Genome Sequence Diversity and Clues to the Evolution of Variola (Smallpox) Virus

Joseph J. Esposito *et al.*

Science **313**, 807 (2006);
DOI: 10.1126/science.1125134

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Eastern Sahara is the onset of semi-arid conditions in the north and semihumid conditions in the south at about 8500 B.C.E. Within only a few centuries, the desert margin shifted up to 800 km north to latitude 24°N, bringing monsoonal rainfall to most of the former desert. Taking into account the west-east gradient of decreasing humidity from the Atlantic Ocean to the Red Sea, this process apparently applied to the entire Sahara.

This fundamental climatic change from terminal Pleistocene hyper-arid desert conditions to savannah-type vegetation and the formation of lakes and temporary rivers resulted in the rapid dissemination of wild fauna and the swift reoccupation of the entire Eastern Sahara by prehistoric populations. Relatively stable humid conditions prevailed over approximately the next 3200 calendar years between 8500 and 5300 B.C.E. Abrupt drying events stated elsewhere in the Sahara may be explained by fading rainfall at a specific latitudinal position at a certain moment, or by dropping local groundwater.

The roughly parallel southward shift of monsoonal precipitation that set in at 5300 B.C.E. can be tracked through the following millennia by the discontinuance of the sedimentary record of aquatic deposits at decreasing latitudes. The geological archives in agreement with the archaeological evidence indicate a gradual desiccation and environmental deterioration of the Eastern Sahara, notwithstanding transitory climatic perturbations that are a common feature of all desert margins. This rather linear process culminates in the present extremely arid conditions, which have not yet reached the extent of the terminal Pleistocene (fig. S2F).

The southward movement of human settlement implied substantial changes in the pattern of behavior and land use as response to regional environmental differences. Most of all, mobility was the key to survival; it has driven prehistoric societies from foraging to a multi-resource economy and specialized pastoralism. The final desiccation of the Egyptian Sahara also had an essential impact on the contemporaneous origin of the pharaonic civilization in the Nile valley. To this day, conflicts in sub-Saharan regions such as Darfur are rooted in environmental deterioration, aggravated by severe demographic growth and man-made desertification. The presented data and conclusions suggest that the climate-controlled desiccation and expansion of the Saharan desert since the mid-Holocene may ultimately be considered a motor of Africa's evolution up to modern times.

References and Notes

1. W. Dansgaard *et al.*, *Nature* **364**, 218 (1993).
2. U. von Grafenstein, H. Erlenkeuser, J. Müller, J. Jouzel, S. Johnsen, *Clim. Dyn.* **14**, 73 (1998).
3. D. Henning, H. Flohn, *Climate Aridity Index Map* (U.N. Doc. A/CONF. 74/31, Nairobi, Kenya, 1977).
4. F. Wendorf, R. Schild, Eds., *Holocene Settlement of the Egyptian Sahara, The Archaeology of Nabta Playa* (Kluwer Academic, New York, 2001).
5. H.-J. Pachur, N. Altmann, in *Palaeogeographic-Palaeotectonic Atlas of North-Eastern Africa, Arabia, and Adjacent Areas*, H. Schandelmeier, P.-O. Reynolds, A.-K. Semtner, Eds. (Balkema, Rotterdam, Netherlands, 1997), pp. 111–125, pl. 17.
6. N. Petit-Maire, "Map of the Sahara in the Holocene 1:5,000,000" (CGMW/UNESCO, Paris, 1993).
7. S. Kröpelin, in *Nordost-Afrika: Strukturen und Ressourcen*, E. Klitzsch, U. Thorweihe, Eds. (Wiley-VCH, Weinheim, Germany, 1999), pp. 448–508.
8. C. V. Haynes Jr., *Geoarchaeology* **16**, 119 (2001).
9. J. C. Ritchie, C. H. Eyles, C. V. Haynes, *Nature* **314**, 352 (1985).
10. H.-J. Pachur, S. Kröpelin, *Science* **237**, 298 (1987).
11. P. Hoelzmann, B. Keding, H. Berke, S. Kröpelin, H.-J. Kruse, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **169**, 193 (2001).
12. H. J. Pachur, B. Wünnemann, *Palaeoecol. Africa* **24**, 1 (1996).
13. K. Nicoll, *Quat. Sci. Rev.* **23**, 561 (2004).
14. F. Gasse, *Quat. Sci. Rev.* **19**, 189 (2000).
15. P. Hoelzmann *et al.*, in *Past Climate Variability Through Europe and Africa*, R. W. Battarbee, F. Gasse, C. E. Stickley, Eds. (Springer, Dordrecht, Netherlands, 2004), pp. 219–256.
16. B. Barich, F. Hassan, *Origini* **13**, 117 (1988).
17. M. M. A. McDonald, in *The Oasis Papers*, C. A. Marlow, A. J. Mills, Eds. (Oxbow, Oxford, 2001), pp. 26–42.
18. F. A. Hassan, *Afr. Archaeol. Rev.* **3**, 95 (1985).
19. K. Neumann, *Afr. Archaeol. Rev.* **7**, 97 (1989).
20. P. M. Vermeersch, *Africa Praehistorica* **14**, 27 (2002).
21. F. Wendorf, in *The Prehistory of Nubia* (SMU Press, Dallas, TX, 1968), pp. 954–995.
22. F. Jesse, *Africa Praehistorica* **16**, 1 (2003).
23. H. Riemer, in *Egyptology at the Dawn of the Twenty-First Century*, Z. Hawass, L. P. Brock, Eds. (American Univ. Cairo Press, Cairo, Egypt, 2003), pp. 408–415.
24. P. M. Vermeersch, *L'Elkabien: Epipaléolitique de la Vallée du Nil Egyptien* (Leuven Univ. Press, Leuven, 1978).
25. B. Gehlen, K. Kindermann, J. Linstädter, H. Riemer, *Africa Praehistorica* **14**, 85 (2002).
26. K. Kindermann, *Archéo-Nil* **14**, 31 (2004).
27. R. Kuper, *Cah. Rech. Inst. Papyrol. Egyptol. Lille* **17**, 123 (1995).
28. P. M. Vermeersch, P. V. Peer, J. Moeyersons, W. V. Neer, *Sahara* **6**, 31 (1994).
29. A. E. Close, *Africa Praehistorica* **14**, 459 (2002).
30. R. Kuper, in *Environmental Change and Human Culture in the Nile Basin and Northern Africa Until the Second Millennium B.C.*, M. Kobusiewicz, L. Krzyzaniak, Eds. (Poznan Archaeological Museum, Poznan, Poland, 1993), pp. 213–223.
31. M. M. A. McDonald, *J. Anthropol. Arch.* **17**, 124 (1998).
32. J. Linstädter, S. Kröpelin, *Geoarchaeology* **19**, 753 (2004).
33. D. Wengrow, in *Ancient Egypt in Africa*, D. O'Connor, A. Reid, Eds. (UCL Press, London, 2003), pp. 121–135.
34. F. Jesse, S. Kröpelin, M. Lange, N. Pöllath, H. Berke, *J. Afr. Archaeol.* **2**, 123 (2004).
35. K. P. Kuhlmann, *Africa Praehistorica* **14**, 125 (2002).
36. R. Kuper, *Antiquity* **75**, 801 (2001).
37. R. Kuper, F. Förster, *Egypt. Archaeol.* **23**, 25 (2003).
38. R. Kuper, *Bull. Soc. Fr. Egyptol.* **158**, 12 (2003).
39. P. B. deMenocal, *Science* **292**, 667 (2001).
40. H. Weiss, R. S. Bradley, *Science* **291**, 609 (2001).
41. P. B. deMenocal, *Quat. Sci. Rev.* **19**, 347 (2000).
42. M. Claussen *et al.*, *Geophys. Res. Lett.* **26**, 2037 (1999).

