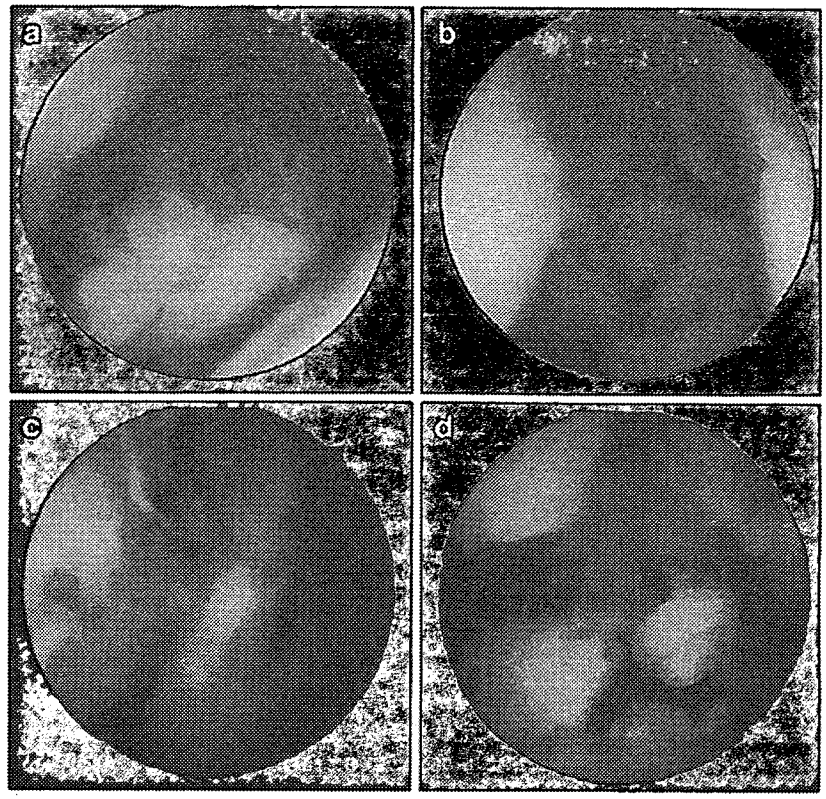


Fig. 2 Arthroscopic finding in shoulder arthroscopy. **a** Synovial proliferation with white fibrous tissue at rotator interval, **b** around the AIGHL, **c** the anterior limb was floated with synovium over the LHB, **d** subacromial bursa

four patients were given 8 mg of MTX (Table 1). Predonin (5 mg/day) was administered in three patients. Prednisolone sodium succinate (20 mg) was used as steroid cover after surgery in two patients. At arthroscopic synovectomy, general anesthesia was administered in the shoulder arthroscopic synovectomy, and lumber anesthesia was administered in the knee and the ankle joints. We used 4.0-mm arthroscope (Smith & Nephew, USA) for the knee and the shoulder joints and 2.7-mm arthroscope (Smith & Nephew) for the ankle joints. We also used shaver apparatus (Smith & Nephew) and VAPR (Johnson & Johnson, USA) to remove synovium for arthroscopic synovectomy.

We observed synovium in articular joints directly by arthroscopy in knee, shoulder, and ankle to analyze how and where synovium proliferate in joints during infliximab treatment. Furthermore, we investigated the change of CRP (mg/dl), DAS28 (CRP), ACR 20, ACR50, ACR70 before and after arthroscopic synovectomy. For data analysis, we used Wilcoxon test of nonparametric statistic test to compare the CRP, DAS28, ACR20, ACR50, ACR70 before and 6 and 50 weeks after surgery.

Results

In the knee arthroscopic finding, we detected high vascular synovium in patellofemoral joint (Fig. 1a) and between the disc and tibial cartilage (Fig. 1b). We shaved and removed the synovium (Fig. 1c), and the joint cavity was made clear and clean after arthroscopic synovectomy (Fig. 1d). In the

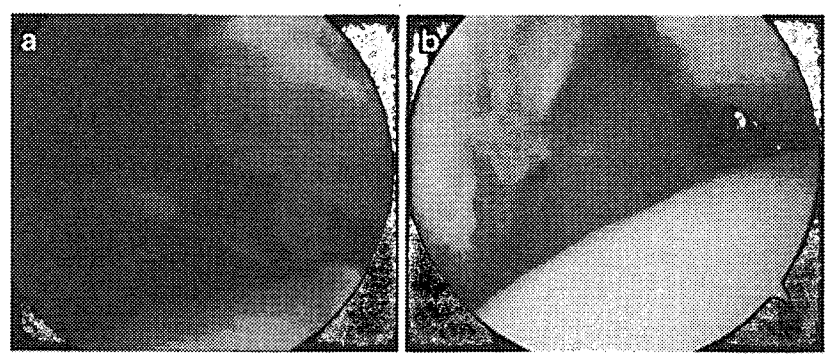


Fig. 3 Arthroscopic finding in ankle arthroscopy. **a** Synovium proliferation between tibiofibular joint, **b** after arthroscopic synovectomy

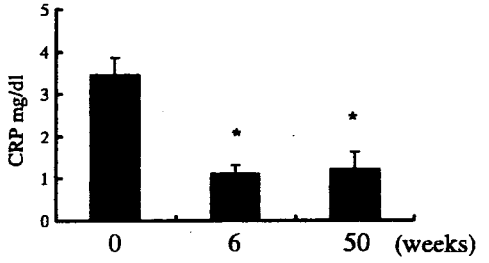


Fig. 4 Serum CRP changes after 0, 6, and 50 weeks by arthroscopic synovectomy

shoulder arthroscopy, we found synovial proliferation with white fibrous tissue at rotator interval (Fig. 2a) and around the anteroinferior glenohumeral ligament (AIGHL) (Fig. 2b). The anterior limb was floated with synovium over the long head of biceps tendon (LHB) (Fig. 2c). The difference between the knee and the shoulder synovium is that shoulder synovium has white fibrous tissue to induce the joint impingement in the glenohumeral joint. In subacromial bursa, the fatty tissue was increased in bursal synovium (Fig. 2d), and we removed those tissues by shaver and VAPR without bony resection for arthroscopic debridement. In ankle arthroscopic finding, we detected synovium proliferation between tibiofibular joint that invades the erosion of tibia (Fig. 3a), and we removed those tissues clearly to see joint cartilage not to induce impingement (Fig. 3b). The average of CRP at preoperation, 3.45±0.4 (2.7–5.6) mg/dl, was changed to 1.12±0.2 (0.6–1.8) mg/dl at 6 weeks later and improved to 1.22±0.4 (0.8–1.9) mg/dl at 50 weeks later. Therefore, the CRP was improved after 6 weeks and continued until 50 weeks by arthroscopic synovectyomy during infliximab treatment (Fig. 4). DAS28 was calculated to be 5.58±0.23 at 0 week, 3.1874±0.47 at 6 weeks, and 2.576±1.49 at 50 weeks after arthroscopic synovectomy (Fig. 5). Therefore, arthroscopic synovectyomy was clinically effective for the patients who tolerated the effect of infliximab. ACR20 was 86%, ACR50 was 57%, and ACR70 was 29% at 6 weeks, and ACR20 was 71%, ACR50 was 42%, and ACR70 was 29% at 50 weeks. Therefore, the efficacy of infliximab and MTX was clinically enhanced by arthroscopic synovectomy.

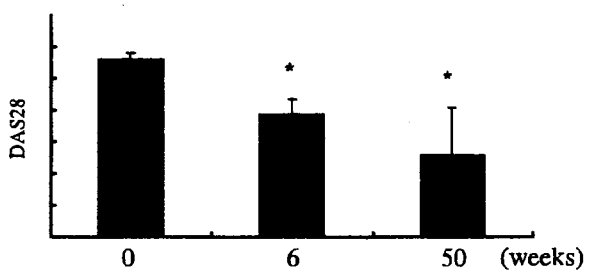


Fig. 5 DAS28 after 0, 6, and 50 weeks by arthroscopic synovectomy

Discussion

Surgical treatment for RA is one the options to be considered if the medical treatment does not respond well. Synovium produces many cytokines and chemokines to develop joint cartilage destruction [6]. Synovectomy is a surgical method used to reduce the cytokines and chemokines especially in joint fluid [7]. There are the two ways to do synovectomy for RA joints, one is arthroscopic synovectomy and the other is open synovectomy. The clinical outcome was much the same in both groups, but there was gradual deterioration, especially after 8 years [8]. RA patients have skin problems such as thin skin that is easily wounded and infected. Arthroscopic synovectomy takes precaution not to get wide wound to induce infection. The patients of infliximab easily get infected, because the function of macrophages was decreased by blocking TNF-α on the surface of macrophages. Therefore, we prefer arthroscopic surgery as a less-invasive method compared with open surgery. As regard to shoulder joint, there is no detailed paper about shoulder arthroscopic synovectomy for RA. We were the first to report shoulder arthroscopic finding and synovectomy especially during infliximab treatment. However, there is a paper that describes infliximab treatment as a reasonable approach for treating early glenohumeral osteoarthritis that had failed to respond to any operative treatment, in which the humeral head and glenoid remain concentric, and where there is still a visible joint space on an axillaries radiograph [9]. We focused our debridement on subacromial bursa so as not to induce impinge pain. After surgery, the range of motion was improved especially in the internal rotation (average from L5 to Th12). We found villous and high vascularity synovitis in arthroscopic finding of effect attenuation cases of infliximab similar to virgins for biological agents.

Synovectomy of the knee in early inflammatory arthritis appears to be successful in decreasing swelling and pain, when the underlying disorders are unresponsive to aggressive medical therapy. Indications for synovectomy may be debated, but generally, the criteria of failure of appropriate medical management for a period of 6 to 12 months in the absence of significant radiographic changes may be accepted as an appropriate indication for consideration of synovectomy. Traditional open synovectomy has been associated with loss of motion in the knee joint suffering from inflammatory arthritis. Multiple reports of the outcome of arthroscopic synovectomy appear to indicate that loss of motion is less of a problem than is seen following an open synovectomy and that the outcome of arthroscopic synovectomy in palliation of the arthritic syndrome is equally effective. A review by Doets et al. [10] concluded that although arthroscopic synovectomy in the setting of RA produced fair or good results in 50% of the cases, one half of the 83 patients in this series had undergone total replacement at a mean interval of 4 years after synovectomy. Matsui et al. [8] similarly reported gradual deterioration in clinical function among patients who had undergone arthroscopic synovectomy for the treatment of RA of the knee. However, arthroscopic synovectomy was associated with less loss of motion and more rapid return of function than open

synovectomy. The advantages of arthroscopic synovectomy therefore appear to be shorter hospitalization and more rapid rehabilitation with less risk of loss of motion than that associated with open synovectomy. Arthroscopic synovectomy in RA may be associated with moderate blood loss. No suction drain is needed in the joint because of less blood loss. Most patients are hospitalized 3–5 days following synovectomy. Although only synovectomy may transiently reduce the swelling associated with inflammatory arthritis of the knee, it does not appear that synovectomy significantly alters the progression of degenerative change in the articular cartilage of the joint [11, 12]. However, it is generally accepted that effects of synovectomy will be kept for about 5 years, and total synovectomy reduces disease activity of the refractory RA following remission [13]. Thus, it is possible to say that improvement of disease activity of given cases might be due to the effects of synovectomy itself. But if we combine anti-TNF-α therapy such as infliximab and arthroscopic synovectomy, the efficacy may be more continuous than synovectomy only. In our data, DAS28 was decreased at 50 weeks after surgery. This clinical improvement may be useful for the continued use of infliximab without adverse events.

