Non-Small Cell Lung Cancer Testing on Reference Specimens: an Italian Multicenter Experience

Francesco Pepe[1†], Gianluca Russo[1†], Alessandro Venuta[1], Claudia Scimone[1], Mariantonia Nacchio[1], Pasquale Pisapia[1], Gaia Goteri[2], Francesca Barbisan[2], Caterina Chiappetta[3], Angelina Pernazza[4], Domenico Campagna[5], Marco Giordano[5], Giuseppe Perrone[6-7], Giovanna Sabarese[7], Annalisa Altimari[8], Dario de Biase[9], Giovanni Tallini[8-10], Daniele Calistri[11], Elisa Chiadini[11], Laura Capelli[11], Alfredo Santinelli[12], Anna Elisa Gulini[12], Elisa Pierpaoli[12], Manuela Badiali[13], Stefania Murru[13], Riccardo Murgia[14], Elena Guerini Rocco[15,16], Konstantinos Venetis[15], Nicola Fusco[15,16], Denise Morotti[17], Andrea Gianatti[17], Daniela Furlan[18], Giulio Rossi[19], Laura Melocchi[19], Maria Russo[1], Caterina De Luca[1], Lucia Palumbo[1], Saverio Simonelli[1], Antonella Maffè[20], Paola Francia di Celle[21], Tiziana Venesio[22], Maria Scatolini[23], Enrico Grosso[23], Sara Orecchia[24], Matteo Fassan[25-26], Mariangela Balistreri[27], Elisabetta Zulato[27], Daniela Reghellin[28], Elena Lazzari[28], Maria Santacatterina[28], Maria Liliana Piredda[28], Manuela Riccardi[29], Licia Laurino[29], Elena Roz[30], Domenico Longo[31], Daniela Petronilla Romeo[31], Carmine Fazzari[31], Andrea Moreno-Manuel[32-34], Giuseppe Diego Puglia[35], Andrey D. Prjibelski[36], Daria Shafranskaya[36], Luisella Righi[37], Angela Listì[37], Domenico Vitale[1], Antonino Iaccarino[1], Umberto Malapelle[1°], Giancarlo Troncone[1].

[1] Department of Public Health, Federico II University of Naples, Via S. Pansini, 5, 80131 Naples, Italy. francesco.pepe4@unina.it; gianluca.russo@unina.it: alessandro.venuta@unina.it; claudia.scimone@unina.it; mariantonia.nacchio@unina.it; pasquale.pisapia@unina.it; maria.russo6@unina.it; caterina.deluca@unina.it; lucia.palumbo@unina.it; saverio.simonelli95@gmail.com; dvitale2506@gmail.com; antiaccc@hotmail.com; umberto.malapelle@gmail.com; giancarlo.troncone@unina.it

[2] Pathological Anatomy Institute, Polytechnic University of Marche Region, Ancona, Italy gaia.goteri@ospedaliriuniti.marche.it; francesca.barbisan@ospedaliriuniti.marche.it.

[3] Department of Pathology, AOU Policlinico Umberto I, Rome, Italy caterina.chiappetta@uniroma1.it

[4] Department of Medico-Surgical Sciences and Biotechnologies, Polo Pontino-Sapienza University, Latina, Italy angelina.pernazza@uniroma1.it

[5] Department of Pathology, San Giovanni-Addolorata Hospital, 00184, Rome, Italy. dcampagna@hsangiovanni.roma.it; mgiordano@hsangiovanni.roma.it

[6] Research Unit of Anatomical Pathology, Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo, 21-00128 Roma, Italy. g.perrone@policlinicocampus.it;

[7] Anatomical Pathology Operative Research Unit, Fondazione Policlinico Universitario Campus Bio-Medico, Via Alvaro del Portillo, 200-00128 Roma, Italy g.sabarese@policlinicocampus.it; g.perrone@policlinicocampus.it

[8] Molecular Pathology, University of Bologna Hospital of Bologna Sant'Orsola-Malpighi Polyclinic, Bologna, Italy annalisa.altimari@aosp.bo.it; giovanni.tallini@ausl.bologna.it

[9] Pharmacy and Biotechnology (FaBiT), Molecular Pathology Laboratory, University of Bologna, Bologna, Italy dario.debiase@unibo.it; giovanni.tallini@ausl.bologna.it

[10] Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy giovanni.tallini@ausl.bologna.it

[11] Biosciences Laboratory, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", 47014 Meldola, Italy daniele.calistri@irst.emr.it; elisa.chiadini@irst.emr.it; laura.capelli@irst.emr.it;

[12] Anatomic Pathology Azienda Sanitaria Territoriale Pesaro-Urbino, Pesaro, Italy alfredo.santinelli@ospedalimarchenord.it; annaelisa.gulini@ospedalimarchenord.it; elisa.pierpaoli@ospedalimarchenord.it

[13] Laboratory of Genetics and Genomics- Pediatric Hospital A.Cao- ASL8 Cagliari, Italy manuela.badiali@aob.it; stefania.murru@aob.it

[14] Experimental Medicine Unit, Department of Biomedical Sciences, University of Cagliari, 09124 Cagliari, Italy murgia.riccardo@gmail.com

[15] Division of Pathology, IEO, European Institute of Oncology IRCCS, Via Giuseppe Ripamonti 435, 20141 Milan, Italy. elena.guerinirocco@ieo.it; konstantinos.venetis@ieo.it; nicola.fusco@ieo.it

[16] Department of Oncology and Hemato-Oncology, University of Milan, Via Festa del Perdono 7, 20122 Milan, Italy elena.guerinirocco@ieo.it; nicola.fusco@ieo.it

[17] Pathology Unit and Medical Genetics Laboratory, Papa Giovanni XXIII Hospital, Bergamo, Italy dmorotti@asst-pg23.it; agianatti@asst-pg23.it

[18] Pathology Unit, "Department of Medicine and Technological Innovation, University of Insubria, Varese, Italy Daniela.Furlan@uninsubria.it

[19] Department of Anatomical Pathology, Fondazione Poliambulanza, 25124 Brescia, Italy giulio.rossi@poliambulanza.it; laura.melocchi@poliambulanza.it

