De novo assembly of human genomes with massively parallel short read sequencing

Mikk Eelmets

Journal Club

06.04.2010

Problem

- DNA sequencing technologies:
  - Sanger sequencing (**500-1000** bp)
  - Next-generation (Illumina Genome Analyzer, Applied Biosystem SOLiD, Helicos BioSciences HeliScope …) (**25-75** bp)

  - De novo assembly VS read mapping onto the reference genome

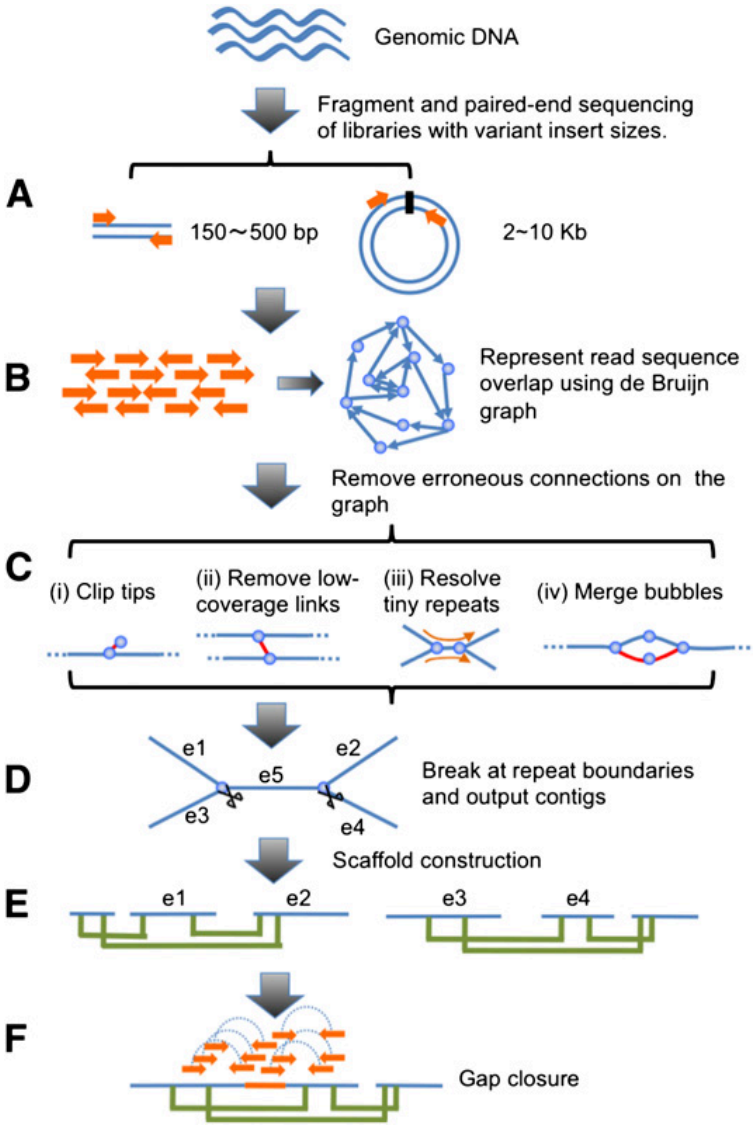
De novo assembly programs

- Overlap-layout strategy (phrap, Celera assembler, ARACHINE, Phusion, RePS, PCAP, Atlas)
- De Bruijn graph data structure (EULER, Velvet, ALLPATHS, EULER-SR, ABySS)
- Overlap and extension (SSAKE, VCAKE, SHARCGS, Edena)

SOAPdenovo

- Short Oligonucleotide Alignmet Program (SOAP) package SOAPdenovo
- Asian and African individual de novo assembly

SOAPdenovo algorithm



Asian individual sequencing data

	Insert size (bp)	Read length (bp)	Phase I data (Gb)	Phase II data (Gb)
Single-end		35	56.3	
		44	15.7	
Paired-end	135	35	38.1	
	440	35	7.6	2.6
		44		20.5
		75		22.1
	2,600	35		1.3
		44		12.3
		75		4.5
	6,000	44		12.3
	9,600	44		6.9
Total			**117.7**	**82.5**
			200.2	

Supplementary table 1. Summary of the produced data for the Asian individual. The short reads were sequenced using Illumina Genome Analyzer (GA) technology. The published phase I data and newly generated phase II data were merged together for *de novo* assembly

