International Standards for Genomes, Transcriptomes, and Metagenomes

Christopher E. Mason,[1,2,3,]* *Ebrahim Afshinnekoo,*[1,2,4] *Scott Tighe,*[5] *Shixiu Wu,*[6] *and Shawn Levy*[7]

[1]*Department of Physiology and Biophysics,* [2]*The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, and* [3]*Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, New York, New York 10065, USA;* [4]*School of Medicine, New York Medical College, Valhalla, New York 10595, USA;* [5]*Advanced Genomics Lab, University of Vermont Cancer Center, Burlington, Vermont 05405, USA;* [6]*Hangzhou Cancer Institute in Hangzhou Cancer Hospital, Hangzhou, China; and* [7]*HudsonAlpha Institute of Technology, Huntsville, Alabama 35806, USA*

Challenges and biases in preparing, characterizing, and sequencing DNA and RNA can have significant impacts on research in genomics across all kingdoms of life, including experiments in single-cells, RNA profiling, and metagenomics (across multiple genomes). Technical artifacts and contamination can arise at each point of sample manipulation, extraction, sequencing, and analysis. Thus, the measurement and benchmarking of these potential sources of error are of paramount importance as next-generation sequencing (NGS) projects become more global and ubiquitous. Fortunately, a variety of methods, standards, and technologies have recently emerged that improve measurements in genomics and sequencing, from the initial input material to the computational pipelines that process and annotate the data. Here we review current standards and their applications in genomics, including whole genomes, transcriptomes, mixed genomic samples (metagenomes), and the modified bases within each (epigenomes and epitranscriptomes). These standards, tools, and metrics are critical for quantifying the accuracy of NGS methods, which will be essential for robust approaches in clinical genomics and precision medicine.

KEY WORDS: genomics, epigenomics, metagenomics, transcriptomics, epitranscriptomics

INTRODUCTION

Over the past 20 yr, advancements in cellular and molecular biology have demonstrated the numerous and complex mechanisms that cells exploit at the molecular level during development, homeostasis, and reproduction.[1] Yet, these complex molecular transactions for storing, processing, and moving DNA, RNA, and proteins, as well as all their intermediates, also create challenges in obtaining precise measures. The intrinsic properties of these complex biomolecules that make them so effective in their cellular role (*e.g.*, long structural repeats, DNA base modifications, cell walls) can also impede the ability for successful extraction, measurement, and ultimately, biologic interpretation. Conversely, other biochemical properties of nucleic acids can uniquely tag them, as in the case of methyl-5-cytosine or other modified DNA/RNA bases, and thus enable an easier path toward isolation and investigation. But in all cases, the means of extraction, preparation, and sequencing define the scope of the biologic phenomenon that can be measured, quantified, and analyzed.

Before 2006, most DNA and RNA sequencing reactions involved only one nucleic acid at a time (DNA or cDNA), which limited the throughput of many scientific questions of genomics, even for single-molecule methods in genotyping.[2] However, with the advent of second- and third-generation sequencing technologies (NGS), billions of templates of DNA or RNA could be assayed at the same time[3, 4] and demonstrated improved performance and information content over microarrays.[5] This created an explosion of novel methods, approaches, techniques, and protocols for the examination of virtually any question in genetics (DNA), transcriptomics (RNA), or combinations of these between multiple kingdoms and species (metagenomics/metatranscriptomics). However, the NGS instruments, chemistries, and algorithms change quickly, often within months, creating numerous challenges for researchers and clinicians trying to finalize a protocol or study. Such a rapid pace of technological change led to calls for NGS standards to be developed, similar to those that were created for microarrays, such as the minimal information about a microarray experiment.[6]

Fortunately, several new biologic and biochemical standards to assess nascent NGS technologies and methods have recently become available, such as the External RNA Control Consortium[7] RNA mixtures, Spike-In RNA Variant Control Mixes (SIRVs; Lexogen, Greenland, NH,

*ADDRESS CORRESPONDENCE TO: Christopher E. Mason, Dept. of Physiology and Biophysics, Weill Cornell Medicine, 1305 York Ave., New York, NY 10021, USA (Phone: 203-668-1448; E-mail: chm2042@med.cornell.edu).

doi: 10.7171/jbt.17-2801-006

USA), and Genome in a Bottle (GIAB) Consortium standard human genome for DNA benchmarking.[8] The testing of these DNA and RNA standards showed differences between the competing NGS platforms, as well as computational methods, which is reminiscent of the same challenges in early microarray work detailed by the Microarray Quality Control (MAQC) Consortium in 2006.[9, 10] Recent efforts have cataloged the technical sources of noise from NGS protocols, including the Sequencing Quality Control Consortium (SEQC) from the U.S. Food and Drug Administration (FDA),[11, 12] Association of Biomolecular Resource Facilities (ABRF)-NGS group,[13, 14] Centers for Disease Control and Prevention's (CDC) Next-Generation Sequencing Standardization of Clinical Testing group,[15] among others (**Table 1**).[16] Whereas many of these studies use human- or other eukaryotic-based models of using or studying nucleic acids, the same requirements are needed for the rapidly growing field of metagenomics and microbiome. For metagenomics and microbiome studies to generate high-quality, robust data, proper reference materials, nucleic acid standards, laboratory reagents, and software will be required.

