Deep whole genome sequencing Japanese Healthy Population

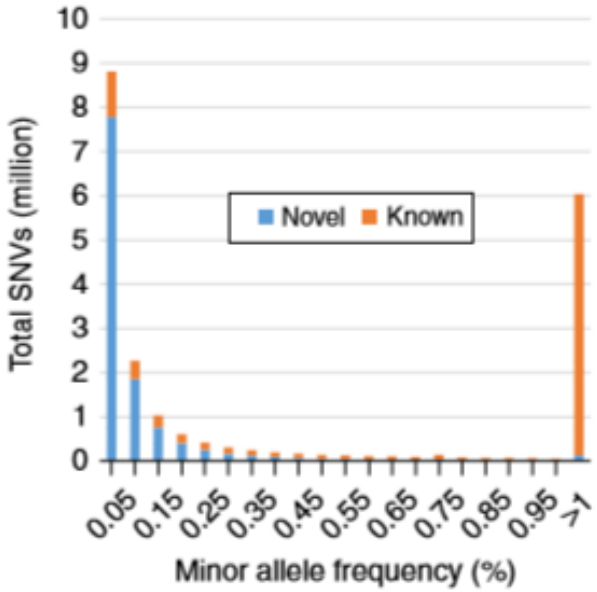
Hiroshi Tanaka
Special Advisor to Executive Director
Tohoku Medical Megabank Organization (ToMMo)
Tohoku University

Whole Genome Sequencing in Tohoku Medical Megabank Project

- Whole genome sequencing (WGS) of 1,070 healthy Japanese individuals
  - executed by PCR-free sequencing
  - more than 30X coverage (average 32.4X) .
- First results of WGS in healthy Japanese
- Very rare as well as novel single-nucleotide variants (SNVs) are identified
  - Totally 21.2 million SNV
  - 12 million novel SNV
- A reference panel of 1,070 Japanese individuals (1KJPN) is constructed
  - From the identified SNVs, we construct 1KJPN, including some very-rare SNVs.
- From this panel, we designed costumed SNP array for Japanese
  - Japonica array
  - 650 thousand SNV

Data Processing and variant discovery

- Material
  - 1344 candidates were selected from biobank
    - Considering traceability of participants' information
    - Quality and abundance of DNA sample for SNP array and WGS
  - 1070 samples were selected by measured results by Omni2.5
    - By filtering out close relatives and outliers
  - Sequenced by Illumina Hiseq2500
    - Using PCR-free protocol
- Variant discovery
  - 21.2 million high confident SNV
  - 12 million novel SNVs
    - After several filtering procedure, high confident SNVs
    - Reference genome: GRCh37/hg19
    - False discovery rate <1.0%



The numbers of novel and known SNVs in each MAF bin.

The novel SNV frequency begins to dominate for lower MAF

Summary of WGS of Japanese individuals and variant detection in autosomes.

Total samples	1,070
Total raw bases	100.4 trillion bases
Mean sequenced depth	32.4 ×

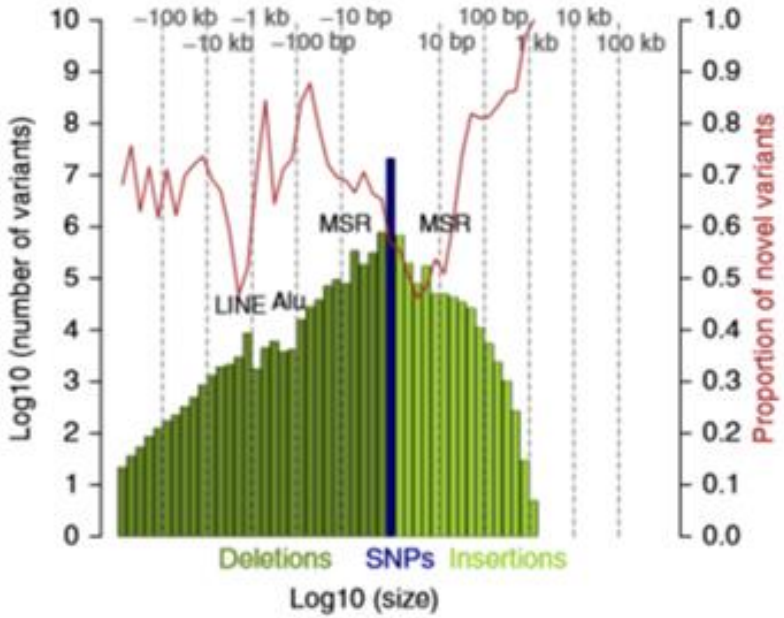
SNVs	High-confidence SNVs
Total	21,221,195
Number of known variants*	9,219,783
Number of novel variants*	12,001,412
Novelty rate	56.55%
Average number per sample	2,716,853
Average individual heterozygosity	1,532,773

Deletions	1 bp ≤ length < 100 bp	100 bp ≤ length
Number of sites overall	1,969,302	47,343
Number of novel variants†	1,429,636	—
Novelty rate	72.60%	—
Number of inframe/frameshift	3,112/4,454	—
Average number per sample	190,857	2,654

Insertions	1 bp ≤ length < 100 bp	100 bp ≤ length
Number of sites overall	1,384,230	9,354
Number of novel variants†	1,037,839	9,354
Novelty rate	74.98%	—
Number of inframe/frameshift	1,577/2,506	—
Average number per sample	159,359	45

