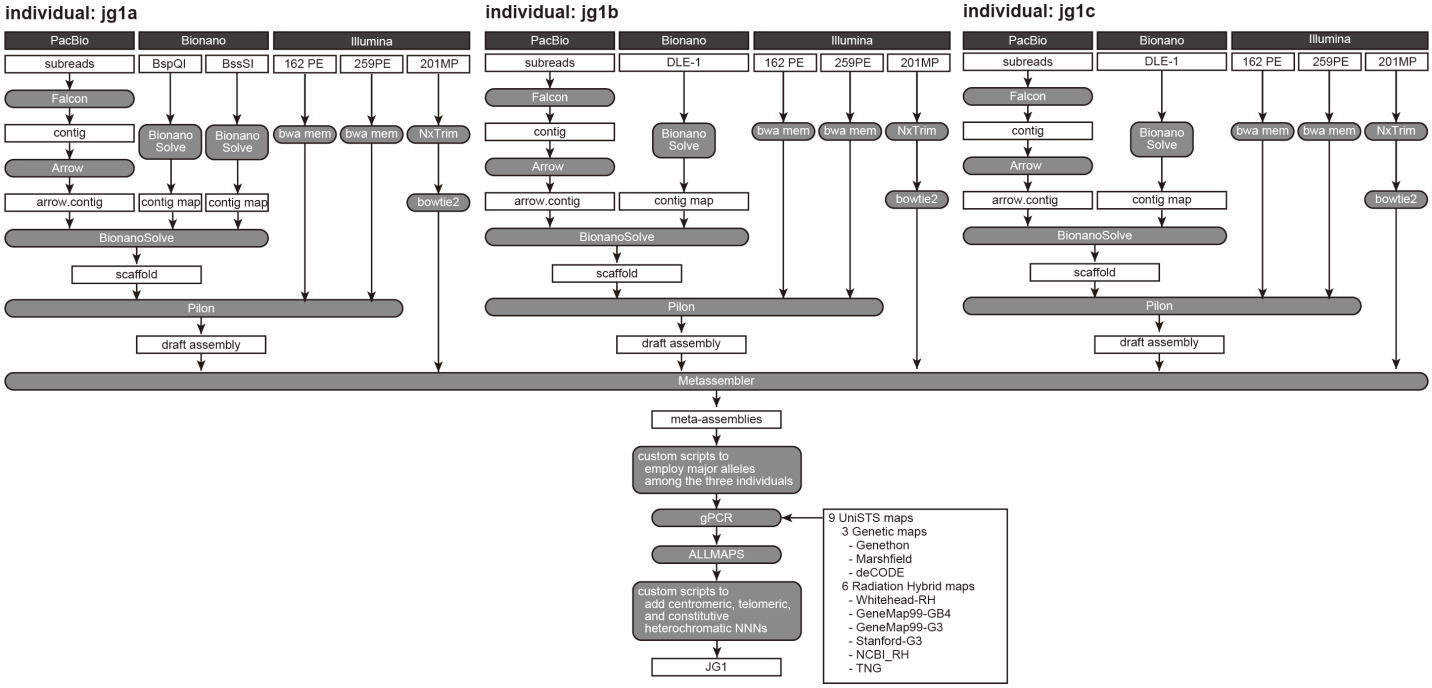
Technical Notes: Computational methods for construction of JG1

- Feb. 25. 2019.
- Jun Takayama

Overview



This note describes information for the construction of **JG1**, the Japanese reference genome sequence. **JG1** was constructed from three de novo assembled haploid genomes from three Japanese male individuals (**jg1a**, **jg1b**, and **jg1c**) through 11 steps as follows.

- **Step 1:** De novo assembly of PacBio subreads by using FALCON software.
- **Step 2:** Polishing the primary contigs of the FALCON output with the PacBio subreads by using the ArrowGrid software.
- **Step 3:** De novo assembly of Bionano molecules by using the Bionano Solve software.
- **Step 4:** Hybrid scaffolding of the polished contigs and the optical contig maps by using the Bionano Solve software
- **Step 5:** Polishing the hybrid scaffolds with Illumina short reads
- **Step 6:** Integrating and filling the gaps of the polished hybrid scaffolds of an individual by those of the other two individuals by using the Metassembler software with the aid of mate-pair short reads.
- **Step 7:** Defining and employing the major allele in the sites that segregate among the three genomes.
- **Step 8:** Ordering and aligning the orientation of the meta-assembled scaffolds by using the ALLMAPS

software.

