applied biosystems

Robust SNP screening methods for *de novo* marker verification and selection

A case study on chicken-marker selection, screening, and genotyping performance

Abstract

Recent advances in whole-genome DNA sequencing have revolutionized our ability to discover genomic variation, thus enabling highly powered genotyping studies. This application note describes a process to screen a large list of single-nucleotide polymorphisms (SNPs), identified in sequencing studies, to enable intelligent selection of the most informative variants for downstream high-throughput genotyping experiments. This process has been successfully applied to development of genotyping marker panels for livestock, aquatic animals, and plants, including diploid and polyploid species. Here we present a case study of this process applied to marker verification and selection in chicken.

Introduction

Large SNP discovery initiatives have confirmed that low-coverage next-generation sequencing (NGS) in many samples is a more powerful *de novo* variant discovery paradigm than deep NGS in fewer samples [1]. However, along with their powerful discovery ability, NGS technologies have been affected by a high proportion of sequence errors and missing data [2], and this is amplified in low-coverage sequencing.

False discovery rates between 6.3% and 7.8% have been reported for NGS platforms from different suppliers [3], equivalent to more than 3,000 false positives in a panel of 50,000 SNPs. At true-positive SNP loci, genotype accuracy decreases as coverage decreases. We have performed extensive verification of millions of SNP genotypes from low- and high-coverage NGS data using our Applied Biosystems™ Axiom™ Genotyping Solution, which itself has a typical concordance of 99.8% to gold-standard reference datasets.

Our verification analysis (Table 1) demonstrated that 30x NGS data had >98% concordance across all genotype classes. Data from other research have since shown that NGS coverage of >40x is required to call genotypes across the genome with acceptably low error rates [4]. Our analysis also showed that 4x NGS SNP discovery data overestimated the major homozygote frequency by miscalling heterozygotes. As a result of this bias, major homozygote concordance looked good compared to the array reference data, but at the expense of a 9.9% error rate for heterozygotes and an 11.6% error rate for minor homozygotes.

ThermoFisher SCIENTIFIC

Sequence and genotype errors in low-coverage NGS data can therefore be a significant source of false and redundant SNPs. Additional markers will also cluster poorly due to incompatibility with the genotyping assay chosen, nearby secondary polymorphisms, or other technical factors.

Table 1. Genotype concordance of low- and high-coverage NGS data, compared to the reference genotype datasets generated on Axiom genotyping arrays.

	Concordance compared to Axiom genotypes	
	4x data	**30x data**
Major homozygote	99.7%	99.9%
Heterozygote	90.1%	99.8%
Minor homozygote	88.4%	98.5%

Impact of false positives and redundant markers on marker panel capability and cost

For a lab developing a genotyping marker panel on arrays or any other technology, the impact of these sources of marker dropout is significant. First, inclusion of false-positive SNPs due to sequencing errors or poorly performing markers wastes time and money, and weakens the power of the marker panel and the studies that use it. Second, at true SNP loci, the error in NGS genotype calls leads to inaccurate allele frequency estimation, incorrect linkage disequilibrium (LD) map construction, and, ultimately, poor marker selection.

A recent publication describing the design and verification of a soybean genotyping array demonstrates the difference in the empirical performance of a marker panel from the original *in silico* design if *de novo* markers are not verified before selection [5]. Starting from a target set of 60,800 SNPs, the authors reported dropout of 4,704 (8%) of the markers due to false positives, monomorphism in the populations studied, or poor clustering performance. When they also included random dropout caused by the manufacturing process of the bead-array technology used, they lost a total of 13,465 (22%) of the content on the array.

Researchers who plan a *de novo* marker verification strategy are likely to reduce sequencing errors, inaccurate allele and LD estimates, poor marker selection, dropout due to false positives, population-specific monomorphs, and poorly performing markers in downstream genotyping experiments. A verification strategy can reduce gaps in coverage, increase power, and alleviate the need to design marker redundancy into a genotyping panel.

Requirements for a robust marker selection strategy

Robust marker panel development requires verification of *de novo* variants prior to selection for genotyping. This will lead to an optimized set of markers that has well-characterized coverage, performance, and population relevance.