References

1. Latosiewicz R, Murawski J (1994) Arthroscopic knee joint synovectomy in the treatment of early stages of rheumatoid arthritis. Rocz Akad Med Bialymst 39:25–30
2. Roch-Bras F, Daures JP, Legouffe MC, Sany J, Combe B (2002) Treatment of chronic knee synovitis with arthroscopic synovectomy: long term results. J Rheumatol 29:1171–1175
3. Maini RN, Breedveld FC, Kalden JR, Smolen JS, Furst D, Weisman MH, St.Clair EW, Keenan GF, van der Heijde D, Marsters PA, Lipsky PE (2004) Sustained improvement over two years in physical function, structural damage, and signs and symptoms among patients with rheumatoid arthritis treated with infliximab and methotrexate. Arthritis Rheum 50:1051–1065
4. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS et al (1988) The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 31:315–324
5. Steinbrocker O, Traeger CH, Battman RC (1949) Therapeutic criteria in rheumatoid arthritis. JAMA 140:659–662
6. Kanbe K, Takagishi K, Chen Q (2002) Stimulation of matrix metalloprotease 3 release from human chondrocytes by the interaction of stromal cell-derived factor 1 and CXC chemokine receptor 4. Arthritis Rheum 46:130–137
7. Kanbe K, Takemura T, Takeuchi K, Chen Q, Takagishi K, Inoue K (2004) Synovectomy reduces cell-derived factor-1 (SDF-1), which is involved in cartilage destruction in osteoarthritis and rheumatoid arthritis. J Bone Joint Surg 86:296–300
8. Matsui N, Taneda Y, Ohta H, Itoh T, Tsuboguchi S (1989) Arthroscopic versus open synovectomy in the rheumatoid knee. Int Orthop 13:17–20
9. Weinstein DM, Bucchieri JS, Pollock RG, Flatow EL, Bigliani LU (2000) Arthroscopic debridement of the shoulder for osteoarthritis. Arthroscopy 16:471–476
10. Doets HC, Bierman BT, Von Soesbergen RM (1989) Synovectomy of the rheumatoid knee does not prevent deterioration. Acta Orthop Scand 60:523–525
11. Smiley P, Wasilewski S (1990) Arthroscopic synovectomy. Arthroscopy 6:18–23
12. McEwen C (1988) The treatment of rheumatoid arthritis: report of results at the end of five years. J Rheumatol 15:764–769
13. Nakamura H, Nagashima M, Ishigami S, Wauke K, Yoshino S (2000) The anti-rheumatic effect of multiple synovectomy in patients with refractory rheumatoid arthritis. Int Orthop 24:242–245

www.nature.com/clinicalpractice/rheum

Mechanisms of Disease: genetics of rheumatoid arthritis—ethnic differences in disease-associated genes

Ryo Yamada and Kazuhiko Yamamoto*

SUMMARY

Large studies on the genetics of common rheumatic diseases, such as rheumatoid arthritis and systemic lupus erythematosus, have identified multiple polymorphisms related to disease susceptibility, including peptidylarginine deiminase 4 (*PADI4*) and protein tyrosine phosphatase N22 (*PTPN22*). Some of the identified genes are associated with multiple autoimmune disorders, and some seem to have unique associations with particular disease entities. Although the molecules encoded by these genes have a primary role in the molecular pathways of autoimmunity, genetic variations and contribution to disease susceptibility seem to vary between ethnic groups. In this Review, we report the findings on genes associated with rheumatoid arthritis and focus on the differences in the frequency of polymorphisms between various ethnic groups.

KEYWORDS ethnic heterogeneity, HLA, *PADI4*, *PTPN22*, SNP

REVIEW CRITERIA

Data for this Review were identified by searching the PubMed database and OMIM for articles with the following terms: "SNP", "single nucleotide polymorphism", "rheumatoid arthritis", "systemic lupus erythematosus", "lupus", "association", "WGA", "PTPN22", "PADI4", "ethnic", "ethnicity", "heterogeneity". The authors focused mainly on association studies reported in articles published in 2000 onwards.

R Yamada is Associate Professor in the Laboratory of Functional Genomics, at the Human Genome Center, Institute of Medical Science, The University of Tokyo and Visiting Associate Professor in the Unit of Human Disease Genomics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Japan. K Yamamoto is Professor in the Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, and Head of the Laboratory for Rheumatic Diseases, SNP Research Center, RIKEN, Japan.

Correspondence
*Department of Allergy and Rheumatology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
yamamoto-tky@umin.ac.jp

Received 29 August 2006 **Accepted** 11 July 2007
www.nature.com/clinicalpractice
doi:10.1038/ncprheum0592

INTRODUCTION

Rheumatic diseases, such as rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE), are characterized by altered inflammatory responses and impaired autoimmune processes. Although the mechanistic etiology of rheumatic diseases is not completely known, familial studies indicate that they are among the most complex genetic disorders. Twin and familial studies in patients with RA have shown a rate of recurrence of disease in a monozygotic twin of approximately 12–15% and, based on these data, MacGregor and colleagues estimated heritability of RA to be about 60%.[1]

Multiple genes are thought to be involved in the genetic susceptibility to rheumatic diseases. The contribution of individual genes to rheumatic disease susceptibility could, therefore, be low. It is not the case, however, that every gene with a disease-associated polymorphism makes only a small contribution to the pathological processes of RA.

RA and SLE have been the principal targets of genetic variation studies in rheumatology. Genetic polymorphisms of the human genome have been investigated in many studies, and novel evidence on the genetic background of rheumatic diseases has been accumulated; this evidence indicates that an ethnic heterogeneity of genetic factors exists for rheumatic disorders. The genetic risk factors for RA and their ethnic heterogeneity are discussed in this Review.

EPIDEMIOLOGY OF RHEUMATOID ARTHRITIS

The prevalence of RA is difficult to calculate precisely, because of clinical heterogeneity and the presence of multiple subclinical conditions. Data from epidemiologic studies, however, suggest a prevalence of 1% or less in most ethnic groups, but reports vary:[2,3] a relatively high prevalence rate of more than 2% has been reported in some Native American groups[4] and a very low prevalence of less than 0.3% has been reported in east Asian, southeast Asian and African populations.[5]

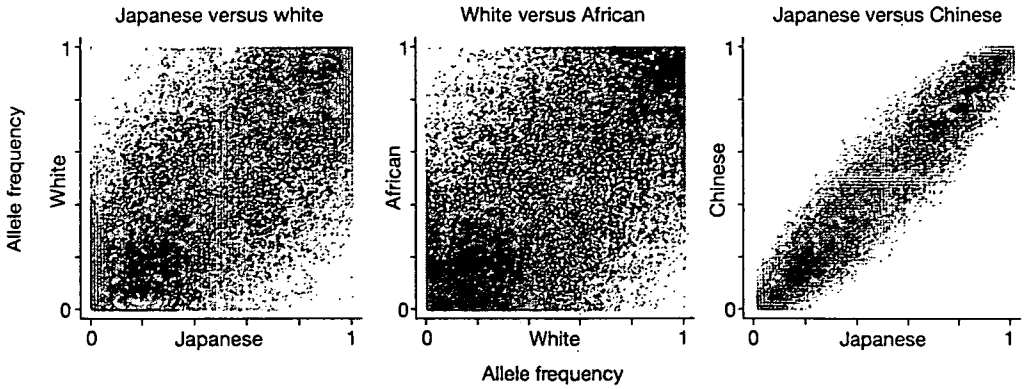


Figure 1 Co-plots of allele frequency of common SNPs in the HapMap project (a subset of SNPs in the genome) to show heterogeneity among ethnic groups. African populations included study samples from the following locations: Yoruba in Ibadan, Nigeria (black African); Utah, USA, from the Centre d'Etude du Polymorphisme Humain (white); Beijing, China (Chinese); and Tokyo, Japan (Japanese). (**A**) A weak correlation was observed between the Japanese and white populations. (**B**) The correlation between the white and black African populations was also weak. (**C**) A strong correlation was observed between Chinese and Japanese populations.

Although differences in study design and the demographic compositions of the populations studied might explain some of the difference in reported prevalence, other risk factors for developing RA should also be considered; genetic and environmental factors are known to have roles. Environmental factors can be classified as regional environmental factors, including geography, climate, and endemic microbes, and sociocultural environmental factors (i.e. lifestyle), such as smoking[6] and nutritional intake. Genetic factors overall can be classivied as fundamental common genetic factors for RA and ethnicity-specific genetic factors.

HETEROGENEITY OF GENES AMONG DIFFERENT ETHNIC GROUPS

In 2003, the Human Genome Project announced the completion of the human genome sequence of 3 billion base pairs.[7,8] Subsequently, heterogeneity of genome sequences within species has been an area of major interest.[9] In 2006, 6,197,786 single nucleotide polymorphisms (SNPs) were validated and registered in the National Center for Biotechnology Information SNP database (NCBI dbSNP build 126), of which 4,116,991 were annotated in the genic region.[10] This large group of polymorphisms is believed to represent one of the basic reasons for the notable diversity among human beings. In addition, the distribution of these polymorphisms has been demonstrated to vary between ethnic groups. As is seen in Figure 1, a greater correlation of allele frequency of SNPs is seen between Chinese and Japanese people than between white and Japanese or between white and black African people. Multiple ethnic populations share some of these genetic variations, but substantial variability is seen in the frequency and monomorphic or polymorphic status of alleles. Less than 60% of the common SNPs evaluated by HapMap phase I were polymorphic in all three ethnic sample sets.[11] Polymorphisms in genes associated with rheumatic disease have also been reported to vary according to allele frequency in different ethnic groups (Table 1).[12] Similarly, common SNPs throughout the genome and SNPs relevant to RA vary among ethnic groups. Interestingly, the differences between SNPs in white and Asian study populations were greater than the differences in SNPs observed between Chinese and Japanese study populations.

ETHNIC HETEROGENEITY OF SUSCEPTIBILITY-ASSOCIATED POLYMORPHISMS

Various genetic study designs have been used to determine genetic susceptibility in RA. Serologic studies in the 1970s and 1980s identified an association between human leukocyte antigen (HLA) serotypes and RA.[13] These findings were confirmed by the later discovery of an association of DNA polymorphisms with HLA-DR antigens. In the interim period, multiple family-based genome scans were done using

www.nature.com/clinicalpractice/rheum

Table 1 Allele frequency of rheumatic disease-associated variants in three ethnic groups

Gene name	Polymorphism	Frequency by population		
		White	Japanese	Black
PADI4	padi_89	0.38	0.40	0.49
	padi_90	0.42	0.40	0.46
	padi_92	0.38	0.39	0.48
SLC22A4	padi_104	0.31	0.33	0.35
	slc2F1	0.09	0.29	0.04
	slc2F2	0.09	0.29	0.04
	Leu503Phe(C1672T)	0.39	0.001	0.04
SLC22A4/A5	haplotype TC	0.39	0.000	0.02
PDCD1	PD-1.3A	0.43	0.00	0.03
PTPN22	Arg620Trp	0.08	0.00	0.02

genome-wide microsatellite markers, and some successfully identified susceptible loci and susceptible genes.[14–17] In addition, multiple candidate genes were tested for an association with various diseases before the human genome sequence was completed. Only a limited number of genes, however, have been replicated.[18–20] The completion of the human genome sequence in 2003 and the development of new techniques for high-throughput genotyping has enabled large-scale, SNP-based scans aimed at identifying complex genetic traits. An introductory genome scan using SNPs has identified several autoimmunity-associated genes.[21–23] Here, we restrict our discussion to the findings on the *HLA-DR* family and the *PTPN22*, *PADI4*, and *FCRL3* genes, because their association with RA has been investigated in multiple ethnic populations.

Susceptibility to rheumatoid arthritis

HLA-DR gene family

In the human genome, the HLA region is the most heterogeneous, and many diseases have been reported to be associated with this area. Various studies have tested the associations with the *HLA-DR* genes, a highly polymorphic gene group within the HLA region, and multiple ethnic groups have been reported to have alleles in the *HLA-DRB* gene associated with susceptibility to RA.[24] Different sets of affected alleles across the *HLA-DR* genes are seen in different ethnic groups, some of which are associated with RA. Most alleles associated with RA susceptibility share a common string of amino acid residues in the epitope-recognition part of the molecule (shared epitope hypothesis), although some do not.[25] Currently, all *HLA-DRB* alleles with the shared epitope are believed to provide RA-prone antigen recognition, and to increase the risk of developing RA, as well as the likelihood of progressing into a destructive phenotype. Most African American patients with RA, however, do not carry the shared epitope in any *HLA-DR* genes,[26] although the occurrence of RA, the rate of adverse events and response to therapy in this population do not differ from those in other ethnic groups.

Although all the mechanisms related to the shared epitope in RA are not yet known, there is little doubt that the HLA-DR molecule is involved in the pathogenesis of RA in populations with alleles associated with RA susceptibility. The HLA-DR molecule probably has an important role in the pathogenesis of RA in African American patients as well. Since the association between RA and *HLA-DRB* was identified, extensive studies have been, and are still being, done to clarify the mechanisms involved, and important results have been reported.[27] Although *HLA-DRB* is one of the most important targets of genetic studies in RA, including those on the ethnic heterogeneity of genetic background, further work on non-HLA genes has shown that heterogeneity among ethnic groups for non-HLA genes can contribute to the determination of subtypes of RA along with *HLA-DRB*.[28]

The PTPN22 gene

PTPN22 encodes an intracellular tyrosine phosphatase in T cells, B cells, and various hematopoietic cells. A candidate gene study of type 1 diabetes

mellitus showed that a missense SNP (Arg620Trp) in the gene was associated with this disorder in multiple populations of European descent.[29] Subsequently, multiple autoimmune diseases in white populations were reported to be associated with this SNP, including SLE[30] and RA.[19,31,32] The molecular mechanisms that determine susceptibility to autoimmune disorders from this SNP are thought to be a decrease in binding between *PTPN22* and *CSK* (c-*src* tyrosine kinase) and an increase in intrinsic enzymatic activity.[33]

The association between this SNP and multiple autoimmune disorders has been replicated in white populations, but the polymorphism has not been not detected in Asian study populations.[12] As with the *HLA-DRB* polymorphisms and RA, Asians have specific distributions of genetic variations, although none has yet been identified in the *PTPN22* gene.