- **Step 9:** Modify the length of consecutive Ns in the telomeric, centromeric, and constitutive heterochromatic region.
- **Step10:** Mask the putative pseudo-autosomal region 1 (PAR1) in the chromosome Y by finding the homologous region in the chromosome X.
- **Step 11:** Shift the start position of the mitochondrial genome sequence in JG1 (chrM) to that of the revised Cambridge Reference Sequence (rCRS).

Detail

Step 1: De novo assembly of PacBio subreads by using the FALCON software.

PacBio subreads were assembled by using FALCON with the following configurations:

```
length_cutoff = 9000
length_cutoff_pr = 15000
genome_size = 3200000000
pa_HPCdaligner_option = -v -dal128 -t16 -e.75 -M16 -l4800 -k18 -h480 -w8 -s100 -T1
ovlp_HPCdaligner_option = -v -dal128 -M24 -k24 -h1024 -e.96 -l2500 -s100 -T1
pa_DBsplit_option = -x500 -s400
ovlp_DBsplit_option = -s400
falcon_sense_option = \
    --ouput_multi --min_idt 0.70 --min_cov = 4 --max_n_read 200 --n_core 1
overlap_filtering_setting = --max_diff 60 --max_cov 60 --min_cov 0 --n_core 12
```

Step 2: Polishing the primary contigs from the FALCON output with the PacBio subreads by using the ArrowGrid software.

We ran the ArrowGrid shell scripts with some modifications in the following shell scripts: `arrow.sh`, `consensus.sh`, and `merge.sh`. In the `arrow.sh`, we modified the variable for the number of jobs to the smaller one between the number of contigs and the number of `bam` files in the `input.fofn` file. We also modified the `arrow.sh` to optimize qsub options to our UGE settings. We modified `consensus.sh` and `merge.sh` to deal with our modification in the `arrow.sh`, and optimized the `qsub` command options.

Step 3: De novo assembly of Bionano molecules by using the Bionano Solve software.

We obtained two optical map data (BspQI and BssSI) from the individual **jg1a**, and one optical map data (DLE-1) from **jg1b** and **jg1c**. In both cases, the Bionano optical maps were assembled in two steps: the

rough assembly step and the full assembly step to assemble the genome sequence as independently as possible from the reference, GRCh38. For the rough assembly step of individual **jg1a**, we ran the `pipelineCL.py` software with the following settings:

```
-T 128
-j 8
-f 0.2
-i 0
-b ${data_dir}/Molecules.bnx
-l ${workdir}
-t ${bindir}/Solve3.1_08232017/RefAligner/6700.6902rel/avx/
-a ${bindir}/Solve3.1_08232017/RefAligner/6700.6902rel/ \
    optArguments_nonhaplotype_irys.xml
-C ${workdir}/clusterArguments_${ver}.xml
-V 0
-A
-z
-u
-m
```

For the full assembly step of individual **jg1a**, we ran it with the following settings:

```
-T 128
-j 8
-f 0.2
-i 5
-b ${data_dir}/Molecules.bnx
-l ${workdir}
-t ${bindir}/Solve3.1_08232017/RefAligner/6700.6902rel/avx/
-a ${bindir}/Solve3.1_08232017/RefAligner/6700.6902rel/ \
    optArguments_nonhaplotype_irys.xml
-C ${workdir}/clusterArguments_${ver}.xml
-V 0
-r ${rough_assembly_output}/exp_mrg0/EXP_MRG0A.cmap
-y
-m
```

For the rough assembly step of individual **jg1b** and **jg1c**, we ran the `pipelineCL.py` software with the following settings:

```
-l ${workdir}
-t ${bindir}/Solve3.2.1_04122018/RefAligner/7437.7523rel/avx/
-C ${bionano_workdir}/clusterArgumentsBG_saphyr_phi_${ver}.xml
-b ${data_dir}/all.bnx
-f 0
-i 5
-a ${bindir}/Solve3.2.1_04122018/RefAligner/7437.7523rel/ \
    optArguments_nonhaplotype_DLE1_saphyr_human.xml
-V 0
-N 4
-R
```

For the full assembly step of individual **jg1b** and **jg1c**, we ran it with the following settings:

```
-l ${workdir}
-t ${bindir}/Solve3.2.1_04122018/RefAligner/7437.7523rel/avx/
-C ${bionano_workdir}/clusterArgumentsBG_saphyr_phi_${ver}.xml
-b ${data_dir}/all.bnx
-f 0
-i 5
-a ${bindir}/Solve3.2.1_04122018/RefAligner/7437.7523rel/ \
    optArguments_nonhaplotype_irys.xml
-V 0
-N 4
-r ${rough_assembly_output}/exp_mrg0/EXP_MRG0A.cmap
-y
```

- The `-T` and `-j` options were set variously for computational efficiency.

