Chapter 12

Analysis of Haplotype Sequences

Sally S. Lloyd, Edward J. Steele and Roger L. Dawkins

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61794

Abstract

In this era of whole-genome, next-generation sequencing, it is important to have a clear understanding of the concept of "haplotype". We show here that most of the important regions of the genome can be described in terms of polymorphic frozen blocks (PFB). At each PFB, there are numerous, even hundreds, of alternative ancestral haplotypes. Haplotypes, not genes, can be regarded as the principal unit of inheritance. We illustrate how sequence data can be analysed to reveal and define these ancestral haplotypes.

Keywords: Ancestral haplotypes, Polymorphic frozen blocks, Genomic evolution

1. Introduction

Comparative analyses of haplotype sequences allow many efficiencies. It is not surprising that there are many enthusiastic claims. Haplotypes, by any of many definitions, offer opportunities to understand the inheritance of polymorphic traits and their regulation. The most useful are markers of extensive complex polymorphic sequences of evolutionary significance even when the functional components, whether coding or noncoding, are yet to be elaborated.

Substantial advances became possible with the elucidation of genomic structure and function more than 20 years ago and long before recent advances in sequencing technology [1] and bioinformatics [2]. It became clear that haplotypes, *not genes*, can be regarded as the principal unit of inheritance.

This chapter evaluates some competing strategies and illustrates the power now available through NGS.

2. Haplotype terminology

A review of current literature reveals a staggering collection of terms synonymous with haplotypes, as listed in Table 1.

Ancestral haplotypes
Conserved extended haplotypes
Linkage groups
Linkage disequilibrium haplotypes
Hapmaps
Haplogroup
Haplobanks
Haploblocks
Haplotype block

Table 1. Terminology

Even if it were possible to define the various neologisms, it seems certain that confusion will remain until there is recognition of the conceptual background.

We introduced the term *ancestral haplotypes* to emphasise the persistence of the founding pool [3, 4]. Such haplotypes are conserved over thousands of generations; they allow identification of remote ancestors and their contributions to the creation of individual members of the species with their diseases. Unfortunately, others use the same term in different ways and even in the opposite sense, that is, to refer to *the single* original haplotype which is presumed to have mutated to give rise to all the so-called variants now present. Indeed, as just one example of the problem, the reader has to be able to interpret the following: "we identified all nonredundant haplotypes with a frequency of ≥10% and consisting of at least 10 SNPs, which are likely to represent the nonrecombinant descendants from a single ancestor" [5].

To yet further confound matters, increasingly, the term *haplotype* is being used to describe any combination of alleles or markers, such as SNPs, without regard to their reproducibility, inheritance, polymorphism or biological significance. Currently, there are conflicting methods of detection. The problems appear to be increasing as ephemeral concepts diverge and as claims for better approaches focus on just one or another competing technology or bioinformatic package.

Several other aspects are clear.

- Linkage groups relate to closely linked loci but do not define haplotypes.
- Linkage disequilibrium is affected by relative frequencies and therefore fails to detect rare haplotypes.
- Trios can be misleading since the coverage of the family is limited.
- Haplobanks. The Tokunaga group has established some important principles with the intention of establishing haplotype-matched pluripotential stem cell banks [6]. Unfortu-

nately, and amazingly, there is now uncertainty as to how to define the haplotypes. For example, a recent paper urges international collaboration to avoid fragmentation [7]. It would be wise to avoid neologisms and such redefinitions without clarity of meaning.

3. Definitions and concepts

In the presequencing era, there was a clear understanding of what was meant by the term *haplotype*: Combinations of alleles at different loci segregating together in multigenerational family studies [8]. Some seem unaware of this long history and have had to rediscover the concept [2].

The implications were apparent at least 50 years ago: a specific allele A1 at locus A is inherited together with a specific allele B1 at an adjacent, “closely linked” locus B [9]. The fact that these two alleles segregated together through multiple generations was unexpected and lead to controversy but, in retrospect, clearly implied that

1. The two alleles were encoded on the same chromosome, whether paternal or maternal.
2. The two loci were closely linked.
3. Recombination was rare.
4. The two loci arose by duplication.
5. Duplication is associated with polymorphism.

The repeated cosegregation of alleles came to be known as a haplotype: from άπλφούς = single [9].

It is worth emphasizing that it was the cosegregation as haplotypes through “phased” multigenerational families (rather than “unphased” populations) which foretold the later demonstration that there was a continuous haplospecific sequence. It is also pertinent, with the benefit of hindsight and in view of recent confusion, that the haplotypes, defined in one family, occurred in other families of similar remote ancestry raising the radical possibility of conservation beyond that expected from close linkage alone. In other words, recombination is patchy and does not necessarily disperse the components of duplications, even after thousands of meioses. The issue of linkage disequilibrium and the limits of LD mapping are considered below.

The implications of haplotypes, as listed above, became even clearer as the HLA A and HLA B locus alleles and then HLA DR alleles were defined during the 1970s. However, in this case, the loci were widely separated. Over time, it became clear that each of the A-B and B-DR haplotypes were some 800 kb in length. Patently, close linkage could not explain these haplotypes; either there was selection for *cis* interaction or there was suppression of recombination [3, 4].

Through their studies of diseases, the Alper–Yunis group discovered that the B-DR haplotypes contained specific alleles at duplicated loci which had no structural or functional relevance to HLA (i.e. complement and 21 hydroxylase loci) but which happen to be located within the major histocompatibility complex [10–16]. Thus, *cis* interaction alone could be rejected as the sole explanation.

The importance of discovery through disease was illustrated at a meeting held in 1982 [3, 4]. As shown in Table 2, it was disease associations which allowed the initial discovery of ancestral haplotypes; note, these three disease-associated haplotypes could have only been discovered through their associations. Two share DR3 and two share B18 but the frequencies differ. Thus, the three haplotypes cannot be detected by linkage disequilibrium.

Designation	A	Cw	B	Bf	C2	C4A	C4B	DR	Disease
8.1	1	7	8	S	C	Q0	1	3	MG, SLE, IDDM
18.2	–	–	18	F1	C	3	Q0	3	IDDM
18.1	25	–	18	S	Q0	4	2	2	C2 deficiency

MG = myasthenia gravis, SLE = systemic lupus erythematosus, IDDM = insulin-dependent (type 1) diabetes mellitus.

Adapted from ref. [4]

Table 2. MHC haplotypes and disease associations

Once the numerous other ancestral haplotypes were defined, multigenerational family studies identified cosegregating combinations of multiple alleles at separated loci, i.e. haplotypes stretching over nearly 2 Mb from HLA A to DR. A haplotype was defined by the alleles “inherited *en bloc* from one parent and implies the transmission of all of the chromosomal segment” from one generation to the next [4].

When haplotypes defined in one family were compared with those identified in apparently unrelated families, sharing was immediately apparent. There were specific combinations of alleles at all the numerous unrelated loci as these were defined and typed. However, and increasingly relevant today, as summarized in refs. [3, 4, 17, 18]:

1. The combinations observed are *not* a simple function of allele frequencies; only some of the components inherited *en bloc* are in linkage disequilibrium.

2. Many haplotypes are rare combinations of frequent alleles at some loci but rare alleles at other loci.

3. Very few alleles are entirely haplospecific.

4. Haplotype frequencies are often less than 1%.

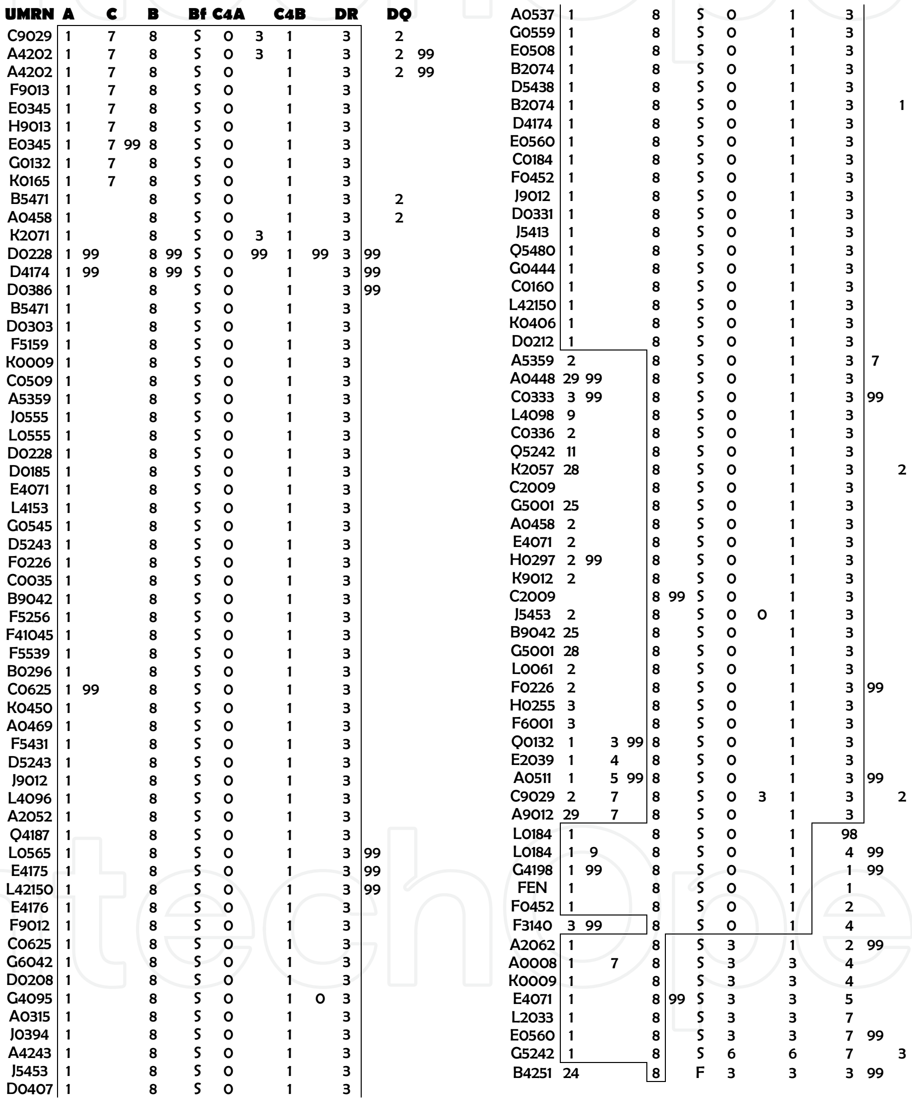
5. The same haplotypes are found in multiple, apparently unrelated, families.

6. Many of these nonrandom combinations are associated with a disease (such as systemic lupus erythematosus) or function (such as TNF production).

7. With a few dramatic exceptions (such as 21 hydroxylase and C2 deficiency carried by what we now call the 47.1 and 18.1 ancestral haplotypes), the individual alleles do not explain the haplospecific effects on disease and function.

8. Penetrance is low. That is to say, the haplotypes are *sine qua non* in that they permit particular diseases and functions but only in the presence of other genetic, infectious, environmental, hormonal and age-related factors.

9. Recombination is rare and difficult to demonstrate even within multigenerational families with the potential to confirm a meiotic recombinant. Nevertheless, over the life of an ancestral haplotype—say 10, 000 meioses—there have been recombinations which have resulted in shuffling between ancestral haplotypes [18, 19].



