Data Quality Control
NGS and Genotype Array Data

Suzanne M. Leal, Ph.D.
Sergievsky Family Professor of Neurological Sciences
Director of the Center for Statistical Genetics
Columbia University
sml3@Columbia.edu

1

DNA Collection

- Blood samples
  - For unlimited supply of DNA
    - Transformed cell lines
      - Is expensive
    - Whole genome amplification
      - Allows for the creation of large amounts of DNA from initial small DNA sample
        - » Perform WGA on each sample three or more times and use pooled samples
      - Can experience lower call rates and higher genotyping error rates
      - Not recommend for whole genome sequencing or copy number variant (CNV) analysis
- Buccal Swabs
  - Small amounts of DNA
  - DNA not stable
- Saliva (Origene collection kit)

Measurement of DNA Concentrations

- Nanodrop
- Picogreen

2

Effect of Genotyping Error – Same Error Rates for Cases and Controls

- For family-based association studies - Trios
  - Can increase both type I and II error
- Population based studies
  - Increases type II error only

Quantitative Traits
If genotyping error is not correlated with trait values type II errors will be increased

3

Effects of Genotyping Error – Different Error Rates for Cases and Controls

- Cases and controls are sequenced/genotyped
  - At different times
  - Different institutions
  - Or one group, e.g., case or control, is predominately sequenced/genotyped in the same batch
- Can lead to different genotyping error rates in cases and controls
  - In this situation both type I and II error can be increased
- If sequencing/genotyping cases and controls
  - Randomize cases and controls so they are spread evenly across batches

Quantitative Traits
If genotyping error is correlated with trait values, it will also increase type I and II errors, e.g., individuals with elevated systolic blood pressure are genotyped in one batch and those with systolic blood pressure within the normotensive range in another batch

4

Genotype SNPs (~20-96) before Exome or Whole Genome Sequencing

- Genotype markers which can be used as DNA fingerprint
- Allows for Assessment of DNA quality
- Aids in determining the the genetic sex of study subjects
  - To aid in identification of potential sample swaps
- Detects cryptic duplicates
- For family data
  - Aids in determining close familial relationships
    - Non-paternity
    - Sample swaps
    - Cryptic relationships

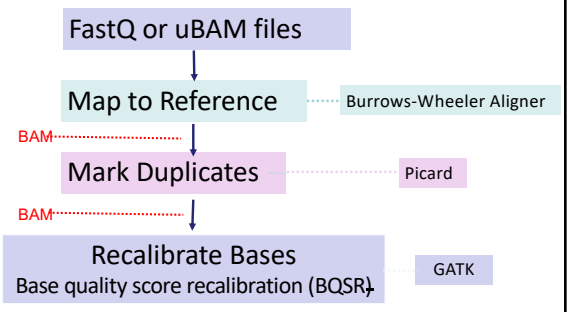
5

Detecting Genotyping Errors

- Duplicate samples genotyped using arrays to detect inconsistencies
  - Can use duplicate samples that are inconsistent to adjust clusters to improve allele calls
    - Will not detect systematic errors
- Usually generated only for genotype array data
  - Due to expense, duplicate samples are usually not generated for exome or whole genome sequencing studies

