European Journal of Human Genetics (2009) 17, 147–150

www.nature.com/ejhg

LETTERS

On Jim Watson's *APOE* status: genetic information is hard to hide

European Journal of Human Genetics (2009) **17**, 147–149; doi:10.1038/ejhg.2008.198; published online 22 October 2008

The recent publication and release to public databases of Dr James Watson's sequenced genome,[1] with the exception of all gene information about apolipoprotein E (ApoE), provides a pertinent example of the challenges concerning privacy and the complexities of informed consent in the era of personalized genomics.[2] Dr Watson requested that his ApoE gene (*APOE*) information be redacted, citing concerns about the association that has been shown with late onset Alzheimer's disease (LOAD), which is currently incurable and claimed one of his grandmothers.[3]

In this letter, without any 'analysis' of Dr Watson's genome, and thus respecting Dr Watson's wishes for *APOE* risk status anonymity, we highlight the challenges concerning the privacy and the complexities of informed consent by pointing out that the deletion of the *APOE* gene information only may not prevent accurate prediction of Dr Watson's risk for LOAD conveyed by *APOE* risk alleles. Specifically, linkage disequilibrium (LD) between one or multiple polymorphisms and *APOE* can be used to predict *APOE* status using advanced computational tools. Therefore, simply blanking out genotypes at known risk factors is generally not sufficient if the aim is to hide genetic information at these loci.

The major *APOE* risk for LOAD is generally assumed to come from the $\varepsilon_2/\varepsilon_3/\varepsilon_4$ haplotype system, with the ε_4 allele increasing risk for the disorder and the ε_2 allele being protective.[4] The $\varepsilon_2/\varepsilon_3/\varepsilon_4$ haplotype system is defined by two nonsynonymous single nucleotide polymorphisms (SNPs) in *APOE* exon 4. One is a C/T SNP (rs429358) that encodes either arginine (C) or cysteine (T) in the ApoE at amino acid 112. The second site defining this haplotype system is a C/T SNP (rs7412), which again encodes arginine (C) or cysteine (T) at ApoE amino acid 158. The allelic compositions of the commonly investigated rs429358-rs7412 haplotypes are T-T for ε_2, T-C for ε_3, and C-C for ε_4. The effects of these coding variants on ApoE function are well defined.[5] A recent meta-analysis of LOAD risk in Caucasians (clinic/autopsy cohorts) indicated odds ratios (OR) of 15.6 (95% CI, 10.9–22.5) and 4.3 (95% CI, 3.3–5.5) for *APOE* ε_4 homozygotes and $\varepsilon_4/\varepsilon_3$ heterozygotes respectively, compared to ε_3 homozygotes.[6] The meta-analytic odds ratios in population-based Caucasian samples were 11.8 (95% CI, 7.0–19.8) and 2.8 (95% CI, 2.3–3.5), respectively.[6] In a large Rotterdam (Netherlands), population-based prospective study of people aged 55 years or above, it was estimated that 17% of the overall risk of AD could be attributed to the ε_4 allele, with 3% (95% CI, 0–6%) of cases attributed to the $\varepsilon_4/\varepsilon_4$ genotype, and 14% (95% CI, 7–21%) to the $\varepsilon_4/\varepsilon_3$ genotype.[7]

A recent investigation of LD for 50 SNPs in and surrounding *APOE* in 550 Caucasians identified multiple SNPs in the *TOMM40* gene ~15 kb upstream of *APOE,* and at least one SNP in the other surrounding genes *LU*, *PVRL2*, *APOC1*, *APOC4* and *CLPTM1* were associated with LOAD risk.[8] In particular, the C allele of SNP rs157581 in *TOMM40* is in strong LD ($r^2>0.6$) with the C allele of rs429358 in *APOE*, which defines the ε_4 allele. For an additive (allelic) logit model, the OR for the presence of ε_4 *versus* the status of LOAD was estimated to be 4.1, whereas the OR for LOAD status using the alleles of rs157581 was 2.9.[8] Furthermore, using data sets such as those of Yu *et al*[8] and SNPs identified in the surrounding regions of *APOE* in Dr Watson's sequence, haplotype phasing software could be utilized to easily and accurately predict Dr Watson's *APOE* risk haplotype status.

