Genotypic Data and Sample Quality Control: Workflow Development and Implementation in Wheat

Umesh Rosyara, PhD
Genomic Breeding Scientist
Global Wheat Program, CIMMYT international

Oct 15th 2018, GOBii webinar
Outline

- Relevance of Quality Control (QC): objectives, and use cases

- Workflow of marker QC
- Marker QC: Genotype calls, batch, marker, sample quality
- Sample QC: Identity, purity

- Wheat QC Panel: Development and evaluation
- GOBii QC tool development: GOBii-DART collaboration
Cheated on seeds and denied justice, a Chinese farmer takes his own life

Tom Lasseter - McClatchy Newspapers
May 25, 2010 04:09 PM
Updated September 18, 2013 06:47 PM

DENG ZHUANG, China — Peng Gonglin wasn’t an important man. He lived in a bare concrete house in a small village where women stoop beside ponds to scrub clothes in buckets and the men often harvest crops by hand.

When his rice fields came up empty last October, Peng had no influence and little cash. The 43-year-old farmer had spent almost all of his family’s savings and borrowed more to lease the land and buy seeds.

County experts in the central province of Henan tested the seeds he’d planted and determined that he’d been sold inferior goods. Peng begged for financial or legal help from the local agricultural bureau and its county seed station.

He took what remained of his family’s money and tried to bribe two local officials to intervene. They accepted the meals, massages and prostitutes, but they did nothing in return, according to a letter he later wrote.
control samples and the samples from centenarians were analyzed in slightly different ways, no efforts to filter out errors caused by different platforms.
Case 1: Paternity

What proportion of plant in this plot are true hybrids from X and Y?

- **Line breeding** – Confirm the accuracy of the pedigree

- **Hybrid wheat** - Degree of hybridity is related to acceptance by farmers and yield?
Case 2: Purity assessment

- Are globally disseminated CIMMYT lines pure enough as many are direct releases?
- What is proportion of off-type plants due to mixtures?
- F7:F8 - Off-type versus segregation due to residual heterozygosity?

What proportion of plants are true-to-type?
Case 3: Genetic identity

- Are CIMMYT improved varieties adopted in its target countries?
- Is a variety in farmer’s field still the same as the breeders selected lines?
- What is genotypic descriptor for a cultivar to claim intellectual property?
- Recent Example: Some line discovered in Brazil resistant to wheat blast, expected to be an off-type of a distributed CIMMYT wheat line.
Impact Assessment

50% of world wheat has CIMMYT lineages

Wheat production sites

Fingerprinting based assessment is helpful for impact assessment in developing countries due to:
- Poor maintenance of record
- Variety renaming (often to their convenience)
- Farmer to farmer exchange of seeds
- Seed purity and mixtures
Case 4: Studies of Gene Discovery

- What is a good or failed marker?
- What is a good or failed sample?
- Does a sample belong within a mapping population?
- Is the identified QTL true or false?
Case 5: Marker-assisted / Genomic Selection

- Has an individual the desired marker allele or haplotype?

- Are the predicted values correct (dependent on genotyping error rates and sample identity)?
Genotypic data and Sample QC

Genotypic data QC: Quality Control of Genotypic Data

Quality control (QC)

Sample / Varietal QC: Quality control using Genotypic Data

Quality Assurance (QA)
Genotypic data and Sample QC

**Product: Genotypic data**
- Sample collection
- DNA extraction
- Genotyping
- Filtering and QC
- Data delivery or upload to data base

**Product: Seed**
- Crossing
- Segregating generations
- International trials
- National trials and cultivar release
- Seed for planting

Quality control (QC)
Data Collection and Use Flow

Sample collection ➔ Sample tracking ➔ Genotyping ➔ Researcher/Genotyping service unit ➔ Data Curator ➔ Database (GOBii) ➔ GOBii-data extractor ➔ User

Placement of sample and blanks and checks

QC by the service provider on genotypic calls, batch QC

QC for marker and sample properties, identity

QC for marker and sample properties, identity

Filtering samples and markers, identity, across platform QC
Genotypic data QC

Pre-QC data

Sample Quality

Sample Call Rate

Sample Relatedness (GD, F1 ped/line)

