Genome-wide approaches and technologies to assess human variation

Claudia Gonzaga-Jauregui [1,2]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA.

[2]Center for Human Disease Modeling, Duke University, Durham, NC, USA (current affiliation).

Abstract

Current genome-wide technologies allow interrogation and exploration of the human genome as never before. Next-generation sequencing (NGS) technologies, along with high resolution Single Nucleotide Polymorphisms (SNP) arrays and array Comparative Genomic Hybrization (aCGH) enable assessment of human genome variation at the finest resolution from base pair changes such as simple nucleotide variants (SNVs) to large copy-number variants (CNVs). The application of these genomic technologies in the clinical setting has also enabled the molecular characterization of genetic disorders and the understanding of the biological functions of more genes in human development, disease, and health. In this review, the current approaches and platforms available for high-throughput human genome analyses, the steps involved in these different methodologies from sample preparation to data analysis, their applications, and limitations are summarized and discussed.

PeerJ PrePrints | https://peerj.com/preprints/147v1/ | v1 received: 9 Dec 2013, published: 9 Dec 2013, doi: 10.7287/peerj.preprints.147v1

Next-generation massively parallel sequencing for high-throughput human genome analysis

Since the beginning of the Human Genome Project, sequencing of the human genome has driven the development of technologies and methods to discover the variation within it. While the sequencing of the original human haploid consensus reference genome cost an estimated $2.7 billion US dollars [1], the subsequent development of better and more efficient sequencing machines and methodologies, and later the development of massively parallel sequencing and next-generation sequencing technologies, dramatically reduced both the cost and time to sequence personal human genomes. In addition, the development and refinement of targeted capture methods and reagents for exome sequencing has resulted in a rapid increase in the number of human exomes analyzed dramatically expanding our knowledge of human genetic variation. Concurrently, bioinformatic algorithms and tools have been developed to manage and analyze the tremendous amount of data generated.

The process to sequence a human genome or exome is now relatively straightforward and the methodological differences arise mainly from the preferred capture, amplification, and sequencing platforms used. In summary, the process can be reduced to four different steps (Figure 1): 1) DNA preparation, 2) library construction, 3) sequencing, and 4) analysis.

1) **DNA preparation.** Human genomic DNA can be isolated from different sources; generally peripheral blood is preferred as the starting biological material. However, available reagents to stabilize other biological fluids such as saliva have proven to be useful when a blood draw is not possible or insufficient, providing an adequate yield and quality of DNA for sequencing when

PeerJ PrePrints | https://peerj.com/preprints/147v1/ | v1 received: 9 Dec 2013, published: 9 Dec 2013, doi: 10.7287/peerj.preprints.147v1

collection is done properly. Extraction of DNA from tissue biopsies, preferably fresh tissues, can be performed by first digesting the tissue using proteinase K. DNA extraction from formalin-fixed, paraffin-embedded (FFPE) tissues is, although possible, often suboptimal in yield and molecular weight integrity. The DNA yield for tissue biopsies is lower due to the amount of starting available material and there is a risk of DNA degradation during extraction and purification. Current next-generation technologies allow the preparation of sequencing libraries with as little as 1 ug of genomic DNA.

2) Library Preparation. After extraction and purification, genomic DNA is fragmented by mechanical methods, such as nebulization or sonication, into fragments of ~200-400 bp. Sonication is usually preferred over nebulization because the amount of input DNA is less and the fragment size is more consistent. Fragment ends are enzymatically repaired and adaptors, which can be barcoded, are ligated to the ends. For whole-genome sequencing, the whole-genome shotgun library is amplified and subjected to next-generation sequencing. For exome sequencing, or other targeted approaches, target capture and enrichment is implemented prior to amplification. Prior to massively parallel sequencing, human genomic sequencing already used target enrichment approaches, but these were laborious and not highly scalable methods such as PCR for specific segment amplification or cloning of discreet genomic segments using bacterial vectors and including fosmid and BAC library construction. The currently used targeted capture methodologies were originally developed using oligonucleotide probes covalently bound to a solid array glass slide designed to specifically bind target regions or the exons of the target genes [2-5]. Later, solution based capture was developed [6,7] in which target fragments specifically hybridize to biotinylated probes that are then pulled down using streptavidin coated magnetic beads. During exome capture hybridization, it is important to block repetitive DNA, using human Cot-I DNA which is added in excess in the hybridization solution, in order to avoid nonspecific cross-hybridization. The fragmented genomic DNA hybridizes to the complementary probes either on the array or to the biotinylated oligonucleotide probes in solution; any non-target fragments that do not hybridize are later washed away and consequently not captured for subsequent sequencing. The capture efficiency is dependent on the target fragment length, sequence complexity, and GC content of the region. Solution-based capture is cheaper and more scalable than microarray-based capture; thus most commercially available exome capture reagents use solution based capture [8].

Importantly, different targeted capture designs exist based upon the genome/gene annotation(s) used for design. Additionally, the above mentioned methodologies are well suited for enriching the “whole exome” with capture libraries of ~50 Mb. However if a more targeted approach is desired, such as those for gene panels or specific regions, approaches such as molecular inversion probes (MIPs) and other multiplex or modified PCR-based amplification of targets can be used to enrich for the desired regions on a reduced scale [9]. Originally MIPs were developed, improved and applied for high-throughput multiplex SNP genotyping [10, 11]. The current MIPs technology relies on the specific design of ~70-mer capture probes. The MIP structure is composed of a common linker sequence flanked by homologous targeting arms that hybridize upstream and downstream to the genomic region of interest. A synthesis reaction follows in which a DNA polymerase copies the target sequence using the upstream targeting arm as an extension primer. After extension, the 5’ end is then ligated to the downstream targeting arm and the probe is circularized. Further post-capture library amplification, barcoding and sequencing adaptor ligation can later be performed using the common MIP linker sequence [12-14]. Current MIPs designs and approaches have proven effective at capturing ~55,000 targets or ~6 Mb [13, 14].

Most of the current sequencing technologies rely on the amplification of the template to be sequenced in order to form clusters of clonally amplified molecules termed “polonies”; which derives from PCR and colony referring to the original bacterial colonies needed to amplify DNA BACs for sequencing. There are currently two predominant approaches for pre-sequencing library amplification. Emulsion PCR amplification is performed in a water-oil emulsion that contains the captured fragment library, dNTPs, polymerase, and beads with oligonucleotide primers complementary to the adaptors initially ligated to the DNA fragments. In the test tube, each of the spheres formed by the water-oil emulsion will perform as an individual isolated PCR reaction in which the template fragments will be clonally amplified. These beads will then be washed and cross-linked or spread into a slide or solid platform in which the sequencing reaction will be performed. The second approach is a solid-phase amplification, in which pairs of oligonucleotide amplification primers are covalently bound to a solid phase and the template amplification takes place by bridge amplification of the target fragment using a pair of primers and generating clusters of clonally amplified target molecules.

