Supplemental Material: Estimating error models for whole genome sequencing using mixtures of Dirichlet-multinomial distributions

Steven H. Wu [1], Rachel S. Schwartz [1,3], David J. Winter [1], Donald F. Conrad [4], Reed A. Cartwright [1,2,*]

[1]The Biodesign Institute, and [2]School of Life Sciences, Arizona State University, Tempe, AZ, 85281, USA; [3]Department of Biological Sciences, The University of Rhode Island, Kingston, RI 02881, USA; [4]Department of Genetics, Department of Pathology and Immunology, Washington University School of Medicine, Saint Louis, MO, 63110, USA

[*]To whom correspondence should be addressed. Tel: +1 480-965-9949; Email: cartwright@asu.edu

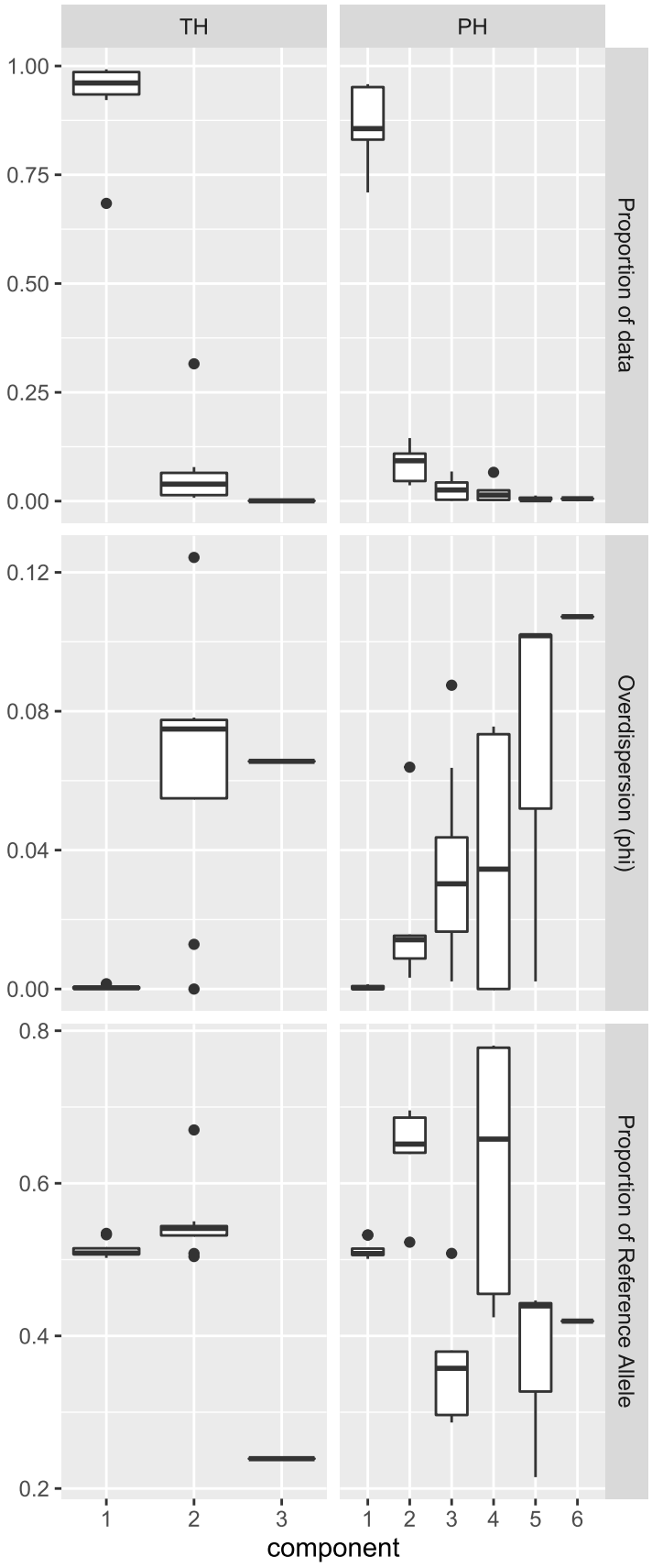

Figure S1: **Comparison of the parameter distributions between PH and TH datasets in the CEU data.** Boxplots summarize the proportion of data in each component, the overdispersion parameter for each component, and the proportion of reference allele in each component across different years and different chromosome regions. Most data is found in the major component of the model, but this is more true of the TH data. The overdispersion of the data increases as the proportion of data in the component decreases, with the major component having almost no overdispersion. The major component contains approximately equal proportions of the reference an alternate allele, with other components having widely disparate proportions of each allele. Thus, the major component, comprising the majority of the data is similar to a binomial without bias.

