SUPPLEMENT ARTICLE

Genomic Approaches to the Study of HIV-1 Acquisition

Amalio Telenti[1] and Paul McLaren[2]

[1]Institute of Microbiology, University Hospital Center, University of Lausanne, Lausanne, Switzerland; [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Host genome studies are increasingly available for the study of infectious disease susceptibility. Current technologies include large-scale genotyping, genome-wide screens such as transcriptome and silencing (silencing RNA) studies, and increasingly, the possibility to sequence complete genomes. These approaches are of interest for the study of individuals who remain uninfected despite documented exposure to human immunodeficiency virus type 1. The main limitation remains the ascertainment of exposure and establishing large cohorts of informative individuals. The pattern of enrichment for *CCR5* Δ*32* homozygosis should serve as the standard for assessing the extent to which a given cohort (of white subjects) includes a large proportion of exposed uninfected individuals.

The discussion on the protective factors that allow a small number of individuals to avoid infection despite repeated exposure is increasingly centered on (1) the nature of the genetic factors that may underlie this trait, (2) the technical approaches used to identify them, and (3) the study population and its defining characteristics.

In the present commentary, we discuss these aspects under the common theme of integrating large-scale data sets that are mainly but not solely obtained from genomic studies. The emphasis here is on understanding the importance of study design and defining the phenotype of human immunodeficiency virus (HIV) resistance. These are essential aspects for the success of these studies; in fact, cohort issues, not technological limits, are increasingly recognized as limiting progress in the field.

GENETIC FACTORS UNDERLYING RESISTANCE TO HIV INFECTION

The identification of the highly protective *CCR5* Δ*32* allele demonstrated the importance of genetic factors in determining susceptibility to HIV in white subjects. However, no similar factor has been identified in other populations despite extensive evidence for similar traits of resistance to HIV. The research community has addressed this specific question through a candidate gene approach (data compiled at the HIV-Pharmacogenomics Web site [1]). The study populations have generally included exposed uninfected individuals, mother-child pairs, or infected individuals and the uninfected general population, to compare allelic frequencies. The proposed variants thus studied have been supported by unequal evidence. Variants or haplotypes of *CCL5*, *CCL2-CCL7-CCL11*, *CCL3*, and *DEFB1* have been associated with differential susceptibility to infection or transmission in >1 study. There is controversy regarding the role of variants of *KIR*, *CXCR1*, *DARC*, *CCL3L1*, *CXCL12*, *CD209*, *CLEC4M*, *MBL2*, and *ABCB1* in susceptibility to infection. There is insufficient evidence or lack of confirmation concerning a role for variants of *IL10*, *IL18*, *IRF1*, *IL4R*, *PTPRC*, and *APOBEC3B*. It is clear that identifying genetic factors underlying resistance requires a large well-defined study population with a clear phenotype (see below), and the unbiased identification of candidate regions using novel tech-

Potential conflicts of interest: none reported.
Financial support: Swiss National Science Foundation (A.T.).
Supplement sponsorship: This article is part of a supplement entitled "Natural Immunity to HIV-1 Infection," sponsored by the Bill and Melinda Gates Foundation and the University of Manitoba.
Reprints or correspondence: Dr Amalio Telenti, Institute of Microbiology, CHUV, 1011 Lausanne, Switzerland (amalio.telenti@chuv.ch).

The Journal of Infectious Diseases 2010; 202(S3):S382–S386

0022-1899/2010/202S3-0014$15.00
DOI: 10.1086/655969

nologies that provide information on likely mechanisms of protection.

Genome-wide association studies (GWASs) constitute the first level of assessment of the impact of common genetic variants in complex traits, such as HIV resistance. Common variation is defined as a single-nucleotide polymorphism (SNP) found in ⩾5% of individuals in a given population. GWASs have been used to identify variants modulating viral load or disease progression in individuals who are already HIV positive [2–7]. These studies are consistent in underscoring the importance of the *MHC* region at the genome scale and of the *CCR5-CCR2* locus. Overall, the common variants identified to date explain <15% of the observed variance in viral load or progression in infected individuals. The unexplained variation may represent yet undiscovered genetic factors as well as viral strain or environmental factors.

