Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool

Thomas Hoenen, Allison Groseth, Kyle Rosenke,
Robert J. Fischer, Andreas Hoenen,
Seth D. Judson, Cynthia Martellaro,
Darryl Falzarano, Andrea Marzi,
R. Burke Squires, Kurt R. Wollenberg,
Emmie de Wit, Joseph Prescott, David Safronetz,
Neeltje van Doremalen, Trenton Bushmaker,
Friederike Feldmann, Kristin McNally,
Fatorma K. Bolay, Barry Fields, Tara Sealy,
Mark Rayfield, Stuart T. Nichol, Kathryn C. Zoon,
Moses Massaquoi, Vincent J. Munster,
Heinz Feldmann

Rapid sequencing of RNA/DNA from pathogen samples obtained during disease outbreaks provides critical scientific and public health information. However, challenges exist for exporting samples to laboratories or establishing conventional sequencers in remote outbreak regions. We successfully used a novel, pocket-sized nanopore sequencer at a field diagnostic laboratory in Liberia during the current Ebola virus outbreak.

Disease outbreaks in resource-limited or remote areas pose unique challenges to outbreak responses. These challenges are exemplified by the ongoing Ebola virus (EBOV) outbreak in West Africa that began in 2014 (1) and is unprecedented in its size and duration. Correspondingly, the magnitude of the international response, encompassing ≈50 Ebola treatment units (ETUs) and >2 dozen diagnostic laboratories, has been equally unprecedented.

Author affiliations: Friedrich-Loeffler-Institut, Greifswald–Insel Riems, Germany (T. Hoenen, A. Groseth); National Institutes of Health, Hamilton, Montana, USA (T. Hoenen, A. Groseth, K. Rosenke, R.J. Fischer, S.D. Judson, C. Martellaro, D. Falzarano, A. Marzi, E. de Wit, J. Prescott, D. Safronetz, N. van Doremalen, T. Bushmaker, F. Feldmann, K. McNally, V.J. Munster, H. Feldmann); Independent scholar, Aachen, Germany (A. Hoenen); University of Saskatchewan, Saskatoon, Saskatchewan, Canada (D. Falzarano); National Institutes of Health, Bethesda, Maryland, USA (R.B. Squires, K.R. Wollenberg, K.C. Zoon); The Liberian Institute for Biomedical Research, Charles Ville, Republic of Liberia (F.K. Bolay); Centers for Disease Control and Prevention, Atlanta, Georgia, USA (B. Fields, T. Sealy, M. Rayfield, S.T. Nichol); Ministry of Health and Social Welfare, Monrovia, Republic of Liberia (M. Massaquoi)

These laboratories often are operated under improvised field conditions to keep them close to active, sometimes remote transmission sites (2,3).

Rapidly obtaining genome sequences during disease outbreaks is crucial for clarifying patterns of virus evolution, monitoring the validity of diagnostic assays, and investigating transmission chains (4,5). Further, rapid results may help determine the efficacy of sequence-dependent countermeasures, such as siRNAs or antibody treatments. In the past, obtaining timely genome sequences has been difficult because of political and logistical obstacles that limited the export of samples to laboratories capable of performing these analyses. As an example, during the first year of the outbreak in West Africa, only 2 reports of genome sequences from patients were published (1,6). Similarly, establishing conventional Sanger or next-generation sequencing technologies in affected countries is logistically challenging because of the size and weight (≈40 to ≈100 kg) of the necessary equipment, the high potential for transport damage related to the sensitive optics many of these machines incorporate, limitations on supportive infrastructure, and complex sample processing procedures. An additional challenge is the required installation or calibration of sequencing machines, which often has to be done by field engineers employed by the manufacturers, who may be reluctant to send their employees into outbreak areas. However, Kugelman et al. recently reported the successful deployment of an Illumina MiSeq, a well-established, conventional next-generation sequencing platform (Illumina, San Diego, CA, USA), to West Africa; the platform became operational in February 2015 (5).

Seeking a platform that would be more rapidly deployable and reliable under field conditions, we established protocols and evaluated the feasibility of nanopore sequencing technology under outbreak conditions using a pocket-sized (≈10 × 4 × 2 cm, 75 g) MinION sequencing device (Oxford Nanopore Technologies, [https://www.nanoporetech.com/]). Because of its small size, this device can easily be transported into remote locations; furthermore, it requires no special setup or calibration procedures and can be operational immediately after arrival in an outbreak area. Further, data turnaround is very rapid, and consequently, nanopore sequencing is being developed as a rapid diagnostic tool for management of outbreaks of various diseases (7,8). The MinION device senses individual DNA molecules based on modulation of ion currents across nanopores as the molecules are passing through. These modulations are

dependent on the physical properties of the nucleotides and allow determination of the nucleotide sequence (*9*).

The Study

To facilitate sequencing of the RNA genome of EBOV, we developed and tested an approach based on reverse transcription PCR, in which whole virus genomes were amplified in overlapping fragments (Figure 1, panels A, B; online Technical Appendix, http://wwwnc.cdc.gov/EID/article/22/2/15-1796-Techapp1.pdf). This approach was first validated in a regular laboratory setting in the Rocky Mountain Laboratories of the National Institutes of Health (NIH) by using blood samples from nonhuman primates experimentally inoculated with EBOV strain Makona-Gueckedou-C07 (*10,11*). This validation showed that sequencing information was obtainable for the complete genome with an average of 7,038 reads at every nucleotide position (read depth; online Technical Appendix Figure 1, panel A). We observed no sequence differences when comparing the consensus sequence derived from these data to those obtained by using Sanger sequencing (online Technical Appendix Figure 1, panel B). Furthermore, by analyzing linearized plasmid DNA of known sequence, we

established the accuracy of the MinION device as ≈84% for a single read (online Technical Appendix Figure 1, panels C, D). On the basis of this information, and the fact that read depth can compensate for miscalled nucleotides in individual reads by piling up reads covering the same region, we determined the theoretical probability for ≥1 miscalled base (TPMB) in a complete MinION-sequenced EBOV genome to be <5% when the read depth is ≥33 at all positions (online Technical Appendix Figure 1, panels E and F).

