

Multi-omic dataset of patient-derived tumor organoids of neuroendocrine neoplasms

Nicolas Alcala^{1,*†}, Catherine Voegelé¹, Lise Mangiante^{1,2}, Alexandra Sexton-Oates¹, Hans Clevers^{3,4,†}, Lynnette Fernandez-Cuesta¹, Talya L. Dayton^{3,4,†,§}, and Matthieu Foll^{1,*†}

¹Rare Cancers Genomics Team (RCG), Genomic Epidemiology Branch (GEM), International Agency for Research on Cancer/World Health Organization (IARC/WHO), Lyon 69008, France

²Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

³Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and UMC Utrecht, 3584 CT Utrecht, The Netherlands

⁴Oncode Institute, Hubrecht Institute, 3584 CT Utrecht, The Netherlands

*Correspondence address. Nicolas Alcala, 25 avenue Tony Garnier CS 90627 69366 Lyon Cedex 07, France. E-mail: alcalan@iarc.who.int, and Matthieu Foll, 25 avenue Tony Garnier CS 90627 69366 Lyon Cedex 07, France E-mail: follm@iarc.who.int

†These authors jointly supervised this work.

‡Current address: Roche Pharmaceutical Research and Early Development, Basel, Switzerland.

§Current address: European Molecular Biology Laboratory (EMBL) Barcelona, Barcelona, Spain.

Abstract

Background: Organoids are 3-dimensional experimental models that summarize the anatomical and functional structure of an organ. Although a promising experimental model for precision medicine, patient-derived tumor organoids (PDTOs) have currently been developed only for a fraction of tumor types.

Results: We have generated the first multi-omic dataset (whole-genome sequencing [WGS] and RNA-sequencing [RNA-seq]) of PDTOs from the rare and understudied pulmonary neuroendocrine tumors ($n = 12$; 6 grade 1, 6 grade 2) and provide data from other rare neuroendocrine neoplasms: small intestine (ileal) neuroendocrine tumors ($n = 6$; 2 grade 1 and 4 grade 2) and large-cell neuroendocrine carcinoma ($n = 5$; 1 pancreatic and 4 pulmonary). This dataset includes a matched sample from the parental sample (primary tumor or metastasis) for a majority of samples (21/23) and longitudinal sampling of the PDTOs (1 to 2 time points), for a total of $n = 47$ RNA-seq and $n = 33$ WGS. We here provide quality control for each technique and the raw and processed data as well as all scripts for genomic analyses to ensure an optimal reuse of the data. In addition, we report gene expression data and somatic small variant calls and describe how they were generated, in particular how we used WGS somatic calls to train a random forest classifier to detect variants in tumor-only RNA-seq. We also report all histopathological images used for medical diagnosis: hematoxylin and eosin-stained slides, brightfield images, and immunohistochemistry images of protein markers of clinical relevance.

Conclusions: This dataset will be critical to future studies relying on this PDTO biobank, such as drug screens for novel therapies and experiments investigating the mechanisms of carcinogenesis in these understudied diseases.

Keywords: organoid, cancer, neuroendocrine neoplasm, genomics, transcriptomics, quality control

Key points:

- Tumor-derived organoids are revolutionary experimental resources to test biological hypotheses and treatment options.
- We have generated the first multi-omic dataset for neuroendocrine tumor organoids of the lung and for the rare neuroendocrine tumors of the pancreas and small intestine (ileum).

Data Description

Context

Organoids are 3-dimensional experimental models that summarize the anatomical and functional structure of an organ [1, 2]. Organoids are revolutionizing fundamental and medical research by allowing us to recapitulate human physiology better than an-

imal models, as well as developmental biology contrary to traditional cell cultures [2]. Patient-derived tumor organoids (PDTOs) have been successfully derived for tumors, providing the experimental tools to model disease progression and the preclinical models for personalized treatment testing [3–5]. Although a promising experimental model, PDTOs have currently been developed only for a fraction of tumor types, focusing on the most frequent cancers and those easiest to culture, leaving rare cancers without appropriate experimental models.

We have recently described one of the very first patient-derived organoid biobanks for the rare and understudied neuroendocrine neoplasms [6]. Neuroendocrine neoplasms are rare tumors that can arise in multiple body sites, predominantly in the lung and gastrointestinal tract [7–9]. Neuroendocrine neoplasms are further classified into neuroendocrine tumors (NETs) and neuroendocrine carcinomas (NECs). NETs are themselves subdivided into grades (ranging from 1 to 2 or 3 depending on the organs), while NECs are subdivided into small cell and large cell (LCNEC).

Received: September 13, 2023. Revised: December 18, 2023. Accepted: February 12, 2024

© World Health Organization, 2024. All rights reserved. The World Health Organization has granted the Publisher permission for the reproduction of this article. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 IGO License (<https://creativecommons.org/licenses/by/3.0/igo/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

While small cell carcinomas are more common (e.g., 15% of lung tumors), have benefited from more studies, and have dedicated treatment options [10], the best treatment option for LCNEC is still unclear [11], and although most NETs progress slowly and have a good prognosis, a subgroup of tumors metastasize and relapse [12].

We report here the multi-omic dataset (whole-genome sequencing [WGS] and RNA sequencing [RNA-seq]) of the neuroendocrine neoplasm PDTO biobank described in [6] (see Table 1). The dataset contains PDTOs of the lung ($n = 12$; 6 grade 1, 6 grade 2) and small intestine ileum ($n = 6$; 2 grade 1 and 4 grade 2), as well as LCNEC of the lung ($n = 4$) and pancreas ($n = 1$). This dataset includes longitudinal sampling of the organoids (2 to 3 time points) and sequencing of the matched parental tumor for most samples (21/23, either primary tumors or metastases). Along with raw and processed data, we provide quality controls for each technique and scripts to run a complete molecular analysis. We also report hematoxylin and eosin-stained (H&E) slides for parental tumors and organoids, brightfield images of organoids, and immunohistochemistry images of neuroendocrine markers (chromogranin A, synaptophysin, CD56, and proliferation marker Ki67) and the EGFR protein. This unique dataset will provide a reference for future research on the understudied neuroendocrine neoplasms.

Methods

Sample collection

PDTO lines of the biobank described in [6] were established from surgical resections or biopsies, put in culture, and expanded. PDTOs periodically underwent passaging, a process by which organoids are subcultured to allow future growth [13]; passage time varied from a week to several months depending on the growth rate ([6], Fig. 2). H&E stainings were performed and samples underwent an independent pathological review, and immunohistochemistry of common neuroendocrine markers (chromogranin A, synaptophysin) were performed to confirm the tumoral neuroendocrine nature of the parental tumors and PDTOs. See [6] for a detailed description of the protocol and the GigaDB repository associated with this article for digital versions of H&E stainings and immunohistochemistry.

Extraction

For each tumor or PDTO, DNA and RNA were extracted from the same sample using the QIAGEN All Prep DNA/RNA Mini kit.

Sequencing

WGS

Whole-genome sequencing was performed by the Utrecht Sequencing Facility. After DNA quality control, genomic DNA (0.5–1 μ g) was used to prepare the whole-genome sequencing library, using the Illumina TruSeq DNA Nano Kit. Libraries were then sequenced on a Novaseq 6000 platform, as paired-end 150-bp reads, with a target average coverage of 30 \times for normal samples and 60 \times to 90 \times for tumor tissue and PDTOs.

