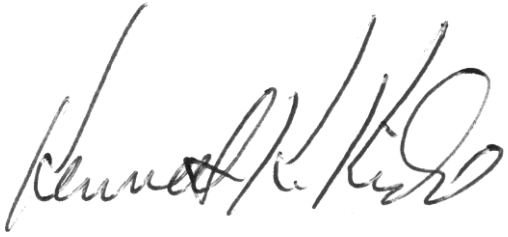
Final Summary Overview
National Institute of Justice
Grant # 2016-DN-BX-0162
Starting 01-01-2017 Ending 03-31-2019
Project Title:
"Enhancing and Sustaining the ALFRED-FROG-kb Forensic Resource"

Kenneth K. Kidd (PI)
Professor Emeritus of Genetics
Senior Research Scientist in Genetics
Email: Kenneth.Kidd@yale.edu
Telephone: 203-785-2654
Department of Genetics
Yale University School of Medicine

Submitted by K.K. Kidd on June 19, 2019
Signature of submitting official



PROJECT OBJECTIVES

As stated in the title, the project was to enhance the pre-existing forensic resource by adding more data to the underlying database and improving the web interfaces providing access to the data.

Background

ALFRED (the ALlele FREquency Database) was put online in 2000 and has grown from a few thousand allele frequency tables (one population for one polymorphism) to over 66 million. The web interface has been elaborated over the years to allow multiple different approaches to searching. In 2011 a separate pilot web interface, FROG-kb (Forensic Resource/Reference on Genetics-knowledge base), was placed online to allow users to use the database for forensic calculations.

Massively parallel sequencing (MPS) has expanded the use of single nucleotide polymorphisms (SNPs) for multiple aspects of forensic work. These uses include SNPs for individual identification (IISNPs), ancestry inference (AISNPs), and microhaplotypes for lineage identification (LISNPs). Numerous SNP and microhaplotype panels are being developed and are starting to be used in conjunction with STR analyses. Also, commercial multiplex panels based on MPS are becoming available and are beginning to be typed and published routinely for various population samples. Evaluating the results of using these SNP panels in forensic practice requires a large reference database on allele frequencies of SNPs in multiple populations from around the world. With this broad objective, the overall goals of this project remain the same: to enhance the already existing large database ALFRED (https://alfred.med.yale.edu/alfred/) and FROG-kb (https://frog.med.yale.edu/FrogKB/) both in content and functionality.

The two-major accomplishments of the project include (1) greatly increasing the quantity of curated data--specifically the data related to forensic panels in the database and (2) improving the functionality provided by the unique FROG-kb and ALFRED interfaces.

Specific objectives

A general specific objective was to add new database functionalities and provide the users with an even more user-friendly, intuitive interface to FROG-kb. The data-related specific objectives of the project included 1) to enhance the database content by making available curated allele frequency data in ALFRED from various sources with a specific emphasis on including markers of forensic interest, 2) to define new forensic SNP sets in ALFRED, 3) to add new SNP panels to FROG-kb, and 4) to update the reference populations used in the calculations for the existing SNPs whenever possible.

PROJECT SUBJECTS

No individuals were studied as part of this project. Allele frequency data that had already been collected for population samples on large numbers of anonymous individuals were entered into the database.

PROJECT DESIGN AND METHODS

The initial FROG-kb web site <https://frog.med.yale.edu/FrogKB/> was introduced in 2011 as a prototype to support forensic use of single nucleotide polymorphisms (SNPs). In the years since then ongoing research into how the database and web interface can better serve forensics has resulted in extensive redesign of the database interface and functionality (Kidd et al., 2018). FROG-kb provides a freely accessible web interface that facilitates forensic practice and can be useful for teaching and research. The site has functional enhancements, extensive new documentation, and new reference panels of SNPs with new curated data. FROG-kb focuses on single nucleotide polymorphisms (SNPs) and provides reference population data for several published panels of individual identification SNPs (IISNPs) and several published panels of ancestry inference SNPs (AISNPs). For each of the various marker panels with reference population data, FROG-kb calculates random match probabilities (RMP) and relative likelihoods of ancestry for a user-entered genotype profile (either completely or partially specified). Example genotype profiles are available and the new User's Manual presents interpretation guidelines for the calculations. The extensive documentation along with ongoing updates makes FROG-kb a comprehensive tool facilitating use of SNPs in forensic practice and education. An overview of the new FROG-kb with examples and material explaining the results is in the publication on the revised FROG-kb (Kidd et al., 2018).

