Published in final edited form as:

Trends Genet. 2013 October ; 29(10): 593–599. doi:10.1016/j.tig.2013.07.006.

Finding the lost treasures in exome sequencing data

David C. Samuels^{1,*}, Leng Han^{2,*}, Jiang Li³, Sheng Quanguhu³, Travis A. Clark⁴, Yu Shyr³, and Yan Guo³

¹Center for Human Genetics Research, Vanderbilt University, Nashville, TN, 37232, USA

²Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, TX, 77030, USA

³Center for Quantitative Sciences, Vanderbilt University, Nashville, TN, 37232, USA

⁴Vanderbilt Technology for Advanced Genomics, Vanderbilt University, Nashville, TN, 37232, USA

Abstract

Exome sequencing is one of the most cost-efficient sequencing approaches for conducting genome research on coding regions. However, significant portions of the reads obtained in exome sequencing come from outside of the designed target regions. These additional reads are generally ignored, potentially wasting an important source of genomic data. There are three major types of unintentionally sequenced read that can be found in exome sequencing data: reads in introns and intergenic regions, reads in the mitochondrial genome, and reads originating in viral genomes. All of these can be used for reliable data mining, extending the utility of exome sequencing. Large-scale exome sequencing data repositories, such as The Cancer Genome Atlas (TCGA), the 1000 Genomes Project, National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project, and The Sequence Reads Archive, provide researchers with excellent secondary data-mining opportunities to study genomic data beyond the intended target regions.

Keywords

mitochondria; exome capture; virus; virus integration; mtDNA copy number; unmapped read

The rise of exome sequencing

Next-generation sequencing (see Glossary) has substantially decreased the cost of sequencing and has become the tool of choice for genomic studies. One of the most popular new sequencing approaches is exome sequencing (Figure 1), in which the coding regions of the full genome are targeted, captured, and sequenced. The exome represents approximately 1–1.5% of the human genome with approximately 50 million bp, but it accounts for over 85% of all mutations that have been identified in Mendelian disorders [1]. As a result, exome sequencing is currently an attractive and practical approach for the investigation of coding variations [2,3].

Targeted resequencing enables the enrichment of specific sequences from a whole-genomic library. Exome sequencing is an example of this approach, whereby the complete coding region of the genome is enriched for sequencing. However, many of the captured DNA

fragments still derive from outside the targeted regions (Figure 2). As a result, intronic and intergenic regions may be sequenced, including promoters, conserved noncoding sequences, untranslated regions (UTR), miRNA target sites, and other potentially functional regions. In a typical exome-sequencing study, approximately 40–60% of the reads are off target [4–6] and all or most of these off-target reads are usually ignored. This practice does not utilize the full potential of exome-sequencing data, because it overlooks a large amount of potentially useful data. Recent studies [5,7–10] have shown that off-target reads can be of good quality and can provide useful insights.

Reads aligned outside the target regions

There are three major exome-sequencing capture kits currently in broad use: Illumina TruSeq, Agilent SureSelect, and NimbleGen SeqCap EZ. All three platforms start with whole-genomic libraries made from fragmented genomic DNA and use biotinylated oligonucleotide baits complementary to the design targets to enrich for exons and other vendor-specific content. The target regions for these three exome capture kits vary and range from 37.6 to 62.1 million bp. The capture kits can enrich just the exome, exons plus 3' and 5' UTRs, and other content. The kits also differ in their target regions, bait length, bait density, and the molecule used for capturing.

Other capture techniques, including array-based, multiplex PCR, selector-probe (HaloPlex), and molecular-inversion probe (MIPs), methods are also available. The capture efficiency varies by capture method. For example, one group [11] (using the NimbleGen 2.1M array-based capture kit) reported having 64.5% of sequenced bases outside the target regions and 31.9% of the reads more than 500 bp away from the target regions; another group [1] (using Agilent 244K microarrays for target enrichment) reported over 50% of sequenced bases outside the target regions. The capture efficiency of the three major exome capture kits has been reported by multiple studies. For Agilent SureSelect, the capture efficiency is between 42% and 58% [4–6]; for Illumina TrueSeq, it is between 45% and 46% [5]; and for NimbleGen SeqCap EZ it is between 50% and 53% [4,6]. Although a capture efficiency of less than 50% can be misinterpreted as failure of the sequencing method, the raw number of reads mapped to the target regions and the median depth of the target regions are more informative parameters to measure the success of the capture method. The unmapped fraction of reads can be anywhere from 5% to 19% [5] and it is related to many factors, such as the type of capture kit, DNA quality, aligner settings, and the completeness of the reference sequence used for the alignment. There is also variability introduced during library preparation and sequencing. Even repeat sequencing of a sample can generate different metrics of capture efficiencies [6].

