_3^2$ distribution, for a complete data and another dataset with missing data.

Based on 10000 datasets, we find that type I error rate of our test is controlled nearly at 5% level of significance for each of the different sample sizes, missing data schemes, and percentages of missing omics data (Tables 2 and S1). Hence, our test statistic is conservative in controlling false positives and is useful for $p$-value computation in a real dataset. To examine the performance of the test, we calculate statistical power under different missing data schemes and for various percentages of missing omics data based on 1000 datasets (Tables 3 and S2). To find whether the power of the test increases with an increase in sample size, we calculate power based on 5% cut-off points from 10000 datasets generated under $H_0$, for different percentages of missing omics data corresponding to each missing data scheme. This would keep the type I error rate fixed exactly at 5% level to make a uniform power comparison. Table 3 demonstrates a substantial increase in power for every combination of missing data with an increase in sample size. Under each missing data scheme, when the percentage of missing data increases, the power decreases. When there is no missing data the power would be maximum. Thus, our test is consistent.

As mentioned earlier, we now observe from Table 3, that TiMEG is more affected (as evident from the drop in statistical power) by (1) a certain percentage of missing gene expression data than the same percentage of missing methylation data and (2) missing both omics information for a subset of individuals than missing gene expression or methylation data on any subset of equivalent size. For instance, let us consider a fixed sample size of 200 (say) and a fixed percentage of missing omics data 40% (say). From Table 3 we note that the power of missing only gene expression data (0.866) is less than that of missing only methylation data (0.936). Clearly, the power of missing both omics (0.793) is less than the minimum of the above two. Such a difference in power is biologically expected (which is reflected in the complete-case analysis as well) because alteration in gene expression is more informative than any other change in the DNA sequence. Thus, for a subset of individuals, no information on any of the two mentioned omics causes more loss of information compared to the presence of at least one of them. So, at the individual level, if possible, it is better to collect at least one observation from gene expression or methylation. Thus, provided there exists a choice, less percentage of missing gene expression is preferred than that of methylation because gene expression data is more informative.

When miscellaneous percentages of data are missing, we observe (from Table 3) a similar phenomenon as above. For the fixed overall percentage of missing omics data (40%) and the fixed sample size (200), we consider three combinations such as (1) 20% individuals have both omics missing, another 10% individuals have only gene expression missing, and another 10% individuals have only methylation missing, (2) 10% individuals have both omics missing, another 20% individuals have only gene expression missing, and another 10% individuals have only methylation missing, and (3) 10% individuals have both omics missing, another 10% individuals have only gene expression missing, and another 20% individuals have only methylation missing. The powers in the above three combinations are 0.858, 0.867, and 0.907. So we observe that even in a miscellaneous missing scenario with such marginal difference in missing percentage of the omics, TiMEG is able to differentiate between statistical powers in accordance with the biological expectation. These results indicate that our method is able to capture all available information corresponding to every individual under study and its performance is robust to the percentage and scheme of missing omics data.

In Table 3, we also include a comparison of the statistical power of TiMEG with a complete-case analysis. Note that, the performance of TiMEG is clearly better than the complete-case for moderate missing percentages. For lower percentages both of them give comparable powers. Besides, we compute powers using mean imputation (MI) but they are less than TiMEG and type I errors are too much inflated (Tables S1 and S2). For k-nearest neighbour (KNN) imputation, powers are not stable (Table S2). We observe that as the percentages of missing increases, the power decreases but after some point the power increases. The powers clearly fluctuate as the missing percentages were increased more than 40%. This could be typically due to the uncertainty introduced by using the imputation technique when the missing data is moderately large[34].

*Run time.* We compare the computation time of TiMEG with other methods in Table 4. All programs are run in a Mac (OS Big Sur, version 11.5.1) laptop with Apple M1 chip having 8 GB RAM. We find that the maximum expected time per run for TiMEG is less than all other methods. Moreover, unlike other methods analysis time for TiMEG is consistent.

*Performance evaluation.* To evaluate the predictive performance of our method, we assess the prediction accuracy of our estimation by tenfold cross-validation (CV). We also compare TiMEG with commonly used imputation based methods such as KNN[39], MI, and also actual dataset without missing omics data. For TiMEG we first generate a dataset that has a pre-assigned missing omics data structure and divide it into two parts, test set and training set. Observations with no missing data are then split into 10 blocks. One block is selected as a test set and all remaining individuals form a training set. Based on the first training set, we find an estimated $\beta$ coefficients (using Eq. 1 in Methods) and classify the individuals in the corresponding test set to cases and controls. We repeat this procedure for all 10 test sets and calculate average prediction accuracy, specificity, and sensitivity for the dataset. For each pre-assigned missing omics data structure, we generate 100 datasets and perform a tenfold CV on each of them. Using them, we compute the specificity and sensitivity of our method for different thresholds of classification. Next, on the same generated dataset, we apply mean and KNN imputation methods to determine specificity and sensitivity using a tenfold CV. We provide four receiver operating characteristic curve (ROC) graphs each depicting better performance of TiMEG in comparison to other imputation methods. Also, we compare them with ROC on full data (Fig. 3). Interestingly, we observe that our method classifies an individual into a case or control group as efficiently as having full data.

Moreover, we find that the mean prediction accuracy for classifying an individual to case or control group is more than 80% under all missing omics schemes and for different missing percentages (Supplementary Table S3). We observe that median prediction accuracy remains the same under all scenarios except when the percentage of missing data is very high (Fig. 4). Although the deviation of median prediction accuracy for an extremely high percentage ($\sim$ 80%) of missing values compared to that for no missing data is small, the higher dispersions indicate fluctuations of the prediction accuracies. This implies that the number of false-positive and false-negative classifications fluctuates for these extreme missing scenarios. We illustrated this increase in dispersion through the plot of false-positive rate (1 - Specificity) versus misclassification rate (($1 -$ prediction accuracy)/100) under different percentages of missing omics data of different missing data schemes (Figs. 5, S1, S2, S3). It is evident that for comparatively smaller percentages of missing data, the dispersions are much less. Only for extreme conditions of missing data, the dispersion is slightly higher. Thus, we find that our method provides a robust estimate of the parameters under different missing omics schemes with a reasonable missing percentage and has high predictive power.

**Application to a real dataset.** We applied our proposed method to a dataset on Tuberous Sclerosis Complex (TSC) patients and healthy controls (phs001357.v1.p1)[29] obtained from the database of Genotypes and Phenotypes (dbGaP). TSC is a rare genetic disorder that causes the growth of non-cancerous (benign) tumours in the brain and other vital organs like kidneys, heart, skin, etc., and in some cases leads to significant health problems[40].

After processing raw data (see Methods) from brain tissues, we obtained 8036 gene expression data on 27 cases and 7 controls, methylation data at 481470 CpG sites on 22 cases and 7 controls, and 1298477 whole-genome genotype data on 38 cases and 7 controls. We got data on all three omics for the control individuals. But only 12 case individuals had complete omics information. 9 other case individuals had all omics data except gene expression data, another group of 13 patients had no methylation data, and 4 patients had neither gene expression nor methylation data. Phenotype or disease status, covariates (such as age and gender), and genotype data were available for all cases and controls.

Since control samples are only a few, we considered those genes that have no missing gene expression value in controls, while in the case samples we allow up to 50% missing gene expression value. Next, we find the SNPs that are within 2000 bp upstream and downstream of each gene. We considered these SNPs as the cis-SNPs to the gene. Moreover, if any methylation site is associated with a gene, information such as the corresponding gene name and chromosome number are known from dbGaP. So, methylation sites in the vicinity of a gene are considered cis-CpG sites corresponding to the gene. After filtering the data, we find the number of unique genes containing at least one cis-SNP and one cis-CpG site reduces to 1691 and the total number of trios (comprised of one gene expression with one cis-genotype and one cis-CpG site corresponding to the gene) is 1184436. To identify the trios associated with the disease, we perform our test for all the mentioned trios, followed by Benjamini-Hochberg (BH) multiple corrections (across all tests).

