Opinion

Polymorphism in regulatory gene sequences

N A Mitchison

Address: Department of Immunology, Windeyer Institute of Medical Science, University College London Medical School, Cleveland Street, London W1P 6DB, UK. E-mail: n.mitchison@ucl.ac.uk

Published: 20 December 2000

Genome ***Biology*** 2000, **2(1)**:comment2001.1–2001.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/2/1/comment/2001

Abstract

The extensive polymorphism revealed in non-coding gene-regulatory sequences, particularly in the immune system, suggests that this type of genetic variation is functionally and evolutionarily far more important than has been suspected, and provides a lead to new therapeutic strategies.

Considerable attention recently has focused on polymorphisms and their potential subtly to alter protein function in ways that might prove biologically or clinically important. But increasing numbers of polymorphisms are also being identified in the regulatory regions of genes, and here I consider the significance of these. Reflecting my perspective as an immunologist, this article takes as its starting point the posy of recently discovered polymorphisms associated with autoimmunity, atopy and resistance to infection that are shown in Table 1. I use the word 'posy' to indicate that, in spite of my effort to make the survey comprehensive, it may well be incomplete. To make sense of the distribution of polymorphisms, I focus here on 'introvert' genes, which have been defined as genes encoding proteins that handle self molecules, as opposed to 'extrovert' genes, which encode proteins that handle foreign molecules that enter the body from outside [1,2]. In the former, non-coding variation predominates, as distinct from extrovert genes, in which the reverse applies, at least in the sample surveyed here.

The collection of genes in Table 1 is of miscellaneous origin. Some members, such as the polymorphisms in the major histocompatibility complex (MHC) class II promoters, emerged from a decades-long search for the genes responsible for variation in immune function and disease susceptibility. Tpm1 (T cell phenotype modifier-1) is a major determinant of Th1/Th2 balance of T-helper-cell subsets that falls into the same category. Others, such as the genes for interleukin (IL)-1, IL-4, IL-10 and tumor necrosis factor-α (TNFα), were first identified because of their functional importance, and were then scrutinized for allelic variations associated with disease. Other instances of polymorphism - in interferon γ (IFNγ) and IL-5Rα, for example - were first identified by microsatellite-based genome searches (quantitative trait loci analysis). This last category is the one most likely to grow in future as it has no prior bias and, therefore, provides objective sampling of disease-associated genes [3]. It has been validated by the concordance between searches carried out in humans and animals [4]. The *ACE1* gene, encoding the angiotensin I-converting enzyme involved in blood pressure regulation, is also included in Table 1 because the biological impact of its polymorphisms are so well understood. Substitutions in the regulatory regions of *ACE1* are known to have a larger effect than those in the coding regions.

The preponderance of non-coding polymorphism

Non-coding polymorphisms make up a clear majority of the polymorphisms listed in Table 1. Furthermore, the coding polymorphisms include some special cases. Thus, that in IL-2 is of minor interest because of its limited disease association. Those in Tpm1 and TGFβ are, in a sense, regulatory, as discussed below. The coding polymorphisms of the chemokine receptors were presumably selected by infection [5], and thus fall into the extrovert category that has otherwise been excluded from this survey.

Some, but not all, of the examples shown in Table 1 have a functional phenotype, defined by an effect on the level of transcription and/or translation, and are thus known to be regulatory. The various types of regulatory variation are

