Chapter 23

Using Droplet Digital PCR to Analyze Allele-Specific RNA Expression

Nolan Kamitaki, Christina L. Usher, and Steven A. McCarroll

Abstract

Genome-wide association studies have discovered thousands of common alleles that associate with human phenotypes and disease. Many of these variants are in non-protein-coding (regulatory) regions and are believed to affect phenotypes by modifying gene expression. In any organism with a diploid genome, such as humans, measuring the expression of each allele of a gene provides a well-controlled way to identify allelic influences on that gene's expression. Here, we describe a protocol for precisely measuring the allele-specific expression of individual genes. This method targets the nucleotide differences between the two alleles of a gene within an individual and measures the "allelic skew," the extent to which one allele is expressed more than the other. We cover the design of effective assays, the optimization of reactions, and the interpretation of the resulting data.

Key words Allele-specific expression, Allelic skew, Allelic imbalance, Droplet digital PCR, Digital PCR, mRNA expression, Assay design

1 Introduction

Genome-wide association studies (GWAS) have mapped hundreds of human traits to thousands of common variants [1]. Less than 30% of these variants are nonsynonymous or are in linkage disequilibrium with a variant that is [2], suggesting that much of the variation driving phenotypic differences does so not by altering protein function, but by altering gene expression. Furthermore, systematic substitution of DNA bases in known enhancers has shown that the vast majority of these variants modestly affect expression (less than twofold change) and very few abolish or greatly increase expression [3]. Measuring these modest effects on gene expression requires techniques that can both ascertain modest differences in expression and can filter through the noise created by environmental influences, genomic background, and other *trans*-acting effects.

George Karlin-Neumann and Francisco Bizouarn (eds.), *Digital PCR: Methods and Protocols*, Methods in Molecular Biology, vol. 1768, https://doi.org/10.1007/978-1-4939-7778-9_23,

Once a candidate gene has been selected, usually by hypotheses about the mechanisms and genetic architecture of a phenotype, we can compare the expression of the two alleles within the same biological sample, thus avoiding many sources of noise. In general, a person's two alleles are subject to the same environmental and *trans*-acting genetic influences, which are potentially strong enough to obscure modest genetic effects and can be difficult to control for in study designs that assay total expression in genetically diverse humans living in variable environments [4].

In the absence of allele-specific regulatory effects, the two alleles will be expressed at a ratio of 1:1, but local *cis*-acting genetic and epigenetic variation can cause one allele to be expressed more than the other. Measuring this difference requires a technology capable of measuring the two alleles precisely and with equal sensitivity. Techniques that take this measurement after PCR amplification—including RNA-seq and pyrosequencing—run the risk of confusing amplification effects with real genetic effects. Even a subtle difference of 1% in amplification efficiency will, after 30 PCR cycles, result in a 1.35-fold relative difference in the abundance of the two alleles in the amplified material, an effect size that would exceed that of most enhancer SNPs. In addition, digital measurement, which counts the number of transcripts with each allele in a sample, is preferable to analog measurement, which estimates ratios, since digital measurement allows the statistical significance of allelic skews to be estimated for each person.

The ability to digitally count the abundance of alleles in an unamplified sample is therefore critical. 1-Step RT-droplet digital PCR (RT-ddPCR) accomplishes this by partitioning the individual RNA transcripts of a gene into separate reaction compartments (droplets) before amplification/detection [5]. Within these ~20,000 droplets, the RNA is reverse-transcribed into cDNA and amplified using a molecular assay that is composed of: a pair of primers that will amplify both alleles, a fluorescence probe targeting the reference allele, and a probe (with a different fluorophore) targeting the alternate allele. The number of RNA molecules originating from each allele is quantified by counting the number of droplets that fluoresce with the corresponding probe colors. Provided that the positive and negative droplets are clearly and consistently distinguished, ddPCR is robust to differences in the amplification kinetics of the alleles (a clear advantage over pyrosequencing and real-time PCR). Droplet digital PCR is also more resistant to the effects of nonspecific binding of the probes to the opposite allele, allowing the determination of the correct allele in each droplet as long as each probe has greater affinity for its respective allelic target.

Here, we share our protocol for measuring the allelic skew of a gene using ddPCR. We cover how to design a successful allele-specific assay, how to screen and optimize that assay, and finally,

how to use it in ddPCR. We focus on assay design in humans, but the protocol is readily adaptable to any species with a diploid genome and substantial heterozygosity. It is also readily adapted to measure the expression levels of paralogous genes in a paralog-specific manner.

However, this method has limitations in that it can only assay those individuals heterozygous for the transcribed reporter SNP, and historic recombination between the reporter SNP and the functional variant may introduce uncertainty about the direction of effect. Thus, the clearest overall picture may emerge when this strategy is combined with another that assays total expression across a large cohort [6].

2 Materials

2.1 Locus-Specific Analysis Reagents

1. 18 μM forward primer; 18 μM reverse primer; 5 μM 5' FAM-labeled, 3' Iowa Black ZEN-quenched probe (*see* **Note 1**) that targets the reference allele; and 5 μM 5′ HEX-labeled, 3′ Iowa Black ZEN-quenched probe that targets the alternate allele. Store at −20 °C away from light (*see* **Note 2**). Vortex and centrifuge before use.

2.2 Biological Sample

1. RNA samples (100 fg–100 ng of total RNA) from individuals heterozygous for the reporter SNP (*see* **Note 3**). They need not be heterozygous for the candidate functional variant. Store at −80 °C in RNase-free ddH_2O. Vortex and centrifuge before use.

2.3 ddPCR Components and Equipment

1. 20× assay mix from Subheading 2.1.
2. One-Step RT-ddPCR Advanced Kit for Probes (Bio-Rad). Store at −20 °C. Invert to mix before use.
3. Droplet Generation Oil for Probes (Bio-Rad). Store at room temperature.
4. Droplet Reader Oil (Bio-Rad). Store at room temperature.
5. DG8™ Cartridges for droplet generation (Bio-Rad).
6. DG8™ Gaskets for droplet generation (Bio-Rad).
7. QX200™ Droplet Digital PCR System: droplet generator and cartridge holders, droplet reader, and QuantaSoft reader software (Bio-Rad).
8. Rainin multichannel pipettors and corresponding tips for pipetting 20 μL and 40 μL volumes (*see* **Note 4**).
9. Half-skirted Eppendorf 96-well PCR plates for droplet thermal cycling and reading.