*Interpretation of a TiMEG trio.* For each significant trio, one or more of its components are associated with the disease. However, we are more interested to observe whether TiMEG is able to identify loci with moderately low
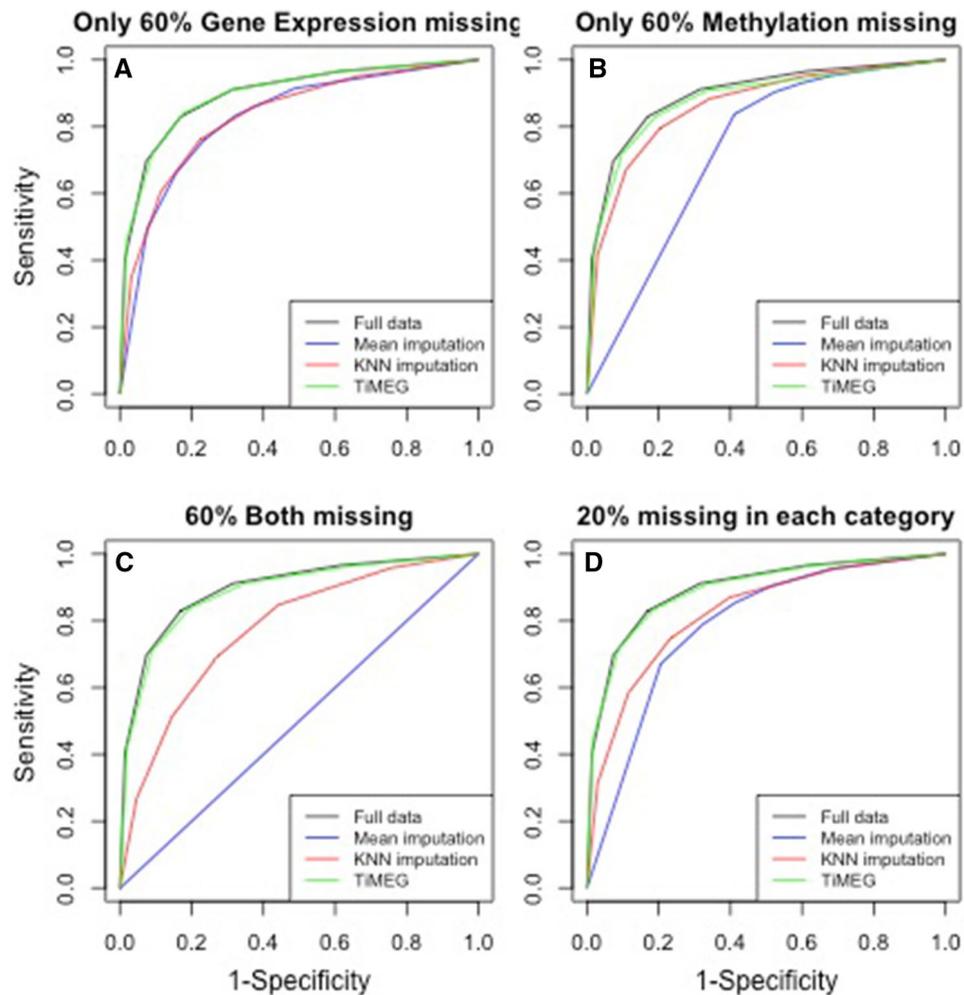
**Figure 3.** Plot of ROC graphs depicting situations with (**A**) only gene expression missing for 60% individuals, (**B**) only methylation missing for 60% individuals, (**C**) both omics missing for 60% individuals, and (**D**) both omics missing for 20%, only gene expression missing for 20% and only methylation missing for 20% individuals.

effect sizes that are missed by single-omic analysis. Therefore, we find those combinations where TiMEG shows association but separate analyses do not. We see that TiMEG successfully captures weaker signals that remain unidentified in single-omic analysis. The probable reason might be that single-locus from any omics data is unlikely to account for much of the variability in the phenotype. Moreover, it is often indicated that an increase in the sample size might capture the loci with moderate or low effect on the disease, but in that case, multiple testing burden also increases resulting in missing out true signals. But our method reduces false-negative associations by efficiently incorporating correlated omics information to decipher the significant signals associated with the disease. Particularly in this article, TiMEG tests if there is any effect of at least one of the components of the trio on the phenotype, but it is also able to test the effect of a single omic locus or combination of any two omics (see Methods). Emphatically, testing a single omic locus using TiMEG will provide greater insight than traditional single-omic analysis because of incorporating additional information from other available omics in the integrated model.

*Functional annotation of TSC genes.* It is well known that mutation in either of the two tumor-suppressor genes viz. *TSC1*[41] and *TSC2*[42], that code for hamartin and tuberin proteins respectively are responsible for TSC. The hamartin/tuberin heterodimer encoded by the interaction of *TSC1* and *TSC2* gene products, function in complex pathways[43]. *TSC*1/2 genes and hence the hamartin/tuberin complex plays a fundamental role in the regulation of phosphoinositide 3-kinase (PI3K) signaling pathway[44] that inhibits the mammalian target of rapamycin (mTOR) through activation of the GTPase activity of Rheb[45].

Using TiMEG we obtained 170 unique genes (see Supplementary Table S4) from 3283 significant trios (https://github.com/sarmistha123/TiMEG), to be associated with the disease risk. These trios are significant due to the combined effect of all its components but none of the single-omic analysis could identify any of the corresponding components. Among the contents of this list, there exits a trio corresponding to gene *TSC1* with cis-genotype
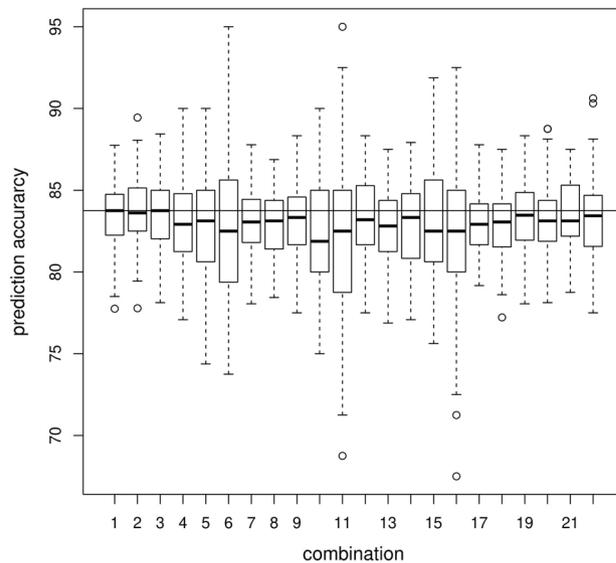
**Figure 4.** Boxplot of prediction accuracy from tenfold CV based on 100 datasets each with 200 cases and 200 controls under different missing omics data structure. The black horizontal line indicates median prediction accuracy for the datasets with no missing information. Each boxplot (from left to right) signifies one combination each viz. no missing information, only 10%, 20%, 40%, 60%, 80% gene expression missing respectively, only 10%, 20%, 40%, 60%, 80% methylation missing respectively, 10%, 20%, 40%, 60%, 80% of both gene expression and methylation missing respectively, 5% of both missing along with 5% of only gene expression missing, 5% of only gene expression missing along with 5% of only methylation missing, 5% of both missing along with 5% of only methylation missing, 10% of both missing along with 10% of only gene expression missing, 10% of both missing along with 10% of only methylation missing, 10% of both missing along with 5% of only gene expression missing and another 5% of only methylation missing.