Table 1

Disease-associated polymorphisms

Protein affected	Disease	Non-coding/coding (N-C/C)	Functional phenotype	Human/animal (H/A)
IL-1/IL-1RA	Periodontal [17], osteoarthritis [18] and others	N-C 5'	Yes	H
IL-2	IDDM [19]	C	No	A (not H)
IL-4	Asthma [20]	N-C 5' and intron	No	H
IL-5Rα	Atopy [21]	N-C 5' and splice variant	Yes	A
IL-6	Juvenile idiopathic arthritis [22]	N-C 5'	Yes	H
IL-10	SLE [23]	N-C 5'	Yes	H
IL-13	Asthma [24]	C	No	H
IFNγ	Asthma and atopy [25]	Intron	No	H
TNFα	RA, SLE and others (Several groups)	N-C 5'	Yes	H
TGFβ	Transplant rejection [26]	C (signal sequence)	Yes	H
FcγRIIB1	SLE [27]	N-C 5'	Yes	A
Insulin	IDDM [28]	N-C 5'	Yes	H
RANTES	Atopic dermatitis [29]	N-C 5'	No	H
Nramp1	HIV, TB, JRA [30]	N-C 5'	Yes	H, A
Nramp2	Anemia [31]	C and intron	Yes	H, A
ACE1	Hypertension [32,33]	N-C 5' and intron (and C)	Yes	H, A
CCR5, CCR2	AIDS [34] AIDS [5]	N-C 5' C	Yes No	H
MHC class II	Many diseases Collagen-induced arthritis [4]	C N-C 5'	Yes Yes	H
IL-12R/Tpm1 (transcription factor?)	Cutaneous leishmaniasis [35]	C?	Yes	A

shown in Figure 1. Most of the examples of *cis*-regulatory variation occur upstream of the coding sequence, as shown in the figure. TGFβ provides an example of the less common, coding form of *cis*-regulatory polymorphism that occurs in the signal sequence; this gene has other *cis*-regulatory variants upstream, which have a weaker phenotype. In addition, as is thought likely in the case of Tpm1, a polymorphism can be *trans*-regulatory if it occurs in the coding sequence of a transcription factor. Mutations, rather than polymorphisms, of this type cause rare congenital diseases such as cleidocranial dysplasia (OMIM #119600) and Rubinstein syndrome (OMIM #180849) [6], by mutations in the CBFA1 transcription factor and the CBP cotranscription factor, respectively. In some patients with cleidocranial dysplasia, such mutations have not been detected, and these cases may possibly result from mutation in the CBFA1 binding site, which would put them into the *cis*-regulatory category.

Most of the examples shown in Table 1 involve the orchestration of the immune response by tightly regulated cytokines and their receptors, where regulatory polymorphism might be expected to play an unusually important part. Reassuringly, increasing evidence indicates that *cis*-regulatory variation also predominates in the evolution of body shape, as recently reviewed [7]. Carroll's review [7] cites the central role of *cis*-regulatory elements in the evolution of modern maize from its ancient progenitor teosinte - as predicted [1].

It is striking how rich the immune system turns out to be in introvert polymorphism, considering the great importance to it of extrovert polymorphism - most notably in the MHC. Indeed, the genetics of the immune system was concerned in the past almost exclusively with the extrovert function of antigen-recognition. Furthermore, genetic variation of the extrovert type is a major feature of the evolution of the host-parasite relationship, extending far beyond the immune system. In the well-studied case of resistance to malaria, it includes the genes encoding the hemoglobins, glucose-6-phosphate dehydrogenase, and the Duffy blood group. Further examples of parasite-selected coding polymorphisms

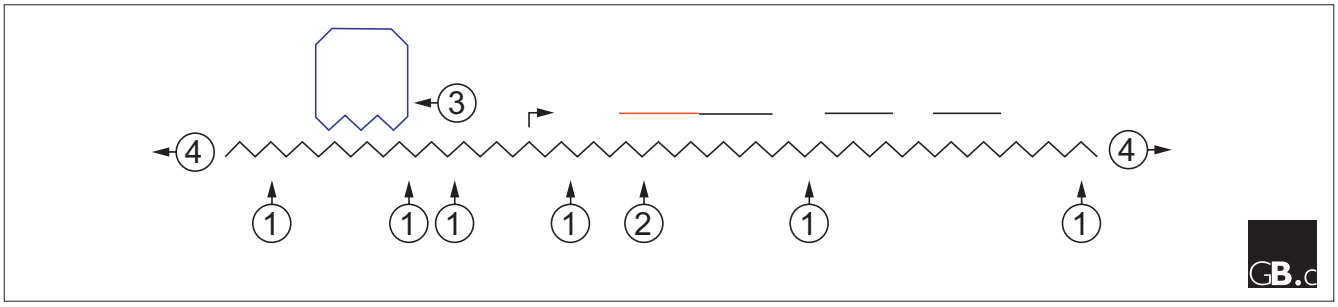

Figure 1
The types of regulatory genetic variation so far found. (1) *cis*-regulatory, non-coding; (2) *cis*-regulatory, coding; (3) *trans*-regulatory, coding or non-coding (although no instance of polymorphism in the expression of a transcription factor has yet been reported); (4) *cis*-regulatory, non-coding, very long distance (position effects). The transcription factor is shown in blue, signal sequence shown in red.

