Witnesses to a Sea Change in Sequencing Capability

Three luminaries provide their perspective on the impact of high-throughput and population sequencing in clinical research and the role they will play in the future of medicine.

Introduction

Sequencing technologies have far surpassed the expectations of Drs. Carlos Bustamante, Stephen Kingsmore, and John Mattick. Had you asked them at the beginning of their careers if one day we could sequence a whole human genome in a day, their responses would have been, respectively: "Crazy talk!", "Absolutely not." and "Not in my wildest dreams."

Although the pace of sequencing innovations surprised them, each was quick to adopt next-generation sequencing (NGS), and now population sequencing, to advance their research and translational efforts. As Professor of Genetics and Biomedical Data Science, and founding Director of the Stanford Center for Computational, Evolutionary and Human Genomics, Dr. Bustamante is using population sequencing to understand genetic variances in ancient and ethnic subpopulations. In his new role as President and CEO of the Rady Children's Institute for Genomic Medicine, Dr. Kingsmore is using it to develop the evidence base for genomic medicine in children. As the Executive Director of the Garvan Institute of Medical Research, Dr. Mattick is leading efforts to leverage population sequencing data for research and clinical applications.

iCommunity spoke with Drs. Bustamante, Kingsmore, and Mattick about how their teams are using high-throughput whole human genome and population sequencing to advance research and translational studies, the need for databases that merge "omics" and phenotypic data, and the challenges of transforming this information into a format that's useful in a clinical environment.

Q: What was sequencing technology like when you first became a scientist?

John Mattick (JM): My first memories of sequencing are peering at bands on autoradiograms. It was the early days of molecular biology. We were cloning and sequencing genes. We thought we were hotshots. We could only read a couple of 100 bases from the gels before the bands were too tight to distinguish. We would assemble a sequence that was 1–2 kilobases long and each would be a separate paper. Looking back, it seems so primitive.

Stephen Kingsmore (SK): My sequencing experience began with radioactive p32 labeling, and agarose and polyacrylamide gels. A great sequencing reaction was 150 nucleotides and that took most of the day to do.

Carlos Bustamante (CB): I became a scientist as automated sequencers were being developed, so I performed a little manual sequencing and then a large amount of sequencing on first-generation sequencers. My first experience was as an intern at The Smithsonian where they had just set up the Laboratory of Molecular Systematics. At the time, sequencing a couple of genes from multiple individuals was a huge deal.



From left to right: Dr. Carlos Bustamante is Professor of Genetics and Biomedical Data Science, and founding Director of the Stanford Center for Computational, Evolutionary and Human Genomics; Dr. Stephen Kingsmore is President and CEO of the Rady Children's Institute for Genomic Medicine; and Dr. John Mattick is Executive Director of the Garvan Institute of Medical Research.

Q: How has your approach to sequencing changed as the tools improved?

CB: In the beginning, we treated every piece of data as if it was precious. When Celera began performing early exome sequencing, they performed PCR on 200,000 samples, and sequenced 39 people across 20,000 genes. I thought, "This is a data set! We've waited a long time for this." We stopped what we were doing and spent 4–5 years studying the 39 exomes, and wrote 8–9 papers analyzing the data in different ways. That mindset has been flipped on its head. We're now generating data quickly and continually with NGS, and then worrying about what it means.

> "The only way to have accurate variant information is for hundreds of thousands of genomes to be available so that we can assess the frequency of every variant that we see."

Q: When next-generation sequencing (NGS) tools were introduced, how quickly did you incorporate them into your research studies?

CB: NGS quickly became a critical tool for our studies. We were part of the macaque and orangutan genome projects, where we analyzed polymorphism data. We were also one of the original analysis groups for the 1000 Genomes Project, designing the sampling in the Americas, determining the value of 2×–4× sequencing, and the bounds of variance frequencies.

SK: We began using NGS systems soon after they were on the market. Those were exciting days. We converted our mail room into an NGS lab. Not much was known about the human genome, so we were discovering new things in every study we performed.