10. Pierceable, heat-sealable foil seals (e.g., Bio-Rad Pierceable Foil Heat Seal).
11. Plate sealer capable of sealing for 5 s at 180 °C (e.g., Bio-Rad PX1™ Plate Sealer).
12. Thermal cycler, preferably one capable of a 4–12 °C hold and a heated lid.

2.4 Web Resources

1. UCSC genome browser (hg19): http://genome.ucsc.edu/cgi-bin/hgGateway.
2. 1000 Genomes browser: http://browser.1000genomes.org/index.html.
3. SNAP (SNP annotation and Proxy Search): https://www.broadinstitute.org/mpg/snap/.
4. Genotype-Tissue Expression (GTEx) Project: http://www.gtexportal.org/home/.
5. IDT OligoAnalyzer: http://www.idtdna.com/calc/analyzer.
6. Primer3Plus primer design tool: http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/ [7].

3 Methods

3.1 Selection of Reporter SNPs for Assays

The first step (*see* Fig. 1, Subheading 3.1) is to identify one or more reporter variants (transcribed heterozygous sites) in the gene of interest, which is generally chosen through hypotheses about the mechanisms of disease. These reporter variants are generally single nucleotide polymorphisms (SNPs) within the target gene's RNA transcript. The chosen SNP will be used just as a reporter of a transcript's chromosome-of-origin; it need not be the variant causing the allelic skew, nor necessarily correlated via linkage disequilibrium with that variant, though it may be useful in downstream analysis to know the relationship between the reporter and causal SNP.

Using the UCSC genome browser, enter the gene name of interest and select the assembly hg19 to focus on that locus. Under the "Variation" header at the bottom of the page, change the "1000G Ph1 Vars" track display setting to "full" and click "refresh." Any exonic variant appearing in this track, as well as those in the 5′UTR and 3′UTR, can be used (*see* Fig. 2). In addition, intronic SNPs can sometimes be utilized, especially if RNA-seq data indicates that the corresponding sequence is detected at a meaningful level in total RNA (*see* **Note 5**), reflecting that sequence's presence in cells for an appreciable period after transcription and before splicing.

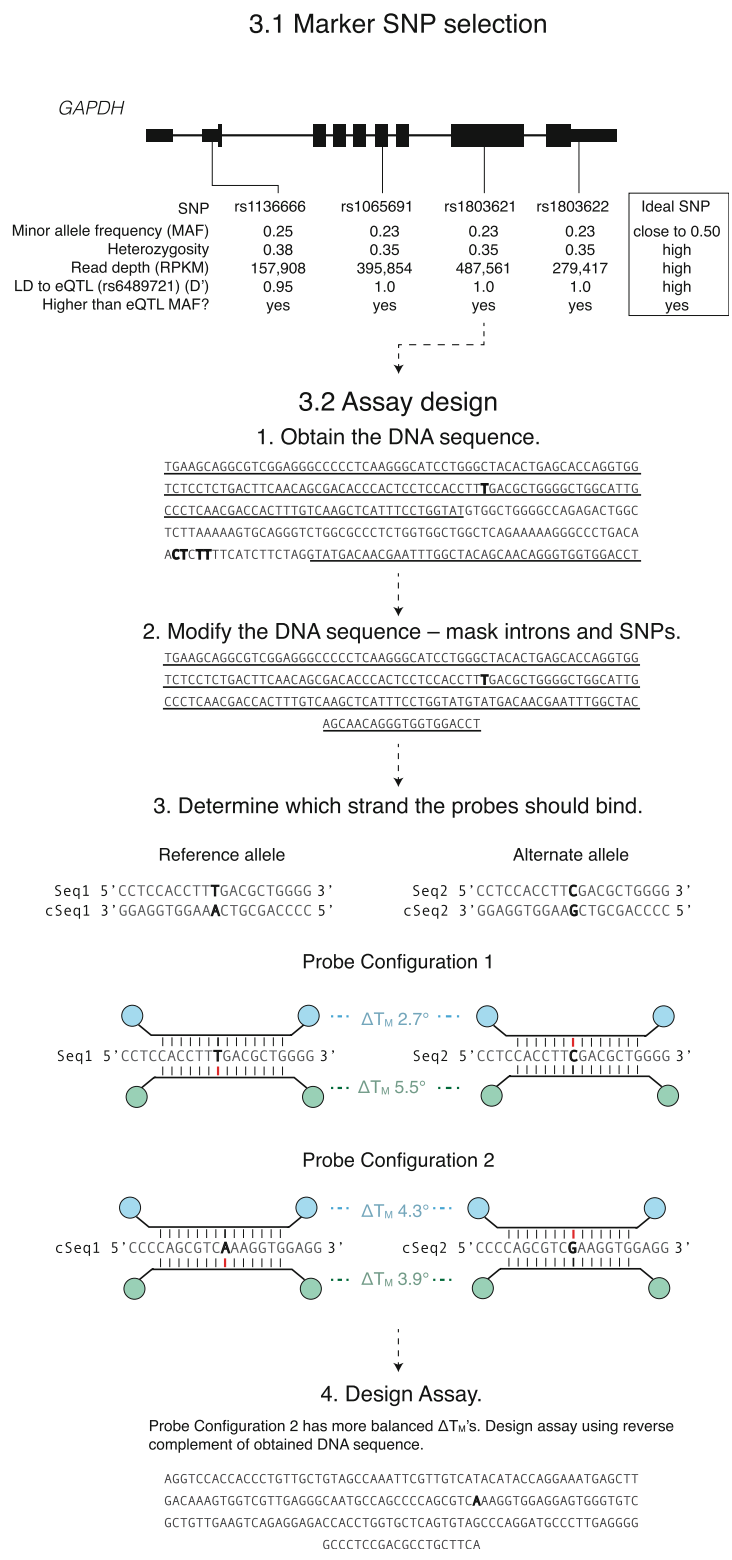


Fig. 1 A schematic following the steps of assay design for the *GAPDH* gene, as referenced sequentially in the chapter text