*kgp*7096367 and cis-methylation site *cg*19350728 that shows no association in any of the single-omic analysis but their combined effect is significant. TSC2 gene is excluded from our analysis because of a mismatch between probe id and HGNC IDs (See Methods).

We use David software[46,47] to identify the functional annotations of the identified genes. Based on David's group enrichment score, we obtained 5 clusters of our genes. The cluster with the maximum group enrichment score is associated with the serine/threonine kinase pathway. This pathway has a strong functional relation with TSC disease, as it is known that mutations in *TSC1/2* genes impair the inhibitory function of the hamartin/tuberin complex, leading to phosphorylation (activation) of ribosomal protein S6 kinase beta-1 (S6K1), a serine/threonine kinase which is a downstream target of mTOR[45].

Another cluster is associated with the zinc-finger protein pathway. Recent findings have highlighted the importance of the zinc-finger family and its involvement in tumorigenesis[48]. Interestingly, this pathway has some special implications in terms of brain tissues. Protein associated with Myc (called Pam) that is abundantly expressed in the brain, is associated with the tuberin/hamartin complex[49]. The C terminus of Pam containing the RING zinc-finger motif binds to tuberin[49]. Besides, Pam is a highly conserved nuclear protein that interacts directly with the transcriptional-activating domain of Myc (a protooncogene that plays an important role in the regulation of cellular proliferation, differentiation, and apoptosis and can contribute to tumorigenesis)[50] and regulates mTOR signaling[51].

Studies have revealed that TSC receives inputs from at least three major signaling pathways (PI3K-Akt-mTOR, ERK1/2-RSK1, LKB1-AMPK) in the form of kinase-mediated phosphorylation events that regulate its function as a GTPase activating protein (GAP)[52]. But only two genes viz. *TSC*1/2 are widely known to be responsible for the disease. Therefore, we searched whether any of our significant genes belong to the same pathway as that of *TSC1* gene. Using David software we identified genes *ACACA* and *CREB5* in our list of significant genes, that occur in two pathways to which *TSC1* belongs.

Studies show that *TSC1* deficiency elevated *ACACA* expression and fatty acid synthesis, leading to impaired epigenetic imprinting on selective genes; tempering *ACACA* activity was able to divert cytosolic acetyl-CoA for histone acetylation and restore the gene expression program compromised by *TSC1* deficiency[53]. *CREB5* encodes CREB protein that serves as a transcriptional activator of Rheb. Rheb acts as an immediate activator of mTOR and in turn promotes tumorigenesis independently of *TSC2*[54]. Moreover, we identified other genes such as *JAK3*, *GNG4*, *FGFR2*, *EFNA2*, *LAMC2* that belong to one of the pathways as that of *TSC1*.

*Implication of TiMEG.* It is important to note that these genes could not have been identified by single-omic analysis. Data integration of different omics led to these findings even when the sample size is small and individual-level data is not available on all omics. Thus, TiMEG holds the potential to understand the genetic architec-
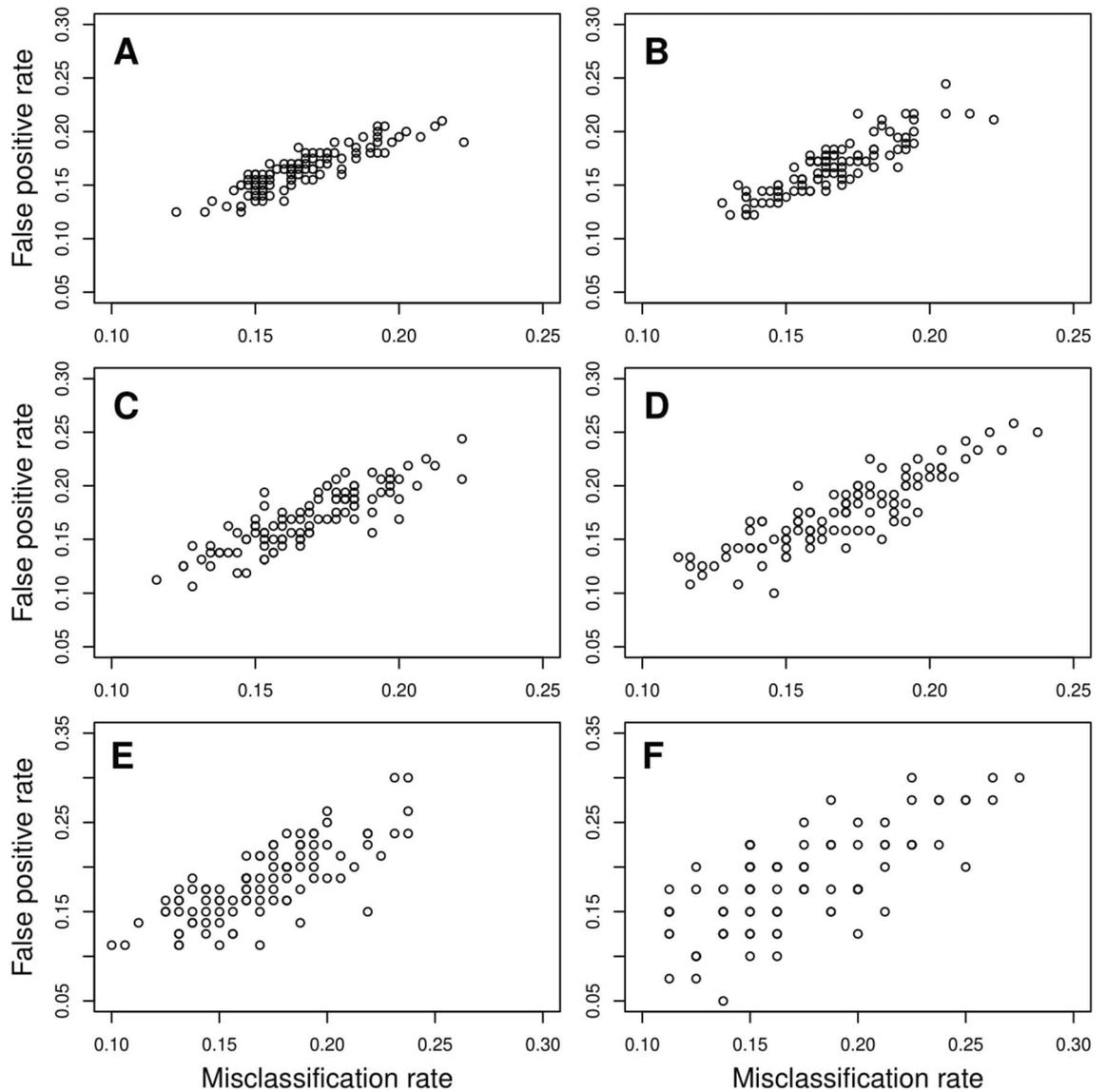
**Figure 5.** Plot of Misclassification rate vs False positive rate (1-Specificity) for only gene expression missing. (**A**) depicts no missing data scenario while (**B**–**F**) respectively depict 10%, 20%, 40%, 60% and 80% only gene expression data missing scenarios.