DATA ANALYSIS

The project itself involved no data analysis other than summary statistics of the numbers of SNPs and SNP panels added to the database and interfaces; however various statistical analyses for users have been embedded into the web interfaces:

1. Calculation of Fst (Sewall Wright's fixation index) and the average heterozygosity for each SNP in ALFRED with data for more than one population (almost all polymorphisms by design--SNPs with data on only one population are generally not entered).

2. Population-specific probabilities and likelihood ratios for a multi-locus genotype entered for a panel in FROG-kb.

3. Counting functions to provide the user with information on how many SNPs or populations have been retrieved by a search.

FINDINGS

What was accomplished under the goals?

This project includes both ongoing and completed activities addressing the specific goals of the project. A key ongoing component of this project involves accumulating forensic related material from various data sources (data from Kidd Lab, the literature, collaborators, and submissions), curating the information for entry into ALFRED, and then implementing the relevant data into FROG-kb. Through these efforts new SNP sets and additional reference population samples for the existing SNP sets are added into FROG-kb. Many of the panels in FROG-kb have more than 70 reference populations and one panel has as many as 164 reference populations (Kiddlab set of 55 AISNPs). Each of these panels has supporting population data. URL links exist to pages in ALFRED for more details and allele frequency data tables for specific populations. Data exist and are accessible through ALFRED not only for the populations used for calculation but also for each SNP-population combination for which data are available. Various interface updates increasing the accessibility and user-friendliness of the database were added during this project period. Details on the FROG-kb updates and with material for better interpretation of the statistical results in FROG-kb was published (Kidd et al., 2018).

Major activities

The major activities and accomplishments include:

1) the content of the ALFRED database increased with the addition of more than 28 million frequency tables (each table is the allele frequency for one population for one polymorphism). ALFRED now has data on **66.7 million** frequency tables.

2) new forensic panels and additional reference populations for existing panels added to the underlying database--ALFRED. Where applicable, panels in FROG-kb have been updated with additional reference populations.
3) interface updates enhanced the functionalities and the output results from the statistical calculations in the FROG-kb database interface.
4) url links to external databases that are obsolete or not active from the site and population description pages are being updated in ALFRED.

Significant results or key outcomes

Increase in ALFRED frequency tables: One of the primary objectives of the project is to accumulate population reference data for common sets of markers. At the beginning of this grant we had included **37 million** allele frequency tables in the ALFRED database involving **664,140** polymorphisms and **723** populations. As of December 2018, the database contains **66,726,252** frequency tables involving **664,708** polymorphisms and **762** populations. These numbers reflect the focus on having a more extensive population representation for the polymorphisms already in the database. Much of the newly added data consists of data on markers already defined in the database for populations that were already defined. The new populations include the twenty-six 1000 Genomes populations (1KGP); we have entered 1KGP data for SNPs in the forensic SNP panels to provide a single source for reference data on those markers. We have not attempted to enter general 1KGP data.

One large recent addition to ALFRED includes markers typed using Illumina Quad 610/Illumina Quad 370 on 1,076 individuals from 30 populations with geographical coverage spanning from the Baltic Sea to Lake Baikal. We identified and entered ~320,000 markers that overlap with the 650K marker set (markers already in ALFRED). Up to **114** populations from various regions of the world now have coverage for this dataset of 320,000 markers.

Microhaplotype data: Allele frequencies for 165 microhaplotypes on 83 population samples including 57 Kidd Lab populations and the twenty-six 1KGP population samples (Kidd KK et al., 2017, Bulbul et al.,

2018) are all entered in ALFRED. The data for 26 microhaps for DNA mixture analysis published recently in Chen et al. (2018a) have all been entered in ALFRED. The publication did not provide the haplotype frequencies for all the phase-3 1KGP samples they studied so we downloaded the 1KGP frequency data, phased the data (Phase 2-1-1 software) and have entered the data in ALFRED for the 26 1KGP population samples. Data on 7 microhaps for ancestry analysis from Chen's group (Chen et al., 2018b) are also entered in ALFRED. There are now 198 microhap loci for which haplotype frequency data can be obtained quickly in ALFRED using the keyword search term 'microhap' or 'clicking on 'microhap' under the pull-down search menu. Between 83 and 96 populations have haplotype frequency data for most of the loci. The entire set of microhaps data is available for download as a text file under the 'Downloads' option -> 'Summaries' tab from ALFRED homepage.