Step 4: Hybrid scaffolding of the long read-polished contigs and the optical contig maps by using the Bionano Solve software.

Hybrid scaffolding for **jg1a** was performed to integrate single set of primary contigs from the FALCON output and the two sets of optical contig maps (BspQI and BssSI) from the Bionano Solve output. To perform two-enzyme hybrid scaffolding, we ran the `runTGH.R` R script with the options below:

```
-N ${jg1a}-p_ctg.arrow.fa
-b1 ${bionano_workdir_BspQI}/contigs/exp_refineFinal1/EXP_REFINEFINAL1.cmap
-b2 ${bionano_workdir_BssSI}/contigs/exp_refineFinal1/EXP_REFINEFINAL1.cmap
-e1 BSPQI
-e2 BSSSI
-O ${jg1a_hybridscaffold_output_path}/${jg1a_output_file_name}
-R ${bindir}/Solve3.2.1_04122018/RefAligner/7437.7523rel/avx/RefAligner
${bindir}/Solve3.2.1_04122018/HybridScaffold/04122018/TGH/ \
    hybridScaffold_two_enzymes.xml
```

Hybrid scaffolding for **jg1b** or **jg1c** was performed to integrate single set of primary contigs from the FALCON output and the single set of optical contig maps (DLE-1) from the Bionano Solve output. To perform single-enzyme hybrid scaffolding, we ran the `hybridScaffold.pl` Perl script with the options below:

```
-n ${arrow_workdir}/${jg1b_or_jg1c}-p_ctg.arrow.fa
-b ${bionano_workdir}/contigs/exp_refineFinal1/EXP_REFINEFINAL1.cmap
-c ${hybridscaffold_workdir}/hybridScaffold_DLE1_config.tmem.xml
-r ${bindir}/Solve3.2.1_04122018/RefAligner/7437.7523rel/avx/RefAligner
-o ${workdir}
-B 2
-N 2
-f
-e ${bionano_workdir}/contigs/auto_noise/autoNoise1.errbin
```

Step 5: Polishing the hybrid scaffolds with Illumina short reads

To further reduce the potential error in the scaffolds, we utilized Illumina WGS short reads of 162-bp and 259-bp paired end (PE) reads; both have ~30X read depth. We mapped both 162-bp PE reads and 259-bp PE reads to the hybrid scaffolds by using BWA MEM software with the option `-t 22 -K 1000000`. The resulting alignment file was coordinate-sorted and compressed with Picard tools `SortSam` command. The resulting bam files from 162-bp PE and 259-bp PE reads were merged with Picard tools `MergeSamFiles` command. The merged bam files were then split to each scaffold by using Samtools `view` command to reduce the computational task for the downstream processes. Then, each scaffold was polished by using Pilon software with option `--threads 22 --diploid --changes --vcf --tracks`. The polished fasta file for each scaffold was then merged to a single multi-fasta file.

Step 6: Integrating and filling the gaps of the polished hybrid scaffolds of an individual by those of the other two individuals by using the Metassembler software with the aid of mate-pair short reads.

The three polished hybrid scaffolds were then undertaken to be integrated with Metassembler. Because Metassembler integrates the two assemblies in an asymmetric manner (that is, the order of the two arguments matters), we performed twelve meta-assemblies; that is, (a + (b + c)), (a + (c + b)), ((a + b) + c), ((a + c) + b), (b + (a + c)), (b + (c + a)), ((b + a) + c), ((b + c) + a), (c + (a + b)), (c + (b + a)), ((c + a) + b), ((c + b) + a), where “x + y” means meta-assemble x and y assemblies in this order.

For each round of meta-assembly, we aligned the two assemblies by using Nucmer software with the option `--maxmatch -c 50 -l 300`. The resulting delta file was filtered by delta-filter software with the option `-1`. The resulting delta file was converted to coords format by show-coords command with `-clrTH` option. Short mate-pair reads were classified into four categories—mp, pe, se, and unknown—by using NxTrim software, and the resulting reads with the correct mate-pair orientation from all three individuals were mapped with the bowtie2 software with `--minins 1000 --maxins 16000 --rf` option. The sam output file from bowtie2 was then processed with mateAn command with `-A 2000 -B 15000` option, meaning the range of insert length was 2 kb to 15 kb. The alignment information from nucmer and the mate-pair mapping information were integrated with asseMerge command with `-i 5 -c 6` options. Finally, the metassem output was converted to the fasta format with the meta2fasta command.