| | SS | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **100** | | | | **150** | | | | **200** | | | |
| *missing* % | **TiMEG** | **CC** | **KNN** | **MI** | **TiMEG** | **CC** | **KNN** | **MI** | **TiMEG** | **CC** | **KNN** | **MI** |
| (0,0,10) | 2.348 | 1.510 | 1.643 | 1.540 | 2.572 | 2.221 | 2.437 | 2.378 | 2.625 | 3.053 | 3.312 | 3.176 |
| (0,0,80) | 2.177 | 0.355 | 1.546 | 1.864 | 2.195 | 1.270 | 2.384 | 2.589 | 2.495 | 1.946 | 3.103 | 3.253 |
| (0,10,0) | 0.185 | 1.437 | 1.668 | 1.659 | 0.222 | 2.300 | 2.580 | 2.448 | 0.260 | 3.063 | 3.334 | 3.132 |
| (0,80,0) | 0.189 | 0.370 | 1.674 | 1.833 | 0.226 | 1.155 | 2.363 | 2.494 | 0.257 | 1.946 | 3.183 | 3.260 |
| (10,0,0) | 2.496 | 1.506 | 1.737 | 1.683 | 2.581 | 2.235 | 2.613 | 2.455 | 2.762 | 3.100 | 3.433 | 3.295 |
| (80,0,0) | 2.430 | 0.388 | 3.237 | 1.909 | 2.451 | 1.237 | 2.433 | 2.681 | 2.645 | 1.911 | 3.223 | 3.298 |

**Table 4.** Expected average computation time (in seconds) per run based on 100 simulations. *missing*% $(m_1, m_2, m_3) \equiv (m_1\%$ both missing, $m_2\%$ only methylation missing, $m_3\%$ only gene expression missing); *SS*: sample size for case (or control).

ture of a disease aetiology by combining three different omics data, even in presence of missing omics data, and when the sample size is not humongous.

Thus, if the sample size is not huge, studying single-omic data to find any new gene that is susceptible to the disease risk is difficult because most of the known diseases have already been studied extensively. Moreover, scientists nowadays are interested in developing drug targets with genetic evidence of disease association as they are much more likely to get approved[55]. So, identification of new disease-associated genes or biomarkers is immensely important to understand the relation of disease with various genes in the pathways. Downstream/detailed investigation of these biomarkers could provide a better understanding of the disease aetiology and hence discover the best drug targets that might lead to the successful development of novel drugs.

## Discussion

Multi-omics data integration elucidates the understanding of the genetic architecture of diseases and complex traits by incorporating additional information from different types of genomic data. But the presence of missing values poses a major challenge. This is more crucial when the sample size is limited and/or the percentage of missing data is large. Sometimes, these missing values occur due to biological reasons such as degradation of RNA or other technical issues. But when resources are limited, more often these assays are not repeated for the missing omics data. Typically, genotype data, being less expensive than gene expression and DNA methylation assays, are available for the entire sample. One option to analyse such data is using a sub-sample for which data are available for all omics. Such a complete case analysis loses a great deal of information. Again, imputation might induce bias arising due to the genetic diversity of reference data[15,24]. On the other hand, different types of omics data may be correlated and associated with a disease, directly or indirectly. Thus, integrating evidence from the inter-relationship among omics data provides additional information for biomarker identification.

We propose TiMEG, a tool for the identification of biomarkers integrating genotype, gene expression, and DNA methylation in presence of missing data under the case–control paradigm. Based on a likelihood approach, TiMEG is able to capture weaker signals that are often missed by single-omic analysis, by efficiently combining the information on interdependence among multiple omics data. Rather than imputing the missing data before the analysis, TiMEG accumulates information on the missing data by estimating the parameters in the likelihood function containing incomplete omics data. For calculating the likelihood function for incomplete data, we evaluate the conditional distribution of the response variable given the available information. This information not only includes the available omics data but also the inter-relationship among different omics. Moreover, our method has the ability to tackle an unrestricted number of missing transcriptomic and epigenetic data. Asymptotic distribution of our test statistic derived under the null hypothesis of no association will lead to the fast calculation of $p$-values compared to computation-intensive techniques. Moreover, the normal approximation of the sigmoid function[56] in the evaluation of the test statistic reduces the computation time to a great extent. Thus, TiMEG could be promptly applied by the end-users on real datasets.

Simulation results confirm consistency of the test, robust performance in terms of prediction accuracy of estimation in tenfold CV, controlled type I error rate, and high statistical power. Moreover, as the percentage of missing values increases, the power of the test decreases as expected. Our method also shows robust performance. Simulation study confirms that for moderately high percentages of missing data, the power and tenfold prediction accuracy of estimation are close to that of no missing data. Simulation results also indicate that reduction in power of the test is not substantial for extremely large missing percentages. Besides, the median prediction accuracy is nearly the same under all scenarios except when the percentage of missing data is very high (Fig. 4) but, the mean prediction accuracy of classifying an individual to case or control group is nearly the same (Table S3). Only for extreme percentages of missing omics data, fluctuations in misclassification rate are slightly higher (Figs. 5, S1, S2, S3). Thus, one of the major advantages of TiMEG lies in its applicability to moderately large missing percentages and limited sample size for the identification of biomarkers. Another advantage is that it identifies a combination of multiple omics loci (or trio) as biomarkers. So, even when anyone or all the components of a trio have some small effect (but significant when combined) on the disease, TiMEG detects it. This is because, it is able to integrate multiple omics loci with small effects together, such that their combined effect on the disease is moderately large.

More often, wet-lab researchers encounter missing data in multiple omics assays, when data are collected on both patients and matched controls. One example of such an experiment is by Martin et al.[29]. We applied our method to their real dataset related to TSC that we obtained from dbGaP. The dbGaP data had genotype for all individuals (both TSC patients and healthy controls) but gene expression and/or methylation data were missing for a number of TSC patients. Although mutations in TSC1 and TSC2 genes are widely known to be responsible for the occurrence of TSC disease, several studies illustrated evidence of factors other than mutations in these genes, to be involved in the aetiology of the disease. Our method could identify a few more TSC associated genes at a much smaller sample size combining different omics data. Some of the identified genes have been previously reported to actively participate in the TSC disease causation[54,57].

Although TiMEG tests a trio for possible association with a disease, it could be extended to test multiple SNPs and multiple methylation sites along with gene expression. But, this will increase the number of parameters in the model. One possibility is to replace multiple SNPs and methylation values with some combined value or score, for example, the median of the methylation values under study etc. We plan to extend TiMEG for the accommodation of multiple SNPs and CpG sites as future work. We have not considered any interaction effect among different omics on the phenotype in this model. So, another extension of this work would be considering the interaction effect. Moreover, extending TiMEG to accommodate mutations instead of SNPs is important for experiments related to cancer. Some experiments collect data on quantitative phenotypes. TiMEG could be applied in such cases by dichotomising the quantitative phenotype but it would lose information. Therefore, it

is important to develop a method for quantitative phenotypes, which might not be straightforward. However, another strength of TiMEG is that it can test the effect of a single omic locus or combination of any two omics immediately. Such tests of single omic locus would provide greater insight than traditional single-omic analysis due to the additional insights from other omics data.

Moreover, application of TiMEG to the available data from public repositories might enhance the understanding of the disease by identifying different biomarkers. A detailed functional analysis of the significant association signals might facilitate understanding of the intricate genetic architecture of disease and therefore, translate the potential stored in the genomic data to develop targeted therapies and aid in precision medicine research.

