Abstract ID:
119P

NGS for precision medicine in non-small cell lung cancer: Challenges and opportunities

M. Ivanov[1,2], E. Novikova[3], E. Telysheva[3], P. Chernenko[4], V. Breder[4], K. Laktionov[4], A. Baranova[5,6], V. Mileyko[1,2]

[1] Laboratory of molecular medicine, Institute of chemical biology and fundamental medicine, Novosibirsk, RU; [2] Atlas Oncology Diagnostics, Moscow, RU; [3]Laboratory of molecular biology and cytogenetics, Russian Scientific Center of Radiology and Nuclear Medicine, Moscow, RU; [4]Department of clinical biotechnologies, N. N. Blokhin Russian Cancer Research Center, Moscow, RU; [5]School of Systems Biology, George Mason University, VA, USA; [6]FSBI Research Centre for Medical Genetics, Moscow, RU

Background

Rapid entering of NGS into the space of clinical diagnostics is complicated by a number of loosely defined unknowns and challenges, including but not excluding:

Bioinformatic pipelines currently producing discordant variant calls

Coverage required for confident detection, data quality control, benchmarking and overall validation of the robustness of NGS techniques

Higher sensitivity of the technology results in producing extensive amount of false positive mutations due to FFPE artifacts which remain unseen with conventional methods

Incidental findings that may be revealed by NGS and are hard to interpret in terms of their clinical relevance and lack a consensus on whether they should be communicated to physician or the patient at all

Bulky diagnostic yield results not only in the complexity of the analysis but also rigorious validation with is often complexed due to shortcomings of the orthogonal methods

Here we describe particular obstacles we encountered while analyzing the clinical NGS dataset obtained using the TruSeq Amplicon - Cancer Panel (TSACP) and possible solutions to the problems presented

Methods

Patient selection

Twelve archived clinical tumor specimens from twelve lung cancer patients treated at Blokhin Russian Cancer Research Centre (RCRC) in 2014-2015 were randomly selected from respective existing registry

Thirteen samples was retrospectively randomly selected from a collection of Russian Scientific Center of Roentgenology and Radiology (RSCRR). Specimens of the latter set had already been screened for the presence of EGFR mutations, hence, this set of samples was enriched with EGFR positive patients by design.

Sequencing and data analysis

Genomic DNA was extracted from formalin fixed, paraffin embedded (FFPE) tissues.

Sequencing libraries were prepared with the TSACP (Illumina, San Diego, USA), according to the manufacturer's pr1otocol. Pooled libraries were sequenced using MiSeqDx (llumina) with a 2 × 150 paired-end sequencing design

Bowtie-2 with the following Varscan2, Strelka and Scalpel accompanied with the in-house software as well as Somatic Variant Caller (SVC, Illumina) were used for data analysis.

Mutation verification

Following NGS sequencing, EGFR and KRAS (including codons 12, 13) mutations in samples from the RCRC set as well as KRAS mutations from RSCRR set were validated either by Sanger sequencing or Real-Time PCR

Baseline patient characteristics

Total patients		25
Gender		
	Male	14 (56%)
	Female	11 (44%)
Age (standard deviation)		59 (54-63)
Histological type		
	Adenocarcinoma	11 (44%)
	Squamous cell cancer	10 (40%)
	Adenosquamous carcinoma	2 (8%)
	Unknown	2 (8%)
Primary tumor size (T1/2/3/4/unknown)		7/10/5/1/2
Regional lymph nodes (N0/1/2/3/unknown)		6/9/4/0/2
Distant metastasis (M0/1/unknown)		20/3/2