Summary of the assembly

Data set	Step	Sequence depth	N50 (bp)	N90 (bp)	Total length	Genome coverage	Gene coverage
Asian genome	Contig	52×	1050	205	2,146,837,026	80.3%	93.4%
	Scaffold (135&440bp PE)	26×	17,331	3838	2,510,643,840	80.3%	93.4%
	Scaffold (+2.6 kb PE)	5×	103,474	21,431	2,718,204,301	80.3%	93.4%
	Scaffold (+6 kb PE)	4×	230,544	47,127	2,800,570,159	80.3%	93.4%
	Scaffold (+9.6 kb PE)	2×	446,283	78,405	2,874,204,399	80.3%	93.4%
	Contig after gap closure		7384	1376	2,457,434,692	87.4%	95.5%
African genome	Contig	40×	886	185	2,098,284,706	79.8%	87.7%
	Scaffold (200bp PE)	40×	4474	936	2,375,357,508	79.8%	87.7%
	Scaffold (+2 kb PE)	4×	61,880	5994	2,696,443,788	79.8%	87.7%
	Contig after gap closure		5909	1004	2,367,973,949	85.4%	89.2%

All read sequences were used in contig assembly, while paired-end libraries with different insert sizes were used step-by-step additively on scaffold construction. N50 of contig or scaffold was calculated by ordering all sequences, then adding the lengths from longest to shortest until the summed length exceeded 50% of the total length of all sequences. N90 is similarly defined. NCBI build 36.1 was used as the reference genome and RefSeq was used as the gene set to evaluate genome and gene region coverage. Since both genomes were sequenced of male individuals, chromosomes X and Y only have half-sequencing depths of the autosomes, and hence were excluded in calculation genome and gene coverage. For calculating scaffold N50 and total length, the intrascaffold gaps were included.

