ARTICLE

A Definitive Haplotype Map as Determined by Genotyping Duplicated Haploid Genomes Finds a Predominant Haplotype Preference at Copy-Number Variation Events

Yoji Kukita,[1,5] Koji Yahara,[6] Tomoko Tahira,[1] Koichiro Higasa,[1,7] Miki Sonoda,[2] Ken Yamamoto,[2] Kiyoko Kato,[3,4] Norio Wake,[4] and Kenshi Hayashi[1,*]

The majority of complete hydatidiform moles (CHMs) harbor duplicated haploid genomes that originate from sperm. This makes CHMs more advantageous than conventional diploid cells for determining haplotypes of SNPs and copy-number variations (CNVs), because all of the genetic variants in a CHM genome are homozygous. Here we report SNP and CNV haplotype structures determined by analysis of 100 CHMs from Japanese subjects via high-density DNA arrays. The obtained haplotype map should be useful as a reference for the haplotype structure of Asian populations. We resolved common CNV regions (merged CNV segments across the examined samples) into CNV events (clusters of CNV segments) on the basis of mutual overlap and found that the haplotype backgrounds of different CNV events within the same CNV region were predominantly similar, perhaps because of inherent structural instability.

Introduction

Copy-number variations (CNVs) are common in the human genome.[1,2] Many are shared across populations, with some differences in frequency,[3,4] and may be involved in the etiology of disease.[5,6] For example, causative involvement of CNVs that alter the dosage of genes related to neurodevelopment has been reported in neurological diseases such as autism and schizophrenia.[7] Thus, further refinement of CNV profiles in various populations and the use of such information in GWAS of various complex diseases is a promising, but not yet fully exploited, area of study.[6]

Here we evaluated CNVs and SNPs in complete hydatidiform moles (CHMs), using a high-density DNA array hybridization system. The advantages of CHMs over conventional diploid cells for determining haplotype structures marked with SNPs and CNVs are as follows: (1) their haplotypes can be read directly by genotyping, and no phase determinations are needed; (2) they uniformly display genome-wide homozygosity, which allows CNVs to be detected with a greater signal-to-noise ratio; and (3) they do not have heterozygous sites of overlapping CNVs, which are often problematic to resolve from diploid data.[3]

The definitive haplotype map of Asian genomes presented here should complement the HapMap Project, in which Asian haplotypes were inferred from the genotypes of randomly collected individuals with the use of an assumed population model. The phasing accuracy of these haplotypes was shown to be lower than that for those of European descent or Africans, which were determined mainly with the use of a Mendelian inheritance rule of trios.[8,9] We also found a haplotype preference for recurrent CNV events; this was in contrast to SNPs, another type of genome diversity, which can be viewed as independent random mutational events.

Material and Methods

Samples

CHM tissues and leukocytes were collected from the mother, with the informed consent of each donor in a nationwide (24 prefectures) effort supported by the Japan Association of Obstetricians & Gynecologists and approved by the institutional review board (Ethical Committee of Kyushu University). Genomic DNA was extracted with a QIAamp DNA Mini Kit (QIAGEN) and diluted to 50 ng/μL with TE (10 mM Tris-HCl, 0.1 mM EDTA, pH 7.6). The DNA concentration was determined with the use of a PicoGreen dsDNA Assay Kit (Molecular Probes). All DNA samples were examined by electrophoresis on 1% agarose gels to confirm a lack of significant degradation. Samples were prescreened with the use of 17 microsatellite loci, and those that showed genome-wide homozygosity and were essentially free from contamination by the maternal genome were subjected to further analysis.[10]

Array Hybridization

DNA array hybridization to Affymetrix Genome-Wide Human SNP Array 6.0 chips (0.9 million SNPs and 0.9 million nonpolymorphic probes) was performed according to the manufacturer's instructions. After hybridization, the arrays were washed and stained with the use of a GeneChip Fluidics Station 450 (Affymetrix). Scans were performed with a GeneChip Scanner 3000 7G (Affymetrix). Output data files (CEL files) were generated with

[1]Division of Genome Analysis, Research Center for Genetic Information, Kyushu University, Fukuoka 812-8582, Japan; [2]Division of Molecular Population Genetics, Kyushu University, Fukuoka 812-8582, Japan; [3]Division of Molecular and Cell Therapeutics, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-8582, Japan; [4]Department of Gynecology and Obstetrics, Graduate School of Medical Sciences, Kyushu University, Fukuoka 812-8582, Japan; [5]Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases, Osaka 537-8511, Japan; [6]Division of Life Science System, Fujitsu Kyushu Systems Inc., Fukuoka 814-8589, Japan
[7]Current address: Division of Statistical Analysis, SNP Research Center, RIKEN, Yokohama 230-0045, Japan
*Correspondence: khayashi@gen.kyushu-u.ac.jp
DOI 10.1016/j.ajhg.2010.05.003.

GeneChip Operating Software (Affymetrix) and analyzed with the Genotyping Console (GTC 3.0.1, Affymetrix).

Five CHMs and one diploid sample were also analyzed with the use of Illumina Human1M-duo BeadChips, which interrogate 1.2 million loci, in accordance with the manufacturer's instructions (see Table S1, available online, for examined samples). The BeadChips were scanned with the BeadArray Reader (Illumina) and analyzed with BeadStudio software (Illumina) with the use of default parameter settings.

SNP Genotyping

The SNPs of the CHMs were genotyped with the Birdseed v2 module of the GTC, together with data from 45 HapMap-JPT samples (CEL files obtained from Affymetrix) that were required to obtain three genotype clusters (two homozygotes and one heterozygote). The intensity data were quantile normalized and subjected to genotyping with a confidence threshold of 0.1. The contrast quality control (QC) scores were greater than 3.9 for all CHMs, and the mean value of the scores far surpassed the recommended mean passing score of 1.7, indicating that the quality of all of the CEL files was sufficiently high to resolve the signals into three genotype clusters (Table S1).

The mean rate of homozygosity calls for 100 CHMs was 99.0% (minimum: 95.1%), and the mean rate of heterozygosity calls was 0.3% (maximum: 2.8%) (Table S1). Call rates and some QC values from the HapMap samples used in this study are shown in Tables S2 and S5.

CNV Status Called at the Single-Marker Level

The CNV status of each Affymetrix marker was assigned with the use of modules in GTC. A reference model file was created with the data from 100 CHMs. The median absolute pairwise differences (MAPDs) were less than 0.307 (Table S1), indicating that the variability of signal intensities along the chromosomes was acceptable.

In the interpretation of the Affymetrix data, the copy-number status of each marker in a particular sample was measured with the use of a *$\log_2$ ratio* value, which is the logarithm of the marker's signal intensity relative to a reference value (in this case, the median of all 100 CHM intensities). Thus, the definition of normal (i.e., *$\log_2$ ratio = 0*) was democratic rather than canonical (i.e., one copy per haploid). This means that the status of a marker could be called normal in a particular sample, even if it was not canonically normal (or vice versa), when the majority of the samples were at a CNV status for that marker in the canonical sense.

In the interpretation of the Illumina data, the indicator of copy-number status (*$\log_2$ RR*) of a marker was calculated with BeadStudio software, with the use of reference values supplied by Illumina (Human1M-Duov3_B.egt). These reference values were determined from clusters of signal intensities from selected HapMap samples and represent the expected signal intensities of markers with a canonically normal copy-number status.[11]

Selection of Shared Markers between the Affymetrix and Illumina Data Sets

Markers shared between the two data sets (Affymetrix SNP Array 6.0 and Illumina 1M-Duo) in the study shown in Figure 1 were identified by their rs numbers after several steps of filtration. Specifically, rs numbers of Affymetrix SNP markers on both the autosomes and the X chromosome were obtained from the Affymetrix annotation file (GenomeWideSNP_6.na26.1.annot.csv). If more than one marker was indicated for the same rs number, the marker with the largest Affymetrix number was chosen. The rs numbers of the Illumina markers were obtained from the UCSC Genome Browser (snpArrayIllumina1M.txt.gz). The Illumina markers were filtered such that the ID did not begin with "cnvi" and was not assigned to chromosomes "Y," "XY," or "MT." We conducted a BLAST search of the remaining markers against the reference human genome (hg18), and markers with no hits, a single hit not at the indicated positions, or multiple hits were removed. The intersection of markers, based on the rs numbers of the two filtered marker sets, was taken as shared.

Initial Detection of Candidate CNV Segments

Segmental evaluation of the copy-number states of the Affymetrix markers was performed with the GTC program, with some changes made to the parameters. This program is designed to analyze diploid samples and assigns copy-number states as integers from 0 to 4 to segments of two or more consecutive markers by interpreting the *$\log_2$ ratios* on the basis of a hidden Markov model (HMM). Our CHM samples were duplicated haploids, however, and odd copy numbers were not expected to occur. For the sake of practicality, we collected two sets of candidate CNV segments (CNVSs) by changing the parameters in the HMM. For relaxed conditions, we used the default values of expected *$\log_2$ ratios* (−2, −0.552, 0, 0.339, and 0.543) for each of the copy-number states (0, 1, 2, 3, and 4, respectively). For stringent conditions, we changed the expected $\log_2$ ratios to (−3, −2, 0, 0.543, and 0.8). For both conditions, segmental copy-number states called as 0 or 1 were translated to "deletion," and copy-number states of 3 or 4 were translated to "amplification" (Figure S2). Candidate CNVSs containing centromeric gaps were divided into two segments, assuming that the gaps always had a normal copy-number state.

Preliminary studies with quantitative PCR (qPCR) (data not shown) indicated that copy-number assignments for segments carrying three markers or less could be falsely positive. Incomplete digestion by the restriction enzymes during probe preparation can lead to false signals for the markers on the involved fragments. Therefore, the candidate CNVSs obtained under both conditions were filtered so that they carried four or more markers and overlapped with at least two restriction fragments, which were judged according to the Affymetrix annotation data.

The candidates obtained under relaxed conditions were further filtered for removal of the segments with a mean $\log_2$ ratio between −1 and 0.5. These threshold values were empirically determined from the results shown in Figure S2. The filtered candidate CNVSs obtained under both conditions were then merged to define CNVSs.

Validation of CNV Status by qPCR

qPCR was performed with the StepOne real-time PCR system (Applied Biosystems). Primer3[12] was used to design primers to amplify 90–120 bp fragments positioned within chosen CNV regions (CNVRs) (Table S10; see the following subsection for the definition of CNVRs). Reactions were prepared in a total of 20 μl containing Power SYBR Green PCR Master Mix (Applied Biosystems) and 10 ng of genomic DNA. The cycling conditions were as described in the manufacturer's guidelines. The amplification profiles were normalized with the use of a product from LINE-1 elements.[13] The copy number in each sample at the examined

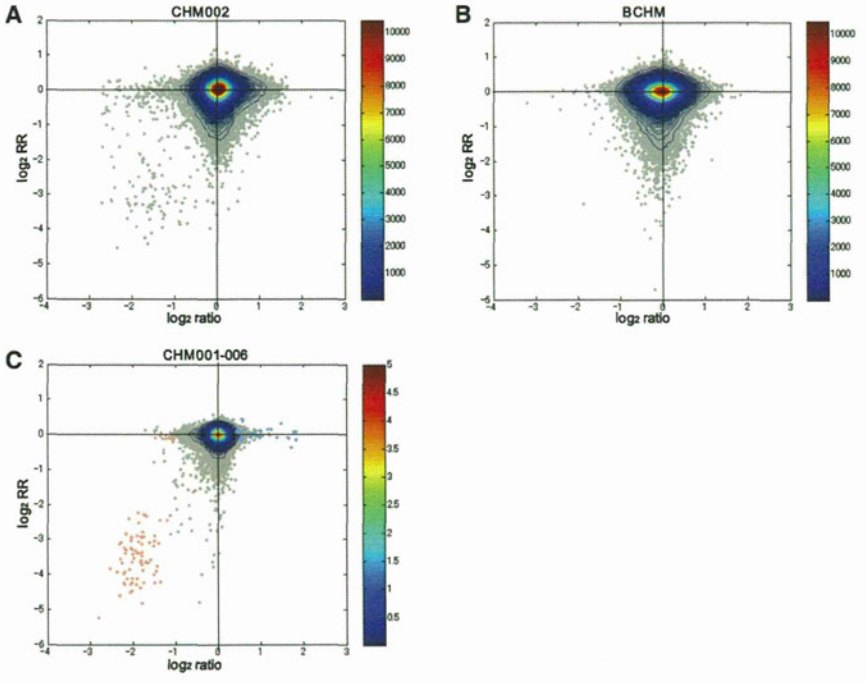

Figure 1. Comparison of CHM and Diploid Samples in the Detection of Copy-Number Status at the Marker Level
(A and B) A CHM sample (A) and a diploid sample (B) were compared on the basis of their relative signal intensities of 280K SNP markers that were common to both the Affymetrix SNP Array 6.0 (log_2 *ratio*, *x* axis) and the Illumina Human 1M-duo (log_2 *RR*, *y* axis) arrays.
(C) CNV segments and normal bins were determined for five CHMs (CHM001, CHM002, CHM003, CHM005 and CHM006; see Table S1) as described in the text, and CNV segments (red for deletion and blue for amplification) or bins (gray) were plotted according to the means of the log_2 *ratios* and log_2 *RRs* for the included markers.

locus was calculated from three replicate reactions with the use of the comparative threshold cycle (Ct) method.[14]

The positions analyzed on the reference genome (hg18) were: chr1:232772797-232772913 (CNVR84), chr3:3898625-3898743 (CNVR184), chr3:101512697-101512816 (CNVR221), chr3:114104343-114104462 (CNVR226), chr5:107704382-107704501 (CNVR402), chr7:26269751-26269868 (CNVR534), chr8:142926423-142926542 (CNVR712), chr11:5228832-5228946 (CNVR833), chr11:119967281-119967399 (CNVR880), chr13:21553526-21553644 (CNVR954), chr19:15862386-15862535 (CNVR1192), chrX:16399969-16400088 (CNVR1288), chr5:143388542-143388661 (CNVR412), chr9:10397271-10397390 (CNVR721), chr10:120166429-120166546 (CNVR822), chr1:40739157-40739274 (CNVR21), chr4:98394328-98394447 (CNVR315), chr12:89016000-89016119 (CNVR936).