RNA-seq

RNA sequencing was performed by the Utrecht Sequencing Facility. After RNA quality control, libraries were prepared using the Illumina TruSeq Stranded mRNA polyA Kit. Libraries were sequenced either on a Nextseq 2000 or an Illumina Novaseq 6000 (RRID:SCR_016387), as paired-end 150-bp reads.

Data processing

All data processing was performed using the workflows developed by the rare cancers genomics team of the International Agency for Research on Cancer/World Health Organization [14], as detailed in [15] and [16]. The workflows are written in the popular domain-specific language nextflow [17]. All software dependencies are contained in conda environments and containerized with Docker and Singularity (containers available online [18, 19]).

WGS

Raw reads were mapped to reference genome GRCh38 using workflow *alignment-nf* v1.2 [20]. This workflow first maps reads (software bwa-mem2 v2.0 [21, 22]), then marks duplicates (software samblaster, v0.1.26 [23]), and finally sorts reads (software sambamba, v0.7.1 [24]).

RNA-seq

Raw reads were mapped to reference genome GRCh38 with annotation gencode v33 using the workflow *RNAseq-nf* v2.4 [25]. This workflow removes adapter sequences (wrapper Trim Galore v0.6.5 [26] for software cutadapt [27]), maps reads (software STAR v2.7.3a [28]), marks duplicated reads (software samblaster, v0.1.25), and finally sorts reads (software sambamba, v0.7.1).

Alignments were then postprocessed using 2 workflows to improve their quality. Workflow *abra-nf* v3.0 [29] performs local realignment using software ABRA2 (v2.22 [30]), and *BQSR-nf* v1.1 [31] performs base quality score recalibration using gatk (v4.0.5.1 [32]).

Variant calling from WGS

Single-nucleotide variants were called on all WGS samples using software Mutect2 from GATK4 (v4.2.0.0 [33, 34]) with workflow *mutect-nf* v2.2b [35], as described in [6]. Resulting variant calling format (VCF) files were normalized using bcftools v1.10.2 [36] (workflow *vcf_normalization-nf* v1.1 [37]) and annotated using ANNOVAR v2020Jun08 (workflow *table_annovar-nf* v1.1.1 [38]). Indels and multinucleotide variants were additionally filtered using the intersection of Mutect2 and strelka2 [39] calls (workflow *strelka2-nf* v1.2a [40]), in order to reduce false positives that are more frequent in indel calls due to the difficulty of detecting such variants with short-read sequencing.

Variant calling from RNA-seq

Variants were called on all RNA-seq samples using software Mutect2 from GATK4 (v4.2.0.0 [33, 34]) with workflow *mutect-nf* (branch *RNAseq*) [35] in RNA-seq and tumor-only modes. The RNA-seq mode incorporates a preprocessing step to fix CIGAR strings (removing NDN elements and ensuring that mapping quality 255 is not used as some mappers like STAR can do) and GATK4's SplitNCigarReads method that splits reads with Ns in their CIGAR string, in order to improve variant calling quality. Resulting VCF files were normalized using bcftools v1.10.2 [36] (workflow *vcf_normalization-nf* v1.1 [37]) and annotated using ANNOVAR v2020Jun08 (workflow *table_annovar-nf* v1.1.1 [38]). For samples that also had WGS data, RNA-seq-detected variants were classified as somatic or germline based on the WGS variant calls described above.

Quality control

For each omic technique, quality controls (QCs) of the samples were performed at each step.

Table 1: Sample summary

ID	Primary site	Tumor type	WGS	RNA-seq	Normal sample (ID)	Tumor sample (ID)	Organoid passages (IDs)
LCNEC1	Pancreas	LCNEC	Yes	Yes	Blood (PANEC1N)	Primary (PANEC1T)	4 (PANEC1Tp4), 14 (PANEC1Tp14)
LNET2	Lung	NET (G1)	Yes	No	Normal-derived organoid passage 7 (LNET2Np7)	Primary (LNET2T)	12 (LNET2Tp12), normal-derived organoid passage 12 (LNET2Np12)
LCNEC3	Lung	LCNEC	Yes	Yes	Tissue (LCNEC3N*)	Primary (LCNEC3T)	17 (LCNEC3Tp17.2), 24 (LCNEC3Tp24)
LCNEC4	Lung	LCNEC	Yes	Yes	Normal-derived organoid passage 6 (LCNEC4Np6)	Primary (LCNEC4T)	7 (LCNEC4Tp7), 24 (LCNEC4Tp24)
LNET5	Lung	NET (G1)	Yes	Yes	Blood (LNET5N)	Primary (LNET5T)	4 (LNET5Tp4), 7 (LNET5Tp7), 2 (LNET5Tp2.2) [†]
LNET6	Lung	NET (G1)	Yes	Yes	Tissue (LNET6N)	Primary (LNET6T)	1 (LNET6Tp1)
mSINET7	Small intestine (ileum)	NET (G2)	Yes	Yes	Blood (SINET7N)	Mesenteric metastasis (SINET7M)	2 (SINET7Mp2)
mSINET8	Small intestine (ileum)	NET (G2)	Yes	Yes	Blood (SINET8N)	Ovary metastasis (SINET8M)	2 (SINET8Mp2)
mSINET9	Small intestine (ileum)	NET (G2)	Yes	No	Blood (SINET9N)	Mesenteric metastasis (SINET9M)	1 (SINET9Tp1)
LNET10	Lung	NET (G2)	Yes	Yes	Blood (LNET10N)	Primary (LNET10T)	4 (LNET10Tp4)
mLCNEC11	Lung	LCNEC	No	Yes	None	Brain metastasis (LCNEC11M)	3 (LCNEC11Mp3)
mSINET12	Small intestine (ileum)	NET (G2)	No	Yes	None	Mesenteric metastasis (SINET12M)	1 (SINET12Mp1 and SINET12Mp1.3) [‡]
LNET13	Lung	NET (G1)	No	Yes	None	Primary (LNET13T)	1 (LNET13Tp1)
LNET14	Lung	NET (G1)	No	Yes	None	Primary (LNET14T)	1 (LNET14Tp1)
mLNET15	Lung	NET (G2)	No	Yes	None	Skin/soft tissue metastasis (LNET15M)	2 (LNET15Mp2)
LNET16	Lung	NET (G2)	No	Yes	None	Primary (LNET16T)	2 (LNET16Tp2)
mLNET16	Lung	NET (G2)	No	Yes	None	Metastasis to the ribcage (LNET16M)	1 (LNET16Mp1)
LNET18	Lung	NET (G2)	No	Yes	None	None	2 (LNET18Tp2, from primary)
LNET19	Lung	NET (G1)	No	Yes	None	Primary (LNET19T)	2 (LNET19Tp2)
mLNET20	Lung	NET (G2)	No	Yes	None	Paravertebral Th1 metastasis (LNET20M)	2 (LNET20Mp2)
mSINET21	Small intestine (ileum)	NET (G1)	No	Yes	None	Paravertebral Th1 metastasis (SINET21M)	2 (SINET21Mp2)
mSINET22	Lung	NET (G1)	No	Yes	None	Paravertebral Th1 metastasis (SINET22M)	2 (SINET22Mp2)
mLCNEC23	Unknown	LCNEC	No	Yes	None	None	3 (LCNEC23Mp3, from paravertebral Th1 metastasis)

For the normal samples, only WGS was performed.