43. Fieldwork was supported by the Deutsche Forschungsgemeinschaft (DFG), the Supreme Council of Antiquities (SCA) in Egypt, the National Corporation for Antiquities and Museums (NCAM) and the Geological Research Authority (GRAS) in Sudan, and the Centre National d'Appui à la Recherche (CNAR) in Chad. This paper was written during an invited visit at St. John's College, Cambridge. We thank J. Alexander, Cambridge; P. M. Vermeersch, Leuven; D. Verschuren, Gent; and the reviewers for critical reading of the manuscript and constructive suggestions as well as K. Kindermann, H. Riemer, G. Wagner, and all other team members for their assistance in the institute and on the many expeditions in the Sahara.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1130989/DC1
Materials and Methods
Figs. S1 and S2
Table S1
References

7 June 2006; accepted 12 July 2006
Published online 20 July 2006;
10.1126/science.1130989
Include this information when citing this paper.

Genome Sequence Diversity and Clues to the Evolution of Variola (Smallpox) Virus

Joseph J. Esposito,[1*†] Scott A. Sammons,[1*] A. Michael Frace,[1*] John D. Osborne,[1*‡] Melissa Olsen-Rasmussen,[1*] Ming Zhang,[1§] Dhwani Govil,[1] Inger K. Damon,[2] Richard Kline,[2] Miriam Laker,[2‖] Yu Li,[2] Geoffrey L. Smith,[3] Hermann Meyer,[4] James W. LeDuc,[2] Robert M. Wohlhueter[1]

Comparative genomics of 45 epidemiologically varied variola virus isolates from the past 30 years of the smallpox era indicate low sequence diversity, suggesting that there is probably little difference in the isolates' functional gene content. Phylogenetic clustering inferred three clades coincident with their geographical origin and case-fatality rate; the latter implicated putative proteins that mediate viral virulence differences. Analysis of the viral linear DNA genome suggests that its evolution involved direct descent and DNA end-region recombination events. Knowing the sequences will help understand the viral proteome and improve diagnostic test precision, therapeutics, and systems for their assessment.

Before eradication was declared in 1980, the *Orthopoxvirus* (OPV) variola virus (VARV) caused from ~1 to 30% case-fatality rates (CFRs) of smallpox, a strictly human disease. The infection began with a prodrome of systemic aches and a fever that peaked in about a week. As the fever broke, an oropharyngeal enanthema developed, followed immediately by an exanthema, a skin rash constituting an end stage of centrifugally distributed virus-filled pustules that felt "shotty," as if each contained a

pellet (*1*). Infectious VARV materials are in two secure repositories authorized by the World Health Organization (WHO); one is at the U.S. Centers for Disease Control and Prevention (CDC), and the other is at Russia's State Research Center of Virology and Biotechnology (VECTOR) (*2*–*4*). Between the two repositories, there is little overlap of isolates. Because a low-risk–high-consequence threat of smallpox by terrorism exists (*5*), the WHO oversees biosafety level 4 live VARV research aimed at developing accurate diagnostic tests, therapeutics, and systems to assess these countermeasures, which have been advocated by the U.S. Institute of Medicine (*6*).

To advance the research effort, we undertook the present comparative genomics study to investigate features of the sequence diversity and evolutionary relationships of 45 temporally, geographically, and epidemiologically varied viral isolates (table S1). Most isolates date to the WHO Intensified Smallpox Eradication Programme (1967 to 1980), a campaign for eliminating endemic disease in Africa, Asia, and South America and preventing smallpox spread back into Europe and elsewhere (*1*). For this study, we sequenced 43 isolates, including the previously described BSH75_banu (*7*, *8*), which we resequenced, and sequenced cowpox virus CPXV-GER91, vaccinia virus VACV-AC2000, and Tatera gerbilpox virus TATV-DAH68 (table S1). We also used reported genome sequences of VARV IND67_mah and BRZ66_gar (*9*, *10*). Forty-four VARV isolates are at the CDC; IND67_mah is at VECTOR.

Many of the isolates are associated with a CFR, an epidemiological estimate that enables classification of outbreak patterns roughly as minor (mostly less than 1% CFR) or major (usually more than 10% CFR). CFRs are generally reasonable estimations of VARV innate virulence, considering that inconsistencies—including case-patient health, age, health care quality, and immune status—affected such calculations. Low-CFR outbreaks of mild smallpox known as "alastrim" first became obvious in the Americas at the turn of the 20th century, which was about when another minor form called "amaas" appeared in Africa (*1*). It was not until the 1960s that a VARV growth-ceiling temperature test (*11*) enabled phenotyping that differentiated alastrim virus from amaas and other VARVs, thereby validating that at least two subspecies existed.

