RESEARCH

Squeegee: A Computational Contamination Detection Tool for Metagenomic Samples

Correspondence:
Full list of author information is available at the end of the article
*Equal contributor

Abstract

Contaminant sequences in metagenomics can have great impact on metagenomic studies, especially in low biomass environment. Common methods to remove contamination from samples requires experimental negative controls, which is time consuming and expensive. Here we present Squeegee, a computational contamination detection tool that identifies contaminants in metagenomic samples by searching shared organisms from multiple samples which used same DNA extraction kit. This tool have the potential to reduce the usage of experimental negative control for metagenomics studies.

Keywords: contamination; DNA extraction kit; metagenomics

Background

In metagenomic studies, researchers are often interested in genetics sequences that are presented in metagenomic environment. One commonly used methods is 16S rRNA gene sequencing, since 16S rRNA gene is highly conserved in bacteria and can be amplified and used as a marker gene for taxonomic classification [1, 2, 3, 4, 5, 6]. The other widely used technique is whole-genome shotgun sequencing, where all DNA sequences in the community are fragmented and sequenced [1, 2, 7, 8, 9]. However, the result from both of those methods can be affected by contamination, which are the sequences that appears in the data but not in the original samples [2].

Contamination can be brought in from different source. The external source includes human bodies, laboratory environment, and kits and reagents used for collecting samples [2, 1, 10, 11, 12, 13, 14, 15, 16] .The internal source of contamination is often caused by mixing up of different samples during the experiment process, including sampling and sequencing [2, 15, 17]. Contaminating sequences have also made there way into public reference databases [18].

Studies have shown that contaminant in DNA extraction kits is ubiquitous, and can have critical impacts on metagenomic studies, especially for low-bio mass environments [19, 1]. In a recent nasopharyngeal microbiota study on new born babies conducted in Thailand, contaminants found in DNA extraction kit caused bias on their initial data analysis[2, 20].

While researchers are taking extra precaution in processing the samples during metagenomic experiments, computer scientists have also build different computational models to identify and remove the contaminating sequencing from the datasets. A recent published software called Recentrifuge uses a score-oriented comparative approach to identify and remove contaminant from the samples. [19]. Experimental controls are required in order to perform contamination removal with

Recentrifuge. Another statistical tool to identify and remove the contamination from metagenomic dataset called Decontam was introduced in the year 2018. Decontam included a combination of frequency-based approach and a prevalence-based approach [2]. Multiple sequencing runs on the same sample is required to perform the frequency-based analysis and negative control samples is required to preform prevalence-based analysis [2].

Using experimental negative control samples and performing computational contamination identification and removal is effective [2, 19]. However, generating experimental negative control samples can be expensive. Researchers have to perform extra experiments and do extra sequencing runs on empty samples to generate those controls. Those extra work means that people must spend resources, including time and money. Since the contamination compositions of DNA extraction kits and other lab reagents are ubiquitous and distinctable, we have a hypothesis that the contaminants from the DNA extraction kit would share similar characteristics, such as composition and relative abundance of each contaminant, with other studies/samples that uses the same kit. We should be able to find contaminants as shared organisms with similar compositions in samples from different metagenomic environment if they use the same DNA extraction kit while the sequencing depths reach a certain level.

Implementation

To test our hypothesis, we implemented a computational contamination detection tool, Squeegee, which is able to identify potential contaminant at species level by performing taxonomic classification and search for shared organisms among multiple samples. The workflow of the pipeline is shown in Figure 1.

In order to generate an accuracy compositions of contaminants among the samples, the users must collect sequencing data from multiple metagenomic samples. Those samples should be independent from each other and representing different metagenomic environment.

Squeegee first filters the samples by the number of the reads within the sample. A threshold of minimum number of reads is set by the user and the small datasets that contains too few reads are filtered out since those dataset is unlikely to contain complete information of contaminants and potentially cause more false negative.

The tools then perform taxonomic classification using Kraken2, a taxonomic classifier using exact k-mer matching approach. Kraken2 would match each k-mer for all reads from each of the studies to the lowest common ancestor of all reference genomes containing the given k-mer [21].