The PADI4 and FCRL3 genes

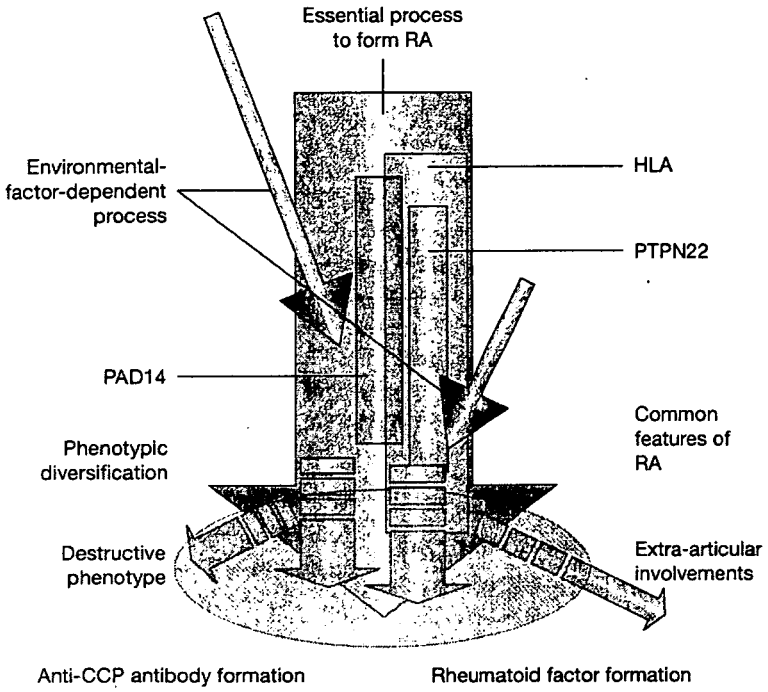
We have reported associations between *PADI4*[22] and RA and between *FCRL3*[23] and RA, SLE, and autoimmune thyroiditis during gene-based, large-scale linkage disequilibrium mapping of SNPs. *PADI4* encodes an enzyme that catalyzes peptidyl arginine to peptidyl citrulline. In mammals, this reaction is carried out only by *PADI4* and its isozymes, and the reaction products are recognized by anticitrullinated peptide antibody, one of the most RA-specific autoantibodies.[34] Although the role of peptidyl citrullination by peptidyl arginine deiminase enzymes and citrullinated peptides in the pathogenesis of RA is not clear, the function of these enzymes seems to make the corresponding genes reasonable candidates for RA-specific genes. Initially, we reported that the *PADI4* gene has haplotypes consisting of multiple coding SNPs associated with RA in Japanese individuals. The SNPs in the coding region were reported to be associated with messenger RNA stability. Contrary to the case of *PTPN22*, the haplotypes were observed with similar frequencies in Asian and white populations.[35] In association studies of *PADI4*, however, the population frequencies differed substantially, with all but one white population showing no distinct association between this gene and RA.[35–41]

Several meta-analyses have also summarized the aforementioned studies, and the results indicate that the genetic risk associated with this haplotype seems to occur in Far East populations, and is much weaker or was nonexistent in European populations.[36,41] Conversely, antibodies to peptidyl citrulline are universally recognized as an important marker in RA. A pathway of citrullination by *PADI4* could, therefore, be a common pathological pathway in RA, even in white populations. Another gene, *FCRL3*, has similar heterogeneity among ethnic groups and is associated with susceptibility to rheumatic diseases. *FCRL3* encodes a B-lymphocyte-specific membrane molecule with speculative function as a cofactor of B-cell receptors. *FCRL3* has a promoter SNP that disrupts the nuclear factor κB-binding motif and was originally reported to be associated with RA, SLE, and autoimmune thyroiditis in Japanese individuals.[23] *FCRL3* was reported to be dominantly expressed in centrocytes in the germinal center of lymph nodes, and its polymorphism was associated with autoantibody production. The gene variation was identified in Korean and British populations with similar frequency to *PADI4*.[23,42] The association of *FCRL3* and RA disease susceptibility shown by another Japanese group was, however, much weaker[43] than the association of *PADI4* with RA, and was not replicated in studies of white populations.[42,44]

Explaining ethnic heterogeneity of susceptibility

Given the complexity of the pathogenesis of RA, the interactions of many cells and molecules clearly have decisive roles in disease development and progression. Results of genetic studies imply that genes with disease-associated polymorphisms are involved in disease-related pathological processes. The outcomes of these pathological processes would, in turn, interact further and form a phenotype of RA. The pathological processes involved in RA can be classified into at least three categories: the essential processes needed to form RA itself; environment-dependent processes in certain settings (e.g. smoking); and, finally, processes yielding specific phenotypes of RA (e.g. rheumatoid nodules and vasculitis).

We believe that there are several essential processes in the pathogenesis of RA, such as lymphocyte activation processes, including those involving *PTPN22* and the *HLA-DR* genes. Consequently, as antibodies to citrullinated proteins are universally recognized as the most specific autoantibody in RA,[34] citrullination by *PADI4* also seems to be an essential pathological process for RA pathogenesis (Figure 2).



Most polymorphisms in known disease-associated genes might cause moderate changes in the intensity of the pathological processes, but do not completely negate existing reactions or create new reactions. As discussed above, there is extensive heterogeneity in the allele frequencies of common SNPs (Figure 1), and it is common for allele frequencies of disease-associated polymorphisms to differ among ethnic groups. For example, the polymorphism identified in *PTPN22* in white populations has not yet been detected in Japanese populations. This lack of could, therefore, be essential. The same is true for *HLA-DR* in African American patients with RA. If this is indeed the case, the same citrullination process used by *PADI4*, which was identified as being associated with RA in Asian patients, might also be important in white patients. The polymorphism might not be detected as being strongly associated with the disease in white populations, however, possibly because of different backgrounds of genetic influences.

Ethnic heterogeneity in a disease entity is one way to explain the ethnic differences of disease-associated polymorphisms. Given the variable clinical presentations of RA, however, most rheumatologists would probably agree that the heterogeneity of RA itself, regardless of ethnicity, is likely to have a larger role in the pathogenesis of the disease than differences due to ethnicity. The prevalence of RA does not differ greatly between populations, and no apparent ethnicity-specific environmental factors have been identified. Thus, it seems reasonable that epidemiologic differences among populations are due to a mixture of genetic and environmental factors that have not been identified so far. The polymorphisms and functionality of *PTPN22* and *PADI4* need to be investigated to the same level as the *HLA-DR* gene family before the heterogeneity of their association with RA susceptibility can be established.

CONCLUSIONS

Further to the advances in high-throughput genetic studies, we are now entering an exciting era of applying genome-wide association analyses to common diseases.[45] In the near future, multiple genes associated with rheumatic disease will be identified. Several factors can, however, lead to false conclusions, including inadequate study sample sizes or incorrect population stratification. Thus, both original and replication studies need to be conducted carefully, giving due consideration to ethnic heterogeneity in genes associated with disease susceptibility. A polymorphism that is replicated frequently in certain populations could be involved in the primary pathological process of the disease. After careful evaluation of ethnic differences in disease entity, environmental factors, and genes, it is important to use

Genetic Epidemiology 31: 709–726 (2007)

A Novel Method to Express SNP-Based Genetic Heterogeneity, Ψ, and Its Use to Measure Linkage Disequilibrium for Multiple SNPs, D_g, and to Estimate Absolute Maximum of Haplotype Frequency

Ryo Yamada* and Fumihiko Matsuda

Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Single nucleotide polymorphisms (SNPs) are important markers to investigate genetic heterogeneity of population and to perform linkage disequilibrium (LD) mapping. We propose a new method, Ψ, to express frequency of 2^{N_s} haplotypes for N_s di-allelic SNPs. Using the new expression of haplotype frequency, we propose a novel measure of LD, D_g, not only for SNP pairs but also for multiple markers. The values of D_g for SNP pairs were revealed to be similar to values of conventional pairwise LD indices, D' and r^2, and it was revealed that D_g quantitated components of LD that were not measured by conventional LD indices for SNP pairs. Also we propose a distinct method, D_g-based absolute estimation, to infer the absolute maximum estimates of haplotype frequency. The result of the D_g-based absolute estimation of haplotype frequency for SNP pairs were compared with the conventional expectation-maximization (EM) algorithm and reported that the new method gave better inference than the EM algorithm which converged infrequently to a local extreme. *Genet. Epidemiol.* 31:709–726, 2007.

*Correspondence to: Ryo Yamada, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.
Received 24 September 2006; Revised 27 February 2007; Accepted 29 March 2007
Published online 16 May 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20235

INTRODUCTION

DNA molecules are chemical compounds that carry genetic information in living organisms. As carriers of genetic information, they have to be homogeneous enough to maintain identity of species and successful reproduction (speciation hypothesis) [Rieseberg 2001; Noor et al., 2001; Navarro and Barton, 2003; Rieseberg and Livingstone, 2003]. On the contrary, they carry heterogeneity that is believed to give benefit to species for survival in ever-changing environments (adaptation) [Hartl and Clark, 1997a]. Heterogeneity of DNA sequence in a population is balanced between pressures in the two directions, toward the homogeneity or clonality and toward the heterogeneity or randomness.

To clarify the heterogeneity of DNA sequence population, let us assume two extreme populations. In one population, all the DNA molecules are completely identical, in other words, a clonal population. In the other one, any nucleotide in the DNA molecules is randomly selected. In this random population, any site in the DNA molecules is polymorphic, and no association exists between sites or between sets of sites. We define the randomness of polymorphism of individual sites as "allele frequency-randomness" and the randomness between sites or between sets of sites as "inter-site randomness". Because only a small fraction of nucleotide sites in a species is polymorphic, the inter-site randomness is observed only among the polymorphic sites. The pressures toward the clonality of individual sites include selection by favorable phenotypes, unsuccessful reproduction through excessive discrepancy between sequences from gametes and genetic drift due to finite effective population size with unbalanced transmission. On the contrary, mutation and recombination of the genetic material to descendent increase the allele frequency randomness and the inter-site randomness [Hartl and Clark, 1997b].

Since the genome sequences from multiple organisms were determined, one of major research targets has been the investigation of intraspecies variations [Collins et al., 2003]. As reported in *Drosophila*, recombination rate seemed correlated with the presence of nucleotide polymorphisms [Aquadro et al., 1994]. Therefore the allele frequency-randomness of individual nucleotides and the inter-site randomness should be quantitated together. Unfortunately nucleotide diversity, the most popular parameter to quantify heterozygosity at the nucleotide level does not take into account the LD [Hartl and Clark, 1997c]. On the other hand, LD indices, such as D' and r^2 as well as multiallelic D' quantitate only allelic inter-site dependency [Devlin and Risch, 1995; Zapata, 2000].

Because single nucleotide polymorphisms (SNPs) are the most common variations [Kidd et al., 2004] and the LD mapping using SNPs is a very promising method to investigate genetic background of various phenotypes, [Morton, 2005], we propose a new method, Ψ, to express the allele frequency-randomness of individual sites and the intersite randomness for diallelic markers together by a uniform expression. The introduction of Ψ enables us to propose D_g as a novel measure of LD for multiple SNPs, which will be beneficial for genetic studies because only limited measures of LD for multiple loci are currently available, such as ε [Nothnagel et al., 2002], are now limited.

Also using Ψ, we propose a new method to infer the absolute maximum estimate of haplotype frequency. Although the conventional expectation-maximization (EM) algorithm is known to give a local maximum periodically [Nin, 2004], Ψ-based method enables us to overcome the shortcoming and to infer the global maximum of haplotype frequency for SNP pairs, we compared both methods and verified that the conventional EM scarcely converges to a local maximum.

INTRODUCTORY EXAMPLES

Before we give generalized expressions of Ψ and D_g, some examples are introduced.

INCOMPLETENESS OF PAIRWISE LD INDEX

Assume three SNP sites, $S = \{s_A, s_B, s_C\}$, and their eight haplotypes

$H = \{"ABC", "ABc", "AbC", "Abc", "aBC", "aBc", "abC", "abc"\}$.

When their haplotype frequencies are

$$F_1 = \{f_{ABC}, f_{ABc}, f_{AbC}, f_{Abc}, f_{aBC}, f_{aBc}, f_{abC}, f_{abc}\}$$
$$= \{0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125\}$$

or

$$F_2 = \{0.25, 0, 0, 0.25, 0, 0.25, 0.25, 0\}$$

pairwise r^2's [Devlin and Risch, 1995] of three SNP pairs for both F_1 and F_2 are 0, because frequencies of four haplotypes for each pair are $\{0.25, 0.25, 0.25, 0.25\}$ for both cases (See Columns 1 "F1" and 2 "F2" in Table I). Although the two cases are apparently different in terms of LD, their pairwise LD index values are the same. This difference implicates that there are components of LD which three pairwise LD indices can not describe. In this article, we propose a LD measure D_g to quantitate all LD components.

CALCULATION OF Ψ AND D_g

Ψ is a novel system to express haplotype frequencies and D_g is a new measure of LD for multiple sites. From haplotype frequencies, Ψ is calculated initially, and D_g is deduced from Ψ.