*One normal tissue for this experiment was excluded due to discordance with the tumor (see Fig. 4).

[†]Two lines were derived for LNET5, one sequenced at passages 4 and 7 (samples LNET5Tp4 and LNET5Tp7) and one at passage 2 (LNET5Tp2.2).

[‡]Two lines were derived for SINET12, each sequenced at passage 1 (samples SINET12Mp1.1 and SINET12Mp1.3).

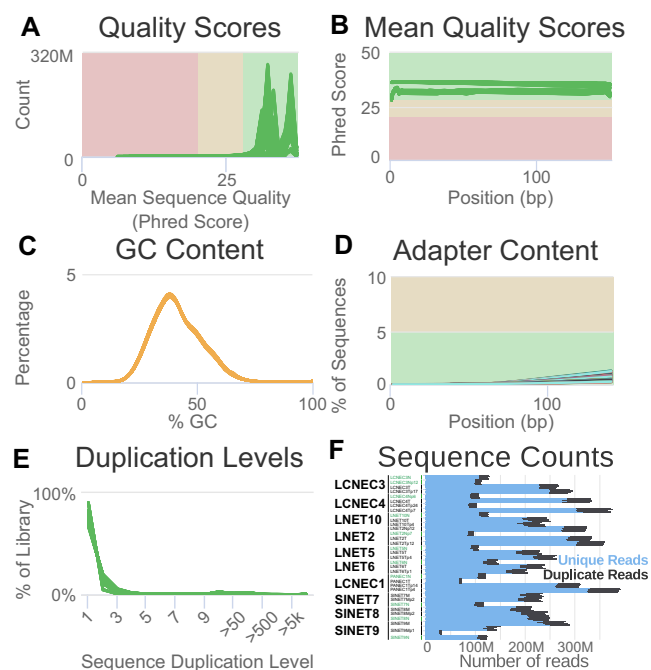


Figure 1: Quality control of the raw WGS data. (A) Distribution of the mean sequence quality of the reads in Phred score. (B) Mean sequence quality score as a function of the position in the read in base pairs (bp). (C) Distribution of the GC content in percentages. (D) Percentage of reads containing a sequence corresponding to the Illumina adapter sequence as a function of the position in the read in bp. (E) Percentage of the library with a given level of duplication. (F) Number of unique and duplicated reads per file. In panels (A–E), each line corresponds to a fastq file, with each of the 34 samples from Table 1 subdivided into 4 sequencing lanes (except SINET9Mp1, subdivided into 8 lanes) and additionally subdivided into 2 read pair files, for a total of $4 \times 2 \times 33 + 8 \times 2 = 280$ files; in panel (F), each horizontal bar corresponds to a file. In (A–E), green lines correspond to files that passed the most stringent quality control filters of software FastQC; orange lines correspond to files that passed a less stringent filter.

Raw reads

Software FastQC (v0.11.9 [41]; [RRID:SCR_014583](#)) was used to check raw reads quality, and software MultiQC (v1.9 [42]; [RRID:SCR_005275](#)) was used to aggregate the QC results across samples and generate interactive plots; all plots from Figs. 1 and 2 were generated by MultiQC from the FastQC outputs. Original MultiQC reports are available in [Supplementary Information \(Files S1–S4\)](#) to allow a free exploration of the QC statistics.

WGS

Raw reads passed quality control filters in all samples. All samples displayed good sequence quality scores (mode above 30 Phred, indicating an error rate below 0.2%), both on average and across all positions in the read (Fig. 1A, B), with samples sequenced later (lower part of Table 1, from LNET5 to LNET10) displaying better scores (highest mode in Fig. 1A). GC content was slightly skewed toward lower values but proved consistent across samples (Fig. 1C), and adapter content (Fig. 1D, less than 5% of sequences with adapter sequence detected) and duplication levels (Fig. 1E, less than 20% of sequences present twice or more) were adequate. The number of reads was consistent between read pairs and consistent with target read depths (Fig. 1F): samples with a target depth of 30 \times —normal, normal-derived organoids, the primary tumor from experiment LCNEC1, and tumor organoid passage 14 from experiment LCNEC3 (LCNEC3Tp14)—having a lower

number of reads ($\sim 4 \times 100$ M reads = 400 M reads) than the others samples ($\sim 4 \times 250$ M = 1,000 M reads), which had a target depth of 90 \times . Note that the metastasis organoid of experiment SINET9 (SINET9Mp1) has been sequenced in 8 lanes, with 4 lanes with a low number of reads (~ 30 M) and 4 additional ones with a larger number (~ 140 M), so the total is comparable with that of the other samples.

RNA-seq

Raw reads passed quality filters after reads trimming for adapter content and quality. All samples displayed good sequence quality scores on average both before and after read trimming (mode above 30 Phred; Fig. 2A, B), with samples sequenced later (lower part of Table 1, from LNET5 to LNET14) displaying better scores (highest mode in Fig. 1A). Six samples displayed lower scores at the end of the reads before trimming (Fig. 2C) but better scores after trimming (Fig. 2D). Indeed, most samples displayed high adapter content before trimming (Fig. 2E), and the trimming step successfully removed them (less than 0.1% in all samples; [Supplementary Information File S2](#)). The trimming step mostly removed less than 5 bp from the read but occasionally could remove up to around 50 bp (Fig. 2F). GC content was consistent across samples (Fig. 2G, H), although the read-trimming step resulted in an excess of reads with high GC content, presumably due to some reads being strongly shortened by the trimming step. Hopefully, in general, the trimming step did not increase much the proportion of short reads (Fig. 2I, J). The number of reads was consistent between read pairs and across sequencing runs both before and after trimming (Fig. 2K, L), and total read numbers for each sample was consistent with the target number of 50 M (25 M pairs): the smallest number, 60.8 M, corresponded to sample PANEC1Tp14.

Alignments

WGS

The software qualimap (v2.2.2b [43]; [RRID:SCR_001209](#)) was called by our workflow *alignment-nf* to generate QC statistics for the WGS alignments in parallel to the data processing (Table 2). All normal and normal tissue-derived organoids displayed a mean coverage $\geq 30\times$, and all tumor and tumor-derived organoids except passage 24 from the organoid of experiment LCNEC4 (sample LCNEC4Tp24) and passage 1 of the organoid of experiment SINET9 (sample SINET9Mp1) had a coverage $\geq 60\times$; all samples displayed at least 65% of the genome with a coverage larger than or equal to 30 \times except LCNEC4Tp24 and (57.4%). Percentages of aligned reads exceeded 99.8% for all samples. Interestingly, some tumor and tumor-derived organoid samples displayed bimodal coverage distributions compatible with variations in copy number state ([Supplementary Information File S3](#)).

