REVIEW PAPER

A beginners guide to SNP calling from high-throughput DNA-sequencing data

André Altmann · Peter Weber · Daniel Bader · Michael Preuß · Elisabeth B. Binder · Bertram Müller-Myhsok

Received: 31 January 2012 / Accepted: 31 July 2012 / Published online: 11 August 2012

Abstract High-throughput DNA sequencing (HTS) is of increasing importance in the life sciences. One of its most prominent applications is the sequencing of whole genomes or targeted regions of the genome such as all exonic regions (i.e., the exome). Here, the objective is the identification of genetic variants such as single nucleotide polymorphisms (SNPs). The extraction of SNPs from the raw genetic sequences involves many processing steps and the application of a diverse set of tools. We review the essential building blocks for a pipeline that calls SNPs from raw HTS data. The pipeline includes quality control, mapping of short reads to the reference genome, visualization and post-processing of the alignment including base quality recalibration. The final steps of the pipeline include the SNP calling procedure along with filtering of SNP candidates. The steps of this pipeline are accompanied by an analysis of a publicly available whole-exome sequencing dataset. To this end, we employ several alignment programs and SNP calling routines for highlighting the fact that the choice of the tools significantly affects the final results.

Introduction

The initial sequencing of the entire human genome with its first draft published in 2001 was an effort that could only be accomplished by large research consortia, and still required a decade of time and large financial resources (Consortium 2004; Lander et al. 2001; Venter et al. 2001). The resulting blueprint of the human genome facilitated a number of follow-up technologies such as (in their current forms) genome-wide association studies and genome-wide gene expression profiling using micro arrays. These technologies enable us to investigate the molecular biology underlying diseases and other hereditary traits. The latest technological advancement along this line, namely next generation of sequencing (NGS), allows to routinely sequence and re-sequence the whole genome of single individuals in a single laboratory within a couple of weeks and at comparably low cost. Feasibility aspects not only include the essential sequencing power but also the required computational capacities along with the necessary bioinformatics tools for evaluating raw genetic data. NGS is also referred to as high-throughput DNA sequencing (HTS), a more general term which we will use through-out the manuscript as it also includes future generations of sequencing technologies.

The sequence data for the human genome project were produced using the traditional capillary-based Sanger sequencing technology generating readouts of 500–1,000

Electronic supplementary material The online version of this article (doi:10.1007/s00439-012-1213-z) contains supplementary material, which is available to authorized users.

A. Altmann · B. Müller-Myhsok
Statistical Genetics, Max Planck Institute of Psychiatry,
Kraepelinstrasse 2-10, 80804 Munich, Germany

Present Address:
A. Altmann (✉)
Functional Imaging in Neuropsychiatric Disorders Laboratory,
Stanford University School of Medicine, 780 Welch Road,
Suite 105, Palo Alto, CA 94304, USA
e-mail: altmann@stanford.edu

P. Weber · D. Bader · E. B. Binder
Molecular Genetics of Affective Disorder,
Max Planck Institute of Psychiatry, Munich, Germany

M. Preuß
Genetic Epidemiology, Institut für Medizinische Biometrie und
Statistik, University of Lübeck, Lübeck, Germany

nucleotides. The workflow of today's most widely applied HTS platforms entails the fragmentation of the DNA to be sequenced into smaller segments. The nucleotide sequence of these fragments is then determined either by synthesis or ligation. Here, the employed nucleotides are modified such that a light signal indicating the identity of the base is emitted upon their integration; this is achieved, for e.g., by chemoluminescence or fluorescence. An image-capturing device records the light signals produced along the growing second strand. The synthesis or ligation at a single strand of DNA is, however, not sufficient for emitting sufficiently strong light signal that can be recorded. As a consequence, the single fragments are first amplified and fixed on a medium to form colonies prior to the synthesis or ligation step. Light signals emitted by these colonies can now be recorded by image capturing devices and analyzed for determining the nucleotide sequence. In principle, the available platforms differ in the way the colonies are formed and amplified and also how the nucleotide sequence is determined in the end (see Metzker 2010 for a review on HTS technologies). The high throughput is achieved by sequencing millions of these colonies in parallel. Compared to traditional Sanger sequencing, the readouts, often simply referred to as "reads", generated by the platforms are considerably shorter [e.g., 75 nucleotides for the SOLiD platform (Shendure et al. 2005), 150 nucleotides for the Illumina platform (Fedurco et al. 2006), 500 nucleotides for 454 pyrosequencing platform (Margulies et al. 2005)] and contain more sequencing errors. Moreover, each platform introduces sequencing errors that are characteristic for its sequencing workflow. Hence, compared to Sanger sequencing, HTS produces much more sequences, but of much shorter length and inferior quality; this has a tremendous impact on how the resulting readouts have to be processed in a downstream analysis.

The technology of HTS is subject to an ongoing development and the widely applied technologies of the current generation are about to be replaced by more modern approaches (the "next–next generation") aiming at eliminating some of the current technical problems. Modern methods move away from optical systems and towards systems relying, for instance, on nanotechnology (Clarke et al. 2009), semiconductors (Rothberg et al. 2011), and microscopy (Tanaka and Kawai 2009). See Schadt et al. (2010) for a review on novel HTS technologies.

For several applications (see below), the short length of the DNA sequence imposes computational challenges for the detection of specific sequence variations, such as longer insertions and deletions as well as inversions. The limitation of the short read length can be circumvented using protocols that allow the generation of read pairs with a known distance between these pairs, typically referred to as *insert length* and a known orientation with respect to the reference sequence. Depending on the protocol used, these pairs are referred to as paired-end sequencing or mate-pairs.

Having a tool for sequencing massive amounts of DNA enables us to investigate almost any question that is associated with the genetic sequence. First, it allows us to determine the nucleotide sequence of a target region [e.g., all exonic regions or the whole exome (Yi et al. 2010)] or the complete genome (Wang et al. 2008) and to identify known as well as novel single nucleotide polymorphisms (SNPs) in the sequenced region. Furthermore, paired reads facilitate the investigation of larger structural variants such as inversions, deletions, and insertions (see Xi et al. 2010 for a review). Moreover, by converting mRNA into cDNA, we have the possibility to examine gene expression and identify novel transcripts, splice variants, and to quantify expression levels of even lowly expressed genes, e.g. (Wang et al. 2009). Additional possibilities are the determination of RNA secondary structure (Kertesz et al. 2010), de novo assembly of genomes and transcriptomes (Robertson et al. 2010) as well as DNA sequences binding to specific proteins, such as histones or transcription binding factors [CHiP-Seq (Barski et al. 2007)]. Hence, high throughput sequencing platforms can be regarded as the Swiss pocketknife of molecular biology; here, we refer to the flexibility of such a sequencing platform rather than to its actual size. Each of the aforementioned applications requires a cleverly designed laboratory procedure to phrase the biological question into a problem that can be solved by sequencing DNA. In addition, efficient bioinformatics algorithms are necessary for analyzing the generated genetic data and to answer the biological question.