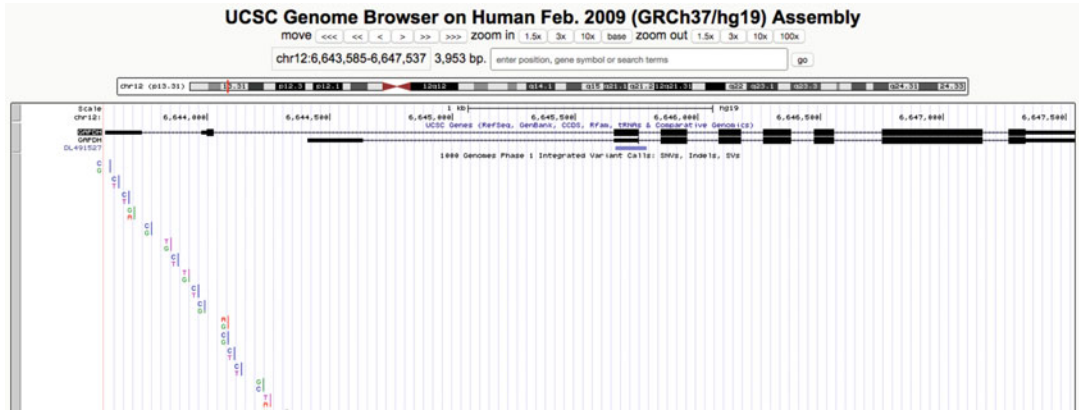


Fig. 2 A screenshot of the UCSC genome browser

1. Find the parameters needed for evaluating potential reporter SNPs. There are two main components to consider: the number of individuals that are heterozygous for the reporter SNP and the abundance of transcription at that SNP. Note that due to the evolutionary pressure on coding sequence variation, many genes may have only one or two common, high read-depth SNPs that are amenable for evaluating allelic skew.
   (a) Because homozygotes cannot be evaluated, determine or estimate the number of heterozygotes among the available RNA-sample donors. If genotype data are already available, this can be measured directly. If not, heterozygosity can be estimated from population-level data from a 1000 Genomes Project population sample whose ancestry matches your study population. Obtain the minor allele frequency (MAF) of the SNP using a source, such as the 1000 Genomes browser, and calculate the heterozygosity—$2\times(\text{MAF})\times(1\text{-MAF})$, which is derived from the Hardy–Weinberg equilibrium value 2pq. The greater the heterozygosity, the more samples will be usable.
   (b) Determine or estimate the level of expression of the potential reporter SNPs; this will be useful for determining the amount of RNA input and selecting a SNP that is substantially expressed, despite alternative splicing or potentially being in an intron (*see* **Note 5**). The average number of sequencing reads, termed read depth and often reported as RPKM or FPKM (reads or fragments per kilobase per million reads), that contain the exon/intron in which the reporter SNP resides can be used as a rough proxy for expression. The read depth can be found for any exon across a range of tissues at the GTEx website by searching for a gene and positioning the mouse cursor over the exon in the row corresponding to the desired

tissue type. The higher the read depth, the more the exon containing the reporter SNP is expressed, and the more informative that SNP will be given less input RNA.

(c) (**Optional**) If there is a prior hypothesis that a particular functional variant is generating allele-specific expression differences, then the linkage disequilibrium between this variant and any potential reporter SNPs should be calculated using the 1000 Genomes browser or SNAP in the population most genetically similar to that of the RNA-sample donors (e.g., samples from donors of Caucasian ancestry should be evaluated using European-ancestry populations such as CEU). A reporter SNP with a high D' (a measure of recombination) and an MAF greater than the candidate functional variant's MAF will be most helpful for validating that candidate and determining its direction of effect (*see* Subheading 3.5, **step 2** and **Note 6**).

2. Determine which potential reporter SNP(s) will be the most informative. A general guideline for ranking these SNPs is (heterozygosity)×(expression_level), which assigns roughly equal weighting to the number of heterozygotes, which influences sample size, and expression level, which influences measurement precision. With ample RNA (>20 ng) and moderate to high expression of the gene (>10 RPKM), the number of heterozygous samples available will be more important than the level of expression. Conversely, when a gene is expressed at low levels or a smaller quantity of usable RNA is available for analysis, it may be more important to select variants with higher expression levels. Select the top reporter SNP(s) based on these criteria. If there are multiple suitable reporter SNPs, it is useful to select two or more with low D' to each other, to increase the likelihood of having a reporter SNP with high D' to the functional variant (*see* Subheading 3.1, **step 1c**).

3.2 Assay Design

1. Obtain the DNA sequence flanking the reporter SNPs for assay design (*see* Fig. 1, Subheading 3.2, **step 1**). Enter the SNP coordinates into the UCSC genome browser and navigate to that genomic region. The browser should be zoomed in on that variant.

   (a) Obtain the surrounding DNA sequence by selecting View > DNA from the tool bar.

   (b) Under the options, add 100 extra bases upstream and 100 extra bases downstream.

   (c) Check the "Mask Repeats" option and select "Mask to N." This blocks the primer-design software from using

sequences that are repeated many times in the human genome.

(d) Hit the "extended case/color options" button and check both the "Underline/Human mRNAs" option and the "Bold/Common SNPs" option.

(e) Hit submit.

The resulting sequence will be 201 bp long, with the target SNP in the middle, all SNPs in bold, and the exonic mRNA sequence underlined. If not all of the sequence is underlined, increase the "extra bases" parameter (possibly up to 1000s of bases) until a total of 100 bases on either side of the SNP are underlined. Repeat-masked sequence does not count in this total.

2. Change the DNA sequence to exclude introns and SNPs (*see* Fig. 1, Subheading 3.2, **step 2**). Unless designing a reporter SNP assay to an intronic SNP, remove the intronic sequence (the sequence that is not underlined), leaving the mRNA transcript. To prevent the primer-design software from designing primers that bind to polymorphic bases, change all the bolded SNPs (except the target SNP) to N (*see* **Note 7**).

3. Determine which strand the probes should bind (*see* Fig. 1, Subheading 3.2, **step 3**). Selection of the best strand (plus or minus strand) for an assay increases the extent to which the correct allele-specific probe will outcompete the other probe when amplifying the target allele and is done by balancing each probe's ΔT_M (delta melting temperature = correct match T_M—mismatch T_M). Thus, the ΔT_M's for a pair of probes binding to one strand of both allelic sequences will be compared to a pair of probes binding to the other strand. To determine all possible ΔT_M's, select the sequence from 10-bp upstream to 10-bp downstream of the reporter SNP (Seq1). Copy/paste the sequence and change the reporter SNP to its alternate allele (Seq2). Obtain the reverse complement sequence for both (cSeq1 and cSeq2). Use IDT OligoAnalyzer with these settings for all following tests:

(a) Target type: DNA.

(b) Oligo Conc: 0.05 μM.

(c) Na^+ Conc: 50 mM.

(d) Mg^{++} Conc: 3 mM.

(e) dNTPs Conc: 0 mM.