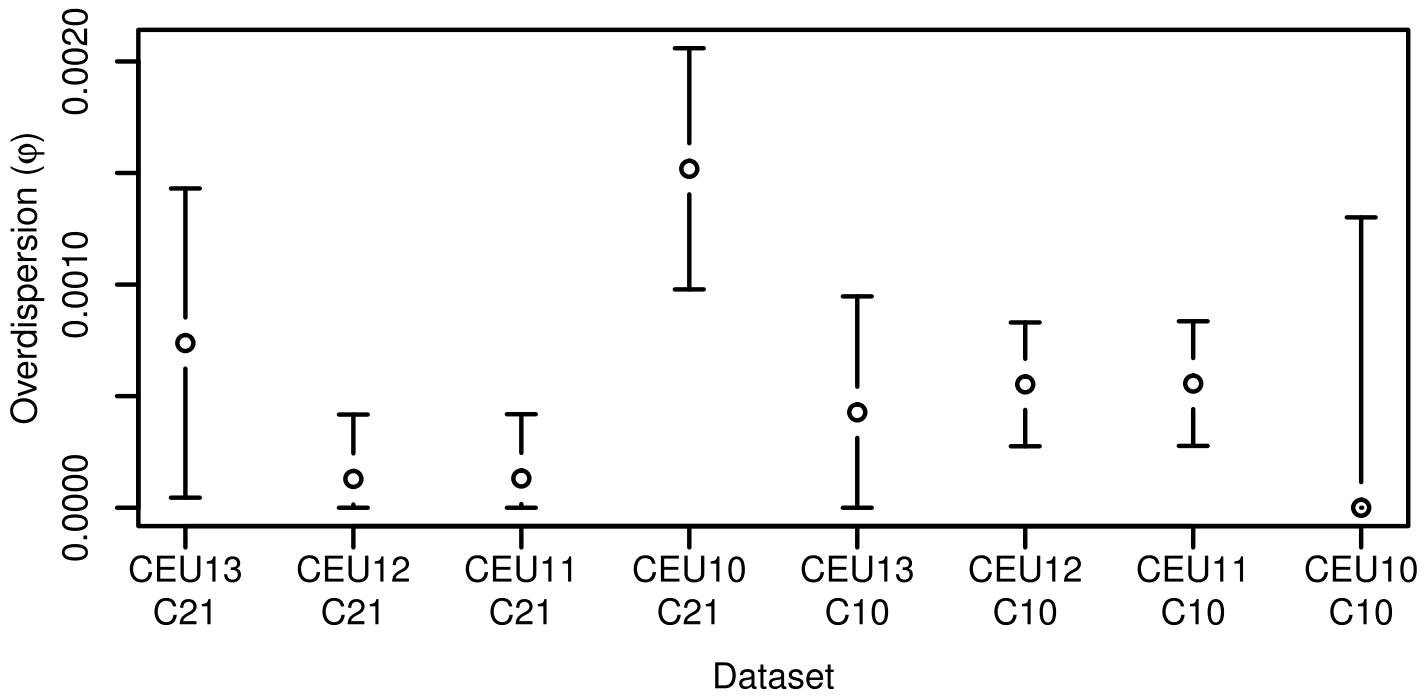

Figure S2: Confidence interval for φ from the major component for different TH datasets. Most of the estimated φ overlap in chromosome 21 and all of them in chromosome 10. Confidence intervals calculated from the EM algorithm.