In the following sections, we provide a general outline of a bioinformatics pipeline that analyzes HTS data. In this manuscript, we focus on the determination of SNPs, an application that is closely related to traditional array-based genome-wide association studies. The initial steps of the analysis are in essence the building block for the analysis of any HTS data. We will not concentrate on the probabilistic background of the SNP calling methods, which is already covered in detail in the review by Nielsen et al. (2011), instead we aim at providing a hands-on guide to calling SNPs from HTS data. To this end, we will illustrate the data processing along the pipeline with a whole-exome sequencing dataset obtained from the 1,000 genomes project (Consortium 2010) (sampleID: NA12287; runID: ERR034546; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12287/sequence_read/). The dataset was generated using the Illumina platform (Illumina HiSeq 2000) by Beijing Genomics Institute and comprises 60,244,693 paired-end reads with a read length of 90 bases. For the enrichment of the exonic regions prior to sequencing, the NimbleGene SeqCap EZ Exome v2.0 (http://www.niblegene.com) assay was used. For a review on exon

Table 1 List of selected software tools for building a SNP calling pipeline

Name	Description	Link
Ibis (Kircher et al. 2009)	Base calling; Illumina	http://bioinf.eva.mpg.de/Ibis/
naiveBayesCall (Kao and Song 2011)	Base calling; Illumina	http://bayescall.sourceforge.net
Pyrobayes (Quinlan et al. 2008)	Base calling; 454	http://bioinformatics.bc.edu/marthlab/PyroBayes
Rsolid (Wu et al. 2010)	Base calling; SOLiD	http://rafalab.jhsph.edu/Rsolid
FastQC	Quality control, all platforms, graphical user interface, excellent tool for beginners	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
PRINSEQ (Schmieder and Edwards 2011)	Read trimming	http://prinseq.sourceforge.net/
SolexaQA (Cox et al. 2010)	Quality control, mainly Illumina, few basic options visualizing Phred quality scores	http://solexaqa.sourceforge.net
BFAST (Homer et al. 2009)	Mapping hash-based, supports color reads	http://bfast.sourceforge.net/
Bowtie (Langmead et al. 2009)	Mapping using BWT transform, version 2.0 does currently not support SOLiD color reads, older versions do not support gapped alignment, fastest read mapper	http://bowtie-bio.sourceforge.net
BWA (Li and Durbin 2009)	Mapping using BWT transform, support for all platforms, supports gapped alignment	http://bio-bwa.sourceforge.net
MAQ (Li et al. 2008)	Mapping hash-based, SNP calling	http://maq.sourceforge.net
SHRiMP2 (David et al. 2011)	Mapping hash-based, supports color reads	http://compbio.cs.toronto.edu/shrimp/
SOAP (Li et al. 2009b)	Mapping hash-based, quality recalibration, SNP calling and more	http://soap.genomics.org.cn
Stampy (Lunter and Goodson 2011)	Mapping hash-based with optional speedup using BWA, no support for color reads, highlight: combines the speed-up from BWT and the sensitivity from hash-based aligners	http://www.well.ox.ac.uk/project-stampy
Picard	Manipulation of SAM files, viewing, sorting, filtering, duplicate removal, generates statistics and more., more options and statistics available than in SAMtools	http://picard.sourceforge.net
SAMtools (Li et al. 2009a)	Manipulation of SAM files, viewing, sorting, filtering, SNP calling and more. Likely the fastest tool form SAM file manipulation	http://samtools.sourceforge.net
SMRA (Homer and Nelson 2010)	Local re-alignment in color space	http://srma.sourceforge.net/
GATK (McKenna et al. 2010)	Quality recalibration, local realignment, SNP calling, SNP filtering and much more. Extremely powerful tool for post-processing alignments and calling SNPs, requires in-depth study of the manual	http://www.broadinstitute.org/gatk/
Beagle (Browning and Yu 2009)	Software package for the analysis of large scale genetic data sets, supports, e.g., genotype calling, imputation, phasing	http://faculty.washington.edu/browning/beagle/beagle.html
VCFtools (Danecek et al. 2011)	Manipulate VCF files, generate statistics on SNPs (e.g., Ts/Tv ratio)	http://vcftools.sourceforge.net
ANNOVAR (Wang et al. 2010)	Annotation and filtering of variants, command line, easy to used and fast	http://www.openbioinformatics.org/annovar
SequenceVariantAnalyzer (Ge et al. 2011)	Annotation of variants, graphical user interface	http://www.svaproject.org/
IGV (Robinson et al. 2011)	Visualization of alignments, graphical user interface	http://www.broadinstitute.org/igv/
GenomeView (Abeel et al. 2012)	Visualization of alignments, graphical user interface	http://genomeview.org/
SAVANT (Fiume et al. 2010)	Visualization of alignments, graphical user interface, extendable by user-contributed modules	http://genomesavant.com/
Galaxy (Goecks et al. 2010)	Workflow automation: web-based platform for data intensive biomedical research	http://galaxy.psu.edu/
SunGridEngine (Gentzsch 2001)	Workflow automation: Open source (until version 6.1) batch-queuing system, now commercial as Oracle Grid Engine; an open source version is maintained under the name Open Grid Scheduler	http://gridscheduler.sourceforge.net

enrichment methods and their advantages, we refer to Teer and Mullikin (2010).

Table 1 provides a list of tools used or referred to during the analysis along with a link to the corresponding web resource. Of note, we restricted our selection to some of the most popular and widely used tools. Table S1 in the supplementary material provides the options for the tools used during the analysis.

The SNP calling pipeline

The SNP calling pipeline comprises seven steps that are also visualized in a flow chart in Fig. 1. The first step of the pipeline is termed base calling and evaluates the images taken during the sequencing process and generates the short reads. This is followed by an initial quality control of the generated reads succeeded by the alignment of the reads to a reference sequence and a post-processing of the alignment. These steps are shared by nearly all HTS applications. The remaining three steps are more specific to the SNP calling pipeline, namely quality score recalibration, SNP calling and filtering of SNP candidates.

Step 0: base calling

As mentioned above, an image-capturing device records the light signals generated by the synthesis or ligation processes at the newly generated strands. After acquisition of the image data, these recorded signals have to be converted into nucleotide bases. In case of the SOLiD platform, the light signals encode neighboring dinucleotides simply referred to as colors. In order to distinguish between nucleotide sequences and colors, the terms *base space* and *color space* are utilized. Furthermore, statistical models provide a measure of certainty of each base call in addition to the nucleotide itself. These statistical models base their error estimate on information such as signal intensities from the recorded image, the number of the sequencing cycle and distances to other sequence colonies. These certainties are usually expressed as Phred-like quality scores, i.e., the decadic logarithm of the expected error probability of the base call:

$$Q_{\text{phred}} = -10 \times \log_{10} P(\text{error})$$

Using this formula, an error probably of 5 % translates into a Phred score of about 13. This step is termed as *base calling* and is usually automatically performed by the sequencing platform itself. Again, each sequencing platform has to solve challenges unique to the underlying sequencing methodology. Thus, the base calling step is specialized for each platform. In the recent past, however, several algorithms have been proposed, which report improvements in sequence quality with respect to the manufacturer's algorithms [Illumina (Kao and Song 2011; Kircher et al. 2009), Roche 454 (Quinlan et al. 2008), SOLiD (Wu et al. 2010)]. Unfortunately, altering or improving the manufacturer's base calling algorithm requires in-depth knowledge of the standard sequencing and analysis workflow of the platform as well as its manipulation, which is not easily done. Thus, most users still rely entirely on the base calling algorithms provided by the sequencing platforms.

