# *Supplementary Text*

## *The TrioPhase Algorithm*

To formally describe TrioPhase, we will introduce some notations. We assume that we are given a set of trios, $t_1, t_2, ..., t_n$, where each trio consists of a mother, father and child, which will be denoted by $m_i, f_i$ and $c_i$ respectively. We consider the case where we genotype *k* SNPs. We denote the genotypes of the trio $t_i$ by the vector $(m_{i1}, m_{i2}, ..., m_{ik}), \in \{0,1,2\}^k$ for the mother, $(f_{i1}, f_{i2}, ..., f_{ik}) \in \{0,1,2\}^k$ for the father, and $(c_{i1}, c_{i2}, ..., c_{ik}) \in \{0,1,2\}^k$ for the child. For each trio $t_i$, two pools are genotyped, the mother-child pool $p_i^{mc}$, and the father-child pool $p_i^{fc}$. The allele frequency of each of the pools is denoted by the vectors $(p_{i1}^{mc}, ..., p_{ik}^{mc}) \in [0,1]^k$, and $(p_{i1}^{fc}, ..., p_{ik}^{fc}) \in [0,1]^k$ respectively. Theoretically, if no noise is introduced, we expect to have the following equation satisfied:

$$p_{ij}^{mc} = (m_{ij} + c_{ij})/4$$
$$p_{ij}^{fc} = (f_{ij} + c_{ij})/4$$

In practice, these equations are only satisfied approximately, due to different chemical characteristics of each of the SNPs, and due to the fact that the pools do not consist of equimolar samples. Thus, the algorithm aims in finding for each pool a quantity ratio $q$, which describes the ratio between the DNA quantity of the parent and the child. Furthermore, for each SNP, we model the effect of the SNP on the resulting allele frequency reads by a polynomial of degree three. For SNP $i$ we estimate two parameters, $a_i$ and $b_i$. The algorithm assumes that if the real frequency in SNP $i$ were

$x$, then the read frequency would be $g_i(x) = a_i x^3 + b_i x^2 + (1 - a_i - b_i)x$. Putting it all together, the algorithm assumes that the following equations hold approximately:

$$p_{ij}^{mc} \approx g_i\left(\frac{q_i^m m_{ij} + c_{ij}}{2(q_i^m + 1)}\right) = a_i\left(\frac{q_i^m m_{ij} + c_{ij}}{2(q_i^m + 1)}\right)^3 + b_i\left(\frac{q_i^m m_{ij} + c_{ij}}{2(q_i^m + 1)}\right)^2 + (1 - a_i - b_i)\left(\frac{q_i^m m_{ij} + c_{ij}}{2(q_i^m + 1)}\right)$$

$$p_{ij}^{fc} \approx g_i\left(\frac{q_i^f f_{ij} + c_{ij}}{2(q_i^f + 1)}\right) = a_i\left(\frac{q_i^f f_{ij} + c_{ij}}{2(q_i^f + 1)}\right)^3 + b_i\left(\frac{q_i^f f_{ij} + c_{ij}}{2(q_i^f + 1)}\right)^2 + (1 - a_i - b_i)\left(\frac{q_i^f f_{ij} + c_{ij}}{2(q_i^f + 1)}\right)$$

The algorithm has to learn the parameters $f_i, m_i, c_i, a_i, b_i, q_i^m$ and $q_i^f$ from the data. In order to do so, we search for a set of parameters so that the accumulated error introduced to the above equations is minimized in the least squares sense. Formally, we search for set of parameters that minimizes the following:

$$\lambda(\{a_i\},\{b_i\},\{q_i^m\},\{q_i^f\},\{f_i\},\{m_i\},\{c_i\}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left[\left(p_{ij}^{mc} - g_j\left(\frac{q_i^m m_{ij} + c_{ij}}{2(q_i^m + 1)}\right)\right)^2 + \left(p_{ij}^{fc} - g_j\left(\frac{q_i^f f_{ij} + c_{ij}}{2(q_i^f + 1)}\right)\right)^2\right].$$

If all parameters are fixed, except for $\{f_i\},\{m_i\},\{c_i\}$, it is easy to minimize $\lambda$ by enumerating over all 15 possible assignments for each trio and SNP (see Table 1). To find the minimum value of $\lambda$ in the entire parameter space, we use the following iterative algorithm.