Enter Seq1 and select the Tm Mismatch tool. Change the generated complementary sequence to cSeq2. Record the DELTA T_M. Repeat for Seq2, cSeq1, and cSeq2 with editing the generated complementary sequence to cSeq1, Seq2, and Seq1 respectively. If the DELTA T_M values when entering Seq1

and Seq2 are more equal than those for cSeq1 and cSeq2, proceed with the sequence already obtained. If cSeq1 and cSeq2 are more equal, then proceed with the reverse complement of the 201-bp mRNA segment from the previous step. If there were multiple potential reporter SNPs, prioritize on those with a higher sum of DELTA T_M for Seq1 and Seq2 or cSeq1 and cSeq2.

4. Design the primers and probe to the reference allele using the Primer3Plus primer design tool. Enter the 201-bp mRNA sequence obtained in Steps 1–3 into the box at the top of the webpage. Check "Pick hybridization probe (internal oligo)" under the input sequence. Under "General Settings," click the "Mispriming/repeat library" pull-down menu above the sequence box and choose "HUMAN." The following conditions are for an assay that will be run with an annealing temperature around 60 °C, but depending on local sequence surrounding the SNP (for example, when this sequence is AT-rich), it may be necessary to use a different annealing temperature (*see* **Note 8**).
   (a) Under "General Settings" set:
      - The Primer Size as Min: 18, Opt: 20, Max: 27.
      - The Primer Tm as Min: 57, Opt: 60, Max: 63.
      - The Primer GC% as Min: 20, Opt: 50, Max: 80.
      - The Concentration of divalent cations to 3.8.
      - The Concentration of dNTPs to 0.8.
   (b) Under "Advanced Settings" set:
      - The Table of thermodynamic parameters to "SantaLucia 1998".
      - The Salt correction formula to "SantaLucia 1998".
   (c) Under "Internal Oligo" set:
      - The Hyb Oligo size as Min: 15, Opt: 20, Max: 27.
      - The Hyb Oligo Tm as Min: 64, Opt: 65, Max: 70.
      - The Hyb Oligo GC% as Min: 20, Opt: 50, Max: 80.
      - The Hyb Oligo Mishyb Library to "HUMAN".
   (d) Force Primer3Plus to design the probe over the target SNP by setting the Hyb Oligo Excluded Region to "1,80 121,80".
   (e) Click "Pick Primers" and select a probe and primer set in which the probe does not start with a G (5′ G) and the probe overlaps the mutation of interest with at least four bases on either side.

(f) If no assay designs pass these design restrictions, relaxing the parameters can be attempted. However, designs requiring significant changes may require trying a different reporter SNP if one is available (*see* **Note 9**).

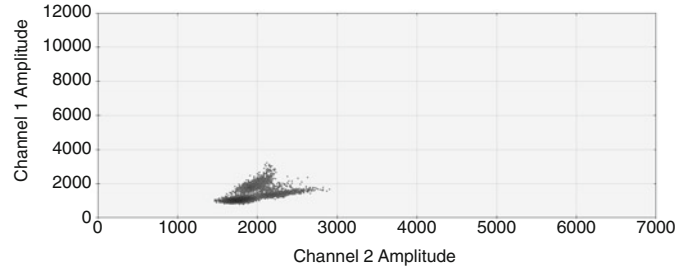
5. Design the probe to the alternate allele. Within the mRNA sequence, change the target SNP to the alternate (nonreference) allele. Copy and paste the left primer from Step 4 into the box below "Pick left primer, or use left primer below:" and copy and paste the right primer from Step 4 into the box below "Pick right primer, or use right primer below." Using the same primer design settings as Step 4, select a probe that does not start with a 5′ G and overlaps the mutation of interest with at least four bases on either side. Note, succeeding in the design of the first allele does not guarantee success with the other allele.
6. Ensure that the assay targets the correct region and is not predicted to have off-target amplification. Use the UCSC BLAT tool (under the "Tools" menu at the top of the page) to check that the two primers and reference-allele probe match only the region of interest perfectly (*see* **Note 10**). View the region with the "Common SNPs" track displayed to ensure the primers and probes do not bind over other SNPs. In addition, check whether the primers and probes are likely to bind each other by using the "Hetero-Dimer" option in the right menu bar of IDT's OligoAnalyzer; ΔGs lower than −7 should be avoided.
7. Order the primers and probe from your usual oligonucleotide supplier. For example, the 5′ FAM and 5′ HEX probe fluorophores, both with the Iowa Black 3′ ZEN quencher from Integrated DNA Technologies. When making the 20× assay mix, please note that the proportion of primers to probes is different than in qPCR.

3.3 RT-ddPCR for Determining Allelic Imbalance

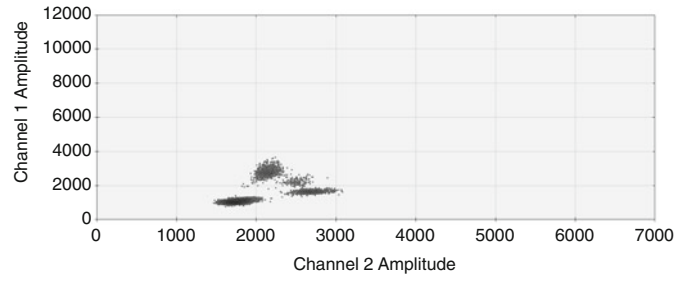
1. (**Optional**) Before running the assay on samples of interest, and particularly if the DELTA TM values from Subheading 3.2, **step 3** were low, it may be useful to run trial reactions with varying annealing temperatures using heterozygous genomic DNA (*see* Fig. 3) and with varying RNA inputs to optimize cluster separation (*see* **Notes 16** and **17**).
2. Thaw the RNA and the reagents of the One-Step RT-ddPCR Advanced Kit for Probes on ice for 30 min. Vortex and centrifuge briefly.
3. For each reaction, assemble:

   6.25 μL Supermix (One-Step RT-ddPCR Advanced Kit for Probes).