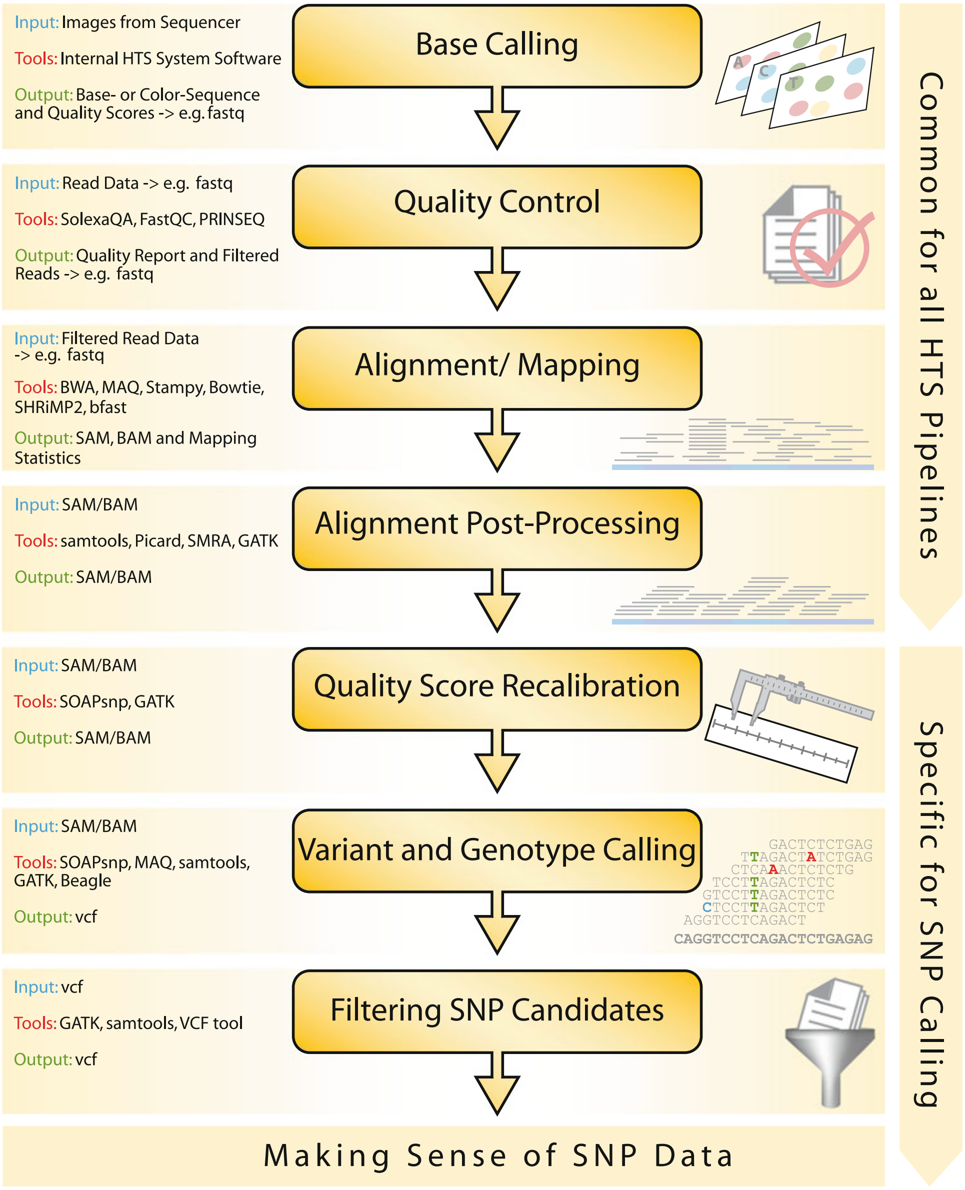
Step 1: quality control

Most platforms provide the read data directly in a flat file format such as FASTQ (Cock et al. 2009) or at least provide tools for conversion of the native output format into the quasi-standard FASTQ. Checking the quality of the generated sequence data is the first step in the pipeline that deals with the actual sequence data in base or color space.

The distribution of the quality scores at each sequence position is one of the most interesting quality parameters for the overall quality of the run. Typically, the base calling software of the manufacturer already provides initial overview on the data quality. For a more thorough quality overview, freely available tools such as SolexaQA (Cox et al. 2010) or FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) can be applied. By default, SolexaQA generates three figures: the average error probability per position in the read (Fig. 2), a histogram of the maximal read length without a single base quality below a specified threshold, and an overview about the average quality per tile (Figure S1a and b, respectively, in the supplementary material). FastQC can handle data from all current HTS platforms and provides a user-friendly graphical interface that allows, among other things, to check for over-represented sequences, deviation from the expected GC content, distribution of nucleotides per read position, thus allowing a fast identification of problems that can occur during sample preparation and sequencing.

The output of SolexaQA for our example data indicates that the error probability increases with increasing read length. This behavior is typical for HTS platforms. As a consequence, read trimming is often applied to increase the number of mappable reads by removing bases at the end of the read that are likely to contain sequencing errors. Hence, read trimming may be of particular value in settings where every aligned read is precious for the analysis. The trimming can be carried out either explicitly by a tool such as the DynamicTrim module provided by SolexaQA or implicitly by the alignment algorithms used in the downstream process. Regarding the quality of the raw reads, there are noticeable differences between the platforms.

Fig. 1 Workflow of the SNP calling pipeline


Input: Images from Sequencer
Tools: Internal HTS System Software
Output: Base- or Color-Sequence and Quality Scores -> e.g. fastq
Base Calling
Input: Read Data -> e.g. fastq
Tools: SolexaQA, FastQC, PRINSEQ
Output: Quality Report and Filtered Reads -> e.g. fastq
Quality Control
Input: Filtered Read Data -> e.g. fastq
Tools: BWA, MAQ, Stampy, Bowtie, SHRiMP2, bfast
Output: SAM, BAM and Mapping Statistics
Alignment/ Mapping
Input: SAM/BAM
Tools: samtools, Picard, SMRA, GATK
Output: SAM/BAM
Alignment Post-Processing
Input: SAM/BAM
Tools: SOAPsnp, GATK
Output: SAM/BAM
Quality Score Recalibration
Input: SAM/BAM
Tools: SOAPsnp, MAQ, samtools, GATK, Beagle
Output: vcf
Variant and Genotype Calling
GACTCTCTGAG
TTAGACTATCTGAG
CTCAAACTCTCTG
TCCTTAGACTCTC
GTCCTTAGACTCTC
CTCCTTAGACTCT
AGGTCCTCAGACT
CAGGTCCTCAGACTCTGAGAG
Input: vcf
Tools: GATK, samtools, VCF tool
Output: vcf
Filtering SNP Candidates
Making Sense of SNP Data
Common for all HTS Pipelines
Specific for SNP Calling

Illumina reads, for instance, undergo a quality control by the manufacturer's software. In case of the SOLiD platforms, no quality control is provided. This platform relies on the fact that reads of insufficient quality will not align to the reference sequence. However, it makes sense to discard reads with a mean quality score below 10 for the SOLiD platform to reduce mapping time. Again this can be done either explicitly via tools such as PRINSEQ (Schmieder and Edwards 2011) or implicitly via the employed aligner, e.g., SHRiMP2 (David et al. 2011).