Forensic SNP panel data from various publications--new reference panels and additional populations added to existing panels in FROG-kb: The current FROG-kb interface has 4 SNP sets under the IISNP component, 14 panels under the AISNP component and one SNP panel under the PISNP (phenotype informative SNPs) component. Three IISNP panels have 70 to 75 reference populations available for calculation and one panel has 28 reference populations. Of the 14 AISNP panels 4 of them have over 100 reference populations for calculations, 6 panels have 70-98 reference populations and 4 panels have 34 to 53 reference populations. Additional SNP sets not included yet in FROG-kb database are available through the ALFRED SNP set page. The IISNP tab under the ALFRED-SNP sets have 6 additional panels for a total of 10 IISNP sets and the AISNP sets have 8 more panels for a total of 22 AISNP panels.

The quality and completeness of the printable output of statistical calculations improved in FROG-kb: The FROG-kb interface has been completely redesigned as part of this project and updated to include 'Time and date stamps', and an option for the user to supply a "case number" to be input and printed on all results generated from FROG-kb calculations. The printable version of the population likelihoods based on a selected panel is furnished with 'Computed Date', 'Printed Date', and the 'Case Number'.

This additional information provides identifiable parameters that makes the record unique and we strongly believe this will eventually be an important feature when the records are used for DNA evidence.

Improved computational time: The computational time for likelihood calculations for panels with more than 100 reference populations most of the time exceeded one minute. This issue was revisited; queries populating and accessing the backend database as well as the logic in the computation code were optimized. Now on a working day (considering the high internet traffic), for a panel with more than 100 SNPs and reference populations the computational time is < 25 seconds.

Frequency Download: The 'Frequencies Download' function now allows download of genotype frequency and allele frequency tables for all the II and AI SNP panels and the downloaded file is up-to-date. If a new reference population is added to a panel for the calculations, the downloaded file will have the corresponding frequencies that were used.

Examples: The 'Examples' function format has been updated. The files that are used for running the pre-selected data entry page can now be downloaded. The file format is compatible with the format required for the 'File Upload' function within the 'Data Entry' page. We have also included additional example files from three different populations to the Kidd Lab panel of 55 AISNPs.

Accomplishments during extension period: 01/01/2019 to 03-31-2019

After funding ended on 12/31/2018 for data curation, loading and checking new data, and implementing new features, effort was directed to migrating the back-end database of ALFRED and FROG-kb from Yale's institutional Oracle server to a local SQL server database system situated at the Yale Center for Medical informatics (YCMI). This was done to reduce the cost of supporting continued maintenance of static versions of ALFRED and FROG-kb. The Yale Center for Medical Informatics (YCMI) has agreed to provide ongoing financial support for the web-interfaces and the back-end database. As a result, external users have an unchanged URL link to ALFRED and FROG-kb since the formal end of NIJ funding. In addition, a stand alone batch likelihood computation tool was created during the January-

March 2019 time frame as an upgrade to the existing likelihood computation function that is available on the FROG-kb web interface. We are preparing a short paper describing this software tool for submission to a forensic journal. The downloadable software tool (called FrogAncestryCalc) is currently in a testing phase. The software distribution of this new tool will include data for five of the AISNP panels in the FROG-kb interface which have the largest number of reference populations currently available.

Literature cited in this report

Kidd, K. K., Soundararajan, U., Rajeevan, H., Pakstis, A. J., Moore, K. N., & Ropero-Miller, J. D. (2018). The redesigned Forensic Research/Reference on Genetics-knowledge base, FROG-kb. Forensic Sci Int Genet, 33, 33-37. doi: 10.1016/j.fsigen.2017.11.009

Kidd, K. K., Speed, W. C., Pakstis, A. J., Podini, D. S., Lagace, R., Chang, J., Soundararajan, U. (2017). Evaluating 130 microhaplotypes across a global set of 83 populations. Forensic Sci Int Genet, 29, 29-37. doi: 10.1016/j.fsigen.2017.03.014

