Ambiguous specification of EGFR mutations compounded by nil or negligible fragmented gene counts and erroneous application of the Kappa statistic reiterates doubts on the veracity of the TEP-study

Sandeep Chakraborty,
R - 44/ 1, Celia Engineers, T. T. C Industrial Area, Rabale, Navi Mumbai, 400701, India.

Abstract

The use of RNA-seq from tumor-educated platelets (TEP) as a 'liquid biopsy' source [1] has been refuted recently (http://biorxiv.org/content/early/2017/06/05/146134, not peer-reviewed). The TEP-study also mentioned that mutant epidermal growth factor receptor (EGFR) was 'accurately distinguished using surrogate TEP mRNA profiles', which is contested here. It is shown that only 10 out of 24 (a smaller sample set, original study has 60) non-small cell lung carcinoma (NSCLC) samples here has any expression at all. Even there the number of reads (101 bp) are [1, 4, 1, 14, 9, 1, 2, 19, 21, 6], and do not even add up to one complete EGFR gene (about 6000 bp). EGFR mutations have been painstakingly collated in www.mycancergenome.org/content/disease/lung-cancer/egfr. In stark contrast, the TEP study has no specification of the EGFR mutant used. The TEP study found EGFR mutations in 17/21 (81%), and EGFR wild-type in 4/39 (10%) for NSCLC samples (Table S7, reflected in Fig 3, Panel E in percentages). A major flaw is the assumption that a non "EGFR wild-type" is a "EGFR mutant" since cases zero with EGFR reads (which are almost half of the samples) could be either. The application of the Kappa statistic to this data is erroneous for two reasons. First, the Kappa statistic does not handle "unknowns", as is the case for samples with zero expression. Secondly, 'interobserver variation can be measured in any situation in which two or more independent observers are evaluating the same thing' [2]. The 90% (Fig 3, Panel E) is just the percentage of samples (35/39) that are not "EGFT WT" in one observation. It is not qualified to be in the Kappa matrix, where it translates to 35, leading to a Kappa=0.707, which implies "substantial agreement" [2]. The other observation (looking for EGFR mutation) is in a different set. To summarize, this work reiterates negligible expression of EGFR reads in NSCLC samples, and finds serious shortcomings in the statistical analysis of subsequent mutational analysis from these reads in the TEP-study.

Introduction

Tumor tissue biopsy, the gold standard for cancer diagnostics, pose challenges that include access to the tumor, quantity and quality of tumoral material, lack of patient compliance, repeatability, and bias of sampling a specfic area of a single tumor [3]. This has resulted in a new medical and scientific paradigm defined by minimal invasiveness, high-efficiency, low-cost diagnostics [4], and, whenever possible, personalized treatment based on genetic and epigenetic composition [5]. The presence of fragmented DNA in the cell-free component of whole blood (cfDNA) [6], first reported in 1948 by Mandel and Metais, has been extensively researched for decades, with extremely promising results in certain niches [7]. Additionally, cfDNA derived from tumors (ctDNA) [8] have tremendous significance as a cancer diagnostic tool [9], and for monitoring responses to treatment [10]. However, detection of ctDNA, and differentiation with cfDNA, remains a challenge due the low amounts of ctDNA compared to cfDNA [11].

Recently, tumor-educated blood platelets (TEP) were proposed as an alternative source of tumor-related biological information [1, 12]. The hypothesis driving the potential diagnostic role of TEPs is based on the interaction between blood platelets and tumor cells, subsequently altering the RNA profile of platelets [13,14]. The study showed using RNA-seq data that tumor-educated platelets (TEP) can distinguish 228 patients with localized and metastasized tumors from 55 healthy individuals with 96% accuracy [1]. As validation, this study reported significant over-expression of MET genes in non-small cell lung carcinoma (NSCLC), and HER2/ERBB2 [15] genes in breast cancer, which are well-established biomarkers. The TEP-study also mentioned that mutant epidermal growth factor receptor (EGFR) was 'accurately distinguished using surrogate TEP mRNA profiles' [1]. EGFR, a trans-membrane glycoprotein, is responsible for triggering several signal transduction cascades implicated in more aggressive tumor phenotypes [16–18].

Previously, the TEP-study was refuted by an analysis of a subset of the samples (yet to be peer-reviewed) based on the absence of evidence of MET-overexpression in NSCLC samples [19]. Here, it is demonstrated that EGFR reads are equally negligible in the NSCLC samples. Next, it is noted that the kind of EGFR mutants are not explicitly specified (there are several EGFR mutations implicated in NSCLC - www.mycancergenome.org/content/disease/lung-cancer/egfr). Lastly, the Kappa statistic used is shown to be flawed for several reasons. In summary, this work reiterates the shortcomings of the TEP-study, which proposes the use platelets as a source for "liquid biopsy" [1].

