Genes and Genomes: Impact on Medicine and Society

Genes, Genomes, and Medicine
October 16, 2003

Roy M. Anderson, Ph.D., Imperial College of Science, Technology and Medicine, London, UK
The Evolution of Pathogen Genomes

Introduction by Andrew R. Marks

Andrew R. Marks: So, now for the last speaker in this session. Roy Anderson is professor of infectious disease epidemiology and head of the Department of Infectious Disease Epidemiology at Imperial College, Faculty of Medicine, University of London. His seminal studies have examined the transmission, evolution, and control of infectious disease agents in human communities, animal livestock, and natural ecosystems. His work has been the basis for designing childhood immunization programs, strategies for controlling the spread of AIDS, and control of Mad Cow Disease in Great Britain. Today he will discuss how genetics in pathogens and their hosts, which are us, determine the mortality of the major infectious diseases, including malaria, AIDS, TB, and the emerging disease SARS. Welcome to Columbia.

Origins of Population Genetics

Roy M. Anderson: Well thank you very much. It is a real pleasure to be here, and a very warm happy birthday on your 250th anniversary. Coming from London University, I very much understand the problems of being a university in a very, very major city with all its disadvantages in one sense, but the many, many advantages of the sort of liveliness and the cultural environment that New York presents.

I was very surprised to walk into your university this morning and see many flags with the royal crown on, and clearly a little bit of thought, thinking about history, explains that. But I was also drawn to think that perhaps this had something to do with "Genes and Genomes" because clearly royalty throughout Europe are quite an interesting case for study in terms of inbreeding coefficients.
My association with Columbia is very, very recent. I've had the privilege to serve on the advisory board for your new Earth Institute, under the directorship of Jeff Sachs.

And what I'd like to talk to you today about actually, I was asked to address the problems of population genetics and genetic epidemiology, I'm going to very briefly refer to that, but I want to move more to the genomes of pathogens and how one studied genetic diversity, and in particular I'm going to talk about two problems which are new and evolving problems. One is the description of genetic diversity and understanding its linkage with biological characteristics, and the second one is a more technical area, it's the beginnings of understanding how we can simulate genome evolution for simple organisms within computers.

If we think about the origins of population genetics, really these were laid down—the techniques and much of the statistical methodology and indeed mathematical methodology—was laid down in the first half of the twentieth century, particularly by R. A. Fisher at Cambridge, but also very importantly by C. C. Lee here in the United States. The term **genetic epidemiology** first emerged somewhere between ’54 and ’70 —there is some debate about who was the person who first coined it—but the second half of the twentieth century really saw the very, very rapid growth in human genetics, and much of the early growth was related to observational studies, family studies, and the development of DNA markers.

We haven't—it's not the sort of place to talk about the statistical methodologies, but there are some very, very important developments in this area. We've heard repeatedly today about you can let the computer do this, that and the other. Well that's not strictly true. Some people have to develop rather clever algorithms containing a set of biological assumptions to instruct the computer to do x, y, and z, and there's quite a lot of theoretical developments in that area when you are dealing with huge linear sequences.

Ah, this is an interesting concept—I hope this is not persistent. This is the translation from a PC to a Mac. There is a picture there which was a copy of the front of the *Journal of Science*, and it was in 2001 and it was essentially a cartoon about different animals talking to each other at a cocktail party, saying, "Well, have you had your genome sequenced?" What I want to mention today is that we've clearly been focused on the human genome, but it's very, very important to remember that the genomes of most of the important human pathogens have now been sequenced. Sydney Brenner mentioned phage. If you think about very shortly after that, a series of very small viruses were sequenced, one of the first was foot-and-mouth virus, and if you think about the SARS epidemic recently, something like two weeks after the identification of the etiological agent, the four-genome sequence is out for this new virus. The only organisms where the genomes are not available at present—but they will be in the coming two years—are the larger pathogens such as the protozoa, the *falciparum* malaria, and the helminths, the worms, which have very, very big...
genomes. But I'm sure that these will tumble out in the coming years. And just as for humans, these genomes, we have complete sequence, dense marker maps of the genome, and increasingly sophisticated methods of analysis.