[20] Genetics and Molecular Biology Unit, Santa Croce e Carle Hospital, 12100 Cuneo, Italy maffe.a@ospedale.cuneo.it

[21] Molecular Pathology, AOU Città della Salute e della Scienza di Torino10126 Turin, Italy pfranciadicelle@cittadellasalute.to.it

[22] Candiolo Cancer Institute, FPO-IRCCS, 10060 Candiolo, Italy tiziana.venesio@ircc.it

[23] Molecular Oncology Lab, Fondazione Edo ed Elvo Tempia, Biella, Italy maria.scatolini@fondazionetempia.org; enrico.grosso@fondazionetempia.org

[24] Pathology Division, S. Antonio and Biagio Hospital, Alessandria, Italy sorecchia@ospedale.al.it

[25] Department of Medicine - DIMED, University of Padua, Padova, Veneto, Italy. matteo.fassan@unipd.it

[26] Veneto Institute of Oncology - IOV - IRCCS, Padova, Italy matteo.fassan@unipd.it

[27] Surgical Pathology Unit, University Hospital of Padua, Padua, Italy Mariangela.balistreri@apd.veneto.it; Elisabetta.zulato@aopd.veneto.it

[28] Department of Pathology, San Bortolo Hospital, Vicenza, Italy daniela.reghellin@aulss8.veneto.it; elena.lazzari@aulss8.veneto.it; maria.santacatterina@aulss8.veneto.it; marialiliana.piredda@aulss8.veneto.it

[29] Department of Pathology, Azienda Ulss3 Serenissima, Ospedale dell'Angelo, Venice, Italy. Manuela.riccardi@aulss3.veneto.it; licia.laurino@aulss3.veneto.it

[30] Pathology Unit, La Maddalena Clinic for Cancer, Palermo, Italy roz@lamaddalenanet.it

[31] UOSD di Anatomia Patologica dell'Azienda Ospedaliera Papardo, Messina, Italy domenicolongo81@yahoo.it; romeopetra@libero.it; cfazzari@hotmail.it

[32] Molecular Oncology Laboratory, Fundación Investigación Hospital General Universitario de Valencia, 46014 Valencia, Spain. andrea.morenomanuel@gmail.com

[33] TRIAL Mixed Unit, Centro Investigación Príncipe Felipe-Fundación Investigación Hospital General Universitario de Valencia, 46014 Valencia, Spain. andrea.morenomanuel@gmail.com

[34] Centro de Investigación Biomédica en Red Cáncer, CIBERONC, 28029 Madrid, Spain andrea.morenomanuel@gmail.com

[35] Institute for Agricultural and Forest Systems in the Mediterranean, National Research Council (ISAFOM-CNR), 95128 Catania, Italy giuseppediego.puglia@cnr.it

[36] Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia. andrewprzh@gmail.com; dariashafranskaya@gmail.com

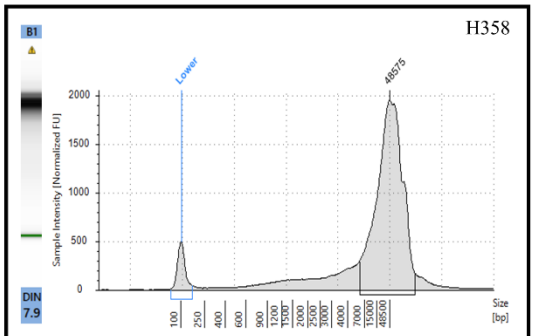
[37] Department of Oncology, University of Turin, San Luigi Hospital, Orbassano (TO), Italy. luisella.righi@unito.it; angela.listi@unito.it

† These Authors contributed equally.
° Corresponding author

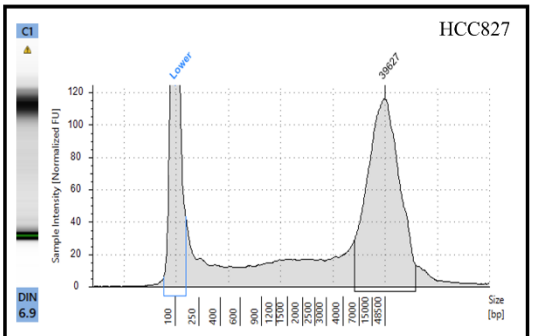
Correspondence to Prof. Umberto Malapelle, Department of Public Health, University Federico II of Naples; umberto.malapelle@unina.it; tel: +390817463674.

Supplementary Figure 1: Cell lines pellet-derived nucleic acids quantifications and qualifications by using TapeStation 4200 platform (Agilent, Santa Clara, CA, USA).

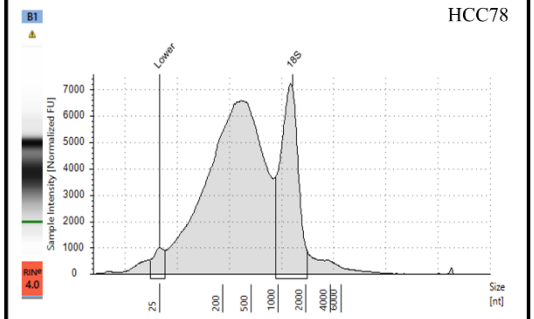
Abbreviations: Conc. (*Concentration*); DIN (*DNA Integrity Number*); ng (*nanograms*); RINe (*RNA Integrity Number equivalent)* µl (*microliters*).



