Opinion

The gentle art of gene arrangement: the meaning of gene clusters

John Trowsdale

Address: Department of Pathology, Tennis Court Road, Cambridge CB2 1QP, UK, and Cambridge Institute of Medical Research, Cambridge CB2 2XY, UK. E-mail: jt233@mole.bio.cam.ac.uk

Published: 22 February 2002

Genome ***Biology*** 2002, **3(3)**:comment2002.1–2002.5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/3/comment/2002

Abstract

Genome sequence comparisons reveal that some sets of genes are in similar linkage groups in different organisms while other sets are dispersed. Are some linkage groups maintained by chance, or is there an advantage to such an arrangement? Some insights may come from large clusters of genes, such as the major histocompatibility complex which includes many genes involved in immune defense.

With only rare exceptions, we all have the same order of genes along our chromosomes. Given that this order is maintained in members of the same species, to permit pairing at meiosis, is it of any other relevance? In other words, are genes arranged in optimal groupings, or could they just as well be scattered randomly, as long as they were coupled to the appropriate regulatory elements? Clustering of functionally related genes can be inferred from studies of conserved linkage groups in diverse prokaryote genomes [1], but is the same true of eukaryotes? Studies of transgenic animals reveal that some introduced genes become expressed in the appropriate tissues, but these experiments tell us nothing about the subtle advantages that may accrue from millions of years of chance reshuffling of the genome between speciation events. Evidence from a cursory comparison of the mouse and human genome sequences is consistent with at least one reordering of genes - one major break in synteny - occurring every million years. The enormous time scale of evolution means that selection can work on even very small margins, and a minor increase in fitness - say, 0.5% - can provide a significant long-term advantage. It therefore seems unlikely that gene order escapes optimization under the scrutiny of natural selection.

What sort of selective advantages can be proposed for gene clusters? Expression of genes at the appropriate place and time in development and differentiation could be coordinated by linkage, as it is in the *Hox* gene cluster for example [2]. Genes could also be linked to facilitate functional interaction of the products of polymorphic alleles (discussed below). A linked arrangement could facilitate sequence exchange, as occurs in gene conversion, when one continuous nucleotide stretch within the genome is replaced with a similar stretch from a related, non-allelic gene present in the same genome. In addition, a consistent order is essential for the assembly of somatically rearranged genes, such as those for immunoglobulins, T-cell receptors, or similar diversifying molecules such as the protocadherins [3]. Genes that are imprinted may also be tightly clustered, one of the best examples being the *Igf2* group of loci; in this case, clustering might facilitate the establishment and maintenance of the epigenetic marks that are crucial for imprinting [4].

The availability of multiple human genome sequences and the comparison of these with sequences from other vertebrate genomes will help to elucidate the significance of gene order on a wider scale. There is already evidence from such data that genes with high levels of expression are concentrated into genomic patches [5]. Genes encoding proteins of the immune system are perhaps of particular relevance, because they are constantly subject to intense selection for disease resistance as a result of interactions with pathogens. Some immune-system genes have undergone repeated duplication; some result from the innovative use of pre-existing gene modules encoding protein domains [6]; and some, such as the major histocompatibility complex (MHC), are extensively polymorphic. Plasticity in immune-system gene evolution may be essential for defense against pathogens that can themselves evolve extremely rapidly. This article considers some aspects of the evolutionary

history of gene clustering in the MHC and its consequences, and whether these insights can be extended to other parts of the genome.

Features of the MHC

MHC class I and class II molecules are expressed on antigen-presenting cells, where their role is to bind short peptides derived from pathogens. The peptides are presented at the cell surface to T cells, which have receptors that are produced by gene rearrangement; antigen presentation to T cells results in appropriate action being taken by the immune system in dealing with a pathogen. The MHC is characteristic of some sets of immune-system genes that are referred to as being in clusters, and 40% of expressed loci in the MHC - which spans around 4 megabases of the genome - are related to the immune system. These include multiple loci encoding antigen-presenting class I and class II MHC molecules, as well as several genes involved in processing the antigens for loading onto class I and class II molecules. As shown in Figure 1, the MHC includes genes for complement components (C2, C4 and factor B) as well as for molecules involved in modulating immune responses, such as tumour necrosis factor (TNF).

