scaling up of biology, the lagging of a well-funded European central database infrastructure undermines the core of European medical and biological research: the easy, continued access to a rising tide of high-density data. Without flourishing, well-accessible resources, which are being actively codeveloped in parallel to other regions in the world, we will not gain the required momentum in turning data into insights. And it is the insights that will be fuelling the engine of European biotech innovation■

Gert-Jan B van Ommen is at the Center for Medical Systems Biology and the Center of Human and Clinical Genetics, Leiden University Medical Center, PO Box 9503, 2300 RA Leiden, The Netherlands.
E-mail: gjvo@lumc.nl

References

1 International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**: 931–945.
2 Eichler EE, Clark RA, She X: An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet* 2004; **5**: 345–354.
3 She X, Jiang Z, Clark RA *et al*: Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 2004; **431**: 927–930.
4 Johnson ME, Viggiano L, Bailey JA *et al*: Positive selection of a gene family during the emergence of humans and African apes. *Nature* 2001; **413**: 514–519.
5 Menashe I, Man O, Lancet D, Gilad Y: Different noses for different people. *Nat Genet* 2003; **34**: 143–144.
6 Sebat J, Lakshmi B, Troge J *et al*: Large-scale copy number polymorphism in the human genome. *Science* 2004; **305**: 525–528.
7 Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ: Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004; **36**: 861–866.
8 Iafrate AJ, Feuk L, Rivera MN *et al*: Detection of large-scale variation in the human genome. *Nat Genet* 2004; **36**: 949–951.
9 Schmitz J, Martin J, Terry A *et al*: The DNA sequence and comparative analysis of human chromosome 5. *Nature* 2004; **431**: 268–274.

Evolutionary Genetics

Genetics of lactase persistence – fresh lessons in the history of milk drinking

Edward Hollox

European Journal of Human Genetics (2005) **13,** 267–269.
doi:10.1038/sj.ejhg.5201297
Published online 15 December 2004

Most people cannot drink milk as adults without the symptoms of lactose intolerance, and most lactose intolerance is due to absence of the lactase enzyme in the gut. This presence/absence is a genetic polymorphism commonly called lactase persistence/nonpersistence, depending on whether or not lactase activity persists from childhood into adulthood.[1] In Northern Europe, lactase persistence is common and many people not only drink milk, but culturally it is seen as a healthy and nutritious food. How this happened is now becoming clearer.

Lactase nonpersistence is the ancestral state, and lactase persistence only became advantageous after the invention of agriculture, when milk from domesticated animals became available for adults to drink. As expected, lactase persistence is strongly correlated with the dairying history of the population. This genetic ability to digest milk has been regarded as a classic example of gene-culture coevolution, where the culture of dairying creates a strong selective advantage to those who can drink milk as adults, for only they can nutritionally benefit from the milk. A recent paper confirmed this link by analysing the diversity in bovine milk protein genes and showing that the highest gene diversity (and by implication the largest historical population size) is in cows from areas of the world where dairy farming is practised and the people are lactose tolerant.[2] In humans, epidemiological analysis has shown that the cultural development of dairying preceded selection for lactase persistence.[3] Since dairying is thought to have originated around 10 000 years ago, the selective pressure has been only for the past 400 generations. Despite this short time, there is suggestive evidence of recent positive selection: lactase persistence is associated with one haplotype, which is very common only in northern Europeans, and is distant from the ancestral haplotype.[4,5] Discovery of the possible molecular basis of this polymorphism – a single nucleotide change 14 kb away from the gene, has allowed further analysis of genetic variation associated with lactase persistence/nonpersistence.[6–8]

Proving that the lactase gene has been under recent positive selection in Northern Europe is difficult. As it is a recent regulatory change, codon-based methods that examine the different substitution patterns across a gene are not suitable. Instead, methods relying on allele frequency must be used – which are vulnerable to the fact that frequency patterns produced by selection can also be produced by demographic processes such as changes in population size and genetic drift. A statistic called 'relative extended haplotype homozygosity' (REHH) has been developed, which relies on the fact that a selected haplotype (ie a haplotype

on which a relatively recent beneficial mutation has occurred and has risen to high frequency) will have an extended range of linkage disequilibrium (LD) compared with other haplotypes in the population.[9] This is because the selected haplotype is young, and hence there has not been enough time for recombination to break it down. We infer that this young haplotype has been driven to a high frequency by positive selection. It is not an ideal method: since it relies on the length of linkage disequilibrium on one haplotype in relation to the frequency of that haplotype, it may be vulnerable to different sampling strategies that could alter the apparent frequency of that haplotype.[10] Allele-specific recombination rates could also produce a similar effect. Nevertheless, since it compares variation on different haplotypes across the same region, it is less vulnerable to demographic changes than other population genetic measures.

