scientific reports

Check for updates

OPEN

# The human "contaminome": bacterial, viral, and computational contamination in whole genome sequences from 1000 families

Brianna Chrisman[1]✉, Chloe He[2], Jae-Yoon Jung[3], Nate Stockham[4], Kelley Paskov[2], Peter Washington[1] & Dennis P. Wall[2,3]✉

**The unmapped readspace of whole genome sequencing data tends to be large but is often ignored. We posit that it contains valuable signals of both human infection and contamination. Using unmapped and poorly aligned reads from whole genome sequences (WGS) of over 1000 families and nearly 5000 individuals, we present insights into common viral, bacterial, and computational contamination that plague whole genome sequencing studies. We present several notable results: (1) In addition to known contaminants such as Epstein-Barr virus and phiX, sequences from whole blood and lymphocyte cell lines contain many other contaminants, likely originating from storage, prep, and sequencing pipelines. (2) Sequencing plate and biological sample source of a sample strongly influence contamination profile. And, (3) Y-chromosome fragments not on the human reference genome commonly mismap to bacterial reference genomes. Both experiment-derived and computational contamination is prominent in next-generation sequencing data. Such contamination can compromise results from WGS as well as metagenomics studies, and standard protocols for identifying and removing contamination should be developed to ensure the fidelity of sequencing-based studies.**

In the last decade, next-generation sequencing has become a commonly used tool in nearly every area of biology and has drastically changed the fields of human genomics[1,2], metagenomics[3,4], and pathogen surveillance[5,6]. Additionally, open-source access to many bioinformatics tools[7,8], benchmarking studies on the efficacy of computational pipelines[9–11], and improvements to laboratory procedures[7,8] have made many next-generation sequencing use cases reliable or nearly reliable enough to be used clinically[12].

As the number of NGS studies grows with the growing diversity of sequencing sites and protocols, many studies have found various sources of contamination in publicly available NGS data. Such studies have found bacterial contamination in laboratory reagents and sequencing kits[13,14], and common cross-contamination across samples[15]. In both genomic analysis of a single organism and metagenomics studies, such contamination can have critical impacts on downstream analysis. In whole-genome sequencing studies, bacterial contamination can result in false alignments and erroneous downstream variant calls[16,17]. In studies of the microbiome, contamination can distort the estimation of microbial abundance of different genera[13,18]. This is especially an issue for studies of microbiota that may have low microbial abundances, where even low levels of contamination may render metagenomics analysis inaccurate[19,20].

In addition to improving laboratory protocols to reduce contamination[21,22], several tools have been developed to identify contamination in next-generation sequencing data[23–25]. These tools rely on either sequencing a reagent-only or blank sample to determine baseline contaminant levels of microbes, or rely on a measurement of total on-target DNA in a sample and assume an inverse relationship between on-target DNA biomass and contaminants. While such tools have improved the reliability of several microbiome studies, their assumptions can break down in sequencing experiments with many different confounding variables not included in the controls[26], very low abundance microbes[18], or when contaminate mass is comparable to sample mass[27]. For one particularly controversial microbiome topic, despite the numerous decontamination techniques available, researchers have

[1]Department of Bioengineering, Stanford University, Stanford, USA. [2]Department of Biomedical Data Science, Stanford University, Stanford, USA. [3]Department of Pediatrics (Systems Medicine), Stanford University, Stanford, USA. [4]Department of Neuroscience, Stanford University, Stanford, USA. ✉email: briannac@stanford.edu; dpwall@stanford.edu

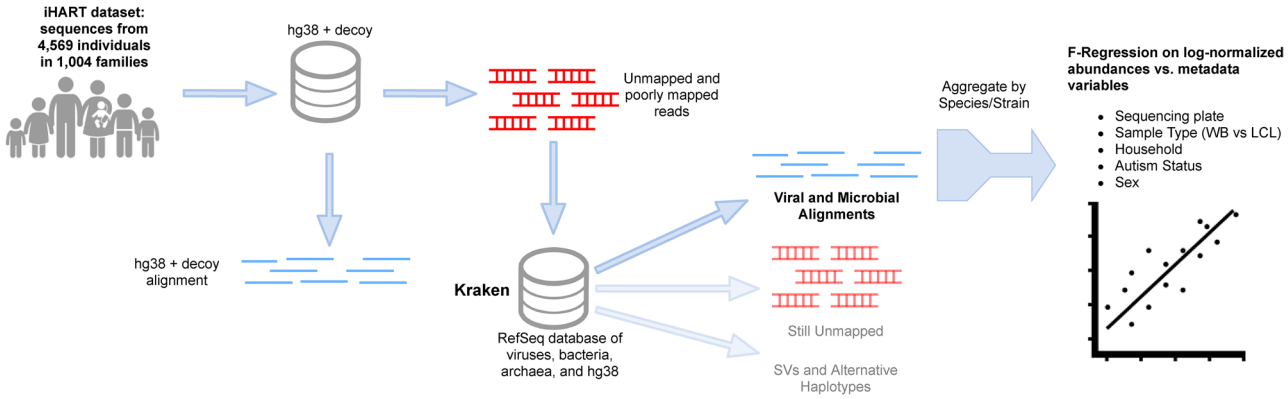


**Figure 1.** The general pipeline of the study: Reads from the iHART dataset that were unmapped or poorly aligned to GRCh38 were extracted and reclassified to a database of viruses, bacteria, archaea using Kraken. An F-regression was then performed on bacterial and viral counts against various sample-level metadata.

been unable to agree on whether there is in fact a core blood microbiome or if all studies claiming so have simply been detecting contaminants[28–32].