The genes for the class I and class II histocompatibility antigens are the most polymorphic in the genome; some of their loci are represented in the population by hundreds of alleles. This polymorphic variation is mainly restricted to changes in the peptide-binding grooves that are directly responsible for antigen presentation, although some areas of the genome immediately adjacent to the highly polymorphic loci are also extremely variable, a phenomenon that may be due to 'hitch-hiking' during evolution. To function effectively in defense against pathogens, MHC molecules must bind large numbers of peptides with high affinity but low specificity. Inspection of the codon usage over the variable exons of class I and class II genes shows a high ratio of non-synonymous to synonymous changes, consistent with selection accounting for the variation [7]. Alignment of multiple allelic sequences reveals a patchwork of changes, elements of which are shared in various combinations by different alleles. As the peptide-binding grooves of the encoded class I and class II molecules are made up of pockets that bind various peptide side-chains, short sequences corresponding to pockets of different shape and charge may be shuffled amongst different allotypes or isotypes.

Various mechanisms have been proposed to account for variation within the MHC peptide-binding groove sequences [8]. There is little, if any, evidence for markedly increased mutation frequency as a factor in promoting the high level of variation, and inspection of aligned sequences is not consistent with frequent point mutation. This is to be expected, given that mutations could throw up a pocket that is not suitable to accommodate a range of appropriate peptides. Other proposed mechanisms all involve sequence exchange, which can be achieved by allele conversion (double crossover at meiosis) or recombination (single crossover) within a locus. Exchange of sequences between different loci, which has been demonstrated conclusively in mouse class I sequences, most likely involves gene conversion. There is some disagreement about whether gene conversion would have a homogenizing or a heterogenizing influence on allele diversity.

The reason for the high level of variation in MHC sequences is generally assumed to entail selection for disease resistance. Ignoring genetic drift, two main selection mechanisms - which are not mutually exclusive - have been proposed: frequency-dependent selection, in which the fitness of one phenotype is dependent on the relative frequency of other phenotypes in the population; and heterozygote advantage, or balancing selection, in which heterozygotes have a greater fitness than either of the respective homozygotes [9]. The MHC region is linked to more diseases than any other section of the genome, including most, if not all, autoimmune conditions, but there is in fact limited evidence for a strong link with infectious disorders, and the diseases responsible for selection have not been identified. Although this appears surprising at first, it could be explained by some models of the role of the MHC in 'herd resistance' to infections, where resistance is dependent on the presence of a pool of variation across the population rather than on a single resistant allele. Other studies have come up with entirely unexpected explanations of selection for variation in the MHC, which rely on there being general benefits of population diversity. For example, mice may be able to maximize out-breeding by choosing mates that are as different as possible from themselves, based on the smell of urine, which in turn depends on the mate's MHC (H-2) allele. Recent data suggest a related phenomenon in sticklebacks, which attempt to maximize the number of class I loci in a mate [10]. In this regard, it is interesting to note that some fish have an arrangement of class I genes in a reiterated string that could facilitate rapid expansion within the genome [11].

The MHC is characterized by regions of linkage disequilibrium (LD), where alleles are most frequently co-inherited without recombination between generations, and which can extend for several million base-pairs. The stretches of the genome with high LD, separated by short recombining intervals, have been referred to as 'polymorphic frozen blocks' [12]; recombination is then restricted to 'hotspots' [13]. It has recently been appreciated that a similar modular structure may apply to the genome in general [14]. The now 'lumpy' model of the genome as a series of conserved haplotypes is more amenable to mapping multigenic disorders, because there are fewer possible haplotypes in the population and diseases are linked to longer genomic regions - although these longer regions also make more difficult the precise identification of the individual polymorphisms underlying a particular disease.

