RosBREED's Community Breeders' Page

Predicting Performance in Fruit Crop Breeding

By Cameron Peace, MAB Pipeline Team Leader

Crossing, selection advancement, new cultivar release, cultivar adoption, and even new cultivar management decisions can be facilitated using knowledge of distributions and their accompanying probabilities to predict the likelihood of achieving performance targets. Breeding decisions already rely on these observed and predicted distributions, whether you draw them out or not. Comparisons between two or more selections or between selections and market-leading cultivars support decisions about which selections to advance and which to discard. Statistical tests use the attributes of the distributions to determine whether the performance of one individual is significantly different from another or whether they just are samples from the same distribution.

Frequency distributions

During the MAB in Action workshop on 30 July 2012, frequency and probability distributions were often used to help elucidate concepts in merging DNA information with conventional breeding. Likewise, many statistical calculations within RosBREED's Breeding Information Management System rely on them. Wise use of these statistical instruments can support breeding decisions, from crossing to cultivar release and beyond, through formulating performance predictions for populations and individuals.

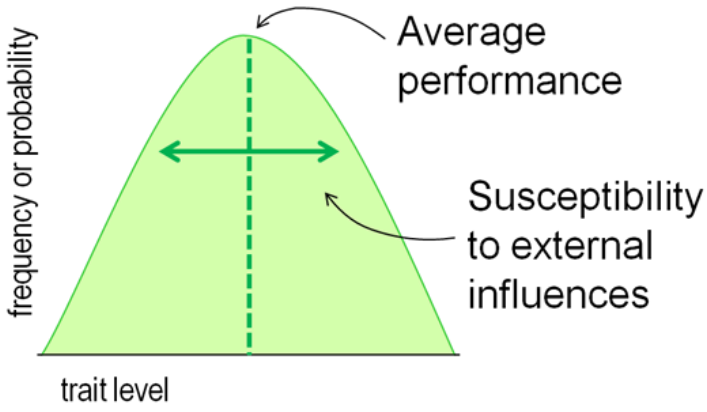

Figure 1

The illustration in Figure 1 shows a typical distribution for a single trait measured multiple times. The trait level or performance, e.g., for fruit size or sweetness, is displayed on the X-axis, increasing from left to right. The multiple measurements may result from growing an individual (cultivar, advanced selection, or seedling) at different sites, from observations made on the same plant during different years, from repeated measures (e.g. multiple fruit) in one growing condition, or from many individuals in a segregating population. The frequency of each measurement level is displayed on the Y-axis. The pertinent features of distribution curves are mean, standard deviation, and shape. Shape most often is conceptualized as bell-shaped although other distributions can be observed and modeled.

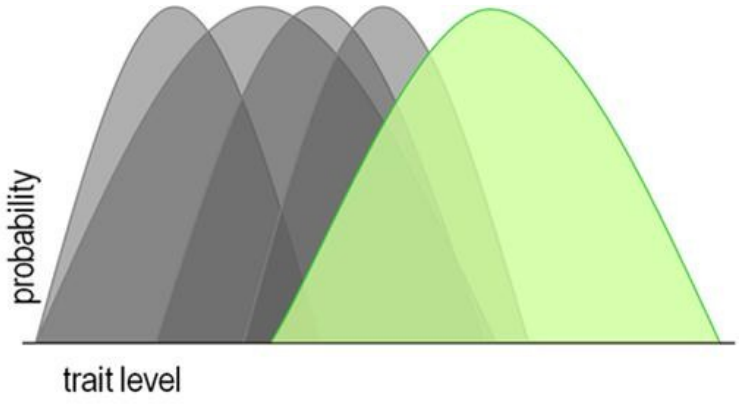

Figure 2

The compelling interest of a breeder should be to know the genetic value that underlies differences in performance distributions among individuals or families. Each individual, e.g., a cultivar, advanced selection, or seedling, will have its own performance distribution (Figure 2). The distribution for an individual with the best performance among those shown in Figure 2 (assuming a higher trait level is desired) appears in green, being better than its peers – perhaps other cultivars – which are skulking in the background in gray. A similar set of distributions could instead represent the performances of several families.

The middle of the curve, the mean, is the average performance potential. The performance potential is the observed or expected performance that an individual (or family) can achieve due to its genetic value when grown under various combinations of fixed effects like location and orchard management system (things you can control and account for) and random effects like weather that may be outside of our control. For an individual (or family), the amount of variation around the mean, shown as the standard deviation, reflects the influence of non-genetic factors on the genetic value of the individual or family. For a family, each observational or predicted datapoint represents the performance of a seedling (phenotypes), and the frequency distribution includes both genetic and non-genetic variation in the family.

