Supplementary Data

1 Supplementary Data 1: Trio K-means for Disomic Data

This section contains the details of the disomic trio K-means algorithm. The trio K-means method uses the following iterative procedure.

Step 1: Start with a set of initial centroids.

The initial centroids are $\{C^{(0)}_{AA}, C^{(0)}_{AB}, C^{(0)}_{BB}\}$.

Step 2: At the k+1 step, update all three observations in a family as described in the following.

Assume we have two alleles, A and B. For disomic family trios, there are 15 possible genotype combinations that agree with Mendelian segregation rules (see Table 1 of the original paper). Let g_1, g_2 and g_3 be the possible genotypes for parent 1, parent 2 and the child. For all combinations of g_1, g_2 and g_3 shown in Table 1, we calculate $D_{g_1,g_2,g_3} = d(x_1, C^{(k)}_{g_1}) + d(x_2, C^{(k)}_{g_2}) + d(x_3, C^{(k)}_{g_3})$, where C^k_g's are the estimated group centers from the k^{th} step and $d(x_i, C^k_j)$ is the squared Euclidean distance between the observed value x_i and the center for j^{th} genotype group, $C^{(k)}_j$. Family members are then assigned to the genotypes $\tilde{g_1}$, $\tilde{g_2}$ and $\tilde{g_3}$ that minimize D_{g_1,g_2,g_3}.

Step 3: Iterate until convergence.

Note that our trio K-means procedure assumes that all the family information is correct, and no Mendelian errors are acceptable. We discuss this assumption further in the discussion section of the paper.

2 Supplementary Data 2: A Model for Trisomic Trios

Xu and others proposed a basic model for genotype data of trisomic trios (Xu et al., 2004). We base our family-based methods for trisomic trio data on this model. The following is a brief description of the model.

Assume a SNP marker with two alleles marked A and B. There are nine possible mating types (*i.e.*, different combinations of parental genotypes) as shown in Table S1. The nondisjoining parent (NDJP) is the parent that contributes two copies of the chromosomes and the correctly dsijoining parent (CDJP) is the parent that contributes only one copy of the chromosome. Because only a small portion of the trisomic conceptuses survive to term, we can only observe the disease status of these trisomic individuals. Hence it is impossible to separate the two events, survival to term and affected with the disease. Therefore, the association parameters in the model are defined as the following,

$$\begin{aligned}
w_0 &= \text{probability of survival and affectedness of a conceptus with genotype } AAA,\\
w_1 &= \text{probability of survival and affectedness of a conceptus with genotype } AAB,\\
w_2 &= \text{probability of survival and affectedness of a conceptus with genotype } ABB,\\
w_3 &= \text{probability of survival and affectedness of a conceptus with genotype } BBB.
\end{aligned}$$

The map parameter used in this model is h, which is defined as the probability that the two chromosomes contributed by the NDJP are reduced to homozygosity (duplicates of the same parental chromosome). Given the parental data, the probability of a living affected trisomic offspring's genotype depends only on the h and the w's. For example, for mating type AB (NDJP) × AA (CDJP), the CDJP must contribute an A. If h=0, *i.e.*, the two chromosomes are not reduced to homozygosity, then the NDJP contributes AB; if $h = 1$, *i.e.*, the two chromosomes are reduced to homozygosity, then NDJP contributes either AA or BB, with half of a chance each. Therefore, given the parental genotypes shown above,

$$\begin{aligned}
&Pr(\text{the affected child is } AAA \mid \text{NDJP is } AB \text{ and CDJP is } AA)\\
&= Pr(\text{NDJP contributes } AA) \times Pr(\text{the child survives to term and is diseased}|\text{the child is } AAA)\\
&= \frac{h}{2} \times w_0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&Pr(\text{the diseased child is } AAB \mid \text{NDJP is } AB \text{ and CDJP is } AA)\\
&= Pr(\text{NDJP contributes } AB) \times Pr(\text{the child survives to term and is diseased} \mid \text{the child is } AAB)\\
&= \frac{1-h}{2} \times w_1,
\end{aligned}$$

and

$$\begin{aligned}
&Pr(\text{the diseased child is } ABB \mid \text{NDJP is } AB \text{ and CDJP is } AA)\\
&= Pr(\text{NDJP contributes } BB) \times Pr(\text{the child survives to term and is diseased} \mid \text{the child is } ABB)\\
&= \frac{h}{2} \times w_2.
\end{aligned}$$

These probabilities are normalized so that they add up to 1 for each mating type, and listed in the 5^{th} column of Table S1. This example is the 5^{th} mating type shown in Table S1.