RNA-seq

Software RSeQC (v3.0.1 [44]; [RRID:SCR_005275](#)) was called to check alignment quality in parallel to the data processing by workflow *RNAseq-nf*. For all samples, the number of known junctions (i.e., junctions annotated in the gencode v33 annotation file) was stable when resampling subsets of 75% to a 100% of the reads (all lines plateau in Fig. 3A), indicating a good saturation and suggesting that the sequencing depth was sufficient to detect known junctions. In contrast, the number of novel junctions (i.e., junctions not in the annotation file) was increasing slowly as a function of the percentage of reads resampled but did not completely saturate (no complete plateau in Fig. 3B). This indicates that we

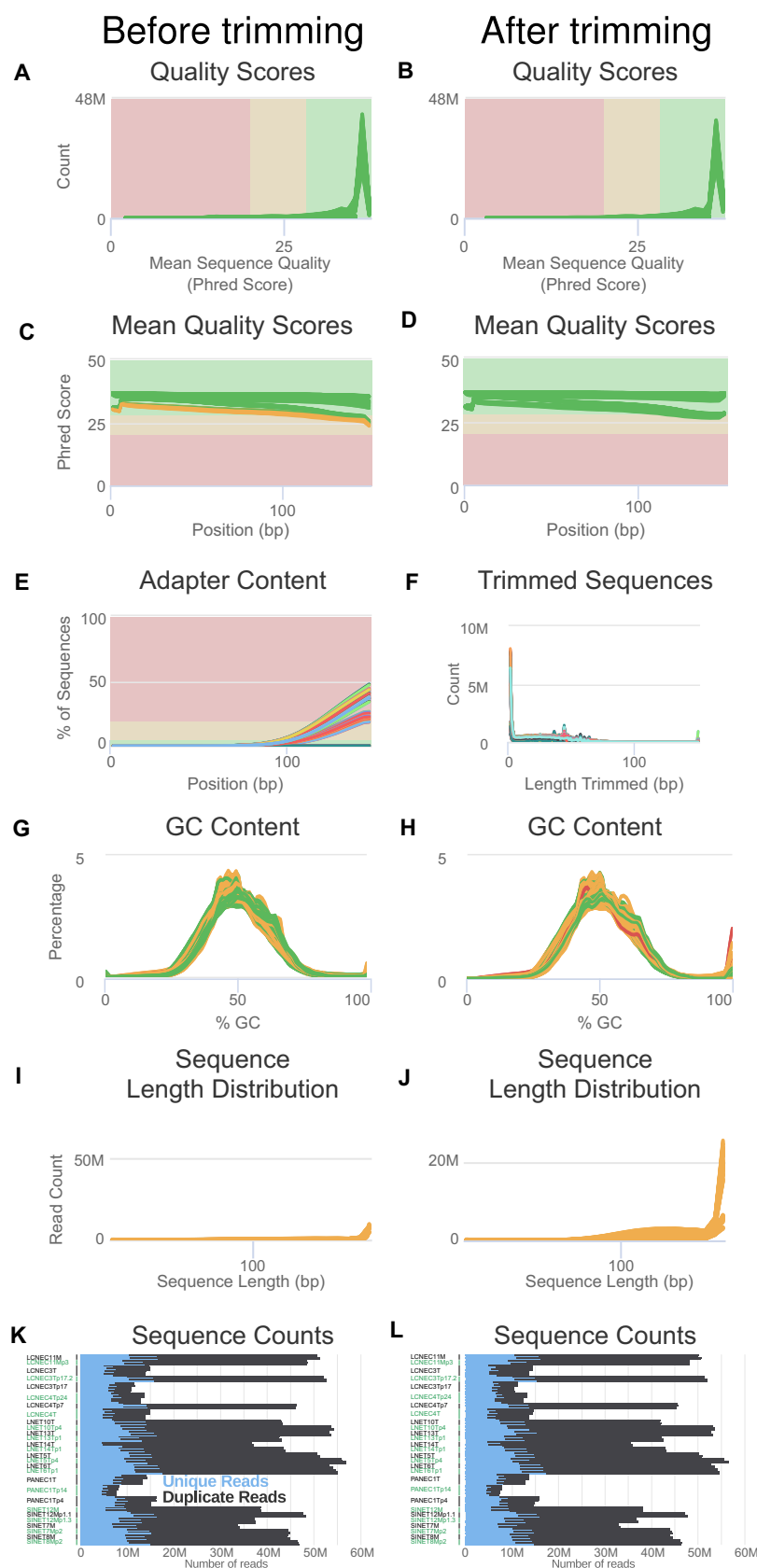


Figure 2: Quality control of the raw RNA-seq data. Panels (A), (C), (E), (G), (I), and (K) correspond to controls before read trimming for quality and adapter content by wrapper Trim Galore for software cutadapt; panels (B), (D), (F), (H), (J), and (L) correspond to controls after read trimming. Figure legends for panels (A–E) and (G–L) follow that of Fig. 1. (F) Distribution of the length of the reads trimmed by software cutadapt, for each file (colored lines). In panels (A–J), each line corresponds to a fastq file, with each of the 10 nonnormal samples from Table 1 divided into 2 or 4 sequencing lanes and further subdivided into 2 read pair files, for a total of $2 \times 2 \times 21 + 4 \times 2 \times 7 = 140$ files; in panels (K) and (L), each horizontal bar corresponds to a file.

Table 2: Quality control of the WGS alignments

Sample Name	% GC	≥30x	≥50x	Coverage	% Aligned
PANEC1N	42%	84.6%	16.2%	41.0x	99.9%
PANEC1T	42%	93.8%	91.7%	104.0x	99.8%
PANEC1Tp4	42%	93.9%	93.3%	127.0x	99.9%
PANEC1Tp14	42%	93.6%	90.9%	89.0x	99.8%
LNET2Np7	41%	67.2%	1.9%	33.0x	99.9%
LNET2Np12	41%	93.4%	92.9%	104.0x	99.9%
LNET2T	41%	93.3%	92.9%	109.0x	99.9%
LNET2Tp12	42%	93.4%	93.1%	115.0x	99.8%
LCNEC3N	41%	82.5%	11.1%	39.0x	99.9%
LCNEC3Np12	42%	82.5%	11.7%	38.0x	99.9%
LCNEC3T	41%	93.9%	90.0%	89.0x	99.9%
LCNEC3Tp17	42%	93.5%	90.9%	90.0x	99.8%
LCNEC4Np6	42%	69.7%	2.7%	34.0x	99.9%
LCNEC4T	41%	93.0%	88.1%	102.0x	99.8%
LCNEC4Tp7	42%	91.7%	87.9%	102.0x	99.9%
LCNEC4Tp24	42%	51.9%	11.3%	30.0x	99.9%
LNET5N	41%	68.9%	2.5%	33.0x	99.9%
LNET5T	42%	91.9%	77.6%	68.0x	99.9%
LNET5Tp4	42%	93.3%	87.6%	75.0x	99.9%
LNET6N	42%	86.6%	22.8%	43.0x	99.9%
LNET6T	42%	93.0%	83.1%	72.0x	99.9%
LNET6Tp1	42%	90.2%	76.8%	61.0x	99.9%
SINET7N	42%	77.2%	4.9%	36.0x	99.9%
SINET7M	41%	92.8%	83.5%	73.0x	99.9%
SINET7Mp2	42%	92.8%	85.7%	69.0x	99.9%
SINET8N	42%	93.1%	91.7%	75.0x	99.9%
SINET8M	41%	92.6%	81.2%	64.0x	99.9%
SINET8Mp2	42%	93.0%	85.5%	70.0x	99.9%
SINET9N	42%	84.4%	7.2%	38.0x	99.9%
SINET9M	41%	93.0%	90.0%	81.0x	99.9%
SINET9Mp1	42%	90.2%	49.1%	49.0x	99.9%
LNET10N	42%	86.4%	9.5%	39.0x	99.9%
LNET10T	42%	93.0%	90.1%	71.0x	99.9%
LNET10Tp4	42%	93.0%	87.1%	66.0x	99.9%

probably detected the most abundant novel junctions but that some low-abundance novel junctions were probably not detected.

