HHS Public Access

Author manuscript

Biotechniques. Author manuscript; available in PMC 2015 March 29.

Published in final edited form as:

Biotechniques. ; 56(3): 134–141. doi:10.2144/000114146.

Controlling for contamination in re-sequencing studies with a reproducible web-based phylogenetic approach

Benjamin Dickins[#1,2,†], **Boris Rebolledo-Jaramillo**[#1,3], **Marcia Shu-Wei Su**[2], **Ian M. Paul**[4], **Daniel Blankenberg**[1], **Nicholas Stoler**[3], **Kateryna D. Makova**[2], and **Anton Nekrutenko**[1]

[1]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA

[2]Department of Biology, Penn State University, University Park, PA

[3]Interdisciplinary Graduate Program in BioSciences, Penn State University, University Park, PA

[4]Department of Pediatrics, Penn State College of Medicine, Hershey, PA

[#] These authors contributed equally to this work.

Abstract

Polymorphism discovery is a routine application of next-generation sequencing technology where multiple samples are sent to a service provider for library preparation, subsequent sequencing, and bioinformatic analyses. The decreasing cost and advances in multiplexing approaches have made it possible to analyze hundreds of samples at a reasonable cost. However, because of the manual steps involved in the initial processing of samples and handling of sequencing equipment, cross-contamination remains a significant challenge. It is especially problematic in cases where polymorphism frequencies do not adhere to diploid expectation, for example, heterogeneous tumor samples, organellar genomes, as well as during bacterial and viral sequencing. In these instances, low levels of contamination may be readily mistaken for polymorphisms, leading to false results. Here we describe practical steps designed to reliably detect contamination and uncover its origin, and also provide new, Galaxy-based, readily accessible computational tools and workflows for quality control. All results described in this report can be reproduced interactively on the web as described at http://usegalaxy.org/contamination.

Address correspondence to Kateryna Makova, Penn State University, University Park, PA, kmakova@bx.psu.edu; or Anton Nekrutenko, Penn State University, University Park, anton@bx.psu.edu..

[†]Present address: School of Science and Technology, Nottingham Trent University, UK

Supplementary material for this article is available at www.BioTechniques.com/article/114146.

Author contributions

B.D. developed the original ideas for using allele frequency distributions and phylogenetic clustering to detect and trace contamination. He also performed most of the experimental work related to collection, preparation, and sequencing of samples. B.R.J. performed analysis of the datasets reported in this manuscript and developed the Get FASTA from Variants Table tool, Phylorelatives Tool, and MAF Boxplot tool. M.S.W.S. helped perform additional experimental work. I.M.P. enabled sample collection. D.B. developed the Naive Variant caller and performed necessary adjustments to the Galaxy framework. N.S. developed the Variant Annotator Tool. K.D.M. conceived the study, oversaw the analysis, and wrote the manuscript. A.N. developed analysis approaches and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Keywords

re-sequencing; contamination; next-generation sequencing; Galaxy; reproducibility

Very high depth of coverage can be achieved for a moderate cost using high-throughput sequencing technologies. This allows identification of very low frequency variants in re-sequencing studies dealing with complex non-diploid mixtures represented by viral, bacterial, and organellar genomes, as well as genetically abnormal samples such as altered genomic DNA isolated from malignant lesions. However, the power to detect rare variants is also the Achilles' heel of these approaches in that contamination and carryover among the sequenced samples cannot be easily distinguished from true genetic variants. This is especially relevant with multiplexing approaches where large numbers of highly similar samples are handled simultaneously. Here we illustrate how to detect warning signs of sample contamination, describe best practices for re-sequencing study design, and provide readily usable computational workf lows aimed at detecting these artifacts.

A typical re-sequencing experiment entails processing of multiple samples that are expected to differ at relatively few sites. These may include bacterial isolates, viral DNA samples, or, in this study, human mitochondrial DNA (mtDNA). Human cells contain various numbers of mitochondria, each harboring a number of circular genomes (1,2). The individual genomes often differ from each other at a few polymorphic sites that display the whole possible range of allele frequencies (a phenomenon called heteroplasmy). As the role of mtDNA in the etiology of human disease is now well established (i.e., mtDNA mutations contribute to over 200 known diseases) (3), the need to reliably identify heteroplasmic sites becomes more urgent with the realization that most disease-causing mtDNA mutations exist as heteroplasmies, and their clinical manifestations depend on the relative proportion of normal to mutant alleles (4–6). This proportion can change dramatically during oogenetic bottlenecks, frequently leading to the increase of disease-causing alleles in offspring (7–11). Thus detection of even low frequency variants becomes critical.

