in genes encoding several other molecular components of the excitatory synapse, and in particular the postsynaptic density, have been identified in other human brain diseases, including autism[14,15].

The findings of Endele *et al.*[3] also shed light on the molecular function of NMDARs at the synapse. For example, a missense mutation of *GRIN2A* causes a more severe phenotype of early onset epileptic encephalopathy as compared to the null alleles of *GRIN2A,* which are associated with generalized epilepsy and various degrees of learning difficulties. This suggests that monoallelic expression of *GRIN2A* is less detrimental than the expression of a mutated subunit.

Expected benefits
Many mutations of structural and functional components of the synapse have previously been identified[13–15.] The application of unbiased screening technologies such as massively parallel sequencing will likely further increase the pace of mutation and new gene discovery. There will be benefits from having a comprehensive catalog of human mutations of genes encoding synaptic proteins linked with detailed clinical outcomes. Precise molecular diagnosis will provide benefits for affected individuals and their families, and researchers will be driven to explore clinically and therapeutically relevant biological questions. Undoubtedly, the interplay of multiple genetic and environmental factors will complicate our investigation of the relationships between the genes, mutations and disease, but there is reason to have optimism that research into the genetic causes of intellectual disability and epilepsy will eventually translate into improved management and therapy.

1. Ropers, H.H. *Annu. Rev. Genomics Hum. Genet.* **11**, 161–187 (2010).
2. Mefford, H.C. *Genet. Med.* **11**, 836–842 (2009).
3. Endele, S. *et al. Nat. Genet.* **42**, 1021–1026 (2010).
4. Sakimura, K. *et al. Nature* **373**, 151–155 (1995).
5. Kutsuwada, T. *et al. Neuron* **16**, 333–344 (1996).
6. Gécz, J. *et al. Curr. Opin. Genet. Dev.* **16**, 308–316 (2006).
7. Tang, Y.P. *et al. Nature* **401**, 63–69 (1999).
8. Wang, D. *et al. PLoS ONE* **4**, e7486 (2009).
9. Reutlinger, C. *et al. Epilepsia* **51**, 1870–1873 (2010).
10. van Spronsen, M. & Hoogenraad, C.C. *Curr. Neurol. Neurosci. Rep.* **10**, 207–214 (2010).
11. Wu, Y. *et al. Proc. Natl. Acad. Sci. USA* **104**, 18163–18168 (2007).
12. Motazacker, M.M. *et al. Am. J. Hum. Genet.* **81**, 792–798 (2007).
13. Poot, M. *et al. Eur. J. Hum. Genet.* **18**, 39–46 (2010).
14. State, M.W. *Nat. Genet.* **42**, 478–479 (2010).
15. Laumonnier, F. *et al. Am. J. Hum. Genet.* **80**, 205–220 (2007).

Genome-wide association studies coming of age in rice

Richard M Clark

A new study reports the next-generation sequencing of 517 rice genomes, each to approximately onefold coverage. By leveraging sequence information across rice lines and by imputing missing genotypes, a haplotype map (HapMap) was constructed and used for genome-wide association studies in this major crop.

To meet the food needs of the human population, plant breeders select for agronomically important traits, such as crop yield. These traits typically vary continuously, and a focus for plant geneticists has been to identify quantitative trait loci (QTLs) that affect this variation. Finding the genes and DNA variants underlying these loci provides insights into genetic pathways on the one hand and on the other hand provides targets for marker-assisted selection to introduce favorable genetic variants into breeding programs. To date, most crop QTLs have been identified using biparental crosses or association studies with candidate genes. These efforts have yielded many successes[1] but have sampled only a small number of genetic variants likely to be of agronomic importance[2]. On page 961 of this issue, Bin Han and colleagues[3] now report the construction of a comprehensive HapMap for rice (*Oryza sativa*) that they used for genome-wide association studies (GWAS) for 14 agronomic traits. Strong associations could be identified for some traits,

Richard M. Clark is at the Department of Biology, University of Utah, Salt Lake City, Utah, USA.
e-mail: clark@biology.utah.edu

and compared to human GWAS, the percentage of variation explained in this study by these associations was generally much higher[3,4].