Identified somatic mutation of clinical importance

EGFR mutations		11 (44%)
	Exon 19 deletion	4 (16%)
	Exon 19 insertion	1 (4%)
	Exon 20 insertion	1 (4%)
	L858R	3 (12%)
	T790M	1 (4%)
	Gene amplification	2 (8%)
KRAS mutations		
	G12X	5 (20%)
	A146X	1 (4%)
PIK3CA mutations		
	E545K	1 (4%)

challenge #1

Data analysis

Baseline sequencing results characteristics

Reads count per sample (95% CI)		1,4 mln (1,25-1,54)
Aligned reads percent (95% CI)		0,92 (0,9-0,94)
Aligned reads pairs percent (95% CI)		0,92 (0,89-0,94)
Sequencing depth (95% CI)		1932 (1735-2128)
Mutations identified		
	With allele frequency > 1%	5484
	With allele frequency > 10%	784
Unique mutations identified		
	With allele frequency > 1%	4076
	With allele frequency > 10%	312

Across the samples, the overall mean coverage was 2084x and median - 2016x and the frequency of amplicon drop-out was at 0.5%, indicating that coverage resolution was high enough to identify somatic point mutations and short indels of low (up to 1%) mutant allele frequency as well as the copy number variations.

Rare EGFR exon 19 insertion was identified in one patient, which may be associated with EGFR TKI sensitivity. This mutation was not detcted employing the default software (Illumina Somatic Variant Caller) due to misalignment near the end of the reads and was successfully identified with custom pipeline, comprising Bowtie2 in conjunction with Strelka, Varscan2 and Scalpel pipelines

The second false-negative result was inframe deletion in exon 19 that which was compounded with single nucleotide variant, a complex mutation p.Glu746_Ser752delinsAlaPhe was mislabeled as two frameshift mutations p.Gly746fs and p.Ser752fs

