Journal of *Personalized Medicine*

MDPI

Article

Maximizing Small Biopsy Patient Samples: Unified RNA-Seq Platform Assessment of over 120,000 Patient Biopsies

P. Sean Walsh *, Yangyang Hao, Jie Ding, Jianghan Qu, Jonathan Wilde †, Ruochen Jiang, Richard T. Kloos, Jing Huang and Giulia C. Kennedy

Veracyte, Inc., 600 Shoreline Court, South San Francisco, CA 94080, USA
* Correspondence: sean@veracyte.com
† Current address: Glympse Bio, Inc., 35 Cambridgepark Drive, Cambridge, MA 02140, USA.

Abstract: Despite its wide-ranging benefits, whole-transcriptome or RNA exome profiling is challenging to implement in a clinical diagnostic setting. The Unified Assay is a comprehensive workflow wherein exome-enriched RNA-sequencing (RNA-Seq) assays are performed on clinical samples and analyzed by a series of advanced machine learning-based classifiers. Gene expression signatures and rare and/or novel genomic events, including fusions, mitochondrial variants, and loss of heterozygosity were assessed using RNA-Seq data generated from 120,313 clinical samples across three clinical indications (thyroid cancer, lung cancer, and interstitial lung disease). Since its implementation, the data derived from the Unified Assay have allowed significantly more patients to avoid unnecessary diagnostic surgery and have played an important role in guiding follow-up decisions regarding treatment. Collectively, data from the Unified Assay show the utility of RNA-Seq and RNA expression signatures in the clinical laboratory, and their importance to the future of precision medicine.

Keywords: RNA sequencing; lung cancer; interstitial lung disease; thyroid cancer; small biopsy; diagnostics; precision medicine

check for updates

Citation: Walsh, P.S.; Hao, Y.; Ding, J.; Qu, J.; Wilde, J.; Jiang, R.; Kloos, R.T.; Huang, J.; Kennedy, G.C. Maximizing Small Biopsy Patient Samples: Unified RNA-Seq Platform Assessment of over 120,000 Patient Biopsies. *J. Pers. Med.* **2023**, *13*, 24. https://doi.org/10.3390/jpm13010024

Academic Editor: Giorgio Malpeli

Received: 11 October 2022
Revised: 6 December 2022
Accepted: 20 December 2022
Published: 22 December 2022

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Large-scale efforts such as The Cancer Genome Atlas (TCGA) have contributed to the identification of genes and pathways that drive various disease states [1,2]. Many of these efforts have focused on using DNA to elucidate the role of genomic variants in disease progression and treatment [3,4].

The clinical utility of DNA markers in diagnosing disease and selecting treatment is undisputable; however, there are limitations to the information that can be derived from DNA alone. Notably, variants present in DNA may not be expressed in the tissue of interest. Available literature describing disease-associated DNA variants varies widely across disease types, and emerging evidence suggests that tissue-restricted, allele-specific, or imprinted expression of certain genes may modify the influence of genetic variants on disease manifestation [5–9]. In addition, the number of cancers for which the canon of known DNA mutations is sufficient to diagnose disease with reasonably high sensitivity and specificity is limited, leading to underdiagnosis. The use of pan-cancer DNA panels has been a valuable step forward in guiding treatment for patients harboring relevant mutations [10,11]; however, until early diagnosis is improved, many patients will be unable to benefit from subsequent treatment.

The diagnostic shortcoming of DNA variants has led to the widespread development and success of diagnostic algorithms, such as classifiers trained using machine learning on RNA-sequencing (RNA-Seq) derived features [12–17]. The potential benefits of applying RNA exome sequencing are wide-ranging. Unlike quantitative reverse transcription PCR (RT-qPCR) and microarrays, which are closed platforms for detecting and quantitating a predefined set of transcripts, use of an open system that interrogates the entire RNA

exome through sequencing allows an extensive repertoire of variation to be detected and quantitated without a priori knowledge of the transcripts of interest [18]. Furthermore, in addition to quantifying expression through gene counts, sequencing of RNA transcripts enables the detection and identification of both known and novel translocation (e.g., fusion) events. Single-nucleotide variants (SNVs), and insertions and deletions (indels) that are present in germline or somatic DNA, can also be detected by RNA-Seq if the gene is expressed in a particular specimen (i.e., expressed variants) [18–20]. Additionally, the mitochondrial genome can be interrogated by measuring quantitative expression of transcripts, as well as variant detection [21,22].

While whole-transcriptome or RNA exome profiling is often performed in research and discovery efforts, such a complex assay platform is more difficult to implement in a clinical diagnostic setting, owing to robustness, cost and/or throughput issues. Furthermore, addition of a separate DNA assay platform to an RNA workflow to derive variant information adds operational complexity and cost. As a result, the use of targeted, closed platforms often prevails.

In establishing the Unified Assay, we have implemented a comprehensive workflow wherein exome-enriched RNA-Seq assays are performed on patient specimens submitted to a Clinical Laboratory Improvement Amendments (CLIA)-certified laboratory for diagnostic testing (Figure 1). Although only the results from clinically validated classifiers, using a subset of all possible features, are delivered to the clinician in a patient report, the underlying RNA exome profile of each sample is stored in a data repository that can be interrogated further in a variety of ways (e.g., for drug target/biomarker profiling and/or development of new classifiers).

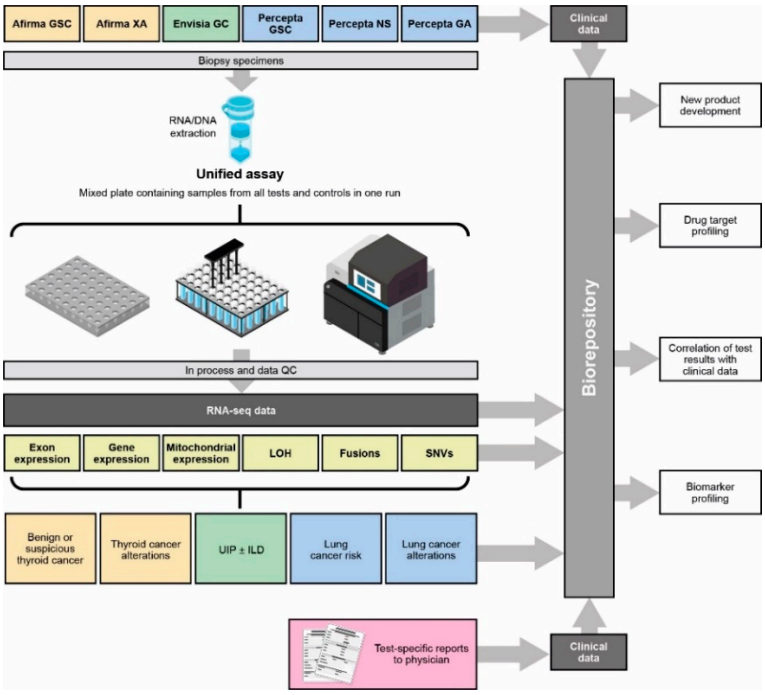


Figure 1. Unified Assay workflow. GA, genomic Atlas; GC, genomic classifier; GSC, Genomic Sequencing Classifier; ILD, interstitial lung disease; LOH, loss of heterozygosity; NS, nasal swab; SNV, single-nucleotide variant; UIP, usual interstitial pneumonia; XA, Expression Atlas.

Here, we show how the Unified Assay can be successfully deployed as a clinical application, as exemplified in the results from a large repository of small biopsy specimens processed via an automated and clinically validated RNA-Seq assay.

2. Materials and Methods

2.1. Machine Learning-Based Classifiers

Sequencing data were analyzed via the following advanced machine learning-based classifiers (all developed/licensed by Veracyte, Inc., South San Francisco, CA, USA) that are applied to the diagnosis of thyroid cancer, interstitial lung disease (ILD), and lung cancer, and may be applicable to other disease indications: the Afirma® Genomic Sequencing Classifier (GSC), which uses RNA exome sequencing information derived from a fine-needle aspiration (FNA) biopsy to classify cytologically indeterminate thyroid nodules as Afirma GSC Benign or Suspicious [12,22–25]; the Afirma Xpression Atlas (XA), which provides genomic alteration information from FNA samples with an Afirma GSC Suspicious result, or cytologically suspicious for malignancy or malignancy nodule [26]; the Envisia® Genomic Classifier (GC), a diagnostic test that uses a 190-gene expression signature to differentiate between usual interstitial pneumonia (UIP) and non-UIP subtypes in lung transbronchial biopsy (TBB) samples from patients with ILD [13]; the Percepta GSC, which uses RNA exome sequencing information from bronchial brushing samples for classification of lung cancer risk among bronchoscopy nondiagnostic samples [16]; the Percepta Nasal Swab (NS) classifier that uses RNA exome sequencing information from nasal (inferior turbinate) samples to classify the risk of lung cancer in patients with previously detected lung nodules [27]; and the Percepta Genomic Atlas (GA) test, which was developed to detect lung cancer-associated fusions and MET exon-skipping events in diagnostic TBB specimens and transbronchial needle aspirates (TBNAs) [28]. Beyond expression signatures, interrogation of RNA-Seq data in these samples revealed specimens with rare and novel genomic events, including gene fusions, mitochondrial variants, and loss of heterozygosity (LOH).

2.2. Sample Processing

Freshly collected biopsy samples were placed into sample tubes containing an RNA preservative (RNAprotect; Qiagen, Germantown, MD, USA) ahead of shipping to the laboratory at refrigerated temperatures (2–8 °C), where they were frozen prior to processing. Sample extraction methods were developed and optimized for each biopsy and tissue type, as described previously [12,13,16,24,25]. Briefly, RNA from thyroid nodule FNAs, lung nodule TBNAs, lung nodule TBBs, lung TBBs, and lung tissue were extracted with the AllPrep Micro Kit (Qiagen, Germantown, MD, USA). RNA was extracted from bronchial brushes of upper airway and nasal swabs using Qiazol and the miRNeasy Mini kit (Qiagen, Germantown, MD, USA). RNA was quantitated using the Promega Quantifluor kit (Promega, Madison, WI, USA) and a Tecan fluorometer (Tecan, San Jose, CA, USA). RNA quality was assessed by Bioanalyzer (Agilent, Santa Clara, CA, USA).