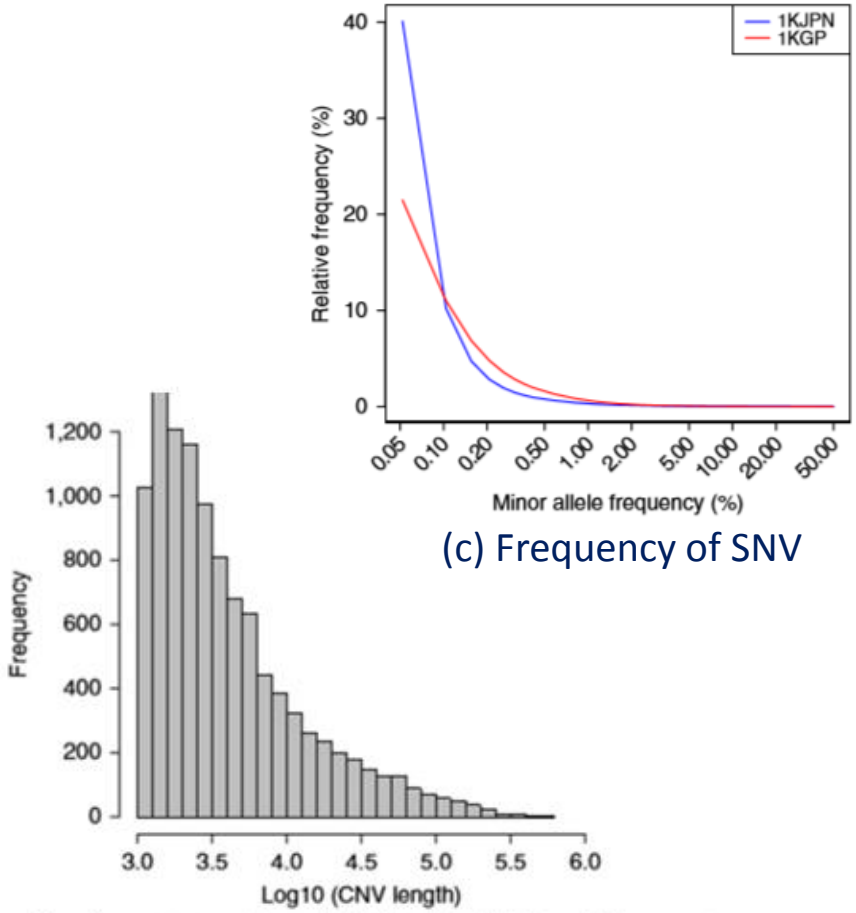
Copy number Variants 25,923

Statistics of Indel and SNV



The size-frequency spectrum of SNVs, deletions and insertions discovered by high-coverage sequencing in 1KJPN. Novelty rates are shown by the red line. Peaks corresponding to long interspersed elements (LINE), Alu and microsatellite repeat (MSR) are shown.

(a) Size-frequency of Del, SNP, Ins



(c) Frequency of SNV

Size-frequency spectrum of CNVs estimated from high-coverage sequencing data in the genic regions in 1KJPN.

(b) Size-frequency of CNV

Japonica Array

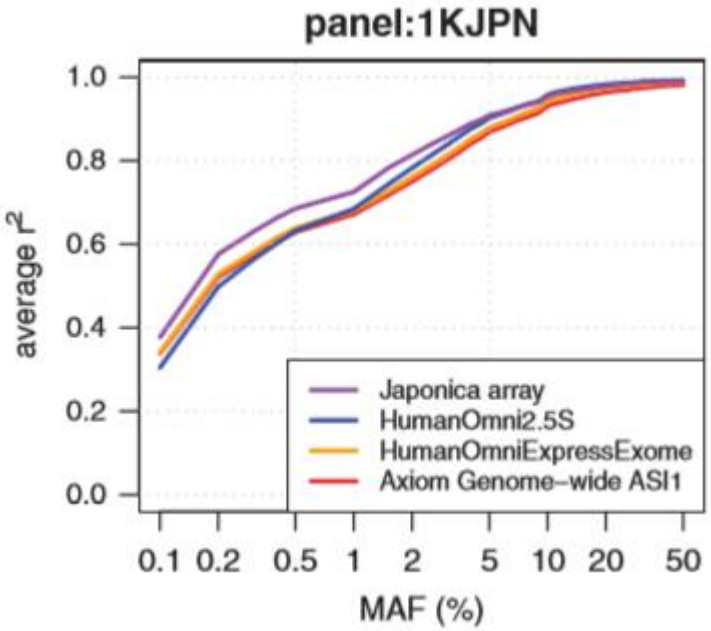
- Novel custom-made SNP array, based on the 1KJPN panel, for whole-genome imputation of Japanese individuals.
- The array contains 659, 253 SNPs
  - tag SNPs for imputation,
  - SNPs of Y chromosome and mitochondria,
  - SNPs related to previously reported genome-wide association studies and pharmacogenomics.
- Better imputation performance for Japanese individuals than the existing commercially available SNP arrays
  - Common SNPs (MAF>5%), the genomic coverage of the Japonica array (r^2>0.8) was 96.9%
  - Coverage of low-frequency SNPs (0.5%<MAF⩽5%) :67.2%,
- High quality genotyping performance of the Japonica array using the 288 samples in 1KJPN;
  - Average call rate 99.7%
  - Average concordance rate 99.7% to the genotypes obtained from high-throughput sequencer.

Japonica Array

Category of SNPs on the Japonica array

Category	*Number of SNPs*[a]	*Array occupancy rate*
Tag SNPs (including X chromosome)	638 269	96.8%
Pharmacogenomics markers	2028	0.31%
Y chromosome	275	0.04%
Mitochondria	70	0.01%
NHGRI GWAS catalog	10 798	1.64%
HLA	3906	0.59%
Untaggable functional SNPs	3990	0.61%
Total	659 253	—

Abbreviations: GWAS, genome-wide association studies; SNP, single nucleotide polymorphism.
[a]Some SNPs are overlapped among categories.



Japonica array (96sample)