Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Clin Cancer Res. Author manuscript; available in PMC 2022 March 15.

Published in final edited form as:
Clin Cancer Res. 2021 September 15; 27(18): 5049–5061. doi:10.1158/1078-0432.CCR-20-3251.

Cross-site concordance evaluation of tumor DNA and RNA sequencing platforms for the CIMAC-CIDC network

Zexian Zeng[1,2,*], **Jingxin Fu**[1,3,*], **Carrie Cibulskis**[4,^], **Aashna Jhaveri**[1,^], **Curtis Gumbs**[5,6,^], **Biswajit Das**[7,^], **Beatriz Sanchez- Espiridion**[8,9,^], **Sylvie Janssens**[10,^], **Len Taing**[11], **Jin Wang**[1,3], **James Lindsay**[1], **Tomas Vilimas**[7], **Jianhua Zhang**[5,6], **Collin Tokheim**[1,2], **Avinash Sahu**[1,2], **Peng Jiang**[12], **Chunhua Yan**[13], **Dzifa Yawa Duose**[8,9], **Ethan Cerami**[1], **Li Chen**[7], **David Cohen**[1], **Qingrong Chen**[13], **Rebecca Enos**[10], **Xin Huang**[1], **Jack J Lee**[14,15], **Yang Liu**[1,2], **Donna S Neuberg**[1], **Cu Nguyen**[13], **Candace Patterson**[4], **Sharmistha Sarkar**[8,9], **Sachet Shukla**[16,11], **Ming Tang**[1,2], **Junko Tsuji**[4], **Mohamed Uduman**[17,1], **Xiaoman Wang**[1], **Jason L Weirather**[17,1], **Jijun Yu**[1], **Joyce Yu**[1], **Jianjun Zhang**[18], **Jiexin Zhang**[19], **Daoud Meerzaman**[13,%], **Magdalena Thurin**[20,%], **Andrew Futreal**[5,6,%], **Chris Karlovich**[7,%], **Stacey B Gabriel**[4,%], **Ignacio Ivan Wistuba**[8,9,&], **X. Shirley Liu**[1,21,&], **Catherine J. Wu**[11,22,4,&]

[1]Department of Data Science, Dana Farber Cancer Institute, Boston, MA, USA.

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[3]Clinical Translational Research Center, Shanghai Pulmonary Hospital, School of Life Science and Technology, Tongji University, Shanghai, China.

[4]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[5]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[6]Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[7]Molecular Characterization Laboratory, Frederick National Laboratory for Cancer Research, Frederick, MD, USA

[8]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Correspondence to: Catherine J Wu, MD, Professor of Medicine, Harvard Medical School, Dana 520C, 450 Brookline Avenue, Boston, MA 02115, Tel: 617-632-5943, cwu@partners.org, X. Shirley Liu, PhD, Professor, Department of Data Science, Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute and Harvard University, CLS11022, 3 Blackfan Circle, Boston 02115, Tel: 617-632-2444, xsliu@ds.dfci.harvard.edu.
*equal first author contributions;
^co-second authors,
%equal contributors;
&co-senior authors

Competing interest statement:
X.S.L. is a cofounder and board member of GV20 Oncotherapy and its subsidiaries, SAB of 3DMed Care, consultant for Genentech, stockholder of BMY, TMO, WBA, ABT, ABBV, and JNJ, and received research funding from Sanofi and Takeda. C.J. Wu holds equity in BioNtech.

Data Availability
All human WES and RNA-seq data presented in this article have been deposited at The database of Genotypes and Phenotypes (dbGaP) under accession number phs002295.v1.p1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[9]Division of Pathology and Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[10]The Emmes Company, LLC, Rockville, MD, USA.

[11]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA.

[12]Cancer Data Science Lab, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.

[13]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, Maryland, USA.

[14]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[15]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[16]Translational Immunogenomics Laboratory, Dana-Farber Cancer Institute, Boston, MA, USA.

[17]Center for Immuno-Oncology, Dana Farber Cancer Institute, Boston, MA, USA.

[18]Department of Thoracic/Head and Neck Medical Oncology and Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[19]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[20]Cancer Diagnosis Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD

[21]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA.

[22]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA.

Abstract

Purpose: Whole-exome (WES) and RNA-sequencing (RNA-seq) are key components of cancer immunogenomic analyses. To evaluate the consistency of tumor WES and RNA-seq profiling platforms across different centers, the Cancer Immune Monitoring and Analysis Centers (CIMACs) and the Cancer Immunologic Data Commons (CIDC) conducted a systematic harmonization study.

Experimental Design: DNA and RNA were centrally extracted from fresh frozen (FF) and formalin-fixed paraffin-embedded (FFPE) non-small cell lung carcinoma (NSCLC) tumors and distributed to three centers for WES and RNA-seq profiling. In addition, two 10-plex HapMap cell-line pools with known mutations were used to evaluate the accuracy of the WES platforms.

Results: The WES platforms achieved high precision (> 0.98) and recall (> 0.87) on the HapMap pools when evaluated on loci using > 50X common coverage. Non-synonymous mutations clustered by tumor sample, achieving an Index of Specific Agreement above 0.67 among replicates, centers, and sample processing. A DV200 > 24% for RNA, as a putative pre-sequencing RNA quality control (QC) metric, was found to be a reliable threshold for generating consistent

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

expression readouts in RNA-seq and NanoString data. MedTIN > 30 was likewise assessed as a reliable RNA-seq QC metric, above which samples from the same tumor across replicates, centers, and sample processing runs could be robustly clustered and HLA typing, immune infiltration, and immune repertoire inference could be performed.

Conclusions: The CIMAC collaborating laboratory platforms effectively generated consistent WES and RNA-seq data and enable robust cross-trial comparisons and meta-analyses of highly complex immuno-oncology biomarker data across the NCI CIMAC-CIDC Network.

Keywords

WES; RNA-seq; harmonization; immunogenomic; NSCLC

Introduction

The Cancer Immune Monitoring and Analysis Centers - Cancer Immunologic Data Commons (CIMAC-CIDC) Network (https://cimac-network.org/) is an NCI Cancer Moonshot initiative that provides cutting-edge technology and expertise in genomic, proteomic, and functional molecular analysis to enhance clinical trials in cancer immune therapies. CIMACs serve as the main units of the Network for correlative studies in clinical trials involving cancer immunotherapy, functioning as platforms for deep molecular characterization of tumor and immune profiling using state-of-the-art analytically-validated and standardized platforms. The CIDC, hosted by Dana-Farber Cancer Institute (DFCI), is dedicated to providing a bioinformatics infrastructure for CIMACs as well as to build a biomarker database. The CIMACs work collaboratively with the CIDC to enable data standardization and the development of uniform analysis pipelines across studies within the Network.

Given the biological complexity of most immunotherapy strategies, data generated from cross-site clinical trials are often confounded by technical variations or artifacts. Objective quality control standards are indispensable for minimizing the variations due to differences in reference genomes, gene models, analytical algorithms, and processing pipelines. Harmonization of center-specific protocols and assay performance is necessary to establish standard operating procedures to overcome the variability of methods and data collection (1–5). In addition, assay harmonization is expected to facilitate objective interpretation and data comparison across different studies and multiple sites, thereby achieving a unified network for cross-trial comparisons and meta-analyses.

WES and RNA-seq data provide a wealth of information for understanding tumor immune responses in clinical studies (6–12). WES can provide a comprehensive characterization of tumor mutations, from which neoantigens, mutational burden, and clonality can be inferred (6–8). Accumulating evidence has suggested the usage of tumor mutation burden (TMB) and tumor neoantigen load as biomarkers for cancer immunotherapy response (7,13–16). Likewise, RNA-seq provides a powerful tool to define response-driving factors from the tumor microenvironment (TME) (9,10). Advanced computational methods are now making it possible to utilize RNA-seq data to estimate the composition of the tumor immune infiltrates (17–21) and infer infiltrating immune B and T cell receptor

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

repertoires (22,23). These immunologic characterizations have yielded valuable insights, with the potential to guide immunotherapy (9,11,12,17,22,24,25). Formalin fixation of tissue sample remains the standard protocol for tissue preservation in the clinical arena (26,27). Successful use of formalin-fixed paraffin-embedded (FFPE) derived material in next generation sequencing (NGS) applications has been reported (28–30). However, data evaluating whether sequencing data generated from FFPE can be used to robustly estimate immunologic characteristics, such as immune gene expression, neoantigens, HLA typing, immune infiltration, and immune repertoires, is lacking. In addition, many existing studies to date have not consistently used matched normal samples as germline comparison and therefore somatic mutation detection could not be rigorously evaluated (26,27,31–33).

In this study, the CIMACs and CIDC performed a cross-site harmonization of WES and RNA-seq data generated from three centers (A, B, and C). They include, but not necessarily in this order, the MD Anderson Cancer Center (MDACC), the Broad Institute of Harvard and MIT, and the Molecular Characterization (MoCha) Lab at the Frederick National Laboratory for Cancer Research. Here, we describe the CIMAC-CIDC harmonization strategy for evaluating DNA and RNA sequencing data generated across distinct platforms and tissue preparation methods. Moreover, we discuss the key metrics needed for successful harmonization within and among the three sites.