Alignment scores were good, with more than 25 M mapped read pairs (50 M reads) for all samples, and from 4 M to 7 M unmapped reads, mainly due to reads being too short or having too many mismatches (Fig. 3C). The distribution of the alignments within annotated regions matched our expectations, with most reads (≥80%) aligning to exons (≥50%), 3' UTR (~20-25%), and 5' UTR (~3%) (Fig. 3D).

Data validation

Sample matching

We used software NGSCheckMate (cloned from the GitHub repository [45] revision 10799087bdf4b990add5b5e536f87c47bbdb688; RRID:SCR_022994) to check that samples from the same experiment indeed came from the same individual, in both WGS and RNA-seq simultaneously, using our workflow NGSCheckMate-rf v1.1[45]. The sample-matching algorithm correctly identified all experiments except one (Fig. 4). The WGS normal-derived organoid sample from experiment LCNEC3 (LCNEC3Np12_WGS in Fig. 4) was found not to match other LCNEC3 samples, suggesting a possible sample swap, and thus excluded from further analyses. Also, the RNA-seq tumor sample for the late-passage organoid of experiment LCNEC3 (sample LCNEC3Tp17_RNA in Fig. 4) was found to better match experiment LNET2 and thus excluded from the subsequent analyses. Finally, 2 samples were found to par-

tially match LNET15 and LNET16, suggesting contamination, and also excluded (UNKN00 and UNKN01).

Sex validation

We validated the sex reported in the clinical data using the multi-omic data. For the WGS data, we used the proportion of reads aligned to the sex chromosomes to assess whether samples clustered by sex (Fig. 5A). We found that all samples clustered by sex except for the normal of experiment LCNEC3 (sample LCNEC3Np12), which clustered with females despite other samples from the experiment clearly clustering with males. This further supports the sample matching reports that suggest that this sample does not match the rest of the experiment. For the RNA-seq data, we compared the total expression level on the sex chromosomes, using the variance-stabilized read counts as a quantification of gene expression (vst function from R package DESeq2 v1.26.0 [46]) (Fig. 5B). We find that samples from the same sex cluster together for all experiments, suggesting concordance with the clinical data.

Small variant calls from RNA-seq

We classified small variants called from RNA-seq in 241 known neuroendocrine neoplasm driver genes (see from Table S4 from reference [6]) as somatic or germline, using a random forest (RF) algorithm [47] (R package randomForest v4.7-1.1 [48]; Fig. 6) and a similar approach as we recently did to classify mutations in tumor-only WGS [49]. After filtering out nonexonic, synonymous, and nonsynonymous mutations with a REVEL score [50] below 0.5, and mutations not in the list of 241 drivers, we were left with 2,430 variants. Among them, 1,174 variants were in samples with WGS data available and their somatic status was thus known.

We used 10 features in the RF model. One feature was directly informative about the potential germline status and came from a public database: the frequency of the allele in human populations from the ExAC database excluding cancers from The Cancer Genome Atlas (TCGA) (feature ExAC_nontcga_ALL). Four features were informative about the alignment and came from the sequencing data themselves: the median distance from the end of the read (feature MPOS), the likelihood ratio score of variant existence (feature TLOD), the coverage at the position (feature DP), and the allelic fraction of the alternative allele (RNA.AF). Finally, the other features were informative about the pathogenicity of the variant and came from public databases: the REVEL score of pathogenicity (feature REVEL), the presence in the COSMIC 92 database (feature cosmic92_coding_nonnull), the presence in the COSMIC 92 database in a lung tumor (feature cosmic92_coding_lung), the InterVar annotation (feature InterVar_automated; with levels “.”, “Uncertain_significance,” “likely_pathogenic,” and “Pathogenic”), and the exonic function of the variant (missense, nonsense, inframe or frameshift insertion, etc).

The RF algorithm was trained and tested on the 1,174 variants with known status (1,148 germline, 26 somatic) called in 22 samples from 8 experiments (Fig. 6A). Note that although the data are imbalanced, we chose to keep this imbalance in the training set to force the algorithm to take into account the fact that most variants are not somatic, and thus having a very good specificity is key to avoid large false discovery rates. We used leave-one-out cross-validation at the experiment level (8 folds), excluding all samples from one same experiment from the model fit at each iteration in order to avoid overfitting due to the inclusion of variants from the same individual but different samples (e.g., LCNEC3T and LC-