Population Structure analysis

Batch Effect Analysis

Average call rate and MAF

Compare plate vs rest

Post-QC data

Marker Quality

Marker Genotyping Call Rate

Revising Genotype Calls

Duplicated Sample Concordance

MAF

Segregation Distortion Checking
Use of Controls

- **Blanks** – mix-up of plates
- **Technical replicates** – Reproducibility and error rate estimation
- **Universal check** – across platform comparison
- **Parents** – paternity confirmation, mendelian error rate
Genotyping

• Genotyping error rate versus Call Rate

• Next-Gen Sequencing (NGS) Genotyping: Read depth in multiplexed samples and cost per sample
NGS genotyping error rate

• Coverage and depth:
  Low-coverage sequencing > low cost per sample > many individuals

• Calling algorithms / pipelines:
  ‘genotype likelihoods’ — which incorporate errors that introduced in base calling, alignment and assembly with prior information [eg. allele frequencies and LD pattern]

• Sequencing technology
  error rate of base calls can reduce genotyping error rate (Illumina <1% sequencing error)
Genotyping error rate in SNP arrays

- **Array associated error rate** - Genotyping errors are caused by poor sample quality, High call rate indicator of good sample quality

- **Clustering algorithm associated error** – some markers are not properly clustered causing genotyping error or high missing data

- Genotyping error rate is usually < 1%
Paralogs, Homologs and Homeologs

**Diploid (2 copies of gene)**

- **Norm Theta**
  - AA
  - AB
  - BB

**Hexaploid (6 copies of gene)**

- **Norm Theta**
  - AA
  - AB
  - BB

- **Norm R**
  - AAAA
  - ABB
  - BB

- **Norm Theta**
  - AAAAB
  - AABBB
  - ABBBB
  - BBB
  - BBB
  - BBB

20
Null alleles ($\phi$) and Presence-Absence Variants

Low call rate

Dominant / Deletions / Null alleles ($\phi$)
Estimation of Genotyping error rates

Two discordant genotypes

\[ e_{ij}(AA) = f(!AA) \text{ at } j \text{ when } AA \text{ at } i \]
\[ \text{total (AA) at } i \]

\[ e_{ij}(AB) = f(!AB) \text{ at } j \text{ when } AB \text{ at } i. \]
\[ \text{total (AB) at } i \]

\[ e_{ij}(BB) = f(!BB) \text{ at } j \text{ when } BB \text{ at } i. \]
\[ \text{total (BB) at } i \]

\[ e_{ij} = 1 - \left( \frac{1}{N} \sum_{k=1}^{N} n_k \right) \]
\[ n_k = \begin{cases} 
1(G^i_k = G^j_k) \\
0(G^i_k \neq G^j_k) 
\end{cases} \]

six potentially different error rates

Where N indicates total SNPs, \( G^i_k \) is genotype called on SNP k for sample i and \( G^j_k \) is genotype called on SNP k for sample j
\[2(1 - \alpha - \beta)(\alpha + \beta) + \alpha\beta\]

when \(\alpha = \beta\)
Genotyping Error rate sensitivity

GWAS for rate variants
Haplotype Construction
Imputation
GWAS for frequent alleles
Genomic selection (low $h^2$)
Genomic selection (high $h^2$)
Biparental QTL mapping

High
Low
Batch Quality

ANOVA

\[ y = a + \beta x + e \]

- \( y \) = call rate
- \( x \) = batch
- \( e \) = error

Systematic problem with a batch
Marker Quality

• **Call rate** (completeness and genotype quality)
  - random vs nonrandom

• **Duplicate sample concordance** – whether duplicate sample give same results (reproducibility)

• **Consistency with Mendelian transmission** – inheritance error (accuracy) and segregation distortion
Call rate at each locus

GBS data

\[ \mu - 2\sigma \quad \mu - \sigma \quad \mu \]