After having validated this approach, MinION devices were taken to the Centers for Disease Control and Prevention (CDC)/NIH field laboratory that provided diagnostic support for ETUs in Monrovia, Liberia, during August 2014–May 2015. All equipment and reagents necessary for sequencing could be easily transported as checked luggage by a single person on a commercial carrier. In Liberia, temperatures in the laboratory area used for sequencing ranged from 28 to 32°C, necessitating the use of an improvised heat sink for the devices, which consisted of a metal plate of ≈30 × 30 cm (online Technical Appendix Figure 2, panels A, B). Under field conditions, we initially failed to produce complete genomes with high confidence because of problems with
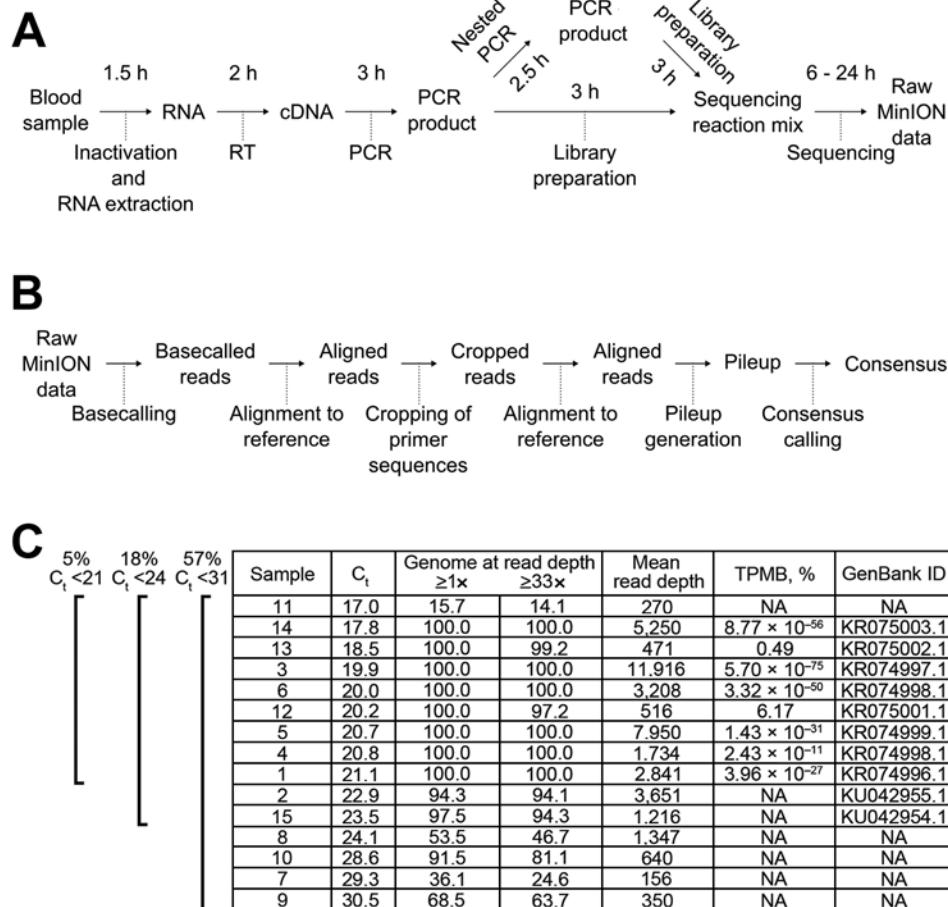
Figure 1. MinION sequencing. A) Experimental and B) bioinformatics workflows. Times indicated are the approximate duration for each procedure. RT, reverse transcription. C) Sequencing results showing Ebola virus load (expressed as C_t value), percentage of the genome with a minimum read depth of ≥1 or ≥33, mean read depth, theoretical probability for a miscalled base (TPMB), and GenBank accession numbers of complete and nearly complete genomes. Brackets at left indicate percentage of Ebola virus–positive patient samples below each of the 3 cutoff cycle threshold (C_t) values used in this study (C_t <21, <24, <31). Sample 8 was from an oral swab; all others were from blood. NA, not available.

Sample	C_t	Genome at read depth ≥1×	≥33×	Mean read depth	TPMB, %	GenBank ID
11	17.0	15.7	14.1	270	NA	NA
14	17.8	100.0	100.0	5,250	8.77×10^{-56}	KR075003.1
13	18.5	100.0	99.2	471	0.49	KR075002.1
3	19.9	100.0	100.0	11.916	5.70×10^{-75}	KR074997.1
6	20.0	100.0	100.0	3,208	3.32×10^{-50}	KR074998.1
12	20.2	100.0	97.2	516	6.17	KR075001.1
5	20.7	100.0	100.0	7.950	1.43×10^{-31}	KR074999.1
4	20.8	100.0	100.0	1.734	2.43×10^{-11}	KR074998.1
1	21.1	100.0	100.0	2.841	3.96×10^{-27}	KR074996.1
2	22.9	94.3	94.1	3,651	NA	KU042955.1
15	23.5	97.5	94.3	1.216	NA	KU042954.1
8	24.1	53.5	46.7	1,347	NA	NA
10	28.6	91.5	81.1	640	NA	NA
7	29.3	36.1	24.6	156	NA	NA
9	30.5	68.5	63.7	350	NA	NA

C) 5% C_t <21   18% C_t <24   57% C_t <31

PCR yields (online Technical Appendix Figure 3, panels A, B). However, by implementing a second PCR step, we circumvented this problem and obtained high quality complete genome sequences for 8 of 9 high-virus load samples (cycle threshold ≤21) (Figure 1, panel C; online Technical Appendix Figure 4, panel A). In lower virus load samples, we could obtain only incomplete genome sequences; however, even in those samples regions for which sequencing information was available generally showed high read depths (online Technical Appendix Figure 4, panel B), suggesting that further optimization of PCRs might also allow complete coverage for these samples. Furthermore, even incomplete genome sequences can provide valuable information during an outbreak, allowing analysis of individual genes and the tracing of transmission chains (*12*).