Materials and Methods

Sample preparation and sequencing

Two mixed HapMap cell-line pools with well-characterized mutational profiles were used as truth data for the evaluation of WES (Supplementary Table S1) (34). Matched formalin-fixed paraffin-embedded (FFPE) tumor, fresh frozen (FF) tumor, and peripheral blood mononuclear cells (PBMC) from eight patients with NSCLC of squamous cell carcinoma histology were also studied; the tumors were collected between the years 2012 and 2015. Ethical approval for this study was obtained under a lab protocol (ProtocolLAB90–020) and was reviewed by The University of Texas MD Anderson Cancer Center Institutional Review Board. All samples used in this study and paper were obtained from patients consented under an IRB-approved informed consent. For the tumor specimens, percent tumor content, quantity and quality of DNA and RNA were assessed at the originating center for sample preparation before distribution to all three centers for WES and RNA-seq (Supplementary Tables S2 and S3). All samples were sequenced to at least 200X mean target coverage (Supplementary Table S4) for WES and at a minimum depth of 50M paired-end fragments for RNA-seq following each center's quality control criteria (Supplementary Table S5). As an alternative to RNA-seq, RNA from FFPE samples were profiled by NanoString using the nCounter® PanCancer Immune Profiling Panel (35). RNA quantity (ng) was determined with the qubit fluorometer while the quality was determined with the TapeStation by measuring the DV200 (Supplementary Table S6). Center C distributed the extracted RNA from the macro-dissected and non-macro-dissected samples to each site. All NanoString data were analyzed with NanoString's nSolver (v4.0).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Centralized data processing by the CIMAC-CIDC bioinformatics pipeline

After sequencing, raw data were transferred to the CIDC for centralized analyses using CIDC common pipelines (Supplementary Fig. S1 and S2). Reference files for human hg38 (GRCh38.d1.vd1) were obtained from the NCI Genomic Data Commons (GDC). For the WES analysis, the CIMAC-CIDC platform incorporated the Sentieon (2018.08.05) workflow for read alignment and variant calling, in which read alignment was performed with BWA (36) (0.7.15-r1140). Aligned and recalibrated BAM files were subjected to somatic mutation calling using Sentieon's TNsnv algorithm. Low-quality mutations were filtered by VCFtools (37) (0.1.16), and remaining somatic mutations were annotated by VEP (38) (v91). For RNA-seq analysis, read alignment was performed with STAR (39) (v.2.4.2a). RNA-seq quality control (QC) examination was performed on the aligned BAM files using RSeQC (40). Expression levels were quantified by SALMON (41) (v.0.14.0). Batch effect removal was performed with Limma (42) (3.42.2). The immune cell repertoires were inferred from aligned BAM files using TRUST4 (22) (v0.1.2). Expression profiles were subjected to immune infiltration estimation using Immunedeconv (43), which integrates six state-of-the-art estimation algorithms, including TIMER (17), xCell (18), MCPCounter (19), CIBERSORT (20), EPIC (21), and quanTIseq (44). Patient HLA types were estimated from both RNA-seq and WES using Optitype (45) (1.3.2). Expression profiles, somatic mutations, and HLA types from WES were integrated for neoantigen prediction using pVAC-Seq (24) (4.0.10) with NetMHC (46) (v4.0), that leveraged information from both binding affinity and eluted ligand data (46).

Statistical analysis for assay and sample harmonization

ISA (index of specific agreement), defined as 2 * Jaccard / (1 + Jaccard), was used to measure mutation agreements. ISA between samples was used as it has the potential to address downward bias in platform agreement on mutation detection when the true mutations are not prespecified (47). RNA-seq harmonization was assessed by the correlation level between replicates, sample tissue type (FF, FFPE), and sequencing centers. Spearman correlation was used to measure the agreements, since it is a more robust measure in settings when the data deviate from a Gaussian distribution and is less influenced by outliers (47). Hierarchical clustering was performed on the ISA or Spearman correlation coefficient derived distance matrix with the average linkage to measure sample similarities.

WES and RNA-seq harmonization baseline and concordance evaluation

The Cancer Genome Atlas (TCGA) lung cancer (NSCLC) cohort (519 adenocarcinomas) was retrieved, processed, and analyzed to establish reference data from which to assess the agreement of mutations from different callers (48). Mutation calls made by different TCGA-approved mutation callers (MuSE, MuTect2, SomaticSniper, VarScan2) on identical WES raw data were found to have an ISA concordance between 0.22 to 0.90 (mean = 0.71). It has been noted that although there is no uniform criterion of "acceptable" agreement, a correlation of greater than 0.7, 0.8, and 0.9 can be considered as having adequate, good, and excellent correlation, respectively [cite the co-submitted CIMAC-CIDC Network overview manuscript]. Of the published studies, depending on the sequencing depth, mutation allele frequency, mutation calling tools, and sample processing, there is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a large variation in the reported concordance level. The concordance levels between FF and FFPE have been reported to be around 70% in previous studies (1,49). Therefore, if the different WES platforms applied to the same DNA sample yielded mutation calls with similar or higher ISA, these sequencing platforms were considered reasonably harmonized. Mutation agreement assessment was performed on the overlapping exon regions to ensure that data generated by different capture kits across centers were comparable. Mutation concordance in cancer driver genes was evaluated, wherein 50 cancer driver genes from Ion AmpliSeq Cancer Hotspot Panel (v2) and the 310 lung cancer oncogenes from COSMIC database (50,51) were selected.

From the same TCGA NSCLC samples, we evaluated the pair-wise correlations among RNA-seq data to create a harmonization baseline. Since tumors of the same cancer type are expected to have similar gene expression levels, we set minimum acceptance criteria of the RNA-seq platform harmonization at 0.94, which is the top 95% Spearman correlation coefficient of the studied TCGA samples (Supplementary Fig. S3). If the analysis of the same RNA-seq data by different transcriptome platforms revealed the samples to have expression levels with similar or higher correlation than TCGA baseline, then these RNA-seq platforms were considered reasonably harmonized. Secondary analyses, including expression-based immune cell infiltration estimated by TIMER (17), xCell (18), MCPCounter (19), CIBERSORT (20), EPIC (21), and quanTIseq (44), immune repertoires estimated by TRUST4 (22), and HLA typing inferred by Optitype (45) were evaluated for their concordance. The Spearman correlation coefficient, the proportion of overlapped unique CDR3s, and the Jaccard index were used as concordance metrics for the immune cell infiltration estimates, immune cell repertoires, and HLA types, respectively.

Results

Central sample preparation and distributed sequencing

We generated data from two sample formats: (i) HapMap cell line pools (n=2); and (ii) non-small cell lung cancer (NSCLC) tumors with squamous cell carcinoma histology (n=8). DNA from two HapMap cell line pools (xx and yy), each consisting of a mixture of 10 well-characterized HapMap cell lines, was equally mixed at Center C (Fig. 1A) (34). In addition, DNA and RNA were centrally extracted from matched fresh frozen (FF) tumor and formalin-fixed paraffin-embedded (FFPE) tumor of eight NSCLC patients at Center B (Fig. 1B). For tumor samples, germline DNA was also extracted from matched peripheral blood mononuclear cells (PBMC) from the corresponding patients. Library preparation and sequencing were performed on two different days as technical replicates in all three centers (Fig. 1A and 1B). For both WES and RNA-seq, the capture kits used per sequencing center were distinct (Fig. 1C and 1D; Supplementary Table S7). For WES-seq, the overlap target regions between kits was increased if we focused on the exons (overlap region increased from 59.4% to 88.7%) (Fig. 1C).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CIMAC genomic platforms achieved high precision and recall in WES calling from the HapMap cell line pools

Utilizing the known mutations and allele fractions in the HapMap cell line pools, we evaluated key determinants of WES platform harmonization. Despite the inherent complexity of the assays and independent protocol development between sites, the WES data generated at different sites and replicates had comparable read coverage and variant allele frequency (VAF) (Fig. 2A). The sequencing data for the HapMap samples were highly concordant between technical replicates for all mutations, as well as for non-synonymous mutations only (ISA >0.874 and ISA >0.875, respectively) (Fig. 2B). These results suggested that potential technical bias and variation introduced during library preparation and sequencing within centers are acceptable.

We next examined the extent to which there was agreement in mutational burden among the three centers. Agreement assessments were performed on the overlap exon regions to ensure that data generated by different capture kits were comparable (Fig. 1C). Upon comparison of mutations called between center-specific data and ground truth data, we obtained an ISA of 0.827 and 0.817 for the xx and yy pools, respectively (Fig. 2C). Mutation agreement among the centers was further investigated as a function of coverage and VAF. At each coverage and VAF cutoff, agreement was evaluated based on the likelihood that a mutation would be detected in common by at least two centers. Overall, a higher level of concordance was observed with increased VAF and with greater in-depth coverage (Fig. 2D). Specifically, a VAF of 10% and 50X coverage cutoff yielded a 95% likelihood that a mutation would be called in common by at least two centers (Fig. 2D). The truth data provided by the HapMap pools gave us an opportunity to evaluate cross-site data variability and reproducibility. In evaluating the VAFs derived from the overlap target regions with common coverage greater than 50X, precision was greater than 0.98 and recall was greater than 0.87 at 10% VAF at all three centers (Fig. 2E). Altogether, the clustering results, the high precision and the high recall in WES called from the HapMap cell line pools lead us to conclude that the CIMAC genomic platforms and the CIDC analysis pipelines have been adequately optimized to ensure reliability and reproducibility in data generation.

