Supplementary Information

Detection of somatic single nucleotide variations (SNVs)

To predict somatic SNVs, the alignment results were classified so as to construct three datasets. Dataset 1 included paired-end reads with both ends aligned uniquely and with proper spacing and orientation. Dataset 2 included paired-end reads that aligned uniquely for at least one read and exhibited proper spacing and orientation of the reads. Dataset 3 included dataset 2 and paired-end reads for which both ends aligned uniquely but with improper spacing, orientation, or both. Dataset 1 likely includes false positive somatic SNVs because of the low sequence depth of the non-tumor genome, and dataset 3 likely includes false positives due to misalignments of the sequence reads. To reduce the number of false positives, the following filters were applied to these three datasets, and concordant somatic SNVs among the three datasets were selected: (i) a mapping quality score of 20 was used as a cutoff value for read selection; (ii) base quality scores of 10 and 15 were used as cutoff values for base selection for the tumor and non-tumor genomes, respectively; (iii) SNVs were selected when the frequency of the non-reference allele was at least 15% in the tumor genome and 3% in the non-tumor genome; (iv) SNVs located within 5 bp from a potential insertion or deletion were discarded; (v) SNVs with a root mean square mapping quality score of less than 40 for reads covering the SNV were discarded; (vi) when there were three or more SNVs within any 10 bp window, all of them were discarded; (vii) SNVs with a consensus quality score less than 20 as calculated by SAMtools (version 0.1.5c) were discarded;

(viii) when a base with a consensus quality score less than 20 was located within 3 bp of an SNV, the SNV was discarded; and (ix) for the tumor genome, SNVs found in at least two sequence reads with at least one base with a quality score greater than 30 were selected. By comparing the predicted nucleotide variations in the tumor and non-tumor genomes, somatic SNVs that occurred only in the tumor genome were identified. If the positions of somatic SNVs were not covered in the non-tumor genome by at least six sequence reads, these somatic SNVs were discarded. After predicting somatic SNVs, these additional filters were applied: (x) somatic SNVs registered in dbSNP were removed if the read depth of the somatic SNV site in the non-tumor genome was less than 10; and (xi) in the non-coding region, tandem repeat regions detected by the Tandem Repeat Finder program and repetitive regions within 1 Mb of a centromeric or telomeric sequence gap detected by the Repeat Masker program were excluded. We randomly selected 92 predictions of somatic substitutions. We examined these 92 substitutions by Sanger sequencing of both the tumor and normal genomes and validated 78 as somatic. Of the remaining 14, seven could not be sequenced due to the surrounding repetitive sequences, six could not be validated and one was validated as germline variation. Therefore, the prediction accuracy of our detecting somatic substitutions was estimated as 91.8% (78/85).

Detection of somatic short insertions/deletions

To reduce the number of false positives, the following filters were applied to the three

datasets described above in the somatic SNV section in order to select concordant somatic indels among these three datasets. (i) Indels found in at least six sequence reads for the tumor and in at least one sequence read for non-tumor genomes, were selected; (ii) indels were selected when the frequency of the indel allele was at least 20% in the tumor genome; (iii) indels with a root mean square mapping quality score of less than 40 for the reads covering the indel were discarded for the tumor genome; and (iv) only indels that had the best SNP quality scores in any 10 bp window were selected for the tumor genome. By comparing the predicted indels in the tumor and non-tumor genomes, somatic indels that occurred only in the tumor genome were identified. If the positions of somatic indels were not covered in the non-tumor genome by at least ten sequence reads, these somatic indels were discarded. After predicting somatic indels, the following additional filters were applied: (v) somatic indels registered in dbSNP were removed if the read depth of the somatic indel site in the non-tumor genome was less than 15; (vi) if indels in the non-tumor genome were found within the 5 and 10 bp flanking regions of a potential somatic indel site in coding and non-coding regions, respectively, this somatic indel was removed; and (vii) in the non-coding region, the tandem repeat regions detected by the Tandem Repeat Finder program and the repetitive regions detected by the Repeat Masker program were excluded. We randomly selected 34 somatic indels. We tested these 34 indels by Sanger sequencing of both the tumor and normal genomes and validated 17 as somatic alterations. Of the remaining 17, seven could not be sequenced due to the surrounding repetitive sequences and ten could not be validated. Therefore, the prediction accuracy of our approach for detecting somatic

indels was estimated to be 63.0% (17/27).

Detection of somatic structural alterations

Fifty bp paired-end reads were used for rearrangement analysis since they contain longer spacers than 100 bp paired-end reads. Therefore, 100 bp paired-end reads were cut to generate 50 bp paired-end reads. To detect structural alterations, paired-end reads for which both ends aligned uniquely to the human reference genome, but with improper spacing, orientation, or both, were considered.

First, paired-end reads were selected based on the following filtering conditions:

(i) sequence reads with mapping quality scores greater than 37;

(ii) sequence reads aligned with two mismatches or less.