Sequence depth effect on genome assembly

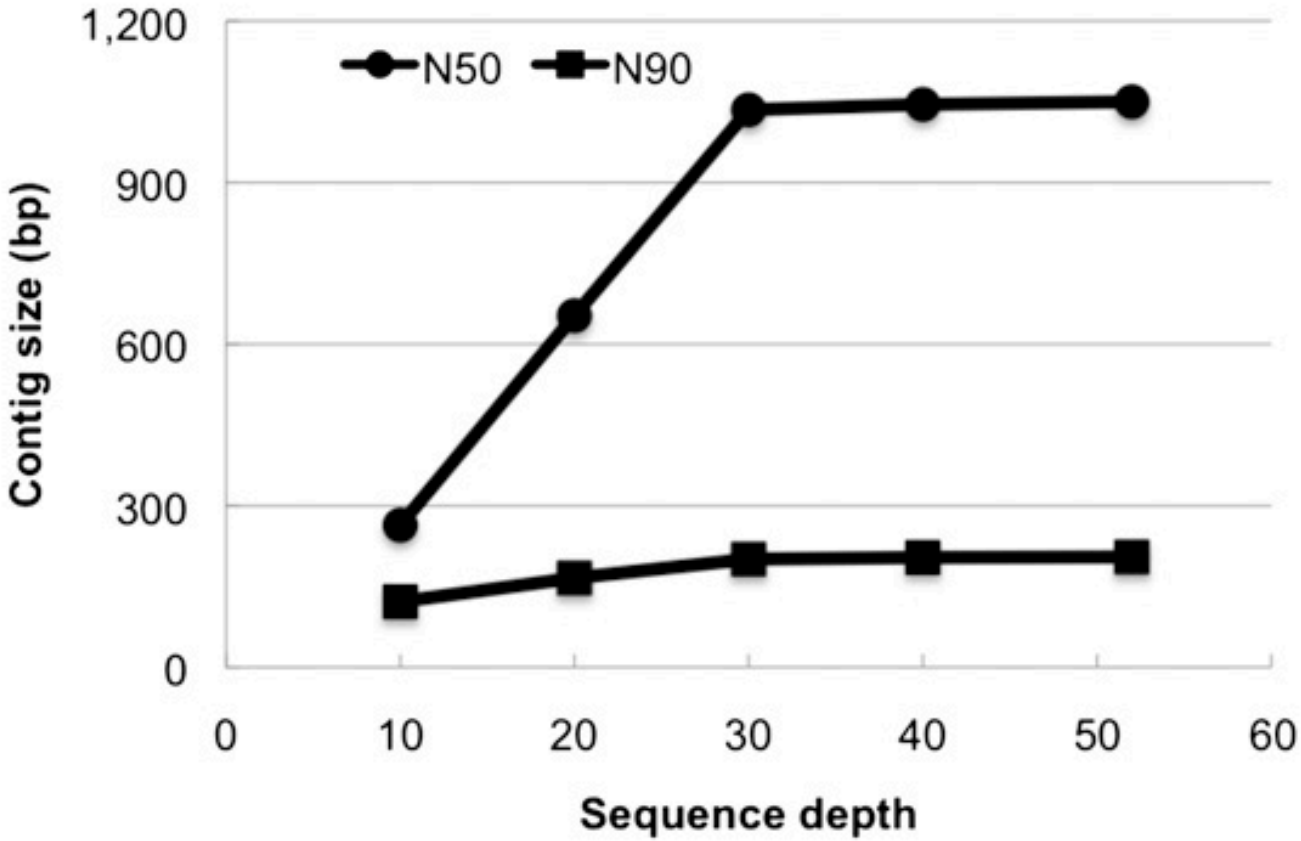


Figure 4. N50 and N90 size of assembled contigs by different sequence depths. We sampled subsets of randomly selected reads from the Asian genome data for de novo assembly of contigs. The same K-mer (K = 25) size was used for all the assemblies.

Computational complexity

Step	Human African			Human Asian		
	Peak memory (Gb)	No. of CPUs	Time (h)	Peak memory (Gb)	No. of CPUs	Time (h)
Preassembly error correction	96	40	22	96	40	24
Construct de Bruijn graph	140	16	8	140	16	10
Simplify graph and output contigs	62	1	3	108	1	6
Remap reads	43	8	2	74	8	4
Scaffolding	23	1	4	15	1	3
Gap closure	35	8	1	53	8	1
Total	140	—	40	140	—	48

Table 4. Statistics of computational complexity at each assembly step. The assemblies were performed on a supercomputer with eight Quad-core AMD 2.3 GHz CPUs with 512 Gb of memory installed, and used the Linux operating system.

Conclusion

SOAPdenovo make it possible for building reference genome sequences for novel species in a more efficient and cost-effective way.

Reference

Ruiqiang Li, Hongmei Zhu, Jue Ruan, et al.

" De novo assembly of human genomes with massively parallel short read sequencing.“