2.3. RNA-Seq

RNA was sequenced using the Illumina TruSeq RNA Access kit (Illumina, San Diego, CA, USA). The amount of RNA used for library preparation varied by sample type, i.e., 15 ng for thyroid nodule FNAs and lung TBB samples from ILD patients, and 50 ng for all other lung and nasal samples. Library preparation was performed on a custom, automated Hamilton STAR platform (Hamilton, Reno, NV, USA) and sequenced on the NextSeq 500/550 platform (Illumina). The scripts for automated liquid handling were optimized to accurately aspirate and dispense low reagent volumes (as low as 2 µL), thus enabling use of biopsy specimens yielding very low amounts of total RNA (15 ng). All total RNA samples were size fragmented (~200–300 bp) prior to cDNA synthesis and adaptor ligation. Sixteen PCR cycles were used to amplify all cDNA species (derived from all transcripts in the total RNA sample) using primers complementary to the generic adaptor sequences. The cDNA libraries were then enriched for exon content via two rounds of hybridization to probes designed to capture all human exons. Sequencing was performed using the NextSeq 500/550 platform via 2 × 75 bp paired-end protocols, to an average read depth of 25M paired-end reads per sample.

2.4. Bioinformatics Pipeline

RNA-Seq data were processed through the bioinformatics pipeline as described previously [12,14,22,29]. Briefly, binary base call (bcl) files were converted to fastq files using Illumina bcl2fastq and aligned to the human genome using STAR aligner [30]. Samples passing prespecified quality metrics were analyzed further. High-throughput sequencing was used to generate gene-level expression data, including mitochondrial chromosome genes, quantified by discrete counts [31], and the read count data normalized using DESeq2 [32] for stabilizing highly variable genes and variation in sequencing depth. SNV were detected using GATK V3.3 HaplotypeCaller, following best practice for RNA-Seq data [33]. Gene fusions were detected using STAR-Fusion [34].

2.5. Downstream Analysis

Expression-level data and variant- and fusion-calling results were passed to the Afirma GSC, Afirma XA, Envisia GC, Percepta GSC, Percepta NS, and Percepta GA, and scores and indication-specific predictions were generated by each classifier and biomarker detector. Data processing and visualization were performed with R [35]. Specifically, multiple R packages were used for visualization: ggplot2 [36], networkD3 [37], pheatmap [38], waterfalls [39], circlize [40], and treemapify [41].

2.6. Fusion Identification

Fusion calls were annotated by FusionAnnotator (part of the STAR-Fusion tool) and identified as having both splice sites matching a known reference splice site (ONLY_REF_SPLICE) or having one or both splice sites not matching a known reference splice site (INCL_NON_REF_SPLICE). The reference splice sites are reference exon junctions as provided by the reference transcript structure annotation GRCh37_gencode_v19. Fusions were further labeled as being either within chromosomes (intrachromosome) or between chromosomes (interchromosome).

2.7. Mitochondrial Gene Expression and Hürthle Classification

Transcripts from all 13 protein-coding mitochondrial genes were captured by the RNA-Seq assay; these were critical input features for the Hürthle Cell Index classifier, an integral component of the Afirma GSC [22].

2.8. LOH Score Determination

For each chromosome, the LOH score was calculated as the proportion of variants with variant allele frequency (VAF) < 0.2 or > 0.8 among heterozygous variants, excluding those in human leukocyte antigen (HLA) or immunoglobulin genes. Genome-level LOH scores were calculated similarly using variants across chromosomes 1 to 22.

2.9. Copy Number Variation (CNV) Identification

To identify CNVs, we employed jointly regulated blocks (JRBs) methodology, whereby increased and decreased JRB expression identified amplification and deletion events, respectively [42].

3. Results

3.1. Patient Samples

The Unified Assay was used to generate RNA-Seq data from deidentified clinical samples collected during the development and implementation of the aforementioned tests [12–14,16,22,24,25,27–29,43–45] and processed through the CLIA-certified laboratory. The RNA exome biorepository described herein is comprised of 120,312 clinical samples across 3 major clinical indications, where each clinical sample was processed and sequenced for 26,268 genes (Table 1).

Table 1. Clinical samples [27,45–49].

Indication	Development Samples, n	Classifier (CLIA) Samples, n
Thyroid (Afirma GSC)	634	109,912
ILD (Envisia GC)	359	3025
Lung (Percepta GSC)	311	5521
Lung (Percepta NS)	356	
Lung (Percepta GA)	194	

Development samples are those used in Research and Development to generate RNA-Seq data to train and develop the classifiers. CLIA samples are patient samples collected and tested once a test is commercially available. CLIA, Clinical Laboratory Improvement Amendments; FNA, fine-needle aspiration; GA, Genomic Atlas; GC, Genomic Classifier; GSC, Genomic Sequencing Classifier; ILD, interstitial lung disease; NS, nasal swab; TBB, transbronchial biopsy.

3.2. Use of an "Open" Platform and a Large Cohort Identifies Known and Novel Fusions, Including Rare and Unexpected Variant Combinations

Use of the Unified Assay enables analysis of the number and types of fusions both across and within chromosomes. Fusions containing known reference splice sites are of particular interest. As shown in Figure 2A, the highest rates of such fusions were observed in malignant lung cancer samples, with a mean of 1.04 intrachromosome and 0.55 interchromosome fusions per sample (Supplementary Figure S1). This rate far exceeded the number of fusions seen in Percepta GA benign samples and in Percepta GSC and Percepta NS samples. Across 159 malignant lung cancer samples, we identified 253 fusions comprising 232 individual fusion pairs, with most individual fusions detected in a single sample. However, several interesting fusions were identified in multiple samples, including *TAOK1-PIPOX* and *LPIN2-MYOM1*, which were each found in 3 samples and have been observed previously in breast and lung cancer TCGA datasets, respectively [50]. Known lung cancer-associated fusions, including *BRAF* and *RET*, were also identified.

ILD is not typically associated with the presence of gene fusions. However, the number of fusions identified per sample within the Envisia ILD set was slightly higher than that observed in the Afirma thyroid set. Accordingly, we explored whether these fusions are relevant to ILD or simply alterations present in normal populations. Within the CLIA patient sample set of 3025 Envisia samples, the most common fusion with a known reference splice site was *TFG-GPR128*, which was also common among Afirma CLIA samples (1.8% across both tests). *TFG-GPR128* has previously been described to have a prevalence of 2% in a healthy population [51], suggesting that this fusion is unlikely to be associated with ILD. Other, rarer fusions were also identified, including *GAS7-MYH1* and *ZCCHC8-RSRC2*, each of which were identified in 4 ILD samples and have been reported previously in cancer samples (*GAS7-MYH1* in lung and stomach adenocarcinoma, and *ZCCHC8-RSRC2* in multiple cancers) [52]. Interestingly, expression of *MYH1* has previously been found to be upregulated in the bleomycin mouse model of idiopathic pulmonary fibrosis (IPF) [53]. Overall, although fusions are rare in ILD, these data indicate the potential for revealing molecular information that may be relevant to disease manifestation.