The efficiency of the targeted capture enrichment step can be easily assessed by qPCR, testing a few target loci in the initial non-amplified input DNA versus the amplified captured DNA and comparing their C_T values. However, this QC step does not provide information on the specificity and sensitivity of the capture method, just the efficiency of the enrichment [5].

3) Sequencing. Sequencing technologies can be divided into two main categories based on the enzyme that they use: i) sequencing by ligation, using a DNA ligase; and ii) sequencing by synthesis, using a DNA polymerase. Sequencing by synthesis is the most commonly used approach and includes Sanger dideoxy sequencing. For sequencing by synthesis next-generation technologies, the distinctions between methods relate to the output signal that is detected when the nucleotide incorporation occurs. We will review Sanger first generation DNA sequencing and the most common and widely used massively parallel next-generation sequencing technologies. Detailed reviews of additional next-generation sequencing technologies are available [15, 16].

Sanger dideoxy sequencing. Sanger dideoxy sequencing [17] remains the gold standard for DNA sequencing due to its high accuracy and read length of ~1 kb; however, the cost, time and scalability of Sanger sequencing make it unfeasible for large-scale sequencing. Originally, Sanger dideoxy sequencing was developed using radiolabeled chain terminating dideoxy nucleotides (ddNTPs) that were individually included in four separate sequencing reactions along with normal unlabeled deoxynucleotides and when incorporated would stop the polymerization reaction; the dideoxy nucleotides competitively inhibit the synthesis reaction of DNA polymerase I. The four separate polymerization reactions were electrophoresed through polyacrylamide gels and the amplified template fragments migrate by molecular weight due to differences in the number of nucleotides the primer was extended. The fragments that stopped first due to the addition of a given ddNTP would be shorter and migrate faster, while longer fragments would migrate more slowly; in this way the sequence of the DNA template could be deduced. Modifications of Sanger dideoxy sequencing came with the utilization of four-color fluorescently-labeled dideoxy nucleotides instead of radioactive ones. These allowed for all four chemistries and capillary electrophoreses to be run simultaneously in the same lane using laser detection to determine the interrogated base; when the truncated fragments pass through the sequencer, the laser excites the fluorophore and the signal of each of the four fluorophores is detected and recorded in a chromatogram.

Sequencing by Oligonucleotide Ligation and Detection. Library amplified fragments are bound through adaptors to a sequencing flow cell slide. Sequencing by ligation is initiated by the hybridization of a first of five universal primers and then by adding a pool of fluorescently labeled 8-mer oligonucleotides that are labeled depending on their two last base pairs. This produces sixteen different dinucleotide combinations labeled by four different fluorophores on their 5' end. During the sequencing reaction, only the oligonucleotide that is complementary to the template strand will hybridize, bind and be ligated to the nascent strand by a DNA ligase. Four-color imaging is performed by exciting each of the fluorophores and detecting the fluorescent signal across all the spots in the flow cell slide. Silver ions are flushed in order to cleave the recently ligated oligonucleotides releasing the fluorophore and leaving the 5'-PO end of the oligonucleotide free to bind the next one. The cycle is repeated nine times, after which the universal primer with the extended strand is stripped. The next universal primer is used to start the next cycle of sequencing, which again is repeated five times. There are five universal primers used. The sequencing is said to be performed in color space, which for downstream analysis requires mapping to a color-space reference sequence in order to infer the nucleotide sequence [18, 19].

Pyrosequencing. After emulsion PCR library amplification, the beads are arrayed into a picotiter plate (PTP) that contains millions of micro wells large and deep enough (44 um x 55 um) only to hold a single bead containing a single amplified molecule per well. Smaller beads with sulphurylase and luciferase enzymes attached, necessary for the later pyrosequencing reactions, are flushed and allowed to diffuse into the wells and cover the target beads. Each of the dNTPs is individually flushed one at a time through the PTP and they diffuse through each of the sequencing wells. When the DNA polymerase incorporates a nucleotide, a pyrophosphate group is released which will be converted by sulphurylase into ATP to phosphorylate luciferase into luciferin. The light produced by luciferin due to the specific incorporation of a nucleotide in that cycle will be recorded by the CCD camera in the machine, producing an output known as a flowgram or pyrogram. The height of the peak is proportional to the bioluminescence signal intensity which in turn is proportional to the number of incorporated nucleotides in that cycle. However, this is both an advantage and disadvantage of the system, as the specificity of the incorporated base is greater but the detector can be saturated by the signal if more than 6 nucleotides are incorporated in the same cycle, making it inaccurate for sequencing homopolymer tracts. Between cycles, there is a wash with apyrase, an enzyme that degrades any remaining unincorporated nucleotides and ATP produced from the previous cycle [20-23].

Reversible terminators. Sequencing by synthesis using reversible terminators uses clusters generated by bridge amplification on an eight lane flow cell slide. The sequencing cycle starts with flushing a mixture of four fluorescently labeled 3'-modified nucleotide terminators and an engineered DNA polymerase that is able to incorporate these modified nucleotides. If the nucleotide is complementary to the next base in the primed template, it will be added by the polymerase; the extension will be blocked and the fluorescent signal derived by laser excitation of each of the fluorophores will be detected by a high resolution camera. After imaging, the terminating group of the modified nucleotides is cleaved along with the fluorophore allowing the regeneration of the 3'-OH for the addition of the next specific nucleotide and starting a new cycle. The presence of this terminator is key to this technology's chemistry as it ensures that no additional or nonspecific nucleotides are added in the same cycle, allowing that just one nucleotide per template is imaged per cycle. All the four-color images are processed in order to derive the actual nucleotide sequence [24, 25].

Semiconductor Ion Sequencing. The most recent next-generation sequencing by synthesis technology is based on detecting the hydrogen ion that is released during the DNA synthesis reaction by a very sensitive pH meter – a microchip sensor. After template amplification by emulsion PCR, template bound acrylamide beads are loaded into the semiconductor chip's wells. Nucleotides are allowed to flow through the chip one at a time. When the DNA polymerase incorporates the next complementary dNTP, the reaction produces pyrophosphate and hydrogen due to the hydrolysis of the triphosphate of the incorporating nucleotide. The hydrogen ion released produces a change in pH proportional to the number of dNTPs incorporated in that given nucleotide flow cycle that can be detected by a tantalum oxide coated sensor, which provides increased proton sensitivity. The 0.02 pH change per nucleotide that is incorporated is registered by the sensor, then converted into a voltage signal that is finally digitalized to a sequence output [26, 27].

