OPEN ACCESS
Freely Available Online

JOURNAL OF AGRONOMY RESEARCH
ISSN NO: 2639-3166

Research

DOI : 10.14302/issn.2639-3166.jar-18-1987

Scientific and Technological Interventions for Attaining Precision in Plant Genetics and Breeding

Prem Narain[1,*]

[1]Professor and Independent Researcher, 29278 Glen Oaks Blvd. W., Farmington Hills MI 48334-2932.

Abstract:

The scientific and technological interventions for attaining precision in plant genetics and breeding since Mendel's discovery of genetic laws have been critically reviewed in terms of cloning technology and reverse genetics, chip technology, genetically modified organisms and CRISPR-based gene editing technology. Their roles in further refining the plant genetics and breeding practices particularly their exploitation in creating variations and their use for development of superior genotypes in model crops like wheat and rice have been discussed. It is stressed how such interventions could prove to be promising for meeting future crop improvement program in terms of climate change, bio-fortification, imaging technology, statistics, big data revolution and deep learning**.**

Corresponding author: Prem Narain, Professor and Independent Researcher, 29278 Glen Oaks Blvd. W., Farmington Hills MI 48334-2932, E-mail: narainprem@hotmail.com

Citation: Prem Narain (2018) Scientific and Technological Interventions for Attaining Precision in Plant Genetics and Breeding . Journal of Agronomy and Research - 1(1):5-21. https://doi.org/10.14302/issn.2639-3166.jar-18-1987

Keywords : Plant genetics and breeding; Mendel; Cloning; Reverse genetics; Markers; Genetically modified organisms; CRISPR-based editing; QTL mapping; Association mapping; eQTL; Marker-assisted breeding.

Received: Feb 5, 2018 **Accepted:** Mar 19, 2018 **Published:** Mar 30, 2018

Editor: Paramanandham Joothi Pillai, Assistant Professor, Department of Zoology and Wildlife Biology, AVC College (Autonomous), India.

Introduction

For ages since the domestication of agriculture about 10,000 years ago plant breeding was regarded as an art rather than a science to manipulate the crop species for improving their characteristics to benefit production. Breeders were using methods to improve the economic traits like yield etc. by selection and hybridization to incorporate desirable traits from one variety to another. However their success was limited due to the fact that whatever variation they exploited was derived from wild relatives with scant natural variation. They practiced a form of mass selection in which plants of superior phenotype were selected and seeds of such plants were planted during the next season. Without knowing the genetic basis of such practices the approach was hit and trial in getting superior progeny generation. This led to varieties which are now termed as *landraces* – locally adopted lines in a particular region. Today plant breeding is a full-fledged science with the ability to create variation according to their needs as well as to expedite the breeding process considerably to evolve new varieties. Behind this progress lies the scientific and technological interventions like fundamental discoveries of Mendel's gene in garden peas in 1900, phenomena of linkage and crossing over in 1910, production of mutation by X-rays in 1920, double-helix DNA in 1953, recombinant DNA technology, transgenesis and genetically modified organisms (GMOs) in 1970s including cloning technology and reverse genetics, epigenetic modifications due to DNA methylation and histone proteins, mapping of genes for quantitative traits with the help of markers and marker-assisted selection in 1980s, capability to sequence the whole genome in 2010s and lately CRISPR-based gene editing technology in 2012. In fact, transgenesis, QTL mapping, molecular marker-assisted breeding, gene sequencing etc. have introduced *precision* in the plant breeding process and have given rise to what is now termed as *molecular breeding*.

Mendelian Genetics Era

With the advent of Mendelian genetics, Danish botanist Wilhelm Johannsen, who coined the word *gene*, developed pure-line breeding theory to generate true-breeding (homozygous) lines through repeated self-pollination. He also stressed the role of environment in the inheritance of quantitative traits. Instead of bulking the seeds from different parent plants for mass selection and picking the best plants from the resultant crop, *progeny row* selection was adopted by sowing the seeds of single parent in separate rows and picking those parents whose progeny means were high. The progeny mean is subject to much smaller environmental variance and helps in picking the plants with the best genotypes. The identification of superior genotypes thus becomes more rigorous and fruitful in producing better progeny. One aspect worthy of being mentioned is the role of heritability – the fraction of total observed variability in an economic characteristic that is attributable to genetic causes – in taking decisions on selection procedures for genetic improvement. The heritability can be estimated from observed correlations between relatives [1]. In plant genetics, however, the total observed variability is enhanced due to the existence of genotype x environment interactions giving lower heritability compared to one without such interactions. Stability parameters – common varietal effects across environments – and environment specific deviations (interactions) are then used to study such problems [2-3].

The discovery of linkage in 1910 by TH Morgan and of mutation by X-rays in 1920 by HJ Muller led plant breeders to increase diversity (variation) in their material as well as to expedite the breeding process. The former tool of linkage and crossing over became of fundamental importance in QTL mapping and marker assisted breeding as we will see in the sequel. The mutagenic effect of X-rays and ionizing radiation as well as other chemical agents opened avenues for mutation breeding as a tool for release of hundreds of improved cultivars. In particular, semi-dwarf stature of plants, preventing lodging of plants in the field, was developed by mutation breeding in several crop plants like barley, wheat, rice, and sunflower. In particular, Dr. Norman Borlaug did extensive experiments in Mexico crossing different strains of wheat to come up with crossing stubby-stalked dwarf wheat with high-yield varieties that resulted in extremely high yields provided sufficient dose of fertilizer was applied to enable the plant to hold up in the field under the weight of large clusters of grain. Agronomists used the same device to breed semi-dwarf rice plants. Such dwarf varieties of wheat and rice when planted in other countries made a tremendous difference in the crop productivity and led in due course to the phenomenon of *green revolution*. Dr. Borlaug was awarded Nobel Peace Prize for this success in 1970. A country like India plagued with deficit food production for ages moved, in 1960s, to a surplus state capable of exporting food due to the success of green revolution in that country. Of course it demanded the use of inorganic fertilizers, irrigation and pesticides. States of Punjab, Haryana and Western Uttar Pradesh endowed with irrigation capability contributed significantly to this revolution. However, with the excess use of high dose fertilizers and pesticides over time, the gains of green revolution were not sustainable, the strategy becoming environment non-friendly. A more holistic approach to transcend the green revolution with an ‘Evergreen Revolution’ by adopting a comprehensive farming systems approach that considers land, cultivar improvement, water, biodiversity, and integrated natural resource management was later advocated [4].

Subsequent development of recombinant DNA technology has shown that if the green revolution were to occur now this process would have been very quick. Dwarfing gene identified from any model crop can be injected in the plant cells of the desired crop to produce shorter plants that can carry greater amount of grain without any concomitant effect of lodging. This *hastening of the genetic process* is the hallmark of transgenic technology innovation achieved in 1970s.

DNA, Clones, and Reverse Genetics

Soon after the discovery of the double-helix DNA, detailed knowledge of genetic material started accumulating. Sensitive techniques of isolating and analyzing genetic material in the laboratory were developed around a crucial attribute of the material–the ability to replicate–as well as the universality of genetic code. The sequence of DNA (deoxyribonucleic acid) letters (4 types of nucleotides – the A, C, G, and T representing respectively four chemical units or bases: adenine, cytosine, guanine, and thymine, their pairing being A binding to T and C to G) in the nucleus of each cell of the organism constitutes the basic genetic entity. But it is not known which DNA letter affects which part of the body and in what way. However we do know how the 4 letter alphabet of the language of DNA is transformed into the 20 letter alphabet of the language of proteins. The *genetic code* consists of a system of successive triplets of nucleotides along the DNA, known as *codons*, which code for successive amino acids of a corresponding polypeptide chain of a protein or enzyme.

