Gene dosage effects seeded, retained and amplified

Almost all of the genes within the boundaries of the large CNVs had changes in their expression levels consistent with the given change in gene dosage. For example, *GTF2I* expression levels were approximately three times as high in iPSCs derived from patients with the duplication (Williams-Beuren region duplication syndrome) than in cells derived from patients with the deletion (WBS). However, this was not the case for all genes in the region, which should make for very interesting follow-up analyses.

At the genome-wide level, there were several hundred differentially expressed genes (DEGs), suggesting a network effect emanating from the CNV at 7q11.23. Some of these expression changes were cell-type specific. Interestingly, many of the affected molecular pathways already showed signs of dysregulation in the iPSC state. This transcriptional dysregulation had a tendency to become amplified in the more differentiated cell states, such that the retained DEGs would often possess a lineage-specific function. For example, DEGs and Gene Ontology (GO) categories related to axon formation emerged more clearly in NPCs, as did those for synapses in NCSCs and for smooth muscle tissue in MSCs. Changes in GTF2I genome-wide binding patterns also indicated lineage-specific effects of the large CNV. Interestingly, overlap with the DEG patterns was limited, pointing to a more indirect effect of this transcription factor on transcription networks. The authors were able to identify a few specific target genes as potentially relevant for the disease phenotypes, namely *PDLIM1* (which acts in neurites and is associated with cardiovascular defects), *MYH14* (involved in hearing) and, in particular, *BEND4* (another transcription factor, involved in neural processes). The targeting of *BEND4* by GTF2I provides a glimpse at how the effects of the large CNV might ripple across longer distances on the molecular network.

Toward a new research paradigm

This study shows that the somewhat notorious variance between iPSC lines[14,15] is not too severe to mask clear effects of large CNVs, some of which are tissue specific and related to disease. There are some deviations between expression levels and discrete dosage changes at the DNA level that remain unexplained, but such discrepancies point the way to interesting studies to be done on the epigenetic level.

The study identified a few specific genes that can be further explored in the context of WBS and Williams-Beuren region duplication syndrome, although, before too much effort is expanded on these individual genes, it would be prudent to explore the effects of the 7q11.23 CNVs in disease-relevant terminal cell differentiation states, not just in precursor cells. In such terminal differentiation states, it will then also be possible and important to carry out functional analyses such as neurophysiological measurements in relevant neuronal subtypes.

Larger cohorts will be used once iPSC methods have sufficiently evolved, and there will have to be an accounting for the typically rather large degree of phenotypic variance between patients with nearly identical large CNVs. But the general approach is sound and already quite reliable. It will be very interesting to see what parallels and differences are shown by the studies that are well under way in laboratories around the world using the same concept for other large CNVs.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Iafrate, A.J. *et al. Nat. Genet.* **36**, 949–951 (2004).
2. Sebat, J. *et al. Science* **305**, 525–528 (2004).
3. Scherer, S.W. *et al. Nat. Genet.* **39**, S7–S15 (2007).
4. 1000 Genomes Project Consortium. *Nature* **491**, 56–65 (2012).
5. Malhotra, D. & Sebat, J. *Cell* **148**, 1223–1241 (2012).
6. Southard, A.E., Edelmann, L.J. & Gelb, B.D. *Pediatrics* **129**, 755–763 (2012).
7. Adamo, A. *et al. Nat. Genet.* **47**, 132–141 (2015).
8. Tordjman, S. *et al. Behav. Genet.* **37**, 61–78 (2007).
9. Nestler, E.J. & Hyman, S.E. *Nat. Neurosci.* **13**, 1161–1169 (2010).
10. Pérez Jurado, L.A. *et al. Am. J. Hum. Genet.* **59**, 781–792 (1996).
11. Urbán, Z. *et al. Am. J. Hum. Genet.* **59**, 958–962 (1996).
12. Osborne, L.R. *Mol. Genet. Metab.* **67**, 1–10 (1999).
13. Merla, G. *et al. Hum. Genet.* **128**, 3–26 (2010).
14. Thatava, T. *et al. Mol. Ther.* **21**, 228–239 (2013).
15. Cahan, P. & Daley, G.Q. *Nat. Rev. Mol. Cell Biol.* **14**, 357–368 (2013).