As a result of the implementation of these standards, several key technical developments have emerged. First, such titrated controls enable normalization methods that can help ameliorate the impact of library preparation, guanine-cytosine (GC) bias, and other batch effects, even among completely different sequencing technologies.[17, 18] Moreover, the latest sequencing technologies have been benchmarked against these standards, enabling more comprehensive views of the human transcriptome or genome. Specifically, this includes a greater coverage of the genome—information that retains the co-occurrence of genetic variation (phasing) and resolution of high-repeat areas, insertion-deletions elements (indels), or segmental duplications with long-read sequencing technologies from Pacific Biosciences (PacBio; Menlo Park, CA, USA) and Oxford Nanopore Technologies (ONT; Oxford, United Kingdom). Indeed, long-read NGS methods have enabled full-length cDNA sequencing for even very complex transcriptomes, creating the highest-ever resolution of splicing events in eukaryotic cells.[12, 19] Likewise, long-read, whole-genome sequencing has revealed that 85% of the genetic variation, >50 bp, is missing from most previously studied genomes.[20] Finally, shotgun metagenomics sequencing has revealed that 30–70% of DNA does not match any known species in the current database.[21, 22] These studies highlight the vast amount of genetic information remaining to be discovered and the limitations of genomic databases, as well as the requirement for reference controls and standards, such as those of the Microbiome Quality Control (MBQC) Consortium.[23] As such, for each of the above assays, a measure of experimental signal/noise, positive controls (Table 1), and negative controls will be essential for robust reporting of new findings.

Finally, just as new DNA, RNA, and metagenomic variations are discovered, many new approaches have also emerged to study modified DNA bases (epigenetics) and RNA bases (epitranscriptomics)[2] or their mixtures from metagenomes (epimetagenomics). The study of phased (coincident on the same molecule) marks for epigenetics in DNA involves epialleles or haplo-epitypes.[24, 25] Such phased information about modified bases can enable discovery of coincident RNA modifications in the epitranscriptome,[26–28] including on nanopore technologies, such as the MinION (ONT).[29] For metagenomic studies, longer reads can assemble complete genomes and also help with resolving species ambiguity from an admixed environment or clinical samples.[30] The genetic resolution of molecular techniques stands at an unprecedented state and will likely continue to revolutionize our understanding of biology and function. Nonetheless, all of these methods necessitate appropriate NGS controls and international standards to ensure their accurate measurement and quantification, which will enable their improved application to clinical settings (such as with fecal microbiome transplant therapy) or discovery-based work and even improve sequencing methods implemented for work in more exotic settings, such as microgravity,[31] analog planetary environment (such as Antarctica), and eventually, environments beyond Earth.[32]

STEPS IN NGS SAMPLE PREPARATION

Recent advances in NGS sample preparation can be divided into 3 main areas: extraction, library preparation, and automation. Ideally, extraction of a sample's nucleic acid fractions would involve little or no degradation and enable complete profiling of the entire length of the molecule. Whereas each of these steps can be performed separately, newer protocols, technologies, and methods are being developed that enable them all to be performed in near-immediate succession, opening up an extraordinary "era of single-cell, multi-omic biology" and ultra-high-throughput genomics.

Extraction

A wide range of approaches can be used for lysing cells or tissues and then for their extraction, but they are often tailored for a specific type of assay planned for an experiment.[33] As such, there are tradeoffs for extraction of samples that will change, depending on the assay and experiment being planned. For example, the physically rigorous methods (*e.g.*, bead-beading, hot phenol) needed

TABLE 1

Molecular standards for assessing library, sequencing, and analysis methods in DNA, RNA, and metagenomics

Acronym	Group	Type	Agency/group	Web site(s) for consortiums, data sets, methods, and/or materials
Genome/epigenome				
GIAB	Genome in a Bottle	DNA and cells	NIST	http://jimb.stanford.edu/giab/
Nex-StoCT	Next-Generation Sequencing: Standardization of Clinical Testing II	DNA	CDC	http://www.cdc.gov/ophss/csels/dlpss/Genetic_Testing_Quality_Practices/ngsqp.html
GeT-RM	Genetic Testing Reference Materials Coordination Program	DNA	CDC	http://wwwn.cdc.gov/clia/Resources/GetRM/default.aspx
RSBP	Registry of Standard Biologic Parts	DNA	iGEM	http://parts.igem.org/Main_Page
SEQC/SEQC2	Sequencing Quality Control Consortium	DNA	FDA	http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/ http://www.nature.com/nbt/collections/seqc/index.html
Epitranscriptome/transcriptome				
MAQC/MAQC2	Microarray Quality Control Consortium	RNA	FDA	http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/ http://www.nature.com/nbt/focus/maqc/index.html
ABRF-NGS	Association of Biomolecular Resource Facilities-Next-Generation Sequencing	RNA	ABRF	http://www.abrf.org/index.cfm/group.show/NextGenerationSequencing%28NGS%29.75.htm http://www.biotech.cornell.edu/news/abrf-next-generation-sequencing-study-webinar
GEUVADIS	Genetic European Variation in Health and Disease	RNA	EU	http://www.geuvadis.org
ERCC	External RNA Control Consortium	RNA	NIST	http://www.nist.gov/mml/bbd/ercc.cfm https://www.lifetechnologies.com/order/catalog/product/4456740
ERCC2	External RNA Control Consortium 2	RNA	NIST	http://www.nist.gov/mml/bbd/ercc2.cfm
SIRV	Spike-In RNA Variant Mixes	RNA	Lexogen	https://www.lexogen.com/sirvsrelease/
Metatranscriptome/metagenome				
MBQC	Microbiome Quality Control Consortium	Meta	MBQC	www.mbqc.org
IMMSA	International Metagenomics and Microbiome Standards Consortium	Meta	NIST	http://www.nist.gov/mml/bbd/microbial_metrology/immsa-mission-statement.cfm
BEI	International Human Microbiome Standards	Meta	NIAID	https://www.beiresources.org/Catalog/otherProducts/HM-782D.aspx
IHMS	International Human Microbiome Standards	Meta	Meta	www.microbiome-standards.org/
BiOMICs	Bio-OMICS Mixed Kingdom DNA Standard	Meta and cells	Zymo	http://www.zymobiomics.com/
ATCC	International Metagenomics and Microbiome Standards Consortium	Meta	ATCC	http://www.atcc.org/products/all/CCL-186.aspx
EMP	Earth Microbiome Project	Meta	EMP	http://earthmicrobiome.org/
XMP	eXtreme Microbiome Project	Meta	XMP	http://extrememicrobiome.org/