Single molecule sequencing. The next-next-generation of sequencing technologies involves single-molecule sequencing. The first of these technologies performs real-time single-molecule sequencing, in which individual DNA polymerases are attached to the bottom of nanophotonic platforms (zero-mode waveguide, ZMW detectors) that can sensitively detect the binding of fluorescently phospho-linked dNTPs to the nascent strand in real time. The template DNA is diluted so that only one molecule will be sequenced by one polymerase in each of the wells. When the nucleotide is in the active site of the polymerase, a pulse of fluorescence in the specific wavelength is detected by the ZMW detector. Once the new correct nucleotide is covalently bound by the DNA polymerase, the fluorophore is released, the pulse ends and the recently incorporated nucleotide is left free for the next dNTP incorporation. The processivity of the DNA polymerase allows the sequencing of several hundreds of base pairs using this technology [28, 29].

2) **Analysis.** Although none of the next-generation sequencing technologies has reached the accuracy of classic Sanger dideoxy sequencing, they compensate by several fold redundancy of sequencing and essentially oversampling of the same genomic region thereby reducing the

noise and error background. However, these massive amounts of data generated by the next-generation sequencing technologies pose different analytical challenges as current algorithms must process information from millions of short sequence reads and deduce variant information contained in these in the context of a complex genome.

After the chemistries and ‘wet bench’ sequencing process, the data generated by the different technologies is exported into sequence files which generally contain the sequence of each of the reads generated plus some encoded quality information for that read. These sequence files are assembled into contiguous genomic sequence and mapped to the reference genome sequence.

Through the use of current next-generation sequencing technologies, most of the human genome sequencing projects are in fact re-sequencing projects, meaning that they rely on a haploid reference genome sequence assembly to map and align the sequence reads produced from any individual personal genome and identify variants determined by differences from the haploid human genome reference. Because of the inability to assemble an entire genome from short read sequences without a reference scaffold, all the individual sequencing read data points that do not map to the reference genome used are generally discarded along with duplicate and low quality reads.

There are different alignment algorithms which can be used for mapping sequence reads to the reference genome. Mapping algorithms vary in their approaches and how exhaustive their mapping is, which reflects both the accuracy and computational speed with which they can be implemented. Alignment algorithms can be broadly divided into those that build a ‘hash’ or associative array of either the reference genome or the sequence reads to use as seeds or anchors for the alignment, once the seeds have been aligned to the reference genome, a smaller local Blast–like alignment is performed in order to extend the alignment and ensure more accurate mapping. The second group of algorithms is formed by those that utilize the Burrows-Wheeler transform (BWT) algorithm in which the reference genome is sorted, reordered, and indexed for more efficient access and read alignment, which makes these algorithms faster [30, 31]. The output of these algorithms is generally a sequence alignment/map (SAM) file. This file is generally very large as it contains the mapping of each read to the reference and its quality. Therefore, for better and more efficient handling of this information, SAM files are converted into binary alignment files (BAM) which can be more readily accessed, read and handled by the computer [32]. From the BAM file, information about each base in the genome can be extracted and this is deposited into a Pileup file. This file reports for every position in the genome the base observed by the pilling up of all the reads at that specific position (Figure 2). However, the majority of the bases in the genome will be identical to the reference, therefore of main interest are those positions that are different. After obtaining the pileup for the whole genome, the variable positions are extracted into another file in a process known as variant calling. It is important to evaluate these variant calls for their quality, number of reads across the position and number of reads that called the variant, strand bias, and likelihood of being a true variant, in order to ensure that the majority are true positives. The quality of a variant call can be generally assessed by its pileup information. There are several algorithms that perform variant calling from next-generation sequencing data, they provide genotype data and can be used for quality filtering [33-35]. Once variants are "called', these can be analyzed and filtered through different approaches and using different criteria depending on the purpose of the study. In general, whether the variants detected map

PeerJ PrePrints | https://peerj.com/preprints/147v1/ | v1 received: 9 Dec 2013, published: 9 Dec 2013, doi: 10.7287/peerj.preprints.147v1

to genic or intergenic regions, if they are coding or non-coding, as well as whether they represent previously observed polymorphisms or novel variants, as well as the allele frequency spectrum in populations are all important parameters to consider. This is achieved through a process known as variant annotation. When annotating variants, several available databases are queried in order to gather as much information as possible regarding that specific genomic position or coordinate in order to assess the functional impact of the identified variant. Some of the most widely used databases included in several annotation pipelines are:

Population Polymorphism Databases: The database of single nucleotide polymorphisms (**dbSNP**) was established in 1998 by the National Center for Biotechnology Information (NCBI) in order to store and catalogue single nucleotide polymorphisms (SNPs) as the most common form of genetic variation between individuals [36, 37]. Since its creation, the database has been greatly expanded to include simple nucleotide variants (SNVs) both SNPs and Indels. Initially the database was populated by the SNPs discovered during the HGP, later by the additional variants discovered during the HapMap Project, and most recently by variants of the Thousand Genomes Project and the myriad of whole genome and exome sequencing projects of the last lustrum. The **1000 (Thousand) Genomes Project (TGP)** was initiated with the purpose of cataloguing most of the polymorphic and low-frequency variation amongst human genomes, including SNVs and copy-number variants, by sequencing >1000 genomes and exomes of different populations around the world using NGS technologies [38-40]. The most recent data release of the **NHLBI Exome Sequencing Project (ESP6500)** contains SNP variants identified through whole exome sequencing of 6503 samples from multiple NHLBI ESP cohorts [41]. In addition, most large-scale sequencing groups use their own internal variant databases to annotate for in-house variant frequencies and genotypes observed, which helps to control for technical replicate errors and specific population variation.

Conservation scores: **PhyloP** computes p-values of nucleotide conservation based on a tree model of neutral evolution and multi-species alignments [42, 43]. The likelihood ratio test (**LRT**) considers all possible ancestral sequences to estimate the 'deleteriousness' of a particular substitution based on a comparative genomic approach across 32 vertebrate species and assuming neutrality from synonymous changes and treating all nucleotide substitutions equally [44]. The Genomic Evolutionary Rate Profiling (**GERP**) approach aims to identify evolutionary constrained regions that have lower nucleotide substitution rates, which may reflect past purifying selection, through the sequence analysis and alignments of 29 mammalian species [45]. **PhastCons** attempts to identify evolutionarily conserved regions through multiple species alignments (46 placental mammals) based on a phylogenetic Hidden Markov Model (phylo-HMM) that uses statistical models for unequal nucleotide substitution rates [46].

Functional prediction algorithms: The 'Sorting Tolerant From Intolerant' (**SIFT**) algorithm predicts the functional effect of an amino acid substitution based on the conservation of that amino acid residue in the protein through multiple sequence alignment of closely related proteins from PSI-BLAST [47]. The scores and predictions given by SIFT range from (1 = tolerated to 0 = damaging); the amino acid change is predicted to be damaging if the score <= 0.05, and tolerated if the score > 0.05.

