OPINION

The importance of phase information for human genomics

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol and Nicholas J. Schork

Abstract | Contemporary sequencing studies often ignore the diploid nature of the human genome because they do not routinely separate or 'phase' maternally and paternally derived sequence information. However, many findings — both from recent studies and in the more established medical genetics literature — indicate that relationships between human DNA sequence and phenotype, including disease, can be more fully understood with phase information. Thus, the existing technological impediments to obtaining phase information must be overcome if human genomics is to reach its full potential.

Advances in DNA-sequencing technologies have made it possible to efficiently characterize large segments of, if not entire, individual human genomes[1–4]. Sequencing the genomes of members of the same family[4], from individuals with and without a particular disease[5], or from individuals sampled randomly from the population[6], can lead to insight into the role of both common and rare DNA sequence variants in mediating phenotypic expression. However, most studies of this kind typically involve sequencing DNA samples that contain both the maternally and the paternally derived DNA associated with the homologous chromosomes inherited by an individual. As such, they essentially ignore the phase of the DNA in those samples — that is, they ignore the unique nucleotide content of the two homologous chromosomes an individual possesses, referred to as an individual's 'diplotype'. Human genome-related initiatives, such as the International HapMap Project and the 1000 Genomes Project, have considered the importance of haplotyping. However, this is usually in the service of assessing, through linkage-disequilibrium measures, the likelihood that variants at one genomic position indicate the presence of variants at neighbouring positions. Rarely does contemporary consideration of phase information concern the molecular physiological consequences of having variants uniquely distributed across two homologous chromosomal copies of a genomic region[7].

The dearth of phased human genomic data is primarily due to the computational complexity associated with, and the lack of cost-effective approaches for, obtaining phase information. Well-established phenomena such as compound heterozygosity in monogenic disorders support the importance of phase information for relating genotype to phenotype. In addition, recent studies have described settings in which the characterization of the specific nucleotides on each homologous copy of a gene or genomic region inherited by an individual is essential for understanding phenotypic expression[4,8–11]. Here, we discuss these studies and consider specific instances in which the specific set of variants on each homologous chromosome contributes to phenotypic expression and disease states. We also briefly describe other settings in which phase information is important for human genomics research. We provide an overview of current methods for obtaining phase information, and discuss their limitations and prospects for future improvement. We also coin the term 'diplomics' to refer to scientific investigations that leverage phase information in order to understand how molecular and clinical phenotypes are influenced by unique diplotypes. We ultimately argue that diplomic investigations will be key to the design and conduct of future functional genomic studies, as well as large-scale human DNA-sequencing initiatives.

Diplotype is important for function

To understand the importance of phase information in human sequencing studies, it is necessary to understand the settings in which the balance of *cis*- and *trans*-acting variants on the two homologous copies of a genomic region affect phenotypic expression (FIG. 1). A number of recent studies have used high-throughput DNA sequencing to investigate how nucleotide variation affects gene function in a way that depends on which chromosome homologue this variation is located[12–14].

Widespread allele-specific expression. The ability of a cell to selectively express a gene on a single chromosome while the gene on the homologous chromosome is silenced is a well-characterized phenomenon in diploid cells. This effect can be caused by, but is not necessarily limited to, nucleotide variation or methylation at the locus that regulates or harbours the affected gene. Recent studies have indicated that such allele-specific expression (ASE) is widespread in humans. Two groups recently used RNA sequencing to study how *cis*-acting sequence variation influences gene expression[10,11]. Both groups showed that 1–5% of human genes are influenced by *cis*-acting DNA sequence variants (known as expression QTLs, or eQTLs) in the contexts that they tested. Most heterozygous *cis*-acting eQTLs resulted in one copy of the gene being expressed at a higher level than its homologous copy — hence exhibiting ASE. There are a number of possible biological mechanisms responsible for ASE. Kasowaki *et al.*[8], for example, showed that the binding strengths of two transcription factors (TFs) exhibit wide variation at ~25% of specific TF target sequences across different individuals. Differences in binding strength across individuals were frequently associated with the existence of genetic variants in these binding regions. Such differences in binding strength were not only shown to be correlated with differences in the expression levels of genes