SNPs outside the exonic regions

Many functional elements are located outside the exonic regions [12–15]. Although the role of introns was unclear for many years, several studies have now established some functional significance for introns [11,16–19]. For example, a study [20] identified two mutations within the core promoter of the telomerase reverse transcriptase in 50 of the 70 melanomas examined. Intergenic regions comprise approximately 70% of the human genome. A previous study [5] showed that approximately 50% of the identified single nucleotide polymorphisms (SNPs) from exome sequencing were in the intended target regions, that 27% of the SNPs identified were in the flanking regions (within 200 bp) of the target regions, and that the remaining 24% of the SNPs were in regions >200 bp away from the target regions.

Although exome sequencing is not designed to identify regulatory SNPs in intronic and intergenic regions, off-target reads from this type of experiment should not be discarded *a priori*. One of the best examples of the usefulness of these off-target data is a study of Tibetans in high altitude [11], which found a pair of intronic SNPs in endothelial PAS domain protein 1 (EPAS1) with the greatest Tibetan-Han frequency difference. The authors specifically noted that these SNPs were outside the intended target regions of the exome sequencing, drawing attention to the potential value of these reads.

These and other studies demonstrate that reliable SNPs can be identified through off-target reads captured by exome sequencing [5], suggesting that it is worth searching for such SNPs even though the experiment was not designed to find them. However, it has been observed that the SNP false positive rate increases as the reads align further away from the captured regions [5]. Thus, more stringent filter criteria, such as depth and genotyping quality score, need to be applied for the SNPs outside the captured regions to achieve the same quality as SNPs inside the captured regions, due to the higher error rate associated with off-target reads. For example, the transition:transversion ratio is commonly used as a quality measurement for SNPs identified through exome sequencing [5,21,22]. To achieve the same transition:transversion ratio for SNPs outside target regions when comparing with SNPs inside the target regions, stronger filters, such as higher depth, are required [5]. Another artifact of exome sequencing is the pseudogene effect, where some intergenic regions are sequenced to abnormally high depth (>1000). This anomaly seems to be consistent regardless of the type of capture kit used [5]. It has been speculated that such phenomena are caused by homologies of pseudogenes. The most commonly used SNP detection framework, Genome Analysis Tool Kit (GATK) [22], developed by the Broad Institute, suggests that SNPs in such regions should be ignored.

The mitochondrial genome in exome sequencing

Mitochondria have an important role in cellular energy metabolism, free radical generation, and apoptosis [23,24]. mtDNA is a maternally inherited 16 569-bp closed-circle genome that encodes two rRNAs, 22 tRNAs, and ten poly-peptides. Dysfunctions in mitochondrial function are an important cause of many neurological diseases [25] and drug toxicities [26,27], and may contribute to carcinogenesis and tumor progression [28,29]. Furthermore, the mitochondrial genome is a fundamental tool for human population genetics and has had a critical role in mapping the migration of humanity across the globe [30–33].

