cancers, and *TP53*, which was mutated only rarely in non-dysplastic Barrett's esophagus (2.5%) but frequently in high-grade dysplasia (72%) and invasive cancer samples (69%). It should be noted that the relatively small size of the discovery cohort and the observation that additional genes varied in their mutation frequency between stages, but at a level of statistical significance that did not survive correction for multiple testing, suggest that increased sample size might lead to the identification of additional genes useful in discrimination.

Opportunities for translation

Although the presence of mutated *TP53* in high-grade dysplasia and its predictiveness of subsequent esophageal adenocarcinoma have been known for a number of years[12], a particularly exciting aspect of the research reported by Weaver *et al.*[4] relates to the application of this knowledge to a developing non-endoscopic clinical test—the Cytosponge[13]. This is a dissolvable capsule that contains a compressed mesh attached to a string; after the capsule is swallowed and allowed to dissolve in the stomach, the expanded mesh is withdrawn. The authors established the ability of the Cytosponge to collect cells in which *TP53* mutations could be reliably detected, despite the low and variable allele fractions for mutant DNA (ranging from <1% to 36%). They further demonstrated the ability of this approach to identify the presence of high-grade dysplasia with 86% sensitivity (19 mutations detected in 22 individuals with high-grade dysplastic Barrett's esophagus) and 100% specificity (0 mutations detected in 67 individuals with non-dysplastic Barrett's esophagus or with no Barrett's esophagus).

How might this approach be translated into improved clinical care? In the current paradigm, there is a substantial 'risk gap' between the large number of persons in the general population who may have recurrent symptoms of gastroesophageal reflux and/or other risk factors for esophageal adenocarcinoma and the much smaller number in whom relevant mutations and other genomic abnormalities have already occurred, marking these individuals as having significant absolute risk of developing the malignancy. Performing invasive endoscopy in (potentially) all persons with recurrent reflux symptoms, who comprise about 20% of the adult population, is an expensive proposition with very low yield. A non-endoscopic test, such as use of the Cytosponge with *TP53* mutation assays, would provide the primary care provider with an important tool to bridge this risk gap (**Fig. 1**). For example, the office-based test could be offered only to selected individuals, such as those considered to be at elevated risk on the basis of a panel of known risk factors and blood-based biomarkers, and only those with evidence of *TP53* mutations would be referred to secondary care providers. This strategy would potentially result in many fewer individuals undergoing endoscopy, but those who did would be at substantially higher risk, representing an appropriate population for more invasive tests and intensive prevention and treatment efforts. This technology has not yet matured but is getting closer. Inexpensive and reliable assays available outside of a research setting are needed, in concert with prospective studies and clinical trials. The inclusion of endpoints other than high-grade dysplasia also should be considered[5]. However, as recent simulation modeling predicts that the incidence of esophageal adenocarcinoma will continue to increase for two more decades, with a cumulative number of cause-specific deaths of approximately 160,000 in the United States alone[14], progress in this area cannot come too quickly.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Ferlay, J. *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide, http://globocan.iarc.fr/ (2013).
2. Kroep, S. *et al. Am. J. Gastroenterol.* **109**, 336–343 (2014).
3. Vogelstein, B. *et al. Science* **339**, 1546–1558 (2013).
4. Weaver, J.M.J. *et al. Nat. Genet.* **46**, 837–843 (2014).
5. Reid, B.J., Li, X., Galipeau, P.C. & Vaughan, T.L. *Nat. Rev. Cancer* **10**, 87–101 (2010).
6. Wang, K.K. & Sampliner, R.E. *Am. J. Gastroenterol.* **103**, 788–797 (2008).
7. Li, X. *et al. Cancer Prev. Res. (Phila.)* **7**, 114–127 (2014).
8. Corley, D.A. *et al. Gastroenterology* **145**, 312–319 (2013).
9. Gordon, L.G. *et al. Gastrointest. Endosc.* **79**, 242–256 (2014).
10. Dulak, A.M. *et al. Nat. Genet.* **45**, 478–486 (2013).
11. Agrawal, N. *et al. Cancer Discov.* **2**, 899–905 (2012).
12. Galipeau, P.C. *et al. PLoS Med.* **4**, e67 (2007).
13. Kadri, S.R. *et al. Br. Med. J.* **341**, c4372 (2010).
14. Kong, C.Y. *et al. Cancer Epidemiol. Biomarkers Prev.* **23**, 997–1006 (2014).