## Methods

**Model.** Figure 1 provides a general missing data structure of multiple omics in different studies. The effect of this structure on the identification of biomarkers is much more prominent in studies with a limited sample size compared to large consortium data. The objective of TiMEG is to identify disease-associated biomarkers by integrating individual-level information from genetic (SNP), transcriptomic, and epigenomic data along with their phenotype (disease status), and covariates in such scenarios. So, TiMEG explores the effect of multi-omics data on disease status and the inter-relation among multiple omics for biomarker identification. To illustrate the scenario, we consider $n$ individuals with a known binary qualitative phenotype. For individual $i\,(i=1,2,\ldots,n)$, let $Y_i$, $G_i$, $M_i$ and $E_i$ denote respectively phenotype, genotype, methylation, and gene expression, and $\boldsymbol{X}_i$ denote the vector of $J$ covariates like age, gender, and other environmental variables. We denote $Y_i=-1$, for controls and $Y_i=1$, for cases. Conventionally, $G_i$ takes value 0, 1, 2 depending on the number of minor alleles present. Let **M** and **E** denote the vectors of continuous values for $n$ individuals and $X$ be a matrix of order $n \times J$ that includes covariate values for all $n$ individuals. Based on the aforementioned omics data, we have proposed the following model for the $i^{th}\,(i=1,\ldots n)$ individual.

$$P(Y_i=1)=\sigma\left(y_i(\beta_0+\beta_e E_i+\beta_m M_i+\beta_g G_i+\boldsymbol{\beta}_x' \boldsymbol{X}_i)\right) \tag{1}$$

$$M_i=\alpha_0+\alpha_g G_i+\epsilon_{1i} \tag{2}$$

$$E_i=\gamma_0+\gamma_g G_i+\gamma_m M_i+\epsilon_{2i} \tag{3}$$

where $\sigma(x)=\frac{1}{1+e^{-x}}$. We assume that $G_i \sim Bin(2,p)$ where, $p$ is the probability of occurrence of a minor allele and $(\epsilon_{1i},\epsilon_{2i}) \sim N_2(0,0,\sigma_1^2,\sigma_2^2,\rho)$ where $\sigma_1^2$ and $\sigma_2^2$ denote the variances of $\epsilon_{1i}$ and $\epsilon_{2i}$ respectively, and $\rho=Cor(\epsilon_{1i},\epsilon_{2i})$. Here, we have considered a likelihood based approach for estimation of the parameters in Eqs. 1 − 3. Denote the set of all parameters as, $\theta=(\beta_0,\beta_e,\beta_m,\beta_g,\boldsymbol{\beta}_x,\alpha_0,\alpha_g,\gamma_0,\gamma_g,\gamma_m,\sigma_1^2,\sigma_2^2,p)'$. So, our joint likelihood function for the full data (i.e. when there is no missing observation), becomes:

$$
\begin{aligned}
f_\theta(\boldsymbol{y},\boldsymbol{E},\boldsymbol{M},\boldsymbol{G}|X) &= \prod_{i=1}^{n} f_\theta(y_i,E_i,M_i,G_i|\boldsymbol{X}_i) \\
&= \prod_{i=1}^{n} \left\{ f_\theta(y_i|E_i,M_i,G_i,\boldsymbol{X}_i)f_\theta(E_i|M_i,G_i,\boldsymbol{X}_i)f_\theta(M_i|G_i,\boldsymbol{X}_i)f_\theta(G_i|\boldsymbol{X}_i) \right\} \\
&= \prod_{i=1}^{n} \left\{ \sigma(y_i(\beta_0+\beta_e E_i+\beta_m M_i+\beta_g G_i+\boldsymbol{\beta}_x'\boldsymbol{X}_i)) \binom{2}{G_i} p^{G_i}(1-p)^{2-G_i} \right. \\
&\quad \left. \times \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2\sigma_2^2}(E_i-\gamma_0-\gamma_g G_i-\gamma_m M_i)^2} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(M_i-\alpha_0-\alpha_g G_i)^2} \right\}
\end{aligned} \tag{4}
$$

As discussed earlier, genetic variants are available for a large population but transcriptomic and epigenomic data tend to be missing due to various reasons. So, we assume that genotype, phenotype, and covariate data are available for the whole population while varying percentages of either gene expression or methylation or both the omics are missing (Table 1). In the following section, we have introduced different schemes of missing values across multiple platforms.

**Missing values scheme.** We suppose that among $n$ individuals, $n_1$ individuals have complete data on all omics, phenotype, and covariates, for $n_2$ individuals only gene expression is missing, $n_3$ has only methylation values missing, and $n_4$ individuals neither have data on gene expression nor on methylation. Thus, depending on the missing data type(s), we can consider three schemes of missingness (Table 1). In each case, we have written the appropriate likelihood function. For that, we need to consider the following lemma (for proof see Appendix A, Supplementary Material).

**Lemma 1** *If $\phi(x)$ is the p.d.f. of a standard normal distribution, i.e. $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $-\infty<x<\infty$, then*

$$\int_{-\infty}^{\infty} \phi\left(\frac{\alpha x-\beta}{\sigma_1}\right).\phi\left(\frac{\gamma-\delta x}{\sigma_2}\right)dx=\phi\left(\frac{\alpha\gamma-\beta\delta}{\sqrt{\alpha^2\sigma_2^2+\delta^2\sigma_1^2}}\right).\frac{1}{\sqrt{\frac{\alpha^2}{\sigma_1^2}+\frac{\delta^2}{\sigma_2^2}}} \tag{5}$$

*where $\alpha$, $\beta$, $\gamma$, $\delta$, $\sigma_1^2$ and $\sigma_2^2$ are constants.*

Moreover, while deriving the likelihood functions, we approximate the logistic function in Eq. (4) by cumulative distribution function of a normal variable[56]. This approximation relation is given by:

$$\sigma(\nu) = \int_{-\infty}^{\nu} \phi\left(\frac{z}{\beta}\right) dz \text{ where } \beta = \frac{\pi}{\sqrt{3}} \text{ and } \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{6}$$

*Scheme 1: Only partial gene expression data are missing.* Consider a situation where phenotype, genotype, covariates, and methylation data are available for all $n$ individuals but, gene expression data are missing only for $n_2$ individuals. This indicates that $n_3 = n_4 = 0$, and $n_1 = n - n_2$. Based on this missing observation scheme, we need to write the likelihood function using Lemma 1. But before that, we have introduced a few notations for the sake of lucidity.

$Z_i = (1, X_i, G_i, M_i, E_i)'$, $w = (\beta_0, \boldsymbol{\beta}_x', \beta_g, \beta_m, \beta_e)'$, $Z_{i,o} = (1, X_i, G_i, M_i)'$, $Z_{i,m} = E_i$, $w_o = (\beta_0, \boldsymbol{\beta}_x', \beta_g, \beta_m)'$. Now, to rewrite the likelihood function as in (4), we need a precise expression for $P(y_i|Z_{i,o})$, as given in the following result. Note that Result 1 (for Proof see Appendix B, Supplementary Material) is related to only $n_2$ individuals for whom gene expression data are not available.

**Result 1** *Using the model (1–3),*

$$P(y_i|Z_{i,o}) = \sigma\left(\frac{y_i\beta(\beta_0\boldsymbol{\beta}_x'X_i + \beta_g G_i + \beta_m M_i + \beta_e\mu_0)}{\sqrt{\beta^2 + y_i^2\beta_e^2 + \sigma_2^2}}\right) \tag{7}$$

*where $\mu_0 = \gamma_0 + \gamma_g G_i + \gamma_m M_i$, for each $i \in S_{-E}$, the set of $n_2$ individuals for whom gene expression data are not available.*

Now without any loss of generality, we have assumed that for the first $n_1$ individuals all data are available whereas the last $n_2$ individuals do not have gene expression data. Hence, using Result 1, the likelihood function (4) can be written under the scheme 1 as:

$$L(\theta|\boldsymbol{y}, \boldsymbol{E}, \boldsymbol{M}, \boldsymbol{G}, X)$$

$$= \prod_{i=1}^{n_1}\left\{\sigma(y_i(\beta_0 + \beta_e E_i + \beta_m M_i + \beta_g G_i + \boldsymbol{\beta}_x'X_i))\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2\sigma_2^2}(E_i - \gamma_0 - \gamma_g G_i - \gamma_m M_i)^2}\right\}$$

$$\times \prod_{i=1}^{n}\left\{\binom{2}{G_i}p^{G_i}(1-p)^{2-G_i}\right\}\prod_{i=1}^{n}\left\{\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2\sigma_1^2}(M_i - \alpha_0 - \alpha_g G_i)^2}\right\} \tag{8}$$

$$\times \prod_{i=n_1+1}^{n}\left\{\sigma\left(\frac{y_i\beta(\beta_0 + \beta_g G_i + \boldsymbol{\beta}_x'X_i + \beta_m M_i + \beta_e\mu_0)}{\sqrt{\beta^2 + \beta_e^2\sigma_2^2}}\right)\right\}$$

where $n_2 = n - n_1$.