1. Initialization: We initially set $a_j = b_j = 0,\ q_i^f = q_i^m = 1$.

2. We fix the values of $\{q_i^m\}$ and $\{q_i^f\}$, and search for the values of $\{a_j\}$ and $\{b_j\}$. For each $j = 1,...,k$ we solve the following optimization problem:

$$\text{Minimize} \quad \lambda_j = \sum_{i=1}^{n} \left[ \left( p_{ij}^{mc} - g_j\left( \frac{q_i^m m_{ij} + c_{ij}}{2(q_i^m + 1)} \right) \right)^2 + \left( p_{ij}^{fc} - g_j\left( \frac{q_i^f f_{ij} + c_{ij}}{2(q_i^f + 1)} \right) \right)^2 \right]$$

$$\text{such that :} \quad -0.5 \le a_j \le 0.5$$

$$-0.15 - \frac{3a_j}{2} \le b_j \le 0.15 - \frac{3a_j}{2}$$

The last two constraints ensure that the values of $g_j(x)$ are close to $x$ for $x = 0, 0.25, 0.5, 0.75, 1$. This problem is easy when the genotype values are known. Since this is not the case, we enumerate over all values of $\{a_j\}$ and $\{b_j\}$ in the above range, with increments of 0.01. For each such value, we calculate the value of $\lambda_j$ by enumerating over all 15 genotype values for each SNP, and we keep the parameters that achieve the minimum value.

3. We fix the values of $\{a_j\}$ and $\{b_j\}$, and we search for the values of $q_i^m$ and $q_i^f$. For each $i = 1, ..., n$ we solve the following optimization problem:

$$\text{Minimize} \quad \mu_i = \sum_{j=1}^{k} \left[ \left( p_{ij}^{mc} - g_j\left( \frac{q_i^m m_{ij} + c_{ij}}{2(q_i^m + 1)} \right) \right)^2 + \left( p_{ij}^{fc} - g_j\left( \frac{q_i^f f_{ij} + c_{ij}}{2(q_i^f + 1)} \right) \right)^2 \right]$$

$$\text{such that :} \quad 0.9 \le q_i^m,\ q_i^f \le 1.1$$

We enumerate over all values of $q_i^m$ and $q_i^f$ in the given range, with increments of 0.01. For each such value, we calculate the value of $\mu_i$ by enumerating over all 15 genotype configurations for every SNP, and we keep the parameters that achieve the minimum.

3. We repeat steps 3 and 4 until the value of $\lambda$ does not decrease (or only a negligible decrease has been achieved).

<u>Tradeoff between discordance rate and no-call rate</u>

For each trio, we calculated a score, which is the sum of squares of deviations from the equations above. When one of the pools of the trio was considered a no-call by the allelotyping software, or when the signal to noise ratio of the two alleles was small (< 5), we added 5 to the score. The trios were then sorted according to their score. For a missing data threshold *t*, we remove the first *t* trios from the dataset, and compute the discordance for the remaining trios. This results in a tradeoff between the missing data rate and the discordance rate (see Table 2).

*The Greedy Algorithm for Resolving Missing Data (Confidence Intervals)*

When a short region is being considered, a greedy algorithm is used to infer the haplotypes from the confidence intervals. The confidence intervals can be viewed as a set of constraints on the possible haplotypes configurations of the trio. For instance, if the region consists of two SNPs, and the confidence intervals are [2,3],[4,4] for the first SNP, and [0,1],[2,3] for the second SNP, then the four haplotypes of the trio may either be (11,10,10,00) or (11,11,10,00), where the haplotypes are ordered by the order given in Table 1. The greedy algorithm attempts to assign four haplotypes to each of the trios, such that the constraints set by the confidence intervals are satisfied. In each step, the algorithm searches for the haplotype that can be assigned to the maximal number of trios, and assigns this haplotype to these trios. Thus, each step of the algorithm results in a partial assignment of haplotypes to the trios. Initially, there are no haplotypes assigned to any of the trios. Once a haplotype is assigned to a trio, this assignment cannot be changed or removed. The algorithm terminates when all trios have been assigned four haplotype.

