Genetical genomics: the added value from segregation

Ritsert C. Jansen and Jan-Peter Nap

The recent successes of genome-wide expression profiling in biology tend to overlook the power of genetics. We here propose a merger of genomics and genetics into 'genetical genomics'. This involves expression profiling and marker-based fingerprinting of each individual of a segregating population, and exploits all the statistical tools used in the analysis of quantitative trait loci. Genetical genomics will combine the power of two different worlds in a way that is likely to become instrumental in the further unravelling of metabolic, regulatory and developmental pathways.

Recent advances of microarray expression profiling demonstrate the power of the high-throughput study of multiple genes for unravelling disease resistance[1] and developmental processes in plants[2,3]. Also, proteomics[4,5] and metabolomics[6] aim at the profiling of many gene products in parallel. Statistical and bioinformatical analyses of these profile data reveal genes and gene regulation events by either (non)hierarchical cluster analysis[7], referenced or supervised classification approaches[8,9] or correlation-based analyses[10]. In these genomics approaches, one typically compares two states, such as mutant versus wild type[1], induced versus uninduced[1], or healthy versus diseased, possibly under multiple conditions[1] or over multiple time points[10].

Here, we outline a concept for a strategy, coined 'genetical genomics', which will allow geneticists to take more advantage of genomics[11]. The strategy uses the genetic variation between related individuals in a segregating population and adds the analytical tools available for molecular markers to the analysis of genome-wide expression profile data. In principle, genetical genomics can generate substantial additional insight into the function and interrelation of gene products and gene action from any method of expression profiling based on RNA, protein or metabolites. Although any segregating population is suitable, one type of population could have an advantage over another for practical reasons, such as the ease of production and maintenance of a certain population. Genetical genomics is likely to be most straightforward for self-compatible plants such as *Arabidopsis* and maize. For such plants, a large pedigree of segregating F2 and F3 progeny, recombinant inbred lines (RILs), backcross (BC) progeny, or near isogenic lines (NILs) can readily be