UMRN	A		C		B		Bf	C4A		C4B		DR		DQ	
C9029	1		7		8		S	O	3	1		3		2	
A4202	1		7		8		S	O	3	1		3		2	99
A4202	1		7		8		S	O		1		3		2	99
F9013	1		7		8		S	O		1		3			
E0345	1		7		8		S	O		1		3			
H9013	1		7		8		S	O		1		3			
E0345	1		7	99	8		S	O		1		3			
G0132	1		7		8		S	O		1		3			
K0165	1		7		8		S	O		1		3			
B5471	1				8		S	O		1		3		2	
A0458	1				8		S	O		1		3		2	
K2071	1				8		S	O	3	1		3			
D0228	1	99			8	99	S	O	99	1	99	3	99		
D4174	1	99			8	99	S	O		1		3	99		
D0386	1				8		S	O		1		3	99		
B5471	1				8		S	O		1		3			
D0303	1				8		S	O		1		3			
F5159	1				8		S	O		1		3			
K0009	1				8		S	O		1		3			
C0509	1				8		S	O		1		3			
A5359	1				8		S	O		1		3			
J0555	1				8		S	O		1		3			
L0555	1				8		S	O		1		3			
D0228	1				8		S	O		1		3			
D0185	1				8		S	O		1		3			
E4071	1				8		S	O		1		3			
L4153	1				8		S	O		1		3			
G0545	1				8		S	O		1		3			
D5243	1				8		S	O		1		3			
F0226	1				8		S	O		1		3			
C0035	1				8		S	O		1		3			
B9042	1				8		S	O		1		3			
F5256	1				8		S	O		1		3			
F41045	1				8		S	O		1		3			
F5539	1				8		S	O		1		3			
B0296	1				8		S	O		1		3			
C0625	1	99			8		S	O		1		3			
K0450	1				8		S	O		1		3			
A0469	1				8		S	O		1		3			
F5431	1				8		S	O		1		3			
D5243	1				8		S	O		1		3			
J9012	1				8		S	O		1		3			
L4096	1				8		S	O		1		3			
A2052	1				8		S	O		1		3			
Q4187	1				8		S	O		1		3			
L0565	1				8		S	O		1		3	99		
E4175	1				8		S	O		1		3	99		
L42150	1				8		S	O		1		3	99		
E4176	1				8		S	O		1		3			
F9012	1				8		S	O		1		3			
C0625	1				8		S	O		1		3			
G6042	1				8		S	O		1		3			
D0208	1				8		S	O		1		3			
G4095	1				8		S	O		1	O	3			
A0315	1				8		S	O		1		3			
J0394	1				8		S	O		1		3			
A4243	1				8		S	O		1		3			
J5453	1				8		S	O		1		3			
D0407	1				8		S	O		1		3			
A0537	1				8		S	O		1		3			
G0559	1				8		S	O		1		3			
E0508	1				8		S	O		1		3			
B2074	1				8		S	O		1		3			
D5438	1				8		S	O		1		3			
B2074	1				8		S	O		1		3		1	
D4174	1				8		S	O		1		3			
E0560	1				8		S	O		1		3			
C0184	1				8		S	O		1		3			
F0452	1				8		S	O		1		3			
J9012	1				8		S	O		1		3			
D0331	1				8		S	O		1		3			
J5413	1				8		S	O		1		3			
Q5480	1				8		S	O		1		3			
G0444	1				8		S	O		1		3			
C0160	1				8		S	O		1		3			
L42150	1				8		S	O		1		3			
K0406	1				8		S	O		1		3			
D0212	1				8		S	O		1		3			
A5359	2				8		S	O		1		3	7		
A0448	29	99			8		S	O		1		3			
C0333	3	99			8		S	O		1		3	99		
L4098	9				8		S	O		1		3			
C0336	2				8		S	O		1		3			
Q5242	11				8		S	O		1		3			
K2057	28				8		S	O		1		3		2	
C2009					8		S	O		1		3			
G5001	25				8		S	O		1		3			
A0458	2				8		S	O		1		3			
E4071	2				8		S	O		1		3			
H0297	2	99			8		S	O		1		3			
K9012	2				8		S	O		1		3			
C2009					8	99	S	O		1		3			
J5453	2				8		S	O	O	1		3			
B9042	25				8		S	O		1		3			
G5001	28				8		S	O		1		3			
L0061	2				8		S	O		1		3			
F0226	2				8		S	O		1		3	99		
H0255	3				8		S	O		1		3			
F6001	3				8		S	O		1		3			
Q0132	1		3	99	8		S	O		1		3			
E2039	1		4		8		S	O		1		3			
A0511	1		5	99	8		S	O		1		3	99		
C9029	2		7		8		S	O	3	1		3		2	
A9012	29		7		8		S	O		1		3			
L0184	1				8		S	O		1		98			
L0184	1	9			8		S	O		1		4	99		
G4198	1	99			8		S	O		1		1	99		
FEN	1				8		S	O		1		1			
F0452	1				8		S	O		1		2			
F3140	3	99			8		S	O		1		4			
A2062	1				8		S	3		1		2	99		
A0008	1		7		8		S	3		3		4			
K0009	1				8		S	3		3		4			
E4071	1				8	99	S	3		3		5			
L2033	1				8		S	3		3		7			
E0560	1				8		S	3		3		7	99		
G5242	1				8		S	6		6		7		3	
B4251	24				8		F	3		3		3	99		

Adapted from ref. [18].

Figure 1. Historic recombinations of AH 8.1. The HLA-B8 allele is carried by one ancestral haplotype marked by A1, Cw7, B8, BfS, C4AQ0, C4B1, DR3. All the haplotypes in data set 1 carrying HLA-B8 are represented. These haplotypes have been sorted so that haplotypes that carry all alleles of 8.1 from HLA-A to DR are shown at the top of the figure, followed by haplotypes that extend from HLA-B to DR. Telomeric recombinants are shown at the bottom. The boxed areas represent those portions of the 8.1 ancestral haplotype that are carried by unrelated B8-containing haplotypes. Vertical lines approximately indicate the region where historical recombination has occurred.

Some of these points are illustrated in Figure 1. It can be seen that subjects with B8 can be listed to show conservation but also historic recombinations between HLA A and B, between C4B and DR, and between HLA B and Bf.

By the mid-1990s, and long before the rediscoveries of the 2000s [2], such analyses led to the conclusion that there are polymorphic frozen blocks (PFB), as illustrated in Figure 2.

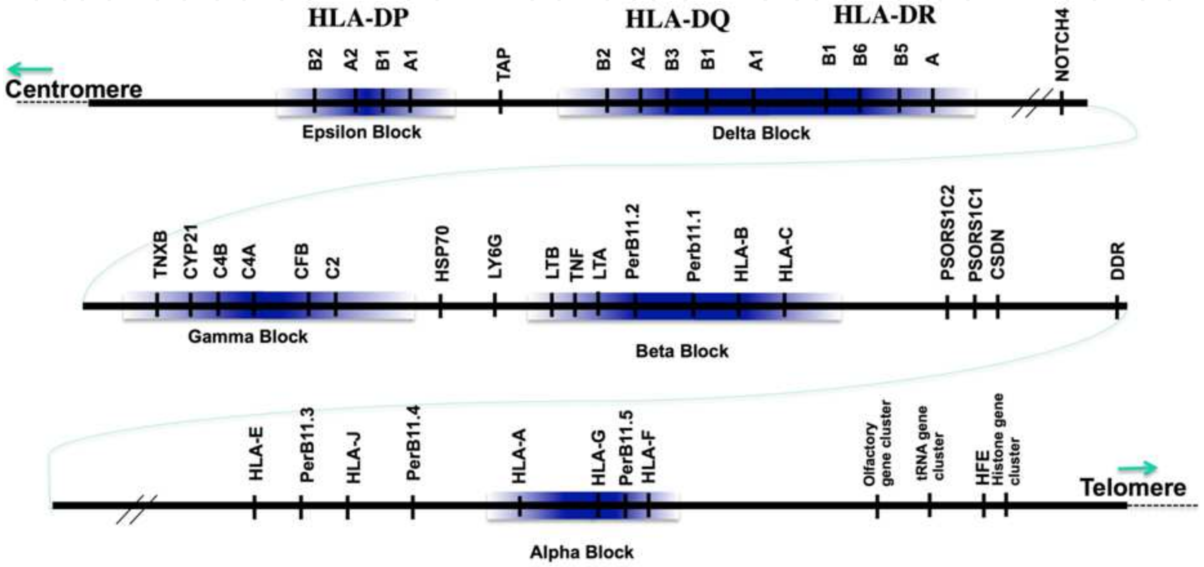
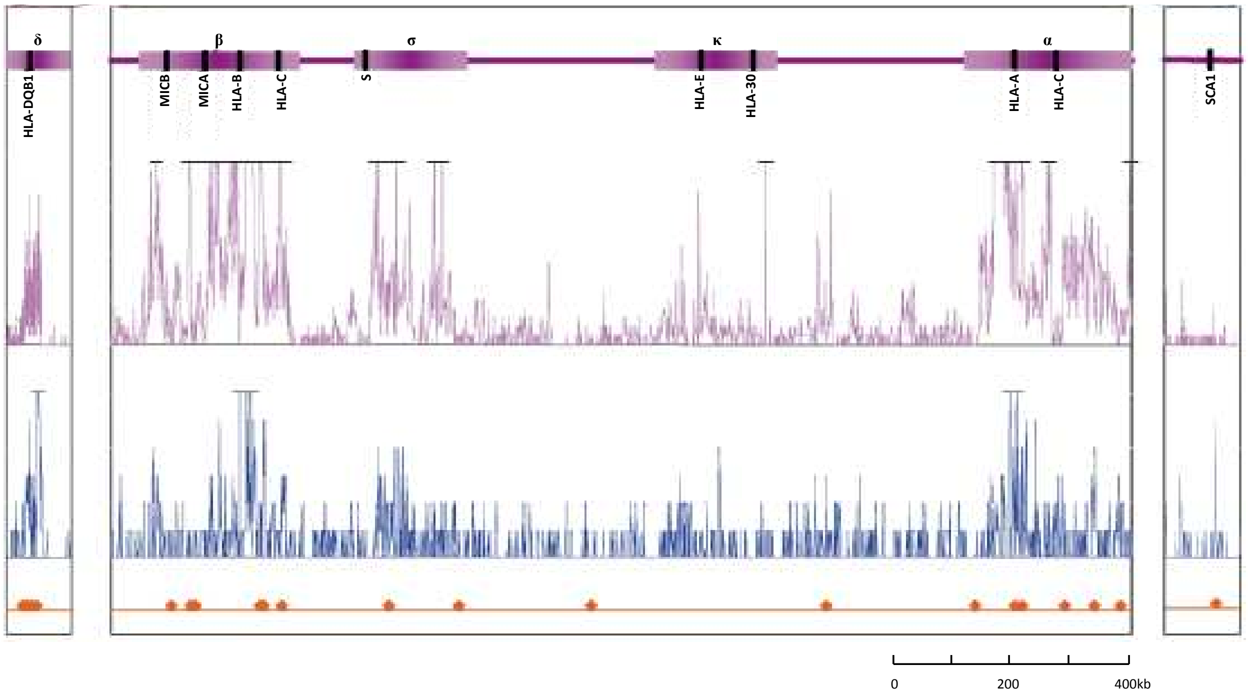


Figure 2. Ancestral haplotypes and polymorphic frozen blocks within the human major histocompatibility complex. Each ancestral haplotype has its own unique DNA sequence which includes single nucleotide polymorphisms (SNPs), copy number variations, segmental duplications, insertion and deletion events (indels) including retroviral and retroviral-like elements (RLEs). The full length is approximately 4 Mb. Higher degrees of diversity indicated by shading define polymorphic frozen blocks (PFB). Recombination occurs far more frequently between, rather than within, these blocks. Mutations within blocks are effectively suppressed. Adapted from refs. [17, 20] and [21]. Reproduced with permission from ref. [22].

PFB throughout the genome are the latter-day equivalents of loci. Sequences which define ancestral haplotypes are the equivalent of alleles. The diversity is multifactorial with contributions from reiterative speciation as follows [17]:

- Retroviral integration
- Duplication
- Indels
- Polymorphism

These elements all contribute to the haplospecificity of the sequence of ancestral haplotypes as shown in Figure 3. Similar distribution of diversity has been found by many others [5, 17, 19, 20, 23, 24]. The same patterns are also found in primates [25].



Adapted from ref. [26].

Figure 3. Sequence diversity is packaged as polymorphic frozen blocks (PFB). SNPs and indel occur in similar locations within PFB. (a) The SNP profile after removing indels. Peaks higher than 20 SNPs per 1000 nucleotides are truncated. (b) The location of indels. Peaks higher than six indels per 1000 nucleotides are truncated. (c) The position of indels greater than 100 nucleotides.

4. Use of ancestral haplotypes

Here, we illustrate the potential of sequence analysis, if designed to identify conserved, extended, ancestral haplotypes. The utility depends very largely on the concept behind the analysis. However, it also depends upon the genomic region actually sequenced and whether it is possible to interpret the patterns in the context of the heterogeneous architecture of the genome. Within PFB, there will be a multitude of alternative sequences to compare. In the genome between these blocks, there is much less diversity with long stretches of monomorphic sequence. Thus, the recent fashion for identifying homozygosity [27, 28], without regard to diversity, shifts the focus to less informative regions of the genome. Of course, by way of explanation for the fashion, homozygosity within PFB is much more difficult to find; the most common ancestral haplotypes with frequencies of 0.1 will be homozygous in only 1% of the general population. Until high-throughput NGS became available, it was necessary to examine disease panels or consanguineous families.

The conceptual background is summarised in the following figures which contrast two approaches. *Population genetics* teaches that free recombination effectively prevents the packaging of polymorphism. The reality, designated here as *quantal genomics*, emphasises clustering and conservation of polymorphism. Each haplotype is a specific sequence which regulates expressed genes by *cis*, *trans* or *epistatic* interaction. The whole sequence is conserved. Linkage disequilibrium, when it occurs, is simply a reflection of this conservation which includes haplotypes with alleles which are relatively common in one haplotype when compared with others. Each is ancestral, in the sense that they are shared by apparently unrelated families separated by hundreds or even thousands of generations. It follows that the polymorphisms are actively conserved and could not be a consequence of recent mutation.