More formally, in order to decide which haplotype to assign next, we define a score to each possible haplotype. For a haplotype $h$, and a trio $j$, we define the score $\sigma_j(h)$ as the maximum number copies of $h$ that can be assigned to trio $j$, without violating the constraints imposed by the confidence intervals. Note that if the trio is already partially assigned by $l \leq 4$ haplotypes, then the number of copies of $h$ that can be assigned to $j$ is at most $4-l$ (i.e. $\sigma_j(h) \leq l-4$). The total score of h is then defined as $\sigma(h) = \sum_j \sigma_j(h)$, where the sum is taken over all trios. Thus, the score of h represents the total number of copies of h that can be assigned to the trios. In each step of the algorithm we pick the haplotype h with maximum score. We then assign $\sigma_j(h)$ copies of h to trio j.

When the algorithm terminates, four haplotypes are assigned to each trio. It is then easy to find the genotypes information from the haplotype information.

When a longer region is considered, the greedy algorithm is performed for overlapping short regions. Each SNP is then covered by more than one window. For each such window, the greedy algorithm sets a genotype value at this SNP. The final genotype value reported by the algorithm is found by a majority vote over the genotype assignments for that SNP over all the overlapping windows containing it.

## *Supplementary Figure Legends*

**Supplementary Figure 1:** Allelic frequencies of pools for the "best-performing" SNP assay, for rs2560643. The x-axes correspond to pool index, ordered by increasing pool allelic frequency. Top: the pools' raw allelic frequency as estimated by MassARRAY genotyping (open boxes) versus Triophase-corrected allelic frequencies (filled circles). Bottom: the pools' known allelic frequencies (large grey circles) versus Triophase-assigned frequency bins (small dark circles). No errors in pool frequency estimation were made.

**Supplementary Figure 2:** Allelic frequencies of pools for the "worst-performing" SNP rs4382469. The x-axes correspond to pool index, ordered by increasing pool allelic frequency. Top: the pools' raw allelic frequency as estimated by MassARRAY genotyping (open boxes) versus Triophase-corrected allelic frequencies (filled circles). Bottom: the pools' known allelic frequencies (large grey circles) versus Triophase-assigned frequency bins (small dark circles). Seven errors in pool frequency estimation were made (lone small dark circles)

**Supplementary Figure 3:** An overview of raw (left) versus Triophase-corrected (right) allelic frequencies of pools. The average standard deviation in allelic frequency estimation was 0.045 ± 0.025.

# *Supplementary Tables*

**Supplementary Table 1:** The pool composition of the trios in pedigree 66 from Coriell. Each row corresponds to one pool of two individuals. In the pedigree, individuals 12547 and 12548 are the parents of individuals 12549,12551,12552,12553,12554 and 12555. Together with the grandparents, this results in a set of eight trios.

| Pool ID | Sample #1 ID | Sample #2 ID |
|---|---|---|
| 1 | 12547 | 12549 |
| 2 | 12547 | 12551 |
| 3 | 12547 | 12552 |
| 4 | 12547 | 12553 |
| 5 | 12547 | 12554 |
| 6 | 12547 | 12555 |
| 7 | 12547 | 12556 |
| 8 | 12547 | 12557 |
| 9 | 12548 | 12549 |
| 10 | 12548 | 12551 |
| 11 | 12548 | 12552 |
| 12 | 12548 | 12553 |
| 13 | 12548 | 12554 |
| 14 | 12548 | 12555 |
| 15 | 12548 | 12558 |
| 16 | 12548 | 12559 |

| | |
|---|---|
| | Father |
| | Paternal grandparents |
| | Mother |
| | Maternal grandparents |
| | Children |