Adding DNA information to frequency distributions

DNA information also can be added to these distributions. Imagine there is a trait controlled by a single locus with very high heritability, such as flesh color (white/yellow) in peach. If you were to record performance of a diverse set of cultivars as either white or yellow, you'd get just two vertical lines because all observations would be just one or the other performance level for flesh color (Figure 3a). If instead you record the degree of yellowness quantitatively with a colorimeter, you'd see some variation around each mean but the two distributions wouldn't overlap (Figure 3b). Now let's say there is a DNA test that can distinguish the two functional genotypes (**Y-** for white and **yy** for yellow) – you will be very confident that an unseen tree with the **Yy** genotype (shown by the DNA test) will be white-fleshed and not yellow, whether you recorded color simply by eye or through a fancy gizmo. You will be confident because the distributions are well separated. What about when two types for a trait have distributions that are closer together?

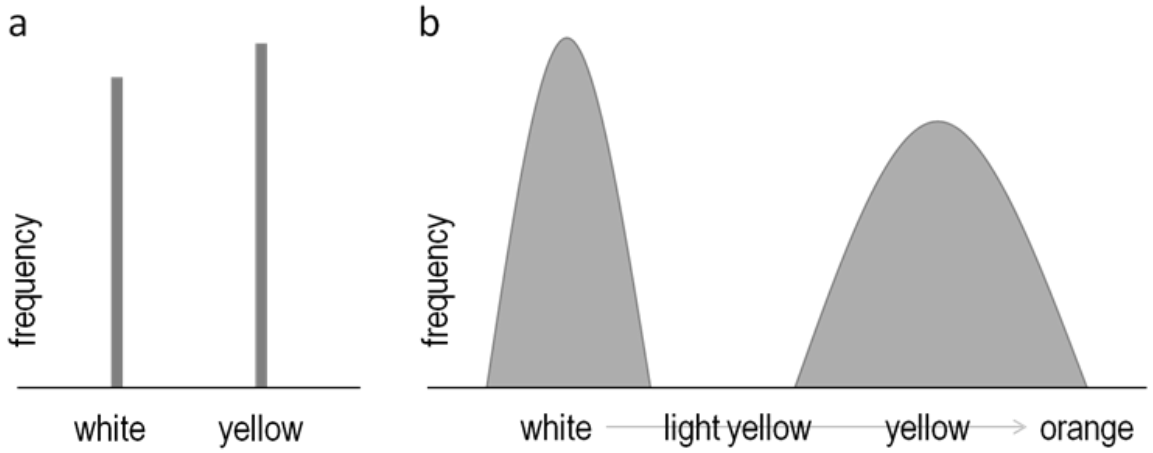

Figure 3

Consider a quantitative trait locus (QTL) with an allele in a selection population that is estimated to explain 30% of the observed (phenotypic) variation. What does that look like as a pair of distributions, where one group carries the allele and the other group doesn't? How much do the distributions overlap? Quite a lot (Figure 4a). In the earlier case of peach flesh color, the QTL explained almost all of the variation, perhaps 90% - enough that the have and have-nots can be distinguished by eye (and thus we call it a Mendelian Trait Locus, a special type of QTL). In the case of the 30% QTL, the blurring of the two genotypes, due to unaccounted-for factors like other variable influencing trait loci or environmental effects, reduces confidence in the distinction. Although the distinction is no longer high, it's better than a 50/50 guess. It's a 65/35 guess - at least helpful. Given that many traits have a heritability not exceeding 30%, conventional Rosaceae breeding takes many guesses that are 65/35 and worse. If the 30% QTL was for a 30% heritable trait (sweetness is a good example), a DNA test for it would be an excellent tool because in just that simple assay it would capture everything that phenotypic evaluation (typically more laborious and often years later) would be able to distinguish.

Adding thresholds make distributions really come alive! You breeders deal with selection thresholds for traits all the time. Placed on a performance distribution, a threshold allows the *frequency or probability of achieving the threshold* to be calculated. Considering our 30% QTL situation, a threshold that is achieved by half of the individuals carrying the desirable allele (shown in blue, Figure 4b) is only achieved by about 1 in 10 of the non-carriers (shown in green). Calculating such probabilities allows objective consideration of alternative choices for achieving breeding goals. If one choice gives you a 50% chance of success for a selection target and another choice gives you a 10% chance, you can objectively determine the relative amount of resources to devote to each.