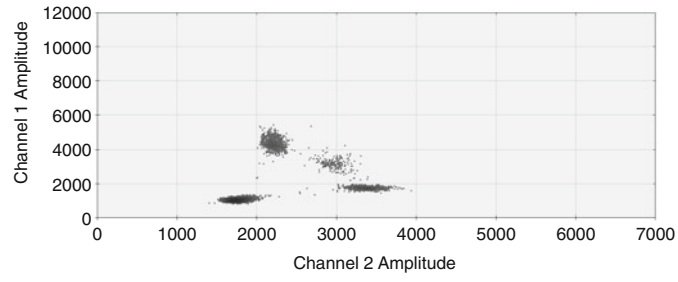
a. 60°C Annealing Temperature



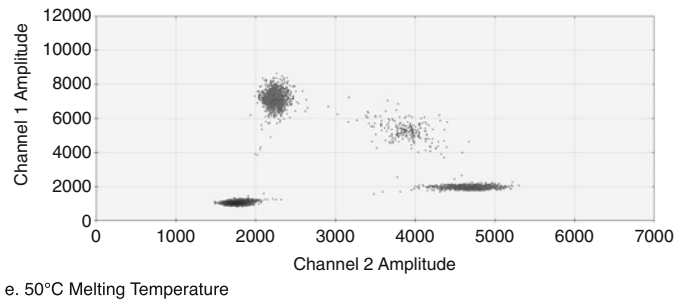
b. 58°C Annealing Temperature



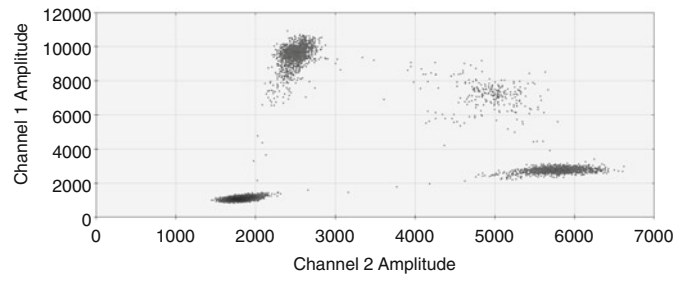
c. 56°C Annealing Temperature



d. 53°C Annealing Temperature



e. 50°C Melting Temperature



2.5 μL Reverse Transcriptase (One-Step RT-ddPCR Advanced Kit for Probes).

1.25 μL 300 mM DTT (One-Step RT-ddPCR Advanced Kit for Probes).

μL of 20× assay targeting the SNP.

variable input of RNA (100 fg–100 ng per reaction) (*see* **Note 3**).

<13.75 μL of DEPC-treated water.

25.0 μL total volume (*see* **Note 11**).

*Some may find it helpful to run a control sample with heterozygous DNA to confirm where the droplets cluster. In addition, because the allelic imbalance statistics can be run using a single reaction, duplicates of a sample are not necessary unless the RNA is rare, necessitating combining several reactions to have enough positive droplets to make a definitive call.

4. Vortex and spin the plate to collect the liquid at the bottom of wells. Keep the plate protected from light until droplet generation, and allow the reactions to equilibrate to room temperature for 3 min prior to droplet generation (*see* **Note 12**).
5. Place a DG8 cartridge into the QX200 droplet generation cartridge holder (Bio-Rad).
   (a) Pipette 20 μL of the PCR mix into the middle row of the cartridge. Only push down to the first stop when ejecting liquid to ensure there are no air bubbles in the sample (*see* **Note 13**). Using a Rainin multichannel pipettor with Rainin tips is preferred at this stage (*see* **Note 4**).
   (b) Pipette 70 μL of oil into the bottom row of the cartridge using a new set of tips. Always be sure to pipette the oil after the samples. The top row (where droplets are outputted) is left empty.
   (c) Place a DG8 rubber gasket over the cartridge.
6. Place the cartridge holder with cartridge and gasket into the QX200 droplet generator. Close the generator and droplets will be formed.
7. When the triangles on the button on the droplet generator lid return to being lit solid green, remove the cartridge. Carefully remove its gasket and transfer the droplets in the top row to a clean, half-skirted Eppendorf plate. It is important that the pipetting at this stage is slow and careful with the pipette

Fig. 3 A temperature gradient run using the same sample and assay. Note how the clusters are intermingled in (**a**), yet increase in separation with increasing annealing temperature in **(b)**, **(c)**, and **(d)** until an optimal temperature is reached in **(e)**. If this gradient were run with values below 50 °C, we would see the clusters start to intermingle again as we move away from the optimal annealing temperature

oriented at 45 degrees, otherwise the droplets may shear. Afterward, discard the gasket and cartridge (*see* **Note 14**).

8. After all droplets are made, seal the droplet plate with a foil seal by heating the top to 180 °C for 5 s. Check to make sure that the outlines of the wells are visible through the foil, indicating a good seal.
9. Thermal cycle as follows:

   42–50 °C for 60 min (*see* **Note 15**).

   95 °C for 10 min.

   40 cycles of 95 °C for 30 s followed by 60 °C (or identified optimal annealing temperature; *see* **Note 16**) for 1 min.

   98 °C for 10 min.

   4 °C hold.

   Use a 2.5 °C per cycle ramp rate for all steps and set to 40 μL volume. Droplets can be stored protected from light at 4 °C after cycling for up to 24 h before reading.
10. Set up a template on the QX200 droplet reader computer. Open QuantaSoft. Under "Template" in the top left corner, select "New" in order to fill in a new plate map. Double click on the first well. In the "Sample" box, under "Experiment," select any of the "ABS" experiments, and then, under "Supermix," select "One-Step RT-ddPCR Kit for Probes." In the "Target 1" box, enter the name of the FAM assay in the "Name" field and select "Ch1 Unknown" from the "Type" menu. In the "Target 2" box, enter the name of the HEX or VIC assay in the "Name" field and select "Ch2 Target" from the "Type" menu. Select all wells of the plate that will contain samples that are using the same assays. Click the blue "Apply" button in the top window. The name of each sample may also be entered. Once finished, click "OK" and save the template.
11. Read the droplets on the QX200 droplet reader. Put the plate into the plate holder of the QX200, then place the black plate holder on top and click the silver tabs on either side down into place, making sure the A1 well is in the top left corner. Close the lid of the QX200. In QuantaSoft on the QX200 computer, make sure the template created in Step 9 is loaded, then click "Run." On the popup menu that appears, select "FAM/HEX" or "FAM/VIC," depending on the pair of fluorophores used, then click "OK."

3.4 Data Finalization and Quality Control

1. After the ddPCR is complete, perform a well-by-well visual inspection of droplet clusters in QuantaSoft by clicking "Analyze" on the leftmost menu followed by "2D Amplitude." Verify that there is clear separation between the positive and negative clusters for both the reference and alternate allele

fluorophores. Some droplets between the positive and negative clusters are acceptable, but if the clusters themselves overlap, it will undermine the accuracy of the measurements. If all the wells have bad cluster separation, first try optimizing the annealing temperature and RNA concentration (*see* **Notes 16** and **17,** Fig. 4). If the assay is still not generating clean data, redesign the assay to the same or a different reporter SNP (*see* Subheading 3.1).