Transformation of F_i to Ψ and Ψ plots. $\Psi = \{\psi_{subset}\}$ for SNP trio S has eight elements, each of which corresponds to a subset of

$$S: \{\{\phi\}(\text{empty}), \{s_A\}, \{s_B\}, \{s_C\}, \{s_A, s_B\}, \{s_A, s_C\}, \{s_B, s_C\}, \{s_A, s_B, s_C\}\}.$$

The Ψ elements for each subset are defined with frequency of their alleles (f_x: frequency of haplotype or allele "x") as below:

$$\psi_\phi = 1,$$
$$\psi_{s_A} = f_A - f_a,$$
$$\psi_{s_B} = f_B - f_b,$$
$$\psi_{s_C} = f_C - f_c,$$
$$\psi_{s_A,s_B} = f_{AB} - f_{Ab} - f_{aB} + f_{ab},$$
$$\psi_{s_A,s_C} = f_{AC} - f_{Ac} - f_{aC} + f_{ac},$$
$$\psi_{s_B,s_C} = f_{BC} - f_{Bc} - f_{bC} + f_{bc},$$
$$\psi_{s_A,s_B,s_C} = f_{ABC} - f_{ABc} - f_{AbC} + f_{Abc} - f_{aBC} + f_{aBc} + f_{abC} - f_{abc}.$$

Table I gives frequencies of three sites, three site-pairs and trio as well as the corresponding Ψs. For the case of F_1, $\Psi = \{1, 0, 0, 0, 0, 0, 0, 0\}$ and for F_2, $\Psi = \{1, 0, 0, 0, 0, 0, 0, 1\}$. The first seven elements of Ψ for both cases are the same and the last element, ψ_{s_A,s_B,s_C}, represents the difference between the two.

It is noticed that there is a rule of alternation of positive/negative signs in the expression of Ψ's with frequency of haplotypes, implicating that the expression will be generalized for more nucleotide sites as shown later in this article. Also it is noticed that the introduced variables are in the hierarchic and nested structure. In other words, the elements of Ψ for three sites are consisted of the elements of Ψ's for subsets of sites with an additional component specific to the trio. The nested structure is visualized as Ψ plots in Figure 1(a) with detailed description of its structure. The Ψ plots tell that the difference between F_1 and F_2 is located into the row on the bottom, that stands for the trio. Ψ for the trio contains Ψ for SNP pairs. The four circles in the upper left of Ψ plots, corresponding to $\psi_\phi, \psi_{s_A}, \psi_{s_B}$ and ψ_{s_A,s_B}, create a rhombus that is Ψ plot for the site-pair. Two other rhombuses for site-pairs, $\{s_A, s_C\}$ and $\{s_B, s_C\}$ exist in the Ψ plot of the trio as well. All circles except for ψ_{s_A,s_B,s_C} are a part of at least one of three rhombuses for site-pairs. Figure 1(b) shows structure of Ψ plots of one to seven sites. The Ψ plot of n sites contains Ψ plots of $n-1$ sites in its inside. All the elements in the Ψ plot of n sites except one circle at the bottom are a part of at least one of Ψ's of $n-1$ sites as shown in Figure 1(b).

Calculation of D_g from Ψ and D_g plots. D_g quantitates LD for division patterns of multiple SNPs. For example, a division, $(s_A, s_B, s_C) \rightarrow \{(s_A, s_B), (s_C)\}$ means a division of three SNPs into a SNP pair (s_A, s_B)

TABLE I. Haplotype frequencies and their ψ and D_g

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Column ID	F1	F2	Clone1	Clone2	1 SNP	2 SNPs in absolute LD	2 SNPs in LE	2 SNPs in partial LD	3 SNPs in absolute LD-1	3 SNPs in absolute LD-2	3 SNPs in LE	3 SNPs in partial LD-1	3 SNPs in partial LD-2	3 SNPs in partial LD-3
f_{ABC}	0.125	0.25	1	0	0.6	0.6	0.42	0.6	0.5	0.6	0.336	0.3	0.3	0.3
f_{ABc}	0.125	0	0	1	0	0	0	0	0	0	0.084	0	0	0
f_{AbC}	0.125	0	0	0	0	0	0.18	0.3	0	0	0.144	0	0	0
f_{Abc}	0.125	0.25	0	0	0	0	0	0	0	0	0.036	0.2	0.2	0.3
f_{aBC}	0.125	0	0	0	0.4	0	0.28	0	0	0	0.224	0.3	0.2	0.2
f_{aBc}	0.125	0.25	0	0	0	0	0	0	0	0	0.056	0	0	0
f_{abC}	0.125	0.25	0	0	0	0.4	0.12	0.1	0	0	0.096	0	0	0
f_{abc}	0.125	0	0	0	0	0	0	0	0.5	0.4	0.024	0.2	0.3	0.2
f_A	0.5	0.5	1	1	0.6	0.6	0.6	0.9	0.5	0.6	0.6	0.5	0.5	0.6
f_a	0.5	0.5	0	0	0.4	0.4	0.4	0.1	0.5	0.4	0.4	0.5	0.5	0.4
f_B	0.5	0.5	1	1	1	0.6	0.7	0.6	0.5	0.6	0.7	0.6	0.5	0.5
f_b	0.5	0.5	0	0	0	0.4	0.3	0.4	0.5	0.4	0.3	0.4	0.5	0.5
f_C	0.5	0.5	1	0	1	1	1	1	0.5	0.6	0.8	0.6	0.5	0.5
f_c	0.5	0.5	0	1	0	0	0	0	0.5	0.4	0.2	0.4	0.5	0.5
f_{AB}	0.25	0.25	1	1	0.6	0.6	0.42	0.6	0.5	0.6	0.42	0.3	0.3	0.3
f_{Ab}	0.25	0.25	0	0	0	0	0.18	0.3	0	0	0.18	0.2	0.2	0.3
f_{aB}	0.25	0.25	0	0	0.4	0	0.28	0	0	0	0.28	0.3	0.2	0.2
f_{ab}	0.25	0.25	0	0	0	0.4	0.12	0.1	0.5	0.4	0.12	0.2	0.3	0.2
f_{AC}	0.25	0.25	1	0	0.6	0.6	0.6	0.9	0.5	0.6	0.48	0.3	0.3	0.3
f_{Ac}	0.25	0.25	0	1	0	0	0	0	0	0	0.12	0.2	0.2	0.3
f_{aC}	0.25	0.25	0	0	0.4	0.4	0.4	0.1	0	0	0.32	0.3	0.2	0.2
f_{ac}	0.25	0.25	0	0	0	0	0	0	0.5	0.4	0.08	0.2	0.3	0.2
f_{BC}	0.25	0.25	1	0	1	0.6	0.7	0.6	0.5	0.6	0.56	0.6	0.5	0.5
f_{Bc}	0.25	0.25	0	1	0	0	0	0	0	0	0.14	0	0	0
f_{bC}	0.25	0.25	0	0	0	0.4	0.3	0.4	0	0	0.24	0	0	0
f_{bc}	0.25	0.25	0	0	0	0	0	0	0.5	0.4	0.06	0.4	0.5	0.5
ψ_ϕ	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ψ_{s_A}	0	0	1	1	0.2	0.2	0.2	0.8	0	0.2	0.2	0	0	0.2
ψ_{s_B}	0	0	1	1	1	0.2	0.4	0.2	0	0.2	0.4	0.2	0	0
ψ_{s_C}	0	0	1	−1	1	1	1	1	0	0.2	0.6	0.2	0	0
ψ_{s_A,s_B}	0	0	1	1	0.2	1	0.08	0.4	1	1	0.08	0	0.2	0
ψ_{s_A,s_C}	0	0	1	−1	0.2	0.2	0.2	0.8	1	1	0.12	0	0.2	0
ψ_{s_B,s_C}	0	0	1	−1	1	0.2	0.4	0.2	1	1	0.24	1	1	1
ψ_{s_A,s_B,s_C}	0	1	1	−1	0.2	1	0.08	0.4	0	0.2	0.048	0	0	0.2

Genet. Epidemiol. DOI 10.1002/gepi

TABLE I. Continued.

Column ID	1 F1	2 F2	3 Clone1	4 Clone2	5 1 SNP	6 2 SNPs in absolute LD	7 2 SNPs in LE	8 2 SNPs in partial LD	9 3 SNPs in absolute LD-1	10 3 SNPs in absolute LD-2	11 3 SNPs in LE	12 3 SNPs in partial LD-1	13 3 SNPs in partial LD-2	14 3 SNPs in partial LD-3
$d_g((s_A, s_B) \rightarrow \{(s_A), (s_B)\})$	0	0	0	0	0	1	0	0.286	1	1	0	0	0.2	0
$d_g((s_A, s_C) \rightarrow \{(s_A), (s_C)\})$	0	0	0	0	0	0	0	0	1	1	0	0	0.2	0
$d_g((s_B, s_C) \rightarrow \{(s_B), (s_C)\})$	0	0	0	0	0	0	0	0	1	1	0	1	1	1
$d_g((s_A,s_B,s_C) \rightarrow \{(s_A, s_B), (s_C)\})$	0	1	0	0	0	0	0	0	0	0	0	0	0	0.2
$d_g((s_A,s_B,s_C) \rightarrow \{(s_A, s_C), (s_B)\})$	0	1	0	0	0	1	0	0.286	0	0	0	0	0	0.2
$d_g((s_A,s_B,s_C) \rightarrow \{(s_A, s_B), (s_C)\})$	0	1	0	0	0	1	0	0.286	0	0	0	0	0	0
$d_g((s_A,s_B,s_C) \rightarrow \{(s_B, s_C), (s_A)\})$	0	1	0	0	0	1	0	0.286	0	0.194	0	0	0	0.2

Cells with 1 or −1 in ψ or d_g are shadowed.

and a single SNP s_C. There are seven division patterns for three sites:

$$(s_A, s_B) \rightarrow \{(s_A), (s_B)\},$$
$$(s_A, s_C) \rightarrow \{(s_A), (s_C)\},$$
$$(s_B, s_C) \rightarrow \{(s_B), (s_C)\},$$
$$(s_A, s_B, s_C) \rightarrow \{(s_A, s_B), (s_C)\},$$
$$(s_A, s_B, s_C) \rightarrow \{(s_A, s_C), (s_B)\},$$
$$(s_A, s_B, s_C) \rightarrow \{(s_A), (s_B, s_C)\},$$
$$(s_A, s_B, s_C) \rightarrow \{(s_A), (s_B), (s_C)\}.$$

The first three divides a SNP pairs into two single SNPs. The next three split a SNP trio into a SNP pair and a single SNP. The last divides a SNP trio into three single SNPs.

The elements of $D_g = \{d_g((\text{subset}_i) \rightarrow \{(\text{subset}_{ji}), (\text{subset}_{jz}), \ldots, (\text{subset}_{jk})\})\}$, correspond to divisions. We define D_g as below. Of note when the denominator of either expression in the parenthesis is zero, take the other value.

$$d_g((s_A,s_B) \rightarrow \{(s_A),(s_B)\})$$
$$= \max\left(\left(1 - \frac{\psi_{s_A,s_B}+1}{\psi_{s_A} \times \psi_{s_B}+1}\right), \left(1 - \frac{\psi_{s_A,s_B}-1}{\psi_{s_A} \times \psi_{s_B}-1}\right)\right),$$

$$d_g((s_A,s_C) \rightarrow \{(s_A),(s_C)\})$$
$$= \max\left(\left(1 - \frac{\psi_{s_A,s_C}+1}{\psi_{s_A} \times \psi_{s_C}+1}\right), \left(1 - \frac{\psi_{s_A,s_C}-1}{\psi_{s_A} \times \psi_{s_C}-1}\right)\right),$$

$$d_g((s_B,s_C) \rightarrow \{(s_B),(s_C)\})$$
$$= \max\left(\left(1 - \frac{\psi_{s_B,s_C}+1}{\psi_{s_B} \times \psi_{s_C}+1}\right), \left(1 - \frac{\psi_{s_B,s_C}-1}{\psi_{s_B} \times \psi_{s_C}-1}\right)\right),$$

$$d_g((s_A,s_B,s_C) \rightarrow \{(s_A,s_B),(s_C)\})$$
$$= \max\left(\left(1 - \frac{\psi_{s_A,s_B,s_C}+1}{\psi_{s_A,s_B} \times \psi_{s_C}+1}\right), \left(1 - \frac{\psi_{s_A,s_B,s_C}-1}{\psi_{s_A,s_B} \times \psi_{s_C}-1}\right)\right),$$

$$d_g((s_A,s_B,s_C) \rightarrow \{(s_A,s_C),(s_B)\})$$
$$= \max\left(\left(1 - \frac{\psi_{s_A,s_B,s_C}+1}{\psi_{s_A,s_C} \times \psi_{s_B}+1}\right), \left(1 - \frac{\psi_{s_A,s_B,s_C}-1}{\psi_{s_A,s_C} \times \psi_{s_B}-1}\right)\right),$$

$$d_g((s_A,s_B,s_C) \rightarrow \{(s_B,s_C),(s_A)\})$$
$$= \max\left(\left(1 - \frac{\psi_{s_A,s_B,s_C}+1}{\psi_{s_B,s_C} \times \psi_{s_A}+1}\right), \left(1 - \frac{\psi_{s_A,s_B,s_C}-1}{\psi_{s_B,s_C} \times \psi_{s_A}-1}\right)\right),$$

$$d_g((s_A,s_B,s_C) \rightarrow \{(s_A),(s_B),(s_C)\})$$
$$= \max\left(\left(1 - \frac{\psi_{s_A,s_B,s_C}+1}{\psi_{s_A} \times \psi_{s_B} \times \psi_{s_C}+1}\right), \left(1 - \frac{\psi_{s_A,s_B,s_C}-1}{\psi_{s_A} \times \psi_{s_B} \times \psi_{s_C}-1}\right)\right).$$