Because the mitochondrial genome is almost all coding sequence, it fits every reasonable definition of the exome. However, mtDNA is not targeted in any of the currently used exome-sequencing methods. Instead, mtDNA sequence can be extracted from exome-sequencing data [2,10]. The average coverage of the mitochondrial genome from exome sequencing is approximately 100, easily surpassing the average coverage of even the targeted genomic regions [10]. The relatively high coverage of mtDNA is due to the high copy number of mtDNA per cell, on the order of hundreds to several hundred thousand copies per cell, depending on the tissue type [34]. This should be contrasted to techniques that specifically target the mitochondrial sequence, which can produce an average depth of tens of thousands of reads across the mitochondrial genome [35–38]. Given that cells typically contain a very large number of copies of mtDNA, mixtures of wild type and mutant mtDNA (heteroplasmy) can range almost continuously from 0 to 100%. Pathogenic mtDNA mutations are typically heteroplasmic in an individual, with asymptomatic carriers of the mutations having a low heteroplasmy level of the pathogenic mutation [39]. An average read depth of only approximately 100 means that, although polymorphisms can be accurately determined, the identification of heteroplasmic mtDNA variations is limited to those present in >10% of the mtDNA molecules in the sample. However, these are likely to

be the most clinically relevant cases, again pointing to the potential utility of analyzing these sequences. Researchers have started to infer mitochondria mutation information from exome-sequencing data. The best examples are The Cancer Genome Atlas (TCGA) project, where all mtDNA somatic mutations were inferred from exome-sequencing data. For example, the current somatic mutation results for breast cancer in TCGA [40] contain exome-sequencing data from 776 tumors and report 325 mtDNA somatic mutations derived from off-target reads from the exome-sequencing data.

An important complication in aligning DNA reads to the mitochondrial genome is the presence of nuclear copies of the mitochondrial genomes (nuMTS) [41,42]. nuMTS can cause ambiguity about whether a read should map to the nuclear or the mitochondrial genome. The simplest way to obtain the mitochondrial genome is to align the raw reads against the mitochondrial reference genome directly and then filter out the nonaligned reads, thus ignoring the nuMTS. The disadvantage of this approach is that the reads that do derive from the nuMTS may introduce false heteroplasmic variability in the mtDNA sequence. A middle approach is to align the reads against both the nuclear and mitochondrial genomes simultaneously. When a read has multiple locations to which it may be mapped, aligners such as BWA [43] will randomly choose among the possible locations. This has the disadvantage of treating the nuMTs and the mitochondrial genome equally, ignoring the very large copy number difference. The effect of this choice will be that many of the reads coming from the mtDNA will be falsely aligned to the nuMTS, causing an artificially high coverage of the nuMTS and an artificially low coverage of the mtDNA. A third choice gives precedence to the nuMTs by first aligning reads against the nuclear genome and then aligning only the nonaligned reads to the mitochondrial genome. This approach will have the most extreme misalignment of true mtDNA reads to the nuclear DNA (potentially leading to false SNP calls in the nuclear DNA), which will lower the coverage of the mitochondria genome and decrease the chance of detecting true variants. The third approach is also the most conservative and time consuming, involving two alignment processes and leaving no chance of misaligning any nuMTS reads to the mitochondria genome. The second approach is the most balanced approach between time consumption and misalignment rate and has been implemented in MitoSeek [44] which can be used to extract mitochondria mutation and heteroplasmy information from exome-sequencing data.

mtDNA copy number is highly variable and has been suggested to be associated with many diseases, including cancer [45–48]. Thus, it is an important mitochondrial statistic that can be derived from exome-sequencing data. Traditional methods for evaluating mtDNA copy number involve quantitative (q)PCR [49]. A more recent method has been developed that relies upon a sequencing-based assay of mtDNA copy number that draws on the unbiased nature of next-generation sequencing and incorporates techniques developed for RNA expression profiling [50]. Although the authors claim that this assay reports absolute mitochondria copy number, we argue that the amount of library constructed will affect the copy number count. For example, it has been shown that the fraction of captured mitochondrial sequences in exome-sequencing data is proportional to the relative abundance of the corresponding mitochondrial genome in the original total DNA extract [10]. Based on this observation, we conclude that relative, but not absolute, mtDNA copy number is detectable through exome-sequencing data. The mtDNA copy extracted from exome-sequencing data can be useful when studying tumor samples for conducting association tests with phenotypes such as tumor stage and metastasis stage. The recently developed software MitoSeek [44] also computes relative mtDNA copy number from exome-sequencing data.

