Public Health Genomics

Public Health Genomics 2012;15:303–311
DOI: 10.1159/000336546

Published online: June 20, 2012

Using Genomic Data to Make Indirect (and Unauthorized) Estimates of Disease Risk

D.R. Nyholt

Queensland Institute of Medical Research, Brisbane, Qld., Australia

Key Words
Association studies · Genetic testing · Genome-wide association studies · Indirect risk estimation · Linkage analysis

Abstract
The number of genetic factors associated with common human traits and disease is increasing rapidly, and the general public is utilizing affordable, direct-to-consumer genetic tests. The results of these tests are often in the public domain. A combination of factors has increased the potential for the indirect estimation of an individual's risk for a particular trait. Here we explain the basic principals underlying risk estimation which allowed us to test the ability to make an indirect risk estimation from genetic data by imputing Dr. James Watson's redacted apolipoprotein E gene *(APOE)* information. The principles underlying risk prediction from genetic data have been well known and applied for many decades, however, the recent increase in genomic knowledge, and advances in mathematical and statistical techniques and computational power, make it relatively easy to make an accurate but indirect estimation of risk. There is a current hazard for indirect risk estimation that is relevant not only to the subject but also to individuals related to the subject; this risk will likely increase as more detailed genomic data and better computational tools become available.

Principles of and Current Approaches to Indirect Risk Estimation

Variation in most human traits and diseases are now viewed as having a genetic component. However, in most human inherited diseases, the biochemical defects are unknown. In other words, the genetic loci causing the disease are not known. This situation requires genomic screening to localize the gene or genes of interest.

Between 1980 and mid-2000, a process known as positional cloning or gene mapping by means of linkage analysis was predominantly used to isolate the genes associated with specific diseases. Many genes were identified for Mendelian diseases, where a single highly penetrant gene effect results in characteristic and well-defined transmission of the disease within a family. Examples of such diseases for which the genes were identified include Huntington's disease [1, 2], Duchenne muscular dystrophy [3], cystic fibrosis [4–6], and neurofibromatosis type 1 [7].

The goal of linkage analysis is to localize a disease or trait locus (T) associated with a given disease. Consider a nuclear family with parents and an affected offspring. A marker locus (M) (a DNA nucleotide or sequence of nucleotides that is known to have multiple alleles in the population, but, unlike a trait locus, it is not necessarily associated with the expression of a phenotype) with a known location in the genome is genotyped in the nuclear family. If T and M are unlinked (i.e. either sufficiently far

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1662–4246/12/0155–0303$38.00/0

Accessible online at:
www.karger.com/phg

Dr. Dale Nyholt
Queensland Institute of Medical Research, Locked Bag 2000
Royal Brisbane Hospital
Herston, QLD 4029 (Australia)
Tel. +61 7 3362 0258, E-Mail dale.nyholt@qimr.edu.au

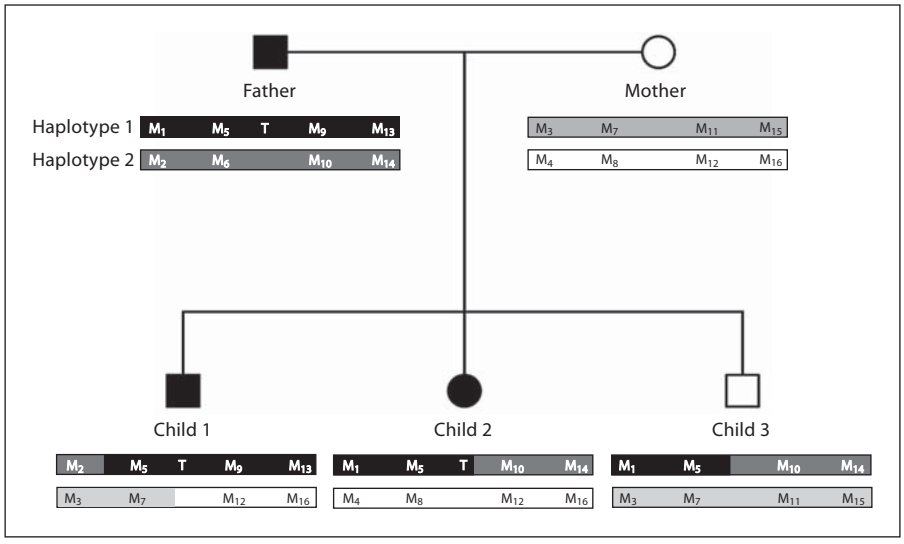

Fig. 1. Segregation of a completely dominant trait locus (causative mutation) and 4 markers in a nuclear family. Square symbols represent males, round symbols represent females; solid symbols represent affected individuals and open symbols represent unaffected individuals. The set of alleles (haplotypes) for each individual are represented by the shaded bars under each symbol.

apart on the same chromosome or on different chromosomes), alleles from both loci segregate independently during meiosis. Therefore, they will be transmitted independently from parents to affected offspring. However, if T and M are linked (i.e. close together on the same chromosome with recombination), the segregation of alleles from both loci during meiosis is not independent: instead, a certain allele from the T will tend to segregate jointly with a certain allele from the M within the family. It is this co-segregation of trait and marker alleles from parents to affected offspring that we aim to detect with a linkage test statistic. The more consistent this pattern is across many meiotic divisions (i.e. parent-offspring transmissions), the lower the recombination fraction between the 2 loci (i.e. the closer T is to M) and, as a result, the stronger the evidence for linkage provided by a linkage test.