Historically, Sanger sequencing has the sensitivity to detect minor alleles at ~10%–20% frequency (12). Application of Illumina technology has driven the detection threshold down to ~1%–2% (13), and increases in sequencing depth combined with application of dynamic likelihood approaches for variant detection promise to drive the detection threshold below 1%. At this detection level, it becomes critical to separate true signal from contamination, which can have multiple sources. For example, Illumina points out that insufficient flushing of HiSeq instruments between runs can lead to a sample carryover rate of 0.05%–0.1%. Additional contamination at this very low detection threshold is also highly likely due to sample handling, including pipetting, gel excision, and airborne droplets produced during opening and closing of PCR strips.

Materials and methods

Ethical approval

This study was approved by the Human Subjects Protection Office of the Penn State College of Medicine.

DNA isolation

Blood was collected from the finger using a BD Microtainer contact-activated lancet (catalog # 366593 or 366594; BD, Franklin Lakes, NJ) and was preserved in a BD Microtainer Tubes with K2E (catalog # 365974) until DNA extraction. DNA was isolated using QIAGEN DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany) in either the low-throughput microtube-based format (catalog # 69504 or 69506) or the 96-well plate format (catalog # 69581 or 69582). During high-throughput extractions, alternate columns on the plate were left empty to minimize the risk of cross-contamination from the use of multichannel pipettes. DNA was eluted using the kit buffer AE and stored at -20°C. DNA extraction from buccal cells was carried out according to the method detailed in Reference 14. Buccal cells were collected by scraping the inside of the mouth with ten cotton swabs on plastic sticks. These swabs were placed in Slagboom buffer (0.1 M NaCl, 10 mM Tris-HCl pH 8.0, 10 mM EDTA, 0.5% SDS) with Proteinase K (0.2 mg/ mL). After storage at room temperature, samples were sorted into a pseudo-random order (separating family members) before DNA extraction was carried out. Proteins were removed using an organic de-proteinization reagent (ORPR), and DNA was precipitated with isopropyl alcohol. The DNA was re-suspended in 250 μl of TE buffer and stored at -20°C or below.

mtDNA amplification

Whole mitochondrial DNA was amplified with two sets of primers: L* 2 8 1 7 (5′-GCGACCTCG-GAGCAGA AC-3′) and H*11570 (5′-GTAGGCAG ATGG AGCT TGTTAT-3′); L10796 (5′-CCACTGACATGACTTTCCA A-3′) and H3370 (5′-AGAATTTTTCGTTCGGTA AG-3′). This produced 2 overlapping products, each ~9 kb in size. These primers are based on those described in our previous publication (13) and also by Tanaka et al. (15) except that L*2817 and H*11570 have been modified to improve amplification success. The PCR amplification was performed in 50 μl with 10 μl (blood-derived) or 2 μl (cheek-derived) DNA, 0.2 mM dNTPs (PCR grade; Roche Applied Science, Pleasanton, CA), 0.84 units Expand High Fidelity PCR Enzyme mix (Roche Applied Science), 1 × buffer including 1.5 mM Mg^{2+}, and 0.2 μM each forward and reverse primer (Integrated DNA Technologies, Inc., Coralville, IA). PCR reactions were carried out in 8-well strips.

Thermal cycling conditions consisted of a progression of two cycles. After an initial denaturation step of 94°C for 2 min, the first cycle was 94°C for 15 s, 62.3°C for 30 s, and 68°C for 8 min for 10 repeats. The second cycle was 94°C for 15 s, 62.3°C for 30 s, and 72°C for 8 min (plus 5 s per cycle) for 20 repeats. The terminal extension step consisted of 72°C for 8 min. After visualizing aliquots by gel electrophoresis, two overlapping amplicons from each individual were mixed in approximately equimolar proportions (based on 2-D densitometry estimates). Mixed amplicons for each individual were cleaned up using

column purification (QIAGEN's QIAquick). The Qubit dsDNA BR assay (Invitrogen, Carlsbad, CA) was used to quantitate samples after mixing using a Qubit 2.0 fluorometer (Invitrogen).