Using this updated protocol, we achieve a sustained capacity of 4 full-length genomes per day for a single person conducting the laboratory work using 2 MinION devices (Figure 1, panels A, B). However, with the exception of the first 2 sequencing runs, bioinformatics analysis during this mission was mainly completed after returning to the NIH, to maximize the time for raw data acquisition (online Technical Appendix).

Phylogenetic analysis of the complete genomes generated in Monrovia, Liberia, showed them being clearly distinct from Sierra Leone or early Guinea sequences of EBOV-Makona (online Technical Appendix Figure 5) but clustering well with all other sequences found in samples from Liberia. These results suggest that EBOV in Liberia resulted from a single introduction or a limited number of introductions with genetically similar viruses. When analyzing the obtained full-length sequences and comparing them to a consensus sequence from the outbreak (*13*), we observed few mutations, most in noncoding regions or synonymous mutations (Figure 2); none affected siRNA target sequences or the diagnostic targets used in the CDC/NIH laboratory.

Using Bayesian analysis including these sequences, we estimated the nucleotide substitution rate during the outbreak at 1.36×10^{-3}, consistent with recently published values (*5,13–15*). In a root-to-tip-analysis, the sequences we obtained showed substitution rates comparable to other sequences from the outbreak (online Technical Appendix Figure 6). Overall, these data suggest that EBOV has remained relatively stable genetically during the outbreak.
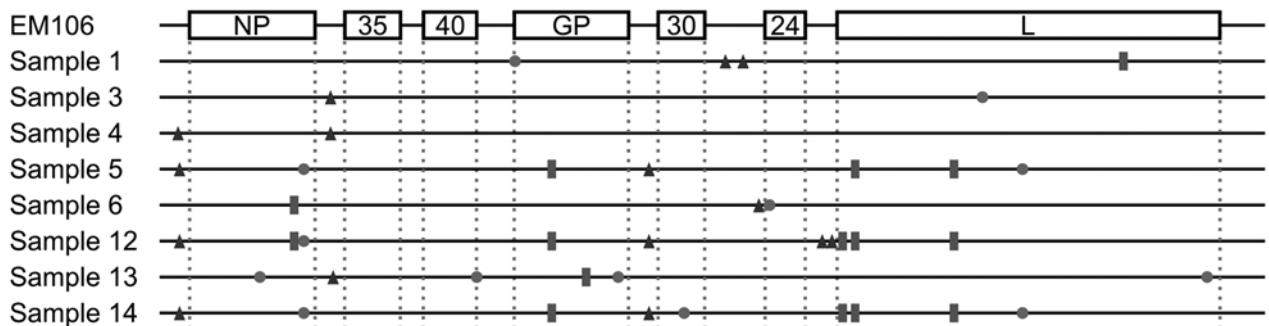
Conclusions

We found that, because of the device's small size and comparatively modest resource requirements, nanopore sequencing has tremendous potential for use in remote and resource-limited areas, and its implementation could revolutionize the capacity of public health professionals to perform sequencing during future disease outbreaks. Although we used a directed approach to sequencing, approaches not dependent on prior pathogen identification (i.e. for diagnostic use of the MinION device) are currently being developed (*7*) and will even further increase this technology's usefulness in future outbreaks.

Figure 2. Observed mutations in the 8 fully nanopore-sequenced Ebola-positive blood samples compared to a reference sequence from June 2014 (SLI/Makona-EM106, GenBank accession number KM233036.1). Squares indicate nonsynonymous mutations, circles indicate synonymous changes, and triangles indicate changes in noncoding regions. NP, nucleoprotein; GP, glycoprotein.

T.H., A.G., K.R., and J.P. are participants in the ONT MinION Access Programme, and received some of the MinION devices and Flow Cells used for this study free of charge or at reduced cost. T.H. was invited by Oxford Nanopore Technologies to present part of this work at the London Calling meeting in London, UK. The deidentified patient specimens used in this manuscript were collected as part of public health surveillance and not as human subjects research. The NIH Office of Human Subjects Research has determined that federal regulations for the protection of human subjects do not apply.

Dr. Hoenen was a visiting postdoctoral fellow at the Laboratory of Virology of the National Institutes of Health in Hamilton, Montana, USA, for major phases of this study, and has recently joined the Friedrich-Loeffler-Institut on the island of Riems, near Greifswald, Germany, where he is head of the Laboratory for Molecular Biology of Filoviruses. His research focuses on the molecular biology of filoviruses and their interactions with host cells, with the ultimate goal of finding new antivirals against filoviruses and other hemorrhagic fever viruses.