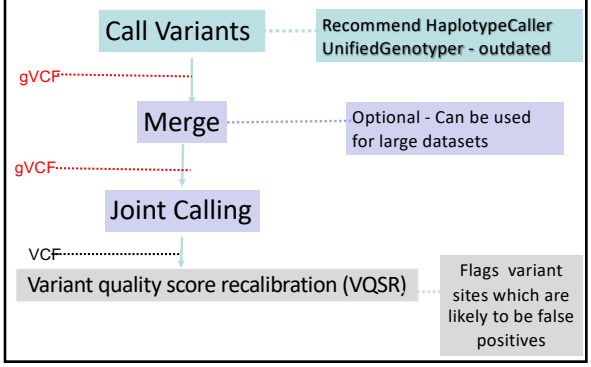
6

Variant Calling Pipeline -Step 1 Preprocessing



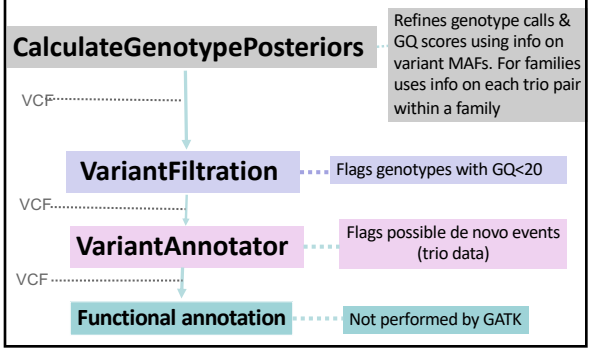
7

Variant Calling Pipeline-Step 2 Variant Discovery



8

Variant Calling Pipeline - Step 3 Call Set Refinement



9

A Short List of Additional Software to Detect Genetic Variation

- Exome data copy number variation
  - CoNIFER (Copy Number Inference From Exome Reads)
    - Krumm et al. 2012
  - XHMM
    - Fromer et al. 2014
- WGS data structural variation
  - MetaSV
    - Mohiyuddin et al. 2015
  - LUMPY
    - Layer et al. 2014

10

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

11

Working with gVCF Files

- Instead of obtaining VCF files
- Can obtain gVCF files to perform joint calling and complete the GATK pipeline
  - A whole genome gVCF
    - ~1 Gigabyte
      - 1/100th the size of a BAM file for one individual

12

Influences on Sequence Quality

- DNA quality
  - Age of sample
  - Extraction method
  - Source of sample
    - e.g., blood, skin punch, buccal
- Sequencing machines (read length)
- Median sequencing depth
- Alignment
- Variant calling method used
  - Single nucleotide variants and insertion/deletions
  - Structural variants

13

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
- Protocols for data QC are still in their infancy
  - No set protocols for QC
- QC is data specific
  - Dependent on read depth
  - Batch effects
  - Availability of duplicate samples
  - etc.

14

NGS Data Quality – Removal of Genotype Calls and Samples

- Sequence depth of coverage
  - DP_variant
    - High DP could be an indication of copy number variants
      - Which can introduce false positive variant calls
        - » Due to down sampling in GATK maximum DP is 250
  - DP_genotype
    - Concerned if depth is too low or too high
      - Low insufficient reads to call a variant site
    - Remove genotypes with low read depth, e.g., DP≤8
- Genotype quality (GQ) score
  - Removal of sites with low genotype quality core, e.g., GQ≤ 20

15

NGS Data Quality – Removal of Genotype Calls and Samples

- Sequence depth of coverage
  - DP_variant
    - High DP could be an indication of copy number variants
      - Which can introduce false positive variant calls
        - » Due to down sampling in GATK maximum DP is 250
  - DP_genotype
    - Concerned if depth is too low or too high
      - Low insufficient reads to call a variant site
    - Remove genotypes with low read depth, e.g., DP≤8
- Genotype quality (GQ) score
  - Removal genotypes with a low genotype quality core, e.g., GQ≤ 20

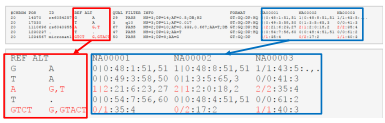
16

VCF Example

17

Variants with more than 2 Alleles

- Genetic analysis tools are usually developed to analyze variant sites that are diallelic
- Some sites may have >2 alleles
- The alleles at these sites need to be split
  - New loci are made each multi-allelic site each with only 2 alleles
    - bcftools
- Multiallelic sites can have higher error rates compared to diallelic sites



18

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, e.g., African American and European American
  - Related individuals should be removed from the sample before testing for deviations from HWE

19

NGS Data Quality Control

- GATK - Variant Quality Score Recalibration (VQSR)
  - Used to determine variant sites of bad quality
    - Variant site is a false positive call
- 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

20

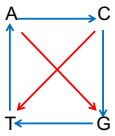
NGS Data Quality Control

- Values which are used for DP (genotype), GQ, and missing data cut offs are based upon
  - Concordance rates
    - If there are duplicate samples are available
  - Ti/Tv ratios
    - By individual
    - By batch
    - Entire data set
  - Amount of data removed
    - QC can remove substantial amounts of data which should be avoided
      - e.g., >15% of variant sites