**Approaches to Genetic Epidemiology**

Now what is genetic epidemiology? I mean in some senses it's very obvious; it's the study of the role of inherited factors in disease etiology, and for infectious diseases, for example, this involves two genomes: that of the host laid upon that of the pathogen, and very complex and subtle interactions happen there that we understand very, very little about at present. In the early years this was largely concerned with family-based studies, or monogenic disorders, but if you think about recent years you think about increasingly a focus on complex multifactorial diseases; arthritis is one, and there are many others. I'm not quite as optimistic as others that we will be able to make rapid progress in this area. If one has six or seven or more genes involved, then genome scanning and linkage studies are going to be exceedingly difficult. In the United Kingdom recently the Medical Research Council approved the funding of a project called BioBank, which is obtaining DNA from half a million people, adults, and then following in detail their clinical and medical histories over time. And the debate really centers on, is that sample size big enough if you've got a very complex disease with many genes involved? And that's far from clear, very far from clear.

If you're interested in infectious diseases, you look with fascination and some wonderment at the increasing number of noninfectious-disease problems that are beginning to be linked with an infectious event or a persistent infection. We have the extraordinary example of gastric ulcers and *Helicobacter*, but more recently some really interesting other examples, one of which is atopic, in particular asthma and the occurrence of a past exposure, which leaving serological markers of Hepatitis A virus. Much remains to be unraveled here, but one suspicion is that infectious agents—particularly something like asthma and arthritis, which involves the immune system—infectious agents may have a very, very important role to play.

Now the two strategies to these linkage problems, there is the sort of brute-force approach of genome-wide screens, using markers, etcetera. An alternative strategy to the genome screening is that of candidate gene studies based on a knowledge of a gene product involved in a particular disease. And it's important to remember in history that many more genes have been discovered by identifying the protease involved in the disease and working backwards to determine the genes that encode. So this in Sydney Brenner's term was sort of going backwards, but it is important to remember that that is where the majority of our information has come from, and it requires a biological understanding of pathogenesis of some detail to lead you to that particular gene target. And clearly in both of these areas the new technologies offer extraordinary scope for rapid progress, as we've seen so clearly today.
If we think about genetic epidemiology and its coming out of population genetics, familial aggregation of disease, the backbone of this approach, but I should stress that increasingly huge population-based studies—where you have differences in the instances of disease, controlling for other environmental factors—are playing an increasingly important role. And in fact, Cori in the last talk talked about this example in New Zealand. Iceland, many of you will be familiar with, a population of a quarter of a million, and a company DeCode, has been involved in studying and extracting genetic material from consenting children and adults, for a nationwide screening where you can link genetic background, hopefully, with the occurrence of certain disease types. It’s helped enormously in Iceland by the existence of a complete genealogy from the founders of that society. However, I was a postdoc in Oxford at a time when there was a series of studies called the Otmoor studies in villages around Oxford on human genetics, in part influenced by Walter Bodner. And what was so remarkable is that genealogies are often false; in other words, something like 10–20 percent of births do not come from the parents in that family. So genealogies may or may not be true, but it is a very powerful tool within Iceland.

Now many population geneticists—Richard Peto is a good example, and our own group also—have been looking with longing eyes towards China where you have a population of over 1 billion, and at the moment a very adherent population to directives, as we saw so clearly in the SARS epidemic, and clearly many very, very important epidemiological genetic studies are going to be carried out in China and are already underway over the coming decades.

Technology is a very important aspect in population genetics. Just to pick on one technology—whole genome microarrays. Increasingly in the pathogen area, we have these designed for very specific problems. One illustration might be drug resistance in tuberculosis, mycobacterium tuberculosis, which facilitates in the research laboratory extremely rapid screening and understanding of the genotype or phenotype of a particular organism. It is interesting to note, however, that these technologies have not made rapid entry into the public-health arena, so if you look across CDC in the United States and equivalent bodies in Europe, you will not find the rapid take-up of these technologies, as there should be. They offer extraordinary scope, even given problems of accuracy.