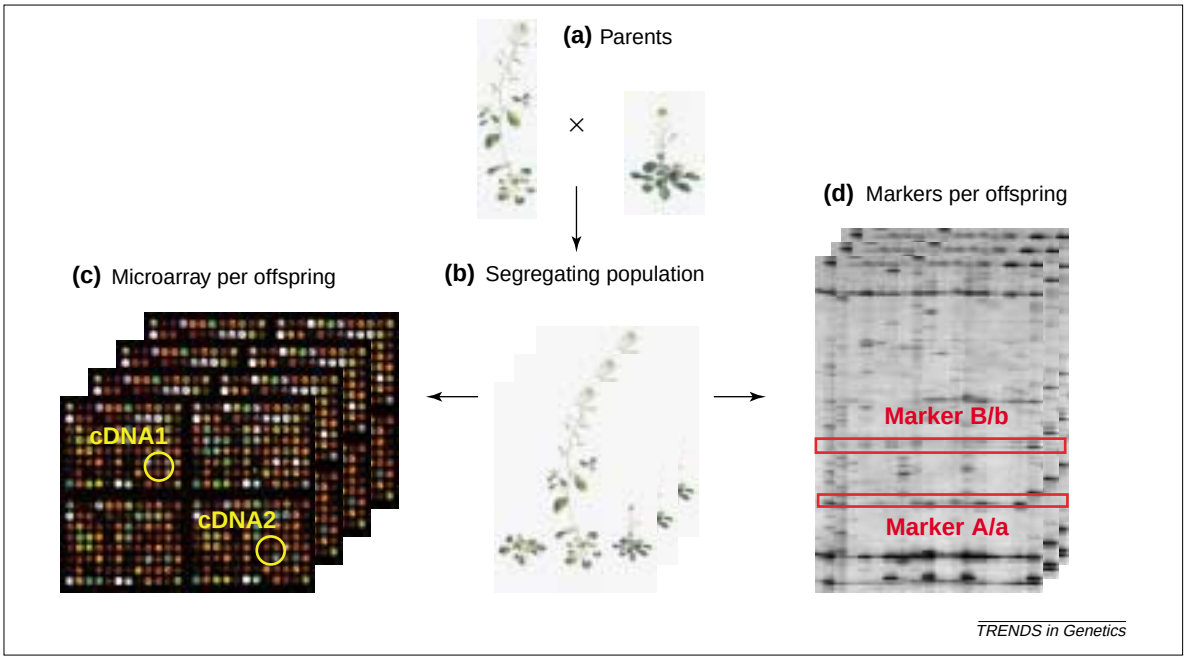

Fig. 1. Expression profiling in combination with molecular marker analysis of a segregating population makes it possible to use quantitative trail loci (QTL) analysis for identification of influential genes and gene products. The example shown here and in Figs 2 and 3 concerns simulated data of both microarray profiling and molecular marker analysis of a hypothetical segregating recombinant inbred line (RIL) of *Arabidopsis*. (a,b) To generate a segregating population, two *Arabidopsis* ecotypes are crossed to generate an F1 offspring (a). One or more F1 plants are self-crossed for a number of generations to generate a RIL segregating population (b). In principle, any segregating population is suitable. (c,d) To analyse the segregating population, each individual of the population is used for analysis by microarray profiling (c), and molecular marker analysis (d). A parent or parental mixture can be used as control in microarray profiling. Any expression profiling method is suitable. The analysis does not need to be genome-wide, although that would add to the reliability of the results obtained. The molecular marker map should at least consist of about one marker of any nature [e.g. amplified fragment length polymorphism (AFLP), single nucleotide polymorphism (SNP), etc.] per about 5–10 cM. The two possible homozygous states at a marker are labelled by a capital and a small character.

R.C. Jansen*
J.P. Nap
Plant Research International BV, PO Box 16, NL-6700 AA Wageningen, The Netherlands.
*e-mail: r.c.jansen@plant.wag-ur.nl

http://tig.trends.com 0168–9525/01/$ – see front matter PII: S0168-9525(01)02310-1