In addition, even if genotypes for non-*APOE* SNPs conveying LOAD risk are not listed in Dr Watson's sequence (ie, because of low sequence coverage), as in the case of *TOMM40* SNP rs157581, it would be straightforward to predict Dr Watson's *APOE* risk status by exclusively using publicly available data, such as HapMap data. Specifically, although the LOAD high-risk *APOE* SNPs rs429358 and rs7412 and *TOMM40* SNP rs157581 are not in the HapMap, a recent genome-wide association screen using 502 627 SNPs performed in 1086 histopathologically verified LOAD cases ($n=664$) and controls ($n=442$), identified HapMap SNP rs4420638, located in the *APOC1* gene 14 kb downstream of the *APOE* ε_4 allele, which has a powerful association with LOAD.[9] Indeed, the association between LOAD and the G allele of rs4420638 ($P=1\times10^{-39}$) is similar to the association with the *APOE* ε_4 allele (rs429358 C allele) itself ($P=1\times10^{-44}$), with additive allelic ORs of approximately 4 and 5, respectively.[9,10] Coon *et al*[9] report strong LD between rs4420638 and rs429358 at $D'=0.86$, which implies an r^2 of approximately 0.60 based on Caucasian allele frequency estimates for these SNPs listed in dbSNP.

We note that Dr Watson received genetic counseling and after being made aware of the privacy risks associated with public data broadcast, Dr Watson decided to share his personal genome by releasing it into a publicly accessible

scientific database (for full details concerning Dr Watson and *Protection of human subjects*, *Returning research results to research participants*, and *Data release and data flow,* see Box 1 of Wheeler *et al*[1]). Nevertheless, during the preparation of this Letter, we contacted Dr Watson and colleagues in December 2007 and February 2008 informing them of the possibility of inferring his risk for LOAD conveyed by *APOE* risk alleles using surrounding SNP data. As a consequence, the online James Watson Genome Browser (JWGB) has nominally removed all data from the 2-Mb region surrounding *APOE*.

To demonstrate our point that genetic information is hard to hide, without contravening Dr Watson's wishes for *APOE* risk status anonymity (see Box 1 of Wheeler *et al*[1]), we utilized SNP genotypes identified in Dr J Craig Venter's genome sequence.[11] Furthermore, Dr Venter's sequence data reports that he is heterozygote for both the LOAD high-risk *APOE* SNP rs429358 (T/C) and *APOC1* SNP rs4420638 (A/G). Briefly, genotype imputation was performed using the MACH (version 1.0.16) computer program,[12] HapMap (CEU)-phased haplotype data (encompassing 144 SNPs) and Dr Venter's genotypes listed for the 200-kb region surrounding rs4420638 (encompassing all 144 HapMap SNPs). Following the two-step approach outlined in the MACH online tutorial and after excluding Dr Venter's genotype data for rs4420638 and all *APOE* SNPs, we were able to correctly impute Dr Venter's rs4420638 genotype as A/G. The posterior probabilities for Dr Venter's rs4420638 genotype being A/A, A/G or G/G were estimated to be 0.008, 0.992 and 0.000, respectively. The high accuracy of Dr Venter's imputed rs4420638 genotype exemplifies the utility of imputing *APOE* genetic risk for LOAD.

Finally, although the deletion of 2 Mb is likely excessive for the surrounding *APOE* region (based on reported LD), as more detailed characterization of the human genome comes to light, it will become even more necessary to redact substantial regions surrounding identified genetic risk variants to avoid the indirect, though accurate, estimation of genetic risk such as those we detail above. For example, in a recent study, using gene expression profiling of Epstein–Barr virus-transformed lymphoblastoid cell lines of all 270 individuals genotyped in the HapMap Consortium, Stranger *et al*[13] reported many instances of the most significant SNP associated with gene expression being located often 100 s of kb and up to 1 Mb outside of the gene transcript, with additional, less significant SNPs, although still useful in estimating risk, being located even further from the gene. Moreover, the potential for indirect estimation of risk will further increase as additional and more detailed genome-wide association studies are performed (which identify new risk loci) and individual human genomes are sequenced.