A rice GWAS
GWAS require both phenotypic data and genome-wide genotypes. Huang *et al.*[3] phenotyped 517 Chinese land races of rice (locally propagated, inbred lines) for 14 agronomic traits, such as heading date (flowering time) and tillering (**Fig. 1a**). The authors then used Illumina next-generation sequencing[5] to generate approximately onefold sequence coverage per rice line. Alignment of short, next-generation sequence reads to reference genome sequences has already been used extensively to detect SNPs[5,6] (**Fig. 1b**). To date, however, deep read coverage has generally been used to overcome the effect of sequencing errors that can be mistaken as SNPs when few reads cover a position. To overcome this challenge, Huang *et al.*[3] made use of the low-coverage sequence data across lines, where the combined coverage was ~508-fold, and detected 3.6 million SNPs.

On a per line basis, however, a major challenge was missing genotypes (**Fig. 1b**), which limit the statistical power of association tests.

Thus, Huang *et al.*[3] exploited the finding that linkage disequilibrium (LD) decayed over >100 kb on average in their study population (see also refs. 7,8). Because LD, or the nonrandom correlation among (typically nearby) variants, is extensive in rice, Huang *et al.*[3] could infer missing genotypes with high confidence by applying an imputation algorithm. As assessed with known sequences, missing genotypes were reduced from ~60% to less than 3% with a prediction accuracy of >98%.

With the resulting sequence data, Huang *et al.*[3] tested for genotype-phenotype associations in a subset of 373 lines of the *indica* subspecies. Applying several statistical methods, one of which effectively reduces false positive associations due to patterns of relatedness common in plant study populations (population structure and kinship)[9], clear peaks of association were observed for a subset of traits, with peaks often harboring small sets of genes. Some of these associations are almost certainly true, as several peaks included genes already shown to underlie variation in the respective traits. In other cases, association signals were less clear, and additional studies will be required to validate candidate regions and causal

variants (for example, with genetic-cross approaches). Although variable by trait, on average, strong associations explained about 36% of the phenotypic variation[3]. This is markedly greater than that observed in many human studies[4] and is more similar to findings from a GWAS for 107 phenotypes in *Arabidopsis thaliana*[10]. In part, this likely reflects the adaptive relevance, either ecological or agronomic, of the traits examined in these studies[10].

HapMaps and next-generation sequencing

For future studies in rice, as well as in other organisms, the potential of low-fold, next-generation sequencing for genome-wide discovery and genotyping of genetic variants is immense. In species including human and *A. thaliana*, HapMaps were constructed in large part with expensive microarray-based methods that suffered from substantial ascertainment biases[11–13]. By comparison, parallel and (comparatively) comprehensive SNP discovery and genotyping by sequencing and imputation is a highly attractive alternate method (although for heterozygous organisms, the challenge will be greater). This is especially true as next-generation sequencing costs are falling, and library barcoding[14], which Huang *et al.*[3] employed, can enable sequencing of more than a hundred samples per sequencing run[15].