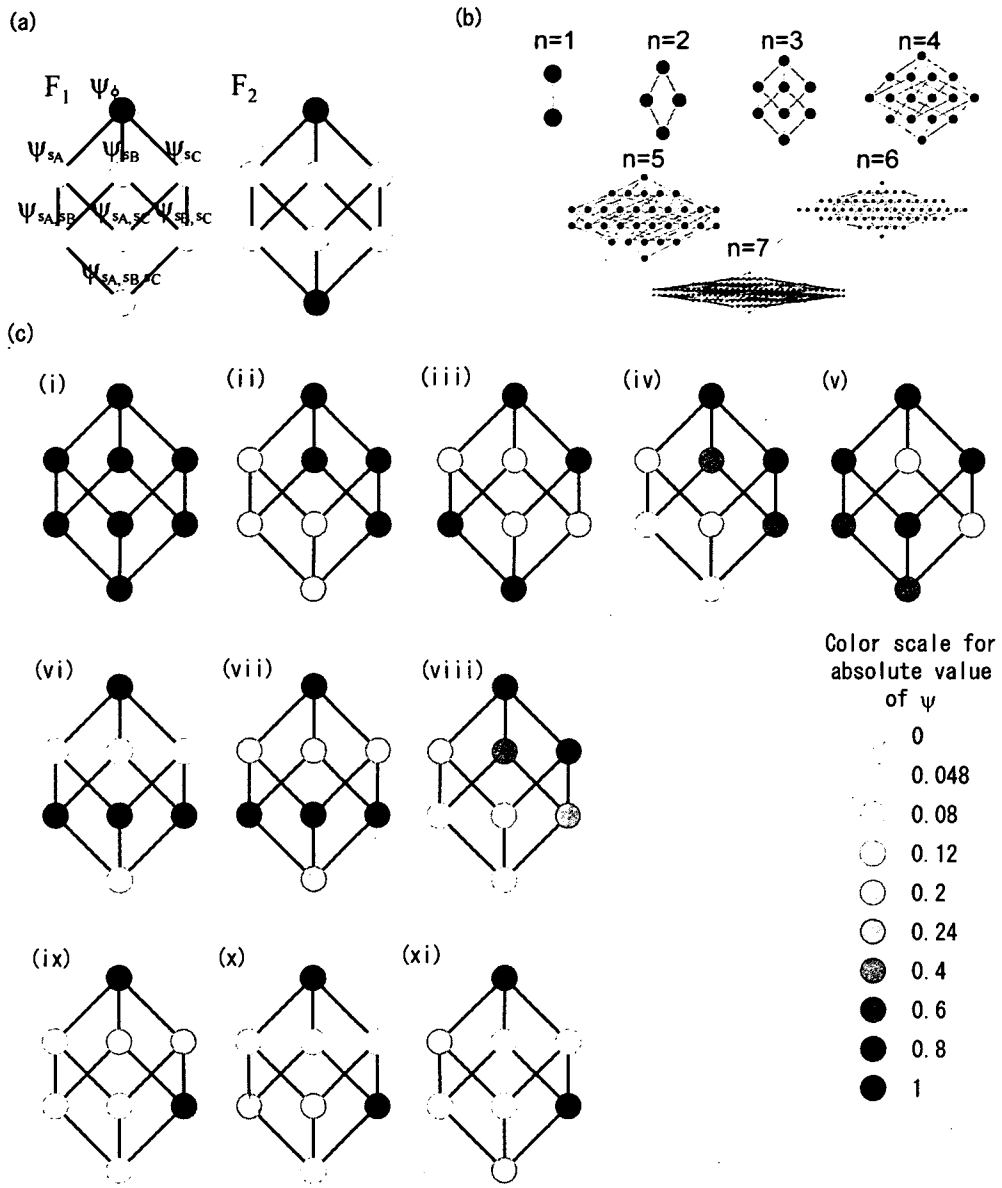


Fig. 1. (a) ψ plot of two haplotype frequency sets, F_1 and F_2 are drawn. ψ plots for three sites are consisted of four rows. The top row is for the empty set, the second top row is for three single sites, the third top row is for three site-pairs and the bottom row is for the site-trio. The circle on the left in the third top row for ψ_{s_A,s_B} is connected to two circles in the second top row, ψ_{s_A} and ψ_{s_B}, representing the relation of $\{s_A\} \subset \{s_A, s_B\}$ and $\{s_B\} \subset \{s_A, s_B\}$. The bottom row has one white (F_1) or black (F_2) circle, corresponding to the site-trio. It is connected to three circles in the third top row, because every site-pair is a subset of the trio. The circles are connected when numbers of elements of two subsets are different by one and the smaller subset is a subset of the larger. Black circles represent ψ value being 1 and white 0. (b) Power sets for 1–7 element set are drawn. They are also ψ plots of clones with $n = 1, \ldots, 7$ are shown. Every ψ plot has one circle at the top as the empty set and one circle at the bottom corresponding to the self subset. (c) Various ψ plots are displayed. The corresponding haplotype frequencies and D_g values are shown in Table I (columns 3 to 14). ψ values were shown in gray scale in (a) and (c).

For the case of F_1, $D_g = \{0,0,0,0,0,0,0\}$ and for F_2, $D_g = \{0,0,0,1,1,1,1\}$. The elements for the divisions of SNP pairs into single SNPs are 0 for both cases, which corresponds to the fact r^2 of three SNP pairs are 0. For F_1 all the other elements of D_g are also 0, indicating that the three sites are truly in linkage equilibrium (LE). On the other hand, the last four elements of D_g for F_2 are different from zero. These four elements represent components of LD in these three sites that can not be described by pairwise LD indices but should be described by taking account of LD for the trio.

Because the number of divisions into subsets becomes very large when the number of sites is

increased, we propose to choose a part of the elements of D_g for visual presentation of D_g and they are plotted into two triangles. One triangle is consisted of d_g's for divisions of all the site-pairs into single sites (pairwise triangle). The other triangle is consisted of d_g's for divisions of all the subsets of sites whose elements are in tandem into single sites (tandem triangle). In case of four sites, the pairwise triangle is consisted of $d_g((s_1,s_2) \to \{(s_1),(s_2)\})$, $d_g((s_1,s_3) \to \{(s_1),(s_3)\})$, $d_g((s_1,s_4) \to \{(s_1),(s_4)\})$, $d_g((s_2,s_3) \to \{(s_2),(s_3)\})$, $d_g((s_2,s_4) \to \{(s_2),(s_4)\})$ and $d_g((s_3,s_4) \to \{(s_3),(s_4)\})$. The tandem triangle is consisted of $d_g((s_1,s_2) \to \{(s_1),(s_2)\})$, $d_g((s_2,s_3) \to \{(s_2),(s_3)\})$, $d_g((s_3,s_4) \to \{(s_3),(s_4)\})$, $d_g((s_1,s_2,s_3) \to \{(s_1),(s_2),(s_3)\})$, $d_g((s_2,s_3,s_4) \to \{(s_2),(s_3),(s_4)\})$ and $d_g((s_1,s_2,s_3,s_4) \to \{(s_1),(s_2),(s_3),(s_4)\})$. Because the pairwise triangle and the tandem triangle share d_g's for divisions of the site-pairs in tandem, the corresponding parts of the two triangles are overlapped and displayed as shown in Figure 2(a). D_g plots for F_1 and F_2 with three sites are shown in Figure 2(b).

MORE EXAMPLES OF Ψ AND D_g FOR THREE SNPS

Now Ψ and D_g are calculated in additional examples with three sites. When all three sites are monomorphic (See column 3 "Clone1" and column 4 "Clone2" in Table I and Fig. 1(c)–(i)), all $|\psi|$'s are 1 and all d_g's are 0. When one of three sites are polymorphic (See column 5 "1 SNP" in Table I and Fig. 1(c)–(ii)), $|\Psi|$'s for the subsets containing the polymorphic site are not 1. All d_g's are 0. Three examples with two SNPs are shown in column 6 "two SNPs in absolute LD", column 7 "two SNPs in LE" and column 8 "two SNPs in partial LD", and Figure 1(c)–(iii),(iv),(v)). When two SNPs are in the absolute LD, a black circle of ψ_{s_B,s_C} in Figure 1(c)–(iii) represents the allelic association. All d_g's are 0 when two SNPs are in LE. When two SNPs in LD, $d_g((s_A,s_B) \to \{(s_A),(s_B)\})$ stands for the strength of LD between the two SNPs. (Columns 6 and 8 of Table I).

Columns 9 and 10, "3 SNPs in absolute LD-1" and "3 SNPs in absolute LD-2" in Table I and Fig. 1(c)–(vi),(vii) present examples where all the three sites are polymorphic and only two haplotypes exist. These two differ in their allele frequencies. The former has two haplotypes with the same frequency and the latter's haplotypes have different frequency. All three SNP pairs are in their absolute LD ($r^2 = 1$). Ψ plots distinguish these two by gray color in the circles for three single SNPs and the trio. The difference between two examples is observed in $d_g((s_A,s_B,s_C) \to \{(s_A),(s_B),(s_C)\})$.

When three SNPs are in LE as shown in column 11 of Table I, all d_g's are 0 (Fig. 1(c)–(viii)). The gradation in gray scale from top to bottom in the Ψ plot is a feature of LE throughout the sites. Columns 12, 13 and 14 are the examples with partial LD in three polymorphic sites. All of them have the same four haplotypes out of eight and their frequencies are 0.2 or 0.3. s_A and s_B are in the absolute LD for all the three examples, as their $d_g((s_B,s_C) \to \{(s_B),(s_C)\}) = 1$. Their difference appears in the distribution of non-zero values in their D_g and their Ψ plots (Fig. 1(c)–(ix),(x),(xi)). Their D_g plots are shown in Figure 2(c).

Estimation of haplotype frequency using Ψ from unphased genotype data. When unphased genotype data of SNP pairs are given, frequencies of four haplotypes have to be estimated, and Ψ transforms this estimation into a monovariate problem.

Example: when genotype counts are observed for two SNPs, allele frequencies of both SNPs are calculated based on their own genotype counts. They give $\psi_{s_A} = f_A - f_a$ and $\psi_{s_B} = f_B - f_b$. In order to estimate frequencies of all four haplotypes, ψ_{s_A,s_B} is the only variable. Therefore estimation of haplotype frequencies of SNP pairs turns to be the same with maximal likelihood estimation of monovariate, ψ_{s_A,s_B}. Once ψ_{s_A,s_B} is estimated, $\{f_{AB}, f_{Ab}, f_{aB}, f_{ab}\}$ can be given by

$$f_{AB} = \tfrac{1}{4}(\psi_{s_A,s_B} + \psi_{s_A} + \psi_{s_B} + \psi_\phi),$$
$$f_{Ab} = \tfrac{1}{4}(-\psi_{s_A,s_B} + \psi_{s_A} - \psi_{s_B} + \psi_\phi),$$
$$f_{aB} = \tfrac{1}{4}(-\psi_{s_A,s_B} - \psi_{s_A} + \psi_{s_B} + \psi_\phi),$$
$$f_{ab} = \tfrac{1}{4}(\psi_{s_A,s_B} - \psi_{s_A} - \psi_{s_B} + \psi_\phi).$$

This topic is discussed in the section "Usage of Ψ for haplotype frequency inference".

NOTATIONS

SITES

- Consider a set of DNA sequences with the same length n. The sites are not necessarily polymorphic.
- Let $S(n)(1_{st})$ denote the first set of n sites. All the sites are potentially diallelic although some of them can be monomorphic;

$$S(n)(1_{st}) = \{s_1, s_2, \ldots, s_n\}.$$

- Let $\mathrm{Pow}(S(n)(1_{st}))$ denote a power set of $S(n)(1_{st})$ (Fig. 1(b)) [Weisstein, 2006a].

$$\begin{aligned}\mathrm{Pow}(S(n)(1_{st})) = \{&S(0)(1_{st}),\\ &S(1)(1_{st}), S(1)(2_{nd}), \ldots, S(1)(n_{th}),\\ &S(2)(1_{st}), S(2)(2_{nd}), \ldots, S(2)({}_nC_{2_{th}}),\\ &S(3)(1_{st}), \ldots, S(3)({}_nC_{3_{th}}), \ldots,\\ &S(n-1)(1_{st}), \ldots, S(n-1)({}_nC_{n-1_{th}}),\\ &S(n)(1_{st})\}.\end{aligned}$$

- $S(i)(j_{th})$ represents the j_{th} subset with i sites in $\mathrm{Pow}(S(n)(1_{st}))$, $(i = 0,1,\ldots,n; j = 1,2,\ldots,{}_nC_i)$. $S(0)(1_{st})$ is an empty set and the last element of $\mathrm{Pow}(S(n)(1_{st}))$ is $S(n)(1_{st})$ itself.