can be expected, for instance in the host proteins that are exploited by intracellular bacteria such as *Listeria* and *Salmonella*. Extrovert genes vary not only between alleles (for example, the MHC and the nutrient-handling plant allozymes) but also between gene duplicates (as in lymphocyte receptors for antigens, olfactory receptors and plant avirulence receptors). Yet in spite of this remarkable range, it is hard to believe that the extrovert genes comprise more than a minor part of the total genome.

Heterozygote advantage - or not?

To what extent are polymorphisms in the regulatory regions subject to natural selection? They could in principle be neutral, transient (reflecting allele replacement or population mixing) or balanced (through increased fitness of the heterozygote). Balanced polymorphism is potentially valuable for the insight it provides into therapeutic approaches. Ideas that have been proposed for balance of non-coding polymorphisms include: a low rate ACE promoter for regulation of systemic blood pressure and a high-rate one for local wound healings; a constitutive insulin promoter for negative selection in the thymus and a regulated one for insulin production in the pancreas; a low-rate cytokine promoter for resistance to some types of parasite and a high-rate one for resistance to others; and slow-rate MHC class II promoter for T-cell Th2 responses and a high-rate one for Th1 responses. The common theme is that the alleles are expressed in different circumstances. As differentiation proceeds, one allele gets transcribed in one type of cell, while the other gets transcribed in another. These ideas seem reasonable enough, but so far lack firm support. Little is known about differential allelic transcription in different cell types. The various mechanisms proposed would all provide selective advantage for heterozygotes, and would thus increase the flexibility of the system.

The relative fitness of heterozygotes has been a major theme of population genetics at large. As shown in Figure 2, genetic variation of any trait can be plotted as a bell-shaped curve that relates the magnitude of the trait to the frequency of individuals. At the curve's center, the genetics are predominantly multifactorial and, on the basis of the present data, predominantly within *cis*-regulatory sequences. At its tails, consideration of Mendelian inheritance in man suggests that the reverse tends to apply. Furthermore, at the center the variation is likely to result substantially from polymorphism (that is, substitutions where the frequency of the rare allele exceeds 1% and which cannot therefore result simply from mutation). The polymorphisms will, in part, be sustained by heterozygote advantage; the extent to which they do so is unknown, but the important point here is that the stronger the heterozygous phenotype, the higher the likelihood that balancing selection is involved. The role of heterozygote fitness at the tails of the curve is more complicated. Some of the more common monofactorial disease genes are thought to spread as a result of heterozygote advantage. Well known examples are the hemoglobin and glucose-6-phosphate dehydrogenase alleles that mediate resistance to malaria, and the cystic fibrosis mutations that do the same for enteric infection [8-10]. My colleagues and I have argued from the preponderance of immunodeficiencies that are X-linked that these examples may be misleading and that the majority of 'recessive' autosomal monofactorial deficiencies may in fact have heterozygotes with reduced fitness [11]. Our reasoning is that at long-standing population equilibrium, the loss of even slightly disadvantageous autosomal alleles would have a significant impact because of the relatively high frequency of heterozygotes, and this in turn should reduce the frequency of the homozygotes responsible for the disease. This effect would not apply to X-linked disease, where heterozygous females are only slightly more common than diseased males, and would thus explain the predominance of X-linkage among the deficiency genes. The question remains open, as once again experimental and epidemiological evidence is lacking. What a shame that more attention is not paid to these heterozygotes.