JM: I've been an early adopter of new genomics technologies for many years. Along with Craig Venter, I was one of the first customers for the Molecular Dynamics Megabase sequencer. The Garvan Institute was one of the first three institutions to acquire a HiSeq X™ Ten System.

Q: How did your early sequencing work inform the focus of your current studies?

CB: Early on, we saw polymorphism and variation in genes of interest. In my PhD thesis, I analyzed the largest genome data set at that time, which consisted of 25 Drosophila genes sequenced across multiple individuals and 15 Arabidopsis genes sequenced across multiple plants. We were looking at amino acid differences and the accumulation of good and harmful mutations. From that moment on, I started thinking about creating a large data set of human sequences so that we could analyze it in the same way.

SK: At the National Center for Genome Resources, we used early NGS to sequence transcriptomes and then the genomes of plants and pathogens, and then began sequencing human samples. Several of us realized that the studies we were performing in a research setting would soon begin to impact medical care. After looking around the country, three of us moved to Children's Mercy Hospital in Kansas City to establish one of the first pediatric genomics medicine centers and began performing translational research. I'm now at the Rady Children's Institute for Genomic Medicine where we're taking that a step further, focusing on implementation of genomic systems medicine at scale in the largest children's hospital in California.

JM: High-throughput sequencing had a huge impact on the appreciation of the transcriptional complexity of the human genome. NGS accelerated our ability to dive into the transcriptome, enabling us to explore the extraordinary world of non-protein coding transcripts, which pour out from the genome in precise patterns in different cells and tissues during development. I now think of the human genome as the .ZIP file extraordinaire. The transcriptional complexity of the human genome is at least an order of magnitude more complex than the genome itself, and it can be unzipped in different ways, with different expression and splice patterns of coding and noncoding RNAs in different cells at different times. We would have had no way to explore this world without high-throughput sequencing.

Q: How are you using NGS today?

CB: NGS has opened up new avenues in population genomics. I remember being at a Cold Spring Harbor meeting and realizing that the 1000 Genomes Project should include admixed genomes. People questioned it, but I believed that to analyze and perform transethnic and multiethnic studies we needed to figure out how to make sense of an admixed genome.

> "In the new world of genomics, every student, post doc, laboratory, and department will need to have the ability to handle and analyze Big Data."

One of the reasons we became involved in the Clinical Genome Resource (ClinGen) Consortium was to aggregate clinical genetic testing data and chip away at the variant of uncertain significance (VUS) rate, which is higher in certain ethnic minority groups simply because there haven't been as many of these sequences analyzed. NGS made it inexpensive and easy to follow up on these genome-wide association study (GWAS) hits. Each amino acid change we found was a smoking gun. It became clear that we needed to broaden ethnic representation in human DNA studies if we really wanted to develop genomic medicine that benefitted everybody.

SK: We're focusing on whole-genome sequencing (WGS) because it's the ultimate molecular test. WGS is also faster and we've worked with Illumina to develop a method that allows us to decode and analyze an entire human genome in 26 hours.[1] It's our plan to offer rapid WGS to every undiagnosed child in our neonatal and pediatric intensive care units (NICU and PICU) by the middle of next year, and to perform clinical research studies to define clinical utility and cost-effectiveness of genomic medicine in pediatric inpatient and outpatient settings.

> "Our biggest challenge is learning how to share population sequencing data."

Q: What are the HiSeq X Systems enabling you to study?
CB: Population sequencing is the culmination of what I've always wanted to do—analyze many human genomes. We're performing large population sequencing studies, using them as the baseline to answer important population genetic questions, and analyzing the results to inform new approaches to clinical medicine. For example, we're conducting a preeclampsia study in Peru using both a mixture of large-scale genotyping and sequencing, looking at altitude adaptation as it's linked to preeclampsia.

