NATURE|Vol 456|11 December|doi:10.1038/nature07629

Next-generation genetics in plants

Magnus Nordborg[1] & Detlef Weigel[2]

Natural variation presents one of the fundamental challenges of modern biology. Soon, the genome sequences of thousands of individuals will be known for each of several species. But how does the genotypic variation that will be observed among these individuals translate into phenotypic variation? Plants are in many ways ideal for addressing this question, and resources that are unmatched, except in humans, have now been developed.

When it comes to dissecting complex traits, genetic studies of plants have always been at the forefront. Work with cereals provided the first demonstrations that segregation at multiple loci could give rise to a continuous distribution of phenotypes[1,2], and it was in peas that individual quantitative trait locus (QTL) effects were first inferred using markers[3]. Many genes responsible for QTL effects have now been cloned in plants, mainly as a result of the persistent application of classical genetic techniques. And it is not just the 'flagship' of plant biology, *Arabidopsis thaliana*, that has proved useful in such studies. Some of the earliest successes came from studying crop plants, such as maize (corn), rice and tomato[4–7].

The genetic maps of many organisms are now becoming increasingly dense, and the cost of genotyping is decreasing. For these reasons, it has become almost routine to identify the genes (or QTLs), and even the individual nucleotides (QTNs), that cause particular phenotypic effects, using a process that starts with linkage mapping (conventional genetic mapping based on the idea that the farther apart two linked genes are, the more likely a recombination event between them will be) in populations derived from crosses between divergent strains. Notably, despite the extensive genetic screens for mutants of *A. thaliana* affecting a wide variety of phenotypes, the above approach has enabled plant geneticists to identify numerous QTL genes, which were not previously known to have an impact on the trait examined (see refs 8–11 for examples).

High-resolution linkage mapping is slow and labour intensive, however, and for all the success stories, there are probably at least as many cases in which QTL cloning efforts were abandoned (and therefore not reported in the literature) because of difficulty with fine mapping, complex genetic architecture and so on. Therefore, there is great interest in developing an alternative technique — genome-wide association (GWA) mapping — which looks for associations between phenotypes of interest and the DNA sequence variants present in an individual's genome, as assessed by determining an individual's genotype at the positions of hundreds of thousands of single nucleotide polymorphisms (SNPs). GWA studies provide much higher resolution than linkage mapping because they involve studying a natural population rather than the offspring of crosses, and associations in natural populations are typically on a much finer scale because they reflect historical recombination events. Here we describe the recent progress in realizing the potential of GWA mapping in plants, and we then discuss the rapid advances that are expected during the next year or two.

Considerations for GWA mapping

Many plants are well suited for GWA studies, in particular species that self-fertilize, such as *A. thaliana* and rice (*Oryza sativa*), and species that can be clonally propagated, such as switchgrass (*Panicum virgatum*) and grape (*Vitis vinifera*). This is because after lines of these species have been densely genotyped or completely sequenced, the plant genetics community can analyse an unlimited number of traits in genetically identical material. This ability to examine individuals with identical genotypes greatly improves statistical power when studying phenotypes with complex genetics, especially when the phenotypes are modified by the environment.

With association mapping, however, the structure of the population can be a strong confounding factor (discussed later), especially for traits that are important in adaptation to the local environment. There are statistical solutions to this problem, but it is important to recognize that these do not always work and that linkage mapping in controlled crosses (which is robust to confounding) is sometimes necessary, perhaps as a complement to GWA mapping (Fig. 1).

Another consideration is that these two mapping techniques — association mapping and linkage mapping — also differ in terms of how the genetic architecture of the trait affects statistical power (Fig. 2). The importance of a particular allele in phenotypic variation across a population depends on its frequency, as well as on its effect. Thus, GWA mapping has low power to detect rare alleles, even if these alleles have a large phenotypic effect (Fig. 2c). Conversely, alleles that are identified by mapping in crosses between two essentially randomly chosen parents can have a large effect but might not be important from an evolutionary perspective because they are rare. In other words, mapping QTLs by using crosses might bias researchers towards identifying rare (and often perhaps deleterious) alleles that have large effects but little relevance to most of the phenotypic diversity found in nature.