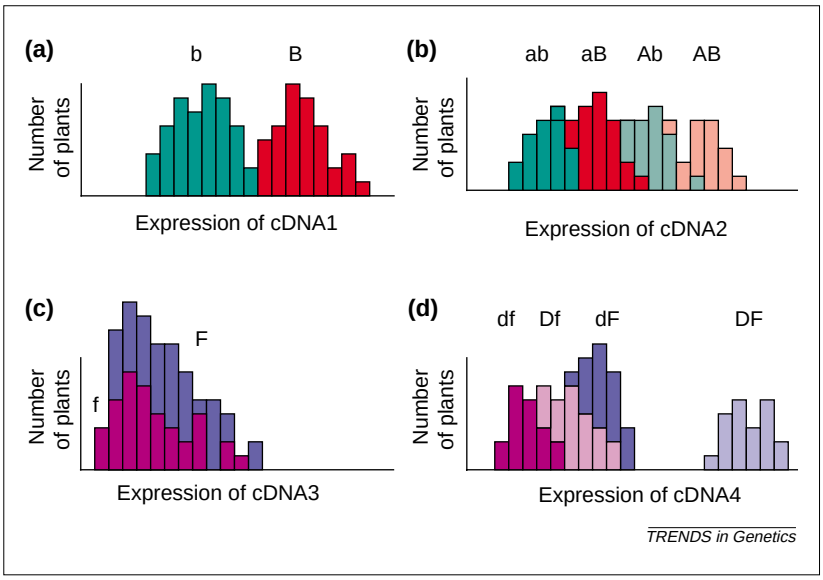

Fig. 2. Combined analysis of expression profile and genetic map data from Fig. 1. Four hypothetical cDNAs are given as example. For each, the microarray data is plotted in a histogram and analysed in combination with the molecular marker data, using multifactorial analysis of variance. The types of allele present for the markers A/a, B/b, D/d and F/f are differentiated by colour. The distribution of expression for cDNA1 shows a qualitative expression (a). Consequently, this is a marker in itself. The distribution of expression of cDNA2 (b) gives a quantitative profile, and its components can be resolved by grouping the segregating alleles at the markers A and B. The distribution of expression of cDNA3 (c) also gives a quantitative profile, but its components cannot be resolved despite the segregation of the F/f alleles of marker F. The quantitative expression profile cDNA4 (d) can be resolved based on D/d and F/f. Thus, marker F contributes to information about other cDNAs, whereas F by itself is not informative. Standard one-way analysis of variance of the profile of cDNA2 would distinguish Ab and AB (light shades) from ab and aB (dark shades), but a two-way analysis of variance is more appropriate and allows distinction between the four classes ab, aB, Ab and AB. For cDNA4, the same is true for the classes df, Df, dF and DF. In the case of cDNA2, there is no epistatic interaction, whereas the position of DF relative to the other three classes in case of cDNA4 indicates such an interaction. The multifactorial analysis of variance can easily be extended to three or more markers. It will allow assessing the significance and the strength of the actions and interactions of gene products.

obtained from a cross between two widely separated inbred lines (Fig. 1a,b). However, the concept can easily be extended to segregating pedigrees of laboratory animals, livestock and man.

About a decade ago molecular markers promised to revolutionize the unravelling of quantitative traits into multiple quantitative trait loci (QTLs), and an elaborate toolbox for statistical QTL mapping is now available[12,13]. The expression profiling of all individuals in a segregating population (Fig. 1c,d) makes it possible to treat the expression profile of each gene in the population as a quantitative trait. Combined with a genetic map, QTL mapping methodology allows the multifactorial dissection of the expression profile of a given RNA (cDNA), protein or metabolite into its underlying genetic components and their genetic map positions (Figs 1–3).

In the analysis of expression profile data, the expression of some genes can turn out to be effectively qualitative, whereas that of other genes will be quantitative. A qualitative profile would establish a marker in itself (Fig. 2a). If present in sufficient numbers, such qualitative profiles will immediately allow the construction of a genetic map. It is anticipated, however, that due to environmental and multigenic variation, the majority of expression profiles will be quantitative (Fig. 2b–d). By the appropriate multifactorial analysis of variance, such profiles can be resolved into the contributing loci on the molecular marker map (Fig. 3a).

A major source of 'experimental noise' in current expression profiling is the large variation between plants of the same genotype, despite well-controlled environmental conditions[3,6,14]. In microarray profiling, the difference in expression level needs to be fourfold or larger to be detected reliably[3]. In a segregating RIL population of 100 diploid individuals, each allelic state of each gene is replicated 50-fold, although in different genetic backgrounds caused by variations of all other genes. A 50-fold replication will reduce the standard error of the 'experimental noise' in profiling data about sevenfold [$\sqrt{(\text{replication})} = \sqrt{(50)} = 7.1$]. Such a reduction increases the power for the detection of contrasts considerably and will help the functional interpretation of profiling data. In view of the current experiences with variation between genetically identical plants, it is clearly preferable to reduce experimental noise further by profiling pooled material from several plants per RIL or from several F3 plants per F2, without the need for more profiling experiments.

In segregating populations, the inherent variation in genetic background between individuals can give additional experimental noise due to additive effects and/or epistatic interactions, but current tools of QTL analysis allow filtering out of such experimental noise[12,15]. Therefore, the 50-fold replication obtained by profiling a segregating population (of 100 diploid progeny) can be about equivalent to the replication of the profile of the same genotype.

When the segregating population has been screened for a phenotypic trait of interest and shows QTLs for that trait, the combined approach will help identify the gene(s) responsible for such QTLs. If a particular phenotype is of interest, the genetics of segregating populations could be used in more simple ways, such as in selective genotyping or bulked segregant analysis[12]. In such cases, however, the power of multifactorial analysis is lost. The analysis zooms in on preselected regions of the genome and genes with an important influence can be missed.

