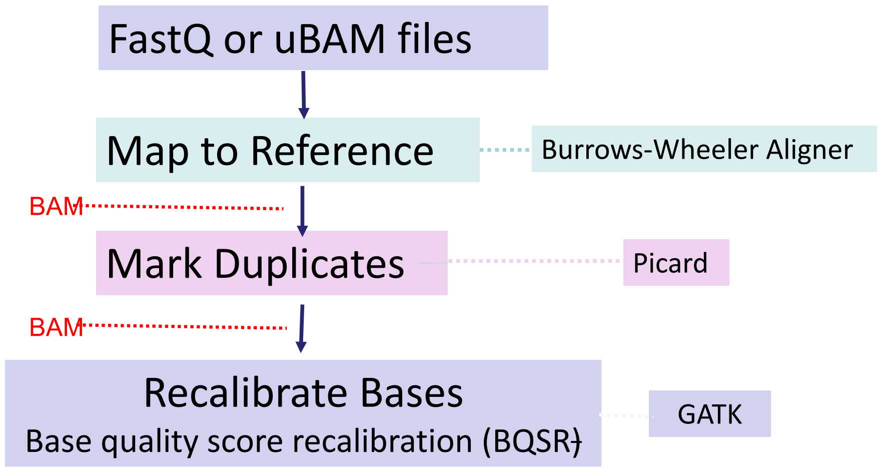
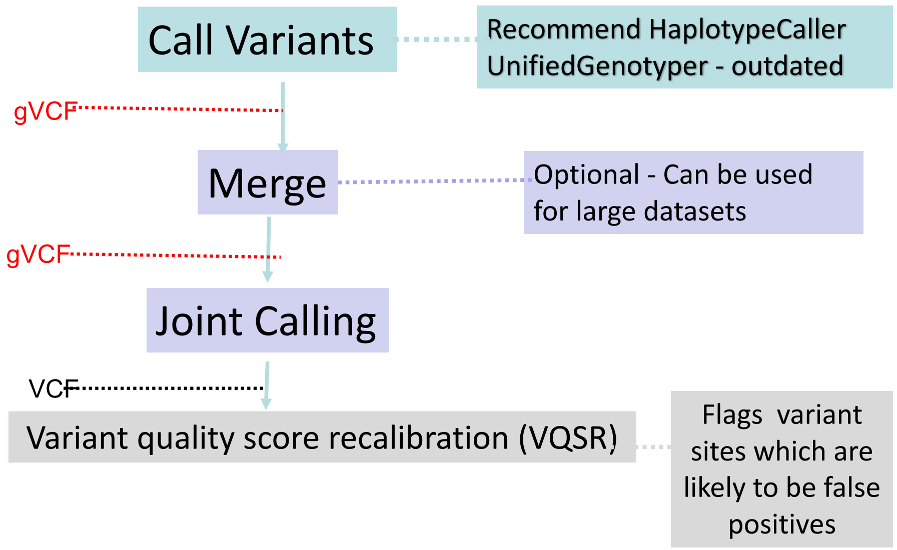
NGS Data Quality Control

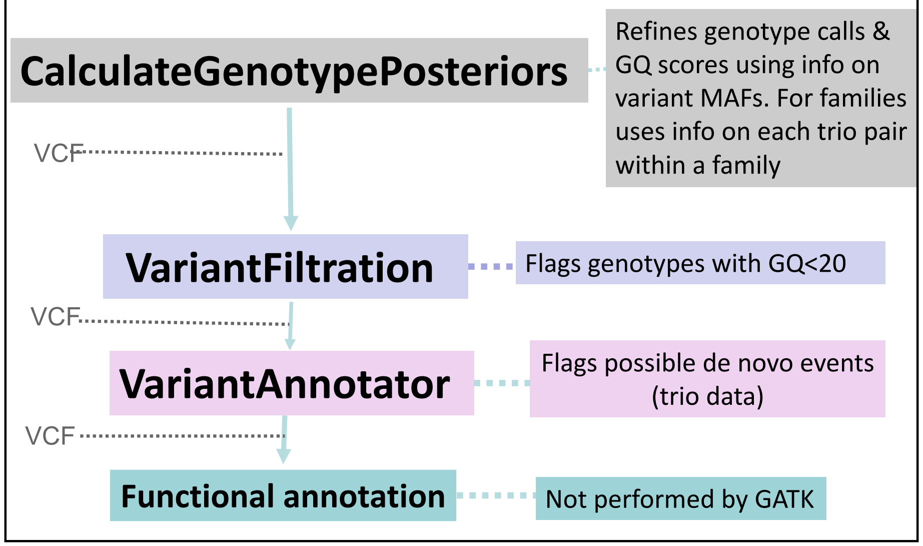
Variant Calling Pipeline -Step 1 Preprocessing