Resources for GWA mapping

The development of reference resources for genomic polymorphism in *A. thaliana* mirrored the progress of the International HapMap Project, which identified SNPs in the human population worldwide[12,13], and resources are now at a stage similar to those for humans. To build these resources, a pilot study was carried out to investigate the basic pattern of polymorphism in *A. thaliana* (that is, the level of variability, extent of linkage disequilibrium, and structure of the population): this involved direct sequencing of 1,400 loci in 96 individuals[14]. Then, a subset of 20 of these individuals was selected for resequencing using a high-density microarray[15]. The SNPs uncovered by this study[15] led to the design, in conjunction with Affymetrix, of a genotyping microarray with 250,000 probes for the purpose of GWA studies of *A. thaliana*[16]. This microarray is now being used to genotype about 1,300 naturally inbred lines (see Genomic polymorphism data in *Arabidopsis thaliana*, http://walnut.usc.edu/2010), which will be distributed to the community for extensive phenotyping. The project is likely to be a model for analogous projects in other species, including one that is underway to

[1]Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, USA. [2]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany.

study rice (see the International Rice Functional Genomics Consortium, http://irfgc.irri.org).

As was the case for the International HapMap Project in humans, many approaches used for the study of *A. thaliana* and rice have already been rendered obsolete by technical advances. Future efforts to uncover SNPs will use next-generation sequencing approaches (such as Illumina's sequencing technology and Applied Biosystems's SOLiD System, which are already commercially available) rather than the microarray-hybridization technology that was used to construct the first-generation haplotype maps of *A. thaliana*[15] and rice, a technology that was more costly and less precise than newer approaches and was highly biased. In addition, rapidly decreasing costs mean that sophisticated schemes that have been used to select the most informative SNPs for genotyping are increasingly becoming irrelevant. By the time that researchers had determined that 140,000 tag SNPs (a subset of informative SNPs) would suffice to cover the *A. thaliana* genome, there was no economic reason not to genotype all 250,000 known high-quality SNPs that were not singletons (which are SNPs that have been detected only in single individuals and whose predictive power for other SNPs is therefore unknown)[16].

The importance of population structure

What, therefore, is the prospect of pinpointing individual genes with GWA approaches? It is well known that demography affects linkage disequilibrium. One example is that there is more linkage disequilibrium in Europeans than in Africans, reflecting humanity's African origins[12,13]. Another is that for wild *A. thaliana*, linkage disequilibrium is more extensive in North America than in Europe, consistent with the plant having been introduced into North America only after Europeans settled there[14,16]. In both cases, the probable explanation is that there was a bottleneck in colonization, with recombination not yet having had enough time to whittle down linkage disequilibrium among the alleles present on the limited number of founder chromosomes.

It is perhaps not as widely recognized that, in the presence of population structure, the genetic architecture of a trait in a sample of individuals depends on how the sample was assembled. For example, GWA mapping immediately reveals the importance of the gene *FRIGIDA* in the variation in flowering time among *A. thaliana* strains from the northwestern parts of continental Europe (where common loss-of-function alleles are an important determinant of early flowering) but not from central Asia (where no single loss-of-function allele is particularly frequent[17]). If variation in a trait is caused by numerous alleles of a single gene (as opposed to a small number of frequently occurring alleles), then researchers carrying out a GWA scan using global samples run the risk of concluding that there is no major locus for this trait (Fig. 2). This is essentially another facet of the problem with population structure that was mentioned earlier: the importance of a particular allele always depends on the reference population, and it is far from clear which population is meaningful from an evolutionary perspective.

Much attention has been given to population structure being a strong confounding factor in association studies, especially for traits that are important in local adaptation (such as flowering time in plants or skin colour in humans). Studies of maize and *A. thaliana*[18–22] have been at the forefront of identifying this problem and indicating statistical solutions. Application of one of these strategies[20] has already led to the identification of a major locus in maize that controls concentrations of provitamin A — an important trait, particularly for people with limited access to a diverse diet[23].

Combining association mapping and linkage mapping

A clear solution to the problem posed by population structure is to complement GWA studies of natural populations with linkage mapping of experimental populations, taking advantage of the increased resolution of the former and the robustness to confounding of the latter, a strategy that has been successfully applied to *A. thaliana*[11,22,24]. When studying human genetics, however, controlled crosses are not possible, so the solution is to use the transmission-disequilibrium test (TDT)[25], which uses the transmission of alleles from parents to offspring to verify linkage.