21

Transition/Transversion (Ti/TV) Ratios

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

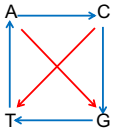


Transition
Transversion

22

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



Transition
Transversion

23

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 DP< 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

24

Sequence Data QC Overview

- 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
      - e.g., $p<5 \times 10^{-8}$, $p<5 \times 10^{-15}$

25

Data Clean – Assessing Sex Chromosomes

- When data is collected on study subjects they are asked about their gender/sex and not their genetic sex
  - Differences in gender/sex and genetic sex can be due to
    - Sample swaps
    - Study subjects who are not cisgender
- Some study subjects may have neither a XX nor XY karyotype
  - Turner syndrome X0
  - Klinefelter syndrome XXY

26

Data Clean – Assessing Chromosomal Sex

- Study subjects labeled as females with an excess of homozygous genotypes on the X chromosome can denote
  - That their genetic sex is male
  - Turner Syndrome

27

Data Clean – Assessing Chromosomal Sex

- Study subjects labeled as males with an excess of heterozygous SNPs* on the X chromosome can denote
  - That their genetic sex is female
  - Klinefelter syndrome
- Note: Individuals who are XY will also be heterozygous for markers in the pseudoautosomal regions
- Availability of Y chromosome data
  - Can greatly aid in determining genetic sex and if an individual has Turner or Klinefelter syndrome

*Both genetic males and females have two alleles for each locus on the X chromosome in the datafile, although males are hemizygous

28

Data Clean – Assessing Sex Chromosomes

- Individuals whose labeled gender/sex does not match their genetic sex are removed from the analysis
- This observation may be due to a sample swap
  - When samples are swapped
    - Phenotype data will be incorrect
      - e.g., may be a case when labeled as a control

29

Checking for Duplicate and Related Individuals

- Duplicate samples are sometimes included in a study as part of quality control to detect inconsistencies
  - Will not detect systematic errors
  - Usually not included in exome and whole genome sequencing studies
  - Intentional duplicates can easily be removed before data quality control
- Cryptic duplicates (unintentional)
  - DNA sample aliquoted more than once
  - Individual ascertained more than once for a study
    - e.g. The same individual undergoes the same operation more than once and is ascertained each time
- Individuals who are related to each other may participate in the same study
  - Unknown to the investigator
  - Or be part of the study design

30

Duplicate and Related Individuals Need to be Identified

- For duplicate samples
  - Only one can be retained
- For related individuals
  - PCA is performed first with unrelated individuals and related individuals are then projected onto the PCs of unrelated individuals
  - Mixed-models need to be used to analyze the data if related individuals are included*
    - Case-Control
      - Generalized linear mixed models (GLMM)
    - Quantitative traits
      - Linear mixed models (LMM)
  - If not type I error rates can be increased

*If only a few related individuals in sample, may wish to remove them or use LMM/GLMM to control type I errors. Must use LMM/GLMM if related individuals are included in the dataset. If possible, opt for LMM/GLMM since it can help to control type I error due to other types of structure in the data, even when no closely related individuals are included in the analysis.

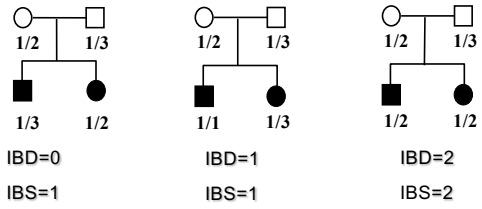
31

Identifying Duplicate and Related Individuals

- Duplicate and related individuals can be detected
  - By examining **Identity-by-State** (IBS) adjusted for allele frequencies (p-hat) between all pairs of individuals within a sample
  - Identify-by-descent (IBD) sharing can be estimated

32

Identity by Descent (IBD)/Identity-by-State (IBS)



33

IBD Sharing Estimated Pairwise for all Individuals in a Samples

- PLINK (Purcell et al. 2007)
- Uses sequence (or genotype array) data to check IBD
  - Prune markers to remove those in LD
    - e.g., r^2<0.1
- P-hat is calculated using the "population" allele frequency
- Used to approximates IBD sharing
- IBD is the number of alleles of alleles which are shared between a pair of individuals
  - Can either share 0, 1, and 2 alleles