Some of the implications are illustrated in Figures 4 and 5.

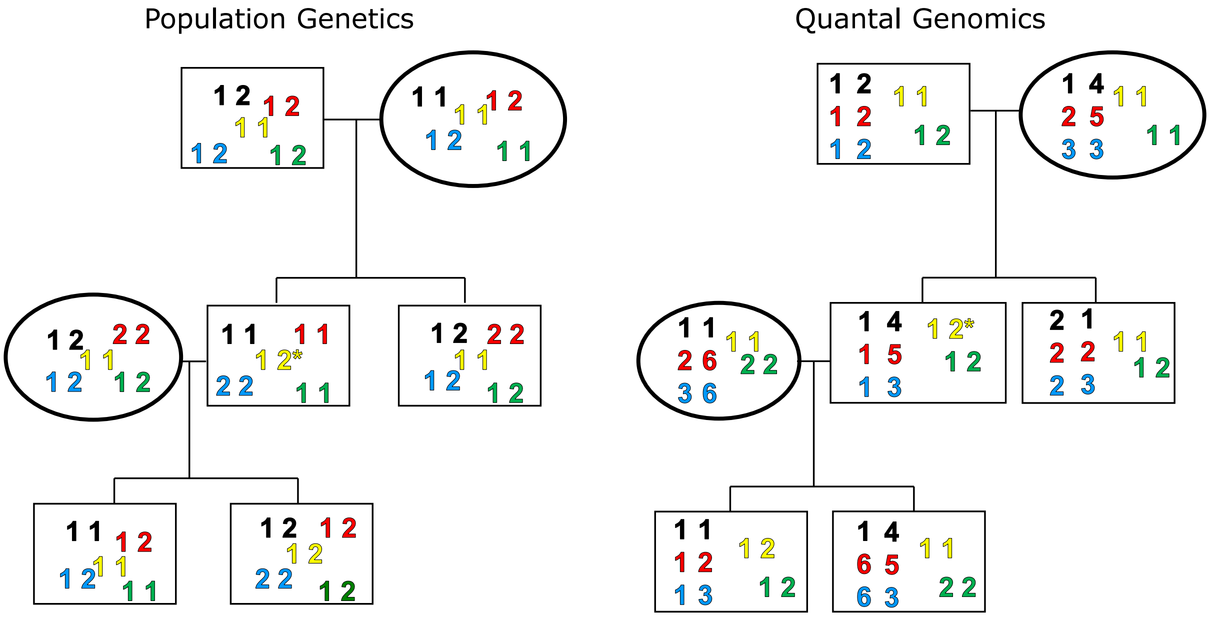


Figure 4. Importance of clustering functional genes. Colours represent loci and numbers represent alleles at those loci. On the left is the basis of the infinitesimal model used in population genetics. Loci are biallelic and can be homozygous or heterozygous. Free recombination occurs between loci and alleles segregate independently. On the right, loci are within polymorphic frozen blocks (PFB), shown by alignment of loci. Alleles within PFB segregate *en bloc*, forming haplotypes, which are inherited intact through many generations. Important genes are carried within PFB, conserving their *cis* interactions. Loci within PFB have multiple alleles, allowing for a greater degree of polymorphism clustered within the block. There can be hundreds of ancestral haplotypes for each PFB. *Trans* interactions between haplotypes increase the diversity expressed in the population. The loci shown in green and yellow are outside the PFB and follow a pattern of inheritance similar to population genetics. *De novo* mutations are indicated by asterisk—on the right the mutations occur at loci outside of conserved PFB and will have little if any consequence because truly important differences are encoded within PFB. Monogenic diseases or traits are the partial exceptions. On the left, mutations can occur at any loci but are generally assumed to occur at loci that were monoallelic. They may or may not be important, depending upon frequency, context, repair and heritability. Adapted with permission from ref. [22].

By 1987, it was clearly established that each ancestral haplotype has a specific content of genomic features such as duplications and indels. These too are actively conserved and can themselves be used as signatures for haplotypes of hundreds of kilobases and even megabases. These observations were very difficult to explain in terms of any form of neo-Darwinism, natural selection, random errors or population genetics as taught then and today. Rather, we realised, the genome is not actually homogeneous but partitioned into protected quanta or PFB [17, 22, 26, 29].

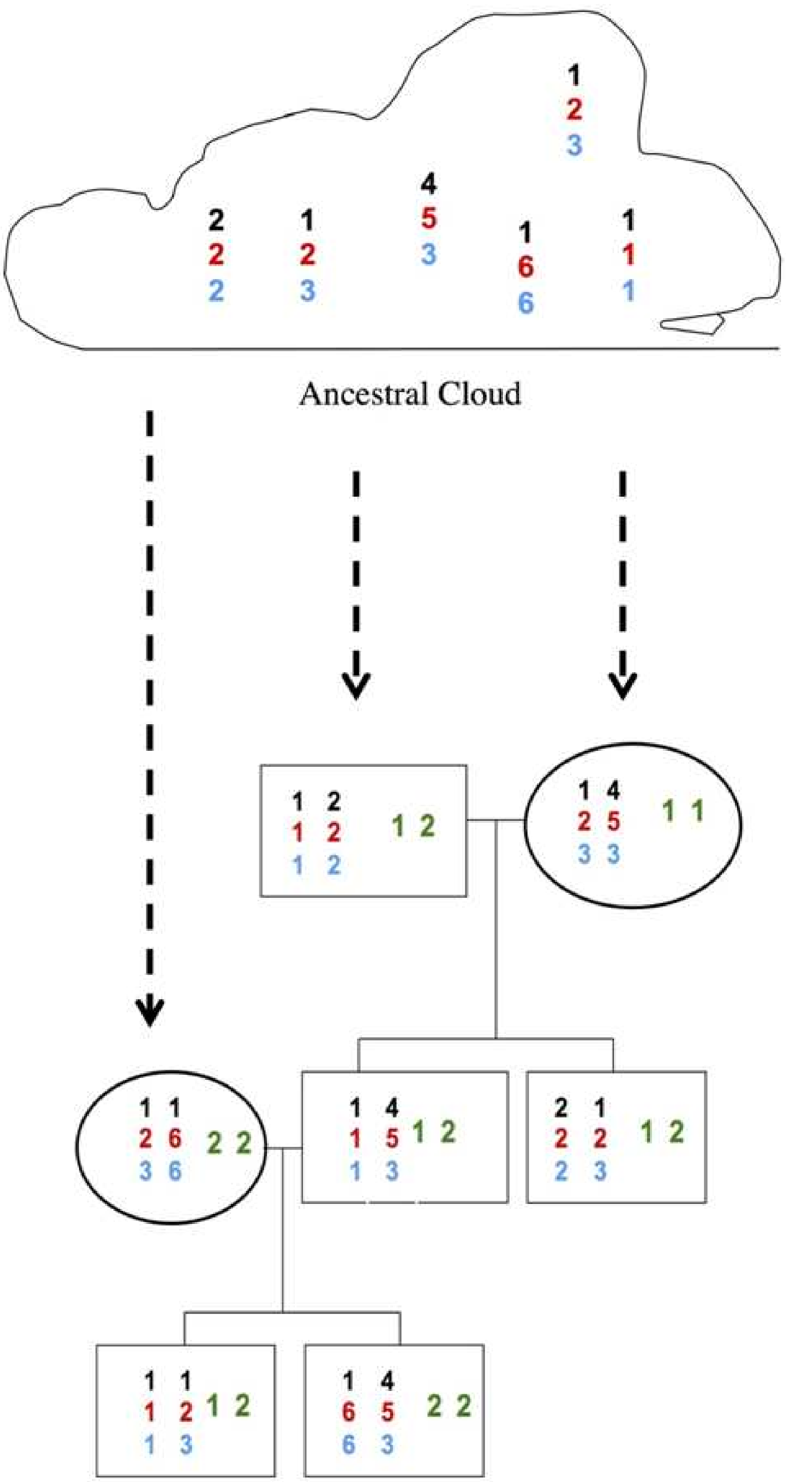


Figure 5. Modern haplotypes are derived from the deep past—they are ancestral haplotypes.

5. Sequencing of critical genomic regions

By 1992, there was sufficient sequencing to confirm the earlier prediction that each ancestral haplotype is actually a frozen sequence.

Haplotype	Geometric element at CL1	Length	Geometric element at CL2	Length
57.1	$(TC)^{12}(TG)^{6}(TC)^{14}(TG)^{3}(TC)^{12}$	94	TA $(TC)^{18}$ TT $(TC)^{9}$	58
18.2	$(TC)^{14}$	28	Deleted	
8.1	$(TC)^{28}$	56	$(TC)^{15}$ TG $(TC)^{6}$ TG $(TC)^{8}$ TG $(TC)^{5}$	96
7.1	$(TC)^{12}(TG)^{6}(TC)^{14}(TG)^{3}(TC)^{12}$	94	$(TC)^{14}$ TG $(TC)^{6}$ TG $(TC)^{8}$ TG $(TC)^{5}$	94

Adapted from ref. [30].

Table 3. Haplospecific geometric elements. Ancestral haplotypes have specific sequence signatures at each of the duplicons. Note in 18.2, the duplication did not occur or has been deleted.

We now know that examples of the 8.1 ancestral haplotype are almost identical over megabases [31, 32].

We illustrate the differences between different haplotype sequences in Figure 6. It can be seen that there are certain sites where haplotypes differ. Importantly, haplospecificity is conferred by the whole sequence rather than single nucleotide polymorphisms. For example, reading from left to right, 8.1 and 18.2 differ in T/G but not A/G, etc. Note also that some of the differences are due to indels. Of critical importance is accurate, unmolested sequencing over kilobases, as is now possible through NGS. It is clear, however, that assembly is hazardous especially in areas of duplication and polymorphism. Note also, that there is no justification for regarding *one* particular sequence as the reference. Rather, it is necessary to compare each output with a library of known sequences within each PFB.

The number of differences depends on which haplotypes are compared (see Table 4). Two of the most common Caucasian haplotypes, 8.1 and 7.1, differ by a hundred positions, representing approximately 1% nucleotide diversity. The most different haplotypes are 18.2 and 7.1, having 2.5% nucleotide diversity. Interestingly, these haplotypes are different functionally; 18.2 permits insulin-dependent diabetes mellitus whereas 7.1 is protective.

AH Haplotype	44.2	62.1	7.1	44.1*	8.1
44.2	0				
62.1	187	0			
7.1	249	221	0		
44.1*	73	154	227	0	
8.1	224	219	101	204	0
18.2*	184	130	250	137	245