Pathogen DNA and integration sites

Finally, it is important to consider the portion of reads from exome sequencing that does not map to the reference genome. Some of these reads may represent viral DNA, as either free viral DNA or as viral genomes that have been incorporated into the genome of a host. Detecting viral DNA is of particular importance due to the important role of viral DNA integration into the host genome in initiating cancer. Many viruses integrate into the genome of their host cells to replicate and, therefore, mutagenesis caused by viral infection may be quite common. Typically, viruses trigger tumor development by altering host genes or by suppressing the immune system of the host, causing inflammation over a long period of time. Most viruses lack clearly identifiable oncogenes capable of cellular transformation and instead mediate oncogenic transformation through a process termed insertional mutagenesis (IM). The molecular mechanisms of viral IM can vary, but most involve viral insertion within tumor suppressor genes or upregulation of cellular oncogenes in close proximity to the site of viral integration via *cis* and *trans* effects of promoter and enhancer sequences within the viral long terminal repeats (LTRs). Known oncogenic viruses [such as the hepatitis B virus (HBV) for liver cancer and the human papillomavirus (HPV) for head and neck cancer and ovarian cancer] are estimated to cause 15–20% of all cancers in humans [51,52]. Understanding the viral integration pattern of cancer-associated viruses may uncover novel oncogenes and tumor suppressors that are associated with cellular transformation.

Viral genomes have been detected using high throughput-sequencing technology [53–57]. The idea of using off-target reads to detect viruses was introduced a few years ago. In general, viruses can be detected through exome sequencing either by detecting viral genome sequences that have been integrated into the host DNA or by inadvertently capturing the viral sequence itself. The presence of HPV [8] and HBV [58,59] has been detected through analysis of exome-sequencing data. Tools for detecting virus sequence through exome-sequencing data have also been developed. For example, PathSeq was developed to identify viruses through sequencing data of human samples [7]. VirusSeq was developed to identify viral sequences using exome-sequencing or RNAseq data [8]. Most recently, ViralFusionSeq was developed [60] to discover viral integration events and to reconstruct fusion transcripts at single-base resolution. Theoretically, bacteria can also be detected in exome-sequencing data provided they are present. For example, PathSeq [7] is designed to capture both bacterial and viral sequences.

One of the challenges associated with identifying viral sequences through exome-sequencing data is the rapid mutation rate of some viruses. DNA viruses have a mutation rate of between 10^{-6} to 10^{-8} mutations per base per generation, and RNA viruses have an even faster mutation rate of 10^{-3} – 10^{-5} per base per generation [61]. There are two possible solutions for identifying viral sequences with a high mutation frequency. First, the number of allowed mismatches per read can be increased. The typical read length of exome sequencing is from 75 to 100 bp. The default mismatch allowed per read for most popular aligners such as BWA [43] and Bowtie [62] is usually two. Allowing more mismatches during the viral genome alignment can alleviate the problem caused by the fast viral mutation rate. Second, a virus reference panel can be created that includes all known variations of a targeted virus. Although this method can increase the alignment time, it is more likely to be accurate than simply allowing more mismatches. However, it does have the disadvantage of potentially failing to detect viral strains that have evolved significantly from the strains in the reference panel. Another challenge associated with virus detection in exome-sequencing data is the potential homology between the reference human genome and viral genomes, similar to the problem of nuclear genome copies of the mitochondrial genome described in the previous section. One conservative approach to solving this

problem is to use only reads unmapped to the human genome for the viral genome alignment.

The location of virus integration into the host genome may have a role in disease etiology [63–66]. However, identifying the sites of virus integration using exome-sequencing data is challenging. For paired-end read data, a single DNA fragment will have sequence reads on both ends. During alignment, discordant pairs can be detected in which one read is aligned to the viral genome whereas its mate is aligned to the human genome, a good indicator of a possible intervening integration site. To find the exact integration site, read-through reads (in which the break point lies within a read) need to be examined. Existing structural variant detection tools, such as BreakDancer [67], can be used to detect integration sites if the viral genome reference is added to the human genome reference before alignment. VirusSeq [8] detects integration sites by first identifying discordant read pairs and then clustering the discordant read pairs that support the same integration event. By contrast, ViralFusionSeq [60] uses a more sophisticated model to detect breakpoints that support viral fusion. Many viral fusion sites have been identified through exome sequencing. For example, in a study of liver cancer, HBV integration was observed in 70 out of 81 liver cancer samples [58]. Furthermore, HBV viral integration sites have also been identified through exome sequencing in a separate liver study [59].