In summary, hiding genetic information in an otherwise fully disclosed genome sequence is not straightforward because of the availability of genomic data in the public domain that can be used to predict the missing data. We believe the potential for such indirect estimation of genetic risk has considerable relevance to concerns about privacy, confidentiality, discriminatory and defamatory use of genetic data, and the complexities of informed consent for both research participants and their close genetic relatives in the era of personalized genomics.

Acknowledgements

This study was supported by Australian NHMRC Grants 389892, 339462 and 442915 and Australian Research Council Grant DP0770096.

Conflict of interest

None declared.

Web Resources

The URL for data presented here are as follows:
James Watson Genome Browser (JWGB), http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/
James Watson Genome Browser (JWGB); local copy installation download, ftp://jimwatsonsequence.cshl.edu/jimwatsonsequence/gbrowse/
Dr J Craig Venter's genome sequence, http://huref.jcvi.org/
MACH (version 1.0.16) computer program, http://www.sph.umich.edu/csg/abecasis/MACH
HapMap (CEU) phased haplotype data (encompassing 144 SNPs), http://www.hapmap.org/cgi-perl/gbrowse/hapmap_B35/
Dr Venter's genotypes (downloaded on June 19, 2008), ftp://ftp.jcvi.org/pub/data/huref/HuRef.InternalHuRef-NCBI.gff
MACH online tutorial, http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html

Dale R Nyholt*[,1], Chang-En Yu[2] and Peter M Visscher[1]
[1]*Genetic Epidemiology and Queensland Statistical Genetics Laboratories, Queensland Institute of Medical Research, Brisbane, QLD, Australia;*
[2]*Division of Gerontology and Geriatric Medicine, Department of Medicine, Geriatric Research, Education, and Clinical Center, Veteran Affairs Puget Sound Health Care System, University of Washington School of Medicine, Seattle, WA, USA*

**Correspondence: Dr DR Nyholt, Genetic Epidemiology and Queensland Statistical Genetics Laboratories, Queensland Institute of Medical Research, Brisbane, Queensland QLD 4006, Australia. Tel: 61 7 3362 0258; Fax: 61 7 3362 0101; E-mail: daleN@qimr.edu.au*

References

1 Wheeler DA, Srinivasan M, Egholm M *et al*: The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008; **452**: 872–876.
2 McGuire AL, Caulfield T, Cho MK: Research ethics and the challenge of whole-genome sequencing. *Nat Rev Genet* 2008; **9**: 152–156.
3 Check E: James Watson's genome sequenced – discoverer of the double helix blazes trail for personal genomics. *Nature*

News 2008. doi:10.1038/news070528-10: http://www.nature.com/news/2007/070528/full/news070528-10.html
4 Farrer LA, Cupples LA, Haines JL *et al*: Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 1997; **278**: 1349–1356.
5 Raber J, Huang Y, Ashford JW: ApoE genotype accounts for the vast majority of AD risk and AD pathology. *Neurobiol Aging* 2004; **25**: 641–650.
6 Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE: Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet* 2007; **39**: 17–23.
7 Slooter AJ, Cruts M, Kalmjin S *et al*: Risk estimates of dementia by apolipoprotein E genotypes from a population-based incidence study: the Rotterdam Study. *Ann Neurol* 1998; **55**: 964–968.
8 Yu CE, Seltman H, Peskind ER *et al*: Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. *Genomics* 2007; **89**: 655–665.
9 Coon KD, Myers AJ, Craig DW *et al*: A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry* 2007; **68**: 613–618.
10 Reiman EM: In this issue: entering the era of high-density genome-wide association studies. *J Clin Psychiatry* 2007; **68**: 611–612.
11 Levy S, Sutton G, Ng PC *et al*: The diploid genome sequence of an individual human. *PLoS Biol* 2007; **5**: e254.
12 Li Y, Abecasis GR: Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet* 2006; **S79**: 2290.
13 Stranger BE, Nica AC, Forrest MS *et al*: Population genomics of human gene expression. *Nat Genet* 2007; **39**: 1217–1224.