Biological differences between tumors are much greater than platform / process-specific differences on WES

To validate the robustness of the center-specific reagents, protocols, data-transferring procedures, and range of acceptance criteria, we performed additional evaluations using the FF and FFPE NSCLC samples. Of note, deamination of nucleotides causes C:G>T:A changes in FFPE tissue samples and can produce false positives during next-generation sequencing (NGS) (52,53). DNA from matched FF and FFPE NSCLC tumors and PBMC in the corresponding patients were subjected to WES. Although generated at different centers using distinct protocols, the coverage, VAF calls, and non-synonymous mutation loads were comparable across replicates, centers, and sample preparations (FF vs. FFPE) (Fig 3A and 3B). Mutations called from FF and FFPE were comparable with an overlap rate of ~85% in the 50-gene panel (Materials and Methods) (Fig. 3C). Furthermore, nucleotide changes shared similar distributions between mutations derived from FF and FFPE tissues (Fig. 3D), suggesting that the deamination effect was not a dominant bias in mutation call from the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

FFPE specimens. Of note, the similar mutational signature patterns between FF and FFPE were consistent with observations previously detected in large cohorts of whole-genome sequencing data (1,54). These data together suggested that the mutations obtained from FFPE tissues collected in clinical settings are comparable to FF samples.

Across multiple studies conducted over the years, no single best strategy has been identified for somatic mutation calling from cancer specimens (55,56). Variant allele frequency, sequencing depth, and sequencing technique are multiple factors that determine whether a variant can be detected (55). We found that although the NSCLC specimens were sequenced and processed at different centers as technical replicates, the mutations clustered by patient with concordance levels above 0.67 among all samples (Fig. 3E, Supplementary Fig. S4). Of note, The ISAs we reported were based on the lowest ISA among samples generated in different replicates, centers, and sample preparation (FF vs. FFPE) for the same tumor. The vast majority of ISAs (96.1%) we obtained is greater than 0.7, with a median of 0.81, which outperformed previously reported concordances between FF and FFPE samples (1,49). To evaluate the key determinants of harmonization, we further evaluated agreement in mutation burden as a function of coverage and VAF. Overall, filtering by increasing VAF and coverage yielded fewer mutations and higher accuracy (Fig. 3F). For FF and FFPE, a cutoff at 10% VAF and 50X coverage resulted in a 93% and 87% likelihood for a mutation to be called in common by at least two centers (Fig. 3F). The high concordance level indicated to us that technical differences between replicates, centers, and sample preparation (FF vs. FFPE) were much smaller than the biological differences across tumors.

RNA-seq data generated from NSCLC are comparable between replicates, sample preparations, and centers

For the RNA-seq data generated across centers, concordance was assessed by correlating the gene expression among replicates, by sample preparation (FF, FFPE) and per sequencing center. Supplementary Fig. S5 shows the clustering result of the 150 samples based on Spearman correlation coefficient of log-transformed expression data. Multiple medTIN cutoffs were evaluated to determine the minimum cutoff at which the RNA samples could harmonize. Of note, medTIN score is a post-sequencing quality control metric to measure RNA integrity and RNA degradation (57). At a medTIN cutoff of > 50, the resultant 91 samples clustered by patient, with a minimum Spearman correlation above 0.94 among samples from the same tumor, thereby achieving a concordance level consistent with the pre-specified TCGA-based acceptance criteria (Materials and Methods; Fig. 4A). In addition, a cutoff of medTIN >30 was tested, and the resultant 134 samples achieved concordance levels above 0.90. Although the concordance level did not satisfy the pre-specified criterion (0.94, based on TCGA NSCLC data), samples still clustered by patient regardless of replicates, centers, or sample preparations (FFPE vs. FF) (Fig. 4B). Together, these data suggested that medTIN > 50 could be used as a post-sequencing quality control (QC) criterion to ensure all the samples cluster by tumors and to meet the pre-specified concordance cutoff, while medTIN > 30 could be used to ensure that all samples cluster by tumors.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To determine a set of criteria for the generation of reliable RNA-seq data, we further evaluated other QC metrics, including DV200 and exon mapping rate. DV200, a pre-sequencing quality metric, was highly associated with medTIN score (Spearman correlation = 0.63) (Fig. 4C). While the manufacturer has recommended that samples with DV200 > 30% usually yield better RNA-seq data quality, our data showed concordance amongst samples even at DV200 > 24% (Fig. 4B). Using DV200 > 24% as cutoff, we could rule out the samples that did not cluster by tumor ID (Fig. 4C). Exon Mapping Rate (EMR), another commonly used quality control metric to quantify the percentage of reads mapping to exon regions, was also associated with medTIN score ($p = 0.04$). However, we did not find EMR as a useful QC metric for ruling out outliers (Fig. 4D). These analyses together showed that DV200 of 24% is an effective pre-sequencing QC metric for the generation of RNA-seq data.

We next performed simulation studies to investigate whether the read number or gene number is a key determinant for successful harmonization. We down-sampled the data from FFPE samples of Center C from 113M paired-end reads to 50M. Using the expression profiles derived from the down-sampled reads, we could cluster the samples by tumor with concordance levels above 0.97 (Supplementary Fig. S6). These high correlations suggested to us that 50M paired end reads was an adequate read number to yield concordance. The effects of gene number on the harmonization were evaluated as well. The top 3,000 most variable genes were selected based on variance distribution in the log-transformed expression profile. When the 3,000 most variable genes were used for clustering, the lower-bound correlation level decreased to 0.88 (Supplementary Fig. S7). Despite the decreased concordance result, the samples still clustered by tumor ID. In addition, the clustering result was better than the baseline derived from TCGA NSCLC samples using the 3,000 most variable genes (0.88 vs. 0.85) (Supplementary Fig. S3).

QC metrics were evaluated to determine optimal cutoffs to generate acceptable secondary immunogenomic characteristics from RNA-seq data, including HLA typing, immune cell infiltration, and immune repertoire. Attempts were made to evaluate the sample data quality across different medTIN scores. All samples from the same tumor were inferred to have identical HLA type when a medTIN cutoff of 50 was used. In contrast, seven samples (8.3%) were noted to have off-target HLA typing when a medTIN cutoff of 30 was used. Agreement between FF and FFPE, measured by the Jaccard index, was similar whether a medTIN cutoff of 50 or 30 was used (1.0 and 0.99, respectively). Overall, FF and FFPE samples clustered by tumor in both medTIN 50 and 30 cutoff groups (Fig. 5A and 5B), and matched FF and FFPE samples per tumor shared similar immune infiltration patterns (Fig. 5C and 5D). The average Spearman correlations between the FF and FFPE samples were 0.88 and 0.87 for the medTIN 50 and 30 cutoff groups, respectively. When we examined the immune repertoires estimated from the RNA-seq data using TRUST4 (22), in which the inferred CDR3 clonotypes included *TCRA*, *TCRB*, *TCRD*, *IGH*, *IGK*, and *IGL*, the immune repertoires inferred from the matched FF and FFPE were highly concordant amongst samples from the same tumor regardless of medTIN cutoff (50 or 30; Fig. 5E and 5F). The immune repertoire clonality correlation between FF and FFPE was slightly higher in the medTIN 50 cutoff group, compared to the medTIN 30 group (Rho 0.58 vs. 0.55) (Fig. 5G and 5H). Overall, when the cutoff of medTIN was above 30, immune cell

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

infiltration and repertoires mostly clustered on a per-patient basis; HLA typing estimation clearly distinguished between tumors. Together, these results suggested that the quality of secondary immunogenomic characteristics were acceptable when inferred from RNA-seq data with medTIN above 30 (or, equivalently, DV200 > 24%).

The NanoString platform was evaluated for its potential to serve as an alternative approach for transcriptome profiling in cases of low-quality RNA samples. RNA extracted and processed from seven patients with NSCLC of squamous cell carcinoma histology were subjected to the NanoString PanCancer Immune Profiling Panel for transcriptomic quantification (35) at Centers B and C. DV200 cutoffs were evaluated by hierarchical clustering to determine the minimum cutoff at which the NanoString-generated data could harmonize between different sample processing (macro-dissected and non-macro-dissected) and centers. Overall, the majority of samples with DV200 below 24% failed to cluster by patient (15 of 20 failed, 75%) (Supplementary Fig. S8), a few samples with DV200 above 24% failed to cluster as well (3 of 44 failed, 6.81%). In summary, while NanoString gene expression data can be generated even from samples with very low DV200 that failed to produce RNA-seq libraries, our hierarchical clustering analysis indicates that the quality of such NanoString data originating from samples with very low DV200 may not be reliable.