**Rapidly Evolving Infectious Diseases**

Now I'm now going to turn for the majority of this talk to the two problems I want to introduce to you, and that's genetic diversity to start with, and I'm going to focus in particular on bacteria here which present many, many interesting problems.

First in passing, I make the observation that the list of disease, infectious disease, and a human genetic influence—in other words, a gene or a product for
a gene—is growing exceedingly rapidly. But this is hardly surprising for those of you who've ever done experiments with pathogens with mice in the laboratory; everybody knows that you use highly inbred strains as choice, and ideally you'd like cloned mice, because as soon as you move a pathogen from one inbred strain to another, you have totally different pathogenesis. Rolf Zinkernagel's system is a very good example of that. So we know that genetics matter enormously in determining the typical pattern of infection. So this list is huge and the origins started way back, probably with the blood-group disorders related to malarial infection.

If we look at infectious disease today, there's been a resurgence in Western medical schools in students interested in this discipline. It is worth reminding ourselves that if we look at the world in total that the leading cause of premature mortality remains infection, sadly. In our own developed and industrialized societies, of course, that is not true today. But in the majority of the world's population these are the agents of the most important source of mortality.

If you think about our world today, there are three factors that tell us that evolution is going to be speeded up in terms of the emergence of new infectious agents. The first is a very obvious thing. Our internationally mixing world, our increasingly global mixing world, means that agents pass quickly from high-density populations to new ones. A rough order of how our mixing has changed, we probably had a thousandfold increase in effective epidemiological contact rates between different countries between 1980 into 2003. Now that's an epidemiological observation, but remember that's also an evolutionary-biology observation, because transmission from one host to the other gives the organism an opportunity for evolution. So concomitant with that transmission increase, there is the huge increase in the evolutionary opportunities. If we look within countries, again we've had extraordinary changes in the pattern of mixing within human societies.

We've also had other changes which are relevant. There is obviously increased population size, there are improved technologies for detecting new pathogens, but there are these factors, pathogens like big, highly dense populations, and if we look at the growth in what the United Nations called megacities—that's cities over 10 million people—we see that there is going to be considerable growth, and there has been considerable growth, particularly in Asia. It's also of interest to note that the fastest growth in air traffic between different cities has also occurred in Asia over the last ten years. So these Asian—the origins of new influenza strains or the coronavirus is not entirely by accident, there are epidemiological reasons why that might happen.

The SARS Coronavirus

So remember we've got all these genomes, and we have the capability for studying genetic diversity in some detail. And remember also in certain fields
there is repetitive whole-genome sequencing; HIV is one, but there are a number of others, where in other words, sequential whole-genome sequencing—
influenza is another good example—enables you to discover or to investigate
 genetic diversity not just in single genes but across a whole range of them.

The coronavirus, this new one that we suffered this year, is an interesting example—this is another interesting case when PowerPoint transferred from one system to another—underneath that red bar is the new coronavirus, and this was Malik Peiris’ first pass based on the whole sequence in looking at the relatedness of this new virus to other viruses. And you can see the red bar falls between the avian, bovine, and murine. This was actually wrong in hindsight, there’s another beast called the civet cat in which the virus, coronavirus, exists in, that is very, very closely related to this new human form. But this question is still very open, where this virus came from, and a lot more viruses need to be isolated from other mammalian species, particularly in these animal markets such as exist in Qangdong Province. So evolution is constantly happening, usually with combination events where two viruses enter the same cell, and there’s a jumbling of the genomes.