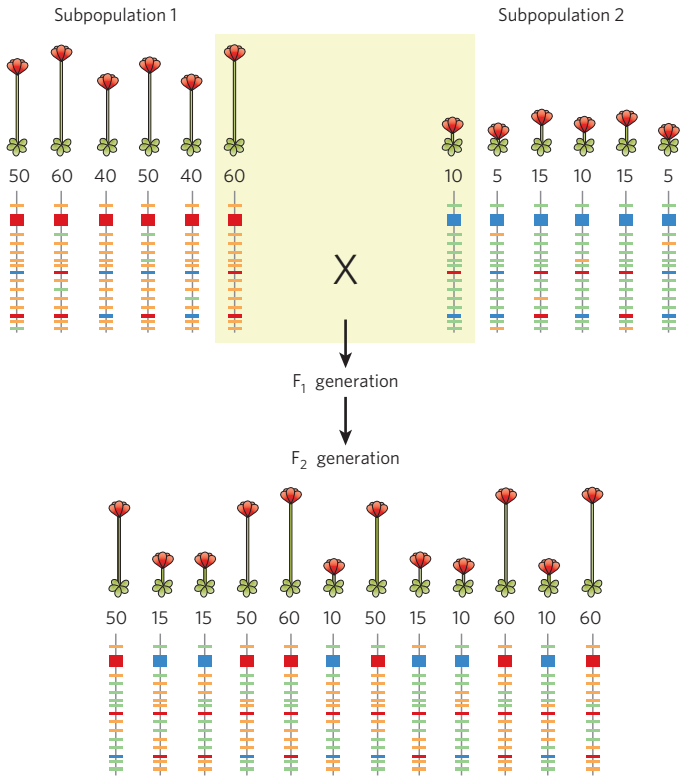

Figure 1 | GWA mapping is ineffective if there is strong genetic differentiation between subpopulations (that is, if there is structure in the population). In this example, two subpopulations of plants are depicted, one tall and one short (as illustrated and indicated by the numerical measurement), together with a schema of the genotype of each plant. The presence of red alleles increases the height of a plant, whereas blue alleles decrease the height; one locus has a major effect, and two have a minor effect. The many background markers (orange and green) are mostly exclusive to a specific subpopulation but are also strongly associated with height, even though they are not causal. By crossing the plants (shaded area) and generating an experimental population of F_2 generation or recombinant inbred lines, any linkage disequilibrium between background markers and causal markers is broken up, and the causal loci can then easily be mapped, albeit with relatively poor resolution.

For maize, by contrast, Ed Buckler and colleagues have pioneered a distinct approach, which is called nested association mapping[26]. GWA studies such as those underway in humans and *A. thaliana* would, at least for the next couple of years, be prohibitively expensive in maize, because its genome is larger than that of humans, is more polymorphic and has less-extensive linkage disequilibrium. Instead, 5,000 recombinant inbred lines (RILs) have been derived from separate crosses of a common standard genotype with 25 genetically diverse lines. The founder lines will be sequenced, whereas the RILs will be genotyped only with sufficient density to identify the ancestral founder at each point in the genome, resulting in a haplotype map that is essentially complete for each of the 5,000 RILs. Because crossing over during RIL formation is limited, such mapping can be accomplished with relatively high accuracy by using a moderate number of markers[26]. This mapping approach is conceptually similar to those applied to a heterogeneous stock of laboratory mice[27] or the Collaborative Cross[28] (a resource that is being generated with the aim of obtaining 1,000 RILs from eight standard mouse strains) (see page 724).

The nested-association-mapping design therefore, in effect, relies on the experimental crosses to map genes — without the confounding effects of population structure — to only a few, but still relatively large, genomic regions. Within these mapping intervals, allele sharing across the founder lines is exploited to achieve the resolution of GWA mapping. It is easy to see how this strategy could be applied to *A. thaliana* by the appropriate selection of subsets of lines from the many available RIL populations.

