SRSLY NGS LIBRARY PREPARATION WITH TWIST EXOME PANEL ENRICHMENT FOR ctDNA SNP PROFILING

INTRODUCTION

Cell-free DNA (cfDNA), found circulating in blood, originates predominantly from dying cells. In healthy individuals the vast majority of cfDNA derives from hematopoietic myeloid and lymph cells undergoing apoptosis. However, in individuals with cancer, a variable fraction of the cfDNA derives from tumor cells undergoing apoptosis and/or necrosis. This tumor derived fraction of cfDNA is known as circulating tumor DNA (ctDNA). The amount of ctDNA found in an individual with cancer depends on numerous parameters such as tumor growth rate, metastasis, and overall tumor size[1].

Detection and serial monitoring of ctDNA in cancer patients, through Next-Generation Sequencing (NGS), shows promise for assessing disease progression, response to treatment, and even early detection of tumors. However, the depth of sequencing necessary to detect ctDNA across the whole genome with high sensitivity and low error rates is cost-prohibitive. For this reason, cfDNA researchers increasingly employ some variation of a panel enrichment procedure. In a typical workflow, the tumor itself is first biopsied and its mutation spectrum profiled in order to identify somatic mutations thought to be tumor-derived. Second, NGS library generation from the patient's cfDNA is followed by targeted enrichment. The NGS library prep captures the breadth of molecules present in the cfDNA and the targeted panel enriches for the small genomic fraction of the genome that is of clinical interest. These enriched cfDNA libraries can now be cost-effectively sequenced to high depths in order to observe extremely low allele fraction of ctDNA that may reside within cfDNA[2-4].

Claret Bioscience's proprietary NGS library preparation kit, SRSLY, is a fast, simple, and efficient ligation-based single-stranded DNA library preparation method optimized for cfDNA analysis[5]. Twist Bioscience's Human Core Exome Enrichment Panel features a streamlined workflow and allows for uniform enrichment and increased sequencing efficiency of the human exome. Here we show that SRSLY can be combined with Twist Bioscience's Human Exome Enrichment Panel in order to profile and quantify ctDNA within cfDNA derived from cancer patients' plasma samples.

METHODS

Sample procurement and DNA preparation

Normal adjacent tissue, tumor biopsy, plasma and buffy coat from two stage III lung cancer donors were commercially purchased from two biorepositories – Cureline Inc, Brisbane CA and Proteogenex Inc, Inglewood CA.

The plasma was prepared according to Claret Bioscience's standard operating protocol. Briefly, whole blood was collected in 10ml Streck Blood Collection tube and immediately spun at 1800G for 10 minutes at 4°C to separate plasma. The plasma was cleared of debris with a second spin at 16000G for 10 minutes at 4°C. The cleared plasma was frozen in 2ml aliquots and sent to Claret Bioscience. Upon receipt the frozen plasma was thawed at 37°C for 5 minutes and spun at 12000 G for 10 minutes to remove cryoprecipitates. cfDNA was extracted from 2 ml of thawed plasma using the Qiagen QIAamp minelute ccfDNA kit following the manufacturer's instructions and eluted in 60 µl water.

Tumor and adjacent normal tissues were obtained from the donor and shipped frozen to Claret Bioscience. ~50-100 mm^3 of each tissue type was used for genome DNA and RNA extraction using the Qiagen AllPrep DNA/RNA/Protein mini kit following the manufacturer's protocol.

NGS library preparation and exome panel enrichment

One hundred nanograms per sample of DNA extracted from the fresh frozen normal adjacent and 50 ng/sample of tumor biopsy tissues were prepared for NGS using the

New England Biolabs NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina® and manufacturer's instructions. Cell-free DNA extracted from the plasma sample was prepared for NGS using Claret Bioscience's SRSLY™ PicoPlus Library Prep Kit and manufacturer's instructions specific for the low fragment retention protocol. Due to low extraction yields, 7.5 ng of cfDNA was input for lung cancer sample 1 and 2.5 ng of cfDNA was input for lung cancer sample 2. Post library construction, libraries made from the normal and tumor tissue DNA were pooled together and enriched using Twist Biosciences Human Core Exome Panel and manufacturer's instructions. Likewise, the cfDNA libraries were pooled together post library construction and enriched using Twist Biosciences Human Core Exome Kit and manufacturer's instructions. Exome enriched libraries were then sequenced to a depth of greater than 130x on an Illumina® HiSeq X at a read length of 2 x 151 bp following manufacturer's instructions.