This epidemic of this new pathogen had the most extraordinary effects in terms of loss to the economy, particularly within Asian countries, WHO took the brave step of issuing travel directives which limited greatly, not as a sort of order but in terms of voluntary limitation of people traveling to certain parts of the world. The sort of postmortem of this event tells us that we were extremely lucky. First, if these major introductions had occurred in North America and Great Britain or Europe, I'm not sure we would have controlled this particular virus, for the simple reason that the draconian measures that were put in place by China, by Singapore, by Hong Kong, I don't think we could've put in place in Western societies where politicians are concerned with reelection.

There's another very important biological reason why we were lucky, and this is a more genetic one. This virus—ah, interesting, we're getting variants—that picture . . . there was a picture of the SARS coronavirus [inaudible] micrograph—never mind, the important information is not in the picture. This virus had trouble replicating in the human body, and a very, very unusual situation pertained where peak viral load, in other words, peak infectiousness, occurred significantly after the onset of clinical symptoms, such as acute temperature. In other words, if you were to implement isolation and quarantine, it would work with this virus. If you look at viremia in clinical epidemiology of influenza A and the onset of clinical symptoms, peak infectiousness coincides with the onset of clinical symptoms. So in other words you've got a chunk of infectiousness before the patient is aware they're sick. With this beast, very fortunately, because of its poor replicative ability in a human host, we were lucky; it was not infectious, so isolation and quarantine worked exceedingly well.
Now this continued evolution occurred in China, much of it not understood, because not enough sampling of coronaviruses has taken place, and we await with some trepidation what might happen this autumn to see whether this virus has indeed increased its evolutionary transmissibility.

The other aspect of this virus—this is one of the most pathogenic I've ever seen, the case-fatality rate in the over-60s was over 60 percent. Case-fatality rate in the under-24s was essentially zero. So that's a mortality rate which is very much linked to age. And remember influenza, which we regard as a highly pathogenic organism, has a much, much lower case-fatality rate than this beast. So we were right to be duly concerned internationally.

**Genetic Diversity in Pathogens**

Now how do we understand and sort of interpret and study genetic diversity in these pathogens? There are a variety of approaches, and this perhaps relates a little bit to Eric Lander's comments about the information that is available. Multilocus sequence typing is one approach used for bacteria, where in an international effort via Web-based information entry, you sequence certain gene fragments on some housekeeping genes, not the ones expressing surface antigens, and you enter these sequences into a common international database with information on where that isolate came from, and if possible clinical information on the pathogenesis induced in that patient. So what you're trying to do is to build up a huge databank on gene sequence and the clinical pattern of disease created by that isolate.

Now this database approach has now been instigated for a wide range of very important human bacteria, and some of them are listed here, ranging from sexually transmitted agents through to agents that cause pneumonia, etcetera. Now the problems with these databases, and some of you would probably pick this out instantly, is that bacteria are dreadful organisms to study at the population-genetic level because they have very, as it were, promiscuous genomes. They're constantly recombining; they constantly suffer the problem of mobile genetic elements, where huge chunks of genetic information is moving around. Your gut is a very good illustration; we've only typed probably 10 percent of the bacterial species in the human gut; there are many, many tens of thousands of species present. There's enormous movement of genetic material between different species. Recombination was always regarded as rare in viruses. It's very interesting, now that we've got whole-genome sequencing for viruses what is coming out time and time again is recombination is very, very frequent in viruses as well. HIV is a very good example; recombination frequency is exceedingly high.

Now the trouble is when you have these international databases is that trees are extremely difficult to interpret, and this is one called *Streptococcus pneumoniae*, the pneumonia-causing bacteria, or one of them, and if you look on your left-
hand side, you'll see the phylogenetic tree of relatedness between all these strains, and they're exceedingly difficult to interpret.

Now the question is, can we get a reliable intraspecies phylogeny for a bacterial species, given the problems of recombination and mobile genetic elements, or can we only hope to look at the most recent clonal groups and to try and associate those, the genetic diversity there, with occurrence or nonoccurrence of serious disease?