Results and discussion

Null or negligible fragmented gene counts of EGFR

A smaller subset (24 out of 60) of NSCLC samples were used here. A large number of samples have zero counts (14 out of 24) (Table 1). Furthermore, the number of reads in other samples are negligible. These are 101 bp reads, and do not even add up to one complete gene (Fig 1). The differentiation of mutants from wild-type (WT) is a challenging task from such scanty reads.

Verification of the computation of the Kappa statistic for EGFR mutations

These formulas are obtained from [2].
pE = [(n1/n) * (m1/n)] + [(n2/n) * (m2/n)]
pE = (21*21 + 39*39) / 3600 = .545
pO = (17+35)/60 = 52/60 = 0.866
pO - pE = 0.866 - .545 = 0.32166
1 - pE = 1 - .545 = 0.455
Kappa = (pO - pE)/(1 - pE) = 0.32166 / 0.455 = 0.70694505494

Thus, this exactly corroborates the value (0.707) reported in Table S7 in the TEP-study [2].

Erroneous application of the Kappa statistic in the EGFR study:

However, the application of the Kappa statistic is seriously incorrect for two reasons.

1. The assumption that a sample that is not a "EGFR wild-type" is a "EGFR mutant" is seriously flawed in view of a large number of samples having zero EGFR reads (see above). The TEP-study found 4 "EGFR WT" in 39 samples, and automatically assumed that the rest 35 are EGFR mutants - this is clearly not correct for the samples with zero counts (Table 1). The Kappa statistic is a binary formula - and does not handle a "do not know" scenario.

2. 'Interobserver variation can be measured in any situation in which two or more independent observers are evaluating the same thing' [2]. Here, the sets are apparently mutually exclusive with 21 and 39 elements, adding up to 60 NSCLC samples (although there is no explicit information).

To further emphasize the fuzziness, consider Fig 3, Panel E in [1] (which is essentially the Kappa matrix in percentages). The diagonal has 81% for both agreeing on "EGFR mutants" and 90% agreeing on "EGFR wild-type". Agreement or disagreement does not makes sense if one is looking at different sets. The 90% is just the percentage of samples that are not "EGFT WT". It is not agreement or disagreement - and thus, not qualified to be in the Kappa matrix.

Conclusion:

Statistical terminologies can often act as a smokescreen. Statistics needs to be supported by raw data. Just providing P-values and Kappa statistics can be misleading, and restrict future verification.

Absence of evidence is not evidence of absence in this case. In the absence of even a single reads, searching and failing to find a wild-type sequence does not imply the presence of a mutant. Furthermore, there is a flippancy in describing the EGFR mutant, since there can be many EGFR mutants - and the assumption that all NSCLC samples have the same mutation is biologically improbable.

This raises serious doubts on using TEP as a possible 'liquid biopsy' candidate. Essentially, it refutes the hypothesis that platelets carry enough RNA-seq from tumors to make it viable as a diagnostic method. A review found it 'surprising' that although 'the tumor type was the predominant factor for the actual platelet conditioning, tumor metastasis did not significantly impact on them when compared to samples from patients without metastasis' [14]. The excitement surrounding the fact that '2016 marked the first approval of a liquid biopsy test in oncology to assist in patient selection for treatment' [20] should be tempered, and a cautious approach adopted [21, 22] with reports of 'broken promises' [23].

Materials and methods

A kmer-based version (KEATS [24]) of YeATS [25–29] was used to obtain gene counts from transcripts in the RNA-seq data. A BLAST search suffices to demonstrate the absence of MET genes in the lung cancer samples. The Kappa statistic computation has been done based on formulas provided in [2].

Competing interests

No competing interests were disclosed.

References

1. Best MG, Sol N, Kooi I, Tannous J, Westerman BA, et al. (2015) Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer cell 28: 666–676.

2. Viera AJ, Garrett JM, et al. (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37: 360–363.

3. Vendrell JA, Mau-Them FT, Béganton B, Godreuil S, Coopman P, et al. (2017) Circulating cell free tumor dna detection as a routine tool forlung cancer patient management. International Journal of Molecular Sciences 18: 264.

4. Han X, Wang J, Sun Y (2017) Circulating tumor dna as biomarkers for cancer detection. Genomics, proteomics & bioinformatics .

5. Sorber L, Zwaenepoel K, Deschoolmeester V, Van Schil P, Van Meerbeeck J, et al. (2016) Circulating cell-free nucleic acids and platelets as a liquid biopsy in the provision of personalized therapy for lung cancer patients. Lung Cancer .

6. Jiang P, Lo YD (2016) The long and short of circulating cell-free dna and the ins and outs of molecular diagnostics. Trends in Genetics 32: 360–371.