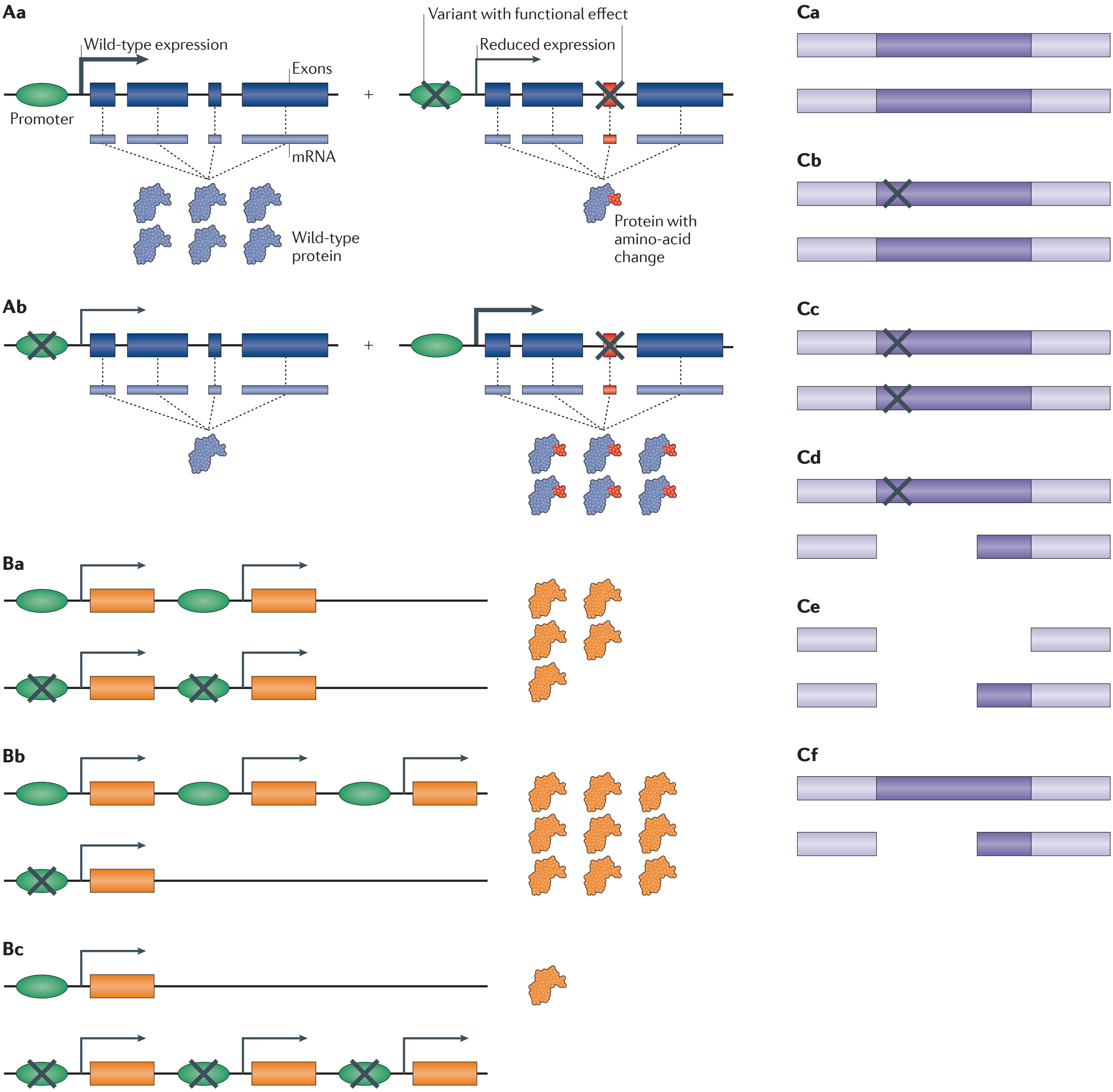

Figure 1 | **The distribution of variants between homologous chromosomes can affect gene function. A** | Distribution of variants that affect regulation and protein function, showing the two homologous gene segments in a single diploid individual. **Aa** | In this case, the leftmost homologue does not contain variation that influences either the expression or the structure of the encoded protein. By contrast, the rightmost homologue contains sequence variation in the promoter that reduces overall expression of the gene and exonic sequence variation that upsets the amino-acid sequence of the encoded protein. **Ab** | Here, the variants in the promoter and exonic sequence are distributed between different homologues. The combination of these homologues in a single individual can lead to haploinsufficiency if the homologue that does not have a functional variant cannot compensate for the affected homologue. If it can compensate, the overall functioning of the gene could be normal, owing to both the downregulation of the aberrant protein and the normal expression of the wild-type protein. **B** | Potential functional effects of haplotypes involving structural variants. Scenarios are shown involving copy-number variants and point mutations in a diploid setting. The possibilities depicted in parts **Bb** and **Bc** reflect increased and decreased overall gene expression, respectively, relative to that in **Ba**. **C** | Unmasking of deleterious mutations through gene deletion. A genomic region is shown that harbours a gene that is often either partially or completely deleted and that also harbours functionally relevant point mutations. **Ca** | Neither homologous copy of the gene harbours a variant. **Cb** | One of the gene homologues carries a point mutation. **Cc** | Both gene homologues carry a point mutation. **Cd** | One of the gene homologues carries a deletion and the other carries a point mutation. **Ce** | Both of the gene homologues carry a deletion. **Cf** | One of the gene homologues carries a deletion. Each situation could produce a different phenotype; for example, in part **Cd** the deletion depicted could unmask the deleterious effect of the point mutation on the other chromosome.

Table 1 | **Example clinical conditions and disorders influenced by compound heterozygosity in single genes**

Disease	**Gene names**	**Mutations implicated in compound heterozygosity**	**Refs**
Blistering skin	*COL7A1*	G2316R, G2287R	59
Cerebral palsy	*PROC*	N2I, S181R	60
CMT	*SH3TC2 KARS*	Y169H, R954X, L133H, Y173SfsX7	9,61
Deafness	*GJB2*	Additive effect of multiple reported recessive and dominant mutations	62
Haemachromatosis	*HFE*	H63D, 2282Y	63
Mediterranean fever	*MEFV*	E14Q, M694I. M694I alone is associated with a mild phenotype	64
Miller syndrome	*DHODH*	G152R, G202A	4
Paraganglioma	*SDHB*	V110F and splice donor c. 200 + 7 A > G	65
Hyperphenylalaninaemia	*PAH*	Multiple PAH variants explained non-*PKU* hyperphenylalaninaemia cases when acquired as compound heterozygote	66
FBPase deficiency	*FBP1*	G164S, 838ΔT	67
Ataxia-telangiectasia	*ATM*	Attenuated phenotype: D2625E, A2626P and splice site c.496+5 G>A	68
Glycogen storage type II	*GAA*	R600C and splice site c.546G>T. Splice variant has reduced expression	69
Chondrodysplasias	*DTDST*	T266I, 340ΔV	70
Turcot's syndrome	*PMS2*	1221ΔG, 2361ΔCTTC	71

CMT, Charcot–Marie–Tooth neuropathy; FBPase, fructose-1,6-bisphosphatase; PAH, phenylalanine hydroxylase.

associated with the TF target sites, but also to have clear segregation in families — therefore exhibiting heritability — thus confirming the genetic origins of the variation in gene expression levels[8,15].

Epigenetic changes in a genomic region can also influence gene expression in a chromosome- or allele-specific manner. Zhang *et al.*[16] studied whole-genome methylation and gene-expression patterns in 153 adult cerebellum samples as a function of the existence of inherited DNA sequence variants. They identified a number of highly significant associations between apparently *cis*- and *trans*-acting SNPs and specific methylation patterns. Many of the SNPs that influenced methylation, and so exhibited allele-specific methylation (ASM), also influenced the expression levels of particular genes. ASM may also influence disease susceptibility, as suggested by Steffanson and colleagues in a study of genetic variants associated with type 2 diabetes[17]. Other studies suggest that ASE or ASM may be widespread even across different cells within an individual[18,19], although the degree to which this heterogeneity can be attributed to the effect of heterozygous *cis*-acting variants is an open question.

Studies showing widespread ASE and ASM make it clear that the specific DNA sequence and/or epigenetic context associated with each of the two homologous copies of a gene or regulatory element influences the function of these elements in their combined, diploid state. Importantly for the focus of this Opinion article, the effect of ASE and ASM on gene function is likely to be compounded if there are other forms of variation in the same gene (FIG. 1A). A case in point is that of chromogranin A (*CHGA*), in which common variation in the promoter region has been shown to affect expression and result in ASE. In addition, coding variants have been identified that alter cholinergic inhibition owing to encoded structural deformations that they induce in proteins[20]. Simply cataloguing the genotypes by combining sequence information from the two chromosomes and ignoring whether heterozygous variants are in *cis* or *trans* with other variants would provide incomplete knowledge of an individual's phenotype with respect to both gene expression and protein function. Thus, the haplotype combinations (diplotype) that an individual possesses are paramount to understanding whether an inhibitory allele is overexpressed or underexpressed relative to the normal allele. Such phenomena are discussed further below in the context of complex disease.