Continued

TABLE 1

(Continued)

Acronym	Group	Type	Agency/group	Web site(s) for consortiums, data sets, methods, and/or materials
MGRG	Metagenomics Research Group	Meta	ABRF	http://blog.abrf.org/
MetaSUB	International Metagenomics and Metadesign	Meta		http://www.metasub.org

NIST, National Institute of Standards and Technology (Gaithersburg, MD, USA); iGEM, International Genetically Engineered Machine (Cambridge, MA, USA); EU, European Union; BEI, BEI Resources (Manassas, VA, USA); NIAID, National Institute of Allergy and Infectious Diseases, U.S. National Institutes of Health (Bethesda, MD, USA); Meta, MetaGenoPolis (Jouy-en-Josas, France); Zymo, Zymo Research (Irvine, CA, USA); ATCC, American Type Culture Collection (Manassas, VA, USA).

for extraction of nucleic acids from plants or bacteria with thick or multiple cell walls unfortunately create DNA/RNA fragments that are shorter than in *in situ* lysis of cells. More gentle methods based on magnetic bead separation, electrophoresis separation, or gel plugs can create longer (>50–100 kb) fragments of DNA for sequencing but necessitates wide-mouth pipettes and very slow pipetting to maintain long, intact nucleic acids.[34] The more rapid extraction methods or column-based extraction are often simpler and faster but lead to shorter (<10 kb) fragments and in many cases, reduced recovery from hard-to-lyse cells.

However, new developments in NGS chemistries, serial dilutions, or microfluidic manipulation of cells have created multiple means by which to lyse, purify, and prepare samples for sequencing in effectively one step. For example, single-cell profiling leveraging Φ29 amplification has been shown to create full-length cDNA synthesis from low-input samples or single cells[35] and includes a library preparation step. Moreover, the VolTRAX microfluidic system, proposed by ONT, may eventually be a system to lyse, prepare, and sequence a drop of blood from one reaction or be used in the field at a remote location, although only a prototype currently exists. [36]

Finally, it is worth noting that the final elution and purification step in many extraction protocols will determine the molecules that can be examined.[37] For example, to purify polyadenylated (polyA) RNA from total RNA, 70% ethanol is normally used, but for smaller RNAs (microRNAs, circular RNAs), the elution is often at 90–100% ethanol. As such, fresh ethanol should always be made for each RNA purification; otherwise, even a slight amount of evaporation of the ethanol can change the extraction efficiency and thus, change the measure of the underlying biology.

Library Preparation

Library preparation for RNA and DNA sequencing can use a variety of approaches, including ligation-based, transposase-based, or tagging approaches. For most commonly used protocols, the purified templates of DNA/cDNA are end repaired, followed by adapter ligation, further size selection, and (potentially) PCR amplification. The use of transposases in sample preparation has combined these steps of fragmentation and tagging (dubbed tagmentation),[38] which can significantly save time for library preparation in genomics and epigenomics. A variety of transposons is commercially available (*e.g.*, Tn5), and all are known to have some degree of bias in representing their targets' genome.[39] Several new techniques for enzymatic fragmentation and rapid library preparation have also emerged, which also eliminate the step of mechanical fragmentation (*e.g.*, Covaris, Woburn, MA, USA). This "fragmentase" reaction is a time-dependent, linear-response reaction that works by enzymatically shearing the DNA, with shorter fragments resulting from a longer reaction time. The advantage to this protocol is that it does not require a separate fragmentation step for the DNA/cDNA molecules, but it can vary from technician to technician, site to site, or is influenced based on sequence context. It is also a time- and concentration-dependent reaction. [40]

For RNA sequencing, advances have been made at the bulk RNA and single-cell level. First, to ensure that a full transcript is captured, just performing cDNA synthesis may not always be sufficient or appropriate. For instance, a previous study by Li and colleagues[13] has shown that the use of an antibody to enrich for transcripts with a 5′ guanosine cap (5′ antibody), combined with a polyA-priming step, created the most even and complete coverage of all annotated genes. Furthermore, the ribo-depletion step for RNA library preparation was able to rescue even degraded RNA (by sonication, enzymatic digestion, and heat), as the full 5′ to 3′ "evenness" of coverage of the gene could be recreated. In this case, phasing formation is lost, but if one's goal is to ensure that all exons from a gene are queried for their expression, rather than just the 3′ end, then ribo-depletion methods can serve to rescue RNA samples degraded by heat, enzymes, or mechanical forces.[13]

For the detection of epigenetic marks, amplification cannot be used during library preparation, as the native