2. Evaluate and adjust the software's automatic calling for each droplet cluster in each well. Droplets in the top left corner of the 2D amplitude plot are FAM+, droplets in the bottom right corner are VIC+ (or HEX+), droplets in the top right corner are double positive, and droplets in the bottom left corner are double negative. Though ddPCR is usually highly accurate for assays targeting independent loci (as in CNV analysis, in which the two probes interrogate different nucleic acid sequences), for allele-specific expression, where the two sequences of interest are highly similar, the signal in the two fluorescence channels is often not independent, due to the probes sometimes binding the opposite allele. The lack of rectangularity in the cluster locations (in two-dimensional fluorescence space) may

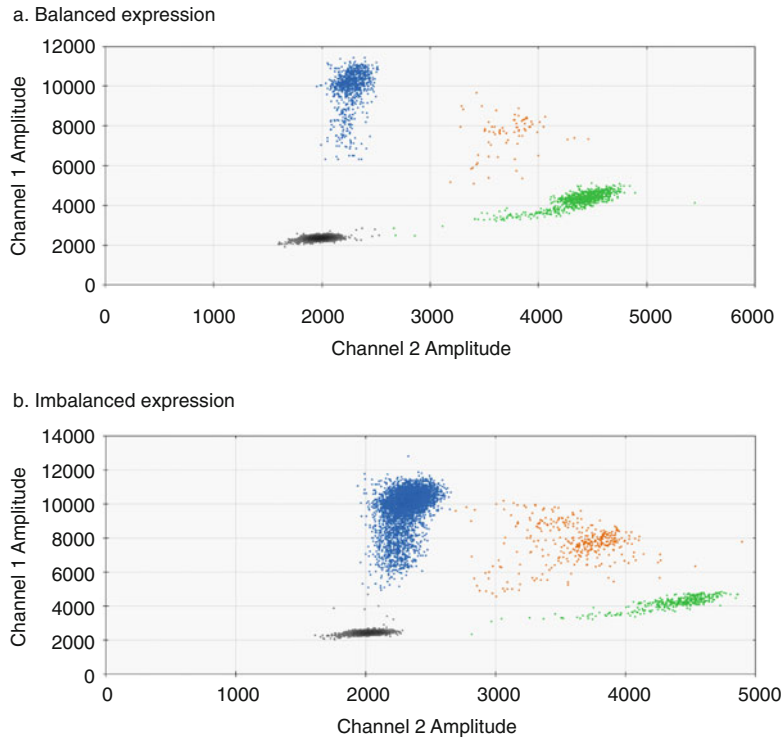


Fig. 4 (**a**) A sample where the alleles (blue and green) appear to have equal expression. (**b**) A sample where the alleles have 90%:10% expression (blue:green)

complicate automated clustering and require interactive visual clustering by the user. To adjust cluster assignment in an unbiased fashion, highlight all the wells containing the same assay on the plate view in upper left and demarcate cluster boundaries identically across all samples on the 2D amplitude plot by using the "Threshold" or "Lasso" tools. If the "Status" column is set to "Check," click anywhere in the amplitude plot to get the software to recognize the data (*see* **Note 18**).

3. Export the data by selecting the wells and clicking "Export CSV." The columns "Ratio" and "FractionalAbundance" correspond to different measures of allelic skew, (Allele A/Allele B) and Allele A/(Allele A + Allele B), respectively. The columns following these with labels such as "PoissonFractionalAbundanceMax" and "PoissonFractionalAbundanceMin" represent the upper and lower 95% confidence intervals for Allele A/(Allele A + Allele B), respectively (*see* **Note 19**).

3.5 Using the Observed Allelic Imbalance to Map Functional Variants

1. Using the "PoissonFractionalAbundanceMax" and "PoissonFractionalAbundanceMin" values, we can determine which samples have significant allelic imbalance. Specifically, these values define the 95% confidence interval around the estimate of allelic imbalance recorded under "FractionalAbundance" and thus identify any samples that do not overlap 0.5 (which represents equal relative expression from each allele). The limit of detection in terms of effect size is, of course, dependent on the level of expression and RNA input, but for a moderately expressed gene with 10–20 ng of RNA (roughly 3000 or more positive droplets total), we can confidently determine effects down to a 10% difference in expression. Ideally, all imbalanced samples will display ratios with consistent absolute distances from 0.5, suggesting a single common variant affecting expression across the population. A wide variance in the magnitude of distance from 0.5 suggests multiple common influences acting on this gene, making each effect difficult to identify in isolation.
2. Map the functional SNP by correlating the presence of imbalance with each putative functional SNP's heterozygosity. In short, samples that have balanced expression of the reporter SNP are homozygous for the functional variant(s), while samples that are imbalanced are heterozygous (*see* Fig. 5a). This can be used to map the functional variant by finding the regional SNPs that are always homozygous when there is balanced expression and heterozygous when there is imbalanced expression (*see* Fig. 5b) (*see* **Note 20**). If all the imbalanced samples are imbalanced in the same direction, this provides not only increased power to find the functional variant, but also allows the easy determination of which allele of the functional SNP is associated with higher expression (*see* **Note 21**). However, this

