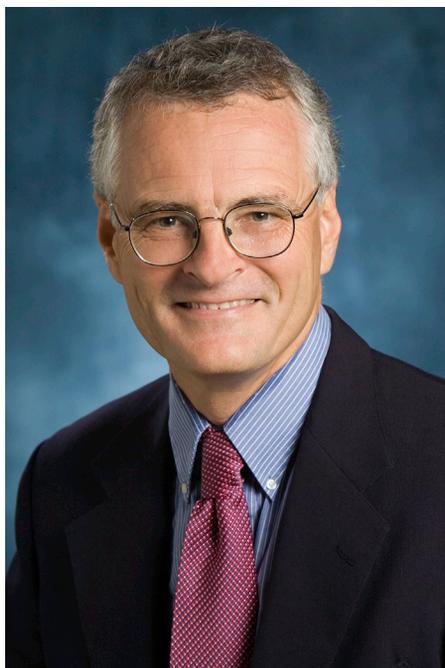


2014 Curt Stern Award Introduction: Gonçalo Abecasis¹Michael Boehnke^{2,*}

Good morning and welcome. On behalf of The American Society of Human Genetics (ASHG), it is my great pleasure to introduce our colleague Gonçalo Rocha Abecasis as recipient (together with Mark Daly) of the 2014 Curt Stern Award.

The Stern Award is presented annually for outstanding research in human genetics in the first 10 years of a research career. The award honors the memory of Curt Stern, a German American geneticist. Stern was a remarkable scientist: he completed his PhD at age 21, was the first to demonstrate physical crossover of homologous chromosomes in *Drosophila*, carried out important research on low-dose radiation safety, and wrote an influential human genetics textbook on which many of us grew up.

Gonçalo Abecasis is another remarkable scientist. He was born in Portugal, grew up in Macau—explaining his penchant for probability—and did his graduate training with Lon Cardon and Bill Cookson at Oxford University in the UK.

I first met Gonçalo at the 2000 ASHG meeting in Philadelphia when he was still a graduate student. I was looking

to hire a junior faculty member, and Lon Cardon encouraged us to meet. As Gonçalo described his dissertation research to me—an efficient method for evaluating probabilities for genetic data in families—I knew in 2 min that the method he was describing was impossible. It took me 5 more min to decide it was not only possible but also a major advance in the field. I recovered quickly enough to recruit Gonçalo to our faculty in biostatistics at Michigan, where he rose quickly through the ranks and was promoted to professor in 2009 at the age of 33.

Gonçalo is that rare individual who is able to identify the essential problems facing our field and to devise ideal solutions. He has made fundamental contributions to complex-disease genetics, including family studies, quantitative-trait mapping, association analysis, and sequence analysis. In each of these areas, he has developed statistically rigorous and computationally efficient methods and computer programs that are now standard tools used by thousands of geneticists worldwide.

Gonçalo's work in genotype imputation particularly stands out. Genotype imputation uses the insight that all of us are related and that some of us detectably share segments of our genomes. Imputation uses this fact to infer genotypes at a dense set of genetic variants in individuals in whom only a sparse set of variants are experimentally genotyped (for example, in a genome-wide association study [GWAS]) by taking advantage of the dense set of genotypes available on a reference set (for example, from the HapMap Project or 1000 Genomes Project). The result is that by genotyping hundreds of thousands of variants, one can infer the genotypes at (tens of) millions of variants. These imputed genotype data, which cost nothing but computer time, provide better genome coverage and increased power to detect disease associations and have been fundamental to the success of modern GWASs and sequencing studies.

In addition to performing methods work, Gonçalo has a keen interest in understanding the genetic basis of health and disease. He is a leader in studies that have identified hundreds of loci for macular degeneration, asthma, psoriasis, type 2 diabetes, and lipid levels and played a leading role in the 1000 Genomes Project. Currently, he is a leader of the Haplotype Reference Consortium, which is bringing

¹This article is based on the address at the meeting of The American Society of Human Genetics (ASHG) on October 20, 2014, in San Diego, CA, USA. The address was given by Aravinda Chakravarti because Michael Boehnke could not be present to deliver these remarks. The audio of the original address can be found at the ASHG website.

²School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, USA

*Correspondence: boehnke@umich.edu

<http://dx.doi.org/10.1016/j.ajhg.2014.11.016>. ©2015 by The American Society of Human Genetics. All rights reserved.

together whole-genome sequence data to create an improved human reference assembly.

Gonçalo is an outstanding communicator, a wonderful collaborator, a gifted teacher, and a generous mentor. He has trained 20 doctoral students and postdoctoral fellows, who have taken important positions in academia and industry.