The concept of data verification using an alternative technology is not new in science. For example, the literature features hundreds of papers that describe the use of real-time PCR to verify differentially expressed genes discovered by microarrays. However, verification is surprisingly rare in NGS studies, although increasingly recommended until the limitations and biases of the technology are better understood [6].

The verification strategy should accomplish 6 key objectives:

Identification of as many *de novo* variants as possible

- NGS has the ability to discover vast numbers of *de novo* variants. On a highly parallel genotyping platform, a maximum number of possible putative variants should be taken through verification. This would give the widest choice of markers and best opportunity to optimize the final marker panel to the target application.

Identification and removal of erroneous SNPs

- By verifying *de novo* variants from NGS on an orthogonal genotyping technology, sequencing errors can be rapidly identified to avoid any chance of selecting them as markers. High accuracy and low error rate are required to enable confident verification.

Identification and removal of poorly performing markers

- The same scalable technology should be used for verification and downstream genotyping. The marker panel selected after verification contains 100% high-performing markers that will continue to perform well in the final genotyping experiments because the technology and assay chemistry are consistent.

Generation of accurate genotypes in the reference sample set

- Accurate genotyping enables accurate estimates of allele frequency and LD maps. This significantly increases the robustness of SNP selection to cover genomic regions without gaps or wasteful marker redundancy.

Adequate power across all study populations

- Often, the number of samples that are included in NGS discovery is limited by available budget. This can result in underrepresentation of some populations or a reduced diversity of population in downstream studies. The danger is that markers can look informative in the discovery set but be monomorphic in important populations in the broader diversity set.
- Smaller discovery sample sets also reduce the power to obtain accurate population-specific allele frequencies and LD maps, especially for rarer variants.
- The verification experiment should also be designed, when necessary, to expand the diversity and size of the discovery sample set to provide sufficient power for population-specific verification of variants.

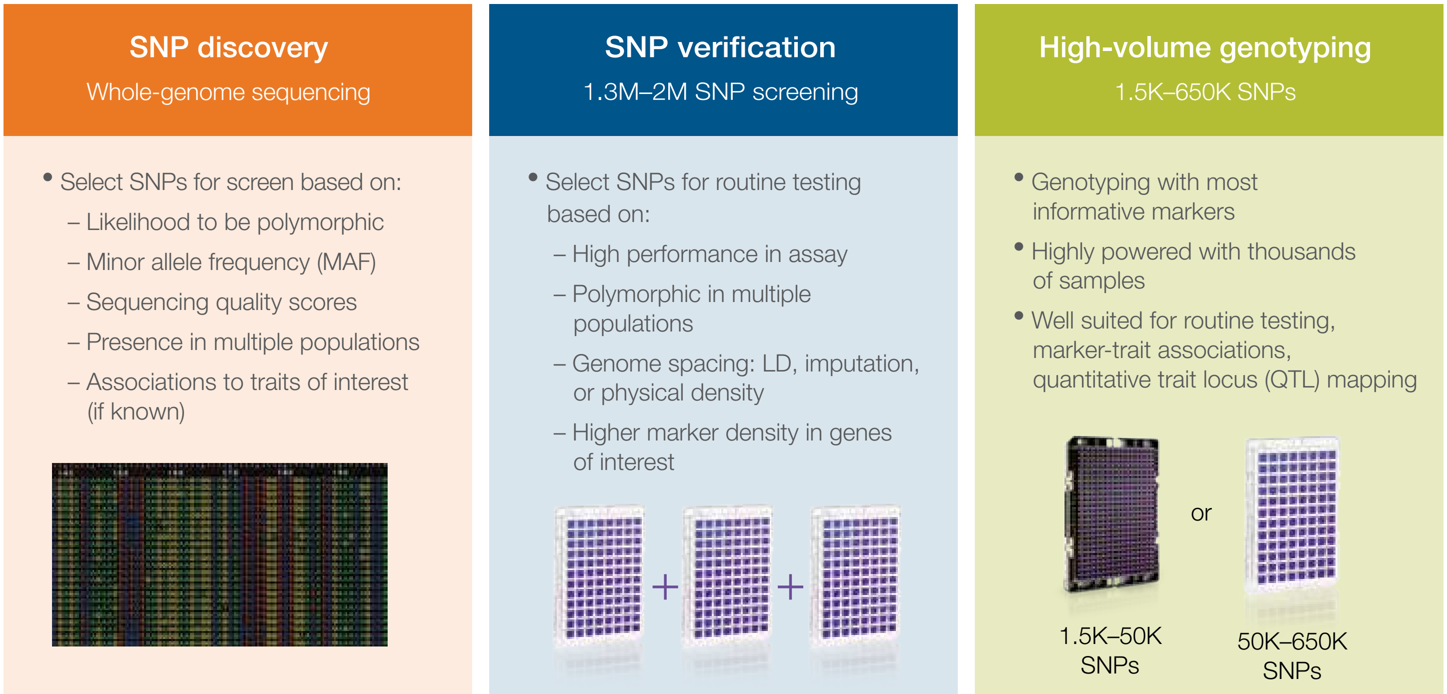
Technical portability of selected markers