Towards sequence-based genomic selection of cattle

Michel Georges

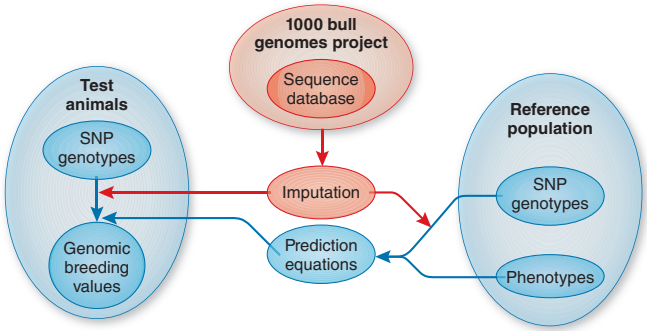
An international effort, the 1000 bull genomes project, aims to resequence the genomes of a large number of key ancestor bulls of the most important domestic cattle breeds. A new study reports on the first results of this important initiative based on the analysis of the first 234 bovine whole-genome sequences.

A reference sequence for the bovine genome was first reported in 2009 (ref. 1). Now, on page 858 of this issue, Ben Hayes and colleagues[2] describe the generation and analysis of the whole-genome sequences of 234 bulls representing four major breeds and including elite sires that have contributed a substantial proportion of the chromosomes segregating in the present population. Variant calling uncovered 26.7 million SNPs and 1.6 million insertions/deletions (indels), 79% of which are new. Concerns are often voiced that intense selection in animals and plants may lead to potentially detrimental erosion of domestic diversity. Surprisingly, the nucleotide diversity of present-day domestic cattle appears to be larger than that found in humans. It is expected that the 1000 bull genomes project will have a major impact on how we select cattle and how much genetic progress we can achieve.

Michel Georges is at the Unit of Animal Genomics, Groupe Interdisciplinaire de Génoprotéomique Appliquée–Research (GIGA-R) and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium.
e-mail: michel.georges@ulg.ac.be

From pedigree- to genetic-based selection

During the last decade, cattle breeding has undergone a once-in-a-lifetime revolution. Since 2005, more than 750,000 animals have been genotyped with genome-wide SNP arrays. This information has been used to predict



Debbie Maizels/Nature Publishing Group

Figure 1 Toward sequence-based genomic selection. In blue, the classical process of genomic selection generating prediction equations from the joint analysis of 10,000–50,000 SNP genotypes and phenotypes recorded in a large (>20,000 individuals) reference population is shown. These equations can then be used to predict genomic breeding values for test animals from their SNP genotypes alone. The red area shows that the sequence database of the 1000 bull genomes project allows for imputation of genotypes for millions of additional DNA variants for both reference and test animals to generate more robust prediction equations and genomic breeding values.

individual breeding values using a process referred to as genomic selection (GS). Before the advent of GS, identifying elite dairy sires relied on a tedious progeny-testing scheme that took 6–7 years and cost approximately $35,000 per bull tested. It is now possible to evaluate the genetic merit of a preimplantation embryo with comparable accuracy for less than $100 (provided that a reference population is available). The basic principles of GS were laid out in a landmark publication by Meuwissen *et al.*[3]. GS was adopted very rapidly in dairy cattle and has been gaining traction in beef cattle, pig, poultry and plant breeding. The efficiency of GS supports the quasi-infinitesimal architecture of the majority of agronomically important quantitative traits, which has illuminated discussions about the genetic architecture of common complex diseases in human[4]. The variance component and Bayesian methods used for GS are receiving increasing attention in human genetics and may pave the way toward innovative diagnostic approaches[5].

Daetwyler *et al.*[2] devised a clever scheme to enhance the value of this gigantic GS genotype database. By providing sequence data corresponding to at least eightfold genome coverage of a minimum of 25 animals, contributors gain access to the variant calls of the complete 1000 bulls genome project data set. This data allows them to impute sequence information on their SNP-genotyped populations, enabling them to perform sequence-based genome-wide association studies (GWAS) and GS (**Fig. 1**).

