Helicobacter pylori: After the Genomes, Back to Biology

Antonello Covacci and Rino Rappuoli

IRIS, Chiron Vaccines Srl, Via Fiorentina 1, 53100 Siena, Italy

The race to sequence the human genome generated a global emotional wave which escaped the scientific community and involved the media, the politicians, the economic world, and the general public. The sequencing of several bacterial genomes also generated a high interest. A few years after publication of the genomes we are back at the bench, performing *ad hoc* experiments and asking ourselves what the genomic wave meant and how it changed our lives. The genome of *Helicobacter pylori*, the bacterium which causes peptic ulcer and gastric cancer, was published back in 1997 (1), and it is very old in genomic terms (*H. pylori* was the fourth bacterial genome to be published after the one of *Haemophilus influenzae*, *Mycoplasma genitalium*, and *Methanococcus jannaschii). H. pylori* was also the first bacterium for which the genomes of two different strains were determined (2), and the first one for which maps of protein–protein interactions were published (3). In spite of this, our knowledge is still limited and pregenomic experiments are still needed to unravel the secrets of how this bacterium causes disease. A paper describing the first application of signature tagged mutagenesis (STM) to identify virulence factors of *H. pylori*, published in this issue (4), provides us with an opportunity to think about the biology in the postgenomic era, the role of the pregenomic techniques in general, and also what the new findings mean for *H. pylori*.

New Insights for H. pylori Colonization and Virulence. Kavermann et al. collected 960 mutants of *Helicobacter pylori*, one of the largest collections described in literature, and tested them for colonization into a suitable animal model (Mongolian gerbils) by using STM to identify genes essential for survival within the animal host. In the original STM developed for *Salmonella* (5), a library of mutants, obtained by transposon mutagenesis and each tagged by a unique sequence, were used in pools to infect an animal model. The mutants which did not survive the infection in vivo were then identified by the absence of the tag. Given the difficulty of *H. pylori* genetic manipulation, a library large enough to represent the entire genome could not be isolated; however, although incomplete, the one obtained allowed finding of some of the known virulence factors and also of some new ones. Among the known factors, most of the genes involved in flagella biosynthesis and four of the genes involved in the urease synthesis were found. The experiment also identified eight hypothetical proteins, and several annotated proteins whose role in colonization is not obvious. It is possible that many of the newly identified proteins do not have a direct role in colonization and that they are only necessary for the changes of metabolism required to survive in the new environment. The role of genes essential for in vivo survival by exerting an indirect role will remain difficult to explain until we have a full understanding of the metabolic chart of this bacterium within its host. A new essential factor identified in the paper that is likely to have a direct role is a collagenase. This type of proteases are known to digest the extracellular matrix, thus promoting the invasive growth of malignant tumors and the invasion of tissues by some bacteria. In the case of *Helicobacter*, collagenase is unlikely to play this role since there is no extracellular matrix in the area colonized by the bacterium. In this case it is more likely that a collagenase may digest the mucous layer, which otherwise would be too thick to allow the passage of the bacterium.

The experiment did not identify other known virulence factors such as the BabA, AlpA, AlpB, HopZ, HpA, and other adhesins. This is probably explained by the fact that a bacterium uses multiple copies of several different adhesins making the absence of one of them not relevant in animal models. In fact, STM has never identified a bacterial adhesin.

The absence of the *cag* pathogenicity island among the factors necessary for colonization in vivo is more difficult to explain. Indeed, this 40 Kb region coding for a type IV secretion system that injects the CagA protein into host cells, is known to be necessary for virulence, and to play an essential role in the induction of peptic ulcer disease and cancer. In addition, in a recent paper (6), this region was shown to be involved in the early phase of colonization (first 10 d). Perhaps the early effects of *cag* in colonization cannot be detected by the STM technique which has a time-frame of three weeks and uses a simple scheme like presence/absence of bacteria instead of measuring the level of attenuation. Moreover, it should be considered that the authors used as a recipient strain an organism possessing an incomplete *cag* region. It is a pity that such a labor intense experiment was performed with a bacterium known to be defective for such a major virulence tract of *H. pylori*. Failure to

Address correspondence to R. Rappuoli, IRIS, Chiron Vaccines Srl, Via Fiorentina 1, 53100 Siena, Italy. Phone: 39-0577-243414; Fax: 39-0577-278508; E-mail: rino_rappuoli@chiron.it

 J. Exp. Med. © The Rockefeller University Press • 0022-1007/2003/04/807/5 $8.00
Volume 197, Number 7, April 7, 2003 807–811
http://www.jem.org/cgi/doi/10.1084/jem.20030063

use the most clinically relevant strain may prevent the detection of major virulence factors and also decreases the general implications of the entire work, because one does not know to what extent the findings may apply to the real pathogen.