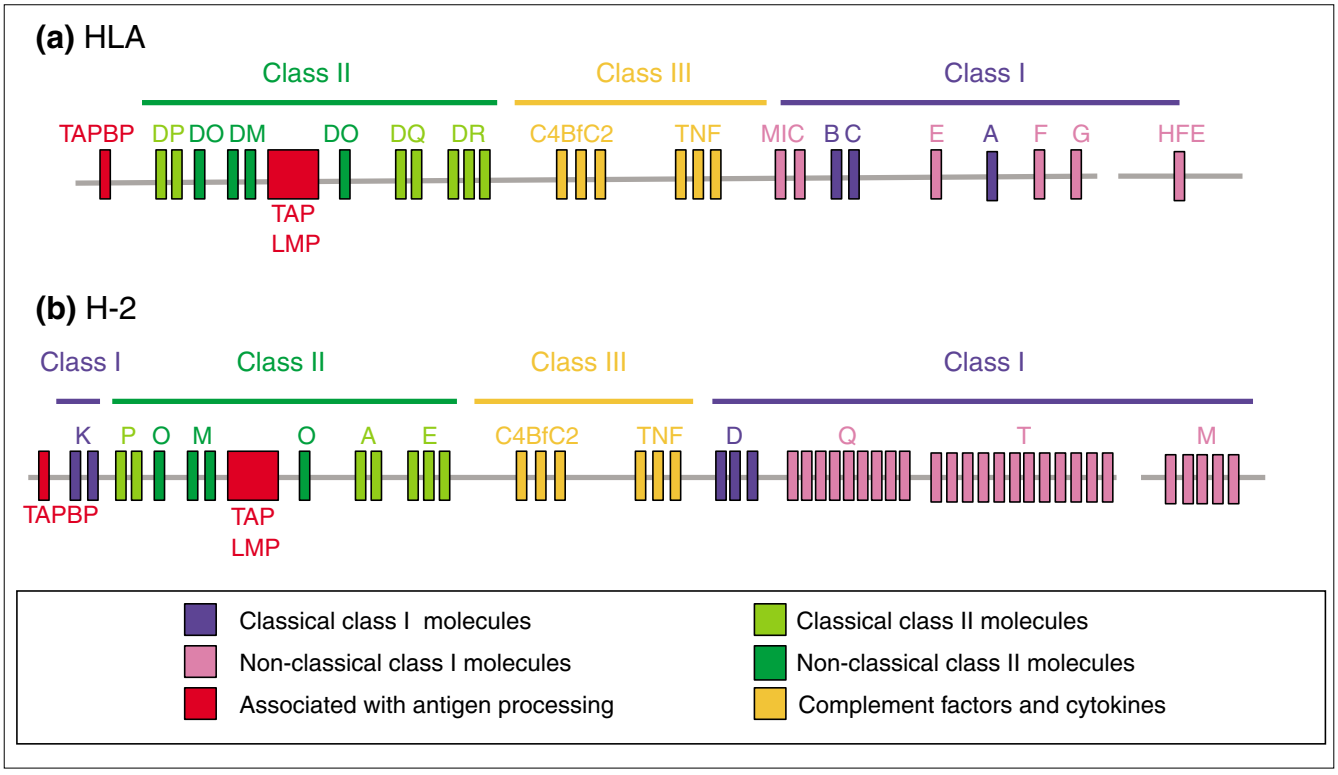

Figure 1
The MHC regions of **(a)** humans (HLA) and **(b)** mice (H-2). Only some of the key immune-system genes are shown, out of the more than 200 loci within the MHC. Historically, the MHC has been divided into three regions - class I, class II and class III. Most of the important phenotypes, such as transplant rejection, are associated with the class I and class II loci; these genes are highly polymorphic and some of them have over 300 alleles. Bf, C2 and C4 are complement proteins; LMP, low-molecular weight polypeptide; MIC, MHC class I chain-related; TAP, transporter associated with antigen-processing; TAPBP, TAP-binding protein (tapasin); TNF, tumor necrosis factor.