Bulbul, O., Pakstis, A. J., Soundararajan, U., Gurkan, C., Brissenden, J. E., Roscoe, J. M., Kidd, K. K. (2018). Ancestry inference of 96 population samples using microhaplotypes. Int J Legal Med, 132(3), 703-711. doi: 10.1007/s00414-017-1748-6

Chen, P., Yin, C., Li, Z., Pu, Y., Yu, Y., Zhao, P. Chen, F. (2018a). Evaluation of the Microhaplotypes panel for DNA mixture analyses. Forensic Sci Int Genet, 35, 149-155. doi: 10.1016/j.fsigen.2018.05.003

Chen, P., Zhu, W., Tong, F., Pu, Y., Yu, Y., Huang, S., Li, Z., Zhang, L., Liang, W., Chen, F. (2018b). Identifying novel microhaplotypes for ancestry inference. Int J Legal Med. doi: 10.1007/s00414-018-1881-x

What opportunities for training and professional development has the project provided?

Much effort was put into training a new assistant curator on the different aspects of data entry and curation. The newly hired curator learned to use the program developed by our programmer for systematic automated data entry following curation. This was utilized for the high throughput dataset entered in ALFRED. The new curator was also trained to assist the senior curator with the organization of data in the spreadsheet format required for the upload of data into ALFRED. The assistant curator was trained to enter sample descriptions and population descriptions via the curator interface. The curators involved in the project interact with the programmer and other members of the project team to understand the informatics and population genomics involved in ALFRED and FROG-kb databases. All individuals

involved in the project benefited from 'cross-training' created by the frequent, and necessary, interaction with other staff members.

How were the results disseminated to communities of interest?

The broad objective of the project was to assemble and make available otherwise widely dispersed data in an integrated fashion. Besides the Kidd et al. (2018) publication mentioned earlier which reports on the redesign of FROG-kb implemented during the award period, the large number of users attests to the successful accomplishment of that objective. The usage currently reflects access from individuals at educational institutions, forensic researchers, and others aware of the resource.

For the 6 month period from January to June 2018 the average number of users each month was around 500 for the FROG-kb database. About forty-seven percent of the users have internet addresses in the United States and 26 percent of the users are from China. Other high-ranking user countries include: Russia, Canada, Germany, United Kingdom and Spain. On average, about 15 visitors per day accessed the FROG-kb database.

For the 2-month period covering mid-July 2018 to mid-September 2018, the ALFRED database had an average of about 1500 users per month. About thirty-eight percent of the users had internet addresses in the United States and 10 percent of the users were from China. Other high-ranking user countries include: Germany, United Kingdom, India, Canada and Russia. An average of about 180 visitors per day accessed the database. The number of visitors is determined by the IP addresses. If a request from an IP address comes after some time (timeout) since the last request from this IP, it is considered to belong to a different visitor. The timeout is set to 30 minutes by default.

The ALFRED and FROG-kb databases are highlighted in all papers we publish in peer-reviewed journals; we have published 16 such forensic-related papers in 2016-2019. Various aspects of the work supported by this project were also presented at numerous talks and posters at international meetings and visits to university laboratories.

Presentations on forensic applications of SNPs

2nd Annual Genetics in Forensics Congress, London, U.K., March 13-15, 2017. K.K. Kidd presented the keynote address titled: Highly informative microhaplotype loci are ready for implementation using massively parallel sequencing (MPS).

10th ISABS conference on Forensic and Anthropologic Genetics, Dubrovnik, Croatia, June 19-24, 2017. K.K.Kidd presented talk: "Microhaplotypes are ready for implementation in casework".

K.K. Kidd co-author on poster presentation by Ozlem Bulbul—"Improving ancestry distinctions among Southwest Asian populations".

Green Mountain DNA Conference, 10th annual meeting, Burlington VT, July 24-26, 2017. K.K. Kidd presented talk: "Microhaplotypes are coming of age".

Cambridge Healthtech Institute sponsored 2nd Annual Next Generation Summit for DNA Forensics; August 15-18, 2017 Grand Hyatt, Washington, DC. K.K. Kidd presentation: "Microhaplotypes, SNP-Based Markers for Typing by Massively Parallel Sequencing (MPS)".