Variant Calling Pipeline-Step 2 Variant Discovery


Call Variants
Recommend HaplotypeCaller
UnifiedGenotyper - outdated
gVCF
Merge
Optional - Can be used for large datasets
gVCF
Joint Calling
VCF
Variant quality score recalibration (VQSR)
Flags variant sites which are likely to be false positives

Variant Calling Pipeline - Step 3 Call Set Refinement


CalculateGenotypePosteriors
Refines genotype calls & GQ scores using info on variant MAFs. For families uses info on each trio pair within a family
VCF
VariantFiltration
Flags genotypes with GQ<20
VCF
VariantAnnotator
Flags possible de novo events (trio data)
VCF
Functional annotation
Not performed by GATK

Variant Calling

- BAM files are large and take considerable resources
  - Storage is expensive
  - One 30x whole genome is ~80-90 gigabytes
  - A small study of 1,000 samples will consume 80 terabytes of disk space
- The cost of cloud computing to call variants
  - (Souilmi et al. 2015)
  - $5 per exome
  - $50 per genome
    - For 1,000 samples
      - $5,000 exome
      - $50,000 genome

Working with gVCF

- Instead of obtaining VCF files
- Can obtain gVCF files to perform joint calling and the rest of the GATK pipeline
  - A whole genome gVCF
    - ~1 Gigabyte
      - 1/100th the size of a BAM file for one individual
- Need additional information on how variants were called
  - e.g. HaplotypeCaller or UnifiedGenotyper
    - Not valid to use Unified Genotyper

Influences on Sequence Quality

- DNA quality
  - Age of sample
  - Extract method
  - Source of sample
- Sequencing machines (read length)
- Median sequencing depth
- Alignment
- Variant calling method used
  - SNVs and Indels

NGS Data Quality Control

- Extremely important to perform before data analysis
  - Poor data quality can increase type I and II errors
  - Due to inclusion of false positive variant sites or incorrect genotype calls
- Sequence quality can be influenced by
  - DNA quality
  - Sequencing machines (read length)
  - Sequencing depth
  - Alignment
  - Variant Calling
    - SNVs and Indels
- Protocols for data QC are still in their infancy
  - No set protocols for QC
- QC which has to be performed is <u>data specific</u>
  - Dependent on read depth
  - Batch effects
  - Availability of duplicate samples
  - etc

NGS Data Quality – Removal of Genotype Calls and Samples

- Sequence read genotype depth (GD)
  - Concerned if GD is too low or too high*
    - GD too low insufficient reads to call a variant site
    - GD too high can be an indication of copy number variants which can introduce false positive variant calls
      - *Due to down sampling in GATK maximum GD is 250
    - Remove genotypes with low read depth, e.g. GD<8
  - Genotype quality (GQ) score
    - Removal of sites with low genotype quality core, e.g. GQ< 20
- Remove individuals who are missing genotype calls/variant sites, e.g. > 10%
  - To remove individuals with bad quality data who can potentially have incorrect genotype calls
- If using different capture arrays use the intersect of the arrays

NGS Data Quality – Removal of Genotype Calls and Samples

- Removal of sites with missing data
  - e.g. missing > 10% of genotypes
- Removal of "novel" variant sites which only occur in one batch and the alternative allele is observed multiple times or the minor allele frequency (MAF) is high in overall sample
- Removal of sites that deviate from Hardy-Weinberg Equilibrium (HWE)
  - Must be performed by population if the study consists of more than one ancestry group, e.g. African American and European American
  - Related individuals should also be removed from the sample before testing for deviations from HWE