Duplications, deletions and chromosome inequivalence. There is a growing literature on the existence and effect of different numbers of copies of entire genes or parts of genes in individual genomes[21,22]. Knowledge of the number of functioning copies of a gene in a single human genome is crucial for determining the potential phenotypic effect of such copy-number variations (CNVs). However, it might be just as important to know how those gene copies are distributed across the two sets of chromosomes in each cell. For example, heterozygous *cis*-acting sequence variations may exist in the surrounding regulatory regions of these gene copies and so influence their function. Thus, the specific combination of gene copies and *cis*-regulatory variants on each chromosomal homologue may dictate the function of those gene copies (FIG. 1B). In this context, it is known that many cancers have somatically acquired 'amplifications' in the form of increased copies of particular genes[23]. Many of these genes have also been found to possess point mutations that influence the function of particular copies[23], which may, in turn, influence tumorigenesis[24]. Understanding the phenotypic effects of deletions also requires knowledge of how variation is partitioned between chromosomes. An example is the phenomenon of 'unmasking' potentially deleterious mutations in one copy of a gene when the homologous copy is deleted[25] (FIG. 1C).

Diplotypic effects and disease

In addition to the influence of haplotype-specific *cis*-acting variation on gene function in cellular and molecular physiological settings, there have been many documented instances in which specific diplotypes influence disease and clinically relevant phenotypes. We describe examples of such cases below.

Compound heterozygosity. Human disorders often exhibit subtle variation in their phenotypic manifestations. Many studies investigating the genetic mechanisms that underlie

Table 2 | **Example studies assessing the effect of combinations of unique gene-specific diplotypes on a complex phenotype**

Gene	Phenotype assessed	Genetic basis	Refs
ADRB2	Response to asthma therapy	Complex promoter and coding-region haplotypes at the *ADRB2* locus alter receptor expression	72
HG1	HGH expression	Non-additivity of the effects of 16 *HG1* SNPs with individual effects, depending on haplotype context	73
FANCD2	Breast cancer	If at least one copy of a specific *FANCD2* haplotype is present, carriers are at fourfold risk	74
IL1B	IL-1β activity	Individual SNPs in the *IL1B* promoter have either an upregulatory or downregulatory effect depending on haplotype context	75
PRKAG3	LDL cholesterol	Homozygotes for specific alleles in a specific *PRKAG3* diplotype exhibited the highest LDL cholesterol of all the frequent diplotypes	76
ATM	Non-small-cell lung cancer	On the basis of haplotype and diplotype analyses, a specific diplotype at the *ATM* locus confers risk	77
MDR1	Multiple myeloma	Protective effects were identified in heterozygotes and homozygotes for a specific diplotype at the *MDR1* locus	78
NPAS3	Schizophrenia and bipolar disorder	Combinatorial action of haplotype pairs was associated with overall susceptibility	79
ADIPOQ	Rosiglitazone response	A specific diplotype at the *ADIPOQ* locus exhibited stronger association with enhanced response than other diplotypes	80

HGH, human growth hormone; IL-1β, interleukin-1β; LDL, low-density lipoprotein.

this variation, especially in the context of monogenic, overtly Mendelian disorders, have implicated the phenomenon of compound heterozygosity (TABLE 1). Compound heterozygosity occurs when the two homologous copies of a genomic region each harbour unique sequence variants, but at different positions in that region. These variants are thought to perturb the function of the two homologous copies of a gene in different ways, with their combined molecular effects resulting in a phenotype that is distinct from that seen if one homologous gene carries both deleterious variants[26]. Thus, in settings in which compound heterozygosity may have a role, merely knowing that an individual is heterozygous for mutations or variants at relevant loci is not enough: knowledge about the specific diplotype is essential.

Additional instances of clinically relevant compound heterozygosity have been uncovered in large-scale human sequencing studies. For example, Roach *et al.*[4] sequenced the genomes of a pair of siblings with two apparently recessive disorders, Miller syndrome and primary ciliary dyskenesia, and also sequenced the genomes of their parents. Sequence information from the siblings was phased by tracking the transmission of variants from parents to offspring, although not all variants could be unequivocally determined as maternal or paternal in origin. For Miller syndrome, two variants at different positions in the same gene, one on the maternally inherited homologue of the gene and one on the paternally inherited homologue, were proposed to influence the disease. Other instances of compound heterozygosity occur in the context of the 'two hit' model of cancer, in which an individual inherits a disruptive cancer-susceptibility variant in one homologue of the gene and then develops a disruptive somatic mutation at a different position in the other homologue. This leads to dysfunction in both gene copies and a potential tumorigenic effect[26]. It is unclear how often the phenomenon of compound heterozygosity is likely to affect different diseases. However, the fact that there are many known instances in which it does so suggests that studies that use sequencing to identify variants that influence a disease need to take this possibility into account, a task that clearly requires phase information.

Complex diplomic phenomena in common disease. Documented instances of compound heterozygosity have typically involved low-frequency, highly penetrant alleles. It is unclear how such effects relate to the higher-frequency alleles of low effect size that have been shown to contribute considerably to many complex, common disorders over the past few years[27]. Despite this, some researchers have begun to consider the influence of haplotypic effects in the context of genome-wide association studies investigating common disorders that may reflect compound heterozygosity[28,29]. In addition, there is growing evidence for the involvement of specific diplotypes, involving combinations of multiple *cis*-acting variants — some in regulatory regions and some in coding regions — in giving rise to phenotypic effects that contribute to common diseases. The principles discussed above and illustrated in FIG. 1 are also likely to apply in such settings. TABLE 2 summarizes a range of recently documented instances and we describe some specific examples below.

Two groups identified a strong association between systemic lupus erythematosus (SLE) and haplotypes that contain variants in the protein-coding region of the gene tumour necrosis factor α-induced protein 3 (*TNFAIP3*)[30,31]. Two additional haplotype blocks located ~200 kb upstream and downstream of the *TNFAIP3* coding region also showed strong independent signals for association with the disease but were not in linkage disequilibrium with the variants in the coding-region haplotype. The findings raised an important question about how these variants modify autoimmune disease susceptibility in different haplotype conformations. Although neither of the studies explicitly investigated how the variants directly interacted when in *cis* confirmation, they did provide indirect evidence that the specific diplotype is important.