To identify the factors responsible for the remaining genetic variation, necessary steps forward include conducting a meta-analysis of all currently available GWASs, a step that has proved highly successful in other fields, as in understanding the genetic structure of diabetes or human height. Here, meta-analyses of genome-wide SNP data on >100,000 individuals have yielded >100 validated genetic loci influencing those traits [8]. Second, new technology (see below) increasingly allows the interrogation of less common and rare genetic variants (<5% in the population). Again, the model of diabetes genetics illustrates the point by describing a scenario in which common and rare variants may both occur at the same genetic loci yet differ in their power of the association and effect on the trait. Third, the frequent call for candidate gene approaches illustrates the intellectual tension between analyses that are agnostic to the nature of the genes involved and those guided by current biological and functional knowledge. The new element that has animated the debate is the possibility of identifying candidate genes or pathways by applying advanced technologies that interrogate the complete genome—for example, genome-wide RNA silencing and transcriptome screens. Figure 1 illustrates the genomic pipeline that integrates the various approaches.

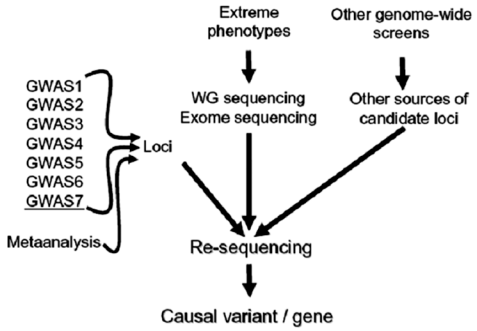

Figure 1. Summary diagram of the genome pipeline. Genome-wide association studies (GWASs) are now processive techniques that have been applied to the analysis of >10,000 individuals infected with human immunodeficiency virus. The individual studies can be pooled through meta-analysis to identify additional genetic variants. Sequencing of the complete human genome or the coding genome (exome) are new steps being applied to a limited number of individuals considered highly informative (eg, elite controllers, rapid progressors, exposed uninfected individuals), with the goal of identifying rare mutations. Other genome-wide screens, such as transcriptome or silencing RNA analyses, may contribute to lists of candidate genes that may be prioritized for sequencing. All approaches will generally lead to in-depth targeted resequencing in more individuals. WG, whole-genome.

TECHNICAL APPROACHES TO IDENTIFYING GENETIC FACTORS

GWAS analysis. Genotyping at the genomic scale generally implies the assessment of 500 to 1 million SNPs per individual. The SNPs assessed by the most widely used commercial platforms have been selected to represent common variation in humans and do not directly interrogate all known variants. More SNPs are needed when the study population is of African ancestry. In addition to HIV, GWASs have been completed for hepatitis C virus [9–12], *Mycobacterium leprae* infection [13], and malaria [14]. In all cases, the approach has shed light on pathogenesis. Of particular relevance to HIV pathogenesis and resistance, current GWASs are mostly limited to populations of European ancestry. Generating such data in other populations, particularly in Africa, has the potential to identify novel loci. However, such studies involve technical challenges, because more SNPs need to be genotyped owing to shorter haplotype lengths.

Whole-genome analysis. Current GWASs have been criticized for their inability to interrogate rare variants, and for the limited amount of genetic variance currently explained. Thus, there has been growing interest in prioritizing the analysis of rare mutations that may carry a significant effect for the individual who carries them and may collectively explain part of the remaining unexplained variance [8]. Investigators are working on the various competing possibilities: lowering the cost for whole-genome sequencing to make its application more feasible in large samples, concentrating resequencing efforts on the complete exome of an individual (the 2% sequence in the genome that is coding) [15], or increasing the density of the coverage through a process known as imputation [16]. The step of imputation assigns additional variants to an individual based on haplotype data from well-characterized external data sets (HapMap phase 3 or the complete genome sequencing of >1000 humans from several populations [17]) that thus do not need to be directly typed. At a different scale, resequencing is increasingly used in the follow-up of loci identified in GWASs; for HIV, this is being done across the *MHC* region.