Integrated DNA and RNA analyses revealed important immunogenomic features in NSCLC

Transcriptomics is a critical adjunct to genomics when interrogating patient tumors for actionable alterations (58). We therefore explored the potential of utilizing matched WES and RNA-seq to derive reliable cancer immunogenomic characteristics across centers and sample preparations (FF vs. FFPE). Analysis of the somatic mutations among the samples highlighted the consistent detection of multiple known recurrently mutated drivers of NSCLC across replicates, centers, and sample preparation (Fig. 6a, Materials and Methods). The majority (48 of 51, 94%) of mutated cancer driver genes were confirmed to have high expression levels (Fig. 6a, left panel). *TP53* was the most frequent mutated cancer driver gene (6 of 7 samples), consistent with the TCGA squamous cell lung carcinomas data (48). These results suggested that the CIMAC-CIDC analysis pipeline can reliably identify cancer driver mutations across replicates, centers, and preparations.

With the matched WES and RNA-seq data available, HLA typing derived from both assays were compared. The four-digit-level of accuracy for HLA typing was inferred using Optitype (45) on both assay data. Overall, HLA typing inferred from WES and RNA-seq was highly concordant for both FF and FFPE. The Jaccard index of HLA typing between the two platforms was 1.0 and 0.98 for FF and FFPE. HLA typing inferred from WES and RNA-seq clustered by tumor, suggesting the CIMAC-CIDC genomics platforms could generate reliable HLA typing regardless of sample preparations, sequencing centers, and sequencing platforms (Fig. 6B). The number of neoantigen calls, using the pVAC-Seq analysis pipeline (24), was performed on both FF and FFPE samples, and was highly associated with mutation burden in both FF (Rho = 0.71) and FFPE (Rho = 0.66) across centers.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Discussion

The CIMAC-CIDC network undertook an effort to establish harmonized platforms for the genomic analyses of clinical specimens from immunotherapy trials including WES and RNA-seq. This study provides a roadmap for how to harmonize diverse sequencing assays that may employ different chemistries and data analysis pipelines. The cross-center concordance evaluation assessed the factors that contributed to the discrepancies and those that facilitate sample harmonization. During the harmonization process, each participating center evaluated and confirmed the validity of center-specific reagents, standards, analytical methods, protocols, and data-reporting procedures throughout assay development and implementation. The discrepancies in somatic mutation calls and expression levels were found to be acceptable between replicates, sample preparation (FF vs. FFPE), and centers. Overall, this study demonstrated the feasibility of leveraging the resources available at different facilities to achieve high throughput at an acceptable level of consistency.

In this harmonization effort, CIMAC-CIDC rigorously evaluated sequencing data generated from multiple assays, including WES, RNA-seq, and NanoString. This study affirmed multiple key determinants to achieve sequencing-assay harmonization, including (1) use of rigorously validated assays at all centers, (2) focus on the overlap of capture regions, (3) use of a common data analysis pipeline, and (4) application of appropriate metrics for reporting the data, including a requirement for 50x coverage and 10% VAF for WES data, and medTIN > 30 and DV200 > 24% for RNA-seq. Studies have reported concordance level of somatic mutation calls generated from different sequencing centers and different pipelines (59,60). In our study, we leveraged the replicated sequencing data (3 centers × 2 replicates × 2 processing (FF and FFPE)) and the HapMap data to systematically evaluate the key determinants for harmonization. In addition to evaluating somatic mutation calls, we have investigated a set of criteria to harmonize the expression profile, immune infiltration, CDR3 immune repertoire, and neo-antigen calls. Together, these evaluation efforts will provide an analysis roadmap for our multi-site sequencing data harmonization.

One caveat in our study is the limited number of samples evaluated, namely two HapMap cell lines and eight NSCLC tumor samples. This study aimed to establish protocols for WES and RNA-seq library preparation and QC metrics to allow reliable and robust cross-site data generation. We have replicated the WES and RNA-seq in twelve aliquots (3 centers × 2 replicates × 2 processing (FF and FFPE)) for each of the eight NSCLC samples and each of the two mixed HapMap cell pools. The high replicate number allowed us to robustly identify factors introducing variability and to set up criteria to ensure comparable results across CIMAC sites while using a centralized analysis pipeline.

WES provides the opportunity to evaluate a spectrum of somatic alterations, whereas RNA-seq provides cell immunological phenotypes, including tumor immune infiltration, HLA typing, and immune repertoire. Here we report a harmonization effort carried out by CIMAC-CIDC to ensure reproducible WES and RNA-seq results both between centers and sample preparation (FF vs. FFPE) that meet the minimum pre-specified quality control criteria. The high level of concordance found supports interpretability of data sets across CIMACs and studies and will facilitate development of a database for secondary analyses.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

These efforts are particularly important and relevant in an era when evidence-based precision medicine is becoming more prevalent.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

Scientific and financial supports for the CIMAC-CIDC Network are provided through the National Cancer Institute (NCI) Cooperative Agreements U24CA224319 (to the Icahn School of Medicine at Mount Sinai CIMAC), U24CA224331 (to the Dana-Farber Cancer Institute CIMAC), U24CA224285 (to the University of Texas MD Anderson Cancer Center CIMAC), U24CA224309 (to the Stanford University CIMAC), and U24CA224316 (to the CIDC at Dana-Farber Cancer Institute). This work is also supported by grants from the Lung SPORE P50CA070907, NIH CCSG Award (CA016672), MD Anderson Cancer Center Support Grant (CA016672), National Cancer Institute under contract HHSN261200800001E. Additional support is made possible through the NCI CTIMS Contract HHSN261201600002C (to the Emmes Company, LLC).

Scientific and financial supports for the PACT projects are made possible through funding support provided to the FNIH by: AbbVie Inc., Amgen Inc., Boehringer-Ingelheim Pharma GmbH & Co. KG., Bristol-Myers Squibb, Celgene Corporation, Genentech Inc, Gilead, GlaxoSmithKline plc, Janssen Pharmaceutical Companies of Johnson & Johnson, Novartis Institutes for Biomedical Research, Pfizer Inc., and Sanofi.

We thank Anita Giobbie-Hurder, Mickey Williams, Rajesh Patidar, Chip Stewart, and Gad Getz for their valuable discussions and suggestions. We also acknowledge helpful suggestions from Helen Chen, Holden Maecker, and Sacha Gnjatic.

References

1. Robbe P, Popitsch N, Knight SJL, Antoniou P, Becq J, He M, et al.Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. Genet Med2018;20(10):1196–205 doi 10.1038/gim.2017.241. [PubMed: 29388947]
2. Pepin MG, Murray ML, Bailey S, Leistritz-Kessler D, Schwarze U, Byers PH. The challenge of comprehensive and consistent sequence variant interpretation between clinical laboratories. Genet Med2016;18(1):20–4 doi 10.1038/gim.2015.31. [PubMed: 25834947]
3. Vrijenhoek T, Kraaijeveld K, Elferink M, de Ligt J, Kranendonk E, Santen G, et al.Next-generation sequencing-based genome diagnostics across clinical genetics centers: implementation choices and their effects. Eur J Hum Genet2015;23(9):1142–50 doi 10.1038/ejhg.2014.279. [PubMed: 25626705]
4. Vail PJ, Morris B, van Kan A, Burdett BC, Moyes K, Theisen A, et al.Comparison of locus-specific databases for BRCA1 and BRCA2 variants reveals disparity in variant classification within and among databases. J Community Genet2015;6(4):351–9 doi 10.1007/s12687-015-0220-x. [PubMed: 25782689]
5. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, et al.Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. American journal of human genetics2016;98(6):1067–76 doi 10.1016/j.ajhg.2016.03.024. [PubMed: 27181684]
6. Melendez B, Van Campenhout C, Rorive S, Remmelink M, Salmon I, D'Haene N. Methods of measurement for tumor mutational burden in tumor tissue. Transl Lung Cancer Res2018;7(6):661–7 doi 10.21037/tlcr.2018.08.02. [PubMed: 30505710]
7. Wang Z, Duan J, Cai S, Han M, Dong H, Zhao J, et al.Assessment of Blood Tumor Mutational Burden as a Potential Biomarker for Immunotherapy in Patients With Non-Small Cell Lung Cancer With Use of a Next-Generation Sequencing Cancer Gene Panel. JAMA Oncol2019;5(5):696–702 doi 10.1001/jamaoncol.2018.7098. [PubMed: 30816954]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