Graham and colleagues also studied another potential SLE gene, interferon regulatory factor 5 (IRF5)[32–34], which also harbours multiple coding and non-coding variants that exhibit associations with autoimmune diseases. Three separate variants were identified within the IRF5 coding region that disrupt IRF5 function through different mechanisms: abnormal splicing of exon 1b, a 10-residue deletion in exon 6, and disruption of a cleavage and polyadenylation specificity factor (CPSF) site[33]. Again, an important question is how the distribution

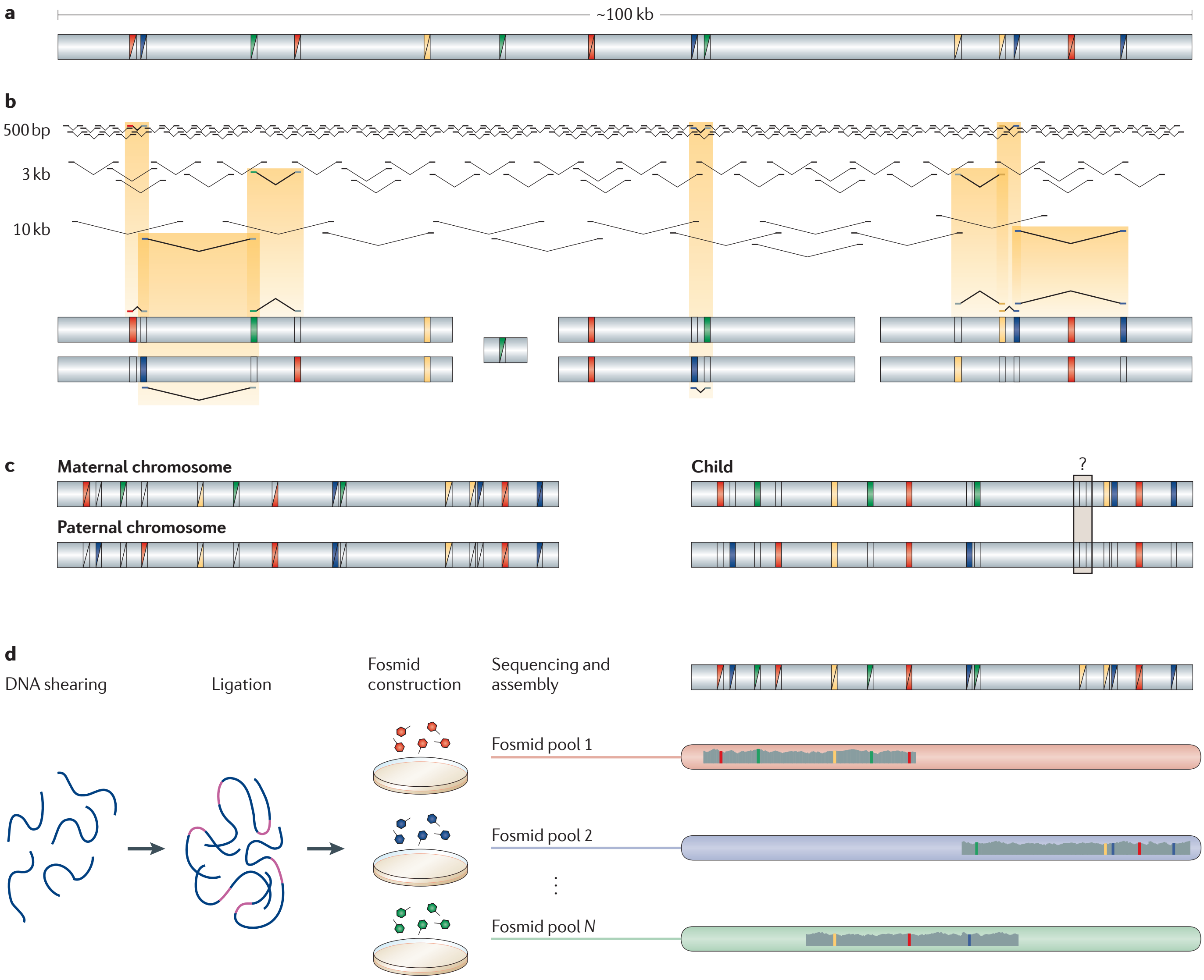

Figure 2 | **Strategies for empirical haplotype reconstruction.** **a** | A hypothetical 100 kb stretch of sequence harbours multiple variants compared with the human reference, as designated by the coloured squares. Variants can be homozygous (solid coloured squares) or heterozygous (split coloured squares). **b** | Sequence reads from libraries of multiple insert sizes can be leveraged to link heterozygous sites together. Informative reads are highlighted and displayed a second time against the diploid reconstruction. The assembly consists of blocks of sequence with gaps arising when variants fall outside the distance of the insert sizes used for sequencing. **c** | Parental information allows for the separation of chromosomal variants except in instances in which both parents are heterozygous, as demonstrated by the black box in the child's assembly. **d** | Laboratory-based methods such as the sequencing of fosmid pools allow for the separation of homologous chromosomes. DNA is sheared, ligated with fosmid vector sequence, packaged and transfected into the bacterium *Escherichia coli*. Pools of fosmid sequence — each containing only a small fraction of the total genome broken into ~40 kb segments — are sequenced independently. The sequenced libraries are then mapped and assembled for phase reconstruction.

of these variants across the two homologous copies of IRF5 in an individual affects overall IRF5 function. For example, the combination of a variant in a splice site and a CPSF mutation on the same chromosome may have a more attenuated effect than if the two variants are on different chromosomes, because in the former case the existence of one functional gene copy with neither variant may compensate for the affected copy with two mutations. Interestingly, Graham and colleagues, and others, have identified further associations implicating additional *cis*-acting regulatory variants in SLE susceptibility[33–35].

A recent example of a complex setting implicating *cis*-acting variants along with structural or repetitive sequences on single chromosomes involved the study of mutations that cause facioscapulohumeral muscular dystrophy[36]. Here, the contraction of microsatellite repeats has a phenotypic effect only when variants that modify the stability of the double homeobox 4 (*DUX4*) transcript are on the same chromosome as the repeats.