More alarming than the presence of contamination in individual sequencing studies is the presence of contamination in reference databases. Studies have identified human DNA contamination in non-primate reference genomes[33], millions of contaminate sequences in GenBank[34], and human contamination in bacterial reference genomes that has created thousands of spurious protein sequences[35]. Such contamination risks compromising the findings of any genomics study, even if the researchers properly decontaminated or controlled for contaminants.

In order to better understand patterns of contamination in human whole genome sequencing, we analyzed sequences from the iHART dataset[36]. Originally curated to study genetic determinants of autism, the iHART dataset contains whole genome sequences from blood samples from children with autism, their siblings, and their parents, but also stands as an invaluable genomics resource due to its unique family structure[37,38]. iHART was sequenced at the New York Genome Sequencing Center, a common site for large sequencing studies, using commonly followed storage, prep, and sequencing protocols[36], making it a good model dataset to understand common sequencing issues. In addition to its unique family structure, the iHART collection contains both whole blood (WB) samples and lymphoblastoid cell lines (LCLs), and contains experimental batch information such as sequencing plate. By realigning reads from the iHART collection that were unmapped or poorly mapped to the human reference genome to a collection of viral, bacterial, and archael sequences, we are able to identify particular signatures of contamination that are unique to metadata variables.

We confirm the presence of many contaminating microbes that have been noted in other studies, including *Mycoplasma*, *Burkholderia*, *Bradyrhizobium*, *Mezorhizobium*, and *Variovorax*. We note that several microbes are strongly associated with cell type, suggesting that the LCL and WB storage pipelines may have differential effects on contamination signatures, and sequencing plate, suggesting that batch contamination can be a major risk to sequencing studies. Finally, we show that over 100 bacteria falsely associate with sex, indicating that reads from poorly catalogued regions of the sex chromosomes inaccurately map to bacterial contigs. We extract the offending k-mers that contribute to these mismappings, and suggest that researchers performing metagenomics pipelines on low microbial load environments filter their reads to remove such reads.

## Results

**Viruses and bacteria commonly found in WGS.** Following the basic pipeline shown in Fig. 1, Kraken2, a k-mer-based read classifier, classified many reads as belonging to bacteria and viruses (Fig. 2). The median number of reads per sample was $7.6x10^8$ [$6.3x10^8$ - $8.9x10^8$ ]. Of the median $1.2x10^7$ [$8.9x10^6$ - $1.7x10^7$] unmapped or poorly mapped reads per sample, a median of 37% [25%–44%] still matched to GRCh38 better than any other organisms, 0.03% [0.01%–0.08%] were reclassified as viruses, 21% [15%–42%] as bacteria or archaea, 9% [2%–17%] mapped ambiguously to organisms from multiple kingdoms, and 27% [19.2%–40%] remained unmapped. Although some reads were classified ambiguously (with their set of k-mers matching equally well to sequences from multiple kingdoms), most reads were able to be classified to the species or strain level (58% [44%–77%] of bacterial, viral, and archaeal reads that Kraken reclassified were reclassified to the species/strain level). Therefore, we aggregated our reads by their lowest taxonomic classification.

We saw two main categories of viruses in the unmapped read space: DNA viruses likely originating from the human virome (such as human herpesviruses 6 and 7 as well as torque teno viruses), and common reagents used in the sequencing pipeline (such as lambda phage and herpesvirus 4). Phi X lambda phage is used as a spike-in for GC content in Illumina sequencing pipelines as well as to calibrate sequencing machines[39]. Herpesvirus 4, or Epstein Barr Virus (EBV) is used to immortalize LCLs[40]. Other phages or relatives to herpesvirus 4 are likely due to either mismappings, or commercial contamination, which we discuss more in the "Discussion". Although the median number of reads belonging to viruses was small, our samples showed a wide range of viral read counts spanning over 4 orders of magnitude. The main contributors to this are lambda phage, which has a large variance across samples and EBV, in which unsurprisingly LCL samples have much higher read counts over whole