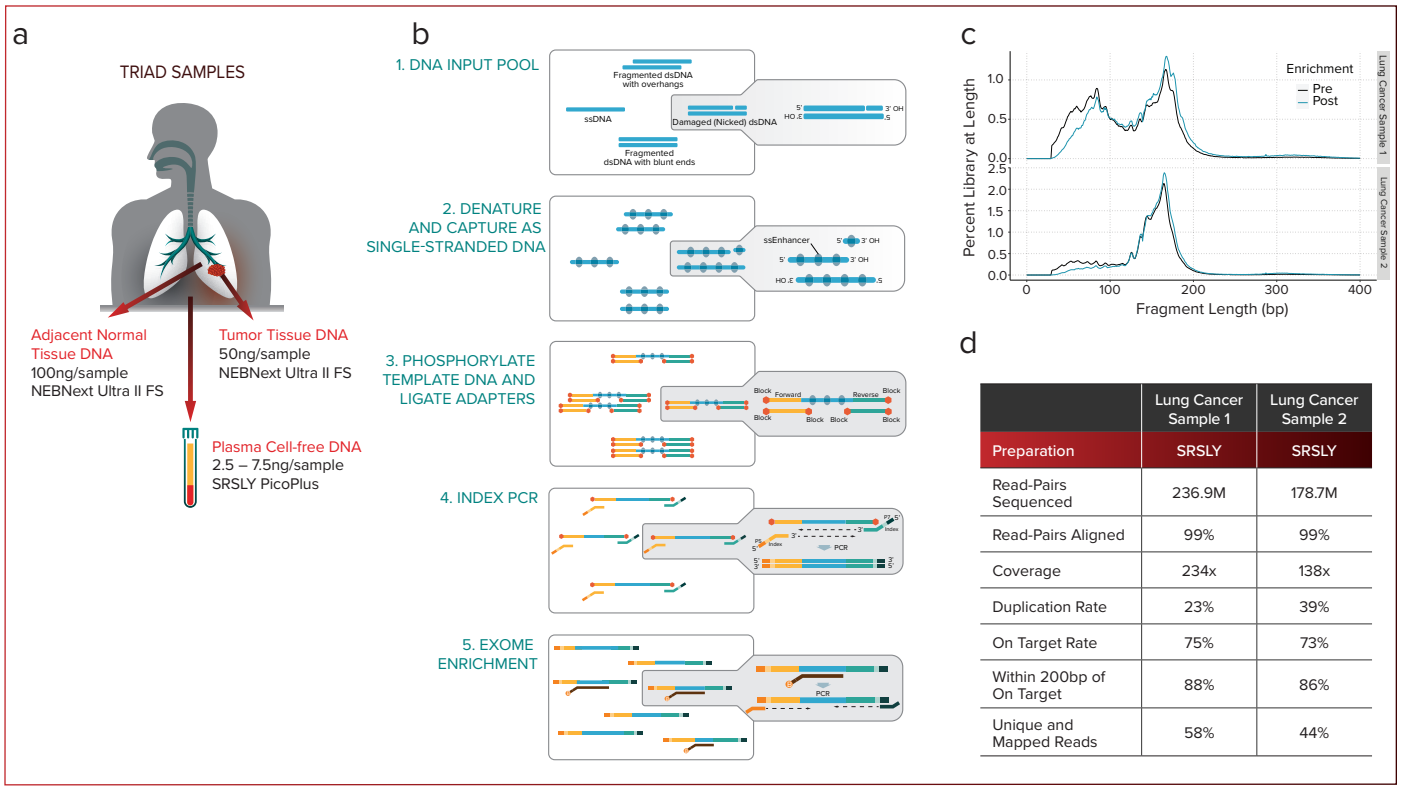
Informatic Processing

Sequence data was converted to FASTQ with bcl2fastq v 2.20.0.422. Tumor and normal data were processed with bcBio docker vm version 1.1.7 with docker image ID 9b6548521f2f, using genome build hg19, variant regions as downloaded from the Twist website, the ensemble variant caller with two passes including VarDict, Mutect2, and Strelka2. All other parameters were set to the defaults. Cell free DNA samples were trimmed with SeqPrep2. Reads were aligned to the hg19 reference with BWA v0.7.15-r1140 with sampe and samse for paired and merged reads, respectively. Duplicates were marked and removed with GATK version 2.17.1.

RESULTS

Experimental overview, cfDNA molecular insert distribution, and sequencing metrics

To showcase the quality of data produced by combining SRSLY cfDNA libraries with the Twist Human Core Exome Panel in reference to ctDNA variant analysis we generated exome enriched NGS libraries from the plasma extracts of two individuals with stage III lung cancer. To detail the mutation profile in the exome of the tumor and account of somatic single nucleotide polymorphisms (SNPs) we generated exome enriched NGS libraries from the tumor and normal adjacent tissue DNA extracts (Figure 1a).



	Lung Cancer Sample 1	Lung Cancer Sample 2
Preparation	SRSLY	SRSLY
Read-Pairs Sequenced	236.9M	178.7M
Read-Pairs Aligned	99%	99%
Coverage	234x	138x
Duplication Rate	23%	39%
On Target Rate	75%	73%
Within 200bp of On Target	88%	86%
Unique and Mapped Reads	58%	44%

Fig 1. Experimental overview, mapped molecular insert lengths, sequencing metrics. (a) Experimental overview of tissue/samples obtained and NGS libraries prepared from each patient. (b) Overview of the SRSLY protocol combined with the Twist enrichment panel for the preparation of cfDNA libraries. (c) Mapped molecular length insert distributions for the SRSLY cfDNA libraries from both patients pre and post enrichment. (d) Table of salient sequencing metrics for the SRSLY cfDNA libraries from both patient samples post exome enrichment.