The ‘Polymorphism Phenotyping’ (**PolyPhen**) algorithm predicts the possible functional impact of an amino acid substitution based on the protein structure, phylogenetic conservation and sequence information calculating a naive Bayes posterior probability that the mutation might be damaging [48]. The score reported by PolyPhen reports the probability of the mutation being damaging when it is not damaging over the probability of the mutation being damaging when it is actually damaging; therefore, the scores range from ‘problably damaging’ (score=0) to ‘benign’ (score=1).

MutationTaster utilizes a Bayes probabilistic algorithm to predict the functional impact of a given nucleotide change, either SNPs or small indels, based on a training set of known disease causing mutations and common polymorphisms. This algorithm calculates the probability of the change being a polymorphism or a damaging mutation and reports back p-values (not scores) of the prediction being correct [49].

Disease/phenotype related databases: The Online Mendelian Inheritance in Man (**OMIM**) database is a compendium of human diseases and phenotypes of genetic or suspected genetic etiology [50]. To date there are about 3,700 genetic phenotypes or diseases for which the molecular cause is known; however there are still at least ~4,000 phenotypes with suspected genetic basis for which the responsible gene(s) remain unknown. The Human Gene Mutation Database (**HGMD**) is a catalogue of known specific mutations reported to be associated with particular genetic diseases or phenotypes [51]. Currently, HGMD includes more than 134,000 single mutations associated to genetic diseases. Unfortunately, some of these mutations are based on single case reports with insufficient support for pathogenicity of the mutation; new exome data in thousands of individuals support the contention that such variants are not damaging or causing the disease they were reported to be associated with as they reach a certain frequency (>2 – 5%) within normal populations.

The Pharmacogenomics Knowledge database (**PharmGKB**) documents reported, well characterized variants and polymorphisms related to the metabolism of a wide variety of drugs and compounds. It is a valuable resource to inform analysis of potential medically actionable variants associated with drug metabolism and suggested dosage and guidelines for the utilization of common medications depending on an individual's genotype [52].

Other databases and resources: Other useful information resources to interpret variants in novel genes can be provided by pathway and protein-protein interaction network databases, such as the Kyoto Encyclopedia of Genes and Genomes (**KEGG**) [53, 54] for biological pathways, and the Human Integrated Protein-Protein Interaction rEference (**HIPPIE**) [55] or the Search Tool for the Retrieval of Interacting Genes/Proteins (**STRING**) [56] databases for protein-protein interactions data. These databases can provide information on which gene products directly interact with each other or function upstream or downstream of other known disease associated genes and help elucidate possible functions and roles of novel genes. Additionally, one would like to be informed of expression profiles of the genes of interest across different tissues or throughout development [57, 58]; and phenotypes of mutant model organisms for the genes of interest [59].

Array Comparative Genomic Hybridization (aCGH) as the gold standard for Copy-Number Variant detection

Next-generation sequencing is allowing the identification of the vast majority of simple nucleotide variants (SNVs) in personal genomes, from single nucleotide changes to small insertion/deletion variants. However, one additional, larger in scale and highly important source of variation in genomes, both polymorphic and rare is structural and copy number variation (CNV). Structural variation refers to segments of the genome that differ in copy-number, architecture and/or orientation between genomes; within this, copy-number variants (CNVs) are regions in the genome that are present in more or less copies than the expected diploid state. It is now widely recognized that structural variation and CNVs contribute largely to human genomic variation, both benign polymorphic and pathogenic [60].

Array Comparative Genomic Hybridization (aCGH) remains the gold standard to identify CNVs in the genome. It is based on the competitive hybridization between a test or proband's DNA and a control DNA. First, the samples can be digested (or not, depending on the protocol) using an *Alu*I/*Rsa*I enzyme mix in order to digest the DNA into fragments of ~500 bp in size. Labeling is performed by priming with random nonamers and then adding fluorescent cyanine dyes, either Cy3 or Cy5, to the test DNA and the control DNA, respectively, that are incorporated by a nick-translation reaction using a Klenow fragment polymerase. Competitive hybridization is then performed on an array platform which contains probes to interrogate the whole genome or specific regions for several hours depending on the protocol. After extended hybridization, the array is washed and scanned using lasers that excite each of the fluorophores and the signal intensity of each probe in the array is measured. The images obtained can then be processed, merged, and analyzed. If the signal intensity for a particular probe is greater when exciting the fluorophore of the test DNA, it indicates that more test DNA was bound and there is a copy-number gain of that probe in the test DNA. Conversely, if the signal intensity for a different probe is greater when exciting the flourophore of the control DNA, then this indicates a copy-number loss of that region in the test DNA. If the signal intensity for both fluorophores is about the same for a given probe, then it is assumed that there is no copy-number change between the test and control DNAs for the region being interrogated by the oligonucleotide probe [61, 62].

Initially, the grid array platforms used BACs to tile the genome and interrogate for CNVs, however this made the estimation of CNV size and boundaries difficult, inaccurate and overestimated. With the development of the aCGH platforms, and a reference human genome sequence to design probes, came the tiling of the genome using covalently bound oligonucleotide (~60mers) probes that could be chosen to tile the whole genome or specific regions at higher resolution; the major constraint in genome resolution afforded being the number of probes utilized per array design; i.e. the number of ‘pixels’ used for genome resolution. These currently range from as few as 60,000 probes to as many as 4.2 million probes depending on the platform used. The performance of the oligonucleotide probes can vary depending on their GC content and therefore it is usually necessary to have a change in the same direction in the signal of at least five continuous probes in order to be able to detect a signal representing a true CNV by aCGH.

aCGH has proven to be very accurate and successful at identifying large CNVs, however it has its limitations including: i) not being able to resolve CNVs of less than 5kb genome-wide, ii) relying on a “control” DNA, which by itself will have CNVs of its own of unknown significance, iii) not being able to detect copy-number changes of more than 4 due to the signal’s dynamic range, iv) relying on a reference sequence on which the array design is based and not being able to detect other types of structural variation such as insertions, inversions and balanced translocations. Currently, copy number variation can be deduced from NGS data based on the coverage distribution of single reads across the genome or target regions and the difference between the number of reads across (i.e. read-depth-of-coverage) a given region and the genome-wide average. Alternatively, CNVs and some structural variants can be inferred from NGS data through the library preparation, sequencing and analysis of paired-end reads. Paired-end sequencing (PE) reads are generated from both ends of a fragment. Because the distance between the two ends is known based on the reference genome assembly, it is possible to use this information to estimate the presence of insertion, deletions, or copy number variants in the region [63]. As it is also the case for aCGH, inference of deletions is in general more robust than the identification of copy-number gains. In addition, sequencing data can potentially provide CNV breakpoint data that can help in further understanding the mechanisms of CNV formation in the human genome [64].