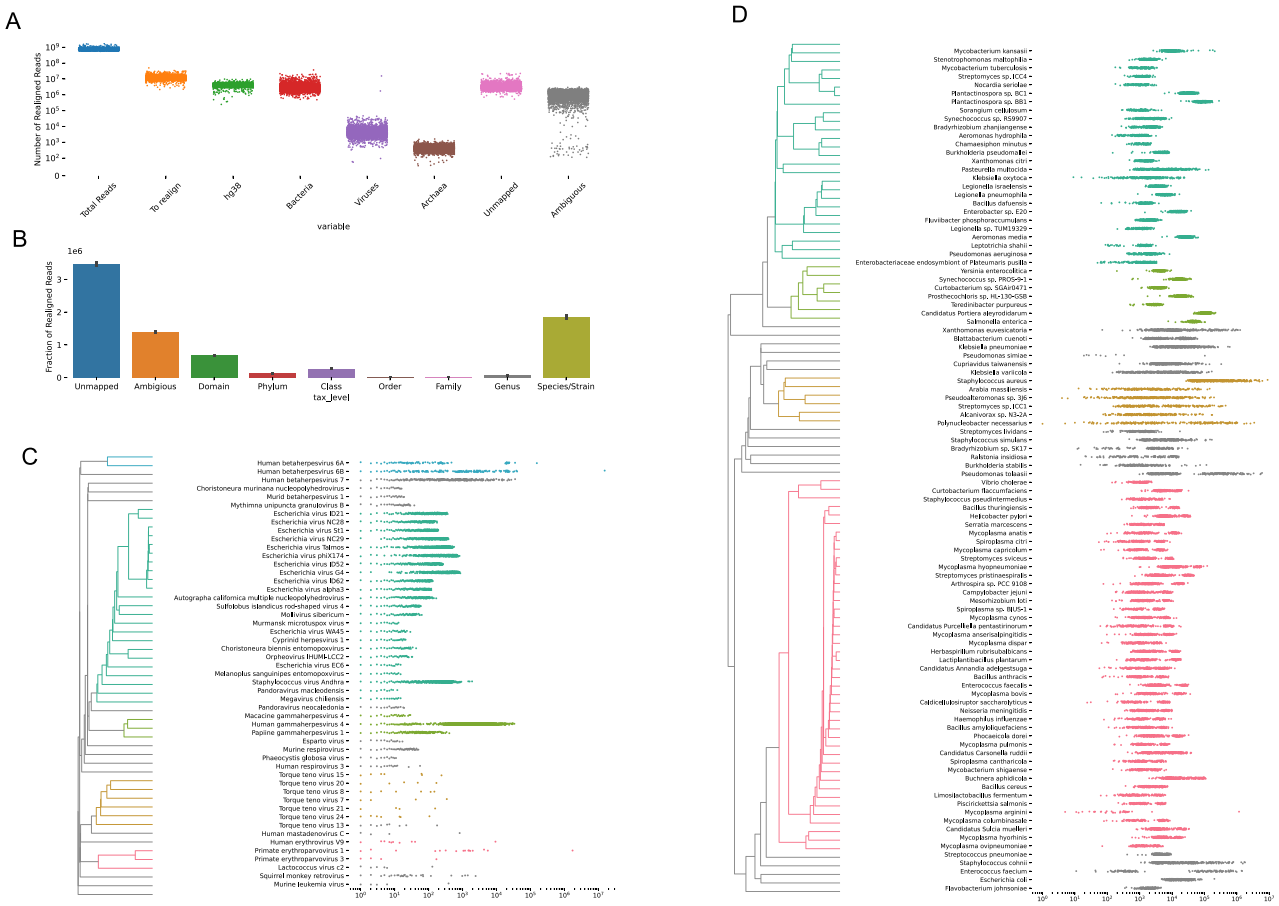


**Figure 2.** (**A**) Number of reads originally unmapped or poorly aligned to GRCh38 that Kraken2 classified as belonging to human, bacterial, viral, and archaeal sequences. (**B**) Taxonomic levels of Kraken2 classifications. While a significant fraction of reads were classified as ambiguous, Kraken2 was able to classify the majority of reads down to the species or strain level. (**C**) The top 50 most abundant viruses based on read count, clustered by Spearman correlation of sample abundances. Each point represents a sample's read count. (**D**) Top 100 most abundant bacteria, archaea, and lower eukaryotes based on read counts and clustered by Spearman correlation of sample abundances.

blood samples. Human herpesviruses 6A, 6B, and 7 also have large variances across samples, likely depending on whether an individual has a latent infection, active infection, or inherited chromosomally integrated human herpesvirus (iciHHV)[41].

We found many bacteria that were highly abundant in our samples. Notably, the top 100 most abundant bacteria also appeared in >90% of our samples, and most appeared in 100% of samples. Although it is possible these bacteria are part of the natural blood microbiome, it seems far more likely these bacteria are due to contamination during the sample collection, storage, and sequencing pipelines. First of all, theoretically, small traces of true bacteremia originally found in a blood sample should have been removed during sterilization steps in sample collection and prep, particularly during the immortalization step in the LCL samples. Importantly, nearly all bacteria found had a strong association with sequencing plate or cell type rather than household, indicating that it is probably the experimental pipeline that is driving the bacteria abundances. We discuss this in detail in the next section. Finally, the types of bacteria and their abundances have profiles more similar to common water and oral cavity contaminants than the bacteria observed in even the controversial blood microbiome studies. In particular, we find many species of *Mycoplasmsa*, *Bradyrhizobium*, *Mycobacterium*, *Staphylococcus*, *Streptomyces*, *Streptococcus*, and *Pseudomonas* (Fig. 2C). Such bacteria are common water contaminants or either commonly found in human respiratory and oral cavities, and likely originated from reagent contamination or contamination introduced by a human experimenter. Many of the same contaminants we found were also found in other large scale WGS or metagenomics studies. We elaborate further on in the "Discussion".

**Sample type and sequencing plate influence microbial contamination profile.** Using an forward F-regression, we found that sample type (LCL vs WB) and sequencing plate strongly influenced the abundances of many bacterial contaminants (Fig. 3). In particular, several species of *Achromobacter*, *Bradyrhizobium*, and *Burkholderia* were more abundant in whole blood samples (Fig. 3A), and several species of *Psuedomonas*, *Streptomyces*, and *Xanthomonas* were more abundant in LCL samples (Fig. 3B). Species of *Acidovorax*, *Bradyrhizobium*, *Mesorhizoium*, and *Variovorax* had different abundances according to sequencing plate (Fig. 3C).