Transcriptome analysis. Large-scale analysis of expression patterns are targeting both tissue and cell populations from the

in vivo HIV setting, as well as cell populations exposed to in vitro perturbations [18, 19]. Transcriptome analysis is also shedding light on the dynamics of infection in pathogenic and nonpathogenic models of SIV infection [20–22]. These analyses use microarrays covering >20,000 genes and regularly identify one-third to two-thirds of the genes as being expressed in the tissue or cell of interest. Against this level of coverage, the new RNA deep sequencing approaches allow the quantitation of more rare transcripts, splice isoforms, and small RNAs [23]. RNA deep sequencing also provides sequence variation information (eg, SNPs) and is the most powerful approach to the study of samples that are enriched for transcripts of interest, in particular those obtained by chromatin immunoprecipitation. On the other side of the scale, a number of technologies aim at the high-throughput analysis of a limited number of transcripts (eg, <500), representing signatures of a particular state or full pathways [24]. As a rule of thumb, deep sequencing may need >1 μg or 1 million cells; microarray analysis, >100 ng or 100,000 cells; and next-generation digital gene expression technology, 10 ng or 10,000 cells. Economy and speed also decrease with decreases in the number of transcripts targeted. Expression analysis is increasingly directed to the identification of micro-RNAs relevant to antiviral defense. Metabolomics and proteomics are also used in analysis of in vivo and in vitro response to HIV or other relevant viruses [25].

Silencing-RNA screening. Four types of silencing RNA (siRNA) screening tests have been used to identify host factors required for HIV replication in vitro (reviewed in [25] and [26]). The results have shown some of the limitations of the technology for generating reproducible lists of candidate genes for HIV dependency factors. Whereas 1000 genes were identified by the different screens, only 34 were shared by ⩾2 studies. However, detailed meta-analysis and network analysis of the data can point to a number of genes and pathways shared across studies—for example, the nuclear pore machinery, the mediator complex, a number of key kinases, and components of the nuclear factor–κB complex—which should be considered key elements of the cellular machinery supporting viral replication [26]. It is unknown, however, how these genes, or variations thereof, may relate to in vivo susceptibility to infection. Resequencing efforts have been initiated by several groups to answer this question.