34

Identifying Duplicate and Related Individuals

- Monozygote twins and duplicate samples will share 100% of their alleles IBD
  - IBD=2 is 1.0 (can be lower due to genotyping error)
- Siblings and child-parent pairs will share 50% of their alleles IBD
  - For parent-child IBD=1 is 1.0 (IBD=0 is 0 & IBD=2 is 0)
  - For sibs IBD=1 is ~0.50 (IBD=0 is ~0.25 & IBD=2 is ~0.25)
    - For more distantly related individuals the IBD measure will be lower

35

Identifying Duplicate and Related Individuals

- KING [Kinship-based INference for Gwas (*Manichaikul et al. 2010*)] can also be used to identify duplicate and related individuals
  - KING is more robust to population substructure and admixture
    - Prune markers for LD (e.g., r^2<0.1)
  - Provides kinship coefficients
    - Duplicate samples
      - Kinship coefficient equals 0.5
    - Siblings
      - Kinship coefficient equals 0.25

36

UK Biobank Related Individuals > Kinship Coefficient 0.0625

White European

# of Relatives	# of relatives
1	86089
2	18491
3	3691
4	707
5	165
6	40
7	9
8	5
9	1
10	11
11	2
12	2
16	1
19	1
25	1
30	1
3985	1

African

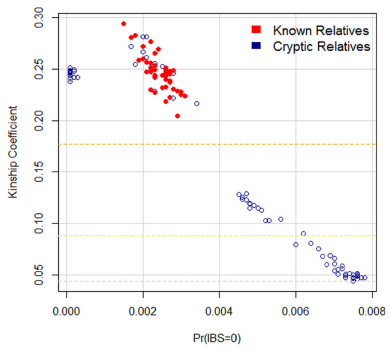
# of relatives	# of individuals
1	715
2	153
3	26
4	10
5	3
6	5
7	5
8	4
9	1
10	4
11	2
13	3
17	2
19	3
20	2
21	1
23	1
.	.
.	.
.	.
390	1
391	1
393	1
396	1

Asian

# of relatives	# of individuals
1	743
2	115
3	33
4	4
5	4

37

King Graphical Output



38

Multiple Individuals observed that are distantly "Related"

- If individuals in sample come from different populations
  - e.g., individuals from the same population within the sample will have inflated p-hat values due to incorrect allele frequencies
    - Incorrectly appear to be related to each other
- "Relatedness" amongst many individuals can also be observed when batches are combined if they have different error rates
  - Individuals from the same batch appear to be related
- DNA contamination can cause "relatedness" between multiple individuals

39

Principal Components Analysis (PCA) / Multidimensional Scaling (MDS)

- Can be used to identify outliers
- Population substructure
  - Individuals from different ancestry
    - e.g., African American samples included in samples of European Americans
- Batch effects
- Use a subset of markers which have been LD pruned
  - Only very low levels of LD between marker loci
    - e.g., $r^2<0.1$
  - MAF cutoff dependent on sample size
    - e.g MAF> 0.01
      - Can use lower MAF for large sample sizes

40

Principal Components Analysis (PCA) / Multidimensional Scaling (MDS)

- Unrelated individuals are used to generate PC plots
  - Related individuals are projected onto to the PC plots
- Plot 1st component vs. 2nd component
  - Additional PCs should also be plotted
    - e.g.. PCs 1-10
- Mahalanobis distance can be used to determine outliers
  - e.g., <1

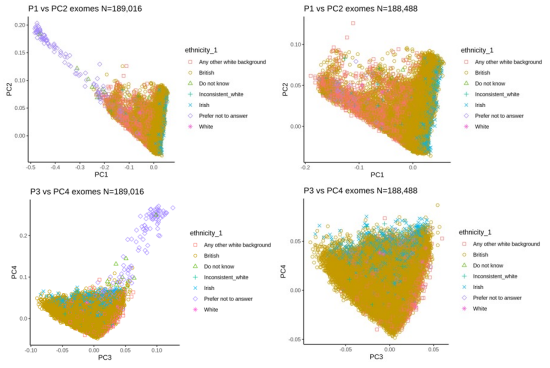
41

PCA/MDS Can be Used to Identify Outliers

- Individuals of different ancestry
  - e.g., African American samples included with European Americans samples
  - Can use samples from HapMap/1000 genomes to help to determine the ancestry for samples that are outliers
    - Should not include HapMap/1000 genomes samples when calculating components to control for population substructure/admixture
- Batch effects