Figure 1 shows a completely dominant T (i.e. if you have a copy of the causative mutation, you always express the trait) and 4 markers segregating in a hypothetical nuclear family. Child 1 inherited 3 marker alleles of haplotype 1 {M_5, M_9, M_{13}} and 1 allele of haplotype 2 {M_2} from his father, indicating a recombination occurred between the 1st and 2nd marker; while he inherited 2 alleles of haplotype 1 {M_3, M_7} and 2 alleles of haplotype 2 {M_{12}, M_{16}} from his mother, indicating a recombination occurred between the 2nd and 3rd marker. From her father, child 2 inherited 2 alleles of haplotype 1 {M_1, M_5} and 2 alleles of haplotype 2 {M_{10}, M_{14}}, indicating a recombination between the 2nd and 3rd marker; while she inherited all 4 alleles of haplotype 2 {M_4, M_8, M_{12}, M_{16}} from her mother. From his father, child 3 inherited 2 alleles of haplotype 1 {M_1, M_5} and 2 alleles of haplotype 2 {M_{10}, M_{14}}, indicating a recombination between the 2nd and 3rd marker; while he inherited all 4 alleles of haplotype 1 from his mother {M_3, M_7, M_{11}, M_{15}}.

In this example, although the precise location of recombination between the linkage markers remains unknown, careful examination of the haplotypes each child inherited from the father with respect to disease status allows the T to be mapped. Briefly, the recombination observed in child 1 requires that T lies somewhere to the right of the 1st marker; the recombination in child 2 requires T to lie somewhere to the left of the 3rd marker; while the recombination in child 3 requires T to lie anywhere except between the 1st and 2nd marker. By process of elimination, the T must lie between the 2nd and 3rd marker.

Although the rate of meiotic recombination varies across the genome, all of the approximately 3 billion DNA base-pairs of the human genome can be efficiently screened for linkage using approximately 400 highly polymorphic microsatellite markers (DNA sequence that contains mono-, di-, tri-, or tetra-nucleotide tandem repeats, with different numbers of repeats coded as different alleles) or approximately 6,000 single nucleotide polymorphisms (SNPs).

Once significant linkage is found to a particular region, which is typically 10–20 million base-pairs (Mb) in size, it is further examined (fine-mapped) by testing a denser set of markers, including plausible causative mutations within known genes across the region. In addition to testing the fine-mapping markers for stronger evidence for linkage, they are also tested for association to the trait.

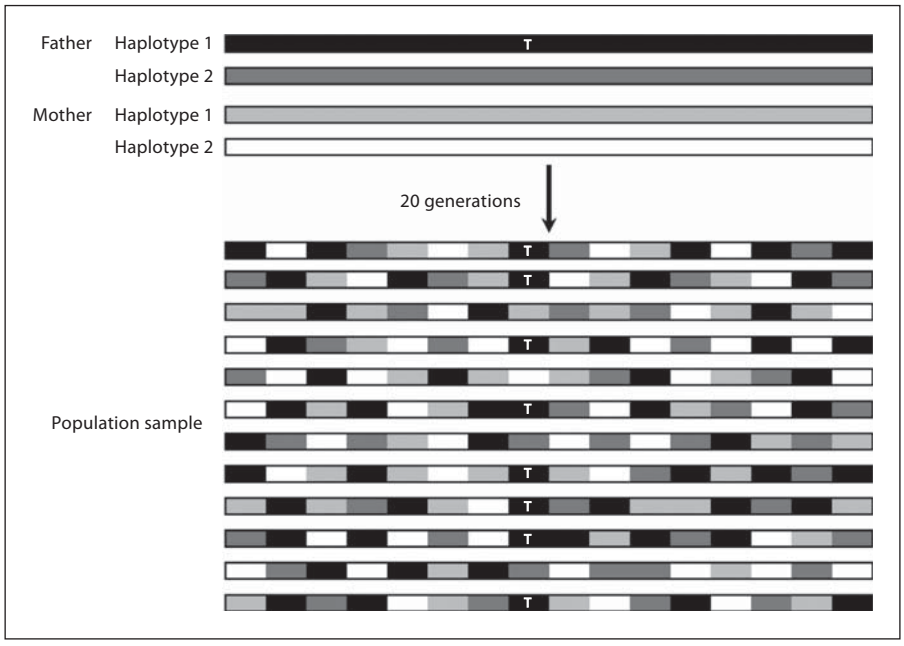

Fig. 2. Basis of association studies. Through successive generations, recombination will cause the T to be separated from the specific alleles of its original haplotype. Adapted from [28].