Table 4. Pairwise differences between haplotypes. Total differences between each pair of haplotypes in the 9277 bp region at HLA-B.

```
44.2    60 C A G A T G 39 A G G A C C A G G G T G 16 T G G T G T 14 A G A G G C A G 11 A G G G A 29 T C T T G G 17 T G G T T C T G T G G C C 29 C A - A T A T A C A A C T T T A T G 13 A A C T T G 41
62.1    // . . . G . . // . . C . . . . . . A . . // . . C . . . // . . G . . . . . // . . . . . // . . . . . . // . . . . . . . . . . . . . // . . - . . . . . . . . . . . . . . . // . . T . . . //
7.1     // . . T G . . // . . . . . . . . . . . . // . . . C . . // . . G . . . . . // . . . . . // . . C C . . // . . C . . . . . . A . . . // . . G . . . . . . G . . . . . C . . // . . T G . . //
44.1*   // . . . G . . // . . C . . . . . . A . . // . . C . . . // . . G . . . . . // . . A . . // . . . . . . // . . . . . . . . . . . . . // . . - . . . . . . . . . . . . . . . // . . T . . . //
8.1     // . . T G . . // . . . . . . . . . . . . // . . . C . . // . . G . T . . . // . . . . . // . . C C . . // . . C . . . . . . A . . . // . . G . . . . . . G . . . . . . . . // . . T G . . //
18.2*   // . . . G . . // . . C . . . . . . A . . // . . C . . . // . . G . . . . . // . . A . . // . . . . . . // . . . . . . . . . . . . . // . . - . . . . . . . . . . . . . . . // . . T . . . //

44.2    T G T G G G C - - G T 55 C T G C C 62 C A T C C 31 T T A A A C A G A G T 55 G G T T C C C A A T C T 9 A T A G G 6 G G C G G 31 A A C C C 9 A G T A T T A T G C 37 G G T G A 12 G A T G T G
62.1    . . C . . . . G A . . // . . A . . // . . G . . // . . C . . . . . G . . // . . C . . . . . . . . . // . . . . . // . . T . . // . . . . . // . . A . . . . . . . // . . . . . // . . C . . .
7.1     . . C . . . . G A . . // . . A . . // . . G . . // . . . . . . . . G . . // . . C . . . . . . . . . // . . . . . // . . . . . // . . . . . // . . . . . . . . . . // . . C . . // . . C A . .
44.1*   . . . . . . . G A . . // . . . . . // . . G . . // . . C . . . . . G . . // . . C . . . . . T - . . // . . . . . // . . . . . // . . . . . // . . . . . . . C . . // . . . . . // . . C . . .
8.1     . . C . . . T G A . . // . . A . . // . . . . . // . . . . . . . . G . . // . . . . . . . . . . . . // . . G . . // . . . . . // . . T . . // . . . . . . . . . . // . . . . . // . . C A . .
18.2*   . . . . . . . G A . . // . . . . . // . . G . . // . . C . . . . . G . . // . . C . . . . . T - . . // . . . . . // . . . . . // . . . . . // . . . . . . . C . . // . . . . . // . . C . . .

44.2    9 G G T G C T C 42 C A G C C 25 T C G A C 9 C C A G G 17 A G A T G 16 G G G T T 21 A C G T G G G G C C 25 T G G T C 19 T G G A G G T 36 T C A G T 7 T T T G T G C C T C A T C C G T G C T T G
62.1    // . . . . T . . // . . . . . // . . A . . // . . . . . // . . . . . // . . T . . // . . A . . . . . . . // . . A . . // . . . . . . . // . . . . . // . . . . . . . . . . . . . . A . . . . . .
7.1     // . . . . . . . // . . . . . // . . A . . // . . T . . // . . G . . // . . . . . // . . . . . . . A . . // . . . . . // . . - - - . . // . . . . . // . . G . . . . T . . . . . . A . . . C . .
44.1*   // . . . . T . . // . . . . . // . . A . . // . . . . . // . . . . . // . . T . . // . . . . . . . . . . // . . . . . // . . . . . . . // . . G . . // . . . . . . . . . . . . . . A . . . . . .
8.1     // . . A . . . . // . . A . . // . . A . . // . . T . . // . . G . . // . . . . . // . . . . . . . A . . // . . . . . // . . . . . . . // . . . . . // . . . . . . . . . . . . . . A . . . . . .
18.2*   // . . . . T . . // . . . . . // . . A . . // . . . . . // . . . . . // . . T . . // . . . . . . . . . . // . . . . . // . . . . . . . // . . G . . // . . . . . . . . . . . . . . A . . . . . .

44.2    11 C T C A T 12 A C T T C 32 C T G A A 27 C C C A G C T 5 A C T T G C C T T C C T G G T 16 G A A T G 7 T G C C C C C T C C T C 45 C A C C C 12 A T C T T 21 T C G T C 7 C A C C A 60 A C G T C C
62.1    // . . . . . // . . C . . // . . A . . // . . A G - . . // . . C C . . . C . . . . A . . // . . G . . // . . T . . . . . . T . . // . . T . . // . . T . . // . . . . . // . . T . . // . . . . . .
7.1     // . . . . . // . . C . . // . . A . . // . . . . . . . // . . C . . . . . . . . . A . . // . . G . . // . . . . . . . . . . . . // . . T . . // . . T . . // . . . . . // . . . . . // . . . C . .
44.1*   // . . . . . // . . C . . // . . A . . // . . A G - . . // . . C C . . . . . . . . A . . // . . G . . // . . . . . . . . . . . . // . . T . . // . . T . . // . . . . . // . . . . . // . . . C . .
8.1     // . . G . . // . . C . . // . . A . . // . . . . . . . // . . C . . . . . . . . . A . . // . . G . . // . . . . . . . . . . . . // . . T . . // . . T . . // . . A . . // . . . . . // . . A . . .
18.2*   // . . . . . // . . C . . // . . A . . // . . A G - . . // . . C C . . . . . . . . A . . // . . G . . // . . . . . . . . . . . . // . . T . . // . . T . . // . . . . . // . . . . . // . . . C . .

44.2    11 C A C G C 9 G G T A A 27 C T T G A 20 G C T T C C C T C A T C C C T C A C C 22 G C A G T 24 G T T A T 6 G C C C T 47 G G T C A G A T G C A 31 C C G C A 24 T C T G T 19 T G C T C 11 C A G - - A C
62.1    // . . . . . // . . . . . // . . . . . // . . C . . . . . . G . . . . . T G . . // . . C . . // . . C . . // . . T . . // . . C . . . . C . . . // . . . . . // . . C . . // . . . . . // . . A A C . .
7.1     // . . . . . // . . A . . // . . C . . // . . C . . . . . . G . . . . . . G . . // . . C . . // . . C . . // . . T . . // . . C . . . . C . . . // . . . . . // . . C . . // . . T . . // . . A A C . .
44.1*   // . . T . . // . . A . . // . . C . . // . . C . . . . . . G . . . . . . G . . // . . C . . // . . C . . // . . T . . // . . C . . . . C . . . // . . A . . // . . . . . // . . . . . // . . . A C . .
8.1     // . . . . . // . . . . . // . . . . . // . . C . . . . . . G . . . . . . G . . // . . C . . // . . C . . // . . T . . // . . C . . . . C . . . // . . . . . // . . . . . // . . . . . // . . A A C . .
18.2*   // . . T . . // . . A . . // . . C . . // . . C . . . . . . G . . . . . . G . . // . . C . . // . . C . . // . . T . . // . . C . . . . C . . . // . . A . . // . . . . . // . . . . . // . . . A C . .

44.2    16 G A T A T 7 T C C C T 28 A A A A A 118 T A G T G 34 G C T A C 18 C T G A T T T 156 C T G C A 50 A G T G C 31 T C T A C 83 G A T G G 51 T C A T G A 54 A G A A G 28 A G G G C 48 C G T G T
62.1    // . . C . . // . . T . . // . . C . . // . . . . . // . . C . . // . . - - - . . // . . . . . // . . . . . // . . . . . // . . C . . // . . . G . . // . . G . . // . . . . . // . . . . .
7.1     // . . C . . // . . T . . // . . C . . // . . A . . // . . . . . // . . - - - . . // . . A . . // . . . . . // . . G . . // . . C . . // . . . . . . // . . . . . // . . A . . // . . . . .
44.1*   // . . C . . // . . T . . // . . C . . // . . . . . // . . . . . // . . . . . . . // . . A . . // . . . . . // . . . . . // . . . . . // . . . . . . // . . . . . // . . . . . // . . C . .
8.1     // . . C . . // . . T . . // . . C . . // . . . . . // . . . . . // . . - - - . . // . . A . . // . . . . . // . . . . . // . . C . . // . . C . . . // . . . . . // . . . . . // . . . . .
18.2*   // . . C . . // . . T . . // . . C . . // . . . . . // . . . . . // . . . . . . . // . . . . . // . . G . . // . . . . . // . . C . . // . . . G . . // . . G . . // . . . . . // . . . . .

44.2    7 C C G C C C T G G 30 C C G C T 55 G A C G 50 C C G T C 98 C T C T C 121 A C A T A 44 C A G G A 21 C A A G T 19 A C G G T G G A C A C G G G G G T G G G C 120 T C A G T 69 A G T G A 7 C T A A A
62.1    // . . A . . . . . . // . . A . . // . . . . // . . A . . // . . T . . // . . . . . // . . A . . // . . G A . // . . A . . . . . . . . A A . . . . . . . . // . . G . . // . . . . . // . . G . .
7.1     // . . . . . . . . . // . . A . . // . . G . // . . A . . // . . . . . // . . G . . // . . A . . // . . G . . // . . A . . . . . . . . A A . . . . . . . . // . . . . . // . . C . . // . . . . .
44.1*   // . . . . . . A . . // . . A . . // . . A . // . . . . . // . . . . . // . . . . . // . . . . . // . . . . . // . . . . . . . . . . . . . . . . . . . . . // . . . . . // . . . . . // . . . . .
8.1     // . . . . . . A . . // . . A . . // . . G . // . . . . . // . . . . . // . . G . . // . . A . . // . . G . . // . . A . . . . . . . . A A . . . . . . . . // . . . . . // . . . . . // . . . . .
18.2*   // . . . . . . . . . // . . . . . // . . G . // . . . . . // . . . . . // . . . . . // . . A . . // . . G . . // . . . . . . . . . . . . . . . . . . A . . // . . G . . // . . . . . // . . G . .
```

```
44.2       6 A C C A C  26 T C T C C  17 C C A C T G C C C C A C C C A C C C C C A G A C C T G C C A C C C C A C C  69 T G C C T  87 C C G C C C C C A T C A  52 T G C C C   6 G G G A C  12 G G G C C  21 C T C G T
62.1      // . . G . .  // . . . . .  // . . . . . . . . . . . . . . . . . . . . . . . . . C . . . . . . . . . . .  // . . G . .  // . . A . . . . . . C . .  // . . T . .  // . . . . .  // . . A . .  // . . . . .
7.1       // . . . . .  // . . C . .  // . . T . . . . . . . . . . . . . . . . . . . . . . C . . . . . . . . . . .  // . . . . .  // . . A . . . . . . C . .  // . . T . .  // . . C . .  // . . . . .  // . . A . .
44.1*     // . . . . .  // . . . . .  // . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  // . . . . .  // . . . . . . . . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .
8.1       // . . . . .  // . . . . .  // . . . . . . . . . . . . . . . . . . . . . . . . . C . . . . . . . . . . .  // . . . . .  // . . A . . . . . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .
18.2*     // . . . . .  // . . . . .  // . . T . . . . . . . C T . . - - - - - - - - - - . C . . . C . . . . G . .  // . . G . .  // . . A . . . . . . C . .  // . . T . .  // . . . . .  // . . . . .  // . . . . .

44.2       4 G A C C C  32 C A A G G 124 C A C C A A G G T G A   4 T C C G T C  57 T G C G T  11 T G G A T G C T G C T T C  93 G A G C T  18 A G C A G  64 G A C G G  11 G A C G G  11 G A A G G  28 G G - - - - - -
62.1      // . . A . .  // . . G . .  // . . T . . . . . C . .  // . . T . . .  // . . . . .  // . . . . . . . . . . . . .  // . . . . .  // . . T . .  // . . T . .  // . . T . .  // . . G . .  // . . G C T C C A
7.1       // . . . . .  // . . G . .  // . . T . . . . . . . .  // . . . A . .  // . . . . .  // . . C . . . . . . . . . .  // . . A . .  // . . . . .  // . . . . .  // . . . . .  // . . G . .  // . . - - - - - -
44.1*     // . . . . .  // . . . . .  // . . . . . . . . . . .  // . . . . . .  // . . . . .  // . . . . . . . . . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . - - - - - -
8.1       // . . . . .  // . . . . .  // . . . . . . . . . . .  // . . . . . .  // . . . . .  // . . . . . . . . . . A . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . G . .  // . . - - - - - -
18.2*     // . . A . .  // . . . . .  // . . T . . . . . . . .  // . . . . . .  // . . T . .  // . . . . . . . . . . . . .  // . . . . .  // . . T . .  // . . T . .  // . . T . .  // . . G . .  // . . G C T C C A

44.2      - - - - - G C  16 A A G A G 163 T C G G T  58 A A G C C  20 T G A C C  27 A C A T G  51 C T C T C A T G G G A C  80 G C C T G G A C G C  15 T C T C C T A  34 A G G G A 105 C A C G G  39 C A G T A T T C T
62.1      G A A G G . .  // . . C . .  // . . . . .  // . . . . .  // . . G . .  // . . . . .  // . . . . . . . . . . . .  // . . . . . . . . . .  // . . . . . . .  // . . . . .  // . . . . .  // . . C . . . C . .
7.1       - - - - - . .  // . . . . .  // . . A . .  // . . A . .  // . . G . .  // . . G . .  // . . T . . . . . . . . .  // . . . . . . . . . .  // . . . . . . .  // . . . . .  // . . . . .  // . . . . . . . . .
44.1*     - - - - - . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . . . . . . . . . . .  // . . . . . . . . . .  // . . . . . . .  // . . . . .  // . . . . .  // . . . . . . . . .
8.1       - - - - - . .  // . . . . .  // . . A . .  // . . . . .  // . . G . .  // . . G . .  // . . . . . . . . . C . .  // . . G - - . . T . .  // . . . . . . .  // . . . . .  // . . . . .  // . . . . . . . . .
18.2*     G A A G G . .  // . . . . .  // . . . . .  // . . . . .  // . . G . .  // . . . . .  // . . . . . . . . . . . .  // . . . . . . . . . .  // . . - - - . .  // . . A . .  // . . T . .  // . . . . . . . . .

44.2      36 A T C C C   6 G T C C T  57 A A G G G 102 C C C G C G C G C T G C A G C G T C T C  15 G T A T C   6 G C G A C   7 A C A G G C  14 C T C A G C T   6 C C A C A  36 G C G G T   4 G C C G C   9 G C T C A
62.1      // . . . . .  // . . . . .  // . . . . .  // . . . . . . . . . . . . . . . . . . . .  // . . . . .  // . . C . .  // . . . . . .  // . . . C A . .  // . . T . .  // . . C . .  // . . . . .  // . . . . .
7.1       // . . T . .  // . . - . .  // . . C . .  // . . A . . . . . . . C . . . . T . G . .  // . . . . .  // . . C . .  // . . T C . .  // . . . C . . .  // . . T . .  // . . C . .  // . . G . .  // . . G . .
44.1*     // . . . . .  // . . . . .  // . . . . .  // . . . . . . . . . . . . . . . . . . . .  // . . . . .  // . . . . .  // . . . . . .  // . . G T C . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .
8.1       // . . . . .  // . . - . .  // . . . . .  // . . . . . . . . . . C . . . . . . G . .  // . . . . .  // . . C . .  // . . G T . .  // . . G T C . .  // . . . . .  // . . . . .  // . . . . .  // . . G . .
18.2*     // . . . . .  // . . . . .  // . . . . .  // . . . . . . . . . . . . . . . . . . . .  // . . G . .  // . . C . .  // . . G T . .  // . . . . . . .  // . . . . .  // . . . . .  // . . . . .  // . . . . .

44.2      36 C G T C C T G G T C A T A C C  41 A T C C T C T G G A T G A T G  12 C C C G G  13 C C C G T C C C C C - G A  91 G T C T G  16 T C G G G G - - G C  36 G T A C G  28 C C G G G  36 G C G G A G C G C G G
62.1      // . . G A . . . . . . . . G . .  // . . . . . . . . . . G . G . .  // . . A . .  // . . . . . . . . . . - . .  // . . T . .  // . . A . . . G G . .  // . . . . .  // . . A . .  // . . C . C . . A G . T
7.1       // . . . A . . . . . . . . G . .  // . . G . . . . . . . G . G . .  // . . . . .  // . . A . . . . . . . - . .  // . . . . .  // . . A . . . A G . .  // . . . . .  // . . . . .  // . . C . C . . A G . T
44.1*     // . . . . . . . . . . . . . . .  // . . . . . . . . . . . . . . .  // . . . . .  // . . . . . . . . . . - . .  // . . . . .  // . . . . . . - - . .  // . . . . .  // . . . . .  // . . . . . . . . . . .
8.1       // . . . A . . . . . T . . G . .  // . . G . . . . . . . G . G . .  // . . . . .  // . . . . . . . . . . C . .  // . . . . .  // . . A . . . A G . .  // . . C . .  // . . . . .  // . . C . C . . A G . T
18.2*     // . . G A . . . . . . . . G . .  // . . . . . . . . . . G . G . .  // . . . . .  // . . . . . . . . . . C . .  // . . . . .  // . . A . . . G G . .  // . . . . .  // . . . . .  // . . C . C . . A G . T

44.2      T G C G C A G G T T C T C T C G G T A A G T C T G T G T G T T G G T C T T G G A G   6 G T C T C C C  20 T C C T G   7 C A T G G   4 C G C G G C T C C T T C C T   6 C G T G G  21 A C A G C G T G T C G  12
62.1      . C . . . . . . C . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  // . . . . . . .  // . . . . .  // . . . . .  // . . G . . . G . . A . . . .  // . . C . .  // . . T . G . . . . . .  //
7.1       . C . . . . . . C . . . . . . . . . C . . . . . . . . C C . G . . C . . . . T . .  // . . G . T . .  // . . . . .  // . . C . .  // . . . . . . . . . . C T . .  // . . C . .  // . . T . G . . . . . .  //
44.1*     . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  // . . . . . . .  // . . . . .  // . . . . .  // . . . . . . . . . . . . . .  // . . . . .  // . . . . . . . . . . .  //
8.1       . C . . . . . . C . . . . . . . . . C . . . . . . . . . . . . . . . . . . . A . .  // . . G . T . .  // . . . . .  // . . C . .  // . . . . . . . . . . C T . .  // . . C . .  // . . T . . . . . . . .  //
18.2*     . C . . . . . . C . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  // . . G . T . .  // . . T . .  // . . C . .  // . . G . . . . . . G . . . .  // . . C . .  // . . T . G . . . C . .  //

44.2      A C G G T G A  31 C A T G G C G G T G T A G A  24 G G G G C G A G  40 C C C C G G G  25 T C T C T C C T C C C C A C A  12 C T C C C G A C C C C G C A C T C A C C G G C  18 C A C T G C C C C C C A G
62.1      . . T . C . .  // . . . . . . . . . . . . . .  // . . . . . . . .  // . . . . . . .  // . . G . . . . . . . . . . . .  // . . G . . . . . . . . . . . . . . . . . . . .  // . . G G . T . . . . G . .
7.1       . . T . A . .  // . . C . . A . . . . . . . .  // . . . . T . . .  // . . T . A . .  // . . G . . . . . . . . G G . .  // . . . . . . . . . . - . . . . . . . . . . . .  // . . G G . . . G . . G . .
44.1*     . . . . . . .  // . . . . . . . . . . . . . .  // . . . . . . . .  // . . . . . . .  // . . . . . . . . . . . . . . .  // . . . . . . . . . . . . . . . . . . . . . . .  // . . . . . . . . . . . . .
8.1       . . T . A . .  // . . . . . . . . . . . C . .  // . . . . . . . .  // . . T . A . .  // . . G . . . . . . . . G G . .  // . . . . . . . . . . - . . . . . . . . . . . .  // . . G G . . . G . . G . .
18.2*     . . T . A . .  // . . C . . A . . . . . G . .  // . . . - . A . .  // . . . . . . .  // . . G . . . . . . . . . . . .  // . . G . . . . . . . . . . . . . . . . . A . .  // . . . . . . . . . . . . .
```