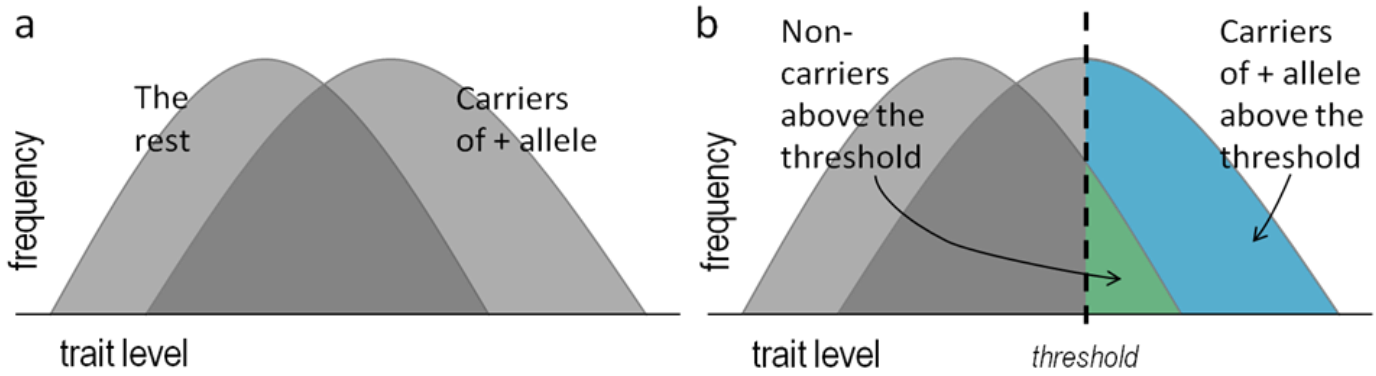

Figure 4

A real example: skin color in apple

The *Rf* locus in apple is strongly associated with the degree of red coloration (blush and/or stripes) on the fruit skin. RosBREED's observed phenotypic distributions for the basic three SNP-determined functional genotypes of RfRf, Rfrf, and rfrf (Figure 5) lead to useful predictions.