References

1. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of Zaire Ebola virus disease in Guinea. N Engl J Med. 2014;371:1418–25. http://dx.doi.org/10.1056/NEJMoa1404505
2. Grolla A, Jones S, Kobinger G, Sprecher A, Girard G, Yao M, et al. Flexibility of mobile laboratory unit in support of patient management during the 2007 Ebola-Zaire outbreak in the Democratic Republic of Congo. Zoonoses Public Health. 2012;59 (Suppl 2):151–7. http://dx.doi.org/10.1111/j.1863-2378.2012.01477.x
3. Allaranga Y, Kone ML, Formenty P, Libama F, Boumandouki P, Woodfill CJ, et al. Lessons learned during active epidemiological surveillance of Ebola and Marburg viral hemorrhagic fever epidemics in Africa. East Afr J Public Health. 2010;7:30–6.
4. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, et al. Molecular Evidence of Sexual Transmission of Ebola Virus. N Engl J Med. 2015. http://dx.doi.org/10.1056/NEJMoa1509773
5. Kugelman JR, Wiley MR, Mate S, Ladner JT, Beitzel B, Fakoli L, et al. Monitoring of Ebola Virus Makona Evolution through Establishment of Advanced Genomic Capability in Liberia. Emerg Infect Dis. 2015;21:1135–43. http://dx.doi.org/10.3201/eid2107.150522
6. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345: 1369–72. http://dx.doi.org/10.1126/science.1259657
7. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 2015;7(1):99. http://dx.doi.org/10.1186/s13073-015-0220-9
8. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biol. 2015;16:114. http://dx.doi.org/10.1186/s13059-015-0677-2
9. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol. 2015;33:296–300. http://dx.doi.org/10.1038/nbt.3103
10. Hoenen T, Groseth A, Feldmann F, Marzi A, Ebihara H, Kobinger G, et al. Complete genome sequences of three Ebola virus isolates from the 2014 outbreak in west Africa. Genome Announc. 2014;2:e01331–14. http://dx.doi.org/10.1128/genomeA.01331-14
11. Marzi A, Feldmann F, Hanley PW, Scott DP, Gunther S, Feldmann H. Delayed disease progression in cynomolgus macaques infected with Ebola virus Makona strain. Emerg Infect Dis. 2015;21:1777–83. http://dx.doi.org/10.3201/eid2110.150259
12. Christie A, Davies-Wayne GJ, Cordier-Lasalle T, Blackley DJ, Laney AS, Williams DE, et al. Possible sexual transmission of Ebola virus – Liberia, 2015. MMWR Morb Mortal Wkly Rep. 2015;64:479–81.
13. Hoenen T, Safronetz D, Groseth A, Wollenberg KR, Koita OA, Diarra B, et al. Virology. Mutation rate and genotype variation of Ebola virus from Mali case sequences. Science. 2015;348:117–9. http://dx.doi.org/10.1126/science.aaa5646
14. Tong YG, Shi WF, Liu D, Qian J, Liang L, Bo XC, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015;524:93–6. http://dx.doi.org/10.1038/nature14490
15. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. Nature. 2015;524:97–101. http://dx.doi.org/10.1038/nature14594

Address for correspondence: Thomas Hoenen, Friedrich-Loeffler-Institut, Südufer 10, 17493 Greifswald–Insel Riems, Germany; email: thomas.hoenen@fli.bund.de

Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool

Technical Appendix

Detailed Materials and Methods

Blood and Virus Samples

Deidentified frozen blood samples or viral transport medium (VTM; from oral swabs) from deceased patients that had tested positive for Ebola virus between November 2014 and January 2015 were thawed and 140 µL of blood or VTM was inactivated in a portable glovebox at the diagnostic laboratory on the ELWA campus in Liberia, Monrovia. RNA was then extracted by using the QIAamp viral RNA mini kit (QIAGEN, Valencia, CA, USA) following the manufacturer's instructions, but with 2 additional AW1 wash steps.

RT-PCRs and Cleanup

To avoid any potential impact on the diagnostic services, all work related to sequencing was performed in a physically separate dedicated area by personnel not involved in diagnostics. Two-step RT-PCRs were performed as previously described (*1,2*), but using 5 µL of RNA instead of 1 µL of RNA as starting material. For the RT primer this sequence was used: CGGACACACAAAAAGAAAGAAG. For the second PCRs, 0.8 µL of the PCR product from the first PCR was used as input material without further purification (each PCR product of the first PCR step served as template for 2-second PCR reactions). Primer pairs and cycling conditions are provided in Technical Appendix Tables 1 and 2. After the second PCR, DNA was purified by using Agencourt AMPure XP PCR purification beads (Beckman Coulter, Krefeld, Germany). All 24 RT-PCR products from 1 sample were purified in a total of 2 purification reactions: 40 µL of RT-PCR product from each of the 12 reactions was pooled, and 720 µL of Agencourt beads was added. Samples were incubated for 5 minutes and then placed on a magnetic Agencourt SPRIStand (Beckman Coulter) for 5 minutes. The supernatant was carefully aspirated off the pellet, leaving 10–20 µL of supernatant on the pellet. Then 1.2 mL of 70% ethanol was added to the samples without disturbing the pellet, and they were incubated on the

magnet for 30 seconds. The supernatant was then aspirated completely, and a second wash was performed with 800 µL of ethanol. After the second wash step, the tubes were briefly (≈3 seconds) spun down in a tabletop centrifuge, placed back onto the magnet, and any additional supernatant was removed by using a P10 pipette. The pellets were air-dried on the magnet for ≈2 minutes, before being removed from the magnet and carefully resuspended in 60 µL of EB buffer (QIAGEN). Samples were incubated for 10 minutes, then they were placed back onto the magnet, incubated for 1 minute to allow beads to pellet, and 50 µL of supernatant was carefully removed. The eluates from the 2 purifications (corresponding to 1 patient sample) were then pooled and used for library preparation and sequencing. For visualization, 5 µL of pooled PCR products were run on a 1% agarose gel, stained for 3 minutes in 100× FastBlast (BioRad), rinsed and destained in water, and documented by using an IPhone 4 with a white laptop screen serving as a light box.

MinION Runs

Library preparation was done by using the Genomic DNA Sequencing Kit SQK-MAP004 (Oxford Nanopore Technologies, Oxford, UK [ONT]) following the manufacturer's instructions. Samples were analyzed on a MinION sequencing device using R7.3 FlowCells (ONT) connected to a laptop running the MinKNOW software 0.48.2.12 (ONT). Internet connectivity was provided through a cellular network (Novafone Inc., Monrovia, Liberia) by using a wireless 4G router. Base calling was done by using the ONT Metrichor software version 2.25.1. Due to restrictions in personnel (i.e., the fact that most of the sequencing work in Liberia was done by a single person) and the need to perform the base calling by using a cloud-server, which required upload of the primary data via a 4G cellular network in Liberia, the bioinformatics aspect of the work was done on-site only for the first 2 sequencing runs. After we had demonstrated that this is in principle feasible, we decided to save the remaining raw data temporarily on a portable hard drive, and base calling and the subsequent bioinformatics analysis was done after return to the Rocky Mountain Laboratories, to maximize the generation of raw data. Base calling algorithms that can be run locally, without the need for an internet connection, are currently being developed, and by including a person dedicated to the bioinformatics work on future outbreak missions it should be possible to do this aspect of workflow rapidly on-site, with the same capacity than the laboratory work.