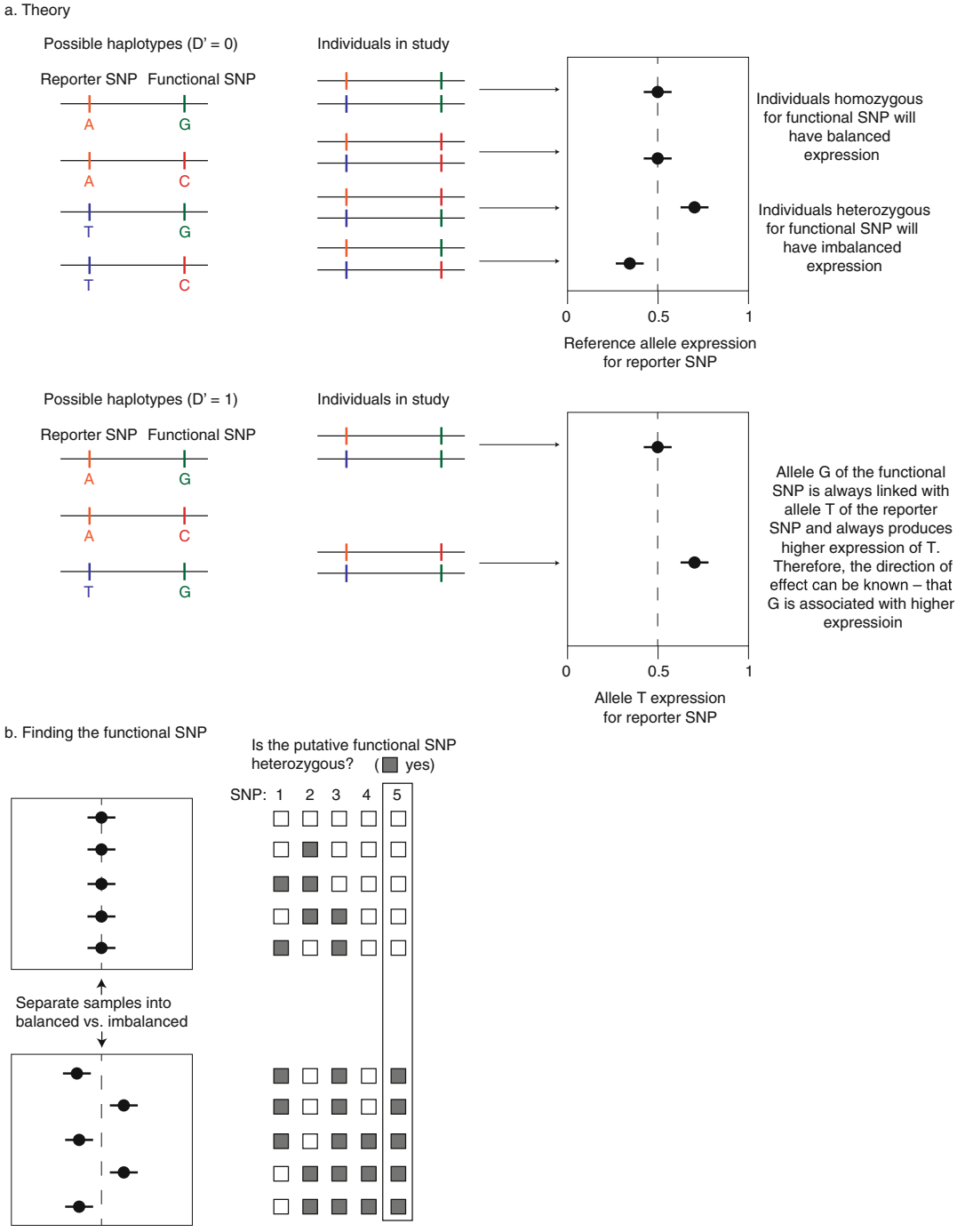


Fig. 5 (**a**) The possible haplotypes for a reporter and functional variant with a D' of 0 or 1, and the pattern of imbalance those configurations would yield. (**b**) The technique used to map the functional variant by correlating the presence of imbalance with heterozygosity

knowledge can also be obtained by determining the chromosomal phase molecularly in the DNA samples using methods we have previously described [8]. A chi square test should be used to determine the significance of the putative functional SNP.

4 Notes

1. We have found that standard TaqMan probes generally perform well, but MGB or LNA probes may also be used. FAM and HEX probes can be ordered through Integrated DNA Technologies, and VIC probes can be ordered through Life Technologies. HEX and VIC fluorescence is read in the same channel in ddPCR, so either an HEX or VIC probe should be used with a FAM probe.
2. For ease of use, 20× mixes can be stored at 4 °C for short periods (~1 month).
3. Because the level of expression across different genes varies by orders of magnitude, there is not a single RNA input concentration that will work for all genes. We generally start with an input of 10 ng and adjust the concentration up or down until we obtain 10–40% positive droplets in ddPCR, which limits the likelihood of the clusters overlapping by limiting the size of those clusters, and provides enough double-negative droplets to confidently call clusters. This may require large amounts of RNA for genes with low expression. Note that this sample concentration is lower than most ddPCR assays.
4. Low-quality pipette tips can shred droplets during droplet generation, likely because they shed tiny pieces of plastic into the reaction. Though not strictly required, Rainin pipettors and tips are preferred when working with droplets.
5. At minimum, the SNP has to be within the transcript of the gene and preferably in an exon that is expressed in all isoforms (unless only one isoform is of interest). We have found that intronic SNPs can also be used, especially when they reside in long introns and are far from the 3′ ends of such introns, ensuring slower degradation rates [9]. Assays on intronic SNPs always require a total-RNA sample. Such assays also tend to require more RNA input and are most feasible for highly expressed genes (RPKM > 25). The GTEx website has expression information for introns as a downloadable file under the "Datasets" tab on "Download" page after registering for an account. The current release is "GTEx_Analysis_V4_RNA-seq_Flux1.6_intron_reads.txt.gz" at the time of writing. However, the data is given as read counts rather than equivalent RPKM on each gene page and needs to be further manipulated

for direct comparison. Briefly, divide each row by the length of the intron given in the first column, divide each column by the number of unique mapped reads in that library (found in "GTEx_Data_V4_Annotations_SampleAttributesDS.txt" under "SMMPPDUN" column label), and multiply by one billion (to correct for base/kilobase and read/(million read) conversion).

6. When the reporter SNP is more common than the proposed functional variant, particularly if $D' = 1$, it is possible to determine or validate the proposed functional variant by correlating the heterozygosity of the other variants to the allelic imbalance observed at the reporter. However, if the reporter SNP and actual functional variant have $D' = 1$ and have similar allele frequencies (i.e., when $r^2 = 1$), then there will be a limited set of possible genotypes and most samples heterozygous for the reporter SNP will have imbalance. Without any samples that are balanced and heterozygous, mapping is nearly impossible (because there is no balance/imbalance vs. heterozygosity to correlate). If prioritizing on finding a reporter SNP with $D' = 1$ to the functional variant, then try to find a reporter SNP where the allele linked to the minor allele of the functional variant is twice as common as the functional variant's minor allele frequency. For example, if the functional variant has an MAF = 0.05, then a reporter SNP with $D' = 1$ and overlapping reporter SNP allele frequency of 0.5 will yield 50% usable reporter SNP heterozygotes, but with 45%:5% being balanced:imbalanced. A reporter SNP with $D' = 1$ and overlapping reporter SNP allele frequency of 0.1 will yield fewer heterozygotes (18%), but more observations of imbalance (9%:9% being balanced:imbalanced), thus increasing the power of mapping.
7. SNPs in the primer or probe binding sites can prevent or hinder amplification. The sites of common SNPs must be excluded from the sequence used to design assays.
8. A temperature of 55 °C is another annealing temperature to consider for assay design at loci with lower nearby GC content. Use the following settings, particularly if design selection at 60 °C is difficult, while keeping the other settings constant:
   (a) Under "General Primer Picking Conditions" set:
      - The Primer Size as Min: 18, Opt: 20, Max: 27.
      - The Primer Tm as Min: 55, Opt: 55, Max: 57.
      - The Primer GC% as Min: 20, Opt: 50, Max: 80.