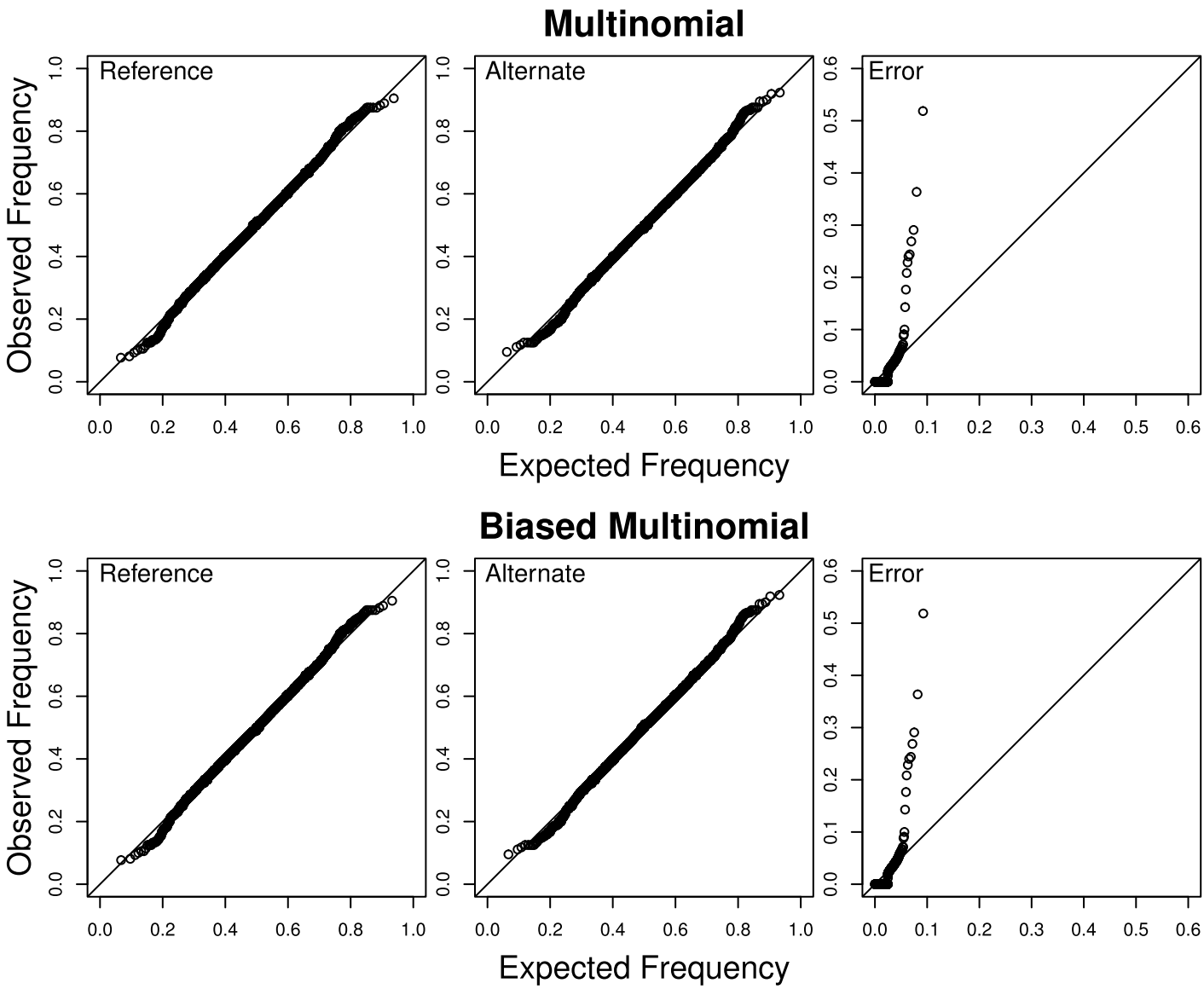

Figure S3: **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU13 Chr10 TH.** QQ plots evaluating the fits of several model to this dataset. The quantiles of the observed read count frequencies are calculated from the datasets, and the quantiles of the expected read count frequencies are estimated from the fitted model. A model that fits the data well produces points that fall along the diagonal.