It is anticipated that eventually genomic sequencing using next-generation technologies will provide accurate information on smaller CNVs, in addition to the larger CNVs readily detected by aCGH, and potentially other structural variants like inversions and novel sequence insertions. However for that to occur, current and novel algorithms for alignment mapping and *de novo* assembly need to be improved and developed. The interpretation of mechanisms for generating complex rearrangements and deducing their structure from short read sequences, and *de novo* assembly algorithms for larger genomic segments and entire genomes, will have to await further developments and refinements.

Single Nucleotide Polymorphism (SNP) detection for genome-wide genotyping and CNV detection

One of the initial types of genomic variation that was attainable for genome-wide testing were Single Nucleotide Polymorphisms (SNPs), which by definition are, in most cases, bi-allelic positions in the genome that differ in one nucleotide among individuals and are present in at least 1% of the general population. SNP genotyping has been widely used for genome-wide association studies (GWAS) with the intention to find common SNP associations to common complex diseases. SNP detection platforms perform allelic discrimination to interrogate the polymorphic position through different approaches, which can be primer extension by single base incorporation, mismatch hybridization, ligation, and enzymatic cleavage [65]. The primer extension approach can utilize a common primer that can detect either allele or allele-specific primers; primer anneals to the contiguous region next to the interrogated SNP and the nucleotide corresponding to the SNP is incorporated in an allele-specific PCR reaction and identified by either mass spectrometry or fluorescence. In the mismatch hybridization approach, allele-specific oligonucleotide probes are printed and arrayed in a solid support. Genomic DNA is digested, PCR amplified, fragmented, labeled and hybridized to the array. Each SNP is independently tested by several probes that differ just at the SNP position and can discriminate between each of the two alleles by fluorescence detection. Another mismatch based approach utilizes allele-specific primers that have a fluorophore and a quencher and a common primer for PCR amplification. During the primer extension step, the polymerase with 5' exonuclease activity can cleave the perfectly matched primer freeing the reporter fluorophore from the quencher's proximity and genotype is detected by the flourescent signal emitted. Other mismatch based approaches using allele-specific primers tagged differently have also been developed in order to detect SNP genotypes by other methods besides fluorescence, such as mass spectrometry or flow cytometry. For ligation-based methods, two allele-specific oligonucleotide probes are used in addition to a common oligonucleotide that binds adjacently to the SNP site. When one of the allele-specific probes binds to the SNP site perfectly, a DNA ligase will ligate the specific probe with the common oligonucleotide. The ligated allele-specific products can then be detected depending on what was used to tag the different probes, most commonly fluorescent dyes. A variation of this approach is using longer linear oligonucleotide probes whose ends are equivalent to the allele-specific and the common probes. The approach is the same, but when allele-specific binding and ligation occurs the probe gets circularized. This circular probe can then be amplified by rolling circle DNA replication or PCR; the genotype can be 'called' using fluorescently labeled primers or by fluorescent labeling during the amplification step with subsequent array hybridization. Enzymatic cleavage methods rely on the specificity of DNA endonucleases to recognize specific (which for SNP genotyping are allele-specific) sequences and cleave the DNA evidencing the genotype. Region of interest PCR amplified products can be incubated with specific restriction enzymes and genotypes can be detected by differences in fragment sizes. However, this method is generally low-throughput and limited by the nucleases sequence recognition repertoire. A variation of the enzymatic cleavage approach uses two fluorescently labeled allele-specific probes with an additional common “invader” probe that is complementary to the 3’ end of the SNP region. When hybridization to the SNP site occurs, the presence of the invader probe creates an overhang of the allele-specific oligonucleotide that is recognized by a nuclease, which cleaves it and releases the fluorophore for genotype detection.

Currently, SNP detection platforms are mostly array based and aimed for high-throughput, genomewide genotyping with varying number of SNP test probes depending on the design of preference. These methodologies also use a combination of two or more of the previously described allele discrimination approaches. Additionally, current SNP arrays designs also include CNV detection probes or inversely, aCGH designs include SNP array probes which allow for the detection of both CNVs and absence of heterozygosity (AOH) in the test sample [66].

SNP arrays genotype data can be analyzed based on the B-allele frequency (BAF), which in a diploid state presents three possible states namely AA=0, AB=0.5, and BB= 1. These data can also be visualized in a B-allele frequency plot. When the BAF deviates from the tri-modal expected distribution, this can indicate allelic imbalance that can relate to genomic events such as copy number variants, regions of absence of heterozygosity, or uniparental disomy, in the same assay [67, 68].

Conclusions

No current genome-wide technology or platform provides information about all types of variation, from SNVs to structural variation including small CNV in the 100bp to 5kb range, at high resolution with high specificity [63]. However, continued developments in sequencing technologies and analysis promise to deliver better, faster and more high-throughput sequencing technologies to assess the complete picture of human genomic variation, spearheading the development of improved and new methods for data analysis, not only to process the peta (10^^{15}) amounts of data produced, but also to make biological sense of this information. The experimental and technical approaches for human genomic variation discovery will most probably not be the main limiting factor in assessing it, but instead our understanding and our ability to derive biologically relevant lessons and conclusions from such massive data will remain the premier challenge.

References:

1. http://www.genome.gov/11006943

2. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA; Direct selection of human genomic loci by microarray hybridization; *Nat Methods.* 2007 Nov;4(11):903-5.

3. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME; Microarray-based genomic selection for high-throughput resequencing; *Nat Methods.* 2007 Nov;4(11):907-9.

4. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR; Genome-wide in situ exon capture for selective resequencing; *Nat Genet.* 2007 Dec;39(12):1522-7.

5. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, Richard McCombie W, Hannon GJ; Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing; *Nat Protoc.* 2009;4(6):960-74.

6. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C; Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing; *Nat Biotechnol.* 2009 Feb;27(2):182-9.

7. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, Kitzman J, Wu YQ, Newsham I, Richmond TA, Jeddeloh JA, Muzny D, Albert TJ, Gibbs RA; Whole exome capture in solution with 3 Gbp of data; *Genome Biol.* 2010;11(6):R62.

PeerJ PrePrints | https://peerj.com/preprints/147v1/ | v1 received: 9 Dec 2013, published: 9 Dec 2013, doi: 10.7287/peerj.preprints.147v1

8. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR; A comparative analysis of exome capture; *Genome Biol.* 2011 Sep 29;12(9):R97.

9. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ; Target-enrichment strategies for next-generation sequencing; Nat Methods. 2010 Feb;7(2):111-8.

10. Hardenbol P, Banér J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U, Davis RW; Multiplexed genotyping with sequence-tagged molecular inversion probes; Nat Biotechnol. 2003 Jun;21(6):673-8.

11. Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, Brundage T, Boudreau A, Chow S, Eberle J, Erbilgin A, Falkowski M, Fitzgerald R, Ghose S, Iartchouk O, Jain M, Karlin-Neumann G, Lu X, Miao X, Moore B, Moorhead M, Namsaraev E, Pasternak S, Prakash E, Tran K, Wang Z, Jones HB, Davis RW, Willis TD, Gibbs RA; Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay; Genome Res. 2005 Feb;15(2):269-75.

12. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J; Multiplex amplification of large sets of human exons; *Nat Methods.* 2007 Nov;4(11):931-6.

13. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J; Massively parallel exon capture and library-free resequencing across 16 genomes; *Nat Methods.* 2009 May;6(5):315-6.

14. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, Munson J, Hiatt JB, Turner EH, Levy R, O'Day DR, Krumm N, Coe BP, Martin BK, Borenstein E, Nickerson DA, Mefford HC, Doherty D, Akey JM, Bernier R, Eichler EE, Shendure J;

Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders; *Science* 2012 Dec 21;338(6114):1619-22.

15. Mardis ER; Next-generation DNA sequencing methods; Annu Rev Genomics Hum Genet. 2008;9:387-402.

16. Metzker ML; Sequencing technologies - the next generation; Nat Rev Genet. 2010 Jan;11(1):31-46.

17. Sanger F, Nicklen S, Coulson AR; DNA sequencing with chain-terminating inhibitors; PNAS 1977 Dec;74(12):5463-7.

18. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM; Accurate multiplex polony sequencing of an evolved bacterial genome; Science. 2005 Sep 9;309(5741):1728-32.

19. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, et al; Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding; Genome Res. 2009 Sep;19(9):1527-41.

20. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P; Real-time DNA sequencing using detection of pyrophosphate release; Anal Biochem. 1996 Nov 1;242(1):84-9.

21. Ronaghi M, Uhlén M, Nyrén P; A sequencing method based on real-time pyrophosphate; Science. 1998 Jul 17;281(5375):363, 365.

22. 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; Genome sequencing in microfabricated high-density picolitre reactors; Nature. 2005 Sep 15;437(7057):376-80.

23. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM; The complete genome of an individual by massively parallel DNA sequencing; Nature. 2008 Apr 17;452(7189):872-6.

24. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ; Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators; Proc Natl Acad Sci U S A. 2006 Dec 26;103(52):19635-40.

25. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al; Accurate whole human genome sequencing using reversible terminator chemistry; Nature. 2008 Nov 6;456(7218):53-9.

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;

An integrated semiconductor device enabling non-optical genome sequencing; Nature. 2011 Jul 20;475(7356):348-52.

27. Merriman B, R D Team IT, Rothberg JM; Progress in Ion Torrent semiconductor chip based sequencing; Electrophoresis. 2012 Dec;33(23):3397-417.

28. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S; Real-time DNA sequencing from single polymerase molecules; Science. 2009 Jan 2;323(5910):133-8.

29. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA; Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology; PLoS One. 2012;7(11):e47768.

30. Flicek P, Birney E; Sense from sequence reads: methods for alignment and assembly; Nat Methods. 2009 Nov;6(11 Suppl):S6-S12.

31. Bao S, Jiang R, Kwan W, Wang B, Ma X, Song YQ; Evaluation of next-generation sequencing software in mapping and assembly; J Hum Genet. 2011 Jun;56(6):406-14.

32. 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 Aug 15;25(16):2078-9.

33. Nielsen R, Paul JS, Albrechtsen A, Song YS; Genotype and SNP calling from next-generation sequencing data; Nat Rev Genet. 2011 Jun;12(6):443-51.

34. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, Yu F; A SNP discovery method to assess variant allele probability from next-generation resequencing data; Genome Res. 2010 Feb;20(2):273-80.

35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA; The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data; Genome Res. 2010 Sep;20(9):1297-303.

36. Sherry ST, Ward M, Sirotkin K; dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation; Genome Res. 1999 Aug;9(8):677-9.

37. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K; dbSNP: the NCBI database of genetic variation; Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

38. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X, Chen Y, Challis D, Clarke L, Ball EV, Cibulskis K, Cooper DN, Fulton B, Hartl C, Koboldt D, Muzny D, Smith R, Sougnez C, Stewart C, Ward A, Yu J, Xue Y, Altshuler D, Bustamante CD, Clark AG, Daly M, DePristo M, Flicek P, Gabriel S, Mardis E, Palotie A, Gibbs R; 1000 Genomes Project; The functional spectrum of low-frequency coding variation; Genome Biol. 2011 Sep 14;12(9):R84.

39. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HY, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stütz AM, Urban AE,

Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO; 1000 Genomes Project; Mapping copy number variation by population-scale genome sequencing; Nature. 2011 Feb 3;470(7332):59-65.

40. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA; An integrated map of genetic variation from 1,092 human genomes; Nature. 2012 Nov 1;491(7422):56-65.

41. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/)

42. Siepel A, Pollard KS, Haussler D; New methods for detecting lineage-specific selection. Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB) 2006: 190–205.

43. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al; A high-resolution map of human evolutionary constraint using 29 mammals; Nature. 2011 Oct 12;478(7370):476-82.

44. Chun S, Fay JC; Identification of deleterious mutations within three human genomes; Genome Res. 2009 Sep;19(9):1553-61.

45. Cooper GM, Stone EA, Asimenos G; NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A; Distribution and intensity of constraint in mammalian genomic sequence; Genome Res. 2005 Jul;15(7):901-13.

46. Margulies EH, Blanchette M; NISC Comparative Sequencing Program, Haussler D, Green ED; Identification and characterization of multi-species conserved sequences; Genome Res. 2003 Dec;13(12):2507-18.

47. Kumar P, Henikoff S, Ng PC; Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm; Nat Protoc. 2009;4(7):1073-81.

48. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR; A method and server for predicting damaging missense mutations; Nat Methods. 2010 Apr;7(4):248-9.

49. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D; MutationTaster evaluates disease-causing potential of sequence alterations; Nat Methods. 2010 Aug;7(8):575-6.

50. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), URL: http://omim.org/

51. Cooper DN, Ball EV, Krawczak M; The human gene mutation database; Nucleic Acids Res. 1998 Jan 1;26(1):285-7.

52. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE; Pharmacogenomics Knowledge for Personalized Medicine; Clinical Pharmacology & Therapeutics 2012; 92(4): 414-417.

53. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M; KEGG: Kyoto Encyclopedia of Genes and Genomes; Nucleic Acids Res. 1999 Jan 1;27(1):29-34.

54. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M; KEGG for integration and interpretation of large-scale molecular data sets; Nucleic Acids Res. 2012 Jan;40(Database issue):D109-14.

55. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA; HIPPIE: Integrating protein interaction networks with experiment based quality scores; PLoS One. 2012;7(2):e31826.

56. Search Tool for the Retrieval of Interacting Genes/Proteins (**STRING**): http://string-db.org/

57. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI; BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources; Genome Biol. 2009;10(11):R130.

58. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A; Gene expression atlas at the European bioinformatics institute; Nucleic Acids Res. 2010 Jan;38(Database issue):D690-8.

59. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE; the Mouse Genome Database Group; The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse; Nucleic Acids Res 2012; 40(1):D881-86.

60. Stankiewicz P, Lupski JR; Structural variation in the human genome and its role in disease; Annu Rev Med. 2010;61:437-55.

61. Pinkel D, Albertson DG; Comparative genomic hybridization; Annu Rev Genomics Hum Genet. 2005;6:331-54.

62. Brady PD, Vermeesch JR; Genomic microarrays: a technology overview; Prenat Diagn. 2012 Apr;32(4):336-43.

63. Alkan C, Coe BP, Eichler EE; Genome structural variation discovery and genotyping; Nat Rev Genet. 2011; 12(5):363-76.

64. Onishi-Seebacher M, Korbel JO; Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond; Bioessays. 2011; 33(11):840-50.

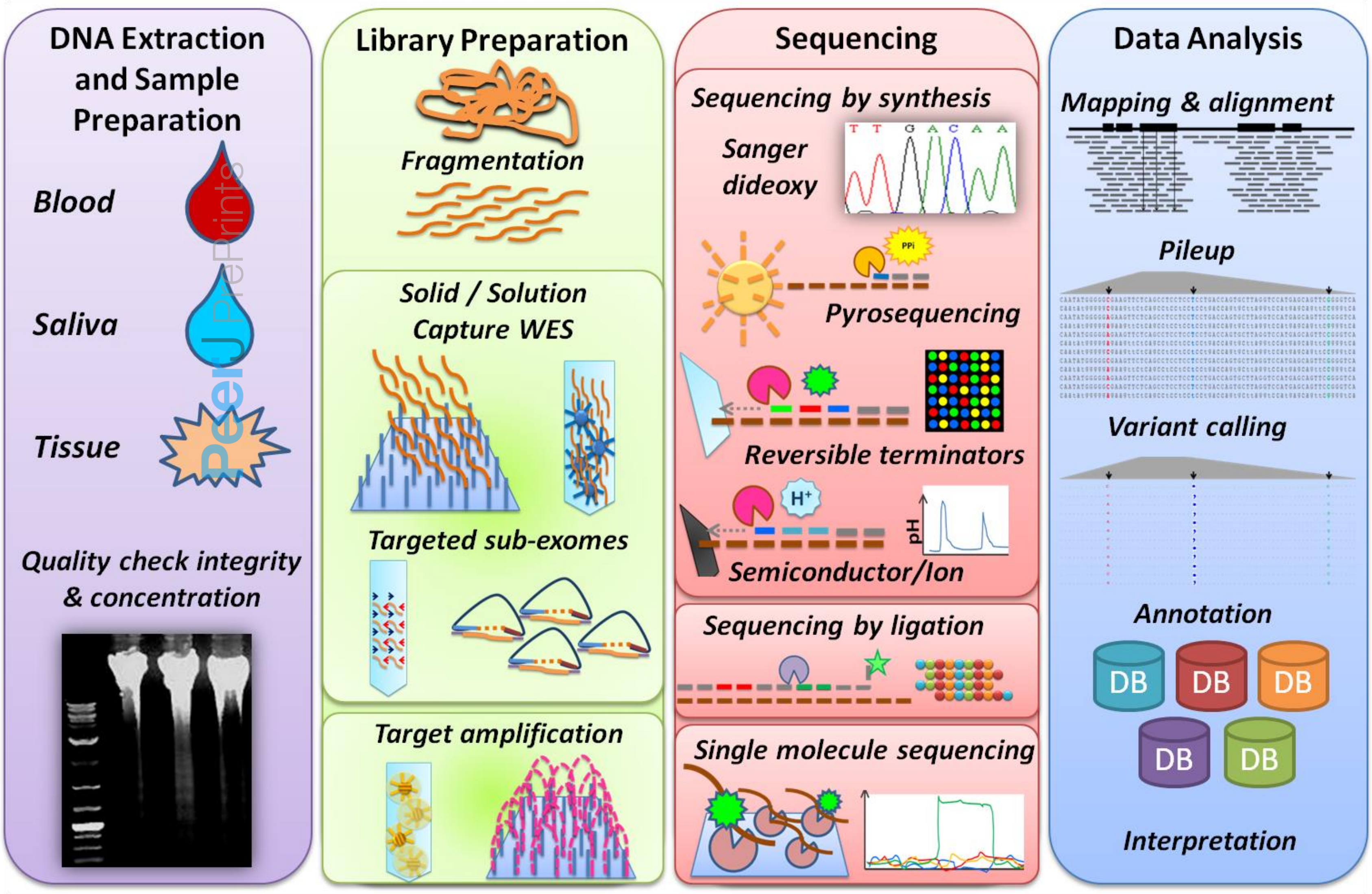
65. Kim S, Misra A; SNP genotyping: technologies and biomedical applications; Annu Rev Biomed Eng. 2007;9:289-320.

66. Schaaf CP, Wiszniewska J, Beaudet AL; Copy number and SNP arrays in clinical diagnostics; Annu Rev Genomics Hum Genet. 2011 Sep 22;12:25-51.

67. Bruno DL, White SM, Ganesamoorthy D, Burgess T, Butler K, Corrie S, Francis D, Hills L, Prabhakara K, Ngo C, Norris F, Oertel R, Pertile MD, Stark Z, Amor DJ, Slater HR; Pathogenic aberrations revealed exclusively by single nucleotide polymorphism (SNP) genotyping data in 5000 samples tested by molecular karyotyping; J Med Genet. 2011 Dec;48(12):831-9.