In West Africa, outbreaks with age-adjusted intermediate CFRs were observed, which motivated attempts, with some success, to differentiate isolates into three classes by using the thermal test and other tests (*1*, *12*–*14*), but eradication ensued and taxonomic interest faded. Despite some discordance, CFRs are all that remain for understanding the innate virulence of different isolates, so we used CFRs to help us recognize putative encoded proteins that might mediate virulence differences between the isolates. The associated CFRs of some isolates are discussed in the supplemental text (*15*).

The VARV genome. VARV contains a linear genome DNA of ~186 kilobase pairs (kbps) with covalently closed ends adjoining small inverted terminal repeat (ITR) regions that flank the coding region sequences (CRS) (Fig. 1) (*7*, *8*, *16*, *17*). Depending on the isolate, software predictions indicate that the CRS of the 45 isolates contain 196 to 207 open reading frames (ORFs), using a low-end cutoff of ≥50 codons (tables S1 and S2), which indicates that the isolates have essentially the same functional gene content. Like all poxviruses, ORFs are closely spaced and nonoverlapping, and there is no mRNA splicing (*18*); however, unlike other poxviruses, VARV lacks ORFs in the ITR region (fig. S2) (*16*, *17*). Depending on the isolate, nearly 90% of predicted ORFs coincide with entire ORFs predicted for other OPVs, and the rest appear to be ≥50-codon segments of entire or partial ORFs predicted for other OPVs, mainly the archetype VACV (fig. S1, annotation).

Intragenomic sequence diversity. Most VARV DNA preparations show, at nucleotide positions ~14,000 and ~161,000 (Fig. 1), respectively, a different number of T and AT repeats (table S3), which suggests that DNA replication involves site-specific viral DNA polymerase stuttering or intragenomic recombination. The T variance might provide frame-shift modulation at a polylysine site encoded by ORF-16, a homolog of a VACV ORF for an intracellular kelchlike protein that somehow alters Ca^{++}-independent adhesion of infected cells to the extracellular matrix (*19*). Ubiquitination of a variable lysine site might diversify an antigenic determinant. The AT variance precedes ORF-192, a VACV-WR ORF homolog for an interleukin-1β binding protein. VACV mutants deficient in this protein induce fever in mice infected intranasally and are more virulent than the parental VACV (*20*). The variance might ensure an ORF-192 deficiency in VARV.

Intergenomic sequence diversity. To investigate intergenomic sequence diversity, we aligned the 45 VARV CRS into a multiple alignment. We calculated diversity (π) of nucleotide sites across the alignment by counting single-nucleotide polymorphisms (SNPs) in columns with no gaps, and to make maximal use of the alignment, we scored insertions-deletions (indels, i.e., gaps) of ≥1 base as a single polymorphism (*15*). The π values derive from the 990 possible pairs of the 45 VARV CRS and represent 1782 specific SNPs and 4812 specific indels. By sliding a window of a defined site length and site step size across the alignment, we plotted (π) midpoint values (Fig. 2A). The diversity distribution agrees with general observations that the central CRS of poxviruses mainly specify conserved proteins essential for virus replication and that the terminal CRS encode for more divergent proteins, including those modulating host range and virulence (*18*). To estimate the extent of diversity of one ORF from another in the genome, we also divided the π value of each ORF of the 45 VARVs by its average number of codons to obtain π^{rel} (fig. S1). In a

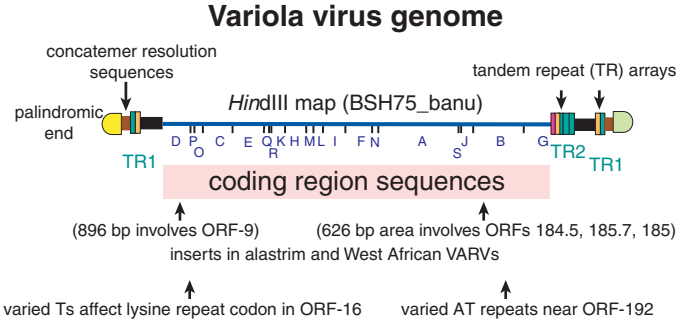

Fig. 1. VARV genome architecture. The VARV genome, depicted on the BSH75_banu *Hin*dIII map (*40*), is a linear ~186-kbp covalently closed DNA. The ~60-bp end loops (sequences not determined) are palindromes (light green or yellow) distal to concatemer resolution sequences (brown). Tandem repeat (TR) array units (aqua, orange, or magenta) flank the CRS (pink), which contain all the ORFs. TR arrays are within regions of inverted terminal repeats (ITRs) that contain no genes (fig. S2). TRs contain 69 to 70 bp units (aqua), partial units of 54 bp (orange) or 22 bp (magenta). The area has segments of nonrepeated sequences (black). The extra TR arrays (e.g., TR2) make the DNA topographically asymmetrical, although some isolates have a symmetrical appearance (fig. S2). Other areas indicated include those containing repeated Ts or ATs found in several isolates (table S3) and those showing gene-loss areas involving ORF-9 and ORFs 184 to 185 (fig. S3).

[1]Biotechnology Core Facility Branch, Division of Scientific Resources, National Center for Preparedness, Detection, and Control of Infectious Diseases¶ and [2]Poxvirus and Rabies Branch,¶ Division of Viral and Rickettsial Diseases, National Center for Zoonotic, Vector-Borne, and Enteric Diseases,¶ Coordinating Center for Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA. [3]Imperial College London, St. Mary's Campus, London, W2 1PG, UK. [4]Bundeswehr Institute of Microbiology, Munich, Germany 80937.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jesposito@cdc.gov
‡Present address: Northwestern University, Chicago, IL 60611, USA.
§Present address: University of Goettingen, Goettingen, Germany 37077.
‖Present address: 324 North Kern Street, Ridgecrest, CA 93555, USA.
¶Unit designations pending agency reorganization.

sense, π^{rel} describes the extent of branching of a phylogram of an ORF.

To establish the diversity between any two isolates of interest, we separately summed the SNPs and the indels for each of the 990 VARV pairs of CRS (table S4). Relative to the size of the VARV genome, all pairs show low numbers, consistent with low diversity, which increases the likelihood that sequence-based detection methods will effectively identify a re-emergent VARV. To understand diversity further, we used sliding window plots of π midpoints of VARV pairs of interest, such as between isolates from the same outbreak or those that differ geographically or temporally (Fig. 2, B to G).