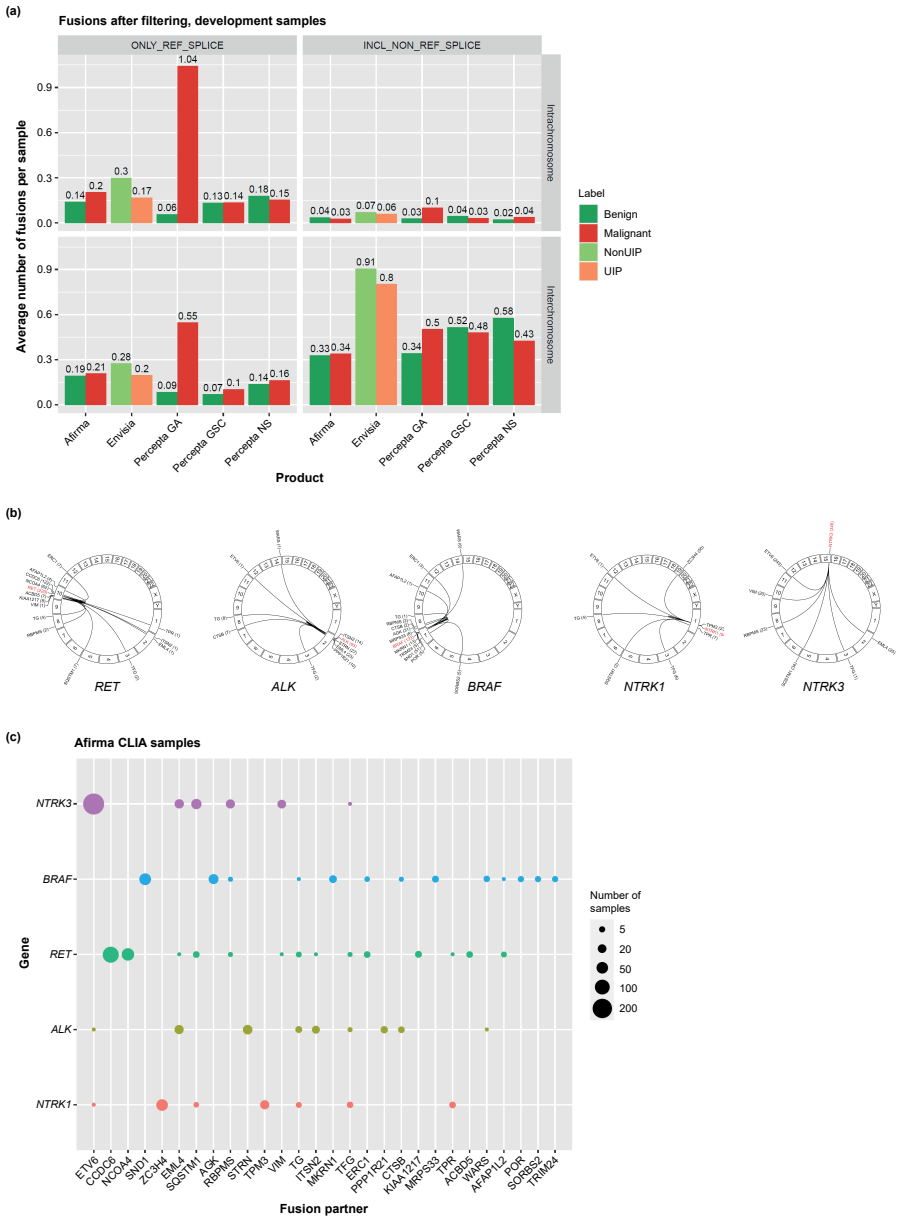


Figure 2. Fusion discovery across Unified Assay samples. (**a**) Number of fusions identified per sample across thyroid (Afirma), ILD (Envisia), and lung (Percepta) samples; (**b**,**c**) canonical fusions identified in thyroid Afirma CLIA samples. The following fusion calls were filtered out: having <3 junction reads; involving HLA, hemoglobin, or mitochondrial genes as one or both fusion partners; being between two immunoglobulin genes; and being called in >10% of samples from any one product with CLIA samples. ONLY_REF_SPLICE = having both splice sites matching a known reference splice site. INCL_NON_REF_SPLICE = having one or both splice sites not matching a known reference splice site. CLIA, Clinical Laboratory Improvement Amendments; GA, Genomic Atlas; GSC, Genomic Sequencing Classifier; HLA, human leukocyte antigen; NS, Nasal Swab; UIP, usual interstitial pneumonia; XA, Expression Atlas.

In total, more than 50,000 fusions (20,527 intrachromosomal and 29,743 interchromosomal fusions with known reference splice sites) containing 36,862 unique fusion pairs were identified in the Afirma CLIA sample set, comprising 109,912 FNAs. Although the number of fusions with known reference splice sites present in thyroid cancer samples was lower than observed for lung cancer samples, it is widely accepted that gene fusions have an important role in thyroid cancer, and drugs have been approved that target some of these (e.g., *RET* and *NTRK*) [54,55]. Evaluation of thyroid FNAs with the Unified Assay identified

all fusion proteins described for thyroid cancer in TCGA [50], the most common fusion pair being *PAX8-PPARg*, with other common fusion partners including *RET*, *NTRK1/3*, *ALK*, and *BRAF*. However, while the TCGA study was limited to papillary thyroid carcinoma (PTC) [1], the Unified Assay cohort included all thyroid nodule subtypes (including benign and malignant neoplasia) and was 200× larger than the TCGA study, thus enabling an unbiased analysis of all possible fusion partners for these kinases. Examination of the 109,912 FNAs revealed 353 samples with *NTRK3*, 105 with *NTRK1*, 112 with *ALK*, 274 with *RET*, and 303 with *BRAF* fusions. A subset of fusions with common fusion partners are shown in Figure 2B. *RET* and *BRAF* fusions were the most diverse and included a high level of inter- and intrachromosomal rearrangements, consistent with studies in other cancers [56,57]; however, examination of our larger cohort identified considerably more *BRAF* fusion partners. Along with the observed diversity, no one *BRAF* fusion partner dominated; *SND1/BRAF* and *AGK/BRAF* were the 2 most common fusions. *RET* fusions had the second largest diversity among fusions; however, unlike *BRAF* fusions, *RET* fusions had just 2 frequent partners: *CCDC6* (*RET/PTC1*) and *NCOA4* (*RET/PTC3*). Only 5 *NTRK3* fusion partners were observed in more than two samples, predominated by *ETV6/NTRK3* fusions. Across the different kinase fusions, common partners were observed, with *EML4*, *SQSTM1*, *ERC1*, *ETV6*, and *VIM* all being observed to be fused with multiple kinase partners. Each had a dominant kinase partner, but was also observed with secondary, and sometimes tertiary, fusion partners, indicating that some genes can participate in driver fusions across multiple kinase partners (Figure 2C). Full details of fusions identified using the Unified Assay are provided as supplementary material (Supplementary file S1).

3.3. Measuring Mitochondrial Gene Expression Identifies Challenging Subtypes

The capabilities of the Unified Assay extend beyond the measurement of canonical nuclear gene expression, as illustrated by assessment of mitochondrial gene expression across the multitude of patient specimens, revealing both expected and unexpected variations. As shown in Figure 3A, differences in mitochondrial gene expression were observed across different disease states, including lower expression in ILD TBBs compared with lung bronchoscopy brushing and biopsy samples. The application of mitochondrial content for diagnostic purposes is illustrated by the classification of thyroid tumors into the challenging subtype known as Hürthle cell neoplasms, which are characterized by an abundance of malfunctioning mitochondria [58]. We observed significantly higher expression of mitochondrial genes in thyroid FNA samples classified as Hürthle cell-positive based on the Afirma classifier (Figure 3B).

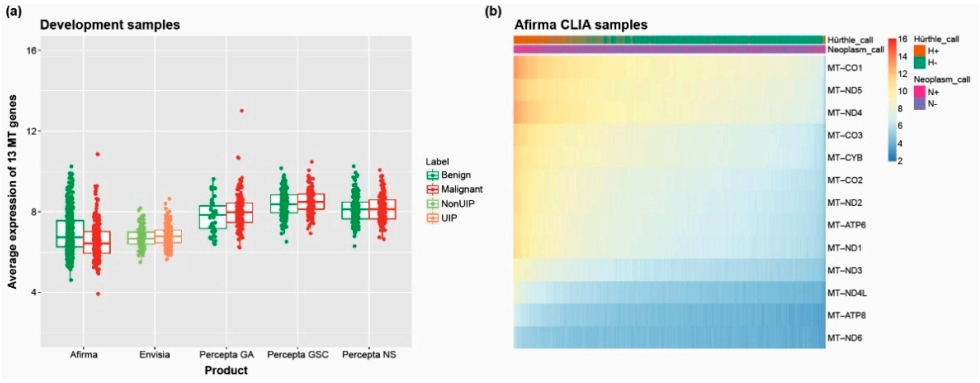


Figure 3. Mitochondrial expression across Unified Assay samples. (**a**) Expression of 13 mitochondrial genes across thyroid (Afirma), ILD (Envisia), and lung (Percepta) samples; (**b**) expression of mitochondrial genes in the subset of thyroid FNAs. Numbers next to the color bar represent the color scale of normalized gene expression, whereby blue indicates low expression values, yellow indicated intermediately expressed genes, and red represents highly expressed genes. CLIA, Clinical Laboratory Improvement Amendments; GA, Genomic Atlas; MT, mitochondrial; UIP, usual interstitial pneumonia.

3.4. Loss of Heterozygosity Can Be Measured by the Unified Assay

We exploited the ability to sequence exonic SNVs to identify blocks in expressed transcripts where observed variant allele frequencies were close to 0 or 1, enabling detection of copy number abnormality in the form of LOH. Examination of 194 lung cancer specimens indicated numerous instances of LOH, with the highest LOH scores observed in chromosomes 9 and 13, and the lowest in chromosomes 2 and 20 (Figure 4A,B). Interestingly, at least one lung cancer sample with high LOH could be identified for all 22 chromosomes. A range of LOH scores was observed across malignant lung cancer samples, across alteration (SNV/CNV)-positive/-negative samples, and within and across chromosomes (Figure 4B; see Supplementary Materials [Supplementary file S2] for LOH scores and SNV/CNV calls for these samples). However, in cases where more than one specimen was tested per patient, consistent LOH scores were observed across samples from the same patient. A high degree of LOH was observed across multiple chromosomes in samples without detected CNV or other common alterations, and LOH was detected in chromosomes beyond those identified by a targeted CNV assay. Finally, in addition to LOH scores, the Unified Assay detected specific CNVs within each chromosome (Figure 4C).