A shared threshold of the value between 0 and 1 is set by the user. A shared organism is defined as the organism with presence rate greater or equal than the shared threshold. For example, while the shared threshold is set to 0.5, the shared organisms among the samples are organisms that presented in half or more than half of all samples. Squeegee searches those shared organism based on the taxonomic classification of the samples.

Results

To validate the precision and recall of Squeegee, we performed the analysis on a series of metagenomic datasets on human microbes. We then compared the contam-

inants identified by Squeegee to the experimental control samples to evaluate how well Squeegee could reconstruct the contamination profile among those experiments.

A total number of 90 metagenomic datasets of varies types were used in the analysis, including gingiva/buccal samples, stool samples, rectum samples, posterior fornix samples, placenta samples, and breast milk samples. The sequences from the metagenomic samples were extracted with the same DNA extraction kit and sequenced in the same lab. The size of the datasets varies from few thousand reads up to over 30 million reads. A total of 11 negative controls for the experiments went through all the steps of the DNA extraction without addition of sample. NEED MORE DETAILS ON HOW DATA-SETS WERE GENERATED. Since all samples are collected from human, all human reads were removed from the analysis in order to remove bias.

To evaluate out method, we first performed taxonomic classification on all negative controls. The union of the taxa (with human taxa removed) from those experimental negative control served as a truth set of contaminants. After Squeegee identifies the candidate contaminants, recall is calculated by comparing candidate contaminants and the truth set. Each contaminant in the truth set is weighted by its relative abundance in the experimental negative controls while calculating recall. Table 1 shows the recall of Squeegee with different read threshold and shared threshold at both species level and genus level.

Since samples are collected from different environment, the composition of the samples varies among each other, and precision at read level become less meaningful. Therefore, the precision of the tool was evaluated at the taxon level by measuring the ratio of number of false positive taxa and the total number of candidate contaminants. Table 2 shows the precision of Squeegee with different read threshold and shared threshold at both species level and genus level.

USE MORE FORMAL DEFINITION? MATH EQUATIONS?

MAYBE REPLACE SOME TABLES WITH FIGURES?

The result shows that the parameters of the tool effects recall and precision of the analysis. An Increase of the shared threshold indicates more strict condition for a taxa to be identified as potential contaminant, result in the decrease of recall and increase of precision. High read count threshold would filtered out samples with low sequencing depth. Since contaminants are less likely to be detected from the samples with lower read count, therefore increasing read count threshold would cause recall to increase and precision to decrease. Squeegee performs better at genus rank than at species rank.

At default setting of read count threshold = 1000, and shared threshold = 0.5, Squeegee achieved both high recall (92.60%) and precision (98.77%) at species rank. With this configuration, Squeegee identified 571 species as potential contaminants. Although the number of potential contaminants is relatively small compare to 3882 contaminant species in the truth set, (3) those candidates cover the majority of the contamination, and only contains 7 false positive taxa.

Discussion

We implemented a tool for detecting potential contaminants in certain DNA extraction kit in metagenomic samples by searching shared organisms found across

samples that are collected in different environment but processed with the same DNA extraction kit. One limitation of using Squeegee is that the samples used in the analysis should fundamentally have different compositions of organisms. Using samples collected in the same or similar metagenomics environment might cause more the false positives.

The average sequencing depth of the samples used in the analysis is also a critical factor for both recall and precision of the result. Using deeply sequenced datasets is recommended and often yields more accurate results. Squeegee gives the user the option to filter out low read count samples, as well as the option to tweak the criteria identifying of candidate contaminants by changing shared threshold. Based on the task, user can either make the contamination detection more sensitive, or identify potential contaminants with higher confidence.

In reality, control experiments are essential and unavoidable in scientific studies. Squeegee gives us the potential to cut down the cost of control experiments in metagenomics researches, as well can be used as a validation tool to check to correctness of the control set. Squeegee is also suitable for preliminary experiments which limited number of control experiments are conducted. It would help to provide a more complete view of the contamination profile along with few experimental control samples.

Conclusions

The result of the study have shown that contaminant sequences from the same source, such as DNA extraction kit and other reagents used during the sequencing process, can be found across multiple samples. Therefore, we developed Squeegee, a computational tool which is able to identify majority of the contaminants with a low false positive rate, without using experimental control sample as reference.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section ...

Acknowledgements
Text for this section ...