Cloning means copying a given gene, a segment of DNA, usually done by putting it in the laboratory version of an *E. Coli* microbe so that, as the bacteria multiply, so do the copies of the gene. When we mix viruses with bacteria that is grown on a petri dish of nutrient *agar*, the areas where viruses have killed the bacteria results in a clearing in the lawn of bacteria on the petri dish – a killing zone of the bacteria known as a *plaque*. The plaques contain millions of virus particles and therefore millions of copies of the original DNA fragment of the gene. The *E. Coli* bacteria resists the virus attack by enzymatically cutting up the DNA of the invading virus into small pieces. Such enzymes are called *restriction enzymes*, a basic tool of genetic analysis. This gives pieces of DNA ending with a single-stranded end protruding from the DNA duplex. These are complimentary and can be joined by another enzyme *ligase* using the property of pairing of the nucleotide bases. Such molecular tools are used in the laboratory to develop what is known as *genomic DNA library*. There is another type of DNA library known as *complementary DNA library*. In this case we use another type of genetic material found in the cells known as *messenger RNA (mRNA)*. These are used to carry out the genome’s orders to make proteins. However mRNA is by its nature transitory and unstable. By using another enzyme called *reverse transcriptase* the RNA can be copied into a stable form of DNA known as *complementary DNA (cDNA)*. The library is developed by isolating mRNAs at work in a tissue, converting them into cDNA fragments, and inserting these fragments into a *plasmid*, a small ring of DNA that carries the instructions for a bacterium. When the bacteria are infected with the plasmids, millions of copies of cDNA are produced. Because each bacterium contains a different segment of cDNA, when it replicates i.e. divides into daughter cells, both the mother and the daughter cells would contain the same DNA fragment. Of the two libraries, the cDNA library exploits the way that Nature's copy editor turns the whole genetic code into a much smaller stretch of mRNAs that represents only the subset of genes-the coding genes-required for a specific cell or tissue type. This helps in gene hunting.

Techniques that are used to manipulate the genetic material in the test tube lead to another phenomenon known as *reverse genetics*. In classical genetics we observe the phenotype and *infer* the genotype on the basis of the results of mating two individuals differing in their genotypes. We conclude that the observed difference between them was due to pre-existing mutation as one can see in the results of Mendel's experiments. In reverse genetics we take a fragment of DNA, the role of which in the life of the organism could be known or unknown, and mutate it in the test tube. After reintroducing it back into the cell where it gets integrated into the chromosomes, we can see the consequence, if any, on the phenotype of the organism. That is, we go from genotype to phenotype, a process reverse to the one used in classical genetics.

Transgenic Technology

It is a derivative of recombinant DNA technology that gave birth to plant genetic engineering involved in creating plants with desired characteristics by inserting useful genes from a wide range of living sources, not just from within the crop species or from closely related plants. It is a man-made technique but based on the principles followed in nature - the ingenious genetic engineering of soil bacterium *agrobacterium tumefacieen* of injecting its own DNA and integrating it with those of the plant with crown gall disease. This gave a clue to the researchers how to *artificially* insert into *agrobacterium*'s plasmid a desired gene for transferring it to the plant cell. When this genetically modified bacterium infected a host plant it would insert the chosen gene into the plant's chromosome which would hereafter be called a *genetically modified organism*. This technique however got refined in 1980s by the invention of "gene gun" in which the desired gene is affixed to tiny gold or tungsten pellets and these are fired carefully like bullets into the cell. By 1990 the scientists succeeded in using the gun to shoot new genes into corn and genetically modified corn was born-the first GM crop.

This technology provides the means for identifying and isolating genes containing specific characteristics in one kind of organism and for moving copies of those genes into another quite different organism which will then also have these characteristics. It has enabled plant breeders to generate more useful and productive crop varieties and in a much shorter time than the cumbersome traditional cross-pollination and selection techniques. Genetically modified crops like corn, soybean, rapeseed oil, cotton, rice are now planted in about 170 million hectares globally.

In the case of Bt cotton, with the transfer of the cloned Bt (*Bacillus thuringiensis*) gene in the cotton plants by genetic engineering, the plants produce their own biocides and kill the caterpillars of the insects (*lepidopteran*) that cause damage to the crop (bollworm attack). It is a chemical protection of the crop, the plant cells being the delivery system. So while the quantity of the insecticides for spraying used in the traditional approach is considerably reduced leading to lower input costs to the farmer and protection of the ecosystem, the strategy might create problems in the internal machinery of the plant itself. But the experimental evidence is to the contrary.

A transgene is a segment of DNA containing a gene sequence that has been isolated from one organism and is introduced into a different organism. It is an assembly of three parts - a *promotor*, an *exon*, and a stop sequence. The promotor is a regulatory sequence that will determine where and when transgene is to be active. The exon is a protein coding sequence usually derived from the cDNA for the protein of interest (vide cDNA library discussed in previous section). All the three parts are typically combined in a bacterial plasmid with the coding sequence being chosen from transgenes with previously known functions.

The dichotomy of GM and non-GM crops seems to be superfluous. All improved varieties are genetically modified; only the methods of obtaining them could be different. As against such man-made genetic modifications in domesticated crops, there is also genetic modification in wild populations by natural selection following the principles put forth by Charles Darwin. This leads to evolution of varieties where the selection is of stabilizing type favoring phenotypes near the mean of the population. In the man-made case, the so called artificial selection, extreme phenotypes (increased yield) are favored that might involve some loss of fitness. Artificial selection practiced by breeders since the advent of agriculture about 10,000 years ago, when even genetic principles were not known, could produce modifications (genetic) of the desired type. It is said that Darwin got the clue to his theory of evolution by natural selection from the results of artificial selection practiced in domesticated species. It was later when Mendel gave the laws of heredity that natural selection got a genetic basis of operating on genetic variation created by mutation and recombination. Genetic modification is therefore at the root of all this process whether natural or man-made.

CRISPR-based Gene Editing Technology

CRISPR/Cas9 is a system consisting of a CRISPR (*clustered regularly interspaced short palindromic repeats*) molecule and an enzyme Cas9 of the cell. The former could be programmed to target a specific section of the DNA by loading it with its matching RNA sequence (guide RNAs i.e. sgRNA) and the latter could function as a powerful pair of molecular scissors to cut the matched section of the DNA. The repeat sequences of 29 nucleotides are separated by various 32 nucleotides spacer sequences. Soon after cleavage of the targeted sequence, the body can either repair itself on its own – non-homologous end joining (NHEJ) - or scientists can patch in a corrected sequence – homology-directed repair (HDR). If done in sex-cells, the changes will be passed on to future generations. It is a very recent biotechnological tool that is revolutionizing plant breeding practices by modifying targeted DNA sequences within plant genomes particularly in crops like rice and wheat. It is much like what we do in a word processor by ‘cut’ and ‘paste’ functions.