Consensus Calling

All consensus calling was done by using an Ubuntu 14.04 linux environment running under Oracle VM VirtualBox 4.3.20 (Oracle.com). FASTA sequences were extracted from the fast5 files returned by Metrichor by using Poretools 0.5.1 (*3*), and were aligned to a consensus of sequences previously observed in the West African outbreak (*2*) by using lastal 393 (*4*). In this alignment, sequences corresponding to primer sequences incorporated in the PCR product were identified and cropped by using last2fasta.pl (see Bioinformatics Scripts section), and the resulting FASTA file with cropped sequences was realigned by using lastal. The resulting alignment was converted into a SAM file and a pileup was constructed by using Samtools 0.1.19 (*5*). Nucleotide counts for each position were extracted from the pileup by using pileup2nucl.pl, and the consensus sequence was identified by using callnucl.pl.

Calculation of a Theoretical Probability for a Miscalled Base

To estimate the effect of read-depth on the overall reliability of the data, a theoretical probability for a miscalled base (TPMB) was calculated. This value was based on the observation that in our hands plasmid DNA with a known sequence could be sequenced with an accuracy of 84.13%. For a single nucleotide position the TPMB was then calculated for a read-depth from 1 to 170 as the sum of probabilities for at least 50% miscalls using the binomial probability formula (results of these calculations are shown in Technical Appendix Figure 1 panel E). For read depths greater than 170, the TPMB was approximated using regression analysis of these data, as shown in Technical Appendix Figure 1, panel E. For calculating the TPMB across a whole EBOV genome, TPMBs of 2 neighboring nucleotides t_1 and t_2 were combined as $t_{1,2} = (t_1 \times [1-t_2]) + ([1-t_1] \times t_2) + (t_1 \times t_2)$, to give the probability that at ≥ 1 base in the dinucleotide is being miscalled. Then, TPMBs of 2 neighboring dinucleotides were combined in the same fashion, and this process was continued for increasingly larger fragments of the genome, until TPMBs of all nucleotides across the whole length of the genome had been considered. Factors that were not taken into consideration in this estimation were errors introduced by the PCR-amplification steps, or the possibility of a non-random distribution of errors. However, it has to be noted that no obvious non-random distribution of errors in our sequencing data was observed.

Phylogenetic Analysis

For Bayesian coalescent phylogeny, 296 Ebola virus cDNA genomes were aligned by using ClustalX2 (*6*) and this alignment was inspected and manually improved. Sample collection

dates were added to the sequence identifiers to allow serial coalescent analysis. This multiple sequence alignment was input into BEAST v1.8.2 (*7*) to calculate a Bayesian coalescent phylogeny. For this analysis we used the HKY substitution model (*8*), a lognormal relaxed uncorrelated clock model (*9*), and the Bayesian skygrid tree model (*10*). For the clock model the CTMC rate reference prior (*11*) was specified. Four independent 40 million generation runs were performed, of which 3 ran to completion without fatal errors. The 3 successful runs converged to roughly identical parameter estimates so the run with the highest ESS values was used to estimate the phylogeny. To estimate the phylogeny the first 10% of the MCMC samples were discarded as burn-in and a maximum clade credibility tree was derived from the remaining 9001 trees.

For Root-To-Tip analysis, Bayesian analysis was performed by using MrBayes 3.2.5 (http://mrbayes.sourceforge.net/), with a with general time-reversable substitution model with a percent of site invariant and gamma-distributed rate heterogeneity across sites. The analysis was terminated after 2,028,000 generations, as it was determined that it had converged. Root-to-tip distances were calculated by using TreeStat v1.8.2 (http://tree.bio.ed.ac.uk/software/treestat/), and the scatter plot and linear regression were performed in R v3.2.2.

References

<jrn>1. Hoenen T, Groseth A, Feldmann F, Marzi A, Ebihara H, Kobinger G, et al. Complete genome sequences of three ebola virus isolates from the 2014 outbreak in west Africa. Genome Announc. 2014;2:e01331–14. **PMID: 25523781** http://dx.doi.org/10.1128/genomeA.0331-14</jrn>

<jrn>2. Hoenen T, Safronetz D, Groseth A, Wollenberg KR, Koita OA, Diarra B, et al. Virology. Mutation rate and genotype variation of Ebola virus from Mali case sequences. Science. 2015;348:117–9.**PMID: 25814067** http://dx.doi.org/10.1126/science.aaa5646</jrn>

<jrn>3. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics. 2014;30:3399–401.**PMID: 25143291** http://dx.doi.org/10.1093/bioinformatics/btu555</jrn>

<jrn>4. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21:487–93.**PMID: 21209072** http://dx.doi.org/10.1101/gr.113985.110</jrn>

<jrn>5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.**PMID: 19505943** http://dx.doi.org/10.1093/bioinformatics/btp352</jrn>

<jrn>6. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–8. **PMID: 17846036** http://dx.doi.org/10.1093/bioinformatics/btm404</jrn>

<jrn>7. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73. **PMID: 22367748** http://dx.doi.org/10.1093/molbev/mss075</jrn>

<jrn>8. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22:160–74. **PMID: 3934395** </jrn>

<jrn>9. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4:e88. **PMID: 16683862** http://dx.doi.org/10.1371/journal.pbio.0040088</jrn>

<jrn>10. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2013;30:713–24. **PMID: 23180580** http://dx.doi.org/10.1093/molbev/mss265</jrn>

<jrn>11. Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. Can J Stat. 2008;36:355–68. http://dx.doi.org/10.1002/cjs.5550360302</jrn>