Step 2: alignment/mapping

The next step in the processing pipeline for almost all applications is the alignment of the reads to a reference sequence, i.e., the human genome in our case. The requirement for aligning several million short reads, which contain small deviations (e.g., SNPs, indels) and sequencing errors, to a reference sequence or a database of sequences has brought forth a number of efficient algorithms. In addition, some algorithms may be fine-tuned for optimal compatibility to specific sequencing platforms.

Briefly, two approaches are commonly used for solving the task. The first one applies the lossless Burrows–Wheeler transform (BWT) (Burrows and Wheeler 1994) for efficient data compression. Other algorithms rely on hashing to accelerate the alignment step. The use of hashing allows quick access to the information on the location of subsequences in the reference sequence. Hash-based aligners either hash the reads, e.g., Eland (part of the Illumina's CASAVA suite), or the reference sequence,

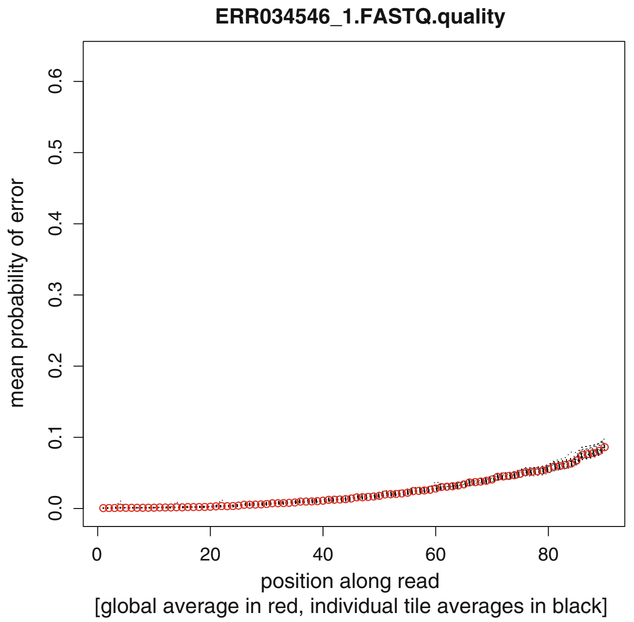


Fig. 2 Output of SolexaQA for the first mate of the pair of the sample data. This plot represents one of the three figures provided by SolexaQA. The *x* axis refers to the position in the read (i.e., the sequence cycle). The *y* axis depicts the mean probability of a sequencing error. *Red circles* indicate the mean value for the whole dataset, whereas *filled black circles* correspond to the mean for each tile. Of note, the error probability increases with increasing sequence length. This behavior is typical and common to all currently used HTS platforms

e.g., SOAP (Li et al. 2009b). See Li and Homer (2010) for a recent survey on mapping methods.

In general, the choice of alignment tool and the corresponding settings significantly affect the outcome. This holds especially true for SNP calling, as wrongly aligned reads may result in artificial deviations from the reference. These deviations in turn may falsely be classified as SNPs in the downstream processing. We will demonstrate the dependence of the results on the aligner by employing multiple tools for the alignment step. For other applications, misaligned reads may be less critical than they are for variant calling. These include mainly quantitative analyses such as gene expression profiling, where the number of reads aligning to a gene is related to its expression level. Here, the sequence content of the read is only of secondary interest. Nevertheless, misalignments may still distort the inferred expression levels.

In this work, we will focus on freely available software. We used the following algorithms for mapping our example reads: Bowtie (Langmead et al. 2009) and its successor Bowtie2, BWA (Li and Durbin 2009), MAQ (Li et al. 2008), and stampy (Lunter and Goodson 2011). The reads were aligned against the version 18 of the human genome as provided by UCSC (Kent et al. 2002). Bowtie cannot perform gapped alignment; hence, one cannot find short indels with this aligner. Its successor Bowtie2 included a number of useful features such as gapped alignment. However, currently, Bowtie2 cannot be applied to align SOLiD color-space reads. Stampy is a hash-based aligner and also incompatible with SOLiD color reads; it can be executed in a hybrid mode that uses the much faster BWA aligner for improving the runtime of the alignment process. We compare stampy in normal and hybrid mode (stampyBWA). The aligners employed here are only a small selection of the freely available and widely used methods. Two further widely applied hash-based aligners, which are also capable of dealing with reads in color space, are SHRiMP2 (David et al. 2011) and BFAST (Homer et al. 2009).

Table 2 depicts the time required on a single CPU for aligning the paired-end reads of the example data to the human genome. To further accelerate the alignment step, many of the algorithms can be easily distributed on a larger number of CPUs (e.g., Bowtie and BWA) by simply adding a parameter at execution of the program. In contrast, for other aligners, the input data have to be split in separate files for parallelizing the alignment (e.g., MAQ), which typically requires manual preprocessing. When stampy is executed in hybrid mode, one can parallelize the BWA process via a parameter. For instance, the allocation of ten CPUs instead of a single CPU for the BWA part in stampy reduces the time from 6,300 to 3,100 min in our example.

The clear advantage of the BWT-based algorithms (here: BWA, Bowtie, Bowtie2) over the hash-based algorithms (here: MAQ, stampy, stampyBWA) is the processing speed (see Table 2). However, BWT algorithms are not as sensitive as hash-based aligners, and therefore may introduce mapping biases in regions with high variability (see Lunter and Goodson 2011 for a detailed sensitivity/specificity analysis). In our opinion, approaches like stampy are a good compromise as they combine the sensitivity of hash-based alignment with the speed-up gain introduced by the BWT approach.

Furthermore, not only the choice of the alignment algorithm is essential but also its parameter settings. Clearly, if one allows only perfect matches between read and reference, the downstream analysis will not find any differences between the reference and the sequenced genome, thus no SNPs can be detected. Conversely, allowing many mismatches between reference and read may promote many wrong alignments and result in a high number of false-positive SNPs in the downstream analysis. Hence, maximizing the number of aligned reads at all costs is not the best strategy. Selecting the best number of accepted mismatches is also highly depended on the species. Specimen of *Mus musculus* strains, for instance, can deviate quite significantly from the available reference (Keane et al. 2011). Human samples, on the other hand, tend to be less variable (Consortium 2010). Unfortunately, this statement does not hold for the entire genome. The major histocompatibility complex (MHC), for instance, shows high variability between human individuals. Is it thus generally very challenging to perform good alignments in this region.

Once the reads have been aligned to the reference, many algorithms allow to store the result in the sequence alignment/map (SAM) format (Li et al. 2009a). Briefly, the SAM format stores information about each aligned read, in particular, the position on the reference contig, the orientation of the read, quality of the alignment and potential further alignment possibilities of the read. In case the aligner's output is in a different format, third party tools may be available for a conversion into the SAM format. The SAM format and its binary version, the BAM format, are by now a quasi-standard for storing the result of the alignment step. Hence, many downstream processing tools rely on the SAM/BAM format. Moreover, tools that provide an efficient manipulation of mappings stored in SAM and BAM format have been published. The most widely applied toolkits are SAMtools (Li et al. 2009a), GATK (McKenna et al. 2010), and Picard (see Table 1).