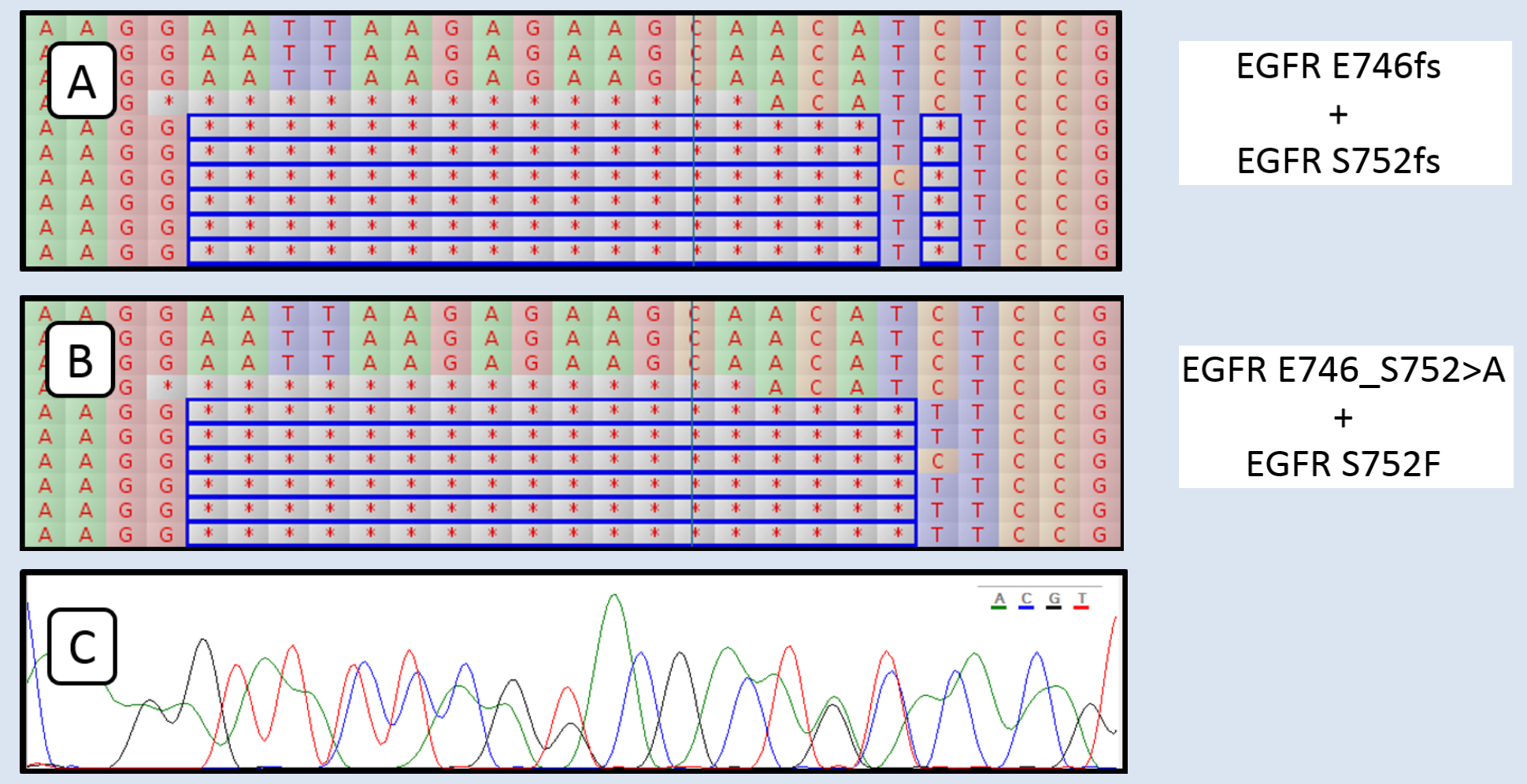

Fig 1. False-negative result with standard pipelines due to misalignment

Another patient with EGFR G719 mutation harbored frameshift mutation in the 717th codon, which would eliminate EGFR activation by G719 mutation. In this case detection of signle G719 mutation would falsely indicate at EGFR TKI sensitivity.

Conclusion #1

Despite relatively small size of the study, a single false positive and two false negative calls of the EGFR mutations were uncovered, indicating that bioinformatic pipelines remain the major sensitivity limitating stage

Aiming at eventual replacement of conventional molecular diagnostics with high-throughput NGS-based methods of NGS, we should remember that commonly used analytic pipelines may remain inefficient when dealing with insertions and deletions, thus, justifying the need for clinically validated pipelines.