Because there is concern (*21*) that biotechnology enables construction of dangerous pathogens from genetic material of naturally occurring organisms, we plotted π midpoints and calculated the number of SNPs and indels that distinguish VARV from other OPVs, including isolates of monkeypox (MPXV), camelpox (CMLV), and gerbilpox (TATV) (Fig. 2, H to J). The results indicate that just a few thousand mutations by one of several rapid methods (*22*) could convert such OPV DNAs into VARV DNA. Recently, infectious VACV was recovered by rescuing its genome from a bacterial artificial chromosome containing its DNA (*23*). In theory, one could use sequences to synthesize long oligonucleotides (*24*) to reconstruct VARV DNA. Such rapidly advancing technology makes it important to understand sequence diversity and the proteome so as to develop and maintain countermeasures against malefic-use–created pathogens.

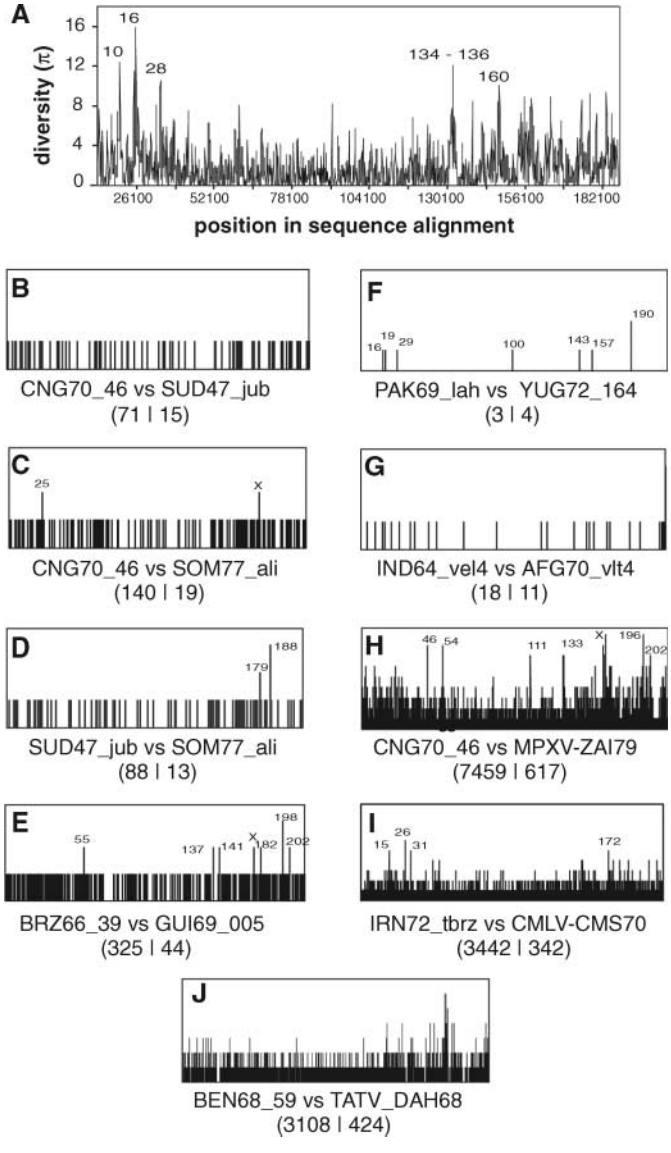

Fig. 2. Diversity of coding region sequences. (**A**) Polymorphic sites (*15*) determined by plotting diversity (π) midpoint values using a sliding window length of 200 sites moved in steps of 100 across an alignment of 45 VARV CRS. (**B** to **J**) Maps of π midpoint values [axis titles same as in (A)], using a window length of 12 sites moved in steps of 12, show loci separating viral pairs that differ by CFR, geographically, and/or temporally [SNP and indel values (table S3) are within parentheses]. (B) CNG70_46 (10% CFR) and SUD47_jub (<1% CFR). (C) CNG70_46 and SOM77_ali (0.4% CFR). (D) SUD47_jub and SOM77_ali. (E) BRZ66_39 (0.8% CFR) and GUI69_005 (8% CFR). (F) Pairing of isolates from the start and end of the 1969–1972 reintroduction of smallpox (16% CFR) into the Mideast and Yugoslavia (*1*) indicates a genetically very stable virus spread through hundreds of people. (G) Alignment of CRS of AFG70_vlt4 and IND64_vel4, which was isolated 5 years before the reintroduction, supports phenotypic and epidemiologic data implicating IND64_vel4 in the 1969 to 1972 Mideast smallpox resurgence (*41*). (H to J) Polymorphic sites that distinguish (H) CNG70_46 from Congo MPXV-ZAI79, (I) IRN72_tbrz from Iran CMLV-CMS70, and (J) BEN68_59 from TATV_DAH68. ORF number and X denote highly polymorphic sites within or outside ORFs, respectively.

The encoded proteome. Knowing the putative proteins and their diversity provides a framework for gaining insight into the actual proteome, mechanisms of antiviral drugs, systems to model the human response to VARV infection, and methods of immunological detection and treatment. Therefore, to compare the isolates, we selected one isolate as the comparator, namely BSH75_banu, and determined the amino acid sequence identity of each putative protein of each isolate to their BSH75_banu ortholog. The resultant data (table S2) were substantial, so we selected 10 of the most divergent VARVs (tables S4 and S5) to illustrate (fig. S1, left panel) that, with few exceptions, no more than about four amino acids in a protein distinguish the isolates. Consistent with π^{rel} and π, the main differences are in proteins encoded by terminal CRS.

Phylogenetic inference. To gain insight into VARV evolution, we aligned a 65-kbp section (Fig. 2A, nucleotide positions ~56,000 to ~121,000) of the mid-CRS of the 45 isolates and reconstructed an unrooted tree (Fig. 3A), which assumes the molecular clock hypothesis that mutation rates are equivalent along all branches of a tree (*15*, *25*). In addition, because poxviruses, including VARVs, can recombine readily (*26*, *27*) and recombination can confound tree reconstruction and an understanding of lineage, we reconstructed a phylogram (Fig. 3B) rooted using two outgroup OPVs, CMLV-CMS70 and TATV-DAH68, to rule out the molecular clock hypothesis (*25*); the phylogram also includes 18 VARVs representing the unrooted tree.