New insight into a complex plant–fungal pathogen interaction

Peter J Balint-Kurti & James B Holland

The coevolution of plants and microbes has shaped plant mechanisms that detect and repel pathogens. A newly identified plant gene confers partial resistance to a fungal pathogen not by preventing initial infection but by limiting its spread through the plant.

Plants, like other higher organisms, display a tremendous diversity of associations with microbes. At one end of the spectrum is mutualism, as observed between nitrogen-fixing rhizobacteria and their legume hosts. At the other extreme is the parasitism of obligate biotrophs, such as the rust fungi that colonize and derive nutrients from living host plant tissue, and of necrotrophs, such as the botrytis fungi that kill and rot plant tissues as a means to obtain their nutrients. Endophytic fungi grow systemically throughout a plant without causing obvious disease symptoms and in some cases appear to benefit their hosts by producing compounds that inhibit insect herbivory[1] or contribute to stress tolerance[2]. On page 151 of this issue, Mingliang Xu and colleagues identify a plant gene, *ZmWAK*, that regulates the interaction between maize and the fungus *Sporisorium reilianum*, causal agent of head smut disease[3]. *S. reilianum* infects the roots of maize seedlings and grows systemically, often causing only subtle observable effects on host morphology or physiology[4,5] until flowering, when the fungus forms spectacular large black sori filled with teliospores that replace the ears and tassels of the host (**Fig. 1**).

Peter J. Balint-Kurti is in the Plant Science Research Unit, US Department of Agriculture–Agricultural Research Service (USDA-ARS) and the Department of Plant Pathology, North Carolina State University, Raleigh, North Carolina, USA. James B. Holland is in the Plant Science Research Unit, US Department of Agriculture–Agricultural Research Service (USDA-ARS) and the Department of Crop Science, North Carolina State University, Raleigh, North Carolina, USA.
e-mail: peter_balintkurti@ncsu.edu

Plant defense responses

Plant pattern recognition receptors recognize general microbial features (pathogen-associated molecular patterns), triggering a basal defense response sufficient to prevent most pathogen infections. Pathogen-derived 'effector' proteins