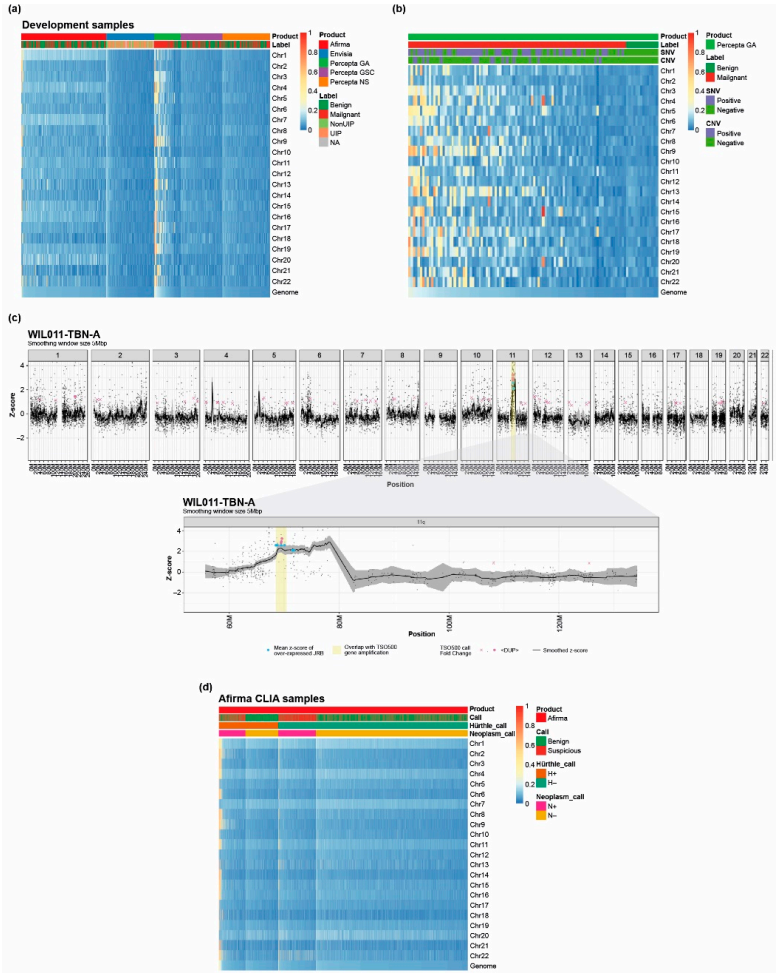


Figure 4. Chromosome and genome-wide LOH across diseases and tissues. (**a**) Chromosome and genome-wide LOH across lung (Percepta), thyroid (Afirma), and ILD (Envisia) samples; (**b**) chromosome and genome-wide LOH across specific lung cancer specimens; (**c**) comparison of results using TSO500 and expression-based CNV identification in an NSCLC-adenocarcinoma sample. The center of the highlighted region on chromosome 11 is gene CCND1, for which a duplication was identified by both TSO500 and expression-based CNV identification; (**d**) chromosome and genome-wide LOH across specific thyroid cancer specimens. Numbers next to the color bar represent the color scale of LOH

scores, whereby blue indicates low LOH scores, yellow indicates intermediate LOH scores, and red represents high LOH scores. Expression-based CNV calls were made by the JRB identification method [42], with increased expression JRBs identified as amplification events and decreased expression JRBs as deletion events. Chr, chromosome; CLIA, Clinical Laboratory Improvement Amendments; CNV, copy number variation; GA, Genomic Atlas; GSC, Genomic Sequencing Classifier; ILD, interstitial lung disease; JRB, Jointly Regulated Block; LOH, loss of heterozygosity; NA, not available; NSCLC, non-small cell lung cancer; NS, Nasal Swab; SNV, single-nucleotide variant; TBNA, transbronchial needle aspiration; TSO500, TruSight Oncology 500; UIP, usual interstitial pneumonia.

We also found LOH in Hürthle classifier-positive thyroid specimens, consistent with previous findings (Figure 4D) [22,59,60]. Examination of Afirma CLIA specimens identified 1063 samples with genomic LOH levels > 0.2, including 973 samples that were Hürthle-classifier positive. High LOH, including near haploidization, has been observed previously in hepatocellular carcinoma [22,59,60], and was also observed in our training set [24]. Patient samples with LOH events on chromosomes 5, 7, 12, and 20 were rare, whereas other chromosomes showed high levels of chromosome-wide LOH across multiple patients.

In contrast to lung and thyroid samples, high LOH was not observed across 359 ILD samples (Figure 4A); the same observation was also true for a larger sample size of ~3000 Envisia ILD samples (data not shown).

3.5. Scanning the Transcriptome of Clinical Samples for Drug Targets Reveals Potentially Useful Therapeutic Information

The immune response to cancer and other diseases is a key area of interest in both therapeutic and biomarker development. Therefore, we used the Unified Assay to investigate the expression of programmed death-ligand 1 (PD-L1) and other related genes. PD-L1 (CD274) was expressed in both benign and malignant thyroid and lung cancer samples, as well as within the nasal epithelium (Percepta NS), upper airway (Percepta GSC), and ILD samples (Envisia), with higher expression in benign versus malignant lung cancer samples and in non-UIP versus UIP ILD samples (Figure 5A). Expression of genes encoding other B7 family ligands and receptors, and additional key components of this signaling system, also varied across the development sample sets for each assay (Figure 5B). Expression of genes encoding two B7 family ligands, *B7-H4* (*VTCN1*) and *B7-H7* (*HHLA2*), showed a large degree of variation across different disease states. Expression of *HHLA2* was lower in thyroid cancer and lung cancer samples compared with other respiratory-associated samples, and higher in UIP versus non-UIP ILD samples (Figure 5C).

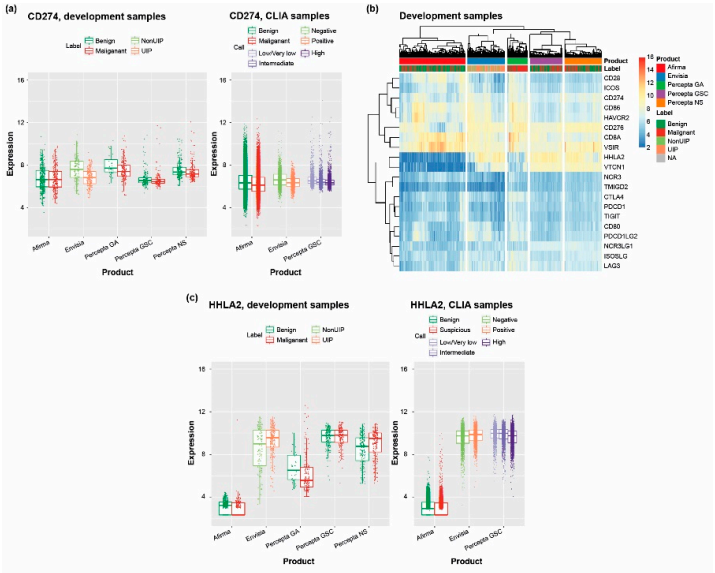


Figure 5. RNA and gene expression across samples used in the development of each test. (**a**) RNA expression of CD274 (PD-L1); (**b**) gene expression of potential drug targets (focusing on B7 ligands

and receptors), whereby blue indicates low expression values, yellow indicated intermediately expressed genes, and red represents highly expressed genes; (**c**) RNA expression of HHLA2. CLIA, Clinical Laboratory Improvement Amendments; GA, Genomic Atlas; GSC, Genomic Sequencing Classifier; HHLA2, HERV–H LTR-associating protein 2; NA, not applicable; NS, Nasal Swab; UIP, usual interstitial pneumonia; PD-L1, programmed death-ligand 1.

4. Discussion

4.1. Establishment of an Innovative Platform for Clinical Diagnostic Testing

The Unified Assay described herein provides a wealth of data that can be used to support accurate diagnoses and/or identify risk factors in challenging diseases, guide treatment decisions, and have the potential to enable further discoveries that can unlock future advances in patient care and the management of various diseases.

Although RNA-Seq has enormous clinical potential, this technique has yet to be widely deployed in clinical laboratories due to both assay and analysis complexity, including the variable nature of sample quality, assay reproducibility, and robustness of the analysis pipeline [18,61]. Notwithstanding these challenges, the data presented herein show the powerful utility of RNA-Seq in the clinical diagnostic laboratory.

Several clinical tests across different diseases and biopsy types have been run on the Unified Assay platform, with each test having been analytically and clinically validated, and shown to be robust to normal laboratory variation [12,13,16,24,25,44]. In addition, the data can be examined in a variety of ways. A summary of these features and other key properties of the Unified Assay platform is provided in Table 2.

Table 2. Key properties of the Unified Assay platform.

Properties	Benefits over Current Diagnostic Tools
Open platform	• No a priori knowledge about transcripts needed
Enrichment-based approach	• Greater read depth for the key coding regions • Ability to measure a wide range of genomic features beyond nuclear gene expression
Wider dynamic range than microarrays and RT-qPCR for detecting expression differences	• Detection and semi-quantitative measurement of known and novel variants in transcribed genes • Improved detection of biologically meaningful genomic variants • More accurate identification of patients most likely to benefit from targeted therapies
Use of the same analy tically validated clinical assay for both research and CLIA samples	• Eliminates the need for bridging of methods to targeted assays • Availability of data for quality control analysis • Application to all phases of product development • Lower cost and faster turnaround time
Detection and quantitation of both nuclear and mitochondrial transcripts	• Increased diagnostic information • More accurate identification of patients most likely to benefit from targeted therapies
Creation of complex RNA signatures for diagnosis as well as providing prognostic and/or predictive information using all transcripts as features in machine learning	• Potential for more accurate diagnosis • Means to avoid unnecessary diagnostic surgery

Table 2. *Cont.*

Properties	Benefits over Current Diagnostic Tools
Detection and quantitation of known and novel translocations/fusions	• Discovery of fusions that may be critical to tumorigenesis/disease etiology • Identification of potential therapeutic targets to slow or halt disease progression
Chromosome and genome-level LOH measurements, and identification of specific CNVs	• Can assess genomic instability • May facilitate development of novel anticancer drugs
Collection of full transcriptome data on every patient sample, creating a large biorepository for Pharma to mine	• Maximizes the data accessible from a small biopsy • Integration with clinical data facilitates identification of patients who may benefit from targeted treatments • Identification of rare events may reveal new drug targets • Allows large-scale biomarker profiling

CLIA, Clinical Laboratory Improvement Amendments; CNV, copy number variation; LOH, loss of heterozygosity; RT-qPCR, real-time quantitative polymerase chain reaction.

4.2. Use of an Enrichment-Based Approach to Facilitate Diagnosis and Treatment Decisions

The Unified Assay uses an enrichment-based approach whereby, following cDNA amplification, the resulting PCR products are hybridized to a cocktail of oligonucleotide probes designed to enrich for all exons of 26,268 genes. Further customization can be achieved by tiling extra enrichment probes over key exons and adding probes in non-coding regions (e.g., untranslated regions, introns, intergenic regions). Compared with total or nonenriched RNA-Seq, this exon-enrichment approach leads to approximately 5-fold lower sequencing costs and the ability to obtain significantly greater read depth for the key coding regions (exons), while retaining the ability to measure a wide range of genomic features (e.g., SNVs, fusions, mitochondrial gene expression, LOH), all from a single test, on a preciously small biopsy sample. Unlike amplicon-based RNA-Seq, which requires locus-specific PCR primer design [62], this enrichment approach can detect novel genomic events such as fusion partners not previously known to exist, and not anticipated by any bespoke assay design.