Step 7: Defining and employing the major allele in the sites that segregate among the three genomes.

The three polished hybrid scaffolds were mapped to the twelve meta-assembled genome sequences by using minimap2 and variants were called by using paftools `call` command. After normalizing the way of variant calling by using bcftools `norm` command, variants shared by two of the three genomes were detected by using bcftools `isec` command, and were regarded as major allele, and employed in the **JG1** sequence by using bcftools consensus command. For multi-allelic sites among the three genomes, one allele was randomly chosen and employed in the **JG1** sequence by using the same software pipeline as the major alleles described above.

Step 8: Ordering and aligning the orientation of the meta-assembled scaffolds by using the ALLMAPS software.

The meta-assembled scaffolds were marked with the STS markers of three genetic and six radiation-hybrid maps (the Genethon, the Marshfield, and the deCODE genetic maps; and the GeneMap99-G3, GeneMap99-GB4, TNG, NCBI_RH, Stanford-G3, Whitehead-RH maps) by using an electronic PCR simulation software, gPCR, with `-S -D` (`-S` : show amplicon sequence, `-D` : show direction of marker) options. The STS marker sequences and their map position were fetched from the UniSTS database.

The electronic PCR results were converted to bed format files and coordinates of some radiation hybrid maps were scaled to about 2,000. These maps were merged by using ALLMAPS mergebed command. Then the resulting merged bed was used for the ALLMAPS `path` command with the option `--gapsize=10000` to order, align the orientation, and link scaffolds with 10,000 Ns. The weights for the

three genetic maps were set to 5, whereas those for the six radiation hybrid maps were set to 1 in the `weights.txt` file.

Step 9: Modify the length of Ns in the telomeric, centromeric, and constitutive heterochromatic region length.

The physical length of short arms of the acrocentric chromosomes 13, 14, 15, 21, and 22 were obtained from the Table 4 of Morton et al., (1991) (ref. 22). The relative length estimates of constitutive heterochromatin regions in the chromosome 1, 9, and 16 were obtained from ref. 17 and ref. 18. The relative length estimate of heterochromatin segment of the Y chromosome was obtained from ref. 19-21. These relative lengths were converted to the absolute length (Mb) by using the physical length of chromosomal arms in the Table 4 of ref. 22.

The length of consecutive Ns for each chromosome is provided in the Table. For chromosomes 8 and 11, in which their centromere-specific sequence repeats were found in the midst of a scaffold—which were found by aligning the LinearCen1.1 sequences with minimap2 software—no centromeric Ns were inserted. For the other chromosomes, 3 Mb consecutive Ns were inserted, instead of 10 kb Ns inserted by the ALLMAPS software, between the two scaffolds flanking the centromere.

The position of centromere was inferred from the Whitehead-RH and GeneMap99-GB4 maps, in which weak radiation was used to break chromosomes and hence the centromeric or constitutive heterochromatin region could be inferred from the region sparsely covered by STS markers.

Step10: Mask the putative pseudo-autosomal region 1 in the sub-telomeric region in the chromosome Y by finding the region matching with those in the chromosome X.

To find the pseudo-autosomal regions, we aligned the chromosome Y and the chromosome X both from JG1 by using minimap2 with option `-cx asm5` and vice versa. The alignment started from the terminal region of chromosome Y and ended at the 2.26 Mb. This region was regarded as the PAR1 region. The other regions such as PAR2 and XTR was ambiguous for unknown reasons. Then we masked the putative PAR1 region by using BEDtools maskfasta command.