NGS Data Quality Control

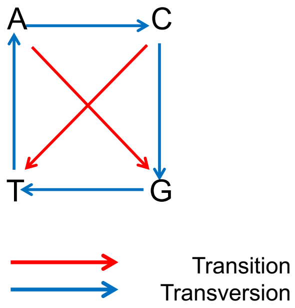
- Variant Quality Score Recalibration (VQSR) or
  - GATK
- Used to determine variant sites of bad quality
- However even after this step
  - Concordance of duplicates (when available) and
  - and Ti/Tv ratios are often low
- Additional QC steps needs to be performed

NGS Data Quality Control

- Values which are used for GD, GQ, and missing data cut offs are based upon
  - Concordance rates
    - if there duplicate samples are available
  - Ti/Tv ratios
    - For individuals
    - Entire sample
  - Removal of batch effects
    - As evaluated by multidimensional scaling (MDS) or
    - Principal components analysis (PCA)
  - Amount of data removed
    - QCl can remove substantial amounts of data which should be avoided
      - e.g. >15% of variant sites

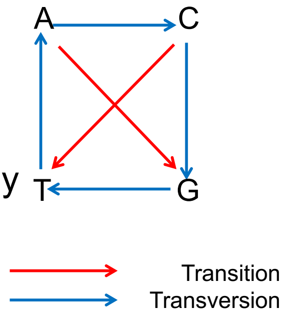
Transition/Transversion (Ti/TV) Ratios

- Transition
  - Purine → Purine
  - Pyrimidine → Pyrimidine
- Transversion
  - Purine → Pyrimidine
  - Pyrimidine → Purine



Transition/Transversion (Ti/TV) Ratios

- Ti/Tv Ratios
  - Whole genome ~2.0
  - Exome novel ~2.7
  - Exome known ~3.5
- Ti/Tv ratios can be calculated by
  - Sample or
  - Dataset
- Ti/Tv ratios can be evaluated for subsets of data
  - e.g. by batch



Example -Project Description

- 1,667 Samples
- Seven cohorts
- Two sequencing centers
  - Center 1
    - Two capture arrays
      - NimbleGen V2Refseq 2010 (CA1): 1082
        - Batch 1 and 3
      - NimbleGen bigexome 2011 (CA2): 234
        - Batch 2
  - Center 2
    - One capture array
      - Agilent SureSelect
        - Batch 4
- Four batches
- No intentional duplicate samples

Example Project Description

- Intersection of the three capture arrays used
  - NimbleGen V2Refseq 2010
    - Batch 1 and 3
  - NimbleGen bigexome 2011
    - Batch 2
  - Agilent Sure Select
    - Batch 4
- Sequencing machine
  - Illumina HiSeq
- Sequence alignment
  - BWA
- Multi-sample variant calling
  - GATK