8. Budczies J, Allgauer M, Litchfield K, Rempel E, Christopoulos P, Kazdal D, et al.Optimizing panel-based tumor mutational burden (TMB) measurement. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO2019;30(9):1496–506 doi 10.1093/annonc/mdz205.
9. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al.Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nature medicine2018;24(10):1550–8 doi 10.1038/s41591-018-0136-1.
10. Patel SP, Kurzrock R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. Mol Cancer Ther2015;14(4):847–56 doi 10.1158/1535-7163.Mct-14-0983. [PubMed: 25695955]
11. Pastor F, Berraondo P, Etxeberria I, Frederick J, Sahin U, Gilboa E, et al.An RNA toolbox for cancer immunotherapy. Nat Rev Drug Discov2018;17(10):751–67 doi 10.1038/nrd.2018.132. [PubMed: 30190565]
12. Sullenger BA, Nair S. From the RNA world to the clinic. Science (New York, NY)2016;352(6292):1417–20 doi 10.1126/science.aad8709.
13. High TMB Predicts Immunotherapy Benefit. Cancer Discov2018;8(6):668 doi 10.1158/2159-8290.Cd-nb2018-048. [PubMed: 29661758]
14. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al.Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science (New York, NY)2018;362(6411) doi 10.1126/science.aar3593.
15. Hellmann MD, Nathanson T, Rizvi H, Creelan BC, Sanchez-Vega F, Ahuja A, et al.Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. Cancer cell2018;33(5):843–52.e4 doi 10.1016/j.ccell.2018.03.018. [PubMed: 29657128]
16. Chan TA, Yarchoan M, Jaffee E, Swanton C, Quezada SA, Stenzinger A, et al.Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO2019;30(1):44–56 doi 10.1093/annonc/mdy495.
17. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al.TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res2017;77(21):e108–e10 doi 10.1158/0008-5472.CAN-17-0307. [PubMed: 29092952]
18. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol2017;18(1):220 doi 10.1186/s13059-017-1349-1. [PubMed: 29141660]
19. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al.Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol2016;17(1):218 doi 10.1186/s13059-016-1070-5. [PubMed: 27765066]
20. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al.Robust enumeration of cell subsets from tissue expression profiles. Nature methods2015;12(5):453–7 doi 10.1038/nmeth.3337. [PubMed: 25822800]
21. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife2017;6 doi 10.7554/eLife.26476.
22. Li B, Li T, Pignon J-C, Wang B, Wang J, Shukla SA, et al.Landscape of tumor-infiltrating T cell repertoire of human cancers. Nature genetics2016;48(7):725–32 doi 10.1038/ng.3581. [PubMed: 27240091]
23. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al.Antigen receptor repertoire profiling from RNA-seq data. Nature biotechnology2017;35(10):908–11.
24. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al.pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med2016;8(1):11 doi 10.1186/s13073-016-0264-5. [PubMed: 26825632]
25. Szolek AHLA Typing from Short-Read Sequencing Data with OptiType. Methods Mol Biol2018;1802:215–23 doi 10.1007/978-1-4939-8546-3_15. [PubMed: 29858812]
26. Munchel S, Hoang Y, Zhao Y, Cottrell J, Klotzle B, Godwin AK, et al.Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. Oncotarget2015;6(28):25943. [PubMed: 26305677]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

27. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al.Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic acids research2010;38(14):e151 doi 10.1093/nar/gkq510. [PubMed: 20525786]
28. Kerick M, Isau M, Timmermann B, Sültmann H, Herwig R, Krobitsch S, et al.Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. BMC Med Genomics2011;4:68 doi 10.1186/1755-8794-4-68. [PubMed: 21958464]
29. Adams MD, Veigl ML, Wang Z, Molyneux N, Sun S, Guda K, et al.Global mutational profiling of formalin-fixed human colon cancers from a pathology archive. Mod Pathol2012;25(12):1599–608 doi 10.1038/modpathol.2012.121. [PubMed: 22878650]
30. Zeng Z, Vo A, Li X, Shidfar A, Saldana P, Blanco L, et al.Somatic genetic aberrations in benign breast disease and the risk of subsequent breast cancer. npj Breast Cancer2020;6(1):1–11. [PubMed: 31934613]
31. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, et al.Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med2014;20(6):682–8 doi 10.1038/nm.3559. [PubMed: 24836576]
32. Astolfi A, Urbini M, Indio V, Nannini M, Genovese CG, Santini D, et al.Whole exome sequencing (WES) on formalin-fixed, paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST). BMC Genomics2015;16:892 doi 10.1186/s12864-015-1982-6. [PubMed: 26531060]
33. Oh E, Choi Y-L, Kwon MJ, Kim RN, Kim YJ, Song J-Y, et al.Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. PloS one2015;10(12):e0144162–e doi 10.1371/journal.pone.0144162. [PubMed: 26641479]
34. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al.Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol2013;31(11):1023–31 doi 10.1038/nbt.2696. [PubMed: 24142049]
35. Cesano AnCounter(®) PanCancer Immune Profiling Panel (NanoString Technologies, Inc., Seattle, WA). J Immunother Cancer2015;3:42 doi 10.1186/s40425-015-0088-7. [PubMed: 26674611]
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics2009;25(14):1754–60. [PubMed: 19451168]
37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al.The variant call format and VCFtools. Bioinformatics2011;27(15):2156–8. [PubMed: 21653522]
38. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al.The ensembl variant effect predictor. Genome biology2016;17(1):122. [PubMed: 27268795]
39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.STAR: ultrafast universal RNA-seq aligner. Bioinformatics2013;29(1):15–21. [PubMed: 23104886]
40. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics2012;28(16):2184–5. [PubMed: 22743226]
41. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature methods2017;14(4):417. [PubMed: 28263959]
42. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al.limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research2015;43(7):e47–e. [PubMed: 25605792]
43. Sturm G, Finotello F, List M. Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data. Methods Mol Biol2020;2120:223–32 doi 10.1007/978-1-0716-0327-7_16. [PubMed: 32124323]
44. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al.Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med2019;11(1):34 doi 10.1186/s13073-019-0638-6. [PubMed: 31126321]
45. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics (Oxford, England)2014;30(23):3310–6 doi 10.1093/bioinformatics/btu548.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

46. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. Nucleic acids research2008;36(Web Server issue):W509–12 doi 10.1093/nar/gkn202. [PubMed: 18463140]
47. Shih JH, Greer MD, Turkbey B. The problems with the kappa statistic as a metric of interobserver agreement on lesion detection using a third-reader approach when locations are not prespecified. Academic radiology2018;25(10):1325–32. [PubMed: 29551463]
48. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature2014;511(7511):543–50 doi 10.1038/nature13385. [PubMed: 25079552]
49. Hedegaard J, Thorsen K, Lund MK, Hein AM, Hamilton-Dutoit SJ, Vang S, et al.Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PloS one2014;9(5):e98187 doi 10.1371/journal.pone.0098187. [PubMed: 24878701]
50. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al.COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research2011;39(Database issue):D945–50 doi 10.1093/nar/gkq929. [PubMed: 20952405]
51. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer2018;18(11):696–705 doi 10.1038/s41568-018-0060-1. [PubMed: 30293088]
52. Kim S, Park C, Ji Y, Kim DG, Bae H, van Vrancken M, et al.Deamination Effects in Formalin-Fixed, Paraffin-Embedded Tissue Samples in the Era of Precision Medicine. The Journal of molecular diagnostics : JMD2017;19(1):137–46 doi 10.1016/j.jmoldx.2016.09.006. [PubMed: 27840062]
53. Prentice LM, Miller RR, Knaggs J, Mazloomian A, Aguirre Hernandez R, Franchini P, et al.Formalin fixation increases deamination mutation signature but should not lead to false positive mutations in clinical practice. PLoS One2018;13(4):e0196434 doi 10.1371/journal.pone.0196434. [PubMed: 29698444]
54. Nagahashi M, Shimada Y, Ichikawa H, Kameyama H, Takabe K, Okuda S, et al.Next generation sequencing-based gene panel tests for the management of solid tumors. Cancer Sci2019;110(1):6–15 doi 10.1111/cas.13837. [PubMed: 30338623]
55. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al.Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome medicine2013;5(10):91. [PubMed: 24112718]
56. Kim SY, Speed TP. Comparing somatic mutation-callers: beyond Venn diagrams. BMC bioinformatics2013;14(1):189. [PubMed: 23758877]
57. Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, et al.Measure transcript integrity using RNA-seq data. BMC bioinformatics2016;17(1):58. [PubMed: 26842848]
58. Adashek JJ, Kato S, Parulkar R, Szeto CW, Sanborn JZ, Vaske CJ, et al.Transcriptomic silencing as a potential mechanism of treatment resistance. JCI Insight2020;5(11).
59. Craig DW, Nasser S, Corbett R, Chan SK, Murray L, Legendre C, et al.A somatic reference standard for cancer genome sequencing. Scientific reports2016;6:24607 doi 10.1038/srep24607. [PubMed: 27094764]
60. Arora K, Shah M, Johnson M, Sanghvi R, Shelton J, Nagulapalli K, et al.Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. Scientific reports2019;9(1):19123 doi 10.1038/s41598-019-55636-3. [PubMed: 31836783]
61. Calogero RA, Carrara M, Beccuti M, Cordero F, Calogero MRA, Biobase D, et al.Package ‘chimera’. 2015.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Translational relevance statement

Given the biological complexity of immunotherapy strategies, data generated from cross-site clinical trials are often confounded by technical variations or artifacts. The Cancer Immune Monitoring and Analysis Centers (CIMACs) function to interface collaboratively with the Cancer Immunologic Data Commons (CIDC) to enable data standardization and the development of uniform analysis pipelines across clinical trials within the network.