Table: The number of consecutive Ns in the telomeric, centromeric, and constitutive hetrochromatin regions.

chr	pter (bp)	cen (bp)	qter (bp)	reference
1	10,000	30,000,000	10,000	17, 18, 22
2	10,000	3,000,000	10,000	
3	10,000	3,000,000	10,000	
4	10,000	3,000,000	10,000	
5	10,000	3,000,000	10,000	
6	10,000	3,000,000	10,000	
7	10,000	3,000,000	10,000	
8	10,000	-	10,000	
9	10,000	30,000,000	10,000	17, 18, 22
10	10,000	3,000,000	10,000	
11	10,000	-	10,000	
12	10,000	3,000,000	10,000	
13	16,000,000	-	10,000	22
14	16,000,000	-	10,000	22
15	17,000,000	-	10,000	22
16	10,000	20,000,000	10,000	17, 18, 22
17	10,000	3,000,000	10,000	
18	10,000	3,000,000	10,000	
19	10,000	3,000,000	10,000	
20	10,000	3,000,000	10,000	
21	11,000,000	-	10,000	
22	13,000,000	-	10,000	
X	10,000	3,000,000	10,000	
Y	2,260,577	3,000,000	30,000,000	19, 20, 21, 22

Step 11: Shift the start position of the mitochondrial genome sequence in JG1 (chrM) to that of the revised Cambridge Reference Sequence (rCRS).

We aligned the twelve meta-assembled scaffolds onto the GRCh38 whose mitochondrial sequence is employed from the rCRS by using minimap2 with option `-cx asm5` to find a scaffold that corresponds to the mitochondrial genome. We found a scaffold of 16,568 bp in length that corresponds to the mitochondrial sequence, which has a shifted start site because the mitochondrial genome is circular. We cut the latter part of the sequence and put them at the first part of the sequence to match the start site.

Software Versions

- FALCON build ver. falcon-2017.11.02-16.04-py2.7-ucs2.tar.gz
- ArrowGrid GitHub commit tag: `81b03f1`
- BionanoSolve
    - v3.1 for assembly of the individual **jg1a**
    - v3.2 for assembly of the individual **jg1b**, **jg1c** and for hybrid assembly
- BWA mem v0.7.17
- Picard v2.18.4
- Samtools v1.8
- Pilon v1.22
    - with modification to fix the issue reported https://github.com/broadinstitute/pilon/issues/48
- Metassembler v1.5
    - with modification of the type of `totalBases` variable in the `CEstat.hh` from int to long to deal with large genomes
- NxTrim v0.4.3
- Bowtie2 v2.3.4.1
- Mummer v4.0.0beta2
- Minimap2 v2.12
- BCFtools v1.8
- Seqtk v1.3
- gPCR v2.6a
- ALLMAPS v0.8.12
- BEDtools v2.27.1

Resources

- UniSTS STS marker primer sequence ftp://ftp.ncbi.nih.gov/pub/ProbeDB/legacy*unists/UniSTS*human.sts
- UniSTS STS marker map position

ftp://ftp.ncbi.nih.gov/pub/ProbeDB/legacy*unists/UniSTS*MapReports/Homo_sapiens

Hardware and environment

- ToMMo Super Computer
  - overview: https://sc.megabank.tohoku.ac.jp/en/outline
  - hardware environment: https://sc.megabank.tohoku.ac.jp/en/hardware
  - software environment: https://sc.megabank.tohoku.ac.jp/en/software

Operators

- Operation was done by the following members (including supervisors; in alphabetical order)
  - Takamitsu Funayama, Chinatsu Gocho, Kengo Kinoshita, Satoshi Makino, Yasunobu Okamura, Shu Tadaka, Jun Takayama, Gen Tamiya, Kazunori Waki, Kenji Yano.

References

1. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13: 1050–1054.
2. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; Genomics (q-bio.GN). (arXiv:1303.3997v2 [q-bio.GN])
3. Picard tools available from http://broadinstitute.github.io/picard/
4. Bionano Solve available from https://bionanogenomics.com/support-page/bionano-solve/
5. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9.
6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
7. Wences AH, Schatz MC. Metassembler: Merging and optimizing de novo genome assemblies. Genome Biol. Genome Biology; 2015;16: 1–10.
8. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–9.
9. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14: 1–14.
10. Li H. Minimap2: pairwise alignment for nucleotide sequences. 2017;34: 3094–3100.
11. Koren, S, Walenz, BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. 2017;27: 722-736
12. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10: 563–569.