challenge #2

Sample quality

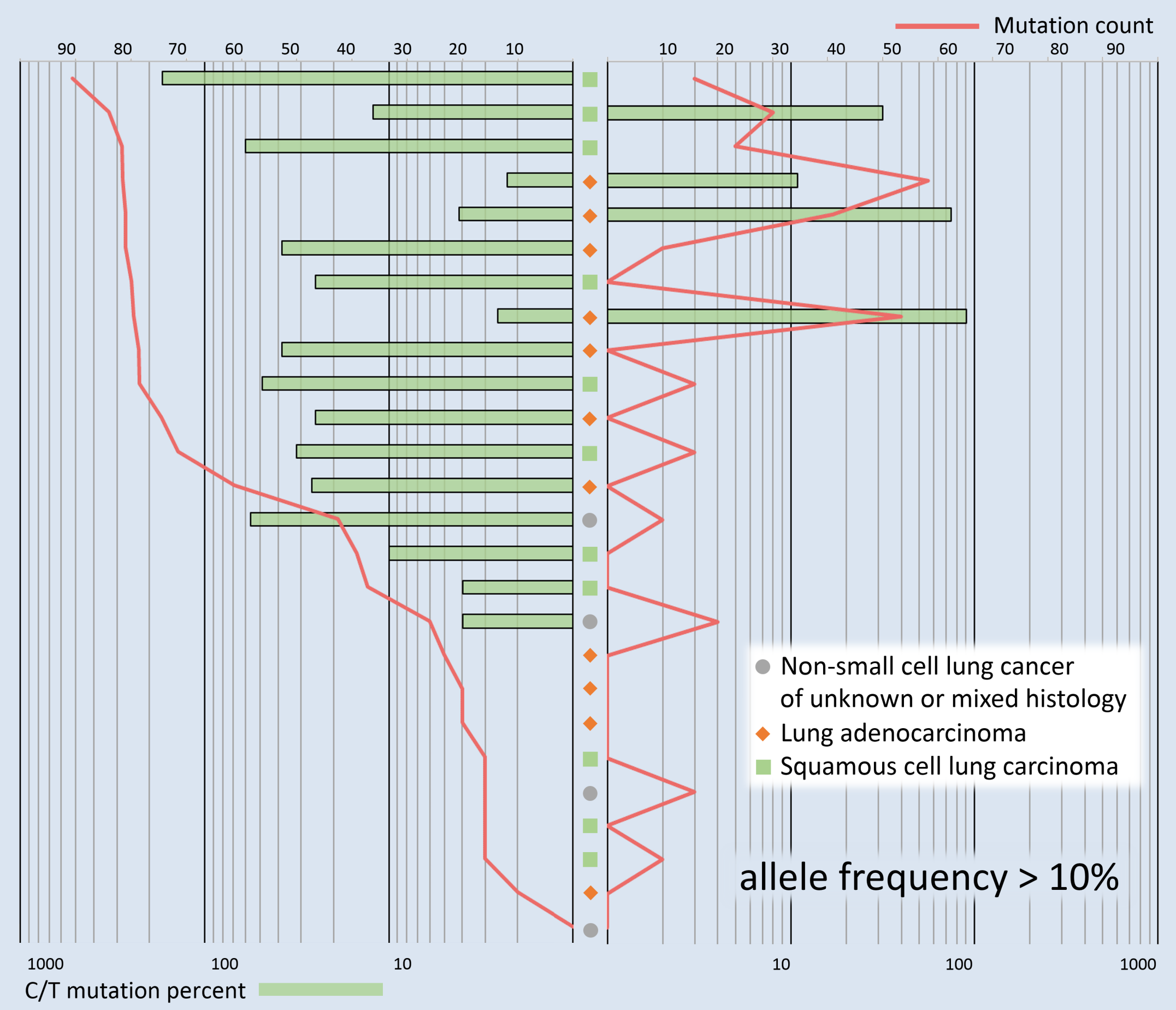

Fig 2. The description of potential artifacts detected in FFPE samples. Mutation counts (solid line) and C>T base substitution rates (bar histogram) at two allele frequency range across all samples were sorted according to the total mutation counts for given specimen. There was a trend of increasing C>T substitution rate along with total mutation count. Four samples did not follow this trend. Only these samples and none of the others showed very high mutation counts at allele frequencies of more than 10% with the accordingly high C>T substitution rate in this range. Pre-normalization library concentrations were low for all four samples in question.

According to TCGA data, both lung adenocarcinoma and squamous cell lung cancer specimens harbor approximately 9 mutations per Mb, or, on average, 0.36 mutations per 40 kb, or the size of the TSACP panel

In our experiment, we detected a bit over 500 variants per 40 kb of sequenced DNA in each sample, on average.

In both histological types of lung cancer samples described in TCGA, the proportion of C>T substitutions was at approximately 20%, while in experimentally assessed DNA specimens extracted from FFPE blocks the percentage of C>T base substitutions was substantially higher ($p < 0.01$)

Four out of 25 FFPE specimens displayed high mutation rate with average allele frequencies of more than 10%, with eight, seventeen, forty and fifty-six of these highly prevalent mutations detected in same DNA sample, respectively. A majority of these highly frequent mutations were C>T substitutions.

Conclusion #2

Clear clinical need of identification of the low-prevalent mutations remains, at least in part, unmet due to use of FFPE for the tumor specimen long-term preservation.

It is tempting to dismiss FFPE artifacts by applying simple allele frequency threshold. An assessment of pre-normalization concentrations of libraries may be part for the sample-specific calculation of the detection cut-offs.