Co-authored 2 presentations at ISFG 2017--27th Congress of the International Society for Forensic Genetics; Aug 28-Sep 2, 2017, Seoul, South Korea.
"An ancestry informative assay for the Asia Pacific region" C. Phillips, D. McNevin, M. Eduardoff, M. dela Puente, A. Heidegger, Catarina Xavier, D. Power, W. Parson, K. Kidd, C. Santos, R. Lagacé, S. Wootton, M. Barash, M.C.A. De Ungria, S.R. Hong, H.Y. Lee, C. Oz, E. Peters, N.A. Soliven, N. Tuitoga, S. Olson, M.V.Lareu, R. Daniel.

"Microhaplotypes for ancestry prediction" Fabio Oldoni, Rebecca Hart, Kelly Long, Kathrina Maddela, Selena Cisana, Moses Schanfield, Sharon Wootton, Joseph Chang, Robert Lagace, Ryo Hasegawa, Kenneth Kidd, Daniele Podini.

Kenneth K. Kidd, slide presentation at NIJ-sponsored symposium, American Academy of Forensic Sciences annual meeting, Seattle WA, Feb. 20, 2018. "Microhaplotypes analyzed by massively parallel sequencing (MPS) are valuable forensic tools".

Kenneth K. Kidd, Andrew J. Pakstis, Willian C. Speed, Haseena Rajeevan, Usha Soundararajan, Sharon Wootton, Joseph Chang, Robert Lagace, poster presentation, Gordon Research Conference on Forensic Analysis of Human DNA, Newry, Maine, June 2018, "More refined ancestry inference with SNPs and microhaplotypes.

Fabio Oldoni, Rebecca Hart, Nyra Rashad, Keylie Gibson, Sharon Wootton, Robert Lagace, Ryo Hasegawa, Joseph Chang, Moses S. Schanfield, Kenneth Kidd, Daniele S. Podini; presentation at Criminalistics session of American Academy of Forensic Sciences annual meeting, Seattle WA, Feb. 19-24, 2018. "An evaluation of a novel Massively Parallel Sequencing (MPS) 74-microhaplotype panel for biogeographic ancestry prediction".

Sharon C. Wootton, Joseph P. Chang, Chien-Wei Chang, Sheri J. Olson, Daniele S. Podini, Kenneth K. Kidd, presentation at Criminalistics session of American Academy of Forensic Sciences annual meeting, Seattle WA, Feb. 19-24, 2018. "Massively Parallel Sequencing (MPS) of Short Tandem Repeats (STRs) and microhaplotypes for mixtures".

Publications 2016-2019

U. Soundararajan, L. Yun, M. Shi, K.K. Kidd, 2016. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Science International: Genetics* 23:25-32.

Cai-Xia Li, Andrew J. Pakstis, Li Jiang, Yi-Liang Wei, Qi-Fan Sun, Hong Wu, Ozlem Bulbul, Ping Wang, Long-Li Kang, Judith R. Kidd, Kenneth. K. Kidd, 2016. A panel of 74 AISNPs: Improved Ancestry Inference within Eastern Asia. *Forensic Science International: Genetics*. 23:101-110.

Lotfi Cherni, Andrew J. Pakstis, Sami Boussetta, Sarra Elkamel, Sabeh Frigi, Houssein Khodjet-El-Khil, Alison Barton, Eva Haigh, William C. Speed, Amel Ben Ammar Elgaaied, Judith R. Kidd, Kenneth K. Kidd, 2016. Genetic variation in Tunisia in the context of human diversity worldwide. *American Journal of Physical Anthropology* 161:62-71.

Ozlem Bulbul, Lotfi Cherni, Houssein El Khil Khodjet, Haseena Rajeevana, Kenneth K. Kidd, 2016. Evaluating a subset of ancestry informative SNPs for discriminating among Southwest Asian and Circum-Mediterranean populations. *Forensic Science International: Genetics*. 23:153-158.