Common inversion polymorphisms and rare microdeletions at 15q13.3

European Journal of Human Genetics (2009) **17,** 149–150; doi:10.1038/ejhg.2008.189; published online 15 October 2008

Sharp *et al*[1] recently described microdeletions at 15q13.3 associated with mental retardation and seizures. These deletions are between Prader–Willi/Angelman break points BP4 and BP5 and include the nicotinic acetylcholine α7 receptor gene (*CHRNA7*). The authors also report a common inversion polymorphism in this region, one orientation of which they suggest might predispose to the microdeletions by non-allelic homologous recombination (NAHR).

15q11–q14 has many segmental duplications, which we have extensively characterised in the human sequence database (Build 36) mainly from one individual.[2] Duplicons of around 300 kb associated with *CHRNA7* and its partial duplication *CHRFAM7A* are in opposite orientation (Figure 1a, black arrows), as are a pair of adjacent

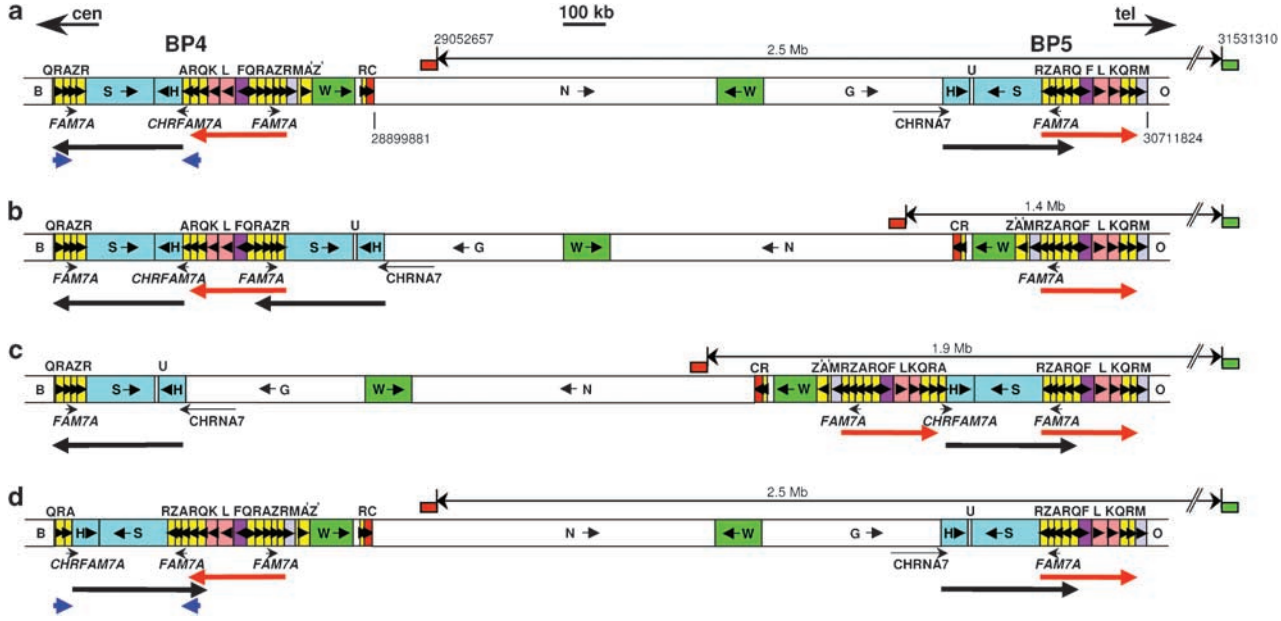

Figure 1 Potential and actual inversion polymorphisms affecting *CHRNA7*. Duplicated segments are shown in the same colour and letter, and unique segments in white, as used previously[5]. (**a**) Database structure from BP4-BP5, showing three pairs of inverted repeats (red, black and blue arrows). (**b**, **c**) Predicted structures for inversion due to NAHR between red (**b**) or black (**c**) arrows. (**d**) Likely structure for confirmed inversion due to NAHR between blue arrows. ▬ Positions of metaphase FISH probes, which are closer together after inversion reported by Sharp *et al*[1]. Coordinates on chromosome 15 build 36 are shown for the above probes[1] and nearby segment junctions.[5]