Considering the new data with those already known, we propose the scheme shown in Fig. 1 to describe the essential steps in *H. pylori* pathogenesis. During colonization, the flagella, the urease and the collagenase are all necessary to propel the bacteria, buffer the acidic pH, and soften the mucus, respectively. Once the bacteria reach the epithelial layer, they adhere to the cells using BabA, AlpA, AlpB, HopZ, HpA, and other adhesins. They then damage the tissue by releasing toxins such as the vacuolating cytotoxin (VacA), by injecting other toxins such as CagA into host cells, and by releasing factors such as the neutrophil activating protein (NAP), which activates inflammatory cells such as neutrophils and mast cells. Recently, important progresses were made in our understanding of the mode of action of CagA. The *cag*A gene was cloned and recognized as a virulence factor associated with disease progression long before a function was assigned to the protein (7). Subsequently, this gene was found to be part of a large pathogenicity island coding for a secretory system which injects the CagA protein into host cells. Once inside the cells, the protein triggers signal transduction acting as a growth factor and alters the integrity of the epithelium by acting at the level of the apical junctional complex (AJC; unpublished data). Today we are beginning to understand some of the molecular mechanisms by which CagA causes tissue damage, which during a long-lasting chronic infection may lead to cancer (Fig. 2).

Inside the host cell, CagA subverts the cellular functions by interacting directly with the host protein ZO-1. This is part of the AJC, and the interaction disrupts the integrity of the epithelium. Inside the host cell CagA is also recognized by the c-Src and Lyn tyrosine kinases which phosphorylate a tyrosine in a repeated EPIYA motif. Once tyrosine-phosphorylated, CagA binds SHP-2 that is recruited to the membrane and triggers a signal cascade which results in cell scattering and proliferation, a phenotype that resembles the one induced by growth factors. A similar phenotype is also induced by nonphosphorylated CagA after binding Grb-2. CagA has no homologues in any database, structural predictions do not suggest any particular function and the site for phosphate transfer is not comprised within the lists of functional motifs. Genes like *cag*A have little chance to be identified by selection methods in the absence of a biological assay, but once assays have been established they can be applied to a wide range of pathogens for the detection of tyrosine-phosphorylated molecules, for cell shape elongation, or for junctional activity in MDCK monolayers.

Genomics Adds Information but Progress in Knowledge Is Slow. Genetics has become more powerful in the postgenomic era and today the relevant biological questions can