- Following verification, the technology platform used must reliably transfer the selected markers into the final genotyping panel. For example, bead-array technologies are known to randomly drop markers from the marker panel during manufacture, and there is no control over which markers are lost. Knowledge of the missing markers only emerges after manufacture, by which time it is too late to repair the gaps. Dropout rates of 5.7% [7] and 14.5% [5] have been variously reported, while suppliers allow for as much as 20.8% [8]. Ideally, the genotyping platform must be able to take all selected markers and represent them in the final panel with 100% reliability. Failure to do this increases the risk of coverage gaps that can only be mitigated by building wasteful redundancy into the panel design.

Strategy for robust *de novo* marker verification and selection

With scientists studying livestock, aquatic animals, and plant species, we have established a robust *de novo* marker verification strategy that enables optimized selection of markers for either whole-genome or targeted-genotyping panels on Applied Biosystems™ Axiom™ myDesign™ Custom Genotyping Arrays (Figure 1). This strategy consists of the following steps:

- Marker discovery
  - Perform whole-genome sequencing and alignment
  - Select SNPs and indels based on quality metrics and likelihood of being polymorphic (~10 million SNPs)
- Marker verification
  - Genotype sample set of representative diversity
  - Select SNPs and indels based on genomic position and coverage, population relevance
- High-volume or routine genotyping
  - Design a smaller, more cost-effective array for routing genotyping
  - Genotype a larger sample set on the best-performing and most-informative SNPs



SNP discovery

Whole-genome sequencing

- Select SNPs for screen based on:
  - Likelihood to be polymorphic
  - Minor allele frequency (MAF)
  - Sequencing quality scores
  - Presence in multiple populations
  - Associations to traits of interest (if known)

SNP verification

1.3M–2M SNP screening

- Select SNPs for routine testing based on:
  - High performance in assay
  - Polymorphic in multiple populations
  - Genome spacing: LD, imputation, or physical density
  - Higher marker density in genes of interest

High-volume genotyping

1.5K–650K SNPs

- Genotyping with most informative markers
- Highly powered with thousands of samples
- Well suited for routine testing, marker-trait associations, quantitative trait locus (QTL) mapping

Figure 1. Process for SNP discovery with DNA sequencing, SNP verification with a large screening experiment on multiple arrays, and high-volume genotyping for downstream discoveries and routine testing.

SNP discovery

For effective variant discovery, a diverse population of samples (multiple breeds, lines, etc.) should be sequenced to increase genetic variability and ensure that polymorphisms between populations are identified. Upon completion of DNA sequencing, the sequences are aligned to a reference genome. In cases of *de novo* sequencing and assembly where there is no reference genome, sequences are joined together where they overlap. Reads are then assembled into larger fragments, generating long contigs. In either case, SNPs and indels can be identified based on sequence mismatches at given locations. For polyploid species, separate assembly of homologs may be necessary so that the subgenomes are not confounded [9].