13. bcftools available from http://samtools.github.io/bcftools/
14. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J. ALLMAPS: Robust scaffold ordering based on multiple maps. Genome Biol. 2015;16: 3.
15. seqtk available from https://github.com/lh3/seqtk
16. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.
17. Lundgren R, Berger R, Kristoffersson U. Constitutive heterochromatin C-band polymorphism in prostatic cancer. Cancer Genet Cytogenet. 1991;51: 57-62.
18. Podugolnikova OA, Blumina MG. Heterochromatic regions on chromosomes 1, 9, 16, and Y in children with some disturbances occurring during embryo development. Hum Genet. 1983;63: 183-8.
19. Erçal MD, Bröndum-Nielsen K. Length polymorphism of heterochromatic segment of the Y chromosome in boys with acute leukemia. Acta Paediatr Jpn. 1995;37: 614-616.
20. Petković I. Variability of euchromatic and heterochromatic segment of the Y chromosome in men with malignant tumors and in a control group. Cancer Genet. Cytogenet. 1984; 13, 29-36.
21. Petković I, Nakić M, Tiefenbach A, Cepulić M, Konja J, Feminic-Kes R, Rajić L. Heterochromatic segment length of Y chromosome in 55 boys with malignant diseases. Cancer Genet. Cytogenet. 1987;25: 351-353.
22. Morton, NE. Parameters of the human genome. Proc Natl Acad Sci USA. 1991; 88: 7474–7476.
23. O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. NxTrim: Optimized trimming of Illumina mate pair reads. Bioinformatics. 2015;31: 2035–2037.
24. Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, Fang N, Hadley D, Harris M, Hussain S, Lee R, Maratukulam A, O'Connor K, Perkins S, Piercy M, Qin F, Reif T, Sanders C, She X, Sun WL, Tabar P, Voyticky S, Cowles S, Fan JB, Mader C, Quackenbush J, Myers RM, Cox DR. An STS-based radiation hybrid map of the human genome. Genome Res. 1997;7: 422-433.
25. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tomé P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJ, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Fox S, Gelling S, Green L, Harrison P, Hocking R, Holloway E, Hunt S, Keil S, Lijnzaad P, Louis-Dit-Sully C, Ma J, Mendis A, Miller J, Morissette J, Muselet D, Nusbaum HC, Peck A, Rozen S, Simon D, Slonim DK, Staples R, Stein LD, Stewart EA, Suchard MA, Thangarajah T, Vega-Czarny N, Webber C, Wu X, Hudson J, Auffray C, Nomura N, Sikela JM, Polymeropoulos MH, James MR, Lander ES, Hudson TJ, Myers RM, Cox DR, Weissenbach J, Boguski MS, Bentley DR. A physical map of 30,000 human genes. Science. 1998;282:744-746.
26. Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature. 1996;380: 152-154.
27. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet. 1998;63: 861-869.
28. Agarwala R, Applegate DL, Maglott D, Schuler GD, Schäffer AA. A fast and scalable radiation hybrid map construction and integration strategy. Genome Res. 2000;10: 350-364.
29. Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, Brady SD, Bushard JM, Bustos VI, Chu A, Chung TR, De Witte A, Denys ME, Dominguez R, Fang NY, Foster BD,

Freudenberg RW, Hadley D, Hamilton LR, Jeffrey TJ, Kelly L, Lazzeroni L, Levy MR, Lewis SC, Liu X, Lopez FJ, Louie B, Marquis JP, Martinez RA, Matsuura MK, Misherghi NS, Norton JA, Olshen A, Perkins SM, Perou AJ, Piercy C, Piercy M, Qin F, Reif T, Sheppard K, Shokoohi V, Smick GA, Sun WL, Stewart EA, Fernando J, Tejeda, Tran NM, Trejo T, Vo NT, Yan SC, Zierten DL, Zhao S, Sachidanandam R, Trask BJ, Myers RM, Cox DR. A high-resolution radiation hybrid map of the human genome draft sequence. Science. 2001;291: 1298-1302.

30. Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, Slonim DK, Baptista R, Kruglyak L, Xu SH, Hu X, Colbert AM, Rosenberg C, Reeve-Daly MP, Rozen S, Hui L, Wu X, Vestergaard C, Wilson KM, Bae JS, Maitra S, Ganiatsas S, Evans CA, DeAngelis MM, Ingalls KA, Nahf RW, Horton LT Jr, Anderson MO, Collymore AJ, Ye W, Kouyoumjian V, Zemsteva IS, Tam J, Devine R, Courtney DF, Renaud MT, Nguyen H, O'Connor TJ, Fizames C, Fauré S, Gyapay G, Dib C, Morissette J, Orlin JB, Birren BW, Goodman N, Weissenbach J, Hawkins TL, Foote S, Page DC, Lander ES. An STS-based map of the human genome. Science. 1995;270: 1945-1954.
31. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. A high-resolution recombination map of the human genome. Nat Genet. 2002;31: 241-247.
32. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, and Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res 24: 697-707