Importance of phase in other settings

In addition to the importance of phase information in resolving how combinations of variants uniquely situated on each homologous genomic region may affect diploid gene function, there are other settings in

Table 3 | **Example methods and software for haplotyping and phasing**

Method name	Data type[*]	Comments	Refs
Hapi	Pedigree genotype[‡]	Dynamic programming-based haplotype assembly	81
ZRBA	Pedigree genotype	Zero-recombination block partition algorithm	82
He *et al.*	Sequencing reads[§]	Dynamic programming-based haplotype assembly	58
HapCUT	Sequencing reads	Max-Cut-based algorithm applicable to arbitrary length reads and insert sizes	56
HASH	Sequencing reads	Markov chain Monte Carlo algorithm for haplotype assembly	57
SHAPE-IT	Genotype[‖]	Tree representation of hidden Markov model	83
Beagle	Genotype	Fast and accurate algorithm for phasing using a haplotype-cluster model	84
HaploRec	Genotype	Uses frequencies of haplotype fragments for phasing	85
fastPHASE	Genotype	Haplotype-clustering model for phasing large data sets	86
HAP	Genotype	Imperfect phylogeny approach	87
PL-EM	Genotype	Expectation-maximization algorithm combined with partition-ligation	88
Merlin	Pedigree genotype	Uses sparse gene-flow trees to reduce computing requirements	89
Phase	Genotype	One of the most accurate phasing method available but slow on large data sets	90
Allegro	Pedigree genotype	Uses multiterminal binary decision diagrams for large pedigrees	91
Arlequin	Genotype	Expectation-maximization algorithm for few markers	92
CRIMAP	Pedigree genotype	One of the first pedigree haplotyping programs	93

[*] Provides the setting in which the method was developed. [‡] Corresponds to family-based phasing with genotype data. [§] Corresponds to assembly algorithms for DNA-sequencing read data. [‖] Refers to settings involving unrelated individuals with genotype data to be phased on linkage-disequilibrium data.

which phase information is important[37]. For example, in the context of human population genomic studies, Nievergelt *et al.* demonstrated that greater differentiation of human populations can be obtained by exploring within- and across-population haplotype diversity than by focusing on multilocus genotype diversity[38]. In terms of cataloguing human genetic variation, Shendure and colleagues have shown that resolving the existence of structural variants within genomes can be enhanced greatly if phase information is considered[37]. Studies of the evolution of genomes across species can be enhanced by comparing individual chromosomes[39]. Finally, classical transplantation studies often exploit haplotype matching to determine optimal host–donor relationships[40].

Approaches for diplotyping

Given the importance of knowing the unique nucleotide content associated with each of the two homologous copies of a genomic region for assessing diploid gene function, it is important to consider how this knowledge can be obtained for any individual or group of individuals. There are several approaches for determining phase from DNA sequence and genotype data (FIG. 2). These approaches can be broadly classified in two categories. First, there are methods that leverage genotype information from individuals of either the same population or the same family as a 'target' individual whose genome is to be phased. Second, there are methods that physically separate the nucleotide content and unique variants on each homologous chromosome. Importantly, although laboratory and computational methods have the potential to phase or separate two homologous chromosomes, only methods that leverage genotype data from parental lineages can determine whether a particular phased chromosomal copy was inherited from an individual's mother or father. Knowledge of the specific parental origins of chromosome regions, rather than just the nucleotide content of chromosome homologues, may be of use in the context of parent-of-origin effects such as epigenetic imprinting, as recently demonstrated for type 2 diabetes[17].

Methods that use information from other individuals. Using information from parents or other relatives is a powerful approach to phasing an individual and has been used in many, if not most, classical family-based human genetic-mapping studies used to identify genomic regions harbouring disease-predisposing variants. Pedigree-based mapping methods such as those that calculate the logarithm of odds (LOD) or that use the transmission disequilibrium test (TDT) track, for example, the transmission of a putative disease-causing variant and a genetic marker together on a single chromosome from generation to generation. Thus, these strategies heavily depend on phase information in the genomic regions of interest. The same approach has been applied to dense genotype data generated by SNP arrays[41], as well as whole-genome sequencing (FIG. 2c); for example, in the study by Roach *et al.*[4], discussed above, in which the genomes of two siblings with different Mendelian disorders were sequenced[4]. Roach *et al.* reported that by sequencing the parents of the two target individuals, they could separate as much as 96.8% of the genome into maternally and paternally inherited chromosomal or haplotypic complements. Leveraging parental information to phase genomes provides excellent accuracy and demonstrates the added benefit that current family-based genome-sequencing studies will be able to exploit. However, for population- or case–control-based studies this strategy would entail a substantial increase in costs associated with the need to sequence the additional genomes of relatives in addition to those of the target individuals.

The use of genotype data from a larger set of unrelated individuals to phase a target individual can provide a cost-effective method for separating homologous chromosomes with respect to common variants. This approach is based on shared ancestry of the target individual and the larger set of individuals so that linkage-disequilibrium patterns between variants can be exploited in haplotyping the target individual[42,43]. However, this approach assumes the availability of genotypes from additional individuals of the same or a similar population as the target and, although the definition of 'similar' is often vague, genotype data from individuals of an appropriate population might not be available.

Population-based approaches also assume that there are reliable statistical and computational techniques available to conduct the phasing. Most population-based phasing methods (and related genotype-imputation methods[44]) can produce reliable haplotypes for moderately long stretches of a chromosome. Human genetics research has a long history of efforts to refine probabilistic phasing methods that leverage data on relatives, entire pedigrees or population linkage-disequilibrium data[45,46] (TABLE 3). However, these methods are notorious for 'switching error' inaccuracies, which arise when chromosomal segments have been phased accurately, but their connections to each other to form larger haplotypes or contigs are incorrect[47]. Deeper catalogues of genetic variation across many populations may reduce switching errors, but they might be hard to eliminate entirely owing to variation in recombination rates and the genetic diversity within and across human populations. Another problem with the population approach is that it requires the larger set of individuals to have been genotyped previously. As a consequence, these individuals may not be useful for phasing rare variants possessed by the target individual, because rare variants are not likely to have been observed (or may not even exist), and so genotyped, among the larger set of individuals. Hence, reliable linkage-disequilibrium information about those variants might not be available to facilitate phasing. Finally, the population-based phasing approach obviously could not work for private variants possessed only by the target individual. This caveat may be of increasing importance in future studies, as shifts in emphasis begin to focus on understanding rare and even *de novo* variation and its role in human diseases. In this context, private variants, or variants private to a specific population not previously studied, are unlikely to be accurately phased using data sets such as those associated with the 1000 Genomes Project, given their focus on specific populations[48].