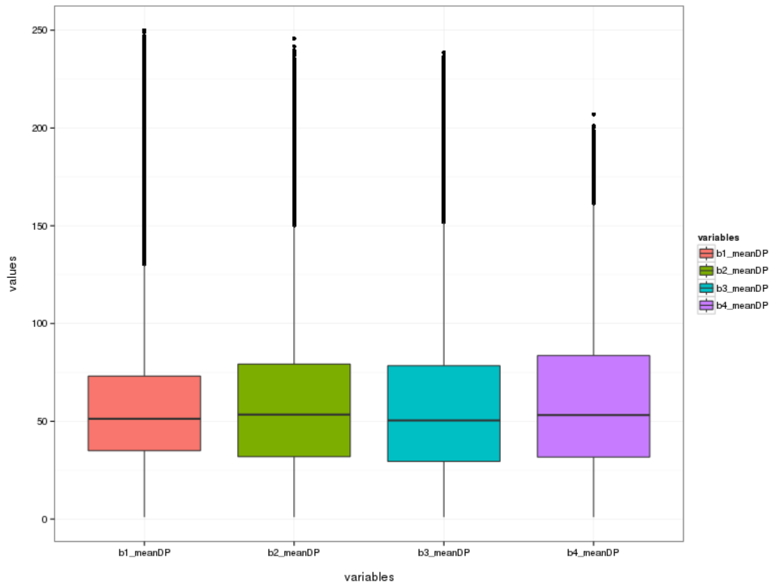
Sequence Data QC Overview

- Variant and genotype call level
  - Evaluation of batch effects
- Genotype call level – Removal of genotype calls
  - Low or high depth of coverage GD < 8
  - Low genotype quality score GQ < 20
- Removal of individual samples
  - >20% missing data
    - After taking the intersect of capture arrays
  - Samples without phenotype information
- Variant level – removal of variant sites
  - Low call rate
    - i.e., missing call rate > 10%
  - "Novel" variant sites observed ≥2 only in a single batch
  - Deviation from Hardy-Weinberg-Equilibrium
    - Population specific
    - Unrelated individuals
      - p<5 x 10^{-8}

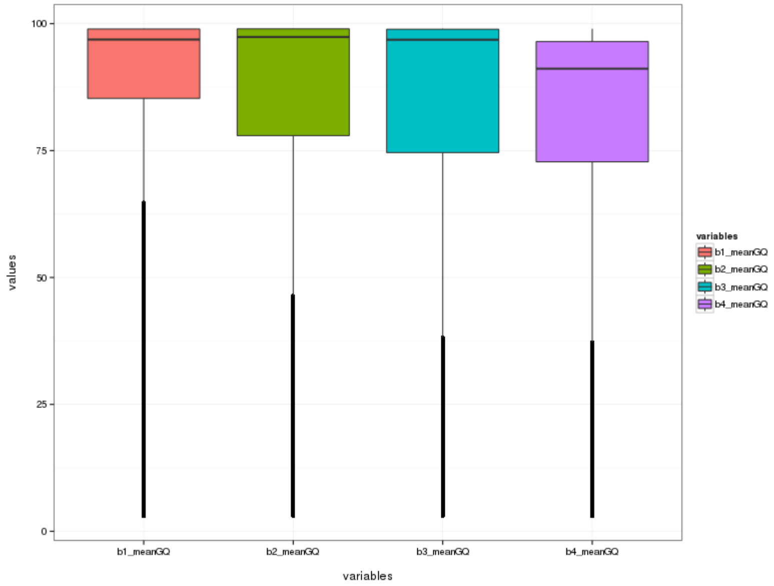
Sequence Data QC Overview

- Detection of sample outliers
  - Perform multidimensional scaling (MDS) to detect outliers
    - Due to population substructure/admixture and batch effects
    - Remove effects by
      - Additional QC
      - Removal of outliers and\or
      - Inclusion of MDS or PCA components in the association analysis
- Evaluate sex of individuals based upon X and Y chromosomal data
  - Sample mix-ups
  - Individuals with Turner or Klinefelter Syndrome
- Evaluate samples for cryptically related individuals and duplicates
  - King or Plink algorithm
    - Retain one duplicate of a pair
    - Retain only one individual of a relative group or control for relatedness in the analysis, i.e. mixed models
- Post Analysis - Quantile-Quantile (QQ)plots
  - To evaluate uncontrolled batch effects and population substructure/admixture