We predict that genetical genomics will have most value for the unravelling of genes and gene products that are involved in metabolic and regulatory (e.g. developmental) pathways. The principle is illustrated in Fig. 3. For each gene (cDNA) or gene product analysed in the segregating population, QTL analysis will pinpoint the regions of the genome influential for its expression. If available, the sequence and annotation of that genomic region would be helpful for the identification of the genes involved. Genetical

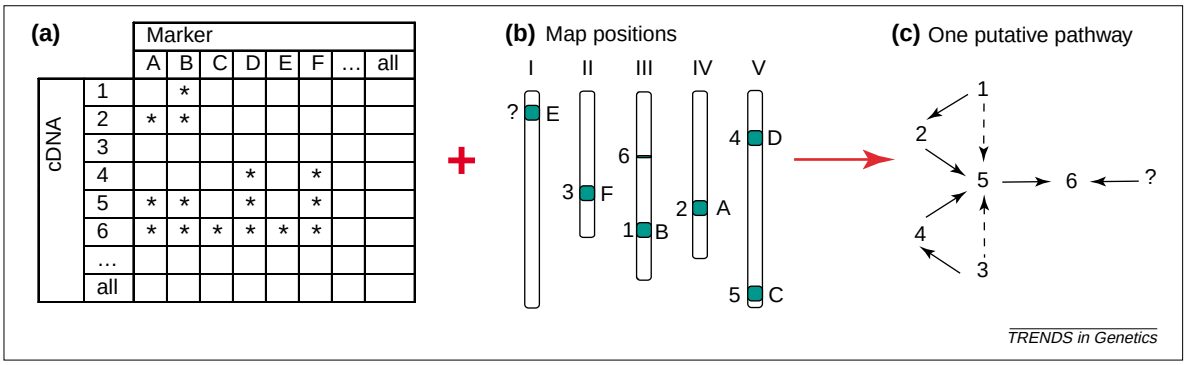

Fig. 3. Pathway reconstruction and pathway 'memory'. The QTL analyses of expression data allow deduction of the relationships of the expression profiled cDNAs. The table (a) gives the hypothetical result of analysing a few of the expression profile data from Fig. 2. The QTL analysis associates the individual cDNAs of the array with markers of the genetic map in a two-way table. Significant differences between the two allelic states of each marker are shown as an asterisk. Given the entire genome sequence of *Arabidopsis*, the genomic position of each cDNA used in the array is known. The position on the genetic map of each marker (in capital letters) and of each cDNA (in numbers) on the five chromosomes of *Arabidopsis* is given (b). In combination, these data can be used to analyse which genes influence the expression of other genes and to deduce in what order this influence is exerted. For example, the expression of a cDNA (cDNA5) is mapped to the regions of the markers A, B, D and F. The cDNA5 itself is located close to marker C. The expression of another cDNA (cDNA6) is mapped to the same four regions, and to two additional regions, C and E. The likelihood of such a fourfold coincidence is very small. Therefore, this indicates that cDNA5 and cDNA6 are likely to operate in the same pathway or network. Because cDNA5 is located in region C and this region has an influence on the expression of cDNA6, the genes encoding these cDNAs are likely to act in series (c). This part of the example illustrates how genetical genomics can reconstruct a step in a pathway or network. The next step is to scan the four regions A, B, D and F for cDNAs whose expression maps to the larger subset of these four regions. A relevant cDNA could map, for instance, to the regions A, B and D. In our example, the cDNA located in region D (cDNA4) maps to the regions D and F. Another cDNA (cDNA2) is located in region A and maps to the regions A and B. Therefore, both cDNA2 and cDNA4 are good candidates for a step in the pathway preceding the involvement of cDNA5 and cDNA6. The same line of reasoning will apply for the next step and concerns the involvement of cDNA1 and cDNA3. This analysis indicates that cDNA4 is acting on cDNA5 and cDNA3 is acting on cDNA4 and cDNA5. However, at this point the possibility that cDNA3 influences both cDNA4 and cDNA5 at the same time cannot be excluded. The putative direct influence of cDNA3 on cDNA5 is given by a broken arrow (c). For clarity, not all putative direct interactions are indicated. There are several special cases to consider in the example given. There are cDNAs that have no differential allele expression at the time of sampling (cDNA3,5,6). Other cDNAs map onto themselves (cDNA1,2,4). Their two alleles apparently have different expression levels, which can be due to a difference in the promoter region. A cDNA can also map to a region that appears not to contain a subsequent candidate gene in the pathway (cDNA6 maps to region E). A reason could be that the causal cDNA from region E, indicated by the question mark (b,c), is missing from the array. With genetical genomics we can still detect and locate this unknown cDNA (e.g. a transcription factor) through the expression of cDNA6, which it has influenced.