REHH was used by Joel Hirschhorn's lab to provide further support for positive selection in Northern Europeans.[8] It confirms that the haplotype carrying lactase persistence is almost identical for nearly 1 Mb, is therefore young and must have been positively selected to reach the observed frequency of 77% in Northern Europeans. Analysis of markers across this region showed very high genetic differentiation between European Americans (dairying) and Asian/African American (nondairying), suggesting that these markers had hitchhiked on the haplotype carrying lactase persistence. By considering the Asian Americans and African Americans to have a diversity representative of a pre-dairying 'European' population, a selection coefficient of 1.4–15% was calculated – consistent with the 5% previously predicted using a gene-culture co-evolutionary model.[11] Did early farmers, who practised mixed farming, really rely on milk so much? There is now genetic evidence that they did, although it is still not clear why milk was so important (for discussion, see Hollox and Swallow[12]).

Most studies for practical reasons have focused on lactase persistence in Europe, but lactase persistence is also common in certain tribes in Africa that have a history of dairying. Is lactase persistence in these people caused by the same mutation – as would seem likely – and has it been under positive selection as well? The first part of this question has been answered by Mulcare *et al*.[13] Their paper shows that the putative causative allele 14 kb upstream from the lactase gene is not at frequencies high enough for it to be the causative allele in Africa, even when the inherent errors in lactose tolerance testing are taken into account. There could be two reasons for this – either the allele is not causative at all and is merely strongly associated with the causative allele, or in Africans lactase persistence is due to another mutation. The first reason is possible, especially given the high LD across the region – many polymorphisms within this region will be strongly associated with lactase persistence just by virtue of being on the same huge haplotype. But functional studies from two groups show that the putative causative allele is a gain-of-function mutation increasing the expression driven from the lactase promoter in reporter gene assays in a human intestinal cell line.[14,15] So what about the second reason – a different causative mutation in Africans? Intuitively, this seems unlikely, but given the powerful selective advantage of being lactase persistent any mutation is very unlikely to be lost by genetic drift. It is possible that another mutation in the same regulatory element, a different element, or even in a *trans*-acting transcription factor may be responsible for lactase persistence in Africans. The answer will only be found by further genetic analysis of this locus in Africans.

As well as examining the role of this polymorphism in human evolution, this work provides an interesting case study for those concerned with finding alleles that confer susceptibility to common disease. In this case, we have a clear clinical phenotype (lactose tolerance) with a very strong well-defined Mendelian genetic component (lactase persistence/nonpersistence polymorphism), and a well-defined 'candidate' gene (LCT, lactase). Despite these factors, the causative polymorphism has proved difficult to discover, and the most likely causative polymorphism is located 14 kb away in an L2 repeat within an intron of another gene. Added to this, if this polymorphism is causative, then it is not the causative polymorphism in all populations. If there is a lesson to be learned from this, it is that the genetics of complex disease are likely to be very complex indeed■

Dr Edward Hollox is at the Institute of Genetics, University of Nottingham, Queen's Medical Centre, Nottingham, UK. Tel: +44 115 922 7815; Fax: +44 115 970 9906; E-mail: ed.hollox@nottingham.ac.uk

References

1 Swallow DM: Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet* 2003; **37**: 197–219.
2 Beja-Pereira A *et al*: Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nat Genet* 2003; **35**: 311–313.
3 Holden C, Mace R: Phylogenetic analysis of the evolution of lactose digestion in adults. *Hum Biol* 1997; **69**: 605–628.
4 Harvey CB *et al*: Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. *Ann Hum Genet* 1998; **62** (Part 3): 215–223.
5 Hollox EJ *et al*: Lactase haplotype diversity in the Old World. *Am J Hum Genet* 2001; **68**: 160–172.
6 Enattah NS *et al*: Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 2002; **30**: 233–237.
7 Poulter M *et al*: The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Ann Hum Genet* 2003; **67**: 298–311.
8 Bersaglieri T *et al*: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004; **74**: 1111–1120.
9 Sabeti PC *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
10 Brookfield JF: Human prehistory: the message from linkage disequilibrium. *Curr Biol* 2003; **13**: R86–R87.
11 Aoki K: A stochastic model of gene-culture coevolution suggested by the 'culture historical hypothesis' for the evolution of adult lactose absorption in humans. *Proc Natl Acad Sci USA* 1986; **83**: 2929–2933.
12 Hollox EJ, Swallow DM: Lactase deficiency – biological and medical aspects of the adult human lactase polymorphism; in King RA, Rotter JI, Motulsky AG (eds): *The Genetic Basis of Common Diseases*. Oxford: Oxford University Press, 2002.
13 Mulcare CA *et al*: The T allele of a single-nucleotide polymorphism 13.9 kb

upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet* 2004; **74**: 1102–1110.