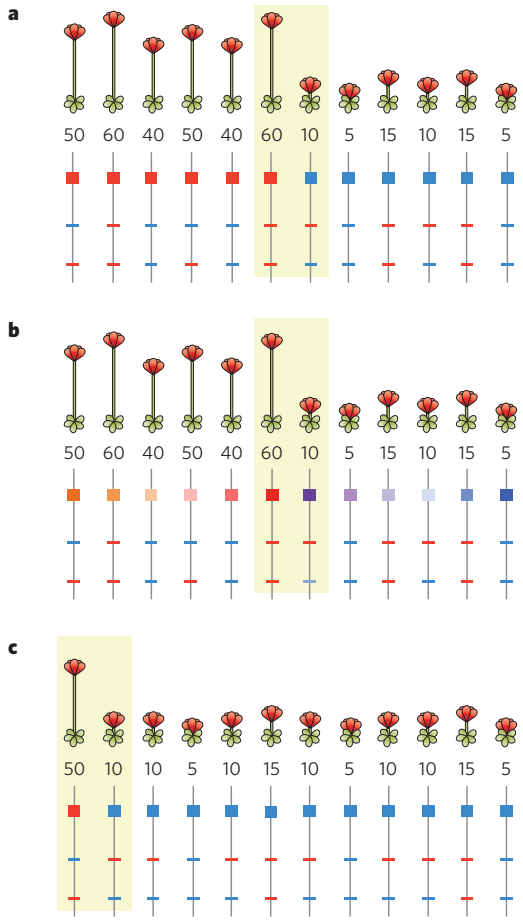

Figure 2 | The performance of GWA mapping depends strongly on the genetic architecture of the trait. Plants are depicted as in Fig. 1. Red alleles increase plant height, whereas blue alleles decrease height; one locus has a major effect, and two have a minor effect. **a**, Provided that there is no confounding effect of population structure, the major causal locus of a trait can easily be mapped by a GWA scan (otherwise an approach such as that in Fig. 1 is needed). **b**, Even if there is a major locus for the trait, the presence of many different alleles of the causal gene (multiple colours) makes naive GWA scans (such as in **a**) ineffective. **c**, GWA scans miss rarely occurring alleles that have a large effect (the top locus) but locate common alleles with smaller effects (the centre locus). In all three cases (**a**–**c**), the causal loci can be identified in experimental crosses (shaded areas) by using linkage mapping instead of GWA mapping, but alleles identified in this way might be rare in the population and therefore not informative for most individuals and not representative of the population.

Future directions

In the past two years, numerous GWA studies of humans have been published[29]. Despite the success of these studies, the identified alleles typically provide an explanation for only a small proportion of the variation in the population[30]. Similar studies in plants, which can be expected to be published in the next two years, will tell the genetics community whether alleles with large effects are more common in plants than was found in human GWA studies.

A serious deficiency in the understanding of plant evolution stems from the inability to compare the genomes of close relatives. The immense value of such studies has been aptly demonstrated in yeasts, *Caenorhabditis*, *Drosophila* and primates. Comparisons of recently diverged genomes allow not only the mapping of conserved genomic elements but also the detection of lineage-specific selection or the identification of recently introgressed segments from relatives. The last is of particular interest in plants, for which interspecific hybrids can often give rise to viable offspring. The closest relative of *A. thaliana* for which a genome sequence is available is papaya, which diverged from *A. thaliana* 70 million years (Myr) ago[31]. This lack of appropriate genome sequence information is ironic, because for years there has been more information about within-species polymorphism in plants than in other non-human organisms[14,15,32]. The plant genetics community is therefore excited that the sequencing of the genomes of *Arabidopsis lyrata* and *Capsella rubella* — close relatives of *A. thaliana* that diverged only about 5 Myr ago and 10 Myr ago, respectively — is almost complete. Similarly, sequencing the genomes of wild progenitors of rice holds the promise of providing important information about the domestication of rice.

Despite the excitement about studying variation between species and between genera, the most revealing insights are likely to come from much more detailed studies of within-species sequence diversity. Until recently, such analyses have focused on relatively minor differences between individuals, such as SNPs, and small insertions and deletions. It has become increasingly clear, however, that individuals, whether maize, *A. thaliana* or humans, can differ by the presence of hundreds of genes and by large genomic rearrangements[1,5,33–35].

One such in-depth study of within-species diversity is the recently launched '1000 Genomes' project (http://1000genomes.org), which seeks to generate a deep catalogue of genetic variation by sequencing the genomes of at least 1,000 individuals from around the world, using next-generation sequencing technologies. We are advocates of a complementary project for *A. thaliana*: 1001 Genomes (http://1001genomes.org). *A. thaliana* is almost uniquely suited to such an effort, given the advantages it offers for analysing population genetics and studying genotype–environment interactions. First, *A. thaliana* genomes are less than one-twentieth the size of human genomes and are much less repetitive in structure, making them less expensive to sequence and much easier to reassemble than human genomes. Second, the genomes of most *A. thaliana* accessions (or strains) are naturally inbred, whereas humans (and most model animals) are heterozygous. Therefore, when using a whole-genome shotgun strategy to sequence *A. thaliana* genomes, the problems of determining haplotype phase (that is, the arrangement of alleles on homologous chromosomes) are avoided, greatly simplifying downstream analyses. Last, for practical and ethical reasons, it will not be possible to use the individuals of the 1000 Genomes project directly in GWA studies, with the exception of measuring phenotypes for cell lines derived from samples from the 1,000 individuals. Instead, the information will need to be used as a resource to impute genotypes in other samples. By contrast, an endless supply of seeds can be produced for any naturally inbred *A. thaliana* accession that is chosen for genome sequencing. Therefore, any number of plants with an identical genotype can be grown for each accession and then phenotyped in as many environments as desired. So the sequence information that is collected can be used directly in GWA studies at many levels, including at the biochemical, metabolic, physiological, morphological and whole-plant-fitness levels.