genomics presents a novel strategy for identifying such candidate gene(s) by combining the QTL information from all genes and gene products that are analysed. It will indicate what portion of the variation in gene expression maps to the genes themselves (*cis*-acting factors), as opposed to other genomic locations (*trans*-acting factors). Candidate genes identified by genetical genomics can involve genes not present in the microarray, genes with very low expression levels, or genes with influential expression at a time (long) before sampling of RNA, protein or metabolite. Analyses could also allow demonstrating fuzzy or epigenetic interactions between genes.

Inevitably, owing to the nature of statistical analysis, some false positives will be identified. Obviously, the importance of any candidate gene must therefore be validated; for example, with the experimental toolbox of functional genomics[16], such as mutation and transformation. The possibility to detect genes with an important expression in the past allows the 'memory' of a pathway to be visualized and to contribute to the reconstruction of pathways. No such results can ever be expected from current pairwise or multiple condition expression profiling.

To our knowledge, the whole-genome data necessary to validate the concept outlined here are not yet available in the public domain. Such data should preferably be generated in a model species, such as *Arabidopsis*, for which the entire genome sequence is known[17], segregating populations and extensive molecular marker maps are available, and expression profiling is in place. We therefore expect that the proposed concept of genetical genomics can be validated in the near future. At present, microarray profiling seems most appropriate for the whole-genome analysis of gene expression[2,3], however the appropriate proteomics[4,18,19] and metabolomics[6] technologies are developing rapidly. In the future, a combination of the different profiling methods is likely to be most informative. Large-scale expression profiling of segregating populations is at present too expensive for most laboratories. The added value outlined above should stimulate the development of appropriate and cost-effective analytical and statistical technologies: the merger of genomics and genetics will indeed combine the better of two worlds.

Acknowledgements
For discussions on and contributions to the concept of genetical genomics, we thank Prof. Gerard Jansen and various colleagues at Wageningen University and Research Centre, in particular Dr Johan van Ooijen and Dr Sjaak van Heusden. Funding was obtained from the Dutch Ministry for Agriculture, Nature Management and Fisheries.

References

1 Maleck, K. *et al.* (2000) The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nat. Genet.* 26, 403–410
2 Girke, T. *et al.* (2000) Microarray analysis of developing *Arabidopsis* seeds. *Plant Physiol.* 124, 1570–1581
3 Aharoni, A. *et al.* (2000) Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *Plant Cell* 12, 647–661
4 Zhu, H. *et al.* (2000) Analysis of yeast protein kinases using protein chips. *Nat. Genet.* 26, 283–289

5 Celis, J.E. *et al.* (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.* 480, 2–16
6 Fiehn, O. *et al.* (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18, 1157–1161
7 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912
8 Brown, M.P.S. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U. S. A.* 97, 262–267
9 Gaasterland T. and Bekiranov, S. (2000) Making the most of microarray data. *Nat. Genet.* 24, 204–206
10 Cohen, B.A. *et al.* (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26, 183–186
11 Johnston, M. and Fields, S. (2000) Grass-roots genomics. *Nat. Genet.* 24, 5–6
12 Lynch, M. and Walsh, B., eds (1998) *Genetics and Analysis of Quantitative Traits*, Sinnauer
13 Jansen, R.C. (2001) Quantitative trait loci in inbred lines. In *Handbook of Statistical Genetics* (Balding, D. *et al.*, eds), pp. 567–597, John Wiley & Sons
14 Mlynarova, L. *et al.* (1996) Approaching the lower limits of transgene variability. *Plant Cell* 8, 1589–1599
15 Zeng, Z-B. *et al.* (1999) Estimating the genetic architecture of quantitative traits. *Genet. Res.* 74, 279–289
16 Pereira, A. (2000) A transgenic perspective on plant functional genomics. *Transgenic Res.* 9, 245–260
17 The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
18 Thiellement, H. *et al.* (1999) Proteomics for genetic and physiological studies in plants. *Electrophoresis* 20, 2013–2026
19 Harry, J.L. *et al.* (2000) Proteomics: capacity versus utility. *Electrophoresis* 21, 1071–1081