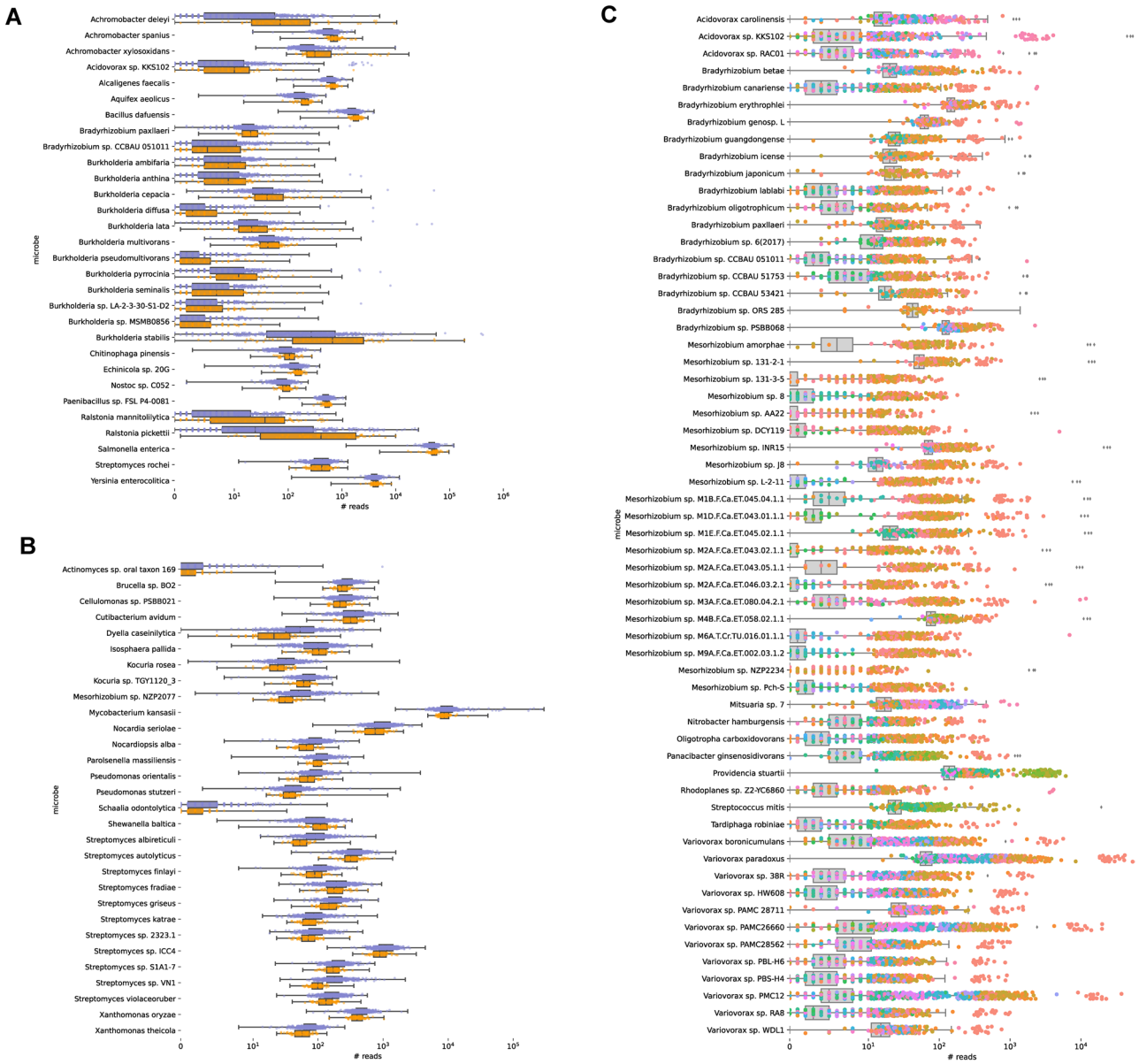


**Figure 3.** (**A**) Top 50 microbes most strongly associated with cell type per F-regression results, and more abundant in whole blood samples (orange) than LCL samples (purple). (**B**) Top 50 microbes most strongly associated with cell type per F-regression results, and more abundant in whole blood samples (orange) than LCL samples (purple). (**C**) Top 100 microbes most strongly associated with sequencing plate. Colors represent sequencing plates that had significantly higher abundances of a given microbe compared to the rest of the population. Samples from sequencing plates without significant enrichment of a microbe are captured in the grey box plots.

**Sex chromosome fragments mismap to bacterial reference genomes.** Upon the F-regression showing many bacteria strongly associated with sex, we hypothesized that this was due to reads from the sex chromosomes being misclassified as bacteria. We show the male and female read counts for the bacteria most strongly associated with sex in (Fig. 4A,B). Furthermore, we found that many bacteria had abundances strongly correlated between fathers and sons from the same nuclear family (an example is shown in Fig. 4D). The Y-chromosome has a notoriously poor reference genome with only about half its sequence present in GRCh38, and also has many repeats. We hypothesize that bacteria with high correlation between father and son read counts are due to repetitive regions in the Y-chromosome being misclassified as bacterial sequences. The number of repeats is passed down the male family line, and thus would be correlated between father and son. Interestingly, we also found that many bacteria had strong correlations between mother/daughter, and father/son (but not between father/daughter or mother/son) (Fig. 4C). An example of this is shown in Fig. 4E). We hypothesize that reads mismapping to these bacteria may be coming from homologous sequences present on both X and Y chromosomes, with more repeats on the Y. Mother/daughter read counts would therefore show a mild correlation, but a