Unlike genetic linkage studies, association (also called linkage disequilibrium (LD)) studies typically do not investigate familial inheritance patterns (although family-based association designs exist). Instead, researchers test whether a particular allele (polymorphism) is correlated with a particular trait in a population sample. For example, whether or not a particular allele occurs at a higher frequency among disease (case) individuals than in non-disease (control) individuals.

Figure 2 demonstrates how, through successive generations, recombination will cause the T to be separated from the specific alleles of its original (ancestral) haplotype. Particular DNA variants that remain together on ancestral haplotypes are said to be in LD. It is LD that provides the genetic basis for most association studies.

If a gene is primarily involved in causing a disease, then alleles of that gene will occur more frequently in individuals suffering from the disorder than in those without the disorder. A positive disease-marker association occurs when an allele (M_1) of a DNA marker occurs more frequently in sufferers than in nonsufferers and can arise for 3 reasons: if allele M_1 is actually the cause of the disease (fig. 3A); allele M_1 does not cause the disease, but is in LD with the actual cause (fig. 3B); and as an artefact of population admixture, resulting in a false positive. These false positives occur in a mixed population; any trait present at a high frequency in an ethnic group will show positive association with any allele that also happens to be more common in that group [8].

Figure 4 demonstrates how comparing marker allele frequencies in a sample of cases to allele frequencies in a sample of controls (i.e. a case-control association study) hopes to detect an overrepresentation of a particular allele. In this example, the increased frequency of allele M_1 in the cases (0.67) compared to the controls (0.50) indicates that allele M_1 is associated with an increased risk for the disease. The relative risk (RR) for M_1 is $0.67/0.5 = 1.34$, while the odds ratio (OR) is $(0.67/0.33)/(0.5/0.5) = 2.03$.

Although linkage mapping has been highly successful for many Mendelian traits, the vast majority of common human genetic traits are not due to a single gene; rather they are due to many genes or genetic variations (polygenic variation), often interacting with environmental factors (thus they are referred to as complex traits). This genetic (locus) heterogeneity, where different combinations of genetic variation (loci) result in the same phenotype, makes gene mapping by means of linkage analysis extremely difficult for common human complex traits, as no one marker will segregate perfectly with affection status within a family or group of families.

The success of linkage mapping in complex traits had been further hampered by the difficulty in collecting enough useful families for a particular trait to provide sufficient power to detect linkage. That is, because only a subset of families are likely to show segregation between

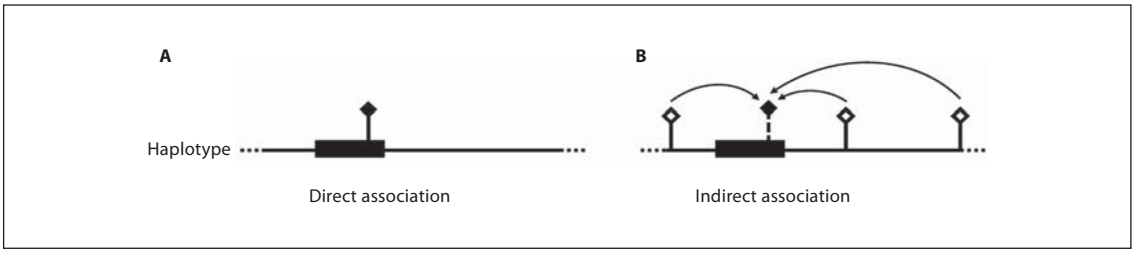

Fig. 3. Testing SNPs for association by direct and indirect methods. In panel **A**, the causative mutation is genotyped and tested directly for association with a trait. For example, a nonsynonymous variation – a mutation that alters the amino acid sequence of a protein (solid diamond) – in a biologically plausible gene (black rectangle). In panel **B**, a subset of known markers (e.g. SNPs) are tested across a region in an attempt to approximately assess all variation (open diamonds). In this situation, the causative mutation is tested for association indirectly, as it is in LD with the genotyped SNPs. Adapted from [29].

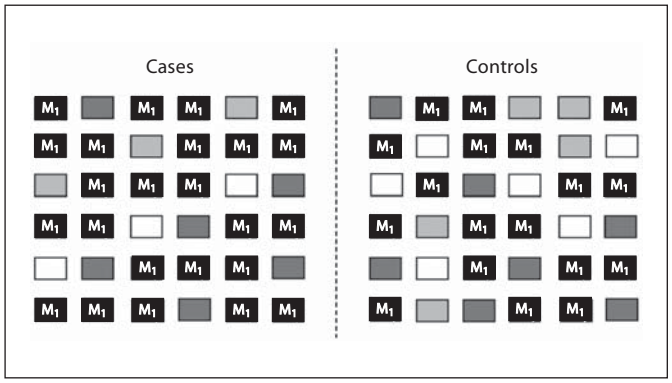

Fig. 4. A case-control association study. The frequency of allele M_1 is overrepresented in the cases compared to the controls.

a particular linkage marker and phenotype, the number of families required is directly proportional to the number of genetic loci underlying the trait. Furthermore, even if one was lucky enough to detect significant linkage, fine-mapping across the implicated region in sufficiently large cohorts (to account for genetic heterogeneity) was an extremely expensive and lengthy process.