42

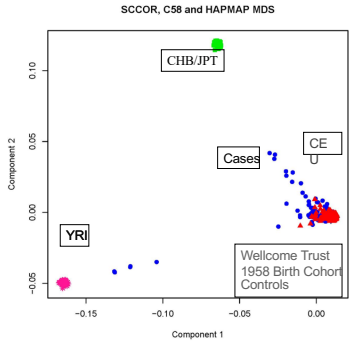
Principal Components Analysis Example



Exclusion of Outliers using Mahalanobis distance (0.997)

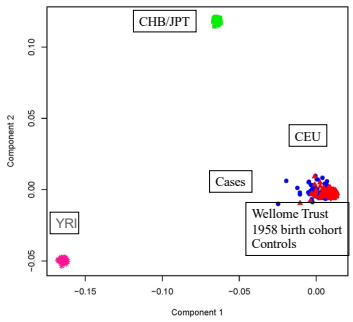
43

Detecting Outliers Using PCA and HapMap Sample



44

Detecting Outliers Using PCA and HapMap Sample



45

Detecting Genotyping Error – Examining HWE

- Testing for deviations from HWE not very powerful to detect genotyping errors
- The power to detect deviations from HWE dependent on:
  - Error rates
  - Underlying error model
    - Random
    - Heterozygous genotypes -> homozygous genotypes
    - Homozygous genotypes ->Heterozygous genotype
  - Minor allele frequencies (MAF)

46

Detecting Genotyping Error – Examining HWE

- Controls and Cases are evaluated separately
  - Deviation found only in cases can be due to an association
- Test for deviation from HWE only in samples of the same ancestry
  - Population substructure can introduce deviations from HWE
- Do not include related individuals when testing for deviations from HWE
  - Can cause deviations from HWE

47

Detecting Genotyping Error – Examining HWE

- What criterion is used to remove variants due to a deviation from HWE
  - GWAS studies have used 5.0×10^{-7} to 5.0×10^{-15}
- Quantitative Traits
  - Caution should be used removing markers which deviate from HWE may be due to an association
    - Remove markers with extreme deviations from HWE and Flag markers with less extreme deviations from HWE
- When performing imputation need to be more stringent in removing variants which deviate from HWE

48

Sequence Data QC Overview

- Remove variant sites that fail VQSR
- Remove genotypes with low DP, GQ scores, etc.
- Remove variant sites with large percent of missing data
- Remove samples with missing large percent of missing data
- Evaluate genetic sex of individuals based upon X and Y chromosomal data
  - Sample mix-ups
  - Individuals with Turner or Klinefelter Syndrome

49

Sequence Data QC Overview

- Evaluate samples for cryptically related individuals and duplicates
  - Use variants which have been pruned for LD
    - e.g., $r^2<0.1$
  - King or Plink algorithm
    - Always remove duplicate individuals
      - Retaining only one in the sample
    - If sample includes related samples use linear mix models (LMM)/Generalized LMM (GLMM) to control for relatedness
      - Best to perform even for data without related individuals
    - If only a few related individuals can retain only one individual of a relative group if not using LMM or GLMM

50

Sequence Data QC Overview

- Detection of sample outliers
  - Perform principal components analysis (PCA) or multidimensional scaling (MDS) to detect outliers
    - Use variants pruned for LD
      - e.g., $r^2<0.1$
    - Use unrelated individuals and then project related individuals onto the PCs
- Due to population substructure/admixture and batch effects
- Remove effects by
  - Additional QC
  - Removal of outliers (can be determined by Mahalanobis distance) and\or
  - Inclusion of MDS or PCA components in the association analysis

51

Sequence Data QC Overview

- Remove/flag variant sites that deviate from HWE in controls
  - HWE should be only be tested in unrelated individuals from the same population
- Post Analysis - Quantile-Quantile (QQ) plots
  - To evaluate uncontrolled batch effects and population substructure/admixture