SRSLY creates complex NGS libraries from cfDNA extracts by rendering all template DNA uniformly single-stranded through heat denaturation. The DNA is then maintained as single-strands throughout the ligation reaction by the inclusion of single-strand enhancer proteins. Next Generation Sequencing adapter ligation is performed utilizing directional splint-adapters that are biochemically blocked on all termini except for the ones that facilitate proper ligation. After ligation, the library is completed with index PCR and enriched using Twist Bioscience's capture probes (Figure 1b).

The length distribution of cfDNA is short (centered around 167 bp) and is the result of mono-nucleosome imparted nuclease protection[6-8]. One of SRSLY's many features is that it allows the user the option to retain a substantial proportion of extremely short fragment length molecules (<100 bp) if desired, in addition to the relatively longer mono-nucleosome length fragments. While not applicable to exome enriched DNA, recent studies examining cfDNA within this short, sub-nucleosome size range show that these fragments can be enriched for useful biological information such as transcription factor binding sites, mitochondrial DNA, and microbial DNA8-11. As a quality control step in our analyses we sequenced our SRSLY cfDNA libraries pre- and post-enrichment. Comparison of the molecular insert distribution of the cfDNA libraries pre and post enrichment show that the exome enriched libraries retain a lower proportion of sub-nucleosomal sized fragments and a higher proportion of the nucleosome sized fragments than the unenriched libraries (Figure 1c). The decreased proportion of sub-nucleosome sized molecules in the enriched libraries could reflect the fact that these short molecules contain biological information not present in the exome.