From associated to causative variants

In a series of four analyses, Daetwyler *et al.*[2] illustrate the utility of the 1000 bull genomes project resource to accelerate the identification of causative variants. The first analysis reports the identification of a likely embryonic-lethal mutation compromising fertility. Previous analyses of SNP genotypes have identified a number of haplotypes for which homozygous individuals are never found in the population, including the Holstein-Friesian (HF) HH3 haplotype on chromosome BTA8 (ref. 6). One animal in the 1000 bull genomes project database was known to carry the HH3 allele. By filtering for variants (i) for which this bull was heterozygous, (ii) that were absent in animals not carrying HH3, (iii) never found in the homozygote state in HF and (iv) absent in breeds other than HF, the authors identified a single, highly damaging phenylalanine-to-serine substitution in the structural maintenance of chromosome protein 2 (SMC2) that is very likely to cause the death of homozygous embryos.

The second analysis identified a mutation causing a dominant form of lethal chondrodysplasia. 'Bulldog' calves have been reported previously in the offspring of Igale, a widely used HF sire. Prior analyses have indicated that Igale was most likely germline mosaic for a dominant *de novo* mutation mapping to BTA5. Daetwyler *et al.*[2] sequenced two affected animals and searched for variants for which both calves were heterozygous but were absent in the 1000 bull genomes project database. They identified a glycine-to-arginine substitution in the α-1 chain of type II collagen (COL2A1), for which Igale was germline mosaic, thereby demonstrating its causality.

The third study identified a mutation that most likely causes the curly coat phenotype that is associated with higher tick and parasite burden. The authors imputed sequence variants from the 1000 bull genomes project database onto a population of 3,222 Fleckvieh (F) sires. They then conducted a sequence-based GWAS using the proportion of curly daughters as the phenotype, which identified two significant associations coinciding with clusters of type I and type II keratins on chromosomes BTA5 and 19, respectively. A potentially damaging lysine-to-asparagine substitution in keratin 27 (KRT27) (*Krt27* causes the wavy coat phenotype when knocked out in mice) was among a set of five most strongly associated SNPs. Analysis of the same five variants in the related Montbeliarde breed indicated that only the p.Asn92Lys variant was associated with curly coat in this breed, thereby strongly supporting its causality.

In the fourth example, Daetwyler *et al.*[2] performed a similar sequence-based GWAS for milk fat content using daughter averages of 3,513 F and 2,327 HF bulls. The top association signal in both breeds corresponded to the previously reported gene *Dgat1* (encoding diacylglycerol O-acyltransferase 1). The second signal shared by both breeds coincided with the location of the gene *Agpat6* (encoding 1-acylglycerol-3-phosphate O-acyltransferase 6), which is known to be involved in lipid biosynthesis and is therefore a strong candidate. The authors pinpointed a cluster of four *Agpat6* promoter variants that are among the most strongly associated in both breeds as plausible causative regulatory variants.

From marker- to sequence-based genomic selection

Beyond these successful illustrations of its immediate utility, the primary motivation of the 1000 bull genomes project was the establishment of a collaborative approach to cost-effectively increase the accuracy and robustness of GS. The accuracy of GS is known to rapidly decrease with increasing genetic distance between the reference population and the animal tested. By performing sequence-based GS, causative variants will be part of the data set, thus potentially increasing accuracy. The jury is still out regarding the actual effectiveness of this approach, which will depend in part on the accuracy of imputation. Daetwyler *et al.*[2] show by cross-validation that with the 234 sequenced animals, the imputation accuracy averages ~80%

for variants with minor allele frequency (MAF) >0.15 yet drops rapidly for those with lower MAF or in regions with poor SNP coverage. This accuracy is unlikely sufficient to improve GS. Moreover, local drops in imputation accuracy uncovered numerous assembly errors, highlighting the need for further improvement of the bovine reference genome. Using *Dgat1* as an example, the authors show that the established causative p.Ala232Lys variant[7] is well imputed in HF (where it is common) but not in F (where it is rare), concomitantly affecting the strength of association with milk fat content in these breeds; however, the authors also show that this accuracy can be markedly improved by increasing the number of sequenced individuals.

What is next?