And this is the work of Brian Spratt in my own institute and cohorts of postdocs with Brian, and they have very much been addressing this question recently, and have come—for most important bacterial pathogens of humans—come to the conclusion that these phylogenetic trees—beyond the end point on the right where you look at clonal complexes—are almost meaningless. Now the reason they've come to that conclusion, as I'm going to show you the techniques later on, is they're beginning to simulate the evolution of genomes in the computer with recombination, and then reconstructing the phylogenetic trees, when in the computer you know exactly what the evolutionary events were. And these evolutionary phylogenetic trees of relatedness are for most bacteria where recombination is frequent almost meaningless past the most recent events. And that's the depressing conclusion.

A New Approach to Genetic Diversity

Now what can you do with the most recent events? Can we have new paradigms for interpreting and looking at genetic diversity? And this is one approach, and I only mentioned one. There are a number of others evolving at the moment; it has a rather trendy title called eBURST, and Ed File has developed this in London. It identifies all the clonal complexes within the multilocus-sequence database. You can define what a clonal complex is; it might be a group of isolates that have at least six to seven alleles in common with any other isolate in the complex, and then you need a simple biological model of how this complex is evolving, and then you need a way of robustly estimating relatedness, and then reflecting the tree.

So you could divide the database into non-overlapping groups, use a defined similarity, and then attempt by various statistical methods to predict the center. And here's some illustrations. Remember, with most bacteria you've got two chromosomes—sorry, you've got double strands, you've got double-locus variants and single-locus variants, and here you can see an attempt from [inaudible] to link in distance terms the relatedness between different isolates.

If you move from a simple example to a more complex one—I'm sorry this hasn't translated either. Let me try another one. Ah, here's one that has. This is Streptococcus pneumoniae from a very large international database where you're searching groups of isolates that are clonally expanding, where within that group
of clonal expansion you've got limited number of genetic changes, you've got clinical data on the seriousness of disease in particular patients. And then by this method—and I'm not showing the linkage of this with the clinical data at the moment, that remains to be done—you can begin to understand more detail about the recent evolution of these organisms.

So it identifies clonal complexes; it provides measures of statistical support for what was the founder of that complex; it's a very conservative approach; and it doesn't attempt to reconstruct what is very difficult to reconstruct, the recombination events in the past.

**Evolution of Influenza A**

Now lastly, to end I want to turn to how one can use modern computational techniques with large computers to simulate the actual evolution of the genome. And I'm going to choose as an example influenza A, for very good reasons—it's a simple virus—and I'm going to look at how one might interpret its evolution.

We have a lot of information epidemiologically on influenza A and B, and its close variants. We have extraordinary information in some systems. The United States doesn't have these. It always surprises me that the CDC in computational terms is a bit backward—they need a heavy dose of quantitative methodologies—it's very surprising given the advances that have been made in the molecular-genetics areas. So here for France, for example, showing my lack of bias towards our European neighbors, is this wonderful spatial evolution of influenza A epidemics, which could be tracked day-by-day, week-by-week, by sentinel survey settings.

Now what we're interested in is evolution. We're interested in three problems. One is evolution—what are the determinants of the pattern of it. The second one is how likely you are to evolve drug resistance—we have two new drugs for influenza A which will be used in the next pandemic of a seriously pathogenic strain, and are used in old people's homes for the vulnerable and in hospital-care settings. And then thirdly the design of effective vaccination programs. I'm only going to talk about one of these problems today, and that's the first one. What are the ecological and immunological determinants of influenza A evolution?

I should remind you that if you take something like measles prior to wide-scale immunization, we always thought that was a highly transmittable infection, average age of infection about 5 in North America and Britain. But if you look at the annual incidence per head of population prior to immunization, it was only about 1.5 percent. By contrast, if you take influenza A, this attack rate is about five times bigger, at least. So influenza A is one of the most transmissible agents that we know in detail.
Interpreting Influenza A Ecology

Now the first study of significance in interpreting ecology was done some years ago in the very early stages of the development of immunological tools, and it was a very clever study by Stuart-Harris. Essentially what he did is he isolated virus from patients and took serum from patients. You type the virus and put it into its various strains, and then you looked how the serum in patients who had different viruses cross-reacted. And here you can see a very important biological problem, you can see that there's not a great deal of cross-reactivity outside of the specific strain. And that began to lead to the notion that once you've had a strain you're probably immune for life and the fact you get influenza A is because the strain has changed.