Results

SNP Genotyping

The CHM samples were examined by hybridization experiments with the Affymetrix SNP Array 6.0. The intensity data generated were then analyzed for SNP genotypes and CNV status with several QC steps as summarized in Figure S1.

We compared the obtained genotype calls with our previous results from 500K arrays using 99 shared CHM samples,[15] and the concordance of homozygous calls was greater than 99.99% (Table S3). Five of the CHMs were also genotyped with Illumina 1M-duo. The genotype concordance of shared SNPs between the Affymetrix and Illumina calls was 99.99% for homozygous calls and 2.05% for heterozygous calls (Table S4). The SNP genotypes of the CHMs were further filtered on the basis of their CNV status, as described later.

Haploid versus Diploid in Detection of CNVs

We next asked how well the CNV status of haploid material could be captured at each marker level by comparing data sets from five CHMs with those of a diploid sample, all of which were analyzed by both Affymetrix SNP Array 6.0 and Illumina 1M-Duo. We chose 280K markers that were shared between the Affymetrix and Illumina data sets (see "Selection of Shared Markers between the Affymetrix and Illumina Data Sets" in Material and Methods), and the signal intensities of each marker determined by the two systems were plotted according to their log_2 *ratio* versus log_2*RR* (see "CNV Status Called at the Single-Marker Level" in Material and Methods).

As illustrated in Figure 1A for a single CHM, a cluster of marker signals was observed in the third quadrant, indicating that the markers in deleted regions were readily recognizable by both systems and were well separated from the majority of the markers with normal copy-number status. Similar results were obtained for all five CHMs examined with both platforms (data not shown). In contrast, such a cluster was virtually absent from the third quadrant when the diploid sample was examined (Figure 1B), clearly demonstrating the advantage of CHM samples over conventional diploid samples in detecting CNVs.

This is in accordance with the expectation that markers deleted in CHMs should have a null copy number and that the intensities of these markers relative to those with a normal copy number should be zero (or close to zero because of the background signal). Most of the deletions in diploid samples are likely to be heterozygous; therefore, their intensities should be around 0.5 relative to markers with a normal copy number. The difference is much more pronounced when the ratios are expressed on a logarithmic scale.

The advantage of CHMs was less evident in the detection of amplifications, especially for the log_2*RR* values.

However, a slight increase in outliers in the first quadrant was discernible when the CHM plot was compared with the diploid plot. Saturation of hybridization is a possible reason for the poor resolution of amplifications and has been reported previously in the case of the Illumina system.[11]

Definition of CNV Segments

The judgment of CNV status at the single-marker level was still ambiguous as evidenced by the continuous distribution of signals between the third-quadrant cluster and the peak of the normal copy signal at the origin seen in Figure 1A. Therefore, CNV status was evaluated by the continuity of markers; that is, by segments. CNVSs were identified with the use of the Affymetrix data only. We removed five CHMs prior to segmental evaluation because visual examination of whole-genome profiles of signal intensities (*log_2 ratios*) indicated that the data for these samples were grossly abnormal at several points. These abnormalities included apparent whole X chromosome amplifications with many heterozygous sites on the chromosome, apparent amplifications of more than 5 Mb in two chromosomes, apparent amplifications of all telomere regions (two samples), and many apparent deletions along G-bands and could be ascribed to poor sample quality, suboptimal hybridization, or atypical CHMs (see Table S1 for a summary of the samples and their QC results).

For the remaining samples, potential CNVSs were identified with the GTC program, which employs a hidden Markov model (HMM), with modifications as detailed in "Initial Detection of Candidate CNV Segments" in the Material and Methods section. In brief, candidate CNVSs collected under relaxed conditions were filtered on the basis of their respective means of *log_2 ratios* and merged with those obtained under stringent conditions to define CNVSs. With the use of these procedures, a total of 8682 CNVSs were identified for the 95 CHMs examined (Figure S1). Of these CNVSs, 822 segments consisted solely of filtered relaxed CNVSs, whereas 407 segments were fusions of two or more stringent segments overlapped with relaxed segments. Filtered relaxed segments that included single stringent segments made up the remaining CNVSs.

To obtain some idea of the false-negative rate for the segment assignment described above, we examined the regions outside the CNVSs. Inter-CNVS regions of the five CHMs examined by both the Affymetrix and Illumina systems were divided into bins. Each bin carried four Affymetrix markers that overlapped by at least two Affymetrix restriction fragments and had three or more Illumina markers. The mean *log_2 ratio* for the Affymetrix markers and the mean *log_2 RR* for the Illumina markers were then calculated for each bin. Figure 1C shows a scatterplot of the bins (gray dots) and the CNVSs identified as described above (red dots for deletions and blue dots for amplifications) in the space of the mean *log_2 ratio* versus mean *log_2 RR*.

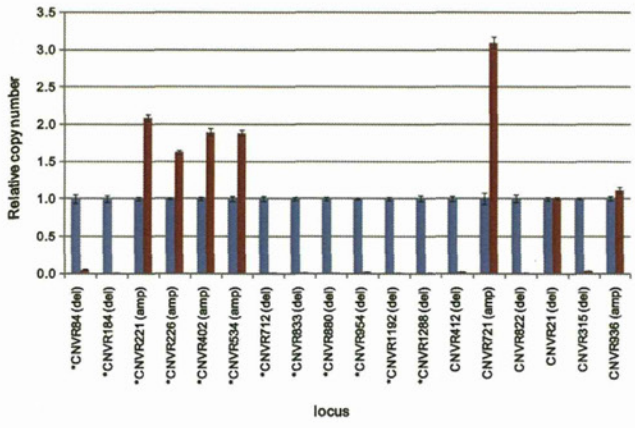

Figure 2. Validation of CNV Segments by qPCR
Twelve singleton CNVRs (asterisks) and six multihit CNVRs were examined by qPCR. Their copy numbers were determined for the samples without copy-number change (blue) or with copy-number change (red). Error bars represent the standard deviation from three determinations. See the text and "Validation of CNV Status by qPCR" in the Material and Methods section. Of the 18 regions examined, copy-number changes were confirmed in 16. See Table S10 for the chromosomal positions of the CNVRs.

As shown in the figure, normal bins that fell within the area of deletions (*log_2 ratio* < -1) comprised an extremely small fraction (0.8×10^{-4}) of the total bins, and most of the bins with a *log_2 RR* less than -2 were clustered near the *y* axis. Approximately 60% of these bins were included in the deletion copy-number polymorphisms (CNPs) that have been described as being common in JPT.[3] This corroborates the characteristics of normal copy number defined by the GTC program (democratic definition) as noted in "CNV Status Called at the Single-Marker Level" in the Material and Methods section. Furthermore, bins with a mean *log_2 ratio* value greater than 0.5 occurred at a very small fraction (5.3×10^{-4}). Thus, we believe that most CNVSs were captured in the present study, with the caveat that CNV status was defined under the assumption that the status of the majority of the samples was normal.

Confirmation of Copy Number by qPCR

Using qPCR, we examined 18 loci within CNVRs (see below for the definition of CNVRs). Twelve of the loci were singletons (copy-number change detected only in one CHM), and of these, eight were at genomic positions that did not overlap with any reported CNVs according to the UCSC database (hg18 DGV StructVarTrack, version 5).[16] The remaining six loci were from six different CNVRs for which multiple CHMs revealed copy-number changes. For each region, two CHMs were examined: one showing a copy-number change and the other showing a normal copy number (control CHM) with respect to the locus.

The qPCR results were interpreted such that fold changes less than 0.5 or greater than 1.4 were considered to indicate a loss or gain of copy number, respectively. Copy-number changes were confirmed for all but two loci (Figure 2). These failures could have been due to fortuitous

amplification in qPCR, possibly because the amplicons overlapped with regions of segmental duplications.[17]

Removal of SNP Genotypes in Deletions followed by Sample QC

In comparing SNP and CNVS data, we noticed that genotypes were called for some SNPs in deleted regions. Because the CHMs examined here contained duplicated haploid material, the SNP genotypes called within deletions were likely false. High rates of heterozygous calls of SNPs with a low (< -0.5) *log_2 ratio*, in contrast to almost entirely homozygous calls for other SNPs, support the conclusion that the majority of the genotypes of the SNPs with low *log_2 ratios* were false (Figure S3). Therefore, we forced genotypes called at a *log_2 ratio* less than -0.5 and those within deletions to be "no call." Approximately 2% of the total SNP calls were rendered "no call" by this filtration step (Figure S1).

Approximately 0.2% of the calls still remained heterozygous, and this could, in principle, be interpreted as evidence that they were in paralogous sequences. The concordance of heterozygous calls for shared SNPs in two comparisons (between Affymetrix SNP Array 6.0 and Affymetrix 500K[15] and between Affymetrix SNP Array 6.0 and Illumina1M-Duo BeadChip), however, were extremely low (1.48% and 2.05%, respectively) (Tables S3 and S4). Therefore, we concluded that error, rather than the presence of paralogous sites, was responsible for the heterozygous calls, and all remaining heterozygous calls were also classified as no calls. After these filtering steps, the call rates of ten CHMs dropped below 95%, and these samples were excluded from further analyses (see Table S1 for QC summary). We also removed one CHM because principal-component analysis revealed that this sample appeared to have exceptionally mixed ancestry and was not suitable as a data source for a typical Japanese population as previously described.[15] As a result of these filtering steps, the call rates of 32,205 SNPs dropped below 85%, and these SNPs were removed (Table S7).

Definitive Haplotype Structures of SNPs and CNVSs

After the refinements described above, the haplotypes of SNPs and CNVSs were definitively delineated on a map containing data from the final 85 CHMs. This map described a total of 875,826 SNPs on autosomes and the X chromosome, 55% of which were 100% called (all 85 CHMs had genotypes) and more than 95% of which were called at least 93% of the time (79 CHMs had genotypes) for the SNPs (Tables S6 and S7).

A total of 6770 CNVSs (4255 deletions and 2515 amplifications) from the 85 CHM samples were included on the map (listed in Table S8). These CNVSs occupied 3.1 Mb per haploid genome (Table S9), in agreement with the previously estimated CNV burden (i.e., equivalent to one half of the value per diploid genome[3]). Approximately 33% of the CNVSs overlapped with segmental duplications, whereas the overlap was 84% in the combined length of CNVSs, indicating that the CNVSs overlapping with segmental duplications were much larger than those without overlap. The large discrepancy between the means and medians of the segment sizes indicates extreme heterogeneity in the size distribution of the CNVSs (Figure S4), especially for those overlapping with segmental duplications.

CNVRs

CNVRs were defined as mergers of CNVSs across the 85 CHMs and given genome-wide numbers that started at CNVR1, located nearest to the terminus of the short arm of chromosome 1. A total of 1336 CNVRs was identified (listed in Table S10), and 582 of these were mergers of two or more CNVSs (multihit CNVRs) (Table S11). More than half of the CNVRs (754, or 56.4%) were singletons, but singletons accounted for only 11.1% of the detected CNVSs, indicating that most of the CNVSs overlapped with one another.

The fact that there is a greater chance of observing multihit CNVRs (i.e., CNV regions consisting of multiple CNVSs) in regions of segmental duplications known to be preferred sites for nonallelic homologous recombination[18] suggests that many of the multiple hits could be attributable to recurrent ancestral events, not an expansion of the results of single-CNV events in the population.

We compared the CNVRs identified here with previously defined CNPs in a Japanese population (JPT-CNPs) that were also identified with the Affymetrix SNP Array 6.0.[3] CNPs have been defined as regions where the copy numbers of included markers tend to vary in a concerted manner among individuals in populations, and they do not overlap with each other.[3] The comparison was limited to CNPs and CNVRs on autosomes with an allele frequency of 2% or higher (two or more segments per regions) for both data sets. We also excluded CNVRs that overlapped with segmental duplications from the comparison, because these CNVRs were often very large and spanned regions where markers were very sparse, making precise coverage of the genome ambiguous. With the use of these criteria, approximately 60% of CNPs found in JPT samples overlapped with our CNVRs, accounting for 40% of our CNVRs (Figure 3A).

These values for the overlap between CNVRs and CNPs were lower than expected (greater than 90%) if CNPs and CNVRs were present at similar frequencies in both the JPT samples and the CHM samples. Part of the reason for this discrepancy could be explained by differences in the definitions of CNVRs and CNPs. The lower threshold in the definition of CNVRs was based on the number of markers (four or greater) in the regions; thus, some CNVRs were short. On the other hand, many of the candidate short regions were filtered out during QC steps in the CNP definition and were likely underrepresented.[3] As a result, approximately 25% of CNVRs were shorter than 2 kb (Figure S4C), whereas less than 8% of CNPs were shorter than that length. It is unknown whether these

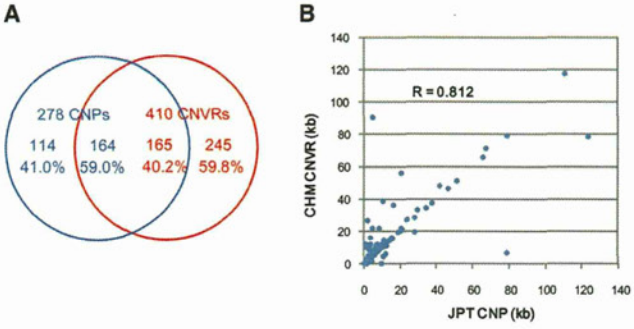

Figure 3. Overlap of CNPs with CNVRs or CNVEs
(A) The overlap of CNVRs (red) and CNPs (blue) reported for JPTs[3] is shown. CNVRs or CNPs on autosomes that were frequent (> 2%) and nonoverlapping with segmental duplications were compared. Values below are percentages in the respective data sets.
(B) The sizes of overlapping CNVRs and CNPs were compared.

differences in the definitions explain most of the discrepancies in the overlapping or not.