14 Olds LC, Sibley E: Lactase persistence DNA variant enhances lactase promoter activity *in vitro*: functional role as a *cis* regulatory element. *Hum Mol Genet* 2003; **12**: 2333–2340.

15 Troelsen JT *et al*: An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology* 2003; **125**: 1686–1694.

Statistical Genetics

Usual suspects in complex disease

CP Ponting and L Goodstadt

European Journal of Human Genetics (2005) **13,** 269–270.
doi:10.1038/sj.ejhg.5201354
Published online 15 December 2004

Surprising observations are made by a recent study scrutinising the aetiology of complex diseases and the genetic mutations that they spring from.

Finding the common DNA variants that contribute significantly to genetic risk for common diseases is a key goal for medical science.[1] However, while distinguishing mutations that cause disease from harmless single nucleotide polymorphisms (SNPs) is difficult enough for the relatively few genetic diseases that are inherited in a simple Mendelian manner, it is a daunting task for complex traits where pathology arises from the interactions of multiple genes with the environment. Point mutations exert a broad spectrum of effects on human health. The most fearsome are those that disrupt development: they cause embryonic morbidity and are seldom observed in postnatal disease; at the other end of the scale are base substitutions under few constraints, such as most common SNPs, and in between fall most of the 1000 or so[2,3] deleterious mutations that are carried by the average person. In a recent article in *Proc. Natl. Acad. Sci. USA*, Thomas and Kejariwal[4] show what types of coding mutations we should expect in complex diseases. They find that these amino-acid changes mostly fall outside of conserved regions and cannot readily be distinguished from the coding sequence variation seen between healthy individuals.

These results are surprising because most amino-acid substitutions associated with Mendelian diseases are of conserved, and thus presumably essential, amino acids.[2,3,5,6] Hitherto, there has been little evidence that this would be any different for complex diseases. Thomas and Kejariwal give three possible explanations for their findings.

First, they suggest that coding mutations in complex disease might cause subtle and almost imperceptible alterations to molecular function. If this were true, researchers proposing molecular dysfunction in complex diseases would not be able to use sequence evolutionary information to prosecute their case. Mutations that are mildly deleterious are difficult enough to substantiate experimentally for Mendelian diseases; for complex traits that are multifactorial, it may well be impossible to detect the knockon effects of such subtle alterations.

Thomas and Kejariwal also entertain the possibility that some of the 37 cases of coding SNPs they examined might not directly contribute to complex diseases. They felt that this explanation was unlikely because their findings remain significant even for a reduced number of cases, those about which they were most confident. Nevertheless, it cannot be discounted that many of these coding SNPs, instead of directly contributing to the complex disease, may merely be closely linked to the real disease-causing polymorphisms. These may lie, for example, in adjacent noncoding sequence that regulates transcription or translation. This suggestion should find favour among those advocating hunts for causative SNPs within regulatory regions, for example, King and Wilson[7] and Prokunina et al.[8]

The authors' final explanation is that lack of conservation may not rule out the functional importance of their disease-associated SNPs if these functions have been acquired only recently in primate evolution. Comparisons with more distantly related mammals might not show conservation if SNP sites have been evolving rapidly under adaptive pressures in our lineage. Adaptive evolution can be detected if K_A/K_S ratios[9] between mouse and human genes are greater than one. However, though the K_A/K_S ratios for complex disease genes are elevated relative to randomly selected genes, the ratios are still much less than one. In any case, the issue here is whether, out of all the many codons in a gene, a particular complex disease-associated SNP has been changing adaptively. This cannot be determined simply by comparing the entirety of one gene from one species with that from another. At best, the jury is still out on this question.

A re-examination of the same K_A/K_S data, however, indicates a functional bias in complex disease genes that was not noted in the original article. Secreted proteins and transmembrane molecules, such as receptors, are greatly over-represented among those encoded by the 32 complex disease genes that Thomas and Kejariwal analysed. These number 8 (25%) and 15 (47%) out of 32, respectively, far more than just the six of each type expected from their frequency ($\sim$20%) in the human genome.[10] Transmembrane and secreted proteins evolve more rapidly than average.[11] This alone would explain much of the elevation in K_A/K_S ratios seen for complex disease genes. Transmembrane and secreted proteins would be prime suspects in complex disease as they tend to have restricted expression profiles: the dysfunction of disease genes commonly afflicts few organs or tissues, rather than being systemic.[12]

Sites mutated in complex disease, according to Thomas and Kejariwal, are distinguished by their very ordinariness, being neither essential over long evolutionary time periods, nor different from