Now today, of course, with genetic information we know a great deal more. And if I just focus for the purposes of this description on the surface glycoprotein, haemagglutinin; it contains about 329 codons, and these are the ones I want to simulate the evolution in. And it's very interesting that 35 percent of replacements occur at a very small fraction of the genome. In other words, that's clear evidence of very heavy selection. We can describe these replacements and selection in very formal mathematical terms, and we can use large computers to simulate that evolution patient-by-patient in a large population.

Here's what happens in reality. If you think about influenza phylogenetic trees, they're like Canadian redwood, they're a sort of straight trunk with little branches coming off. And that's in huge contrast to most other viruses; the most extreme would be HIV. HIV is like a flat-topped acacia with a trunk, and then evolution has gone boomph, like that. So this one is the easiest problem to address in the first instance.

Ah, another picture gone, sorry. In that box up there was the acacia tree of HIV-1, so it was like a flat-topped acacia, and these are the influenza A trees, and remember there's two processes of importance—drift, slow evolutionary change—and then we have antigenic shift, which is the recombination event for an avian strain that causes a dramatic change in the virus, or significant change in the virus.

Now what surprises me at first sight here is given the high transmissibility and high mutation rate, why are these trees so linear? So the scientific question relates to this, this is looking within influenza A. We've got A subtype H3, subtype H1, and then through to B. You can see A subtype H3 is very linear; A subtype H1 is a little bit more diverse; and then B there's been a major bifurcation in the past. And we've got two distinct serotypes at present. These bifurcation events, of course, are clearly very important in vaccine development. So we've got a slender trunk for H3, H1 is more similar, and influenza B reflects this branching.
If we look at any one year, we find the interesting phenomenon that you have many strains present at one time, which is a slight problem for immunization effectively, but you also, if you do a time-series analysis of this data, you find that there are strong correlations here. One being very prevalent dictates that another one will be less prevalent, and that may fluctuate from year to year. And we know something about the rate of evolution for the different major types.

**Computer-Simulated Genome Evolution**

Now to move to what shapes the topology of these phylogenetic trees of genetic diversity. Now the only way I can really think of doing this is that you can't do real-time experiments with humans. You could, of course, with an animal model, but again you'd like large populations of animals like some of the very early studies done in the UK by a variety of people on populations of mice who, in the basement of the London School of Hygiene and Tropical Medicine, introduced ectromelacia virus and let it spread for months on end in this free-running population. But in those days they didn't have the genetic tools to study the evolution of the virus.

So why not use modern computational power to actually explore the evolution of the genome? So mathematical models can play a very important role here, and these have to be structured with some degree of sophistication; they've got to include age and spatial structure—and I'll come to spatial structure as well because that's very important—and then the evolution of the genome.

So we have a very trivial individual framework, susceptible person infected and infectious, recovered, immune to all strains, transient nonspecific immunity—and I'll come to that in a second—and then immune for life to a specific strain so you can circle back and be reinfected with another strain which is significantly distinct.

Now when you start to simulate these evolutionary trees within an artificial genome, what you do is you have a patient who's infected, you're simulating the mutational changes across a set of codons, and then you're replicating that in millions of people, and then you're also simulating transmission within that group. You cannot do population genetics here without including transmission. Much of population genetics assumes arbitrary frequency-dependent functions, totally arbitrary. You must have the real frequency-dependent function which is the nature of population density and the transmission of the virus.