SK: Using the HiSeq X Systems, genomes are much less expensive so we can sequence many more trios. There are 8000 named genetic diseases and we and others feel strongly that NGS is going to transform our ability to identify them. We hope to use the HiSeq X and Illumina SeqLab infrastructure to gradually develop the evidence base to support that.

JM: The Garvan Institute was one of the first institutes to put genomics at the center of its research endeavor, rather than as an extension of conventional molecular biology. With the extraordinary advances in genome sequencing and concomitant cost reductions, it has become feasible economically to leverage population sequencing and put genomics at the center of both research and the clinic.

It's extraordinary how the HiSeq X Systems are enabling translational and research endeavors to merge. We've been collaborating with researchers throughout the world. The HiSeq X Ten Systems are working beautifully.

In addition to studying monogenic diseases, we are using population sequencing for major research programs in cancer, diabetes, osteoporosis, immunological diseases, neurodegenerative and neuropsychiatric diseases, and aging. We're performing cancer stratification studies as part of the International Cancer Genome Consortium (ICGC), and using NGS to decipher the cancer genome and assess the inherited components of familial cancer risk. We are sequencing people with type 1 diabetes to discover genetic differences between those with the condition who do well through life, and those who suffer severe complications later in life, such as renal failure. In our aging studies, we're using population sequencing to study several thousand individuals who have reached old age without any sign of cardiovascular, cancer, cognitive decline, or neurodegenerative disease. We're developing a risk depleted cohort that we can use as a control for studies of populations that do suffer such diseases. Other programs underway using the HiSeq X Ten sequencing capacity are to study populations with cardiac, mitochondrial, and Alzheimer's diseases.

Q: What are the challenges in sharing population sequencing data?
CB: Our biggest challenge is learning how to share population sequencing data. The NIH and other organizations now mandate that researchers share their data. Unfortunately, this is not true for clinical data. Most hospitals have no real tenet to share data. We also live in a world that is interconnected, and that is making patients uncomfortable in sharing information. That's where the efforts of the Global Alliance for Genomics and Health and other entities will be valuable in developing forward-looking consent, privacy procedures, and best practices in data governance and transparency.

> "With the extraordinary advances in genome sequencing and concomitant cost reductions, it has become feasible economically to leverage population sequencing and put genomics at the center of both research and the clinic."

SK: Before we can sequence a genome at Rady Children's Hospital, parents have to give informed consent. Part of that consent process is an agreement for us to be able to post the genome. We de-identify it so there's no information that can tie the genome back to the child or parent, then the information is made available on the National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGaP), a private database. Researchers can obtain access to the data only after applying to NIH and providing a good reason why they need to access the information for their research. It seems to provide a good balance between privacy concerns and the benefit of other researchers being able to study public genomes.

It's unfortunate that not all hospitals have a genome sharing informed consent process in place. Clinical researchers need human whole genome sequence information for benchmarking. They want to see how common a variant is in a genome. The only way to have accurate variant information is for hundreds of thousands of genomes to be available so that we can assess the frequency of every variant that we see.

Q: What is the value in integrating WGS, epigenome, transcriptome, and other genomic and phenotypic data to obtain different genomic snapshots?

CB: There's significant value in performing all kinds of omics profiling, RNA-Seq, methylome sequencing, etc. We still don't understand the regulatory network of the human body. Are we performing and integrating omics data today? I think it's happening slowly and part of that is because it's much easier to sequence than to interpret.

> "There is definitely value in panomics, where we're taking whole-genome data and bringing it together with deep phenome, epigenetic, gene expression, metabolomic, and proteomic data."

SK: There is definitely value in panomics, where we're taking whole-genome data and bringing it together with deep phenome, epigenetic, gene expression, metabolomic, and proteomic data. Sequencing the genome is not the end of the game, but it's a great start. We're starting to understand what we need to deliver precision medicine. For example, we don't know what most of the variants that we see in genomes mean functionally. Therefore, we can't give a confident assessment of whether they could produce a change in a human being. It's clear that we need additional types of data to be able to make those assessments at scale.