Virus detection through exome sequencing has several limitations. First is the obvious limitation that this method can only detect DNA viruses or RNA viruses that are reverse transcribed and have a DNA phase. To detect an RNA virus, RNAseq technology needs to be used [68–72]. Second, it is highly dependent on the amount of reads sequenced. If the depth of exome sequencing is low, the chance of detecting any virus also decreases. Finally, it is impractical for exome sequencing to detect any novel virus, or a virus with variants that have not been previously described. Nevertheless, there are successful examples of detection of viral genomes from exome sequencing, providing another example of the value of reconsidering off-target reads.

Concluding remarks

Exome-sequencing data are now becoming widely available for secondary uses through efforts to encourage data sharing, such as TCGA (currently 15000 exomes) and the NHLBI Exome Sequencing Project (6500 exomes). It was widely predicted that the price of whole-genome sequencing for the human genome would drop to under US\$1000 as early as 2003 [73–75]. However, with currently available technologies, to achieve an average of 30× coverage in whole-genome sequencing still costs over \$5000 a sample, whereas exome sequencing at 30× coverage costs under US\$500. There is always a possibility that an advance in technology will reduce the cost of whole-genome sequencing to a comparable price of exome sequencing. However, the extra cost associated with the data analysis of whole-genome sequencing data is likely to remain significantly higher. The storage and processing time of whole-genome sequencing data can be 10 to 20 times more than that of exome sequencing data. Until these limitations of whole-genome sequencing cost and data storage are overcome, the growing amount of exome data available can be usefully mined for additional research purposes.

Another future development that could impact the types of secondary analysis we have outlined here are improvements in exome capture technology to eliminate or reduce significantly off-target reads. Exome capture technology has been continuously improving since it was introduced. However, the capture efficiency has increased only slightly over the years. Furthermore, the major reason for the increased capture efficiency has been due to the increased size of capture regions rather than improvement of the capture technology itself.

For example, the Agilent SureSelect v1 kit captured 37 Mb of the human genome, whereas the latest SureSelect v5 kit captures 50 Mb. Additionally, the amount of output of sequencing instruments has also increased over the years. The original Illumina GA II platform could output 20 million–25 million reads per lane. The newest Illumina HiSeq 2500 can produce 150 million–200 million reads per lane. Even after multiplexing three to four samples on a HiSeq lane, the amount of reads sequenced per sample is still much higher than that achieved using the GA II machine. Thus, even though the percentage of reads not mapped to target regions might decrease, the raw number of reads not mapped to the target regions might increase due to the increase of machine throughput, suggesting that exome-sequencing data will continue to be good candidates for additional data mining despite technological improvements.

Several tools are now available to mine these data for the ‘lost treasure’ buried in off-target reads. We have summarized here the possibilities and challenges in studying variants outside of the targeted exonic regions. These include mitochondrial variants, as well as viral genomes and virus–host integration sites. However, we note that another possibility for some of the unmapped reads is that they may still belong to the human genome, but may come from genome regions not covered by the current human genome reference, GRCh37. With GRCh38 (scheduled to be released during late 2013), it is likely that some of the previously unmapped reads will be mapped to the new human reference. There are also possibilities that have yet to be discovered, making studying the unmapped reads a potentially fruitful opportunity.

Although we are encouraging researchers to conduct additional data mining using existing data, we would also like to promote good study design. If the goal of the study is to survey all SNPs, then a whole-genome study should be used. If the goal of the study is to examine the mtDNA sequence, then mitochondria-targeted sequencing should be used, and if the goal is to detect the presence of viruses then a virus-specific method should be used. Exome sequencing is a powerful tool, but it is not designed specifically for the additional targets described in this review. However, to get the fullest use of this low-cost sequencing technology, and of the massive amount of exome sequences currently publically available, we should not ignore the unexpected DNA reads, which can comprise as much as half of the data produced by exome sequencing methods. The off-target reads must be subject to stringent quality control and, thus, we recommend an additional validation phase for all important findings observed through off-target reads whenever possible, including the use of targeted resequencing.