(b) Under "Internal Oligo (Hyb Oligo) General Conditions" set:
- The Hyb Oligo size as Min: 15, Opt: 20, Max: 27.
- The Hyb Oligo Tm as Min: 56.5, Opt: 57.5, Max: 59.
- The Hyb Oligo GC% as Min: 20, Opt: 50, Max: 80.

9. Generally the hybridization probe will be the source of design failure due to the constraint that the probe sequence must overlap the variant. To identify the specific requirement (s) that the surrounding sequence does not meet, enter the 21 bp sequence from Subheading 3.2, **step 3** into the box below "Pick hybridization probe (internal oligo), or use oligo below" and click "Pick Primers." The warning message at the top of the design will help troubleshoot the design. If problematic, parameter relaxing might first begin with decreasing "Hyb Oligo Max Mishyb" and increasing "Internal Oligo Max Poly-X," both under the "Internal Oligo" tab. If other settings are leading to design failure, adjustments such increasing annealing temperature range slightly can still lead to a usable assay, but consider designing assays for other reporter SNPs.
10. If the primers are too short for BLAT, use the in silico PCR function ("In-Silico PCR" under the UCSC genome browser "Tools menu"). If the region of interest is in a segmental duplication, one unique and perfect match may not be possible. Try choosing a different reporter SNP if the other duplications are also expressed.
11. Though the ddPCR instructions have 20 μL as the listed final reaction volume before droplet generation, scaling reaction volumes up by 25% ensures fewer air bubbles are introduced in Subheading 3.2, **step 4a**.
12. ddPCR droplets are made in sets of eight samples at a time and read in 96-well plate format, so it is easiest to set up the PCR in a 96-well plate.
13. Pushing the pipette down to the final stop introduces air bubbles, which compromise the number and quality of droplets. If air bubbles are introduced into the sample chamber, manually pop them with a clean pipette tip.
14. An automatic droplet generator (Bio-Rad QX200 AutoDG) and associated consumables can be used instead of manual droplet generation.
15. Reverse transcription temperature can be set to a higher temperature ($\leq$50 °C) to denature any secondary structure that may prevent conversion to cDNA, but using lower temperatures ($\geq$42 °C) can help reduce RNA degradation in solution.

16. Though assays are designed to work best at a particular annealing temperature (60 °C), many SNP assays yield cleaner data at another temperature. Before running the assay on the samples of interest, run a set of reactions containing identical reagents and input sample on a gradient of melting temperatures (*see* Fig. 4). Select the melting temperature with the best cluster separation. An ideal plot will have tight clusters with ample room to draw separating thresholds. The clusters need not lie perfectly horizontal or vertical. In fact, it is expected that the probes will cross-react with the other allele, causing the trajectory of the single-positive droplets to form an acute angle, rather than the normal 90° right angle, but this will not affect the measurements as long as the four clusters are separate.
17. Because the correct identification of droplets containing one or both alleles depends on the competitive binding of the probes to sequences with which they are both similar, cluster separation can be more challenging. This is ameliorated by lower RNA inputs, which yield fewer double-positive droplets. If any two clusters overlap, run a range of reactions with decreasing sample input until there is clear separation. Allele-specific expression analysis benefits from using alleles as natural controls for each other, and will often require less sample input than total expression assays to obtain equivalent power in detecting the same regulatory effect. An RNA sample can also be assayed across multiple wells, with the resulting allele counts then combined across wells to increase the resolution/precision of the analysis.
18. The software only calls wells with 10,000 or more droplets and sets the "Status" column to "Check" for wells with fewer droplets.
19. "FractionalAbundance" and the related confidence interval defined by the min and max columns may better depict the magnitude of deviation as compared to the nonlinear distances between ratios. For example, if one allele generates 1 transcript to the other allele's 2 transcripts, the "Ratio" column would read either 0.5 or 2, depending on the phase between the reporter and functional variants. Whereas "FractionalAbundance" would read either 33.3 or 66.6—depending on phase—and thus more intuitively depict the same effect size.
20. Some samples could have a chromosome containing a rare and/or recent recombination event that breaks the expected LD pattern seen in most individuals of similar ancestry. Because allele-specific imbalance is more resistant to variation arising from genetic and environmental *trans*-acting factors (*see* **1**), a single sample can be confidently called as exhibiting allelic imbalance or not. Based on the allelic imbalance of that sample

resembling that of other samples with equivalent haplotypes on each side of the recombination event, this may suggest or eliminate sets of variants normally segregating on the same haplotype.

21. If all samples with significant allelic imbalance are imbalanced in the same direction, then we automatically know which allele is expressed more highly. This is because the reporter SNP and the functional SNP have a $D' = 1$. Effectively, of four possible combinations between two variants, a $D' = 1$ implies that one or two of them are never seen. The reporter SNP heterozygotes have only two possible genotypes (*see* Fig. 5a), resulting in either balance or imbalance. This is different from the case where D' does not equal 1, where two different genotypes result in imbalance and its not clear from genotypes alone which allele is associated with higher expression.

Acknowledgments

Our understanding of allelic skew has greatly benefited from the work of Tom Mullen and Jim Nemesh in our lab. This work was supported by a grant from the National Human Genome Research Institute (R01 HG006855, to SAM).

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106:9362–9367. https://doi.org/10.1073/pnas.0903103106
2. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073. https://doi.org/10.1038/nature09534
3. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, Ahituv N, Pennacchio LA, Shendure J (2012) Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol 30:265–270. https://doi.org/10.1038/nbt.2136
4. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002) Detection of regulatory variation in mouse genes. Nat Genet 32:432–437. https://doi.org/10.1038/ng992
5. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148:1293–1307. https://doi.org/10.1016/j.cell.2012.02.009
6. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res 24:14–24. https://doi.org/10.1101/gr.155192.113
7. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35:W71–W74. https://doi.org/10.1093/nar/gkm306

8. Regan JF, Kamitaki N, Legler T, Cooper S, Klitgord N, Karlin-Neumann G, Wong C, Hodges S, Koehler R, Tzonev S, McCarroll SA (2015) A rapid molecular approach for chromosomal phasing. PLoS One 10:e0118270. https://doi.org/10.1371/journal.pone.0118270

9. Gray JM, Harmin DA, Boswell SA, Cloonan N, Mullen TE, Ling JJ, Miller N, Kuersten S, Ma YC, McCarroll SA, Grimmond SM, Springer M (2014) SnapShot-Seq: a method for extracting genome-wide, in vivo mRNA dynamics from a single total RNA sample. PLoS One 9:e89673. https://doi.org/10.1371/journal.pone.0089673