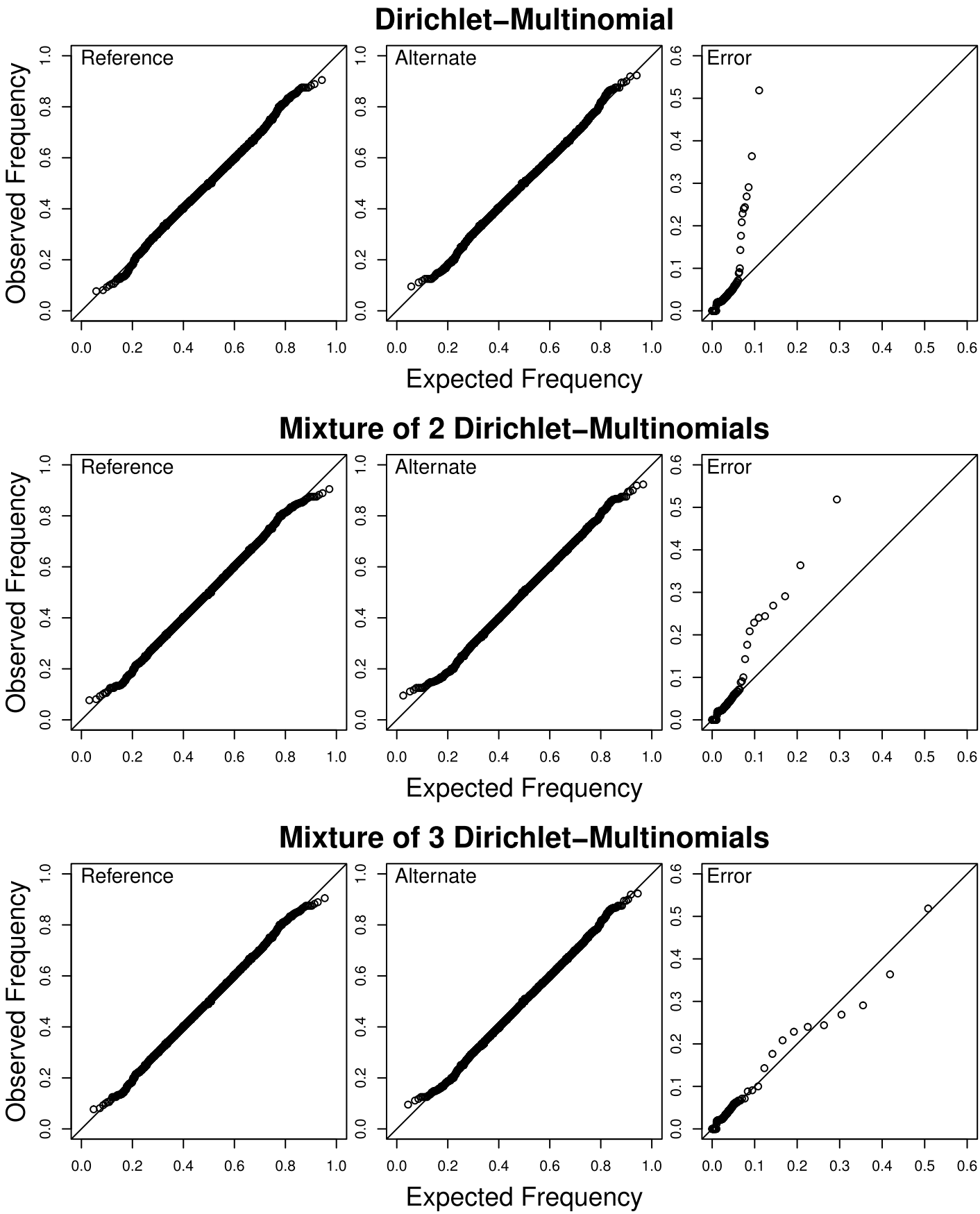

Figure S3 (Continued): **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU13 Chr10 TH.**

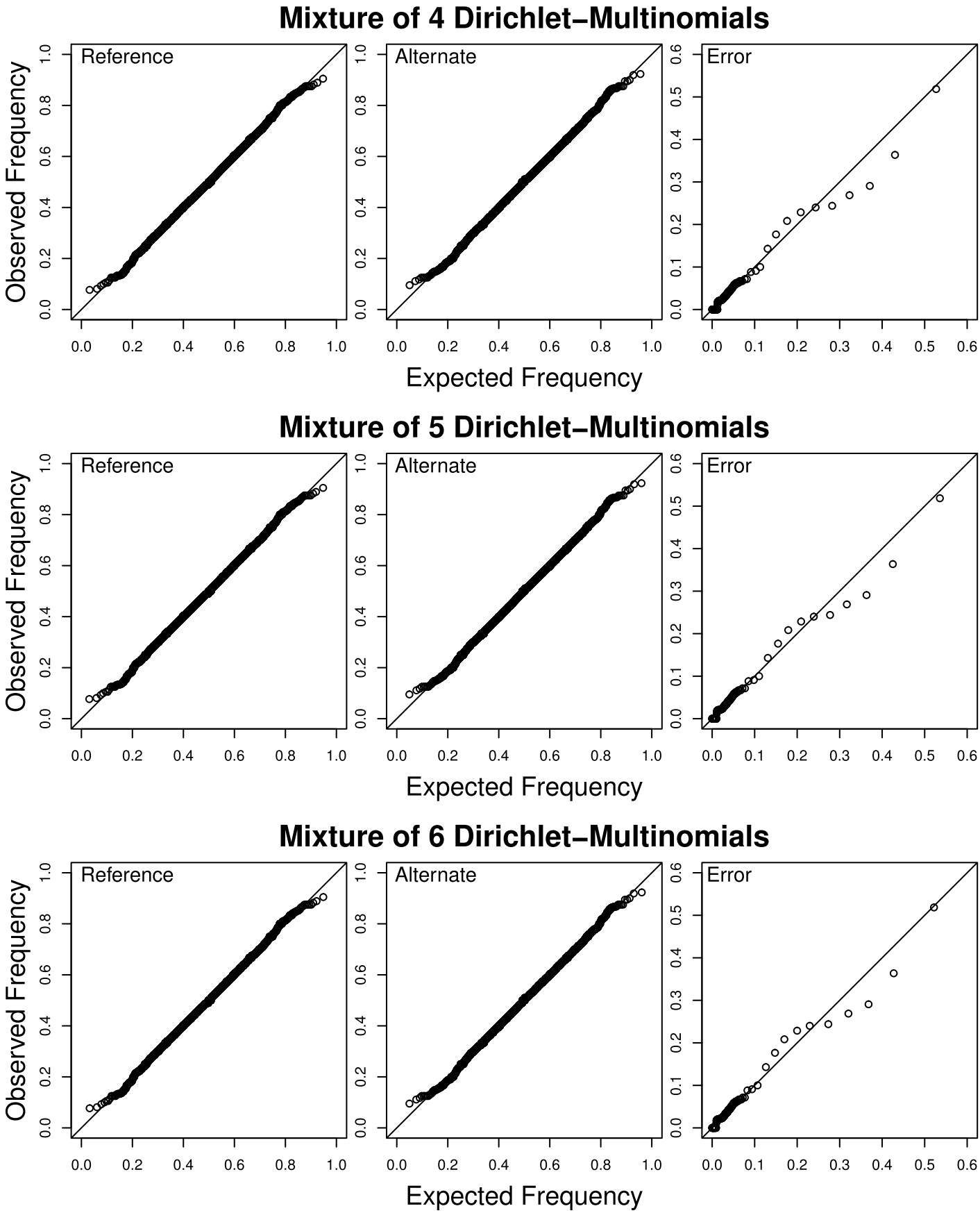

Figure S3 (Continued): **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU13 Chr10 TH.**