A comparison of the sizes of CNVRs with Japanese CNPs that overlapped with each other revealed a high correlation, although with some discrepancies (Figure 3B). Essentially all of the CNVRs with sizes greater than an overlapped CNP were found to contain rare (mostly one), large CNVSs that caused an expansion in the size of the CNVRs.

CNV Events

Visual examination of multihit CNVRs revealed that many of them consisted of two or more clusters of CNVSs with different ends and were likely to have resulted from different ancestral events of segmental deletion or amplification. In an attempt to resolve these events, CNV events (CNVEs) were defined as clusters of CNVSs.[4] Specifically, CNVSs in each CNVR were clustered with the use of a greedy algorithm that consisted of the following steps: (1) groups of CNVSs were determined by their mutual overlap at or above a threshold value; (2) the largest group was identified, and the CNVSs within this group were merged and named a CNVE; (3) the CNVSs belonging to the CNVE were removed, and the procedure was repeated from step 1 until the CNVSs were exhausted. If two or more of the largest groups were found in step 2, the first group identified during the process was adopted. CNVEs were cumulatively numbered, starting from CNVE1 as the first CNVE identified in CNVR1.

By choosing an overlap threshold of 51% of the physical distance, 582 multihit CNVRs were resolved into 1124 CNVEs (listed with allele frequencies in Table S12). Further visual inspection suggested that many of the CNVEs defined here were still heterogeneous and could likely be divided into subevents. We did not attempt to resolve these regions further, due to the difficulty in meaningfully improving event detection because of the extreme bias of marker distribution in or near many CNVRs.

Table 1. Capturing CNVRs and CNVEs by SNPs

Region or Event	Number of Sites	Mean of Max r^2	Fraction Captured[a] at $r^2 \geq 0.5$	Fraction Captured[a] at $r^2 \geq 0.8$
CNVRs	130	0.68	0.70	0.49
CNVEs	164	0.59	0.59	0.41

Common deletion CNVRs or CNVEs (frequency ≥ 5%) without segmental duplications were analyzed for linkage disequilibrium with SNPs that were on SNP Array 6.0, located within 200 kb from the boundaries of regions or events with a minor allele frequency ≥ 5%.
[a] Fractions of CNVRs or CNVEs that were captured by at least one SNP at the indicated r^2 values.

Capturing CNVs by Linkage Disequilibrium with SNPs

We asked how well CNVRs could be captured by linkage disequilibrium with SNP alleles. The examination was limited to common CNVRs (minor allele frequency > 5%) that were deletion changes only and occurred in nonduplicated regions, in order to minimize the effects of possible errors on the definition of CNVSs. As shown in Table 1, approximately one half of the common CNVRs remained uncaptured (maximum $r^2 < 0.8$) by SNP markers on SNP Array 6.0.

McCarroll et al. and Cooper et al. have shown that the capture rate of CNV regions by SNPs was approximately half of the rate of SNPs, when the platform Affymetrix SNP Array 6.0 was used.[3,19] They also showed that scarcity of effective SNP markers in the vicinity of CNVRs relative to other genomic regions was the reason for poor capturing of CNVRs. Our observation was in accordance with these earlier reports.

We found that the capture rate (with a maximum $r^2 > 0.8$) of amplification CNVRs was lower (0.37) than that of deletion CNVRs (0.47, including those in segmental duplications). An altered physical relationship between CNVRs and adjacent SNPs in samples with amplifications (e.g., due to the location of the amplified copy at a chromosomal position different from original position) is among the possible explanations of the lower capture rate. We also found that deletion CNVRs overlapping with segmental duplications showed a lower capture rate (0.30) in comparison to those in unique regions (0.49), most likely because of the scarcity of SNP markers in segmental duplications.[3]

Capture rates can also be reduced if the CNVRs are ancestrally heterogeneous; that is, if they consist of two or more CNVEs that occurred independently. In such cases, each of the CNVEs should be more efficiently captured than the CNVRs; however, we found that the capture rate of the CNVEs was consistently low (Table 1). We also defined CNVEs by reciprocal overlap of CNV segments on the basis of the number of markers rather than physical distance, and essentially the same results were obtained (data not shown). These observations are seemingly the opposite of the anticipated results and can be explained if CNVEs within a CNVR have common haplotype backgrounds.

Haplotype Preference of CNVEs

To test the possibility of haplotype sharing between CNVEs, we chose common deletion CNVRs that did not overlap with segmental duplications, consisted of multiple CNVEs, and had at least one common event (allele frequency 5% or higher). We further restricted the comparison by requiring any two CNVEs to be distinguishable by at least two markers and not allowing any of the CNVRs to contain interrupted CNVEs in any of the samples. The rationale for this restriction was to avoid false haplotype similarity caused by erroneous splitting of single events. A total of 35 CNVEs in 17 CNVRs met these criteria. The similarities in haplotype background between common CNVEs within the same CNVR were then examined.

The haplotypes examined here were those defined by SNPs found within 200 kb of both ends of each CNVR. As a measure of haplotype similarity between two CNVEs in a CNVR, we calculated the mean homozygosity of haplotype pairs between every sample in one CNVE and every sample in the other CNVE (observed between-events homozygosity). The tendency of recurrence of the two CNVEs in particular haplotypes was then evaluated against their occurrence in independent haplotypes (which is the expected between-events homozygosity under the assumption of independent occurrence) by bootstrapping the second events. Specifically, the null distribution of homozygosity was generated from 10,000 sets of haplotype pairs with the assumption that the second CNVEs occurred randomly in any of the observed haplotypes of all samples. The probability densities of the null distributions were obtained by kernel-density estimation with the use of R.[20] The comparison was limited to 26 cases that gave a unimodal probability density of null distributions as judged by visual inspection. The empirical p value for the occurrence of observed homozygosity in the null distribution was then estimated (see footnote of Table 2).

As shown in Table 2, the means of the homozygosity between events were predominantly higher than the means of the null distributions (24 of 26), and the differences were significant for most comparisons (21 of 26, or 12 of 26 after Bonferroni correction), despite the fact that the number of alleles examined was small. These results indicate that the recurrence of CNVEs is strongly dependent on haplotype. The 12 comparisons that showed strong haplotype similarity were between CNVEs in ten CNVRs, and nine of these CNVRs overlapped with CNPs. The CNVRs carrying CNVEs with significantly similar haplotype backgrounds are shown with the use of the UCSC Genome Browser with modification of some lane names for better visualization (Figure 4 and Figure S5). Figure S6 illustrates the haplotype profiles of CNVE samples and non-CNV samples for all of the CNVRs listed in Table 2 (an example is shown in Figure 5). As is evident from the figure, remarkable haplotype sharing between CNVE samples was evident when compared with non-CNV samples, especially near each of the CNVRs, with one exception (CNVR 273; see Figure S6). In this exceptional CNVR, the two CNVEs seemed to have arisen from different haplotypes.

Discussion

We determined the haplotype structures of SNPs and CNVSs in Asian genomes, taking advantage of CHMs and their haploid genomes. SNP haplotypes[8,21] and CNV maps[3,4] have been reported previously with the use of HapMap populations; however, the phasing accuracy of the Asian haplotypes has been shown to be more than 10-fold lower than the phasing accuracy for individuals of European descent and Africans.[9] The high-resolution SNP and CNV definitive haplotype map presented here for a Japanese population is based on the examination of 100 CHMs, which are naturally occurring haploid human samples. Therefore, these haplotypes are definitive, and the phases are accurate.[10]

Recent studies have indicated that the maternal physiological state is responsible for mole formation, whereas the sperm genome does not seem to play a role. Thus, the genomes of CHMs can be regarded as unbiased samples of sperm genomes.[22,23] More than 95% of the CHMs studied here were collected within 13 wks of gestation. In such a short period, these CHMs were unlikely to have been subjected to extensive selection. This is in contrast to cultured cell lines, including some HapMap samples known to carry large CNV segments that probably arose during extensive culturing and were fixed by repeated passaging.[4]

CHM genomes have not been biologically proven to be complete in the sense of being capable of supporting the normal development of individuals. Abnormalities that occur de novo in paternal germ cells may remain unselected, so long as the abnormality does not influence cell growth. Such events, however, are likely to be rare.

We genotyped CHMs by using available high-density DNA arrays, and we determined their CNV structures by using a modification of an available method. The copy-number status of each marker in each sample was judged by its signal intensity relative to the intensity of the majority of the samples, which can yield results that differ from the canonical copy-number status (i.e., one copy per haploid), as mentioned earlier. The Canary algorithm[24] assigns absolute copy numbers of predefined CNPs for each sample;[3] however, this algorithm was developed specifically for diploid samples and could not be directly applied to our haploid samples. Considering this limitation, we analyzed our data by using the Canary analysis module integrated in GTC, assuming that copy numbers of 0 or 1 were deletions and that copy numbers of 3 or 4 were amplifications. As a result, a total of 537 biallelic CNPs were identified, 283 of which overlapped with our biallelic CNVRs. Of these 283 CNPs, 29 were copy-number changes in opposite directions. Thus, approximately 10% of the CNVRs detected were possibly in a copy-number state opposite to the canonical state.

Table 2. Haplotype Preference of CNVEs

CNVR	Chr.	First CNVE	Second CNVE	No. of Pairs	Observed[a]	Null[b]	Difference	p Value
CNVR154	2	CNVE228	CNVE227	75	0.7455	0.6151	0.1304	*0*
CNVR1199	19	CNVE1685	CNVE1684	14	0.8342	0.6704	0.1637	*0*
CNVR1079	15	CNVE1509	CNVE1508	23	0.8737	0.7079	0.1658	*0*
CNVR315	4	CNVE432	CNVE431	40	0.8993	0.6347	0.2646	*0*
CNVR1251	21	CNVE1771	CNVE1770	52	0.9096	0.7458	0.1638	*0*
CNVR219	3	CNVE304	CNVE303	28	0.9165	0.7028	0.2137	*0*
CNVR55	1	CNVE103	CNVE102	17	0.9592	0.7225	0.2366	*0*
CNVR328	4	CNVE448	CNVE449	56	0.7155	0.6316	0.0839	*0.0001*
CNVR1128	16	CNVE1592	CNVE1591	8	0.8284	0.6387	0.1897	*0.0003*
CNVR774	10	CNVE1096	CNVE1095	54	0.8332	0.75	0.0833	*0.0008*
CNVR1251	21	CNVE1770	CNVE1771	52	0.9096	0.7242	0.1854	*0.0008*
CNVR1251	21	CNVE1772	CNVE1771	4	0.9351	0.7148	0.2203	*0.0014*
CNVR633	8	CNVE863	CNVE862	56	0.6975	0.6503	0.0472	0.002
CNVR328	4	CNVE449	CNVE448	56	0.7155	0.641	0.0745	0.0039
CNVR774	10	CNVE1095	CNVE1096	54	0.8332	0.747	0.0862	0.006
CNVR154	2	CNVE227	CNVE228	75	0.7455	0.6234	0.1222	0.0111
CNVR592	7	CNVE796	CNVE795	18	0.7494	0.6779	0.0715	0.0115
CNVR1125	16	CNVE1588	CNVE1587	13	0.6877	0.6396	0.0481	0.0157
CNVR633	8	CNVE862	CNVE863	56	0.6975	0.6376	0.06	0.016
CNVR152	2	CNVE225	CNVE224	81	0.6713	0.6324	0.0389	0.0169
CNVR1251	21	CNVE1771	CNVE1772	4	0.9351	0.7464	0.1886	0.0496
CNVR1202	19	CNVE1690	CNVE1689	11	0.777	0.7462	0.0308	0.084
CNVR592	7	CNVE795	CNVE796	18	0.7494	0.6867	0.0628	0.1904
CNVR152	2	CNVE224	CNVE225	81	0.6713	0.6495	0.0219	0.2153
CNVR273	4	CNVE375	CNVE374	18	0.4741	0.5812	−0.1071	0.9962
CNVR649	8	CNVE912	CNVE911	25	0.5285	0.6107	−0.0823	0.9997

[a] Observed similarity of haplotype backgrounds between CNVEs in the same CNVR, which was measured by the averaged homozygosity of every between-event haplotype pair.
[b] Expected similarity was obtained by bootstrapping to generate null distributions of averaged homozygosity and under the assumption that one of the CNVEs could arise randomly from any of the observed haplotypes. See the text for details regarding the analysis. p values in italics were significant after Bonferroni correction. Additional information on each of the CNVRs and CNVEs is given in Tables S10 and S12.

McCarroll et al. defined CNPs as regions where the copy numbers of included markers tend to vary in a concerted manner among individuals in populations.[3] By definition, CNPs do not overlap, and many of them seem to behave like biallelic polymorphisms. Recently, however, many CNPs have been shown to be resolvable to several different ancestral events.[25,26] Therefore, we attempted to resolve CNVRs into CNVEs by reciprocal overlaps of CNVSs. The resolution was far from perfect, and many of the CNVEs seemed to consist of subevents; however, different origins of ancestral events were evident between different CNVEs.

Comparisons of surrounding haplotypes between CNVEs belonging to the same CNVR revealed that most of the haplotypes were significantly similar. One plausible explanation for this is that the presence of CNVSs induces instability in the region and encourages secondary amplifications or deletions within the same allele, although other explanations are also possible. Although this scenario sounds like a remote possibility, it may not be if one considers the situation of CNVSs in meiosis. During meiosis, CNVSs are almost always paired with normal counterparts (given their low allele frequencies, at least when they are newly formed), and the local instability caused by imperfect asymmetric homologous paring of chromatids may render these sites or their vicinity vulnerable to secondary events such as amplifications or deletions.