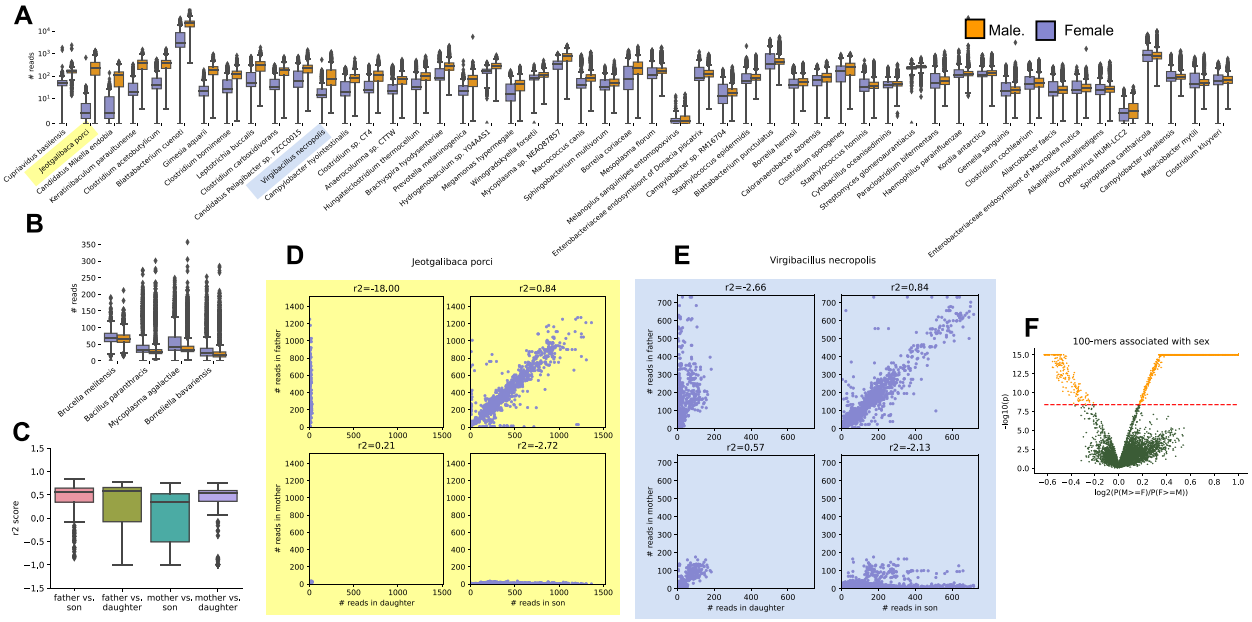


**Figure 4.** (**A**) Bacteria significantly enriched in males, subsetted to the 50 bacteria with the strongest association. Examples of bacteria with distinct inheritance patterns are highlighted in yellow and blue and shown in **D**–**E**. (**B**) Bacteria that were found to be significantly enriched in females (purple). (**C**) Boxplot of correlations between children and parents of different sexes, as measured by the $R^2$ metric of log-normalized counts. (**D**) Example of a bacteria with inheritance patterns illustrative of y-chromosome repeats mismapping to bacterial contigs. (**E**) Example of a less clear inheritance pattern, which we hypothesize results from a sequence present both on the X and Y chromosome in varying numbers of repeats. (**F**) A sub-sampled volcano plot for the association between counts of 100-mers extracted from reads mapping to sex-associated bacteria and sex, using a paired analysis of siblings. The red line represents the Bonferonni-adjusted p-value cutoff, with statistically significant hits in orange.

father/daughter or mother/son correlation would get watered down by the large number of repeats coming from the Y-chromosome.

Regardless of inheritance mechanism, we sought to identify particular sequences leading to these problematic mismappings in order to help other researchers prevent sex-biased errors in future studies. We extracted 100-basepair sequences from the reads that mapped to any of the sex-associated bacteria. We then performed a strict paired analysis using males and females of the same autism phenotype from the same family, analyzing differences in coverage of a given 100-mer. We identified 77,647 100-mers significantly enriched with males and 369 in females (adjusted p-value < 0.05). We make these k-mers publicly available and recommend that studies susceptible to such sex-biased mismappings mask reads containing these problematic sequences.

## Discussion

Many of the contaminants found in our data have been documented in other large scale genomic studies. One study[42] found several phages with abundances correlated to that of phiX, similar to what we found. The authors interpreted this as contamination of the commercial preparation of phiX174. *Mycoplasma* contamination has been found in many cell lines[43], and *Bradyrhizobium* was the most common contaminant found in WGS from the 1000 genomes project[19]. *Staphylococcus*, *Acinetobacter*, *Streptococcus*, and *Pseudomonas* have been identified as possible contaminants in several WGS studies[44–46]. Despite numerous studies cataloguing bacterial contamination in NGS, the same species seem to persist in sequencing contamination.

NGS of large case control cohorts become one of the most popular study designs in studies of human health. Whether hoping to identify variants in the human genome contributing to disease risk, gene expression profiles of particular phenotype, or understand causal effects of various microbiome signatures, one of the first steps in NGS pipelines is to typically to align raw reads to a set of reference databases. Unchecked bacterial contamination in NGS can compromise NGS in a variety of ways: In metagenomics studies, bacterial contamination from laboratory reagents can distort abundance counts of microbes in the samples[14], and in the worst case can lead to spurious associations between disease and microbial signature[47]. In human genetics studies, contamination mismapped to the human reference genome can lead to false variant calls[16], and different amounts of contamination across samples makes it difficult to maintain consistent coverage levels across samples. Decontamination software packages may help with some of these issues but special care must be taken to sequence a control sample for all combinations of sequencing plates and sample storage and prep, as these different experimental parameters have clear differences in bacterial contamination signatures. Meticulous paired study designs controlling for the potential for contamination in different steps of the pipeline (ie sequencing each case/control pair on the same plate, extracting and prepping the samples on the same timeline) may also reduce the risk of contamination causing a

false association between microbial signature and disease. Regardless of study design, in studies of microbiota with low bacterial load, contamination from laboratory regents can limit identification of related low abundance microbes[20]. More needs to be done to understand and mitigate laboratory and reagent sources of contamination.

Poor quality reference genomes pose an additional set of risks for next gen sequencing studies. Misconstructed reference genomes that are actually chimeras of several organisms can result in incorrectly identifying which microbes are present[34] and[35]. Incomplete reference genomes, such as the human Y chromosome, may result in mismappings from reads coming from poorly catalogued sections of genome to satisfactorily similar sequences on well-characterized reference genomes.. We have identified tens of thousands of 100-bp sequences likely originating from the sex chromosomes that mismap to bacterial contigs. It is possible these mismappings are due to poorly constructed bacterial reference genomes that actually contain human DNA sequences[33], mismappings from Y chromosome reads as a result of its incomplete reference, or a combination of both. Regardless, we have made these problematic sequences available in a fasta file format. Many read masking and trimming tools, such as `BBTools`[48,49], `Trimmomatic`[50], and `Cutadapt`[51], can take in a `fasta` file of adaptor or contaminant sequences and remove reads that contain any of the problematic sequences. We recommend metagenomics or other studies performing alignment of reads derived from a human host remove reads with these problematic sequences, in order to reduce potential sex-related artifacts. This is particularly important in studies of microbiota with low bacteria-to-host-DNA ratios, such as the blood microbiome.