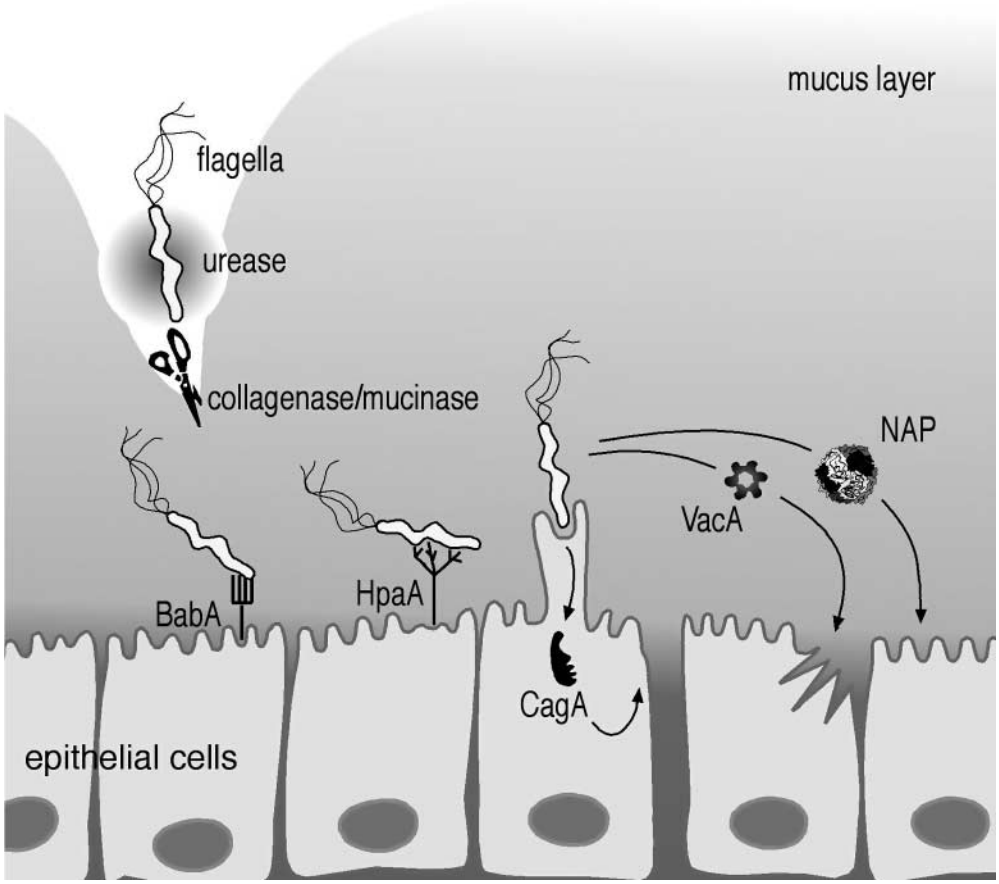

Figure 1. Summary of the main steps involved in *H. pylori* colonization, and of the factors involved. The figure shows the possible role for a new factor (the collagenase) described for the first time in the paper by Kavermann et al. in this issue (reference 4), which is proposed to cut the mucin layer, allowing the passage of the bacterium.

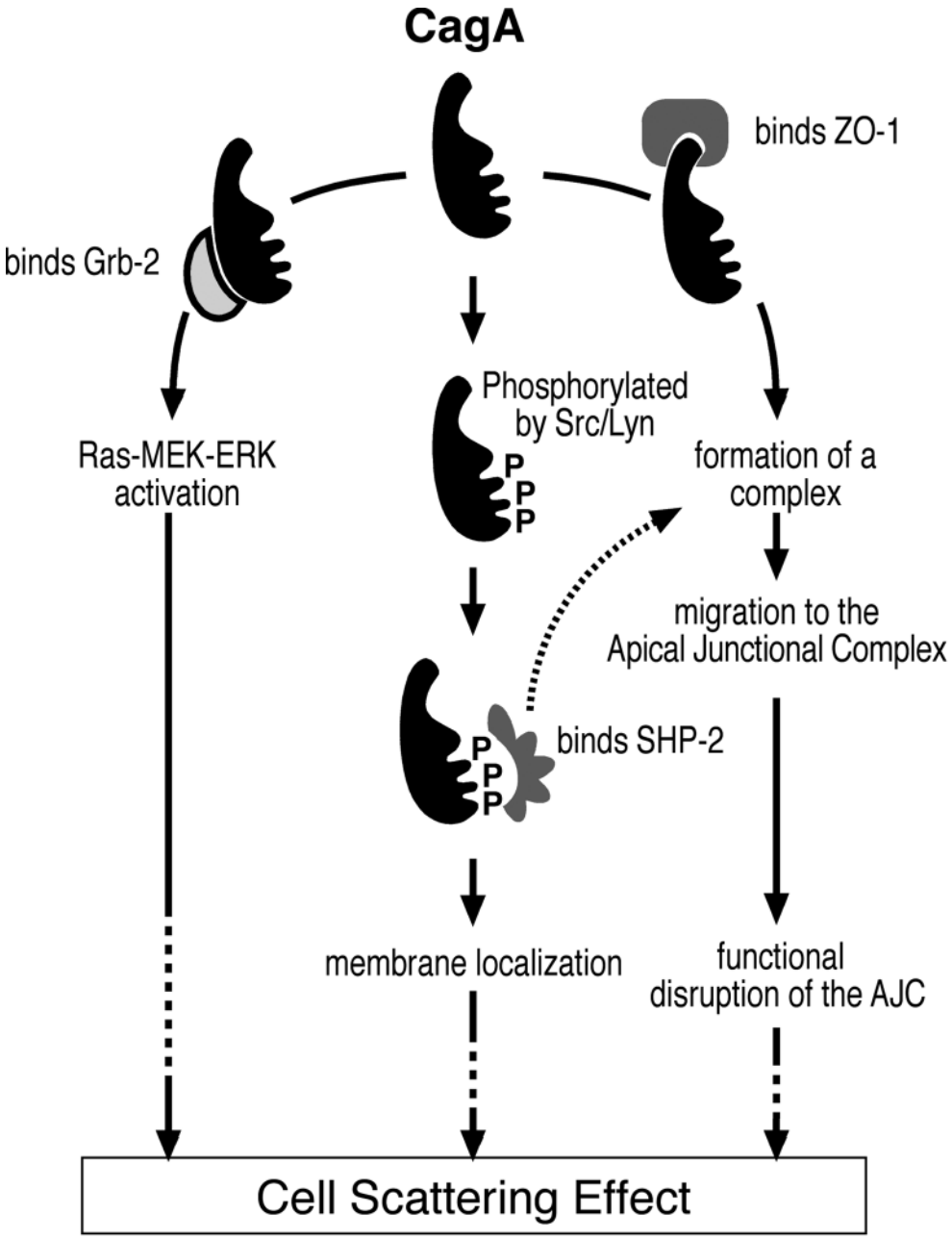

Figure 2. Known interactions of the CagA protein.