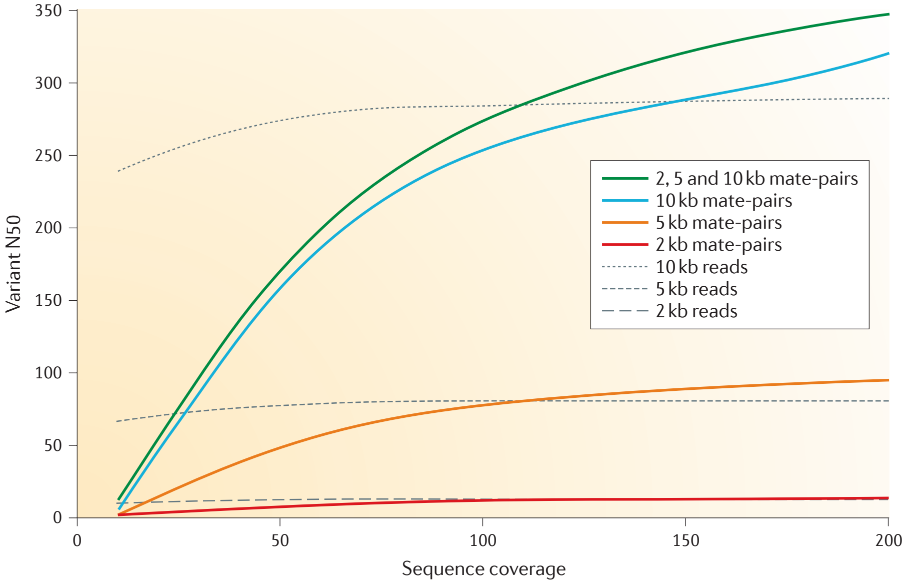

Figure 3 | **Phase reconstruction using mate-pair information.** Simulated 100 bp mate-pair read coverage of various depths (sequence (fold) coverage, *x*-axis) for chromosome 1 of a Yoruban individual. All simulations were done using SNP calls (for chromosome 1) for the Yoruban individual NA19240, obtained from the 1000 Genomes project (released December 2008). Paired-end reads were simulated with the starting position of one read, chosen consistently at random on the chromosome, and the insert length sampled from a normal distribution with a given mean insert length (2, 5 or 10 kb) and standard deviation equal to 10% of the mean. For each simulation experiment, we constructed a graph with nodes corresponding to the heterozygous SNPs and edges corresponding to reads that cover multiple variants. The N50 was calculated using the number of variants in each connected component of this graph that correspond to the phased haplotype blocks. The vN50 is defined as the point at which half of the heterozygous loci of the chromosome are contained in contigs with the vN50 or greater number of variants. Mate-pair libraries outperform reads of the same length because the size distribution of the insert consists of lengths greater than 10 kb, allowing for longer connections than are possible with single reads alone. The software used in the simulation studies is available from the Polymorphism Research Laboratory (see Further information).

Methods based on information from a single individual. The second set of phasing methods works by seeking to resolve the haplotypic arrangement of two or more neighbouring variants empirically from sequence data gathered on a single individual. Such methods provide a direct approach to phasing and can be used to phase *de novo* mutations, which, when combined with knowledge of the parental origins of variants surrounding a *de novo* mutation, can be used to assess, for example, parent-of-origin and paternal age mutation rates, something that is not feasible using other approaches[49,50]. Phasing techniques that physically separate chromosomes fall into two broad categories[51]: separation of complete chromosomes before sequencing, and reduction of the complexity of mixtures of paternally and maternally inherited DNA. Physical separation of entire chromosomes is not trivial because it involves the isolation of chromosomes from a single cell, amplification of the DNA from those isolated chromosomes, and then sequencing. The use of sophisticated microfluidic technologies has recently been applied to this process[40] and represents a substantial improvement over previous methods[52].

Complexity reduction involves the separation of genomic DNA into pools that contain DNA from regions of the genome that are either maternally or paternally derived[53]. A compelling recent example of this approach used 115 fosmid libraries to reconstruct the diploid sequence of the genome of a South Asian individual[37] (FIG. 2d). As an alternative to the use of fosmid libraries, pooled maternal and paternal DNA samples diluted to a point at which only a fraction of a complete genome is present for sequencing could be used. With the proper assessment of the dilutions, each pool will be expected to contain only a single chromosome at any particular region[54]. Cloning- and dilution-based methods for complexity reduction are straightforward and probably within the capabilities of most sequencing laboratories with standard equipment, but result only in large contigs that reflect haplotypic segments of a chromosome that still need to be stitched together to characterize an entire chromosome — a process that could be error prone.

As an alternative, phase can be reconstructed from diploid DNA from a single individual using computational approaches that link partially overlapping DNA-sequencing reads harbouring variants at heterozygous positions[55–58] (FIG. 2b). This approach requires long DNA-sequencing reads or mate-pairs of variable insert size in order to reliably capture multiple heterozygous sites that can be used to assemble reads into larger contigs on the basis of their overlapping nucleotide content[56]. This approach was used in the construction of the first diploid genome[1], although, owing to limitations in the available sequence data and the number of heterozygous positions spanned by the sequencing reads, only ~70% of the genome could be phased. Current sequencing projects that use a limited selection of short insert size, paired-read distances are not well designed for phase reconstruction. Future work should focus on improvements to mate-pair construction and projects that leverage variable insert size libraries, which, coupled with longer reads, should allow reasonably sized haploid contig assemblies (FIG. 3).

Diplomics: a new frontier?

We have emphasized why an understanding of how specific combinations of genetic variants on the two homologous copies of a chromosomal region influence diploid gene function is crucial for human genetic research. There may, however, be other phenomena that reflect the consequences of diploidy that we have not touched on here. For example, differences in the mere lengths of inherited genomes (owing to, for example, copy-number variations, repeat polymorphisms or large indels) may affect DNA packing and epigenomic phenomena. For these reasons, the science of diplomics should receive greater attention in the human genetics community in the future. However, as we have argued, diplomic enquiry requires more sophisticated sequencing and study-design strategies than those in current use. For example, better *a priori* chromosome-separation techniques are needed for human sequencing studies, as are sequencing technologies that generate longer reads to facilitate *de novo* haplotype-based assemblies. We foresee that a re-emergence of family studies will occur to help to resolve important diplomics-related issues, such as those involving complex forms of compound heterozygosity. Finally, in order to fully understand how the diplotypic genomic 'whole' functions over and above its haplotypic 'parts', we believe that more relevant functional assays, perhaps involving the simultaneous introduction of different haplotypic complements into functional assays or transgenic animals, are needed. Ultimately, if collaborative science teams are to make headway in unravelling the secrets of the human genome, especially in refining the functional and clinical effects of human genomic variation, then it makes no sense to ignore one of its most fundamental aspects: its diploid nature.