Fortunately, the completion of the human genome project in 2003 [9] and the International HapMap Project in 2005 [10], together with advances in genotyping technology, made it possible to screen the human genome's approximately 2.4 million (identified by the HapMap project) common SNPs (minor allele frequency, MAF > 0.01) for association. Genome-wide association (GWA) studies take advantage of the SNP organization along chromosomes, the haplotype structure, to identify sets of highly correlated SNPs (i.e. in strong LD) that can be efficiently tested for association with a trait using a subset of so-called tagSNPs. Most GWA studies have tested approximately 300,000 to 1,000,000 common tagSNPs genotyped in many thousands of individuals. Since the first publication in 2005, as of December 2011, more than 1,100 human GWA publications have examined well over 230 traits, finding over 6,750 genome-wide significant SNP associations and approximately 600 unique disease/trait associations (see http://www.genome.gov/gwastudies/ and [11]).

The scatter plot in figure 5 displays results from a GWA study, termed a Manhattan plot. Here the $-\log_{10}$ of the p values generated by the association analysis, using for instance χ^2 tests, is plotted against chromosomal location, allowing interesting association signals to be clearly seen against background signals. The dashed and dotted lines represent the genome-wide *significant* and *suggestive* association thresholds of $p < 5 \times 10^{-8}$ and $p < 1 \times 10^{-5}$, respectively. The plot in figure 5 indicates that genome-wide significant associations were observed on chromosomes 7 and 17, and at least 12 independent genome-wide suggestive associations were observed.

Although the typical effect size of the risk-related SNPs identified through GWA studies is relatively small (OR = 1.1–1.2), the associated SNPs nonetheless offer crucial insight into the biological pathways and mechanisms underlying human complex traits. Moreover, it is not necessary to know all the causal mutations affecting a trait in order to generate a reliable prediction of an individual's risk, and in fact, the first novel statistical methods that summarize an individual's polygenic risk using genome-wide data were developed some 5 years ago [12].

It is also important to appreciate that while the number of loci identified for many traits are still in the single digits, more loci will be identified as more GWA studies are performed. That is, while an individual, well-designed GWA study has only low power to detect *all* underlying loci, it has sufficient power to detect *some* loci.

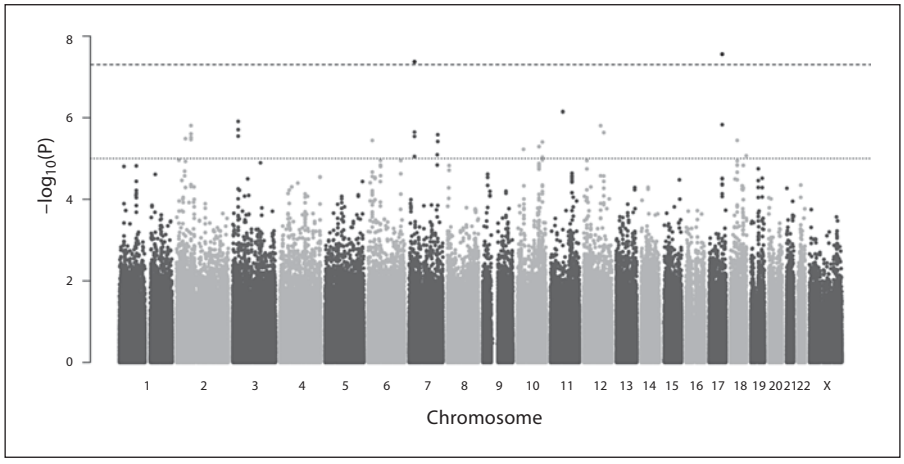

Fig. 5. Manhattan plot of hypothetical GWA results. p values for each SNP analyzed in the GWA study are shown as their $-\log_{10}$ values. Each chromosome is represented in alternating dark and light gray. The dashed and dotted horizontal lines show the thresholds for genome-wide significant and suggestive association, respectively.

Namely, locus heterogeneity combined with random sample variation ensures that particular loci may have a larger observed effect (OR) in one GWA study than in another. As a consequence, the effect sizes may sometimes be upwardly biased relative to the whole population, a feature commonly known as the Winner's curse (also known as the Beavis effect [13]).