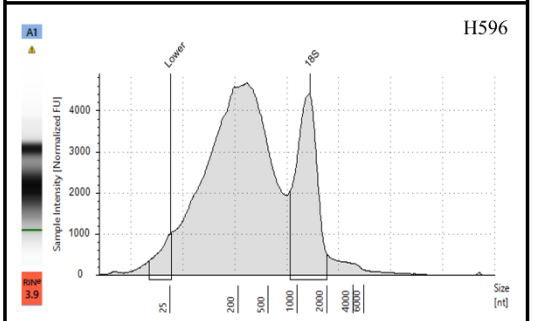
DIN	Conc. [ng/µl]
7.9	115



DIN	Conc. [ng/µl]
6.9	8,90



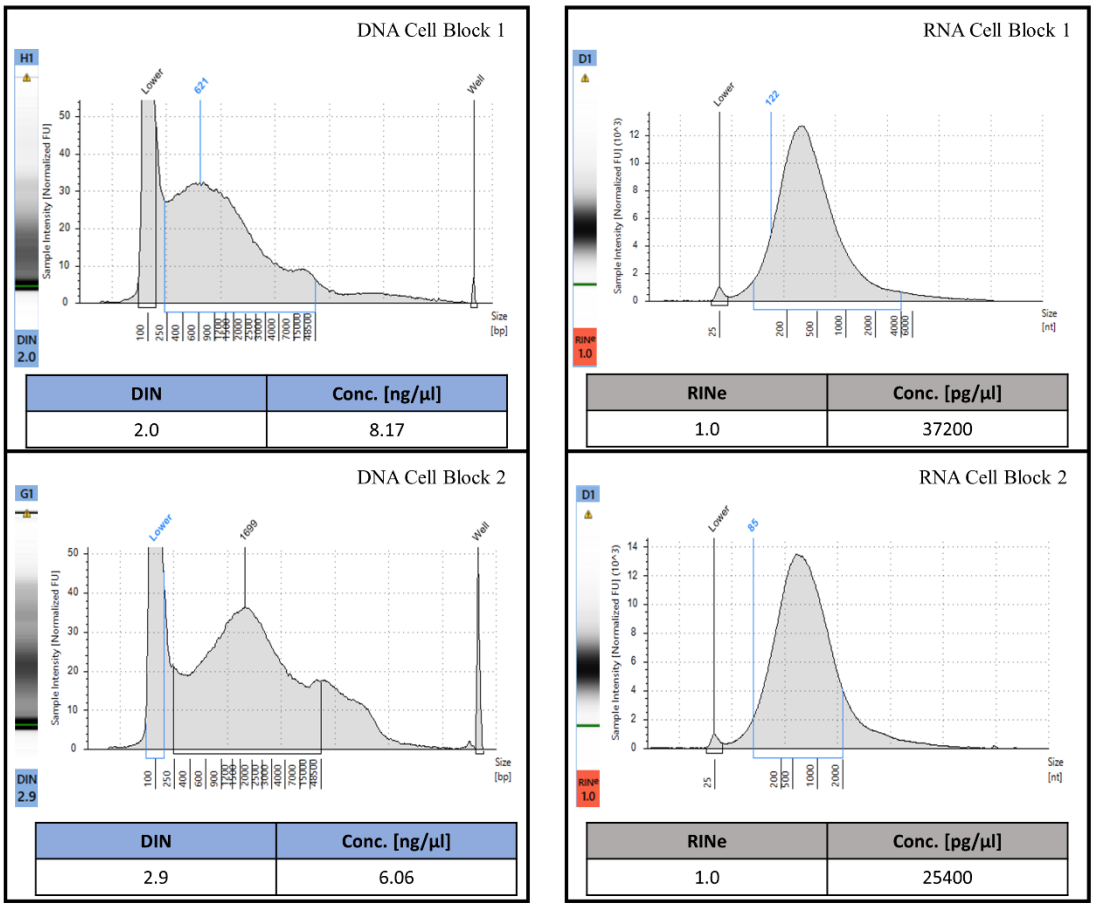
RINe	Conc. [pg/µl]
4.0	15300



RINe	Conc. [pg/µl]
3.9	9480

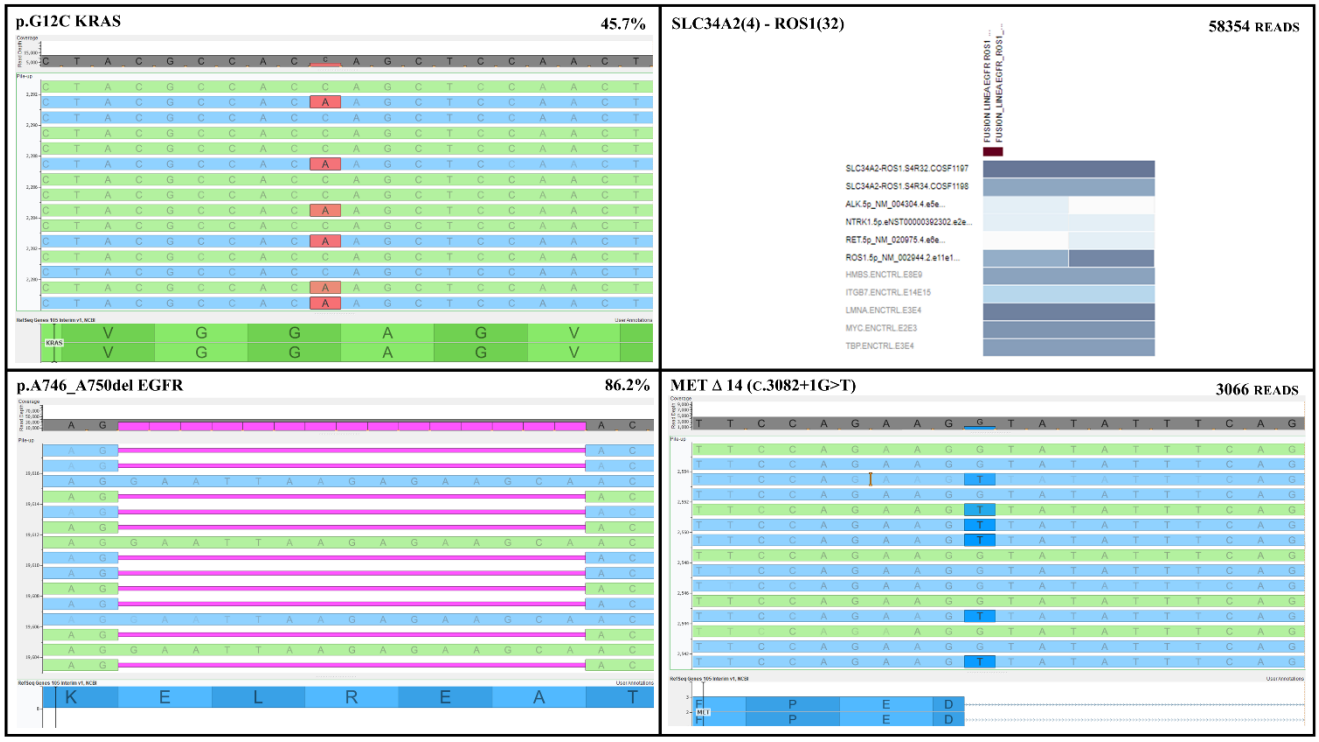
Supplementary Figure 2: DNA and RNA evaluation from cell block specimens of mixed engineered cell lines performed at University of Naples Federico II by using TapeStation 4200 (Agilent, Santa Clara, CA, US) during validation step.

