ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2011) **9**(2); 159–162
doi:10.1017/S1479262111000104

Genomic tools for the analysis of genetic diversity

J. Antoni Rafalski*

DuPont Agricultural Biotechnology Group and Pioneer Hi-Bred International, Wilmington, DE 19880-0353, USA

Abstract

We now understand that many different types of DNA structural polymorphisms contribute to functional diversity of plant genomes, including single nucleotide polymorphisms, insertions of retrotransposons and DNA transposons, including Helitrons carrying pseudogenes, and other types of insertion–deletion polymorphisms, many of which may contribute to the phenotype by affecting gene expression through a variety of mechanisms including those involving non-coding RNAs. These polymorphisms can now be probed with tools such as array comparative genomic hybridization and, most comprehensively, genomic sequencing. Rapid developments in next generation sequencing will soon make genomic sequencing of germplasm collections a reality. This will help eliminate an important difficulty in the estimation of genetic relationships between accessions caused by ascertainment bias. Also, it has now become obvious that epigenetic differences, such as cytosine methylation, also contribute to the heritable phenotype, although detailed understanding of their transgenerational stability in crop species is lacking. The degree of linkage disequilibrium of epialleles with DNA sequence polymorphisms has important implications to the analysis of genetic diversity. Epigenetic marks in complete linkage disequilibrium (LD) with DNA polymorphisms do not add additional diversity information. However, epialleles in partial or low LD with DNA sequence alleles constitute another layer of genetic information that should not be neglected in germplasm analysis, especially if they exhibit transgenerational stability.

Keywords: comparative genomic hybridization; epigenetic; haplotype; next generation sequencing; single nucleotide polymorphism

Introduction

Thirty years ago Botstein *et al.* (1980) introduced the method of constructing genetic maps with DNA markers, known as restriction fragment length polymorphisms (RFLP). This development revolutionized genetic mapping and the analysis of diversity. Subsequent methodological advances, such as development of simple sequence repeat markers (SSRs), random amplification of polymorphic DNA (RAPD) (Williams *et al.*, 1990) and amplified fragment length polymorphisms (AFLP) (Zabeau and Voss, 1993), were enabled by the development of polymerase chain reaction. Development of single nucleotide polymorphism (SNP)-based markers brought a new level of resolution to the analysis of genetic diversity and for most applications superseded other genetic marker categories. More recently, DNA sequencing of partial or complete genomes from multiple individuals has expanded our understanding of the range of intraspecific genetic variation encountered in higher plants (Fu and Dooner, 2002; Yang and Bennetzen, 2009). With the rapid decline in the cost of DNA sequencing and new technological developments, it is certain that genome sequencing of germplasm collection will become accessible, eliminating biases present in existing genotyping methodologies, although it will also impose a significant data analysis overhead, necessitating increased investment in bioinformatics. The proposed 1001 Arabidopsis genomes project

* Corresponding author. E-mail: j-antoni.rafalski@cgr.dupont.com

(http://1001genomes.org/about.html) is a sign of things to come. Beyond DNA sequence, there is a renewed interest in the epigenetic marks, such as cytosine methylation, decorating DNA and chromatin, and potentially influencing the phenotype. We have discussed the impact of these developments on the analysis of genetic diversity.

Intraspecific diversity and the phenotype

Genomic sequencing of diverse genotypes in several plant species demonstrated that in addition to SNPs and SSR polymorphisms, extensive intraspecific differences include large insertions/deletions frequently composed of highly repetitive sequences such as retrotransposons and DNA transposons (Wang and Dooner, 2006), and in some cases also genes (Beló *et al*., 2009; Springer *et al*., 2009). For example, the complement of disease resistance genes may differ between accessions (Chin *et al*., 2001; Yahiaoui *et al*., 2009). Sequences that do not code for proteins may nevertheless affect the phenotype, by supplying enhancers or promoters to nearby genes, or code for small RNAs, which affect expression of other genes by a variety of mechanisms (Chen, 2009). Pseudogenes, which in maize are frequently generated by Helitron transposons, are sometimes transcribed in sense or antisense direction, also affecting gene expression phenotype (Yang and Bennetzen, 2009).

