ПРИЛОЖЕНИЯ
к статье А.А. Бердниковой, И.В. Зоркольцевой, Я.А. Цепилова, Е.Е. Елгаевой
«Импутация генотипов в геномных исследованиях человека»

Приложение 1. Сравнительная характеристика Michigan Imputation Server и TOPMed Imputation Server

	Michigan Imputation Server	TOPMed Imputation Server
Инструменты для импутации	MINIMAC 4; MINIMAC 3	MINIMAC 4
Инструменты для фазирования	Eagle v2.4; Beagle 5.4	Eagle v2.4
Референсные панели	1000G Phase 1 v3 Shapeit2 (GRCh37/hg19); 1000G Phase 3 (GRCh38/hg38) [BETA]; 1000G Phase 3 30x (GRCh38/hg38) [BETA]; 1000G Phase 3 v5 (GRCh37/hg19); CAAPA African American Panel (GRCh37/hg19); Genome Asia Pilot – м (GRCh37/hg19); HapMap 2 (GRCh37/hg19); HRC r1.1 2016 (GRCh37/hg19)	TOPMed r3 (GRCh38/hg38)

Приложение 2. Вывод формулы (1)

Метрика $\hat{R}_d^2$:

$$\hat{R}_d^2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(D_i - 2\hat{p})^2}{2\hat{p}(1-\hat{p})},$$

где N – количество индивидов в выборке, D_i –доза импутированного аллеля для i-го индивида, $\hat{p} = \sum_{i=1}^{N}\frac{D_i}{2N}$ – оценка частоты аллеля. Генотипы в кодировке 0, 1, 2 (альтернативный аллель с частотой ($\hat{p}$) закодирован как 1, референсный с частотой ($1 - \hat{p}$) закодирован как 0).

Метрика $\hat{R}_d^2$ представляет собой отношение дисперсии дозы аллеля и ожидаемой дисперсии при равновесии Харди–Вайнберга. Рассмотрим вывод формулы. Допустим, что число индивидов в выборке равно N, тогда общее количество генотипов тоже равно N. Среднее значение генотипа в выборке при условии равновесия Харди–Вайнберга и бинарной кодировке генотипа будет расписываться следующим образом:

$$\bar{x} = \frac{\sum_1^N x_i}{N} = \frac{0*N_0+1*N_1+2*N_2}{N} = \frac{0*(1-\hat{p})^2N+1*2*\hat{p}(1-\hat{p})N+2*\hat{p}^2N}{N} = 2\hat{p},$$

где 0, 1, 2 – значения генотипов в бинарной кодировке, N_i – количество i-го генотипа в выборке, $(1 - \hat{p})^2$, $\hat{p}(1 - \hat{p})$ и $\hat{p}^2$ – частоты генотипов при равновесии Харди–Вайнберга.

Тогда дисперсия импутированных генотипов, измеряемых в дозах аллеля, при ожидаемом равновесии Харди–Вайнберга будет описываться как:

$$\frac{\sum_1^N(x_i - \bar{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(D_i - 2\hat{p})^2.$$

Ожидаемая дисперсия реальных генотипов при условии равновесия Харди–Вайнберга в таком случае равна

$$\frac{\sum_1^N(x_i - \bar{x})^2}{N} = \frac{(0-2\hat{p})^2 * N_0 + (1-2\hat{p})^2 * N_1 + (2-2\hat{p})^2 * N_2}{N} =$$

$$= \frac{(0-2\hat{p})^2(1-\hat{p})^2N + 2(1-2\hat{p})^2(1-\hat{p})\hat{p}N + (2-2\hat{p})^2\hat{p}^2N}{N} = 2\hat{p}(1-\hat{p}).$$

Приложение 3. Вывод формулы (2)

Метрика $\hat{R}_h^2$:

$$\hat{R}_h^2 = \frac{\frac{1}{2N}\sum_{i=1}^{2N}(H_i-\hat{p})^2}{\hat{p}(1-\hat{p})},$$

где N – количество индивидов в выборке, H_i – вероятность импутированного аллеля в i-м гаплотипе (варьируется от 0 до 1), $\hat{p} = \sum_{i=1}^{2N} H_i/(2N)$ – оценка частоты аллеля.

Метрика $\hat{R}_h^2$ рассчитывается аналогично метрике $\hat{R}_d^2$ (см. Приложение 2). Допустим, что число индивидов в выборке равно N, тогда общее количество генотипов тоже равно N, а общее количество аллелей равно $2N$. Среднее значение генотипа в выборке при условии равновесия Харди–Вайнберга и бинарной кодировке аллелей будет расписываться следующим образом:

$$\bar{x} = \frac{\sum_1^{2N} x_i}{2N} = \frac{0 * N_0 + 1 * N_1}{N} = \frac{0 * (1-\hat{p}) * 2N + 1 * \hat{p}(1-\hat{p}) * 2N}{2N} = \hat{p},$$

где 0, 1 – значения аллелей в бинарной кодировке, N_i – количество i-го аллеля в выборке, $\hat{p}$ и $(1 – \hat{p})$ – частоты аллелей при равновесии Харди–Вайнберга.

Тогда дисперсия импутированных генотипов, измеряемых в вероятностях импутированных аллелей, при ожидаемом равновесии Харди–Вайнберга будет описываться как:

$$\frac{\sum_1^{2N}(x_i-\bar{x})^2}{2N} = \frac{1}{2N}\sum_{i=1}^{2N}(H_i-\hat{p})^2.$$

Ожидаемая дисперсия при равновесии Харди–Вайнберга равна:

$$\frac{\sum_1^{2N}(x_i-\bar{x})^2}{2N} = \frac{N_0(0-\hat{p})^2(1-\hat{p}) + N_1(1-\hat{p})^2\hat{p}^2}{2N}$$
$$= \frac{2N(0-\hat{p})^2(1-\hat{p}) + 2N(1-\hat{p})^2\hat{p}^2}{2N} = \hat{p}(1-\hat{p}).$$