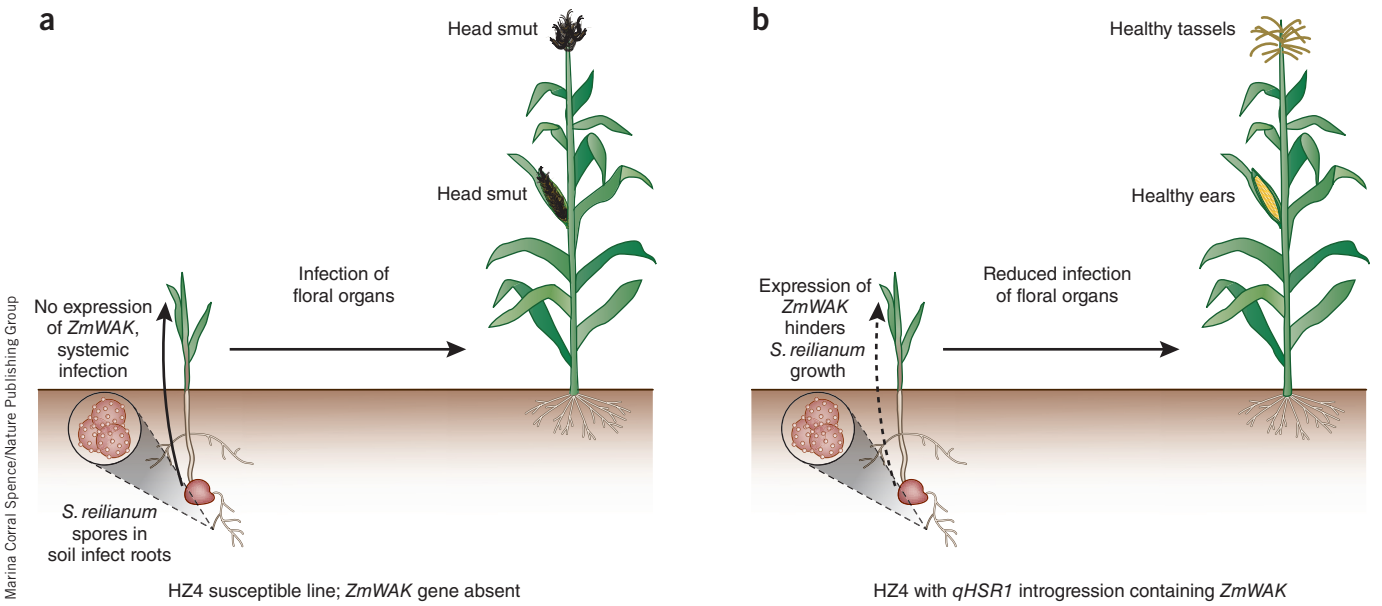

Figure 1 *ZmWAK* regulates the interaction between maize and the fungus *S. reilianum*, the causal agent of head smut disease. *S. reilianum* infects the roots of maize seedlings and grows systemically. At flowering, the fungus forms large black sori filled with teliospores that replace the ears and tassels of the host. Maize lines that lack *ZmWAK* expression are susceptible to systemic infection, whereas expression of *ZmWAK* hinders *S. reilianum* growth.

can suppress this defense response in hosts to which the pathogen is specifically adapted. As a counter-adaptation, plant resistance genes encode proteins that recognize the presence of particular effectors and initiate a localized hypersensitive response that confers immunity[6]. This type of host immunity, characterized by mendelian inheritance, discrete separation of 'resistant' and 'susceptible' plants, and specificity of resistance genes to particular races of pathogens, is relatively well understood[7].

In contrast, quantitative disease resistance (QDR) describes the continuous spectrum of resistance exhibited to some diseases, wherein resistance is conferred by the action of numerous genes, each of which has a small effect on the levels of disease observed[8]. Understanding of the genes and mechanisms underlying QDR in plants is negligible. Only recently have plant pathologists and geneticists identified several genes underlying QDR in a number of plant-pathogen systems[9]. Unlike the genes involved in immunity-type resistance, many of which encode proteins with specific shared motifs and confer resistance by similar mechanisms[7], the few genes for QDR identified thus far represent proteins of diverse function, suggesting that QDR is based on a multitude of mechanisms[8]. But the details of the mechanisms in QDR remain elusive.

Resistance genetics and mechanisms

Zuo *et al.* identify *ZmWAK* as a gene conferring QDR to head smut by resolving a previously identified quantitative trait locus (QTL) to a few genes using genetic fine mapping and then confirm the gene's effect on disease resistance using transgenic complementation. In addition to performing the impressive technical feat of cloning a gene with a quantitative effect on a complex trait, this study describes where and when during development the expression of the gene coincides with the cessation of fungal spread in resistant plants.

Typically, the QTL regions identified by linkage in maize encompass >10 Mb of DNA sequence and hundreds of genes. Higher-resolution mapping is a daunting task because of difficulties in reliably scoring quantitative phenotypes, owing to the subtle effects of QTLs and the influence of the environment (and, in this case, the pathogen) on trait expression. A QTL might also represent the combined effects of multiple genes whose individual effects are even more difficult to pinpoint when they are separated by recombination during fine mapping[10]. Finally, the substantial physical rearrangements among the genomes of different maize lines[11] further complicate the identification of genes underlying QTLs. Indeed, Zuo *et al.* found that the genomic region associated with head smut resistance corresponded to a 152-kb interval containing *ZmWAK* and four other genes in the resistant parent of their mapping population, of which 147 kb (including *ZmWAK*) was absent from the susceptible parent. To prove the causal effect of *ZmWAK* on resistance, Zuo *et al.* showed that transgenic expression of *ZmWAK* in a susceptible background conferred significant levels of head smut resistance and, conversely, that transgenic suppression of *ZmWAK* expression in a resistant background conferred increased susceptibility.

