Genome **Biology**

## OPINION

# Exome sequencing: the expert view

Leslie G Biesecker[1]*, Kevin V Shianna[2] and Jim C Mullikin[3]

### Abstract

To complement our special issue on exome sequencing, *Genome Biology* asked several leaders in the field for their views on this new approach. Leslie G Biesecker (LGB), Jim C Mullikin (JM) and Kevin V Shianna (KVS) discuss the reasons for the popularity of exome sequencing and its contribution to genomics.

## How can exome sequencing contribute to our understanding of the dynamic nature of the genome?

**LGB, JCM:** Exomes are ideal to help us understand high-penetrance allelic variation and its relationship to phenotype. Because exomes focus on exons, which include coding regions of genes, and because most high-penetrance (Mendelian or nearly so) variation is mediated by non-synonymous, frameshifting and canonical splice variation, exomes are ideal for studying the relationship of such variation to health and disease.

**KVS:** Sequencing using any approach is still in its early days, but it is clear that exome sequencing will often lead to the identification of the causative variant for Mendelian diseases. This should not be surprising given that we know that most mutations causing Mendelian disease are exonic. That said, there are clear limitations even for Mendelian disease. Structural variations (SVs), which are also important for Mendelian disease, are not easily detected using an exome approach. How well exome sequencing may do for complex traits is an entirely open question since we do not know what kinds of mutations are important there, but it is possible they are more often regulatory than for Mendelian disease.

## How much has exome sequencing been driven by cost alone?

**LGB, JCM:** Cost is a huge factor – every day we ask ourselves the question, 'Would we rather have six

*Correspondence: leslieb@helix.nih.gov
[1]Genetic Disease Research Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
Full list of author information is available at the end of the article

**BioMed** Central

samples analyzed by whole exome sequencing (WES) or one by whole genome sequencing (WGS)?' Our current, fully loaded price for a WGS is six times that of a WES assay – a ratio that has changed surprisingly little in the past 2 years. Which study one should use depends on the biomedical question that is being asked. If it is primarily a genotype-phenotype question, and the putative variant is high penetrance, then it is crucial to increase our statistical power by increasing our $N$, so exomes provide a big advantage here. If the question is different, it could be that a smaller number of WGS interrogations would be more effective. WES and WGS are tools – one has to select the optimal tool considering the biomedical question and the available resources.

**KVS:** The lower cost of exome sequencing is probably the primary driver for its increased use, but a related and equally important factor is how much longer it takes to generate whole genome sequence data. As the cost of sequencing drops and the data generation per run increases, the cost and time required for WGS will become more similar to that for WES.

## What are the major limitations of exome sequencing?

**LGB, JCM:** We are unable to interrogate many variants that may be important for controlling gene transcriptional regulation or splicing. Also, our current understanding of the genome limits our exome interrogation – nucleotides in regions of the genome not currently recognized to be a gene will be missed by exome approaches. Finally, exomes may not be ideal for understanding structural variation in genomes.

**KS:** The major limitation of exome sequencing may be the inability to comprehensively represent genomic SVs. Many groups have designed algorithms that use a read depth or read pair-based approach for predicting structural variation; however, these approaches are not very efficient at identifying SVs with exome data. Another approach uses a split read method, but this will not be comprehensive and will miss many of the SVs. Another key limitation is that parts of the genome that we do not already recognize as functional are not included. Thus, WES will only find variants when they are in a part of the genome that we are familiar with. If a variant sits in a distal regulatory element and has a major impact on a

trait, it will be completely missed. How important this will turn out to be is yet to be determined.

## What lessons from exome sequencing studies can be applied to whole genome sequencing studies?

**LGB, JCM:** Exomes will be a fantastic platform to build capabilities in many domains. Annotation of variation is easier (but still far from easy) in WES than it is in WGS since a higher proportion of the variation falls on exons by design. If we can build robust annotation pipelines for a WES sequence, we can extend and generalize the lessons learned from that activity into interpretation of intronic and intergenic variation (both point and structural). Also, exomes provide us with low hanging fruit – to dissect the genetic architecture of a trait, culling out potential high penetrance variants from exomes, assessing the remaining heritability, and then tackling that remainder (assuming it is significant) with WGS would be a practical and economical approach. This is a triage approach; WES first then WGS on what remains. This assumes that WES is the obvious first choice for the samples. There are cases, like structural rearrangements, where WGS is the obvious first step. But in that particular example of finding breakpoints, deep WGS is not necessary, one just needs deep physical coverage with large spanning paired-end reads.

**KVS:** The exome sequencing approach has been a cost effective option for sequencing the human genome and has resulted in the identification of many disease-causing variants. The methods used to identify these variants are fully transferable to working with whole genome sequencing data. However, to efficiently and comprehensively work with whole genome sequencing data it will require a new set of bioinformatics tools that are not required for analyzing exome datasets.