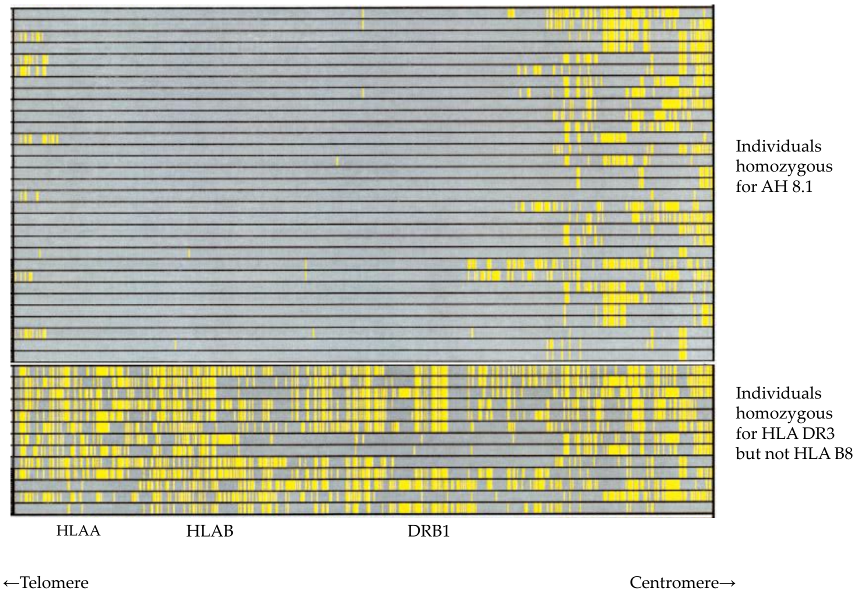
```
44.2   11  GAGGG 9   CCGTGACCCGC 17  GATTCTG 36  CCCCGGCGACACT 14  GACACCCGACAC 14  AAATGGGGACGC 29  CATAGGT 18  TCGTGG 15  AACGC 10
62.1   //  ..C.. //  ........... //  ....T.. //  ............. //  ............ //  ............ //  ....... //  ...... //  ..... //
7.1    //  ..C.. //  ..A.....A.. //  ..C.... //  ..G.......G.. //  ..T.T..A.T.. //  ..G......... //  ..C.A.. //  ..AG.. //  ..... //
44.1*  //  ..... //  ........... //  ....... //  ............. //  ............ //  ............ //  ....... //  ...... //  ..... //
8.1    //  ..C.. //  ..A.....A.. //  ..C.... //  ..........G.. //  ..T.T....... //  ..G......T.. //  ..C.A.. //  ..AG.. //  ..... //
18.2*  //  ..... //  ........... //  ....... //  ............. //  ............ //  ............ //  ....... //  ...... //  ..T.. //

44.2   CATTGCAGACGCGGCCCT 24  CGTCCGGGACC 33  CA-GC 28  TGGAC 45  CAGCCCCT 13  TCCTC 34  TCTCA //  GA-TC //  GATCACGGGGA 59  CTCTA 5
62.1   ..A......AA...T... //  ........... //  ..-.. //  ..C.. //  ..C..G.. //  ..... //  ..... //  ..-.. //  ........A.. //  ..... //
7.1    .................. //  ........... //  ..A.. //  ..... //  ........ //  ..G.. //  ..C.. //  ..C.. //  ..G........ //  ..T.. //
44.1*  .................. //  ........... //  ..-.. //  ..... //  ........ //  ..... //  ..... //  ..-.. //  ........... //  ..... //
8.1    .................. //  ........... //  ..A.. //  ..... //  ........ //  ..G.. //  ..C.. //  ..C.. //  ..G........ //  ..T.. //
18.2*  ..A......AA...T... //  ..C.....G.. //  ..-.. //  ..C.. //  ..C..G.. //  ..... //  ..... //  ..-.. //  ........A.. //  ..... //

44.2   GGCGA 4   TTGTCTGAAGCCACCACACAGAGA 28  TTATT 50  ATCTGCAA 56  TTCCCAGACTATG 5   GTGAC 38  CCCGA 6   CTATATG 28  AGCGGGGTT 19
62.1   ..... //  ........................ //  ..... //  ........ //  ..A.......... //  ..... //  ..T.. //  ..G.G.. //  ......T.. //
7.1    ..T.. //  ..C.......G....G..T..G.. //  ..G.. //  ..T..T.. //  ..........G.. //  ..A.. //  ..T.. //  ..G.G.. //  ...A..... //
44.1*  ..... //  ........................ //  ..... //  ........ //  ............. //  ..... //  ..... //  ....... //  ......... //
8.1    ..T.. //  ..C.......G....G..T..G.. //  ..G.. //  ..T..T.. //  ..........G.. //  ..A.. //  ..T.. //  ..G.G.. //  ...A..... //
18.2*  ..... //  ........................ //  ..... //  ........ //  ............. //  ..... //  ..... //  ....... //  ..G...... //

44.2   CGGGAACC 20  TGGGAGTGC 24  CCGACGCTCAA 21  TCCGCACATCTAAGCAGTATGCATA 7   TGGACACTGGG 22  TT-AA 29  TCTAA 11  AAATA 18  CTGCC
62.1   ........ //  ..C...G.. //  ..AG....... //  ..T.......C.......G...... //  ..A.....T.. //  ..-.. //  ..C.. //  ..... //  .....
7.1    .....G.. //  ..C...... //  ........G.. //  ..T.......C.......G..G... //  ..A.....T.. //  ..T.. //  ..C.. //  ..... //  .....
44.1*  ........ //  ......... //  ........... //  ......................... //  ........... //  ..-.. //  ..... //  ..... //  .....
8.1    .....G.. //  ..C...... //  ........G.. //  ..T.......C.......G..G... //  ..A.....T.. //  ..-.. //  ..C.. //  ..... //  .....
18.2*  ..C..... //  ..C...G.. //  ..AG....... //  ..T...T...C.......G...... //  ..A..C..T.. //  ..-.. //  ..C.. //  ..C.. //  ..A..

44.2   AGACT 12  TAAAATTCTTTCCCC 36  CTCTC 20  CCCCCGCTCC 5   GTCCCCGC 4   CTTTCACTCC 38  ATGTC 23  GTCCT 16  TCATG 12  TTGCAATG 25  TAGATGT
62.1   ..G.. //  ..G.........T.. //  ..G.. //  ..A..-.C.. //  .....T.. //  .......... //  ..A.. //  ..T.. //  ..... //  ..A..G.. //  ..T....
7.1    ..G.. //  ..G.........T.. //  ..G.. //  .......... //  ........ //  ..G....... //  ..A.. //  ..... //  ..... //  ........ //  ..T.C..
44.1*  ..... //  ............... //  ..... //  .......... //  ........ //  .......... //  ..... //  ..... //  ..... //  ........ //  .......
8.1    ..G.. //  ..G.T.......T.. //  ..G.. //  .......... //  ........ //  ..G....... //  ..A.. //  ..... //  ..... //  ........ //  ..T.C..
18.2*  ..G.. //  ..G.........T.. //  ..G.. //  .......... //  ..G..T.. //  .......C.. //  ..... //  ..... //  ..G.. //  ........ //  .......

44.2   194 AG-AA 84  TAAGCTCAATG 6   TGTAA 69  TGTAT 86  TTTAA 84  CCCTTTTATTTATTTATTTATTTT 17  ATATG 35  GCAAA 99  CCCTT 92  GTCACTCCT 8
62.1   //  ..-.. //  ........... //  ..... //  ..... //  ..... //  ........................ //  ..... //  ..G.. //  ..T.. //  ..T...T.. //
7.1    //  ..C.. //  ..GA....G.. //  ..C.. //  ..... //  ..A.. //  ...CCC.T...T...T..----.. //  ..G.. //  ..G.. //  ..T.. //  ..T..CT.. //
44.1*  //  ..-.. //  ........... //  ..... //  ..... //  ..... //  ........................ //  ..... //  ..... //  ..... //  ......... //
8.1    //  ..C.. //  ..GA....G.. //  ..C.. //  ..C.. //  ..A.. //  ..G-----CCCT...T...T.... //  ..G.. //  ..G.. //  ..T.. //  ..T...T.. //
18.2*  //  ..-.. //  ........... //  ..... //  ..... //  ..... //  ........................ //  ..... //  ..... //  ..... //  ..T..CT.. //

44.2   GCGCT 7   ACCGT 105 ATACT 29  TCTTA 43  GTTTG 13  CACTC 30  GCGGG 64  GGCTT 151 CTGCC 36  CCTAT 17  CAATA 133 TATTA 24  AAATTCAATG 192 CAGCC
62.1   ..... //  ..T.. //  ..... //  ..... //  ..... //  ..... //  ..T.. //  ..... //  ..... //  ..G.. //  ..G.. //  ..... //  .......... //  .....
7.1    ..... //  ..... //  ..... //  ..G.. //  ..G.. //  ..T.. //  ..T.. //  ..T.. //  ..... //  ..... //  ..G.. //  ..C.. //  .......G.. //  ..A..
44.1*  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  .......... //  .....
8.1    ..A.. //  ..... //  ..... //  ..G.. //  ..G.. //  ..T.. //  ..T.. //  ..T.. //  ..A.. //  ..... //  ..G.. //  ..... //  ..G....G.. //  ..A..
18.2*  ..... //  ..... //  ..-.. //  ..G.. //  ..... //  ..... //  ..... //  ..... //  ..... //  ..... //  ..G.. //  ..... //  .......... //  .....
```