And here's four examples: one, this is run for fifty years of evolution. B has no short-lived immunity post-recovery from influenza; C has no short-lived immunity, and has a very low transmission rate; D has over-restricted diversity achieved with very specific parameter values; and then E has—reducing the intensity of cross-immunity across the different strains. So slowly you can build a picture of how each biological factor influences the shape of this tree.
And here’s one very trivial example, cross-immunity. Using the information from Stuart-Harris, C is an inverse measure of cross-immunity, if it's 1 it's perfect, if it's 0 there’s no cross-immunity, and you can see how changing the parameter C changes the shape of these evolutionary trees. And these calculations were done by Neil Ferguson and the group, who's a theoretical physicist by background who when interviewed for the post when asked why he wanted to come to biology from a very good theoretical physics, he said, "Biology is much more interesting because there are so many more nonlinear processes." It wasn't the furry and the featherers in the gene sequences, it was the fact that the mathematical nonlinearities excited him, and there are some very significant ones in these problems.

Now I work with trepidation to see whether this functions. The reason space is important—as illustrated in this slide—this is a study of the spatial spread with these genome evolutions going on. And each color is a different strain. And these are two loosely connected patches in the country, or between countries, and so evolution depends on transmission between the patches.

You might ask how on earth can theoreticians describe mixing of human societies. Well, very interesting, one facet of our modern society, namely mobile phones—and you're probably not aware of this—provides an extraordinarily digital record of what you do. This is not so true in North America, but it's very true in Europe with a high density of masts. Every time if your phone is switched on you move between a different mast, a digital imprint is left on the computer there that your phone is logged onto that mast. There are teraflops of data per year, so you need bloody big analysis setups, but you can then map for location X the probability distribution that an individual moves a distance Y. And that spatial probability distribution underpins these spatial transmission events.

And what you see is that frequency-dependent selection operates very much in a spatial domain. So a strain sweeps through a population, you've got heavy frequency-dependent selection with the evolution of a new variant. Then it moves onto another patch. The thing with influenza is it goes through the world in about two years because of its very, very high transmissibility.

Now one can map these epidemiological and genetic things in three dimensions, and here’s an illustration. On the axis going into the graph you see substitutions, which is the genetic information, at a defined set of sites. On the horizontal axis is time, and on the vertical axis is the epidemiological variant, which is the frequency that that strain is in the population. So now we're beginning to get the computational tools to meld the epidemiology with the genetics, where you can study the evolution through time and you can study the epidemiological, the prevalence and instance of infection. These are very, very early beginnings with a very, very simple virus, influenza A, but I see no reason why one shouldn’t
extend this to much more sophisticated systems. We just need rather large computers.

Human Genetics and Pathogen Diversity

So in conclusion, evolution in pathogens, of course, occurs in a very, very fast time scale. And for many of the human ones, the important point is that evolution is speeding up, not slowing down, because of our mixing, because of our population size, and so forth. So they're very good models for understanding evolution in animal populations. We have increasingly sophisticated methods for study, but we very rarely bring all these together in one laboratory or team. And that's so apparent in our centers for disease control or public health authorities.

The need to link the human genetic background with the pathogen diversity I haven't touched today, but it's so obvious. They're two gloves that fit together—pathogenesis against one genetic background for one virus may be totally different with a different virus or against a different genetic background. We've hardly begun to think about how to address those problems. We can do it in the laboratory context, but not in real human populations as yet. But, given the fascination in human genetics and the accumulation of information like the BioBank project, given the ability to isolate in sequence pathogens, the real issue here is sampling, not the lack of technology. We are very, very bad at taking representative samples of pathogens at one point in time internationally.

And lastly I do mention—which I'm sure will come up tomorrow in much more detail—the very important ethical issues for epidemiology that must be resolved. If some current legislation and discussion in the European parliament were to go through, it would almost inhibit a huge chunk of epidemiological research. It's the question of informed consent and the holding of biological samples such as blood, serum, etcetera. So there are very, very important problems there indeed that scientists must play a role in discussing with the public.

Thank you very much for your attention.