This study details the CIMACs-CIDC's harmonization strategy for evaluating DNA and RNA sequencing data generated across distinct platforms and tissue preparation methods. This work also provides a roadmap for harmonizing diverse sequencing assays that employ different chemistries and data analysis pipelines. The key metrics for successful harmonization described herein are expected to further advance inter-laboratory data comparison and database development to facilitate integrative cross-cohort analysis. This work is anticipated to facilitate biomarker development across trial networks and organizations and ultimately address the critically important mission of improving the therapeutic management of cancer patients.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

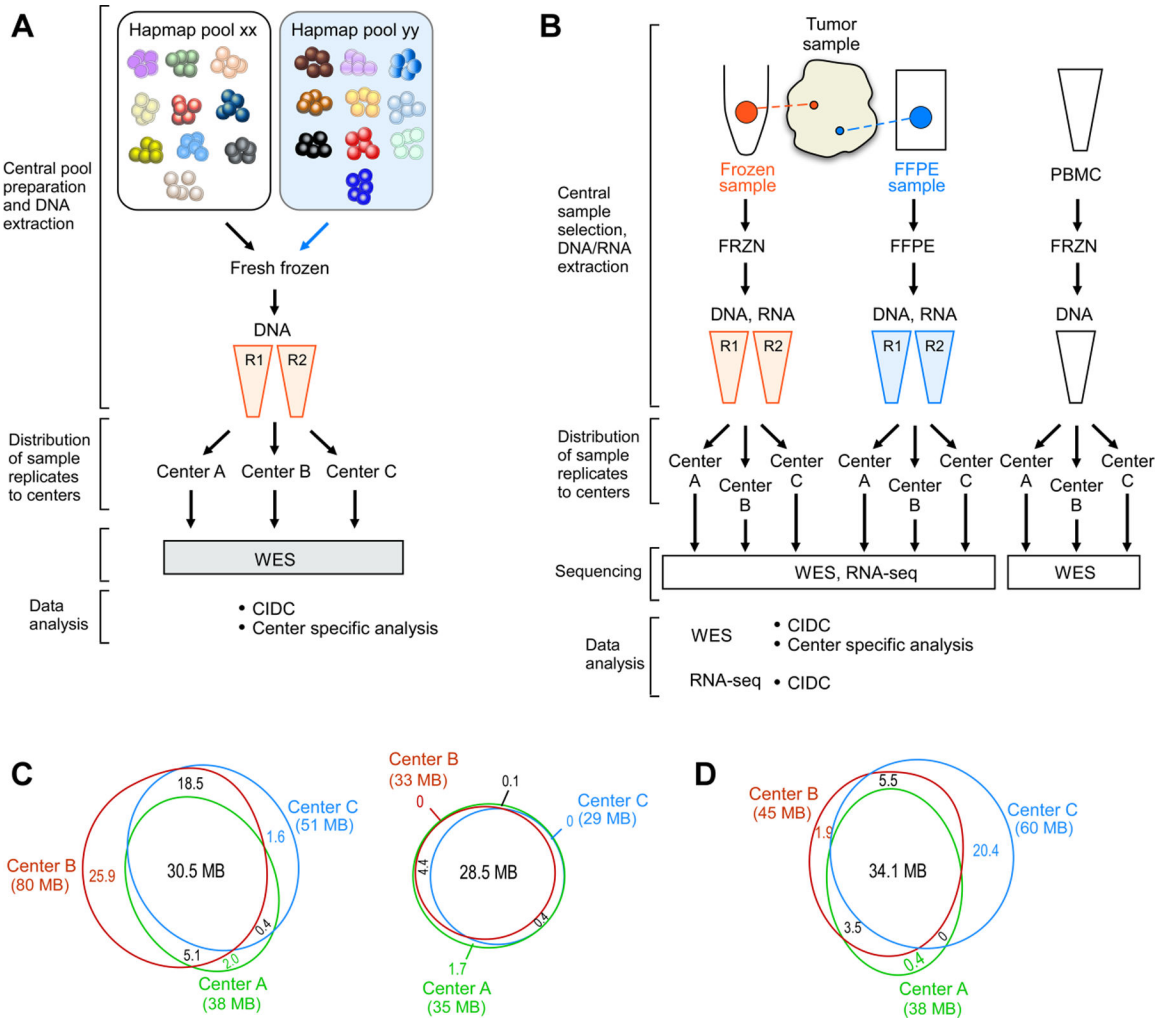

Figure 1. Illustration of study design and capture kits

(A) Two HapMap cell line pools were generated and used to provide 'ground truth' data. The HapMap cell lines xx and yy, each consisting of 10 individual HapMap cell lines mixed in equal proportions, were prepared and processed at Center C and distributed to all three centers for WES. Each HapMap pool was paired with a single cell line as germline control for mutation calls. The sequenced data was transferred to CIDC for centralized analyses. **(B)** Tumor samples from eight patients diagnosed with non-small-cell lung carcinoma (NSCLC) were selected. DNA and RNA extraction were performed by Center B from both fresh frozen (FF) and formalin-fixed paraffin-embedded (FFPE) processing. Germline DNA was also extracted from matched peripheral blood mononuclear cells (PBMC) from the corresponding patients. For all samples, two sets of aliquots were prepared as technical replicates. Extracted DNA and RNA were distributed to the three centers: Center A, B, and C for WES and RNA-seq. **(C)** Overlap of WES target regions between the three centers were evaluated. Left - Venn diagram of overall covered regions from the different centers. Right - Venn diagram of the overlap in exome regions. **(D)** Overlap regions were evaluated on the different RNA-seq capture kits used by the three centers.

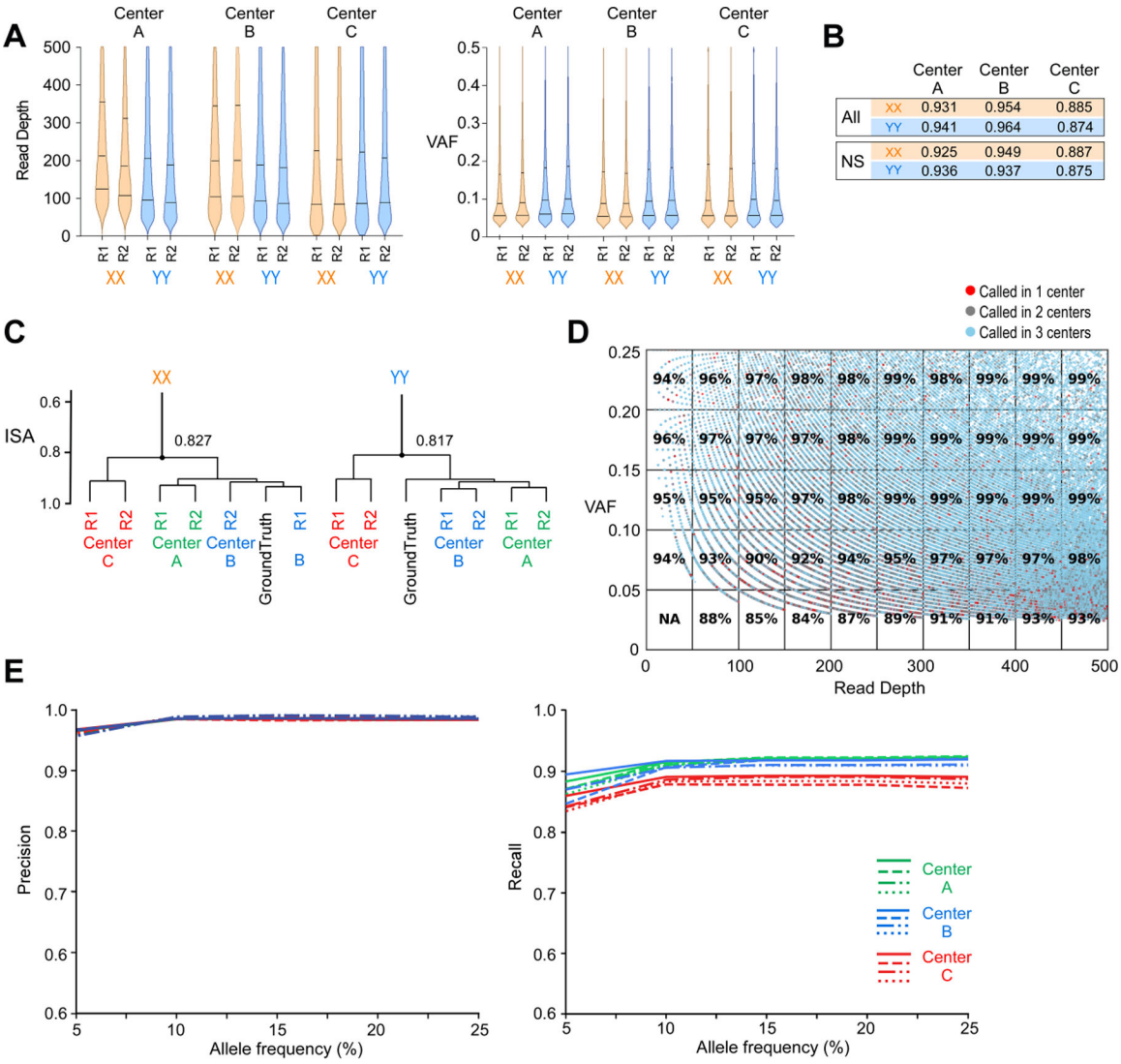

Figure 2. Evaluation and harmonization of somatic mutations identified in the HapMap cell-line pools