Figure 6. Alignment of 9 kb sequence at HLA-B. Sequences of 6 individuals with homozygous ancestral haplotypes were downloaded from UCSC browser [33] at HLA B and aligned using ClustalX2 [34]. For the purposes of illustration only, common sequences were removed and the interruption marked as //. The nucleotides of AH 44.2 are displayed in

the first row. Nucleotides of AH 62.1, 7.1, 44.1*, 8.1 and 18.2* are given only where they differ from AH44.2 and otherwise marked with a dot. Missing nucleotides are marked with a dash and shaded grey. The sequences are described by Horton et al. [24], whereas AH haplotypes have been assigned from the HLA allele types given by Horton, according to Cattley [35].

The degree of conservation of each ancestral haplotype is truly remarkable. For example, Smith et al. [32] found variation at only 11 of 3, 600, 000 positions between HLA-A and DR. Similar findings have been reported by others, including Aly et al. [31], see Figure 7. Mutation and recombination must be suppressed.

Figure 7 illustrates the importance of interpreting nucleotide diversity according to the block structure of the genome. Thus, conservation in the intervening, essentially monomorphic regions, is of minor interest, whereas differences within PFB allow the discovery of evolution, function and disease susceptibility.



Adapted from ref. [31].

Figure 7. Remarkable conservation within 8.1 haplotypes. A total of 656 SNPs spanning 4.8 Mb in the MHC region are depicted. The lower frequency allele (row) for each SNP along each haplotype column is highlighted in yellow. The top group depicts SNP results from 8.1 AH haplotypes (n = 31), the lower group are HLA-DR3, non-B8 haplotypes (n = 13). The 29.9 Mb range between HLA and DRB1 was >99.9% conserved, with only 9 variant alleles of the 10, 768 alleles identified for the 384 SNPs in the 31 8.1 AHs.

The inescapable conclusion is that some parts of the genome have *not* two or three but hundreds of alternative ancestral sequences.

6. Sequence analysis of ancestral haplotypes

The challenge in terms of sequence analysis is to compile a sufficient matrix to be able to recognize each haplotype and its extent. Assume access to multigenerational families with accurate, truly phased but unmolested raw sequences of at least 100, 000 bases:

1. Clustering of these by independent criteria relating to as many as hundreds of distinct ancestral haplotypes.

2. Alignments which take account of haplospecific duplicons, indels and retroviral-like elements (RLE).

3. Functional information to address biological and disease significance.

Given NGS, this approach is now feasible, even if daunting.

Importantly, those regions which are complex because of duplications and indels should be included rather than "corrected" based on the assumption that there is a single reference or "wild" sequence. Some examples are shown in Figure 6.

In designing better algorithms [36], the strategy for comparative analysis will be crucial. In many polymorphic regions, the density of differences can be as high as 1 per 10 bases when different haplotypes are compared but as low as 0 if the haplotypes are the same. It follows that analysis without haplotype assignment will be misleading.

7. Finding polymorphic frozen blocks and their ancestral haplotypes

The best clue to the location of these blocks is segmental duplication [17, 37].

To characterize the PFB, it is helpful to amplify haplospecific geometric elements [30], see also Table 3. Essentially, this approach reveals duplications as seen in Figure 8. McLure developed the approach to find PFB throughout the genome [36]. Paralogous regions are also helpful as shown in Figure 9.

Once identified, we recommend tracking the polymorphism through panels of multigenerational families as illustrated in Figure 10. Although the region is over 10 megabases, recombination was not found. The different haplotypes in the three breeds must have been conserved for at least hundreds of generations and mark differences in function such as the melting point of fat [37].

8. Applications to NGS and the 1000 genomes project

8.1. Mapping PFB from 1000 genomes data

Since it is known that PFB can be mapped by plotting diversity measurements (see Figure 3), we asked whether it would be possible to use data from the 1000 Genomes Project [39] in the same way.

Earlier work was based on haplotypes defined in multigenerational families. Initially, sequences of haplotypes were determined from Sanger sequencing of homozygous cell lines. In contrast, variations in 1000 genomes are determined from NGS for heterozygous and unrelated

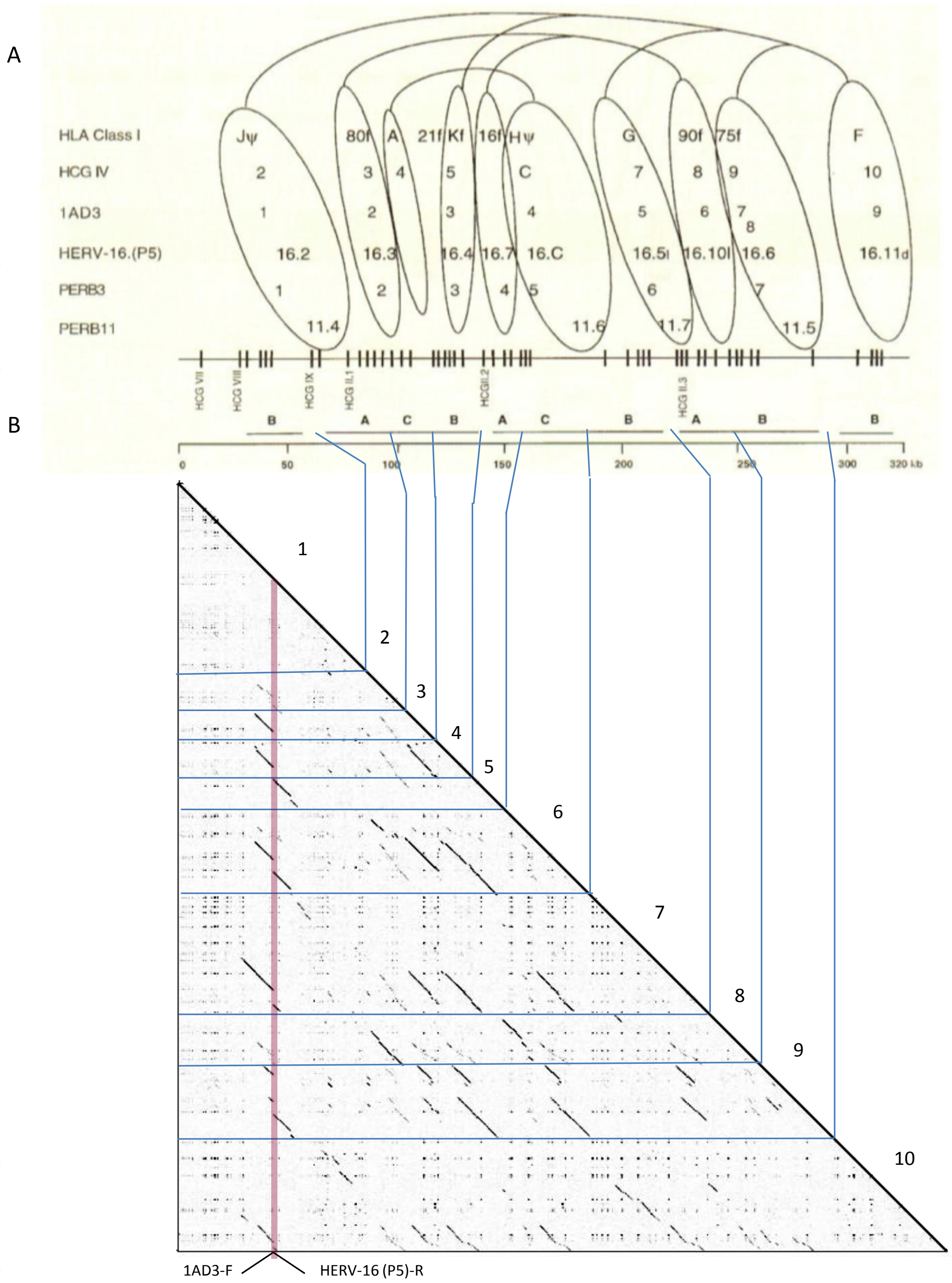
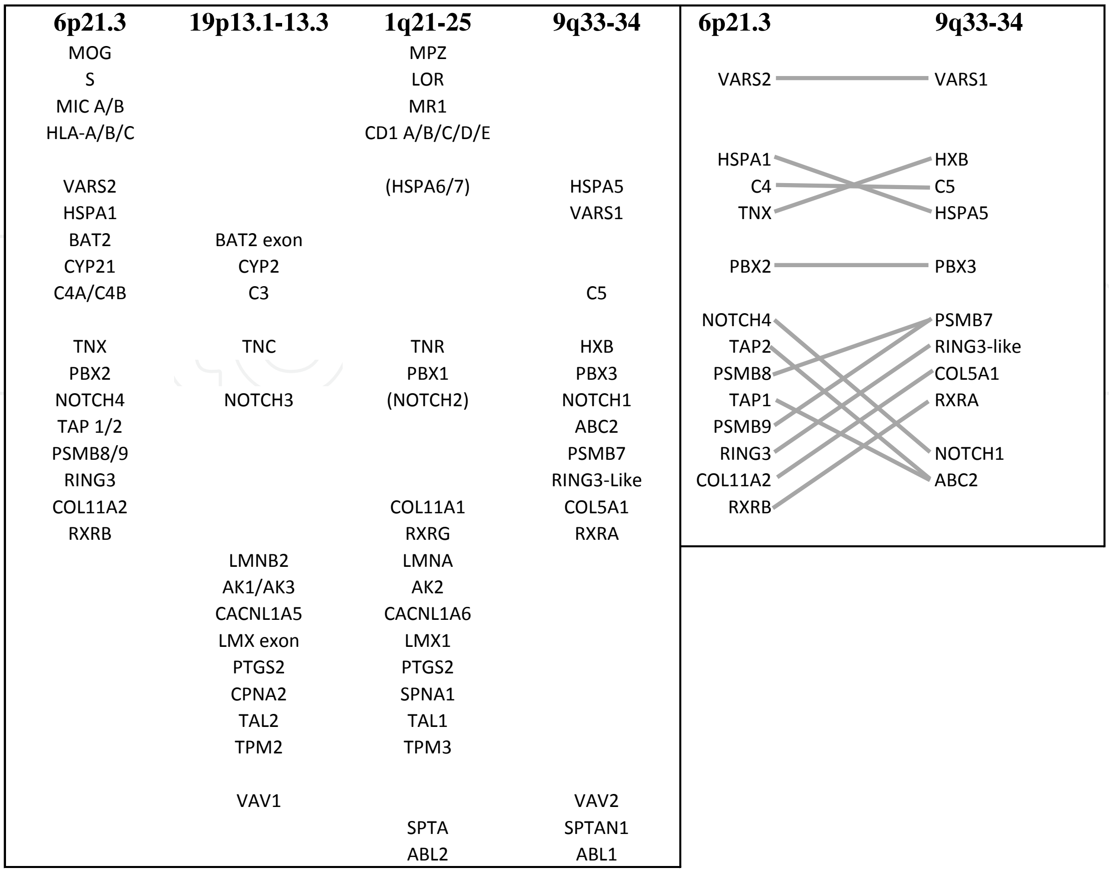


Figure 8. Segmental duplications in MHC alpha block. (a) Gene families and retroelements PERB 11, HLA, HCGIV, AD-3, HERV-16, PERB3 are duplicated to form an ordered pattern within the alpha block of the MHC, indicating that a segment containing multiple genes and retroelements has been duplicated to give 10 duplicons. Full-length duplicons consist of PERB11, HLA, HCGIV, 1AD3, HERV-16 (P5) and PERB3 genes. HLA-80, HLA-A, HIA-K, HLA-16, HLA-90 and HLA-F duplicons lack PERB11 gene. f = fragment, 1 = LTR only, d = discontinuous. ψ = pseudogene. A, B and C represent subgroups of duplicons with greater similarity. (b) A dot plot of the 319 kb genomic sequence encompassing the alpha block was compared against itself. The oblique lines in the plot represent duplications whereas the dots represent retroelements. Lines connect regions of the dotplot to the appropriate duplicons. The primers shown amplify products of different lengths in each duplication. Sequence from GenBank accession number AF055066. Adapted from ref. [17].