Gonçalo has received numerous awards: a Pew Biomedical Sciences Scholarship, the Overton Prize from the International Society for Computational Biology, repeat

membership on the annual Thomson Reuters list of the world's ten most frequently cited scientists, and most recently, membership in the Institute of Medicine of the National Academy of Sciences.

Gonçalo Abecasis is a remarkable scientist who has already made fundamental contributions to human genetics and who will make many more in the future. Please join the ASHG and me in congratulating Gonçalo for his contributions to human genetics and as co-recipient of the 2014 Curt Stern Award.

2014 Curt Stern Award: Adventures in Human Genetics¹Gonçalo R. Abecasis^{2,*}

Thank you very much. I would first like to thank the Awards Committee, The Society, and all of you who are here this morning. For the past 15 years, I have treasured coming to this meeting, which is always full of energy, ideas, and brilliant scientists. To be honest, I find it hard to see myself in Aravinda Chakravarti's and Michael Boehnke's kind description of me, and in the next few minutes, I will probably disprove their claim that I am a good communicator. I stand before you trembling.

If you will indulge me, I would like to tell you about the work I have done in human genetics and about what I have learned in the process. The past several years have been remarkable in that they've shown exponential growth in the size of human genetic studies. While preparing this talk, I reviewed the studies in which I have been involved since attending my first American Society of Human Genetics Annual Meeting. If we use the total number of genotypes characterized as a measure of study size, these studies have grown in size by a factor of about 4 every year, about 1,000 every 5 years, and about 1,000,000 every 10 years. It's easy to be hopeful and imagine that this pace might continue for several years.

For someone working on computational methods, this exponential growth presents many opportunities and challenges. It has also been a lot of fun, and I feel all of us are extremely lucky to witness and participate in this era of human genetics.

I started my career in human genetics at the Wellcome Trust Center for Human Genetics at a time when complex-trait studies were shifting from linkage to association. I originally started working in the lab on developing methods for "high-throughput" SNP discovery and genotyping. At that time, our "high-throughput" methods enabled characterization of perhaps hundreds of genotypes per day. Those rudimentary "high-throughput" methods enabled us to generate data at a pace that made analysis with the then-available tools cumbersome. My first paper, published in *The American Journal of Human Genetics*,¹ described methods for analyzing association in samples that typically had been collected for family linkage studies. A companion computer program was able to handle hundreds or thousands of genetic markers and could examine the evidence for association in samples of families or unrelated individuals. During my time at Oxford, I was greatly aided by great colleagues and two very generous mentors, Professors William Cookson and Lon Cardon. They gave me the opportunity to explore interesting datasets and the flexibility to pursue my own ideas, and they shared their time generously.

Soon after, I worked on a method for rapid pedigree analysis,² which Mike described charitably in his introduction. As I reminisce on this work, I would like to share two impressions with you. First, it is clear that I was greatly aided by Leonid Krulyak and Mark Daly's earlier work³ on GeneHunter. (Mark is the 2014 co-recipient of the Curt Stern Award.) Their work served me as a useful guide and reference, allowing me to validate implementations of my new method and sparing me many sleepless nights. Second, I noticed an evaluation we'd carried out with a 10-SNP dataset generated by Bernard Keavney and colleagues.⁴ At the time, our new very fast and memory-efficient approach enabled analyses of that dataset to be completed in ~40 s. That might have been an advance then, but it also represents a level of performance that is unacceptable now that datasets can include information on millions of genetic markers. Time will continue to put

¹This article is based on the address given by the author at the meeting of The American Society of Human Genetics (ASHG) on October 20, 2014, in San Diego, CA, USA. The audio of the original address can be found at the ASHG website.

²Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

*Correspondence: goncalo@umich.edu

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

<http://dx.doi.org/10.1016/j.ajhg.2015.02.006>. ©2015 by The American Society of Human Genetics. All rights reserved.

our best computational approaches to the test and demand that we continually improve or replace them.

In part because of this early work on computational methods for SNP analysis, I had the privilege of participating in many early efforts characterizing variation at a genomic scale, starting with the International HapMap Consortium. There, I encountered several greats of modern human genetics, including Mark Daly (again) and also Aravinda Chakravarti, Peter Donnelly, and David Altshuler. Aided by a second set of remarkable characters, I soon proceeded to apply methods to characterize genomic variation at scale to study a variety of complex human traits and diseases. Among the remarkable colleagues with whom I worked at the time, a few stand out: Michael Boehnke, Francis Collins, Anand Swaroop, David Schlessinger, Serena Sanna, Joel Hirschhorn, James Elder, Sekar Kathiresan, and Cristen Willer.