SNP array

\[ \mu - 2\sigma \quad \mu - \sigma \quad \mu \]

Definition of failed marker?
Sample Quality

- **Call rate** (poor call rate across markers – poor amplification / DNA quality)

- **Duplicate sample concordance** – reproducibility, confirm that there is no sample handling problem / mistaken identity

- **High inheritance error rate** – the sample has poor quality or mistake in identity
Call rate for each sample

Definition of failed sample?
SAMPLE IDENTITY

• **Correctness:** Checking if the individual genotype is correct and there is no mistake in sample collection, genotyping or data processing

• **Paternity:** Checking if X individual is offspring of A and B

• **Purity and Identity:** finding the most representative sample and avoid heterozygotes/heterogeneous, finding most likely identity of unknown sample
Identity

Probability that two alleles drawn at random from the population will be different

$$GD = \frac{n}{n-1} \left(1 - \sum_i p_i^2\right)$$  \hspace{1cm} \text{Nei (1987)}

The match probability (PM) and power of discrimination (PD)

$$PM = \sum_i G_i^2$$  
$$PD = 1 - PM$$  \hspace{1cm} \text{Fisher (1951)}
Relatedness

Inbreeding \((F)\) - probability of identity by descent (IBD) and relatedness \((r)\) - correlations in marker-based analyses as predictor of kinship

Many measures, e.g. Queller and Goodnight (1989)

\[
\hat{r} = \frac{\hat{r}_{XY} + \hat{r}_{YX}}{2}
\]

\[
\hat{r}_{XY} = \frac{\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd} - 2(p_c + p_d)}{2(1 + \delta_{cd} - p_c - p_d)}
\]

\[
\hat{r}_{YX} = \frac{\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd} - 2(p_a + p_b)}{2(1 + \delta_{ab} - p_a - p_b)}
\]
Kinship coefficient is often ambiguous for establish relationship but can provide some indicator:
siblings = 0.25, parent-offspring =0.25;
Grandparent-grand child=0.125, Uncle-Nephew = 0.125
Excess Mendelian inheritance errors is indicator of wrong parentage

Polymorphism Informative Content (PIC)

\[
PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2
\]

Probability of Excluding Paternity (Q)

\[
Q = \sum_{i=1}^{n} p_i - (1 - p_i + p_i^2)(1 - p_i)^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_i p_j (1 - p_i + p_j^2)(p_i + p_j)
\]
Development of a Wheat QC Panel

Objective: Sample / Cultivar QC (genetic identity, purity)
Guide for Markers Selection

Closely related markers
Distance \( (\text{sample}_i, \text{sample}_{i+1}) \)

Distantly related markers
Distance \( (\text{sample}_i, \text{sample}_{i+1}) \)

Selected markers
Probability of identity (PI) and Power

\[ \text{PI} = \sum_i p_i^4 + \sum_{j \neq i} 2p_ip_j \]

Paetkau and Strobeck, 1994

\[ \text{PI}_{\text{SIB}} = \frac{1}{4} + \frac{1}{2} \sum_i p_i^2 + \frac{1}{2} \left( \sum_i p_i^2 \right)^2 - \frac{1}{4} \sum_i p_i^4 \]

Taberlet and Luikart, 1999; Evett and Weir, 1998

The PD for a set of loci, 1..., L

\[ \text{PD} = 1 - \prod_{l=1}^{L} 1 - \text{PD}_l \]

The graph shows the Power of Discrimination (PD) as a function of the number of markers for different MAF values (0.1, 0.2, 0.3, 0.4, 0.5). The x-axis represents the number of markers, and the y-axis represents the power of discrimination.
Variation within Purelines

\[ p(\text{homozygous}) = \frac{2^s - 1}{2^s} \]

- Empirical estimation of allowable variation within cultivar
- Minimum difference between sister lines

\( I = \) number of loci segregating

Last selection for purity before release
Development and Testing Datasets

The breeding dataset:
n = 46,089 lines and 11,293 GBS markers

The diversity dataset:
1,102 lines and 21,885 markers SNP array
Breeding Panel
\( n=46K, m=12K \), Marker platform: GBS