SNPs are then filtered according to multiple criteria [10] that may include:

- Reference sequence length
- Minimum and maximum read depth
- Consensus base ratio
- SNP quality score
- Presence of nearby SNPs
- Presence of SNPs in multiple populations
- SNPs within exons
- Coding and nonsynonymous SNPs
- Coverage of genes of interest
- Genome-wide coverage based on LD or imputation

Depending on the size of the SNP-screening experiment, more stringent QC metrics may be applied to define the SNP list that will be used for verification.

SNP verification

The SNPs discovered from sequencing must be verified to identify the true SNPs and eliminate false positives and redundancy in the final marker panel. The ideal screen would be maximally powered by genotyping all SNPs across all samples, but this would be a costly experiment. Here we present an economical approach to first screen a large set of polymorphisms across a diverse but smaller set of samples, then genotype a larger set of samples across a selected set of high-value, high-performing markers.

The first stage of the verification will accomplish the goal of identifying a subset of high-performing polymorphic SNPs that show potential for marker–trait associations and other downstream applications. This is accomplished by designing a genotyping screening marker panel on Axiom myDesign Custom Genotyping Arrays, which can include any number of SNPs. The screening arrays typically contain around 2 million SNPs but have contained as many as 8 million SNPs. Since the discovery phase likely resulted in a large list of SNPs (tens of millions), bioinformatic filtering can be applied to select the SNPs that will be used for screening. Our microarray bioinformatics service provides *in silico* design scores that predict the likelihood of success in the genotyping assay. SNPs with the highest scores representing LD blocks—even physical distribution or best genetic coverage across the genome (or genes of interest)—and exonic nonsynonymous SNPs may be selected. SNPs with neighboring polymorphisms within 10 bases are excluded. SNPs that are likely to be polymorphic in multiple populations are often prioritized.

Once a SNP list has been defined, the genotyping screening arrays are designed. Roughly 650,000 SNPs fit on each array, and multiple arrays are usually designed for this step. The number of arrays used for screening experiments has been as large as 12, but the typical screen uses 3 arrays, which enables approximately 2 million genotypes. Axiom arrays are formatted on a 96-array microplate, which enables end-to-end automation of the assay and high-throughput genotyping.