With sequencing costs continuing to decline, the number of publically funded genomic research projects involving bovine resequencing increasing and the number of organizations investing in GS growing, the number of sequenced animals in the 1000 bull genomes project database will increase rapidly. At the time of this writing, this number was already in excess of 1,100, and the sequences of at least 500 other animals are in the pipeline. Open access to publicly funded sequences will ensure that more effective sequence-based GWAS and GS become the norm in livestock genomics, attributable largely to the 1000 bull genomes project.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Gibbs, R.A. *et al. Science* **324**, 528–532 (2009).
2. Daetwyler, H.D. *et al. Nat. Genet.* **46**, 858–865 (2014).
3. Meuwissen, T.H., Hayes, B.J. & Goddard, M.E. *Genetics* **157**, 1819–1829 (2001).
4. Manolio, T.A. *et al. Nature* **461**, 747–753 (2009).
5. Yang, J. *et al. Nat. Genet.* **46**, 100–106 (2014).
6. VanRaden, P.M. *et al. J. Dairy Sci.* **94**, 6153–6161 (2011).
7. Grisart, B. *et al. Genome Res.* **12**, 222–231 (2002).

Fingerprints of Epstein-Barr virus in nasopharyngeal carcinoma

Robert B West

The influence of Epstein-Barr virus (EBV) on cancer is not well understood. High-throughput sequencing of nasopharyngeal carcinoma (NPC) illustrates the influences of EBV on oncogenesis and identifies driver pathways that might be therapeutically useful for NPC treatment.

More than 90% of the world population has been infected with EBV. Its presence has been documented in a number of malignancies, including in NPC[1]. Its true influence on cancer incidence is not well understood, owing in part to ignorance of the inciting events in the vast majority of cancer types. In a new study by De-Chen Lin and colleagues[2], the authors present the first genome-wide sequence analysis of NPC, the only common carcinoma that is frequently infected with EBV (although a subset of gastric carcinomas are also EBV infected). NPC represents an unusual carcinoma with respect to its incidence, biology and treatment, and its study can possibly provide insight into the relationship between neoplasia and this ubiquitous virus. NPC incidence is high in endemic areas[3] (occurring in Southeast Asians, North Africans and Inuit people), comparable to the incidence of melanoma in the United States, and very low in other areas (<1 case per 100,000 people). Another unusual aspect of NPC is a tremendous lymphocytic infiltrate that often obscures the presence of malignant cells in bright-field microscopy. These infiltrating lymphocytes, often outnumbering the cancer cells by several fold, have posed a challenge to genomic studies of NPC, particularly hybridization microarray approaches.

Robert B. West is in the Department of Pathology, Stanford University, Stanford, California, USA.
e-mail: rbwest@stanford.edu

Fingerprints of EBV

Lin and colleagues present a genome-wide view of the changes in NPC by examining 128 cases by exome sequencing, targeted sequencing and SNP array analysis. Their results provide a framework in which to begin subdividing NPCs according to their biological drivers and clinical implications. In a comparative study using several previously sequenced cancers, they find that NPC results in a relatively low level of genomic alteration, similar to that seen in human papillomavirus (HPV)-associated carcinomas. However, NPCs do not have the characteristic APOBEC mutational signature seen in HPV-associated carcinomas. These comparisons provide insight into the fundamental differences in the ways that EBV and HPV lead to cancer.

The study identifies 144 genes recurrently mutated in the 128 cases. Although some mutations had previously been found in NPC, such as ones in *PIK3CA* and *TP53*, seven significantly mutated genes are newly identified in NPC: *BAP1*, *ERBB2*, *ERBB3*, *KRAS*, *NRAS*, *KMT2D* (*MLL2*) and *TSHZ3*. Pathway analysis of genomic events identified the chromatin-modification pathway as among the most frequently affected pathways, with *ARID1A* being the most frequently altered gene in this category. Mutations in this pathway are strongly associated with EBV levels. In this context, it is interesting to note that a study with whole-exome sequencing of gastric cancer[4], the only other common carcinoma with frequent EBV involvement, identified a similar association. The sequencing data for gastric cancer showed a significant difference in the mutation rate of *ARID1A* between EBV-infected gastric cancers (high) and non-EBV-infected gastric cancers (low). Taken together, these findings suggest that altered chromatin remodeling is a general, possibly oncogenic feature of EBV infection in neoplasia. Further work is needed to extend these observations and determine the mechanisms involved.

Other pathways commonly mutated in carcinomas were also found to be altered in NPC (**Fig. 1**). Mutations in genes controlling the G1/S transition were frequently mutated in NPC. Genes in the ERBB–phosphoinositol 3-kinase (PI3K) signaling pathway were also often altered, and cases bearing these mutations had tumors with more aggressive clinical behavior. As in other head and neck carcinomas,