be addressed globally. However, in spite of the enormous amount of information added to databases, real progress in knowledge is rather slow. Complete genomes include more than 95 bacterial species with clinically relevant isolates, the yeast *Saccharomyces cerevisiae*, invertebrates like *Caenorhabditis elegans* and *Drosophila melanogaster*, vertebrates, including *Homo sapiens*, and plants. During the determination of genomic sequences, polished raw sequences linked after gap closure move into the annotation process, that consists in prediction of genetic elements, classification of typical primary sequence patterns as functional motifs or by structural, phylogenetic, or topological criteria. Proteins are eventually indexed into a scheme, or ontology, that recapitulates all the available information. The large majority of proteins are very similar to a Lego toy: blocks of amino acids are stacked to form a polymer and one or more blocks contain the information for subcellular localization, for association into complexes, signatures for catalytic sites, or for acceptor sequences. Most of them are wired into circuits and could perform more than a single operation. We do not have a grammar or syntax for rules for functional prediction and we can infer the secondary structure only to a limited extent. The best we can do is to group them into families, because they have a certain degree of resemblance, and to mark common blocks. If one member of the family was recognized to perform a particular function, we may extrapolate that similar proteins will perform the same function. There is a tendency to overestimate the contribu-

tion of biocomputing, a diffuse sense dominated by the assumption that we do not need more than we have in our primary sequence. Functional prediction or *in silico* discovery is entirely performed by crossing mines of data to identify intersections. The primary source of data is scientific literature. We currently access literature with simple keywords: the best way is to decompose any form of scientific communication (including lab notebooks) into facts, rules, or data and use them in association with other methods.

To analyze the abstract set of data stored in databases and analyzed by computation, miniatures of the entire set of genes can be collated in small surfaces to generate microarrays, which can be used for DNA hybridization or for the evaluation of global transcription patterns. In addition, 2D maps of cell extracts followed by microsequence can identify most expressed proteins and their interactions can be captured by several techniques, the most popular of which is the two hybrid system. While microarrays and proteomics have been slow in delivering novel functional data, they produced spectacular results in recognizing profiles. Perhaps the most remarkable result is the typing of human lymphomas based on expression profiles of genes involved in tumorigenic progression and metastasis (8, 9). Similar remarkable results were the determination of the sequential activation of genes that are relevant for development in Drosophila or Zebrafish, or the identification of single nucleotide polymorphisms or SNPs of selected patients from segregated families that are likely to provide insight in multigene diseases. Ideally, we would like to compare expression data from microarrays. However, most of the data have a nonstandard format that prevent us from properly operating on them.

Microarrays and proteomics are postgenomic experimental techniques but also pregenomic techniques such as STM, IVET, the two hybrid system, and others benefit from knowledge of the genomes. Interestingly, when each of these experimental techniques are applied to the whole genomes, very rarely do they arrive to similar conclusions, even when they start from similar questions. We can consider each of the experimental techniques as a narrow window that allows us to visualize a part of the picture containing the biological information (Fig. 3). Depending on the window used, we see different subsets of the biological picture. In the case of *H. pylori*, the STM filter applied allowed for the visualization of the flagellin genes, the urease, collagenase and missed other important factors. A different portion of the same picture has been visualized by microarrays (10), and another part has been visualized by the genomic approach that generated the complete map of protein–protein interactions (3). Several important virulence factors were missed by all genomic approaches. Retrospective studies show that most of what is referred to as global analysis identify only a minor part of the information, and only rarely the information detected by one approach overlaps with that detected by a different approach (11). The question is how many windows we need to have in order