An alternative is in introduction of additional library preparation steps such as so-called "DNA fixing".

challenge #3

Incidental findings

Gene	Protein sequence variation	Patient	Mutant allele frequency	Variant impact
HNF1A	p.Gly306fs	5	15%	deleterious
TP53	p.Gly272fs	9	15%	deleterious
TP53	p.Val173Met	10	17%	deleterious
MLH1	p.Ser406Asn	12	15%	deleterious
KIT	p.Glu76Asp	12	50%	unknown
TP53	p.206_209del	12	70%	unknown
ABL1	p.Thr243Ile	65	15%	unknown
KRAS	p.Gly12Cys	65	31%	activated
CTNNB1	p.Ser33Phe	89	11%	activated
KRAS	p.Gly12Asp	90	30%	activated
TP53	p.Cys238Tyr	91	44%	deleterious
NOTCH1	p.Leu1600Pro	105	11%	activated
TP53	p.Arg175fs	106	56%	deleterious
HRAS	p.Gly13Val	120	82%	activated
AKT1	p.Glu17Lys	120	4%	activated
TP53	p.Ser215Gly	131	45%	deleterious
ATM	p.Asn856Ile	140	10%	unknown
TP53	p.His214Arg	150	33%	deleterious
TP53	p.Arg337Pro	152	50%	deleterious
VHL	p.Lys171Arg	161	18%	deleterious
TP53	p.Arg248Gln	161	31%	deleterious
TP53	p.Ser185fs	187	17%	deleterious
RB1	p.Ser576fs	187	18%	deleterious
PIK3CA	p.Glu545Lys	193	15%	activated
TP53	p.Tyr205Asp	193	22%	deleterious

Table 1. Identified exonic mutations in non-EGFR genes.

In addition to the mutation in known hotspots, a total of 24 unique somatic mutations were detected in 25 studied samples. Assuming that these mutations were never prospectively validated in NSCLC, and, therefore, never received a designation of actionable item, all of these would have to be classified as incidental findings. Among these were the molecular changes validated as actionable in other types of the tumors, or the mutations that influence the prognosis but not yet targeted by approved medicines and non-hotspot mutations associated with treatment resistance.

Conclusion #3

Reporting of entire mutational spectrum revealed by NGS is questionable, at least until the clinically-driven guidelines on somatic mutations are established.
Being well-discussed for conventional EGFR mutations, some questions on incidental findings are still to be resolved, e.g. allele frequency tresholds and somatic/germline status identification.
To ensure utitlity of incidentalome reporting clinical significance of such findings should be assessed along with standardization of sequencing protocols and NGS data analysis including assay-, disease-, and even sample type-specific adjustment.