However, and more importantly, as more GWA studies are performed and subsequently combined (GWA meta-analysis), more novel loci will be implicated. For example, a simple calculation using Fisher's combined p value method,

$$\chi^2_{2\kappa} = -2\sum_{i=1}^{\kappa} \log_e(p_i)$$

where p_i is the p value for the i^{th} test and k is the number of tests being combined [14], indicates that the combined analysis of 2 similarly powered GWA studies observing the same SNP effect at $p = 1 \times 10^{-5}$ will produce a genome-wide significant association ($\chi^2_4 = 46.05$, $p = 2.4 \times 10^{-9}$). Indeed, combining 3 studies observing a SNP effect at $p = 1 \times 10^{-4}$, 4 studies at $p = 1 \times 10^{-3}$ or 7 studies at $p = 1 \times 10^{-2}$, will produce a genome-wide significant association signal. These results very clearly justify the need for researchers to continue performing and combining GWA studies in order to elucidate the biological pathways and mechanisms underlying human complex traits.

The Public Availability of Genomic Data and the Implications for Indirect Risk Estimation

The continued rapid development of genetic sequencing and genotyping technology has led to an equally rapid increase in genetic variation identification and tools for its characterization. GWA genotyping arrays have now reached 5,000,000 SNPs per individual, while costs for sequencing an individual's whole genome or whole exome (known coding portions of genes) continue to drop at an astounding rate (a times 29,000 price reduction since 2001) and as of this writing are below USD 4,000 and USD 1,800, respectively. A USD 1,000 (whole) genome sequence is expected to be available within the next 2 years.

Additional and more detailed genomic studies are also being developed and performed at a similar rate, with the results, in one form or another, typically available in the public domain (e.g. http://snpedia.com, http://hapmap.ncbi.nlm.nih.gov/, http://www.1000genomes.org/, and http://www.ncbi.nlm.nih.gov/gap). Direct-to-consumer genetic tests are now affordable and widely available to the general public (e.g. http://www.decodeme.com/, http://www.familytreedna.com/, http://www.navigenics.com/, http://www.personalgenomes.org, and https://www.23andme.com/). A particularly striking example concerns the current aim of the Personal Genomes Project (PGP) (http://www.personalgenomes.org) of recruiting 100,000 volunteers 'who are willing to share their genome sequence and many types of personal information with the research community and the general public' (http://www.personalgenomes.org/participate.html).

The first 10 participants in the PGP (the 'PGP-10'), all of whom are named, have shared their DNA sequences, medical records and other personal information with the research community and the general public (http://www.personalgenomes.org/pgp10.html). One of the identifiable PGP-10 participants openly declares in his publicly available medical history summary: 'I have experienced

recurrent bouts of depression and anxiety for the last several years and recently began taking SSRIs [selective serotonin re-uptake inhibitors]' (http://www.personalgenomes.org/public/4.html). To date, more than 1,000 individuals have enrolled in the PGP ('PGP-1K') and allow their data to be available on the internet (http://www.personalgenomes.org/pgp1k.html).

Even if a participant decides to restrict the release of specific information, however, their participation in any of the genomic studies, combined with enhanced genomic knowledge, advances in mathematical and statistical techniques, and increased computational power, increase the potential for almost anyone with access to the internet, a good computer, and persistence to undertake without a person's consent an indirect estimation of that person's risk for a particular trait, and to arrive at a remarkably accurate conclusion.

Research Design and Methods

The goal of the research team was to demonstrate why genetic information is hard to hide once it has been collected, and that an unauthorized, that is, indirect estimation of an individual's risk for a particular trait could be arrived at using standard, currently available computer hardware and software. The research team designed a test for this hypothesis using genetic information publicly available over the internet, a good computer and some persistence.

The team had to identify publicly available genetic material to use. A particularly famous example of such data is the publication and release to public databases of Dr. James Watson's sequenced genome [15], excepting all gene information about apolipoprotein E (ApoE). Dr. Watson, best known as one of the codiscoverers of the structure of DNA in 1953 with Francis Crick, had requested that his apolipoprotein E 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 had already claimed one of his grandmothers [16].

To demonstrate 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. [15]), the research team utilized SNP genotypes identified in Dr. J. Craig Venter's genome sequence, which was also released publicly around the same time [17]. Importantly, Dr. Venter's sequence data redacted neither information around *APOE* nor the information that he is heterozygote for both the LOAD high-risk *APOE* SNP rs429358 (T/C) and for the nearby correlated *APOC1* SNP rs4420638 (A/G) (i.e. the *APOC1* SNP is in strong LD with the *APOE* SNP [18]). We therefore replicated Dr. Watson's sequence redaction in Dr. Venter's data and attempted to infer or, perhaps more correctly, *impute* Dr. Venter's *APOE* status using publicly available data [19]. (For a description of genotype imputation, see fig. 6, 7.)

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. However, we contacted Dr. Watson and colleagues 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) nominally removed all data from the 2 Mb region surrounding *APOE*.

Briefly, genotype imputation was performed using the MaCH (version 1.0.16) computer program [20, 21], 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 2-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 (i.e. 99.2% likelihood of being A/G) exemplifies the utility of imputing *APOE* genetic risk for LOAD [19].