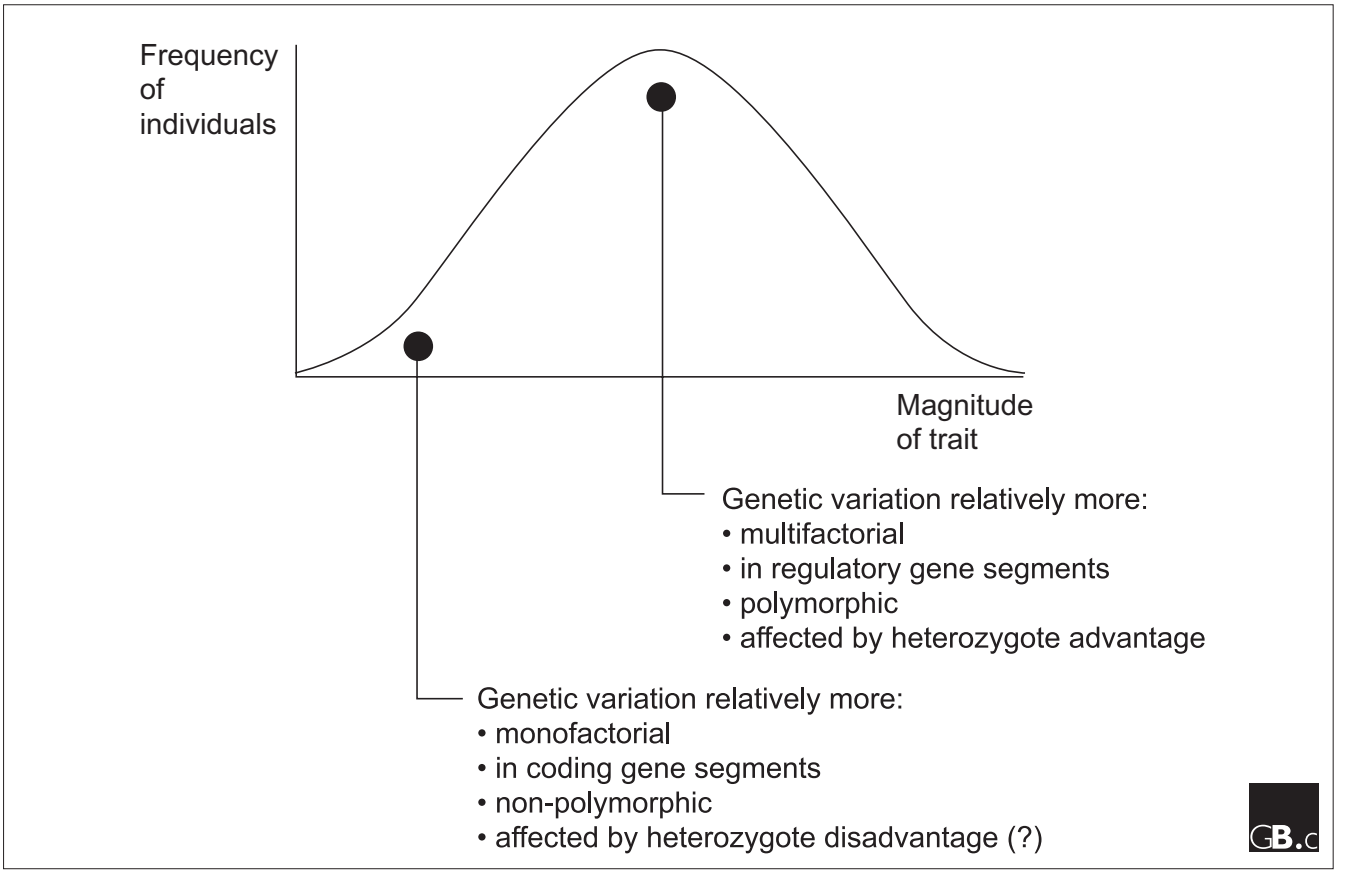

Figure 2
Genetic variation in relation to the bell-shaped curve. The biases shown are discussed in the text in relation to the working of the immune system, but may apply more widely.