References

1. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W.: Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC biology **12**(1), 87 (2014)
2. 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 **6**(1), 226 (2018)
3. Fox, G.c.-a., Stackebrandt, E., Hespell, R., Gibson, J., Maniloff, J., Dyer, T., Wolfe, R., Balch, W., Tanner, R., Magrum, L., *et al.*: The phylogeny of prokaryotes. Science **209**(4455), 457–463 (1980)
4. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., Relman, D.A.: Diversity of the human intestinal microbial flora. science **308**(5728), 1635–1638 (2005)
5. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., *et al.*: A core gut microbiome in obese and lean twins. nature **457**(7228), 480 (2009)
6. Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., *et al.*: Vaginal microbiome of reproductive-age women. Proceedings of the National Academy of Sciences **108**(Supplement 1), 4680–4687 (2011)
7. Riesenfeld, C.S., Schloss, P.D., Handelsman, J.: Metagenomics: genomic analysis of microbial communities. Annu. Rev. Genet. **38**, 525–552 (2004)
8. Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., Nelson, K.E.: Metagenomic analysis of the human distal gut microbiome. science **312**(5778), 1355–1359 (2006)

9. Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., Thomas, B.C., Singh, A., Wilkins, M.J., Karaoz, U., *et al.*: Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature communications **7**, 13219 (2016)
10. Kitchin, P., Szotyori, Z., Fromholc, C., Almond, N.: Avoidance of false positives. Nature **344**(6263), 201 (1990)
11. Meadow, J.F., Altrichter, A.E., Bateman, A.C., Stenson, J., Brown, G., Green, J.L., Bohannan, B.J.: Humans differ in their personal microbial cloud. PeerJ **3**, 1258 (2015)
12. Adams, R.I., Bateman, A.C., Bik, H.M., Meadow, J.F.: Microbiota of the indoor environment: a meta-analysis. Microbiome **3**(1), 49 (2015)
13. Bittinger, K., Charlson, E.S., Loy, E., Shirley, D.J., Haas, A.R., Laughlin, A., Yi, Y., Wu, G.D., Lewis, J.D., Frank, I., *et al.*: Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing. Genome biology **15**(10), 487 (2014)
14. Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., Kelley, S.T.: Bayesian community-wide culture-independent microbial source tracking. Nature methods **8**(9), 761 (2011)
15. Jousselin, E., Clamens, A.-L., Galan, M., Bernard, M., Maman, S., Gschloessl, B., Duport, G., Meseguer, A., Calevro, F., Coeur D'Acier, A.: Assessment of a 16s rrna amplicon illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus. Molecular ecology resources **16**(3), 628–640 (2016)
16. Glassing, A., Dowd, S.E., Galandiuk, S., Davis, B., Chiodini, R.J.: Inherent bacterial dna contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut pathogens **8**(1), 24 (2016)
17. Larsson, A.J., Stanley, G., Sinha, R., Weissman, I.L., Sandberg, R.: Computational correction of index switching in multiplexed sequencing libraries. Nature methods **15**(5), 305 (2018)
18. Breitwieser, F.P., Pertea, M., Zimin, A.V., Salzberg, S.L.: Human contamination in bacterial genomes has created thousands of spurious proteins. Genome research **29**(6), 954–960 (2019)
19. Martí, J.M.: Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLoS computational biology **15**(4), 1006967 (2019)
20. Turner, P., Turner, C., Jankhot, A., Helen, N., Lee, S.J., Day, N.P., White, N.J., Nosten, F., Goldblatt, D.: A longitudinal study of streptococcus pneumoniae carriage in a cohort of infants and their mothers on the thailand-myanmar border. PloS one **7**(5), 38271 (2012)
21. Wood, D.E., Lu, J., Langmead, B.: Improved metagenomic analysis with kraken 2. Genome biology **20**(1), 257 (2019)