## How do exome sequencing studies contribute to our mechanistic understanding of disease?

**LGB, JCM:** The essential contribution of exomes is to enrich, extend, and possibly even complete our search for the heritable basis of Mendelian disease. This would be a stupendous biomedical research accomplishment and potentially lead to a huge improvement in our understanding of the pathophysiology of many diseases, rare and common.

**KVS:** For the simplest cases of disease, such as Mendelian diseases, exome sequencing has led to the discovery of many causative variants. The identification of these variants will greatly increase our understanding of the most basic causes of disease. However, exome studies will have very limited power to identify causative variants in regulatory regions spread across the genome (transcription binding sites, enhancers, and so on). Implementing a WGS approach would allow detection of

variants in these regions, thus increasing our knowledge of disease beyond the coding region of the genome.

## Does exome sequencing have a limited 'shelf life'?

**LGB, JCM:** This is an open question. It is conceivable that exome sequencing, with future refinements and indexing of samples, could remain sufficiently less expensive than WGS that it would be preferable to WGS for certain applications. It will be essential for exome capture kit unit costs to decline significantly as WGS costs fall for exomes to remain competitive.

Even if the ratio of cost differential decreases, currently 6:1, even to near parity in terms of consumables, it may be better to continue with WES due to the other costs that are often ignored. These include sequencing instrument time and compute resources. There the ratio will remain at about 15:1 based on machine time and resulting data volume. Thus, if you can generate 1,000 exomes, the same number of sequencing machines can only produce 67 whole genomes. If you really would like to complete 1,000 samples using WGS in the same timeframe as the WES approach, you will need 15 times more sequencing machines. That is a huge outlay in capital costs, lab space, and so on. Downstream of the sequencing instruments, the data generated for WES are also 1/15 the volume when compared to WGS; thus, the networking and compute infrastructure are greatly simplified. This reason alone may make WES attractive for quite a number of years.

Perhaps the expiry date will arrive when anyone that wants or needs their genome sequenced can send a buccal swab out for WGS for $1,000 and they receive a cloud-computing account with their complete sequence. But even in this scenario, the monthly cost of an account with a WGS versus a WES, if based on data volume, would be 15 times more expensive for the WGS than a WES dataset.

**KVS:** Yes, as soon as the difference in cost between exome and whole genome diminishes (which will be soon) and issues with data management and storage are resolved, whole genome sequencing will be the method of choice. In addition, there will be rapid increases in sequencing technology over the next few years, resulting in the ability to sequence a genome at high coverage in a very short period of time (a few days and possibly hours). When this becomes a reality there will be little demand for an exome sequencing approach.

## How much do you think that future research will be restricted by the IT-related costs of the analysis?

**KVS:** This is the one major advantage of exome sequencing that will be difficult to overcome. The gap between WES and WGS IT costs will surely dwindle over time, but the bottom line is that analyzing data for the

exome will be easier due to the smaller number of required sequence reads (and therefore smaller file sizes). There would need to be a major paradigm shift in how data are analyzed and stored if one were to consider implementing a WGS approach on a population scale due to the substantial IT costs.

## Are there any advantages of whole exome over whole genome sequencing?

**LGB, JCM:** For clinical applications, it may be preferable to have a more delimited dataset (WES) as it generates fewer (though still many) results that cannot be interpreted. Medicolegal liability is a pervasive problem in clinical medicine and there are strong pressures against generating information that has little benefit if it may have liability. We are very far from being able to clinically interpret a genome, or even an exome, but here, more is definitely worse.

### Author details

[1]Genetic Disease Research Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. [2]Genome Analysis Facility, Duke Center for Human Genome Variation, Duke University School of Medicine, B342E LSRC Bldg, Durham, NC 27710, USA. [3]NIH Intramural Sequencing Center and Comparative Genomics Unit, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

### About the contributors

**Leslie G Biesecker** is Chief and Senior Investigator of the Genetic Disease Research Branch at the National Human Genome Research Institute. His research focuses on the clinical and molecular delineation of human genetic diseases. He is involved in developing clinical genomics research and it is during this research that he has become engaged in exome sequencing.

**Kevin V Shianna** is the director of the Genomic Analysis Facility within the Duke Center for Human Genome Variation as well as the director of operations for the Center. He established the Facility for high-throughput genotyping in 2005 and made a transition to next generation sequencing in mid-2007. Since then, the major focus of the Facility and the Center has been to use whole genome and exome sequencing to identify variants associated with human disease.

**Jim C Mullikin** is Acting Director of the NIH Intramural Sequencing Center. He is a computational geneticist who has been a key participant in the International Haplotype Map (HapMap) Project and on the Neanderthal genome project. His research involves him in large-scale medical sequencing at the NIH Intramural Sequencing Center and it is here that he has participated in exome sequencing projects.