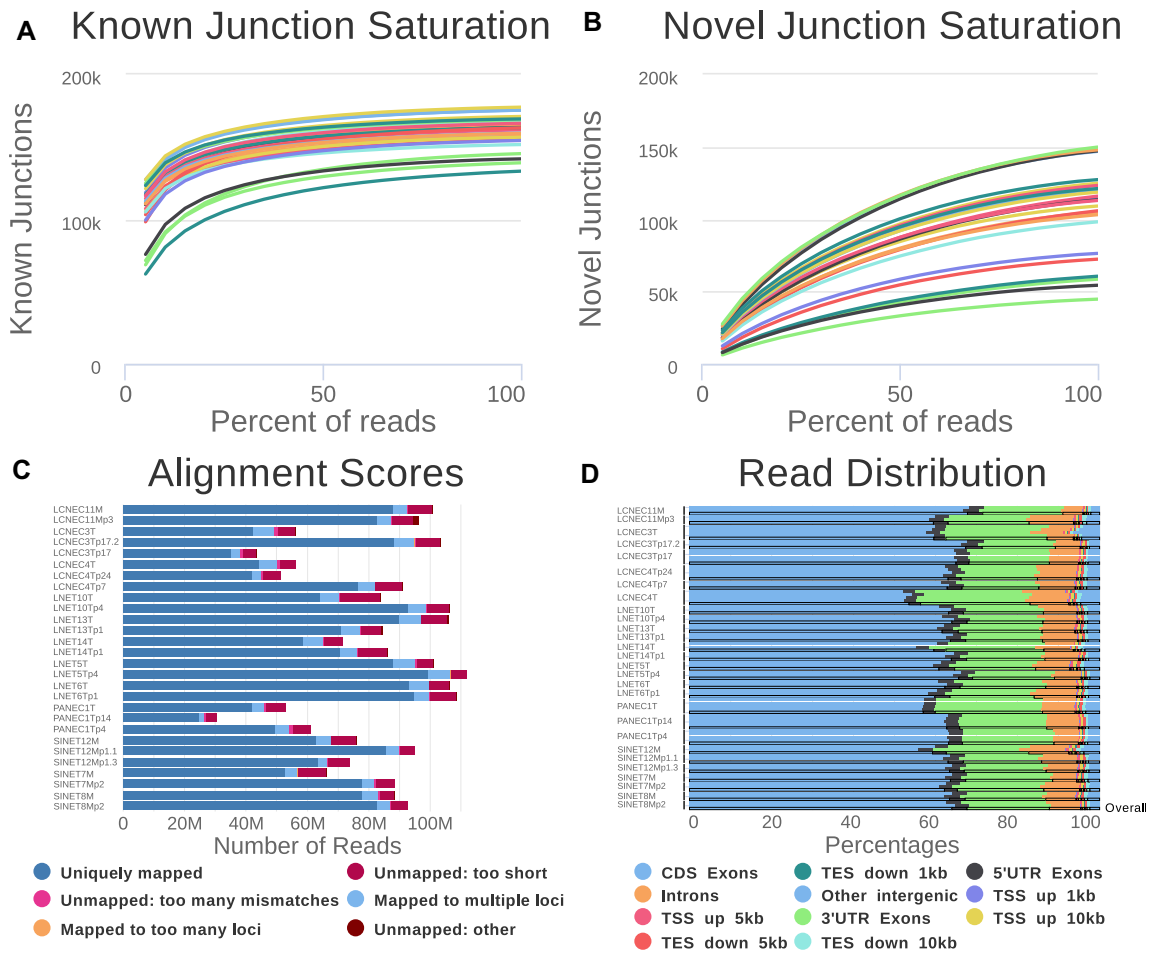


Figure 3: Quality control of the RNA-seq alignments. (A) Number of known junctions identified by software STAR in a subsample as a function of the percentage of reads in the subsample. (B) Number of novel junctions identified by STAR in a subsample as a function of the percentage of reads in the subsample. (C) Number of sequence tags with each alignment score. (D) Distribution of reads among annotated regions.

NEC3Tp17) in the training and test sets. We used 5,000 trees and 3 features per split (the square root of the total number of features as recommended by default) and a minimal node size of 1. We estimated the performance of the model using the receiver operating characteristic (ROC) curve and its area under the curve (AUC, computed using the trapezoid rule), showing the sensitivity as a function of $1 - \text{specificity}$ across different thresholds for the proportion of votes for the somatic class. We also computed the false discovery rate to get a sense of the proportion of variants classified as somatic that would actually be false positives. Once the RF model performance was assessed, we trained a RF model on the full 1,174 variants and predicted the status of the remaining 1,256 variants. See the GitHub repository associated with the article for the complete R script [51]. Note that the same approach allowed us to classify variants called from tumor-only WGS data as somatic or germline with high performance (accuracy greater than 92%; [49]).

We find that we can classify variants as somatic or germline with a balanced accuracy of 86%, with both specificity greater than 98% and sensitivity greater than 73% (AUC = 0.965). Interestingly, although somatic variants are just a fraction of the calls (2%), the high sensitivities and specificities of our RF algorithm allowed us to classify variants with false discovery rates below 50% while still preserving sensitivities above 60% (see Fig. 6B, E–G). We also tested the predictive accuracy of the model fitted on this set

of 1,174 variants from known neuroendocrine neoplasm genes on the set of somatic variants from other recurrently mutated genes in our cohort (Supplementary Fig. S1). We find that the predictive power of the RF model was similar (AUC = 0.90, sensitivity up to 73% with a specificity above 87%).

We evaluated the importance of features for the classification using both the mean decrease in accuracy, which captures how much the model loses accuracy when the feature is excluded, and the mean tree depth at which the feature was observed, with a low value meaning that the feature is used early in the decision trees and thus separates many variants [47, 52] (R package randomForestExplainer v0.10.1). The most important features for the classification were the REVEL score, the TLOD, and the cosmic annotation, while the frequency in the ExAC database was the least important, presumably because all these variants were very rare (Fig. 6C). Indeed, the most representative tree from the RF, computed using the reprotree R package v0.6 using the d2 distance metric between tree predictions [53], relied on these 3 variables, with all alterations present in a lung tumor from the COSMIC 92 database automatically classified as somatic (root of the tree), and TLOD and REVEL score being the most common features used for splitting (Fig. 6D). Of note, using the most important feature alone (the REVEL score) led to a much lower accuracy, consistent with the importance of other features such as TLOD and pathogenic annotations (COSMIC, InterVar).