It is however significant to note that such a system has been derived from a *naturally* occurring defense mechanism first observed to take place in a cup of yoghurt when the bacterium *streptococcus thermophilus* used it to defend themselves from repeated viral infections by providing a type of acquired immunity for it. After viral invasion is repelled the bacterial DNA keeps a genetic record of the viruses infecting them as short repeated sections of the DNA along with short segments of spacer DNA in-between them as snippets of virus’ genes repelled so that when the same virus attempts to again infect the bacteria it would gravitate towards its matching section on the bacterial genome and bind to it. That summons the powerful enzyme *cas9* of the cell to perform the task of snipping the virus out, leaving the bacteria free from infection. Researchers realized then that this trick of bacteria could be used to cut not only viral DNA but any DNA sequence in any organism at a specifically selected gene or genes by altering guide RNAs in combination with enzyme Cas9 to match a targeted gene or genes. The sgRNA is part of a longer RNA molecule that forms a riboprotein with the Cas9 enzyme machinery positioning the Cas9 enzyme to the correct position on the target DNA for cleavage.

In plant breeding this technology can enable the scientists to edit the genomes of superior varieties to produce new varieties in a single generation irrespective of the existing variability and without the need to select favorable combination of alleles. But such an approach requires knowledge of the nucleotide sequence and function of the targeted genome so as to be able to design the appropriate sgRNA and predict the editing outcome. However it has been applied in rice crops by generating mutations at the target sites at nearly 100 % efficiency. A CRISPR/Cas9 mutagensized rice line with enhanced blast resistance was recently released [5]. In wheat however this technology has not been that successful. The first CRISPR/Cas9 mutagenized wheat plants developed had an efficiency of only 5 % [6]. The capability of multiple targeting of sites of this technology can however be useful in wheat due its being a polyploid crop (having more than two sets of chromosomes). Scientists involved in climate change studies recently used genome editing to enhance drought tolerance in maize by editing a previously unidentified promoter to increase expression of the *ARCOSS* gene which down regulates the growth-inhibiting hormone ethylene, enhancing plant growth and yield under drought stress [7]. In tomato flowering time can be manipulated by using CRISPR/Cas9 to generate early–yielding varieties by disrupting the flower-repressing gene *SP5G* [8].

Like GMO the CRISPRized crops also face sociopolitical challenges such as government regulations, public acceptance and adoption by producers such as small farmers. However advantages of genome editing over conventional and earlier transgenic approaches being its low cost, ease of use, lack of transgenes permanently introduced into crop germplasm and high level of multiplexing (editing of multiple targets) can lead to its wide adoption in the near future for increased crop production.

Molecular Markers and Linkage Maps

Soon after the introduction of technology for genotyping molecular markers, the so called chip technology, the methods of plant breeding got a big impetus in increasing precision in the breeding process by incorporating the marker information in the existing approaches of selection and cross breeding. This involves three components (a) Molecular Markers and Linkage Maps, (b) Mapping of QTLs, and (c) Maker-assisted Plant Breeding discussed in this and following three sections.

The role of markers, however, was implicit in earlier studies on quantitative genetics way back to the work of K. Sax who investigated the existence of linkage between the polygenes of a quantitative trait like the weight *of* seeds and a Mendelian gene like the color of the seed [9]. He crossed a strain of dwarf beans*, Phasecolus vulgaris*, having large colored seeds with another whose seeds were small and white. While seed size showed itself to be a continuously variable character, the pigmentation proved to be a single gene difference (*P-p*), the F_2 giving a ratio of 3 colored to 1 white seeded plant. By means of F_3 progeny, the colored F_2 plants were further classified into homozygotes (*PP*) and heterozygotes (*Pp*). The average bean weights in the three classes of F_2 plants were found to be *PP* (30.7), *Pp* (28.3) and *pp* (26.4). Their standard errors showed the difference in seed weight to be statistically significant. Clearly the average weight is associated with the number of *P*-alleles present viz. 2, 1, 0. The pigmentation is thus here synonymous with a marker that is associated with a quantitative trait. Such a marker can be followed through the generations and can serve as a tag for following the quantitative trait provided it is linked with it. This aspect has become of crucial importance to plant geneticists and plant breeders for improving economic traits.

Quantitative traits such as yield of plant, flowering time, pest resistance etc. are complex in nature being controlled by several genes and affected by environmental factors. Quantitative genetics in contrast to Mendelian genetics has developed around such traits with a heavy dose of statistical input [2-3]. Quantitative Trait Loci (QTL) is a segment of DNA and its effects could be either small or large at least in comparison to the environmental modifications. As mentioned earlier the methodology of quantitative genetics has considerably got modified due to the introduction of marker information via chip technology. There are several ways of getting such information as for instance pigmentation in the study of K. Sax [9]. Broadly there are three categories of markers viz. morphological like blood groups, biochemical like allozymes and molecular which are at the DNA level. The last one can be listed as:

Restriction fragment length polymorphism (RFLP)

Random amplified polymorphic DNA (RAPD)

Amplified fragment length polymorphism (AFLP)

Variable number of tandem repeats (VNTR) - that consist of microsatellites (short sequences) termed as short tandem repeats (STR) or simple sequence repeats (SSR) and mini-satellites (long sequences)

Single nucleotide polymorphism (SNP).

Of course the whole DNA sequence is itself an ultimate marker in the process of marker development. These all help in identifying the QTL by looking for association between the trait and the specific one or several markers [10-13]. They are like sign posts or tags. For instance, suppose you go to a new city and are interested in locating the house of your friend whose address you don't know but you do know that the house is in the vicinity of a petrol pump with a known address. Your ability to be successful in the search would depend on the closeness, including direction, of the petrol pump to the house. In the absence of such sign post information you would have a cumbersome task of knocking the doors of each house and enquiring whether your friend lives there.

In addition to the above type of markers we have also what are known as *functional markers* which are superior to above mentioned random DNA markers in that they are located within specific gene regions delimited by QTLs and are therefore completely linked with the QTL alleles. They are derived from functionally characterized sequence motifs affecting phenotypic variation.

The first problem in QTL mapping is to construct a linkage map that indicates the position and relative genetic distances between the chosen markers along each of the chromosomes. The map distance is based on the total number of crossovers between two markers whereas the physical distance between them is in terms the nucleotide base pairs (bp). A centi-Morgan (cM), corresponding to a cross-over of 1%, can be a span of 10 kbs to 1,000 kbs and can vary across species. Linkage maps for several crop species like rice, wheat, maize etc. have been constructed and are used for QTL mapping.

Since the marker genotypes can be followed in their inheritance through generations, they can, as stated above, serve as molecular tags for following the QTL provided they are linked with the QTL. This requires detecting the marker-QTL linkage and if established, estimating the QTL map position on the chromosome along with effect size of the QTLs. However, these problems depend on what sort of experimental populations we have in plant breeding investigations. In crops practicing self-fertilization, populations are derived from a cross between two pure breeding parents, homozygous at all the loci controlling variation in the trait. Such F_1 hybrids are selfed to produce segregating F_2 populations whereas backcross (BC) populations are derived by crossing the F_1 hybrid to one of the parents, usually the recessive ones. Inbreeding from the individual F_2 plants can lead to recombinant inbred (RI) lines which consist of a series of homozygous lines, each containing a unique combination of chromosomal segments from each of the two original parents. It takes around six to eight generations to achieve this type of populations. In species capable of tissue culture such as rice, barley and wheat, plants can be regenerated by inducing chromosome doubling from pollen grains. This leads to production of double haploid (DH) populations. Both RI and DH populations are true breeding lines that can be multiplied and reproduced without any segregation and therefore provide eternal resources for QTL mapping. In cross pollinating species, on the other hand, such simple designs are not possible due to lack of inbreeding. Mapping populations are usually derived from a cross between a heterozygous parent and a haploid or homozygous parent depending upon the plant breeding need.