After exome enrichment we sequenced both SRSLY cfDNA samples to greater than 200M read-pairs on a 2x151 bp HiSeq X sequencing run. Lung cancer sample 1 was sequenced to 234x exome coverage with a duplication rate of 23% and an on-target rate of 75%, which resulted in 58% unique and mapped reads. Lung cancer sample 2 was sequenced to 138x exome coverage with a duplication rate of 39% and an on-target rate of 73%, which resulted in 44% unique and mapped reads (Figure 1d). The increased duplication and lower usable data for lung cancer sample 2 is most likely the result of the limited DNA input. The DNA input for lung cancer sample 2 ended up being 3x less than the input DNA amount used for lung cancer sample 1.

SNP variant calling

After sequencing and mapping we ran tumor-normal analysis with bcBio's pipeline and an ensemble caller of Strelka, Mutect2, and VarSeq on the normal adjacent and tumor biopsy exome enriched libraries for both patient samples. We identified 229 total variant calls for lung cancer sample 1 with 28 classified as high impact and 185 classified as moderate. Within the high impact variants, we found SNP mutations in TP53 (NM_000546.6:c.469G>T), BTK (NM_000061.3:c.1673A>G), and TGFB2 (NM_001135599.3:c.874C>T). We identified 188 total variant calls for lung cancer sample 2 with 20 classified as high impact and 138 classified as moderate. Within the high impact variants, we found SNP mutations in TP53 (NM_000546.6:c.817C>T), SLIT2 (NM_004787.4:c.3684delC), and JAK2 (NM_004972.4:c.3320A>G). By taking the mean variant allele fraction across all the variants in a given sample we estimate that the lung cancer sample 1 tumor biopsy contains a tumor purity of 32% and the lung cancer sample 2 tumor biopsy contains a tumor purity of 36%.

Next we looked at the cfDNA exome data to identify tumor derived molecules circulating in the blood plasma of the lung cancer patents. First, we compared the depth of coverage at the known tumor variant allele sites between the cfDNA libraries and the tumor libraries for both lung cancer patient samples. From the tumor-normal analysis we had 229 and 185 total loci of interest to examine in the cfDNA library for lung cancer patient 1 and 2, respectively. The depth of sequencing and usable data was similar between the tumor and cfDNA libraries for each patient. However, the depth of coverage plots show that for each patient's known variants, the allelic sites are covered roughly half to two-thirds as frequently in the cfDNA

libraries than in the tumor libraries (Figure 2a, b). One hypothesis for this observation is that cfDNA samples do not display uniform genomic coverage but rather display coverage inversely proportional to expression. This is due, in part, because expressed genomic regions are not protected by nucleosomes and therefore are preferentially degraded by circulating nucleases upon cell death and release into the circulatory system[8,12,13]. These results suggest that, as a whole, the cancer genes of interest for each patient are expressed in the tumor.

Finally, to quantify the amount of ctDNA in the cfDNA, we compared the cancer variant allele frequency (VAF) between the cfDNA libraries and the tumor libraries for both patient samples. The vast majority of the VAFs for the tumor library of patient 1 occur in 10% - 30% of reads at each particular locus, while the vast majority of the VAFs in the cfDNA library occur in less than 10% of the reads at each particular locus (Figure 2c). By taking the mean VAF across all the variants (6.5%) and the expectation that the vast number of variant sites are