68. Yau C, Holmes CC; CNV discovery using SNP genotyping arrays; Cytogenet Genome Res. 2008;123(1-4):307-12

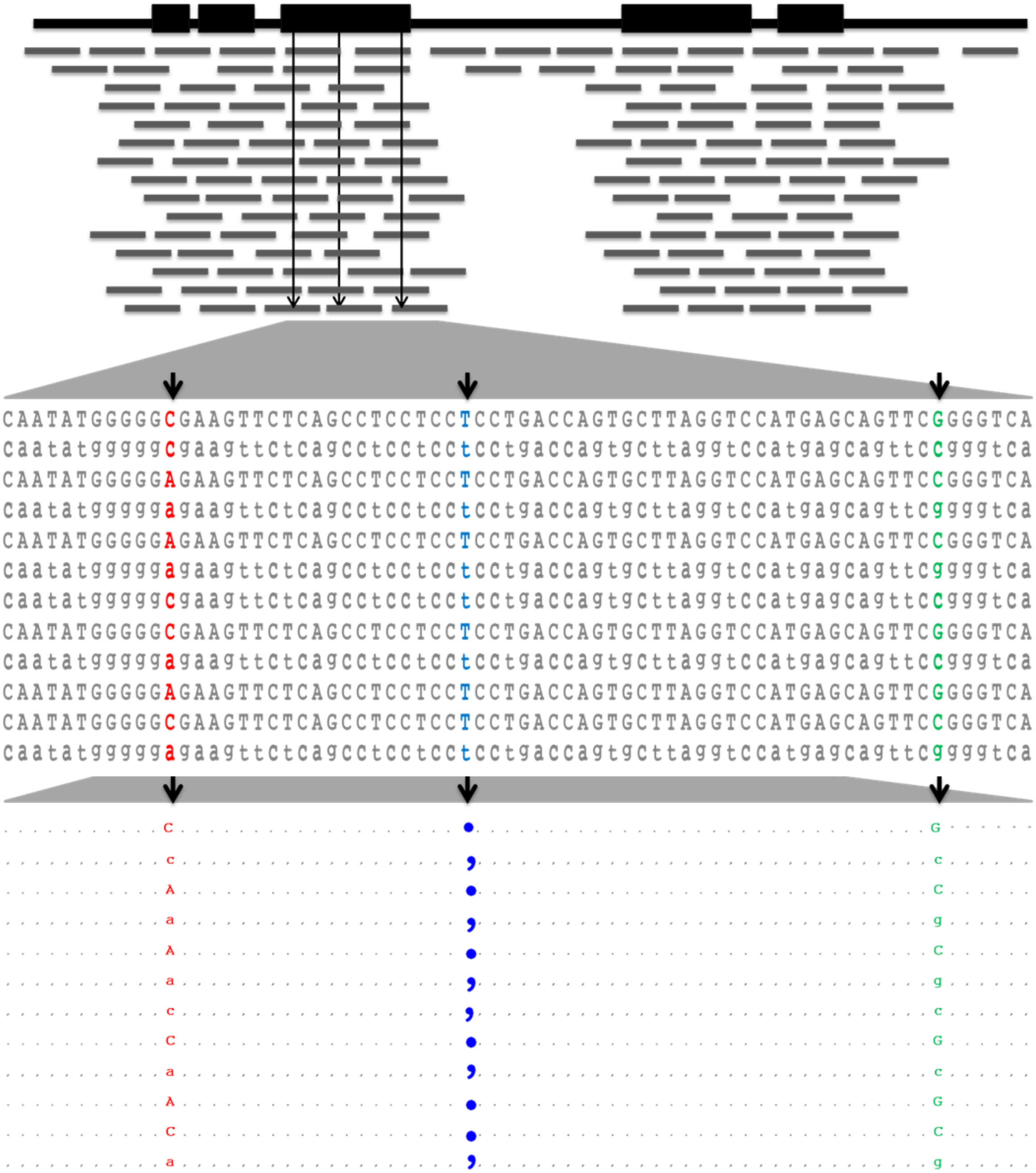
687 **Figure 1**



PeerJ PrePrints | https://peerj.com/preprints/147v1/ | v1 received: 9 Dec 2013, published: 9 Dec 2013, doi: 10.7287/peerj.preprints.147v1

Figure 1. Overview of the methods used for whole-genome/exome sequencing for human genome variation analyses. 1) DNA extraction from sources such as blood, saliva or tissues is performed. 2) DNA is fragmented and samples are prepared for whole-genome sequencing or target enrichment by a capture method that can be on a solid surface (array) or in solution. Alternatively, for more targeted approaches, PCR or molecular inversion probes (MIPs) can be used for target enrichment. 3) A variety of sequencing technologies are available. These can be subdivided according to the enzyme they use to amplify the target sequences and by the output signal that is detected for sequencing. 4) After sequencing, data generated is mapped and aligned to the human genome reference sequence. A pileup of every base and the nucleotide detected at that position is generated and from the file generated, variants that differ from the reference are extracted and annotated using an extensive variety of databases in order to aid with variant interpretation.

Figure 2.


CAATATGGGGGCGAAGTTCTCAGCCTCCTCCTCCTGACCAGTGCTTAGGTCCATGAGCAGTTCGGGGTCA
caatatgggggcgaagttctcagcctcctcctcctgaccagtgcttaggtccatgagcagttccgggtca
CAATATGGGGGAGAAGTTCTCAGCCTCCTCCTCCTGACCAGTGCTTAGGTCCATGAGCAGTTCCGGGTCA
caatatgggggagaagttctcagcctcctcctcctgaccagtgcttaggtccatgagcagttcggggtca
CAATATGGGGGAGAAGTTCTCAGCCTCCTCCTCCTGACCAGTGCTTAGGTCCATGAGCAGTTCCGGGTCA
caatatgggggagaagttctcagcctcctcctcctgaccagtgcttaggtccatgagcagttcggggtca
caatatgggggcgaagttctcagcctcctcctcctgaccagtgcttaggtccatgagcagttccgggtca
CAATATGGGGGCGAAGTTCTCAGCCTCCTCCTCCTGACCAGTGCTTAGGTCCATGAGCAGTTCGGGGTCA
caatatgggggagaagttctcagcctcctcctcctgaccagtgcttaggtccatgagcagttccgggtca
CAATATGGGGGAGAAGTTCTCAGCCTCCTCCTCCTGACCAGTGCTTAGGTCCATGAGCAGTTCGGGGTCA
CAATATGGGGGCGAAGTTCTCAGCCTCCTCCTCCTGACCAGTGCTTAGGTCCATGAGCAGTTCCGGGTCA
caatatgggggagaagttctcagcctcctcctcctgaccagtgcttaggtccatgagcagttcggggtca
C . G
c , c
A . C
a , g
A . C
a , g
c , c
C . G
a , c
A . G
C . C
a , g

PeerJ PrePrints | https://peerj.com/preprints/147v1/ | v1 received: 9 Dec 2013, published: 9 Dec 2013, doi: 10.7287/peerj.preprints.147v1

Figure 2. Schematic representation of the mapping, pileup and variant calling process. Once individual reads are mapped to the reference human genome sequence, a pileup of these reads is generated and every base reported at each aligned position in the genome is reported. In order to facilitate the processing of these data and the files generated, symbols have been assigned to represent a reference base reported in the forward strand (●), a reference base in the reverse strand (,) and capital and small letters to represent specific variant calls either in the forward or reverse strand respectively..