DNA template will lose its chemical marks, and the same restriction applies to RNA modifications after cDNA synthesis. For RNA-based modifications (the epitranscriptome), the action of creating the synthetic strand(s) of cDNA synthesis with unmodified bases will hide RNA base modifications (112 discovered to date),[41] as the RT will not keep the fidelity of the modified base. However, studies have shown that the PacBio,[2] nanpore-based,[42] and Helicos Genetic Analysis System[43] single-molecule sequencers can directly sequence RNA, and potentially the kinetic changes observed during base incorporation could be mapped to reveal these modified bases. Currently, the most reliable method for discovering base modifications includes variations of Methylated-RNA Immunoprecipitation (MeRIP-seq[44]; also called m6A-seq[45]) or cross-linking methods, such as methylation-induced cross-linking immunoprecipitation,[46] and such methods can also reveal the heterogeneity of a sample, such as in DNA methylation patterns.[47] Eventually, these immunoprecipitation-based and bulk methods may be replaced by highly sensitive, third-generation methods that are single molecule, and some evidence has shown this is possible,[2, 28, 48] even though the informatics for interpreting these data is still in its early stages.[49]

Natively phased library preparation methods

To create phased genetic or epigenetic data, several technologies exist that can either directly produce phased information from single molecules or synthetic reads that can create them through automated chemistry and informatics. However, it is worth noting that all sequence reads are phased to some degree, even if the read is only 10 or 50 nt long. Indeed, such short reads have been shown to contain multiple informative cytosine-phosphate-guanosine sites, which create phased genetic information and phased epigenetic haplotypes from bisulfite-treated DNA in whole-genome epigenetic profiling, called whole-genome bisulfite sequencing (WGBS). These epigenetic haplotypes, or "epialleles," create several new types of information: they serve as a way to monitor the clonality of mixed tumor samples, provide loci that harbor "epigenetic evolution," and show promise as a way to stratify a risk for cancer.[24, 25] Notably, these measures of tumor heterogeneity and epigenetic stratification have also shown relevance for glioma, chronic lymphocytic leukemia, and diffuse large B cell lymphoma.[50–52]

However, the ultimate goal in phasing is to create a telomere-to-telomere map for each chromosome[53] or complete circles (for a circular genomes).[54] Fortunately, for many of the single-molecule, long-read technologies, phased information is native to the data. This includes long (>10 kb) reads from nanopore-based methods (ONT and Genia, Santa Clara, CA, USA), phased information from nanochannel or optical mapping approaches (Nabsys, Providence, RI, USA; BioNano Genomics, San Diego, CA, USA), and long reads from single-molecule detection methods (PacBio). For any single-molecule, kinetic-based method (ONT, PacBio, Genia), there is an ability to detect epigenetic states, including methyl-cytosine, methyl-6-adenosize, and hydroxyl-methyl cytosine, with a potential for detecting a variety of other nucleic acid variants, such as those from DNA damage (*e.g.*, 8-oxo-guanosine).[55–59]

Synthetically phased library preparations

Another option for creating phased information is using biochemical methods that maintain relationships among smaller, subhaploid DNA fragments produced during library preparation. Detection of these fragments and the computational tracking of the relationships among them allow very long, haploid fragments to be constructed as synthetic reads. These methods rely on unique tagging or proximity-based methods to create maps of reads. The methods include such hybrid approaches as Nanopore Synthetic-long reads that combine ONT and Illumina (San Diego, CA, USA) data[58] or biochemical approaches[60], such as multiple displacement amplification of subhaploid DNA, followed by indexed library preparation and sequencing[61]; statistically aided, long-read haplotyping; the basis of Moleculo (Illumina)[62]; Contiguity-Preserving Transposition (CPT-seq)[63]; and the 10X Genomics (Pleasanton, CA, USA) technology that introduced the GemCode/Chromium platform. The 10X system uses the same principles as earlier work by diluting DNA into subhaploid fractions and then generating uniquely indexed libraries from the subhaploid fractions.

Where 10X Genomics differs from the earlier work is in the number of possible subhaploid fragments that are created. Whereas prior work used 96- or 384-well plates to generate hundreds or even low thousands of fractions, the 10X Genomics platform creates >1 million oil-encased droplets, called "gems." Each gem has a 16-nt unique tag and usually holds only 1 long DNA fragment. Within each gem, a series of semirandom priming reactions creates 16 nt-tagged molecules that are then transformed to complete sequencing libraries by shearing and ligation-based methods. The very large number of partitions created by the 10X Genomics platform should allow robust sequencing coverage of very long DNA fragments, and indeed public data have shown phased variants as long as 150 Mb. The related CPT-seq method also relies on creation of subhaploid fractions, followed by tagging of adjacent sequences, but rather than using random priming or ligation-based library preparation, it uses preparation methods via T5 transposases and the consideration that transposition does not fragment DNA until the transposase is disrupted by some means, typically detergent. The use of multidimensional indexing in the preparation and amplification steps allows tens of thousands to hundreds of thousands of indexed libraries to be created, provided the appropriate scale of subhaploid fragments can be created. Each of these library types can be sequenced on an Illumina sequencer and combined with a reference genome for phasing. Likewise, the Long-Fragment Read (Complete Genomics, Mountain View, CA, USA) technology uses a limiting dilution to create 2–3 DNA molecules per well, which can then be tagged with customized barcode adapters and mate-pair sequenced to generate the phased information per molecule.[64, 65]