*Scheme 2: Only partial methylation data are missing.* We may have a situation where all types of data are available for $n_1$ individuals and for another group of $n_3$ individuals all types of data except methylation data are available. So here we have $n_2 = n_4 = 0$ and $n_3 = n - n_1$. So, the terms involving $M_i$ are not available for $n_3$ individuals. Again as in the above scheme, we have now introduced a few notations as:

$Z_i = (1, X_i, G_i, M_i, E_i)'$, $w = (\beta_0, \boldsymbol{\beta}_x', \beta_g, \beta_m, \beta_e)'$, $Z_{i,o} = (1, X_i, G_i, E_i)'$, $Z_{i,m} = M_i$, $w_o = (\beta_0, \boldsymbol{\beta}_x', \beta_g, \beta_e)'$. To write down the likelihood function, we have first evaluated the expression for $P(y_i|Z_{i,o})$ as given in Result 2 using Lemma 1. Note that Result 2 (for proof, see Appendix B, Supplementary Material) is related to only $n_3$ individuals for whom methylation data are not available.

**Result 2** *Using the model (1-3), for each $i \in S_{-M}$,*

$$P(y_i|Z_{i,o}) = \sigma\left(\frac{y_i\beta(\beta_0 + \boldsymbol{\beta}_x'X_i + \beta_g G_i + \beta_e E_i + \beta_m\frac{(\alpha_0 + \alpha_g G_i)\sigma_2^2 + \gamma_m(E_i - \gamma_0 - \gamma_g G_i)\sigma_1^2}{\sigma_2^2 + \gamma_m^2\sigma_1^2})}{\sqrt{\beta^2 + \beta_m^2\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2 + \gamma_m^2\sigma_1^2}}}\right) \tag{9}$$

*where $S_{-M}$ is the set of $n_3$ individuals for whom no methylation data are available.*

Now without any loss of generality, we have assumed that for the first $n_1$ individuals all data are available whereas the last $n_3$ individuals do not have methylation data. Hence, using Result 2, the likelihood function (4) can be written under the scheme 2 as:

$$L(\theta|\boldsymbol{y}, \boldsymbol{E}, \boldsymbol{M}, \boldsymbol{G}, X) = \prod_{i=1}^{n_1} \sigma\left(y_i(\beta_0 + \beta_e E_i + \beta_m M_i + \beta_g G_i + \boldsymbol{\beta}'_x \boldsymbol{X}_i)\right)$$

$$\times \prod_{i=1}^{n_1} \left\{ \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2\sigma_2^2}(E_i - \gamma_0 - \gamma_g G_i - \gamma_m M_i)^2} \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(M_i - \alpha_0 - \alpha_g G_i)^2} \right\}$$

$$\times \prod_{i=n_1+1}^{n} \sigma\left( \frac{y_i \beta(\beta_0 + \beta_e E_i + \beta_g G_i + \boldsymbol{\beta}'_x \boldsymbol{X}_i + \beta_m \frac{(\alpha_0 + \alpha_g G_i)\sigma_2^2 + \gamma_m(E_i - \gamma_0 - \gamma_G G_i)\sigma_1^2}{\sigma_2^2 + \sigma_1^2 \gamma_m^2})}{\sqrt{\beta^2 + \beta_m^2 \frac{\sigma_1^2 \sigma_2^2}{\sigma_2^2 + \sigma_1^2 \gamma_m^2}}} \right)$$

$$\times \prod_{i=n_1+1}^{n} \frac{1}{\sqrt{2\pi(\sigma_2^2 + \sigma_1^2 \gamma_m^2)}} e^{-\frac{1}{2(\sigma_2^2 + \sigma_1^2 \gamma_m^2)}(E_i - \gamma_0 - \gamma_g G_i - \gamma_m(\alpha_0 + \alpha_g G_i))^2}$$

$$\times \prod_{i=1}^{n} \left\{ \binom{2}{G_i} p_i^G (1-p)^{2-G_i} \right\} \tag{10}$$

where, $n_3 = n - n_1$

*Scheme 3: Methylation and gene expression data are partially missing.* Lastly, under the most general missing value scheme, we have considered $n_2$ individuals have only missing gene expression values, $n_3$ individuals have only missing methylation values, $n_4$ individuals have both missing gene expression and methylation values, and the rest of the individuals have all types of data. Similarly, as for other schemes, we have now introduced a few notations as:

$\boldsymbol{Z}_i = (1, \boldsymbol{X}_i, G_i, M_i, E_i)'$, $\boldsymbol{w} = (\beta_0, \boldsymbol{\beta}'_x, \beta_g, \beta_m, \beta_e)'$, $\boldsymbol{Z}_{i,o} = (1, \boldsymbol{X}_i, G_i)'$, $\boldsymbol{Z}_{i,m} = (E_i, M_i)'$, $\boldsymbol{w}_o = (\beta_0, \boldsymbol{\beta}'_x, \beta_g)'$. Then, we have evaluated $P(y_i|\boldsymbol{Z}_{i,o})$ as given in Result 3 (for proof, see Appendix B, Supplementary Material) using Lemma 1 in order to write the joint likelihood equation.

**Result 3** *Under the model (1-3), for each $i \in S_{-(E,M)}$,*

$$P(y_i|\boldsymbol{Z}_{i,o}) = \sigma\left( \frac{\beta y_i(\beta_0 + \beta_g G_i + \boldsymbol{\beta}'_x \boldsymbol{X}_i + \beta_e(\gamma_0 + \gamma_g G_i) + (\alpha_0 + \alpha_g G_i)(\beta_e \gamma_m + \beta_m))}{\sqrt{\beta^2 + y_i^2 \beta_e^2 \sigma_2^2 + y_i^2 (\beta_e \gamma_m + \beta_m)^2 \sigma_1^2}} \right) \tag{11}$$

*where $S_{-(E,M)}$ is the set of $n_4$ individuals for whom both expression and methylation data are missing.*

In order to write down the likelihood function, we have assumed without any loss of generality, that first $n_1$ individuals have all data, next $n_2$ individuals have all data except gene expression data, next $n_3$ individuals have all data except methylation data and for the remaining $n_4$ individuals neither gene expression data nor methylation data are available but phenotype, covariates, and genotype data are available for all $n$ individuals. Clearly $n_4 = n - n_1 - n_2 - n_3$.