Genome Res. 2010 20: 265-272

THANK YOU

Unique and repeat sequences clusters in human genome

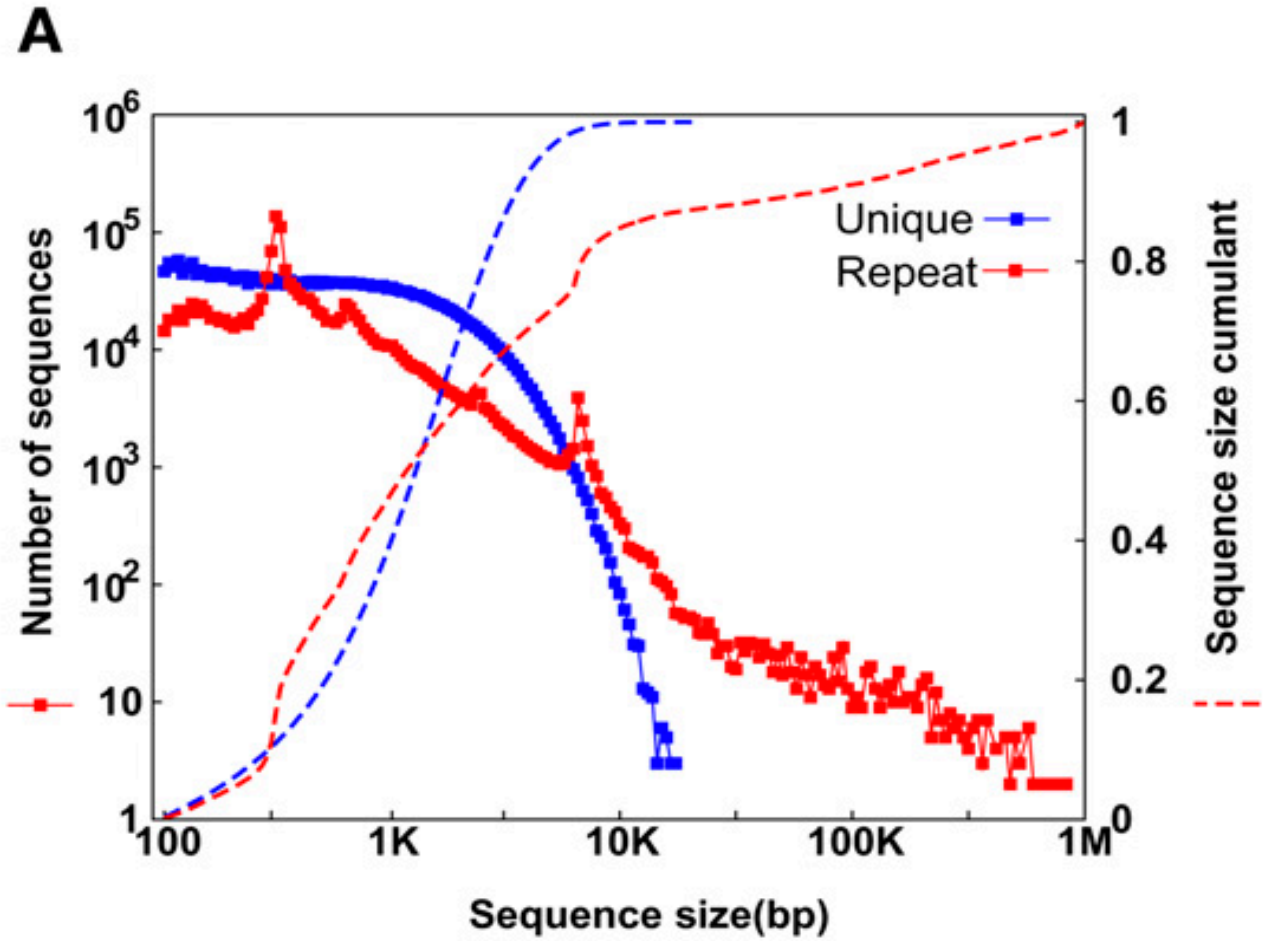


Figure 1. (A) Length distribution of unique and repeat sequence clusters in the human genome. At each chromosomal location, we checked the frequency of the 25-mer in the whole human genome. If it appeared once, we defined it as unique; otherwise it was considered a repeat 25-mer. The regions were then merged as unique clusters and repeat clusters, and those small unique clusters (<100 bp) inside repeat clusters were defined as repeats.