Galaxy pipeline

Our intention was not only to develop a computational methodology for contamination detection but to also make it readily accessible to anyone wishing to test it or apply it to their own studies. To achieve this, we have implemented a number of components described below and incorporated them into our widely used Galaxy platform (http://usegalaxy.org) as described at the following URL: http://usegalaxy.org/contamination.

Naive Variant Caller tool

The Naive Variant Caller tool processes aligned sequencing reads from the BAM format and produces a VCF file containing per position variant calls. This tool allows multiple BAM files to be provided as input and utilizes read group information to make calls for individual samples. User configurable options allow filtering reads that do not pass mapping or base quality thresholds and minimum per base read depth; users can also specify the ploidy and whether to consider each strand separately. In addition to calling alternate alleles based upon simple ratios of nucleotides at a position, per base nucleotide counts are also provided. A custom tag, NC, is used within the Genotype fields. The NC field is a comma-separated listing of nucleotide counts in the form of <nucleotide> = <count>, where a plus (+) or minus (-) character is prepended to indicate strand if the strandedness option was specified.

Variant Annotator tool

The Variant Annotator tool processes the raw variant count data from the Naive Variant Caller tool. SNV counts and allele statistics are reported for each site in a simple tabular format. Data from multiple samples are supported, via sample columns in the input VCF. The first (major) and second (minor) most abundant alleles are reported, along with the frequency of the latter. The user can set a coverage threshold, which is applied to each strand individually. An allele count is computed based on the number of alleles passing a user-supplied frequency threshold. A basic filter for strand bias is applied at this stage, excluding sites where the threshold-passing alleles differ between the strands. At these sites, neither allele count is used, and the tool will instead mark the count as zero.

Get FASTA from Variants Table tool

Provided a table defining the major and minor alleles per position and the length (L) of the target sequence, the Get FASTA from Variants Table tool generates a string of length L where every position is an N nucleotide. Then, position by position in the alleles table, every N is replaced by the inferred major allele nucleotide (or the minor allele at heteroplasmic positions when generating the minor allele sequence). Positions that are not described in the alleles table will remain as N.

Since all sequences share the same length, all the major allele sequences are included into a single file (with proper headers per sample) to create a multiple sequence alignment in FASTA format that can be used for downstream phylo-genetic analyses. In contrast, the

minor allele sequences are recorded as single FASTA files per sample to ease their downstream manipulation. For our purposes, L was set to the length of the Revised Cambridge Reference Sequence (NC_012920), 16,569 nucleotides.

Phylorelatives tool

The Phylorelatives tool takes as input the set of sequences generated by the Get FASTA from Variants Table tool, and reports the closest relatives of the test minor allele sequence in a Neighbor-Joining (NJ) tree (16), along with a picture of the tree, and the resulting NJ tree in Newick format. In addition, the set of sequences used during the analysis is returned as a single multiple sequence alignment FASTA file. This tool uses a combination of R and Python libraries implemented in a Python script. The R package Analysis of Phylogenetics and Evolution (ape)(24) is used to generate the NJ tree. The pairwise distance between the sequences is calculated using the raw model, which is simply the proportion of different sites between the two sequences. Sites with missing information are excluded by default (complete deletion), but this option can be set to pairwise deletion at run time. Also, by default the tool runs 1000 bootstrap replicas and does not root the tree. Options can be set to include a rooting sequence, suppress bootstrap, or change the number of replications. Next, the Python library Dendropy (25) is used to process the resulting tree topology and infer the relatives of the samples. Starting from the leaf node representing the minor allele sequence in question, the tool travels up the tree looking for the closest node whose descendants include at least one major allele sequence. The list of descendants of this node is informed as the relatives of the sample in question (i.e., the closest related samples in the NJ tree). Input minor allele sequences are required by default. However, the tool can disregard the absence of minor allele sequences by setting the option major-alleles-only at run time.