The VARV isolates in both trees branch into three main clades (strains taxonomically), which we denoted A to C (Fig. 3A) and which cluster according to the isolate's origin—West Africa, South America, or Asia, respectively. However, the clade-C Asian VARV cluster contains branches to subclusters, that is, derivatives; mainly, these are isolates (variants taxonomically) from non–West-African African countries, and these variants segregate into viral types of low- or mid-range CFR. In Fig. 3A, we show several CFRs [most from reference (*1*)]; the geographic grouping and strain clustering coincide with low-, mid-, and high-range CFR values. Atypically, the high-range (~30%) CFR IND67_mah stands out as more divergent than other isolates from India. The phylogram confirms epidemiologic data (*1*), which indicates that the German, Yugoslavian, and British isolates we sequenced are due to importations of high-CFR Asian-origin VARVs into Europe.

The phylogenetic method, which enabled systematically organizing by genotype what had been somewhat unsystematically collected isolates, supports late smallpox-era proposals based on comparative phenotyping tests that three distinct VARV classes exist—major, intermediate, and minor. The fact that three CFR classes coincide respectively with the

main genotypic clusters of viral strains and that intermediate and minor variant African types derive from main Asian VARV cluster probably caused many of the inconsistent results reported when researchers tried to correlate CFRs with phenotyping results (*1*, *12*, *14*, *28*).

Both trees (Fig. 3, A and B) show an ancestral node that is closer to and equidistant from clade A and clade B relative to clade C. Essentially, the branches connecting clades A, B, and C superimpose whether or not the tree is rooted, therefore the assumption of equal mutation rates seems nondistortive and tentatively acceptable. To look for further clues to better understand the VARV lineage, which in Newick format is either ((A,B)C) or ((B,A)C), thereby leaving the antecedence of A relative to B unresolved, we noted that the sequences of the indel areas (Fig. 1) in clade A are identical to their counterparts in clade-B. Moreover, we noted that clade C contains truncated versions of these indels. Together, the identity and a deletion within each indel area suggested an investigation for gene loss, which is generally clocklike (*29*), and for recombination, which can cause gene gain or gene loss and thereby markedly influence mutation rates governing branch lengths, which represent sequence distances (*25*).

Gene loss. The total number of nucleotides comprising clade-A and clade-B CRS are similar to each other and greater than the number constituting clade-C CRS (table S1); therefore, in its more distant separation, clade C lost more sequences, which suggests that gene loss played a role in VARV evolution. Additionally, in clade A and clade B the left-end indel includes ORF-9, a truncated version of the VACV-C9L–like ankyrin-repeat protein, which resembles the myxomatosis poxvirus protein M150R, a potential antiinflammatory component that subverts nuclear-factor-κB activation (*30*). Lastly, the clade-A and clade-B right-end indel causes formation of ORFs 184.5, 185, and 185.7, which encode hypothetical proteins (fig. S1 and table S2). These ORFs precede ORF-186, a homolog of VACV ORF-B12R, which encodes a serine-threonine kinase–like protein that does not phosphorylate in reactions that activate another VACV kinase (*31*).

To inspect the sequences containing these ORFs, we stratified the hierarchy of the CRS of the 45 VARVs by using a clade-A or a clade-B genome sequence to query a database containing the 45 VARV genome sequences in addition to genome sequences of selected other OPV strains. The stratifications show some sequence losses have not been at all random, particularly two that are specific for the left and right indel areas (Fig. 1) in clade C (fig. S3, A and B). Depending on which clade is the query sequence, as expected, the hierarchy agrees with the lineage ((A,B)C) or ((B,A)C). To focus on the left-end deletion, we queried the data with a 2-kbp segment of DNA that includes ORF-9 from a clade-A strain. The results (fig. S3C) show the deletion and possible gene loss within the clade-C VARVs, the lineage ((A,B)C), and the putative sources of the insert, possibly acquired by direct descent or recombination. Of sequenced OPVs, ORF-9 might be from TATV-DAH68, CPXV90_gri, CMLV-CMS, CMLV-M96, or the like. The right-end indel (fig. S3D) shows a deletion and possible gene loss within the clade-C VARVs and a lineage ((A,B)C). This indel might be from CPXV-GER91 (*32*) or the like, although TATV-DAH68 and CMLV are potential sources.

When we queried with ORF-184.5-186 sequences from a clade-A VARV, the lineage was ((A,B)C) and the origin of the ORFs appeared to be CPXV-GER91 (fig. S3E). Moreover, ORF-186 showed a distinct area of high sequence identity and another of low identity, which might be relevant to gain or loss of protein-kinase function.

Recombination. To assess whether the inserts were a result of recombination, we explored distance matrixes of multiply aligned CRS by using split-decomposition analysis (*15*), which produces reticulate treelike network graphs if a genetic sequence distance matrix evidences recombination. The algorithm yielded reticulate trees when the full CRS of VARVs representing each clade constituted the alignment input (fig. S4A); however, the network decomposed into a standard phylogram when the alignment was restricted to the mid-CRS (fig. S4A, insert). A split decomposition analysis with different OPV species produced a similar result (fig. S4B and insert). Together, the results infer recombination within the extremities of the CRS and thereby suggest that the end CRS have a different evolutionary history from the mid CRS, which further suggests that poxvirus phylograms might not be exact because of the influence of