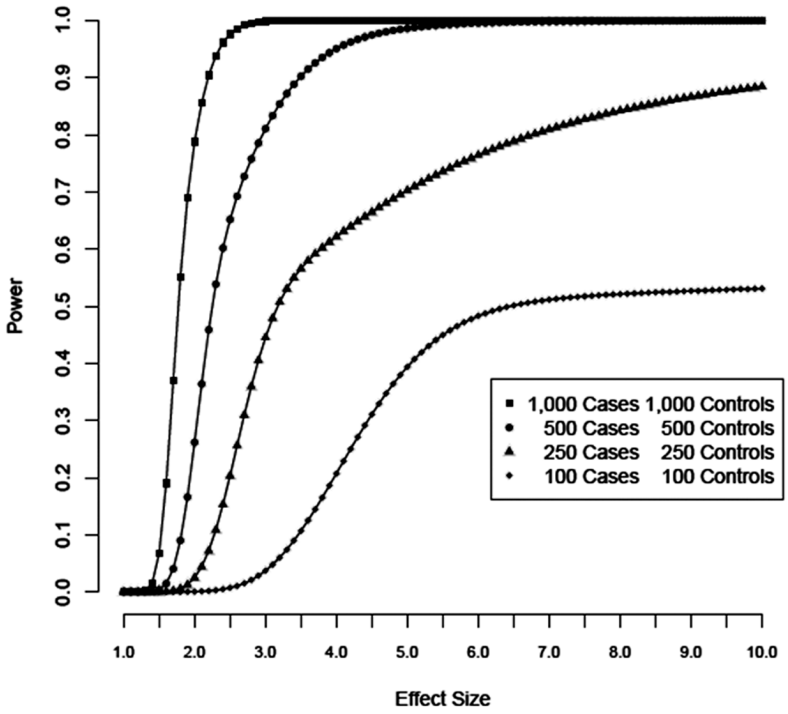

Figure 2. Considerations affecting the power of variant detection over a range of effect sizes in genetic discovery studies in exposed uninfected individuals. The power of genetic discovery is a function of sample size, enrichment of the causal variants in the case population, minor allele frequency in the population, and effect size. Small sample sizes (<500 case subjects) have only modest ability to detect variants with large effects (odds ratio [OR], >5; eg, *HLA-B*5701*–tagging single-nucleotide polymorphisms associated with viral control) at a genome-wide discovery threshold of $P < 5 \times 10^{-8}$. Identification of variants with even greater effects, such as *CCR5* Δ*32* homozygosity in exposed uninfected individuals (OR, >10) would still require >100 case subjects and control subjects. These considerations are of particular importance in light of the modest effects (OR, <2) for validated loci discovered in other complex traits. This power is further reduced by ambiguous phenotype definitions, because the allele frequency differences between case and control subjects would be reduced.

Closing the loop: systems and network biology. Many of the techniques described above may not generate much more than descriptive data that in isolation may fail to identify key genes, pathways, or factors relevant for understanding susceptibility or pathogenesis or for developing vaccines. Systems biology or network biology has been proposed as the approach to integrate high-throughput biological data obtained through clearly designed experiments that include controlled perturbation of the system (eg, exposure of cells to a cytokine), reconstruction of the network and its regulators, silencing of the candidate genes or master regulators, and biological interrogation of the resulting system [27]. In light of the currently available technologies for interrogating the whole genome, such analytical approaches hold the most promise for explaining the biological underpinnings of reduced susceptibility to HIV.

DEFINING CHARACTERISTICS OF A STUDY POPULATION

The success of the various technologies just described for demonstrating the bases of resistance to HIV infection will depend heavily on the study design—in particular, the specificity of the clinical or biological phenotype and the power calculation of the study. Specifically, the relative contributions of cultural, viral, host genetic, and (possibly) environmental cofactors will depend profoundly on the "tightness" of the measurement of the phenotype.

It is important to consider the enrichment of *CCR5 Δ32* homozygosity in hemophiliac individuals as a model for the design of studies in exposed uninfected individuals. The frequency of CCR5 Δ32 homozygosity increases several fold, from 1%–3% in the general white population to frequencies up to 25% in uninfected hemophiliacs, with the highest frequencies in those with severe hemophilia. This is the benchmark of what is to be expected from a factor that confers almost absolute protection in a population with unequivocal repeated exposure to HIV. Therefore, the identification of factors that do not confer absolute protection and the study of a population with poorly defined levels of exposure will be challenging and require significant statistical power. As demonstrated by Figure 2, the power to detect variants with decreasing effect size greatly diminishes as sample sizes are reduced. Thus, even detection of variants with large effect sizes relative to those described by current GWASs (ie, >2) can be achieved only by using large numbers of subjects (>1000 case subjects and >1000 control subjects) or greatly enriching the study population for individuals with the most specific phenotype. In practice, both are advisable to maximize discovery because, even with clear phenotypes, polygenic underpinnings and heterogeneity may reduce the power to detect individual variants. Studies will need to make phenotype definition and cohort assembly a priority; technology will not remediate a poor study design.

Using an alternative approach, researchers who have previously obtained genome-wide data from HIV-1–infected individuals will assess whether infected subjects, unlike those who are exposed but uninfected, are depleted of protective markers. For example, *CCR5 Δ32* homozygosis is reduced to a frequency of ⩽1:1000 among individuals included in HIV GWASs. Here, the comparator is the general population, representing the reference for genetic variant frequencies. A meta-analysis of GWASs and reference populations is currently underway.

CONCLUSIONS

Studies of the infected individual suggest that the genetic basis of human susceptibility to HIV-1 includes common variants, and probably an undefined number of rarer variants. The identification of additional common variants will require the completion of a meta-analysis of GWASs with sufficient power to identify the less common variants and those with less contribution to the phenotype. From this point on, exploring the nature of rarer variants will require a shift in technology to whole-genome or exome sequencing. This is still an unexplored but promising territory in medicine. For the second general path, candidate gene analyses will require feeding from other types of large-scale screens, such as siRNAs and expression analysis, and from more integrative approaches and systems biology [25]. For other disorders (eg, metabolic and autoimmune diseases), the identification of multiple genetic loci and markers allows the first attempts at building genetic scores and proposing interactive models for those diseases.