The information obtained from disease associations may be compared with the much larger body of sequence comparisons emerging from studies using DNA microarrays, which do not (yet) relate to gene function. A recent study [12] found sequence diversity to be almost identical for coding and non-coding regions, but with over twice as much silent as replacement substitution in the coding sequences. These newer data give the same overall impression as the disease associations surveyed here, even though the DNA microarray screen did not distinguish between regulatory and non-regulatory sequences. The study also made comparisons with sequence data from apes, but did not compare coding with non-coding regions. A 15-fold variation in nucleotide diversity across genes (coefficient of variation was 74% for non-coding segments) was noted. This opens up a reasonable prospect of finding functionally significant 'nests' of single nucleotide polymorphisms (SNPs) such as are present in MHC class II promoters [12] - sequences that overlap transcription factor binding sites and that are rich in SNPs. Of particular interest are transcription factor recognition elements such as the cAMP response element (CRE), which occurs in many genes, allowing comparisons of its level of polymorphism in differing circumstances. CRE activates transcription of target genes in response to a diverse array of stimuli, including peptide hormones, growth factors and neuronal activity. Our published [13] and unpublished data on the CRE sequences in MHC gene promoters indicate that polymorphism varies significantly in level between genes with different functions and, as expected, tends to localize not so much in the CRE sequence itself as in its immediate neighborhood.

Care is called for in interpreting these evolutionary patterns. One cannot just extrapolate from a snapshot of contemporary intraspecific variation, even though it is this that provides much of the raw material for long-term change. It may be reassuring to find the importance of *cis*-regulatory elements repeating itself in the evolution of mice, maize and the vertebrate body, as emphasized here, but that does not mean that the course of evolution is so simple.

New avenues for therapy

The importance of non-coding polymorphism sends a strong message to the therapeutic strategists: search the genome for sites of high non-coding polymorphism, identified as

SNP nests, by whatever tools come to hand. They will tell you where nature has found a way of intervening at an important checkpoint. Follow her lead! But it is not quite that simple. The counter argument is that nature tends to conserve functionally important sequences. So the art of therapeutic strategy will be to strike the right balance between these two messages. It would make things easier if we could point to some really important checkpoint that has already been identified by genetics alone. That has not yet been achieved, but current progress in genomics tells us that we may not have long to wait.

Hitherto the collection of human polymorphism data has been a sort of cottage industry, where various groups have chosen to focus on different genes without much rhyme or reason. A more systematic approach can now be formulated, based on whole-genome sequencing. Before long the mouse, and later the chimpanzee, genome will be sequenced. Then the following scenario could be applied: First, proximal promoters in the human genome would be identified via the Eukaryotic Promoter Database [14,15] and by other means. Next, upstream *cis*-regulatory sequences would be identified via their conservation between mouse and human and by other means [16]. Conservation between chimp and human may be too high for this kind of use. Finally, the *cis*-regulatory sequences identified in this way would then be scanned for divergence between human and chimpanzee, thus identifying the sites that have most responded to selective pressure. These would therefore specify candidate checkpoints of importance to the functioning of the body. But before setting out to exploit this information, it would be wise to find out whether these candidates were also sites of polymorphism in humans. Intraspecific and short-term interspecific variation may not always run in parallel, as mentioned above, but it will be reassuring when they do so. One suspects that divergence between mouse and man may be too high for this kind of use, as the differences driven by selection would be submerged in junk variation.

My study of progressive evolutionary divergence in MHC gene promoters in the series *Mus musculus* - *M. pahari* (approximately 4 million years) - rat (approximately 10 million years) - mole rat (over 10 million years) provides, I hope, a rehearsal of this scenario. One can watch the divergence evident in modern laboratory mice, which is located preferentially around the transcription binding sites, becoming gradually swamped by random divergence - presumably junk - as one moves back in time. For humans, the trick will be to find an interspecies comparison where the divergence in *cis*-regulatory sequences is high enough to be informative, without being swamped by the junk. With luck, the comparison with chimp may provide much of what is needed. If not, then another one or two primate genomes may be required - and the additional information that they could provide would surely justify the effort. One might expect the human-chimp comparison to tell us more about checkpoints in the brain than in the cardiovascular system, for instance. Because the genome will not behave in the same way throughout, having a few comparisons to choose from should, in any case, be useful. In summary, one can envisage a royal road running from comparative genomics to the identification of key checkpoints, and then leading on to novel drug discovery.

Acknowledgements

The Leverhulme and Wellcome Trusts supported this work.