SIR2 and aging – the exception that proves the rule

Leonard Guarente

One of the holy grails of medicine is the possibility of an increase in lifespan without a decrease in vitality. However, the causes and processes of human aging are still unclear. One evolutionary theory is that in the post-reproductive stage of life, selective forces decline allowing many vital systems to deteriorate. This suggests that intervention will be difficult, if not impossible. However, molecular geneticists propose an aging process that is programmed (like other developmental processes) and regulated by single genes, meaning that intervention could be possible. Here, I discuss a way of reconciling these two views that could have major implications for healthcare.

A theory of evolutionary biology is that aging is caused by the deterioration of many processes as selective forces wane in the post-reproductive phase of life[1,2]. A corollary of this view is that no single intervention in the aging process should extend lifespan because the decline in vitality affects many processes. However, research from the past decade shows that single-gene mutations can have profound effects on the longevity of laboratory organisms such as yeast[3,4], *Caenorhabditis elegans*[5–7], *Drosophila*[8,9] and mice[10,11]. Thus, molecular geneticists skeptical of the evolutionary view might deduce that aging is programmed in a manner not unlike embryonic development and is therefore subject to regulation by single genes. Here, I describe a third model that is consistent with our studies of Sir2 genes and that reconciles the two extreme views. This notion could have far-reaching implications for strategies of intervention in the aging process.

L. Guarente
Dept of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
e-mail: leng@MIT.edu

In mother cells of *Saccharomyces cerevisiae*, the *SIR2* gene is a key determinant of lifespan: null mutations shorten lifespan and an extra copy of *SIR2* can extend lifespan[12]. These effects appear to derive from the silencing of chromatin in the ribosomal DNA (rDNA) repeats by Sir2p. This silencing reduces rDNA gene expression and suppresses recombination that would generate toxic extrachromosomal rDNA circles (ERCs)[13]. The aging in yeast mother cells results from an asymmetry of cell division, leading to the accumulation of ERCs and perhaps other deleterious molecules.

Aging in *C. elegans* appears to be fundamentally different from yeast aging in that the soma of adults consists solely of post-mitotic cells. Nevertheless, a *SIR2* homolog, *sir-2.1*, seems to regulate lifespan: transgenic worms with extra copies of *sir-2.1* live longer[14]. How does SIR-2.1 regulate aging in worms? Genetic analysis indicates that SIR-2.1 probably functions in the insulin signaling pathway. Mutations in components of this pathway reduce signaling and confer longevity[15–18]. Perhaps SIR-2.1 normally represses one or more components of this pathway.

Why are replicative aging in yeast and post-mitotic aging in worms both regulated by Sir2 genes? An important insight comes from a consideration of the biochemical activity of Sir2 proteins as NAD-dependent protein deacetylases[19–21]. As such, Sir2 proteins might have the ability to monitor metabolic rate, reflected by the amount of available NAD, and couple this status to regulatory events, such as the silencing of chromatin[22]. An important consequence follows from this model: Sir2 genes might regulate aging in many different organisms, even though the molecular events that it controls are disparate (i.e. rDNA silencing in yeast and insulin signaling in worms). Thus, even if the molecular causes of aging are different in different organisms, aging might still be controlled by SIR2 deacetylation of histones or, perhaps, other protein substrates. The common role of Sir2 genes would be coordinating the pace

http://tig.trends.com 0168–9525/01/$ – see front matter PII: S0168-9525(01)02339-3