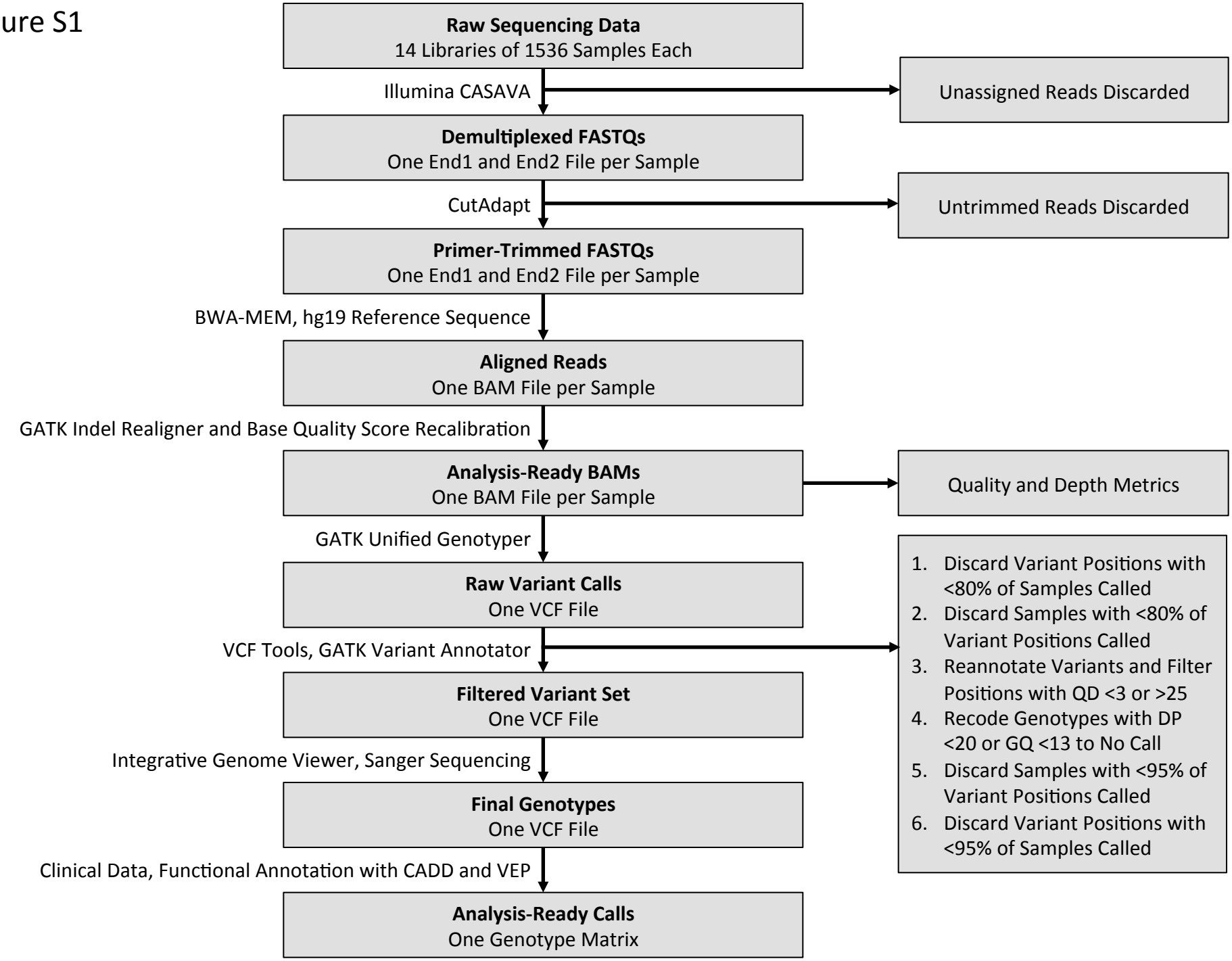
Figure S1


Raw Sequencing Data
14 Libraries of 1536 Samples Each
Illumina CASAVA
Unassigned Reads Discarded
Demultiplexed FASTQs
One End1 and End2 File per Sample
CutAdapt
Untrimmed Reads Discarded
Primer-Trimmed FASTQs
One End1 and End2 File per Sample
BWA-MEM, hg19 Reference Sequence
Aligned Reads
One BAM File per Sample
GATK Indel Realigner and Base Quality Score Recalibration
Analysis-Ready BAMs
One BAM File per Sample
Quality and Depth Metrics
GATK Unified Genotyper
Raw Variant Calls
One VCF File
VCF Tools, GATK Variant Annotator
1. Discard Variant Positions with <80% of Samples Called
2. Discard Samples with <80% of Variant Positions Called
3. Reannotate Variants and Filter Positions with QD <3 or >25
4. Recode Genotypes with DP <20 or GQ <13 to No Call
5. Discard Samples with <95% of Variant Positions Called
6. Discard Variant Positions with <95% of Samples Called
Filtered Variant Set
One VCF File
Integrative Genome Viewer, Sanger Sequencing
Final Genotypes
One VCF File
Clinical Data, Functional Annotation with CADD and VEP
Analysis-Ready Calls
One Genotype Matrix