7. Lo YD, Corbetta N, Chamberlain PF, Rai V, Sargent IL, et al. (1997) Presence of fetal dna in maternal plasma and serum. The Lancet 350: 485–487.

8. Chen XQ, Stroun M, Magnenat JL, Nicod LP, Kurt AM, et al. (1996) Microsatellite alterations in plasma dna of small cell lung cancer patients. Nature medicine 2: 1033–1035.

9. Yi X, Ma J, Guan Y, Chen R, Yang L, et al. (2017) The feasibility of using mutation detection in ctdna to assess tumor dynamics. International Journal of Cancer 140: 2642–2647.

10. Imamura F, Uchida J, Kukita Y, Kumagai T, Nishino K, et al. (2016) Monitoring of treatment responses and clonal evolution of tumor cells by circulating tumor dna of heterogeneous mutant egfr genes in lung cancer. Lung Cancer 94: 68–73.

11. Diaz LA, Bardelli A (2014) Liquid biopsies: genotyping circulating tumor dna. Journal of Clinical Oncology 32: 579–586.

12. Nilsson RJA, Balaj L, Hulleman E, Van Rijn S, Pegtel DM, et al. (2011) Blood platelets contain tumor-derived rna biomarkers. Blood 118: 3680–3683.

13. Bardelli A, Pantel K (2017) Liquid biopsies, what we do not know (yet). Cancer cell 31: 172–179.

14. Feller SM, Lewitzky M (2016) Hunting for the ultimate liquid cancer biopsy-let the tep dance begin. Cell Communication and Signaling 14: 24.

15. Foulkes WD, Stefansson IM, Chappuis PO, Bégin LR, Goffin JR, et al. (2003) Germline brca1 mutations and a basal epithelial phenotype in breast cancer. Journal of the National Cancer Institute 95: 1482–1485.

16. Helena AY, Arcila ME, Rekhtman N, Sima CS, Zakowski MF, et al. (2013) Analysis of tumor specimens at the time of acquired resistance to egfr-tki therapy in 155 patients with egfr-mutant lung cancers. Clinical cancer research 19: 2240–2247.

17. Sos ML, Koker M, Weir BA, Heynck S, Rabinovsky R, et al. (2009) Pten loss contributes to erlotinib resistance in egfr-mutant lung cancer by activation of akt and egfr. Cancer research 69: 3256–3261.

18. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, et al. (2004) Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304: 1497–1500.

19. Chakraborty S (2017) No evidence of met and her2 over-expression in non-small cell lung carcinoma and breast cancer, respectively, raises serious doubts on using rna-seq profiles of tumor-educated platelets as a liquid biopsysource. bioRxiv : 146134.

20. Blumenthal GM, Pazdur R (2017) Approvals in 2016: the march of the checkpoint inhibitors. Nature Reviews Clinical Oncology 14: 131–132.

21. Diamandis EP (2016) A word of caution on new and revolutionary diagnostic tests. Cancer cell 29: 141–142.

22. Best MG, Sol N, Tannous BA, Wesseling P, Wurdinger T (2016) Re: a word of caution on new and revolutionary diagnostic tests. Cancer cell 29: 143.

23. Shee K, Chamberlin M, Varn F, Bean J, Marotti J, et al. (2017). Abstract p6-07-03: Broken promise of liquid biopsy: Plasma dna does not accurately reflect tumor dna in metastatic breast cancer.

24. Chakraborty S (2017) Cataloguing over-expressed genes in epstein barr virus immortalized lymphoblastoid cell lines through consensus analysis of pacbio transcriptomes corroborates hypomethylation of chromosome 1. bioRxiv : 125823.

25. Chakraborty S, Britton M, Wegrzyn J, Butterfield T, Martinez-Garcia PJ, et al. (2015). YeATS-a tool suite for analyzing RNA-seq derived transcriptome identifies a highly transcribed putative extensin in heartwood/sapwood transition zone in black walnut.

26. Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, et al. (2016) The walnut (juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of nonstructural polyphenols. The Plant Journal .

27. Chakraborty S, Britton M, Martínez-García P, Dandekar AM (2016) Deep RNA-seq profile reveals biodiversity, plant–microbe interactions and a large family of NBS-LRR resistance genes in walnut (*juglans regia*) tissues. AMB Express 6: 1.

28. Chakraborty S, Martínez-García PJ, Dandekar AM (2016) Yeatsam analysis of the walnut and chickpea transcriptome reveals key genes undetected by current annotation tools. F1000Research 5.

29. Chakraborty S (2017) Mcf-7 breast cancer cell line pacbio generated transcriptome has˜ 300 novel transcribed regions, un-annotated in both refseq and gencode, and absent in the liver, heart and brain transcriptomes. bioRxiv : 100974.