Abbreviations: Conc. (*Concentration*); DIN (*DNA Integrity Number*); ng (*nanograms*); RINe (*RNA Integrity Number equivalent)* µl (*microliters*).



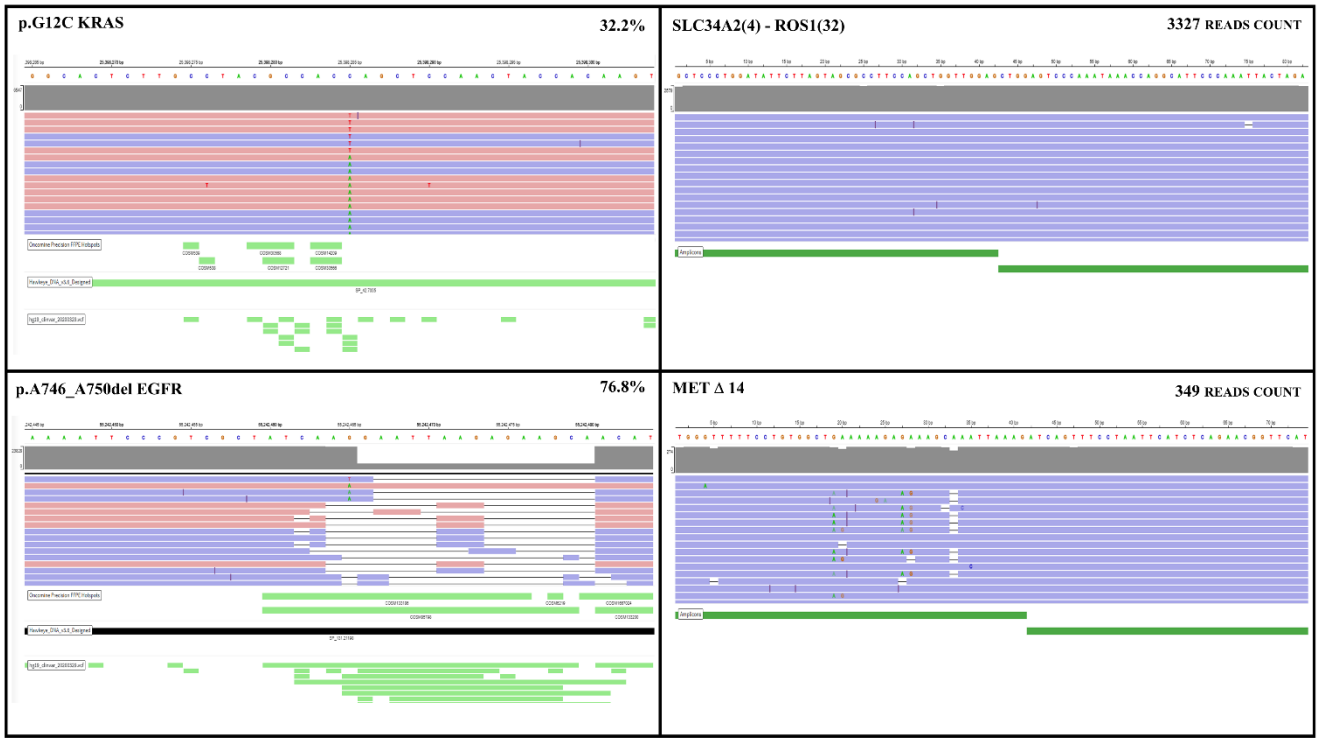
Supplementary Figure 3A: NGS analysis on Cell Block n.1 performed at University Federico II in Naples by using Ion GeneStudio™ S5 Plus System platform (Thermo Fisher Scientifics, Waltham, MA, US).

Abbreviations: EGFR (*Epidermal Growth Factor Receptor*); KRAS (*Kirsten Rat Sarcoma Viral Oncogene Homologue*); MET (*MET proto-oncogene, receptor tyrosine kinase*); ROS1 (*c-Ros Oncogene 1*); SLC34A2 (*Solute Carrier Family 34 Member 2*).



Supplementary Figure 3B: NGS analysis on Cell Block n.2 performed at University Federico II in Naples by using Genexus™ platform (Thermo Fisher Scientifics, Waltham, MA, US).

Abbreviations: EGFR (*Epidermal Growth Factor Receptor*); KRAS (*Kirsten Rat Sarcoma Viral Oncogene Homologue*); MET (*MET proto-oncogene, receptor tyrosine kinase*); ROS1 (*c-Ros Oncogene 1*); SLC34A2 (*Solute Carrier Family 34 Member 2*).



Supplementary Table 1: Growth protocol and referral molecular alteration of commercially available cell lines selected for the study.

Cell line	Molecular Alteration	Complete Growth Medium	Origin
HCC827 (H068)	*EGFR DEL 19 E746 - A750 DELETION*	RPMI-1640 Medium, Fetal Bovine Serum 10%	Human Non-Small Cell Lung Cancer Cells
HCC78	*ROS1 (FUSION)*	RPMI-1640 Medium, Fetal Bovine Serum 10%	Human Non-Small Cell Lung Cancer Cells
H358 (H159)	*KRAS G12C*	RPMI-1640 Medium, Fetal Bovine Serum 10%	Human Non-Small Cell Lung Cancer Cells
H596	*MET Δ 14*	RPMI-1640 Medium, Fetal Bovine Serum 10%	Human Lung adenosquamous carcinoma cell line

Abbreviations: EGFR (*Epidermal Growth Factor Receptor*); KRAS (*Kirsten Rat Sarcoma Virus*); MET (*Mesenchymal Epithelial Transition factor*); ROS1 (*c-ros Oncogene 1*).