challenge #4

Diagnostic yield

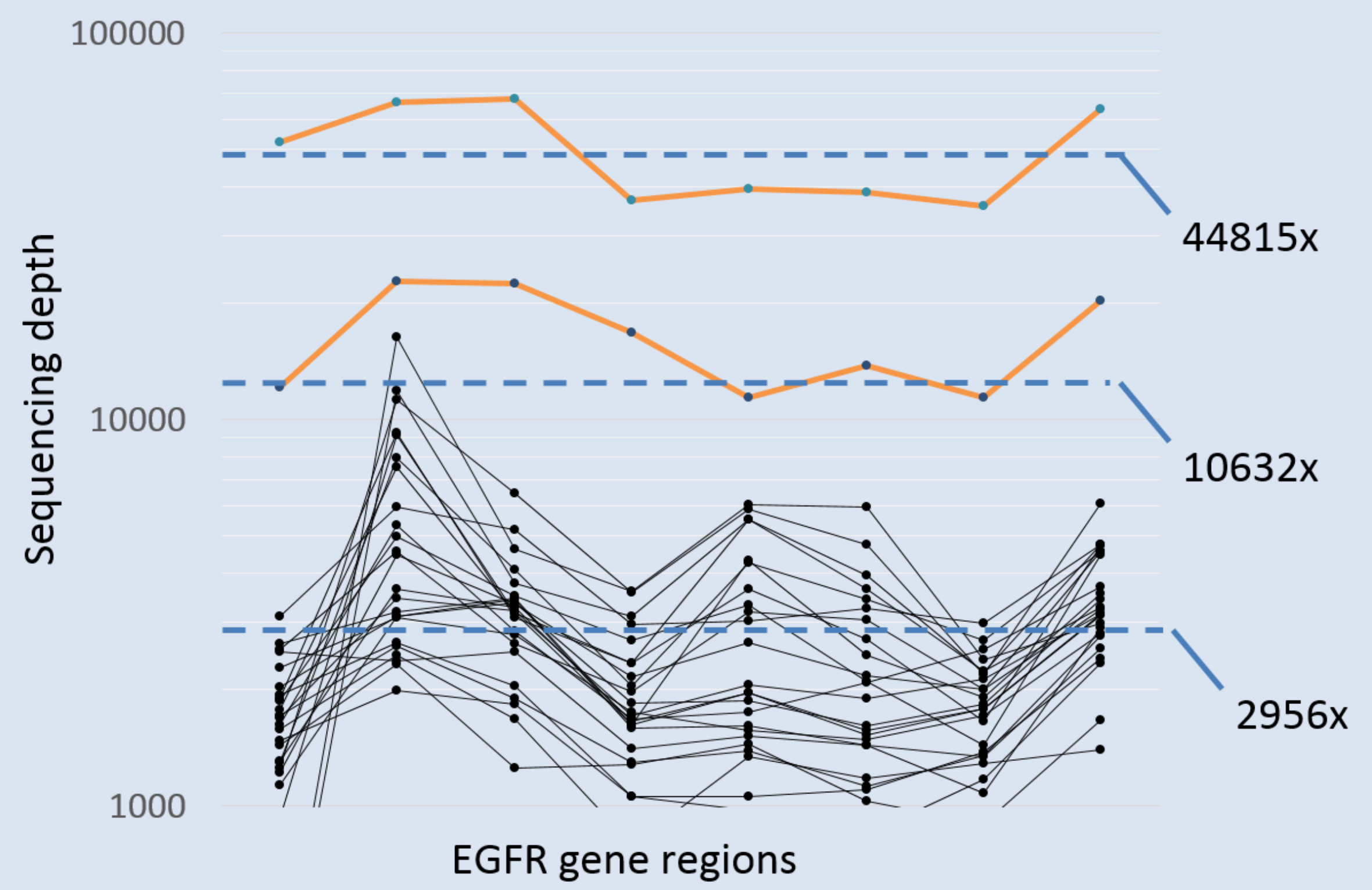

Fig 3. Identified EGFR amplifications.

Even in absence of the sequenced control samples, the Copy Number Variations (CNVs) may be identified by employing bootstrapping that generates control sets by randomly subsampling experimental ones with replacement.

Using this approach, in two specimens an amplification of *EGFR* was detected, with respective increase in its copy number by 5.7 and 17.6 folds.

In another patient, a significant increase in *MET* gene coverage across its five TSACP amplicons was detected. The level of presumable amplification that we observed were low, at 1.65x before normalization and at 1.6x after normalization, thus, indicating that, in specimen of study, amplification of MET had not reached clinical significance (defined as MET:CEP7 >5)

Conclusion #4

NGS source low-level DNA information which further may be analysied with various ways in order to identify mutations of different types. Unless being focused during analysis some alterations may be left unseen, though bearing clinical importance in terms of predictive of prognostic biomarkers.

Acknowledgements

Study was supported by Ministry of Education and Science of the Russian Federation (project RFMEFI60714X0098)

Contacts

atlas oncology diagnostics

Vladislav Mileyko
mileyko@atlas.ru
Atlas Oncology Diagnostics