Another example of indirect risk estimation is one in which the genetic distance of an individual with respect to 2 population samples is used to infer the presence of an individual of known genotype in a sample for which only allele frequencies are known [22]. Such an approach numerically expresses how genetically similar individuals or populations are by comparing their allele frequencies. Although the method is less accurate in practice than originally thought – due to its sensitivity to underlying assumptions [23, 24], such identification could be utilized in forensic science to establish the presence or absence of a person's DNA in a mixture of DNA and used to establish whether an individual was a member of a particular disease cohort of a GWA study [22–24].

Conclusion: An Outlook to the (Near) Future

To date, GWA studies have been based on information learned from the HapMap project and therefore have been concentrated on testing common SNP variation for association with human complex traits. Most recently, the 1000 Genomes project [25], which involves the whole-genome sequencing of hundreds of individuals sampled from broad geographic regions, is generating high-quality haplotypes for >1,000 individuals, including near-complete coverage of SNPs with population MAFs of 1% or more [26]. Variation identified by the 1000 Genomes project, allowed Illumina® to design and release a human exome GWA array that allows inexpensive (<USD 85 per sample) interrogation of approximately 250,000 putative functional exonic variants.

In addition to providing a near-complete catalog of SNPs with MAF $\geq$ 0.01, the 1000 Genomes project is identifying millions of rarer SNPs and cataloging other

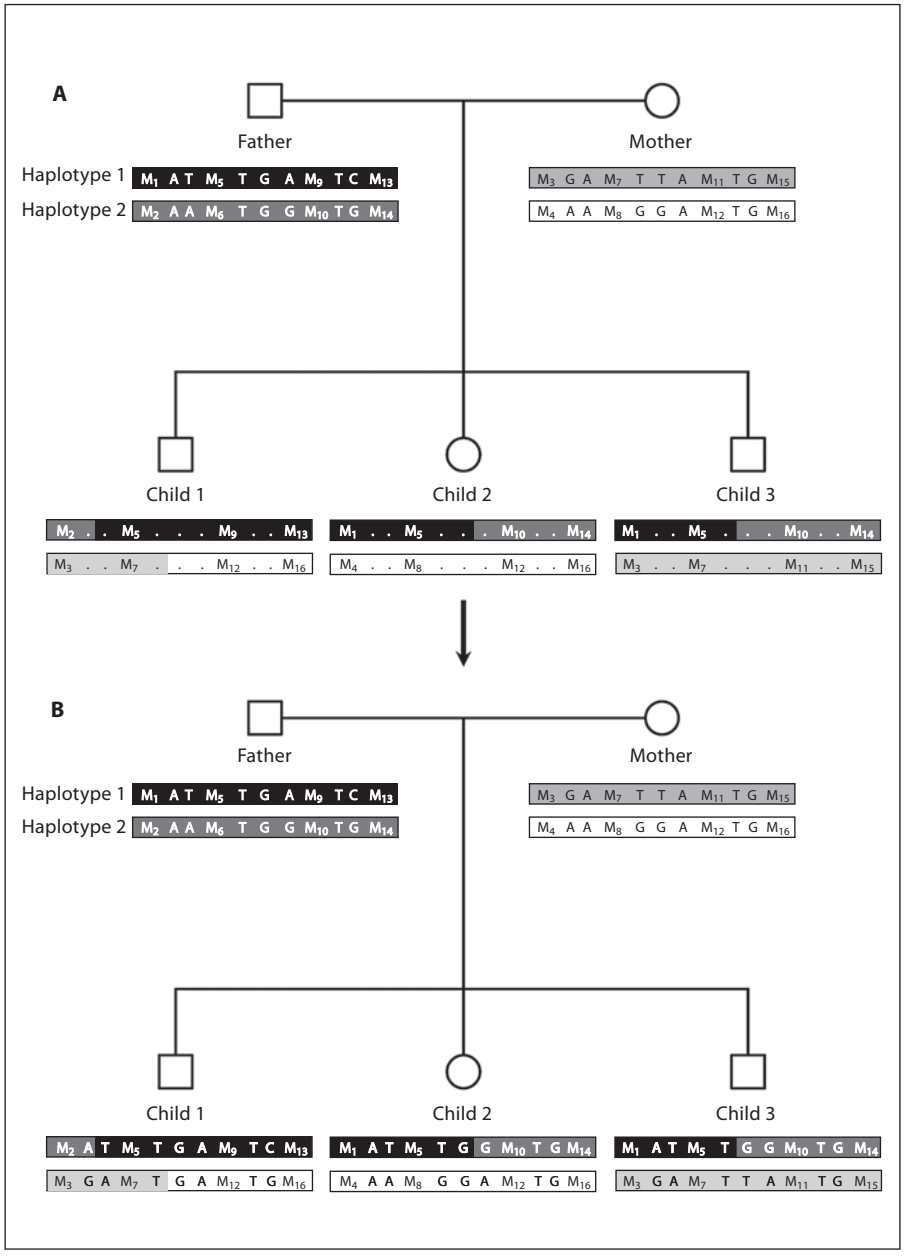

Fig. 6. Genotype imputation within a sample of related individuals. Panel **A** presents a pedigree similar to that in figure 1, but, in addition to the microsatellite markers genotyped in the 2 parents and 3 children, 7 SNPs were genotyped in the parents. Panel **B** shows how the observed SNP genotypes and haplotype information have been combined to fill in the SNP genotypes that were originally missing in the children. Adapted from [20].

