Conta: Methods for detecting trace amounts of contamination

Onur Sakarya, John Lamping, Alexander Blocker, TaeHyung Kim, Saniya Fayzullina, Earl Hubbell, Catalin Barbacioru, and Franz Och.
GRAIL Inc., 1525 O'Brien Dr, Menlo Park, California, 94025, US

Introduction

Next generation sequencing (NGS) assays of cell-free DNA (cfDNA) must achieve high sensitivity and specificity in order to accurately detect circulating tumor DNA, enabling the early detection of cancer. Contaminating DNA from adjacent samples in library preparation plates may compromise specificity, because rare single nucleotide polymorphisms (SNPs) from the contaminant may look like low-frequency somatic mutations. Methods that obtain a signal based on fragment size and methylation status may also be affected by a contaminating sample. Copy number variations (CNVs), pregnancy, and transplants may also generate contamination-like SNP signals in plasma.

Here we present conta, a package for detecting presence of cross-contamination in NGS samples with high reliability. The package includes methods to call putative contamination events based on population minor allele frequencies (MAF) as well as methods to detect source of contamination from possible candidates.

Informative SNPs for Contamination

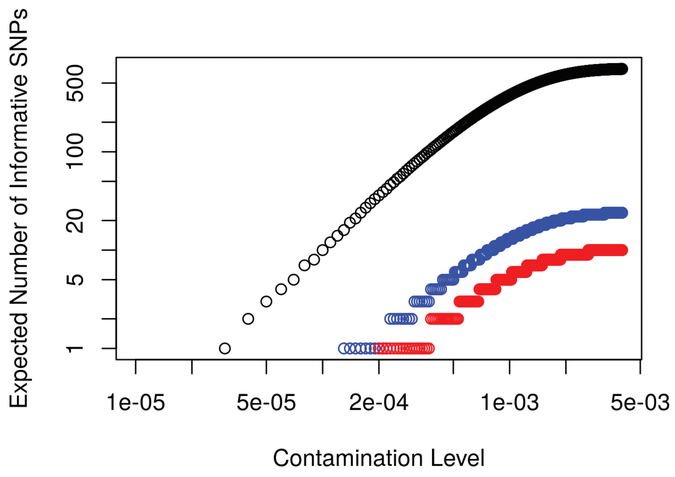

Fig. 1. Expected non-host SNPs observed in host samples by contamination fraction.

In a targeted panel covering 507 genes commonly mutated in cancer (Pedram Razavi et al., 2017), about 14,000 common SNPs (dbSNP version 150) are also sequenced on the targeted regions. When we performed pairwise comparisons of SNPs across 94 healthy individuals sequenced by this panel, we observed that each individual had on average 700 variant alleles that were not present in the other individual. We call these *informative SNPs* in the sense that their presense could signal a contamination event. Figure 1 shows the expected number of informative SNPs observed in a host sample depending on contamination level. When we sequenced a cfDNA sample to an error corrected depth of 3000x, requiring at least two independent reads to make a call, we expect to observe as many informative SNPs as shown by the black circles. If we filter to SNPs that are either present in dbSNP with a functional effect, or recurrent in COSMIC, we expect to observe, on average, 24 SNPs with expected numbers by contamination level shown by the blue circles. If we further filter to SNPs that are either not in dbSNP but have a functional effect, or recurrent in COSMIC, we expect to observe 10 SNPs on average, with expected numbers by contamination level as shown by the red circles.

Methods

Methods to detect contamination events by using SNP allele frequencies were previously developed by Jun et al. (2012), Cibulskis et al. (2011), and Bergmann et al. (2016). To increase the sensitivity and specificity of these methods to detect contamination at < 0.1% levels, we implemented different statistical models to distinguish signal (contamination) from noise, resulting in the conta method.

To detect contamination, we examine the allele frequency signature of informative SNPs in each host sample. First, we model the observed variant frequencies as a linear combination of contamination and a background noise model, and solve the resulting regression problem:

$$V_n = \alpha C + \beta N + \epsilon$$

where V_n are the homozygote variant allele frequencies, negated for non-reference homozygotes, α is the contamination coefficient for a prior contamination probability vector C for each SNP and β is the noise coefficient for a baseline noise vector N calculated across healthy, clean samples. ϵ is the error term for linear regression.

Second, we calculate contamination based on a maximum likelihood approach with the same C and N vectors. Based on these prior probabilities for contamination (C) and noise (N), each SNP is assigned a likelihood based on how well its number of variants are explained by noise alone, or a combination of noise and contamination. Multiple contamination levels are tested to find the one that maximizes the log likelihood across all SNPs:

$$\alpha_{max} = \underset{\alpha}{argmax} \sum_{i=1}^{N} log P(D_i | \alpha)$$

where α is the tested contamination level and D_i is the data including depth of read, observed variant frequencies and error rates for each position. We limit each SNPs likelihood contribution to eliminate outliers, and report a likelihood ratio for the α_{max} and whether it passes a validated threshold (see Limits of detection).

Third, we further inform the likelihood model with P_C calculated from known genotypes of other samples that were processed in the same batch or the same time period in the lab, and check if the maximum likelihood is improved by the knowledge of these genotypes (compared to population frequencies). To note, our method does not require matching normal or tumor samples to calculate genotypes.