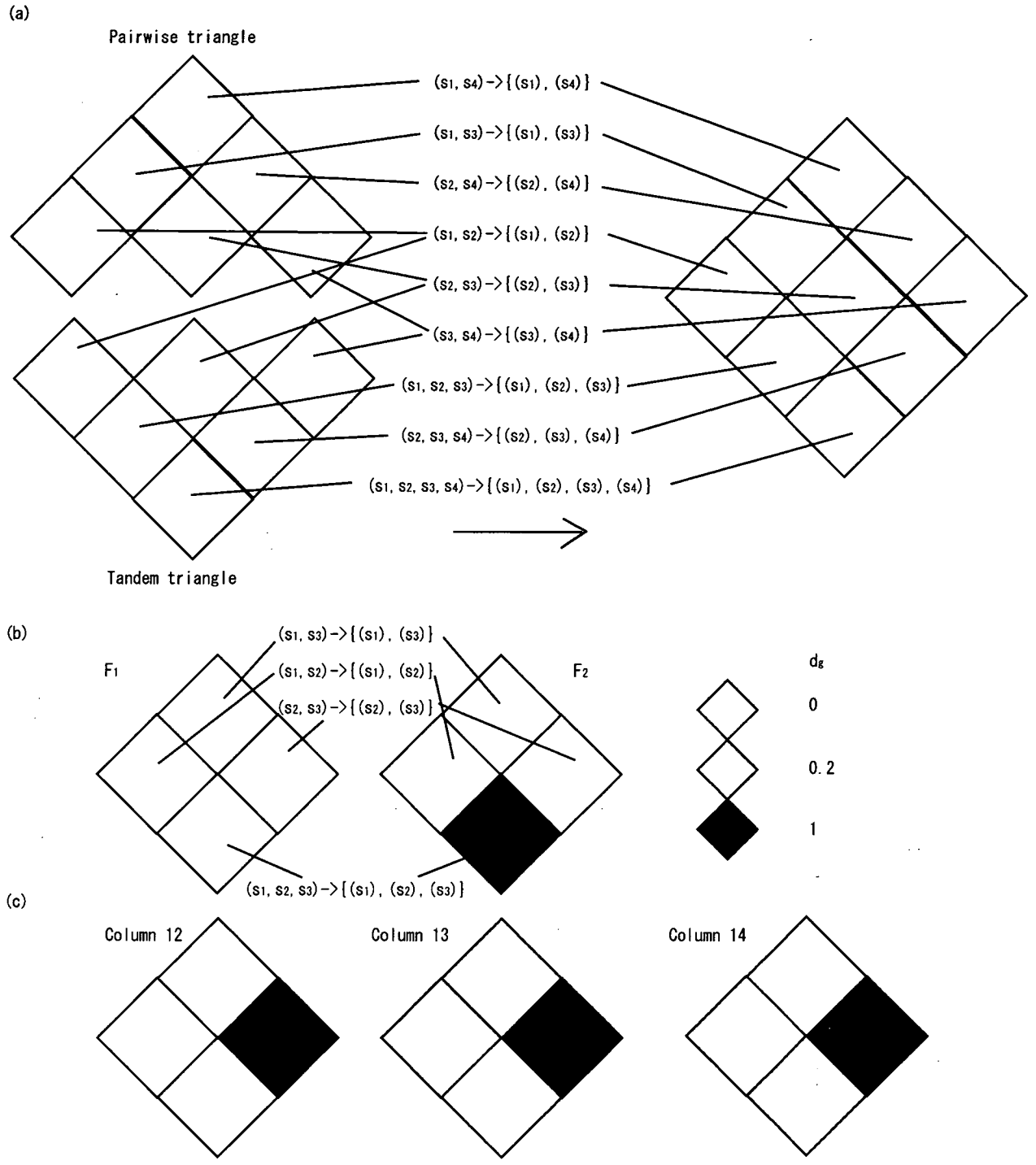


Fig. 2. D_g plots. (a) The pairwise triangle, the upper half of D_g plots, are consisted of squares for the site pairs. The tandem triangle, the lower half of D_g plots, are consisted of squares for the subsets of sites in tandem. The squares for the site pairs in tandem are arranged on the bottom of the pairwise triangle and on the top of the tandem triangle. Therefore they are overlapped in the right drawing. (b) D_g plots for F_1 and F_2. Their difference appears in the lower tandem triangle but not in the upper pairwise triangle. (c) D_g plots for examples of Columns 12, 13 and 14 (Table I).

- Number of subsets which have i elements is ${}_nC_i$. The total number of elements of Pow($S(n)(1)$) is $\sum_{i=0}^{n} {}_nC_i = 2^n$.

HAPLOTYPES

- Let $H(i)(j_{\text{th}})$ and $F(i)(j_{\text{th}})$ denote the 2^i haplotypes and their frequency of $S(i)(j_{\text{th}})$.

$$H(i)(j_{\text{th}}) = \{h_1(i)(j_{\text{th}}), h_2(i)(j_{\text{th}}), \ldots, h_{2^i}(i)(j_{\text{th}})\},$$

$$F(i)(j_{\text{th}}) = \{f_1(i)(j_{\text{th}}), f_2(i)(j_{\text{th}}), \ldots, f_{2^i}(i)(j_{\text{th}})\}.$$

- Let $V(i)(j_{\text{th}})$ denote alternating positive/negative signs for $H(i)(j_{\text{th}})$ as mentioned in the section "INTRODUCTORY EXAMPLES".

$$V(i)(j_{\text{th}}) = \{v_1(i)(j_{\text{th}}), v_2(i)(j_{\text{th}}), \ldots, v_{2^i}(i)(j_{\text{th}})\}.$$

Initially dummy value, 1 or -1, is given to two alleles of individual n sites in $S(n)(1_{\text{st}})$. When a haplotype has even number of sites of value -1, dummy value of the haplotype is 1, and when it has odd number, the value is -1.
Example: Assume $i = 3$ and two alleles of s_1 are A and T, for s_2, G and C and for s_3, C and A where value 1 is assigned to the first allele of each site. Two of three sites (the first and third sites) of haplotype TGA is the allele of (-1), therefore the dummy value of haplotype TGA is 1.

- Partial haplotype. When $S(p_i)(q_{i_{\text{th}}}) \subset S(p_j)(q_{j_{\text{th}}})$ and $h_{k_i}(p_i)(q_{i_{\text{th}}})$ is a part of $h_{k_j}(p_j)(q_{j_{\text{th}}})$, $h_{k_i}(p_i)(q_{i_{\text{th}}})$ is called as a partial haplotype of $h_{k_j}(p_j)(q_{j_{\text{th}}})$in $S(p_i)(q_{i_{\text{th}}})$. Let $u^{p_i,q_{i_{\text{th}}}}_{kj,pj,qjth}$ denote the ordinal number to indicate a partial haplotype in $S(p_i)(q_{i_{\text{th}}})$ for $S(p_j)(q_{j_{\text{th}}})$.

Example: Consider the sample example in the previous bullet. $S(2)(1_{\text{st}}) = \{s_1, s_2\}$ is a subset of $S(3)(1_{\text{st}}) = \{s_1, s_2, s_3\}$, $(S(2)(1_{\text{st}}) \subset S(3)(1_{\text{st}}))$.

$$\begin{aligned} H(3)(1_{\text{st}}) &= \{h_1(3)(1_{\text{st}}), h_2(3)(1_{\text{st}}), h_3(3)(1_{\text{st}}), h_4(3)(1_{\text{st}}), \\ &\quad h_5(3)(1_{\text{st}}), h_6(3)(1_{\text{st}}), h_7(3)(1_{\text{st}}), h_8(3)(1_{\text{st}})\} \\ &= \{\text{"}AGC\text{"}, \text{"}AGA\text{"}, \text{"}ACC\text{"}, \text{"}ACA\text{"}, \text{"}\underline{TG}C\text{"}, \\ &\quad \text{"}\underline{TG}A\text{"}, \text{"}TCC\text{"}, \text{"}TCA\text{"}\}, \\ H(2)(1_{\text{st}}) &= \{h_1(2)(1_{\text{st}}), h_2(2)(1_{\text{st}}), h_3(2)(1_{\text{st}}), h_4(2)(1_{\text{st}})\} \\ &= \{\text{"}AG\text{"}, \text{"}AC\text{"}, \text{"}\underline{TG}\text{"}, \text{"}TC\text{"}\}, \end{aligned}$$

$h_3(2)(1_{\text{st}}) = \text{"}\underline{TG}\text{"}$ is a part of $h_5(3)(1_{\text{st}}) = \text{"}\underline{TG}C\text{"}$ and $h_6(5)(1_{\text{st}}) = \text{"}\underline{TG}A\text{"}$. Then $u^{2,1_{\text{st}}}_{5,3,1_{\text{st}}} = 3$ and $u^{2,1_{\text{st}}}_{6,3,1_{\text{st}}} = 3$.

DIVISION OF A SET OF SITES INTO SUBSETS

- Divisions of a set of sites.

Consider a division of a set of n sites. $S(n)(1_{\text{st}})$ is divided into m non-empty subsets that are mutually exclusive $(S(n)(1_{\text{st}}) = \bigcup_{i=1}^{m} S_i(p_i)(q_{i_{\text{th}}}); i = 1, 2, \ldots, m; p_i \neq 0; n = \sum_{i=1}^{m} n_i; q_i = 1, 2, \ldots, {}_nC_{p_i}; S_i \cap S_j = \{\phi\}$ for any i and j $(i \neq j))$. Let $S(n)(1_{\text{st}}) \rightarrow \{S_1(p_1)(q_{1_{\text{th}}}), \ldots, S_m(p_m)(q_{m_{\text{th}}})\}$ denote this division pattern.
Example: The above mentioned example $S(3)(1_{\text{st}}) = \{s_1, s_2, s_3\}$ is divided into four different division patterns; Into $S(2)(1_{\text{st}}) = \{s_1, s_2\}$ and $S(1)(3_{\text{rd}}) = \{s_3\}$, or into $S(2)(2_{\text{nd}}) = \{s_1, s_3\}$ and $S(1)(2_{\text{nd}}) = \{s_2\}$, or into $S(2)(3_{\text{rd}}) = \{s_2, s_3\}$ and $S(1)(1_{\text{st}}) = \{s_1\}$, or into three single sites $S(1)(1_{\text{st}}) = \{s_1\}$, $S(1)(2_{\text{nd}}) = \{s_2\}$ and $S(1)(3_{\text{rd}}) = \{s_3\}$.

- Division of a haplotype into partial haplotypes.

When $S(n)(1_{\text{st}})$ is divided into m subsets, $H(n)(1_{\text{st}})$ is also divided into their m partial haplotypes, each of which is a haplotype in $S_i(p_i)(q_{i_{\text{th}}})$. $h_k(n)(1_{\text{st}})$, the kth haplotype of length n in $S(n)(1_{\text{st}})$, is expressed as a set,

$$\begin{aligned} h_k(n)(1_{\text{st}}) = (&h_{u^{p_1,q_{1_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_1)(q_{1_{\text{th}}}), h_{u^{p_2,q_{2_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_2)(q_{2_{\text{th}}}), \ldots, h_{u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{\text{st}}}} \\ &(p_i)(q_{i_{\text{th}}}), \ldots, h_{u^{p_m,q_{m_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_m)(q_{m_{\text{th}}})). \end{aligned}$$

where $h_{u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_i)(q_{i_{\text{th}}})$ represents the $u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{st}}$ haplotype in $S_i(p_i)(q_{i_{\text{th}}})$, that is a part of $h_k(n)(1_{\text{st}})$.
Example: when $S(3)(1_{\text{st}}) = \{s_1, s_2, s_3\}$ is divided into $S(2)(2_{\text{nd}}) = \{s_1, s_3\}$ and $S(1)(2_{\text{nd}}) = \{s_2\}$, $h_5(3)(1_{\text{st}}) = \text{"}TGC\text{"}$ is divided into $h_3(2)(2_{\text{nd}}) = \text{"}TC\text{"}$ and $h_1(1)(2_{\text{nd}}) = \text{"}G\text{"}$. Therefore $u^{2,2_{\text{nd}}}_{5,3,1_{\text{st}}} = 3$ and $u^{1,2_{\text{nd}}}_{5,3,1_{\text{st}}} = 1$.

$$\begin{aligned} h_5(3)(1_{\text{st}}) &= (h_{u^{2,2_{\text{nd}}}_{5,3,1_{\text{st}}}}(2)(2_{\text{nd}}), h_{u^{1,2_{\text{nd}}}_{5,3,1_{\text{st}}}}(1)(2_{\text{nd}})) \\ &= (h_3(2)(2_{\text{nd}}), h_1(1)(2_{\text{nd}})). \end{aligned}$$

- Dummy values and division and partial haplotypes.

Dummy value of $h_k(n)(1_{\text{st}})$ is also expressed as,

$$v_k(n)(1_{\text{st}}) = \prod_{i=1}^{m} v_{u^{p_i,q_{i_{\text{th}}}}_{k,n,1_{\text{st}}}}(p_i)(q_{i_{\text{th}}})$$

Example: $v_5(3)(1_{\text{st}}) = v_3(2)(2_{\text{nd}}) \times v_1(1)(2_{\text{nd}}) = (-1) \times (1) = -1$.