types of variation such as insertions/deletions (indels) and structural variants (SVs) [27] – a type of copy number variant. The most recent 1000 Genomes reference data (December 2011) contains approximately 36.6 million SNPs, 3.8 million short indels and 14,000 deletion SVs (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/README.20111111_phase1_integrated_call_set). Moreover, the impending release of the next generation of imputation software will allow these and other types of variation to be inferred utilizing existing common SNP GWA data and 1000 Genomes project reference haplotypes.

Thus, as our knowledge of the human genome increases (identifying new variation associated with risk) and computational tools are further developed (which accurately and efficiently measure genetic risk), the potential for *indirect* estimation of risk will continue to increase.

It is also vital to appreciate that because such risk is inherited, it can be shared among relatives and there-

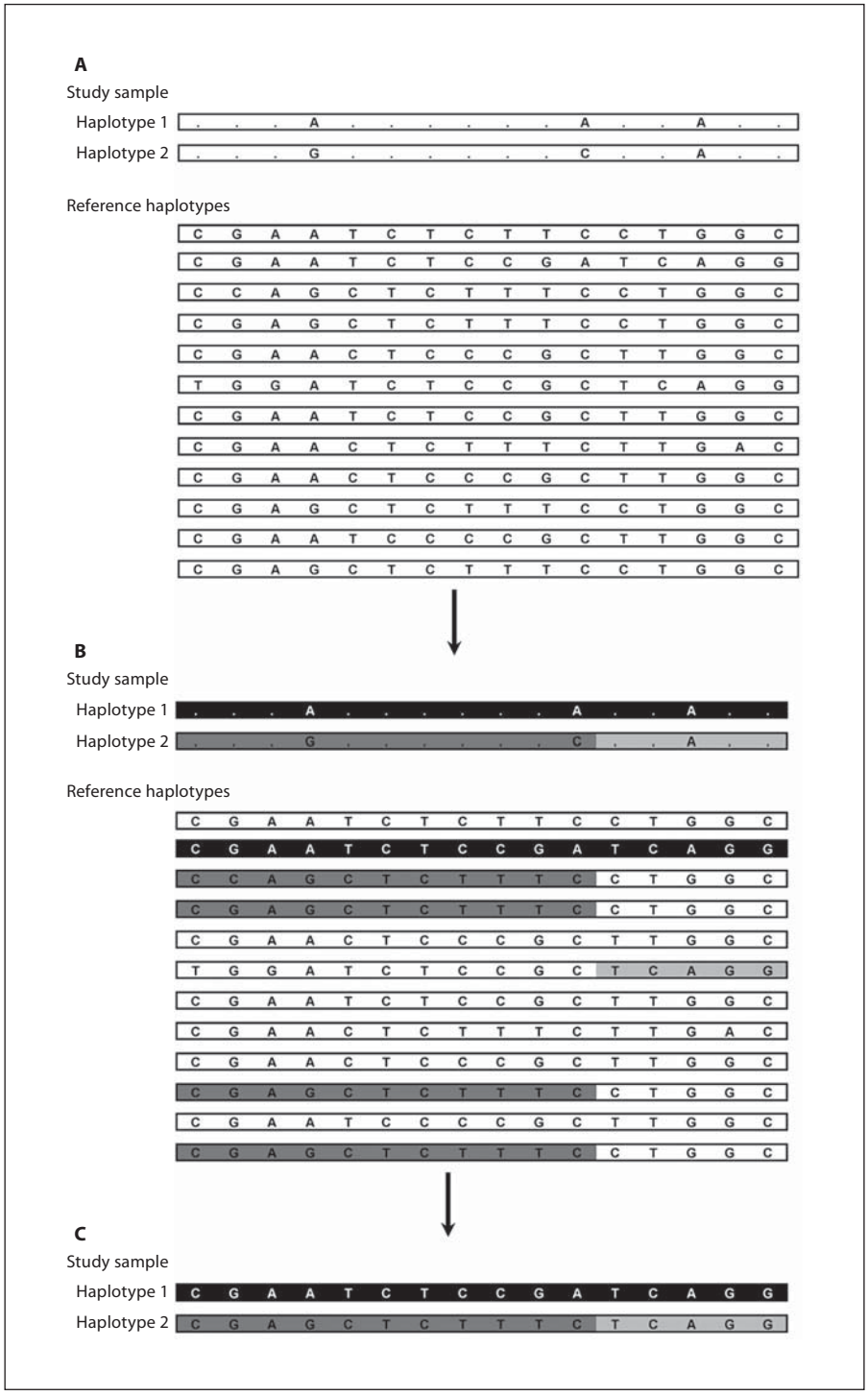

Fig. 7. Genotype imputation within a sample of unrelated individuals. Panel **A** represents an individual (study sample) genotyped at 3 SNPs across a small region (say 100–200 kb) and a collection of haplotypes from a reference sample (e.g. HapMap reference). Panel **B** shows how the observed SNP genotypes and reference haplotypes can be used to predict the study sample's haplotype, and in panel **C**, subsequently fill in the unobserved SNP genotypes in the study sample. Adapted from [20].

fore not only pertains to the specific individual with genetic data in the public domain, but also to their relatives.