Kenneth K. Kidd, 2016. Proposed nomenclature for microhaplotypes. *Human Genomics* **10**:16; DOI**:**10.1186/s40246-016-0078-y

Andrew J Pakstis, Longli Kang, Lijun Liu, Zhiying Zhang, Tianbo Jin, Elena L. Grigorenko, Frank R.Wendt, Bruce Budowle, Sibte Hadi, Mariam Salam Al Qahtani, Niels Morling, Helle Smidt Mogensen, Goncalo E. Themudo, Usha Soundararajan, Haseena Rajeevan, Judith R. Kidd, Kenneth K. Kidd, **2017**. Increasing the reference populations for the 55 AISNP panel: the need and benefits. *International Journal of Legal Medicin*e 131:913-917*;*

Kenneth K. Kidd, William C. Speed, Andrew J. Pakstis, Daniele S. Podini, Robert Lagace, Joseph Chang, Sharon Wootton, Eva Haigh, Usha Soundararajan, 2017. Evaluating 130 microhaplotypes across a global Set of 83 populations. *Forensic Science International: Genetics* 29:29-37.

George Stamatoyannopoulos, A. Bose, A. Teodosiadis, F. Tsetsos, A. Plantinga, N. Psatha, N. Zogas, E. Yannaki, P. Zalloua, K.K. Kidd, B.L. Browning, J. Stamatoyannopoulos, P. Paschou, P. Drineas, 2017. Genetics of the Peloponnesean populations and the theory of extinction of the medieval Peloponnesean Greeks. European Journal of Human Genetics 25:637-645.

Kenneth K. Kidd, Usha Soundararajan, Haseena Rajeevan, Andrew J. Pakstis, Katherine N. Moore, Jeri D. Ropero-Miller, **2018**. The redesigned Forensic Research/Reference on Genetics-knowledge base, FROG-kb. *Forensic Science International: Genetics* 33:33-37

Ozlem Bulbul, Andrew J. Pakstis, Usha Soundararajan, Cemal Gurkan, Jane E. Brissenden, Janet M. Roscoe, Baigalmaa Evsanaa, Ariunaa Togtokh, Peristera Paschou, Elena L. Grigorenko, David Gurwitz, Sharon Wootton, Robert Lagace, Joseph Chang, William C. Speed, Kenneth K. Kidd, 2017. Ancestry inference of 96 population samples using microhaplotypes. *International Journal of Legal Medicine* 132:703-711.

Ozlem Bulbul, William C Speed, Cemal Gurkan, Usha Soundararajan, Haseena Rajeevan, Andrew J Pakstis, Kenneth K Kidd, 2018. Improving Ancestry Distinctions among Southwest Asian Populations. *Forensic Science International Genetics* 35:14-20.

Kenneth K. Kidd, Andrew J. Pakstis, William C. Speed, Robert Lagace, Sharon Wootton, Joseph Chang, 2018. Selecting microhaplotypes optimized for different purposes. *Electrophoresis* doi.org/10.1002/elps.201800092; online June 21, 2018.

Sheng Gu, Hui Li, Andrew J. Pakstis, William C. Speed, David Gurwitz, Judith R. Kidd, Kenneth K. Kidd, 2018. Recent selection on a Class I ADH locus distinguishes Southwest Asian populations including Ashkenazi Jews. *Genes* 9, E452, 16pp; doi:10.3390/genes9090452.

Lindsay Bennett, Fabio Oldoni, Kelly Long, Selena Cisana, Katrina Madella; Sharon Wootton, Joseph Chang, Ryo Hasegawa; Robert Lagace, Kenneth K. Kidd, Daniele Podini, 2019. Mixture deconvolution by massively parallel sequencing of microhaplotypes. *International Journal of Legal Medicine*. 133:719-729.

Sami Boussetta, Lotfi Cherni, Andrew J. Pakstis, Nesrine Ben Salem, Sarra Elkamel, Kenneth K. Kidd, Houssein Khodjet-el-Khil, Amel Ben Ammar ElGaaied (2019) Usefulness of COMT gene polymorphism in North African populations. Gene 696:186-196.

Drineas, Petros; Tsetsos, Fotis; Plantinga, Anna; Lazaridis, Iosif; Yannaki, Evangelia; Razou, Anna; Kanaki, Katerina; Michalodimitrakis, Manolis; Perez-Jimenez, Francisco; De Silvestro, Giustina; Renda, Maria C; Stamatoyannopoulos, John A; Kidd, Kenneth; Browning, Brian; Paschou, Peristera; Stamatoyannopoulos, George (2019). Genetic history of the population of Crete. Annals of Human Genetics. In Press May 1, 2019; online (early view) June 12, 2019.