After the mapping step, it is advisable to check the alignment again. This can be done easily by generating a mapping statistic, i.e., computing the fraction of reads that was successfully mapped to the reference (using, e.g., the *flagstat* command of SAMtools or the CollectAlignmentSummaryMetrics module of Picard). Moreover, when working with paired reads, the fraction of reads that was successfully paired (see Table 3) and the distribution of the insert size are parameters of interest (see Fig. 3). In our case, based on these metrics, we choose a larger insert size of 1,000 bp for the following processing. When analyzing data from target re-sequencing, the enrichment of the reads in the target area compared to the off-target area is of high interest. We found the CalculateHsMetrics module of Picard to be most useful for computing this ratio. In the example-data, we achieved a 34-fold enrichment of reads regardless of the alignment method we used. This indicates a successful enrichment process. In addition to the enrichment, the module provides information on the fraction of the target region that was not covered by any read (about 15 % in our data), the fraction of the target region with a minimum coverage of 10 (73 %), and also the mean coverage in the target region (about 130).

Table 2 Time (min) required for the mapping step by different algorithms on a single CPU (AMD 2.1 GHz)

Mapping algorithm	Time (min)	
	Default insert size	1,000 bp insert
Bowtie	910	780
Bowtie2	880	990
BWA	1,534	1,522
MAQ	14,719	14,848
Stampy	12,254	12,590
StampyBWA	6,362	6,302

The mapping was carried out with two different settings for the expected insert size length (columns). The BWT-based aligners are about one magnitude faster than the hash-based ones

A visual inspection of a whole genome sequencing experiment is usually not realistic. One can, however, use a tool (e.g., the *view* command of SAMtools) for extracting the alignments within a target region and visualize only that specific region in a genome browser such as the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). Figure 4 depicts the alignment of the reads in the SLC6A15 locus using IGV. The visualization reveals that the aligned reads concentrate on exonic regions. Hence, the whole exome sequencing was successful, at least for the inspected region. Moreover, it is interesting to note that the whole exome enrichment process is not very precise, as adjacent intronic regions are highly enriched as well. The coverage in the adjacent intronic regions is, however, dominated by either forward or reverse reads and may thus introduce a bias in the SNP called in that region. Besides IGV, there are further tools allowing visual inspection of alignments, for instance, the software tools GenomeView (Abeel et al. 2012) and SAVANT (Fiume et al. 2010). The latter can even be extended by user-contributed software modules.

In general, fine-tuning the alignment parameters requires some effort. Most of the projects involve a large body of data, thus it saves time and hard drive space to only work with a randomly selected subset of the reads, e.g., 10 %. This also applies to the quality control aspect of step 1.

Step 3: alignment post processing

Prior to the actual variant calling, the algorithms require the alignments to be sorted with respect to their chromosomal position. This can easily be done using tools like SAMtools or Picard. Next, since the PCR used for amplifying the library and adding adapters may introduce artifacts, i.e., reads or read pairs starting at exactly the same position and having the same insert length, respectively, it is common practice to remove or simply mark such PCR artifacts. Again, SAMtools and Picard provide the means for solving this task. The next post-processing step is the removal of all non-unique alignments, i.e., reads with more than one optimal alignment; since in these cases, it cannot be determined from which site the read really originates. And last but not least, it is common to realign reads around small indels. Since, differences in resolving the indels may cause artificial SNPs in the downstream analysis. The GATK software for instance offers the possibility to realign

Table 3 Overview of the fraction of aligned and properly aligned reads for two different settings for the expected insert size

	Default insert size		1,000 bp insert	
	%aligned	%properly	%aligned	%properly
Bowtie	50.13	100.0	94.70	100.0
Bowtie2	99.45	96.52	99.58	99.74
BWA	95.93	96.18	95.93	96.18
MAQ	99.04	52.29	99.45	98.95
Stampy	98.32	96.84	98.25	0.00
StampyBWA	99.52	99.44	99.52	98.81

The fraction of aligned reads (%aligned) is in reference to the total amount of reads, while the fraction of properly aligned reads (% properly) is in reference to the aligned ones. Here properly refers to the fact that both reads are aligned in the expected direction and the expected distance. With the default insert size Bowtie only maps about 50 % of the reads; reads with larger insert sizes are completely rejected. MAQ, on the other hand, maps these reads and reports them as not properly paired. Setting the insert size to 1,000 leads to comparable mapping statistics for all aligners

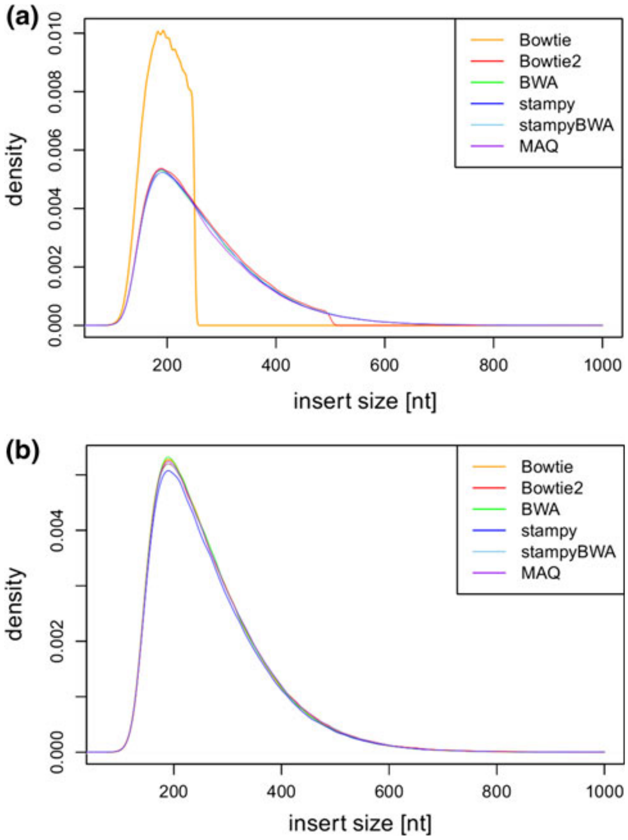


Fig. 3 Distribution of the insert size of the whole exome paired-end alignments. *Different colors* correspond to the various aligners. The default insert size parameter (**a**) appears to be too small as seen in the example of Bowtie; the density distribution appears to be cut after 250 bp. A larger parameter for the insert size of 1,000 bp leads to almost identical distributions for all alignment programs (**b**)