Limits of detection

To train our methods, we built a noise model by sequencing 94 clean cfDNA samples from healthy individuals across our 507-gene targeted panel at 3000× coverage, where contaminated samples called by conta were removed from the model. Then we generated 1,500 mixture samples from a different set of 85 clean samples, choosing two of them at random to titrate in Poisson mixtures ranging from 0.01% to 1%. We used these clean and mixture samples in a 5-fold cross-validation setting to choose optimal thresholds and estimate the test error. We optimized thresholds for 95% specificity across all contamination levels. For the maximum likelihood method with MAF (second method described above), at 95% of specificity and 0.02% contamination level, detection sensitivity was 94.94% (Fig. 2).

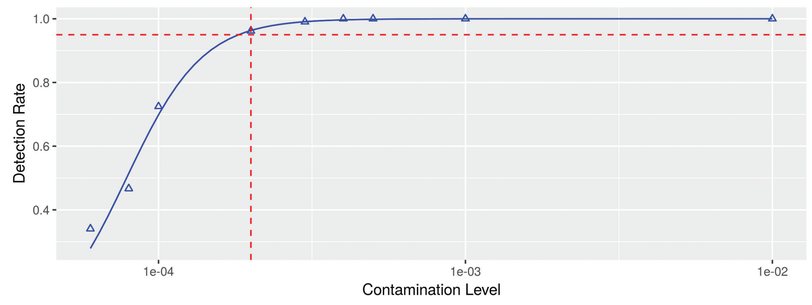

Fig. 2. Cross-validation to measure limits of detection.

In vitro titrations

Next, we titrated cfDNA samples from one normal healthy individual into multiple normal healthy backgrounds. Titrations ranged from 0.01% to 1%, and titrated pairs were processed on separate days to minimize the chance of cross-contamination. Samples were sequenced using the same 507 gene panel mentioned above to a depth about 2000x. Conta called 100 % of titrated samples (N = 40) correctly with 100 % specificity (N = 5). Fig. 3 shows the titrated vs. observed contamination levels (Pearson correlation coefficient = 0.99).

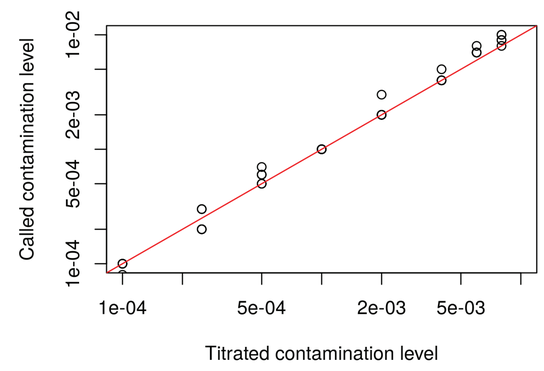

Fig. 3. In-vitro titrations across 5 individuals.

Events in the lab

Using conta, we tracked contamination events in our lab over a one-month period. Overall, we detected contamination >0.025% in 2.9% of the samples with contamination levels as shown in Fig. 4a (n = 310). Previously, we were using 12 unique index pairs across 16 sample batches, and the low-level contamination rates were higher as shown in Fig. 4b (n = 108). The red dashed line in the figure shows the targeted threshold of detection at >0.025%. In comparison, Cibulskis et al. (2011) estimated that a typical cancer project might expect contamination >1.5% in more than 10% of samples. Identifying contamination events accurately is beneficial for improving lab processes. For example, we found that using separate index pairs for each sample in a batch prevents cross-contamination. Many of the contamination events occurred from a sample directly upstream or within close vicinity (e.g. same strip tube).

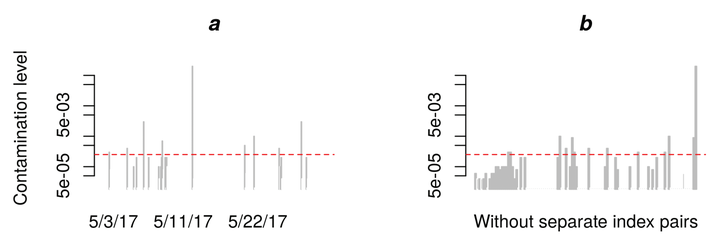

Fig. 4. Contamination calls in the lab.

Conclusions

We built a tool called conta to detect low-level contamination events. We demonstrated conta's ability to detect contamination as low as 0.01% with titrated samples and an in-silico limit of detection at 0.02%. Our method performs accurate detection for both targeted and whole-genome sequencing as tested by both in silico and in vitro titrations. Preventing contamination at low levels requires clean and robust workflows in the lab. Dual indexing of the samples across batches may alleviate post-indexing contamination, but robust detection methods are still necessary to detect and understand any remaining causes of contamination.

References

- Pedram Razavi et al. Performance of a high-intensity 508-gene circulating-tumor DNA (ctDNA) assay in patients with metastatic breast, lung, and prostate cancer. ASCO, 2017.
- Jun G, Flickinger M, Hetrick KN, et al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. American Journal of Human Genetics. 2012;91(5):839-848. doi:10.1016/j.ajhg.2012.09.004.
- Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. 2011;27(18):2601-2602. doi:10.1093/bioinformatics/btr446.
- Bergmann EA, Chen B-J, Arora K, Vacic V, Zody MC. Conpair: concordance and contamination estimator for matched tumor–normal pairs. Bioinformatics. 2016;32(20):3196-3198. doi:10.1093/bioinformatics/btw389.

Acknowledgements

We thank Chenlu Hou, Tom Chien, Eric Scott, K.C. Shashidhar, Collin Melton, Oliver Venn, Darya Filippova, Christopher Chang, Roger Jiang, Tina Truong, Punam Adhikari, Ting-Chun Liu, Hui Xu, Byoungsok Jung, Shivani Nautiyal, Phil Kiefer, Andro Hsu, Felipe Geyer and Vik Bajaj for support with generating data, analysis and writing the abstract.