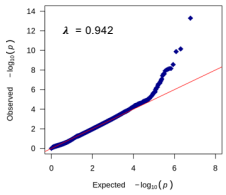
52

QQ Plots - Genome Wide Association Diagnosis

- Thousands of variants/genes are tested simultaneously
- The p-values of neutral markers follow the uniform distribution
- If there are systematic biases, e.g., population substructure, genotyping errors, there will be a deviation from the uniform distribution
- QQ plots offers an intuitive way to visually detect biases
- Observed p-values are ordered from largest to smallest and their $-\log_{10}(p)$ values are plotted on the y axis and the expected $-\log_{10}(p)$ values under the null (uniform distribution) on the x axis

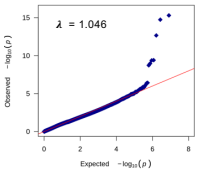
53

QQ Plot of Exome Wide P Values UK Biobank 200K



Hearing aid users

Case N= 6,436
Controls N= 96,601



Problem hearing with background noise

Cases N=65,660
Controls N= 96,601

54

Genomic Inflation Factor to Evaluate Inflation of the Test Statistic

- Genomic Inflation Factor (GIF): ratio of the median of the test statistics to expected median and is usually represented as λ
  - No inflation of the test statistic λ=1
  - Inflation λ>1
  - Deflation λ<1
    - Can be observed when a study is underpowered
- Problematic to examine the mean of the test statistic
  - Can be large if many variants are associated
    - Particularly if they have very small p-values
    - Should not be used

55

Phenotype	Covariate	Mean Chi-Square	GIF (λ)
BP		1.23829	1.16932
BP	Age	1.24119	1.18025
BP	**Age-EV1**	1.09471	**1**
BP	Age-EV2	1.0881	1
BP	Age-EV4	1.08385	1
BP	Age-EV10	1.09582	1.00402
BPI		1.14931	1.08921
BPI	Age	1.15139	1.08113
BPI	Age-EV1	1.05079	1.01148
BPI	**Age-EV2**	1.0428	**1**
BPI	Age-EV4	1.04204	1
BPI	Age-EV10	1.05421	1.01724
BPII		1.17283	1.25664
BPII	Age	1.17583	1.26996
BPII	Age-EV1	1.09874	1.15065
BPII	Age-EV2	1.09904	1.16425
BPII	Age-EV4	1.09502	1.14609
BPII	Age-EV10	1.10046	1.1418
BPII	Sex,Age-EV1	1.05958	1.06424
BPII	**Sex,Age-EV4**	1.05817	**1.05323**
BPII	Sex,Age-EV10	1.06338	1.05581