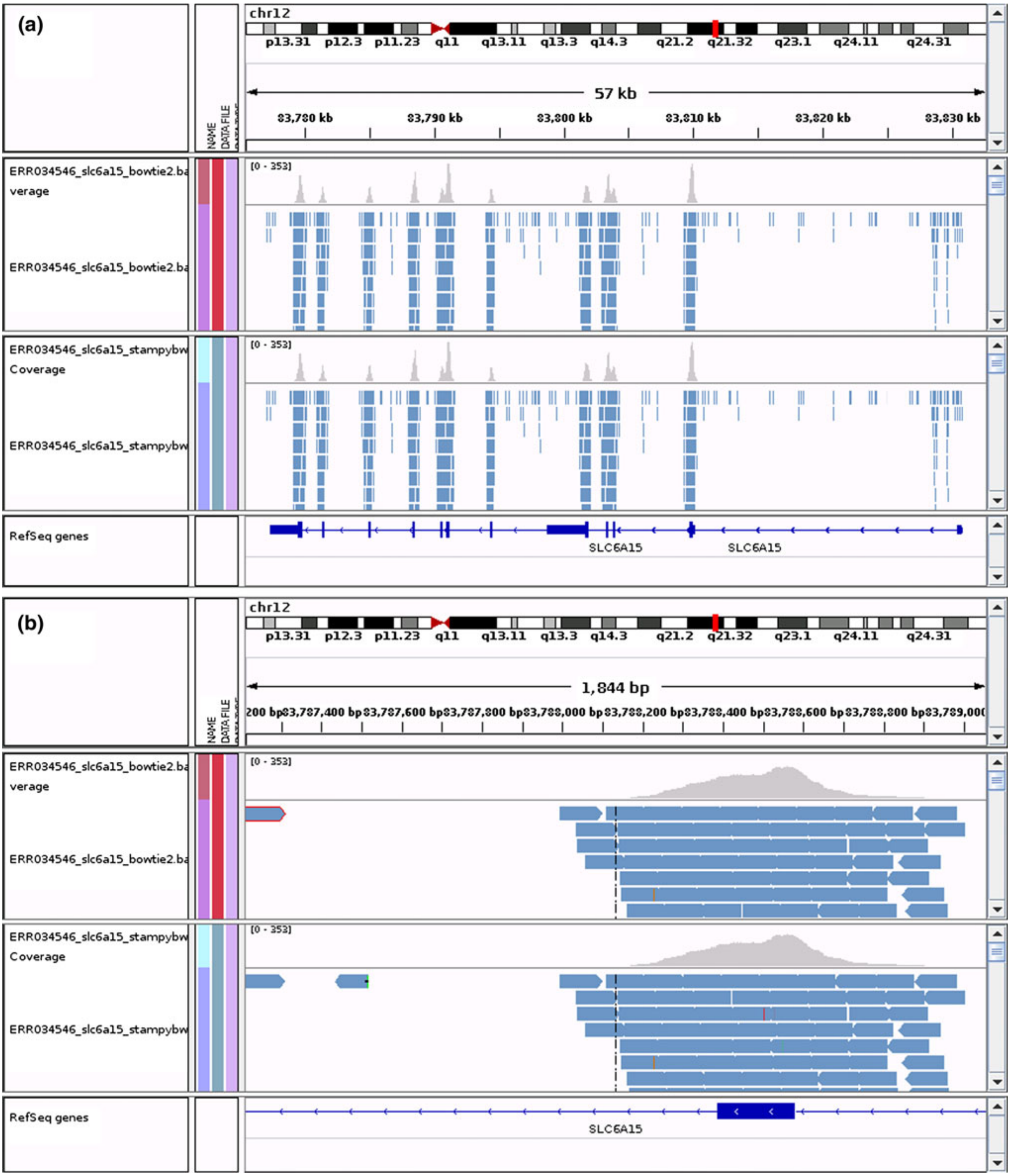
Fig. 4 IGV visualization of the SLC6A15 locus (hg18 coordinates: chr12:83,777,000–83,831,000). Inlet (**a**) shows the view of the entire gene, while inlet (**b**) depicts a region covering a single exon within SLC6A15. In all images the *top most panel* shows the view of the full chromosome with the current position highlighted in *red*; the *bottom panel* provides the exon–intron structure of the gene as listed in the RefSeq database. The *middle panels* visualize the alignments for Bowtie2 (*top*) and stampyBWA (*bottom*). These panels each show the coverage, i.e., number of reads covering a specific region (*upper part*) and the details of single reads or read-pairs (*lower part*). In subfigure (**a**) one can see that most of the reads are concentrated in the area of the exons and only few reads are aligned to intronic regions. Both alignments are very similar; minor differences can be seen in subfigure (**b**). For instance, the read-pair at the *left side*: Bowtie2 aligned only one mate of the pair, while stampyBWA aligned both. Interestingly, whole-exome sequencing is not restricted to the exons only, but also covers the adjacent intronic regions ▶

those reads. The tool SMRA (Homer and Nelson 2010) allows to realign the reads in color space originating from the SOLiD platform. Once these steps are completed, one can proceed to the processing steps of the pipeline that are specific to the SNP calling process.

Step 4: quality score recalibration

Previous works demonstrated that the Phred-like quality scores issued by the sequencing platforms may often deviate from the true error rate (Li et al. 2009b). Having accurate quality scores is essential for the modern SNP calling algorithms, as they integrate the Phred scores of the bases covering the site to be examined into their scoring functions (see step 5). The first software to provide recalibration of quality scores was SOAPsnp (Li et al. 2009b). The approach exploits sites in the reference genome without any reported SNPs. On these sites, SOAPsnp computes the empirical mismatch rate as an estimate for the true base quality. More precisely, for a given machine provided quality score, sequencing cycle (i.e., position of the base in the read) and substitution type (e.g., A→G: A in reference and G in read), it calculates the average mismatch rate with respect to the reference. This mismatch rate is then used as the recalibrated quality score. Based on a similar concept, the GATK software also provides a recalibration function: first, bases are grouped with respect to several features (e.g., raw quality, dinucleotide content); second, for each such category, the empirical mismatch rate is computed and used to correct the raw quality score. GATK's recalibration functionality can be applied to the sequencing data of various platforms. Figure 5 depicts original and recalibrated quality scores using GATK for the alignment with BWA.

Of note, since only little differences are expected from an alternative order, steps 3 and 4 may be swapped.


(a)
chr12
p13.31 p12.3 p11.23 q11 q13.11 q13.3 q14.3 q21.2 q21.32 q23.1 q24.11 q24.31
57 kb
83,780 kb 83,790 kb 83,800 kb 83,810 kb 83,820 kb 83,830 kb
NAME
DATA FILE
ERR034546_slc6a15_bowtie2.ba
verage
ERR034546_slc6a15_bowtie2.ba
[0 - 353]
ERR034546_slc6a15_stampybw
Coverage
ERR034546_slc6a15_stampybw
[0 - 353]
RefSeq genes
SLC6A15
SLC6A15
(b)
chr12
p13.31 p12.3 p11.23 q11 q13.11 q13.3 q14.3 q21.2 q21.32 q23.1 q24.11 q24.31
1,844 bp
200 bp 83,787,400 bp 83,787,600 bp 83,787,800 bp 83,788,000 bp 83,788,200 bp 83,788,400 bp 83,788,600 bp 83,788,800 bp 83,789,000
NAME
DATA FILE
ERR034546_slc6a15_bowtie2.ba
verage
ERR034546_slc6a15_bowtie2.ba
[0 - 353]
ERR034546_slc6a15_stampybw
Coverage
ERR034546_slc6a15_stampybw
[0 - 353]
RefSeq genes
SLC6A15

Step 5: variant and genotype calling

Early variant or SNP calling approaches relied entirely on counting the abundance of high-quality nucleotides at a singe site (e.g., Wang et al. 2008). Recent approaches, however, integrate several sources of information within a probabilistic framework. This procedure facilitates SNP calls in medium to low coverage regions, where for example only five reads are covering the position of the potential SNP. Moreover, these probabilistic approaches provide a natural way for quantifying uncertainty about the variant call. Further details on the statistical models used are available in the Online Supplementary Material.