MAF Boxplot tool

The MAF Boxplot tool takes a table listing heteroplasmic sites per sample and their corresponding minor allele frequency (MAF) values. It generates a boxplot of the MAFs per sample by default. Optionally, it can generate a report including the total number of heteroplasmic sites and the median and median absolute deviation (MAD) of the MAFs per sample. Sites with a maximum of 2 alleles and an MAF ≥2% were selected from the table generated by the Variant Annotator tool. This table was used as an input to the MAF Boxplot tool to generate the graph and text report.

Results and discussion

The first warning sign of sample carryover is an unexpectedly high number of apparent variants. In a recent study utilizing a sequencing service provider, we analyzed a total of 56 mitochondrial DNA samples, representing blood and buccal cells. Normally, we expect a relatively small number of heteroplasmic sites per sample, with maternal transmission evidence and varying MAFs across sites (13,17). The sequencing reads were processed using a previously published workflow (13), which identifies heteroplasmies above a 2% frequency threshold. For example, a site may have 2.5% reads supporting an A allele and 97.5% reads supporting a G allele; in this case, A is the minor allele with frequency 2.5%. After identifying heteroplasmies, we have calculated the distribution of MAFs for each

individual and represented them as box plots in Figure 1. The immediately striking observation was the large number of heteroplasmic sites in many individuals (indicated along the x-axis). The situation in Figure 1 is rather extreme; based on previous studies (13,17) we expect ~0–3 heteroplasmies per individual. The boxplots highlight the fact that individuals with a high number of heteroplasmic sites have a narrow distribution of MAFs. This is indicative of carryover from another sample that differs from the one being analyzed at a number of fixed sites. For instance, if 2 samples differ at 10 non-polymorphic positions (i.e., they belong to 2 different mtDNA haplogroups) and there is carryover between them, the sites will appear as 10 heteroplasmies with identical MAFs. While Figure 1 clearly suggests such a problem, it does not identify the source of the putative contamination.

To understand the direction of carryover, we employed a phylogenetic approach. For each sample, we created an mtDNA sequence in which nucleotides at all detected heteroplasmic sites have been set to the major allele at that site. Applying the NJ phylogenetic tree reconstruction approach (16) to these sequences recapitulated family stratification of the samples as shown in Figure 2. Next, for each suspected instance of contamination such as samples F41M52 and F41M52C1, which have the narrowest distribution of MAFs as per Figure 1, we created another version of the mitochondrial genome by setting each heteroplasmic site to its minor allele nucleotide (termed F41M52_MINOR and F41M52C1_MINOR in Figure 2). Adding these sequences to the phylogenetic reconstruction showed that minor allele sequences for individuals F41M52 and F41M52C1, who belong to family F41, cluster with family F46 instead of F41, suggesting that these samples are contaminated by DNA originating from family F46 individuals (also see Supplementary Figure S1). In the case of this particular re-sequencing experiment, we have tracked the order of samples as they were sent to the sequencing facility. This allowed us to determine that these particular samples were located in adjacent cells on a 96-well plate. (Supplementary Figure S1 demonstrates the analysis of two additional samples with the number of minor alleles falling into a gray zone where one of the samples appears to be contaminated while the other is not.)

While the approaches described above seem to work well for controlling the data quality in re-sequencing experiments, we wanted to integrate them into a workflow that can be reproduced and re-used by others. Reproducibility is particularly important, as even with the latest advances in high-throughput sequencing studies such as those described here remain costly (18). Because the sequencing is often performed outside of the laboratory by an institutional core facility or a commercial sequencing provider, it is necessary to show where the problem occurred. Therefore, being able to run the contamination analysis in a transparent way such that all steps of the process can be reviewed and shared among involved parties becomes critical. The Galaxy platform (www.galaxy-project.org), developed and maintained by our group, is an ideal solution for implementing such a workflow. A Galaxy page at http://usegalaxy.org/contamination provides detailed description of a workflow that performs contamination analysis as described in this paper. This online document also provides original sequencing data that can be used to reproduce the results shown in Figures 1 and 2. By providing a turnkey solution to the detection of contamination and making suggestions for best practices in experiments, we hope to encourage reproducible and accurate studies that fully leverage these novel capabilities.