Rearrangements were then identified using the following analytical conditions:

(i) forward clusters and reverse clusters, which included paired-end reads respectively, were constructed from the end sequences aligned with forward and reverse directions respectively;

(ii) two reads were allocated to the same cluster if their end positions were not farther apart than the maximum insert distance of pair-end library;

(iii) clusters with a distance between the leftmost and rightmost reads that was greater than the maximum insert distance were discarded;

(iv) paired-end reads were selected if one end sequence fell within the forward cluster and the other end fell within the reverse cluster (we hereafter called this pair of forward and reverse clusters as paired-clusters);

(v) if paired-clusters overlapped with other paired-clusters, all of the overlapping paired-clusters were discarded;

(vi) for the tumor genome, rearrangements predicted from paired-clusters which included at least four pairs of end reads and at least one pair of end reads perfectly matched to the human reference genome, were selected;

(vii) for the non-tumor genome, rearrangements predicted by at least one pair of end reads were selected.

By comparing the predicted rearrangements in the tumor and non-tumor genomes, somatic rearrangements that were only detected in the tumor genome were identified.

In total 60,888 rearrangements (33,908 deletions, 7,004 inversions, 6,095 tandem duplications and 13,881 translocations) were predicted in ten tumor genomes and 4,212 rearrangements (961 deletions, 788 inversions, 553 tandem duplications and 1,910 translocations) remained by subtracting rearrangements predicted in the corresponding ten normal genomes. Many rearrangements were subtracted as false-positive rearrangements rather than germline rearrangements since there were many misalignments due to sequence variations between the analyzed genome and the reference genome, or due to the presence of multiple similar sequences in the reference genome. To reduce misalignments, we further applied the following filtering conditions.

(i) Paired-end reads included within paired-clusters were aligned to the human reference genome using the BLASTN program and these false positive rearrangements were removed using the two following analytical conditions.

(ii) If one end sequence was aligned to the region of paired-clusters (the flanking region of the rearrangement breakpoint) and the other end was aligned with proper spacing and orientation, this rearrangement was removed. An expectation value of 10^{-1} was used as a cutoff value for BLASTN so that most variations between the analyzed genome and the reference genome could be removed.

(iii) If at least one end sequence of paired-end reads was aligned to other regions outside of paired-clusters, this rearrangement was removed. An expectation value of 10^{-10} was used as a cutoff value for BLASTN so that only highly homologous sequences in the reference genome could be detected.

Using these BLASTN filters, finally 350 rearrangements (61 deletions, 134 inversions, 91 tandem duplications and 64 translocations) remained. We randomly selected 80 predicted somatic rearrangements (20 deletions, 35 inversions, 5 tandem duplications and 20 translocations) and performed validation analysis. We amplified DNA fragments of the tumor genome containing the breakpoints of these 80 rearrangements and determined the exact breakpoints of 67 rearrangements by Sanger sequencing. All 67 rearrangements were validated as somatic events by analyzing the corresponding the normal genome. Of the remaining 13, 12 could not be amplified or not sequenced due to the surrounding repetitive sequences and only one could not be validated. Therefore, the prediction accuracy of our approach for detecting somatic rearrangement was estimated to be 98.5% (67/68).

Estimation of significantly mutated genes

Since the number of mutations in a gene is influenced by the length of the gene and CpG sites, the probability of the number of protein-altering mutations was calculated under the given mutation rate and gene length using the following methods.

(1) Substitution rate was estimated by dividing the total number of synonymous mutations by the total number of synonymous sites in the genome. Since the substitution rate in CpG sites was much higher than that of other regions, the substitution rate in CpG and non-CpG site was estimated separately. For each gene, the number of nonsynonymous sites and splice sites in CpG and non-CpG site was counted separately and the expected number was calculated by multiplying the substitution rate by the total number of nonsynonymous sites and splice sites.

(2) Coding indel rate was estimated by dividing the total number of coding indels by the total number of coding sites in the genome. For each gene, the expected number was calculated by multiplying the coding indel rate by the coding length.

(3) Rearrangement rate was estimated by dividing the total number of rearrangements by the length of the genome. For each gene, the expected number was calculated by multiplying the rearrangement rate by the gene length.

(4) The expected number of protein-altering mutations was calculated by joining the expected number of nonsynonymous and splice site substitutions in CpG and non-CpG sites, coding indels and rearrangements.

(5) Tests of significance for each gene were performed by assuming a Poisson

distribution. The adjustment by multiple testing was performed using the Benjamini and Hochberg's method.