56

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

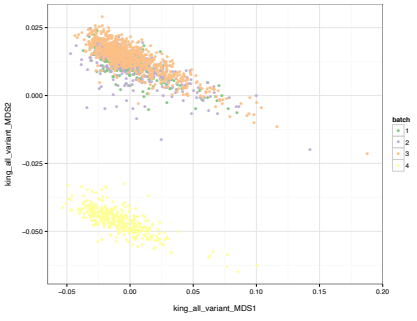
57

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

58

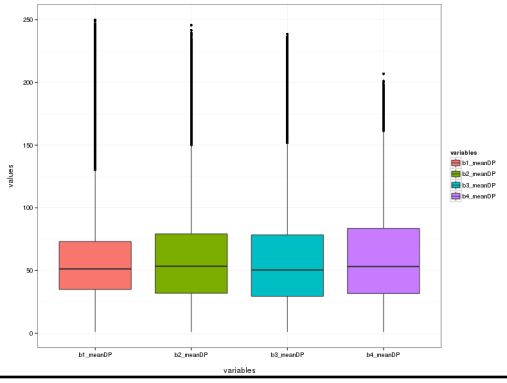
MDS First 2 Components Before QC*



*After VQSR

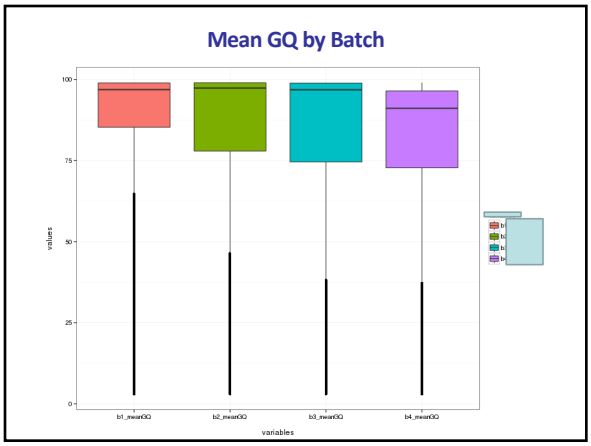
59

Mean GP (genotype) by Batch



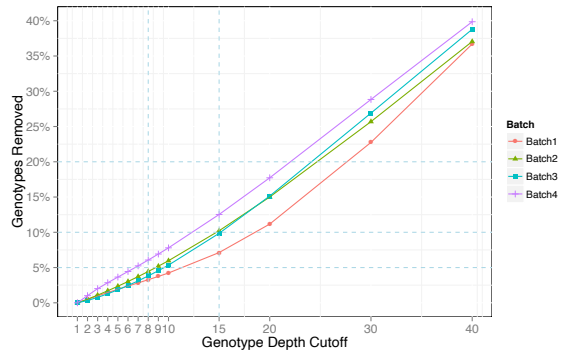
60

Mean GQ by Batch



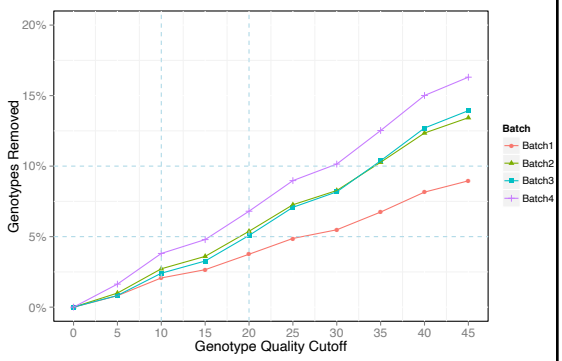
61

Genotypes Removed by DP (genotype) Cut-off by Batch



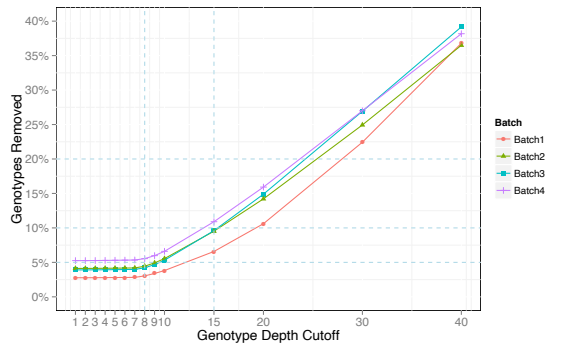
62

Genotypes Removed by GQ Cut-offs by Batch



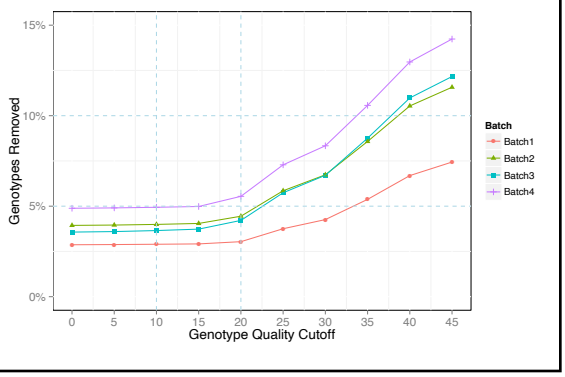
63

Genotypes Removed by DP (genotype) Cut-off by Batch (First removing genotypes with GQ ≤ 20)



64

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



65

Missing Rate Criteria & Sites Removed

	Variant sites removed if missing >10% of their genotypes	Variant sites removed if missing >5% of their genotypes
	Percent of genotype data removed	
Before QC*	2.5%	3.9%
After QC	12.9%	18.3%