3 Supplementary Data 3: Trisomic Trio K-means Method

The trisomic trio K-means algorithm follows the same iteration steps as the disomic trio K-means algorithm. For trisomic trio data, the parents are disomic and the offspring are trisomic. We follow the trisomic model described by Xu *et al*, 2004. There are a total of 18 possible family types. They are listed in Table S1. Details of the model are described in supplementary data 2. NDJP is the abbreviation for the nondisjoining parent that passes 2 alleles to the offspring and CDJP is the abbreviation for the correctly disjoining parent that passes 1 allele to the offspring. Again, at each iteration, we calculate $D_{g_1,g_2,g_3} = d(x_1, C_{g_1}^{(k)}) + d(x_2, C_{g_2}^{(k)}) + d(x_3, C_{g_3}^{(k)})$ as defined in supplementary data 1. Family members are then assigned to the genotypes $\tilde{g}_1$, $\tilde{g}_2$ and $\tilde{g}_3$ that minimize D_{g_1,g_2,g_3}.

Table S1: Eighteen Family Types of a SNP Marker for a Nuclear Family with One Trisomic Offspring.

Family Type	NDJP	CDJP	Child	Probability	
				p	q
1	AA	AA	AAA	1	p_{aa}^2
2	AA	AB	AAA	$(\frac{w_0}{w_0+w_1})$	$p_{aa}p_{ab}$
3			AAB	$(\frac{w_1}{w_0+w_1})$	$p_{aa}p_{ab}$
4	AA	BB	AAB	1	$p_{aa}p_{bb}$
5	AB	AA	AAA	$(\frac{w_0h}{w_0h+2w_1(1-h)+w_2h})$	$p_{ab}p_{aa}$
6			AAB	$(\frac{2w_1(1-h)}{w_0h+2w_1(1-h)+w_2h})$	$p_{ab}p_{aa}$
7			ABB	$(\frac{w_2h}{w_0h+2w_1(1-h)+w_2h})$	$p_{ab}p_{aa}$
8	AB	AB	AAA	$(\frac{w_0h}{w_0h+(w_1+w_2)(2-h)+w_3h})$	$p_{ab}p_{ab}$
9			AAB	$(\frac{w_1(2-h)}{w_0h+(w_1+w_2)(2-h)+w_3h})$	$p_{ab}p_{ab}$
10			ABB	$(\frac{w_2(2-h)}{w_0h+(w_1+w_2)(2-h)+w_3h})$	$p_{ab}p_{ab}$
11			BBB	$(\frac{w_3h}{w_0h+(w_1+w_2)(2-h)+w_3h})$	$p_{ab}p_{ab}$
12	AB	BB	AAB	$(\frac{w_1h}{w_1h+2w_2(1-h)+w_3h})$	$p_{ab}p_{bb}$
13			ABB	$(\frac{2w_2(1-h)}{w_1h+2w_2(1-h)+w_3h})$	$p_{ab}p_{bb}$
14			BBB	$(\frac{w_3h}{w_1h+2w_2(1-h)+w_3h})$	$p_{ab}p_{bb}$
15	BB	AA	ABB	1	$p_{bb}p_{aa}$
16	BB	AB	ABB	$(\frac{w_2}{w_2+w_3})$	$p_{bb}p_{ab}$
17			BBB	$(\frac{w_3}{w_2+w_3})$	$p_{bb}p_{ab}$
18	BB	BB	BBB	1	p_{bb}^2

p_{aa}, p_{ab} and p_{bb} are the population genotype frequencies for genotypes aa, ab and bb respectively. h is the trisomic map parameter.

4 Supplementary Data 4: Trisomic Trio Beta-Mixture Model

This section contains the details of the trisomic trio beta-mixture model for genotype calling, including the derivation of the likelihood, the expectation-maximization (EM) algorithm for the estimation of the parameters and the genotype prediction procedure.