The potential for the indirect estimation of genetic risk has considerable relevance to concerns about privacy, confidentiality, discriminatory, and defamatory use of genetic data as well as relevance regarding the complexities of informed consent for both research participants and their close genetic relatives in the era of personalized genomics.

Acknowledgements

This work was supported by Australian NHMRC Grants 389892, 339462, and 442915 and Australian Research Council Grant DP0770096. The author was supported by the NHMRC Research Fellowship (613674) and the ARC Future Fellowship (FT0991022) schemes. The author gratefully acknowledges Prof. Dr. Peter Dabrock, Prof. Dr. Herbert Gottweis, and Dr. Andréa Vermeer for their support and organization of the PRIVATE Gen Workshop 'Privacy and Post-Genomics Medical Research: Challenges, Strategies, Solutions,' supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF).

References

1 Gusella JF, Keys C, VarsanyiBreiner A, Kao FT, Jones C, Puck TT, Housman D: Isolation and localization of DNA segments from specific chromosomes. Proc Natl Acad Sci USA 1980;77:2829–2833.
2 Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, Young AB, Shoulson I, Bonilla E, Martin JB: A polymorphic DNA marker genetically linked to Huntington's disease. Nature 1983; 306:234–238.
3 Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM: Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the *DMD* gene in normal and affected individuals. Cell 1987;50:509–517.
4 Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC: Identification of the cystic fibrosis gene: genetic analysis. Science 1989; 245:1073–1080.
5 Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS, Tsui LC: Identification of the cystic fybrosis gene: cloning and characterization of cDNA. Science 1989;245:1066–1073.
6 Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan JR, Tsui LC, Collins FS: Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 1989;245:1059–1065.
7 Xu GF, O'Connell P, Viskochil D, Cawthon R, Robertson M, Culver M, Dunn D, Stevens J, Gesteland R, White R, Weiss R: The neurofibromatosis type 1 gene encodes a protein related to GAP. Cell 1990;62:599–608.
8 Lander ES, Schork NJ: Genetic dissection of complex traits. Science 1994;265:2037–2048.
9 International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. Nature 2004; 431:931–945.
10 International HapMap Consortium: A haplotype map of the human genome. Nature 2005;437:1299–1320.
11 Yu W, Yesupriya A, Wulf A, Hindorff LA, Dowling N, Khoury MJ, Gwinn M: GWAS integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. Eur J Hum Genet 2011;19:1095–1099.
12 Wray NR, Goddard ME, Visscher PM: Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 2007;17:1520–1528.
13 Beavis WD: Qtl analyses: power, precision, and accuracy; in Paterson AH (ed): Molecular Dissection of Complex Traits. Boca Raton, CRC Press, 1998, pp 145–162.
14 Fisher RA: Statistical methods for research workers, ed 5, Edinburgh, Oliver and Boyd, 1932.
15 Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872–876.
16 Check E: James Watson's genome sequenced – discoverer of the double helix blazes trail for personal genomics. Nature News. http://www.nature.com/news/2007/070528/full/news070528-10.html.
17 Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC: The diploid genome sequence of an individual human. PLoS Biol 2007;5:e254.
18 Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH, Zismann VL, Beach TG, Leung D, Bryden L, Halperin RF, Marlowe L, Kaleem M, Walker DG, Ravid R, Heward CB, Rogers J, Papassotiropoulos A, Reiman EM, Hardy J, Stephan DA: 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.
19 Nyholt DR, Yu CE, Visscher PM: On Jim Watson's APOE status: genetic information is hard to hide. Eur J Hum Genet 2009;17: 147–149.
20 Li Y, Willer C, Sanna S, Abecasis G: Genotype imputation. Annu Rev Genomics Hum Genet 2009;10:387–406.
21 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–834.
22 Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;4:e1000167.
23 Braun R, Rowe W, Schaefer C, Zhang J, Buetow K: Needles in the haystack: identifying individuals present in pooled genomic data. PLoS Genet 2009;5:e1000668.
24 Visscher PM, Hill WG: The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet 2009;5:e1000628.
25 The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature 2010;467: 1061–1073.
26 Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5: e1000529.
27 Handsaker RE, Korn JM, Nemesh J, McCarroll SA: Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 2011;43: 269–276.
28 Cardon LR, Bell JI: Association study designs for complex diseases. Nat Rev Genet 2001;2:91–99.
29 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6:95–108.