Unmapped reads can constitute up to 10% of WGS data, and usually are thrown out in downstream analysis. With the wealth of WGS data that has been and continues to be generated, this unmapped read space composes several petabytes of data. We, and others[45] have shown that the unmapped read space is a valuable resource for quantifying contamination that might pollute NGS studies. The unmapped read space may also be a valuable resource for better understanding the virome[42,52] as well as host genetic diversity[52–54], especially with the help of a well-characterized contamination profile.

The unmapped read space of WGS contains information on common contaminants of WGS. Contamination profiles depend on primarily cell source type and sequencing plate. Additionally, many sequences from the Y-chromosome mismap to bacterial contigs, creating problematic sex-biased bacteria counts. The unmapped read space is a valuable resource for better understanding ubiquitous contamination patterns in WGS.

## Methods

**Extracting unmapped and poorly aligned reads.** We obtained Whole Genome Sequencing (WGS) data from the Hartwell Autism Research and Technology Initiative (iHART) database, which includes 4842 individuals from 1050 multiplex families in the Autism Genetic Resource Exchange (AGRE) program 1C.

All WGS data from the iHART database have been previously processed using a standard bioinformatics pipeline which follows GATK's best practices workflows.Specifically, We used the iHART WGS collection[36], a dataset of multiplex autism families. Individuals were sequenced at 30 × coverage using Illumina's TruSeq Nano library kits, reads were aligned to build GRCh38 of the reference genome and decoy contigs (GRCh38_full_analysis_set_plus_decoy_hla.fa) using bwa-mem[55], and variants were called using GATKv3.4. We excluded secondary alignments, supplementary alignments, and PCR duplicates from downstream analyses. We extracted reads from the iHART genomes that were unmapped or mapped with low confidence. Low-confidence reads were defined as reads marked as improperly paired and with an alignment score below 100. We used alignment score rather than mapping quality in order to select for reads were likely not true alignments to the human reference genome, rather than for reads that had ambiguous alignments to GRCh38.

**Re-alignment.** We used Kraken2[56] to align the unmapped and poorly aligned reads to a the Kraken default (RefSeq) databases of archaeal, bacterial, human (GRCh38.p13), and viral sequences[57]. These reference databases were accessed on Feb 16, 2021. Kraken2 was run on the unmapped and poorly mapped reads from each sample, using the default parameters. Because Kraken was able to map the majority of reads down to the species or strain level, Kraken classifications were aggregated by species before downstream analysis.

**F-regression.** To analyze the effect of various demographic (such as household, autism status, and sex) and experimental parameters (such as sequencing plate and sample type) on microbial and viral profile, we performed an F-regression analysis. We chose an F-regression because many variables were highly collinear with each other: for example, samples from the same household were nearly always sequenced on the same sequencing plate, autism is much more prevalent in males, and the same sample types were normally collected from households. For each microbe, we built an ordinary least squares (OLS) model, using as our regressor an indicator matrix of sample type, sex, child vs. parent, autism status, sequencing plate, household/family, and sample id, and as our response variable the log-normalized counts of microbes (with pseudo-counts of 1). Using the `statsmodels` library[58], we then ran a forward OLS regression in which we iteratively selected the regressor features that best explained the previous model's residuals, and ceased adding features when the ANOVA score between the previous and new models was no longer statistically significant ($p<.05$)[59].

**Y-chromosome mismapping analysis.** Using the F-regression, we found that many microbes were significantly associated with sex (162 species were enriched in males and 4 species were enriched in females. Hypothesizing that such mismappings were due to mismappings of repetitive regions on the X or Y chromosome, we analyzed inheritance patterns, looking at the correlation between children and parents using the r2 score (as shown in Fig. 4. Furthermore, we sought to identify specific subsequences that cause these problematic bacterial classifications. From the collection of reads that aligned to the bacterial reference contigs associated with sex, we extracted and counted the occurrence of 100-basepair k-mers in every sample. We counted the

100mers using the highly parallel k-mer counter `jellyfish`. We chose 100-basepair k-mers because assuming a uniform distribution of 150b reads across the human genome at 30x coverage and a trivial sequencing error rate, 100 bases is the longest length of a k-mer with over a 99.5% chance of being captured within the 150 bases of at least one read in an individual. To reduce k-mers generated by sequencing error or low frequency genetic variants, we filtered to 100-mers that occurred at least twice in at least two samples. In order to test the null hypothesis that these subsequences show equal occurrences in males and females, we then performed a paired test between males and females siblings within the same family with the same autism status (to stringently weed out ancestry and disease phenotype as confounding variables). We reported the 100-mers that had a Bonferonni-adjusted p-value <.05, and make them publicly available for access in a "'fasta"' format that can easily be access by read trimming and masking tools. These sequences are available at the link described below.

## Data availability

The iHART dataset is available upon reasonable request at http://www.ihart.org/home. The complete list of reference genomes used for Kraken realignment can be found at: http://github.com/briannachrisman/blood_microbiome/public_data/kraken_db. The sequences associated with sex can be found at: http://github.com/briannachrisman/blood_microbiome/public_data/x_sequences.fasta and http://github.com/briannachrisman/blood_microbiome/public_data/y_sequences.fasta.