Technical Appendix Table 1. Primer sequences

First PCR	Second PCR	
fwd: GAGTGCGGACAGTTTCCTTC rev: GATGAATGCTGATGACACACTG	rxn 1	fwd: GGCCAAGCATGGAGAGTATG rev: CAAGCTCGGGGAATGTCAC
	rxn 2	fwd: GGGTGGACAACAGAAAAACAG rev: CAAGCTCGGGGAATGTCAC
fwd: CGAAGCCAAACCCGAAGATG rev: GAGAGCATCTTGCATTGTGTAC	rxn 1	fwd: CCCCTCAATGTGCCCTAATTC rev: GTCGCCTCACAATATCCTTCTAG
	rxn 2	fwd: GCGTAATCTTCATCTCTCTTAG rev: CCATCCTGTCCACCAATTGTC
fwd: CTTGACATCTCTGAGGCAAC rev: GGGTGTGATTTACAGCTAAATGC	rxn 1	fwd: CGAACCACATGATTGGACCAAG rev: CTCATCAGACCTCCGCATTAATC
	rxn 2	fwd: ATATGAGAGAGGACGCCCCC rev: CAGTGAGGATTTATCTGTGGTTAAAC
fwd: GTTAACTTGACATCTCTGCCTTC rev: TGTCGTGAGGATGTACATGATC	rxn 1	fwd: CAAGGCACTATCAGGCAATG rev: CTGCATCAGTCTCTAAGGG
	rxn 2	fwd: CCTCACAAATTCAGTACCAAACG rev: CCGATAGTCCAGCTTATTCG
fwd: CAACCTGGTGGGAAACCATTC rev: TGGACACACAAAAAAGAAGAAATAG	rxn 1	fwd: CCGAGAAAATGAATTGATTTATGAC rev: AGTTAAATGACTTAGCCAGTATGG
	rxn 2	fwd: CCTCACAAATTCAGTACCAAACG rev: TGGACACACAAAAAAGAAGAAATAG
fwd: CGGACACACAAAAAGAAAGAAG rev: CATGGTGAGGTCTCCTGGAG	rxn 1	fwd: CGGACACACAAAAAGAAAGAAG rev: CCTGTTTGCGTTCCTTGACTAT
	rxn 2	fwd: CATGGCAATCCTGCAACATC rev: GATGAATGCTGATGACACACTG
fwd: CCCCTCAATGTGCCCTAATTC rev: CACACGGTAACTGGAGAGC	rxn 1	fwd: GAGACACTCCATCGAATCCAC rev: GGGTCCTCCGTTGCATTGAC
	rxn 2	fwd: GCGTAATCTTCATCTCTCTTAG rev: CTCGATAATTCTCTCTGGATGATG
fwd: ATATGAGAGAGGACGCCCCC rev: GGGTGTGATTTACAGCTAAATGC	rxn 1	fwd: ATATGAGAGAGGACGCCCCC rev: ATGCAGGGGCAAAGTCATTAG
	rxn 2	fwd: GGTGGAAAGTTTATTGGGCTG rev: GGGTGTGATTTACAGCTAAATGC

First PCR	Second PCR		
fwd: TCTCGAACCATTGTGCTTGG rev: TGTCGTGAGGATGTACATGATC	rxn 1	fwd: CCTCACAAATTCAGTACCAAACG rev: CCGATAGTCCAGCTTATTCG	
	rxn 2	fwd: CTGGACAAGTATTTCATGTGCTC rev: CAGCTGTTTGCCTTGGAAAAATG	
fwd: CGAGAAAATGAATTGATTTATGAC rev: TGGACACACAAAAAAGAAGAAATAG	rxn 1	fwd: GAGATCCGTCATTGATACCACAG rev: TGGACACACAAAAAAGAAGAAATAG	
	rxn 2	fwd: ATGCCACACCAAAACCATCTC rev: TGGACACACAAAAAAGAAGAAATAG	
fwd: CTGGACAAGTATTTCATGTGCTC rev: AGTTAAATGACTTAGCCAGTATGG	rxn 1	fwd: CTCCGAATGATTGAGATGGATG rev: TGTCGTGAGGATGTACATGATC	
	rxn 2	fwd: CAACCTGGTGGGAAACCATTC rev: GCCGACTTAAAATTCTCTATTTCC	
fwd: CAAGGCACTATCAGGCAATG rev: CTGCATCAGTCTCTAAGGG	rxn 1	fwd: CAGTTTTGAAGCTGCACTATG rev: GGGTGTGATTTACAGCTAAATGC	
	rxn 2	fwd: CAGTTTTGAAGCTGCACTATG rev: GATATTGTGGTAGTAGATACTCGAG	

Technical Appendix Table 2. Cycling conditions

First PCR			Nested PCR		
	30 min	98°C		30" min	98°C
	15"	98°C		15"	98°C
10 ×	30"	59 −54.5°C (−0.5°C / cycle)	10 ×	30"	59 −54.5°C (−0.5°C / cycle)
	90"	72°C		60"	72°C
	15"	98°C		15"	98°C
30 ×	30"	54°C	30 ×	30"	54°C
	90"	72°C		60"	72°C
	3′	72°C		3′	72°C

Technical Appendix Figure 1. Initial testing of MinION sequencing. A) Read depth plot for Ebola virus Makona. A blood sample from a non-human primate infected with Ebola virus Makona was subjected to the procedure shown in Figure 1, panel A. The read depth for each position in the genome on a log10-scale is shown. B) Comparison of a Sanger chromatogram and MinION read data. A peak intensity chromatogram from Sanger-sequencing and the corresponding number of reads displaying each nucleotide from a MinION sequencing run of the same random 30 nt region (nt 400 to 430) of the sample in panel A are shown. C) Read accuracy. Plasmid DNA with a known sequence was sequenced using the MinION device. The percentage of positions with a given accuracy (% of correct calls) is shown in 2% intervals on a linear scale. D) Frequency of second-most called nucleotide (N2). The percentage of positions with a given frequency for the second-most called nucleotide, compared to the dominant nucelotide, is shown in 1% intervals on a log scale. E) Probability for a miscall as a function of read depth. The probability that at least half of the reads correspond to an incorrect nucleotide for a given read depth is shown. In black a regression curve for read depths between 70 and 170 is shown, together with the corresponding formula which was used to approximate the error probability for large read depths (>170). F) Theoretical probability for a miscalled base as a function of read depth. The probability for at ≥ 1 miscalled nucleotide in a complete genome is shown as a function of the read depth.