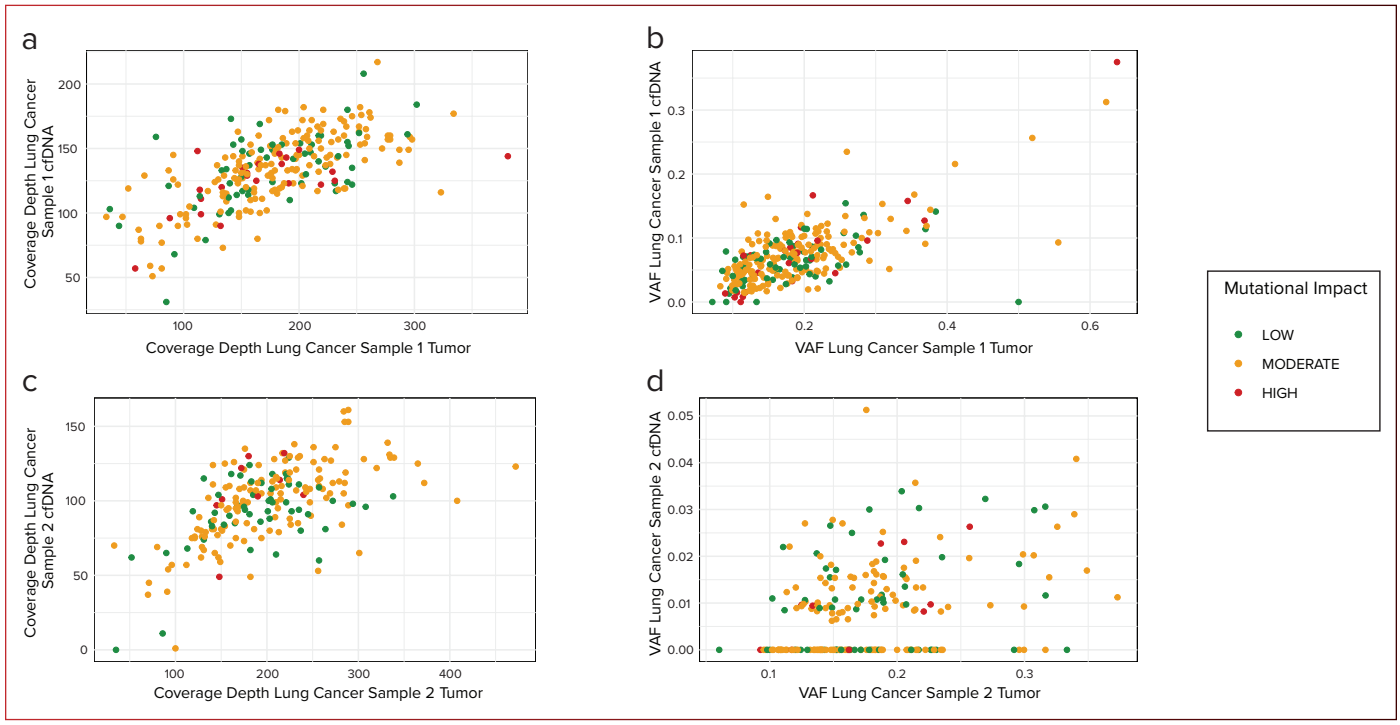


Fig 2. Depth of coverage and VAF calls color coded by impact classification. (a) Depth of coverage for reads occurring at all 229 variant allele loci regardless of whether the read contains the variant allele or not. Cell-free DNA data versus tumor data for lung cancer patient 1. (b) Variant allele fractions for all 229 identified SNPs for lung cancer patient 1. Cell-free DNA data versus tumor data. (c) Same as (a) except for all 188 variant allele loci for lung cancer patient 2. (d) Same as (b) except for patient 2. Mutational Impact - Predicted impact of mutation on protein function.

heterozygous with a copy number of two, we estimate the ctDNA fraction of patient 1 to be 13% of the total cfDNA pool. Patient 2 had a much lower ctDNA fraction than patient 1. The vast majority of the VAFs for the tumor library of patient 2 occur in 2% - 20% of reads at each particular locus, while the vast majority of the VAFs in the cfDNA library occur in less than 5% of the reads at each particular locus (Figure 2d). By taking the mean VAF across all the variants (1%) and the expectation that the vast number of variant sites are heterozygous with a copy number of two, we estimate the ctDNA fraction of patient 2 to be 2% of the total cfDNA pool.

CONCLUSION

Quantification of ctDNA within cfDNA by locating and monitoring SNPs and other mutation types via NGS is clinically promising but fraught with technical challenges. Here we present validation that Claret Bioscience's SRSLY library prep combined with Twist Bioscience's Human Core Exome Panel successfully profiles ctDNA within cancer patients cfDNA samples. First, we showed that Claret Bioscience's SRSLY kit creates complex cfDNA libraries from minimal inputs and retains a substantial proportion of short DNA fragments. Next, we showed that SRSLY libraries are fully compatible with Twist's enrichment

panels and obtains high on-target rates. Finally, we showed that the data generated by combining Claret Bioscience's SRSLY with Twist Bioscience's Human Core Exome Panel can be used to quantify the VAFs of ctDNA within the cfDNA of cancer patients using standard pipelines.