JM: The future of clinical research and medicine will revolve around the integration of Big Data sets. It's more than just individual and amalgamated genomic data sets. Increasingly, these will become merged with transcriptomic, epigenomic, proteomic, and most importantly, phenotypic data to create highly connected, information-rich data sets. Medicine is heading quickly towards Big Data and the acquisition of tens and hundreds of thousands of genome sequences will accelerate this. It's going to change everything.

Q: How important will bioinformatics and databases be in gaining the full value of population sequencing?

CB: From the beginning, it was clear that we would have to marry sequencing with analysis tools to make sense of all the data. By linking and analyzing phenotypic and genotypic information, we can begin to unravel patterns that we can't see from static data. There's an optimism that if we measure phenotypes and exposures in much more rigorous ways, we could collect vast amounts of data to help us nail genetic associations.

JM: I think the bioinformatics framework and databases are central to the whole endeavor. It will integrate genomic data with orthogonal data sets to extract valuable information. The genetic patterns we identify will help inform individual circumstances in the clinic, and through the analysis of the metadata, entire health systems in terms of patterns of disease, co-morbidities, etc.

Population sequencing isn't for the faint hearted. We've invested about $10 million over the last 1–2 years into building the computational pipelines. We have a growing team of 60 people working on the entire assembly pipeline, performing sequencing, assembling data, calling variants and variant difference between populations, and connecting the data with phenotypic data.

In the new world of genomics, every student, postdoc, laboratory, and department will need to have the ability to handle and analyze Big Data. It's not something for specialists at the end of the corridor. It's central to the entire endeavor of research and medicine. It's a data driven world and we're charging into it.

SK: We recognized the value of bioinformatics in a recent study that compared the effectiveness of WGS and traditional genetic testing to identify Mendelian disorders in critically ill newborns.[2] To analyze the data, we developed several novel bioinformatics tools. The paper demonstrates the usefulness of genome sequencing, but we need further evidence of the clinical value of genomics. We'll also need a streamlined method for informing clinicians of the results, not just for diagnosis, but also for how NGS data can inform treatment decisions.

Q: What kinds of databases will be required?

JM: We need national-level genotype/phenotype correlation databases that are maintained by health authorities and can be queried by accredited researchers and clinicians. They'll have to be national databases because there are legal and other contextual requirements that are idiosyncratic to each jurisdiction. Somehow they need to be linked into one global database so that data generated in one country can be used elsewhere and explored in multidimensional ways to advance our understanding of human biology and disease.

> "I think population-scale sequencing in the broadest sense will begin with children, possibly at birth to replace the present Guthrie test."

Q: How long will it take to create these databases?

JM: We can't sequence everyone in the world overnight, but I'm convinced that within a decade we'll have large genomic databases. Genomic data will increasingly become a standard part of medical records. Ideally, we'll have well curated, evidence-based genotype/phenotype correlation databases in the cloud that are maintained and continuously updated national resources.

The initial use will be sequencing individuals with serious genetic disability, because we can diagnose the causative mutation in about half of such cases very quickly. Cancer stratification will be an important area, enabling physicians to determine the molecular

basis of the disease and consequently treat the disease more effectively. The third area will be to detect the genetic markers of adverse drug reactions because that's a huge burden on the hospital systems in every country. We'll be able to predict and avoid a high proportion of those adverse reactions through genomic information.

We're proposing that the Australian health system sequence everyone with developmental and/or intellectual disabilities as a first-line diagnostic. I expect that will become routine over the next 2–5 years. I think population-scale sequencing in the broadest sense will begin with children, possibly at birth to replace the present Guthrie test. The next generation of kids will be the genome generation, with genome sequencing and analysis applied selectively and then more widely as the technology and the value of the information improves.

> "Population sequencing will enable us to uncover and characterize global allele frequencies of clinically actionable variants involved in adverse reactions."

Q: Do you think WGS will become a routine clinical test?