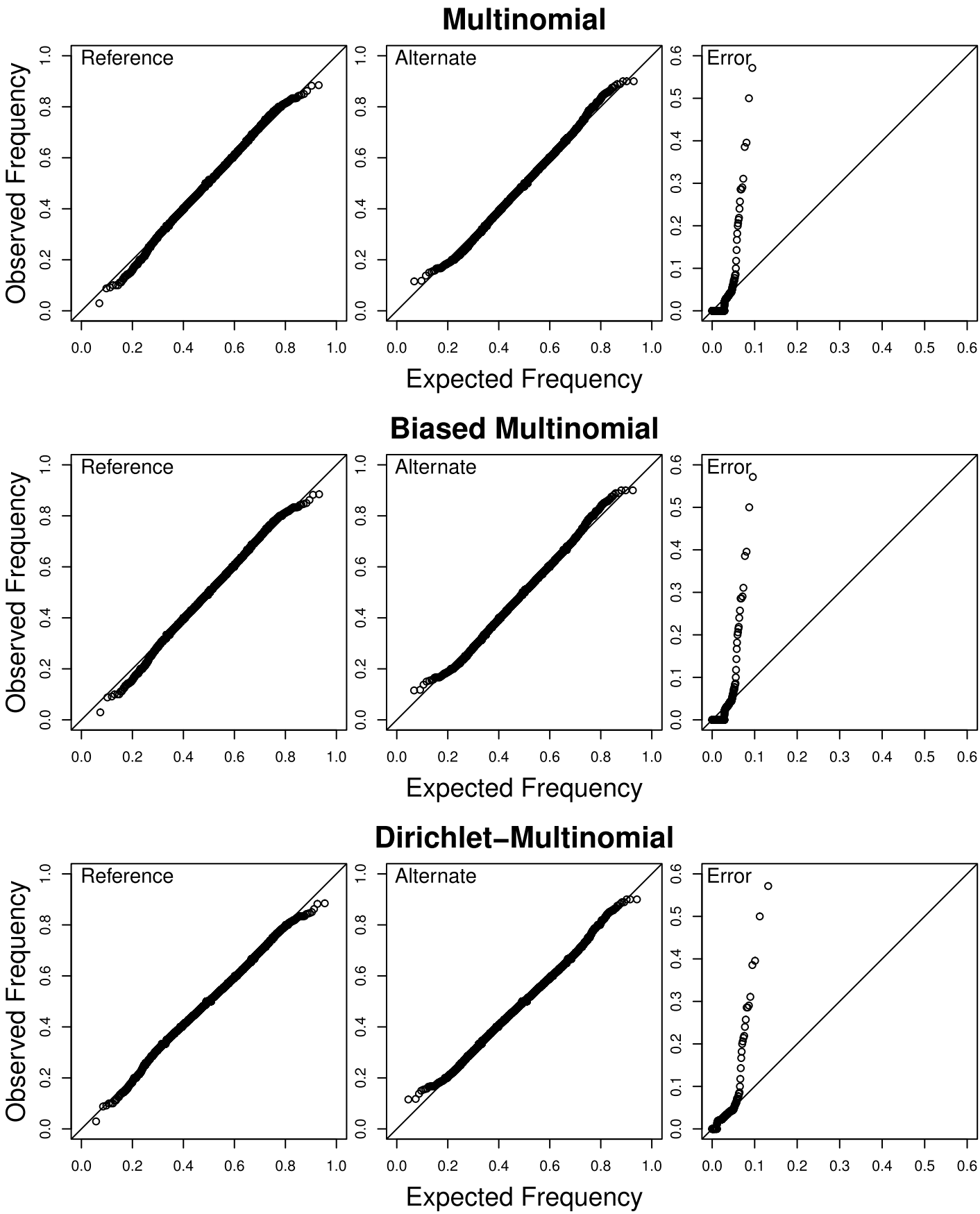

Figure S4: **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU13 Chr21 TH.**