(A) The read depth and VAF of the somatic mutations detected in the HapMap pools xx and yy across the three centers. **(B)** Reproducibility between technical replicates in each center measured by ISA. The evaluations were performed using all mutations and non-synonymous mutations (NS). **(C)** Agreement assessment between centers and the truth data, based on mutation call agreement (ISA concordance score) on overlap of target exons. **(D)** Mutation agreement between the three centers was evaluated as a function of coverage and VAF. Red - mutations only identified in one center; Grey - mutations identified by two centers; Blue - mutations identified by all three centers. Numbers indicate the percentage of mutations called by at least two centers. **(E)** Precision and recall for mutation calling at different VAFs when evaluated against truth data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

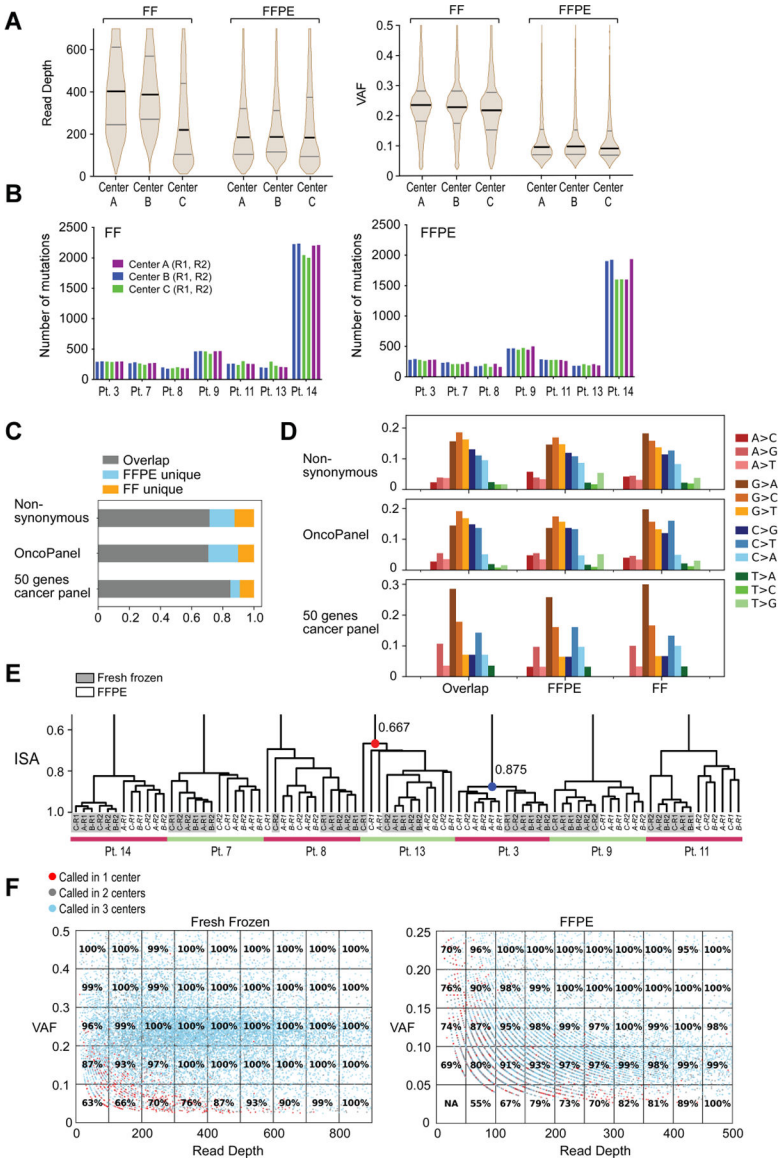

Figure 3. Evaluation and harmonization of somatic mutations identified in NSCLC samples

(A) The read depth and VAF of the somatic mutations generated from the three centers, separated by FF and FFPE samples. **(B)** The non-synonymous mutation loads of NSCLC samples across the three centers, separated by FF and FFPE samples. **(C)** Agreement between somatic mutations derived from FF and FFPE samples. The bars are the proportions of FF- FFPE-unique mutations, and their overlaps. **(D)** Distributions of the nucleotide changes for the FF and FFPE mutations, and their overlaps **(E)** Agreement assessment between mutations generated across replicates, centers, and sample processing (FF and FFPE). Clustering was performed upon the pair-wise mutation call agreements reflected by ISA scores. **(F)** Mutation agreement between the three centers was evaluated as a function of coverage and VAF. Red - mutations identified in one center; Grey - mutations identified in common by two centers; Blue - mutations identified by all three centers. Numbers indicate the percentage of mutations called by at least two centers at the corresponding cutoffs.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

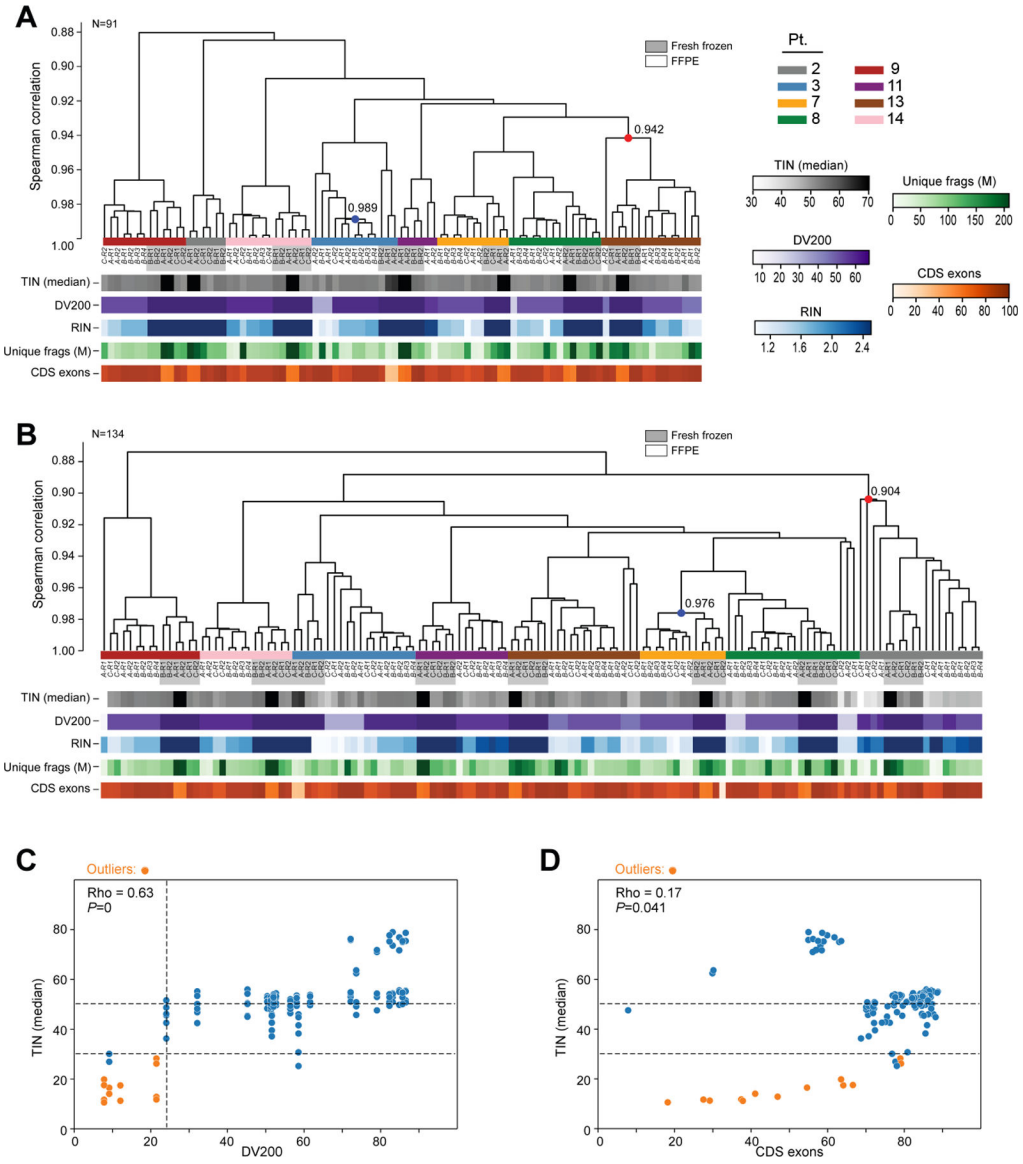

Figure 4. RNA-seq harmonization and quality control matrix evaluation

(A-B) Hierarchical clustering based on Spearman correlation coefficient of log2-TPM values for FF and FFPE samples with **(A)** medTIN > 50 (91 samples), along with QC metrics. **(B)** with medTIN > 30 (DV200 > 24%) (134 samples). **(C)** Scatterplot of medTIN and DV200 scores for the 150 sequenced samples. Outliers (yellow) are the samples that did not cluster by patients. (**D**) Scatterplot of medTIN scores and exome mapping rate for the 150 sequenced samples. Outliers are the samples that did not cluster by patient. Spearman correlation was performed to calculate the association between the two QC metrics in (C) and (D).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