References

1. Mitchison A: **Partitioning of genetic variation between regulatory and coding gene segments: the predominance of software variation in genes encoding introvert proteins.** *Immunogenetics* 1997, **46:**46-52.
2. Mitchison NA, Schuhbauer D, Muller B: **Natural and induced regulation of Th1/Th2 balance.** *Springer Semin Immunopathol* 1999, **21:**199-210.
3. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalayanaraman N, *et al.*: **Characterization of single-nucleotide polymorphisms in coding regions of human genes.** *Nat Genet* 1999, **22:**231-238.
4. Becker KG, Simon RM, Bailey-Wilson JE, Freidlin B, Biddison WE, McFarland HF, Trent JM: **Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases.** *Proc Natl Acad Sci USA* 1998, **95:**9979-9999.
5. Smith MW, Dean M, Carrington M, Winkler C, Huttley GA, Lomb DA, Goedert JJ, O'Brien TR, Jacobson LP, Kaslow R, *et al.*: **Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC), ALIVE Study.** *Science* 1997, **277:**959-965.
6. **Online Mendelian Inheritance in Man** [http://www.ncbi.nlm.nih.gov/Omim/]
7. Carroll SB: **Endless forms: the evolution of gene regulation and morphological diversity.** *Cell* 2000, **101:**577-580.
8. Ashley-Koch A, Yang Q, Olney RS: **Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review.** *Am J Epidemiol* 2000, **151:**839-845.
9. Ruwende C, Hill A: **Glucose-6-phosphate dehydrogenase deficiency and malaria.** *J Mol Med* 1998, **76:**581-588.
10. Pier GB, Grout M, Zaidi T, Meluleni G, Mueschenborn SS, Banting G, Ratcliff R, Evans MJ, Colledge WH: ***Salmonella typhi* uses CFTR to enter intestinal epithelial cells.** *Nature* 1998, **393:**79-82.
11. Mitchison NA, Muller B, Segal RM: **Natural variation in immune responsiveness, with special reference to immunodeficiency and promoter polymorphism in class II MHC genes.** *Hum Immunol* 2000, **61:**177-181.
12. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A: **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.** *Nat Genet* 1999, **22:**239-247.
13. Cowell LG, Kepler TB, Janitz M, Lauster R, Mitchison NA: **The distribution of variation in regulatory gene segments, as present in MHC class II promoters.** *Genome Res* 1998, **8:**124-134.
14. Perier RC, Junier T, Bonnard C Bucher P: **The Eukaryotic Promoter Database (EPD): recent developments.** *Nucleic Acids Res* 1999, **27:**307-309.
15. **Eukaryotic Promoter Database** [http://www.epd.isb-sib.ch]
16. Duret L Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7:**399-406.
17. Kornman KS, Crane A, Wang HY, di Giovine FS, Newman MG, Pirk FW, Wilson TG Jr, Higginbottom FL, Duff GW: **The interleukin-1 genotype as a severity factor in adult periodontal disease.** *J Clin Periodontol* 1997, **24:**72-77.