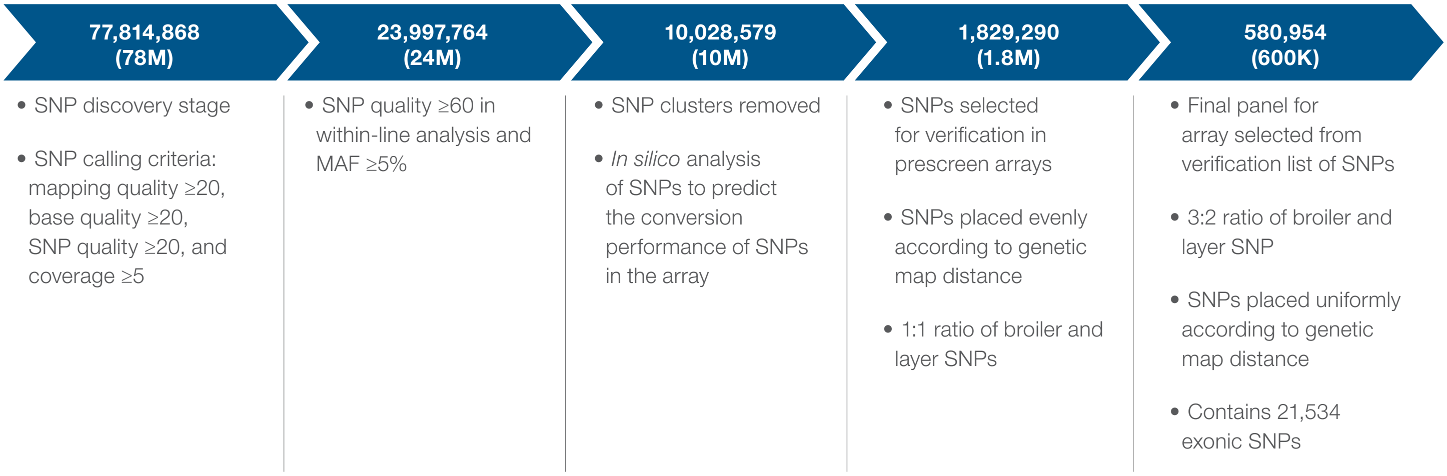
High-volume or routine genotyping

In the second stage, a larger set of samples is genotyped across a smaller set of SNPs and indels. The goals of this stage are two-fold: genetic discoveries (QTL mapping, marker–trait associations, genome-wide association study (GWAS), etc.) and routine testing for ongoing molecular breeding activities. SNPs and indels may be selected for the final array based on the following criteria:

- High call rates in the Axiom assay
- Good genotype cluster separation
- Polymorphic (true positives)
- Informative across populations to be genotyped
- Associations with traits of interest (if known)
- Tagging other variants based on LD
- Imputation of other variants in the genome
- Even spacing across the genome (using genetic map distance or physical distance)

Case study: application to chicken genetics research

The proposed strategy for SNP discovery, verification, and routine testing has been applied to chicken genetic analysis research as described by Kranis A et al. [11]. A consortium of chicken researchers and breeders was interested in developing a high-density genotyping array for multiple breeds and populations of chicken, one of the world's most important farm animals. The group sequenced the chicken genome, compiled a list of potential variants, conducted a SNP-screening experiment using Axiom myDesign Custom Genotyping Arrays, and designed a 600K Applied Biosystems™ Axiom™ Genome-Wide Chicken Genotyping Array (Figure 2).



77,814,868 (78M)

- SNP discovery stage
- SNP calling criteria: mapping quality ≥20, base quality ≥20, SNP quality ≥20, and coverage ≥5

23,997,764 (24M)

- SNP quality ≥60 in within-line analysis and MAF ≥5%

10,028,579 (10M)

- SNP clusters removed
- *In silico* analysis of SNPs to predict the conversion performance of SNPs in the array

1,829,290 (1.8M)

- SNPs selected for verification in prescreen arrays
- SNPs placed evenly according to genetic map distance
- 1:1 ratio of broiler and layer SNPs

580,954 (600K)

- Final panel for array selected from verification list of SNPs
- 3:2 ratio of broiler and layer SNP
- SNPs placed uniformly according to genetic map distance
- Contains 21,534 exonic SNPs

Figure 2. The process used for SNP selection during SNP discovery, SNP verification, and design of Axiom Genome-Wide Chicken Genotyping Array. Source: Kranis A et al. (2013) Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* 14:59.

First, 243 chickens were resequenced. The samples represented 24 lines of broilers, white-egg layers, brown-egg layers, and experimental inbred layers. Samples were pooled to introduce additional variation without incurring experimental costs. Depth of coverage ranged from 8x to 17x. The sequences were aligned to the Gallus_gallus_4.0 reference genome, and 139 million SNPs were identified from resequencing, 78 million of which were present in multiple chicken lines. To select the SNPs with the highest likelihood of conversion, quality control metrics were applied:

- Sequencing SNP quality score ≥60
- MAF ≥0.05
- SNP or indel was previously detected by another platform
- No interfering polymorphisms within 10 bp of one side of SNP and within 4 bp of the other side
- Representation of all breeds and lines (Figure 3)
    - Many of the SNPs appear in multiple lines (these are older variants)
        - 23% common among broilers, layers, and inbred lines
        - 1% common among broilers, white-egg layers, and brown-egg layers
    - Newer variants, appearing in only one line, were also included

Ten million SNPs were selected and submitted to our design team to assign *in silico* design scores to predict likelihood of success in the Axiom assay. These scores were calculated for both forward and reverse strands for each SNP. Roughly 6.6 million SNPs passed both of these design criteria:

- The 16 bp flanking sequence on either side of the SNP should not match another sequence in the genome
- The p-convert value predicts high probability of conversion on the array

Next, SNPs were selected for even spacing across the genome according to genetic map distance, with an equal ratio of SNPs that segregate in layers and broilers, taking into consideration that all 24 lines of chickens were represented.

