Novel insights into mid-sized heterozygous structural variations and DNA methylome in a diploid human genome through SMRT sequencing

講師：森下 真一
東京大学 新領域創成科学研究科 情報生命科学専攻 教授
京都大学化学研究所 客員教授

日時：7月14日（月）2:00 pm～
場所：総合研究実験棟2階 CB207

Direct observation of mid-sized (1-10 kb) heterozygous structural variations (SVs) in human genomes has long posed a technical challenge. The use of single-molecule real-time (SMRT) sequencing is promising because of its ability to output long reads without suffering from GC-bias, although its error rate is relatively high (~15%). We developed a method of correcting errors in SMRT reads in a reasonable amount of time using sufficient Illumina short reads, and used it to process ~30-fold coverage of SMRT reads sequenced from the diploid genome of a Japanese individual. We also developed a method of identifying mid-sized hetero- and homozygous SVs and uncovered 831 mid-sized SVs, including 307 in introns, 6 that duplicated or altered exons, 10 in pericentromeres, and 200 variable number tandem repeats that were expanded in the Japanese genome and were on average ~2000 b long regardless of unit length. The length of a mid-sized SV stemming from nonallelic homologous recombination was not correlated with its breakpoint homology length, which is contrary to the findings of previous studies in larger SVs. We also identified 201 novel Alu insertions into the Japanese sample. We reconfirmed all SVs using ~15-fold coverage of SMRT reads from a biological replicate. Our novel insights demonstrate the importance of exploring mid-sized SVs, which are currently missing from personal genomics.

SMRT sequencing also shows promise to determine the methylation information for regions with low or high copy numbers. Solving this problem by second-generation sequencing is challenging because the read length is insufficient, especially when the repetitive regions are long and nearly identical to each other. To resolve these problems, SMRT sequencing shows potential because it is not vulnerable to GC bias, it has long read lengths, and its kinetic information is sensitive to DNA modifications. However, raw kinetic information at a single CpG site contains some noise, and characterizing the DNA methylation for large size genomes demands prohibiting coverage of SMRT reads. Since hypo-/hypermethylated CpG dinucleotides are often contiguous over a long span in vertebrate genomes, we propose a novel algorithm that combines the kinetic information for neighboring CpG sites and increases the confidence in identifying the methylation statuses of those sites when they are correlated. The sensitivity and specificity of our algorithm were both of >90% for the genome of an inbred medaka (Oryzias latipes) strain within a practical read coverage of <30-fold. With this method, we newly characterized the methylation status of repetitive elements (e.g., the occurrences of ~6-kb-long interspersed nuclear elements (LINEs)), regions of duplicated genes (e.g. , HIST2 cluster) in the human genome, and nearly identical living transposons of length 4682 bp in the medaka genome, which were difficult to observe using bisulfite-treated short reads.

This is a joint work with Shoichiro Oishi, Yuta Suzuki, Koichiro Doi, Hideaki Yurino, Shingo Tomioka, Wei Qu, Jun Yoshimura, Jonas Korlach, Stephen Turner, Jun, Mitsui, Yuji Takahashi, Shoji Tsuji, Tatsuya Tsukahara, and Hiroyuki Takeda.

問い合わせ：化学研究所バイオインフォマティクスセンター生命知識工学研究領域
馬見塚 拓（内 3023、メール mami@kuicr.kyoto-u.ac.jp)