In contrast to these dilution-based methods, other library approaches that preserve structural information of adjacent molecules have emerged. First, Dovetail Genomics (Santa Cruz, CA, USA)uses a variation of the Hi-C method (called Chicago) for studying chromatin proximity.[66, 67] The company's method uses cross-linking of histones to mark the areas of any long DNA fragment (>150 kb) that is in close proximity. Then, the DNA is digested, subjected to religation for mate pairs, and analyzed with the HiRise software. *A priori*, such synthetic histones can mark any organism's naked DNA. This has been shown in human and alligator,[65] although it is unclear how this will perform for circular DNA-like plasmids or high and low GC content. Another example of crosslinking to resolve DNA–DNA interactions came from the Burton lab,[68] which could profile metagenomic samples to resolve which plasmids were co-occurring with the specific DNA from various bacteria. Second, Base4 Innovation (Cambridge, United Kingdom) is a sequencing company that uses the inverse of DNA synthesis, wherein a pyrophosphate (PPi) is released and instead performs pyrophosphorolysis, where the PPi reacts with the 3′ end of a strand of DNA and removes the last nucleotide as a triphosphate. Each serial-severed nucleotide is captured into a microdroplet and is queried in a "cascade reaction" that creates an optical signal, ideally keeping the precise order of the nucleotides from a single, contiguous fragment of DNA. This is conceptually similar to the Exonuclease-Seq (Exo-Seq) that is performed at the entrance to a nanopore[69] but requires an additional chemical reaction to enable the nucleotide to be optically observed.

Automation

Automation and robotics have improved the throughput for most laboratories, but they have also raised questions about the impact on reproducibility. The common advantage of robotic approaches to library preparation is the consistency of yield and library size from the automation and usually an increase in speed or throughput. However, the risks of automating one's workflow also increase the chances of large-scale failure, as so many samples are being processed at once or with a potentially less efficient reaction volume. A notable exception to this is the microfluidic devices that miniaturize the protocols to use very small reaction volumes, which save sample input, as well as enzyme and reagent use. This includes the Fluidigm (San Francisco, CA, USA) C1, as well as the Becton Dickinson (Franklin Lakes, NJ, USA) CLiC Gencell composite liquid cell methods that use noncontact microfluidics mix reagents.

By building on those library preparation techniques described above, several novel approaches have emerged for large-scale automation of library preparation, with an emphasis on approaches to single-cell work. There are myriad methods for microfluidic manipulation of single cells, or samples have become very automated, with platforms, such as the Fluidigm C1,[70] Drop-seq,[71] inDrop,[72] 10X Genomics,[73] and others (**Table 2**), enabling routine single-cell analysis of transcriptomes, exomes, genomes, or targeted sequencing for many (>10,000) cells at once. Additionally, some hybrid protocols have emerged to enable profiling of DNA and RNA from the same cells (G&T-seq),[74] examination of single-cell epigenetic states with single-cell WGBS (scWGBS),[38] or even simultaneous examination of genetic, epigenetic, and transcriptome states (scTrio-seq).[75] All of these methods are being migrated to automated systems, which will create ample opportunity for addressing single-cell research questions. Overall, there are collectively almost 1 dozen means by which to lyse, extract, and prepare samples for sequencing for low-input or single-cell methods (Table 2).

PREPARATION OF METAGENOMES

Bacterial genome assembly was traditionally a very difficult problem that necessitated the expertise of microbiology, biochemistry, and genetics. However, with the current tools and technologies available, the means are widely available to quickly, correctly, and easily close complete bacterial genomes and plasmids with long-read technologies, such as ONT and PacBio. Moreover, the use of single-molecule approaches creates unprecedented opportunities to gauge the genetic and epigenetic states of pathogens or other organisms of interest simultaneously.

Notably, these types of information have already shown some rapid use for understanding outbreaks of disease. For example, during the outbreak of Haitian cholera in 2010, researchers used single-molecule methods and the kinetic information of base incorporation from the DNA polymerase used in PacBio's RS instrument to discern the genetic

TABLE 2

Comparison of technologies for single-cell automated capture, preparation, and sequencing

Source	Instrument	Number of cells	Input cells	Est. cost per run, $	Est. cost per cell, $	UMIs	Cell phenotype	DNA	RNA	ATAC	3′	Full cDNA	Size range, μm
10X Genomics	Chromium	60,000	100,000	10,000	0.17	Yes	No	No	Yes	unk	Yes	No	1–60
Becton Dickinson	FACSseq/BD Precise	96	unk	10,000	104.17	Yes	No	unk	unk	unk	unk	unk	5–100
Becton Dickinson	Resolve	10,000	50,000	10,000	1.00	Yes	Yes	unk	unk	unk	unk	unk	5–100
Bio-Rad–Illumina	ddSeq	10,000	unk	10,000	1.00	unk	No	No	Yes	unk	unk	unk	unk
Drop-seq	Drop-seq	10,000	100,000	1000	0.10	Yes	No	No	Yes	Yes	Yes	No	1–100
Fluidigm	C1	96	5000	1900	19.79	Yes	No	Yes	Yes	Yes	No	Yes	5–10, 11–17, 17–24
Fluidigm	scRRBS	96	5000	1900	45.00	Yes	No	Yes	Yes	Yes	No	Yes	5–10, 11–17, 17–24
Fluidigm	C1–high throughput	800	5000	4000	5.00	Yes	No	Yes	Yes	Yes	Yes	No	5–10, 11–17, 17–24
Fluidigm	Polaris	800	5000	10,000	12.50	No	Yes	No	Yes	No	Yes	Yes	5–10, 11–17, 17–24
inDrop	Custom	10,000	100,000	5000	0.50	Yes	No	No	Yes	No	Yes	No	5–100
RainDance	RainDrop	unk	unk	unk	unk	Yes	No	unk	unk	unk	unk	unk	unk
Qiagen	CellRaft (Cell Microsystems)	44,000	unk	unk	unk	unk	No	unk	unk	unk	unk	unk	unk
WaferGen	ICELL8	1800	40,000	2750	1.53	Yes	Limited	Soon	Yes	unk	unk	Maybe	5–100