Supplementary Table 2: Additional molecular alteration detected by each institution.

Center ID	Mutation	MAF% or Ct
1	**PIK3CA** p.E545K	16,9%
2	**PIK3CA** p.E545K	5,5%
5	**PIK3CA** p.E545K **EGFR** p.S720A **TP53** p.G245C **TP53** p.S241F **TP53** p.V218del	8,3% 5,4% 12,2% 28,8% 62,7%
6	**PIK3CA** p.E545K **TP53** p.G245C **TP53** p.S241F **TP53** p.V218del	6,0% 16,0% 31,0% 54,0%
7	**PIK3CA** p.E545K	8,5%
8	**PIK3CA** p.E545K **TP53** p.V218del	7,4% / 34,4 Ct 61,4%
10	**PIK3CA** p.E545K **TP53** p.G245C **TP53** p.S241F **TP53** p.V218del **MRE11** p.S209F **U2AF1** p.S34F **NOTCH3** p.D139N	8,0% 16,0% 26,0% 54,0% 41,0% 16,0% 30,0%
11	**PIK3CA** p.E545K **TP53** p.G245C **TP53** p.S241F **TP53** p.V218del	13,8% 13,7% 33,3% 48,6%
12	**PIK3CA** p.E545K	20,3%
15	**PIK3CA** p.E545K	7,1%
16	**PIK3CA** p.E545K	27,0%
17	**PIK3CA** p.E545K	19,5%
18	**PIK3CA** p.E545K	8,0%
19	**PIK3CA** p.E545K	21,7%

Abbreviations: Ct (*Cycle threshold*); MAF (*Minor Allele Frequency*); MRE11 (*Meiotic recombination 11 homolog 1*); NOTCH3 (*Neurogenic locus notch homolog protein 3*); PIK3CA (*Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha*); TP53 (*Tumor Suppressor Protein p53*); U2AF1 (*U2 small nuclear RNA auxiliary factor 1*).

Supplementary Material

Library preparation and Sequencing assays

cDNA libraries were made using the SQK-PCS109 cDNA-PCR Sequencing kit (Oxford Nanopore Technologies, UK) according to the manufacturer's protocol. In brief, 50 ng of a mixture of total RNA from two cell lines was denatured at 65 °C for 5 min for oligo-dT (VNP) hybridization, then reverse transcribed using the strand-switching protocol in 20 µl of the reaction mixture as reported in the relevant cDNA-PCR sequencing kit protocol (version: PCS_9035_ v109_revM, Oxford Nanopore Technologies, Oxford, UK). The reaction was incubated at 42 °C for 90 min then with cycles 1 cycle of (80 °C for 10 min) for enzyme inactivation. The double-stranded cDNA was split into four PCR reactions for which we used LongAmp-Taq 2x MasterMix (NEB) and cDNA primer (cPrim) for amplification incubating at 95 °C for 30 s followed by 20 cycles of (95 °C for 15s, 62 °C for 15s, 65 °C for 3min), with a final extension at 65 °C for 6 min. Purification steps were carried out following the manufacturer's instructions by Exonuclease I digestion and affinity purification by AMPure XP beads (Agencourt, Beckman Coulter, Beverly, USA). The purified library was quantified by fluorometric quantitation and 200 fmol was used in the reaction of adapter addition. Then library was mixed with library loading beads and running buffer with fuel mix provided in ONT protocol. The full-length cDNA library was then sequenced on a MinION R9.5 flow cells for 24 h using the High accuracy (HAC) sequencing protocol.

Data analysis

Primary data acquisition and FAST5 files production was performed by MinKNOW, the operating software that drives nanopore sequencing devices. Then secondary analysis and FASTQ files production was performed on a Google Colabs remote server by Guppy using two different basecaller modes: high

speed (fast) and high accuracy (sup). The obtained reads, as reported in tab.1, were used as long-sequencing reads (LR-seq) for downstream analyses.

LR-seq data was aligned to the human GRCh38.p13 reference genome using minimap2 in spliced mode (-x splice). For more accurate spliced alignment GENCODE v36 reference annotation was also provided to minimap2 as input. Resulting alignments were processed with two different fusion discovery methods: LongGF and JAFFAL. In both tools the same GENCODE v36 annotation was used. Additionally, we applied a direct target approach by creating fusion transcripts for the specific known fusion genes from FusionGDB and aligning LR-seq reads with minimap2 in non-spliced mode (-x map-ont).

Results and Discussion

To benchmark and evaluate the possibility of using low-coverage ONT-Nanopore long-read sequencing to detect SVs in different cell types we tested standard full-length PCR cDNA long sequencing assay in two cell lines: lung cancer (NCI-H596) and breast cancer (MCF7). A mixture of total RNA from the above cell lines was sequenced by the Nanopore PCR full-length cDNA protocol and sequencing data was used to search structural variation (SV) by two bioinformatic pipelines LongGF, Jaffal and in the target search approach by MiniMap2. From the ONT-Nanopore sequencing run, we obtain a total of 6,917,552 reads base-called by high speed (FAST) and 6,216,200 reads in super-accuracy (SUP) mode with an average reading length of 700 bp and percentage of reads mapping against GRCh38.p13.genome of 79.95% in the super-accuracy data-set (see Table 1). Selecting only fusion genes arose by gene-gene pairs and supported at least by two long reads, our analysis provided a list of 25 fusion gene candidates by Jaffal and 28 by LongGF (see table 2). Among them, six fusion gene candidates are in common between both pipelines as reported in table 3. Due to the lack of a fully validated gene fusion dataset in

these two cell lines, we have used data from the wide bibliography where frequently are reported discordant data about fusion genes detected. To this end, we identified for MCF7 a set of four fusions previously described as reported in table 4, in our data we can confirm RPS6KB1-VMP1and AC099850.1-VMP1 fusion genes on the other hand we did not find any fusion occurrence about the other three genes couples. About (NCI-H596) cell line all data present in the literature agree with the presence of MET Exon 14 Skipping11 in our data we have 68 reads mapping on Met genes by MiniMap2 and four reads cover the exon 14 region where two reads show exon 14 skipping fig.1. Moreover, this modification was confirmed manually by MiniMap2 and by IsoQuant a tool to identify isoforms from long-read RNA sequencing.