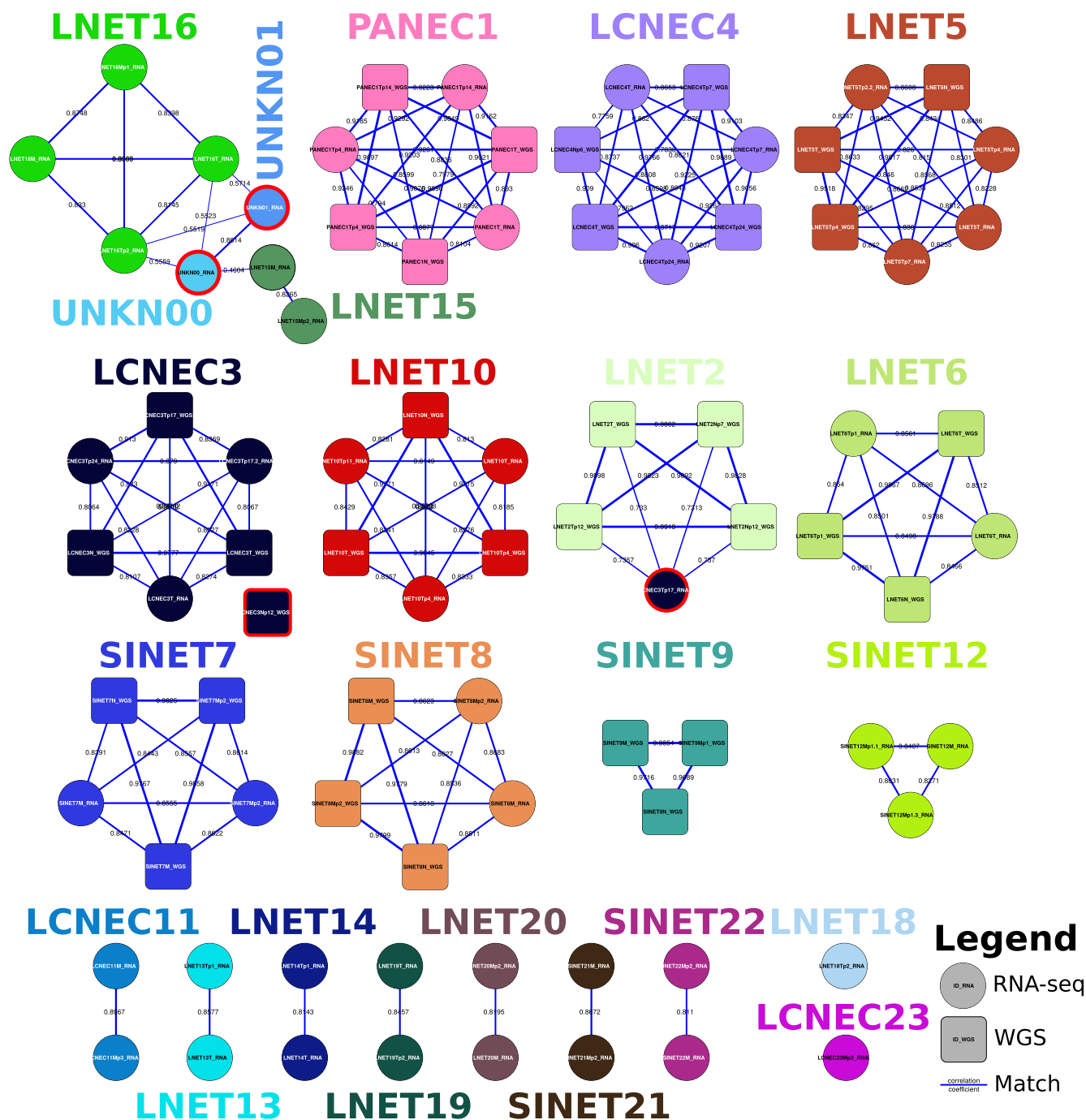


Figure 4: Network of matches between WGS and RNA-seq samples, computed with software NGSCheckmate. Numbers on the edges and edge thickness correspond to the Pearson correlation coefficient r between allelic fractions for the germline SNP panel; colors: experiments (see Table 1); squares: WGS, circles: RNA-seq, red contour: mismatches.

Comparing molecular profiles of PDTOs and parental tumors

We report here all the R scripts used in Dayton et al. [6] to validate that PDTOs faithfully represent their parental tumors (available on the GitHub repository associated with the article [51]). In particular, we provide the code that we used to compare the expression profiles of PDTOs and reference lung and small intestine (SI) NETs and LCNECs with that of PDTOs and their parental tumors (file Fig3B_S3BCE.md [51]). This analysis confirmed the neuroendocrine nature of the PDTOs by showing that they express neuroendocrine markers routinely used in the clinic (>1 transcript per million, TPM in at least 1 of 6 markers). We also provide the code

(file Fig3CD_S3FGHI.md [51]) used in Dayton et al. [6] to demonstrate that pure PDTOs preserve the expression profiles of their parental tumor using dimensionality reduction techniques (Uniform Manifold Approximation and Projection, UMAP). In addition, we provide the code (files Fig4BC_S4BC.md and Fig4D_S4D.md [51]) used to show that PDTOs preserve the genomic profile (small variants, copy number variants, and structural variants) of their parental tumor. To do so, we focused on mutations known to be drivers of neuroendocrine neoplasms [15, 54–60]. Both variants identified with WGS and variants identified with RNA-seq include driver mutations in key recurrently altered LCNEC driver genes

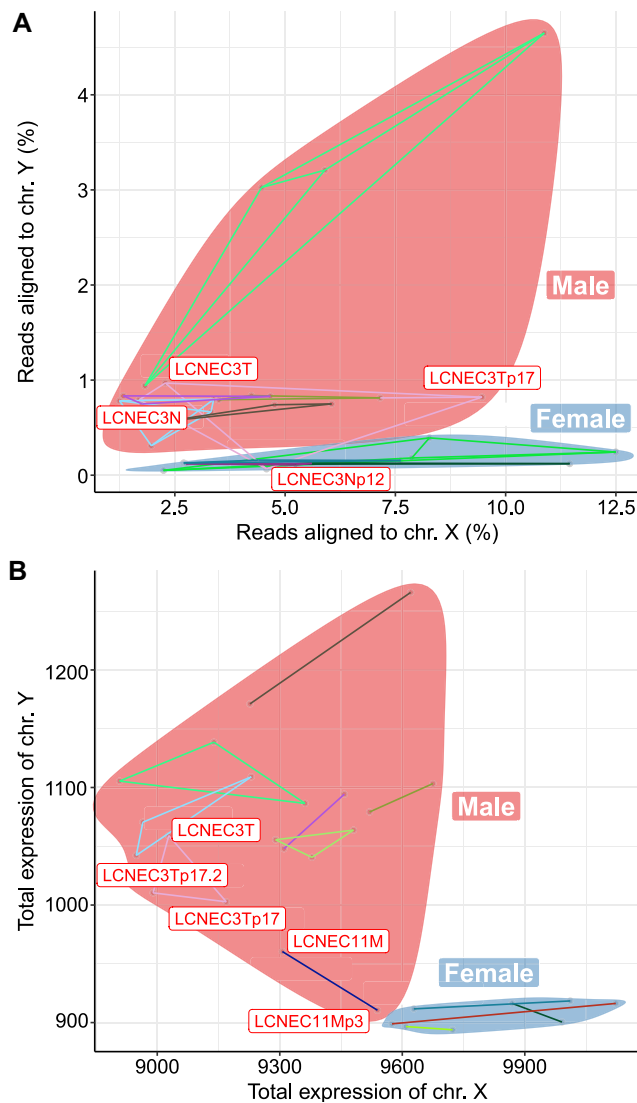


Figure 5: Validation of reported sex. (A) Percentage of reads aligned to chromosomes X and Y in the whole-genome sequencing data. (B) Total gene expression in X and Y chromosomes, in units of variance-stabilized read counts, computed from RNA-seq data. In all panels, samples from each sex are encircled (red: male, blue: female), excluding LCNEC3Np12, which we report as not matching the other samples from the LCNEC3 experiment.

such as *TP53* (mutated in 5/5 LCNECs) and *STK11* (mutated in 3/5 LCNECs). We also identified mutations or structural variants in known driver genes in all but 1 neuroendocrine tumors (17/18), but as previously reported, they involve multiple genes instead of recurrently mutated genes [15, 61]. This confirms that PDTOs recapitulate the genomic profile of neuroendocrine neoplasms.

We also report the R scripts used in Dayton et al. [6] to analyze the temporal evolution of the PDTOs (file Fig5_S5.md [51]). These analyses showed that PDTOs preserve the genetic diversity and clonal architecture of their parents across long periods of time (6 months to more than a year). In particular, the analysis of 2 samples with multiple time points (LCNEC1 and LCNEC4) highlighted that the genetic makeup of the parental tumor is preserved across PDTO passages.

Of note, 1 sample, LCNEC23 was a paravertebral metastasis of an LCNEC of unknown primary. As mentioned in Dayton et al. [6] (fig. 3), the transcriptome of this sample did cluster with the other

LCNEC from the lung and pancreas; in addition, we detected from the RNA-seq 2 high-confidence somatic mutations characteristic of LCNEC: a nonsynonymous *TP53* and a nonsense *PIK3CA* mutation. These molecular results comfort the LCNEC nature of the PDTO, but the overlap between known lung and pancreas LCNEC profiles does not allow to infer the site of origin of the tumor.