Figures

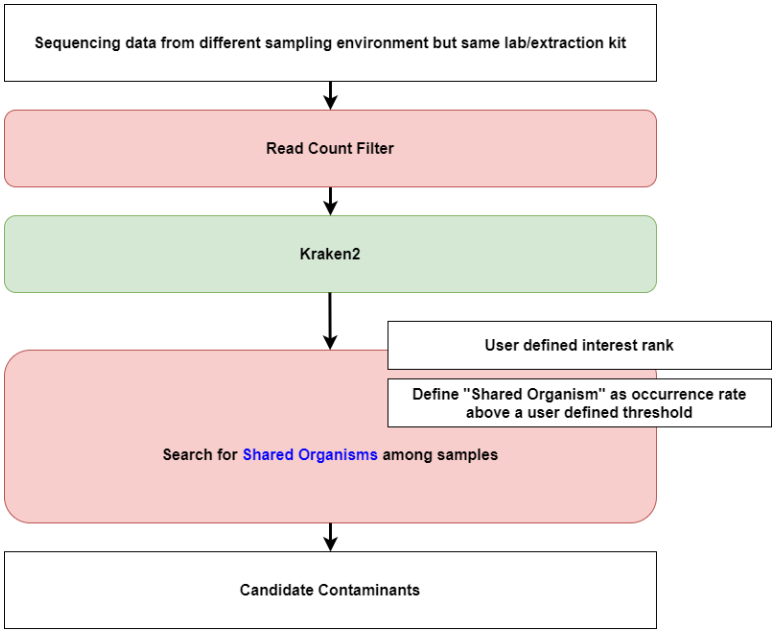

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Maybe includes a krona plot of the combined experimental control with recall

Tables

Table 1 Recall at read level. Weighted Contamination Recover Rate.

Reads	Shared Threshold											Rank
Threshold	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	
1000	99.02%	97.66%	95.52%	92.60%	89.33%	82.00%	77.21%	59.88%	47.54%	40.57%	35.57%	Species
10000	99.15%	97.91%	95.95%	92.97%	89.64%	83.52%	78.25%	58.89%	48.41%	43.55%	35.57%	
100000	99.23%	98.27%	96.46%	93.62%	89.91%	85.49%	79.18%	57.70%	49.24%	43.04%	35.51%	
1000000	99.92%	99.91%	99.89%	99.83%	99.75%	99.63%	99.53%	98.71%	97.49%	96.51%	90.98%	
1000	99.86%	99.66%	99.22%	98.70%	97.80%	95.89%	94.84%	93.95%	88.63%	82.27%	81.08%	Genus
10000	99.89%	99.68%	99.31%	98.78%	97.93%	97.09%	94.52%	93.68%	88.52%	83.07%	81.08%	
100000	99.90%	99.73%	99.42%	99.01%	98.35%	96.09%	94.70%	93.29%	88.47%	82.23%	81.08%	
1000000	99.97%	99.97%	99.97%	99.96%	99.94%	99.93%	99.91%	99.86%	99.73%	99.59%	98.68%	

Table 2 Precision at taxon level.

Reads	Shared Threshold											Rank
Threshold	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	
1000	89.90%	94.79%	96.96%	98.77%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	Species
10000	88.79%	93.57%	96.72%	98.90%	99.45%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
100000	87.65%	91.98%	95.46%	96.96%	98.84%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
1000000	75.27%	75.75%	76.80%	78.08%	78.83%	80.55%	81.58%	84.20%	87.44%	89.65%	93.50%	
1000	91.71%	97.12%	98.19%	99.73%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	Genus
10000	90.86%	95.94%	98.16%	99.50%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
100000	89.58%	94.26%	97.74%	98.72%	99.69%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
1000000	80.82%	81.33%	82.35%	82.98%	83.57%	84.73%	85.67%	87.39%	90.53%	93.22%	96.89%	

Table 3 Number of candidate contaminants.

Reads	Shared Threshold											Rank	Control
Threshold	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85		Pool Size
1000	3058	1996	1217	571	335	187	119	78	34	17	9	Species	3882
10000	3247	2192	1372	634	366	219	123	76	33	22	9		
100000	3402	2494	1608	822	430	245	142	73	36	21	8		
1000000	4966	4928	4815	4666	4577	4360	4186	3702	2867	2299	1077		
1000	1037	763	553	375	254	161	110	86	53	30	24	Genus	1175
10000	1083	812	598	401	271	185	112	83	53	35	24		
100000	1123	889	664	469	323	202	130	81	54	34	24		
1000000	1408	1398	1377	1357	1345	1310	1277	1182	982	855	546		

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.