UMIs, unique molecular index; ATAC, Assay for Transposase-Accessible Chromatin; scRRBS, single-cell Reduced Representation Bisulfite Sequencing; RainDance, RainDance Technologies (Billerica, MA, USA); Cell Microsystems (Research Triangle Park, NC, USA); WaferGen, WaferGen Biosystems (Fremont, CA, USA); unk, unknown.

and epigenetic signature of a strain of *Vibrio cholera* and also predicted its likely source.[76] Another outbreak that occurred in Germany for *Escherichia coli* showed that the virulent bacteria show a very specific profile of methyl-6-adenosine that was implicated in its virulence.[77] This has led to rapid assemblies of many other bacterial genomes with long reads or short and long reads,[78] and from these works, the number of completed bacterial genomes now numbers in the thousands.

However, the type of sample preparation for metagenomics and microbiome research, as with the above methods in genomics and transcriptomics, clearly defines the scope of what can be observed. Historically, work in metagenome and microbiome profiling used 1 of 3 methods: 16S profiling of the variable regions (V4, V5) of the rRNA subunit has historically been used for bacteria/archaea, the 18S rRNA subunit for eukaryotes, and the internal tandem spacer (ITS) sequence for fungi. Yet, work in the past few years has shown how limited these methods are, both for their intended purpose of taxonomic identification and also for their absence of co-occurring and often more informative genetic data about pathogenicity or virulence. Specifically, the 16S marker (even when using multiple V regions) is one of the least effective genes for distinguishing closely related species, and it is not even the best gene for distinguishing distantly related species.[79, 80]

Perhaps most importantly are the distinctions of informative context between shotgun sequencing (metagenomics) and targeted amplicon sequencing (16S, 18S, ITS). The amount and diversity of information that can be obtained from shotgun sequencing are inherently far greater (**Table 3**) for a given sample, but there are nonetheless caveats and challenges to address in sequencing all of the DNA of a sample. For clinical samples, this may create a large proportion of reads from the host, which has privacy and Health Insurance Portability and Accountability Act of 1996 concerns, but also may waste sequencing costs on a target of noninterest. For environmental samples, other contaminants (mammalian or plant DNA) can reduce one's ability to observe the microbial diversity of a sample because of the potential loss of sequencing depth. Moreover, whereas shotgun sequencing allows one to study DNA from all kingdoms of life, it also introduces challenges, such as inherent biases in many current approaches and extraction/library preparation kits toward bacterial species. For all contexts, the controls listed above (Table 1) are essential for discerning the accuracy and presence of the correct species or strains.

CONCLUSIONS

Since 2006, there has been a rapid development of sequencing technologies, along with the antecedent biochemical methods to extract, generate libraries, and automate the preparation/capture of biologic samples. Such a rapid development of sequencing technologies, sample preparation, and computational methods to analyze the data has led to some uncertainty regarding which technology is appropriate, useful, or relevant. Moreover, the continuing emergence of new technologies creates excitement to implement the latest methods but often unknown accuracy. Ideally, new methods, technologies, and protocols need to be benchmarked against known standards and measures (Table 1) to ensure their use and potential improvement over the state-of-the-art methods.

Such standards have never been more paramount, especially as we enter an era of "ubiquitous sequencing" that can highlight the promise and perils of large-scale availability of genomic technologies.[81] Students at almost all ages can implement extraction, library preparation, and automation for genomes, including sequencing, as a part of coursework[82] or anything found in their home or subway.[83] However, without proper physical, library, and computational controls, the data generated may not only be unusable but potentially misleading, as measured species can sometimes be only DNA found in the sampling kit.[84] As such, the use of titrated molecular controls for the instrumentation, extraction, preparation, and sequencing is essential to interpretation of data, both in the lab and in the field.

Nonetheless, when looking at the sample preparation methods for genomics applications across all layers of biology (genome, epigenome, transcriptome, epitranscriptome), it is clear that an era of long-read, fully phased genomes and single-cell methods has just begun and brings considerable opportunity. The largest challenge for implementation is to leverage such information for problems that

TABLE 3

Comparison of metagenomic assay capabilities and limitations

Data type	16S	18S	ITS	Shotgun
Taxonomic classification	Yes	Yes	Yes	Yes
Prokaryotes	Yes	No	No	Yes
Archaea	Yes	No	No	Yes
Eukaryotes	No	Yes	Yes	Yes
Parasites	No	Yes	No	Yes
Plasmids	No	No	No	Yes
Phages	No	No	No	Yes
Human ancestry	No	No	No	Yes
Biosynthetic gene clusters	No	No	No	Yes
Antimicrobial resistance markers	No	No	No	Yes
Kingdom specificity	Yes	Yes	Yes	No
Removal of host DNA	Yes	Yes	Yes	No

actually require it. Although phased genomic data are known to be important for certain types of cancer[85, 86] and even for just understanding the structure of the genome,[87] it still remains unclear how many diseases and areas of the genome will dramatically benefit from such information. Furthermore, although tumor heterogeneity can be measured with unprecedented detail and precision, a single clone that does not require resolution through single-cell approaches may drive some cancers.