SNP-BASED HETEROGENEITY TENSOR Ψ

- We define Ψ for $S(n)(1_{\text{st}})$, which is consisted of 2^n elements, each of which is a value for an element of Pow($S(n)(1_{\text{st}})$).

$$\Psi = \{\psi(i)(j_{\text{th}})\},$$

where $\psi(i)(j_{\text{th}})$ represents a value for an element, $S(i)(j_{\text{th}})$, in Pow($S(n)(1_{\text{st}})$, the j_{th} subset with i sites.

Because the elements of Ψ are arranged in the multi-dimensional structure with indices, i and j, we call Ψ as "SNP-based heterogeneity tensor" (tensor: a multi-dimensional array). [Rowland and Weissltein, 2006].

Further details of Ψ will be defined in the following sections.

SNP-BASED HETEROGENEITY TENSOR Ψ

DEFINITIONS

Here we give basic rules for $\psi(i)(j_{st})$ so that all of Ψ s are defined by a systematic way, that correspond to subsets of SNPs with various size.

- $\psi(0)(1_{st})$ is defined as 1.
- $\psi(i)(j_{st})$; $i>0$ is defined below:

$$\psi(i)(j_{st}) = \sum_{k=1}^{2^i}(v_k(i)(j_{st}) \times f_k(i)(j_{st})) \quad (1)$$

With these definitions, Ψ has the following features:

- When DNA sequence population is a clone, absolute value of all the elements of Ψ is 1 or -1.
- When DNA sequence population is in the limit randomness, all the elements of Ψ, except for $\psi(0)(1_{st})$, are 0.
- Otherwise $\psi(i)(j_{th})$ ranges from -1 to 1 according to the heterogeneity condition of the population.

Ψ GIVES A BASE FOR HAPLOTYPE FREQUENCY SPACE

This section gives a note on relation between haplotype frequencies, $F(n)(1_{st})$ and Ψ. Both $F(n)(1_{st})$ and Ψ have 2^n elements. They are in one-to-one correspondence. This bijective relation can be proven by showing that the determinant of the matrix, that transforms $F(n)(1_{st})$ to Ψ, is different from zero, which will recurrently be proven with Laplace expansion of determinant. (proof not shown) [Weisstein, 2006b].

The transformation from $F(n)(1_{st})$ to Ψ is expressed by

$$\psi(i)(j_{th}) = \sum_{k=1}^{2^i}(v_k(i)(j_{th}) \times f_k(i)(j_{th})), i = 0,1,\ldots,n;$$
$$j = 1,2,\ldots,{}_nC_i; \quad k = 1,2,\ldots,2^i. \quad (2)$$

The reverse transformation from Ψ back to $F(n)(1_{st})$ is expressed by

$$f_k(n)(1_{st}) = \frac{1}{2^n} \times \sum_{p=0}^{n-1}\sum_{q=1}^{{}_nC_p} v_{u_{k,n,1_{st}}^{p,q_{th}}}(p)(q_{th}) \times \psi(p)(q_{th}),$$
$$p=0,1,\ldots,n; \quad q=1,2,\ldots{}_nC_p; \quad k=1,2,\ldots,2^n. \quad (3)$$

Each element of $F(n)(1_{st})$ represents the frequency of one of the 2^n distinct haplotypes. Because the sum of the 2^n elements of $F(n)(1_{st})$ is 1 and fixed, their degree of freedom is 2^n-1. Because $F(n)(1_{st})$ and Ψ are mutually in one-to-one correspondence, the degree of freedom of Ψ should be also 2^n-1. One of the 2^n elements of Ψ is 1 and constant, therefore all the other 2^n-1 elements are mutually independent. This means that the dimension of the haplotype frequency space is 2^n-1 and 2^n-1 elements of Ψ except for $\psi(0)(1_{st})$ consist a base of the space.

LINKAGE DISEQUILIBRIUM AND Ψ

INTER-SITE RANDOMNESS AND INDEPENDENCY

The inter-site randomness is defined for division patterns of a SNP set as follows. Consider a division of $S(n)(1_{st})$ into m mutually exclusive non-empty subsets $\mathrm{Div}: S(n)(1_{st}) \rightarrow \{S_1(p_1)(q_{1_{th}}),\ldots,S_m(p_m)(q_{m_{th}})\}$. In this situation, when the inter-site randomness is at its maximal conformation, the frequency of all haplotypes in $S_i(p_i)(q_{i_{th}})$ is mutually independent. In this condition, $f_k(n)(1_{st}) = \prod_{i=1}^{m} f_{u_{k,n,1_{st}}^{p_i,q_{i_{th}}}}(p_i)(q_{i_{th}})$, and $v_k(n)(1_{st}) = \prod_{i=1}^{m} v_{u_{k,n,1_{st}}^{p_i,q_{i_{th}}}}(p_i)(q_{i_{th}})$. By a simple transformation, Ψ in the maximized inter-site randomness can be expressed by

$$\begin{aligned}\psi(n)(1_{st}) &= \sum_{i=k}^{2^n}(f_k(n)(1_{st}) \times v_k(n)(1_{st})) \\ &= \sum_{k=1}^{2^n}\left(\prod_{i=1}^{m} f_{u_{k,n,1_{st}}^{p_i,q_{i_{th}}}}(p_i)(q_{i_{th}}) \times \prod_{i=1}^{m} v_{u_{k,n,1_{st}}^{p_i,q_{i_{th}}}}(p_i)(q_{i_{th}})\right) \\ &= \prod_{i=1}^{m}\left(\sum_{j=1}^{2_i^p}(f_j(p_i)(q_{i_{th}}) \times v_j(p_i)(q_{i_{th}}))\right) \\ &= \prod_{i=1}^{m}\psi(p_i)(q_{i_{th}}). \end{aligned} \quad (4)$$

Figure 1(c)–(viii) shows an example of $n=3$ in the maximized inter-site randomness LE for all division patterns, in which the equation (4) is satisfied.

GENERALIZED LINKAGE DISEQUILIBRIUM INDEX, D_g

When alleles at the sites are associated on the same chromosome, they are called to be in LD. In LE no allelic association is present. Therefore when the equation (4), $\psi(n)(1_{st}) = \prod_{i=1}^{m}\psi(p_i)(q_{i_{th}})$ for $\mathrm{Div}, S(n)(1_{st}) \rightarrow \{S_1(p_1)\ (q_{1_{th}}),\ldots,S_m(p_m)(q_{m_{th}})\}$, is satisfied, it can be said that $S(n)(1_{st})$ is in LE for the particular division pattern Div. The deviation from the equation (4) represents the degree of LD for $S(n)(1_{st})$ with

respect to the particular Div. Therefore LE and LD are defined for ways to divide a set of sites.

We introduce generalized linkage disequilibrium index, d_g(Div) for Div as:

$$d_g(\mathrm{Div}) = \max\left(\left(1 - \frac{\psi(n)(1_{st}) + 1}{\prod_{i=1}^{m} \psi_i(p_i)(q_{i_{st}}) + 1}\right), \times \left(1 - \frac{\psi(n)(1_{st}) - 1}{\prod_{i=1}^{m} \psi_i(p_i)(q_{i_{st}}) - 1}\right)\right). \quad (5)$$

When denominator of either expression in the parenthesis of "max" is zero, the other value should be selected for d_g(Div).

By this definition, d_g(Div) satisfies:

- d_g(Div) takes a value in the range 0–1.
- D_g(Div) takes zero in LE.

For a pair of two sites and its division into two single SNPs, this is expressed by

$$D_g(\mathrm{Pair}) = \max\left(\left(1 - \frac{\psi(2)(1_{st}) + 1}{\psi(1)(1_{st})\psi(1)(2_{nd}) + 1}\right), \times \left(1 - \frac{\psi(2)(1_{st}) - 1}{\psi(1)(1_{st})\psi(1)(2_{nd}) - 1}\right)\right). \quad (6)$$

For a SNP pair, $F(2)(1_{st}) = \{f_1(2)(1_{st}), f_2(2)(1_{st}), f_3(2)(1_{st}), f_4(2)(1_{st})\}$. For simplicity, $F = \{f_1, f_2, f_3, f_4\}$ will be used hereafter. The numerator of equation (6) is expressed as $\psi(1)(1_{st})\psi(1)(2_{nd}) - \psi(2)(1_{st}) = -((f_1 - f_2 - f_3 + f_4) - ((f_1 + f_2) - (f_3 + f_4)) \times ((f_1 + f_3) - (f_2 + f_4)))$. The right-hand side of the equation is transformed into $-4 \times (f_1 \times f_4 - f_2 \times f_3)$ (Appendix 1). This expression of numerator is proportional to the numerator of other conventional LD indices including D′ and r^2 [Devlin and Risch, 1995], which indicates D_g(Pair) is an appropriate index for SNP pairs. D_g(Pair) has a standardized value of the numerator that takes 1 when $\psi(2)(1_{st}) = \pm 1$, and takes 0 in case of LE. Values of D_g(Pair) are plotted for comparison with other conventional LD indices, D′, r^2 and r in Figure 3. Let A/a and B/b denote alleles of two SNPs. In Figure 3(a), major alleles of two SNPs, A and B, are fixed at 0.8. Frequency of haplotype AB (f_1), is parameterized from 0.6 to 0.8 under the condition where frequency of haplotype Ab (f_2) equals the one of aB (f_3). In Figure 3(b), the frequency of haplotype AB (f_1) is parameterized from 0 to 0.8 under the condition where frequency of haplotype aB is fixed at zero. D_g(Pair) takes the same value with D' and r, when $f_1 \times f_4 - f_2 \times f_3$ is positive and when $f_2 = f_3$. However, when the symmetry of $f_2 = f_3$ is lost, the values of D_g(Pair), D' and r diverge. Both D_g(Pair) and r^2 are 1 when $f_1 + f_4 = 1$, and both converge to zero when $f_1 + f_3 = 1$.

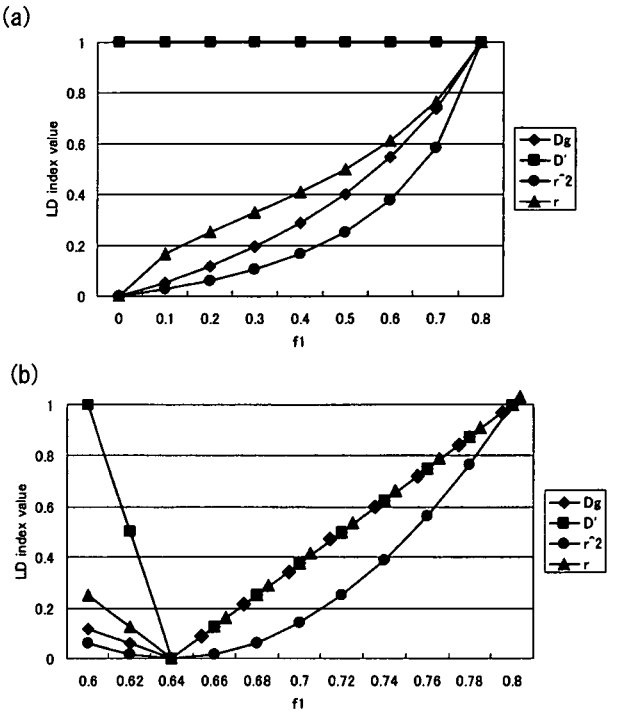


Fig. 3. Plots of D_g, D', r^2 and r. (a) SNP A and SNP B with respectively alleles A, a and B, b. Allele frequencies $P(A)$ and $P(B)$ are fixed at 0.8, and $P(AB)$ is parameterized from 0.6 to 0.8 under the condition of $P(Ab) = P(aB)$. (b) $P(AB)$ is parameterized from 0 to 0.8 under the condition of $P(aB) = 0$.

ADDITIONAL EXAMPLES

In addition to the examples presented in the section INTRODUCTORY EXAMPLES, a few more examples will be helpful.

Ψ FOR TWO SITES

Assume there are two sites, $S(2)(1_{st}) = \{s_1, s_2\}$.
The power set $\mathrm{Pow}(S(2)(1_{st})) = \{\{\phi\}, \{s_1\}, \{s_2\}, \{s_1, s_2\}\}$.

$$H(2)(1_{st}) = \{h_1(2)(1_{st}), h_2(2)(1_{st}), h_3(2)(1_{st}), h_4(2)(1_{st})\} = \{\text{"00"}, \text{"01"}, \text{"10"}, \text{"11"}\}.$$

For a subset $S(2)(1_{st}) = \{s_1, s_2\}$,

$$F(2)(1_{st}) = \{f_1(2)(1_{st}), f_2(2)(1_{st}), f_3(2)(1_{st}), f_4(2)(1_{st})\},$$
$$V(2)(1_{st}) = \{1, -1, -1, 1\},$$
$$\psi(2)(1_{st}) = f_1(2)(1_{st}) - f_2(2)(1_{st}) - f_3(2)(1_{st}) + f_4(2)(1_{st}).$$

The single site frequencies are derived by summing full haplotype frequencies across alleles at the other sites.