Glossary

Bait	the hybridization probe designed to capture effectively the coordinates of the target region to be sequenced. The bait design differs by manufacturer and method. Some methods use baits that tile the target region, whereas others use baits that do not overlap and differ in distance between the baits
Exome sequencing	selectively capturing the exome (coding regions) and other content in a whole-genome library before sequencing. This enables deeper coverage of the genomic region that is enriched in disease-causing variants in a megabase-sized DNA library instead of sequencing a lower coverage gigabase-sized whole-genome library

Next-generation sequencing	high-throughput DNA sequencing using massively parallel reactions generating millions of independent reads. The methodology employs a variety of technologies, including highly parallelized pyrosequencing, sequencing-by-synthesis, sequencing by ligation, and single molecule sequencing methods
Off-target reads	the sequencing reads that do not align to the target region
Oncogenic virus	a virus associated with cancer. The cause of this association is generally due to the insertion of the viral genome into the host genome in a location that disrupts a crucial host gene, leading to the expansion of that cell into a tumor
Reads	the fragments of DNA sequences generated that represent data from a unique fragment of the sequencing library. A typical next-generation sequencing run generates millions of reads per sample
Target regions	the region of interest defined for enrichment. The genomic coordinates of a target region are used to design the capture baits, probes, or primers for enrichment and vary by exome sequencing kit
Unmappable reads	the reads that are not aligned to the human genome

References

1. Ng SB, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet.* 2010; 42:30–35. [PubMed: 19915526]
2. Durbin RM, et al. A map of human genome variation from population-scale sequencing. *Nature.* 2010; 467:1061–1073. [PubMed: 20981092]
3. Fu W, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature.* 2013; 493:216–220. [PubMed: 23201682]
4. Sulonen AM, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* 2011; 12:R94. [PubMed: 21955854]
5. Guo Y, et al. Exome sequencing generates high quality data in non-target regions. *BMC Genomics.* 2012; 13:194. [PubMed: 22607156]
6. Asan, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol.* 2011; 12:R95. [PubMed: 21955857]
7. Kostic AD, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol.* 2011; 29:393–396. [PubMed: 21552235]
8. Chen Y, et al. VirusSeq: software to identify viruses and their integration sites using nextgeneration sequencing of human cancer tissue. *Bioinformatics.* 2013; 29:266–267. [PubMed: 23162058]
9. Larman TC, et al. Spectrum of somatic mitochondrial mutations in five cancers. *Proc Natl Acad Sci USA.* 2012; 109:14087–14091. [PubMed: 22891333]
10. Picardi E, Pesole G. Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods.* 2012; 9:523–524. [PubMed: 22669646]
11. Yi X, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science.* 2010; 329:75–78. [PubMed: 20595611]
12. Djebali S, et al. Landscape of transcription in human cells. *Nature.* 2012; 489:101–108. [PubMed: 22955620]
13. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. [PubMed: 22955616]

14. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012; 22:1760–1774. [PubMed: 22955987]
15. Pei B, et al. The GENCODE pseudogene resource. *Genome Biol.* 2012; 13:R51. [PubMed: 22951037]
16. Alberobello AT, et al. An intronic SNP in the thyroid hormone receptor beta gene is associated with pituitary cell-specific over-expression of a mutant thyroid hormone receptor beta2 (R338W) in the index case of pituitary-selective resistance to thyroid hormone. *J Transl Med.* 2011; 9:144. [PubMed: 21871106]
17. Kawase T, et al. Alternative splicing due to an intronic SNP in HMSD generates a novel minor histocompatibility antigen. *Blood.* 2007; 110:1055–1063. [PubMed: 17409267]
18. Moyer RA, et al. Intronic polymorphisms affecting alternative splicing of human dopamine D2 receptor are associated with cocaine abuse. *Neuropsychopharmacology.* 2011; 36:753–762. [PubMed: 21150907]
19. Rearick D, et al. Critical association of ncRNA with introns. *Nucleic Acids Res.* 2011; 39:2357–2366. [PubMed: 21071396]
20. Huang FW, et al. Highly recurrent TERT promoter mutations in human melanoma. *Science.* 2013; 339:957–959. [PubMed: 23348506]
21. Guo Y, et al. The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics.* 2012; 13:666. [PubMed: 23176052]
22. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011; 43:491–498. [PubMed: 21478889]
23. Andrews RM, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet.* 1999; 23:147. [PubMed: 10508508]
24. Verma M, Kumar D. Application of mitochondrial genome information in cancer epidemiology. *Clin Chim Acta.* 2007; 383:41–50. [PubMed: 17532310]
25. Fernandez-Vizarra E, et al. Impaired complex III assembly associated with BCS1L gene mutations in isolated mitochondrial encephalopathy. *Hum Mol Genet.* 2007; 16:1241–1252. [PubMed: 17403714]
26. Lemasters JJ, et al. Mitochondrial dysfunction in the pathogenesis of necrotic and apoptotic cell death. *J Bioenerg Biomembr.* 1999; 31:305–319. [PubMed: 10665521]
27. Wallace KB, Starkov AA. Mitochondrial targets of drug toxicity. *Annu Rev Pharmacol Toxicol.* 2000; 40:353–388. [PubMed: 10836141]
28. Modica-Napolitano JS, Singh KK. Mitochondrial dysfunction in cancer. *Mitochondrion.* 2004; 4:755–762. [PubMed: 16120430]
29. Chen EI. Mitochondrial dysfunction and cancer metastasis. *J Bioenerg Biomembr.* 2012; 44:619–622. [PubMed: 22892817]
30. Soares P, et al. The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Mol Biol Evol.* 2012; 29:915–927. [PubMed: 22096215]
31. Yao YG, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet.* 2002; 70:635–651. [PubMed: 11836649]
32. Bandelt HJ, et al. Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. *Ann Hum Genet.* 2003; 67:512–524. [PubMed: 14641239]
33. Kong QP, et al. Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet.* 2003; 73:671–676. [PubMed: 12870132]
34. Bogenhagen D, Clayton DA. The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. *J Biol Chem.* 1974; 249:7991–7995. [PubMed: 4473454]
35. Guo Y, et al. The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation. *Mutat Res.* 2012; 744:154–160. [PubMed: 22387842]
36. Tang S, Huang T. Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. *Biotechniques.* 2010; 48:287–296. [PubMed: 20569205]