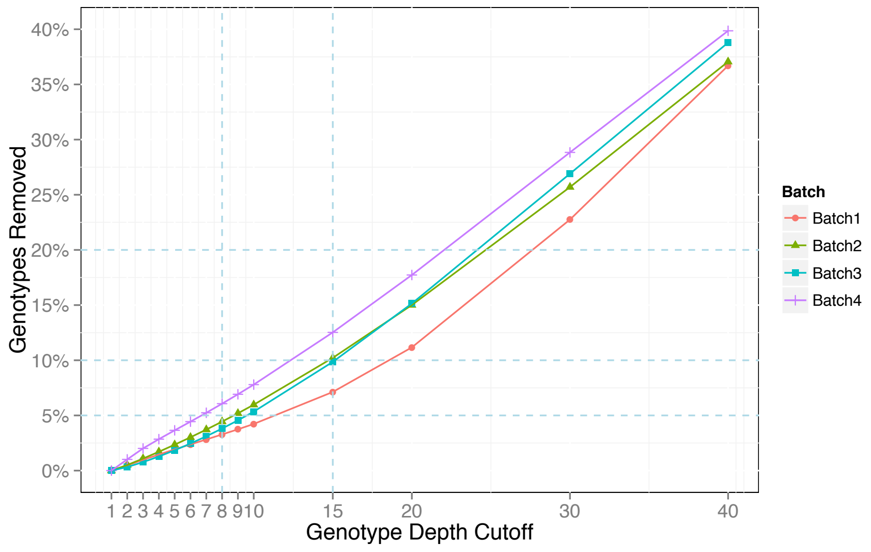
Mean DP by Batch



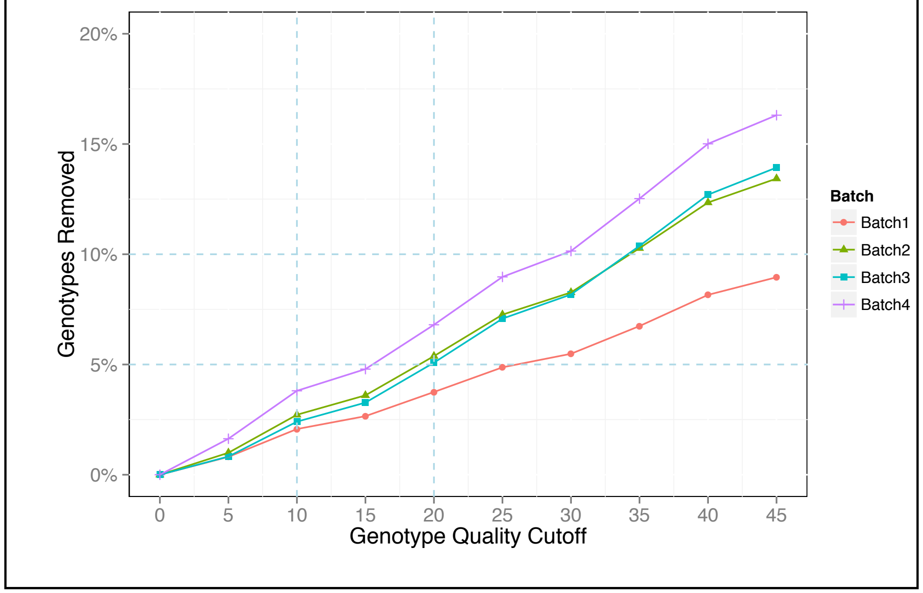
Mean GQ by Batch



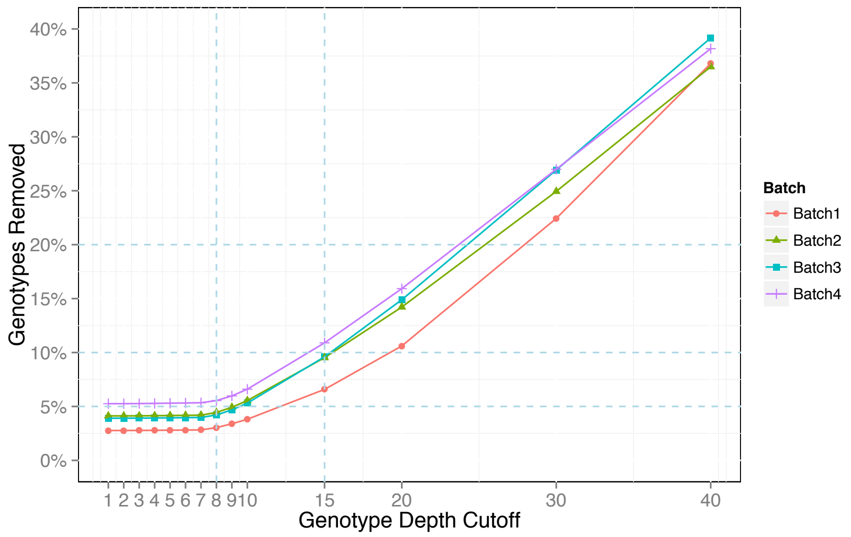
Genotypes Removed by GD Cut-off by Batch