Latest Next Generation Sequencing technologies opened the way to a novel era of genomic studies, allowing us to gain novel insights into multifactorial pathologies such as cancer. In particular, gene fusion detection and comprehension have been deeply enhanced by these methods. Indeed the possibility of third-generation sequencing to overcome all limitations of the high fragmentation of the second-generation NGS-based approach is crucial to resolve long repeats and highlight candidate genes fusion. The accurate and timely detection of tumour-associated SVs can make a difference in the early treatments and for molecular monitoring of disease relapse, as well as determining or predicting patient chemoresponse. Tumour-associated SVs are hardly revealed when they occur in repetitive regions, which account for over half of the human genome. The ability of long-read 3rd-generation sequencing methods, such as ONT-Nanopore, to read through repetitive regions could make it an ideal tool for detecting tumour-associated SVs. This work serves as proof of principle, showing the ability of this sequencing approach to correctly and reliably detect SVs with only hundreds, instead of millions of reads. Moreover, ONT-Nanopore technology has attracted much attention due to the portability and low cost of this novel sequencing instrument, indeed the cost per device is approximately 1,000 € and the small size, like a USB pen, of the MinION nanopore sequencing instrument, offers accessibility to testing in nearly any

setting. In contrast, the instrumentation for 2nd generation methods requires a substantial upfront investment greater than 100,000 € and sufficient lab space for their large footprint, which are prohibitive to many research and clinical labs. We have designed, developed, and verified an mRNA-Seq assay that can be potentially extended to detect gene fusions in tumour specimens. This test in the target approach could be used clinically to aid in diagnosis and to guide targeted therapy decisions for patients with solid and hematologic tumours. The main limitations that probably can restrict the utility of ONT-Nanopore sequencing in clinical testing is a relatively high mismatch and indel error rate but this factor is in rapid improvement.

Candidate fusion genes highlighted in the above cell lines show some differences in fusion reported in the literature, for which frequently contrasting data are reported about the genomic characterization of cancer cell lines. As regards to NCI-H596 cell line, in agreement with literature data, our approach confirmed the identification of Exon 14 skipping in the Met gene. On the other hand, for MCF-7 cell line almost 30,000 articles can be retrieved in PubMed indicating a low level of reproducibility probably as a result of biological variation and different technical approaches used, but the most of studies take into account mainly the technical aspect overlooking the biological bias. Indeed, only few studies highlighted how genomic instability that characterizes cancer cell lines can produce an effect on over-cultured, genomic and epigenetic modification. As a consequence, the genetic heterogeneity in cancer cell lines leads to varying patterns of gene expression as a result of positive clonal selection that is highly sensitive to culture conditions. In other words, much evidence showed that prolonged cell culture is more likely to induce the occurrence of secondary genomic changes, such as copy number variations as well as transcriptomic drifts. This process may have occurred in the cell lines analyzed in the present study and can explain the discrepancy to literature data. Not even STR cell authentication techniques would ensure that “their sub-clone” will behave with sufficient stability and reproducibility, even in case the starting

material is a single batch of cells from a cell bank, limiting the number of passages and cell authentication by short tandem repeat markers (STR). Indeed, in some reports gene expression patterns suggested that MCF-7 cells of the same batch from a cell bank include subpopulations with different genomic backgrounds. Cancer cell lines are a powerful tool for cancer research and as reference material to use in tests validation interlaboratory programs, but their genomic evolution leads to a high degree of variation across cell line strains and generations, which must be considered in experimental design and is mandatory a multi approaching technique in genomic characterization of the starting reference control specimens.

Supplementary Table 3: Full-length cDNA long Sequencing results using FAST or SUP Guppy base-calling mode.

	Total number reads	**Mapped**	**Percentage Mapped (%)**	**Average Reading Length (bp)**
Fast	6,917,552	5,171,856	74.76	671
Sup	6,216,200	4,970,122	79.95	716

Supplementary Table 4: fusion analysis results from Jafall and LongGF application.

Abbreviations: ANXA2 (*Annexin A2*); ATP1A1 (*ATPase Na+/K+ Transporting Subunit Alpha 1*); AXL (*AXL Receptor Tyrosine Kinase*); CALM2 (*Calmodulin 2*); CASKIN2 (*CASK Interacting Protein 2*); CBX3 (*Chromobox 3*); CCDC170 (*Coiled-Coil Domain Containing 170*); CCDC32 (*Coiled-Coil Domain Containing 32*); CDC5L (*Cell Division Cycle 5 Like*); CPNE3 (*Copine 3*); CRTAP (*Cartilage Associated Protein*); DDX27 (*DEAD-Box Helicase 27*); EIF6 (*Eukaryotic Translation Initiation Factor 6*); ESR1 (*Estrogen Receptor 1*); F3 (*Coagulation Factor III*); FKBP14 (*FKBP Prolyl Isomerase 14*); GNL3L (*G Protein Nucleolar 3 Like*); HIF1A (*Hypoxia Inducible Factor 1 Subunit Alpha*); HRH1 (*Histamine Receptor H1*); IFT52 (*Intraflagellar Transport 52*); KIAA0319L (*Dyslexia-Associated Protein KIAA0319-Like*); LAMTOR2 (*Late Endosomal/Lysosomal Adaptor, MAPK And MTOR Activator 2*); LSG1 (*Large 60S Subunit Nuclear Export GTPase 1*); LYRM2 (*LYR Motif Containing 2*); MAPRE1 (*Microtubule Associated Protein RP/EB Family Member 1*); MATN2 (*Matrilin 2*); MT-ATP6 (*Mitochondrially Encoded ATP Synthase Membrane Subunit 6*); MT-CO3 (*Mitochondrially Encoded Cytochrome C Oxidase III*); MT-CYB (*Mitochondrially Encoded Cytochrome B*); MT-ND4 (*Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 4*); MT-ND5 (*Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 5*); MT-ND6 (*Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 6*); NBEAL1 (*Neurobeachin Like 1*); NDUFA4 (*NDUFA4 Mitochondrial Complex Associated*); NSA2 (*NSA2 Ribosome Biogenesis Factor*); ORC6 (*Origin Recognition Complex Subunit 6*); PAIP2 (*Poly(A) Binding Protein Interacting Protein 2*); PHF14 (*PHD Finger Protein 14*); POP1 (*POP1 Homolog, Ribonuclease P/MRP Subunit*); PTBP2 (*Polypyrimidine Tract Binding Protein 2*); QRSL1 (*Glutaminyl-TRNA Amidotransferase Subunit QRSL1*); RP11 (*Pre-MRNA Processing Factor 31*); RPL12 (*Ribosomal Protein L12*); RPL30 (*Ribosomal Protein L30*); RPS6KB1 (*Ribosomal Protein S6 Kinase B1*); S100A11 (*S100 Calcium Binding Protein A11*); S100A6 (*S100 Calcium Binding Protein A6*); SPIN1 (*Spindlin 1*); SPTLC1 (*Serine Palmitoyltransferase Long Chain Base Subunit 1*); SRSF9 (*Serine And Arginine Rich Splicing Factor 9*); SSBP4 (*Single Stranded DNA Binding Protein 4*); SSR2 (*Signal Sequence Receptor Subunit 2*); SUPT3H (*SPT3 Homolog, SAGA And STAGA Complex Component*); TMSB4X (*Thymosin Beta 4 X-Linked*); VMP1 (*Vacuole Membrane Protein 1*); WWP1 (*WW Domain Containing E3 Ubiquitin Protein Ligase 1*); YWHAE (*Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Epsilon*); ZFP64 (*ZFP64 Zinc Finger Protein*); ZNF616 (*Zinc Finger Protein 616*).