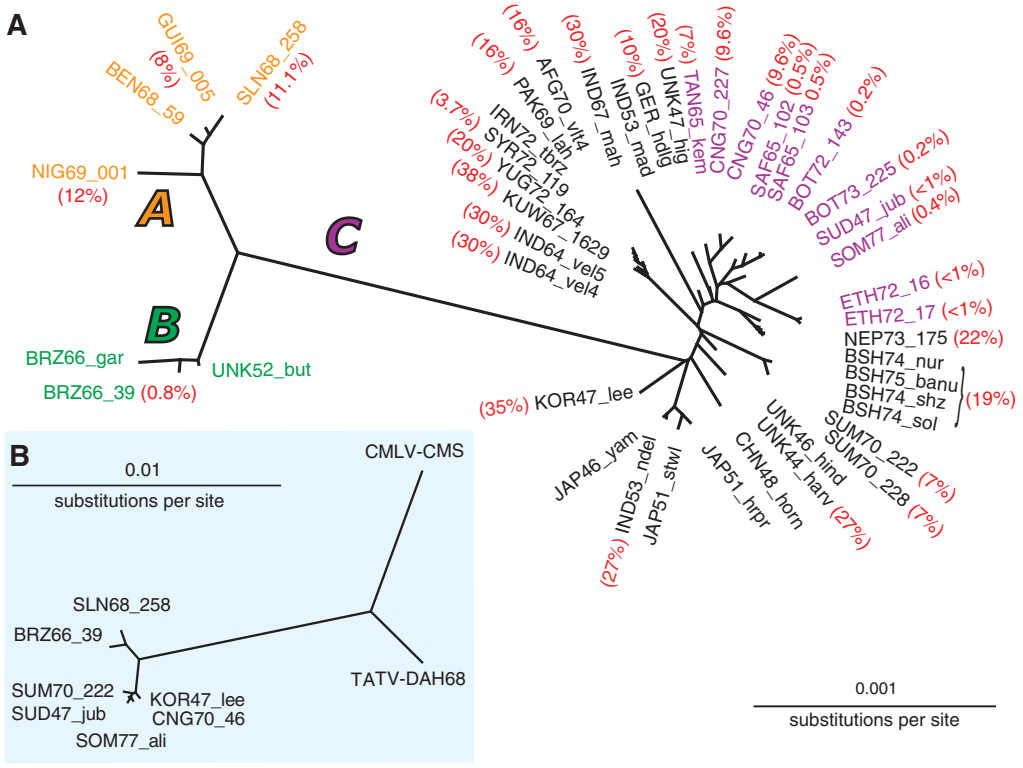

Fig. 3. Phylogenetic relationships. (**A**) An unrooted consensus phylogram from an alignment of 65 kbp of the mid-CRS of 45 VARVs reveals three high-level clades that represent clusters of isolates with origins in West Africa (clade A, orange), South America (clade B, green), and Asia (clade C, purple). The Asian clade contains a subgroup of non–West-African African variants (violet) that diverge into viral types of low- or midrange CFR. CFRs (red) associated with some isolates are indicated, and some of these are discussed in the supporting text (*15*). (**B**) A consensus tree rooted using CMLV-CMS70 and TATV-DAH68 mid-CRS aligned with a subset of VARVs representative of the tree in Fig. 3A.

recombination on mutation rates, and hence on tree branch lengths.

To investigate recombination further, we used sister-scanning assays (*15*) to quantify the significance (*Z* scores) of sequences being recombined. The algorithm slides a window of defined size across an alignment of two putative parents and a putative hybrid and essentially looks for patterns revealing sequence crossover. A significant likelihood of recombination was inferred by using a hypothetical alignment of the CRS of clade-A and clade-C VARVs as the parents and a clade-B VARV as the hybrid (fig. S5A). Similar results appeared using a clade-A VARV and CPXV-GER91 as the parents and a clade-B VARV as the hybrid (fig. S5B). Additional support for recombination became apparent by phylogenetic inference separately using the right-end indel recombined and flanking sequences (fig. S5C). Together the results in figures S3 to S5 suggest that evolution of the 45 VARVs involved gene gain by acquisition of the inserts, if indeed the ORFs represent genes and, if so, gene loss composed of the deletions in clade C compared with the other two clades.

Clade and encoded proteome relationships. Previous reports have compared genes of high-CFR members of clade C—BSH75_banu (CFR ~19%) and IND67_mah (CFR ~30%)—with each other and with genes of other OPVs, including VACV-COP and the low-CFR clade-A member BRZ66_gar (*7*, *8*, *10*, *33*–*36*). In the present study, we ascertained protein differences between low-range CFR clade-B and midrange CFR clade-A strains and between low- and midrange CFR clade-C variants from non–West-African Africa (Fig. 3A). The compilation used data from table S2, which contains the percentage identity between proteins of BSH75_banu and orthologs of the other VARVs. The amino acid sequence identity differences represent non-synonymous changes in codons that differentiate each ortholog from its correlate in BSH75_banu. The results reveal that a consensus of 67 putative ORFs distinguishes the group of the four clade-A strains from the group of the three clade-B strains and that a consensus of 15 putative ORFs distinguishes the mid- from the low-CFR groups of clade-C viral types from Africa (fig. S1, columns showing ORFs with non-synonymous nucleotide changes).

Discussion. In 1999, a U.S. Institute of Medicine report (*6*) advocated the scientific need for live VARV research to improve bioterrorism preparedness because infectious VARV samples might secretly exist (*5*). The report advised research to develop safe, efficacious antiviral drugs and a system for their assessment, and sequencing to refine the accuracy of VARV diagnostic tests and possibly provide additional scientific advances.

The low sequence diversity that we report here for a selection of isolates is reassuring and important from a biodefense perspective, because it suggests a high precision of differentiating VARVs if tracking single- or multi-source outbreaks. The ability to track the virus accurately might be a deterrent in its own right. The low diversity should also aid development of targeted, efficacious antiviral drugs and resolution of the actual proteome to help unravel the mechanism of action of the drug during infection of cell cultures or a model system. The low diversity of CRS indicates that the functional genome is not greatly varied, which improves the odds of understanding the proteome and reproducing smallpox in a system to assess countermeasures.

Treating the entire CRS as a seamless evolutionary unit oversimplifies reality. A main effecter of diversity in a population is natural selection, which causes genes within a genome to mutate. The relative diversity of each ORF (fig. S1, π^{rel}) and the percentage amino acid sequence identity maps (fig. S1, left panel) reveal that the proteins encoded within the terminal CRS vary the most, which indicates that selection pressure generally targeted these proteins the most. The pressure on the terminal regions of the proteome causing deletion and interruption probably drove the antecedent to become a devastating human pathogen.