Mapping of Quantitative Trait Loci (QTL)

The detection of marker-QTL linkage is based on a statistical test of a null hypothesis (H_0) against an alternative hypothesis (H_1). The null hypothesis postulates that there is no QTL in the vicinity of the chosen marker with a known location on a given chromosome and hence no linkage exists between them. This can happen in several ways. The QTL is not on the same chromosome as the marker or it is on the same chromosome but cross over with it at the meiosis occurs with probability ½. If we reject this hypothesis saying that we detect linkage when in fact no QTL is present we commit an error which is termed as *false positive*. On the other hand if we accept the null hypothesis meaning that there is no linkage when in fact a QTL is present we commit another error of *missing* the QTL which is termed as *false negative*. These errors are respectively known as *Type I* and *Type* II errors in the statistical literature pertaining to testing of hypotheses. Including the two other possibilities of *true positive* and *true negative*, the four possibilities are:

Reject H_0 when H_0 is true – false positive (type I error)

Accept H_0 when H_0 is true – true negative

Reject H_0 when H_0 is false – true positive

Accept H_0 when H_0 is false – false negative (type II error)

In statistical testing our strategy is to minimize the probability of committing the error of missing the QTL for a fixed low level of the probability of occurrence of the false positive, usually kept at 5 % level. When H_0 is taken as false, the alternative hypothesis H_1 is regarded as true implying that a QTL is present and the probability of such a contingency is maximized. This provides the power of the test and can be increased by increasing the sample size. It may be noted that the probability of the concerned events can only be determined on postulating the true hypothesis. In general the test statistic is derived by *a likelihood ratio criterion*. This statistic is, in genetic applications, termed as LOD score and is approximately related to a chi-squared distribution.

Broadly there are two approaches to QTL mapping known as (a) candidate gene mapping and (b) genome wide association study (GWAS). In the former a specific genomic region on a given chromosome is chosen to look for the QTL with the help of the markers known to be located in that region. Tests for the presence or absence of the QTL are conducted at several map positions in this region say every 1 cM with the help of the LOD scores. Map positions showing significant values of the LOD score are deemed to contain a QTL. Amongst these the one with the maximum LOD score is chosen to indicate the position of the QTL. However the distribution of the maximum LOD score is not just chi-square due to non-independence of the successive tests, particularly in a dense-marker linkage map. In GWAS, on the other hand, all maker positions on all the chromosomes are tested for the presence or absence of the QTL. This therefore requires a genome-wide threshold for judging the significance. With larger genomes more tests will be performed increasing the probability that a fixed LOD threshold will be exceeded. When we need an experiment-wise significance level of 5 % this means the probability of obtaining a LOD score above the threshold somewhere on the whole genome just by chance to be 5 %. The genome-wide threshold will thus depend on the number and length of the chromosomes as well as on the number of markers on the chromosomes. When few markers are tested per chromosome – the so called sparse map case – a lower threshold is needed at the same genome-wide significance level than when many markers are tested per chromosome – the so called dense map case. An exercise of determining LOD significance thresholds in experimental plant populations was attempted by using large scale simulations [14].

Taking into account the genetic basis there are three major methods of QTL mapping applicable to plant populations. These are (a) single marker analysis, (b) simple interval mapping (SIM), and composite interval mapping (CIM). We consider them below for a double back-cross population segregating for the quantitative trait under study as well as the chosen marker.

Single Marker Analysis

This is the simplest situation wherein for all the sampled plants from the population observations are recorded for the trait under the study and individual plant is genotyped for the marker. The data analysis can be performed either as a *t*-test, or as an analysis of variance (ANOVA) test. We can visualize that with one QTL locus and one marker locus there would be four marker-QTL genotypes whose frequencies would depend on the recombination probability between the two loci. Since the marginal frequency of the two possible marker genotypes is one-half each, the frequency of the QTL genotypes conditional on the marker genotype can be worked out. The expected value of the difference between the observed trait means of the backcross population in each of the two marker groups can be obtained in terms the recombination probability as well as the genetic effects for each of the QTL summed over all the QTLs. The two are confounded and so the null hypothesis being composite can mean either there is no linkage between QTL and marker loci or the QTL genetic effects are zero. This method is highly inefficient since we cannot determine whether a significant marker effect is due to one or multiple QTLs and whether the effect is due to far distantly linked QTLs with large effects or closely linked QTLs with small effects.

Simple Interval Mapping (SIM)

The most popular method is that of s*imple interval mapping* (SIM) [15]. It involves formation of intervals by pairing of adjacent markers and treating them as a single unit of analysis for detection and estimation purposes. It is based on the joint frequencies of a pair of adjacent markers and a putative QTL flanked by the two markers. Suppose markers *A* and *B* are linked with recombination fraction *r* and QTL *Q* is located between them with r_1 recombination from *A* and r_2 from *B*. Then $r=r_1+r_2-2r_1r_2$ approximated as r_1+r_2, on the assumption of no interference and *r* so small that no double crossovers can be assumed. In the classical back cross with three loci each with two alleles, *A-a*, *B-b*, and *Q-q,* the expected frequencies for the eight marker-QTL genotypes can be used to obtain the conditional probabilities of the QTL genotypes given the marker genotypes. By setting up a linear regression model between the trait (Y) and the indicator variable (X) taking the value 1 if the QTL is *QQ* and –1 if it is *Qq*, one can estimate the regression coefficient that defines the allelic substitution effect of this QTL. In such a model, the QTL genotype for a given individual is unknown. X is then a *random* indicator variable with conditional probabilities of obtaining *QQ* or *Qq* at the QTL. This means the observed value is modeled as a mixture distribution with mixture ratios as the *conditional* probabilities. We have, therefore, a situation often referred to as a linear regression with *missing* data. The problem of estimation then involves the use of EM algorithm. By assuming that the character is normally distributed within each of the eight marker-QTL classes with equal variance σ^2, one can set up a likelihood function in terms of unknown parameters, and develop a log likelihood ratio Λ for testing the hypothesis that the QTL is not located in the interval where the log likelihoods are evaluated using the maximum likelihood estimates of the genotypic values for the two QTL genotypes, the variance σ^2 and the recombination fraction r_1 between marker *A* and the putative QTL using iterative procedures based on EM algorithm. This statistic is distributed as χ^2 with 1 d.f. The associated LOD score for the interval mapping is then (½)$\log_{10}$e) Λ

This statistic is evaluated at regularly spaced points, say 1 or 2 cM distance, covering the interval as a function of the presumed QTL position. Repeating this procedure for each interval along the chromosome and plotting the LOD score curve against the interval gives a *QTL likelihood map* that presents the evidence for the QTL at any position in the genome. Presence of a putative QTL is assumed if LOD score exceeds a certain threshold *T* and the maximum of the LOD score function in the map gives an estimate of the QTL position and the gene effects. The mapping of QTL by interval method is widely used in practice. The analysis is done through the software MAPMAKER/QTL. The estimates of QTL effects and its location are asymptotically unbiased if there is *only* one QTL on a chromosome. But if there are two or more QTLs on the chromosome the test statistic for the effect and location will be affected by other QTLs linked to the QTL under test and therefore can result in biased estimates of effect and location. Also some regions not having any QTL can show a significant peak if there are several QTLs in the neighboring regions – a situation known as *ghost* gene phenomenon. This defect of SIM can be overcome by adopting *composite interval mapping* (CIM) discussed in the next section.