If these types of polymorphisms are in linkage disequilibrium with genetic markers used for germplasm characterization (predominantly SNPs and SSRs), then no additional information other than marker genotype is needed to reflect correctly the underlying genetic relationships of accessions. However, if linkage disequilibrium (LD) between markers for germplasm fingerprinting and genic or non-genic large indel polymorphisms breaks down rapidly, direct genotyping of these differences may be necessary by DNA sequencing or other methods such as array comparative genomic hybridization (Beló *et al*., 2009; Springer *et al*., 2009). This is likely to occur in the case of variants, which occurred recently on the background of pre-existing haplotype pattern.

An important issue not always appreciated in the germplasm analysis context is the prevalence of ascertainment bias, which occurs when polymorphic loci are identified (ascertained) in one collection of germplasm, but used to evaluate diversity in another set (Clark *et al*., 2005). For example, a collection of SNP loci identified in a set of cultivated lines will not correctly represent polymorphic loci present in unadapted accessions, leading to incorrect estimates of genetic distances in the latter set of germplasm. Many polymorphic loci in the non-adapted accessions will not be represented in the SNP collection developed from adapted germplasm, and, in turn, some alleles common in adapted material may be rare in non-elite accessions. As a result, genetic distances determined in the ascertainment population may be lengthened in comparison with those in the non-ascertained population (Fig. 1). It is difficult to identify *a priori* an appropriate collection of germplasm for ascertainment (marker discovery), given unbalanced representation of different types of germplasm in many collections. Perhaps, the most appropriate unbiased methodology for germplasm fingerprinting is genotyping by genomic sequencing.

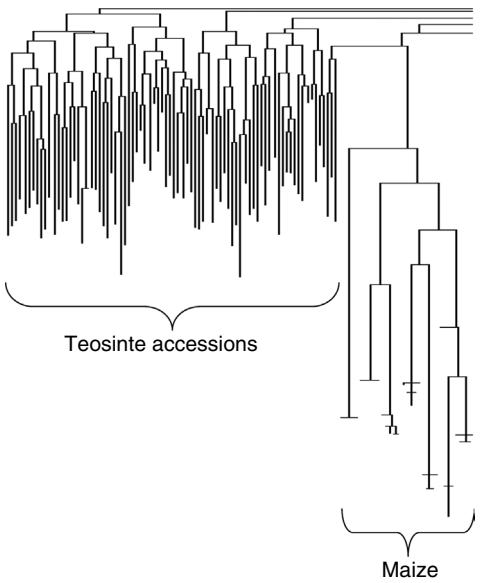

Fig. 1. An example of ascertainment bias. SNP markers ascertained in elite maize inbred collection were used to fingerprint a set of maize and teosinte lines. Genetic distances between maize lines appear much lengthened with respect to those between teosinte accessions, which are foreshortened. Using unbiased genotyping method eliminates this disparity. Data courtesy of Stan Luck (Pioneer Hi-Bred).

The sequencing technology is rapidly approaching the stage where it will become a cost-effective tool for genotyping (Edwards and Batley, 2009; Varshney *et al*., 2009). A number of accessions will be simultaneously sequenced in each lane of the instrument, after appropriate encoding. Depending on the size of the genome, some form of reduced representation analysis (Yuan *et al*., 2003) will probably be necessary to focus the effort on non-repetitive fraction of the genome.

Perspective on epigenotyping of germplasm

It is well established that epigenetic variation encoded by DNA base modifications such as 5-methylcytidine affects phenotype in animals and plants (Peaston and Whitelaw, 2006; Henderson and Jacobsen, 2007; Chandler and Alleman, 2008). Some of the epialleles in plants are

https://doi.org/10.1017/S1479262111000104 Published online by Cambridge University Press

remarkably stable and affect important plant characteristics (Cubas *et al.*, 1999). It is therefore reasonable to propose that a complete characterization of a germplasm accession or a breeding stock should involve not only the description of the genotype but also of the epigenotype. It has recently been demonstrated that recursive selection for a yield component in canola results in plants that are genetically identical but can be distinguished by DNA methylation differences and exhibit significant differences in yield (Hauben *et al.*, 2009). The tools for comprehensive epigenotyping are available and involve chemical deamination of m^5C to U followed by DNA sequencing, enabling single base resolution across the whole genome, albeit at considerable expense (Lister and Ecker, 2009; Lister *et al.*, 2009; Wang *et al.*, 2009). The high throughput sequencing technology, especially rapidly developing single molecule sequencing (Edwards and Batley, 2009), promises to enable comprehensive epigenotyping of germplasm collections in the coming years. Currently, several options exist for epigenotyping of a subset of the genome, for example by excluding repetitive fraction of the genome (Peterson *et al.*, 2002) or capturing specific sequences of interest (Hodges *et al.*, 2009).