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol and Nicholas J. Schork are at The Scripps Translational Science Institute, 3344 North Torrey Pines Road, Suite 300, La Jolla, California 92037, USA.

Ali Torkamani, Eric J. Topol and Nicholas J. Schork are also at the Department of Experimental Medicine, The Scripps Research Institute, 3344 North Torrey Pines Road, Suite 300, La Jolla, California 92037, USA.

Vikas Bansal, Ali Torkamani, Eric J. Topol and Nicholas J. Schork are also at Scripps Health, 3344 North Torrey Pines Road, Suite 300, La Jolla, California 92037, USA.

Ryan Tewhey is also at the Graduate Program, Division of Biological Sciences, The University of California, San Diego, California 92093, USA.

Correspondence to N.J.S.
e-mail: nschork@scripps.edu

doi:10.1038/nrg2950
Published online 8 February 2011

1. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
2. Lifton, R. P. Individual genomes on the horizon. *N. Engl. J. Med.* **362**, 1235–1236 (2010).
3. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
4. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
5. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
6. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
7. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
8. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
9. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
10. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
11. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
12. Morozova, O., Hirst, M. & Marra, M. A. Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet.* **10**, 135–151 (2009).
13. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
14. Tucker, T., Marra, M. & Friedman, J. M. Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* **85**, 142–154 (2009).
15. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
16. Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86**, 411–419 (2010).
17. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
18. Tycko, B. Mapping allele-specific DNA methylation: a new tool for maximizing information from GWAS. *Am. J. Hum. Genet.* **86**, 109–112 (2010).
19. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136–1140 (2007).
20. Wen, G. *et al.* Both rare and common polymorphisms contribute functional variation at CHGA, a regulator of catecholamine physiology. *Am. J. Hum. Genet.* **74**, 197–207 (2004).
21. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41**, 1061–1067 (2009).
22. Wain, L. V., Armour, J. A. & Tobin, M. D. Genomic copy number variation, human health, and disease. *Lancet* **374**, 340–350 (2009).
23. Leary, R. J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl Acad. Sci. USA* **105**, 16224–16229 (2008).
24. Knudson, A. G. Two genetic hits (more or less) to cancer. *Nature Rev. Cancer* **1**, 157–162 (2001).
25. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Rev. Genet.* **11**, 415–425 (2010).
26. Zschocke, J. Dominant versus recessive: molecular mechanisms in metabolic disease. *J. Inherit. Metab. Dis.* **31**, 599–618 (2008).
27. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Rev. Genet.* **10**, 241–251 (2009).
28. Su, Z., Cardin, N., Donnelly, P., Marchini, J. & Control, W. T. C. A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association etudies. *Statistical Sci.* **24**, 430–450 (2009).
29. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *Plos Biol.* **8**, e1000294 (2010).
30. Graham, R. R. *et al.* Genetic variants near *TNFAIP3* on 6q23 are associated with systemic lupus erythematosus. *Nature Genet.* **40**, 1059–1061 (2008).
31. Musone, S. L. *et al.* Multiple polymorphisms in the *TNFAIP3* region are independently associated with systemic lupus erythematosus. *Nature Genet.* **40**, 1062–1064 (2008).
32. Graham, R. R. *et al.* A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nature Genet.* **38**, 550–555 (2006).
33. Graham, R. R. *et al.* Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl Acad. Sci. USA* **104**, 6758–6763 (2007).
34. Harley, J. B. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. *Nature Genet.* **40**, 204–210 (2008).
35. Shimane, K. *et al.* The association of a nonsynonymous single-nucleotide polymorphism in TNFAIP3 with systemic lupus erythematosus and rheumatoid arthritis in the Japanese population. *Arthritis Rheum.* **62**, 574–579 (2010).
36. Lemmers, R. J. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–1653 (2010).
37. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotech.* 19 Dec 2010 (doi:10.1038/nbt.1740).
38. Nievergelt, C. M., Libiger, O. & Schork, N. J. Generalized analysis of molecular variance. *PLoS Genet.* **3**, e51 (2007).
39. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
40. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotech.* 19 Dec 2010 (doi:10.1038/nbt.1739).
41. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40**, 1068–1075 (2008).
42. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
43. Browning, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* **124**, 439–450 (2008).
44. Biernacka, J. M. *et al.* Assessment of genotype imputation methods. *BMC Proc.* **3** Suppl. 7, S5 (2009).
45. Gao, G., Allison, D. B. & Hoeschele, I. Haplotyping methods for pedigrees. *Hum. Hered.* **67**, 248–266 (2009).
46. Salem, R. M., Wessel, J. & Schork, N. J. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum. Genomics* **2**, 39–66 (2005).
47. Andres, A. M. *et al.* Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epidemiol.* **31**, 659–671 (2007).
48. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