Three Axiom myDesign Custom Genotyping Arrays were then designed to interrogate a total of 1.8 million SNPs. Each array contained ~600,000 markers. 282 samples were genotyped, including 32 trios from 3 broiler lines, 4 white-egg layer lines, 5 brown-egg layer lines, and 26 other diverse individuals. The samples were selected to represent the same diversity as the lines sequenced in the previous experiment.

The call rate from the 1.8 million SNP screen was >98%. Over 1.18 million (64.9%) SNPs were polymorphic and exhibited stable Mendelian inheritance and high resolution in the Axiom assay. Next, the final genotyping array was designed based on the following criteria:

- Polymorphic
- Good genotype cluster separation
- High call rates
- Priority for nonsynonymous SNPs in protein-coding regions
- Synonymous SNPs in strong LD with functional mutations
- Representation of >100,000 SNPs in all 24 lines
- Uniform distribution across the genome, based on genetic map distance, for both broilers and layer lines
- A 3:2 broiler-to-layer ratio of representation of SNPs (due to low LD in broilers)

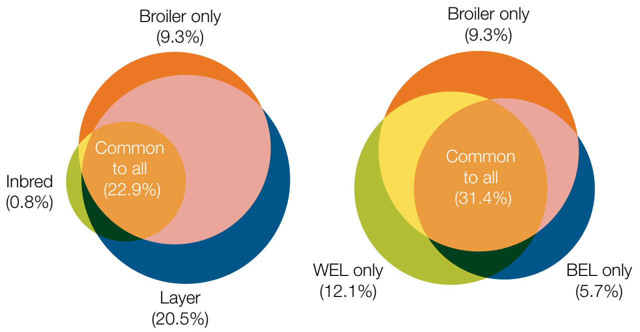


Figure 3. Venn diagrams showing overlap of SNPs in the list that was submitted to our design team for the screening experiment. Source: Kranis A et al. (2013) Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* 14:59. WEL: white-egg layer; BEL: brown-egg layer.

applied biosystems

The resulting 600K Axiom Genome-Wide Chicken Genotyping Array (Cat. No. 902148) is the highest-density chicken genotyping array on the market and the only chicken genotyping array that is openly available to the public. The SNP-screening experiment has enabled researchers to design an array with SNPs that are high-performing and represent a population diversity of 24 lines of chickens, making this product well suited for many high-throughput applications, including GWAS, QTL mapping, marker–trait associations, and genomic selection.

This screening protocol has since been adapted to develop genotyping arrays for diploid and polyploid animal, aquatic, and plant species. This process has enabled development of well-characterized, highly optimized marker panels for downstream genotyping applications.

References

1. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.
2. Luo L et al. (2011) Association studies for next-generation sequencing. *Genome Res* 21(7):1099-1108.
3. Liu DJ et al. (2010) Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am J Hum Genet* 87:790-801.
4. Ajay SS et al. (2011) Accurate and comprehensive sequencing of personal genomes. *Genome Res* 21(9):1498-1505.
5. Song Q et al. (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE* 8(1):e54985.
6. Irizarry R (2012) In: Validating Complex Biology: How Arrays Can Complement Your Next-Gen Data. Science Webinar Series, http://webinar.sciencemag.org/webinar/archive/validating-complex-biology
7. Eeles R et al. (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Gene*t 40:316-321.
8. Illumina Inc. Designing and Ordering iSelect® HD Custom Genotyping Assays. Technical note. www.illumina.com/documents/products/technotes/technote_iselect_design.pdf
9. Byers R et al. (2012) Development and mapping of SNP assays in allotetraploid cotton. *Theor Appl Genet* 124(7):1201-1214.
10. You F M et al. (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* 12:59.
11. Kranis A et al. (2013) Development of a high density 600K genotyping array for chicken. *BMC Genomics* 14:59.

Find out more at **thermofisher.com/microarrays**

ThermoFisher SCIENTIFIC

For Research Use Only. Not for use in diagnostic procedures.