Gene clustering in the MHC

There are several examples of sets of MHC genes for which linkage may be functionally advantageous. The first example operates at a local level, with just two adjacent loci. The human HLA-DQ molecule is a heterodimer of two polymorphic chains, DQα and DQβ. In an individual heterozygous for both the *DQA1* (α) and *DQB1* (β) genes, each α chain should in principle be able to interact with a β chain encoded by the same chromosome (in a *cis* heterodimer) or by the other chromosome (in a *trans* heterodimer), making up four distinct DQ molecules. Not all α and β chains pair efficiently, however. The β chains from *DQw1* haplotypes do not form stable cell-surface heterodimers with *DQw2-4* α chains, and the converse is also true, namely that α chains from *DQw1* do not pair with β chains from the other haplotypes. In most populations studied to date, *DQA* and *DQB* alleles encoding these unstable combinations are rarely found on the same chromosome [15]. In a study of 2,807 chromosomes in 15 diverse populations, no 'forbidden' *cis* combinations were observed. The absence of such haplotypes, with a few exceptions confined to isolated populations, has been confirmed in a number of studies. There are two suggested interpretations of these findings: firstly, that recombination does not occur between *DQA* and *DQB* loci from *DQw1* and *DQw4* haplotypes; or secondly, that recombination does occur but that any resulting chromosomes carrying a mixed haplotype are purged from the population by negative selection. The two loci are less than 50 kilobases apart, so recombination should occur infrequently even on a random basis.

A second example of an advantageous closely linked gene arrangement concerns the *TAP1*, *TAP2*, *LMP2* and *LMP7* loci, which form a tight cluster in the class II region of the MHC. The *LMP2* and *TAP1* loci are arranged head-to-head (5′-to-5′) and are separated by less than 500 base-pairs. The products of these genes are needed to provide peptide antigens for loading onto class I molecules. The expression of the two genes is coordinately controlled from a shared bi-directional promoter, confining them to tight linkage. In addition, there is evidence of a possible functional explanation for the linkage of these antigen-processing genes with those encoding class I molecules. In the rat, both class I molecules and one of the *TAP* genes, *TAP2*, are highly polymorphic; some *TAP2* alleles differ by around 30 amino acids. RT1A class I molecules from rat (a) haplotypes are better at binding peptides with a particular amino-acid constitution,

characterized by the carboxy-terminal residue which is either neutral or positively charged. *TAP*-encoded transporters from these haplotypes are invariably of a type that preferentially provides such peptides. Similarly, class I molecules of the (b) haplotype prefer peptides with a neutral carboxyl terminus and this is mirrored by the type of peptide transported by the TAP2 allele linked to this haplotype in *cis*. This kind of mutually beneficial arrangement is not found in other mammalian species studied, where there are generally few changes in *TAP* alleles, the products of which are functionally identical; this does not necessarily negate the findings in rats but it does suggest that not all species have exploited such an arrangement.

Are there similar reasons to explain why there are so many unrelated (at the sequence level) immune-system genes concentrated in one region of the genome? It has been suggested that allelic variants of some of the genes in the class III region of the MHC, such as TNF, could compensate for overreaching immunostimulatory effects of the products of some of the class I or class II alleles alongside which they are inherited. Indeed, there is a significant group of genes, broadly categorized as modulators of inflammation, that could mediate such a function; these have been designated class IV of the MHC by one group of researchers [16]. This idea resonates well with the extended LD in the MHC, which could act to preserve combinations of alleles of different genes that work well together. It could also explain why the three sets of most highly polymorphic loci - classical class I, class II and the class I-related MIC genes - are all linked. MHC genes might also be clustered to facilitate regulation of expression at a broader level. In support of this idea, it has been observed that the nuclear multi-protein complexes known as promyelocytic leukemia (PML) bodies associate specifically with the MHC chromosomal region in interphase nuclei [17]; this was found to be the case even when the MHC region was translocated to a different chromosome from usual.

Other observations are not fully compatible with the idea that class I and class II linkage is of any biological relevance, however. For example, although all jawed vertebrates studied so far appear to have an MHC, the class I and class II genes are not clustered in one of the largest vertebrate groups, the bony fish. All bony fish species studied so far are exceptions to the canonical MHC arrangement, with class II genes lying outside the main group. There is evidence of a primitive synteny, however, as class I and class II genes are linked in cartilaginous fish such as the nurse shark [18]. It has been proposed that adaptive immunity evolved rapidly in the 30 million years after the emergence of the jawless fish (470 million years ago) and before the emergence of jawed vertebrate fish (440 million years ago; see [19] for review). This would mean that the system evolved at the same time as exposure to novel pathogens as a consequence of the new, jawed-predator lifestyle. The simplest, but not the only, interpretation of the comparative evolution of the MHCs of different species is that clustering of class I and II loci was lost in teleost fish but has remained in other species, including amphibians and birds, for over 400 million years (see [20] for further discussion of this issue).