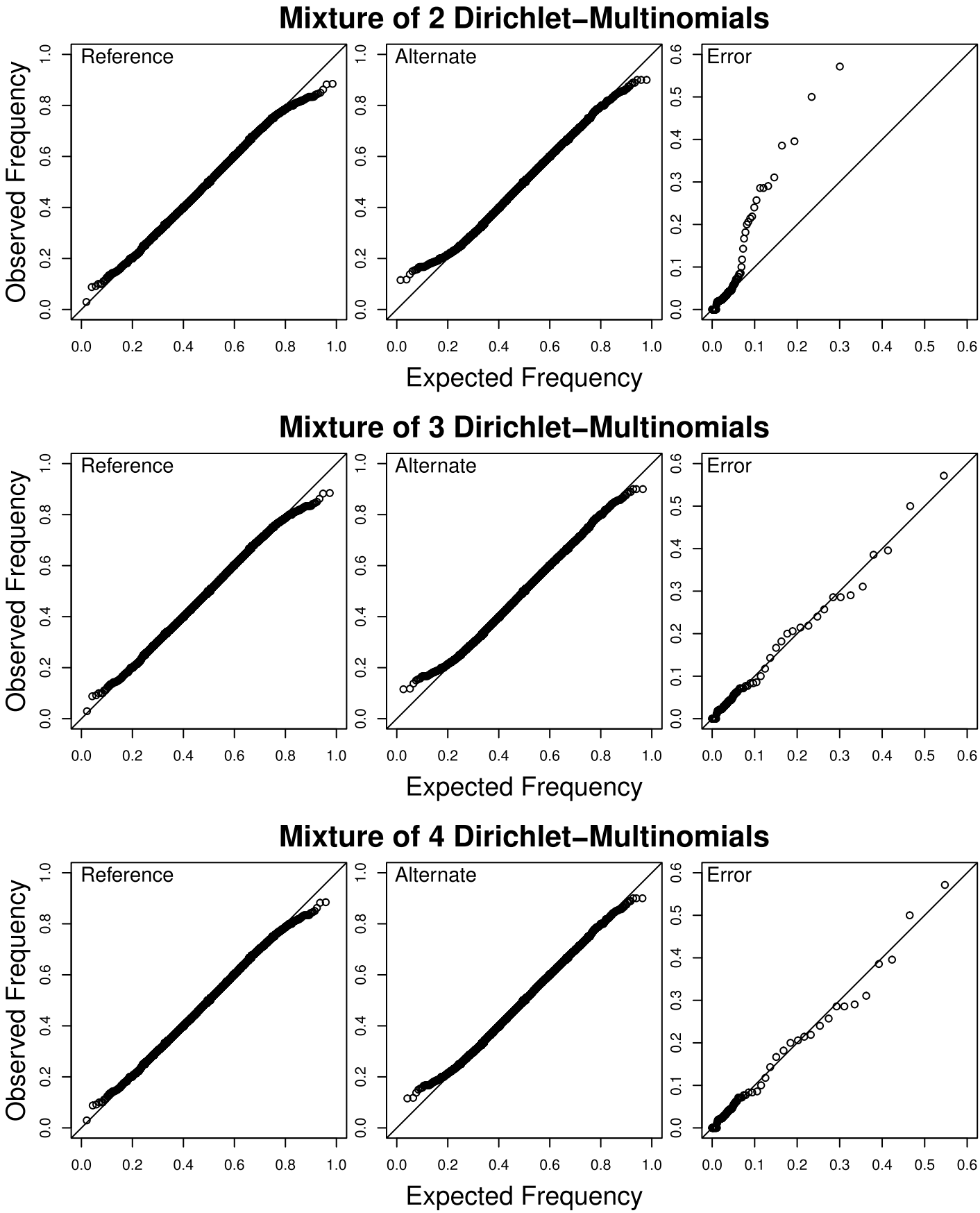

Figure S4 (Continued): **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU13 Chr21 TH.**