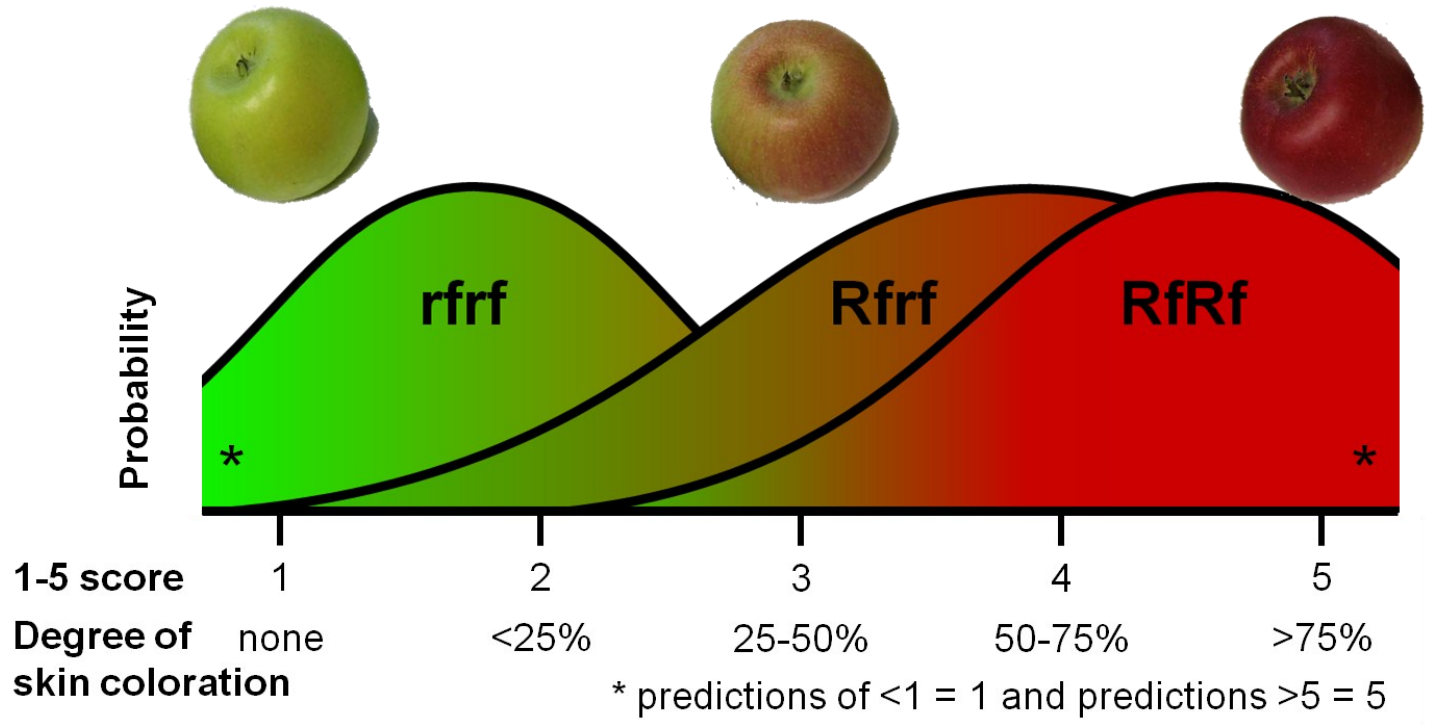

Figure 5

Let's say you are aiming for seedlings with blushes or stripes covering at least half the fruit surface so that you can then focus on selecting for other traits within such well-colored seedlings. Obviously RfRf × RfRf would be best, but...

Q1) What proportion of seedlings from an RfRf × RfRf cross would you expect to achieve your skin color target?
Q2) How much more efficient would such a cross be than RfRf × Rfrf, especially if you had plenty of Rfrf parents but few RfRf?
Q3) How would you even know the functional genotypes of your potential parents?

Probability distributions and jewel polishing to the rescue!

A1) For RfRf × RfRf, 79% of seedlings are expected to have a 4 or 5 on the 1-5 scale described in the diagram. 93% are expected to have at least a 3.5 score.
A2) RfRf × RfRf is 25% more efficient than RfRf × Rfrf (because you would need an estimated 127 seedlings of RfRf × RfRf to achieve 100 that had a score of 4+, while for RfRf × Rfrf you'd need 159 seedlings).
A3) RosBREED's genome scans have provided this information for all individuals of the reference germplasm sets. DNA tests specific for the *Rf* locus (so you would only need to run a single marker) are being finalized now, or you can use the DNA test of Takos et al. (2006) employed by Zhu et al. (2011).

Using* Cross Assist *in a breeding program

Cross Assist, the parent selection decision support module of the Breeding Information Management System, uses this same approach to compare efficiencies among all possible parental combinations of achieving trait target thresholds in the next generation. Probabilities for simultaneously achieving multiple trait targets are multiplied together to give a joint probability. The higher the joint probability, the fewer seedlings in a segregating population would be needed to find some that achieve all the targets, and so the better a cross's predicted efficiency. Currently, *Cross Assist* accesses the burgeoning phenotypic, breeding value, and functional genotype datasets of RosBREED's reference germplasm. While this germplasm does to some extent represent that of any U.S. breeding program of apple, peach, cherry, and strawberry, you might want to tailor the software tool even more to your own germplasm: if so, let us know.

Accuracy of predictions

Accurate predictions are made on relevant and large underlying datasets, analyzed in clever ways. Large datasets are achieved with large collaborations like RosBREED. Relevancy is achieved with breeder involvement to ensure that observed performance reflects target traits under target conditions. We hope to increase the accuracy of predictions as we amass more data, involve more U.S. Rosaceae breeding programs, and develop increasingly sophisticated software-based analytical and display tools. Stick around! Or better yet, hop on board! I predict we can increase the performance of your breeding outputs.

Acknowledgement

Fred Bliss, Scientific Advisory Panel member, provided a helpful revision of this article. For the *Rf* locus predictions, RosBREED's apple breeding team collected the phenotypic data and the Genotyping Team collected the genome-wide SNP data including numerous SNPs across the *Rf* locus.

References cited

Takos AM, Jaffé FW, Jacob SR, Bogs J, Robinson SP, Walker Amanda R (2006) Light-induced expression of a *MYB* gene regulates anthocyanin biosynthesis in red apples. Plant Physiology 142:1216–1232

Zhu Y, Evans K, Peace C (2011) Utility testing of an apple skin color MdMYB1 marker in two progenies. Molecular Breeding 27:525–532