Investigators studying HIV acquisition cohorts of exposed uninfected individuals should design a genomics pipeline based on the progress and shortfalls of the studies already completed in HIV-infected individuals, paying particular attention to issues of power and phenotype definition. The pattern of enrichment for *CCR5 Δ32* homozygosis over the expected frequency in the reference population should be used as a reliable tool to assess the extent to which a given cohort includes a large proportion of exposed uninfected individuals, although this estimate can be applied only to white participants.

References

1. HIV-Pharmacogenomics Web site. http://www.HIV-Pharmacogenomics.org. Accessed 20 August 2010.
2. Fellay J, Shianna KV, Ge D, et al. A whole-genome association study of major determinants for host control of HIV-1. Science **2007**; 317: 944–947.
3. Dalmasso C, Carpentier W, Meyer L, et al. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. PLoS One **2008**; 3:e3907.
4. Le Clerc S, Limou S, Coulonges C, et al. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). J Infect Dis **2009**; 200:1194–1201.
5. Limou S, Le Clerc S, Coulonges C, et al. Genomewide association study

of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). J Infect Dis **2009**; 199:419–426.
6. Fellay J, Ge D, Shianna KV, et al. Common genetic variation and the control of HIV-1 in humans. PLoS Genet **2009**; 5:e1000791.
7. Herbeck JT, Gottlieb GS, Winkler CA, et al. Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. J Infect Dis 2010; 201:618–626.
8. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature **2009**; 461:747–753.
9. Ge D, Fellay J, Thompson AJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature **2009**; 461: 399–401.
10. Suppiah V, Moldovan M, Ahlenstiel G, et al. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet **2009**; 41:1100–1104.
11. Tanaka Y, Nishida N, Sugiyama M, et al. Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet **2009**; 41:1105–1109.
12. Rauch A, Kutalik Z, Descombes P, et al. Genetic variation in IL28B Is associated with chronic hepatitis C and treatment failure: a genome-wide association study. Gastroenterology **2010**; 138:1240–1243.
13. Zhang FR, Huang W, Chen SM, et al. Genomewide association study of leprosy. N Engl J Med **2009**; 361:2609–2618.
14. Jallow M, Teo YY, Small KS, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet **2009**; 41: 657–665.
15. Biesecker LG. Exome sequencing makes medical genomics a reality. Nat Genet **2009**; 42:13–14.
16. Halperin E, Stephan DA. SNP imputation in association studies. Nat Biotechnol **2009**; 27:349–351.
17. 1000 Genomes Web site. http://www.1000genomes.org.
18. Giri MS, Nebozhyn M, Showe L, Montaner LJ. Microarray data on gene modulation by HIV-1 in immune cells: 2000–2006. J Leukoc Biol **2006**; 80:1031–1043.
19. Rotger M, Dang KK, Fellay J, et al. Genome-wide mRNA expression correlates of viral control in CD4+ T-cells from HIV-1-infected individuals. PLoS Pathog **2010**; 6(2):e1000781.
20. Lederer S, Favre D, Walters KA, et al. Transcriptional profiling in pathogenic and non-pathogenic SIV infections reveals significant distinctions in kinetics and tissue compartmentalization. PLoS Pathog **2009**; 5:e1000296.
21. Bosinger SE, Li Q, Gordon SN, et al. Global genomic analysis reveals rapid control of a robust innate response in SIV-infected sooty mangabeys. J Clin Invest **2009**; 119:3556–3572.
22. Jacquelin B, Mayau V, Targat B, et al. Nonpathogenic SIV infection of African green monkeys induces a strong but rapidly controlled type I IFN response. J Clin Invest **2009**; 119:3544–3555.
23. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **2009**; 10:57–63.
24. Geiss GK, Bumgarner RE, Birditt B, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol **2008**; 26:317–325.
25. Telenti A. HIV-1 host interactions: integration of large-scale datasets. F1000 Biol Rep **2009**; 1:71.
26. Bushman FD, Malani N, Fernandes J, et al. Host cell factors in HIV replication: meta-analysis of genome-wide studies. PLoS Pathog **2009**;5: e1000437.
27. Amit I, Garber M, Chevrier N, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science **2009**; 326:257–263.