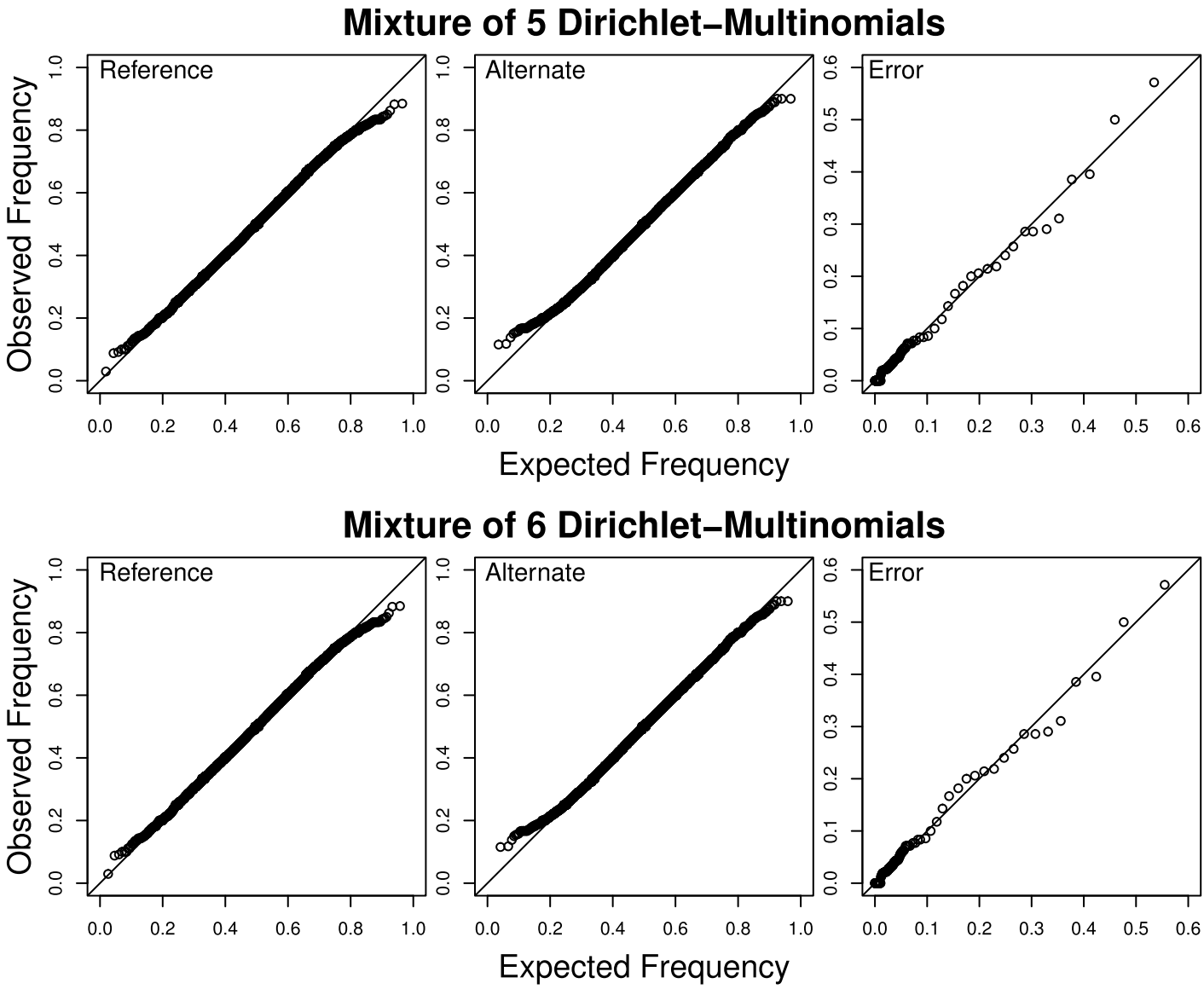

Figure S4 (Continued): **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU13 Chr21 TH.**