With the combination of GWA studies and forward-genetic approaches, it will finally become possible to bridge the genotype–phenotype divide, at least in *A. thaliana*. And, as similar projects are set up to study other species, it will become possible to answer general questions about the molecular genetic basis of evolutionary change. ■

1. Nilsson-Ehle, H. Kreuzungsuntersuchungen an Hafer und Weizen. 1. *Lunds Universitets Årsskrift* **5,** 1–122 (1909).
2. East, E. M. A Mendelian interpretation of variation that is apparently continuous. *Am. Nat.* **44,** 65–82 (1910).
3. Sax, K. The association of size differences with seed coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8,** 552–560 (1923).
4. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386,** 485–488 (1997).
5. Frary, A. *et al. fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289,** 85–88 (2000).
6. Fridman, E., Pleban, T. & Zamir, D. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl Acad. Sci. USA* **97,** 4718–4723 (2000).
This study pinpointed a QTL to a small intragenic region by using recombination mapping.
7. Yano, M. *et al. Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering-time gene *CONSTANS*. *Plant Cell* **12,** 2473–2483 (2000).
8. Bentsink, L., Jowett, J., Hanhart, C. J. & Koornneef, M. Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **103,** 17042–17047 (2006).

9. Mouchel, C. F., Briggs, G. C. & Hardtke, C. S. Natural genetic variation in *Arabidopsis* identifies *BREVIS RADIX*, a novel regulator of cell proliferation and elongation in the root. *Genes Dev.* **18,** 700–714 (2004).
10. Macquet, A. *et al.* A naturally occurring mutation in an *Arabidopsis* accession affects a β-D-galactosidase that increases the hydrophilic potential of rhamnogalacturonan I in seed mucilage. *Plant Cell* **19,** 3990–4006 (2007).
11. Baxter, I. *et al.* Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (*MOT1*). *PLoS Genet.* **4,** e1000004 (2008).
12. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
13. Frazer, K. A. *et al.* A second-generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
14. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3,** e196 (2005).
15. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317,** 338–342 (2007).
This study provided the first genomic view of the pattern of polymorphism in a plant.
16. Kim, S. *et al*. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genet.* **39,** 1151–1155 (2007).
17. Shindo C. *et al.* Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of *Arabidopsis*. *Plant Physiol.* **138,** 1163–1173 (2005).
18. Thornsberry, J. M. *et al. Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genet.* **28,** 286–289 (2001).
19. Flint-Garcia, S. A. *et al.* Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* **44,** 1054–1064 (2005).
20. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet.* **38,** 303–308 (2006).
This study identified an effective way to account for population structure in association mapping, an innovative approach that has had a large impact on the field.
21. Aranzana, M. J. *et al.* Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.* **1,** e60 (2005).
22. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3,** e4 (2007).
23. Harjes, C. E. *et al.* Natural genetic variation in *lycopene epsilon cyclase* tapped for maize biofortification. *Science* **319,** 330–333 (2008).
24. Balasubramanian, S. *et al.* The *PHYTOCHROME C* photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*. *Nature Genet.* **38,** 711–715 (2006).
25. Ewens, W. & Spielman, R. The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.* **57,** 455–464 (1995).
26. Yu, J., Holland, J. B., McMullen, M. D. & Buckler, E. S. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178,** 539–551 (2008).
This report introduced an elegant experimental combination that simultaneously exploits the advantages of mapping in experimental crosses and the increased resolution of association mapping.
27. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genet.* **38,** 879–887 (2006).
28. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genet.* **36,** 1133–1137 (2004).
29. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9,** 356–369 (2008).
30. Iles, M. M. What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet.* **4,** e33 (2008).
31. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452,** 991–996 (2008).
32. Wright, S. I. The effects of artificial selection on the maize genome. *Science* **308,** 1310–1314 (2005).
33. Buckler, E. S., Gaut, B. S. & McMullen, M. D. Molecular and functional diversity of maize. *Curr. Opin. Plant Biol.* **9,** 172–176 (2006).
34. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
35. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).

Acknowledgements Studies of natural genetic variation are supported by the National Science Foundation and the National Institutes of Health (M.N.), and by the German Research Foundation, the German Federal Ministry of Education and Research, the European Union's Sixth Framework Programme, the Human Frontier Science Program and the Max Planck Society (D.W.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to the authors (magnus@usc.edu; weigel@weigelworld.org).