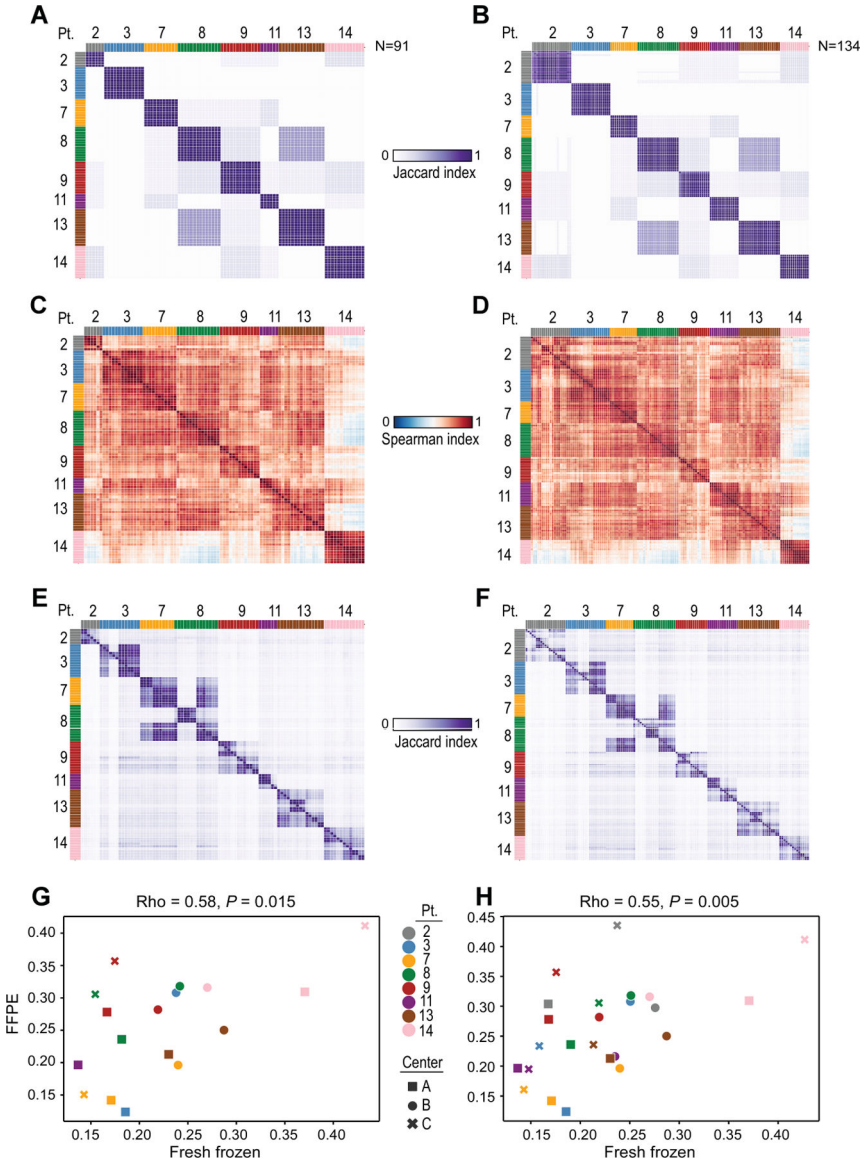

Figure 5. HLA typing, immune cell infiltration, and immune repertoire inferred from RNA-seq of FF and FFPE tumors robustly cluster together

(**A-B**) Clustering assessment between FF and FFPE samples using HLA typing inferred from RNA-seq using the tool Optitype (45). **(A)** medTIN >50 as a cutoff for sample selection, or (**B**) medTIN >30 as cutoff. (**C-D**) Clustering analysis between FF and FFPE using immune infiltration as features. The Infiltration was estimated by immunedeconv (43), which integrates six state-of-the-art estimation algorithms, including TIMER (17), xCell (18), MCPCounter (19), CIBERSORT (20), EPIC (21), and quanTIseq (44). As cutoffs, **(C)** medTIN >50 and (**D**) medTIN >30 were assessed. (**E-F**) Clustering results between FF and FFPE using immune repertoires as features. Immune repertoires were estimated by TRUST4, which is an updated version of the original TRUST(22) to infer CDR3 clonal types for *TCRA, TCRB, TCRD, IGH, IGK*, and *IGL* in tumor immune repertoires, using **(E)** medTIN >50 or (**F**) medTIN >30 as cutoffs. (**G**) Scatterplot of immune repertoire clonality inferred in FF and FFPE tumors (using Spearman correlations).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

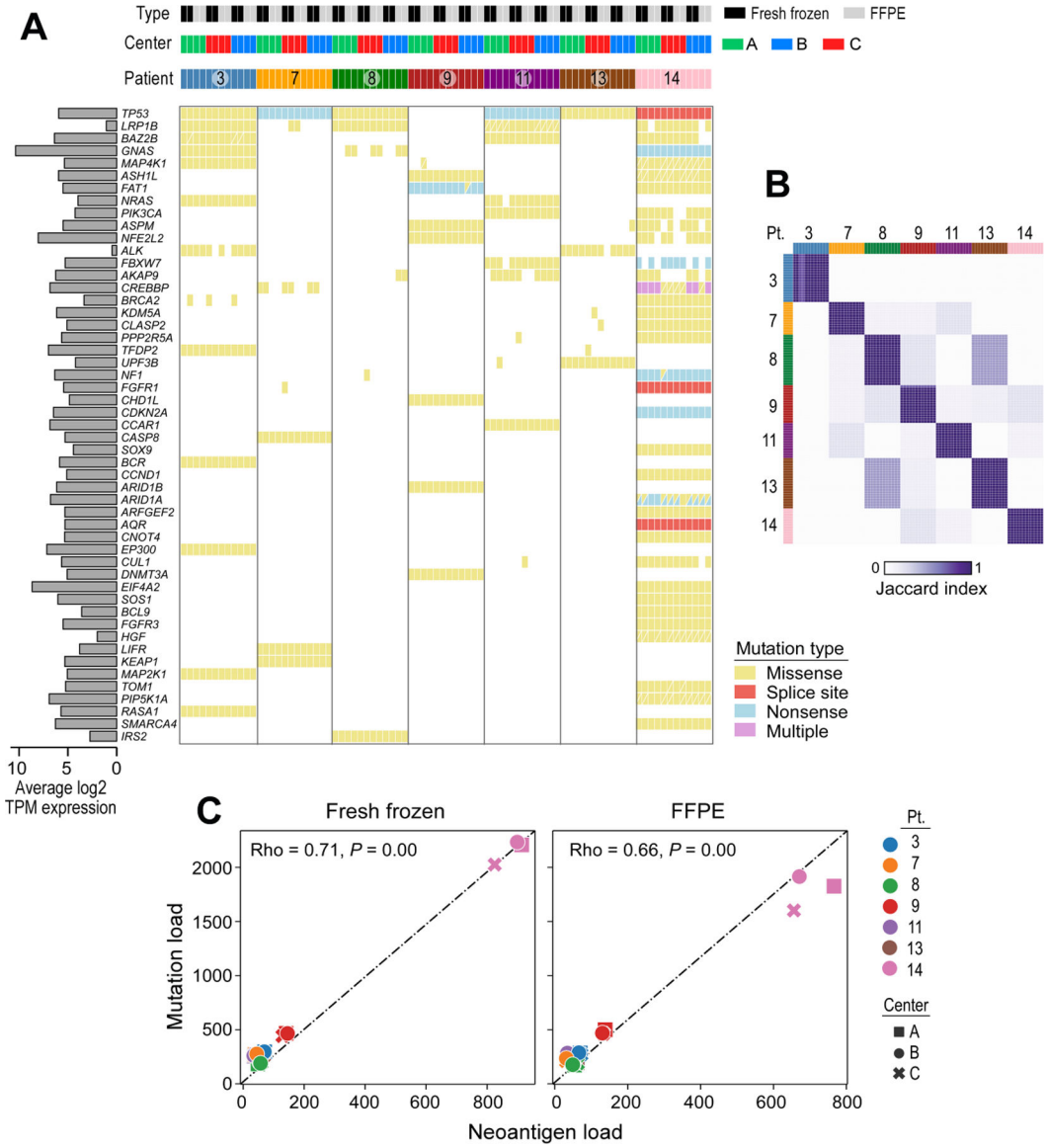

Figure 6: Integrated DNA and RNA analyses in FF and FFPE NSCLC tumors

(A) Co-mutation plot using WES and RNA-seq of the NSCLC tumors. The average log TPM expression is shown on the left panel. Mutations were called by the TnSnv algorithm from the Sentieon pipeline. Average log expression was calculated from SALMON counts. **(B)** HLA types were estimated for both WES and RNA-seq data using the tool Optitype (45). Jaccard index was calculated per patient using RNA-seq (Y-axis) and WES (X-axis) data. **(C)** Comparison of mutation and neoantigen load per patient specimen between FF and FFPE. Neoantigens were inferred by pVAC-Seq (24). Mutation load is the total number of non-synonymous mutations.