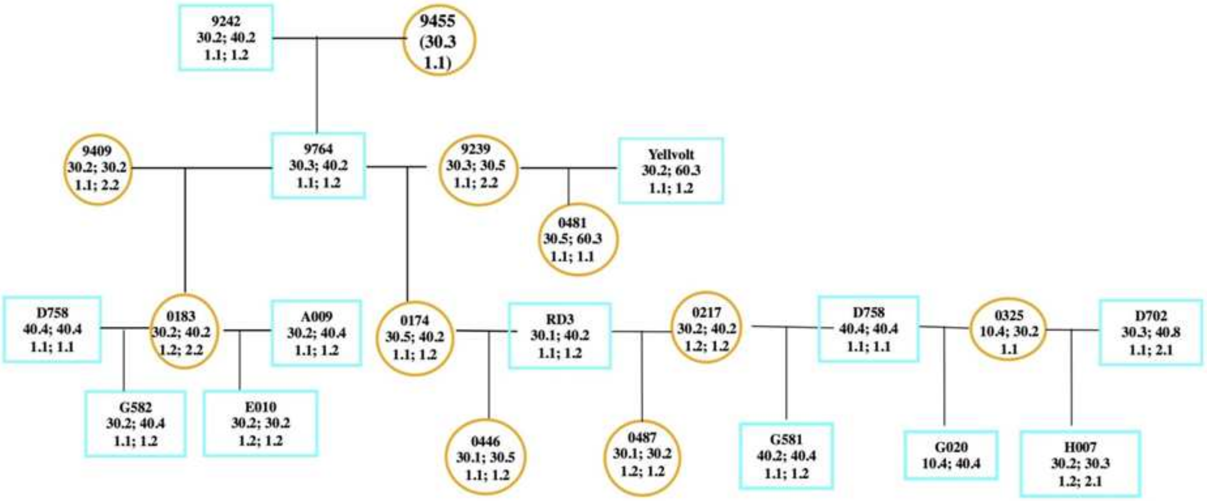
6p21.3	19p13.1-13.3	1q21-25	9q33-34
MOG		MPZ	
S		LOR	
MIC A/B		MR1	
HLA-A/B/C		CD1 A/B/C/D/E	
VARS2		(HSPA6/7)	HSPA5
HSPA1			VARS1
BAT2	BAT2 exon		
CYP21	CYP2		
C4A/C4B	C3		C5
TNX	TNC	TNR	HXB
PBX2		PBX1	PBX3
NOTCH4	NOTCH3	(NOTCH2)	NOTCH1
TAP 1/2			ABC2
PSMB8/9			PSMB7
RING3			RING3-Like
COL11A2		COL11A1	COL5A1
RXRB		RXRG	RXRA
	LMNB2	LMNA	
	AK1/AK3	AK2	
	CACNL1A5	CACNL1A6	
	LMX exon	LMX1	
	PTGS2	PTGS2	
	CPNA2	SPNA1	
	TAL2	TAL1	
	TPM2	TPM3	
	VAV1		VAV2
		SPTA	SPTAN1
		ABL2	ABL1

Figure 9. Paralogous locations of MHC genes. MHC genes are found on four chromosomes: 1, 9, 19 as well as chromosome 6. The arrangements of genes in each of the paralogous groups can be largely explained by duplication with and without inversion events. The genes common to chromosomes 6 and 9 are shown.

individuals. The phasing is an estimate based on ideas inherent in population genetics. It is known that the approach is a risky approximation. For example, artefactual “switch-overs” between haplotypes are misleading [40]. Since the reads tend to be short, such as just hundreds of bases, assembly can be fraught. There is a risk of missing complex polymorphisms and underestimating the number of ancestral haplotypes. Given these problems, we plotted several indices related to the 1000 genomes. The intention was to identify any similarities with the distribution as shown in Figure 3.

Unexpectedly, Figure 11 shows a remarkable correspondence between the classical measurements and our extraction from the 1000 Genomes database. The exception around 31.4 Mb was missed by the NGS reanalysis presumably because it is a region which is rich in complex iterative sequences, as shown in Figure 12.

These results are very encouraging in that the advantages of NGS can be coupled with identification of genomic architecture and therefore targeting of the most informative regions. The similarity, by simply counting the base differences per 10 kb, can be refined and applied to the whole genome. The plot of number of “haplotypes” is also promising, although clearly not indicative of the number of ancestral haplotypes.



Reproduced with permission from ref. [22].

Figure 10. Tracing segregation through three generation families. The alleles at MRIP, now known as myosin phosphatase Rho-interacting protein, are used to designate haplotypes within the 5.5 Mb region of bovine chromosome 19 from SREBF1 to TCAP. Within this region, there are many genes involved in muscle development, growth and fatty acid synthesis. For further details, see Williamson et al. [38].

8.2. Comparing polymorphic sequences of well-characterised PFB

Since there are numerous ancestral haplotypes within a PFB, it is essential to compare as many sequences as possible. An example is shown in Figure 6.

It can be seen that

- Only a minority of sites are informative and these must be selected from the remainder.
- Kilobases need to be examined and reduced 10- to 100-fold, retaining the informative sites.
- Different haplotypes are defined by specific combinations of bases at those informative sites.
- Very few single nucleotide polymorphisms are specific for a particular ancestral haplotype. On the contrary, specific combinations may be best defined by comparison with a library of reference sequences.
- Indels are important: alignments can be misleading.

Thus, although the identification of each of the many haplotype remains challenging, the overall patterns of informative sites are helpful in screening for PFB and for localising haplospecific sequences.

9. Conclusion

In analysing NGS databases, we recommend:

1. Screening for PFB.

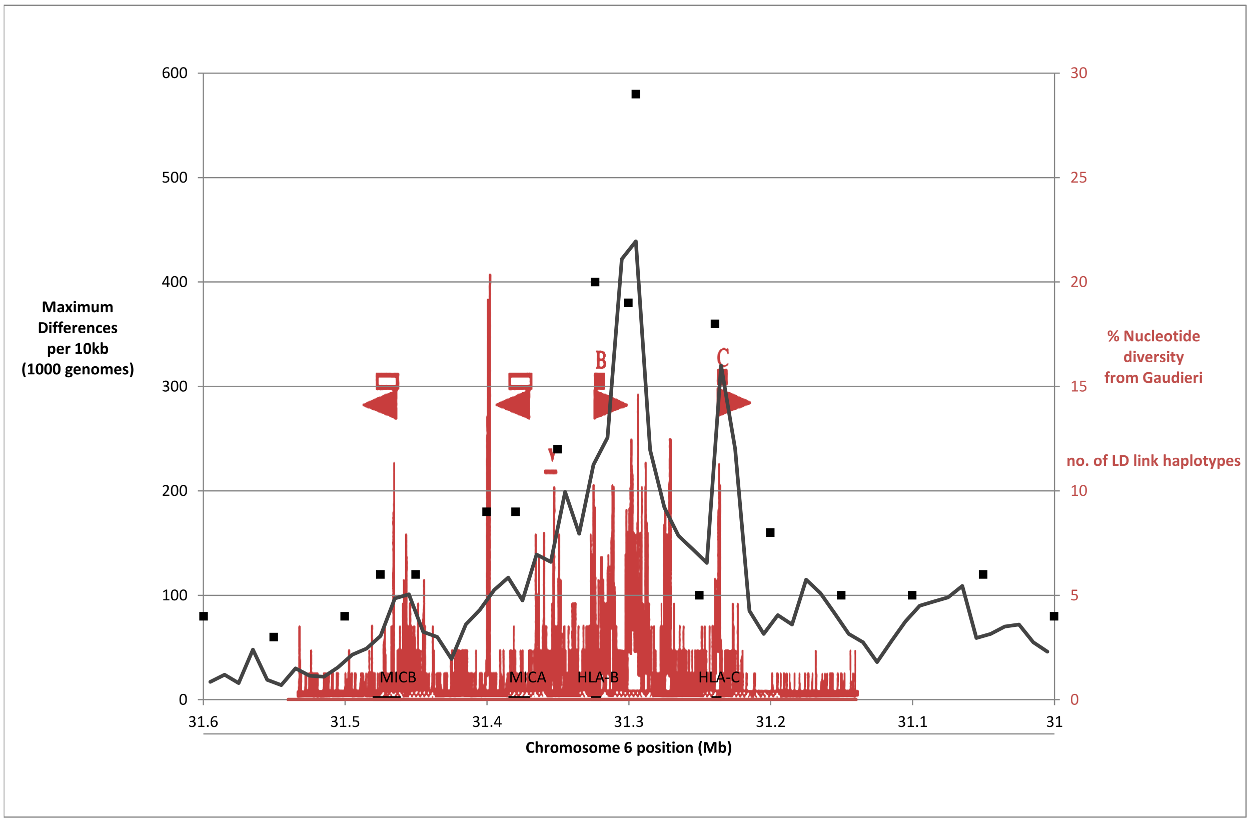


Figure 11. Regions of high sequence diversity within 1000 genomes are similar to previously identified PFB. Imputed haplotypes in the 600 kb region surrounding HLA-B from 553 individuals were downloaded from the 1000 Genomes browser [41]. The population groups chosen were of African, European and Asian origin (ACB, ASW, BEB, CEU, CHB and YRI). The majority of variations recorded in the 1000 Genomes vcf files are SNPs, but some indels up to 174 bp are recorded. For each imputed haplotype, we counted the number of differences from the reference sequence in 10 kb sections. Indels were counted as one difference, irrespective of length. The black curve represents the maximum difference at each 10 kb. The red lines, taken from ref. [42], show the amount of nucleotide diversity between two individual haplotypes, counted in 100 bp sections. Haplotypes compared for this section were 44.1 to 62.1, 44.1 to 8.1 and 8.1 to 14.1. Squares show the number of LD_link [41] "haplotypes", calculated from sets of adjacent variants in 500 bp intervals. LD link requires that variants be biallelic and only takes single nucleotide changes, not indels. Only variants with at least two examples in the CEU and YRI populations were included.

2. Alignment based on the ability to detect multiple, and even hundreds of ancestral haplotypes.

3. Analysis must recognise that haplospecificity is confirmed by many characteristics including RLE, indels, copy number and complex iterative sequences.