The breadth and richness of the data generated by the Unified Assay is exemplified by the Afirma GSC/XA, which uses multiple data features to facilitate diagnosis and treatment decisions for suspected thyroid cancer. For example, detection and quantitation of both nuclear and mitochondrial transcripts increases the information available for diagnostic purposes. Chronic lymphocytic thyroiditis (benign), Hürthle cell adenomas (benign), and Hürthle cell carcinomas (malignant) are observed among thyroid nodules and are characterized by increased abundance of mitochondria relative to non-Hürthle nodules [58]. Differentiating which samples require surgery has been a long-standing clinical challenge. Instead of relying on cytopathology or histopathology diagnoses, which are highly variable (and the latter of which requires surgical resection [63]), we used the presence of mitochondrial genes to develop a Hürthle-specific classifier that uses RNA-Seq expression and mitochondrial transcript data to routinely identify FNAs containing Hürthle cell features. These FNAs are then interrogated by an additional classifier that utilizes RNA-Seq expression and chromosomal level LOH to differentiate neoplastic from non-neoplastic samples [22]. Partnering this process with a 12-classifier expression-based "Ensemble Model", the Afirma GSC renders either a highly accurate result of "Benign", which facilitates clinical observation in lieu of diagnostic surgery, or renders a result of "Suspicious", where the yield of malignancy is significantly increased.

The Unified Assay has also been deployed for other diagnostic tests, including a classifier that can utilize small biopsies to distinguish the UIP pattern associated with IPF from the non-UIP patterns seen in other ILDs [13,29]. Analysis of whole-transcriptome

data from Envisia CLIA samples also enables investigation of other potential biomarkers that may be relevant to ILD diagnosis. While we found no evidence of changes in mitochondrial gene expression or LOH signal across the ILD samples, we did identify several non-canonical fusions whose association with ILD behavior could be explored further. In addition, ILD was found to share some similarities with lung cancer, such as variation in genes involved in cell proliferation and immune signaling, which is notable given the existence of shared risk factors for lung cancer and ILD (e.g., smoking, older age, male sex) [64]. Current lung cancer biomarker testing focuses on specific variants and fusions that are associated with existing therapeutics; however, RNA-Seq data can be used to add to and extend these investigations. Identification of novel fusion partners may help to reveal more patients with fusions involving currently targetable kinases such as *RET* and *ALK*, and, although the biological significance of many such fusions is yet to be determined, they could potentially become the therapeutic targets of tomorrow.

4.3. Identification of Disease-Related Variants by Comprehensive Testing

Robust functional testing of reagents and the use of pre-tested control reagents alongside maintained and validated equipment, within a full suite of standard operating procedures, minimizes technical batch effects and enables the combination and analysis of data sets that were run months, and even years, apart and most importantly, across both development and CLIA samples. Additionally, use of a whole-transcriptome assay yields copious data that can be used for quality control analysis to reduce technical noise and improve detection of biological signals. Quality-control methods include identifying gender through expressed genes and use of expressed single-nucleotide polymorphisms [12], not only to identify potential technical contamination events, but also to link samples from the same patient and confirm accuracy of associated clinical data, which is particularly useful in longitudinal studies.

Eliminating the need for bridging to targeted assays means all important features from development to commercial use could be preserved. For example, there is no limit to the number of genes that can be used in the Unified Assay, whereas targeted assays are limited, e.g., to tens to hundreds of genes. Each targeted assay type is also different and, when bridging to a new platform, not all genes or features will perform equivalently, thus necessitating further rounds of optimization.

Most importantly, however, targeted assays may not be able to generate data on all feature types. By contrast, the use of RNA-Seq enables detection of expressed, and therefore clinically relevant, variants and fusions that are more likely to be important to the etiology of the disease compared with non-expressed variants [8,12,65–67]. Allele-specific expression, or allelic imbalance, has been observed in tumors [8,9,12,65,67,68], where it gives rise to expressed RNA allele frequencies that are different from DNA-based allele frequencies. These include both variant-enriched expression (RNA VAF > DNA VAF) and wild-type–enriched expression (RNA VAF < DNA VAF). Indeed, we previously described detection of dramatically different VAFs when comparing DNA and RNA methods for some FNA samples in which a DNA variant was identified, but no variant was identified by whole-transcriptome RNA-Seq [12]. In these cases, the wild-type allele was predominantly expressed, accounting for the lack of variant detection using RNA methods. Treatment selections are often based on the interpretation of a DNA-based assay; however, if the variant is not expressed, then the treatment may not be efficacious [66]. Conversely, when the presence of a variant allele is contraindicative for therapy (e.g., *KRAS* variant positivity and anti-EGFR treatment in colorectal cancer [69]), variant-enriched expression may negatively impact treatment outcomes. Finally, the importance of expressed variants has been recognized in the identification of genetic disorders that may be caused by splicing alterations [6,70]. Such variants classified as variants of unknown significance based on DNA-seq may be resolved using RNA-Seq in some cases, thereby improving the diagnostic yield. Taken together, these observations suggest that RNA-Seq can play an important role in the identification of critical disease-related variants.

4.4. RNA-Seq–Based Identification of Potential Therapeutic Targets

Use of the Unified Assay provides new insights into both novel and rare fusions. Analysis of more than 100,000 FNA samples, including both benign and malignant thyroid nodules, identified 895 instances of *NTRK*, *RET*, *ALK* and *BRAF* fusions, representing potentially actionable kinase fusions. Interestingly, rare fusions included some involving *RET* and *NTRK3* that had only been described previously in single-case studies of patients exposed to radiation from the Fukushima nuclear disaster (*SQSTM1/NTRK3*; *AFAP1L2/RET*) [71] or atomic bomb fallout (*ACBD5/RET*) [72], and in a single case of salivary gland carcinoma (*VIM/RET*) [73].

Beyond detecting somatic variants that are potential oncogenic drivers, chromosome and genome-wide variants can be used to analyze LOH and CNV in tumors [22]. While less dense than LOH measurements using genomic DNA, genome-wide LOH using expressed genes is nonetheless a powerful tool for assessing genomic instability. Moreover, this measurement is obtained using the same small quantity of RNA that is used to generate count, mitochondrial, and fusion data, and therefore does not require extraction of a separate DNA/RNA sample. We noted potentially relevant patterns of LOH both across and within tumor (lung and thyroid) types. As well as detecting the same CCND1 CNV that was identified in lung cancer samples using a CNV-specific targeted assay, use of the Unified Assay revealed LOH in multiple chromosomes, thus providing a much more complete picture of the changes within these tumor samples.

Measurements of PD-L1 levels are currently used to identify patients most likely to benefit from certain immune checkpoint inhibitors [74], and the observed variable expression of genes encoding other B7 family members and inhibitory receptors in our patient samples highlights their potential as additional biomarkers and/or therapeutic targets. In this regard, HHLA2 (along with other B7 family members) has been proposed as a potential immunotherapy target in lung cancers [75]. Furthermore, while currently available immune checkpoint inhibitors are known to have immune-related side effects, including development of ILD [76], our findings of upregulation of HHLA2 expression in UIP ILD samples could indicate a potential therapeutic target of interest for IPF. With ongoing advances in the development of new immunotherapies, the need to understand the expression and function of such molecular markers in various diseases and tissues is paramount. Expression of genes encoding B7 family members B7-H1 (PD-L1, CD274), B7-DC (PD-L2, CD274, PDCD1LG2), B7-H2 (ICOSLG), and B7-H3 (CD276) in the airway and nasal epithelium has been demonstrated previously [77], which, our findings indicate, can be extended to other B7 costimulatory molecules B7-H4 (VTCN1), B7-H5 (VSIR), and B7-H6 (HHLA2). Moreover, while HHLA2 has previously been identified as being upregulated in IPF compared with normal lung tissues [78], we also show increased expression of HHLA2 within UIP versus non-UIP ILD subtypes. Notably, several other groups have proposed a role for B7 family members in inflammatory disorders, including allergic rhinitis, asthma, and autoimmune diseases such as arthritis [77,79,80], and a possible role for immune checkpoint inhibitors has been considered in IPF [81]. Therefore, our finding of variable expression of genes encoding additional B7 family members between non-UIP and UIP subtypes suggests further investigation into their role in ILD is warranted.

4.5. Realizing the Potential of the Unified Assay

No individual molecular feature described above is sufficient to describe the complexity of disease biology. For example, in the lung cancer data set, LOH scores were different between pathological variants and between CNV-positive and -negative samples. However, combining data on novel expression signatures with expressed fusion, variant, and LOH data may facilitate development of improved biomarker-based diagnostics without the need for additional biopsies or assays. Realizing this potential will require extensive analysis of these large data sets and demonstration of robust and reproducible performance at all stages of the pipeline.

Several key challenges have been identified that contribute to "analysis paralysis" in the development and clinical implementation of RNA-Seq platforms. These include the involvement of complex pipelines encompassing multiple tools that are independently developed, maintained, and licensed, as well as an overabundance of software tools and other options [18]. To address such challenges, carefully designed experiments were conducted to leverage samples with clear biological truth on the genomic alterations under investigation, including technical replicates at each experimental stage. This allowed us to further refine and optimize the pipeline to fit our needs and implement additional, novel normalization techniques to enhance the signal-to-noise ratio. It also enabled us to clearly quantitate all key performance metrics. We have leveraged modern tools such as Docker to modularize each key component to ensure that it is agile and robust, and carefully followed best practice for software development, with comprehensive documentation and thorough validation of each step of the pipeline, as well as the entity of the system. Our pipeline is also deployed on-premises and in-cloud to maximize computational efficiency.