Marker filter
Missing < 0.05

Marker filter
MAF > 0.35

Population Structure
H-K mean with random sampling 10000

Marker clustering
HK-means

Marker selection
Criteria: MAF, Marker cluster, Chromosome

Discrimination Power evaluation
(% of combinations with > 3 marker different)

Diversity Panel
\( n=1,102, m=22K \), Marker platform: 94K array

Marker Filter
Missing < 0.05

Marker filter
MAF > 0.35

Population Structure
H-K-means

Marker clustering
H-K-means

Marker selection
Criteria: MAF, Marker cluster, Chromosome

Discrimination Power evaluation
(% of combinations with > 3 marker different)

KASP array

Array Evaluation for Identity, Purity and \( F_1 \) pedigree verification

Final marker selection and Panel Release
Power of Discrimination

\[ 1 - PD = \frac{\# \text{NonDiscriminated pairs}}{\text{Total Pairs}} \]

Breeding Population

- Proportion of samples with less than 3 markers genotypes are different
- Proportion based on 7,998,000 pair combination

Diversity Population

- Proportion of samples with less than 3 markers genotypes are different
- Proportion based on 576,201 pair combination
Panel Development and Evaluation

- Selected GBS and SNPs converted to KASP assays with total 166 markers (83 from breeding set and diversity set)

- Final set of markers – number of markers, power and cost of assay

- Evaluated for new samples for purity assessment, $F_1$ pedigree verification / Hybridity and identity (initially 744 entries)
Seed Purity

Seed mixture vs segregation due to residual heterozygosity

Segregating Generations

Yield Trials (F₇:F₈)

Seed multiplication

International Nurseries
F₁ verification and hybridity testing

True hybrids vs selfs versus mixture

Degree of hybridity = 90%
Identity Confirmation / Assurance

Gene bank  Breeders/ Pre-breeder’s stock  Gene bank  Breeders/ Pre-breeder’s stock
Quality assurance of near identical lines such as Near Isogenic lines or A/B lines in hybrid wheat production.
Germplasm Profile

- QC panel
- QTL / Gene alleles
- Morphological / Specific traits

Cultivar identification and intellectual property management
GOBii QC tools

Flapjack tools (stand alone)

- Pedigree Verification (F₁s Known Parents)
- Pedigree Verification (Lines with Known Parents)
- Similarity matrix computation
- Graphical genotyping

KDCompute QC integrated with GOBii GDM

- Marker data QC
- Reproducibility
- F₁ pedigree verification
- Graphical genotyping
- Visualization and dynamic filtering
GOBii is developing developing database and tools to enable genomics assisted breeding in public institutions

GOBii- Genotypic data manager (GDM) is genotypic database

DART is developing KDCompute module for QC that is integrated with GOBii - Genotypic data manager (GDM)
Features

• Most of **Marker QC** statistics discussed in this presentation

• **Sample and batch QC** - sample reproducibility when technical replications are provided  graphical genotypes

• **Genetic identity** - Relatedness calculation, $F_1$ pedigree verification

• **Graphics and dynamic filtering** (under development)
Graphical Genotypes and QC statistics
QC statistics

Filters
Few points …

• **QC is subjective** (objective of study, quality requirement / sensitivity, platform)

• **QC is important** (reduce false discovery rate, error in pedigrees, establish genetic identity, level of genetic purity)
Acknowledgements

• Susanne Dreisigacker, Molecular Breeding team, CIMMYT
• Eng Hwa NG, HTPG project
• Liz Jones, Director, GOBii
• Andrew Kowalczyk, DART
• Bhoja Basnet, Hybrid wheat, CIMMYT
• Deepmala Sehegal, Molecular Breeding, CIMMYT
• Ravi Singh, Bread wheat, CIMMYT
• Tom Payne, Wheat Gene Bank, CIMMYT
• Members of Global wheat program / Genetic resource Program CIMMYT
Thank you for your interest!