Technical Appendix Figure 2. Effect of an external heat sink on MinION temperature. A) Improvised external heat sink. The heat sink consisted of a ≈30 × 30 cm metal plate, onto which the sequencing devices were placed. B) Device temperatures with and without external heat sink. Temperature data recorded by the MinION sequencing device from 2 representative 12 hour runs are shown, 1 with the

device sitting on the external heat sink, and 1 on a plastic table, but under otherwise identical conditions (e.g., time of day, external temperature).

Technical Appendix Figure 3. Initial MinION results and optimization of workflow. A) Read depth plot of an initial MinION run. A blood sample was sequenced under field conditions by using the workflow established under laboratory conditions. The read depth for each position in the genome is shown on a log10 scale. B) Optimization of PCRs. Three blood samples were subjected to RT-PCR as outlined in the workflow depicted in Technical Appendix Figure 1, panel A, by using 1 μL RNA as starting material for single PCRs, 5 μL RNA as material for single PCRs, or 5 μL RNA as starting material for nested PCRs. PCR products from each sample were pooled and purified, and 5 μL of the purified products was visualized by gel electrophoresis.

Technical Appendix Figure 4. (A) and (B) Read depth plots of representative samples. The read depths for each position in the genomes of a high (panel A, sample 13) and a low (panel B, sample 9) virus load sample on a log10-scale are shown.

Technical Appendix Figure 5. Phylogenetic analysis of determined full-genome sequences. A Bayesian tree of 296 sequences from the West African outbreak is shown. Branch colors indicate posterior probability as shown in the legend, with terminal branches being shown in black. The x-axis indicates time in years before acquisition of the last sample (March 12th, 2015). Origin countries of the samples are indicated. The blow-out shows sequences from Liberia and Mali. Bold text indicates sequences determined as a part of this study.

Technical Appendix Figure 6. A root-to-tip analysis was performed using for the sequences analyzed in Figure 2. The Liberian sequences obtained by us are highlighted in red, and the blue line shows a linear regression curve of all samples, with the inferred rate of substitution from this regression curve indicated.

Bioinformatics Scripts

Bash script for bioinformatics workflow

```
#!/bin/bash
if [ "$#" -ne 2 ]; then
echo
echo "usage: MinION_CGen <reference_file_without_ending> <dir>"
echo
exit 0
fi
# REFFILE contains reference file name without fasta, SEQDIR contains
directory with files
REFFILE=$1
SEQDIR=$2
echo "extracting all reads"
poretools fasta $SEQDIR > sequences.fasta
echo "generating alignment"
lastdb -Q 0 $REFFILE.index $REFFILE.fasta
```

```
lastal -s 2 -T 0 -Q 0 -a 1 $REFFILE.index sequences.fasta | last-map-probs >
sequences.last
echo "cropping primer fragments from sequences"
cat sequences.last | last2fasta_v4.pl >cropped.fasta
echo "realigning cropped sequences"
lastal -s 2 -T 0 -Q 0 -a 1 $REFFILE.index cropped.fasta | last-map-probs >
cropped.last
echo "generating sorted SAM file"
maf-convert sam cropped.last > cropped.sam
samtools view -T $REFFILE.fasta -bS cropped.sam | samtools sort -
cropped.last.sorted
samtools index cropped.last.sorted.bam
echo "generating pileup"
samtools mpileup -BQ 0 -d 1000000 -f $REFFILE.fasta cropped.last.sorted.bam
>pileup
echo "calling consensus"
cat pileup | pileup2nucl.pl >nucl
cat nucl | callnucl.pl >consensus
```

<u>last2fasta_v4.pl:</u>

```perl
#!/usr/bin/perl
use warnings;
use strict;
# numbers indicate first and last nucleotide in primer
my @startRanges = ([673, 692],
[1330, 1349],
[1981, 2001],
[2662, 2681],
[3313, 3332],
[3943, 3963],
[4609, 4629],
[5291, 5311],
[5923, 5944],
[6570, 6590],
[7223, 7242],
[7867, 7888],
[8520, 8539],
[9168, 9188],
[9780, 9802],
[10454, 10474],
[11150, 11172],
[11823, 11842],
[12471, 12490],
[13102, 13124],
[13725, 13747],
[14398, 14419],
[15052, 15072],
[15702, 15726],
[16324, 16345],
[16962, 16984],
[17620, 17640],
[18233, 18257]);
my @stopRanges = ([747, 766],
[1440, 1458],
[2057, 2076],
[2739, 2760],
[3394, 3414],
[4030, 4048],
[4679, 4700],
```

```perl
[5374, 5393],
[6010, 6032],
[6645, 6664],
[7294, 7312],
[7961, 7981],
[8594, 8617],
[9244, 9262],
[9891, 9913],
[10529, 10549],
[11226, 11251],
[11895, 11916],
[12549, 12571],
[13192, 13216],
[13837, 13855],
[14492, 14512],
[15124, 15143],
[15776, 15798],
[16426, 16447],
[17046, 17069],
[17692, 17715],
[18367, 18388]);
my $line;
my $lineIdx = 0;
my $template;
my $read;
# store next relevant line into $line
sub nextLine
{
while (1)
{
$line = <>;
exit unless defined $line; # end of file
$lineIdx++;
next if $line =~ m/^\s*(#.*)?$/; # blank or comment line
chomp $line;
return;
}
}
# remove insertions from reads
sub remove_insertions
{
my $inspos = index($template,"-");
while ($inspos > -1)
{
$template = substr($template,0,$inspos) . substr($template,$inspos+1);
$read = substr($read,0,$inspos) . substr($read,$inspos+1);
$inspos = index($template,"-");
}
}
while (1)
{
# first line
&nextLine;
unless ($line =~ m/^a\s/)
{
die "unexpected line $lineIdx (expected 'a ...'): $line\n";
}
# second line
```

```perl
&nextLine;
unless ($line =~ m/^s\s+\S+\s+(\d+)\s+(\d+)(?:\s+\S+){2}\s+([ACGT\-]+)/)
{
die("unexpected line $lineIdx (expected 's ... Start Length ...'):"
. "$line\n");
}
my $start = $1;
my $stop = $start + $2;
$template = $3;
$start++;
# third line
&nextLine;
unless ($line =~ m/^s\s+(\S+)(?:\s+\S+){4}\s+(\S+)/)
{
die("unexpected line $lineIdx (expected 's Name ... Sequence ...'):"
. "$line\n");
}
my $name =$1;
$read = $2;
&remove_insertions;
# eventually strip from line end
foreach (@stopRanges)
{
my $first = $_->[0];
my $last = $_->[1];
if ($first <= $stop and $stop <= $last)
{
$read = substr $read, 0, $first - $stop - 1;
last;
}
}
# eventually strip from line beginning
foreach (@startRanges)
{
my $first = $_->[0];
my $last = $_->[1];
if ($first <= $start and $start <= $last)
{
$read = substr $read, $last - $start + 1;
last;
}
}
# filter out hyphens
$read =~ tr/-//d;
# dump fasta record
print ">$name\n";
print "$read\n";
}
```