Other clusters of immune-system genes

In addition to the MHC, other large clusters of immune-system genes are prominent in the human genome. Some are the result of repeated duplication and divergence, which, for trivial reasons, results in linked clusters of related genes. Examples might include the *KIR* superfamily (encoding receptors on natural killer (NK) cells) on chromosome 19q; its genes, like those of the MHC, are polymorphic [21]. Multiple related loci spread up the long arm of the chromosome, including the *LILR*, *SIGLEC* and *CD66* families. Other immunoglobulin superfamily clusters include some sets of cadherin genes that are involved in a novel form of somatic rearrangement [3]. Human chromosome 1 bears large numbers of genes encoding proteins from different branches of the immunoglobulin superfamily, such as a variety of Fc receptors (which mediate the internalization of antibody-containing complexes) as well as class-I-related proteins and NK cell receptors.

Preliminary analysis of some duplicated regions suggests that they are targets for rapid evolutionary turnover [22]. Their dynamic nature provides the possibility of creating fusion genes from juxtaposed 'cassettes'; a continuously reiterated string of highly related genes may also be prone to genetic loss by non-reciprocal recombination. These may be important forces in the evolution of the human genome, particularly in regions such as the *KIR* complex, where there is considerable plasticity in gene arrangement between individuals; the *KIR* loci must have evolved recently as they are restricted to primates. Interestingly, the functionally equivalent murine *Ly49* genes are similarly arranged in a string of loci, the composition of which differs between mouse strains.

The advantages of clustering may be more marked in sets of polymorphic immune-system genes, which are subject to selection for disease resistance, than in most of the genome. As pointed out by Fisher many years ago [23], LD may be facilitated by gene clustering. Inspection of the human genome map reveals a number of potential immune-gene clusters that, like the MHC, could be subject to inheritance as functional haplotypes. The bunching of the genes that regulate the development of tolerance in the immune system is a particular problem in mapping candidate autoimmune loci.

Implications for genomic studies

The first flush of sequence information has been from index species, such as human, mouse and *Drosophil*a. Comparing sequences across many phyla will be invaluable for identifying

conserved elements, such as upstream control regions, as well as conserved linkage groups. It will also be of immense interest to re-sequence some polymorphic regions, such as the MHC, from different individuals. Very few studies of this type have been done, but they will be essential for effective population-based disease-gene mapping [14], especially in regions such as the MHC that are associated with so many diseases. The human MHC sequence available to date has been compiled from different genomic libraries, and future efforts should aim to provide a number of distinct single-haplotype sequences.

The available genome sequences already tell us something about the mechanisms for generating new gene arrangements, however [24]. Ohno proposed tandem duplication and divergence [25], but outside the 'immunogenome' these may not be as common as retrotransposition, which may have been an important mechanism in generating some of the 1,000 or so olfactory receptor genes, for example. Transfer of whole blocks of sequence to new sites has been identified, again using olfactory genes as an example [26]; inversions may also play a major part in sculpting the eukaryotic genome, as in the mouse *t* complex [27,28].

The sequence suggests that the human genome is an untidy mess, but this is possibly illusory. Most teenagers are aware of the dubious advantages of room tidying, and the socialite Quentin Crisp famously remarked that there was no need to do any housework, because after four years the dirt didn't get any worse [29]. By analogy with some of our inner cities, respectable genomic loci live side by side with litter and burnt-out wrecks of pseudogenes and junk DNA. There is a fundamental difference between the untidiness we make and the clutter within our genomes, however: litter is random but genes are arranged in the same, identical 'untidy' order in every one of us. A more apt metaphor for the genome is perhaps the computer disc, which can hold dispersed fragments of data; it can be periodically defragmented to optimize the data and at the same time can be unique yet endlessly and precisely copied. Until we understand the processes that have determined the ordering and clustering of genes in genomes our understanding of gene regulation and evolution will be incomplete.