The similarity of the haplotype backgrounds between CNVEs in the same CNVR has been implicated, although

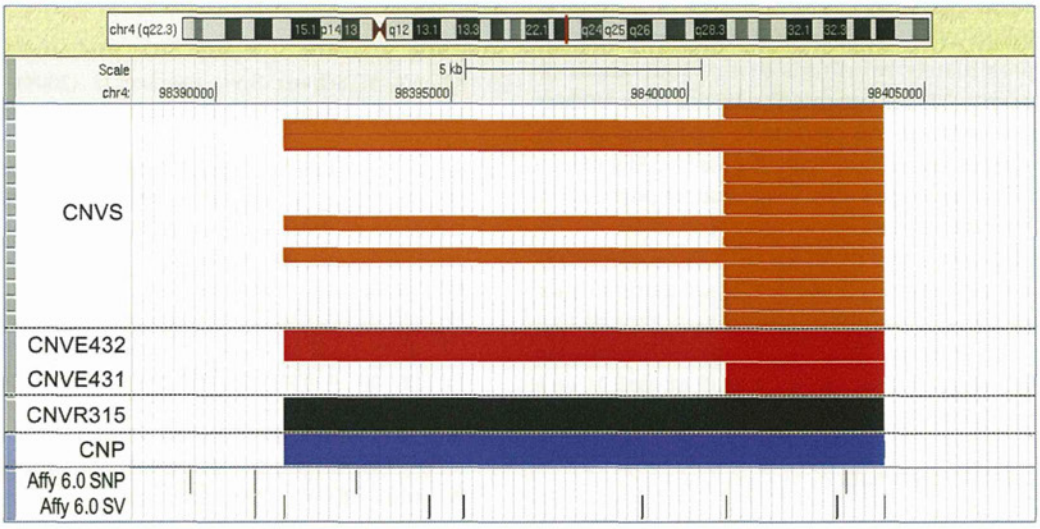

Figure 4. Map View of CNVRs Carrying CNVEs with Significant Haplotype Similarity
An example of a CNVR carrying CNVEs with significantly similar haplotype backgrounds is shown with the use of the UCSC Genome Browser. Other examples are presented in Figure S5. Thin bars in orange indicate the positions of CNVSs in individual CHMs. Thick bars in red, black, and blue represent the positions of CNVEs, CNVRs, and CNPs,[3] respectively. The bottom two lanes show the positions of SNP markers (Affy 6.0 SNP) and CNV markers (Affy 6.0 SV) in the Affymetrix SNP Array 6.0.

not explicitly stated, in previous reports.[3] McCarroll et al. demonstrated that most CNPs could be captured at a high linkage disequilibrium by nearby SNPs if the SNPs used were of sufficiently high density to allow estimation of the capture rate, despite the fact that some of the CNPs were clusters of CNVEs. These findings are most easily understood if haplotype-dependent recurrence of CNVEs is assumed. The possible dependence of CNVE occurrence on preexisting events is in contrast to SNPs, which can be regarded as the result of independent, random events.

The determination of CNV structure with the use of available arrays involves some uncertainty because of the extremely uneven distribution of markers, as noted previously.[3,19] Perhaps significant improvement in the detection of CNVs must await the availability of arrays carrying an unbiased distribution of markers. Recently, Conrad et al. reported an advanced CNV-typing array system that can efficiently detect even small CNVs.[27] With the use of this system, the detection of CNVs in existing materials should be improved; however, this system still suffers

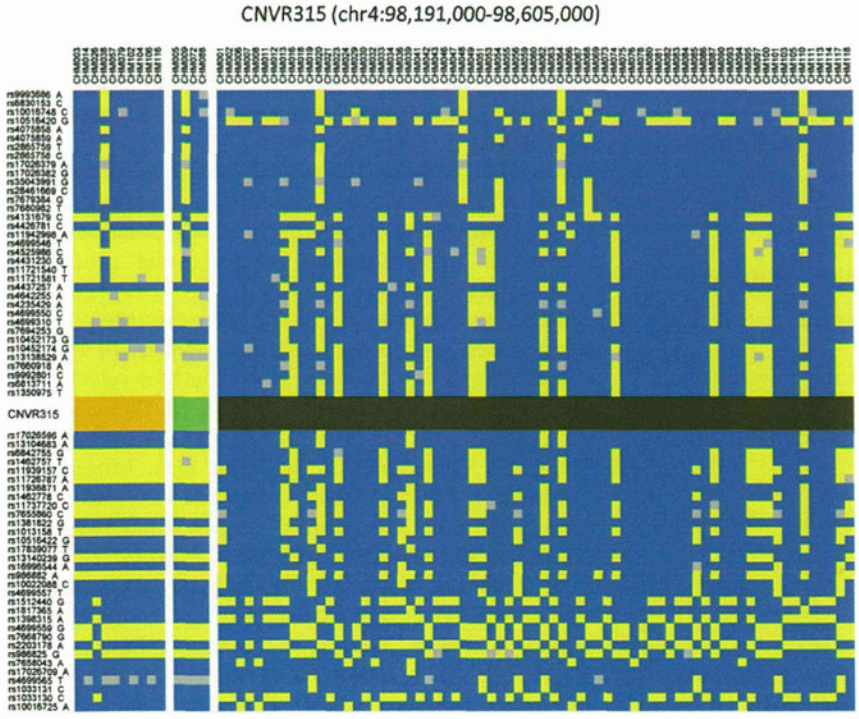

Figure 5. An Example of Haplotype Sharing between CNVEs
Haplotype profiles of CNVE samples (different CNVEs are color-coded by yellow or green in CNVR lines) and non-CNV samples (black in CNVR lines) for CNVR315 are shown. The major and minor SNP alleles are shown in blue and yellow, respectively, and SNPs with no genotype calls are shown in gray. See Figure S6 for the profiles of other CNVRs listed in Table 2.

from the fact that detecting CNVs in the Asian genome is highly inefficient (the number of CNVs detectable in Asians is approximately two-thirds that of individuals of European descent). This is because the initial experiments conducted to determine the markers to be loaded in the typing arrays were carried out with the use of European-descent and African samples, resulting in some population bias in the detection efficiency of the typing array.

Non-hybridization-based methods such as resequencing by new-generation sequencers are obviously among other future approaches. CHM samples provide an exceptional opportunity for effective whole-genome resequencing because CHMs display genome-wide homozygosity and require less sequencing redundancy. Furthermore, the reads can be aligned with greater confidence, unlike resequencing of diploid materials.

Supplemental Data

Supplemental Data include six figures and twelve tables and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

We thank members of the Japan Association of Obstetricians & Gynecologists for their cooperation in collecting mole samples. We also thank Professor Yanagawa (Division of Biostatistics and Infectious Diseases, Kurume University School of Medicine, Kurume, Fukuoka) for help with the statistical evaluation of the haplotype preference of CNVEs. This work was supported by KAKENHI #17019051 (Grant-in-Aid for Scientific Research on Priority Areas "Applied Genomics"), KAKENHI #18710163 (Grant-in-Aid for Young Scientists [B]), and KAKENHI #20681020 (Grant-in-Aid for Young Scientists [A]) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, as well as by a grant from the Osaka Cancer Society.

Received: February 10, 2010
Revised: April 13, 2010
Accepted: May 7, 2010
Published online: May 27, 2010

Web Resources

The URLs for the data and software used herein are as follows:

Affymetrix: Genotyping Console software and annotation files, http://www.affymetrix.com/
Database of Genomic Variants, http://projects.tcag.ca/variation
dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/
Illumina: BeadStudio software and other requirement files, http://www.illumina.com/
R software, http://www.R-project.org
UCSC Genome Browser: genome annotation and SNP array marker information, http://genome.ucsc.edu/

Accession Numbers

The Gene Expression Omnibus (GEO) accession number for the array intensity data reported in this paper is GSE18701.

References

1. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*, 949–951.
2. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004). Large-scale copy number polymorphism in the human genome. Science *305*, 525–528.
3. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.
4. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.
5. Feuk, L., Marshall, C.R., Wintle, R.F., and Scherer, S.W. (2006). Structural variants: changing the landscape of chromosomes and design of disease studies. Hum. Mol. Genet. *15*(Spec No 1), R57–R66.
6. McCarroll, S.A. (2008). Extending genome-wide association studies to copy-number variation. Hum. Mol. Genet. *17*(R2), R135–R142.
7. Cook, E.H., Jr., and Scherer, S.W. (2008). Copy-number variations associated with neuropsychiatric conditions. Nature *455*, 919–923.
8. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al; International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
9. Kidd, J.M., Cheng, Z., Graves, T., Fulton, B., Wilson, R.K., and Eichler, E.E. (2008). Haplotype sorting using human fosmid clone end-sequence pairs. Genome Res. *18*, 2016–2023.
10. Kukita, Y., Miyatake, K., Stokowski, R., Hinds, D., Higasa, K., Wake, N., Hirakawa, T., Kato, H., Matsuda, T., Pant, K., et al. (2005). Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles. Genome Res. *15*, 1511–1518.
11. Peiffer, D.A., Le, J.M., Steemers, F.J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C.A., Belmont, J., et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. *16*, 1136–1148.
12. Rozen, S., and Skaletsky, H.J. (2000). Primer3 on the WWW for general users and for biologist programmers. In Bioinformatics Methods and Protocols: Methods in Molecular Biology, S. Krawetz and S. Misener, eds. (Totowa, NJ: Humana Press), pp. 365–386.
13. Wang, T.L., Maierhofer, C., Speicher, M.R., Lengauer, C., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. (2002). Digital karyotyping. Proc. Natl. Acad. Sci. USA *99*, 16156–16161.
14. Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods *25*, 402–408.
15. Higasa, K., Kukita, Y., Kato, K., Wake, N., Tahira, T., and Hayashi, K. (2009). Evaluation of haplotype inference using definitive haplotype data obtained from complete

hydatidiform moles, and its significance for the analyses of positively selected regions. PLoS Genet. *5*, e1000468.
16. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R., and Scherer, S.W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res. *115*, 205–214.
17. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. *11*, 1005–1017.
18. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. *77*, 78–88.
19. Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E., and Nickerson, D.A. (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat. Genet. *40*, 1199–1203.
20. R Development Core Team. (2008). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
21. International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
22. Murdoch, S., Djuric, U., Mazhar, B., Seoud, M., Khan, R., Kuick, R., Bagga, R., Kircheisen, R., Ao, A., Ratti, B., et al. (2006). Mutations in NALP7 cause recurrent hydatidiform moles and reproductive wastage in humans. Nat. Genet. *38*, 300–302.
23. Slim, R., and Mehio, A. (2007). The genetics of hydatidiform moles: new lights on an ancient disease. Clin. Genet. *71*, 25–34.
24. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. *40*, 1253–1260.
25. Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N.A., et al. (2008). The fine-scale and complex architecture of human copy-number variation. Am. J. Hum. Genet. *82*, 685–695.
26. Pique-Regi, R., Ortega, A., and Asgharzadeh, S. (2009). Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. Bioinformatics *25*, 1223–1230.
27. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al; Wellcome Trust Case Control Consortium. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

Genes and Immunity (2012) 13, 543–548

www.nature.com/gene

ORIGINAL ARTICLE

HLA and SNP haplotype mapping in the Japanese population

H Kitajima, M Sonoda and K Yamamoto

The genes that encode the human leukocyte antigen (HLA) class I and II molecules are highly polymorphic and located in the major histocompatibility complex (MHC) region, where there is a high density of immune-related genes. Numerous studies have identified disease susceptibility in this region; however, interpretation of the results is complicated because of the strong linkage disequilibrium (LD) among HLA alleles and single-nucleotide polymorphisms (SNPs). In this study, we evaluated the correlation between the HLA alleles of 6 loci (*HLA-A, C, B, DRB1, DQB1* and *DPB1*) and 6502 SNPs within 8 Mb of the extended MHC region using 92 Japanese subjects to identify SNP single loci or haplotypes that tag HLA alleles. We found a total of 39 HLA alleles that showed strong LD ($r^2 \geqslant 0.8$) with SNPs, including 11 non-synonymous SNPs in non-HLA genes. In addition, we identified several SNP haplotypes in strong LD ($r^2 \geqslant 0.8$) with eight HLA alleles, which do not possess tag SNPs. Our detailed list of tag SNPs and haplotypes could be utilized for a better understanding of the results obtained by association studies in the Japanese population and for the characterization of the differences in LD structures between races.

Genes and Immunity (2012) **13,** 543–548; doi:10.1038/gene.2012.35; published online 23 August 2012

Keywords: HLA; tag SNP; tag SNP haplotype; disease susceptibility; linkage disequilibrium

INTRODUCTION

The human leukocyte antigen (HLA) class I and II molecules have central roles in the differentiation of T cells in the thymus and in immune responses to foreign antigens in the peripheral lymphoid organs. The genes that encode these molecules are highly polymorphic and are located in the major histocompatibility complex (MHC) region on the short arm of chromosome 6, where immune-related genes are found at a high density with strong linkage disequilibria (LD). Numerous studies have demonstrated associations between HLA alleles and disease susceptibility in various populations.[1] In addition, recent genome-wide association studies (GWASs) have identified single-nucleotide polymorphisms (SNPs) associated with multifactorial diseases in the MHC region (GWAS catalog in the Table Browser in the UCSC Genome Bioinformatics database, http://genome.ucsc.edu/). However, the analyses were not always able to determine the primary susceptibility gene because of the strong LD among HLA alleles and SNPs. Therefore, efforts to clarify the details of the LD structure constructed with HLA alleles and SNPs residing in non-HLA genes would be required for better interpretations of the results obtained by the association studies that identify susceptibility loci in the MHC region.

Using 361 multiethnic samples including African (YRI), European (CEU), Chinese (CHB) and Japanese (JPT) samples, de Bakker *et al.*[2] produced a high-resolution HLA and SNP haplotype map in the extended human MHC region (26–34 Mb of chromosome 6). The study elucidated detailed LD structures in this region and provided information about the correlation between HLA alleles and SNPs. However, in this study, the *DPB1* locus, which is also reported to be associated with several immune-related diseases,[3–9] was not genotyped. Therefore, additional studies including *DPB1* would provide more information in this research area.