For me, one of the remarkable lessons from that time was to learn that by combining data and results across studies, we could rapidly accelerate the pace of discovery. This lesson was helped by the discovery of genotype imputation^{5,6} and by new tools for meta-analysis of genetic association studies.⁷ Genotype imputation uses stretches of chromosome shared between apparently unrelated individuals to reconstruct genotypes at any genetic marker and greatly facilitates comparisons between studies relying on different marker sets. In our first genome-wide association study (GWAS) meta-analysis, aided by imputation, we combined three studies each with little evidence of new loci and discovered 17 loci associated with blood lipid levels.⁸ Through advances in genotyping, imputation, and meta-analysis, these meta-analyses can now include hundreds of thousands of individuals. It is now straightforward to generate very large catalogs of loci associated with complex traits. It is also clear that, even 10 years ago, these catalogs would have been a great achievement and perhaps hard to imagine. Still, most of us are now unsatisfied with simply cataloguing complex-trait loci and believe it is very important to move on to the next phase, which is to translate these catalogs into insights about human disease biology and medicine.

At the same time that interesting things happen in genetics, other interesting things happen in the rest of our lives. I would like to share with you a memory from this time, when we were conducting our early genome-wide analyses of genetic variation and its relationship with many human traits. I visited Mark Daly in Boston, and he invited me to join him for dinner at the end of the day. We walked out of his office, walked through a grocery store near Massachusetts General Hospital, and picked up a truly large cut of salmon. I must have told Mark we'd have trouble finishing such a big fish, but he was undaunted. He drove us home and introduced me to Mary Pat and four charming children. Mark then proceeded to prepare and cook the fish. When dinner was ready, Mark's four wonderful kids helped make the entire salmon disappear quickly.

It was quite an inspiration to see a very successful scientist get home in time to cook dinner and share it with his family.

Studies of human genetic variation have been gradually shifting from arrays to full genome sequencing. One of the first projects to explore the possibilities was the 1000 Genomes Project, which also showed me how different very competitive, very able people could come together to make special things happen. If you know my three project co-leaders—David Altshuler, Gil McVean, and Richard Durbin—you'll know that they are extremely competitive, extremely able scientists. Working together with them made the project a memorable experience for me.

My experience early in the project greatly influenced and, indeed, changed how I think about the analysis of large complex datasets. After assembling all the whole-genome sequence data collected in the first phase of the project,⁹ we set out to identify an optimal strategy for converting the raw sequence data (which consisted of many short sequence fragments with many errors) into high-quality lists of variants and genotypes for the individuals we were studying. I was sure that we would be able to compare several alternative strategies and agree on one superior strategy—ideally, this optimal strategy would be the one developed by my group at the University of Michigan.¹⁰ Instead, we were deadlocked. Each of the analyses carried out by teams at the Broad Institute, Michigan, and the Wellcome Trust Sanger Institute was optimal in some way. Eventually, we resolved the problem not by deciding on which of these strategies was superior but by combining their respective solutions into a consensus or ensemble prediction. Remarkably, this consensus solution was better in all respects than the solutions each of our teams had spent time crafting and optimizing. The advantages of ensemble predictors are common to many areas of computation, biology, and society—but I had not come across them in such a direct way before.

Most recently, we have applied the same methods and ideas that were used for sequencing the 1000 Genomes Project participants to other interesting studies, especially ones where rich health and medical information is available. One of the most interesting of these has been our ongoing study of the genetics of isolated Sardinian populations¹¹ in collaboration with a very able team of Sardinian geneticists led by Francesco Cucca. To reduce cost, we sequenced each genome at low coverage.¹⁰ Although this sequencing strategy is not well suited to the analysis of a single genome, it has a remarkable property: as more and more individuals are sequenced, we are able to interpret available sequence data for each genome better. In practice, this means that as we sequence more genomes, we are able to examine progressively more genetic variants and, at the same time, to more accurately characterize the genotypes at each variant. There is not enough time to describe the results in detail, but given that you will soon hear from William Allan Award winner

Stuart Orkin—who has done remarkable work studying the causes and mechanisms of the thalassemias and sickle cell anemia—I want to highlight one interesting and unpublished observation from our studies of low-density lipoprotein (LDL) cholesterol, a strong risk factor for cardiovascular disease. When we sequenced 2,000 Sardinians and examined the list of loci associated with LDL cholesterol, we observed many of the usual suspects seen in other GWASs (including the loci where *APOB*, *APOE*, *LDLR*, *PCSK9*, and *SORT1* reside). Remarkably, we were also able to show a very strong effect on LDL cholesterol levels for the p.Gln39* variant in *HBB*—a variant that Dr. Orkin will know well as the cause of β -thalassemia on the island. Our analysis suggests that p.Gln39* is also one of the biggest drivers of LDL cholesterol levels in the Sardinian population. Association between the *HBB* locus and LDL cholesterol is not typically seen in GWASs because the loss-of-function alleles, which are typically rare and population specific, cannot be well captured by standard arrays.