18. Moos V, Rudwaleit M, Herzog V, Hölig K, Sieper J, Muller B: **Association of genotypes affecting the expression of interleukin-1β or interleukin-1 receptor antagonist with osteoarthritis.** *Arthritis Rheum* 2000, **43:**2417-2422.
19. Denny P, Lord CJ, Hill NJ, Goy JV, Levy ER, Podolin PL, Peterson LB, Wicker LS, Todd JA, Lyons PA: **Mapping of the IDDM locus Idd3 to a 0.35-cM interval containing the interleukin-2 gene.** *Diabetes* 1997, **46:**695-700.
20. Chouchane L, Sfar I, Bousaffara R, El Kamel A, Sfar MT, Ismail A: **A repeat polymorphism in interleukin-4 gene is highly associated with specific clinical phenotypes of asthma.** *Int Arch Allergy Immunol* 1999, **120:**50-55.
21. Daser A, Koetz K, Bätjer N, Jung M, Rüschendorf F, Goltz M, Ellerbrok H, Renz H, Walter J, Paulsen M: **Genetics of atopy in a mouse model: polymorphism of the IL-5 receptor a chain.** *Immunogenetics* 2000, **51**:632-638.
22. Fishman D, Faulds G, Jeffery R, Mohamed-Ali V, Yudkin JS, Humphries S Woo P: **The effect of novel polymorphisms in the interleukin-6 (IL-6) gene on IL-6 transcription and plasma IL-6 levels, and an association with systemic-onset juvenile chronic arthritis.** *J Clin Invest* 1998, **102:**1369-1376.
23. Eskdale J, Kube D, Tesch H , Gallagher G: **Mapping of the human IL10 gene and further characterization of the 5' flanking sequence.** *Immunogenetics* 1997, **46:**120-128.
24. Heinzmann A, Mao X, Akaiwa M, Kreomer RT, Gao P, Ohshima K, Umeshita R, Abe Y, Braun S, Yamashita T, *et al.*: **Genetic variants of IL-13 signalling and human asthma and atopy.** *Hum Mol Genet* 2000, **9:**549-559.
25. Barnes KC, Freidhoff LR, Nickel R, Chiu YF, Juo SH, Hizawa N, Naidu RP, Ehrlich E, Duffy DL, *et al.*: **Dense mapping of chromosome 12q13.12-q23.3 and linkage to asthma and atopy.** *J Allergy Clin Immunol* 1999, **104:**485-491.
26. Hutchinson IV. **The role of transforming growth factor-beta in transplant rejection.** *Transplant Proc* 1999, **31:**9S-13S.
27. Jiang Y, Hirose S, Sanokawa-Akakura R, Abe M, Mi X, Li N, Miura Y, Shirai J, Zhang D, Hamano Y, Shirai T: **Genetically determined aberrant down-regulation of FcgammaRIIB1 in germinal center B cells associated with hyper-IgG and IgG autoantibodies in murine systemic lupus erythematosus.** *Int Immunol* 1999, **11:**1685-1691.
28. McGinnis RE, Spielman RS: **Linkage disequilibrium in the insulin gene region: size variation at the 5' flanking polymorphism and bimodality among "class I" alleles.** *Am J Hum Genet* 1994, **55:**526-532.
29. Nickel RG, Casolaro V, Wahn U, Beyer K, Barnes KC, Plunkett BS, Freidhoff LR, Sengler C, Plitt JR, Schleimer RP, *et al.*: **Atopic dermatitis is associated with a functional mutation in the promoter of the C-C chemokine RANTES.** *J Immunol* 2000, **164:**1612-1616.
30. Blackwell JM, Searle S: **Genetic regulation of macrophage activation: understanding the function of Nramp1 (=Ity/Lsh/Bcg).** *Immunol Lett* 1999, **65:**73-80.
31. Lee PL, Gelbart T, West C, Halloran C, Beutler E: **The human Nramp2 gene: characterization of the gene structure, alternative splicing, promoter region and polymorphisms.** *Blood Cells Mol Dis* 1998, **24:**199-215.
32. Villard E, Tiret L, Visvikis S, Rakotovao R, Cambien F, Soubrier F: **Identification of new polymorphisms of the angiotensin I-converting enzyme (ACE) gene, and study of their relationship to plasma ACE levels by two-QTL segregation-linkage analysis.** *Am J Hum Genet* 1996, **58:**1268-1278.
33. Challah M, Villard E, Philippe M, Ribadeau-Dumas A, Giraudeau B, Janiak P, Vilaine JP, Soubrier F, Michel JB: **Angiotensin I-converting enzyme genotype influences arterial response to injury in normotensive rats.** *Arterioscler Thromb Vasc Biol* 1998, **18:**235-243.
34. Martin MP, Dean M, Smith MW, Winkler C, Gerrard B, Michael NL, Lee B, Doms RW, Margolick J, Buchbinder S, *et al.*: **Genetic acceleration of AIDS progression by a promoter variant of CCR5.** *Science* 1998, **282:**1907-1911.
35. Guler ML, Gorham JD, Dietrich WF, Murphy TL, Steen RG, Parvin CA, Fenoglio D, Grupe A, Peltz G, Murphy KM: **Tpm1, a locus controlling IL-12 responsiveness, acts by a cell-autonomous mechanism.** *J Immunol* 1999, **162:**1339-1347.