Regarding the antecedent of the presently known VARVs, the OPVs most closely related to VARV are TATV_DAH68 (Fig. 2J) from a gerbil in Dahomey (Benin), Africa, and CMLV. The first credible physical evidence of smallpox is pockmarked Egyptian mummies that date to about 1500 BC (*1*). Given that there are missing links in the OPV lineage because the archived isolates are contemporary specimens and that there are various evolutionary possibilities, it is nevertheless tantalizing to speculate that gerbils and/or the sparse human population of ancient West Africa somehow spread smallpox into the more densely populated Fertile Crescent, perhaps when the Sudan-Sahel region across Africa was less arid. In addition, the appearance of alastrin might be related to the 18th-century slave trade of West African Yorubas tribesmen into Brazil.

Gene loss in general correlates with time (*29*), the inference being that if the ORFs in the indel areas represent real genes, then the encoded proteins might have been under negative selection, probably more by the human population in Asia than in Africa, as suggested by the often higher virulence of Asian VARVs in clade C (*25*, *29*). Increasing virulence reduces spread of a pathogen because, to the detriment of its own population, it kills off or maims the host population. Compared with the less virulent VARVs in clades A and B, the ORF-9 in clade C has a deletion, but deletions in the ortholog M150R are attenuating (*30*), which would constitute positive selection. However, there are examples like the interleukin-1β binding protein, which, if deleted, increases virulence of VACV (*20*).

Our study provides evidence for terminal region intergenomic recombination possibly within the species and with other OPVs (figs. S3 to S5). Considering the genome architecture (Fig. 1), one aspect of recombination that appears to be specific to VARV is the lack of intragenomic transposition in which genes from one end of the DNA appear duplicated and inverted at the opposite end of the DNA, probably through recombination events during replication, as described for other OPVs (*37*–*39*). We noted above that the production of site-specific, variable Ts and ATs might be due to DNA polymerase stuttering or intragenomic recombination. It is unknown whether recombination events in relation to the varied T and AT sites and the lack of VARV transposition have anything to do with each other.

Comparison of the amino acid percents identity between BSH75_banu and other VARV proteins (table S2) invited a determination of which proteins might influence virulence differences between low- and midrange CFR strains from clade A and clade B, respectively, and between clade-C African variant viral types of mid- or low-range CFR (fig. S1, candidate ORFs involved in virulence change). The ORFs of two clade-C Asian VARVs and the ORFs of these viruses and a clade-B virus have been compared previously (*10*, *33*). It is tempting to propose that the identified proteins alone modulate virulence; however, apart from these viral proteins, gene-control elements and yet undefined and undiscovered components could alter VARV innate virulence.

Our search for clues into the evolution of VARV is provocative, but it remains limited, partly because the repository represents a somewhat unsystematic collection. The present value of the viral stocks resides in their usefulness for understanding the VARV genome, the proteome, and the intracellular dynamics of infection, which will facilitate preparing for a natural, accidental, or deliberate release of VARV upon an unprotected world population.

References and Notes

1. F. Fenner, D. A. Henderson, I. Arita, Z. Jezek, I. D. Ladnyi, *Smallpox and Its Eradication* (World Health Organization, Geneva, 1988).
2. World Health Assembly, *Resolution WHA 55.15* (2002).
3. World Health Organization, *Wkly. Epidemiol. Rec.* **74**, 188 (1999).
4. World Health Organization, *Wkly. Epidemiol. Rec.* **77**, 34 (2002).
5. D. A. Henderson *et al.*, *JAMA* **281**, 2127 (1999).
6. Institute of Medicine, *Assessment of Future Scientific Needs for Live Variola Virus* (National Academy Press, Washington, DC, 1999).
7. R. F. Massung *et al.*, *Nature* **366**, 748 (1993).
8. R. F. Massung *et al.*, *Virology* **201**, 215 (1994).
9. S. N. Shchelkunov *et al.*, *Dokl. Akad. Nauk* **328**, 629 (1993).
10. S. N. Shchelkunov *et al.*, *Virology* **266**, 361 (2000).
11. M. Nizamuddin, K. R. Dumbell, *Lancet* **1**, 68 (1961).
12. E. Shafa, *WHO/SE/72. 35. Case-Fatality Ratios in Smallpox* (1972).
13. K. R. Dumbell, F. Huq, *Trans. R. Soc. Trop. Med. Hyg.* **69**, 303 (1975).

14. F. Huq, K. R. Dumbell, *Trans. R. Soc. Trop. Med. Hyg.* **97**, 97 (2003).
15. Materials and Methods are available as supporting material on *Science* Online.
16. R. F. Massung *et al.*, *Virology* **221**, 291 (1996).
17. R. F. Massung, J. C. Knight, J. J. Esposito, *Virology* **211**, 350 (1995).
18. B. Moss, *Poxviridae: The Viruses and Their Replication*, in *Fields Virology*, D. M. Knipe *et al.*, Eds. (Lippincott, Williams, Wilkins, Philadelphia, 2001), vol. 2, pp. 2849–2883.
19. M. Pires de Miranda, P. C. Reading, D. C. Tscharke, B. J. Murphy, G. L. Smith, *J. Gen. Virol.* **84**, 2459 (2003).
20. A. Alcami, G. L. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 11029 (1996).
21. National Research Council, *Biotechnology Research in an Age of Terrorism* (National Academies Press, Washington, DC, 2004).
22. P. Balbas, G. Gosset, *Mol. Biotechnol.* **19**, 1 (2001).
23. A. Domi, B. Moss, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12415 (2002).
24. J. Tian *et al.*, *Nature* **432**, 1050 (2004).
25. M. Salemi, A.-M. Vandamme, *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny* (Cambridge Univ. Press, Cambridge, UK, 2003).
26. H. S. Bedson, K. R. Dumbell, *J. Hyg. (Lond.)* **62**, 147 (1964).
27. M. D. Hamilton, D. H. Evans, *Nucleic Acids Res.* **33**, 2259 (2005).
28. K. R. Dumbell, F. Huq, *Am. J. Epidemiol.* **123**, 403 (1986).
29. B. Snel, P. Bork, M. A. Huynen, *Genome Res.* **12**, 17 (2002).
30. C. Camus-Bouclainville *et al.*, *J. Virol.* **78**, 2510 (2004).
31. A. H. Banham, G. L. Smith, *J. Gen. Virol.* **74**, 2807 (1993).
32. H. Meyer, H. Neubauer, M. Pfeffer, *J. Vet. Med. B Infect. Dis. Vet. Public Health* **49**, 17 (2002).
33. S. N. Shchelkunov, R. F. Massung, J. J. Esposito, *Virus Res.* **36**, 107 (1995).
34. S. N. Shchelkunov, V. M. Blinov, S. M. Resenchuk, A. V. Totmenin, L. S. Sandakhchiev, *Virus Res.* **30**, 239 (1993).
35. S. N. Shchelkunov *et al.*, *Virus Res.* **27**, 25 (1993).
36. S. N. Shchelkunov, S. M. Resenchuk, A. V. Totmenin, V. M. Blinov, L. S. Sandakhchiev, *Virus Res.* **32**, 37 (1994).
37. R. W. Moyer, R. L. Graves, C. T. Rothe, *Cell* **22**, 545 (1980).
38. J. J. Esposito, C. D. Cabradilla, J. H. Nakano, J. F. Obijeski, *Virology* **109**, 231 (1981).
39. D. J. Pickup, B. S. Ink, B. L. Parsons, W. Hu, W. K. Joklik, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6817 (1984).
40. J. J. Esposito, J. C. Knight, *Virology* **143**, 230 (1985).
41. K. R. Dumbell, L. Harper, A. Buchan, N. J. Douglass, H. S. Bedson, *Epidemiol. Infect.* **122**, 287 (1999).
42. We thank J. L. Patton, M. Khristova, D. Carroll, S. Reeder, R. Morey, and R. Galloway for expert technical assistance; C. Chesley for grammatical advice; the Poxvirus Bioinformatics Resource Center (www.poxvirus.org and www.biovirus.org) for poxvirus orthologous clusters (POCs) and viral orthologous clusters (VOCs) source codes; J. Rozas for DnaSP source code; A. J. Gibbs, M. J. Gibbs, and J. S. Armstrong for SiScan; and K. R. Dumbell and F. Fenner for smallpox history information. D.G. is funded by Atlanta Research and Education Foundation, Decatur, GA. Sequence data are deposited under GenBank accession numbers AY313847, DQ437580-DQ437594, and DQ441416-DQ441448.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1125134/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Tables S1 to S6
References