Composite Interval Mapping (CIM)

Although SIM is the method for QTL mapping most widely used with advantage in several practical situations, it ignores the fact that most quantitative traits are influenced by numerous QTLs. This is overcome either by adopting a model of *Multiple QTL Mapping* (MQM) or by combining SIM with the method of multiple linear regression, a procedure known as *composite interval mapping* (CIM) [16]. Consider a segment of chromosome between markers *i* and (*i+1*) using a backcross progeny and set up the same type of linear model as in the section on SIM with X replaced by X_i and *b* replaced by b_i and adding a sum over b_k X_k for the markers other than i-th marker with b_k as the partial regression coefficient of the trait value on the marker *k* and X_k as a dummy variable for marker *k* taking value *1* if the marker has genotype *AA* and *0* for *Aa*. The maximum likelihood procedure is adopted to derive the formula for the relative position of the QTL as well as the likelihood ratio test statistic to obtain the LOD score for the hypothesis under test [17]. Here the regression coefficient under test is a partial regression coefficient conditional on other partial regression coefficients in the model. The hypothesis under test is thus a composite and hence got the name composite interval mapping (CIM). It may be noted that the markers in the CIM model can control the residual genetic background only when they are linked to QTLs. In practice, CIM is implemented using iterative E-M algorithm. For each position of the QTL, the iteration starts with the

E-step : getting the probability of X_i = 1 for QTL being QQ and then performing the

M-step : estimating b_i , *B* and σ^2 for the next round of iteration where *B* is a vector of maximum likelihood estimates of the intercept and partial regression coefficients for all the markers except *i* and (*i+1*) and σ^2 is the variance of the error term.

The advantage of CIM over SIM can be seen in the results of the analysis of mapping body weight QTLs on mouse chromosome X from a backcross population [18]. The CIM analysis achieved a much better resolution than the SIM.

Other Methods

In all the above methods, one uses the approach of maximum likelihood that produces only point estimates of the parameters such as the number of QTLs, their location, and effects. Their incorrect specification leads to distortion of the estimates of locations and effects of QTLs. To address these problems a *Bayesian* approach is often adopted wherein the joint posterior distribution of all the unknown parameters given their *prior* distributions and the observed data is computed [10, 19].

Application to Tomato Crop

The first application of interval mapping in plant breeding was to an inter-specific backcross in tomato [20]. The parents for the back-cross were the domestic tomato *Lycopersicon esculentum* (*E*) with fruit mass 65 gm and a wild South American green-fruited tomoto *L. chmielewskii* (*CL*) with fruit mass 5 gm. A total of 237 back-cross plants, with *E* as the recurrent parent, were grown in the field at Davis, California. Around 5 to 20 fruits per plant were assayed for continuously varying characters like fruit mass, soluble-solids concentration and pH. Around 63 RFLP and 20 isozyme markers spaced at approximately 20 cM intervals and showing polymorphisms between the *E* and *CL* strains were selected for QTL mapping. These markers were a subset from a larger number used for constructing a complete linkage map of tomato with 12 chromosomes on an earlier occasion. The markers were scored for each of the 237 backcross progeny and a linkage map was constructed *de novo* using software MAPMAKER. This map covered all the 12 chromosomes with an average spacing of 14.3 cM. The methods of maximum likelihood and LOD scores were used through the software MAPMAKER-QTL to implement the interval mapping. A threshold *T*=2.4, giving the probability of less than 5% that even a single false positive will occur anywhere in the genome, was used. This corresponds approximately to the significance level for any single test as 0.001. The resulting QTL likelihood maps revealed multiple QTLs for each trait (6 for fruit weight, 4 for concentration of soluble solids and 5 for fruit pH) and estimated their location to within 20-30 cM.

In regard to fruit weight, the above type of investigation was continued with more and more QTL for this trait being identified. At least 28 QTLs controlling the difference in fruit weight between the wild and cultivated tomato were identified, one of them being *fw2.2* on chromosome 2 [21]. By refined mapping this QTL was localized to a narrow chromosomal region of the order of 1/10,000th of the genome. Using a map-based approach, *fw2.2* was cloned and a 19-kb segment of DNA containing it was sequenced. This made it possible to identify a single gene responsible for the QTL effect as *ORFX*. By transforming the wild version of the gene into a cultivated tomato, it was shown that the transformed plants decrease in weight by about 30 % as predicted thus conforming that there are no additional fruit weight QTLs nearby on the chromosome [22]. The gene is expressed early in floral development, controls carpel cell number, and has a sequence suggesting structural similarity to the human oncogene c-H-*ras*p21. Alternations in fruit size, imparted by *fw2.2* alleles, are most likely due to changes in regulation rather than in the sequence and structure of the encoded protein.

The laboratory of Dr. Tanksley at Cornell University, Ithaca, New York also reported the results of another QTL analysis in tomato in which the population under study was derived from a cross between the wild type species *L. pimppinellifolium* with a very low average weight of 1 g. and *L. esculentum* cultivar var. Giant Heirloom with fruit weight in excess of 1,000 g. [23]. They found the same six major loci on chromosomes 1-3 and 11 accounting for as much as 67 % of phenotypic variation in fruit mass as in the previous experiments. The two most significant QTLs detected in this study are *fw11.3* and *fw2.1* on chromosomes 11 and 2 respectively. Both of them affect the fruit size through the control of carpel/locule number.

These investigations seem to counter our belief that genes for quantitative traits have effects so small and their number so large that they cannot be followed individually through generations as we are able to do with Mendelian genes. The effects of *fw2.2* were found to be sufficiently large adding about 17 grams to a tomato.

Association Mapping

The mapping of QTLs in plants based on data collected from pedigrees of populations formed by crossing inbred lines is on a coarser scale because there are not enough recombination events so that recombination probability of less than 1 % cannot be estimated. A QTL detected is therefore likely to refer to several genes in a chromosomal region. The achievable range of resolution in QTL mapping is about 3 cM in large populations. But QTL peaks often extend to more than 20 cM in linkage maps. The approach of population-based association mapping involving linkage disequilibrium (LD) between markers and genes underlying complex traits leads, on the other hand, to more accurate and finer mapping of genes. The key idea is that a trait mutation assumed to have arisen once on the ancestral haplotype of a single chromosome in the past history of the population of interest has had many thousands of recombination events and is therefore passed on from generation to generation together with markers at tightly linked loci resulting in LD. DNA markers close enough (< 2 cM) remain associated with the trait 'gene' for many generations. This approach is also termed as Linkage Disequilibrium Mapping or simply as Association Genetics.

The advantages of the two approaches can often be combined by initially detecting QTL using linkage mapping with moderate number of markers followed by a second-stage of high-resolution association mapping in QTL regions that capitalizes on a high-density marker map [13].

Nested Association Mapping (NAM) in Maize

The benefits of linkage and association mapping can be combined in a single population of maize by adopting a *nested association mapping* (NAM) approach as done in experiments with maize in the laboratory of Dr. Edward S. Buckler at the Department of Plant Breeding and Genetics, Cornell University, Ithaca. The maize NAM population was derived by crossing a common reference sequence strain to 25 different maize lines. Individuals resulting from each of the 25 crosses were self-fertilized for four further generations, to produce 5,000 NAM recombinant inbred lines (RILs).

This population was first used for initial detection of QTL using linkage mapping approach. Subsequently, within each diverse strain, high-resolution association mapping was adopted with a high-density marker map. It is significant to note that within each RIL all individuals are nearly genetically identical – the so-called *immortal genotypes*. This means we can estimate the true breeding value of each line much more accurately by averaging the phenotypic measurements of a given trait taken on several individuals with the same genotype.