In this study, we investigated the correlation between HLA alleles of 6 loci (*A, C, B, DRB1, DQB1* and *DPB1*) and 6502 SNPs within the extended MHC region using 92 Japanese samples. The results, in combination with those from de Bakker *et al.*,[2] could be utilized for association studies in the Japanese population and for the characterization of the differences in LD structures between races.

RESULTS

Identification of pairs of HLA alleles and SNPs in absolute LD

The calculation of the pairwise linkage disequilibrium (LD) revealed diversity in the extent of LD between HLA alleles and SNPs, even within the same HLA locus (Figure 1 and Supplementary Figure 1). For example, the SNPs showing moderate-to-strong LD with A*33:03 were more broadly distributed than SNPs with A*24:02 (blue and red dots in Figure 1). The HLA alleles C*14:03, B*44:03, DRB1*13:02 and DQB1*06:04 that compose a common HLA haplotype together with A*33:03 in the Japanese population also showed a broad distribution of LD SNPs (Supplementary Figure 1), supporting the inheritance of the haplotype together with the surrounding SNP alleles. A total of 29 HLA alleles showed absolute LD ($r^2 = 1$) with at least one SNP (Supplementary Table 1). In the data set used in this study, there were seven types of missense SNPs among SNPs showing absolute LD with HLA alleles, which should be noted in any association studies because these SNPs might directly affect the function of the corresponding genes (Table 1). The combinations of HLA alleles and SNPs were the following: C*04:01 and three missense SNPs (rs2233976: Gly>Arg in *C6orf15*, rs130072: Arg>Gln in *CCHCR1* and rs2073724: Pro>Leu in *TCF19*); C*12:02 and one missense SNP (rs2270191: Val>Met in *C6orf15*); C*14:03 and three missense SNPs (rs2255221: Trp>Cys in *HCP5*, rs11538264: Val>Met in *PRRC2A* and rs11758242: Ser>Tyr in *LY6G5B*); B*44:03 and three missense SNPs (rs2255221, rs11538264 and rs11758242: see above); and B*52:01 and one missense SNP

Division of Genome Analysis, Research Center for Genetic Information, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan. Correspondence: Dr K Yamamoto, Division of Genome Analysis, Research Center for Genetic Information, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan.
E-mail: kyama@bioreg.kyushu-u.ac.jp
Received 23 April 2012; revised 17 July 2012; accepted 27 July 2012; published online 23 August 2012

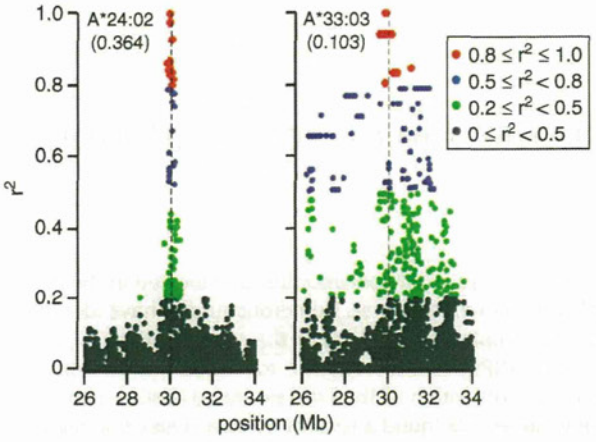

Figure 1. LD between SNPs and HLA-A*24:02 or HLA-A*33:03 in the extended HLA region. The SNPs across the 8 Mb extended HLA region (from 26 to 34 Mb of chromosome 6) showing weak ($0.2 \leqslant r^2 < 0.5$; light green), moderate ($0.5 \leqslant r^2 < 0.8$; blue) and strong ($0.8 \leqslant r^2 \leqslant 1.0$; red) LD were plotted. The allele frequencies of each HLA allele observed in this study are in parentheses. The dashed lines indicate the position of *HLA-A*.

Table 1. List of pairs between HLA allele and non-synonymous SNP in absolute LD ($r^2 = 1$)

HLA allele	*SNP*	*Amino-acid substitution*	*Genes*
C*04:01	rs2233976	Gly > Arg	*C6orf15*
	rs130072	Arg > Gln	*CCHCR1*
	rs2073724	Pro > Leu	*TCF19*
C*12:02	rs2270191	Val > Met	*C6orf15*
C*14:03	rs2255221	Trp > Cys	*HCP5*
	rs11538264	Val > Met	*PRRC2A*
	rs11758242	Ser > Tyr	*LY6G5B*
B*44:03	rs2255221	Trp > Cys	*HCP5*
	rs11538264	Val > Met	*PRRC2A*
	rs11758242	Ser > Tyr	*LY6G5B*
B*52:01	rs2270191	Val > Met	*C6orf15*

Abbreviations: HLA, human leukocyte antigen; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.

(rs2270191: see above). Besides the 29 HLA alleles, there were 10 HLA alleles that showed strong LD with SNPs ($0.80 \leqslant r^2 < 1.0$) (Supplementary Table 2). Among them, A*31:01, C*04:01, C*12:02, B*52:01 and DPB1*05:01 showed LD with missense SNP rs1116221: Glu > Lys in *TRIM31*, rs9263870: Asn > Asp in *HCG27*, rs130075: Arg > Gln in *CCHCR1*, rs130075: see above and rs11551421: Val > Met in *HLA-DPB1*, respectively. We also observed various degrees of pairwise LD between HLA alleles in the Japanese population, which is consistent with previous reports.[10] In this study, four absolute LD pairs of HLA alleles were detected: C*12:02 and B*52:01; C*14:03 and B*44:03; B*13:01 and DRB1*12:02; and DRB1*15:01 and DQB1*06:02 (Supplementary Table 3).

LD between HLA alleles and SNP haplotypes

Among the 53 HLA alleles that did not show absolute LD with any SNPs, 3 HLA alleles exhibited absolute LD with a total of 7 haplotypes (Table 2). Among the three HLA alleles, C*14:02 and B*46:01 do not possess SNPs in strong LD ($0.80 \leqslant r^2 < 1.0$). One missense SNP (rs1116221: Glu > Lys in *TRIM31*) was linked with A*31:01 in the haplotype. Thirteen HLA alleles exhibited strong LD

Table 2. List of pairs of HLA alleles and SNP haplotypes in absolute LD ($r^2 = 1$)

HLA allele	*Haplotype*	*SNPs in haplotype*	*r^2 of each SNP with the HLA allele*
A*31:01	rs2524035(T)-	rs2524035	0.38
	rs2844796(T)-	rs2844796	0.81
	rs2517592(A)	rs2517592	0.25
A*31:01	rs9258883(C)-	rs9258883	0.31
	rs17180570(C)-	rs17180570	0.35
	rs2844792(A)	rs2844792	0.72
A*31:01	rs2524005(T)-	rs2524005	0.35
	rs2517597(A)	rs2517597	0.72
A*31:01	rs12665039(C)-	rs12665039	0.38
	rs2245420(G)-	rs2245420	0.72
	rs12176323(T)	rs12176323	0.81
A*31:01	rs1116221(T)-	rs1116221	0.81
	rs2523979(T)	rs2523979	0.72
C*14:02	rs4947296(C)-	rs4947296	0.38
	rs2442736(G)	rs2442736	0.74
B*46:01	rs3828913(A)-	rs3828913	0.35
	rs9296003(T)-	rs9296003	0.48
	rs9348878(C)	rs9348878	0.21

Abbreviations: HLA, human leukocyte antigen; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.

Table 3. List of HLA alleles showing strong LD ($0.80 \leqslant r^2 < 1.0$) with SNP haplotypes

HLA allele	*SNP*		*Haplotype*	
	Number[a]	*Maximum r^2*	*Number*[b]	*Maximum r^2*
A*02:01	0	—	2	0.94
C*08:01	20	0.87	2	0.86
C*15:02	3	0.85	3	0.92
DRB1*04:06	0	—	1	0.83
DRB1*08:03	1	0.92	2	0.92
DRB1*14:06	0	—	1	0.86
DQB1*03:01	0	—	2	0.86
DPB1*02:01	0	—	6	0.90
DPB1*03:01	0	—	1	0.82
DPB1*09:01	1	0.83	5	0.94

Abbreviations: HLA, human leukocyte antigen; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism. The HLA alleles that possess tag haplotype but not tag SNP are shown in bold face. Note that HLA-C*14:02 and -B*46:01 that possess tag haplotype ($r^2 = 1.0$) but not tag SNP are shown in Table 2. [a]Number of single SNP loci showing strong LD ($0.80 \leqslant r^2 < 1.0$) with the HLA allele. [b]Number of SNP haplotypes showing strong LD ($0.80 \leqslant r^2 < 1.0$) with the HLA allele.

with a total of 45 haplotypes ($0.80 \leqslant r^2 < 1.0$) (Supplementary Table 4). One missense SNP (rs2523989: Val > Ile in *TRIM31*) was linked with A*31:01 in the haplotype, and two missense SNPs (rs2071554: Arg > Gln in *HLA-DOB* and rs2855430: Pro > Leu in *COL11A2*) were linked with DPB1*09:01 in the haplotype. The analyses newly identified tag haplotypes for six HLA alleles (A*02:01, DRB1*04:06, DRB1*14:06, DQB1*03:01, DPB1*02:01 and DPB1*03:01), which possess no SNPs in strong LD ($0.80 \leqslant r^2 < 1.0$) (Table 3). Thus, we identified SNP haplotypes in strong LD with eight HLA alleles, which do not possess tag SNPs. The tag haplotypes for C*15:02 and DPB1*09:01 showed stronger LD than the tag SNPs (Table 3).

Tag SNPs and disease associations

The results of this study reveal that some disease-related HLA alleles possess tag SNPs in the Japanese population (Table 4). It is noteworthy that B*52:01, which denotes susceptibility to ulcerative colitis[15] and aortitis syndrome[16], possesses several tag SNPs located in the *C6orf15*, *PSORS1C1*, *CDSN*, *PSORS1C2* and *CCHCR1* region, suggesting that these non-HLA genes could be considered as candidate genes that are responsible for these diseases.

From the viewpoint of SNPs identified by GWAS in Asian populations, several tag SNPs of HLA alleles have been reported to be associated with multifactorial diseases. For Japanese subjects, the tag SNP rs9263739 of C*12:02 and B*52:01, which is located in the intron region of *CCHCR1*, was reported to be significantly associated with ulcerative colitis.[11] The tag SNP rs11752643 of DQB1*06:04 and DRB1*13:02 was reported to indicate susceptibility to coronary heart disease.[12] The association of the SNP rs987870, which composes the tag haplotype for DPB1*09:01, with pediatric asthma was also reported.[13] For Chinese Han subjects, the SNP rs2281388 showing absolute LD with DPB1*05:01 and the SNP rs4947296 composing the tag haplotype for C*14:02 were reported to be significantly associated with Graves disease[14] (Tables 5 and 6). These results suggest that the HLA alleles should be considered as indicators of susceptibility to these diseases, even in cases of non-immune-related diseases such as coronary artery disease. Thus, information about SNPs that tag HLA alleles broadens the spectrum of candidate genes in GWASs that detect associated SNPs in the HLA regions.

DISCUSSION

This study detected tag SNPs and haplotypes for some disease-related HLA alleles in the Japanese population. Based on this data, additional candidate genes were discovered to provide a better understanding of the pathogenesis of HLA-associated diseases. Although previous studies reported that HLA-B*52:01 is associated with ulcerative colitis[15] and aortitis syndrome[16] in the Japanese population, this study detected strong LD between this HLA allele and the SNPs that are located in the major psoriasis-susceptibility locus (*PSORS1C1*, *CDSN* and *CCHCR1*). The Japanese GWAS reported that rs9263739 in *CCHCR1*, which was found to be in strong LD ($r^2 = 0.94$) with HLA-B*52:01 (Supplementary Table 2), was significantly associated with ulcerative colitis.[11] Elomaa *et al.*[17] reported that *CCHCR1* transgenic mice appeared phenotypically normal and that their skin was histologically indistinguishable from wild-type mice. However, the expression of genes involved in the pathogenesis of psoriasis was changed in the mice. In addition, an epidemiological study reported that psoriasis was associated with ulcerative colitis.[18] Considering that the inflammation of the mucosa and occasionally the submucosa of the colon causes ulcerative colitis, it was necessary to investigate not only HLA-B*52:01 but also the genes in the psoriasis-susceptibility locus to elucidate the pathogenesis of ulcerative colitis.

It was also important to consider the influence of HLA alleles when investigating the genes located in the extended MHC region. This study revealed that the SNP rs11538264: Val > Met,

Table 4. List of HLA alleles, HLA-associated diseases and the tag SNPs in the Japanese population

HLA risk allele (associated disease, reference number)	*Tag SNP*	r^2	*Function of SNP*	*Gene*
HLA-A*02:06	rs2517830	1.0	Unknown	—
(Graves disease[28])	rs7760545	0.95	Unknown	—
	rs6457109	0.85	Unknown	—
HLA-B*51:01 (Behçet's disease[29])	rs2442736	0.91	Unknown	—
HLA-B*52:01	rs2270191	1.0	Missense	*C6orf15*
(Ulcerative colitis[15])	rs3132550	0.94	Intron	*CDSN, PSORS1C1*
(Aortitis syndrome[16])	rs4410768	1.0	Intron	*PSORS1C1*
	rs7757012	0.94	Near gene 5	*PSORS1C2*
	rs12364	0.94	Coding synon	*CCHCR1*
	rs9263739	0.94	Intron	*CCHCR1*
	rs9263749	0.94	Intron	*CCHCR1*
	rs130075	0.94	Missense	*CCHCR1*
	rs35718543	0.94	Unknown	—
	rs28360997	0.94	Unknown	—
	rs28367729	0.84	Unknown	—
	rs2246010	0.84	Unknown	—
	rs2844586	0.84	Unknown	—
	rs28399987	0.83	Near gene 3, near gene 5	*MSH5, SAPCD1*
HLA-DRB1*09:01 (Myasthenia gravis[30])	rs16870207	0.88	Near gene 5	*HLA-DRB5*
HLA-DPB1*05:01	rs9378177	1.0	Intron	*HLA-DPB1*
(Multiple sclerosis opticospinal form[7])	rs11551421	0.96	Missense	*HLA-DPB1*
(Graves disease[3])	rs2068204	0.98	Unknown	—
(Japanese cedar pollinosis[6])	rs10484569	0.98	Unknown	—
	rs2281388	1.0	Unknown	—
	rs9296081	0.98	Unknown	—
	rs6457713	0.98	Unknown	—
	rs9380342	1.0	Unknown	—
	rs9380343	0.98	Unknown	—
	rs12174662	1.0	Unknown	—
	rs6937034	1.0	Unknown	—
	rs9348906	0.98	Untranslated 5	*HLA-DPB2*
	rs9366814	0.91	Intron	*HLA-DPB2*
	rs2235499	0.82	Unknown	—

Abbreviations: HLA, human leukocyte antigen; SNP, single-nucleotide polymorphism. HLA alleles that have been reported to be associated with diseases but do not possess SNPs in strong LD are not shown.