However, in all cases, the implementation of standards and controls for benchmarking have established the veracity of new methods and helped catapult them to broader use. As shown above, the controls for DNA and RNA are now well established and commonly implemented. However, the physical standards for epigenomes and epitranscriptomes are just now being developed, and controls for metagenomes, beyond just bacterial DNA, are also emerging. Finally, there are no standards or current controls for single-cell biology; instead, many groups use peripheral blood mononuclear cells for test runs, but these will harbor different proportions of cells among individuals. Indeed, until a perfectly reproducible, synthetic biology construct is made for prokaryotic and eukaryotic cells, there will always be some biologic noise as a part of the genesis of a standard. Beyond this small aspect of noise, however, are the controls described above, to tease out the other components of technical, technician, site, and laboratory noise. Their removal is essential to eavesdropping more closely on the actual biology, from a single cell to an entire organisms and their ecosystem.

ACKNOWLEDGMENTS

The authors thank Sofia Ahsanuddin for help in formatting the manuscript for submission. The authors also thank the following for funding: Starr Cancer Consortium (Grants I7-A765, I9-A9-071), Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L. and N. Kuggie Vallee Foundation, WorldQuant Foundation, Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX17AB26G), and Bill and Melinda Gates Foundation (OPP1151054).

DISCLOSURES

The authors herein declare that this research was conducted in the absence of any financial or commercial interests that could be potentially regarded as a conflict of interest.

REFERENCES

1. Mason CE, Porter SG, Smith TM. Characterizing multi-omic data in systems biology. *Adv Exp Med Biol* 2014;799:15–38.
2. Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE. The birth of the epitranscriptome: deciphering the function of RNA modifications. *Genome Biol* 2012;13:175.
3. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;11:31–46.
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333–351.
5. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;18:1509–1517.
6. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;29:365–371.
7. Munro SA, Lund SP, Pine PS, et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat Commun* 2014;5:5125.
8. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 2016;3:160025.
9. Chen CY. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Front Microbiol* 2014;5:305.
10. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;24: 1151–1161.
11. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 2014;32: 903–914.
12. Xu J, Su Z, Hong H, et al. Cross-platform ultradeep transcriptomic profiling of human reference RNA samples by RNA-Seq. *Sci Data* 2014;1:140020.
13. Li S, Tighe SW, Nicolet CM, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* 2014;32:915–925.
14. Li S, Łabaj PP, Zumbo P, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* 2014;32:888–895.
15. Gargis AS, Kalman L, Bick DP, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat Biotechnol* 2015;33:689–693.
16. GEUVADIS Consortium. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol* 2013;31:1015–1022.
17. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;32:896–902.
18. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 2012;40:e72.
19. Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci USA* 2014;111:9869–9874.
20. Chaisson MJ, Huddleston J, Dennis MY, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 2015;517:608–611.
21. Afshinnekoo E, Meydan C, Chowdhury S, et al. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst* 2015;1:72–87.
22. Yooseph S, Andrews-Pfannkoch C, Tenney A, et al. A metagenomic framework for the study of airborne microbial communities. *PLoS One* 2013;8:e81862.
23. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. The microbiome quality control project: baseline study design and future directions. *Genome Biol* 2015;16:276.
24. Li S, Garrett-Bakelman F, Perl AE, et al. Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biol* 2014;15: 472.
25. Li S, Garrett-Bakelman FE, Chung SS, et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat Med* 2016;22:792–799.

26. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SJ. Comprehensive analysis of mRNA methylation reveals pervasive adenosine methylation in 3′ UTRs. *Cell* 2012; 149:1635–1646.
27. Saletore Y, Chen-Kiang S, Mason CE. Novel RNA regulatory mechanisms revealed in the epitranscriptome. *RNA Biol* 2013; 10;342–346.
28. Li S, Mason CE. The pivotal regulatory landscape of RNA modifications. *Annu Rev Genomics Hum Genet* 2014;15:127–150.
29. Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv beta* Aug. 12, 2016. Available at: biorxiv.org/content/early/2016/08/12/068809.
30. Miyamoto M, Motooka D, Gotoh K, et al. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics* 2014;15: 699.
31. McIntyre AB, Rizzardi L, Yu AM, et al. Nanopore sequencing in microgravity. *bioRxiv beta* Dec. 10, 2015. Available at: http://biorxiv.org/content/early/2015/12/10/032342.
32. Castro-Wallace SL, Chiu CY, John KK, et al. Nanopore DNA sequencing and genome assembly on the International Space Station. bioRxiv beta Sept. 27, 2016. Available at: http://biorxiv.org/content/early/2016/09/27/077651.
33. Lever MA, Torti A, Eickenbusch P, Michaud AB, Šantl-Temkiv T, Jørgensen BB. A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types. *Front Microbiol* 2015;6: 476.
34. Fisher Scientific. Thermo Scientific™ ART™ Non-Filtered Extended-Length Wide-Bore Genomic pipette Tips. Available at: https://www.fishersci.com/shop/products/thermo-scientific-art-non-filtered-extended-length-wide-bore-genomic-pipette-tips-3/p-3206875.
35. Pan X, Durrett RE, Zhu H, et al. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci USA* 2013;110:594–599.
36. Oxford Nanopore Technologies. VolTRAX™: rapid, programmable, portable, disposable sample processor. https://nanoporetech.com/publications/2016/05/26/voltrax-rapid-programmable-portable-disposable-sample-processor.
37. Zumbo P, Mason CE. Methods for RNA isolation, characterization, and sequencing. In: Genome Analysis: Current Procedures and Applications. Poole, UK: Caister Academic, 2014:Chapt. 2.
38. Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res* 2012;22:1139–1143.
39. Tatsumi K, Nishimura O, Itomi K, Tanegashima C, Kuraku S. Optimization and cost-saving in tagmentation-based mate-pair library preparation and sequencing. *Biotechniques* 2015;58:253–257.
40. Dunham JP, Friesen ML. A cost-effective method for high-throughput construction of Illumina sequencing libraries. *Cold Spring Harb Protoc* 2013;2013:820–834.
41. Cantara WA, Crain PF, Rozenski J, et al. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res* 2011;39:D195–D201.
42. Ayub M, Hardwick SW, Luisi BF, Bayley H. Nanopore-based identification of individual nucleotides for direct RNA sequencing. *Nano Lett* 2013;13:6144–6150.
43. Ozsolak F, Platt AR, Jones DR, et al. Direct RNA sequencing. *Nature* 2009;461:814–818.
44. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SJ. Comprehensive analysis of mRNA methylation reveals pervasive adenosine methylation in 3′ UTRs. *Cell* 2012; 149:1635–1646.
45. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012;485:201–206.
46. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 2015;12:767–772.
47. Landan G, Cohen NM, Mukamel Z, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet* 2012;44:1207–1214.
48. Vilfan ID, Tsai YC, Clark TA, et al. Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J Nanobiotechnology* 2013;11:8.
49. Dominissini D. Genomics and proteomics. Roadmap to the epitranscriptome. *Science* 2014;346:1192.
50. Cancer Genome Atlas Research Network. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 2010;17:510–522.
51. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* 2015;526:525–530.
52. Pan H, Jiang Y, Boi M, et al. Epigenomic evolution in diffuse large B-cell lymphomas. *Nat Commun* 2016;6:6921.
53. Vembar SS, Seetin M, Lambert C, et al. Complete telomere-to-telomere de novo assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing. *DNA Res* 2016;23:339–351.
54. Kurylo CM, Alexander N, Dass RA, et al. Genome sequence and analysis of *Escherichia coli* MRE600, a colicinogenic, nonmotile strain that lacks RNase I and the type I methyltransferase, EcoKI. *Genome Biol Evol* 2016;8:742–752.
55. Feng Z, Fang G, Korlach J, et al. Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLOS Comput Biol* 2013;9: e1002935.
56. Genest PA, Baugh L, Taipale A, et al. Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing. *Nucleic Acids Res* 2015;43:2102–2115.
57. Simpson JT, Workman R, Zuzarte PC, David M, Jonathan L, Timp WD. Detecting DNA methylation using the Oxford Nanopore Technologies MinION sequencer. bioRxiv beta April 4, 2016. Available at: http://biorxiv.org/content/early/2016/04/04/047142.
58. Rand AC, Jain M, Eizenga J, et al. Cytosine variant calling with high-throughput Nanopore sequencing. bioRxiv beta April 4, 2016. Available at: http://biorxiv.org/content/early/2016/04/04/047134.
59. Stranges PB, Palla M, Kalachikov S, et al. Design and characterization of a nanopore-coupled polymerase for single-molecule DNA sequencing by synthesis on an electrode array. *Proc Natl Acad Sci USA* 2016;113:E6749–E6756.
60. Madoui MA, Engelen S, Cruaud C, et al. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 2015;16:327.
61. Kaper F, Swamy S, Klotzle B, et al. Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc Natl Acad Sci USA* 2013;110:5552–5557.
62. Kuleshov V, Xie D, Chen R, et al. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* 2014; 32:261–266.
63. Amini S, Pushkarev D, Christiansen L, et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat Genet* 2014;46:1343–1349.
64. Peters BA, Kermani BG, Sparks AB, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 2012;487:190–195.
65. Peters BA, Liu J, Drmanac R.. Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. *Front Genet* 2015;5:466.

66. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326: 289–293.
67. Putnam NH, O'Connell BL, Stites JC, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 2016;26;342–350.
68. Burton JN, Liachko I, Dunham MJ, Shendure J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)* 2014;4: 1339–1346.
69. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 2009;4:265–270.
70. Arguel MJ, LeBrigand K, Paquet A, et al. A cost effective 5′ selective single cell transcriptome profiling approach with improved UMI design. *Nucleic Acids Res* 2016 [Epub ahead of print]. doi:10.1093/nar/gkw1242
71. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;161:1202–1214.
72. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;161:1187–1201.
73. Coombe L, Warren RL, Jackman SD, et al. Assembly of the complete Sitka spruce chloroplast genome using 10X Genomics' GemCode sequencing data. *PLoS One* 2016;11: e0163059.
74. Macaulay IC, Haerty W, Kumar P, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* 2015;12:519–522.
75. Hou Y, Guo H, Cao C, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* 2016;26: 304–319.
76. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. *N Engl J Med* 2011;364:33–42.
77. Rasko DA, Webster DR, Sahl JW, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* 2011;365:709–717.
78. Bashir A, Klammer AA, Robins WP, et al. A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol* 2012;30:701–707.
79. Lan Y, Rosen G, Hershberg R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome* 2016;4:18.
80. Mao DP, Zhou Q, Chen CY, Quan ZX. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol* 2012;12:66.
81. Erlich Y. A vision for ubiquitous sequencing. *Genome Res* 2015; 25:1411–1416.
82. Zaaijer S, ; Columbia University Ubiquitous Genomics 2015 class, Erlich Y. Using mobile sequencers in an academic classroom. *eLife* 2016;5:e14258.
83. MetaSUB International Consortium. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome* 2016;4:24.
84. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014;12:87.
85. Smith CC, Wang Q, Chin CS, et al. Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia. *Nature* 2012;485:260–263.
86. Li S, Garrett-Bakelman FE, Akalin A, et al. An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics* 2013;14 (Suppl 5):S10.
87. Rosenfeld J, Mason CE, Smith T. Limitations of the human genome reference. *PLoS One* 2012;7:e40294.