Variant sites missing >10% of their data were removed

*After VQSR

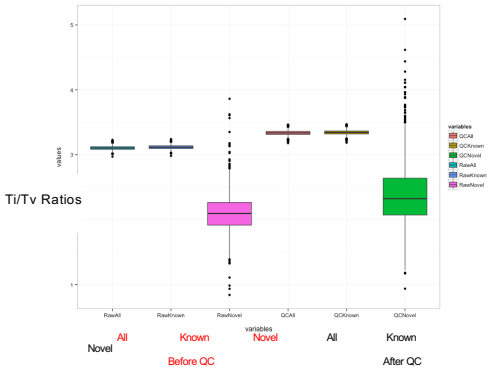
66

Ti/Tv Ratios during QC Process

	Known	Novel	All
Before VQSR	2.95 ± 0.05	1.18 ± 0.29	2.86 ± 0.07
Before additional QC	3.12 ± 0.03	2.01 ± 0.32	3.11 ± 0.03
Genotype QC DP≤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 $p \leq 5 \times 10^{-8}$ N=4,414	3.41 ± 0.04	2.39 ± 0.54	3.40 ± 0.04

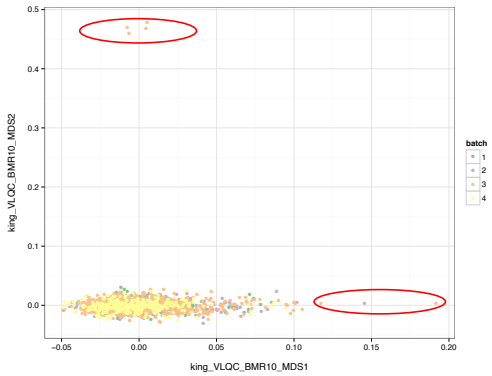
67

Ti/Tv Ratios by Individual Before and After QC



68

MDS First 2 Components After QC



69

Sequence Data QC

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

70

Convenience Controls

- Can reduce the cost of a study
- Genotype data
- Type I error can be increased
  - Ascertainment from different population
  - Differential genotyping error
    - Even if performed at the same facility
- Proper QC can reduce or remove biases

71

Convenience Controls–Sequence Data

- Obtain BAM files and recall cases and control together
  - Can still have differential errors between cases and controls
  - Check variant frequency by variant types in cases and control
    - Synonymous variants should have the same frequencies
    - Would not expect large differences in numbers of variants between cases and controls
- For single variants can compare difference in frequencies with gnomAD but is problematic
  - Differences in frequencies can be due to differences in ancestry and/or sequencing errors
  - Cannot adjust for confounders
    - e.g., sex, population substructure/admixture
- Don't perform an aggregate test using frequency information obtained from databases, e.g., gnomAD, TOPMed Bravo

72

Genotype Array Data
Genotype Data QC – Population Based Studies

- Initially remove DNA samples from individuals who are missing >10% or their genotype data
- For variant sites with a minor allele frequency (MAF)≥0.05
  - Remove variants sites missing >5% of their genotype data
- For variant sites with a MAF<5%
  - Remove variant sites missing > 1% of their genotype data
- The genotypes for variant sites with missing data may have higher genotype error rates

73

Order of Data Cleaning-Genotype Array Data

- Remove samples missing >10% genotype data
- Remove SNPs with missing genotype data
  - If minor allele frequency >5%
    - Remove markers with >5% missing genotypes
  - If minor allele frequency <5%
    - Remove markers with >1% missing genotypes
- Remove samples missing >3% genotype calls
- Check genetic sex of individuals based on X-chromosome markers & Y chromosome marker data (if available)
  - Remove individual whose reported gender/sex is inconsistent with genetic data
    - Could be due to a sample mix-up
- Check for cryptic duplicates and related individuals
  - Used "trimmed data set of markers which are not in LD
    - e.g. r2<0.1
  - Remove duplicate samples

74

Order of Data Cleaning-Genotype Array

- Perform PCA or MDS to check for outliers
  - Use trimmed data set of markers which are not in LD
    - e.g., r2<0.1
  - First with unrelated individuals and then project related individuals on the components
  - Remove outliers from data
    - e.g., Mahalanobis distance
- Check for deviations from HWE
  - Separately in cases and controls
  - Only unrelated individuals
  - If more than one ancestry group
    - Separately for each ancestry group
      - As determined via PCA or MDS
- Examine QQ plots for potential problems with the data
  - e.g., not controlling adequately for population admixture

75