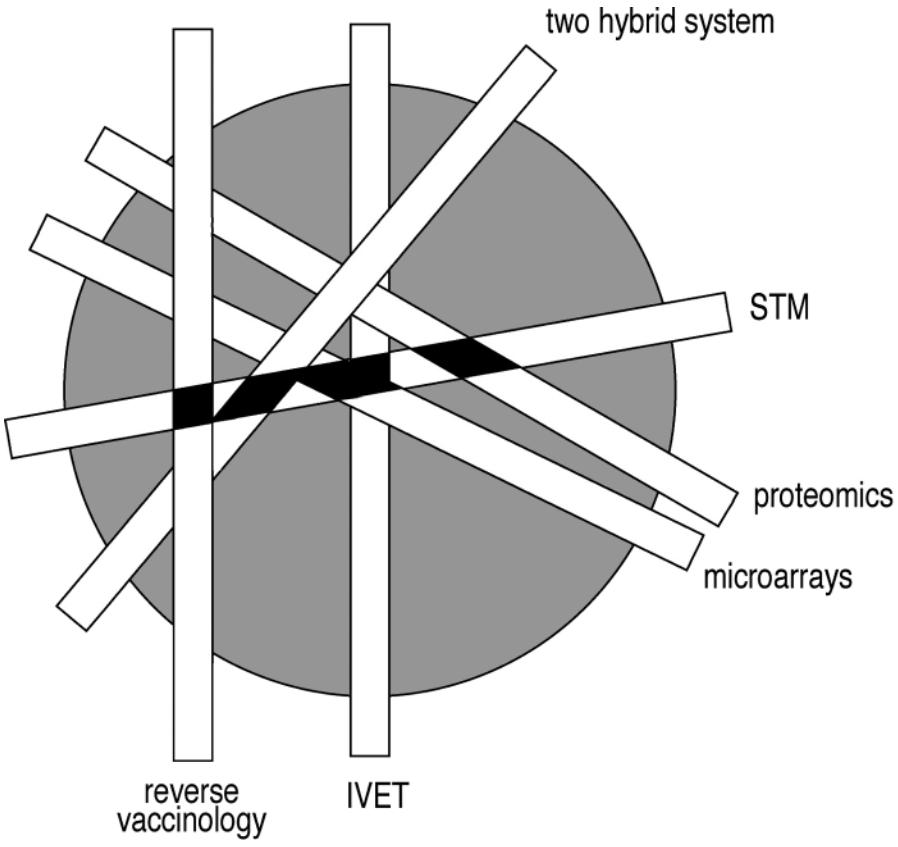

Figure 3. Schematic representation showing the progress of knowledge in the genomic era. Each of the technologies intended to globally study a biological event represents in fact only a small window (drawn in the figure as small rectangular surface) able to visualize a part of the whole picture (represented by the circle). When several of these technologies are applied, they see mostly different parts of the picture with minor areas of overlap.

to visualize the whole picture and, more importantly, whether there are more efficient ways to make progress.

Submitted: 14 January 2003
Revised: 24 January 2003
Accepted: 7 February 2003

References

1. Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty, et al. 1997. The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature*. 388:539–547.
2. Alm, R.A., L.S. Ling, D.T. Moir, B.L. King, E.D. Brown, P.C. Doig, D.R. Smith, B. Noonan, B.C. Guild, B.L. deJonge, et al. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*. 397:176–180.
3. Rain, J.C., L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, et al. 2001. The protein-protein interaction map of Helicobacter pylori. *Nature*. 409:211–215.
4. Kavermann, H., B.P. Burns, K. Angermüller, S. Odenbreit, W. Fischer, K. Melchers, and R. Haas. 2003. Identification and characterization of *Helicobacter pylori* genes essential for gastric colonization. *J. Exp. Med*. 197:813–822.
5. Hansel, M., J.E. Shea, C. Gleeson, M.D. Jones, E. Dalton, and D.W. Holden. 1995. Simultaneous identification of bacterial virulence genes by negative selection. *Science*. 269:400–403.
6. Marchetti, M., and R. Rappuoli. 2002. Isogenic mutants of the cag pathogenicity island of *Helicobacter pylori* in the mouse model of infection: effects on colonization efficiency. *Microbiology*. 145:1447–1456.
7. Covacci, A., S. Censini, M. Bugnoli, R. Petracca, D. Burroni, G. Macchia, A. Massone, E. Papini, Z. Xiang, N. Figura, and R. Rappuoli. 1993. Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. *Proc. Natl. Acad. Sci. USA*. 90:5791–5795.
8. Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 403:503–511.
9. Sauter, G., and R. Simon. 2002. Predictive molecular pathology. *N. Engl. J. Med*. 347:1995–1999.
10. Salama, N., K. Guillemin, T.K. McDaniel, G. Sherlock, L. Tompkins, and S. Falkow. 2000. A whole genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. USA*. 97:14668–14673.
11. Tong, A.H., B. Drees, G. Nardelli, G.D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, et al. 2002. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*. 295:321–324.