Table 5. List of HLA alleles and their tag SNPs that are described in the GWAS catalog[a]

SNP	*HLA allele*	r^2	*Function of SNP*	*Genes*	*Reported trait of SNP*	*Analyzed ethnicity*	*Reference*
rs2860580	HLA-A*11:01	0.87	Unknown	—	Nasopharyngeal carcinoma	Southern Chinese descent	31
rs9263739	HLA-C*12:02, HLA-B*52:01	0.94	Intron	*CCHCR1*	Ulcerative colitis	Japanese	11
rs1265112	HLA-C*04:01	0.83	Intron	*CCHCR1*	Nevirapine-induced rash	HIV-infected Thai	32
rs4418214	HLA-B*13:01, HLA-DRB1*12:02	1.0	Unknown	—	HIV-1 control	Caucasian	33
rs2255221	HLA-C*14:03, HLA-B*44:03	1.0	Missense	*HCP5*	HIV-1 control	African-American	33
rs10484561	HLA-DQB1*05:01	1.0	Unknown	—	Follicular lymphoma	Caucasian	34
rs10484561	HLA-DRB1*01:01	0.92	Unknown	—	Follicular lymphoma	Caucasian	34
rs11752643	HLA-DRB1*13:02	0.87	Unknown	—	Coronary heart disease	Japanese	12
rs11752643	HLA-DQB1*06:04	0.93	Unknown	—	Coronary heart disease	Japanese	12
rs2281388	HLA-DPB1*05:01	1.0	Unknown	—	Graves disease	Chinese Han	14

Abbreviations: GWAS, genome-wide association study; HIV, human immunodeficiency virus; HLA, human leukocyte antigen; SNP, single-nucleotide polymorphism. [a]The GWAS catalog is based on the Table Browser of the UCSC Genome Bioinformatics database (GRCh37/hg19).

Table 6. List of HLA alleles and their tag haplotypes that include SNPs described in the GWAS catalog[a]

Haplotype	*HLA allele*	r^2 *between haplotype and HLA allele*	*SNPs in the haplotype*	r^2 *between SNP and HLA allele*	*Function of SNP*	*Genes*	*Reported trait of SNP*	*Analyzed ethnicity*	*Reference*
rs4947296(C)-rs2442736(G)	HLA-C*14:02	1.0	rs4947296	0.38	Unknown	—	Graves disease	Chinese Han	14
			rs2442736	0.74	Unknown	—	—	—	—
rs3815087(T)-rs9264885(T)	HLA-C*08:01	0.86	rs3815087	0.38	Untranslated	*PSORS1C1*	Stevens-Johnson syndrome and toxic epidermal necrolysis	Caucasian	35
							HIV-1 control	Caucasian	36
			rs9264885	0.79	Unknown	—	—	—	—
rs2395148(A)-rs3763313(C)-rs9268979(T)	HLA-DRB1*08:03	0.92	rs2395148	0.67	Intron	*C6orf10*	Primary biliary cirrhosis	Caucasian	37
							Juvenile idiopathic arthritis	Caucasian	38
			rs3763313	0.25	Near gene 5	*BTNL2*	HIV-1 control	Caucasian	36
			rs9268979	0.22	Unknown	—	—	—	—
rs3129888(C)-rs2187668(T)	HLA-DRB1*14:06	0.86	rs3129888	0.48	Intron	*HLA-DRA*	—	—	—
			rs2187668	0.48	Intron	*HLA-DQA1*	Anti-dsDNA-positive systemic lupus erythematosus	Caucasian	39
							Idiopathic membranous nephropathy	Caucasian	40
							Immunoglobulin A deficiency	Caucasian	41
							Celiac disease	Caucasian	42
							Celiac disease	Caucasian	43
							Systemic lupus erythematosus	Caucasian	44
rs9272219(T)-rs4538747(T)	HLA-DQB1*03:01	0.85	rs9272219	0.70	Unknown	—	Rheumatoid arthritis	Caucasian	45
							Schizophrenia	Caucasian, African-American	46
			rs4538747	0.34	Unknown	—	—	—	—
rs987870(C)-rs2855430(A)	HLA-DPB1*09:01	0.82	rs987870	0.69	Intron	*HLA-DPA1*	Pediatric asthma	Japanese	13
							Systemic sclerosis	Caucasian	47
			rs2855430	0.63	Missense	*COL11A2*	—	—	—

Abbreviations: dsDNA, double-stranded DNA; GWAS, genome-wide association study; HIV, human immunodeficiency virus; HLA, human leukocyte antigen; SNP, single-nucleotide polymorphism. [a]The GWAS catalog is based on the Table Browser of the UCSC Genome Bioinformatics database (GRCh37/hg19).

which is located in *PRRC2A*, was in absolute LD with HLA-C*14:03 and HLA-B*44:03. A microsatellite analysis reported an association between the age-at-onset of insulin-dependent diabetes mellitus and *PRRC2A* in the Japanese population.[19] Although *PRRC2A* is located in the same region as the genes of tumor necrosis factor-α and tumor necrosis factor-β,[20,21] the effects of HLA-C*14:03 and

HLA-B*44:03 should be considered when surveying the causal factors of this disease in future studies.

The GWAS of Japanese pediatric asthma reported a significant association of rs987870, located in the intron of the *HLA-DPA1* locus, with this disease.[13] The significant association signals were distributed between the *HLA-DPB1*, *HLA-DPB2*, *COL11A2* and *RXRB* loci. The authors reported that HLA-DPA1*02:01, which was in strong LD with rs987870 and the HLA-DPB1*09:01 allele, was also significantly associated with the disease. Considering our data showing that rs987870 and the missense SNP (rs2855430) in the *COL11A2* region compose a tag haplotype for HLA-DPB1*09:01 (Table 6), it would be possible to highlight not only the effect of HLA-DPA1*02:01 and HLA-DPB1*09:01 but also that of *COL11A2*.

By comparison with a previous report,[2] the results of this study provide new information about the correlation between HLA alleles and SNPs in the Japanese population. Although the distributions of SNPs that show LD with HLA-A*33:03, A*31:01, C*01:02, C*14:03, B*52:01, DRB1*15:02 or DQB1*06:01, which are the relatively common HLA alleles in the Japanese population, were similar to the previous report, we also detected SNPs in strong LD with these HLA alleles that had not been previously described (Supplementary Figure 1). With respect to HLA-DRB1*13:02, the distribution of SNPs in LD and each LD status are somewhat different from the previous report. Our study showed that SNPs in strong LD with HLA-DRB1*13:02 are distributed in the region from 32 to −33 Mb and that SNPs in LD ($0.2 \leqslant r^2 < 0.8$) are distributed across the entire extended MHC region, whereas the previous study showed that the SNPs in weak or moderate LD with this HLA allele are only distributed in two regions from 26 to 29 Mb and from 31 to 33 Mb. As we obtained similar results in the analysis of HLA-DQB1*06:04, which is in strong LD with DRB1*13:02 (Supplementary Figure 1), our data about the correlation between the SNPs and HLA-DRB1*13:02 and DQB1*06:04 could be applied to Japanese subjects.

Differences in the tag pattern between JPT, CEU and YRI were also indicated in the previous report.[2] For example, HLA-C*07:02 possessed many SNPs in moderate-to-strong LD in CEU and YRI across several Mb, whereas in JPT, the SNPs in strong LD were distributed in a narrow region near *HLA-C* locus.[2] On the other hand, tag SNPs for HLA-C*03:04 were not found in any population samples.[2] Consistent with the results, the SNPs in LD with HLA-C*07:02 are distributed within several kb of *HLA-C*, and HLA-C*03:04 does not possess tag SNP in our Japanese samples (Supplementary Figure 1). With respect to *HLA-DRB1*, we found that the SNPs in LD with DRB1*15:02 are distributed across 5.5 Mb in HLA region (Supplementary Figure 1), whereas in CEU, the SNPs in LD were distributed in a narrow region of this locus.[2] Thus, the tag pattern is likely to differ between populations even in a same HLA allele. This suggests that different SNPs could be associated with the same disease caused by a particular HLA allele in an analysis using different populations.

We utilized 92 Japanese samples and common SNPs; therefore, we could not analyze low-frequency HLA alleles. Recent advances in genomic analyses lead us to reaffirm the significance of the MHC region as a genetic factor in diseases. Genetic dissection of this region using not only common but also rare variants may be required for a comprehensive understanding of its roles in human disease susceptibility.

In conclusion, we identified tag SNPs and haplotypes for several HLA alleles of six HLA loci in the Japanese population. Our data will confer useful information for etiological studies of east Asian populations, specifically the Japanese population, focusing on both HLA and non-HLA genes in the MHC region.

MATERIALS AND METHODS

Subjects

Epstein–Barr virus-transformed B-cell lines derived from 92 healthy Japanese subjects were provided by the Riken Bioresource Center Cell Bank.[22] HLA typing was performed for the six loci (*HLA-A, C, B, DRB1, DQB1* and *DPB1*) by the Luminex microbead method (Luminex, Austin, TX, USA).

SNP genotyping

SNP genotyping was carried out using the Illumina Human1M BeadChip (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. All subjects showed a genotyping call rate >0.99.

Selection of SNPs and HLA alleles

SNPs that showed minor allele frequencies <0.01, call rates <0.97 (missed >3 samples) or Hardy–Weinberg equilibrium test *P*-values <0.001 were excluded from the analyses. HLA alleles with frequencies <0.01 were also excluded. A total of 82 HLA alleles (11, 11, 22, 18, 11 and 9 alleles in *HLA-A, C, B, DRB1, DQB1* and *DPB1*, respectively) and 6502 SNPs within the extended HLA region (the region from 26 to 34 Mb of chromosome 6) were subjected to the analyses. PLINK version 1.07[23] and R version 2.14.0[24] were used for the selection of SNPs and HLA alleles.

Tag SNP and tag haplotype analysis

Initially, we calculated the pairwise LD between the HLA alleles and SNPs using Haploview.[25,26] Then, we assessed tag haplotypes for the HLA alleles that did not have SNPs in absolute LD. In this analysis, we chose SNPs in LD ($r^2 > 0.20$) with the HLA alleles and then carried out an aggressive search for tag haplotypes consisting of two or three SNPs by the 'Tagger' algorithm implemented in Haploview. The base position and function of the SNPs were based on the database of NCBI36/hg18. A total of 35 HLA alleles (4, 2, 13, 10, 4 and 2 alleles in *HLA-A, C, B, DRB1, DQB1* and *DPB1*, respectively) among 85 alleles do not possess tag SNPs or tag haplotypes ($r^2 \geqslant 0.8$). The SNPs reported to be associated with diseases were extracted from the GWAS catalog on the UCSC Genome Bioinformatics Browser (http://genome.ucsc.edu/).[27]

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by KAKENHI (22133003) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

REFERENCES

1 Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet* 2009; **54**: 15–39.
2 de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* 2006; **38**: 1166–1172.
3 Dong RP, Kimura A, Okubo R, Shinagawa H, Tamai H, Nishimura Y *et al.* HLA-A and DPB1 loci confer susceptibility to Graves' disease. *Hum Immunol* 1992; **35**: 165–172.
4 Richeldi L, Sorrentino R, Saltini C. HLA-DPB1 glutamate 69: a genetic marker of beryllium disease. *Science* 1993; **262**: 242–244.
5 Horiki T, Inoko H, Moriuchi J, Ichikawa Y, Arimori S. Combinations of HLA-DPB1 and HLA-DQB1 alleles determine susceptibility to early-onset myasthenia gravis in Japan. *Autoimmunity* 1994; **19**: 49–54.
6 Hori T, Kamikawaji N, Kimura A, Sone T, Komiyama N, Komiyama S *et al.* Japanese cedar pollinosis and HLA-DP5. *Tissue Antigens* 1996; **47**: 485–491.
7 Ito H, Yamasaki K, Kawano Y, Horiuchi I, Yun C, Nishimura Y *et al.* HLA-DP-associated susceptibility to the optico-spinal form of multiple sclerosis in the Japanese. *Tissue Antigens* 1998; **52**: 179–182.
8 Lv N, Dang A, Wang Z, Zheng D, Liu G. Association of susceptibility to Takayasu arteritis in Chinese Han patients with HLA-DPB1. *Hum Immunol* 2011; **72**: 893–896.
9 Ivansson EL, Juko-Pecirep I, Erlich HA, Gyllensten UB. Pathway-based analysis of genetic susceptibility to cervical cancer in situ: HLA-DPB1 affects risk in Swedish women. *Genes Immun* 2011; **12**: 605–614.
10 Saito S, Ota S, Yamada E, Inoko H, Ota M. Allele frequencies and haplotypic associations defined by allelic DNA typing at HLA class I and class II loci in the Japanese population. *Tissue Antigens* 2000; **56**: 522–529.
11 Asano K, Matsushita T, Umeno J, Hosono N, Takahashi A, Kawaguchi T *et al.* A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat Genet* 2009; **41**: 1325–1329.