	LongGF			JAFFAL			
Genes	**Supporting reads number**	**Left Breakpoint**	**Right Breakpoint**	**Genes**	**Supporting reads number**	**Left Breakpoint**	**Right Breakpoint**
DDX27:EIF6	29	chr20:49219540	chr20:35280117	AC099850.1:VMP1	42	chr17:59107591	chr17:59838295
NBEAL1:RPL12	11	chr2:203190798	chr9:127451403	DDX27:EIF6	27	chr20:49219541	chr20:35280118
ESR1:CCDC170	8	chr6:151702004	chr6:151573171	ESR1:CCDC170	8	chr6:151702005	chr6:151573174
MT-ATP6:MT-CO3	7	chrM:8857	chrM:9551	POP1:MATN2	7	chr8:98117390	chr8:980304602
ZFP64:ATP1A1	2	chr20:52052133	chr1:116396734	SPTLC1:RP11-27K13.3	5	chr9:92080016	chr1:117050334
VMP1:IFT52	2	chr17:59839998	chr20:43594754	RPS6KB1:VMP1	2	chr17:59910611	chr17:59838295
VMP1:RPS6KB1	2	chr17:59838294	chr17:59910610	NDUFA4:PHF14	2	chr7:109400016	chr7:111113500
LYRM2:AL139353.1	2	chr6:89636905	chr14:31427611	RP11-96H19.1:RP11-446N19.1	2	chr12:46387972	chr12:46652390
CPNE3:WWP1	2	chr8:86561123	chr8:864354551	YWHAE:ANXA2	2	chr17:1400047	chr15:60364523
CDC5L:SUPT3H	2	chr6:44445658	chr6:453651959	HRH1:PAIP2	2	chr3:11137399	chr5:139363759
MAPRE1:MT-CYB	2	chr20:32848998	chrM:14746	SPTLC1:RP11-27K13.3	2	chr9:92080016	chr1:117032336
FKBP14:AXL	2	chr7:30013853	chr19:41260589	ATP1A1:ZFP64	2	chr1:116396734	chr20:52052132
PHF14:NDUFA4	2	chr7:11111349	chr7:109400013	HIF1A-AS2:SRSF9	2	chr14:61750885	chr12:120464122
SPIN1:S100A6	2	chr9:88477342	chr1:153534626	PTBP2:C8orf76	2	chr1:96751500	chr8:123231757
LAMTOR2:SSR2	2	chr1:156055671	chr1:156016483	AC099850.1:VMP1	8	chr17:59107503	chr17:59838295
PTBP2:C8orf76	2	chr1:96751500	chr8:123231756	AC099850.1:VMP1	2	chr17:59118046	chr17:59838293
PAIP2:HRH1	2	chr5:139363757	chr3:111373938	AC099850.1:VMP1	2	chr17:59107327	chr17:59838295
CBX3:CCDC32	2	chr7:26201766	chr15:40562770	S100A6:SPIN1	2	chr1:153534641	chr9:884773538
GNL3L:NSA2	2	chrX:54561059	chr5:747689683	C9orf78:CRTAP	2	chr9:129828113	chr3:331429448
QRSL1:AXL	2	chr6:106640383	chr19:41260504	C15orf57:CBX3	2	chr15:40562772	chr7:262017609
C9orf78:CRTAP	2	chr9:129828106	chr3:331429400	SSBP4:ORC6	2	chr19:18434406	chr16:46698221
KIAA0319L:MT-ND4	2	chr1:35449948	chrM:11221	F3:LSG1	2	chr1:94530150	chr3:194641620
CALM2:MT-ATP6	2	chr2:47176510	chrM:8778	IFT52:VMP1	2	chr20:43594751	chr17:59840002
RPL30:ZNF616	2	chr8:98041852	chr19:52123918	QRSL1:AXL	2	chr6:106640375	chr19:41260513
S100A11:TMSB4X	2	chr1:152032628	chrX:129769788	FKBP14:AXL	2	chr7:30013858	chr19:41260595
CASKIN2:MT-ND4	2	chr17:75503963	chrM:11229				
MT-ND6:MT-ND5	2	chrM:14414	chrM:13843				
LSG1:F3	2	chr3:194641628	chr1:945301401				

Supplementary Table 5: Common fusion genes found by Jaffal and LongGF.