Unique and repeat sequences clusters in human genome

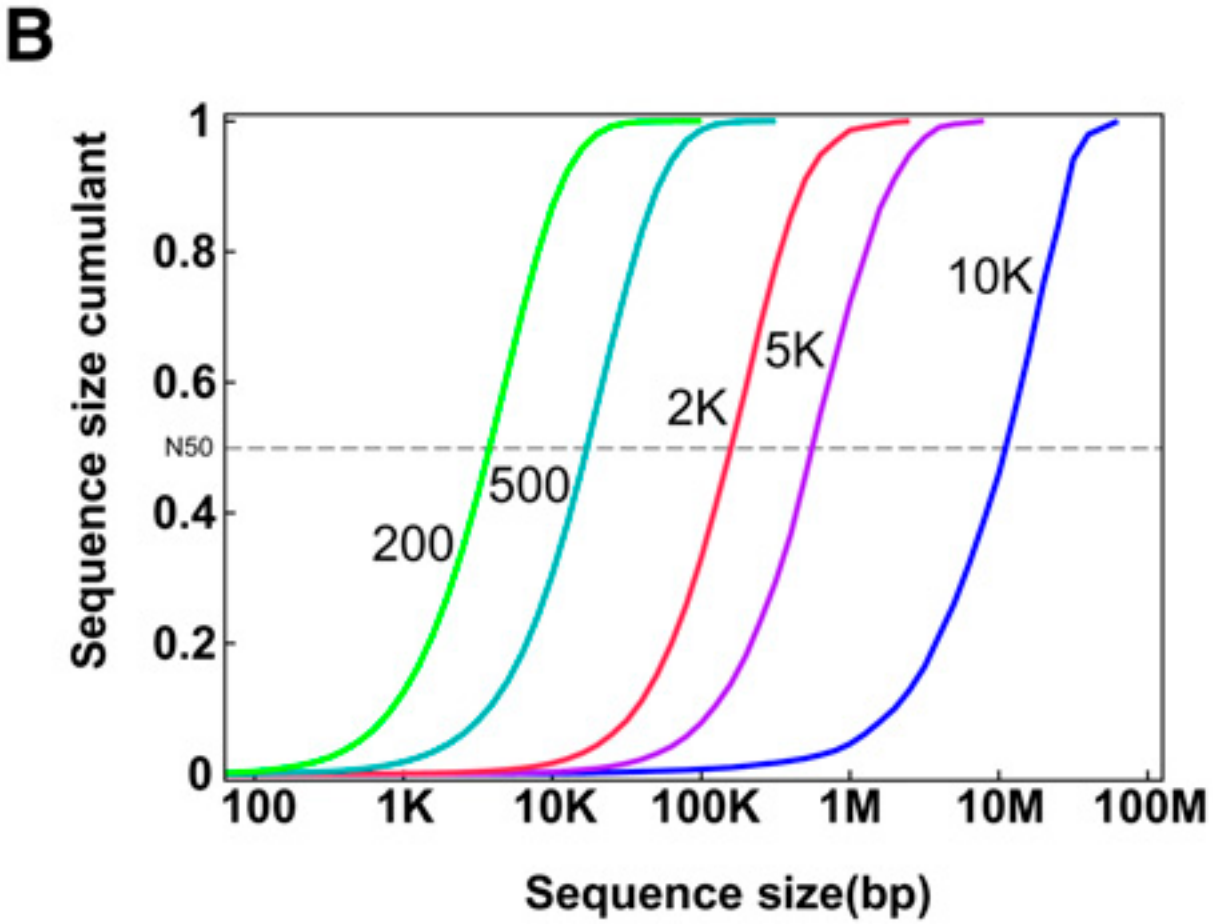


Figure 1. (B) Sequence length distribution of an ideal assembly with each insert-sized paired-ends. The repeat clusters with lengths smaller than the assumed insert size of paired-ends were crossed and the unique clusters were merged. These unique clusters represent the ideal assembly using the paired-ends.

Deletion example

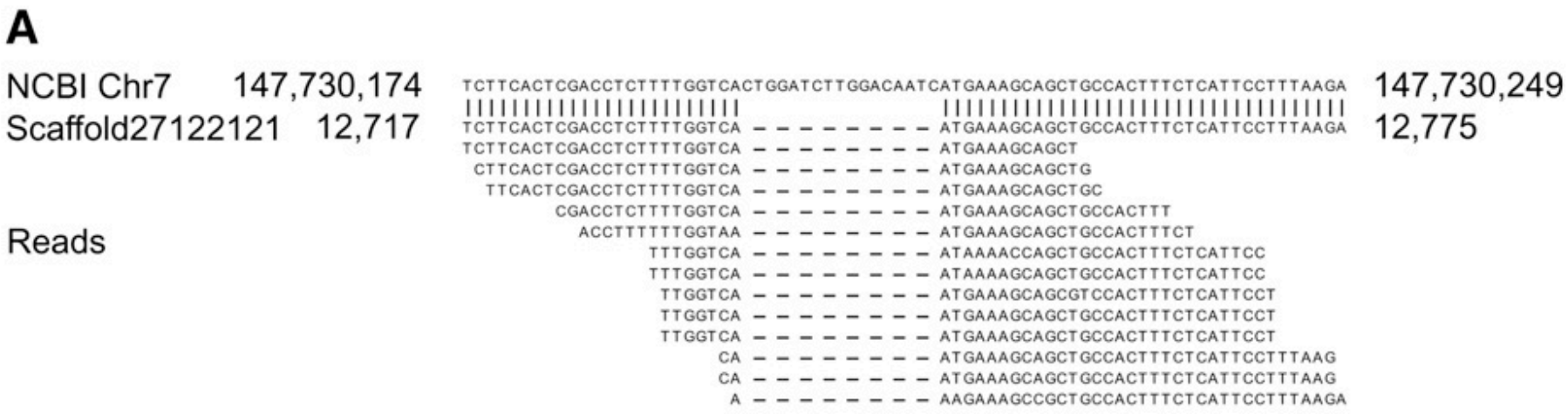


Figure 3. Examples of deletion and insertion identified in the comparison of the assembled individual human genomes and the NCBI reference genome. **(A**) A 17-bp deletion in scaffold27122121 of the African genome located on chromosome 7.