Conclusions

Rapid technological developments are changing our understanding of genetic diversity, by allowing increasingly dense genotyping and identification of types of genetic polymorphisms that were previously not easily accessible to molecular analysis. In the next few years, another step change will occur with the availability of inexpensive genomic sequencing and development of tools for direct probing of epigenetic layer of information (Flusberg *et al.*, 2010). These developments will further enable the understanding of relationship between haplotype defined at the sequence level and phenotypic expression, through the use of association mapping and genome prediction techniques. To fully exploit these developments, we need to better understand the extent of linkage disequilibrium in the germplasm of interest.

Acknowledgements

I appreciate many discussions with Scott Tingey and with all of my professional colleagues at DuPont/Pioneer Hi-Bred Int.

References

Beló A, Beatty MK, Hondred D, Fengler KA, Li B and Rafalski A (2009) Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics* 120: 355–357.

Botstein D, White RL, Skolnick MH and Davis RW (1980) Construction of a genetic map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32: 314–331.

Chandler V and Alleman M (2008) Paramutation: epigenetic instructions passed across generations. *Genetics* 178: 1839–1844.

Chen X (2009) Small RNAs and their roles in plant development. *Annual Review of Cell and Development Biology* 25: 21–44.

Chin DB, Arroyo-Garcia R, Ochoa OE, Kesseli RV, Lavelle DO and Michelmore RW (2001) Recombination and spontaneous mutation at the major cluster of resistance genes in lettuce (*Lactuca sativa*). *Genetucs* 157: 831–849.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH and Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15: 1496–1502.

Cubas P, Vincent C and Coen E (1999) An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* 401: 157–161.

Edwards D and Batley J (2009) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol Journal* 8: 2–9.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J and Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* 7: 461–465.

Fu H and Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences USA* 99: 9573–9578.

Hauben M, Haesendonckx B, Standaert E, Van Der Kelen K, Azmi A, Akpo H, Van Breusegem F, Guisez Y, Bots M, Lambert B, Laga B and De Block M (2009) Energy use efficiency is characterized by an epigenetic component that can be directed through artificial selection to increase yield. *Proceedings of the National Academy of Sciences USA* 106: 20109–20114.

Henderson IR and Jacobsen SE (2007) Epigenetic inheritance in plants. *Nature* 447: 418–424.

Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L, McCombie WR, Wigler M, Hannon GJ and Hicks JB (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Research* 19: 1593–1605.

Lister R and Ecker JR (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Research* 19: 959–966.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B and Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462: 315–322.

Peaston AE and Whitelaw E (2006) Epigenetics and phenotypic variation in mammals. *Mammalian Genome* 17: 365–374.

Peterson DG, Wessler SR and Paterson AH (2002) Efficient capture of unique sequences from eukaryotic genomes. *Trends in Genetics* 18: 547–550.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D and Schnable PS (2009)

Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics* 5(11): e1000734. doi:10.1371/journal.pgen.1000734.

Varshney RK, Nayak SN, May GD and Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* 27: 522–530.

Wang Q and Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proceedings of the National Academy of Sciences USA* 103: 17644–17649.

Wang X, Elling AA, Li X, Li N, Peng Z, He G, Sun H, Qi Y, Liu XS and Deng XW (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell* 21: 1053–1069.

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA and Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic-markers. *Nucleic Acids Research* 18: 6531–6535.

Yahiaoui N, Kaur N and Keller B (2009) Independent evolution of functional Pm3 resistance genes in wild tetraploid wheat and domesticated bread wheat. *Plant Journal* 57: 846–856.

Yang L and Bennetzen JL (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proceedings of the National Academy of Sciences USA* 106: 19922–19927.

Yuan Y, SanMiguel PJ and Bennetzen JL (2003) High-Cot sequence analysis of the maize genome. *Plant Journal* 34: 249–255.

Zabeau M and Voss P (1993) Selective restriction fragment amplification: a general method for DNA fingerprinting. European Patent Application 92402629.7 (publication no. 0 534 858 A1).