Received: 11 January 2022; Accepted: 18 May 2022
Published online: 14 June 2022

## References

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* https://doi.org/10.1038/s41586-019-1879-7 (2020).
2. DiResta, C., Galbiati, S., Carrera, P. & Ferrari, M. Next-generation sequencing approach for the diagnosis of human diseases: Open challenges and new opportunities. *Electron. J. Int. Fed. Clin. Chem. Lab. Med.* **29**(1), 4–14 (2018).
3. Ji, B. & Nielsen, J. From next-generation sequencing to systematic modeling of the gut microbiome. *Front. Genet.* https://doi.org/10.3389/fgene.2015.00219 (2015).
4. Kim, Y., Koh, I. S. & Rho, M. Deciphering the human microbiome using next-generation sequencing data and bioinformatics approaches. *Methods* **79**, 52–59. https://doi.org/10.1016/j.ymeth.2014.10.022 (2015).
5. Moran-Gilad, J. Whole genome sequencing (WGS) for food-borne pathogen surveillance and control: Taking the pulse. *Eurosurveillance* https://doi.org/10.2807/1560-7917.ES.2017.22.23.30547 (2017).
6. Maljkovic Berry, I. *et al.* Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: Approaches, applications, and considerations for development of laboratory capacity. *J. Infect. Dis.* https://doi.org/10.1093/infdis/jiz286 (2020).
7. Da Veiga Leprevost, F. *et al.* BioContainers: An open-source and community-driven framework for software standardization. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btx192 (2017).
8. Kulkarni, N. *et al.* Reproducible bioinformatics project: A community for reproducible bioinformatics analysis pipelines. *BMC Bioinform.* https://doi.org/10.1186/s12859-018-2296-x (2018).
9. D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* https://doi.org/10.1186/s12864-015-2194-9 (2016).
10. Zhao, S., Agafonov, O., Azab, A., Stokowy, T. & Hovig, E. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci. Rep.* https://doi.org/10.1038/s41598-020-77218-4 (2020).
11. Thankaswamy-Kosalai, S., Sen, P. & Nookaew, I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics* https://doi.org/10.1016/j.ygeno.2017.03.001 (2017).
12. Gu, W., Miller, S. & Chiu, C. Y. Clinical metagenomic next-generation sequencing for pathogen detection. *Annu. Rev. Pathol. Mech. Dis.* https://doi.org/10.1146/annurev-pathmechdis-012418-012751 (2019).
13. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* https://doi.org/10.1186/s12915-014-0087-z (2014).
14. de Goffau, M. C. *et al.* Recognizing the reagent microbiome. *Nat. Microbiol.* https://doi.org/10.1038/s41564-018-0202-y (2018).
15. Merchant, S., Wood, D. E. & Salzberg, S. L. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* https://doi.org/10.7717/peerj.675 (2014).
16. Goig, G. A., Blanco, S., Garcia-Basteiro, A. L. & Comas, I. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC Biol.* https://doi.org/10.1186/s12915-020-0748-z (2020).
17. Samson, C. A., Whitford, W., Snell, R. G., Jacobsen, J. C. & Lehnert, K. Contaminating DNA in human saliva alters the detection of variants from whole genome sequencing. *Sci. Rep.* https://doi.org/10.1038/s41598-020-76022-4 (2020).
18. McArdle, A. J. & Kaforou, M. Sensitivity of shotgun metagenomics to host DNA: Abundance estimates depend on bioinformatic tools and contamination is the main issue. *Access Microbiol.* https://doi.org/10.1099/acmi.0.000104 (2020).
19. Laurence, M., Hatzis, C. & Brash, D. E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE* https://doi.org/10.1371/journal.pone.0097876 (2014).
20. Eisenhofer, R. *et al.* Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends Microbiol.* https://doi.org/10.1016/j.tim.2018.11.003 (2019).
21. Reigel, A. M., Owens, S. M. & Hellberg, M. E. Reducing host DNA contamination in 16S rRNA gene surveys of anthozoan microbiomes using PNA clamps. *Coral Reefs* https://doi.org/10.1007/s00338-020-02006-5 (2020).
22. Ji, X. C. *et al.* Reduction of human DNA contamination in clinical cerebrospinal fluid specimens improves the sensitivity of metagenomic next-generation sequencing. *J. Mol. Neurosci.* https://doi.org/10.1007/s12031-019-01472-z (2020).
23. Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M. & Kang, H. M. Correcting for sample contamination in genotype calling of DNA sequence data. *Am. J. Hum. Genet.* https://doi.org/10.1016/j.ajhg.2015.07.002 (2015).
24. Martí, J. M. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1006967 (2019).
25. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* https://doi.org/10.1186/s40168-018-0605-2 (2018).
26. Karstens, L. *et al.* Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems* https://doi.org/10.1128/msystems.00290-19 (2019).
27. Zinter, M. S., Mayday, M. Y., Ryckman, K. K., Jelliffe-Pawlowski, L. L. & Derisi, J. L. Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome* https://doi.org/10.1186/s40168-019-0678-6 (2019).