Insertion example

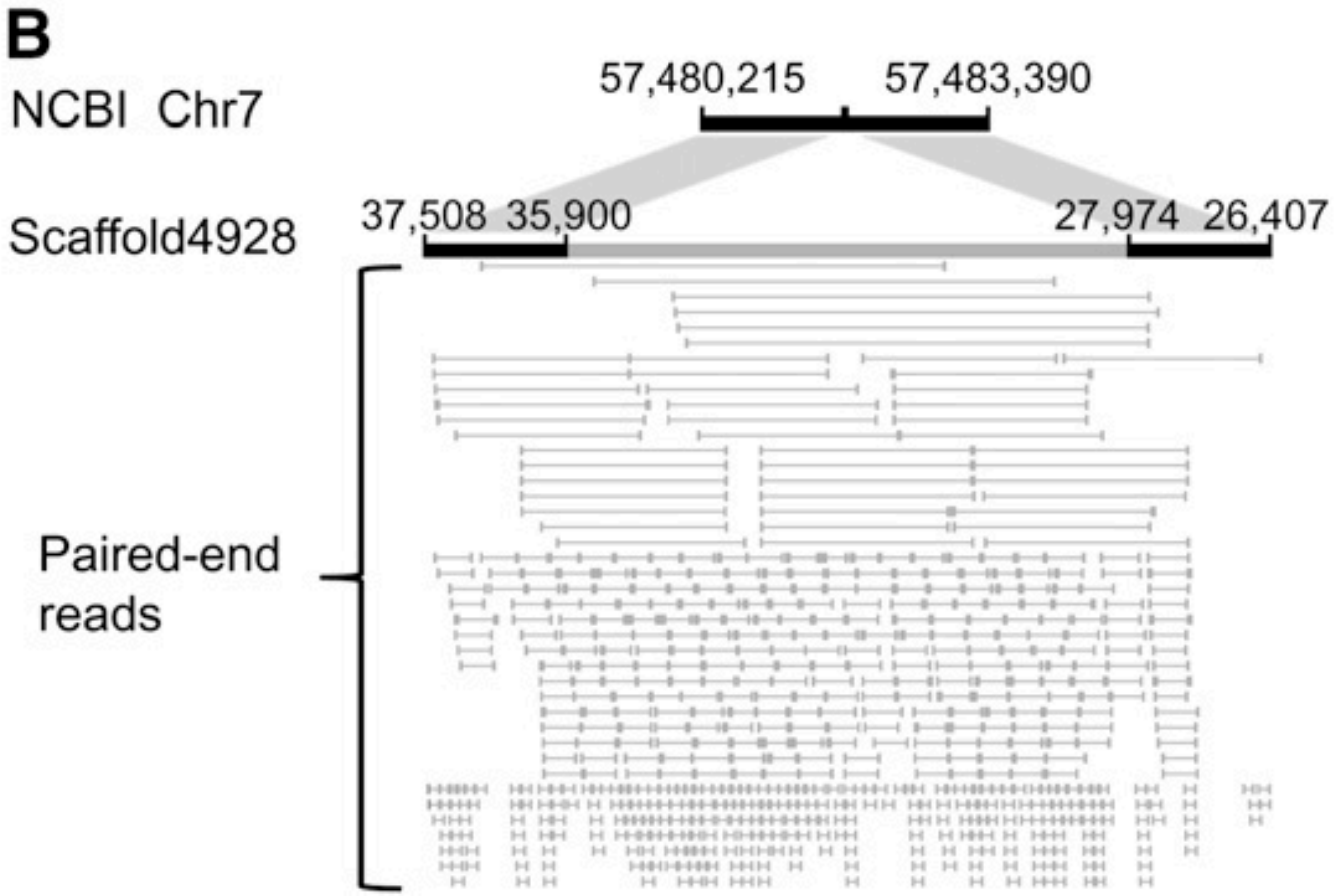


Figure 3. Examples of deletion and insertion identified in the comparison of the assembled individual human genomes and the NCBI reference genome **(B)** A 7926-bp insertion in scaffold4928 of the Asian genome located on chromosome 7. The inserted sequence fragment was validated by a human BAC clone AC153461.2 in GenBank, and also exists in the chimpanzee genome.