4. Analysis may be facilitated by examining paralogous regions which help to define interactions, including epistasis.

5. Validation of results by showing segregation in multigenerational family studies.

6. Confirming biological significance by demonstrating permissive or *sine qua non* associations.

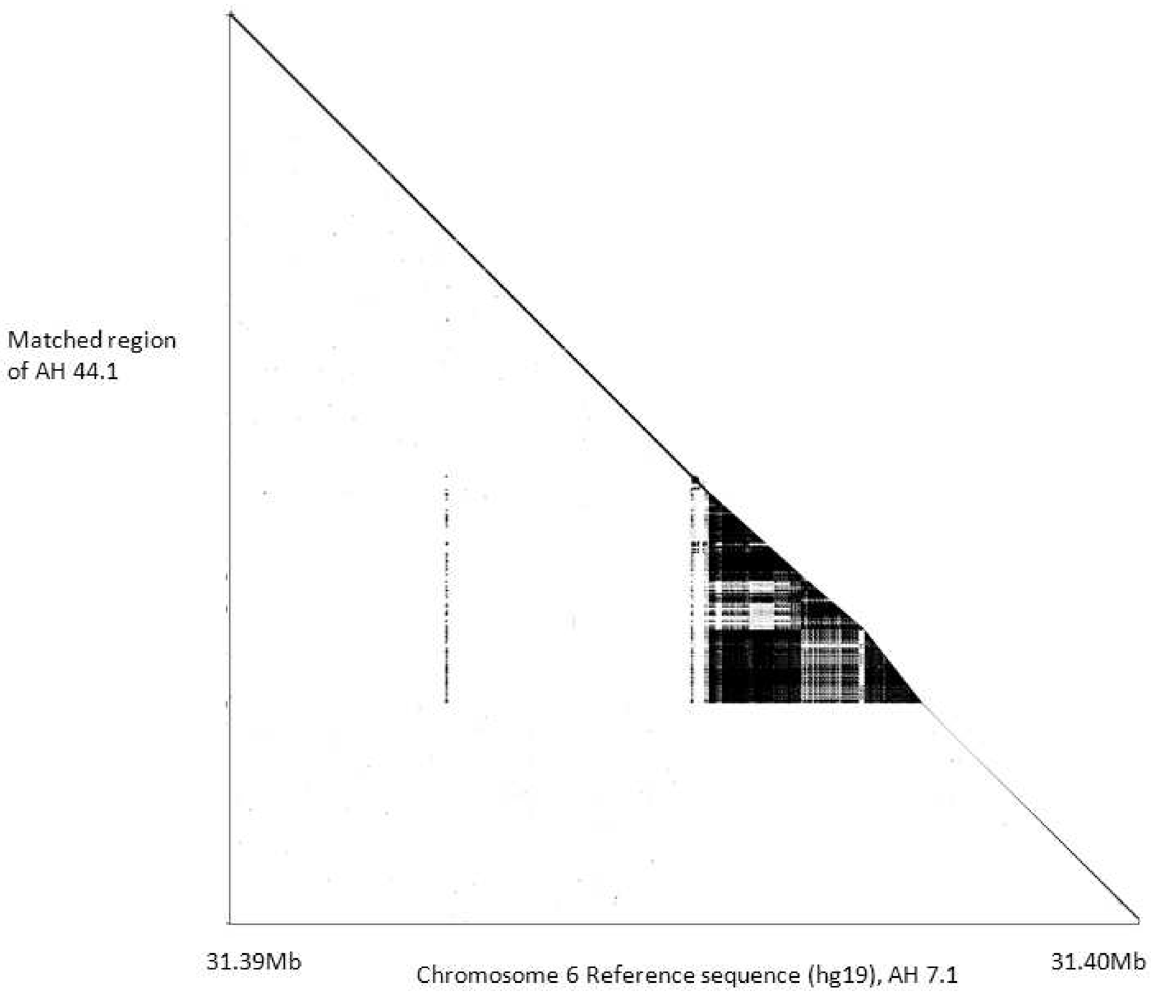


Figure 12. Complex iterative element. Dotplot of a 10 kb region in the MHC between MICA and MICB showing a complex iterative element. Gaudieri [42] shows high nucleotide diversity for this region which was not recorded within 1000 Genomes data. Example sequences for AH 7.1 and AH 44.1 downloaded from UCSC genome browser. Dotplot generated with Gepard [43] using word length 10.

Author details

Sally S. Lloyd[1], Edward J. Steele[1] and Roger L. Dawkins[1,2,3*]

*Address all correspondence to: rldawkins@cyo.edu.au

1 CY O'Connor ERADE Village Foundation, 24 Genomics Rise, Piara Waters, Western Australia, Australia

2 School of Veterinary and Biomedical Sciences, Division of Health Sciences, Murdoch University, Murdoch, Western Australia, Australia

3 Faculty of Medicine and Dentistry, University of Western Australia, Nedlands, Western Australia, Australia

References

[1] Kulski J, Suzuki S, Ozaki Y, Mitsunaga S. In Phase HLA Genotyping by Next Generation Sequencing—A Comparison Between Two Massively Parallel Sequencing Bench-Top Systems, the Roche GS. In: Xi Y, editor. HLA Assoc. Important Dis., InTech; 2014, p. 141–81. doi:10.5772/57556.

[2] Lander ES. Initial impact of the sequencing of the human genome. Nature 2011;470:187–97. doi:10.1038/nature09792.

[3] Dawkins R, Christiansen F, Zilko P, editors. Immunogenetics in Rheumatology: Musculoskeletal Disease and D-Penicillamine. Excerpta Medica. Amsterdam-Oxford-Princeton; 1982.

[4] Dawkins RL, Christiansen FT, Kay PH, Garlepp M, McCluskey J, Hollingsworth PN, et al. Disease associations with complotypes, supratypes and haplotypes. Immunol Rev 1983;70:5–22.

[5] de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 2006;38:1166–72. doi:10.1038/ng1885.

[6] Nakajima F, Tokunaga K, Nakatsuji N. Human leukocyte antigen matching estimations in a hypothetical bank of human embryonic stem cell lines in the Japanese population for use in cell transplantation therapy. Stem Cells 2007;25:983–5. doi:10.1634/stemcells.2006-0566.

[7] Barry J, Hyllner J, Stacey G, Taylor CJ, Turner M. Setting Up a Haplobank: Issues and Solutions. Curr Stem Cell Reports 2015;1:110–7. doi:10.1007/s40778-015-0011-7.

[8] Bodmer WF, Trowsdale J, Young J, Bodmer J. Gene clusters and the evolution of the major histocompatibility system. Philos Trans R Soc Lond B Biol Sci 1986;312:303–15.

[9] Ceppellini R, Curtoni ES, Mattuiz PL, V.Miggiano, Scudeller G, Serra A. Genetics of Leukocyte Antigens: A Family Study of Segregation and Linkage. In: Curtoni ES, Mattiuz PL, Tosi RM, editors. Histocompat. Test. 1967, Munksgaard, Copenhagen: 1967, p. 149–87.

[10] Awdeh ZL, Raum D, Yunis EJ, Alper CA. Extended HLA/complement allele haplotypes: evidence for T/t-like complex in man. Proc Natl Acad Sci U S A 1983;80:259–63.

[11] O'Neill GJ, Pollack MS, Yang SY, Levine LS, New MI, Dupont B. Gene frequencies and genetic linkage disequilibrium for the HLA-linked genes Bf, C2, C4S, C4F, 21-hydroxylase deficiency and glyoxalase I. Transplant Proc 1979;4:1713–5.

[12] O'Neill GJ, Yang SY, Dupont B. Two HLA-linked loci controlling the fourth component of human complement. Proc Natl Acad Sci U S A 1978;75:5165–9. doi:10.1073/pnas.75.10.5165.

[13] O'Neill GJ, Nerl CW, Kay PH, Christiansen FT, McCluskey J, Dawkins RL. Complement C4 is a Marker for Adult Rheumatoid Arthritis. Lancet 1982;320:214. doi:10.1016/S0140-6736(82)91057-1.

[14] Pollack MS, Levine LS, O'Neill GJ, Pang S, Lorenzen F, Kohn B, et al. HLA linkage and B14, DR1, BfS haplotype association with the genes for late onset and cryptic 21-hydroxylase deficiency. Am J Hum Genet 1981;33:540–50.

[15] Alper CA, Awdeh ZL, Raum DD, Yunis EJ. Extended major histocompatibility complex haplotypes in man: role of alleles analogous to murine t mutants. Clin Immunol Immunopathol 1982;24:276–85.

[16] Raum D, Awdeh Z, Yunis EJ, Alper CA, Gabbay KH. Extended Major Histocompatibility Complex Haplotypes in Type I Diabetes Mellitus. J Clin Invest 1984;74:449–54.

[17] Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, et al. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. Immunol Rev 1999;167:275–304. doi:10.1111/j.1600-065X.1999.tb01399.x.

[18] Degli-Esposti MA, Leaver AL, Christiansen FT, Witt CS, Abraham LJ, Dawkins RL. Ancestral Haplotypes: Conserved Population MHC Haplotypes. Hum Immunol 1992;34:242–52. doi:10.1016/0198-8859(92)90023-G.

[19] Gaudieri S, Leelayuwat C, Tay GK, Townend DC, Dawkins RL. The major histocompatibility complex (MHC) contains conserved polymorphic genomic sequences that are shuffled by recombination to form ethnic-specific haplotypes. J Mol Evol 1997;45:17–23.

[20] Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 2013;14:301–23. doi:10.1146/annurev-genom-091212-153455.

[21] Lloyd SS, Bayard D, Lester SA, Williamson JF, Dawkins RL. The Value of Haplotyping. INTERBULL Bull 2013;47:252–5.

[22] Dawkins RL. Adapting Genetics. Dallas, TX: Near Urban Publishing; 2015.

[23] Smith WP, Vu Q, Li SS, Hansen J a., Zhao LP, Geraghty DE. Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. Genomics 2006;87:561–71. doi:10.1016/j.ygeno.2005.11.020.

[24] Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, et al. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 2008;60:1–18. doi:10.1007/s00251-007-0262-2.

[25] Shiina T, Ota M, Shimizu S, Katsuyama Y, Hashimoto N, Takasu M, et al. Rapid Evolution of Major Histocompatibility Complex Class I Genes in Primates Generates

New Disease Alleles in Humans via Hitchhiking Diversity. Genetics 2006;173:1555–70. doi:10.1534/genetics.106.057034.

[26] Longman-Jacobsen N, Williamson JF, Dawkins RL, Gaudieri S. In polymorphic genomic regions indels cluster with nucleotide polymorphism: Quantum Genomics. Gene 2003;312:257–61. doi:S0378111903006218 [pii].

[27] Curtis D, Vine AE, Knight J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet 2008;72:261–78. doi:10.1111/j.1469-1809.2007.00411.x.

[28] Clark AG. The size distribution of homozygous segments in the human genome. Am J Hum Genet 1999;65:1489–92. doi:10.1086/302668.

[29] Lloyd SS, Bayard D, Lester S, Williamson JF, Steele EJ, Dawkins RL. Ancestral Haplotypes, Quantal Genomics and Healthy Beef S. Proceedings, 10th World Congr. Genet. Appl. to Livest. Prod. Ancestral, 2014.

[30] Abraham LJ, Leelayuwat C, Grimsley G, Degli-Esposti M a, Mann A, Zhang WJ, et al. Sequence differences between HLA-B and TNF distinguish different MHC ancestral haplotypes. Tissue Antigens 1992;39:117–21.

[31] Aly T a., Eller E, Ide A, Gowan K, Babu SR, Erlich H a., et al. Multi-SNP analysis of MHC region: remarkable conservation of HLA-A1-B8-DR3 haplotype. Diabetes 2006;55:1265–9. doi:10.2337/db05-1276.

[32] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res 2002;12:996–1006. doi:10.1101/gr.229102.

[33] Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–8. doi:10.1093/bioinformatics/btm404.

[34] Cattley SK, Williamson JF, Tay GK, Martinez OP, Gaudieri S, Dawkins RL. Further characterization of MHC haplotypes demonstrates conservation telomeric of HLA-A: Update of the 4AOH and 10 IHW cell panels. Eur J Immunogenet 2000;27:397–426. doi:eji226 [pii].

[35] Su SY, Balding DJ, Coin LJM. Disease association tests by inferring ancestral haplotypes using a hidden markov model. Bioinformatics 2008;24:972–8. doi:10.1093/bioinformatics/btn071.

[36] McLure CA, Hinchliffe P, Lester S, Williamson JF, Millman JA, Keating PJ, et al. Genomic Evolution and Polymorphism: Segmental Duplications and Haplotypes at 108 Regions on 21 Chromosomes. Genomics 2013;102:15–26. doi:10.1016/j.ygeno.2013.02.011.

[37] Lloyd SS, Valenzuela J, Bayard D, de Bruin S, Gilmour P, Steele EJ Dawkins RL. Heritability of fat melting temperature in beef cattle 2015. In preparation

[38] Williamson JF, Steele EJ, Lester S, Kalai O, Millman JA, Wolrige L, et al. Genomic evolution in domestic cattle: Ancestral haplotypes and healthy beef. Genomics 2011;97:304–12. doi:S0888-7543(11)00037-1 [pii] 10.1016/j.ygeno.2011.02.006.

[39] Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs R A., et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73. doi:10.1038/nature09534.

[40] Machiela MJ, Chanock SJ. LDlink©: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015.

[41] Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, et al. 1000 Genomes Selection Browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res 2014;42:D903–9. doi: 10.1093/nar/gkt1188.

[42] Gaudieri S, Kulski JK, Dawkins RL, Gojobori T. Extensive nucleotide variability within a 370 kb sequence from the central region of the Major Histocompatibility Complex. Gene 1999;238:157–61.

[43] Krumsiek J, Arnold R, Rattei T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 2007;23:1026–8. doi:10.1093/bioinformatics/btm039.