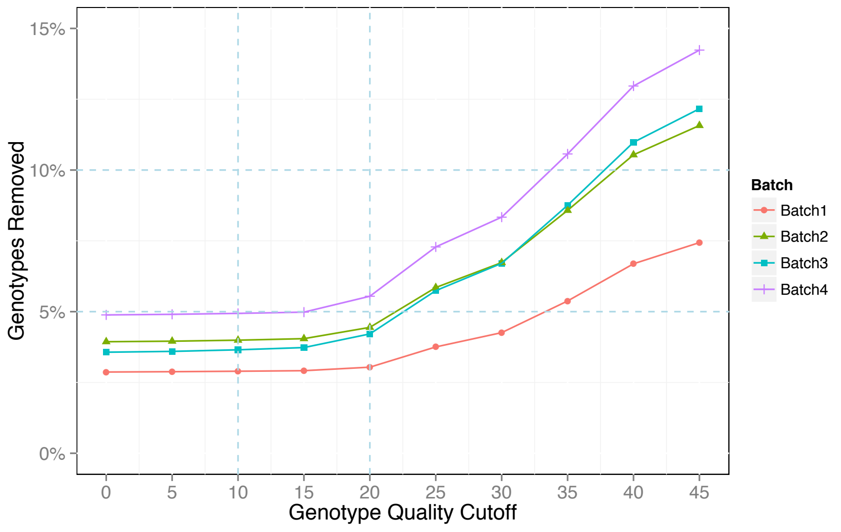
Genotypes Removed by GQ Cut-offs by Batch



Genotypes Removed by GD Cut-off by Batch (First removing genotypes with GQ < 20)



Genotypes Removed by GQ Cut-offs by Batch (First removing genotypes with a GD≤8)



Missing Rate Criteria & Sites Removed

	10%	5%
Before QC*	2.5%	3.9%
After QC	12.9%	18.3%

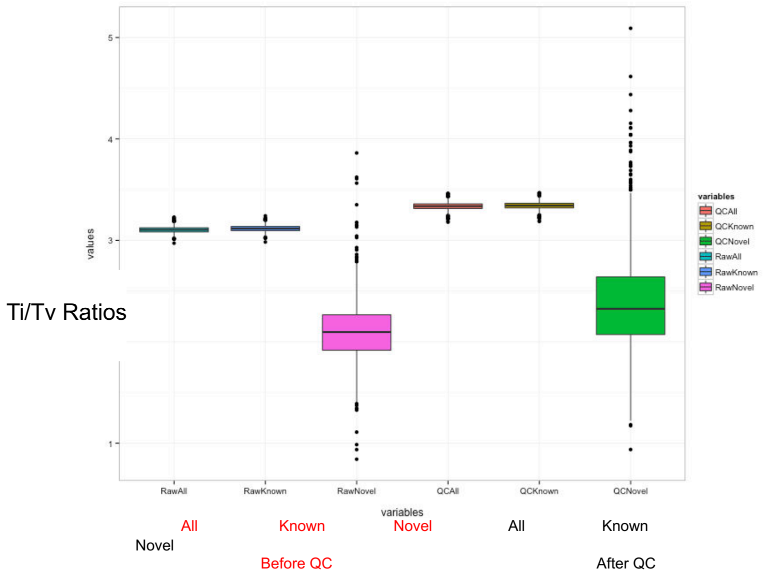
*After VQSR

Variant sites missing >10% of their data were removed

Ti/Tv Ratios during QC Process

	Known	Novel	All
Before VQSR	2.95 ± 0.05	1.18 ± 0.29	2.86 ± 0.07
Before QC	3.12 ± 0.03	2.01 ± 0.32	3.11 ± 0.03
Genotype QC GD<8, GQ <20	3.18 ± 0.04	2.10 ±0.32	3.16 ± 0.03
Remove sites missing >10% genotypes	3.39 ± 0.04	2.42 ± 0.52	3.39 ± 0.04
Remove batch specific novel sites ≥ 2 N=17,835	3.39 ± 0.04	2.41 ± 0.53	3.39 ± 0.04
Remove sites deviating from HWE $p<5x10^{-8}$ N=4,414	3.41 ± 0.04	2.39 ± 0.54	3.40 ± 0.04