One major advantage of the statistical framework is the use of prior probabilities for a SNP at a given position. These prior probabilities can be derived from databases listing of confirmed SNPs such as the dbSNP (Sherry et al. 2001) or by carrying out SNP calling in multiple individuals at the same time. The SNP calling routines implemented in SAMtools and GATK both support the use of multiple sample SNP calling. Further improvements can be achieved by incorporating linkage disequilibrium (LD) information. Here, the same principles that allow the imputation of missing genotypes (see e.g., Howie et al. 2009; Marchini et al. 2007) facilitate more reliable genotype calls. This functionality is implemented in the software Beagle for instance (Browning and Yu 2009). However, when working with whole exome data, large fractions of the LD structure are missing and therefore no improvement can be expected from applying this step.

A thorough review on available SNP calling algorithms is provided by Nielsen et al. (2011) and references therein. For our example data, we applied two SNP calling programs: SAMtools and GATK. Since our example only comprises data from a single individual, we did not make use of the improved accuracy due to multi sample SNP calling.

Step 6: filtering SNP candidates

Filtering is an essential step in reducing the number of false-positive SNP calls. Typically applied filters check for deviations from the Hardy–Weinberg equilibrium (HWE), minimum and maximum read depth, adjacency to indels, strand bias, etc. While filtering might also remove real SNPs from the candidate list, it is an important tool for minimizing SNP calling artifacts. Filtering is provided by GATK, SAMtools (via the script '*vcfutils.pl*') and VCFtools (Danecek et al. 2011). In particular, GATK provides 'best practice' settings for the variant calling pipeline, including SNP candidate filtering. For our example data, we used the SAMtools default filter (see Table 1) and the GATK VariantFiltration. Here, we followed the recommendations of the best practice guidelines version 3 and used a hard filter due to the low sample number, i.e., a single individual. More precisely, SNPs with a quality below 30.0 or a quality per depth below 5.0 or SNPs within a homopolymer of length 6 and more were discarded.

Most SNP calling tools have the option to generate the data in the VCF format (Danecek et al. 2011). The VCF format records for each identified SNP candidate basic information such as the chromosomal position, the reference base, the identified alternative base or bases in case of trialleic SNPs. Furthermore, information on the quality of the SNP call as well as the amount of sequence data available for the call are stored. The VCFtools provide the possibility to easily manipulate VCF files, e.g., merge multiple files and extract SNPs in defined regions.

Table 4 depicts the number of SNPs called with every combination of alignment algorithm and SNP caller after the initial filtering step. Given one alignment algorithm, the SNPs called with the two different tools largely overlap, i.e., about 85 % of the SNPs are shared. However, the number of called SNPs exceeded the expected number by one order of magnitude, this was likely due to reads aligned outside the target region. Thus, we filtered the SNP candidates further using VCFtools. In particular, we required that SNPs reside within the target region of the enrichment assay ±50 bp and that positions with SNPs show a minimum sequencing depth of ten. Table 5 shows the number of SNPs after the second filtering step. Again, about 80 % of the SNPs were identified with both SNP callers from the same alignment. More precisely, GATK generated about 5,000 additional SNP candidates compared to SAMtools. Table 6 shows the impact of the different alignment algorithms given a fixed SNP caller. Here, most SNPs (about 85 %) were found in the alignments produced by the four utilized aligners. Thus, both aligner and SNP caller showed a significant impact on the result. The majority of SNPs, however, was discovered with any combination of the tools used here.

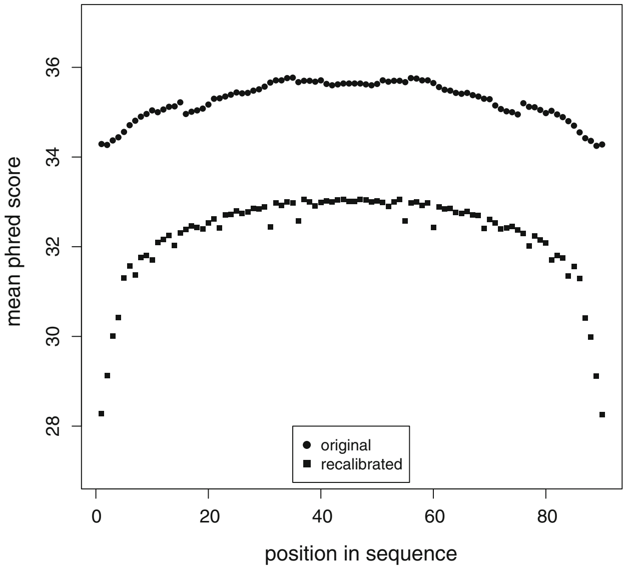


Fig. 5 Average base qualities before and after quality score recalibration for the alignment with BWA. Original quality scores (*filled circles*) vary only a little along the position of the read (i.e., mean scores are between 34 and 36). Recalibrated scores (*filled squares*) are considerably lower (i.e., ranging from 28 at the ends of the read to 33 in the middle of the read), thus indicating an overestimation of the error probability by the manufacturer's base calling software

Making sense of SNP data

The SNP calling process on the whole exome data generated about 24,000 variants. Thus, following up on each variant manually is clearly out of scope; even more so for whole genome re-sequencing studies. Due to this requirement, tools for automated variant annotation have been developed. ANNOVAR (Wang et al. 2010), for instance, offers a command line interface for annotating variants from different species. The tool relies on information available via the UCSC genome browser and offers in addition precomputed scores such as the SIFT-score (Ng and Henikoff 2003) for predicting the likely functional consequences of non-synonymous amino acid exchanges. Furthermore, ANNOVAR allows to easily filter already known SNPs using information from dbSNP or the 1,000 genome project. Table S2 in the Online Supplementary material depicts the result of the functional annotation by ANNOVAR for the SNPs that were called with any combination of aligner and SNP caller. About 50 % of the discovered SNPs are actually exonic variants; the remaining 50 % originate from the region with 50 bp around the target region. A further value of interest when analyzing SNP data is the ratio between transitions and transversions. This statistic can be easily computed using VCFtools. In our case, the Ts/Tv ratio is between 2.1 and 2.4 (2.6 and 2.7) for SNPs called with GATK (SAMtools). Hence, the SNPs called with GATK are closer to the expected ratio of 2, where transitions seemed to be enriched among the SNPs called with SAMtools.

The post-processing and interpreting the generated SNP data are the substantial challenges and the effort associated with this important task should not be underestimated. Apart from ANNOVAR, there are further tools that assist in the interpretation of the SNP data, see for instance the tools listed at http://www.gen2phen.org/wiki/tools-predicting-overal-functional-consequences-snps. An example for annotation software providing a graphical user interface is the sequence variant analyzer (Ge et al. 2011). Moreover, a number of commercial suites in addition to these free tools exist.