In an experiment conducted in 2009, the genetic architecture of flowering time in *Zea mays* (maize) was dissected using NAM. About 1 million plants were assayed in eight environments to map the QTLs. About 29 to 56 QTLs were found to affect flowering time. These were small-effect QTLs shared among the diverse families. The analysis showed, surprisingly, the absence of any single large-effect QTL. Moreover, there was found no evidence of epistasis or environmental interactions. Flowering time controls adaptation of plants to their local environment in the out-crossing species *Zea mays*. A simple additive genetic model predicting accurately the flowering time in this species is thus in sharp contrast to what has been observed in several plant species which practice self-fertilization [24-25]. We may compare this finding with that noted in the section on tomato crop wherein a QTL with a major pronounced effect was detected and cloned.

Mapping QTLs for Gene Expression profile (eQTL)

The advent of DNA chip technology in the form of cDNA and oligonucleotide microarrays provides with huge and complex datasets on gene expression profiles of different cell lines from different organisms. Such gene expression profiles were combined with linkage analysis based on QTL mapping through molecular markers in what has been termed as 'genetical genomics' [26]. Gene expression, in terms of transcript levels, for each individual of a segregating population, are phenotypes that are correlated with markers, genotyped for that individual, to identify the QTLs and their locations on the genome to which the expression traits are linked. Such expression quantitative trait loci (eQTL) studies are similar to traditional multi-trait QTL studies but with thousands of phenotypes. It is also important to note that, underlying the gene expression differences, there are two types of regulatory sequence variation. One is *cis*-regulatory that affects its own expression and the other is *trans*-acting or protein coding that affects the expression of other genes.

Marker-assisted Plant Breeding

There is a vast literature on this topic with several applications in different crops and for different traits with two useful publications [27] [28].

Marker-assisted breeding practices can take different forms depending upon the traits and crops chosen. In particular, one can use markers to select plants at the seedling stage particularly when the trait is expressed at later development stages such as in rice breeding. In such a case the pre-germinated seeds are first sown in nurseries and resulting seedlings are next transplanted into rice fields. The breeder can use the marker to eliminate undesirable plant genotypes at the seedling stage and transplant only selected ones.

Considerable progress has taken place in marker-assisted plant breeding across crops and traits. Four components of this topic are discussed below: (a) *Marker-assisted backcrossing (MABC)*, (b) *Marker-assisted gene pyramiding (MAGP)*, (c) *Marker-assisted recurrent selection (MARS)*, and (d) *Genomic selection (GS)*.

Marker-assisted backcrossing (MABC)

It is the simplest method of MAS widely used in several cereal crops and for various quantitative traits. It has one donor parent (DP) with the desired trait from which to transfer marker genes to a superior cultivar or elite breeding line serving as the recurrent parent (RP) to improve the given trait. Instead of phenotypic performance of the trait, the marker alleles linked with genes of interest are used in crossing and selection. The F_1 of the cross DP x RP is checked for the maker alleles at early stages of growth to eliminate false hybrids and the true F_1 plants are backcrossed to RP. The individuals of resulting BCF_1 population are screened for the markers at the early growth stages and the plants carrying the desired markers (heterozygous) are backcrossed to the RP. This process is repeated for two to four generations depending upon the need. The final backcrossing population BC_4F_1 say is planted after screening the plants with the markers of the desired trait and discarding the plants with homozygous marker alleles from the RP. The plants with required marker alleles are selfed and harvested. The progenies of the BC_4F_2 population are planted, the markers detected, and individuals with homozygous DP marker alleles of target trait are harvested for further evaluation and release.

The MABC has been applied to several traits like disease/pest resistance, drought tolerance and quality in crop species like rice, wheat, maize, barley, pear, millet, soybean, tomato, etc. For instance, in corn, as mentioned earlier, Bt transgene was integrated into various corn genetic backgrounds by using MABC. It was used to select for aroma in rice. In tomato, Tanksley and his colleagues modified MABC strategy to advanced backcrossing QTL (AB-QTL) to transfer resistance genes from wild relative genotype into elite germplasm. It has also been used in other crop species like rice, barley, wheat, corn, cotton and soybean proving thereby the effectiveness of AB-QTL in transferring favourable alleles from the wild to elite germplasm.

Marker-assisted gene pyramiding (MAGP)

The breeding strategy of this technique depends on the number of genes required for improvement of traits, the number of parents that contain the required genes, the heritability of the traits, marker-gene association, duration of the plan and the relative cost. If we have four desired genes existing separately in four lines, pyramiding them can be done either by stepwise back crossing, or by simultaneous/ synchronized backcrossing or else by convergent backcrossing. Let W be a line (RP), superior in all respects except lacking in a trait, genes for which are identified to be in four lines P_1, P_2, P_3, and P_4. In stepwise backcrossing the four genes are transferred in W in order, one-step of backcrossing for each of the four parents in succession till all the four genes have been introgressed into W. In the simultaneous/ synchronized backcrossing W is first crossed to each of the four donor parents to produce four single-cross F_1s. Two of them are crossed to each other to produce two double-cross F_1s which are crossed again to produce a hybrid integrating all the four genes in heterozygous state. It is subsequently crossed back to W until a satisfactory recovery of the RP genome and finalized by one generation of selfing. The third design of convergent backcrossing combines the procedures of the stepwise and synchronized backcrossing.

In applications with rice crop, pyramiding has been achieved for bacterial blight and blast. The cumulative effects of multiple gene pyramiding have been proven in crop species like wheat, barley, and soybean.

Marker-assisted recurrent selection (MARS)

It is a scheme in which genotypic selection and intercrossing are performed in the same crop season for one cycle of selection. It enhances the efficiency of recurrent selection and helps in integrating multiple favorable genes from different sources through recurrent selection based on a multiple-parental population. MARS can also be defined as a recurrent selection scheme that uses molecular markers for the identification and selection of multiple genomic regions involved in the expression of complex traits to assemble the best-performing genotype within a single or across related populations [29].

Genomic selection (GS)

Genomic selection approach was first introduced in 2001 [30]. This new approach requires a sufficiently high marker density such that every QTL affecting the trait would be in linkage disequilibrium with at least one of the markers. It involves estimating the effects of *all* the markers together without testing for significance. It consists of a *training* population (genotyped as well as phenotyped) to model the relationship between phenotypes and molecular markers and a *testing* population (genotyped but not phenotyped) to estimate the breeding values. Using the simulated data set of 1010 genetic markers and 1000 QTLs, four modeling methods – linear regression, best linear unbiased prediction (BLUP), and two Bayesian methods dubbed BayesA and BayesB – were tested for accuracy in predicting the breeding value of an individual genotyped for many alleles. BLUP and the Bayesian methods were able to predict the breeding values with accuracies upwards of 0.73. The estimation of the breeding value was via determining the conditional mean of the breeding value given the genotype at each QTL using a *prior* distribution of QTL effects.