The secret of success is to have many able collaborators, mentees, and students. I certainly wouldn't be here without all their criticism, encouragement, ideas, and hard work. I am truly grateful to each of them.

It is probably premature, but also tempting, to reminisce on the lessons I have learned. I certainly have many more to learn. I share the lessons I have learned in the hopes that they might be helpful to audience members who are starting their careers in our wonderful field of human genetics.

1. *One person and a good idea can make a difference.* As human genetic studies increase in size and complexity, it is sometimes easy to get disheartened and think that one can't compete with much larger teams, studies, and budgets. Instead, we should remember that novel algorithmic insights can often improve computational performance by several orders of magnitude and thus enable even small teams to beat much larger and better resourced teams. Even more important, approaching a problem from a new angle can lead to completely new, previously overlooked answers and insights.
2. *The best students, postdocs, and collaborators generally know something that you don't.* This insight is a bit disconcerting at first. When evaluating a prospective student, postdoc, or collaborator (which you should do carefully!), it is not sufficient to rank them according to what you know. Instead, it is perhaps most important to explore how much of what they know is something you don't and how well you might complement each other. Many accomplishments depend on just the right partnership and would not be possible otherwise.
3. *Take the time to be amazed.* Drop everything, and explore a new idea. Some of my colleagues consider this a personality defect (we complement each

other!), but I believe that on occasion it is important to drop the day-to-day things that one should be doing and instead take the time to explore something new, interesting, and fun. In many ways, that is where the fun of science, of genetics, and of discovery comes from.

4. *Keep learning.* There are so many ideas out there. Perhaps the truth in this statement is most obvious when attending The Society's annual meetings. It is really a privilege to come to this meeting and to see so many developments in human genetics. I really find it surprising that, with all the great science being presented at this year's meeting, I would have been lucky enough to be selected for the 2014 Curt Stern Award.
5. *The most valuable tools and algorithms are often extremely simple.* It's sometimes tempting to think that great discoveries are complicated. I have learned over time that sometimes our most valuable discoveries are very simple. They just need to be applied and tried at the right time.

Looking ahead, it seems clear that we'll sequence thousands of genomes and that we'll discover better computational methods and strategies for managing, analyzing, and interpreting the resulting data. Still, some of the biggest challenges in our field are not about the scale of experiments or computational efficiency. In my view, there remain big challenges in how we enable different scientists to interact with data, understand it, choose powerful study designs, and answer important questions. As the field of human genetics and its applications get more complex and more diverse, I feel that we need to spend quite a bit of time facilitating communication of our findings to experts in other disciplines and across the different subfields in our society.

To conclude, I wanted to take a moment to thank my family. The past 10 years have also been a very interesting time for them. A little over 10 years ago, I married Cristen Willer, whom I met during my doctoral studies at Oxford. Cristen has been wonderfully encouraging, supportive, and patient. Those of you who know her will know that she is also an accomplished geneticist and member of The Society. We now have four kids, and another one is due in March. They are a little smaller than Mark's, so they're probably not as much of a handful, but they really make my days at the office worthwhile. I encourage anybody who is thinking about a career in genetics not to be concerned that it gets in the way of family life. Don't trade. Do both.

Thank you all very much.

Acknowledgments

I am deeply grateful to the members of The Society for this prestigious award. Major support for my research has come from the University of Michigan, the NIH, the Pew Charitable Trusts, GlaxoSmithKline, and the Foundation for the NIH. I would also

like to thank my assistants, Laura Baker and Irene Felicetti, for their valuable help over the years and with this presentation.

References

1. Abecasis, G.R., Cardon, L.R., and Cookson, W.O. (2000). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* *66*, 279–292.
2. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* *30*, 97–101.
3. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* *58*, 1347–1363.
4. Keavney, B., McKenzie, C.A., Connell, J.M., Julier, C., Ratcliffe, P.J., Sobel, E., Lathrop, M., and Farrall, M. (1998). Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum. Mol. Genet.* *7*, 1745–1751.
5. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* *34*, 816–834.
6. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* *39*, 906–913.
7. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* *26*, 2190–2191.
8. Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* *40*, 161–169.
9. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* *467*, 1061–1073.
10. Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* *21*, 940–951.
11. Pilia, G., Chen, W.M., Scuteri, A., Orrù, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* *2*, e132.