12 Takeuchi F, Yokota M, Yamamoto K, Nakashima E, Katsuya T, Asano H *et al.* Genome-wide association study of coronary artery disease in the Japanese. *Eur J Hum Genet* 2012; **20**: 333–340.
13 Noguchi E, Sakamoto H, Hirota T, Ochiai K, Imoto Y, Sakashita M *et al.* Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations. *PLoS Genet* 2011; **7**: e1002170.
14 Chu X, Pan CM, Zhao SX, Liang J, Gao GQ, Zhang XM *et al.* A genome-wide association study identifies two new risk loci for Graves' disease. *Nat Genet* 2011; **43**: 897–901.
15 Aizawa H, Kinouchi Y, Negoro K, Nomura E, Imai G, Takahashi S *et al.* HLA-B is the best candidate of susceptibility genes in HLA for Japanese ulcerative colitis. *Tissue Antigens* 2009; **73**: 569–574.
16 Yoshida M, Kimura A, Katsuragi K, Numano F, Sasazuki T. DNA typing of HLA-B gene in Takayasu's arteritis. *Tissue Antigens* 1993; **42**: 87–90.
17 Elomaa O, Majuri I, Suomela S, Asumalahti K, Jiao H, Mirzaei Z *et al.* Transgenic mouse models support HCR as an effector gene in the PSORS1 locus. *Hum Mol Genet* 2004; **13**: 1551–1561.
18 Yates VM, Watkinson G, Kelman A. Further evidence for an association between psoriasis, Crohn's disease and ulcerative colitis. *Br J Dermatol* 1982; **106**: 323–330.
19 Hashimoto M, Nakamura N, Obayashi H, Kimura F, Moriwaki A, Hasegawa G *et al.* Genetic contribution of the BAT2 gene microsatellite polymorphism to the age-at-onset of insulin-dependent diabetes mellitus. *Hum Genet* 1999; **105**: 197–199.
20 Spies T, Blanck G, Bresnahan M, Sands J, Strominger JL. A new cluster of genes within the human major histocompatibility complex. *Science* 1989; **243**: 214–217.
21 Iris FJ, Bougueleret L, Prieur S, Caterina D, Primas G, Perrot V *et al.* Dense Alu clustering and a potential new member of the NF kappa B family within a 90 kilobase HLA class III segment. *Nat Genet* 1993; **3**: 137–145.
22 Iwakawa M, Goto M, Noda S, Sagara M, Yamada S, Yamamoto N *et al.* DNA repair capacity measured by high throughput alkaline comet assays in EBV-transformed cell lines and peripheral blood cells from cancer patients and healthy volunteers. *Mutat Res* 2005; **588**: 1–6.
23 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
24 R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011 (http://www.R-project.org).
25 de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
26 Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
27 Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004; **32**(Database issue): D493–D496.
28 Takahashi M, Yasunami M, Kubota S, Tamai H, Kimura A. HLA-DPB1*0202 is associated with a predictor of good prognosis of Graves' disease in the Japanese. *Hum Immunol* 2006; **67**: 47–52.
29 Meguro A, Inoko H, Ota M, Katsuyama Y, Oka A, Okada E *et al.* Genetics of Behcet disease inside and outside the MHC. *Ann Rheum Dis* 2010; **69**: 747–754.
30 Matsuki K, Juji T, Tokunaga K, Takamizawa M, Maeda H, Soda M *et al.* HLA antigens in Japanese patients with myasthenia gravis. *J Clin Invest* 1990; **86**: 392–399.
31 Bei JX, Li Y, Jia WH, Feng BJ, Zhou G, Chen LZ *et al.* A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat Genet* 2010; **42**: 599–603.
32 Chantarangsu S, Mushiroda T, Mahasirimongkol S, Kiertiburanakul S, Sungkanuparph S, Manosuthi W *et al.* Genome-wide association study identifies variations in 6p21.3 associated with nevirapine-induced rash. *Clin Infect Dis* 2011; **53**: 341–348.
33 International HIV Controllers Study; Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PI *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 2010; **330**: 1551–1557.
34 Conde L, Halperin E, Akers NK, Brown KM, Smedby KE, Rothman N *et al.* Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat Genet* 2010; **42**: 661–664.
35 Genin E, Schumacher M, Roujeau JC, Naldi L, Liss Y, Kazma R *et al.* Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe. *Orphanet J Rare Dis* 2011; **6**: 52.
36 Fellay J, Ge D, Shianna KV, Colombo S, Ledergerber B, Cirulli ET *et al.* Common genetic variation and the control of HIV-1 in humans. *PLoS Genet* 2009; **5**: e1000791.
37 Hirschfield GM, Liu X, Xu C, Lu Y, Xie G, Lu Y *et al.* Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. *N Engl J Med* 2009; **360**: 2544–2555.
38 Behrens EM, Finkel TH, Bradfield JP, Kim CE, Linton L, Casalunovo T *et al.* Association of the TRAF1-C5 locus on chromosome 9 with juvenile idiopathic arthritis. *Arthritis Rheum* 2008; **58**: 2206–2207.
39 Chung SA, Taylor KE, Graham RR, Nititham J, Lee AT, Ortmann WA *et al.* Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genet* 2011; **7**: e1001323.
40 Stanescu HC, Arcos-Burgos M, Medlar A, Bockenhauer D, Kottgen A, Dragomirescu L *et al.* Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N Engl J Med* 2011; **364**: 616–626.
41 Ferreira RC, Pan-Hammarstrom Q, Graham RR, Gateva V, Fontan G, Lee AT *et al.* Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nat Genet* 2010; **42**: 777–780.
42 Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010; **42**: 295–302.
43 van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007; **39**: 827–829.
44 Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, Garnier S *et al.* Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N Engl J Med* 2008; **358**: 900–909.
45 Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJ. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Hum Mol Genet* 2011; **20**: 3494–3506.
46 Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009; **460**: 753–757.
47 Gorlova O, Martin JE, Rueda B, Koeleman BP, Ying J, Teruel M *et al.* Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. *PLoS Genet* 2011; **7**: e1002178.

Supplementary Information accompanies the paper on Genes and Immunity website (http://www.nature.com/gene)

The EMBO Journal (2012) 31, 1904–1915 | © 2012 European Molecular Biology Organization | All Rights Reserved 0261-4189/12
www.embojournal.org

THE EMBO JOURNAL

NLK positively regulates Wnt/β-catenin signalling by phosphorylating LEF1 in neural progenitor cells

Satoshi Ota[1], Shizuka Ishitani[1], Nobuyuki Shimizu[1], Kunihiro Matsumoto[2], Motoyuki Itoh[3,4] and Tohru Ishitani[1,3,*]

[1]Division of Cell Regulation Systems, Department of Immunobiology and Neuroscience, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan, [2]Group of Signal Transduction, Laboratory of Cell Regulation, Division of Biological Science, Graduate School of Science, Nagoya University, Nagoya, Japan, [3]Unit on Nervous Development Systems, Division of Biological Science, Graduate School of Science, Nagoya University, Nagoya, Japan and [4]Institute for Advanced Research, Nagoya University, Nagoya, Japan

Nemo-like kinase (NLK/Nlk) is an evolutionarily conserved protein kinase involved in Wnt/β-catenin signalling. However, the roles of NLK in Wnt/β-catenin signalling in vertebrates remain unclear. Here, we show that inhibition of Nlk2 function in zebrafish results in decreased Lymphoid enhancer factor-1 (Lef1)-mediated gene expression and cell proliferation in the presumptive midbrain, resulting in a reduction of midbrain tectum size. These defects are related to phosphorylation of Lef1 by Nlk2. Thus, Nlk2 is essential for the phosphorylation and activation of Lef1 transcriptional activity in neural progenitor cells (NPCs). In NPC-like mammalian cells, NLK is also required for the phosphorylation and activation of LEF1 transcriptional activity. Phosphorylation of LEF1 induces its dissociation from histone deacetylase, thereby allowing transcription activation. Furthermore, we demonstrate that NLK functions downstream of Dishevelled (Dvl) in the Wnt/β-catenin signalling pathway. Our findings reveal a novel role of NLK in the activation of the Wnt/β-catenin signalling pathway.

The EMBO Journal (2012) **31**, 1904–1915. doi:10.1038/emboj.2012.46; Published online 28 February 2012
Subject Categories: signal transduction; neuroscience
Keywords: lymphoid enhancer factor-1; nemo-like kinase; Wnt/β-catenin signalling; zebrafish

Introduction

The T-cell factor/lymphoid enhancer factor (TCF/LEF) family of transcription factors regulate Wnt/β-catenin signalling, which controls cell proliferation and fate decision during embryogenesis and adult tissue homeostasis (Logan and Nusse, 2004; Arce *et al*, 2006; Clevers, 2006; Hoppler and Kavanagh, 2007). TCF/LEF transcriptional activity is switched in a manner dependent on Wnt/β-catenin signalling (Logan and Nusse, 2004; Arce *et al*, 2006; Clevers, 2006; Hoppler and Kavanagh, 2007). In unstimulated cells, the levels of cytoplasmic β-catenin, a co-activator of TCF/LEF, are kept low by a degradation complex that includes Axin and glycogen synthase kinase 3β (GSK-3β). This kinase catalyses the phosphorylation of β-catenin, which promotes its ubiquitination and subsequent proteasomal degradation (Logan and Nusse, 2004; Clevers, 2006). In the absence of stimulation, TCF/LEF represses the expression of Wnt/β-catenin signalling-target genes by interacting with transcriptional co-repressors such as histone deacetylase 1 (HDAC1) and Groucho (Cavallo *et al*, 1998; Roose *et al*, 1998; Billin *et al*, 2000; Arce *et al*, 2009). The Wnt/β-catenin signalling pathway is induced when the secreted glycoprotein Wnt binds to the cell surface Frizzled (Fz) receptor and its co-receptor LRP5/6. This Wnt-bound receptor complex recruits the cytoplasmic protein Dishevelled (Dvl), which in turn brings the Axin–GSK-3β complex to the membrane and induces the phosphorylation of LRP6. Phosphorylated LRP6 promotes the dissociation of the β-catenin degradation complex (Davidson *et al*, 2005; Zeng *et al*, 2005, 2008). This series of events result in the accumulation of cytoplasmic β-catenin (Niehrs and Shen, 2010; MacDonald *et al*, 2011). The increased β-catenin concentration drives its migration into the nucleus where it forms complexes with TCF/LEF proteins, which then activate gene expression. However, the mechanism(s) by which TCF/LEF is converted from a repressor to an activator is poorly understood.

Nemo-like kinase (NLK) is an evolutionarily conserved MAP kinase-like kinase that regulates diverse signalling processes via phosphorylation of several transcription factors (Ishitani *et al*, 1999, 2010; Kanei-Ishii *et al*, 2004; Ohkawara *et al*, 2004; Kojima *et al*, 2005; Zeng *et al*, 2007). In *Caenorhabditis elegans*, the NLK homologue LIT-1 regulates POP-1, the *C. elegans* homologue of TCF/LEF (Meneghini *et al*, 1999; Rocheleau *et al*, 1999; Herman, 2001; Siegfried and Kimble, 2002; Siegfried *et al*, 2004). POP-1 represses the expression of genes required for endoderm induction. LIT-1 co-operates with the *C. elegans* β-catenin homologue WRM-1 to promote the phosphorylation and consequent nuclear export of POP-1, resulting in the transcriptional activation of POP-1-repressed genes (Meneghini *et al*, 1999; Rocheleau *et al*, 1999). LIT-1 also functions as a positive regulator of POP-1 in the fate specification of gonadal precursor cells (Herman, 2001; Siegfried and Kimble, 2002; Siegfried *et al*, 2004). However, the mechanism underlying this positive regulation is unclear. The regulation of POP-1 activity by LIT-1 is cell context dependent. The negative regulation of TCF/LEF by NLK has been also observed in human embryonic kidney 293 (HEK293) cells and the cervical epithelioid carcinoma cell line HeLa (Ishitani *et al*, 1999, 2003b). In these cell lines, overexpression of NLK inhibits β-catenin–TCF/LEF complex-mediated transcription via phosphorylation of TCF/LEF. On the other hand, positive regulation of TCF/LEF by NLK has not yet been observed in vertebrates.

*Corresponding author. Division of Cell Regulation Systems, Department of Immunobiology and Neuroscience, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, Fukuoka 812-8502, Japan. Tel.: +81 92 642 6789; Fax: +81 92 642 6790; E-mail: tish@bioreg.kyushu-u.ac.jp

Received: 26 August 2011; accepted: 30 January 2012; published online: 28 February 2012

Mouse NLK is expressed in neural tissues, suggesting that mammalian NLK might play a role in nervous system development. Indeed, mice lacking NLK display various neurological abnormalities (Kortenjann *et al*, 2001). In the present studies, we demonstrate that NLK positively regulates the transcriptional activity of LEF1, a member of the TCF/LEF family, in zebrafish midbrain and mammalian neural progenitor cell (NPC)-like cell lines. We further show that Dvl activates NLK in the Wnt pathway and that NLK promotes the release of HDAC1 from LEF1 by phosphorylating LEF1. Our findings provide evidence that NLK mediates Wnt/β-catenin signalling, and consequently NPC proliferation through LEF1 phosphorylation.

Results

Nlk2 is essential for Wnt/β-catenin signalling in zebrafish midbrain

We used zebrafish as a model animal to investigate the roles of NLK in Wnt/β-catenin signalling *in vivo*. Zebrafish has two *nlk* genes, *nlk1* and *nlk2* (Supplementary Figure S1A and B). Nlk1 protein is more related to *Xenopus laevis* NLK1 (73% identical) than to human NLK (68% identical), while Nlk2 protein is most similar to human NLK (97% identical). Nlk2 and human NLK, but not Nlk1, contain histidine-rich (His-rich) and carboxyl terminal conserved regions (Supplementary Figure S1B). Vertebrate NLK proteins can be classified into two groups by phylogenetic analysis: type-I NLK, which includes *X. laevis* NLK1 and Nlk1, and type-II NLK, which includes mammalian NLK and Nlk2 (Supplementary Figure S1A). Recent studies show that Nlk1 regulates primary neurogenesis, ventrolateral mesoderm formation and brain anterior-posterior patterning in early embryogensis (Thorpe and Moon, 2004; Ishitani *et al*, 2010). We therefore investigated the physiological roles of Nlk2.