An additional challenge associated with the implementation of RNA-Seq platforms is the lack of consensus by regulatory authorities regarding best practices and reference standards for validation [18]. Although RNA-Seq is now established as a research tool in many laboratories and institutions, these assays are not necessarily standardized, or their reproducibility assessed sufficiently to meet the requirements for a clinical assay. Often, the main aim is to control technical variability within a study to support the answering of specific biological questions within that study. The Unified Assay comprises multiple molecular tests, which were developed using FDA design control guidance [82] and run within the Veracyte CLIA-certified laboratory. Afirma GSC, Envisia, and Percepta GSC all have CLIA and New York State Department of Health approval. To demonstrate the technical and clinical performance of the launched assays, various analytical validation, clinical validation, and clinical utility studies have been performed and published in peer-reviewed journals [12,13,16,24]. Controls for each assay, documented and version-controlled procedures, and demonstration of training and competence of staff have also been implemented as part of CLIA requirements. Demonstration of clinical validity is not required by CLIA regulations but is required as part of the FDA's in vitro device approval process and, along with clinical utility data, is key to obtaining reimbursement.

5. Conclusions

The Unified Assay platform described herein shows the realization of the promise of RNA-Seq in the clinic. Use of RNA-Seq data from thyroid nodules has enabled a complex, multifactorial analysis to examine the molecular nature of individual clinical samples. Notably, of the 109,912 Afirma GSC samples, 75,428 patients received a molecular benign diagnosis, potentially allowing for the prevention of unnecessary surgery and, among the 34,484 patients with a molecular suspicious or high-risk cytopathology result, the use of RNA-Seq helped to identify several important driver events that may inform treatment decisions, including *BRAF* mutations, and fusions in *RET* and *NTRK*. The diagnosis of ILD and lung cancer in thousands of patients has similarly benefitted from the development of classifiers based on robust expression data.

Continued application of multi-modal data that can be accessed from RNA-Seq will help to maximize the information that can be obtained from individual patient biopsies and will ultimately support our scientific understanding of disease biology to expedite the development of new diagnostics and medicines. Indeed, the vast size of this biorepository and the capability of finding associated rare alterations in a large cohort has proven useful to identify patients who may benefit from targeted treatments, in the development of new tests, and in providing enhanced visualization and knowledge to patients, clinicians, and therapeutic partners. Ultimately, we endeavor to enable the biorepository as a platform for interactive research collaboration that will allow for the exploration of additional biological characteristics or genomic signatures beyond the fully developed and validated classifiers.

Collectively, these assays show the power and utility of RNA-Seq and RNA expression signatures in the clinical laboratory and their importance to the future of precision medicine.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jpm13010024/s1, Figure S1: Number of fusions identified per sample across benign and malignant lung samples; Supplementary file S1: All filtered fusions; Supplementary file S2: Percepta GA_LOH.

Author Contributions: Conceptualization, P.S.W., J.W., R.T.K. and G.C.K.; Formal analysis, P.S.W., Y.H., J.D., J.Q., J.W., R.J., R.T.K., J.H. and G.C.K.; Writing—original draft, J.W.; Writing—review and editing, P.S.W., Y.H., J.D., J.Q., R.J., R.T.K., J.H. and G.C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was conducted by Veracyte, Inc.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw (>100,000 exome level sequencing file) datasets generated and/or analyzed during the study are impractical to archive; however, the authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

Acknowledgments: Medical writing support was provided by Alligent Europe (a Division of Envision Pharma Inc.), and funded by Veracyte, Inc.

Conflicts of Interest: P.S.W., Y.H., J.D., J.Q., R.J., R.T.K, J.H. and G.C.K. are all employees of, and own stock options in, Veracyte, Inc. J.W. was an employee of Veracyte, Inc. at the time of manuscript preparation, and is currently an employee of Glympse Bio. Veracyte was involved in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

References

1. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **2014**, *159*, 676–690. [CrossRef] [PubMed]
2. Smedley, D.; Smith, K.R.; Martin, A.; Thomas, E.A.; McDonagh, E.M.; Cipriani, V.; Ellingford, J.M.; Arno, G.; Tucci, A.; Vandrovcova, J.; et al. 100,000 genomes pilot on rare-disease diagnosis in health care—Preliminary report. *N. Engl. J. Med.* **2021**, *385*, 1868–1880. [CrossRef] [PubMed]
3. Malone, E.R.; Oliva, M.; Sabatini, P.J.B.; Stockley, T.L.; Siu, L.L. Molecular profiling for precision cancer therapies. *Genome Med.* **2020**, *12*, 8. [CrossRef]
4. Sokolenko, A.P.; Imyanitov, E.N. Molecular diagnostics in clinical oncology. *Front. Mol. Biosci.* **2018**, *5*, 76. [CrossRef] [PubMed]
5. Correia, L.; Magno, R.; Xavier, J.M.; de Almeida, B.P.; Duarte, I.; Esteves, F.; Ghezzo, M.; Eldridge, M.; Sun, C.; Bosma, A.; et al. Allelic expression imbalance of *PIK3CA* mutations is frequent in breast cancer and prognostically significant. *NPJ Breast Cancer* **2022**, *8*, 71. [CrossRef] [PubMed]
6. Cummings, B.B.; Marshall, J.L.; Tukiainen, T.; Lek, M.; Donkervoort, S.; Foley, A.R.; Bolduc, V.; Waddell, L.B.; Sandaradura, S.A.; O'Grady, G.L.; et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **2017**, *9*, eaal5209. [CrossRef]
7. Goovaerts, T.; Steyaert, S.; Vandenbussche, C.A.; Galle, J.; Thas, O.; Van Criekinge, W.; De Meyer, T. A comprehensive overview of genomic imprinting in breast and its deregulation in cancer. *Nat. Commun.* **2018**, *9*, 4120. [CrossRef]
8. Grant, A.D.; Vail, P.; Padi, M.; Witkiewicz, A.K.; Knudsen, E.S. Interrogating mutant allele expression via customized reference genomes to define influential cancer mutations. *Sci. Rep.* **2019**, *9*, 12766. [CrossRef]
9. Liu, Z.; Dong, X.; Li, Y. A genome-wide study of allele-specific expression in colorectal cancer. *Front. Genet.* **2018**, *9*, 570. [CrossRef]
10. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93. [CrossRef]
11. Wang, H.Y.; Deng, L.; Li, Y.Q.; Zhang, X.; Long, Y.K.; Zhang, X.; Feng, Y.F.; He, Y.; Tang, T.; Yang, X.H.; et al. Pan-cancer analysis of tumor mutational burden and homologous recombination DNA damage repair by targeted next-generation sequencing. *Cancer Res. Treat.* **2021**, *53*, 973–982. [CrossRef] [PubMed]
12. Angell, T.E.; Wirth, L.J.; Cabanillas, M.E.; Shindo, M.L.; Cibas, E.S.; Babiarz, J.E.; Hao, Y.; Kim, S.Y.; Walsh, P.S.; Huang, J.; et al. Analytical and clinical validation of expressed variants and fusions from the whole transcriptome of thyroid FNA samples. *Front. Endocrinol.* **2019**, *10*, 612. [CrossRef] [PubMed]