37. He Y, et al. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*. 2010; 464:610–614. [PubMed: 20200521]
38. Ameer A, et al. Ultra-deep sequencing of mouse mitochondrial DNA: mutational patterns and their origins. *PLoS Genet*. 2011; 7:e1002028. [PubMed: 21455489]
39. Falk MJ, Sondheimer N. Mitochondrial genetic diseases. *Curr Opin Pediatr*. 2010; 22:711–716. [PubMed: 21045694]
40. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012; 490:61–70. [PubMed: 23000897]
41. Hazkani-Covo E, et al. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet*. 2010; 6:e1000834. [PubMed: 20168995]
42. Li M, et al. Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. *Nucleic Acids Res*. 2012; 40:e137. [PubMed: 22649055]
43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–1760. [PubMed: 19451168]
44. Guo Y, et al. MitoSeek: extracting mitochondria information and performing high throughput mitochondria sequencing analysis. *Bioinformatics*. 2013; 29:1210–1211. [PubMed: 23471301]
45. Shen J, et al. Mitochondrial copy number and risk of breast cancer: a pilot study. *Mitochondrion*. 2010; 10:62–68. [PubMed: 19788937]
46. Yu M, et al. Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients. *IUBMB Life*. 2007; 59:450–457. [PubMed: 17654121]
47. Tseng LM, et al. Mitochondrial DNA mutations and mitochondrial DNA depletion in breast cancer. *Genes Chromosomes Cancer*. 2006; 45:629–638. [PubMed: 16568452]
48. Bai RK, et al. Mitochondrial DNA content varies with pathological characteristics of breast cancer. *J Oncol*. 2011; 2011:496189. [PubMed: 22028711]
49. Bhat HK, Epelboym I. Quantitative analysis of total mitochondrial DNA: competitive polymerase chain reaction versus real-time polymerase chain reaction. *J Biochem Mol Toxicol*. 2004; 18:180–186. [PubMed: 15452886]
50. Castle JC, et al. DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. *BMC Genomics*. 2010; 11:244. [PubMed: 20398377]
51. Parkin DM. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer*. 2006; 118:3030–3044. [PubMed: 16404738]
52. Morissette G, Flamand L. Herpesviruses and chromosomal integration. *J Virol*. 2010; 84:12100–12109. [PubMed: 20844040]
53. Barzon L, et al. Applications of next-generation sequencing technologies to diagnostic virology. *Int J Mol Sci*. 2011; 12:7861–7884. [PubMed: 22174638]
54. Radford AD, et al. Application of next-generation sequencing technologies in virology. *J Gen Virol*. 2012; 93:1853–1868. [PubMed: 22647373]
55. Chevaliez S, et al. New virologic tools for management of chronic hepatitis B and C. *Gastroenterology*. 2012; 142:1303–1313. [PubMed: 22537437]
56. Li L, Delwart E. From orphan virus to pathogen: the path to the clinical lab. *Curr Opin Virol*. 2011; 1:282–288. [PubMed: 22003378]
57. Capobianchi MR, et al. Next-generation sequencing technology in clinical virology. *Clin Microbiol Infect*. 2013; 19:15–22. [PubMed: 23279287]
58. Sung WK, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet*. 2012; 44:765–769. [PubMed: 22634754]
59. Jiang Z, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res*. 2012; 22:593–601. [PubMed: 22267523]
60. Li JW, et al. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics*. 2013; 29:649–651. [PubMed: 23314323]
61. Drake JW, et al. Rates of spontaneous mutation. *Genetics*. 1998; 148:1667–1686. [PubMed: 9560386]

62. Langmead B, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
63. Gozuacik D, et al. Identification of human cancer-related genes by naturally occurring Hepatitis B Virus DNA tagging. *Oncogene.* 2001; 20:6233–6240. [PubMed: 11593432]
64. Mason WS, et al. Clonal expansion of normal-appearing human hepatocytes during chronic hepatitis B virus infection. *J Virol.* 2010; 84:8308–8315. [PubMed: 20519397]
65. Murakami Y, et al. Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas. *Gut.* 2005; 54:1162–1168. [PubMed: 16009689]
66. Saigo K, et al. Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (MLL4) gene and rearrangements of MLL4 in human hepatocellular carcinoma. *Hum Mutat.* 2008; 29:703–708. [PubMed: 18320596]
67. Chen K, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods.* 2009; 6:677–681. [PubMed: 19668202]
68. Palacios G, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med.* 2008; 358:991–998. [PubMed: 18256387]
69. Nakamura S, et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE.* 2009; 4:e4219. [PubMed: 19156205]
70. Quan PL, et al. Astrovirus encephalitis in boy with X-linked agammaglobulinemia. *Emerg Infect Dis.* 2010; 16:918–925. [PubMed: 20507741]
71. Briesse T, et al. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog.* 2009; 5:e1000455. [PubMed: 19478873]
72. Isakov O, et al. Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics.* 2011; 27:2027–2030. [PubMed: 21666269]
73. Robertson JA. The \$1000 genome: ethical and legal issues in whole genome sequencing of individuals. *Am J Bioeth.* 2003; 3:35–42.
74. Mardis ER. Anticipating the 1,000 dollar genome. *Genome Biol.* 2006; 7:112. [PubMed: 17224040]
75. Bennett ST, et al. Toward the 1,000 dollars human genome. *Pharmacogenomics.* 2005; 6:373–382. [PubMed: 16004555]

Figure 1. Results of a PUBMED search for papers using the term 'exome', through 1 July, 2013 showing the rapid and recent spread of this sequencing method.

Figure 2.

A flow diagram illustrating how off-target reads can be identified from exome-sequencing data. Currently available tools for the analysis of the different types of off-target read are given. Abbreviation: SNP, single nucleotide polymorphism.