While cancer progression monitoring via mutation profiling in cfDNA continues to show promise an alternative approach, called fragmentomics, is emerging. Fragmentomics is built on the theory that the biological degradation of cfDNA is non-random and aims to classify the origin of these molecules based on size, termini motifs, and genomic location. Recent publications in the field of fragmentomics show that ctDNA is on average shorter than healthy cfDNA fragments and that patients with certain cancer types show a preferential fragment termini motif patterns[10,11,14-16]. The goal of fragmentomics is to explore and hopefully discover robust signatures of ctDNA fragments that can be employed as more sensitive biomarkers when compared to SNPs. While not discussed here, in addition to retaining short fragments, SRSLY facilitates fragmentomic analysis of cfDNA by retaining the both the native 5' and 3' of all library molecules.

DECLARATIONS

Ethics statement

The samples were obtained from a commercial biorepositories Cureline. Inc and Proteogenex Inc and therefore the study is not considered Human Subject Research under HHS human subjects regulations (45 CFR 36). However at the site of collection, all patients provided written informed consent for normal tissue, tumor and whole blood collection and genomic analyses with the DNA extracted from these samples. The review of patient medical records for detailed demographic, pathological and treatment information was done under an Ethical Committee Review-approved protocols: Cureline - CU-M-07092015-C-INT, Proteogenex - PG-ONC-2003/1

Data availability

Data availability is determined on a case by case basis. Please contact technicalsupport@claretbio.com if interested.

REFERENCES

1. Cabel, L. et al. Clinical potential of circulating tumour DNA in patients receiving anticancer immunotherapy. Nat Rev Clin Oncol 15, 639-650, doi:10.1038/s41571-018-0074-3 (2018).
2 Christensen, E. et al. Early Detection of Metastatic Relapse and Monitoring of Therapeutic Efficacy by Ultra-Deep Sequencing of Plasma Cell-Free DNA in Patients With Urothelial Bladder Carcinoma. J Clin Oncol 37, 1547-1557, doi:10.1200/JCO.18.02052 (2019).
3 Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. Nat Rev Clin Oncol 10, 472-484, doi:10.1038/nrclinonc.2013.110 (2013).
4 Reinert, T. et al. Analysis of Plasma Cell-Free DNA by Ultradeep Sequencing in Patients With Stages I to III Colorectal Cancer. JAMA Oncol, doi:10.1001/jamaoncol.2019.0528 (2019).
5 Troll, C. J. et al. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. BMC Genomics 20, 1023, doi:10.1186/s12864-019-6355-0 (2019).
6 Jahr, S. et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res 61, 1659-1665 (2001).
7 Quake, S. Sizing up cell-free DNA. Clin Chem 58, 489-490, doi:10.1373/clinchem.2011.174250 (2012).
8 Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell 164, 57-68, doi:10.1016/j.cell.2015.11.050 (2016).
9 Burnham, P. et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Sci Rep 6, 27859, doi:10.1038/srep27859 (2016).
10 Jiang, P. et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci U S A 112, E1317-1325, doi:10.1073/pnas.1500076112 (2015).
11 Jiang, P. & Lo, Y. M. D. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. Trends Genet 32, 360-371, doi:10.1016/j.tig.2016.03.009 (2016).
12 Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A 105, 16266-16271, doi:10.1073/pnas.0808319105 (2008).
13 Kelly, T. K. et al. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res 22, 2497-2506, doi:10.1101/gr.143008.112 (2012).
14 Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385-389, doi:10.1038/s41586-019-1272-6 (2019).
15 Han, D. S. C. et al. The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. Am J Hum Genet 106, 202-214, doi:10.1016/j.ajhg.2020.01.008 (2020).
16 Lapin, M. et al. Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. J Transl Med 16, 300, doi:10.1186/s12967-018-1677-2 (2018).