13. Choi, Y.; Lu, J.; Hu, Z.; Pankratz, D.G.; Jiang, H.; Cao, M.; Marchisano, C.; Huiras, J.; Fedorowicz, G.; Wong, M.G.; et al. Analytical performance of Envisia: A genomic classifier for usual interstitial pneumonia. *BMC Pulm. Med.* **2017**, *17*, 141. [CrossRef] [PubMed]
14. Choi, Y.; Qu, J.; Wu, S.; Hao, Y.; Zhang, J.; Ning, J.; Yang, X.; Lofaro, L.; Pankratz, D.G.; Babiarz, J.; et al. Improving lung cancer risk stratification leveraging whole transcriptome RNA sequencing and machine learning across multiple cohorts. *BMC Med. Genomics* **2020**, *13*, 151. [CrossRef]
15. Grewal, J.K.; Tessier-Cloutier, B.; Jones, M.; Gakkhar, S.; Ma, Y.; Moore, R.; Mungall, A.J.; Zhao, Y.; Taylor, M.D.; Gelmon, K.; et al. Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw Open* **2019**, *2*, e192597. [CrossRef]
16. Johnson, M.K.; Wu, S.; Pankratz, D.G.; Fedorowicz, G.; Anderson, J.; Ding, J.; Wong, M.; Cao, M.; Babiarz, J.; Lofaro, L.; et al. Analytical validation of the Percepta genomic sequencing classifier; an RNA next generation sequencing assay for the assessment of lung cancer risk of suspicious pulmonary nodules. *BMC Cancer* **2021**, *21*, 400. [CrossRef]
17. Walter, W.; Shahswar, R.; Stengel, A.; Meggendorfer, M.; Kern, W.; Haferlach, T.; Haferlach, C. Clinical application of whole transcriptome sequencing for the classification of patients with acute lymphoblastic leukemia. *BMC Cancer* **2021**, *21*, 886. [CrossRef]
18. Byron, S.A.; Van Keuren-Jensen, K.R.; Engelthaler, D.M.; Carpten, J.D.; Craig, D.W. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nat. Rev. Genet.* **2016**, *17*, 257–271. [CrossRef]
19. Piskol, R.; Ramaswami, G.; Li, J.B. Reliable identification of genomic variants from RNA-Seq data. *Am. J. Hum. Genet.* **2013**, *93*, 641–651. [CrossRef]
20. Prodduturi, N.; Bhagwate, A.; Kocher, J.A.; Sun, Z. Indel sensitive and comprehensive variant/mutation detection from RNA sequencing data for precision medicine. *BMC Med. Genom.* **2018**, *11*, 67. [CrossRef]
21. Yuan, Y.; Ju, Y.S.; Kim, Y.; Li, J.; Wang, Y.; Yoon, C.J.; Yang, Y.; Martincorena, I.; Creighton, C.J.; Weinstein, J.N.; et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **2020**, *52*, 342–352. [CrossRef] [PubMed]
22. Hao, Y.; Duh, Q.Y.; Kloos, R.T.; Babiarz, J.; Harrell, R.M.; Traweek, S.T.; Kim, S.Y.; Fedorowicz, G.; Walsh, P.S.; Sadow, P.M.; et al. Identification of Hürthle cell cancers: Solving a clinical challenge with genomic sequencing and a trio of machine learning algorithms. *BMC Syst. Biol.* **2019**, *13*, 27. [CrossRef] [PubMed]
23. Alexander, E.K.; Schorr, M.; Klopper, J.; Kim, C.; Sipos, J.; Nabhan, F.; Parker, C.; Steward, D.L.; Mandel, S.J.; Haugen, B.R. Multicenter clinical experience with the Afirma gene expression classifier. *J. Clin. Endocrinol. Metab.* **2014**, *99*, 119–125. [CrossRef] [PubMed]
24. Hao, Y.; Choi, Y.; Babiarz, J.E.; Kloos, R.T.; Kennedy, G.C.; Huang, J.; Walsh, P.S. Analytical verification performance of afirma genomic sequencing classifier in the diagnosis of cytologically indeterminate thyroid nodules. *Front. Endocrinol.* **2019**, *10*, 438. [CrossRef] [PubMed]
25. Patel, K.N.; Angell, T.E.; Babiarz, J.; Barth, N.M.; Blevins, T.; Duh, Q.Y.; Ghossein, R.A.; Harrell, R.M.; Huang, J.; Kennedy, G.C.; et al. Performance of a genomic sequencing classifier for the preoperative diagnosis of cytologically indeterminate thyroid nodules. *JAMA Surg.* **2018**, *153*, 817–824. [CrossRef]
26. Hu, M.I.; Waguespack, S.G.; Dosiou, C.; Ladenson, P.W.; Livhits, M.J.; Wirth, L.J.; Sadow, P.M.; Krane, J.F.; Stack, B.C.; Zafereo, M.E.; et al. Afirma genomic sequencing classifier and Xpression Atlas molecular findings in consecutive Bethesda III-VI thyroid nodules. *J. Clin. Endocrinol. Metab.* **2021**, *106*, 2198–2207. [CrossRef]
27. Lamb, C.; Rieger-Christ, K.; Reddy, C.; Ding, J.; Qu, J.; Wu, S.; Johnson, M.; Whitney, D.; Walsh, P.; Wilde, J.; et al. A nasal clinical-genomic classifier for assessing risk of malignancy in lung nodules demonstrates accurate performance independent of nodule size or cancer stage. *Chest* **2021**, *160*, A2518–A2519. [CrossRef]
28. Babiarz, J.; Hao, Y.; Manqiu, C.; Bailey, G.; Wilson, D.S.; Krimsky, W.; Sarkar, S.; Bernstein, M.; Lee, H.; Wahidi, M.M.; et al. Detection of actionable molecular alterations through combined DNA/RNA molecular profiling of biopsies collected in early-stage lung cancer at time of diagnosis. *J. Clin. Oncol.* **2021**, *39*, e20546. [CrossRef]
29. Choi, Y.; Liu, T.T.; Pankratz, D.G.; Colby, T.V.; Barth, N.M.; Lynch, D.A.; Walsh, P.S.; Raghu, G.; Kennedy, G.C.; Huang, J. Identification of usual interstitial pneumonia pattern using RNA-Seq and machine learning: Challenges and solutions. *BMC Genom.* **2018**, *19*, 101. [CrossRef] [PubMed]
30. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-Seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
31. Anders, S.; Pyl, P.T.; Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [CrossRef]
32. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef] [PubMed]
33. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef]
34. Haas, B.J.; Dobin, A.; Li, B.; Stransky, N.; Pochet, N.; Regev, A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **2019**, *20*, 213. [CrossRef] [PubMed]

35. R Core Team. R: A Language and Environment for Statistical Computing. Available online: http://www.R-project.org/ (accessed on 8 July 2022).
36. Wickham, H.; Chang, W.; Henry, L.; Pedersen, T.L.; Takahashi, K.; Wilke, C.; Woo, K.; Yutani, H.; Dunnington, D.; RStudio. ggplot2. Available online: https://ggplot2.tidyverse.org/ (accessed on 30 June 2022).
37. Allaire, J.J.; Ellis, P.; Gandrud, C.; Kuo, K.; Lewis, B.W.; Owen, J.; Russell, K.; Rogers, J.; Sese, C.; Yetman, C.J. D3 JavaScript Network Graphs from R. R Package Version 0.4. Available online: https://cran.r-project.org/web/packages/networkD3/networkD3.pdf (accessed on 30 June 2022).
38. Kolde, R. pheatmap: Pretty Heatmaps. Available online: https://rdrr.io/cran/pheatmap/ (accessed on 30 June 2022).
39. Parsonage, H. Waterfalls. R package version 0.1.2. Available online: https://cran.r-project.org/web/packages/waterfalls/index.html (accessed on 30 June 2022).
40. Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **2014**, *30*, 2811–2812. [CrossRef] [PubMed]
41. Wilkins, D. Treemapify 2.5.3. Available online: http://packages.renjin.org/package/org.renjin.cran/treemapify (accessed on 30 June 2022).
42. Seo, J.S.; Ju, Y.S.; Lee, W.C.; Shin, J.Y.; Lee, J.K.; Bleazard, T.; Lee, J.; Jung, Y.J.; Kim, J.O.; Shin, J.Y.; et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.* **2012**, *22*, 2109–2119. [CrossRef] [PubMed]
43. Ali, S.Z.; Siperstein, A.; Sadow, P.M.; Golding, A.C.; Kennedy, G.C.; Kloos, R.T.; Ladenson, P.W. Extending expressed RNA genomics from surgical decision making for cytologically indeterminate thyroid nodules to targeting therapies for metastatic thyroid cancer. *Cancer Cytopathol.* **2019**, *127*, 362–369. [CrossRef] [PubMed]
44. Raghu, G.; Flaherty, K.R.; Lederer, D.J.; Lynch, D.A.; Colby, T.V.; Myers, J.L.; Groshong, S.D.; Larsen, B.T.; Chung, J.H.; Steele, M.P.; et al. Use of a molecular classifier to identify usual interstitial pneumonia in conventional transbronchial lung biopsy samples: A prospective validation study. *Lancet Respir. Med.* **2019**, *7*, 487–496. [CrossRef]
45. Richeldi, L.; Scholand, M.B.; Lynch, D.A.; Colby, T.V.; Myers, J.L.; Groshong, S.D.; Chung, J.H.; Benzaquen, S.; Nathan, S.D.; Davis, J.R.; et al. Utility of a molecular classifier as a complement to high-resolution computed tomography to identify usual interstitial pneumonia. *Am. J. Respir. Crit. Care Med.* **2021**, *203*, 211–220. [CrossRef]
46. Alexander, E.K.; Kennedy, G.C.; Baloch, Z.W.; Cibas, E.S.; Chudova, D.; Diggans, J.; Friedman, L.; Kloos, R.T.; LiVolsi, V.A.; Mandel, S.J.; et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. *N. Engl. J. Med.* **2012**, *367*, 705–715. [CrossRef]
47. Bhorade, S.; Bernstein, M.; Dotson, D.L.; Feller-Kopman, D.; Lee, H.; Choi, Y.; Lofaro, L.; Huang, J.; Whitney, D.; Stevenson, C.; et al. Accuracy of the next generation Percepta Genomic Sequencing Classifier (GSC) for the diagnosis of suspicious intermediate pulmonary nodules. In Proceedings of the American Association for Bronchology and Interventional Pulmonology, Denver, CO, USA, 15–17 August 2019.
48. Silvestri, G.A.; Vachani, A.; Whitney, D.; Elashoff, M.; Porta Smith, K.; Ferguson, J.S.; Parsons, E.; Mitra, N.; Brody, J.; Lenburg, M.E.; et al. A bronchial genomic classifier for the diagnostic evaluation of lung cancer. *N. Engl. J. Med.* **2015**, *373*, 243–251. [CrossRef]
49. Perez-Rogers, J.F.; Gerrein, J.; Anderlind, C.; Liu, G.; Zhang, S.; Alekseyev, Y.; Porta Smith, K.; Whitney, D.; Johnson, W.E.; Elashoff, D.A.; et al. Shared gene expression alterations in nasal and bronchial epithelium for lung cancer detection. *J. Natl. Cancer Inst.* **2017**, *109*, djw327. [CrossRef] [PubMed]
50. Yoshihara, K.; Wang, Q.; Torres-Garcia, W.; Zheng, S.; Vegesna, R.; Kim, H.; Verhaak, R.G. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **2015**, *34*, 4845–4854. [CrossRef] [PubMed]
51. Chase, A.; Ernst, T.; Fiebig, A.; Collins, A.; Grand, F.; Erben, P.; Reiter, A.; Schreiber, S.; Cross, N.C. TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. *Haematologica* **2010**, *95*, 20–26. [CrossRef] [PubMed]
52. Vellichirammal, N.N.; Albahrani, A.; Banwait, J.K.; Mishra, N.K.; Li, Y.; Roychoudhury, S.; Kling, M.J.; Mirza, S.; Bhakat, K.K.; Band, V.; et al. Pan-cancer analysis reveals the diverse landscape of novel sense and antisense fusion transcripts. *Mol. Ther. Nucleic Acids* **2020**, *19*, 1379–1398. [CrossRef]
53. Park, H.J.; Yang, M.J.; Oh, J.H.; Yang, Y.S.; Kwon, M.S.; Song, C.W.; Yoon, S. Genome-wide transcriptional response during the development of bleomycin-induced pulmonary fibrosis in Sprague-Dawley rats. *Toxicol. Res.* **2010**, *26*, 137–147. [CrossRef] [PubMed]
54. Hescheler, D.A.; Riemann, B.; Hartmann, M.J.M.; Michel, M.; Faust, M.; Bruns, C.J.; Alakus, H.; Chiapponi, C. Targeted therapy of papillary thyroid cancer: A comprehensive genomic analysis. *Front. Endocrinol.* **2021**, *12*, 748941. [CrossRef]
55. Yakushina, V.D.; Lerner, L.V.; Lavrov, A.V. Gene fusions in thyroid cancer. *Thyroid* **2018**, *28*, 158–167. [CrossRef]
56. Botton, T.; Talevich, E.; Mishra, V.K.; Zhang, T.; Shain, A.H.; Berquet, C.; Gagnon, A.; Judson, R.L.; Ballotti, R.; Ribas, A.; et al. Genetic heterogeneity of BRAF fusion kinases in melanoma affects drug responses. *Cell Rep.* **2019**, *29*, 573–588. [CrossRef]
57. Ross, J.S.; Wang, K.; Chmielecki, J.; Gay, L.; Johnson, A.; Chudnovsky, J.; Yelensky, R.; Lipson, D.; Ali, S.M.; Elvin, J.A.; et al. The distribution of BRAF gene fusions in solid tumors and response to targeted therapy. *Int. J. Cancer* **2016**, *138*, 881–890. [CrossRef]
58. Kumari, S.; Adewale, R.; Klubo-Gwiezdzinska, J. The molecular landscape of Hürthle cell thyroid cancer is associated with altered mitochondrial function-a comprehensive review. *Cells* **2020**, *9*, 1570. [CrossRef] [PubMed]