Detection of somatic substitutions from pooled DNA

We performed target exon resequencing of pooled DNA from 47 chondrosarcoma samples, 19 corresponding adjacent non-tumor tissue samples, and 41 enchondroma samples. We obtained 486,506 and 325,846 sequence coverage at each nucleotide position on average in the chondrosarcoma and enchondroma samples, respectively. Therefore, the expected mutated allele frequency was 0.0106 (0.5/47) for chondrosarcoma and 0.0122 (0.5/41) for enchondroma, and the expected numbers of supported reads for each mutation were 4,962 for chondrosarcoma and 3,878 for enchondroma. In view of this, we set the following stringent conditions: (i) a mapping quality score of 20 was used as a cutoff value for read selection; (ii) a base quality score of 12 was used as a cutoff value for base selection; (iii) substitutions with allele frequency greater than 0.01 and number of supported reads greater than 2,000 were selected for the tumor genome; and (iv) substitutions with allele frequency greater than 0.0005 and number of supported reads greater than 100 were selected for the non-tumor genome. By comparing the selected substitutions in the tumor and non-tumor genomes, somatic substitutions that occurred only in the tumor genome were selected. If the positions of somatic substitutions were not covered in the non-tumor genome by at least 20,000 sequence reads, these somatic substitutions were discarded. If the selected substitutions were found in dbSNP or in 225 Japanese germline samples sequenced

in-house, these substitutions were removed. To remove rare single nucleotide polymorphisms (SNPs) that were not found in dbSNP or the 225 in-house sequenced germline samples of Japanese, the numbers of rare SNPs were estimated from these 225 germline samples. The copy number of one rare SNP in the *YEATS2* gene was estimated in chondrosarcoma and enchondroma cases. This estimated number of rare SNPs was subtracted from the number of somatic substitutions. To remove false positives due to PCR amplification errors, the numbers of false positives were estimated from the prediction of somatic substitutions, described above, using the non-tumor genome. Two, one, and one false positive in the *COL2A1*, *YEATS2*, and *ACVR2A* genes, respectively, were estimated in the chondrosarcoma cases, and two, three, and one false positives in the *COL2A1*, *YEATS2*, and *ACVR2A* genes, respectively, were estimated in the enchondroma cases. These estimated numbers of false positives were subtracted from the number of somatic substitutions.

Verification of somatic mutations by MassArray system

All candidate 137 SNVs and 13 short indels for *ACVR2A*, *COL2A1* and *YEATS2* and the mutational hot spots for *IDH1* and *IDH2* were further verified in individual case by MassArray system (Sequenom). The primer sets, which include a pair of amplicon primers and an extension primer for each SNV, were designed using the MassARRAY Designer software (Sequenom). The amount of DNA added to the PCR was 10 ng per reaction, which was quantified using the Qubit 2.0 Fluorometer (Invitrogen). A single base extension reaction was performed using iPLEX Pro reagents and an allele-specific

mass difference was determined using the SpectroCHIP arrays placed into the matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometer (Sequenom). Mutations with the allele frequency more than 0.2 were identified reviewing manually the analyzed data by the MassARRAY Typer Analyzer 4.0 software (Sequenom).

Analysis of somatic substitution patterns

The number of somatic substitution patterns, C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G, and T>G/A>C, was counted. Dividing by the total substitution number, their frequency was used for principal component analysis (PCA). PCA was implemented using the R command prcomp with the scaling option on. Similarity of somatic substitution patterns of tumors was examined by permutation test. We selected all possible combinations of sets of the classified tumors, and calculated the average distance between tumors in each set. The P-value was calculated as the proportion of the sets for which average distance was shorter than or equal to the average distance of the classified tumor set. The principal components for which eigenvalues were greater than 1 were used for these permutation tests. The numbers of 96 triplet sequence patterns of somatic substitutions (substitutions with immediate 5' and 3' nucleotides) were also counted. Dividing by the total substitution number, their frequency was used for PCA of the triplet sequence patterns.

Detection of somatic copy number alteration

The average sequence depth was calculated for several window sizes (500, 5,000, 10,000, and 100,000 bp) for both the tumor and non-tumor genomes using only sequence reads that uniquely aligned to the human reference genome. The ratio of standardized average sequence depth between non-tumor and tumor genomes (log2R ratio) was calculated. Copy number alteration regions were defined by segment using the R command for DNA Copy (Andersson et al. 2008) with undo.split="sdundo", undo.SD=3, and trim=0.0001.

Analysis of whole transcriptome sequence data

After removing PCR duplications (same paired-end sequences), 100 bp paired-end reads from RNA sequencing were mapped to known RNA sequences in the RefSeq and Ensembl databases using the Bowtie (Langmead et al. 2009) program. The Bowtie program was performed with the –v 3 option so that three or fewer mismatches were allowed, and with the –a option so that all multiple hits could be detected, since there are many spliced variants in the RNA databases. After selecting the best hits with proper spacing and orientation, the number of reads per kilobase pairs per million reads (RPKM) was calculated.

References for Supplementary Information

Andersson R, Bruder CE, Piotrowski A, Menzel U, Nord H, Sandgren J, Hvidsten TR, Diaz de Ståhl T, Dumanski JP, Komorowski J. 2008. A segmental maximum a posteriori approach to genome-wide copy number profiling. *Bioinformatics* **24**:

751-758.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.