Table 1: **Counts of reads of the EGFR gene obtained using KEATS:** A large number of samples have zero counts. Furthermore, the number of reads in other samples are negligible. Note, that these are 101 bp reads - so, they do not even add up to one complete gene. The differentiation of mutants from wild-type is a challenging task from such meagre counts.

SRR1982781	1
SRR1982780	4
SRR1982791	1
SRR1982772	14
SRR1982770	9
SRR1982790	1
SRR1982795	2
SRR2096501	19
SRR1982777	21
SRR1982771	6
SRR2096517	0
SRR2096502	0
SRR1982756	0
SRR1982759	0
SRR1982762	0
SRR1982761	0
SRR2096503	0
SRR1982782	0
SRR2096516	0
SRR1982793	0
SRR1982765	0
SRR1982787	0
SRR1982792	0
SRR1982760	0

Table 2: **Kappa statistic derivation table for EGFR mutants and wild-type:** The values are provided in Table S7 and also reflected in Fig 3, Panel E (which are percentages) in the TEP-study [1]. The computation of the Kappa statistic [2] corroborates a value of 0.707. However, the application of the Kappa statistic is erroneous since they involve mutually exclusive sets. Furthermore, in the presence of samples with absolutely no EGFR reads, this is not a binary situation wherein the absence of EGFR WT implies the presence of the EGFR mutant. The "35" (90% if you take percentages) is not where both observers agree that they are non-mutants (which gives the Kappa statistic such a good value of 0.707 which implies "Substantial agreement" [2]). It is just the percentage of one observer - it has no place in the Kappa matrix. One can make a matrix out of this data, but it can not be used to compute Kappa values.

	Second	Observer	
	Mut	Non-Mut	Total
First	17 (81%)	4 (10%)	21 (m1)
Observer	4 (19%)	35(90%)	39 (m2)
total	21(n1)	39(n2)	60 (n)

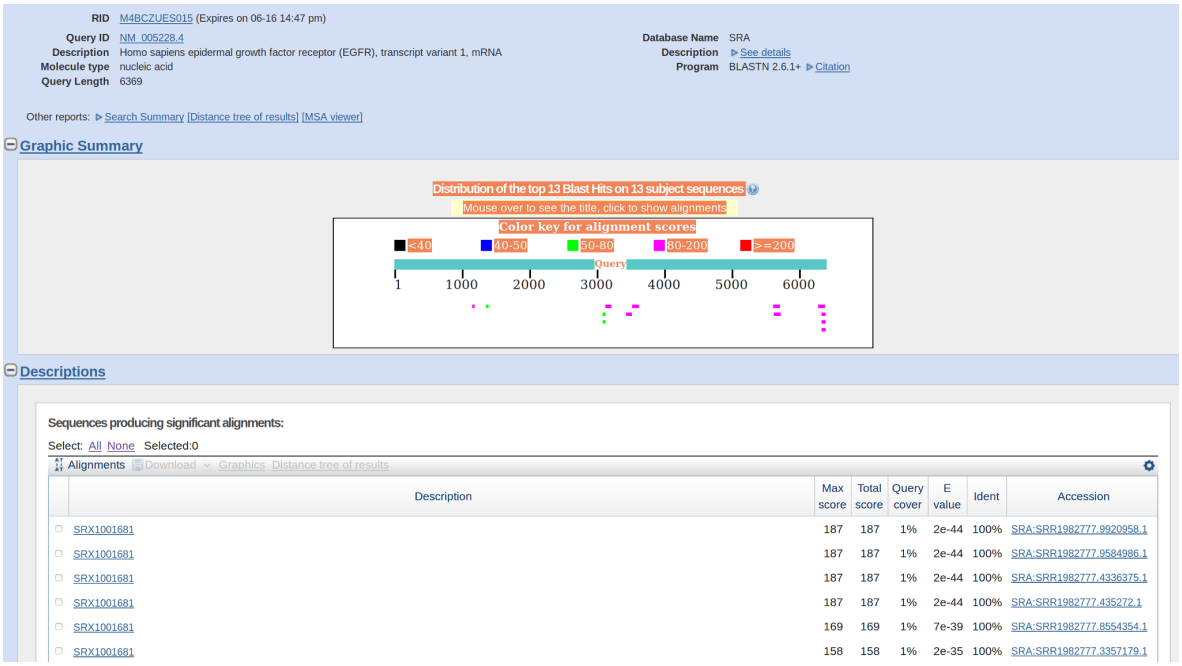

Figure 1: **Low counts of EGFR reads in a NSCLC sample (SRA:SRR1982777) obtained using the online BLAST interface:** NM_005228.4 (EGFR transcript variant 1, mRNA) was used as the query sequence. Identitying EGFR mutations from such a data is bound to be error-prone.