Some susceptible lines carry *ZmWAK*, indicating that the presence of the gene is not sufficient to confer resistance. Zuo *et al.* showed that the expression levels of *ZmWAK* in seedling tissues correlate with resistance to head smut. In seedlings infected with *S. reilianum*, fungal biomass levels in the lower half of the mesocotyl are indistinguishable between near-isogenic lines differing for the resistance allele. Although the fungus is able to colonize tissues higher up in the susceptible isoline, the resistant isoline has less fungus in the upper mesocotyl and almost no fungus in the coleoptiles (**Fig. 1**). Thus, in contrast to most cases of immunity-type resistance where resistance is manifested at the point of pathogen infection[7], the mechanism of resistance to head smut conferred by *ZmWAK* takes effect at a later stage when the fungus exhibits characteristics of endophytic growth in the host. This is remarkable because the plant displays few if any symptoms of disease in the susceptible isoline or of resistance response in the resistant isoline during the period when resistance is being manifested, 1–3 d after infection.

Although this work probably represents the most comprehensive characterization of a QDR-associated gene thus far, many questions remain. Prime among them is how the *ZmWAK* gene product stops the spread of the fungus. *ZmWAK* is related to the *Arabidopsis thaliana AtWAK2* gene, which is believed to perceive pectin[12,13] and regulate osmotic stress and turgor. The authors provide evidence that the ZmWAK protein is also a transducer of extracellular signals and regulates osmotic

stress through its kinase domain. But how this relates to disease resistance remains a mystery. Several other QDR-associated genes encode proteins with kinase domains[9]: do these genes confer resistance through similar mechanisms?

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Saikkonen, K., Gundel, P. & Helander, M. *J. Chem. Ecol.* **39**, 962–968 (2013).
2. Nagabhyru, P., Dinkins, R., Wood, C., Bacon, C. & Schardl, C. *BMC Plant Biol.* **13**, 127 (2013).
3. Zuo, W. *et al. Nat. Genet.* **47**, 151–157 (2015).
4. Matyac, C. *Phytopathology* **75**, 924–929 (1985).
5. Martinez, C., Jauneau, A., Roux, C., Savy, C. & Dargent, R. *Protoplasma* **213**, 83–92 (2000).
6. Jones, J.D.G. & Dangl, J.L. *Nature* **444**, 323–329 (2006).
7. Bent, A.F. & Mackey, D. *Annu. Rev. Phytopathol.* **45**, 399–436 (2007).
8. Poland, J.A., Balint-Kurti, P.J., Wisser, R.J., Pratt, R.C. & Nelson, R.J. *Trends Plant Sci.* **14**, 21–29 (2009).
9. Roux, F. *et al. Mol. Plant Pathol.* **15**, 427–432 (2014).
10. Studer, A.J. & Doebley, J.F. *Genetics* **188**, 673–681 (2011).
11. Fu, H. & Dooner, H. *Proc. Natl. Acad. Sci. USA* **99**, 9573–9578 (2002).
12. Kohorn, B.D. *et al. Plant J.* **60**, 974–982 (2009).
13. Kohorn, B.D. *et al. Plant J.* **46**, 307–316 (2006).

Big data mining yields novel insights on cancer

Peng Jiang & X Shirley Liu

Recent years have seen the rapid growth of large-scale biological data, but the effective mining and modeling of 'big data' for new biological discoveries remains a significant challenge. A new study reanalyzes expression profiles from the Gene Expression Omnibus to make novel discoveries about genes involved in DNA damage repair and genome instability in cancer.