JM: We're close to sequencing being used routinely as part of a medical examination. The cost of sequencing will continue to decrease, making it feasible to perform reanalysis to improve the accuracy of someone's primary genome data, to incorporate epigenomic and transcriptomic data, or to look at somatic variations. The value of sequencing will go up as we get more information about what variation in the genome means in biology and medicine. Higher use of sequencing in medicine is now limited by the richness and quality of the databases that sit behind the analysis of that information.

It's worth noting that the American College of Medical Geneticists (ACMG) has mandated reporting on 56 genes because it can have a significant bearing on a patient's future health. We'll start to see well-validated collections of genes that will be either mandated to report or that organizations working in this space will be confident to report back to clinicians and patients, with the list expanding over time.

SK: We have a rich tradition of newborn screening programs where each baby at birth has a heel stick that's tested for 29 conditions. Several groups around the US are starting to investigate what additional information would be provided if we could replace the heel stick with genome sequencing. We don't know yet.

Q: Is human whole genome data already moving us closer to personalized medicine?

CB: I think genome sequencing is going to end up being a part of routine care and a component of people's electronic health records. It's an interesting time because we're in a bit of a transition phase. Sequencing technology has matured and people are implementing high-throughput sequencing and soon will be performing population sequencing routinely.

We need to come up with a concerted plan for aggregating these data, analyzing them, and translating them into health benefits as quickly as we can. Ultimately, we need to provide the public a good return on the investment.

SK: In the future, sequencing results will inform treatment changes. Traditionally, the diagnostic sphere has been the home of the pathologist and the laboratorian, while medical implementation has been the role of the physician and clinician. In genomic medicine, those two will be fused. That's going to be a challenge because neither side is used to having the other side involved in those tasks or information.

JM: I think the problem is that our understanding of the genome is still limited. Today, we can only accurately report on the impact of some variations in protein-coding sequences. It's a huge effort to assemble enough evidence and data from the literature to confidently call mutations or variations in other parts of the genome that might have medical significance. Large global databases created through population sequencing will support this effort. These databases will contain sequences that reflect a spectrum of mutations and phenotypic characteristics, will enable queries to determine if a new sample reflects the symptoms and mutations of those already in the databases.

> "Particularly in the US, we need population sequencing of ethnic populations that have the worst health outcomes so the negative gap in their care doesn't increase."

Q: How will the data from population sequencing transform medicine?

JM: Population sequencing will have a profound impact on medicine, changing it from the art of crisis management to the science of good health. We now understand that individual genomic variation and our genetic idiosyncrasies affect our present health and contribute to the risk of future disease, whether it's type 2 diabetes, cancer, rheumatoid arthritis, or Alzheimer's disease. In many cases, forewarned is forearmed, enabling clinicians and patients to implement strategies to reduce, avoid, or prepare for these eventualities.

SK: I study rare genetic diseases in children, which are simple genetically. We now have the ability to make rapid diagnoses and so, for the first time, those conditions can be a cost-effective place to develop and manufacture a drug. Our hope is that genomes will increasingly be as valuable with diagnosing complex disease as they are with single-gene disorders. It's going to take a couple of decades to catch up, and population studies will be very

important in closing the gap. One of the things that is exciting about population studies is that we're starting to redefine how we describe diseases based on genetics, rather than based on symptoms.

JM: Population studies will inform the development of therapeutics, especially in identifying the genetics of adverse reactions. There are 100,000 deaths a year in the United States from adverse drug reactions to prescription drugs.[3] In Australia, at least 2–3% of all hospital admissions are due to adverse reactions to prescribed drugs.[4]

CB: For example, Abacavir is an important HIV drug and researchers have identified an HLA variant involved in Abacavir hypersensitivity. Prevalence of the variant is low in Africans and Europeans, but there is a 20% frequency of the mutation in certain populations in India and Asia.[5] If a patient with the variant is given Abacavir once, they become very sick. If they are given it twice, they die. Population sequencing will enable us to uncover and characterize global allele frequencies of clinically actionable variants involved in adverse reactions. The bottleneck is going to be making drug metabolism information understandable by the physicians so they'll know to pick drug A vs. drug B, or to give half the dose or double the dose of a drug.