References

1. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
2. Zakany J, Kmita M, Alarcon P, de la Pompa JL, Duboule D: **Localized and transient transcription of *Hox* genes suggests a link between patterning and the segmentation clock.** *Cell* 2001, **106:**207-217.
3. Wu Q, Maniatis T: **A striking organization of a large family of human neural cadherin-like cell adhesion genes.** *Cell* 1999, **97:**779-790.
4. Reik W, Bowden L, Constancia M, Dean W, Feil R, Forne T, Kelsey G, Maher E, Moore T, Sun FL, Walter J: **Regulation of Igf2 imprinting in development and disease.** *Int J Dev Biol* 1996, **Suppl:**53S-54S.
5. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, *et al.*: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291:**1289-1292.
6. Liu Y, Shaw S: **The human genome: an immuno-centric view of evolutionary strategies.** *Trends Immunol* 2001, **22:**227-229.
7. Ohta T: **Effect of gene conversion on polymorphic patterns at major histocompatibility complex loci.** *Immunol Rev* 1999, **167:**319-325.
8. Marsh SGE, Parham P, Barber LD. *The HLA Factsbook*. Academic; 2000, 398.
9. Parham P, Ohta T: **Population biology of antigen presentation by MHC class I molecules.** *Science* 1996, **272:**67-73.
10. Reusch TB, Haberli MA, Aeschlimann PB, Milinski M: **Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism.** *Nature* 2001, **414:**300-302.
11. Clark MS, Shaw L, Kelly A, Snell P, Elgar G: **Characterization of the MHC class I region of the Japanese pufferfish (*Fugu rubripes*).** *Immunogenetics* 2001, **52:**174-185.
12. Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez P, Kulski J: **Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease.** *Immunol Rev* 1999, **167:**275-304.
13. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the MHC.** *Nat Genet* 2001, **29:**217-222.
14. Goldstein DB: **Islands of linkage disequilibrium.** *Nat Genet* 2001, **29:**109-111.
15. Begovich AB, Klitz W, Steiner LL, Grams S, Suraj-Baker V, Hollenbach J, Trachtenberg E, Louie L, Zimmerman PA, Hill AVS, *et al.*: **HLA-DQ haplotypes in 15 different haplotypes.** In *MHC Evolution, Structure and Function.* Edited by Kasahara M. Tokyo: Springer; 2000: 412-426.
16. Gruen JR, Weissman SM: **Evolving views of the MHC.** *Blood* 1997, **90:**4252-4265.
17. Shiels C, Islam SA, Vatcheva R, Sasieni P, Sternberg MJ, Freemont PS, Sheer D: **PML bodies associate specifically with the MHC gene cluster in interphase nuclei.** *J Cell Sci* 2001, **114:**3705-3716.
18. Ohta Y, Okamura K, McKinney EC, Bartl S, Hashimoto K, Flajnik M: **Primitive synteny of vertebrate major histocompatibility complex class I and class II genes.** *Proc Natl Acad Sci USA* 2000, **97:**4712-4717.
19. Shand R, Dixon B: **Teleost MHC genes: diverse but not complex.** *Mod Asp Immunobiol* 2001, **2:**66-77.
20. Flajnik MF, Kasahara M: **Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system.** *Immunity* 2001, **15:**351-362.
21. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, Beck S, Trowsdale J: **Plasticity in the organization and sequences of human KIR/ILT gene families.** *Proc Natl Acad Sci USA* 2000, **97:**4778-4783.
22. Eichler EE: **Recent duplication, domain accretion and the dynamic mutation of the human genome.** *Trends Genet* 2001, **17:**661-669.
23. Fisher RA: *The Genetical Theory of Natural Selection*. Oxford: Clarendon; 1930.
24. Green ED, Chakravarti A: **The human genome sequence expedition: views from the "base camp".** *Genome Res* 2001, **11:**645-651.
25. Ohno S: *Evolution by Gene Duplication.* Spinger: Berlin; 1970.
26. Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, *et al.*: **Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements.** *Am J Hum Genet* 2001, **68:**874-883.
27. Silver LM: **Genetic organization of the mouse t complex.** *Cell* 1981, **27:**239-240.
28. Huynen MA, Snel B, Bork P: **Inversions and the dynamics of eukaryotic gene order.** *Trends Genet* 2001, **17:**304-306.
29. Crisp Q: *The naked civil servant*. Penguin; 1968.