The optimal coverage depth needed for a given study, however, is far from clear and will depend on experimental aims as well as underlying biology. When read coverage per sample is high, genetic variation will be more fully cataloged, increasing the likelihood of finding causal variants for phenotypes. Nonetheless, deep coverage will be achieved at the sacrifice of obtaining lower-fold coverage for many more samples. An obvious biological consideration that will affect genotype imputation is the extent of LD present; in species or populations where LD decays quickly, higher read coverage will be necessary for accurate imputation. It should be noted, however, that Huang *et al.*[3] successfully applied their imputation approach, with only marginal losses in performance, to regions of the rice genome where LD decays within ~10 kb.
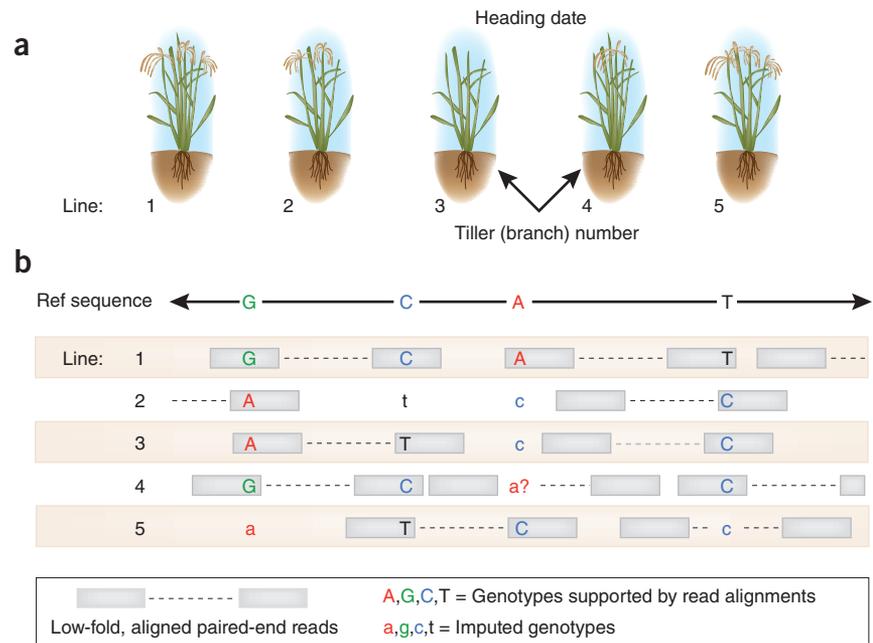

Figure 1 Constructing a rice HapMap for GWAS with low-fold sequencing. (**a**) Lines of rice vary for agronomically important traits, such as heading date (flowering time) and tiller (branch) number. (**b**) Low-fold sequencing and read alignment to the reference (Ref) rice genome sequence is both efficient and sufficient for genome-wide genotyping[3]. Aligned reads (gray boxes) enable SNP discovery in different lines (bases, upper case; consistent patterns of discrepancies in alignments distinguish genetic differences from sequencing errors). Further, imputation 'fills in' missing genotypes in areas of no coverage (bases, lower case), even though factors such as (historical) recombination will inherently lead to uncertainties in some regions (line 4, question mark). For simplicity, only nonoverlapping reads are displayed for each line. Paired-end reads[5] (gray boxes with dashed lines connecting the read pairs) were used by Huang *et al.*[3] and facilitate correct read alignment and discovery of structural variants (but are not formally required).

Against the backdrop of these considerations, a remaining question is whether low-fold sequence coverage for many samples can be used not only to detect SNPs but also more complex polymorphisms, partially overcoming the need for deeper sequence coverage[13]. For instance, can data for a common haplotype be combined, albeit against the background of other segregating sequences, for *de novo* assembly to identify all (or nearly all) genetic variation (for instance, large insertions)? The read data of Huang *et al.*[3] are ideal to begin to address this and similar questions, and answers are also likely to come from efforts of the human 1000 Genomes Project and the analogous project in *A. thaliana*[13].

1. Harjes, C.E. *et al. Science* **319**, 330–333 (2008).
2. Myles, S. *et al. Plant Cell* **21**, 2194–2202 (2009).
3. Huang, X. *et al. Nat. Genet.* **42**, 961–967 (2010).
4. Manolio, T.A. *et al. Nature* **461**, 747–753 (2009).
5. Bentley, D.R. *et al. Nature* **456**, 53–59 (2008).
6. Ossowski, S. *et al. Genome Res.* **18**, 2024–2033 (2008).
7. Mather, K.A. *et al. Genetics* **177**, 2223–2232 (2007).
8. McNally, K.L. *et al. Proc. Natl. Acad. Sci. USA* **106**, 12273–12278 (2009).
9. Yu, J. *et al. Nat. Genet.* **38**, 203–208 (2006).
10. Atwell, S. *et al. Nature* **465**, 627–631 (2010).
11. Clark, R.M. *et al. Science* **317**, 338–342 (2007).
12. Hinds, D.A. *et al. Science* **307**, 1072–1079 (2005).
13. Weigel, D. & Mott, R. *Genome Biol.* **10**, 107 (2009).
14. Cronn, R. *et al. Nucleic Acids Res.* **36**, e122 (2008).
15. Huang, X. *et al. Genome Res.* **19**, 1068–1076 (2009).