Since the invention of gene expression microarray technology almost 20 years ago, numerous mRNA profiling data sets have been generated for diverse biological processes in many organisms. Currently, there are over 30,000 series and 1 million samples of array-based gene expression data deposited in the NCBI Gene Expression Omnibus (GEO) database. In this issue, Rudolf Fehrmann and colleagues comprehensively reanalyzed the expression profiles of 77,840 Affymetrix gene expression data sets from GEO, using principal-components analysis (PCA) to identify 'transcriptional components', which each capture a part of the variance seen in gene expression across samples[1]. Using this test set of samples, the authors developed a method for extracting biological information about the regulatory program of the samples. They then used this method to analyze expression data from 16,172 tumor samples for cancer biology discovery.

The vast amounts of biological big data—genomic, transcriptomic, proteomic and epigenomic—available through public repositories are a potential source for novel biological discoveries. To make these discoveries, however, bioinformatic tools are needed to integrate the different data types and platforms. There have been efforts to create processed public data resources for the scientific community[2–4], which require extensive investment in data collection, curation and processing. There are also studies integrating expression data sets from GEO to make new discoveries. For example, expression compendia integration identified the conditional activity of expression modules in cancer[5], expression outlier analysis predicted the frequent fusion of the *TMPRSS2* and *ETS* transcription factor genes in prostate cancer[6] and mutual information has been used to infer post-translational modulators of transcription factor activity[7]. The current study by Fehrmann *et al.* represents a fresh angle for big data integration and novel discovery[1].

Peng Jiang and X. Shirley Liu are in the Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, Massachusetts, USA.
e-mail: xsliu@jimmy.harvard.edu

Landscape of mRNA profiles

Using PCA, Fehrmann *et al.* identified principal components (PCs), which they refer to as transcriptional components, from public gene expression profiles (**Fig. 1a**). Each PC explained a portion of the total variation in gene expression across samples. Understandably, some of the PCs reflect technical artifacts, and these components can be used to remove batch effects. However, if some of the PCs contain high-coefficient genes that are known to be associated with a certain biological process, then other genes with similarly high PC coefficients might also be involved in this process. The authors used their PCA approach, combined with gene set enrichment analysis, to build a model of the regulatory network of 19,997 genes, which they used to predict the biological function of some genes within the network.

Through painstaking comparison with other methods, the authors demonstrate the superior performance of their PCA approach and make some new discoveries about the gene regulatory network. For example, they found that *FEN1* had coefficients similar to those of *BRCA1* and *BRCA2* across PCs (**Fig. 1b**). *FEN1* was previously known to be involved in DNA repair, and *BRCA1* and *BRCA2* are well known as mediators of homologous recombination in DNA damage repair. Through guilt-by-association analysis, Fehrmann and colleagues predicted that, like *BRCA1* and *BRCA2*, *FEN1* had a function in homologous recombination–mediated repair. The authors experimentally validated this prediction, showing that *FEN1* inactivation impaired homologous recombination–mediated repair in human cells.

The authors also demonstrate the use of their PCA approach to identify somatic copy number alterations (SCNAs) by locating neighboring genes on a chromosome with consistently higher or lower coefficients in one PC (**Fig. 1c**). This approach is based on the finding that coordinated aberrations in expression for nearby genes suggest the presence of SCNAs[8]. The association of PCs with SCNAs was only observed in human samples derived from cancer tissues or cell lines; non-tumor samples and samples from rodents did not show this association. On the basis of these observations, the authors developed a computational method, termed 'functional genomic mRNA' (FGM) profiling that uses non-genetic transcriptional components to correct raw expression data, and they used this method to determine the landscape of genome-wide SCNAs in cancer samples. The authors also derived a genome instability value for each sample, which was used to measure the overall degree of genome-wide SCNA (or total functional aneuploidy). In comparison to a previous study[8], Fehrmann *et al.* had improved power to detect associations with genomic instability, likely owing to