Expression of *nlk2* was observed in head tissues from the late somite stage (Figure 1A), suggesting that *nlk2* is involved in later brain development. To monitor activity of the zebrafish LEF1 homologue, Lef1, we used a transgenic zebrafish line carrying a Wnt/β-catenin signalling reporter construct (TOPdGFP), in which destabilized green fluorescent protein (dGFP) is driven by a promoter containing multiple TCF/LEF-binding sites, and thus indicates tissues where Lef1 is transcriptionally active (Dorsky *et al*, 2002). This transgenic zebrafish exhibited dGFP expression in the developing midbrain from the late somite stage (Figure 1B; Supplementary Figure S2A). As shown in Figure 1C, dGFP expression was observed in the dorsal and lateral marginal regions of the midbrain at 24 and 27 h.p.f. At 30 h.p.f., dGFP was expressed in the entire dorsal midbrain, while expression in lateral dorsal region was decreased. Consistent with this reporter activity, *lef1* and *nlk2* mRNAs were detected by *in-situ* hybridization in the entire dorsal midbrain of zebrafish embryos at 24 h.p.f. (Figure 1D). At 27 h.p.f., their expression levels in the lateral dorsal part of the midbrain were relatively low (Figure 1D) and their expression patterns were similar to that of TOPdGFP at 30 h.p.f. (Figure 1C). dGFP expression was attenuated by knockdown of zygotic Lef1 using a morpholino oligo (MO) that blocks *lef1* splicing (*lef1* spl MO) (Figure 1B; Supplementary Figure S2A and B; Supplementary Table SI), confirming that Lef1 indeed mediates the transactivation induced by Wnt/β-catenin signalling in the developing

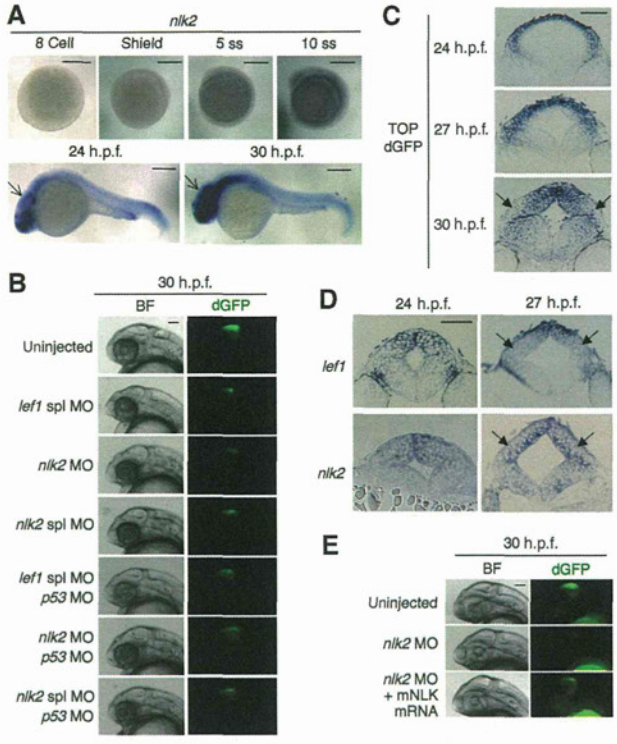

Figure 1 Nlk2 and Lef1 are required for the activation of Wnt/β-catenin signalling in zebrafish developing midbrain. (**A**) Whole mount *in-situ* hybridization staining for *nlk2* in zebrafish embryos at the indicated stage. Scale bar: 250 μm. The expression of *nlk2* in midbrain is indicated with arrows. (**B**, **E**) TOPdGFP-transgenic zebrafish embryos injected with *lef1* spl MO, *nlk2* MO, *nlk2* spl MO, or *p53* MO with or without mouse NLK (mNLK) mRNA, as indicated. Panels show the left side head views of 30 h.p.f. embryos with the anterior to the left. Cells expressing dGFP were visualized by fluorescence microscopy (right panels). Bright-field (BF) images are shown in left panels. Scale bar, 50 μm. Note that mNLK partially rescued *nlk2* MO-induced reduction of TOPdGFP activity ($n = 28$, 50%). (**C**, **D**) *In-situ* hybridization staining for TOPdGFP (**C**), *lef1* (**D**), and *nlk2* (**D**) in the transverse section at the level of midbrain in the indicated stage zebrafish embryos. Scale bar: 50 μm. The lateral dorsal region is indicated with arrows.

midbrain. We next examined the effect of *nlk2* inactivation on TOPdGFP reporter expression in the developing midbrain of zebrafish embryos using a translation-blocking MO against *nlk2* (*nlk2* MO) and an *nlk2* splice-blocking MO (*nlk2* spl MO) (Supplementary Figure S3). We found that, similarly to *lef1* spl MO-injected TOPdGFP fish embryos, the embryos injected with *nlk2* MO or *nlk2* spl MO showed lower TOPdGFP activity than the uninjected embryos at 24 h.p.f. (Supplementary Figure S2A and B; Supplementary Table SI) and 30 h.p.f. (Figure 1B; Supplementary Table SI). Co-injection of a validated MO for *p53* (Robu *et al*, 2007; Tsukada *et al*, 2010; Gerety and Wilkinson, 2011) together with an MO for *nlk2* or *lef1* had no effect on the phenotype induced by MO-mediated knockdown of *nlk2* or *lef1* (Figure 1B; Supplementary Table SI), eliminating the possibility that this phenotype was due to artificial MO-induced p53 activation (Robu *et al*, 2007). Furthermore, the *nlk2* MO-induced reduction of TOPdGFP activity in the midbrain was partially rescued by co-injection with mouse NLK mRNA (Figure 1E). We confirmed that injection of *lef1* spl MO or *nlk2* MO reduced dGFP expression levels in the midbrain but had no effect on the midbrain formation at 27 h.p.f. (Supplementary

Figure S2C). By observing the expression of brain maker genes, we also confirmed that neither injection of *nlk2* MO nor *lef1* spl MO affected the patterning of 24 h.p.f. zebrafish midbrain (Supplementary Figure S4A and B). These results suggest that Nlk2 positively regulates Wnt/β-catenin signalling in the developing zebrafish midbrain.

Nlk2 contribute to midbrain tectum development in zebrafish

To investigate the physiological roles of Nlk2 in the midbrain, we injected MOs for *nlk2* into a transgenic zebrafish line carrying the HuC:Kaede reporter, which expresses the fluorescent protein Kaede in neurons under the control of the neuron-specific *HuC/elavl3* promoter (Sato *et al*, 2006). Injection of either *nlk2* MO or *nlk2* spl MO reduced the size

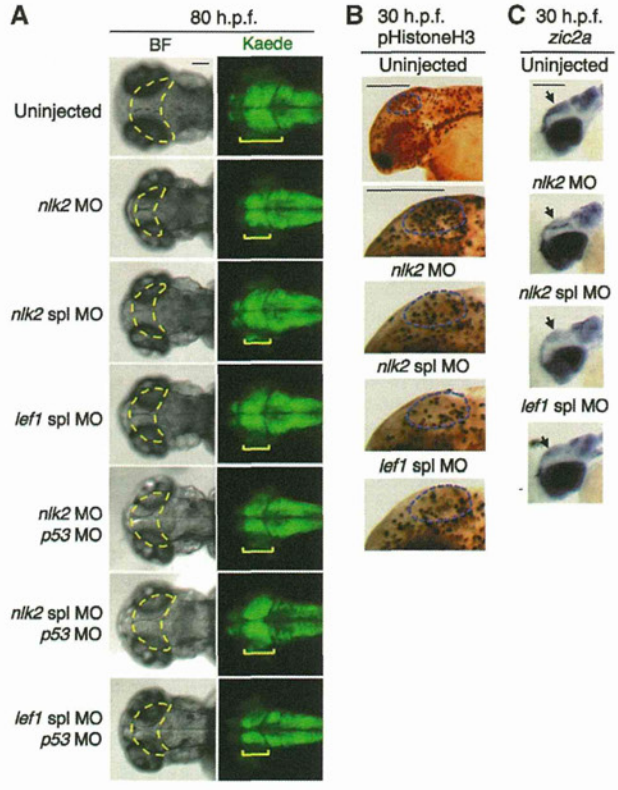

Figure 2 Nlk2 and Lef1 are essential for the midbrain tectum development in zebrafish. (**A**) HuC:kaede-transgenic zebrafish embryos injected with *nlk2* MO, *nlk2* spl MO, *lef1* spl MO, or *p53* MO as indicated. Panels show the dorsal head views of 80 h.p.f. embryos with the anterior to the left. Neurons expressing Kaede were visualized by fluorescence microscopy (right panels). Rectangles indicate the tectum. Bright-field (BF) images are shown in left panels. Broken lines indicate the tectum or presumptive tectal region. Scale bar, 50 μm. (**B**) Knockdown of either *nlk2* or *lef1* decreases the number of proliferating cells in midbrain. Anti-phospho-histone H3 immunostaining of 30 h.p.f. zebrafish embryos injected with *nlk2* MO, *nlk2* spl MO, or *lef1* spl MO, as indicated. Top panels show the left side head views of embryos with the anterior to the left. The other panels show the left side midbrain view of embryos with the anterior to the left. Broken lines indicate the presumptive tectal region. Scale bar, 250 μm. (**C**) Knockdown of either *nlk2* or *lef1* reduces expression of *zic2a* in midbrain. Panels show whole mount *in-situ* hybridization for *zic2a* in 30 h.p.f. embryos. Embryos were injected with *nlk2* MO, *nlk2* spl MO, or *lef1* spl MO as indicated. Panels show the left side head views of embryos with the anterior to the left. Expression of *zic2a* in midbrain is indicated with arrows. Scale bar, 250 μm.

of the midbrain tectum at 80 h.p.f. (Figure 2A; Supplementary Figure S5; Supplementary Table SI) but did not affect the development of the hindbrain (Supplementary Figure S5). Injection of *lef1* spl MO also resulted in a phenotype similar to that observed following injection of *nlk2* MOs. Co-injection of *p53* MO together with an MO for *nlk2* or *lef1* had no effect on the phenotype induced by MO-mediated knockdown of *nlk2* or *lef1* (Figure 2A; Supplementary Table SI). These results suggest that Nlk2 and Lef1 contribute to midbrain tectum development.

We next explored the mechanism by which Nlk2 and Lef1 contribute to tectum development. A previous report has shown that Lef1-mediated Wnt/β-catenin signalling promotes the proliferation of NPCs by activating the transcription of the *zic2a* and *zic5* genes in zebrafish developing midbrain (Nyholm *et al*, 2007). To examine whether Nlk2 contributes to the proliferation of midbrain NPCs, 30 h.p.f. zebrafish embryos were immunostained with anti-phospho-histone H3 antibody, which labels the nuclei of proliferating cells. Knockdown of *nlk2* or *lef1* decreased the numbers of phospho-histone H3-positive cells in the midbrain (Figure 2B; Supplementary Table SI), suggesting that Nlk2 and Lef1 are required for cell proliferation in the developing midbrain. We also found that expression of *zic2a* was decreased in the midbrain, but not in other regions of the brains of 30 h.p.f. embryos injected with *nlk2* MO, *nlk2* spl MO or *lef1* spl MO (Figure 2C; Supplementary Table SI). Using quantitative PCR (qPCR), we confirmed that injection of *nlk2* MO or *lef1* spl MO reduced the expression levels of *zic2a* in the midbrain (Supplementary Figure S2B). Our results suggest that Nlk2 and Lef1 promote the proliferation of NPCs through Wnt/β-catenin signalling in the developing midbrain.

Nlk2 phosphorylates Lef1 in zebrafish

We have previously reported that NLK phosphorylates human LEF1 at Thr-155 and Ser-166 *in vitro* (Ishitani *et al*, 2003b). The Thr residue is conserved between human and zebrafish (Figure 3A). To examine whether Nlk2 phosphorylates Lef1 at Thr-151, we generated an antibody that specifically recognizes phosphorylation of Lef1 at the conserved Thr residue (Figure 3A). This anti-phospho-Lef1 (anti-pLef1) antibody recognized Lef1 when it was co-expressed in mammalian neuro-2a cells with Nlk2, but not with kinase-negative Nlk2 (Figure 3B). These data suggest that Nlk2 phosphorylates Lef1 Thr-151. To verify antibody specificity, we generated a Lef1 mutation, Lef1(T151A), in which Thr-151 was changed to alanine. Lef1(T151A) was not detected by anti-pLef1 antibody when co-expressed with Nlk2 (Figure 3B).

We next investigated whether Nlk2 phosphorylates Lef1 in zebrafish embryos. We performed immunoblotting assays with anti-pLef1 and anti-Lef1 antibodies. Anti-Lef1 detected a protein of about 50 kDa in zebrafish embryo extracts (Figure 3C). In embryos injected with a translation-blocking MO against *lef1* (*lef1* MO) (Ishitani *et al*, 2005), levels of the protein detected by anti-Lef1 antibody decreased (Supplementary Figure S6), confirming that this protein corresponds to Lef1. Western blotting with anti-pLef1 antibody revealed that Lef1 phosphorylation could be detected in 24 h.p.f. embryo extracts (Figure 3C). Injection of *nlk2* MO reduced Lef1 phosphorylation, but had little effect on Lef1 protein levels. Thus, Nlk2 is able to phosphorylate Lef1 in 24 h.p.f. zebrafish embryos.