To supplement our ability to reliably detect contamination with an independent approach, we now routinely employ DNA spike-ins in our experiments. For this purpose, we chose DNA from the high copy number plasmid pUC18, a standard, readily available cloning vector, and double-stranded genomic DNA from bacteriophage ϕX174. These spike-ins lack extensive homology with human mtDNA or with each other and are added prior to the preparation of barcoded libraries in an alternating fashion. Spike-ins allow for straightforward detection of contamination by mapping all reads generated in an experiment against reference sequences, in this case from pUC18 and ϕX174.

Our experience with sequencing at external facilities indicates that contamination is a significant threat affecting outcome in a research study. We have adopted the following set of procedures for performing re-sequencing for rare variant detection in a large number of samples:

1. Utilize two types of spike-ins in a striped layout by adding spike-ins in a sequence (*i*) spike-in 1, (*ii*) spike-in 2, (*iii*) no spike-in. Make every effort to keep samples with the same spike-in from being in physical proximity to each other, such as adjacent wells within a 96-well plate or adjacent tubes in a PCR strip.
2. Physically separate samples expected to have a high degree of sequence homology. For example, in the case of mitochondrial DNA we determine haplogroups for our samples using Sanger sequencing prior to beginning the re-sequencing experiment. We use this information while handling the samples to make sure that samples belonging to similar haplogroups are not adjacent to each other. It is also advisable to sequence the mtDNA of the investigator performing the experiments in order to rule out an additional potential source of contamination.
3. Perform spike-in detection with a sensitive assay prior to sequencing but after library construction. This would avoid additional sequencing costs if contamination is detected.
4. After sequencing, map the reads against the reference genome, as well as sequences of spike-ins, and perform the analysis of distribution of MAFs. Identify suspicious samples with unusually high minor allele counts (≥10).
5. Perform the phylogenetic distribution analysis on suspicious samples to determine the source of contamination.

Our approach relies on the assumption that heteroplasmic sites are rare and exhibit over-dispersed MAFs. Our first method (Figure 1) identifies contamination by visualizing mutation frequency and MAF variation. Contamination is manifested by multiple polymorphic sites with a tight MAF distribution. This approach rapidly identifies the existence of contamination, but not the source. In contrast, previously deployed methods by Li et al. (17) and Avital et al. (19) identify contamination by assigning samples to Phylotree (20)-derived haplogroups (with Avital et al. utilizing Haplogrep) (21). While these methods offer the advantage that contamination can be identified from any source, they are of limited utility when an exhaustive list of haplotypes is unavailable (as might be expected for most heterogenous samples and certainly from samples drawn from recombining populations). Even if relevant databases are established, integrated with an analysis platform, and suitably

maintained, it would be relatively costly to implement a search across a large panel of samples. Furthermore, as the number of possible haplotypes and samples increases, interpretation would become challenging. Our simple approach is therefore more generalizable and scalable.

To determine the source of contamination in a sample flagged by our first method, our second method employs a phylogenetic approach. On the principle that the source and sink of contamination should cluster, we identify the most likely source of contamination for the focal sample. This contrasts with the above-mentioned methods based on haplogroup comparison in that our detection method does not depend on intervening data sets or structures. Li and Stoneking (22) likewise adopted a direct approach by searching all samples to identify those that explained a significant proportion (≥3) of apparent minor allele identities in potentially contaminated samples. Our approach is better suited to large data sets for two reasons. First, Li and Stoneking flag all samples with >5 polymorphic sites (verified using their bias statistics); this fixed threshold might lead to an unsustainable number of comparisons in a large data set (although this could be mitigated if our first method were used to select candidates based on MAF variation). Second, their approach entails repeated pairwise comparisons whereas ours jointly considers all intra-experiment hypotheses regarding the origins of contamination and displays the result in a single graphic.