JM: Drug companies are also beginning to use population sequencing to identify exceptional responders in past drug trials. If they can stratify the population and identify the particular genetic background of responders, they can analyze the biochemical pathways involved. They're not only rescuing failed drugs, they're rescuing responder patients for effective, potentially life-saving treatment.

> "Ultimately, there will be automatic reporting of genomic information into the cloud to and from smart devices. It's going to take us places we haven't even dreamt of."

Q: How important will it be to sequence ethnic subpopulations?
CB: The fact that we have the technology to perform population sequencing is awesome. However, we need a concerted effort so that research continues on ethnic subpopulations. Without one, the focus will remain on sequencing large homogenous populations, like the Finns or Icelanders. Although those efforts are important, their benefits don't translate into all populations. Particularly in the US, we need population sequencing of ethnic populations that have the worst health outcomes so the negative gap in their care doesn't increase. This presents a challenge because there's no high-level initiative to fund these efforts. The US government's Precision Medicine Initiative is a great effort, however it doesn't compare to what the UK and other countries are doing. Particularly China, which sees genomics as one of the major planks of their development program.

Q: What has or will be the impact of the $1,000 genome?
SK: The good news is that the $1000 genome exists for population sequencing. What we need in clinical care is for the cost of rapid genome sequencing to decrease to the $1000 genome level, and that hasn't happened yet.

JM: The $1000 genome was a practical and psychological tipping point. It's changed the way we think about technology and what we believe is possible. It's sparked the integration of clinical and research endeavors in a way that we never anticipated or thought would be possible. People now recognize that we're close to shifting from genomics being used as a research tool, to it becoming an everyday clinical analysis tool.

Q: When you first became a scientist, did you believe there would be a day when human whole-genome sequencing could be performed in a day?
CB: I would have said that it was impossible. Crazy talk!

SK: Absolutely not. Even if you took me back to when I was sequencing with my first Solexa System, I couldn't have anticipated that we would be churning out genomes as quickly as we are.

JM: Not in my wildest dreams. In the second half of the 20th century, we were just cutting our teeth in understanding what DNA looked like, what a gene looked like, and developing primitive genomic analysis tools. At the time, everything we were doing was considered to be leading edge, and it was. Now we are moving at warp speed. The 21st century will be the century of biology and medicine. The integration of NGS with Big Data is still unfolding and will be for the foreseeable future. Ultimately, there will be automatic reporting of genomic information into the cloud to and from smart devices. It's going to take us places we haven't even dreamt of. It's a wonderful and exciting time. We are grateful that companies like Illumina have led the way technologically.

Learn more about the Illumina systems mentioned in this article:

HiSeq X Series, www.illumina.com/systems/hiseq-x-sequencing-system/illumina-seqlab.html

References

1. Miller NA, Farrow EG, Gibson M, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Medicine*. 2015; 7(1) 100. do: 10.1186/s13073-015-0221-8.
2. Willig LK, Petrikin JE, Saunders CJ, et al. Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. *Lancet Respir Med*. 2015; 3 (5):377–387.
3. Preventable Adverse Drug Reactions: A Focus on Drug Interactions. U.S. Food And Drug Administration. www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/DrugInteractionsLabeling/ucm110632.htm. Accessed May 16, 2016.

4. Roughead L, Semple S, Rosenfield E. Literature Review: Medication Safety in Australia. www.safetyandquality.gov.au/wp-content/uploads/2014/02/Literature-Review-Medication-Safety-in-Australia-2013.pdf. Published August 2013. Accessed May 16, 2016.

5. Puthanakit T, Bunupuradah T, Kosalaraksa P, et al. Prevalence of human leukocyte antigen B*5701 among HIV-infected children in Thailand and Cambodia: implications for abacavir use. *Pediatri Infect Dis J*. 2013; 32(3): 252–253.

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com