While this approach has made great advances in animal breeding, its use in plant breeding is just beginning to catch up. A recent review examined the accuracy of genomic prediction for two cereal crops, wheat and maize, and few legume crops like pea, soybean, chickpea, groundnut, and pigeon pea, based on random cross-validation [31]. It includes studies performed on maize and wheat at International Maize and Wheat Improvement Center (CIMMYT), Mexico and on chickpea at International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad in India. An important issue is the incorporation of genotype x environment interaction in genomic selection due to almost universal occurrence of such interactions in plant breeding. Multi-environment trials for assessment of such interactions are used to select high-performing stable lines across environments. Marker- and pedigree-based GBLUP models for assessing these interactions under genomic selection were applied in cotton trials, in genomic selection of extensive wheat gene bank accessions, in genomic selection of Fe and Zn in wheat grain, in genomic selection of bread wheat lines in sites located in diverse ecological zones, in genomic prediction of wheat lines evaluated in Mexico and predicted in locations in South Asia and in genomic selection of extensive field trials in wheat on different continents. In most of the cases genomic selection showed tangible genetic gains.

Future of Crop Improvement by Plant Breeding

The future scenario of plant breeding depends very much on the challenges the agriculture sector will face in times to come. The global population is expected to rise to 9.8 billion by the year 2050 from the present level of 7.6 billion. FAO projects the need of food to grow by about 70 % more which includes additional billion tons of cereals. That means a lot of food needs to be produced in the face of sobering constraints like reduced arable land due to continued urbanization, reduced availability of water, reduced soil health, and above all the impact of climate change which is now a reality compared to its expectations debated in the past. Plant breeders will therefore have to breed better crops that give more yields, are less prone to disease, require less water, are more draught-resistant, and are easier to harvest in addition to their being more nutrient-rich.

How do we achieve such additional food needed? The horizontal expansion, in terms of area under the crop, has already reached a plateau in most of the developing countries. The only way is to go for vertical expansion in terms of productivity which is only possible by scientific and technology innovations. Two issues discussed in this review viz. the genetically modified organisms (GMO) and CRISPR based gene editing technology hold great promise for increased cereal production as demonstrated in developed countries like USA and some developing countries like India but sociopolitical issues are hampering this progress. Of these two ways the CRISPR-associated technique, CRISPR/cas9, is likely to be more receptive to the consumers as well as in not being objectionable to environmental activists as there is no possibility of escape of the transgene to the environment as in GMO since no transgenes are involved, the technique using 'cut' and 'paste' modes as in any word processor.

Climate Change

Plants take energy from the sun, inhale CO_2 from the air and convert water from the soil into sugars and oxygen. With the associated rise in atmospheric CO_2 due to climate change, increased yields are expected due to increased photosynthesis in C_3 category of plants like wheat, rice etc. Such crops may then be packing more of carbohydrates in grains at the expense of protein and other essential elements. This could be detrimental in the long run to produce what are known as junk foods. Experiments by USDA and others are underway to examine such effects to confirm such detrimental effects or otherwise. For C_4 category of plants like maize, sugarcane, sorghum, millets etc. the photosynthetic rates are, however, not expected to increase as much as those in C_3 category.

Biofortification

Genetic improvement of nutritional quality of crops, particularly staple crops, is referred to as *biofortification*. Several efforts at the global level have targeted this aspect. For small shareholder farmers one such program is *HarvestPlus (www.harvestplus.org)* which has resulted in new cultivars with increased levels of iron in beans and millet, increased zinc in rice and wheat and an improved source of vitamin A in cassava, sweet potato, and maize. Bio-fortified foods like orange-fleshed sweet potatoes and golden rice to counter the effects of vitamin A deficiency will therefore need to have more attention of the plant breeding community.

Imaging Technology

Satellite imagery has been in use for decades to collect data on crop health and productivity and are used with ground truth data for farm management decisions. Advanced imaging technology since developed can now enable measurement of diverse crop and plant characteristics in an automated fashion such as by the deployment of drones fitted with several cameras to take large numbers of images of a developing crop [32]. Such data can enable one to know the growth rate of a crop, nitrogen deficiency or disease outbreaks or even to know whether the plants are under stress such as due to drought etc. Hyper-spectral, visible, and near-infrared cameras, capturing hundreds of images, can now routinely be used to assess crop performance as well as measure moisture availability in the soil. Canopy temperature data thus collected can help in identifying drought-tolerant genotypes or even predict biomass yield. Several of such traits are correlated with economic characteristics of the plant and can therefore serve as proxy for them. Canopy temperature is related to root depth as plants that can track water down the soils late in the growing season can better access water. This provides a way for selecting plants with deep roots by selecting for canopy temperature – a form of indirect selection for affecting genetic improvement in deep root. There is however a problem with imagery technology. One cannot take images continuously. So for a developing crop one gets time slots of images in three-dimensional space. If a particular desired time point is missed, one has to reconstruct it. This needs statistical/ mathematical approaches via modeling.

Statistics, Big Data Revolution and Deep Learning

Statistics has been a prerequisite for plant genetics and breeding particularly after discovery of Mendel's laws and their reconciliation with inheritance of quantitative traits exhibiting continuous variation. Quantitative genetics was the outcome of this development. Experimental designs and data analysis in plant breeding have become important tools in the hands of plant breeder [33]. Over time when computers emerged on the scene software development took place at great speed and statistical software for the analysis of plant breeding data became a routine affair. Software like MAPMAKER/QTL etc. has now become standard arsenal in the tool box of a plant breeder. At the same time new technologies like 'chip technology' and others emerged that brought in explosion of data. New statistical methods have been invented to cope with such developments.

The advent of *big data* ushered in a revolution. These are databases that dwarf in size any databases statisticians previously encountered. Newer technologies in several fields like genomics, neurology, social networks, etc. as well as more internet connected devices and machines talking to one another than ever before have led to amassing of unprecedented mountains of information. It is held that around 90 % of all data in the world today has been created over the past two years. The need to store, sift, and make sense of all those petabytes (one million gigabytes) of data has led to massive investments in creating data centers of enormous capacity across the globe. Cloud-computing traffic, the fastest-growing area of data center activity, is expected to grow more than quadruple in a short span of five years. For handling such massive and growing amount of information will need a forward-thinking strategy in addition to using tools like *Hadoop* software framework required for processing large-scale data sets and decidedly recruiting people with the right skills to make sense of it all. And here comes the role of statisticians trained in mathematics, computer science and statistics.

The power of Big Data is being projected from its confines of computer technology and telecommunication to breakthroughs in life sciences and plant genetics and breeding are not left out. Supercomputing power is harnessed to analyze vast amounts of DNA sequencing information that helps in giving life-saving treatments. Now the next big breakthrough might not be found in a test tube but in Big Data.

Statistical challenges with massive data sets occur in several instances. In the analysis of gene expression data sets we encounter high order data matrix with thousands of rows (*p*) representing genes but very small number of columns (*n*) representing samples. This belongs to the general issue known in the statistical literature as 'large *p* and small *n'* problem. Multivariate statistical techniques like principal components, singular value decomposition etc. are often invoked to tackle them. In an another instance on genetic association studies, millions of molecular markers (*p*) are genotyped on a limited number of randomly selected individuals (*n*) on which phenotypic traits are also measured. The problem requires relating the trait values with the millions of predictor values. The usual multiple linear regression method becomes inadequate and sparse regression methods like ridge regression, Lasso, GFLasso or elastic net are needed to make the regression coefficients of irrelevant predictor variables either tend to zero or be exactly at zero so as to reduce the number of predictor variables considerably.

Multiple loci influence complex phenotypic characteristics in plants but genomic selection or GWAS, as discussed earlier, usually focus on single variant (SNP) *at a time* to assess its association to trait and scan millions of SNPs for the trait with appropriate multiple testing corrections. However a complex phenotype may not have a clear pattern of its expression due to possible interactions between the variants themselves or else their interactions with the environment and act in a non-linear fashion. Recent investigations have dealt with such issues by applying deep learning approaches [31].