4.1 Likelihood for complete data

Let $\mathbf{Y_i} = (y_{Ni}, y_{Ci}, y_{Ki})$ denote the observed one-dimensional data for the NDJP, CDJP and child of the i^{th} trio; $\mathbf{G_i} = (g_{Ni}, g_{Ci}, g_{Ki})$ the corresponding genotype vector, where $\mathbf{G_i}$ is unknown. The contribution to the complete-data likelihood function from the i^{th} trio is:

$$\begin{aligned} & L_i(\theta, Y_i, G_i, h_i) \\ = \; & \{Pr(g_{Ni})Pr(g_{Ci})Pr(g_{Ki}|g_{Ni}, g_{Ci})\} \\ & \times \{Pr(y_{Ni}|g_{Ni})Pr(y_{Ci}|g_{Ci})Pr(y_{Ki}|g_{Ki})\}, \end{aligned} \tag{4.1}$$

where the first component, $Pr(g_{Ni})Pr(g_{Ci})Pr(g_{Ki}|g_{Ni}, g_{Ci})$, is the pedigree likelihood; the second component, $Pr(y_{Ni}|g_{Ni})Pr(y_{Ci}|g_{Ci})Pr(y_{Ki}|g_{Ki})$, is the penetrance term. h is the probability that the two alleles contributed by the NDJP are reduced to homozygousity (see supplementary data 2). The parameter vector is $\theta = (p_{\lambda 1}\text{'s}, \alpha_{\lambda 1}\text{'s}, \alpha_{\lambda 2}\text{'s}, \beta_{\lambda 1}\text{'s}, \beta_{\lambda 2}\text{'s})^T$, where $\lambda 1 \in \Lambda 1 = \{AA, AB, BB\}$, and $\lambda 2 \in \Lambda 2 = \{AAA, AAB, ABB, BBB\}$. Therefore, the likelihood for the i^{th} trio is

$$L_i(Y_i, G_i, h_i, \theta) \;\; = \;\; p_{\lambda 1} p_{\lambda 1} Pr(g_{Ki}|g_{Ni}, g_{Ci}) \tag{4.2}$$

$$\times f(y_{Ni},\alpha_{\lambda 1},\beta_{\lambda 1})f(y_{Ci},\alpha_{\lambda 1},\beta_{\lambda 1})f(y_{Ki},\alpha_{\lambda 2},\beta_{\lambda 2})$$

and the corresponding log likelihood is

$$\begin{aligned} l_i(Y_i,G_i,h_i,\theta) &= \log p_{\lambda 1} + \log p_{\lambda 1} + \log Pr(g_{Ki}|g_{Ni},g_{Ci}) \\ &\quad + \log f(y_{Ni},\alpha_{\lambda 1},\beta_{\lambda 1}) \\ &\quad + \log f(y_{Ci},\alpha_{\lambda 1},\beta_{\lambda 1}) \\ &\quad + \log f(y_{Ki},\alpha_{\lambda 2},\beta_{\lambda 2}) \end{aligned} \tag{4.3}$$

4.2 A beta-mixture model

A beta-mixture model is assumed for the penetrance term. In a trisomic trio, the parents are disomic and the child is trisomic. Therefore, two mixture models are needed for the data, one for the parents, and one for the children. Let y be the obseved value for an individual, we assume the following two beta mixture-models for the parents and the children respectively

$$y \sim \sum_{\lambda 1 \in \Lambda 1} \nu_{\lambda 1} f(y,\alpha_{\lambda 1},\beta_{\lambda 1}), \tag{4.4}$$

and

$$y \sim \sum_{\lambda 2 \in \Lambda 2} \nu_{\lambda 2} f(y,\alpha_{\lambda 2},\beta_{\lambda 2}), \tag{4.5}$$

where $\nu_{\lambda 1}$ is the probability of a parent having genotype $\lambda 1 \in \Lambda 1 = \{AA, AB, BB\}$, $\nu_{\lambda 2}$ is the probability of a child having genotype $\lambda 2 \in \Lambda 2 = \{AAA, AAB, ABB, BBB\}$, and

$$\begin{aligned} f(y,\alpha,\beta) &= \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, 0<y<1, \alpha,\beta>0. \end{aligned}$$

4.3 Pedigree likelihood

We follow the model proposed by Xu et al., 2004, as discussed in supplementary data 2. In our real data example, the h's are estimated from the microsatelite marker map already established on this dataset (Feingold, et al. 2000). $h = 1$ when we are sure that the two alleles from the NDJP are reduced to homozygosity, and $h = 0$ when we are sure that the two alleles from the NDJP are not reduced to homozygosity. For the purpose of genotype calling, all w's are set to 1.

4.4 Estimation

The expectation-maximization (EM) algorithm was applied to estimate the model parameters, $\theta = (\nu_{\lambda 1}\text{'s}, \alpha_{\lambda 1}\text{'s}, \beta_{\lambda 1}\text{'s}, \alpha_{\lambda 2}\text{'s}, \beta_{\lambda 2}\text{'s})^T$.