19 January 2006; accepted 30 May 2006
Published online 27 July 2006;
10.1126/science.1125134
Include this information when citing this paper.

REPORTS

Magnetic Fields in the Formation of Sun-Like Stars

Josep M. Girart,[1]* Ramprasad Rao,[2,3] Daniel P. Marrone[2]

We report high-angular-resolution measurements of polarized dust emission toward the low-mass protostellar system NGC 1333 IRAS 4A. We show that in this system the observed magnetic field morphology is in agreement with the standard theoretical models of the formation of Sun-like stars in magnetized molecular clouds at scales of a few hundred astronomical units; gravity has overcome magnetic support, and the magnetic field traces a clear hourglass shape. The magnetic field is substantially more important than turbulence in the evolution of the system, and the initial misalignment of the magnetic and spin axes may have been important in the formation of the binary system.

Magnetic fields are believed to play a crucial role in the formation of stars (*1*, *2*). In the standard model of isolated low-mass-star formation (*3*, *4*), magnetized molecular clouds that are magnetically supported against gravitational collapse (or "subcritical") are expected to slowly form dense molecular cores through ambipolar diffusion. The neutral particles are only weakly coupled to the ions, which couple to the magnetic fields, and can drift toward the center of the cloud. The increasing central mass eventually overcomes the magnetic support and the now supercritical core collapses gravitationally. In this collapse phase, the initially uniform magnetic field is warped and strengthened in the core. The magnetic field is expected to assume an hourglass shape, with an accretion disk formed at the central "pinch" in the field, which corresponds, for an initial cloud size on the order of 10^4 to 10^5 astronomical units (AU), to scales of about 100 AU. At large scales, where the contraction effect is small the magnetic field lines are essentially straight. In this axisymmetric scenario, the process of magnetic braking, which forces the core to rotate at the same angular speed as the envelope, prevents fragmentation of the collapsing core and the formation of multiple stellar systems. However, models with nonaxisymmetric perturbations show that fragmentation can occur on a broad range of scales ($\leq$100 to 10^4 AU), depending on the initial conditions, such as magnetic field strength, rotation rate, and cloud mass–to–Jeans mass ratio (*5*).

The polarization of dust continuum emission provides an opportunity to examine the magnetic field configuration in star-forming regions (*6*). Aspherical spinning dust particles preferentially align themselves with the rotational axis (minor axis) parallel to the direction of the magnetic field. The emission from such grains is partially linearly polarized, with the observed polarization angle perpendicular to the direction of the magnetic field.

NGC 1333 IRAS 4A is a well-studied binary protostellar system in the Perseus molecular cloud complex (*7*). The two protostars, IRAS 4A1 and A2, are associated with molecular outflows directed roughly north-south (*8*). The distance to this cloud complex from Earth is thought to be between 220 and 350 parsecs (pc) (*9*); we adopt the value of 300 pc here. The Perseus complex is an active star-forming region with around 20 young stellar objects within a projected radius of 4×10^4 AU from IRAS 4A. Early polarimetric observations of IRAS 4A (*10*, *11*) did not have enough angular resolution to be able to examine the spatial scales relevant to the putative hourglass. Yet, in the highest angular resolution observations to date, polarimetry at 230 GHz with the Berkeley-Illinois-Maryland Association (BIMA) array showed hints of an hourglass shape in the magnetic field on 3.5″ (1000 AU) scales (*12*).

The Submillimeter Array (SMA) is the first imaging submillimeter interferometer (*13*, *14*), providing arcsecond angular resolution and good continuum sensitivity at frequencies that are higher than those currently observable with any other radio interferometer. We used the SMA to observe NGC 1333 IRAS 4A at 345 GHz at an angular resolution of 1.56″ $\times$ 0.99″ [with a position angle (PA) of 85°]. Our data resolve the continuum peaks of IRAS 4A1 and

[1]Institut de Ciències de l'Espai (CSIC- IEEC), Campus UAB–Facultat de Ciències, Torre C5–Parell 2ª, Bellaterra, Catalunya 08193, Spain. [2]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA. [3]Academia Sinica, Institute of Astronomy and Astrophysics, 645 North Aohoku Place, Hilo, HI 96720, USA.

*To whom correspondence should be addressed. E-mail: girart@ieec.uab.es