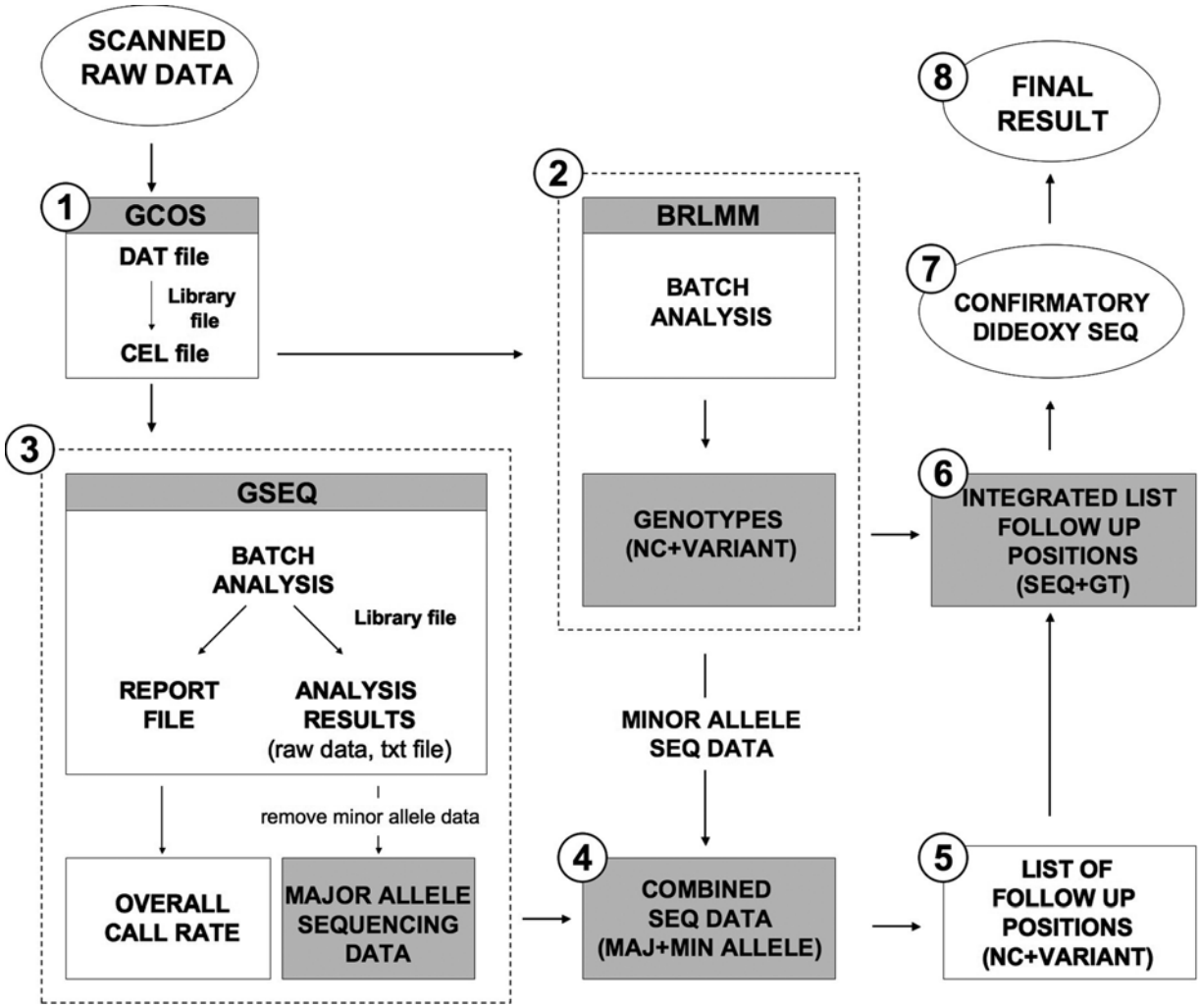

Supplementary Figure. Bioinformatics pipeline. (1) Raw data are initially processed with **GCOS** (Gene Chip Operating system). The resultant .cel file is routed into two separate pipelines, which later converge to give the final list of positions for confirmatory Sanger sequencing follow up. **(2)** Genotyping pipeline. BRLMM **(**Bayesian Robust Linear Modeling using Mahalanobis distance): A batch analysis is performed with a pre-existing set of 40 experiments, which are required for accurate clustering. All homozygous or heterozygous calls in addition to all no-calls are used to trigger the use of the "minor allele sequencing data". **(3)**

GSEQ (Gene Chip Sequence Analysis Software): A batch analysis using a pre-existing set of 20 data sets is used to derive the overall call rate as well as the overall sequencing result. A custom script removes all minor allele sequencing data to minimize unnecessary confirmatory follow up. **(4)** All relevant minor allele sequencing data (triggered by the presence of heterozygous or homozygous variants) is added back after integration with the genotyping calls generated by BRLMM to generate a combined set of sequencing data **(5)**. All No calls (NC) and variant calls derived from the genotyping data as well as the resequencing data sets are integrated **(6)** and routed to confirmatory Sanger sequencing follow up **(7).**