For a subset $S(1)(1_{st}) = \{s_1\}$,

$$F(1)(1_{st}) = \{(f_1(2)(1_{st}) + f_2(2)(1_{st})), (f_3(2)(1_{st}) + f_4(2)(1_{st}))\},$$
$$V(1)(1_{st}) = \{1, -1\},$$
$$\psi(1)(1_{st}) = (f_1(2)(1_{st}) + f_2(2)(1_{st})) - (f_3(2)(1_{st}) + f_4(2)(1_{st})).$$

For a subset $S(1)(2_{nd})=\{s_2\}$,

$$F(1)(2_{nd})=\{(f_1(2)(1_{st})+f_3(2)(1_{st})),(f_2(2)(1_{st})+f_4(2)(1_{st}))\},$$
$$V(1)(2_{nd})=\{1,-1\},$$
$$\psi(1)(2_{nd})=(f_1(2)(1_{st})+f_3(2)(1_{st}))-(f_2(2)(1_{st})+f_4(2)(1_{st})).$$

For a subset $S(0)(1_{st})$,

$$\psi(0)(1_{st})=1.$$

$$\Psi=\{\psi(0)(1_{st}),\psi(1)(1_{st}),\psi(1)(2_{nd}),\psi(2)(1_{st})\},$$

Ψ plots of these cases are shown and explained in Figure 4.

D_g FOR SIX SITES

In the section "INTRODUCTORY EXAMPLES", F_2 was shown to have LD components that are not detected by pairwise LD measures but detected by D_g. In this section, we deal with six site examples, that are consisted of two sets of three sites; $S = \{S_p, S_q\} = \{s_A, s_B, s_C, s_{A'}, s_{B'}, s_{C'}\}$. Haplotype frequencies for the former and the latter three sites are identical with F_2; $F_p = \{f_{ABC}, f_{ABc}, f_{AbC}, f_{Abc}, f_{aBC}, f_{aBc}, f_{abC}, f_{abc}\}$ = {0.25,0,0,0.25,0,0.25,0.25,0} and $F_q = \{f_{A'B'C'}, f_{A'B'c'}, f_{A'b'C'}, f_{A'b'c'}, f_{a'B'C'}, f_{a'B'c'}, f_{a'b'C'}, f_{a'b'c'}\}$ = {0.25,0,0,0.25, 0,0.25,0.25,0}. In the first case of six sites, case1, the haplotypes of the former three sites and the haplotypes of the latter are in one-to-one correspondence;

$$F(\text{case1})=\{f_{ABCA'B'C'}, f_{AbcA'b'c'}, f_{aBca'B'c'}, f_{abCa'b'C'}\}$$
$$=\{0.25,0.25,0.25,0.25\}.$$

The D_g plot of case1 is shown in Figure 5(a). The black square on the bottom representing the division of all the six sites into six single sites, explains LD in the region as a whole. Four black squares in the third row from the bottom, representing site-trios, which are in LD themselves. Three black squares in the third row from the top, representing site-pairs intervened by two sites, are also in LD. In the second case, case2, each haplotype in the former site-set are evenly connected to every haplotype in the latter site-set.

$$F(\text{case2})=\{f_{ABCA'B'C'}, f_{ABCA'b'c'}, f_{ABCa'B'c'}, f_{ABCa'b'C'}, f_{AbcA'B'C'},$$
$$f_{AbcA'b'c'}, f_{Abca'B'c'}, f_{Abca'b'C'},$$
$$f_{aBcA'B'C'}, f_{aBcA'b'c'}, f_{aBca'B'c'}, f_{aBca'b'C'}, f_{abCA'B'C'},$$
$$f_{abCA'b'c'}, f_{abCa'B'c'}, f_{abCa'b'C'}\}$$
$$=\{0.0625,0.0625,0.0625,0.0625,0.0625,0.0625,$$
$$0.0625,0.0625,$$
$$0.0625,0.0625,0.0625,0.0625,0.0625,0.0625,$$
$$0.0625,0.0625\}.$$

D_g plot of this case is shown in Figure 5(b), All the pairwise d_g's are 0. The black square on the bottom representing the division of all the six sites into six single sites, explains LD in the region as a whole. Two black squares for $(s_A,s_B,s_C)\rightarrow\{(s_A),(s_B),(s_C)\}$ and $(s_{A'},s_{B'},s_{C'})\rightarrow\{(s_{A'}),(s_{B'}),(s_{C'})\}$ stands for LD at the trio level.

When the latter three sites are monomorphic for three of four haplotypes in the former set (case3),

$$F(\text{case3})=\{f_{ABCA'B'C'}, f_{ABCA'b'c'}, f_{ABCa'B'c'}, f_{ABCa'b'C'}, f_{AbcA'b'c'},$$
$$f_{aBcA'B'C'}, f_{abCA'B'C'}, f_{abCA'b'c'}\}$$
$$=\{0.0625,0.0625,0.0625,0.0625,0.25,0.25,0.25\}.$$

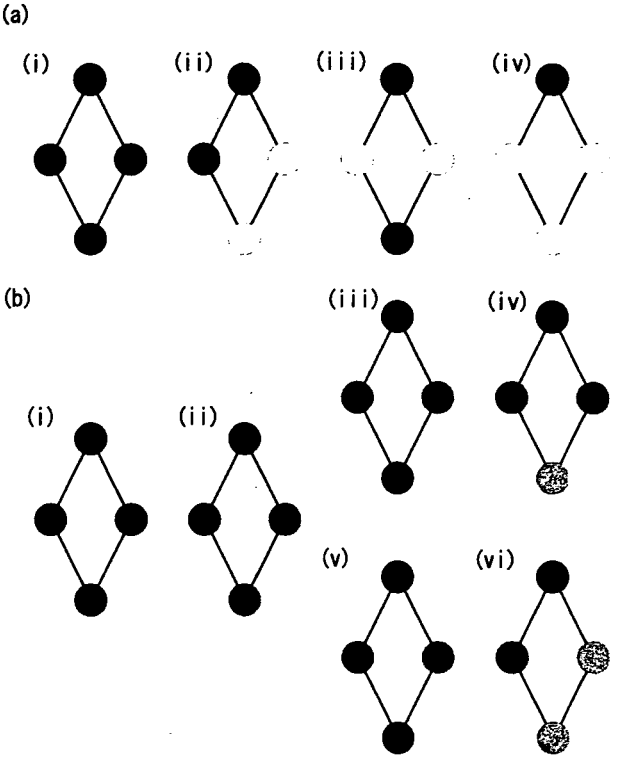


Fig. 4. ψ plots for two sites. (a) Four patterns of ψ where two sites are monomorphic or their allele frequency is 0.5. (a)–(i) $F(2)(1_{st})$ = {1, 0, 0, 0} (a clone). Ψ= {1, 1, 1, 1}. (a)–(ii) One site is monomorphic and the other site is polymorphic and its allele frequency is 0.5. $F(2)(1_{st})$ = {0.5, 0.5, 0, 0}. Ψ= {1, 1, 0, 0}. (a)–(iii) and (a)–(iv) Allele frequencies of both sites are 0.5. (a)–(iii) Absolute LD. $F(2)(1_{st})$ = {0.5, 0, 0, 0.5} and Ψ= {1, 0, 0, 1}. (a)–(iv) LE. $F(2)(1_{st})$ = {0.25, 0.25, 0.25, 0.25} and Ψ= {1, 0, 0, 0}. The black-and-white circles distinguish the four patterns. The absolute LD was indicated by $\psi(2)(1_{st}) = 1$ and LE was by $\psi(2)(1_{st}) = 0$. (b) shows patterns of Ψ where allele frequency of two sites are not necessarily 0.5. (b)–(i) is an example of a clone. When one site is polymorphic and its allele frequency is not 0.5, ψ plot appears like (b)–(ii) ($F(2)(1_{st})$ = {0.6, 0.4, 0, 0}. Ψ= {1, 1, 0.2, 0.2}. When both sites are polymorphic and their allele frequencies are the same but not 0.5, they are in the absolute LD and its ψ plot is (b)–(iii) ($F(2)(1_{st})$ = {0.6, 0, 0, 0.4}. Ψ= {1, 0.2, 0.2, 1}. (b)–(iv) is a plot when two sites are in LE and allele frequencies of both sites are the same ($F(2)(1_{st}$ = {0.36, 0.24, 0.24, 0.16}, Ψ= {1, 0.2, 0.2, 0}). (b)–(v) represents two sites having different allele frequencies and being in LD with $D' = 1$ but their $r^2 \neq 1$. ($F(2)(1_{st})$ = {0.6, 0.2, 0, 0.2}, Ψ= {1, 0.6, 0.2, 0.6}. (b)–(vi) is a plot when two sites have different allele frequencies and they are in LE. $F(2)(1_{st})$ = {0.48, 0.32, 0.12, 0.08}, Ψ= {1, 0.6, 0.2, 0.12}.

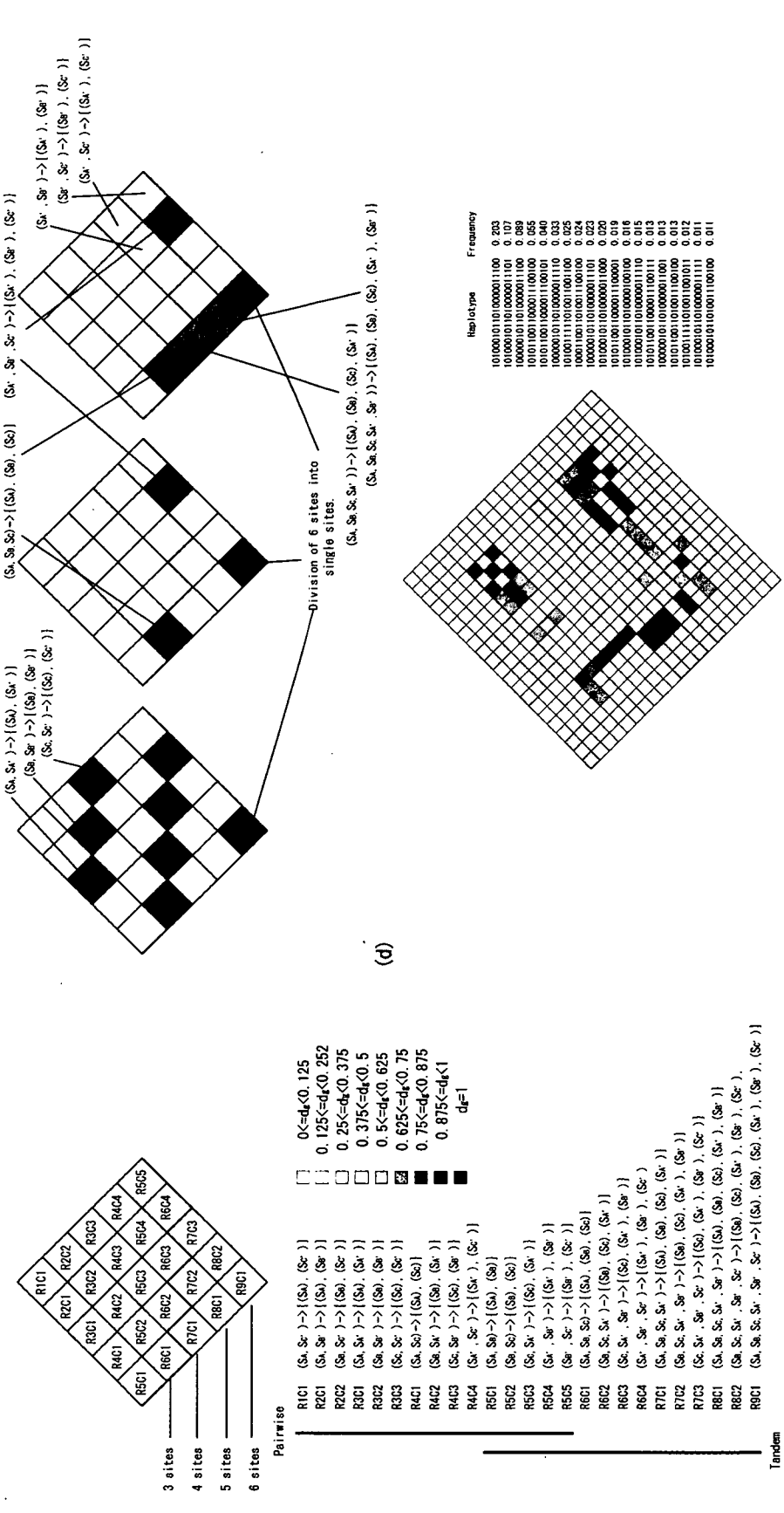


Fig. 5. D_g plots for examples of six sites. The upper halves are d_g's for site-pairs and the lower halves are for sites in tandem. All the three cases are in strong LD, but the components of LD are different each other. (a) Case1. d_g's for three site-pairs separated by two sites, four site-trios in tandem and the whole six site-set indicate strong LD. (b) Case2. All the site-pairs are in LE. d_g's for the two trio-sites on the left and right indicate LD. d_g for division of six sites into single sites also indicates LD. (c) Case3. This plot is similar to (b) with additional colored squares for site-pairs in the right-most three-site segment and for divisions of four tandem sites or five tandem sites containing all the three sites in the left-most. See text for haplotype frequencies of the three examples. (d) D_g plot for 22 site region. The upper pairwise triangle displayed extension of LD in the region, and the lower tandem triangle indicated LD components that were not captured bythe pairwise d_g's.