Using Results 1-3, the likelihood function (4) under scheme 3 can be written as:

$$L(\theta|\boldsymbol{y}, \boldsymbol{E}, \boldsymbol{M}, \boldsymbol{G}, X) = \prod_{i=1}^{n_1} \sigma\left(y_i(\beta_0 + \beta_e E_i + \beta_m M_i + \beta_g G_i + \boldsymbol{\beta}'_x \boldsymbol{X}_i)\right)$$

$$\times \prod_{i=1}^{n_1} \left\{ \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2\sigma_2^2}(E_i - \gamma_0 - \gamma_g G_i - \gamma_m M_i)^2} \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(M_i - \alpha_0 - \alpha_g G_i)^2} \right\}$$

$$\times \prod_{i=n_1+1}^{n_1+n_2} \left\{ \sigma\left( \frac{y_i \beta(\beta_0 + \beta_g G_i + \boldsymbol{\beta}'_x \boldsymbol{X}_i + \beta_m M_i + \beta_e \mu_0)}{\sqrt{\beta^2 + \beta_e^2 \sigma_2^2}} \right) \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(M_i - \alpha_0 - \alpha_g G_i)^2} \right\}$$

$$\times \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left\{ \sigma\left( \frac{y_i \beta(\beta_0 + \beta_e E_i + \beta_g G_i + \boldsymbol{\beta}'_x \boldsymbol{X}_i + \beta_m \frac{(\alpha_0 + \alpha_g G_i)\sigma_2^2 + \gamma_m(E_i - \gamma_0 - \gamma_G G_i)\sigma_1^2}{\sigma_2^2 + \sigma_1^2 \gamma_m^2})}{\sqrt{\beta^2 + \beta_m^2 \frac{\sigma_1^2 \sigma_2^2}{\sigma_2^2 + \sigma_1^2 \gamma_m^2}}} \right) \right.$$

$$\left. \times \frac{1}{\sqrt{2\pi(\sigma_2^2 + \sigma_1^2 \gamma_m^2)}} e^{-\frac{1}{2(\sigma_2^2 + \sigma_1^2 \gamma_m^2)}(E_i - \gamma_0 - \gamma_g G_i - \gamma_m(\alpha_0 + \alpha_g G_i))^2} \right\}$$

$$\times \prod_{i=n_1+n_2+n_3+1}^{n} \sigma\left( \frac{\beta y_i(\beta_0 + \beta_g G_i + \boldsymbol{\beta}'_x \boldsymbol{X}_i + \beta_e(\gamma_0 + \gamma_g G_i) + (\alpha_0 + \alpha_g G_i)(\beta_e \gamma_m + \beta_m))}{\sqrt{\beta^2 + y_i^2 \beta_e^2 \sigma_2^2 + y_i^2 (\beta_e \gamma_m + \beta_m)^2 \sigma_1^2}} \right)$$

$$\times \prod_{i=1}^{n} \left\{ \binom{2}{G_i} p_i^G (1-p)^{2-G_i} \right\} \tag{12}$$

where $\mu_o = \gamma_0 + \gamma_g G_i + \gamma_m M_i + \beta'_x X_i$, $n_4 = n - n_1 - n_2 - n_3$.

Next, for estimating the parameters in each of the likelihood functions, we used the L-BFGS-B method (in R package 'stats'), an iterative algorithm for numerical optimisation to find the maximum likelihood estimates of the parameters. Thus, theoretically, our method is able to incorporate any amount of missing gene expression and methylation data. In this paper, we focused on the identification of disease-associated trios (that is, a combination of the gene along with its cis-genotype and cis-methylation site). Thus, the components of a significant trio are expected to have a joint effect on affection status. Traditional single-omic analysis of each component is likely to miss these loci unless the sample size is humongous and/or the technologies are tremendously improved. More information from multiple omics on each individual, coupled with additional insights from the inter-relationship among the omics, supported the identification of significant loci even at smaller sample sizes compared to large single-omic analysis.

**Hypothesis of interest.** With the objective to identify a trio that may be associated with the disease or phenotype, we formulated the hypotheses of interest as:

$$H_0 : \beta_e = \beta_m = \beta_g = 0 \text{ against } H_1 : \text{not } H_0$$

Rejection of $H_0$ would indicate association with one or more components of the trio with the disease. To test the null hypothesis we adopted likelihood ratio test under a very general likelihood structure under various schemes of missing data. The test statistic for testing $H_0$ would be,

$$\Lambda = -2 \ln \frac{\sup_{\theta \in \Theta_0} L(\theta|data)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta|data)} \tag{13}$$

where $\theta$ is the vector of parameters in the likelihood function $L$, $\Theta_0$ and $\Theta_1$ are the parametric spaces under $H_0$ and $H_1$ respectively. Using standard asymptotic theory, it can be easily shown that the test statistic $\Lambda$ follows $\chi^2$ distribution with 3 degrees of freedom asymptotically under $H_0$. Usually, the sample sizes are considerably large so that we can use the asymptotic distribution of $\Lambda$ under $H_0$ for a real dataset. This reduces a huge computational burden while calculating the $p$-value in order to come to a conclusion.

If interested, one may test the effect of any two components (duos) such as genotype and gene expression, by testing

$$H_0 : \beta_e = \beta_g = 0 \text{ against } H_1 : \text{not } H_0$$

to find whether any combination of a gene and a genotype is associated with the disease. Other alternative hypotheses may be framed as per the objective. But here we considered only the identification of significant trios.

**dbGaP data on TSC.** All the real data on TSC patients and healthy controls have been published previously by Martin et al.[29] and deposited on dbGaP. We obtained publicly available real data from dbGaP (phs001357. v1.p1). All the data are available through a request for external collaboration and upon approval of a letter of intent and a research proposal. Details of how to request controlled-access data for external collaboration is available on the dbGaP website https://urldefense.proofpoint.com/v2/url?u=https-3A__dbgap.ncbi.nlm.nih. gov_aa_wga.cgi-3Fpage-3Dlogin&d=DwIDaQ&c=vh6FgFnduejNhPPD0fl_yRaSfZy8CWbWnIf4XJhSqx8&r= QqDpGi6FCxUEcyvzrkCUIg&m=e3Ku3-dtS10VfFaPnA85WzqdSg7HsMqlS3UVFJ39LyU&s=aDZbfDRAg5 aqiY4ZWXTC5TvbK4WN34r8R8rCALD6KpM&e=. Required ethical consent was obtained from the patients and/or their legal guardians before the data collection by the appropriate authorities. For this work, we analysed raw BAM files for gene expression data, IDAT files for genotype, and methylation data from brain tissues only. Genotypes were generated using Illumina Infinium Omni2.5 SNP arrays, methylation using Illumina Infinium HumanMethylation450 (HM450) BeadArrays, and gene expression using mRNA sequencing (RNAseq) for patient and control samples. For cases and controls, we derived log-normalised count for gene expression data, normalised-beta count for methylation data, and genotype data using Bioconductor package 'DESeq2', 'methylumi', and 'CRLMM' respectively in R software. For each probe ID, we have found its transcription start and end sites according to the human genome assembly 19 (hg19) from the UCSC genome browser using Bioconductor package TxDb.Hsapiens.UCSC.hg19.knownGene[58]. To have the same gene nomenclature across all omics platforms we converted probe IDs to gene names (or HGNC IDs) (using http://hgdownload.cse.ucsc.edu/ goldenPath/hg19/database/refFlat.txt.gz) and soutannotate annotated the SNPs using Bionconductor package 'humanomni258v1aCrlmm in R'.

## Data availability
The codes for TiMEG are available at https://github.com/sarmistha123/TiMEG with detailed directions of its use on the given toy dataset.

## References
1. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genom. Hum. Genet.* **9**, 387–402 (2008).
2. Green, E. D. & Guyer, M. S. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).

