Summary of the autosomal SNP data of Ritidian Beach Cave, Guam

Alex Hübner

Dec 08, 2019

From two samples from Ritidian Beach Cave, Guam, DNA was extracted multiple times and Illumina sequencing libraries were generated (Table 1). The produced sequencing libraries were initially screened using shotgun sequencing before capture-enrichment for both mtDNA and autosomal SNPs from the 1240K array (Haak et al. 2015, @Fu2015) was performed and the enriched libraries paired-end sequenced across multiple sequencing runs.

Table 1: Number of produced Illumina sequencing libraries for Ritidian Beach Cave samples.

name	sampleID	no. of libraries
RBC1	SP4210	13
RBC2	SP4211	11

The sequencing data was processed by the following steps. First, Illumina sequencing adapters were removed using *leehom* (Renaud, Stenzel, and Kelso 2014) and overlapping read pairs were merged. The sequencing data de-multiplexed by assigning reads only to a sample when there were no mismatches between the observed and expected double-index combination of seven basepair length. The sequencing data was subsequently aligned against the human reference genome *hg19* from the 1000Genomes projects with decoy sequences (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa) using BWA *aln* (Li and Durbin 2009) with the settings "-n 0.01 -o 2 -l 16500" to allow for a higher number of mismatches and switch off the seeding algorithm. We filtered the aligned sequencing data and only kept reads for which the paired reads could be merged, had a minimum read length of 35 bp, and a minimal mapping quality of 25. Finally, duplicated reads were identified and removed using DeDup (Peltzer et al. 2016).

Table 2: Summary of the filtering process summarised by library. The percentages of the reads with length >= 35 bp [L35] and the mapping quality >= 25 [MQ25] are based on the raw reads.

sample	libraryID	raw reads	reads w/ L35	reads with L35 [%]	reads with MQ25 [%]	final reads	duplication rate
RBC1	D5864	446659	324102	72.56	26.52	3593	32.97
	F8851	2257410	1461056	64.72	0.01	165	1.01
	F8852	1955908	1283610	65.63	0.01	150	1.01
	F8853	1793979	1061543	59.17	0.01	136	1
	F8854	1720163	1238355	71.99	0.01	102	1
	F8855	2083565	1505265	72.24	0.01	117	1.01
	F8856	2149938	1537783	71.53	0.01	135	1.01
	F8857	1925097	1443207	74.97	0.01	101	1.02
	F8858	1922169	1433610	74.58	0.01	96	1.01
RBC2	D5865	544588	401578	73.74	40.95	23351	9.55
	F8862	1970355	1257609	63.83	0.01	200	1.01
	F8863	2032066	1303693	64.16	0.01	228	1

sample	libraryID	raw reads	reads w/ L35	reads with L35 [%]	reads with MQ25 [%]	final reads	duplication rate
	F8864	2048503	1314263	64.16	0.01	253	1
	F8865	1950996	1408813	72.21	0.03	585	1
	F8866	1490887	1077709	72.29	0.03	410	1
	F8867	1723947	1244039	72.16	0.03	483	1

References

Fu, Qiaomei, Mateja Hajdinjak, Oana Teodora Moldovan, Silviu Constantin, Swapan Mallick, Pontus Skoglund, Nick Patterson, et al. 2015. "An Early Modern Human from Romania with a Recent Neanderthal Ancestor." *Nature* 524 (7564). Nature Publishing Group: 216.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe." *Nature* 522 (7555). Nature Publishing Group: 207.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14). Oxford University Press: 1754–60.

Peltzer, Alexander, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause, and Kay Nieselt. 2016. "EAGER: Efficient Ancient Genome Reconstruction." *Genome Biology* 17 (1). BioMed Central: 60.

Renaud, Gabriel, Udo Stenzel, and Janet Kelso. 2014. "LeeHom: Adaptor Trimming and Merging for Illumina Sequencing Reads." *Nucleic Acids Research* 42 (18). Oxford University Press: e141–e141.