4.5 Estimation of $\nu_{\lambda 1}$'s.

Here we denote the population genotype frequencies, the p_λ's=$\nu_{\lambda 1}$'s, where $\lambda 1 \in \Lambda 1 = \{AA, AB, BB\}$. The sufficient statistic for $\nu_{\lambda 1}$ is

$$S_{1,\lambda 1} = \sum_{i=1}^{n} [1\{g_{Ni} = \lambda 1\} + 1\{g_{Ci} = \lambda 1\}] \tag{4.6}$$

E-step

At the E-step, we calculate $E(S_{1,\lambda1}|Y, \theta^{(t)})$.

$$\begin{aligned} & Pr(g_{Ni} = \lambda 1|Y_i, \theta^{(t)}) \\ = & \frac{Pr(Y_i|g_{Ni} = \lambda 1, \theta^{(t)})Pr(g_{Ni} = \lambda 1, \theta^{(t)})}{\sum_{\lambda1\in\Lambda=\{AA,AB,BB\}} Pr(Y_i|g_{Ni} = \lambda 1, \theta^{(t)}Pr(g_{Ni} = \lambda 1, \theta^{(t)})} \\ = & \frac{\nu_{\lambda1} f(y_{Ni}, \alpha_{\lambda1}, \beta_{\lambda1}) \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci}, g_{Ni} = \lambda 1, \theta^{(t)}) \prod_{\gamma\in\{Ci,Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}{\sum_{g_{Ni}} \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ni}} \nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci}, g_{Ni}, \theta^{(t)}))\} \prod_{\gamma\in\{Ni,Ci,Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})} \end{aligned} \tag{4.7}$$

Similarly

$$\begin{aligned} & Pr(g_{Ci} = \lambda 1|Y_i, \theta^{(t)}) \\ = & \frac{\nu_{\lambda1} f(y_{Ci}, \alpha_{\lambda1}, \beta_{\lambda1}) \sum_{g_{Ni}} \sum_{g_{Ki}} \nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci} = \lambda 1, g_{Ni}, \theta^{(t)}) \prod_{\gamma\in\{Ni,Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}{\sum_{g_{Ni}} \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ni}} \nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci}, g_{Ni}, \theta^{(t)}))\} \prod_{\gamma\in\{Ni,Ci,Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}. \end{aligned} \tag{4.8}$$

Therefore, we can use the above two formulas to calculate

$$\begin{aligned} & E(S_{1,\lambda1}|Y, \theta^{(t)}) \\ = & \sum_{i=1}^{n} \Big[Pr(g_{Ni} = \lambda 1|Y_i, \theta^{(t)} + Pr(g_{Ci} = \lambda 1|Y_i, \theta^{(t)})\Big] \end{aligned} \tag{4.9}$$

M-step

At the M-step, we update $\nu_{\lambda1}$ using the following formula,

$$\nu_{\lambda1}^{(t+1)} = \frac{E(S1_\lambda|Y, \theta^{(t)})}{2n}, \tag{4.10}$$

where n is the number of family trios.

4.6 Estimation of the beta components.

- Estimation of $\alpha_{\lambda1}$'s and $\beta_{\lambda1}$'s

  The part of the log likelihood that involves $\alpha_{\lambda1}$ and $\beta_{\lambda1}$ is

  $$l_{\lambda1} = \sum_{i=1}^{n} 1\{g_{Ni} = \lambda 1\} \log f(y_{Ni}, \alpha_{\lambda1}, \beta_{\lambda1}) + \sum_{i=1}^{n} 1\{g_{Ci} = \lambda 1\} \log f(y_{Ci}, \alpha_{\lambda1}, \beta_{\lambda1}).$$

  E-step

  At the E-step, we calculate

$$\begin{aligned} & E(l_{\lambda1}|Y, \theta^{(t)}) \\ = & \sum_{i=1}^{n} Pr(g_{Ni} = \lambda 1|Y_i, \theta^{(t)}) \log f(y_{Ni}, \alpha_{\lambda1}, \beta_{\lambda1}) \\ & + \sum_{i=1}^{n} Pr(g_{Ci} = \lambda 1|Y_i, \theta^{(t)}) \log f(y_{Ci}, \alpha_{\lambda1}, \beta_{\lambda1}). \end{aligned} \tag{4.11}$$

$Pr(g_{Ni} = \lambda 1|Y_i, \theta^{(t)})$ and $Pr(g_{Ci} = \lambda 1|Y_i, \theta^{(t)})$ can be clculated using equations 4.7 and 4.8 respectively.