    pileup2nucl.pl:

```perl
#!/usr/bin/perl
use warnings;
use strict;
my $refbase;
my $callstring;
my $lineindex = 0;
my $position;
my %count = (
A => 0,
```

```perl
C => 0,
G => 0,
T => 0
);
# get next reference base and callstring
sub nextLine
{
my $line;
my @content;
$line = <>;
&exitprogram unless defined $line; # end of file
@content = split(" ", $line);
$position = $content[1];
$refbase = $content[2];
$callstring = $content[4];
$lineindex++;
return;
}
sub exitprogram
{
$lineindex++;
while ($lineindex<18959)
{
print "$lineindex\tA 0\tC 0\tG 0\tT 0\n";
$lineindex++;
}
exit;
}
sub removeInDels
{
my $croppedstring = "";
my $offset;
my $nextindel = index($callstring,"-");
while ($nextindel>0)
{
$callstring =~ /\-(\d+)/;
$offset = $nextindel + $1 + length($1) + 1;
$croppedstring = $croppedstring . substr($callstring,0,$nextindel);
$callstring = substr($callstring,$offset);
$nextindel = index($callstring,"-");
}
$callstring = $croppedstring . $callstring;
$croppedstring = "";
$nextindel = index($callstring,"+");
while ($nextindel>0)
{
$callstring =~ /\+(\d+)/;
$offset = $nextindel + $1 + length($1) + 1;
$croppedstring = $croppedstring . substr($callstring,0,$nextindel);
$callstring = substr($callstring,$offset);
$nextindel = index($callstring,"+");
}
$callstring = $croppedstring . $callstring;
}
while (1)
{
&nextLine;
while ($lineindex<$position)
```

```perl
{
print "$lineindex\tA 0\tC 0\tG 0\tT 0\n";
$lineindex++;
}
&removeInDels;
$count{'A'} = ($callstring =~ tr/A//);
$count{'C'} = ($callstring =~ tr/C//);
$count{'G'} = ($callstring =~ tr/G//);
$count{'T'} = ($callstring =~ tr/T//);
$count{$refbase} = ($callstring =~ tr/\.//);
print "$lineindex\tA $count{'A'}\tC $count{'C'}\tG $count{'G'}\tT
$count{'T'}\n";
}
```

callnucl.pl:

```perl
#!/usr/bin/perl
use warnings;
use strict;
use List::Util qw[min max];
my $A;
my $C;
my $G;
my $T;
my $depth;
my $separation=1.0;
# get next reference base and callstring
sub nextLine
{
my $line;
my @content;
$line = <>;
exit unless defined $line; # end of file
@content = split(/\s/, $line);
$A = $content[2];
$C = $content[4];
$G = $content[6];
$T = $content[8];
return;
}
MAIN: while (1)
{
&nextLine;
$depth=$A+$C+$G+$T;
if ($depth==0)
{
print "n";
next MAIN;
}
if ($A >= max($C,$G,$T))
{
if ($A > (max($C,$G,$T)*$separation))
{
print "a";
next MAIN;
}
if ($C > (max($G,$T)*$separation))
{
print "m";
next MAIN;
```

```perl
}
if ($G > (max($C,$T)*$separation))
{
print "r";
next MAIN;
}
if ($T > (max($C,$G)*$separation))
{
print "w";
next MAIN;
}
}
if ($C >= max($A,$G,$T))
{
if ($C > (max($A,$G,$T)*$separation))
{
print "c";
next MAIN;
}
if ($A > (max($G,$T)*$separation))
{
print "m";
next MAIN;
}
if ($G > (max($A,$T)*$separation))
{
print "s";
next MAIN;
}
if ($T > (max($A,$G)*$separation))
{
print "y";
next MAIN;
}
}
if ($G >= max($A,$C,$T))
{
if ($G > (max($A,$C,$T)*$separation))
{
print "g";
next MAIN;
}
if ($A > (max($C,$T)*$separation))
{
print "r";
next MAIN;
}
if ($C > (max($A,$T)*$separation))
{
print "s";
next MAIN;
}
if ($T > (max($A,$C)*$separation))
{
print "k";
next MAIN;
}
}
```

```perl
if ($T >= max($A,$C,$G))
{
if ($T > (max($A,$C,$G)*$separation))
{
print "t";
next MAIN;
}
if ($A > (max($C,$G)*$separation))
{
print "w";
next MAIN;
}
if ($C > (max($A,$G)*$separation))
{
print "y";
next MAIN;
}
if ($G > (max($A,$C)*$separation))
{
print "k";
next MAIN;
}
}
if (($A*$separation) <= min($C,$G,$T))
{
print "b";
next MAIN;
}
if (($C*$separation) <= min($A,$G,$T))
{
print "d";
next MAIN;
}
if (($G*$separation) <= min($A,$C,$T))
{
print "h";
next MAIN;
}
if (($T*$separation) <= min($A,$C,$G))
{
print "v";
next MAIN;
}
print "n";
}
```