Not only variant annotation but also the statistics for finding significant associations have to be adapted. As whole genome and whole exome sequencing studies will produce more and more rare SNPs, i.e. SNPs with minor allele frequencies below 1 %, standard statistical approaches do not have sufficient power for finding significant associations with the currently available and realistic sample sizes. Thus, variants with similar characteristics such as SNPs with the same functional annotation, SNPs within the same biological pathway, or SNPs close on genome (Cohen et al. 2004)) are often grouped to proxy variables to increase power. These proxy variables are then subject to statistical significance analysis. Here, variant annotation and statistical analysis go hand-in-hand. Bansal et al. (2010) provide a review on current statistical approaches for analyzing rare variants.

Table 4 Number of SNPs for each combination of alignment algorithm and SNP caller after the initial filtering

	SAMtools	GATK	Common
Bowtie2	236,399	228,818	204,129
BWA	230,754	234,455	199,569
MAQ	248,853	241,234	212,724
StampyBWA	247,855	252,458	211,802

The absolute number varies from 230,000 SNPs to 250,000. The column entitled "common" displays the number of SNPs that were found with both SNP callers using the same alignment. Common refers to identical position and identical genotype call. Here, numbers range from 200,000 to 213,000

Table 5 Number of SNPs for each combination of alignment algorithm and SNP caller after restriction to the target region ±50 bp and min. sequencing depth of 10

	SAMtools	GATK	Common
Bowtie2	25,115	30,519	23,988
BWA	24,645	30,471	23,976
MAQ	25,155	31,130	24,013
StampyBWA	25,166	31,512	24,057

The absolute number varies from 24,600 to 31,500 SNPs. The column entitled "common" displays the number of SNPs that were found with both SNP callers using the same alignment. Common refers to identical position and identical genotype call. Here, numbers are close to 24,000

Conclusion

The SNP calling pipeline, like many other HTS pipelines, involves many different steps. Typically, these steps can be

Table 6 Number of SNPs in common (position and genotype call) between the four alignment algorithms (also used in Tables 4, 5) after restricting to the target region ±50 bp and a min. depth of 10

	SAMtools	GATK
Only in a single alignment	1,792	1,682
Only in two alignments	715	696
Only in three alignments	1,473	1,254
In all four alignments	23,110	29,199

Each column corresponds to one SNP caller; the rows indicate from how many alignments a SNP could be called. The largest fraction of SNPs (about 23,000) was found in all four alignments produced by the algorithms listed in Table 5. Moreover, 22,383 SNPs could be found with any combination of aligner and SNP caller integrated using shell scripts in combination with a queuing system such as the Sun Grid Engine (Gentzsch 2001). Another option that will also allow less computer-experienced users to carry out many HTS tasks is the Galaxy web service (Goecks et al. 2010). Galaxy is a "web-based platform for data intensive biomedical research" and the developers also provide it as a freely accessible web service. For large datasets, however, it is essential to run Galaxy on local computation infrastructure. Setup and integration of new pipelines, however, requires in-depth computer knowledge. An option that we did not address in this review is the use of commercial software products. Many companies have developed software packages that allow basic and advanced analysis of HTS data. These software suites are in general based on a graphical user interface. The ease of use, however, may be compromised by the lack of flexibility. In any case, costs associated with software licenses are not negligible.

We have presented a SNP calling pipeline starting from the sequenced reads to the annotation of the identified variants. The presentation of the pipeline was illustrated by processing a whole exome sample dataset. The results of the example data demonstrated that the choice of the tools and parameters have significant impact on the final result. Thus, the outcome of a SNP calling pipeline is not set in stone. A recommended strategy, for instance, is the use of different aligners and SNP callers for generating independent SNP candidates. Reliable candidates are those appearing in more than one setting.

Working with HTS systems is a truly interdisciplinary effort. While the generation of the data is mainly laboratory-centered, the initial processing of the short read data falls into the domain of bioinformatics and can be automated to a certain degree. The interpretation of the results, however, requires close interaction between biology and bioinformatics in order to derive the maximal insights from the data. The work associated with these three domains—data generation, processing, and interpretation—is non-negligible and requires dedicated resources (including human resources) for guaranteeing a successful completion of the research project.

Further reading

It is essential to stay updated regarding new developments in the field. A virtual online issue of the journal *Bioinformatics* collects articles related to HTS sequence analysis published in the journal (http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html). In general, the algorithm developers provide information on improvements on the corresponding program websites (see Table 1). For instance, GATK comes along with a best practice guideline (http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3), also the 1,000 genomes project provides information on their processing pipelines (http://www.1000genomes.org). Moreover, many of the developers are actively participating on the seqanswers (http://seqanswers.com) forum, a discussion forum and source of information for all matters regarding HTS. Seqanswers is probably the most helpful online resource.

Acknowledgments The work of MP was supported by an intramural grant from the Medical Faculty of the University at Lübeck.

References

Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y (2012) GenomeView: a next-generation genome browser. Nucleic Acids Res 40:e12

Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11:773–785

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837

Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85:847–861

Burrows M, Wheeler DJ (1994) A block-sorting lossless data compression algorithm. Technical Report Digital Equipment Corporation, Palo Alto

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265–270

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872

Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinform 11:485

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

David M, Dzamba M, Lister D, Ilie L, Brudno M (2011) SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics 27:1011–1012

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34:e22

Fiume M, Williams V, Brook A, Brudno M (2010) Savant: genome browser for high-throughput sequencing data. Bioinformatics 26:1938–1944

Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP, Zhu M, Singh A, Allen AS, Goldstein DB (2011) SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics 27:1998–2000

Gentzsch W (2001) Sun Grid Engine: towards creating a computer power grid. In: First IEEE/ACM International Symposium on Cluster Computing and the Grid 2001, pp 35–36

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86

Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS ONE 4:e7767

Homer N, Nelson SF (2010) Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome Biol 11:R99

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529

Kao WC, Song YS (2011) naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. J Comput Biol 18:365–377

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellaker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assuncao JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289–294

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12:996–1006

Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E (2010) Genome-wide measurement of RNA secondary structure in yeast. Nature 467:103–107

Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10:R83

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li H, 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, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009a) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K (2009b) SNP detection for massively parallel whole-genome resequencing. Genome Res 19:1124–1132

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936–939

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814

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

Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods 5:179–181

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I (2010) De novo assembly and analysis of RNA-seq data. Nat Methods 7:909–912

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Tanaka H, Kawai T (2009) Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. Nat Nanotechnol 4:518–522

Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet 19:R145–R151

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C et al (2001) The sequence of the human genome. Science 291:1304–1351

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Yang H (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wu H, Irizarry RA, Bravo HC (2010) Intensity normalization improves color calling in SOLiD sequencing. Nat Methods 7:336–337

Xi R, Kim TM, Park PJ (2010) Detecting structural variations in the human genome using next generation sequencing. Brief Funct Genomics 9:405–415

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Ni Asan P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Luosang Huasang J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y, Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Yang H, Nielsen R, Wang J (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75–78