M-step
At the M-Step, we maximize $E(l_{\lambda 1}|Y, \theta^{(t)})$ using the *nlm* procedure included in the *R*-package to get $\alpha_{\lambda 1}^{(t+1)}$ and $\beta_{\lambda 1}^{(t+1)}$.

- Estimation of $\alpha_{\lambda 2}$'s and $\beta_{\lambda 2}$'s

  The part of the log likelihood that involves $\alpha_{\lambda 2}$s and $\beta_{\lambda 2}$s is

$$l_{\lambda 2} = \sum_{i=1}^{n} 1\{g_{Ni} = \lambda 1\} \log f(y_{Ki}, \alpha_{\lambda 2}, \beta_{\lambda 2}).$$

E-step
At the E-step, we calculate

$$E(l_{\lambda 2}|Y, \theta^{(t)}) = \qquad \sum_{i=1}^{n} Pr(g_{Ki} = \lambda 2|Y_i, \theta^{(t)}) \log f(y_{Ki}, \alpha_{\lambda 2}, \beta_{\lambda 2}).$$

Similar to derivation shown above, we get

$$\begin{aligned} & Pr(g_{Ki} = \lambda 2|Y_i, \theta^{(t)}) \qquad (4.12) \\ = & \frac{Pr(Y_i|g_{Ki} = \lambda 2, \theta^{(t)}) Pr(g_{Ki} = \lambda 2, \theta^{(t)})}{\sum_{\lambda 2 \in \Lambda 2 = \{AAA, AAB, ABB, BBB\}} Pr(Y_i|g_{Ki} = \lambda 2, \theta^{(t)} Pr(g_{Ki} = \lambda 2, \theta^{(t)})} \\ = & \frac{f(y_{Ki}, \alpha_{\lambda 2}, \beta_{\lambda 2}) \sum_{g_{Ni}} \sum_{g_{Ci}} p_{g_{Ni}} p_{g_{Ci}} Pr(g_{Ki} = \lambda 2|g_{Ci}, g_{Ni}) \prod_{\gamma \in \{Ni, Ci\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}{\sum_{g_{Ni}} \sum_{g_{Ci}} \sum_{g_{Ki}} p_{g_{Ni}} p_{g_{Ci}} Pr(g_{Ki}|g_{Ci}, g_{Ni})\} \prod_{\gamma \in \{Ni, Ci, Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}. \end{aligned}$$

M-step
At the M-Step, we maximize $E(l_{\lambda 2}|Y, \theta^{(t)})$ using the *nlm* procedure included in the R-package to get $\alpha_{\lambda 2}^{(t+1)}$ and $\beta_{\lambda 2}^{(t+1)}$.

4.7 Genotype prediction using Bayes rule

Once the parameters are estimated, we can use the Bayes rule to determine the genotypes of the family members. The posterior probability of the family genotype vector $G = (g_N, g_C, g_K)$ given the observed values $Y = (y_N, y_C, y_K)$ is

$$\begin{aligned} & p(G|Y) \qquad (4.13) \\ = & \frac{\nu_{\lambda 1} \nu_{\lambda 1} Pr(g_K|g_N, g_C) f(y_N, \xi_{\lambda 1}) f(y_C, \xi_{\lambda 1}) f(y_K, \xi_{\lambda 2})}{\sum_{j=1:18} \nu_{\lambda 1} \nu_{\lambda 1} Pr(g_{Kj}|g_{Nj}, g_{Cj}) f(y_{Nj}, \xi_{\lambda 1}) f(y_{Cj}, \xi_{\lambda 1}) f(y_{Kj}, \xi_{\lambda 2})}. \end{aligned}$$

Here the $\xi_\lambda = (\alpha_\lambda, \beta_\lambda)^T$ are the parameters for the beta distribution of genotype cluster λ.

5 Supplementary Data 5: Disomic Simulation Study Results

Two simulation studies were conducted to compare the performance of different clustering algorithms. For each simulation study, we simulated 1000 datasets. Each dataset consists of 150 trios. A beta distribution was used in simulating the observations in different genotype groups. Six different clustering methods were applied to these datasets. Three methods treat each individual independently: the K-means clustering method, the regular Gaussian-mixture model for independent data and the regular beta-mixture model for independent data. Three corresponding methods treat the family as a group: the trio K-means method, the trio Gaussian-mixture model and the trio beta-mixture model.