49. Goriely, A. & Wilkie, A. O. Missing heritability: paternal age effect mutations and selfish spermatogonia. *Nature Rev. Genet.* **11**, 589 (2010).
50. Moloney, D. M. *et al.* Exclusive paternal origin of new mutations in Apert syndrome. *Nature Genet.* **13**, 48–53 (1996).
51. Bansal, V., Tewhey, R., Topol, E. J. & Schork, N. The next phase in human genetics. *Nature Biotech.* **29**, 38–39 (2011).
52. Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nature Methods* **7**, 299–301 (2010).
53. Kouprina, N. & Larionov, V. TAR cloning: insights into gene function, long-range haplotypes and genome structure and evolution. *Nature Rev. Genet.* **7**, 805–812 (2006).
54. Paul, P. & Apgar, J. Single-molecule dilution and multiple displacement amplification for molecular haplotyping. *Biotechniques* **38**, 553–559 (2005).
55. Kim, J. H., Waterman, M. S. & Li, L. M. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res.* **17**, 1101–1110 (2007).
56. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 24, i153–159 (2008).
57. Bansal, V., Halpern, A. L., Axelrod, N. & Bafna, V. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.* **18**, 1336–1346 (2008).
58. He, D., Choi, A., Pipatsrisawat, K., Darwiche, A. & Eskin, E. Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics* **26**, i183–i190 (2010).
59. Shimizu, H. *et al.* Epidermolysis bullosa simplex associated with muscular dystrophy: phenotype-genotype correlations and review of the literature. *J. Am. Acad. Dermatol.* **41**, 950–956 (1999).
60. Fong, C. Y., Mumford, A. D., Likeman, M. J. & Jardine, P. E. Cerebral palsy in siblings caused by compound heterozygous mutations in the gene encoding protein C. *Dev. Med. Child. Neurol.* **52**, 489–493 (2010).
61. McLaughlin, H. M. *et al.* Compound heterozygosity for loss-of-function lysyl-tRNA synthetase mutations in a patient with peripheral neuropathy. *Am. J. Hum. Genet.* **87**, 560–566 (2010).
62. Welch, K. O., Marin, R. S., Pandya, A. & Arnos, K. S. Compound heterozygosity for dominant and recessive GJB2 mutations: effect on phenotype and review of the literature. *Am. J. Med. Genet. A* **143A**, 1567–1573 (2007).
63. Aguilar Martinez, P. *et al.* Compound heterozygotes for hemochromatosis gene mutations: may they help to understand the pathophysiology of the disease? *Blood Cells Mol. Dis.* **23**, 269–276 (1997).
64. Nakamura, A., Yazaki, M., Tokuda, T., Hattori, T. & Ikeda, S. A Japanese patient with familial Mediterranean fever associated with compound heterozygosity for pyrin variant E148Q/M694I. *Intern. Med.* **44**, 261–265 (2005).
65. Majumdar, S. *et al.* Compound heterozygous mutation with a novel splice donor region DNA sequence variant in the succinate dehydrogenase subunit B gene in malignant paraganglioma. *Pediatr. Blood Cancer* **54**, 473–475 (2010).
66. Avigad, S. *et al.* Compound heterozygosity in nonphenylketonuria hyperphenylalanemia: the contribution of mutations for classical phenylketonuria. *Am. J. Hum. Genet.* **49**, 393–399 (1991).
67. Moon, S. *et al.* Novel compound heterozygous mutations in the fructose-1,6-bisphosphatase gene cause hypoglycemia and lactic acidosis. *Metabolism* **60**, 107–113 (2011).
68. Dork, T., Bendix-Waltes, R., Wegner, R. D. & Stumm, M. Slow progression of ataxia-telangiectasia with double missense and in frame splice mutations. *Am. J. Med. Genet. A* **126A**, 272–277 (2004).
69. Maimaiti, M. *et al.* Silent exonic mutation in the acid-α-glycosidase gene that causes glycogen storage disease type II by affecting mRNA splicing. *J. Hum. Genet.* **54**, 493–496 (2009).
70. Miyake, A. *et al.* A compound heterozygote of novel and recurrent *DTDST* mutations results in a novel intermediate phenotype of Desbuquois dysplasia, diastrophic dysplasia, and recessive form of multiple epiphyseal dysplasia. *J. Hum. Genet.* **53**, 764–768 (2008).
71. De Rosa, M. *et al.* Evidence for a recessive inheritance of Turcot's syndrome caused by compound heterozygous mutations within the *PMS2* gene. *Oncogene* **19**, 1719–1723 (2000).
72. Drysdale, C. M. *et al.* Complex promoter and coding region β_2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* **97**, 10483–10488 (2000).
73. Horan, M. *et al.* Human growth hormone 1 (GH1) gene expression: complex haplotype-dependent influence of polymorphic variation in the proximal promoter and locus control region. *Hum. Mutat.* **21**, 408–423 (2003).
74. Barroso, E. *et al.* *FANCD2* associated with sporadic breast cancer risk. *Carcinogenesis* **27**, 1930–1937 (2006).
75. Chen, H. *et al.* Single nucleotide polymorphisms in the human interleukin-1B gene affect transcription according to haplotype context. *Hum. Mol. Genet.* **15**, 519–529 (2006).
76. Weyrich, P. *et al.* Role of AMP-activated protein kinase gamma 3 genetic variability in glucose and lipid metabolism in non-diabetic whites. *Diabetologia* **50**, 2097–2106 (2007).
77. Yang, H. *et al.* ATM sequence variants associate with susceptibility to non-small cell lung cancer. *Int. J. Cancer* **121**, 2254–2259 (2007).
78. Maggini, V. *et al.* MDR1 diplotypes as prognostic markers in multiple myeloma. *Pharmacogenet. Genomics* **18**, 383–389 (2008).
79. Pickard, B. S. *et al.* Interacting haplotypes at the *NPAS3* locus alter risk of schizophrenia and bipolar disorder. *Mol. Psychiatry* **14**, 874–884 (2009).
80. Sun, H. *et al.* The association of adiponectin allele 45T/G and -11377C/G polymorphisms with type 2 diabetes and rosiglitazone response in Chinese patients. *Br. J. Clin. Pharmacol.* **65**, 917–926 (2008).
81. Williams, A. L., Housman, D. E., Rinard, M. C. & Gifford, D. K. Rapid haplotype inference for nuclear families. *Genome Biol.* **11**, R108 (2010).
82. Jiang, H. T., Xu, Y., Zhao, Y. Z. & Chen, G. L. A novel algorithm for minimum recombinant haplotyping on pedigrees by zero recombinant block partition. *Interdiscip. Sci.* **2**, 185–192 (2010).
83. Delaneau, O., Coulonges, C. & Zagury, J. F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**, 540 (2008).
84. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
85. Eronen, L., Geerts, F. & Toivonen, H. HaploRec: efficient and accurate large-scale reconstruction of haplotypes. *BMC Bioinformatics* **7**, 542 (2006).
86. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
87. Halperin, E. & Eskin, E. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics* **20**, 1842–1849 (2004).
88. Qin, Z. S., Niu, T. & Liu, J. S. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71**, 1242–1247 (2002).
89. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30**, 97–101 (2002).
90. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
91. Gudbjartsson, D. F., Thorvaldsson, T., Kong, A., Gunnarsson, G. & Ingolfsdottir, A. Allegro version 2. *Nature Genet.* **37**, 1015–1016 (2005).
92. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
93. Lander, E. S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA* **84**, 2363–2367 (1987).

Acknowledgements

This work was supported, in part, by the following research grants: U19 AG023122-01, R01 MH078151-01A1,N01 MH22005, U01 DA024417-01, P50 MH081755-01 and UL1 RR025774, as well as the Price Foundation and Scripps Genomic Medicine. This work is the authors' sole responsibility and does not necessarily represent funding agencies' views.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Nicholas J. Schork's homepage: http://www.stsiweb.org/
Polymorphism Research Laboratory: http://polymorphism.scripps.edu/~vbansal/software/HASH/SimulationCode/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF