GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

ReSeqTools: an integrated toolkit for large-scale next-generation sequencing based resequencing analysis

W. He[1,2], S. Zhao[2], X. Liu[2], S. Dong[2], J. Lv[2], D. Liu[2], J. Wang[1,2] and Z. Meng[1]

[1]School of Bioscience and Bioengineering,
South, China University of Technology, Guangzhou, Guangdong, China
[2]BGI-Shenzhen, Shenzhen, China

Corresponding authors: Z. Meng / J. Wang
E-mail: zzmeng@scut.edu.cn / wangj@genomics.cn

Genet. Mol. Res. 12 (4): 6275-6283 (2013)
Received February 18, 2013
Accepted September 2, 2013
Published December 4, 2013
DOI http://dx.doi.org/10.4238/2013.December.4.15

ABSTRACT. Large-scale next-generation sequencing (NGS)-based resequencing detects sequence variations, constructs evolutionary histories, and identifies phenotype-related genotypes. However, NGS-based resequencing studies generate extraordinarily large amounts of data, making computations difficult. Effective use and analysis of these data for NGS-based resequencing studies remains a difficult task for individual researchers. Here, we introduce ReSeqTools, a full-featured toolkit for NGS (Illumina sequencing)-based resequencing analysis, which processes raw data, interprets mapping results, and identifies and annotates sequence variations. ReSeqTools provides abundant scalable functions for routine resequencing analysis in different modules to facilitate customization of the analysis pipeline. ReSeqTools is designed to use compressed data files as input or output to save storage space and facilitates faster and more computationally efficient large-scale resequencing studies in a user-friendly manner. It offers abundant practical functions and generates useful statistics during the analysis pipeline,

which significantly simplifies resequencing analysis. Its integrated algorithms and abundant sub-functions provide a solid foundation for special demands in resequencing projects. Users can combine these functions to construct their own pipelines for other purposes.

Key words: Next-generation sequencing; Resequencing; Toolkit; Sequence variation

INTRODUCTION

Next-generation sequencing (NGS) techniques have created fascinating opportunities to use sequencing to understand diverse aspects of biology. Higher efficiency, lower cost, and reduced labor required to apply NGS techniques make possible the comprehensive analysis of genomes, transcriptomes, and methylomes. In particular, whole genome resequencing becomes feasible using NGS platforms. Whole genome resequencing is widely employed to characterize genetic variations (Rubin et al., 2010), construct the evolutionary histories of populations (Xia et al., 2009), conduct genome-wide association studies (Huang et al., 2010, 2012), or to construct linkage maps (Huang et al., 2009). In particular, when large-scale resequencing becomes more affordable, resequencing will be widely used to identify the correlations between genotypes and phenotypes. The more species with reference genome sequences constructed, resequencing studies will further be applied to different species.

The total amount of available genomic data is increasing approximately 10-fold every year, a rate much faster than described by Moore's Law for computational processing power (Loh et al., 2012). This holds true for resequencing studies. Because NGS platforms such as Illumina generate data from shorter sequences with a higher error rate (Nielsen et al., 2011), the common resequencing strategy used by these platforms determines larger numbers of sequences, which results in higher depth and coverage to avoid errors caused by mapping short reads as well as technical errors.

In such studies, the individual sequencing output should be high and population sequencing should be even higher, thus generating extremely large volumes of data that create difficult challenges for computational speed and data storage. Therefore, the toolkits/pipeline used to manipulate and analyze the resequencing data should be not only accurate and stable, but also efficient for processing large datasets. They should also compute compressed genomic data to keep pace with data generation. For example, as large resequencing datasets have been generated in the 1000 Human Genomes Project (Abecasis et al., 2010) and the 1001 Genomes Project for *Arabidopsis thaliana* (Cao et al., 2011), customized data analysis of these datasets also requires an appropriate and efficient toolkit/pipeline.

For NGS-based resequencing analysis, particularly for the widely applied Illumina platform, several steps are required. First, the raw reads should be filtered. Because Illumina sequencing uses linked adaptors and PCR amplification before sequencing, reads contaminated with adaptor sequences must be eliminated as well as reads that are duplicated during PCR. Moreover, low quality sequences should also be removed. Available software must process raw data using custom scripts that immediately generate large amounts of data.

In ReSeqTools, we package these functions into the toolkit, and incorporate them in modules. To conserve space on storage media and volatile computer memory, this toolkit directly

uses compressed files as input. After filtering, the reads should be mapped to a reference genome. Numerous tools differing in accuracy and efficiency have been developed to map massive numbers of short reads to the reference genome (Nielsen et al., 2011). Among them, SOAP (Li et al., 2008) and BWA (Li and Durbin, 2009) are widely used (Wang et al., 2008; Xia et al., 2009; Lam et al., 2010; Xu et al., 2012). SOAP maps more efficiently with lower demand for CPU time and memory, particularly in the latest version (SOAP3), which uses graphics processing unit computing to further improve efficiency, although BWA is considered faster, particularly when mapping reads that might be difficult for other software. After mapping, the variations will be detected. For BWA-based mapping results, which are either in the 'SAM' or 'BAM' format, SAMtools (Li et al., 2009a) is available for variant calling and performing downstream analyses. However, for SOAP mapping results, there are only a few independent software packages available for further downstream analysis, and they lack many of the features required for resequencing Our toolkit also provides tools that can be used as alternatives to SAMtools for generating SAM-format mapping results. ReSeqTools is based on the results of SOAP and BWA, which is the two main short-read alignment algorithms, and can be used to detect high quality genetic variations, including single nucleotide polymorphisms (SNPs), small insertion and deletions (InDels), structure variations (SVs), and copy number variations (CNVs).

MATERIAL AND METHODS

Overview of the toolkit

Eight major modules with different functions were designed for ReSeqTools (Table 1) and were organized into a three-layer analysis architecture (Figure 1), which is consistent with processing resequencing data. Fatools, Fqtools, and Filetools were designed for processing raw data and input files that include raw sequencing results and reference genome files, respectively. Fatools provides functions to handle FASTA files of the reference genome as follows: a) summarizes length, proportion of N, GC content, lower case (potential repeats) content; b) finds N regions, extracts regional sequences; c) randomly joins original scaffolds into pseudochromosomes; d) filters short or N-rich sequences, e) delves into sub-files to acquire coding and protein sequences (gff file required); f) sorts files according to sequence ID or length; and g) modifies sequences requiring inversion or complementation.

Table 1. Modules and their functions in ReSeqTools.

Modules	Input format	Introduction	Functions included
Fatools	FQSTA	Provide functions to deal with FASTA file	Stat/findN/extract/regenerate Filter/cut/get/reform/sort/ et al.
Fqtools	FASTQ	Provide functions to deal with fastq file	Stat/filter/rmAdapter/pooling
SOAPtools	SOAP	Dealing with the 'SOAP' format	Split/msort/merge/rmdup View/stat/filter/genotype et al.
Xamtools	SAM BAM	Dealing with the BWA result 'SAM/BAM' format	Split/msort/merge/rmdup View/stat/filter/genotype/ et al.
CNStools	CNS GLF	Dealing with the SOAPsnp (Li et al., 2009c) result CNS/GLF format	Extract/Filter/Addcn/GLF2GPF GLF2Geno/AddRef/Fre/ et al.
Vartools	SOAP	Detect the variation Based SOAP result	SOAPsnp/SOAPindel/ SOAPsv/SOAPcnv/GLFmulti
Formtools	-	Format conversion of regular specified file	Soap2Sam/Xam2Soap/Alg2Fq Maq2Sam/ Fq2Fa/CDS2Pep
Gfftools	GFF	Annotation of the variations	AnoVar/VarType/GenePoly/getCdsPep

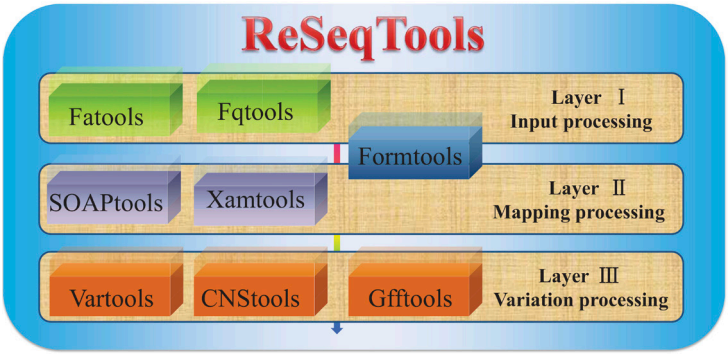

Figure 1. Modules and their organization in ReSeqTools. All the modules in ReSeqTools are organized in three layers, including input data processing, mapping result processing and variation processing.

Fqtools was developed for processing the FASTQ sequence file and performs the functions as follows: a) summarizes the quality and amount of data as well as the GC content; b) filters or trims the reads according to sequencing quality; c) removes reads contaminated with adapter sequences; and d) splits reads according to the index sequence. SOAPtools and Xamtools can be used to process the mapping results and to generate proper input files for the next step of variation calling. SOAPtools can process the SOAP format mapping results, and it includes the functions as follows: a) splits the mapping result according to chromosome; b) sorts the mapping results according to mapping location [msort (Guo et al., 2012) was integrated and modified to allow use of compressed gzip files]; c) displays the mapping results; d) merges sorted mapping results; e) filters the mapping results according to different features; f) counts the depth and genome coverage according to the mapping results; g) removes PCR duplication from the mapping results; h) calls genotype from the mapping results using Bayes's theorem to with the range of prior probability.

Xamtools provides the same functions as SOAPtools but it takes input files in SAM or BAM formats. Formtools was developed to change file format. For example, it can be used to convert between SOAP and SAM/BAM formats. The last bundle of functions provided in ReSeqTools is for variation calling and analysis. Vartools integrates different SOAP mapping result-based variation-calling software, including SOAPsnp (Li et al., 2009c), SOAPindel (Xia et al., 2009), SOAPsv (Wang et al., 2008), SOAPcnv, which was developed to analyze CNVs according to depth distribution, and GLFmulti (Xia et al., 2009). CNStools provides functions to process SOAPsnp result files (in cns or glf format) to filter SNPs, and to acquire genotype information. Finally, Gfftools can be used to annotate the variations according to provided reference gene annotations.

Using the toolkit to customize pipelines

Because comprehensive functions are included in the ReSeqTools package, the resequencing analysis pipeline can be easily constructed by integrating different modules. The current version of ReSeqTools focuses on various aspects of resequencing analysis. Here, we provide examples for one pipeline of variation (SNPs, InDels, SVs and CNVs) detection and a pipeline to identify regions and genes of interest. These examples demonstrate the feasibility of designing customized pipelines using ReSeqTools.

Pipeline for detecting variation

Variations including SNPs, InDels, SVs and CNVs are the major focus of resequencing studies. Resequencing based on NGS short reads is a proven and powerful strategy for detecting these variations (Xia et al., 2009; Lam et al., 2010; Xu et al., 2012).

Individual SNP calling can be conducted using SOAPsnp (Li et al., 2009c), and population SNP calling can be accomplished using GLFmulti (Xia et al., 2009), which integrates the likelihood of genotypes of each individual to produce a pseudo-genome for each site using maximum likelihood estimation. The final SNPs are determined according to their likelihood after filtering and assigning a genotype for each individual (Xia et al., 2009; Lam et al., 2010).

InDels, SVs and CNVs also play important roles in individuals. Thus, detection of those variations would help understand the variation content. SOAPindel, SOAPsv, and SOAPcnv can be used to identify these variations using the ReSeqTools Vartools module. SOAPindel detects small insertion/deletion events based on the gap size used for mapping. SOAPsv detects structure variations according to the insert size of paired end reads. And SOAPcnv detects copy number variations according to the distribution of the depth.

Using these tools and the basic data processing modules in ReSeqTools, the variations-detecting pipeline can be organized into three main steps as follows: 1. preprocessing raw data (Fqtools, Fatools); 2. mapping short reads to the reference genome and preprocessing the mapping result (SOAPtools, Xamtools); and 3. detecting variation and annotation (Varools and GFFtools). One advantage of our analysis is that objective data files can be stored compressed through the entire procedure. More details about the pipeline process are shown in Figure 2 and **Supplementary File 2**.

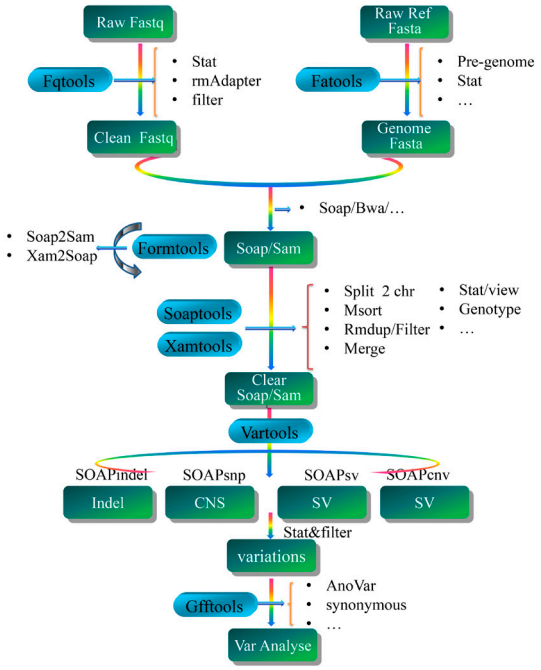

Figure 2. Pipeline for variation detection. An example of routine variation detection pipeline using ReSeqTools was illustrated, including raw data and reference file processing, mapping, and variation calling.

Pipeline for detecting genes under selection

Artificial selection shapes the phenotypes of different individuals and is particularly important for development of agricultural traits in domesticated organisms. A straightforward way to detect genes under selection is to analyze and compare the genetic polymorphisms between domesticated and wild populations (Xia et al., 2009; Lam et al., 2010). Population genetic parameters [θ_π, θ_w and Tajima's D (Tajima, 1983, 1989)] in sliding windows can be applied to identify regions under selection, and the Gfftools module in ReSeqTools calculates such statistics. By comparing these parameters between wild and domesticated populations, genes under artificial selection in domesticated population can be inferred (Lam et al., 2010; Xu et al., 2012). The design of the pipeline for detecting genes under selection in two populations is described in Figure 3.

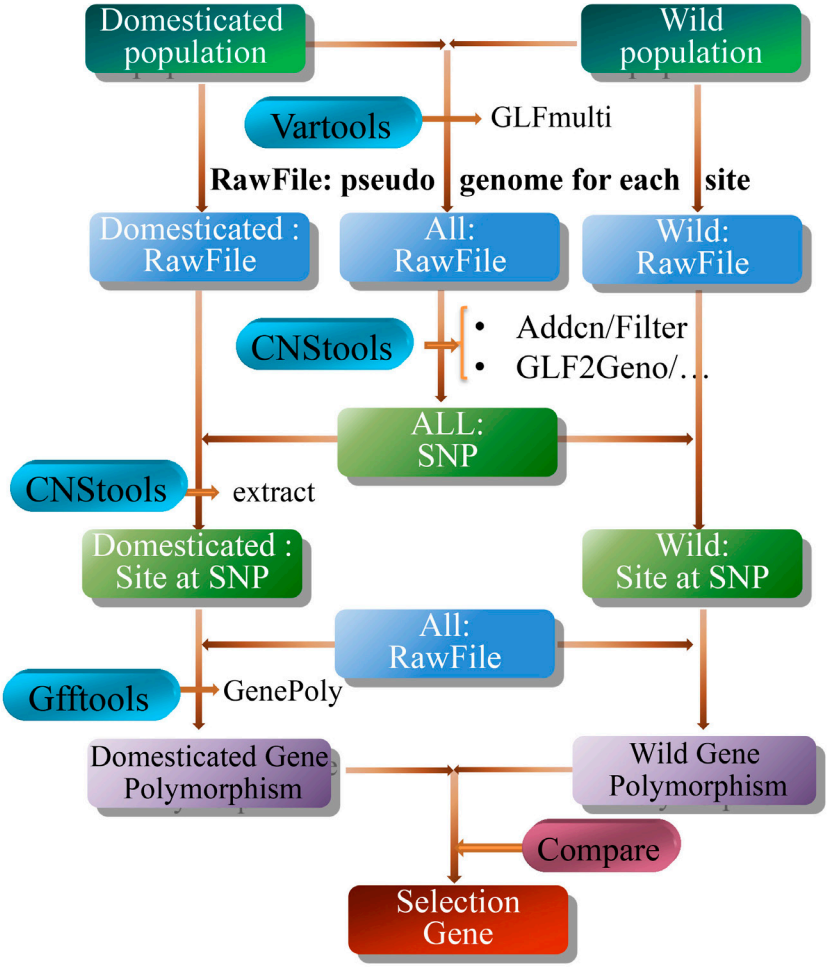

Figure 3. Pipeline for detecting putative genes under selection. Pipeline of identifying genes under selection using ReSeqTools by comparing polymorphisms in two populations.

Developing customized pipelines

ReSeqTools toolkit provides the opportunity to define custom pipelines to accommodate specific data processing other than resequencing analysis. The ReSeqTools framework is available through a distributed source-code repository, which makes it easy for developers to create derived modules and applications. Users can easily modify code and design new functions according to their own needs. Users can construct pipelines for themselves by combining different modules or functions in ReSeqTools.

RESULTS

Evaluation of the toolkit

Example pipeline to detect variation

In this section we applied this pipeline for detecting SNPs on data from the soybean population resequencing project (Lam et al., 2010). Here, we briefly introduce the key procedures shown below, and users can find the detailed procedures and results in resequencing procedures part of the **Supplementary File 1**. The pipeline script can be found in **Supplementary File 2**.

1. Use Fatools to preprocess the soybean genome *Williams 82* (Schmutz et al., 2010).
2. Use Fqtools to filter the raw fqdata to get the clean fqdata.
3. Use SOAP2 (Li et al., 2009b) to map clean fqdata to the soybean genome.
4. Use SOAPtools to preprocess the mapping results.
5. Use Vartools [SOAPsnp (Li et al., 2009c)] to calculate the likelihood of genotypes of each individual.
6. Use CNStools to detect population SNPs.
7. Use Gfftools to annotate population SNPs.

Using this pipeline, we can easily compute the data for the production statistics of the 31 soybeans (**Table S1**). Approximately 8.4% of the raw data was filtered. For each individual, the average mapping depth is approximately 5x and the mapped reads covered >90% of the genome with approximately 93.1% reads mapped (**Table S2**).

The CNStools analysis identified 7,024,922 SNPs and assigned genotypes at those SNP positions to each individual in the population. We also used use Gfftools to annotate these SNPs (**Tables S3** and **S4**).

Comparison with GATK and SAMtools pipeline for SNP calling

We compared ReSeqTools with GATK and SAMtools, two classical resequencing data processing tools, for SNP calling to evaluate the capability of ReSeqTools to deal with large amounts of data in terms of physical space and CPU cost. The input data included 30x-simulated raw reads (Fq format) and 200,000 simulated SNPs in the *A. thaliana* genome. We applied all of the three programs to detect SNPs. More details are shown in **Supplementary File 3**.

From the results presented in Table 2, it is clear that ReSeqTools significantly reduced computational and memory cost for obtaining similar performance compared with the other

two programs. ReSeqTools provides functions that are more detailed for processing data compared with SAMtools and GATK, such as computing mapping coverage and mean depth. Moreover, ReSeqTools uses a more stringent filtering standard to call SNPs, which yields higher specificity at the cost of sensitivity.

Table 2. Performance of ReSeqTools, GATK and SAMtools for SNP calling.

		Peak RAM/mean RAM	Physical space	CPU time	Accuracy SNPs/error SNP
SOAP	ReSeqTools	~3G/0.6G	~21G	~4.2h	188,460/1818
BWA	SAMtools	~3G/1.0G	~28G	~4.1h	171,028/2895
	GATK	~3G/1.0G	~33G	~5.3h	194,010/2647

DISCUSSION

ReSeqTools integrates comprehensive functions for large scale NGS-based resequencing analysis. Most functions in ReSeqTools are practical and efficient according to our tests using a published sequencing dataset. We also use a multithread method to minimize the processing time required by some functions. Considering the large amount of data produced by the NGS platform, all the functions in ReSeqTools can directly use compressed files as input, and the output file can be compressed if a large dataset will be generated.

Using this toolkit, users can perform routine resequencing analysis. Through the classic resequencing pipeline of detecting variation, users can deal with raw data, map the reads to the genome, interpret the mapping results, identify, and annotate the variations.

ReSeqTools requires more physical space and computing resources to detect SNPs in a population, similar to the GATK (McKenna et al., 2010) pipeline. However, ReSeqTools is realizing a new approach for detecting SNPs, which directly reads the mapping result thus saves physical space and computing resources. Moreover, the design of the algorithm and data limitation makes it difficult to accurately detect InDels, SVs and CNVs. We hope to develop new models to improve these variation-calling functions in the future.

Because this toolkit runs on the Linux platform, the interface requires a user who is familiar with Linux. Therefore, to make ReSeqTools convenient and attractive, we provide documentation and a website (http://code.google.com/p/reseqtools/) to enable researchers to perform most kinds of routine analyses. We plan to implement a web-based system for using RefSeqTools and to solicit suggestions from users for improvements.

ACKNOWLEDGMENTS

We thank the ReSeqTools developers and community members that have contributed to this project. Zhe Su, Jinjin Wang, Bing He, JunLi and ShuaiShuai Tai provided their requirements and helped to test the toolkit. Zhengqin Rong and Zhe Su offered technological suggestions. Research supported by the Education Department of BGI-Shenzhen and the Innovative Program for Undergraduate Students.

Supplementary material

REFERENCES

Abecasis GR, Altshuler D, Auton A, Brooks LD, et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-1073.

Cao J, Schneeberger K, Ossowski S, Gunther T, et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* 43: 956-963.

Guo S, Zhang J, Sun H, Salse J, et al. (2012). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* 45: 51-58.

Huang X, Feng Q, Qian Q, Zhao Q, et al. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19: 1068-1076.

Huang X, Wei X, Sang T, Zhao Q, et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42: 961-967.

Huang X, Zhao Y, Wei X, Li C, et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44: 32-39.

Lam HM, Xu X, Liu X, Chen W, et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42: 1053-1059.

Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, et al. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.

Li R, Li Y, Kristiansen K and Wang J (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics* 24: 713-714.

Li R, Yu C, Li Y, Lam TW, et al. (2009b). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966-1967.

Li R, Li Y, Fang X, Yang H, et al. (2009c). SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 19: 1124-1132.

Loh PR, Baym M and Berger B (2012). Compressive genomics. *Nat. Biotechnol.* 30: 627-630.

McKenna A, Hanna M, Banks E, Sivachenko A, et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297-1303.

Nielsen R, Paul JS, Albrechtsen A and Song YS (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12: 443-451.

Rubin CJ, Zody MC, Eriksson J, Meadows JR, et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464: 587-591.

Schmutz J, Cannon SB, Schlueter J, Ma J, et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178-183.

Tajima F (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-460.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.

Wang J, Wang W, Li R, Li Y, et al. (2008). The diploid genome sequence of an Asian individual. *Nature* 456: 60-65.

Xia Q, Guo Y, Zhang Z, Li D, et al. (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* 326: 433-436.

Xu X, Liu X, Ge S, Jensen JD, et al. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30: 105-111.