The datasets we simulated in the first simulation study represent "good" data, because the genotype clusters are well separated. The distributions of the AA and the BB genotype clusters are highly skewed and the distribution of the AB genotype cluster is relatively more symmetric following our experience with SNP array data in general. We then simulated datasets that represent "bad" data. The genotype clusters are less well defined and the distributions of each cluster are wider compared to the "good" data. **In the original paper, we reported the results of the second simulation only.** The detailed results for these two simulation studies are summarized in the following two tables.

Table S2: Simulation study 1: "good" disomic data.

Methods	K-means	Regular Gaussian-mixture model	Regular beta-mixture model	Trio K-means	Trio Gaussian-mixture model	Trio beta-mixture model
Average number of mistakes						
Total	0.253	0.699	0.037	0.127	0.549	0.024
Misscalled heterozygotes	0.25	0	0.019	0.125	0	0.011
Misscalled homozygotes	0.003	0.699	0.018	0.002	0.549	0.013
Number of simulations with						
0 error	775	514	964	883	589	976
1 errors	200	320	35	107	295	24
2 errors	22	122	1	10	94	0
3 errors	3	41	0	0	22	0
4 errors	0	3	0	0	0	0

A total of 1000 datasets were simulated. Each dataset consisted 150 disomic trios. Population genotype frequencies were set at $p_{AA} = 0.2$, $p_{AB} = 0.35$ and $p_{BB} = 0.45$. The beta parameters used in the simulations for the three genotype clusters were $\alpha_{AA} = 1$, $\beta_{AA} = 40$, $\alpha_{AB} = 20$, $\beta_{AB} = 20$, $\alpha_{BB} = 40$, $\beta_{BB} = 1$.

Table S3: Simulation study 2: “bad” disomic data.

Methods	K-means	Regular Gaussian-mixture model	Regular beta-mixture model	Trio K-means	Trio Gaussian-mixture model	Trio beta-mixture model
Average number of mistakes						
Total	14.62	6.65	5.31	7.13	4.42	3.47
Misscalled heterozygotes	14.46	1.38	2.89	6.93	1.09	1.90
Misscalled homozygotes	0.16	5.27	2.42	0.20	3.33	1.57
Number of simulations with						
0 error	0	1	3	3	15	35
1-5 errors	8	349	572	285	705	819
6-10 errors	158	574	394	599	274	145
10-15 errors	449	73	31	108	6	1
>15 errors	385	3	0	5	0	0

A total of 1000 datasets were simulated. Each dataset consisted 150 disomic trios. Population genotype frequencies were set at $p_{AA} = 0.2$, $p_{AB} = 0.35$ and $p_{BB} = 0.45$. The beta parameters used in the simulations for the three genotype clusters were $\alpha_{AA} = 5$, $\beta_{AA} = 40$, $\alpha_{AB} = 10$, $\beta_{AB} = 10$, $\alpha_{BB} = 40$, $\beta_{BB} = 5$.

6 Supplementary Data 6: Trisomic Simulation Study Results

This document provides the detailed results for Figure 4 of the original paper. A total of 1000 datasets were simulated and 150 family trios were simulated in each dataset. A beta distribution was used to simulate the observed values. Population genotype frequencies were set at $p_{AA} = 0.2$, $p_{AB} = 0.35$ and $p_{BB} = 0.45$. The K-means clustering method, the regular beta-mixture model, the trio K-means and the trio beta-mixture model were applied to the simulated datasets. The results of the simulation study are summarized in Table S4.

Table S4: Trisomic simulation study results

Methods	K-means	Regular beta-mixture model	Trio K-means	Trio beta-mixture model
Average number of mistakes				
	4.41	4.18	3.23	1.44
Number of simulations with				
0 error	14	32	35	233
1-5 errors	714	762	825	761
6-10 errors	265	162	138	6
10-15 errors	7	27	2	0
>15 errors	0	15	0	0

The beta parameters used in simulation of the parental genotypes are: $\alpha_{AA} = 2, \quad \beta_{AA} = 20, \quad \alpha_{AB} = 60, \quad \beta_{AB} = 84, \quad \alpha_{BB} = 2, \quad \beta_{BB} = 2$. Those for the four genotype clusters of the children's genotyes are: $\alpha_{AAA} = 2, \quad \beta_{AA} = 20, \quad \alpha_{AAB} = 26, \quad \beta_{AB} = 51, \quad \alpha_{ABB} = 83, \quad \beta_{BB} = 65, \quad \alpha_{ABB} = 2, \quad \beta_{BB} = 20$.