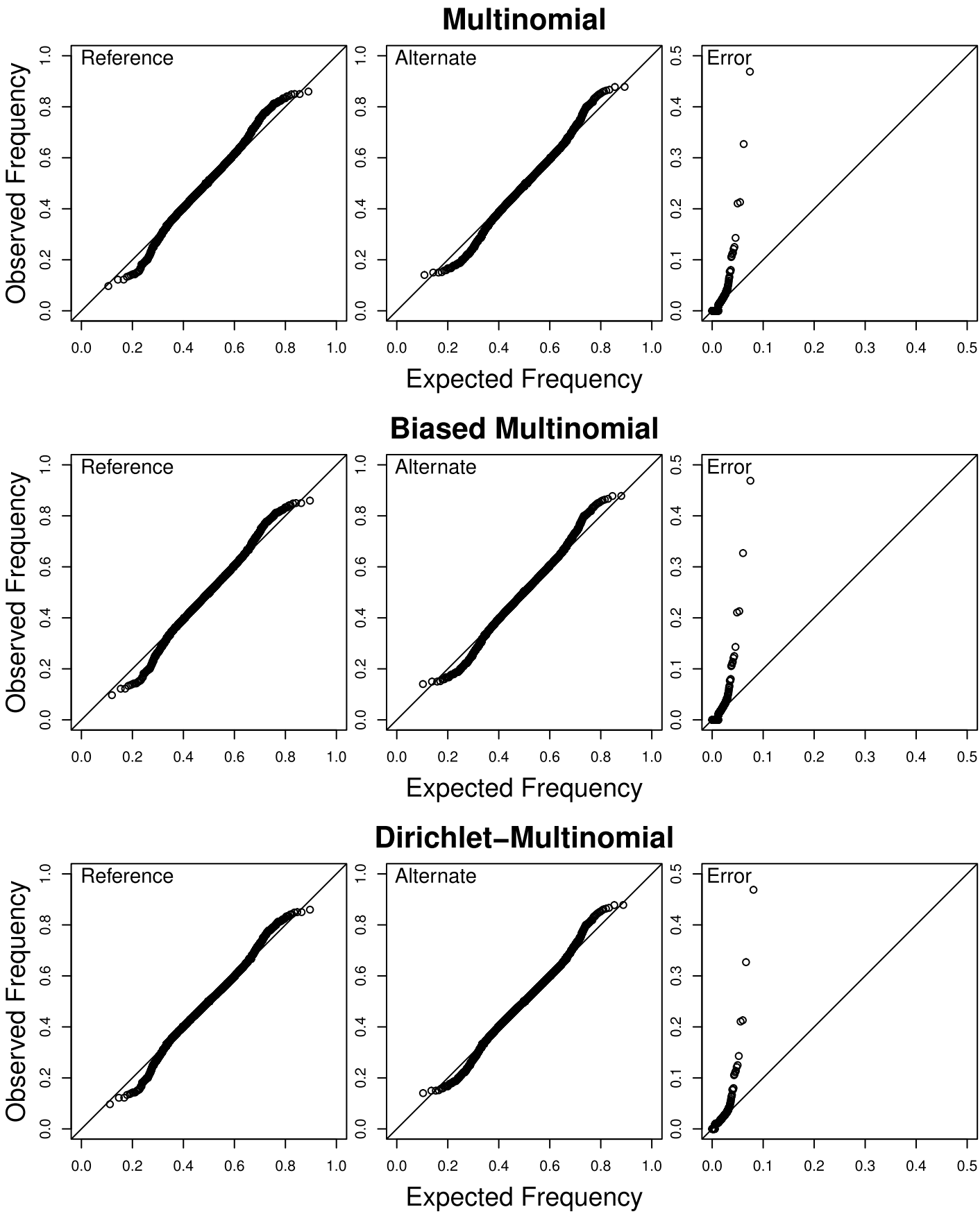

Figure S5: **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU12 Chr10 TH.**

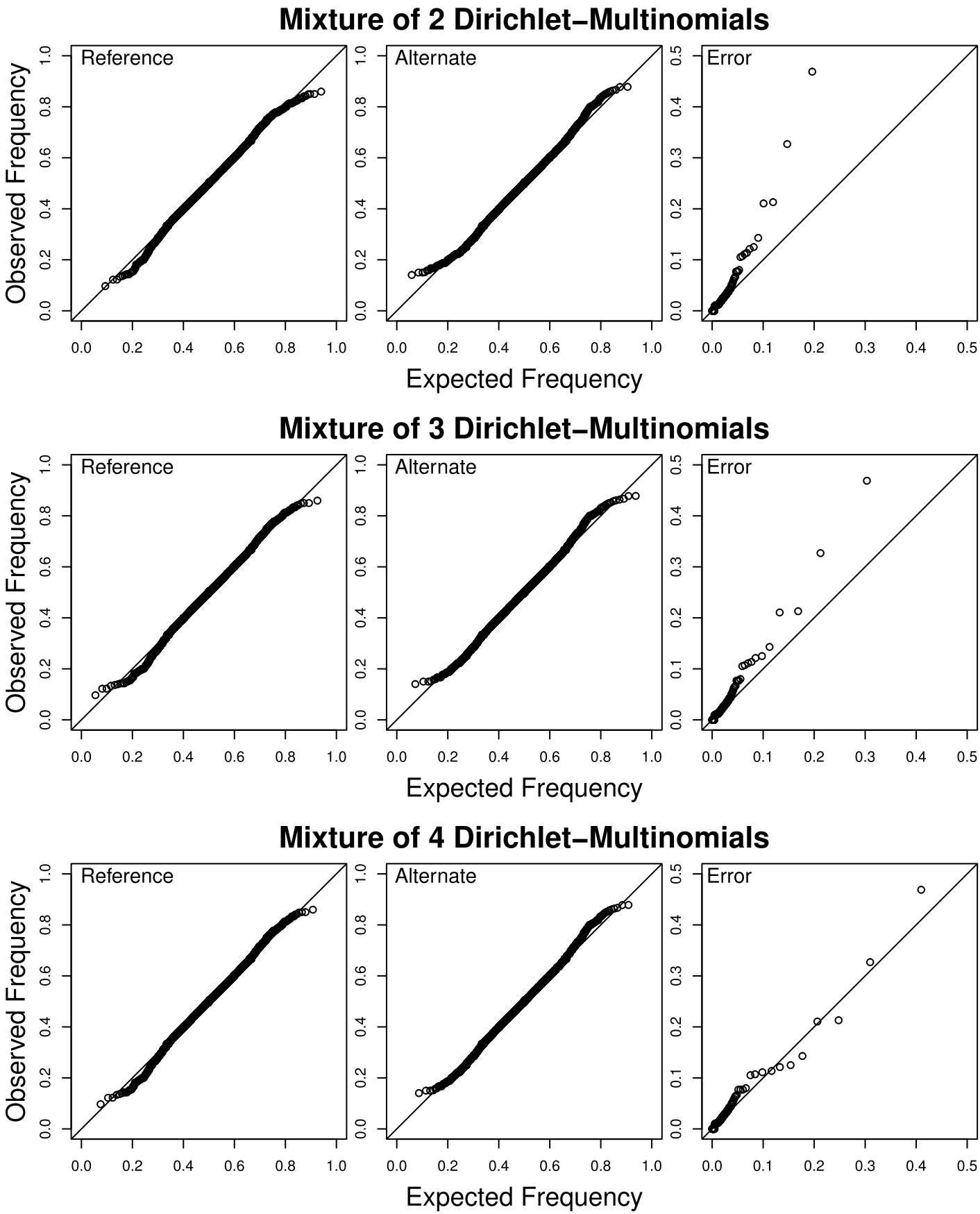

Figure S5 (Continued): **Mixtures of Dirichlet-multinomials provide the best fits to genomic data: CEU12 Chr10 TH.**