3. Blumenthal, G. M., Mansfield, E. & Pazdur, R. Next-generation sequencing in oncology in the era of precision medicine. *JAMA Oncol.* **2**, 13–14 (2016).
4. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).
5. Manolio, T. A. *et al.* Bedside back to bench: Building bridges between basic and clinical genomic research. *Cell* **169**, 6–12 (2017).
6. Consortium, W. T. C. C. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447**, 661 (2007).
7. MacArthur, J. *et al.* The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucl. Acids Res.* **45**, D896–D901 (2017).
8. Schork, N. J. Genetics of complex disease: Approaches, problems, and solutions. *Am. J. Respir. Crit. Care Med.* **156**, S103–S109 (1997).
9. Jansen, R. C. & Nap, J.-P. Genetical genomics: The added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
10. Editorial, N. G. Integrating with integrity. *Nat. Genet.* **42**, 1–1. https://doi.org/10.1038/ng0110-1 (2010).
11. Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. & Tarczy-Hornoch, P. Data integration and genomic medicine. *J. Biomed. Inform.* **40**, 5–16 (2007).
12. Das, S., Majumder, P. P., Chatterjee, R., Chatterjee, A. & Mukhopadhyay, I. A powerful method to integrate genotype and gene expression data for dissecting the genetic architecture of a disease. *Genomics* **111**, 1387–1394 (2019).
13. Balliu, B., Tsonaka, R., Boehringer, S. & Houwing-Duistermaat, J. A retrospective likelihood approach for efficient integration of multiple omics factors in case-control association studies. *Genet. Epidemiol.* **39**, 156–165 (2015).
14. Pineda, S. *et al.* Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. *PLoS Genet.* **11**, e1005689 (2015).
15. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genet.* **47**, 1091–1098 (2015).
16. Richardson, S., Tseng, G. C. & Sun, W. Statistical methods in integrative genomics. *Annu. Rev. Stat. Appl.* **3**, 181–209 (2016).
17. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* **14**, 1177932219899051 (2020).
18. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucl. Acids Res.* **46**, 10546–10562 (2018).
19. Wu, C. *et al.* A selective review of multi-level omics data integration using variable selection. *High-Throughput* **8**, 4 (2019).
20. Yan, K. K., Zhao, H. & Pang, H. A comparison of graph-and kernel-based-omics data integration algorithms for classifying complex traits. *BMC Bioinform.* **18**, 539 (2017).
21. Meng, C. *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).
22. Zarayeneh, N. *et al.* Integration of multi-omics data for integrative gene regulatory network inference. *Int. J. Data Min. Bioinform.* **18**, 223–239 (2017).
23. Boulesteix, A.-L., De Bin, R., Jiang, X. & Fuchs, M. Ipf-lasso: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Comput. Math. Methods Med.* **2017**, 1–33 (2017).
24. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245 (2016).
25. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
26. Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Sci. Data* **6**, 1–4 (2019).
27. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
28. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
29. Martin, K. R. *et al.* The genomic landscape of tuberous sclerosis complex. *Nat. Commun.* **8**, 1–13 (2017).
30. Romero, I. G., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biol.* **12**, 1–13 (2014).
31. Fang, Z. *et al.* Bayesian integrative model for multi-omics data with missingness. *Bioinformatics* **1**, 8 (2018).
32. Lin, D. *et al.* An integrative imputation method based on multi-omics datasets. *BMC Bioinform.* **17**, 1–12 (2016).
33. Henry, A. J., Hevelone, N. D., Lipsitz, S. & Nguyen, L. L. Comparative methods for handling missing data in large databases. *J. Vasc. Surg.* **58**, 1353–1359 (2013).
34. Kaambwa, B., Bryan, S. & Billingham, L. Do the methods used to analyse missing data really matter? An examination of data from an observational study of intermediate care patients. *BMC. Res. Notes* **5**, 330 (2012).
35. Xu, H., Gao, L., Huang, M. & Duan, R. A network embedding based method for partial multi-omics integration in cancer subtyping. *Methods* **192**, 67–76 (2020).
36. Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data* Vol. 793 (Wiley, 2019).
37. Ibrahim, J. G., Chen, M.-H. & Lipsitz, S. R. Bayesian methods for generalized linear models with covariates missing at random. *Can. J. Stat.* **30**, 55–78 (2002).
38. Mukhopadhyay, I., Feingold, E., Weeks, D. E. & Thalamuthu, A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet. Epidemiol.: Off. Publ. Int. Genet. Epidemiol. Soc.* **34**, 213–221 (2010).
39. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525. https://doi.org/10.1093/bioinformatics/17.6.520 (2001).
40. Crino, P. B., Nathanson, K. L. & Henske, E. P. The tuberous sclerosis complex. *N. Engl. J. Med.* **355**, 1345–1356. https://doi.org/10.1056/NEJMra055323 (2006) (**PMID: 17005952**).
41. van Slegtenhorst, M. *et al.* Identification of the tuberous sclerosis gene tsc1 on chromosome 9q34. *Science* **277**, 805–808 (1997).
42. Consortium, E. C. T. S. *et al.* Identification and characterization of the tuberous sclerosis gene on chromosome 16. *Cell* **75**, 1305–1315 (1993).
43. Van Slegtenhorst, M. *et al.* Interaction between hamartin and tuberin, the tsc1 and tsc2 gene products. *Hum. Mol. Genet.* **7**, 1053–1057 (1998).
44. Tee, A. R., Anjum, R. & Blenis, J. Inactivation of the tuberous sclerosis complex-1 and -2 gene products occurs by phosphoinositide 3-kinase/akt-dependent and -independent phosphorylation of tuberin. *J. Biol. Chem.* **278**, 37288–37296 (2003).
45. Tee, A. R., Manning, B. D., Roux, P. P., Cantley, L. C. & Blenis, J. Tuberous sclerosis complex gene products, tuberin and hamartin, control mTOR signaling by acting as a GTPase-activating protein complex toward Rheb. *Curr. Biol.* **13**, 1259–1268 (2003).
46. Sherman, B. T. *et al.* Systematic and integrative analysis of large gene lists using David bioinformatics resources. *Nat. Protoc.* **4**, 44 (2009).
47. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucl. Acids Res.* **37**, 1–13 (2009).
48. Cassandri, M. *et al.* Zinc-finger proteins in health and disease. *Cell Death Discov* **3**, 1–12 (2017).
49. Murthy, V. *et al.* Pam and its ortholog highwire interact with and may negatively regulate the tsc1.tsc2 complex. *J. Biol. Chem.* **279**, 1351–1358 (2004).
50. Guo, Q., Xie, J., Dang, C. V., Liu, E. T. & Bishop, J. M. Identification of a large Myc-binding protein that contains RCC1-like repeats. *Proc. Natl. Acad. Sci.* **95**, 9172–9177 (1998).

51. Han, S. *et al.* Pam (protein associated with Myc) functions as an E3 ubiquitin ligase and regulates TSC/mTOR signaling. *Cell. Signal.* **20**, 1084–1091 (2008).
52. Kwiatkowski, D. J. & Manning, B. D. Tuberous sclerosis: A gap at the crossroads of multiple signaling pathways. *Hum. Mol. Genet.* **14**, R251–R258 (2005).
53. Shi, L. *et al.* TSC1/mTOR-controlled metabolic-epigenetic cross talk underpins dc control of cd8+ t-cell homeostasis. *PLoS Biol.* **17**, e3000420 (2019).
54. Liu, Z. *et al.* Pdk4 promotes tumorigenesis through activation of CREB-RHEB-mTORC1 signaling cascade. *J. Biol. Chem.* **289**, 29739–29749 (2014).
55. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
56. Williams, D., Liao, X., Xue, Y. & Carin, L. Incomplete-data classification using logistic regression. In *Proceedings of the 22nd International Conference on Machine Learning* 972–979 (ACM, 2005).
57. Shah, O. J. & Hunter, T. Turnover of the active fraction of irs1 involves raptor-mTOR-and s6k1-dependent serine phosphorylation in cell culture models of tuberous sclerosis. *Mol. Cell. Biol.* **26**, 6425–6434 (2006).
58. Carlson, M. & Maintainer, B. Txdb. hsapiens. ucsc. hg19. knowngene: Annotation package for txdb object (s). *R package version* **3** (2015).

## Acknowledgements

## Author contributions

S.D. and I.M. conceived of the presented idea, carried out the mathematical derivations, interpreted the results, and wrote the manuscript. S.D. carried out the simulations and real data analysis. I.M. supervised the findings of this work. Both the authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03034-z.

**Correspondence** and requests for materials should be addressed to I.M.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.