Deep learning is a newly emerging *artificial intelligence* (AI) technique in which intricate structures in high-dimensional data are discovered by computational models composed of multiple processing layers to *learn* representations of data with multiple layers of abstraction. The important point to note is that these layers of features are not designed *a priori* by us but are *learned* from data *by the machine* through a general-purpose learning procedure like deep learning. We are thus in the domain of AI. Starting with raw input data the machine transforms it into a representation at the next higher level (somewhat abstract) by composing and using a simple but non-linear module. This is repeated for enough successive transformations through modules to enable the machine to *learn* very complex functions. With multiple non-linear layers, a depth of about 5 to 20, a system can implement extremely intricate functions of its inputs that are simultaneously sensitive to minute details as well as insensitive to large irrelevant variations. The technique of stochastic gradient descent (SGD) through back-propagation procedure is used to *train* the multilayer architectures. Many applications of deep learning use *feed-forward neural network* architectures which learn to map a fixed size input to a fixed size output. The basic units of the network are neurons inspired by human brain. From the statistical angle the deep feed-forward neural networks are recursive generalized linear models (RGLMs), the link function being termed as activation function which adds non-linearity to network's function. Common activation functions are sigmoid, tanh (hyperbolic tangent) and reLU (rectified linear unit). For instance in genomic selection the input layer is each SNP marker which is connected to *all* the markers in the first hidden layer and these are connected to all the markers in the second hidden layer, and so on, up to the output layer, which is used for prediction of the phenotypes.

Emergence of Plant Breeding as a Multidisciplinary Activity

The era of plant breeding as an individual discipline is virtually coming to an end. Inputs from crop agronomy, cell and molecular biology, genetics, entomology, pathology, physiology, nutrition, engineering, economics, statistics and mathematics, computer science, bioinformatics and health science are regular features of a successful plant breeding enterprise. Meaningful cooperation is therefore of utmost importance between expertise of these disciplines.

References:

1. Fisher, RA (1918) On correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc., Edinburgh, 52: 399-433.
2. Narain, P (1988) Quantitative inheritance, pp. 311-332 in *Human Population Genetics* edited by K.C. Malhotra. Indian Statistical Institute, Calcutta.
3. Narain, P (1990). *Statistical Genetics*. New York: John Wiley and Wiley Eastern Ltd., New Delhi. Reprinted in 1993. Published by the New Age International Pvt. Ltd., New Delhi in 1999. Reprinted in 2008.
4. Kesevan, PC and Swaminathan, MS (2008) Strategies and models for agricultural sustainability in developing Asian countries. Philos T Roy Soc B 363 (1492): 877-891.
5. Wang, F, Wang, C, Liu, P, Lei, C, Hao, W et al. (2016). Enhanced rice blast resistance by CRISPR/Cas9–targeted mutagenesis of ERF transcription factor gene *OsERF922*. PLoS ONE 11, e0154027.
6. Wang, Y, Cheng, X, Shan, Q, Zhang, Y, Liu, J et al. (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat. Biotech. 32: 947 -951.
7. Shi, J., Gao, H., Wang, H., Lafitte, H. R., Rayeena, L. A. et al. (2017). ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. Plant Biotechnology J. 15: 207-216.
8. Soyk, S., Mueller, N., Park, S., Schmalenbach, I., Jiang, K. et al. (2016). Variation in the flowering gene SP5G promotes day-neutrality and early yield in tomato. Nature Genetics, 49: 162-168.
9. Sax, K (1923). The association of size differences with seed-coat pattern and pigmentation in *phasealus vulgaris*. Genetics, 8: 552-560.
10. Narain, P. (2003). Evolutionary genetics and statistical genomics of quantitative characters. Proc. Indian National Science Academy, Biological Sciences, B 69(3): 273-352.
11. Narain, P (2009). The genetic architecture of quantitative variation. National Academy Science Letters 32: 1-19.
12. Narain, P (2010a). Quantitative genetics: past and present. Molecular Breeding 26: 135-143.
13. Narain, P (2010b). Statistical genomics and bioinformatics. J. Horticultural Science 5: 85-93.
14. Van Ooijen, JW (1999). LOD significance thresholds for QTL analysis in experimental populations of diploid species. Heredity 83: 613-624.
15. Lander, ES and Botstein, D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.
16. Zeng, ZB (1993) Theoretical basis of separation of

multiple linked gene effects on mapping quantitative trait loci. Proc. National Academy of Sciences USA 90: 10972-76.

17. Narain, P (2016a). Statistical aspects of QTL mapping in experimental population. Doi:10.13140/ RG.2.1.1696.7765 (Research Gate).

18. Dragani,TA, Zeng, Z-B, Canzian, F, Gariboldi, M, Ghilarducci, MT et al. (1995) Molecular mapping of body weight loci on mouse chromosome X. Mammalian Genome 6: 778-81.

19. Narain, P (2005). Mapping of quantitative trait loci (QTL). The Mathematics Student, 74: 7-18.

20. Paterson, AH, Lander, ES, Hewitt, JD, Peterson, S, Lincoln, SE and Tanksley, SD (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphism. Nature, 335: 721-726.

21. Grandillo, S, Ku, HM and Tanksley, SD (1999). Identifying the loci responsible for natural variation in fruit size and shape in tomato. Theor. Appl. Genet., 99: 978-987.

22. Frary, A, Nesbitt, TC, Frary, A, Grandillo, S, Knapp et al. (2000). *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. Science, 289: 85-88.

23. Lippman, Z and Tanksley, SD (2001). Dissecting the genetic pathway to extreme fruit size in tomato using a cross between small-fruited wild species *Lycopersicon pimpinellifolium* and *L.esculentum var. Giant Heirloom*. Genetics, 158:413-422.

24. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ et al. (2009). The genetic architecture of maize flowering time. Science 325: 714–718.

25. Wallace, JG, Larson, GJ, and Buckler, ES (2014). Entering the second century of maize quantitative genetics. Heredity 112: 30-38.

26. Jansen, RC and Nap, Jan-Peter. (2001) Genetical genomics: the added value from segregation. Trends in Genetics, 17: 388-391.

27. Collard, BCY and Mackill, DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil. Trans. R. Soc. B 363: 557-572.

28. Jiang, Guo-Liang (2013) Molecular Markers and Marker-Assisted Breeding in Plants. In: *Plant Breeding from Laboratories to Fields*, Chapter 3: 45-83.

29. Ribaut, JM, de Vicente, MC and Delannay, X (2010) Molecular Breeding in developing countries: challenges and perspectives. Current Opinion in Plant Biology 13: 1-6.

30. Meuwissen,THE, Hayes, BJ and Goddard, ME (2001) Prediction of total genetic value using genome wide dense marker maps. Genetics 157: 1819-1829.

31. Crossa, J, Perez-Rodriguez, P, Cuevas, J, Montesinos -Lopez, O, Jarquin, D. et al. (2017) Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science, 1594: 1-15.

32. Council for Agricultural Science and Technology. *Plant Breeding and Genetics – A paper in the series on The Need for Agricultural Innovation to Sustainably Feed the World by 2050.* Issue Paper 57. CAST, Ames, Iowa, USA.

33. Narain, P (2016b). Experimental design and analysis in plant breeding. Doi:10.13140/RG.2.1.2769.1606 (Research Gate).