Abbreviations: AXL (*AXL Receptor Tyrosine Kinase*); CCDC170 (*Coiled-Coil Domain Containing 170*); CRTAP (*Cartilage Associated Protein*); DDX27 (*DEAD-Box Helicase 27*); EIF6 (*Eukaryotic Translation Initiation Factor 6*); ESR1 (*Estrogen Receptor 1*); FKBP14 (*FKBP Prolyl Isomerase 14*); PTBP2 (*Polypyrimidine Tract Binding Protein 2*); QRSL1 (*Glutaminyl-TRNA Amidotransferase Subunit QRSL1*).

	LongGF			**JAFFAL**		
Genes	**Supporting reads number**	**Left Breakpoint**	**Right Breakpoint**	**Supporting reads number**	**Left Breakpoint**	**Right Breakpoint**
DDX27:EIF6	29	chr20:49219540	chr20:35280117	27	chr20:49219541	chr20:35280118
ESR1:CCDC170	8	chr6:151702004	chr6:151573171	8	chr6:151702005	chr6:151573174
PTBP2:C8orf76	2	chr1:96751500	chr8:123231756	2	chr1:96751500	chr8:123231757
C9orf78:CRTAP	2	chr9:129828106	chr3:33142940	2	chr9:129828113	chr3:33142948
QRSL1:AXL	2	chr6:106640383	chr19:41260504	2	chr6:106640375	chr19:41260513
FKBP14:AXL	2	chr7:30013853	chr19:41260589	2	chr7:30013858	chr19:41260595

Supplementary Table 6: fusion genes reported in MCF7 and verify in our data.

Abbreviations: ARFGEF2 (*ADP Ribosylation Factor Guanine Nucleotide Exchange Factor 2*); BCAS3 (*BCAS3 Microtubule Associated Cell Migration Factor*); BCAS4 (*Breast Carcinoma Amplified Sequence 4*); CARM1 (*Coactivator Associated Arginine Methyltransferase 1*); RPS6KB1 (*Ribosomal Protein S6 Kinase B1*); SMARCA4 (*SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily A, Member 4*); SULF2 (*Sulfatase 2*); VMP1 (Vacuole Membrane Protein 1).

Genes	**Chr**	**Genes**	**Chr**	**Expression**	**Fusion**
BCAS4	20	BCAS3	17	Yes	No
ARFGEF2	20	SULF2	20	No	No
RPS6KB1	17	VMP1	17	Yes	Yes
AC099850.1	17	VMP1	17	Yes	Yes
SMARCA4	19	CARM1	19	No	No

References

1. F. Mitelman, B. Johansson, F. Mertens, *Nat. Rev. Cancer* 2007, *7*, 233

2. M. J. Annala, B. C. Parker, W. Zhang, M. Nykter, Cancer Lett. 2013, 340, 192.

3. M. O. Pollard, D. Gurdasani, A. J. Mentzer, T. Porter, M. S. Sandhu, *Hum. Mol. Genet.* 2018, *27*, R234.

4. Lin B, Hui J, Mao H. Nanopore Technology and Its Applications in Gene Sequencing. Biosensors [Internet]. 2021; 11. doi: 10.3390/bios11070214.

5. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19:R227–40.

6. Liu Q, Hu Y, Stucky A, Fang L, Zhong JF, Wang K. BMC Genomics. 2020 Dec 29;21(Suppl 11):793. doi: 10.1186/s12864-020-07207-4.PMID: 33372596

7. Davidson NM, Chen Y, Sadras T, Ryland GL, Blombery P, Ekert PG, Göke J, Oshlack A. Genome Biol. 2022 Jan 6;23(1):10. doi: 10.1186/s13059-021-02588-5.

8. Sota Y, Seno S, Shigeta H, Osato N, Shimoda M, Noguchi S, Matsuda H.J Bioinform Comput Biol. 2019 Jun;17(3):1940008. doi: 10.1142/S0219720019400080.PMID: 31288642

9. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O.Genome Biol. 2011;12(1):R6. doi: 10.1186/gb-2011-12-1-r6. Epub 2011 Jan 19.PMID: 21247443

10. Paciello G, Ficarra E.BMC Bioinformatics. 2017 Jan 23;18(1):58. doi: 10.1186/s12859-016-1450-6.PMID: 28114882

11. Nosi V, Luca A, Milan M, Arigoni M, Benvenuti S, Cacchiarelli D, Cesana M, Riccardo S, Di Filippo L, Cordero F, Beccuti M, Comoglio PM, Calogero RA.Int J Mol Sci. 2021 Apr 19;22(8):4217. doi: 10.3390/ijms22084217.PMID: 33921709

12. Li H. Bioinformatics. 2018 Sep 15;34(18):3094-3100. doi: 10.1093/bioinformatics/bty191.PMID: 29750242

13. Kleensang A, Vantangoli MM, Odwin-DaCosta S, Andersen ME, Boekelheide K, Bouhifd M, Fornace AJ Jr, Li HH, Livi CB, Madnick S, Maertens A, Rosenberg M, Yager JD, Zhao L, Hartung T.Sci Rep. 2016 Jul 26;6:28994. doi: 10.1038/srep28994.PMID: 27456714

14. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, Strathdee CA, Dempster J, Lyons NJ, Burns R, Nag A, Kugener G, Cimini B, Tsvetkov P, Maruvka YE, O'Rourke R, Garrity A, Tubelli AA, Bandopadhayay P, Tsherniak A, Vazquez F, Wong B, Birger C, Ghandi M, Thorner AR, Bittker JA, Meyerson M, Getz G, Beroukhim R, Golub TR**.** Nature. 2018 Aug;560(7718):325-330. doi: 10.1038/s41586-018-0409-3. Epub 2018 Aug 8.PMID: 30089904

15. Gillet JP, Varma S, Gottesman MM.J Natl Cancer Inst. 2013 Apr 3;105(7):452-8. doi: 10.1093/jnci/djt007. Epub 2013 Feb 21.PMID: 23434901

16. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, Berry A. Nucleic acids research. 2021 Jan 8;49(D1):D916-23.
doi: https://doi.org/10.1093/nar/gkaa1087

17. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU.
Nat Biotechnol. 2023 Jan 2. doi: 10.1038/s41587-022-01565-y. Online ahead of print.PMID: 36593406