59. Ganly, I.; Makarov, V.; Deraje, S.; Dong, Y.; Reznik, E.; Seshan, V.; Nanjangud, G.; Eng, S.; Bose, P.; Kuo, F.; et al. Integrated genomic analysis of Hürthle cell cancer reveals oncogenic drivers, recurrent mitochondrial mutations, and unique chromosomal landscapes. *Cancer Cell* **2018**, *34*, 256–270. [CrossRef] [PubMed]
60. Gopal, R.K.; Kübler, K.; Calvo, S.E.; Polak, P.; Livitz, D.; Rosebrock, D.; Sadow, P.M.; Campbell, B.; Donovan, S.E.; Amin, S.; et al. Widespread chromosomal losses and mitochondrial DNA alterations as genetic drivers in Hürthle cell carcinoma. *Cancer Cell* **2018**, *34*, 242–255. [CrossRef] [PubMed]
61. Stark, R.; Grzelak, M.; Hadfield, J. RNA sequencing: The teenage years. *Nat. Rev. Genet.* **2019**, *20*, 631–656. [CrossRef] [PubMed]
62. Callahan, B.J.; Wong, J.; Heiner, C.; Oh, S.; Theriot, C.M.; Gulati, A.S.; McGill, S.K.; Dougherty, M.K. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* **2019**, *47*, e103. [CrossRef]
63. Thodou, E.; Canberk, S.; Schmitt, F. Challenges in cytology specimens with Hürthle cells. *Front. Endocrinol.* **2021**, *12*, 701877. [CrossRef]
64. Naccache, J.M.; Gibiot, Q.; Monnet, I.; Antoine, M.; Wislez, M.; Chouaid, C.; Cadranel, J. Lung cancer and interstitial lung disease: A literature review. *J. Thorac. Dis.* **2018**, *10*, 3829–3844. [CrossRef]
65. Govindan, R.; Ding, L.; Griffith, M.; Subramanian, J.; Dees, N.D.; Kanchi, K.L.; Maher, C.A.; Fulton, R.; Fulton, L.; Wallis, J.; et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **2012**, *150*, 1121–1134. [CrossRef]
66. Rabizadeh, S.; Garner, C.; Sanborn, J.Z.; Benz, S.C.; Reddy, S.; Soon-Shiong, P. Comprehensive genomic transcriptomic tumor-normal gene panel analysis for enhanced precision in patients with lung cancer. *Oncotarget* **2018**, *9*, 19223–19232. [CrossRef]
67. Rhee, J.K.; Lee, S.; Park, W.Y.; Kim, Y.H.; Kim, T.M. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. *Sci. Rep.* **2017**, *7*, 1653. [CrossRef]
68. Ma, X.; Liu, Y.; Liu, Y.; Alexandrov, L.B.; Edmonson, M.N.; Gawad, C.; Zhou, X.; Li, Y.; Rusch, M.C.; Easton, J.; et al. Pan-cancer genome and transcriptome analyses of 1699 paediatric leukaemias and solid tumours. *Nature* **2018**, *555*, 371–376. [CrossRef] [PubMed]
69. Markman, B.; Javier Ramos, F.; Capdevila, J.; Tabernero, J. EGFR and KRAS in colorectal cancer. *Adv. Clin. Chem.* **2010**, *51*, 71–119. [CrossRef]
70. Kremer, L.S.; Bader, D.M.; Mertes, C.; Kopajtich, R.; Pichler, G.; Iuso, A.; Haack, T.B.; Graf, E.; Schwarzmayr, T.; Terrile, C.; et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **2017**, *8*, 15824. [CrossRef]
71. Iyama, K.; Matsuse, M.; Mitsutake, N.; Rogounovitch, T.; Saenko, V.; Suzuki, K.; Ashizawa, M.; Ookouchi, C.; Suzuki, S.; Mizunuma, H.; et al. Identification of three novel fusion oncogenes, SQSTM1/NTRK3, AFAP1L2/RET, and PPFIBP2/RET, in thyroid cancers of young patients in Fukushima. *Thyroid* **2017**, *27*, 811–818. [CrossRef] [PubMed]
72. Hamatani, K.; Eguchi, H.; Koyama, K.; Mukai, M.; Nakachi, K.; Kusunoki, Y. A novel RET rearrangement (ACBD5/RET) by pericentric inversion, inv(10)(p12.1;q11.2), in papillary thyroid cancer from an atomic bomb survivor exposed to high-dose radiation. *Oncol. Rep.* **2014**, *32*, 1809–1814. [CrossRef]
73. Skálová, A.; Banečkova, M.; Thompson, L.D.R.; Ptáková, N.; Stevens, T.M.; Brcic, L.; Hyrcza, M.; Michal, M., Jr.; Simpson, R.H.W.; Santana, T.; et al. Expanding the molecular spectrum of secretory carcinoma of salivary glands with a novel *VIM-RET* fusion. *Am. J. Surg. Pathol.* **2020**, *44*, 1295–1307. [CrossRef]
74. Davis, A.A.; Patel, V.G. The role of PD-L1 expression as a predictive biomarker: An analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. *J. Immunother. Cancer* **2019**, *7*, 278. [CrossRef]
75. Cheng, H.; Borczuk, A.; Janakiram, M.; Ren, X.; Lin, J.; Assal, A.; Halmos, B.; Perez-Soler, R.; Zang, X. Wide expression and significance of alternative immune checkpoint molecules, B7x and HHLA2, in PD-L1 negative human lung cancers. *Clin. Cancer Res.* **2018**, *24*, 1954–1964. [CrossRef] [PubMed]
76. Delaunay, M.; Cadranel, J.; Lusque, A.; Meyer, N.; Gounant, V.; Moro-Sibilot, D.; Michot, J.M.; Raimbourg, J.; Girard, N.; Guisier, F.; et al. Immune-checkpoint inhibitors associated with interstitial lung disease in cancer patients. *Eur Respir. J.* **2017**, *50*, 1700050. [CrossRef]
77. Kim, J.; Myers, A.C.; Chen, L.; Pardoll, D.M.; Truong-Tran, Q.A.; Lane, A.P.; McDyer, J.F.; Fortuno, L.; Schleimer, R.P. Constitutive and inducible expression of b7 family of ligands by human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **2005**, *33*, 280–289. [CrossRef]
78. Wan, H.; Huang, X.; Cong, P.; He, M.; Chen, A.; Wu, T.; Dai, D.; Li, W.; Gao, X.; Tian, L.; et al. Identification of hub genes associated with COVID-19 and idiopathic pulmonary fibrosis by integrated bioinformatics analysis. *Front. Mol. Biosci* **2021**, *8*, 711239. [CrossRef] [PubMed]
79. Luo, L.; Zhu, G.; Xu, H.; Yao, S.; Zhou, G.; Zhu, Y.; Tamada, K.; Huang, L.; Flies, A.D.; Broadwater, M.; et al. B7-H3 promotes pathogenesis of autoimmune disease and inflammation by regulating the activity of different T cell subsets. *PLoS ONE* **2015**, *10*, e0130126. [CrossRef] [PubMed]
80. Tsuyuki, S.; Tsuyuki, J.; Einsle, K.; Kopf, M.; Coyle, A.J. Costimulation through B7-2 (CD86) is required for the induction of a lung mucosal T helper cell 2 (TH2) immune response and altered airway responsiveness. *J. Exp. Med.* **1997**, *185*, 1671–1679. [CrossRef] [PubMed]

81. Duitman, J.; van den Ende, T.; Spek, C.A. Immune checkpoints as promising targets for the treatment of idiopathic pulmonary fibrosis? *J. Clin. Med.* **2019**, *8*, 1547. [CrossRef]
82. US Food & Drug Administration. Design Control Guidance for Medical Device Manufacturers. Guidance for Industry. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/design-control-guidance-medical-device-manufacturers (accessed on 30 June 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.