Ti/Tv Ratios by Individual Before and After QC



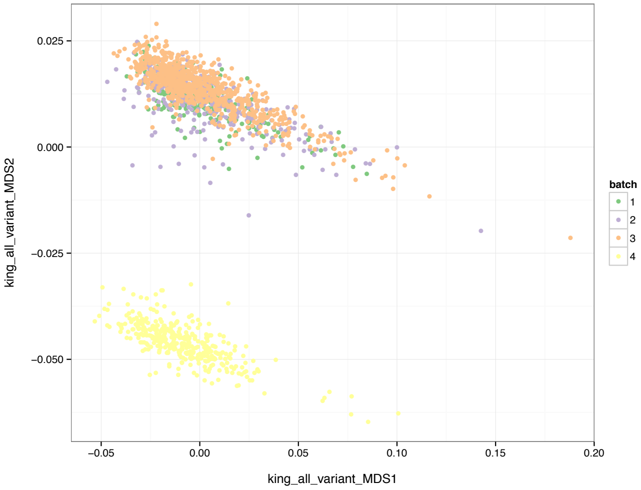
Detecting Outliers Using Multidimensional Scaling (MDS)

- Multidimensional Scaling (MDS) and principal components analysis (PCA) are frequently used to detect outliers
  - MDS & PCA can also be used to control for substructure in the analysis
- Outliers can be cause by
  - Population stratification
  - Population substructure
  - Batch Effects

Sequence Data QC

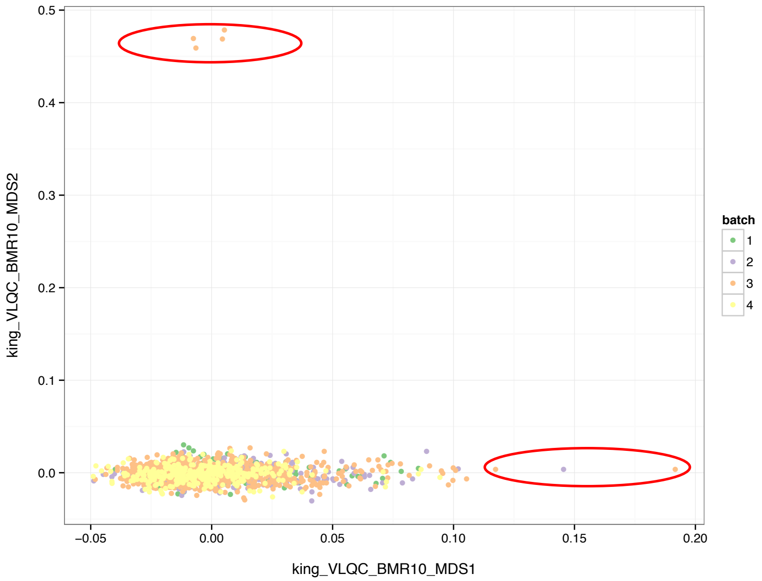
- Batch effects can sometimes be removed with additional QC
- Extreme outliers should be removed
- Additionally MDS or PCA components can be included in the analysis to control for population substructure\admixture and batch effects
  - Unless correlated with the outcome (phenotype)
- Batch effects (dummy coding) may be included as a covariate in the analysis
  - Unless correlated with the outcome (phenotype)

MDS First 2 Components Before QC*



*After VQSR

MDS First 2 Components After QC


king_VLQC_BMR10_MDS2
king_VLQC_BMR10_MDS1
batch