While our methods cannot substitute for careful experimental controls and evaluation of raw data, we do believe they provide a broadly applicable two-step approach. Our first filter encourages the experimenter to consider both the numbers of polymorphic sites and the amount of MAF variation at these sites. By visualizing both sources of information, all suspicious samples can be identified together in a manner that is sensitive to the experimenter's expectations regarding polymorphism in the data set. (Although MAF variation is the more important of these two measures, we feel that a single summarizing statistic would be misleading and prefer a visual approach). Our second method can be selectively deployed by the researcher to identify sources of contamination in particular samples. This method also yields a visualization that represents the relative likelihood of contamination from other samples in the data set. Due to their ease of use, we hope that our computational tools, which are based on these methods, will become useful additions to the quality control toolkit that investigators use to examine next-generation sequencing data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Jessica Beiler for coordinating sample collection, to clinical nurses from Penn State College of Medicine's Pediatric Clinical Research Office for collecting the samples, and to volunteers for donating the samples. Efforts of the Galaxy Team (Enis Afgan, Dannon Baker, Dave Clements, Nate Coraor, Carl Eberhard, Dorine Francheteau, Jeremy Goecks, Sam Guerler, Greg Von Kuster, Jennifer Jackson, Ross Lazarus, and James Taylor) were instrumental for making this work happen. Special thanks to Dave Bouvier for helping to tune the Galaxy tools described in this manuscript. This work was funded by an NIH grant GM0722640 5S2 to KDM and an NIH grant HG004909 to AN, as well as funds from Penn State University and the Huck Institutes for the Life Sciences to AN, KDM. Additional funding was provided, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds and by the Penn State Clinical and Translational Science Institute

(CTSI). The Department specifically disclaims responsibility for any analyses, interpretations or conclusions. This paper is subject to the NIH Public Access Policy.

References

1. Bogenhagen DF. Mitochondrial DNA nucleoid structure. Biochim. Biophys. Acta. 2012; 1819:914–920. [PubMed: 22142616]
2. Legros F, Malka F, Frachon P, Lombès A, Rojo M. Organization and dynamics of human mitochondrial DNA. J. Cell Sci. 2004; 117:2653–2662. [PubMed: 15138283]
3. Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, et al. An enhanced MITOM A P with a global mtDNA mutational phylogeny. Nucleic Acids Res. 2007; 35(Database issue):D823–D828. [PubMed: 17178747]
4. Chinnery PF, Thorburn DR, Samuels DC, White SL, Dahl HM, Turnbull DM, Lightowlers RN, Howell N. The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? Trends Genet. 2000; 16:500–505. [PubMed: 11074292]
5. Jacobs HT. Making mitochondrial mutants. Trends Genet. 2001; 17:653–660. [PubMed: 11672866]
6. DiMauro S. Mitochondrial diseases. Biochim. Biophys. Acta. 2004; 1658:80–88. [PubMed: 15282178]
7. Jenuth JP, Peterson AC, Fu K, Shoubridge EA. Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA. Nat. Genet. 1996; 14:146–151. [PubMed: 8841183]
8. Cao L, Shitara H, Horii T, Nagao Y, Imai H, Abe K, Hara T, Hayashi J, Yonekawa H. The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells. Nat. Genet. 2007; 39:386–390. [PubMed: 17293866]
9. Cree LM, Samuels DC, de Sousa Lopes SC, Rajasimha HK, Wonnapinij P, Mann JR, Dahl HH, Chinnery PF. A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. Nat. Genet. 2008; 40:249–254. [PubMed: 18223651]
10. Wai T, Teoli D, Shoubridge EA. The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. Nat. Genet. 2008; 40:1484–1488. [PubMed: 19029901]
11. Cree LM, Samuels DC, Chinnery PF. The inheritance of pathogenic mitochondrial DNA mutations. Biochim. Biophys. Acta. 2009; 1792:1097–1102. [PubMed: 19303927]
12. Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C. The Pedigree Rate of Sequence Divergence in the Human Mitochondrial Genome: There Is a Difference Between Phylogenetic and Pedigree Rates. Am. J. Hum. Genet. 2003; 72:659–670. [PubMed: 12571803]
13. Goto H, Dickins B, Afgan E, Paul IM, Taylor J, Makova KD, Nekrutenko A. Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biol. 2011; 12:R59. [PubMed: 21699709]
14. Freeman B, Smith N, Curtis C, Huckett L, Mill J, Craig IW. DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. Behav. Genet. 2003; 33:67–72. [PubMed: 12645823]
15. Tanaka, M.; Hayakawa, M.; Ozawa, T. Methods in enzymology. Vol. 264. Elsevier; 1996. Automated sequencing of mitochondrial DNA.; p. 407-421.[Methods in Enzymology]
16. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 1987; 4:406–425. [PubMed: 3447015]
17. Li M, Schonberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am. J. Hum. Genet. 2010; 87:237–249. [PubMed: 20696290]
18. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat. Rev. Genet. 2012; 13:667–672. [PubMed: 22898652]
19. Avital G, Buchshtav M, Zhidkov I, Tuval Feder J, Dadon S, Rubin E, Glass D, Spector TD, Mishmar D. Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins. Hum. Mol. Genet. 2012; 21:4214–4224. [PubMed: 22736028]

20. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 2009; 30:E386–E394. [PubMed: 18853457]
21. Kloss-Brandstätter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum. Mutat. 2011; 32:25–32. [PubMed: 20960467]
22. Li M, Stoneking M. A new approach for detecting low-level mutations in next-generation sequence data. Genome Biol. 2012; 13:R34. [PubMed: 22621726]
23. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, Kivisild T, Torroni A, Villems R. A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. Am. J. Hum. Genet. 2012; 90:675–684. [PubMed: 22482806]
24. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20:289–290. [PubMed: 14734327]
25. Sukumaran J, Holder MT. DendroPy: A Python library for phylogenetic computing. Bioinformatics. 2010; 26:1569–1571. [PubMed: 20421198]

Method summary

Cross-contamination among samples in experiments involving next generation sequencing can readily masquerade as low level polymorphisms. In this manuscript we present a robust approach for identifying this type of contamination and tracking down its source. In addition, we provide a ready-to-use web platform for performing the analyses described here.

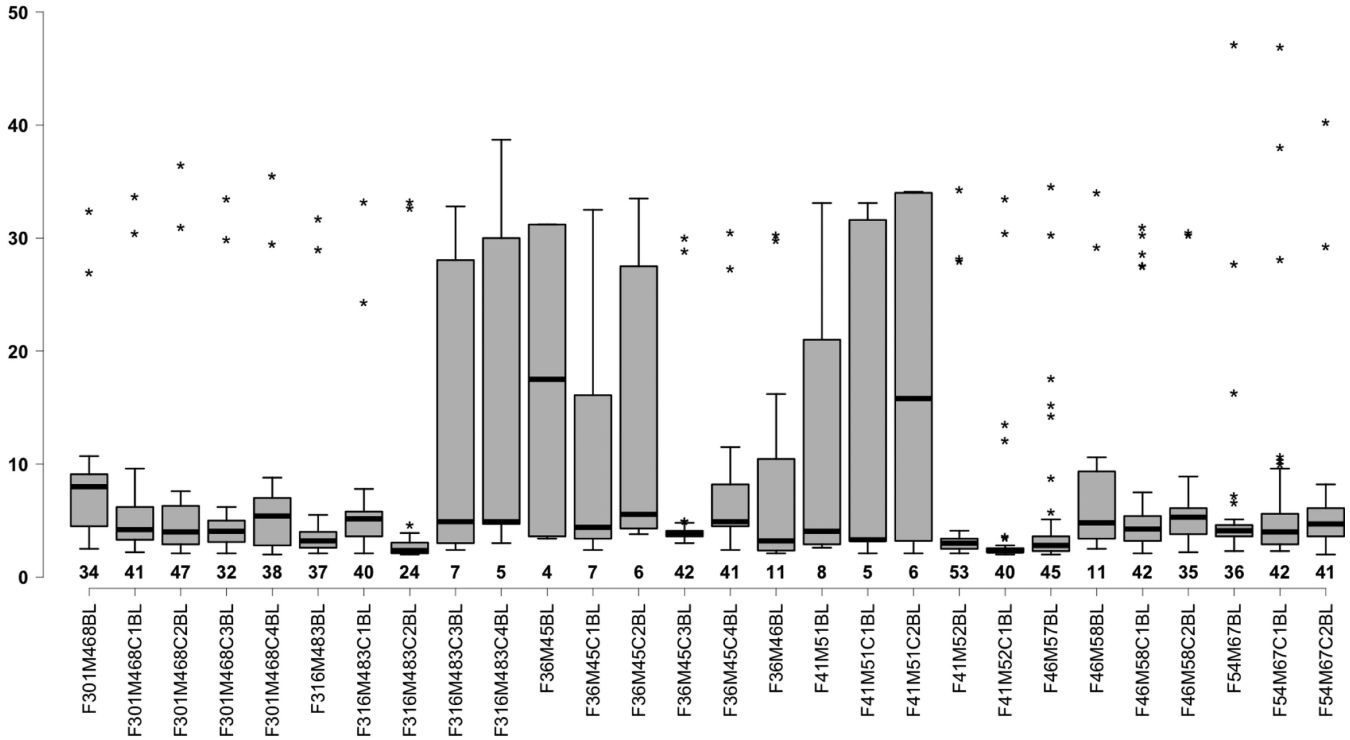

Figure 1. Boxplot summarizing the distribution of allele frequencies across samples in a contaminated mitochondrial re-sequencing study

The x-axis represents individual samples, with the numbers above each sample name indicating the number of detected heteroplasmies. The y-axis represents minor allele frequency.

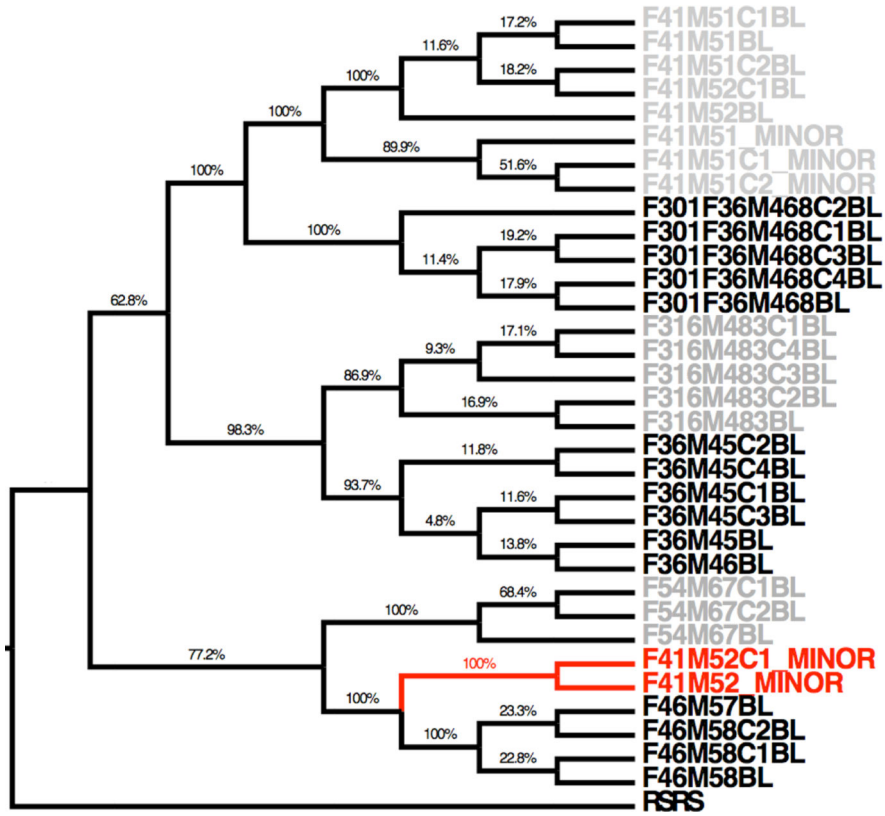

Figure 2. Phylogenetic analysis of minor allele consensus sequences for samples F41M52 and F41M52C1

Minor allele consensus sequences are shown in red on the background of major allele consensus sequences from all samples (shown in black). The numbers above branches and line thickness reflect bootstrap support (from 1000 iterations). Alternating black and gray lettering signifies distinct families used in the study. RSRS: a hypothetical version of mtDNA designed for rooting of phylogenetic trees (23).