28. Castillo, D. J., Rifkin, R. F., Cowan, D. A. & Potgieter, M. The healthy human blood microbiome: Fact or fiction?. *Front. Cell. Infect. Microbiol.* https://doi.org/10.3389/fcimb.2019.00148 (2019).
29. Païssé, S. *et al.* Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion* https://doi.org/10.1111/trf.13477 (2016).
30. Schierwagen, R. *et al.* Trust is good, control is better: Technical considerations in blood microbiome analysis. *Gut* https://doi.org/10.1136/gutjnl-2019-319123 (2020).
31. Schierwagen, R. *et al.* Circulating microbiome in blood of different circulatory compartments. *Gut* https://doi.org/10.1136/gutjnl-2018-316227 (2019).
32. Hornung, B. V. H., Zwittink, R. D., Ducarmon, Q. R. & Kuijper, E. J. Response to: 'Circulating microbiome in blood of different circulatory compartments by Schierwagen et al.'. *Gut* https://doi.org/10.1136/gutjnl-2019-318601 (2020).
33. Longo, M. S., O'Neill, M. J. & O'Neill, R. J. Abundant human DNA contamination identified in non-primate genome databases. *PLoS ONE* https://doi.org/10.1371/journal.pone.0016410 (2011).
34. Steinegger, M. & Salzberg, S. L. Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* https://doi.org/10.1186/s13059-020-02023-1 (2020).
35. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* https://doi.org/10.1101/gr.245373.118 (2019).
36. Ruzzo, E. K. *et al.* Inherited and de novo genetic risk for autism impacts shared networks. *Cell* **178**, 850–866. https://doi.org/10.1016/j.cell.2019.07.015 (2019).
37. Paskov, K. *et al.* Estimating sequencing error rates using families. *BioData Mining* **14**, 1–10. https://doi.org/10.1186/s13040-021-00259-6 (2021).
38. Chrisman, B. *et al.* Analysis of sex and recurrence ratios in simplex and multiplex autism spectrum disorder implicates sex-specific alleles as inheritance mechanism. In *Proceedings: 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, 1470–1477, https://doi.org/10.1109/BIBM.2018.8621554 (2019).
39. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by illumina Phix control. *Standard. Genom. Sci.* https://doi.org/10.1186/1944-3277-10-18 (2015).
40. Sugimoto, M., Tahara, H., Ide, T. & Furuichi, Y. Steps involved in immortalization and tumorigenesis in human B-lymphoblastoid cell lines transformed by Epstein–Barr virus. *Cancer Res.* https://doi.org/10.1158/0008-5472.CAN-04-0079 (2004).
41. Pantry, S. N. & Medveczky, P. G. Latency, integration, and reactivation of human herpesvirus-6. *Viruses* https://doi.org/10.3390/v9070194 (2017).
42. Moustafa, A. *et al.* The blood DNA virome in 8000 humans. *PLoS Pathog.* https://doi.org/10.1371/journal.ppat.1006292 (2017).
43. Nikfarjam, L. & Farzaneh, P. *Prevention and Detection of Mycoplasma Contamination in Cell Culture* (Springer, 2012).
44. Ashokan, A. *et al.* Case report: Identification of intra-laboratory blood culture contamination with *Staphylococcus aureus* by whole genome sequencing. *Diagn. Microbiol. Infect. Dis.* https://doi.org/10.1016/j.diagmicrobio.2019.02.016 (2019).
45. Sangiovanni, M., Granata, I., Thind, A. S. & Guarracino, M. R. From trash to treasure: Detecting unexpected contamination in unmapped NGS data. *BMC Bioinform.* https://doi.org/10.1186/s12859-019-2684-x (2019).
46. Strong, M. J. *et al.* Microbial contamination in next generation sequencing: Implications for sequence-based analysis of clinical samples. *PLoS Pathog.* https://doi.org/10.1371/journal.ppat.1004437 (2014).
47. Robinson, K. M., Crabtree, J., Mattick, J. S., Anderson, K. E. & Hotopp, J. C. Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. *Microbiome* https://doi.org/10.1186/s40168-016-0224-8 (2017).
48. Bushnell, B. BBTools suite (2014).
49. Clum, A. *et al.* DOE JGI metagenome workflow. *mSystems* https://doi.org/10.1128/msystems.00804-20 (2021).
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btu170 (2014).
51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* https://doi.org/10.14806/ej.17.1.200 (2011).
52. Laine, V. N., Gossmann, T. I., Van Oers, K., Visser, M. E. & Groenen, M. A. Exploring the unmapped DNA and RNA reads in a songbird genome. *BMC Genomics* https://doi.org/10.1186/s12864-018-5378-2 (2019).
53. Hasan, M. S., Wu, X. & Zhang, L. Uncovering missed indels by leveraging unmapped reads. *Sci. Rep.* https://doi.org/10.1038/s41598-019-47405-z (2019).
54. Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* https://doi.org/10.1038/ng.3801 (2017).
55. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. http://arxiv.org/abs/1303.3997 [q-bio.GN] (2013).
56. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* https://doi.org/10.1186/s13059-019-1891-0 (2019).
57. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv1189 (2016).
58. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-011 (2010).
59. Bendel, R. B. & Afifi, A. A. Comparison of stopping rules in forward "stepwise" regression. *J. Am. Stat. Assoc.* https://doi.org/10.1080/01621459.1977.10479905 (1977).


## Acknowledgements

Thank you to The Hartwell Foundation for supporting the creation of the iHART database and the Simons Foundation for additional support for genome sequencing. We thank the New York Genome Center for conducting sequencing and initial quality control of the iHART dataset. We thank Amazon Web Services for their grant support for the computational infrastructure and storage for the iHART database. This work has been supported by grants from The Hartwell Foundation and the NIH (U24 MH081810, R01MH064547, NS101158, NS070911, NS101665, NS095824, S10OD011939, P30AG10161, R01AG17917, and U01AG61356 ) and from the Stanford Precision Health and Integrated Diagnostics Center and from the Stanford Bio-X Center.

## Author contributions

B.S.C. wrote the manuscript. B.S.C., C.H., and J.J. wrote and developed the source code. D.P.W., N.S., K.P., and P.W. contributed to the conception and design of the study, facilitated collection and sequencing of autism dataset, and participated in the analysis of the results. All authors read and approved the final manuscript.

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to B.C. or D.P.W.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.