Reuse potential

We describe here some of the very first multi-omic datasets for patient-derived tumor organoids of pancreatic, small intestine (ileum), and pulmonary neuroendocrine neoplasms, in particular including the first lung neuroendocrine tumor organoids. Because such low-grade tumors are difficult to cultivate *in vitro*, there is currently a lack of adequate experimental systems for these tumors, and we expect the biobank associated with the data presented here to be the basis for future experimental studies—either fundamental or treatment oriented—on neuroendocrine neoplasms across body sites. The multi-omic dataset we provide here constitutes the molecular fingerprints of these experimental models and will be key to investigate oncogenic processes responsible for tumor initiation and progression and to link drug responses to molecular features to design future personalized treatments.

To facilitate future studies, we used the exact same data processing as in our previous studies of neuroendocrine neoplasms [15, 16] and other rare cancers [62], in particular using rigorous RNA-seq expression quantification with containerized software and operating systems (see Methods section). To ease future studies, we make the expression matrix publicly available (file `gene_expression_PDTOs_parents.tsv` in reference [51]). In addition, we provide all R scripts to analyze the data [51].

Note that the slow passage time of low-grade PDTOs makes them appropriate models to study the biology of neuroendocrine tumors but challenges their use for drug testing. This is particularly true of small intestine NETs, which were only short-term cultures that did not grow past 4 passages. Finally, as noted in most molecular studies of PDTOs [63], one of the main differences between PDTOs and their parental tumors is the absence of a microenvironment. Future work would ideally focus on creating co-cultures of PDTOs and immune cells to remedy this shortcoming.

Conclusion

We have shown that our multi-omic dataset is of high quality and can be easily reused. Given the rarity of neuroendocrine tumors from the lung, pancreas, and small intestine, past genomic studies each only reported data for a handful of samples, limiting the potential discoveries. For example, for lung NETs, 29 WGS and 39 RNA-seq were reported in [61], 3 WGS and 20 RNA-seq in [15], and 30 RNA-seq in [64]; for small intestine NETs, for example, 81 RNA-seq with no WGS were reported in [65] and 7 RNA-seq in [66]. As a result, the primary tumors and metastasis sequencing data we report here (10 samples with WGS, 21 with RNA-seq) alone are very valuable and should be combined with other datasets in future studies to provide enough power to discover informative molecular features for diagnosis, prognosis, and treatment. In addition, we report a unique multi-omic dataset generated from patient-derived tumor organoids, which will allow all researchers working on our biobank to test hypotheses regarding the molecular features associated with drug responses and thus advance research on personalized treatments for these understudied diseases.

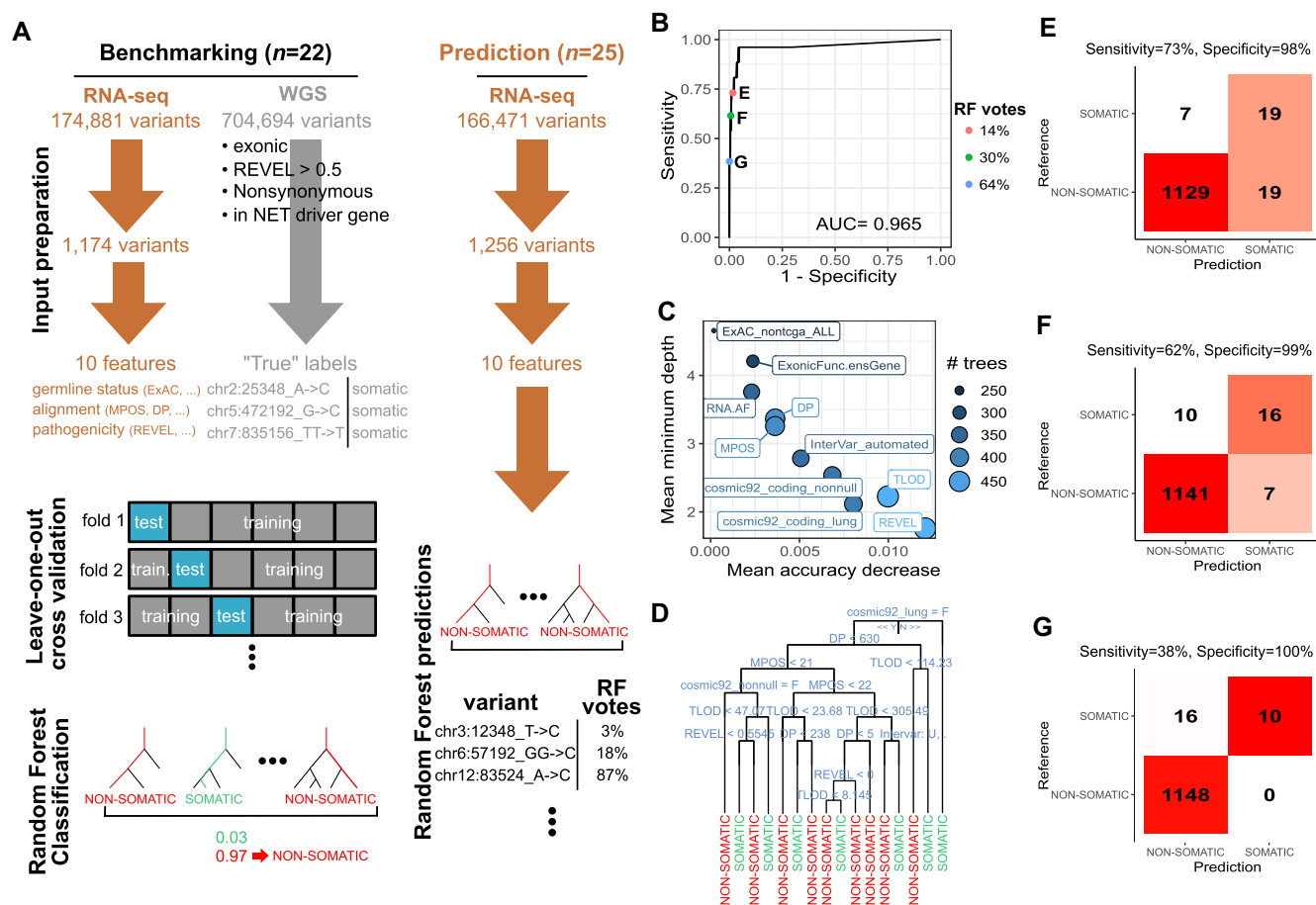


Figure 6: RF classification of variants as somatic or germline from RNA-seq data. (A) Schematic of the RF training, test, and prediction. (B) ROC curve. (C) Feature importance for classification accuracy. Mean accuracy decrease: mean difference in accuracy between trees with the feature and trees without the feature; high values indicate important features. Mean minimum depth: tree depth (1: root, value >1: leaves) of the first time the feature is used for classification, averaged across all trees; low values indicate features often used at the root and thus particularly important. (D) Representative tree of the RF. At each split, the split condition is written above; the left branch corresponds to a Yes and the right branch to a No. Final decision (SOMATIC or NON-SOMATIC) is represented by the leaves. (E–G) Confusion matrix for different levels of sensitivity and specificity. Reference: somatic status assessed from whole-genome sequencing data. Prediction: somatic status predicted from RNA-seq data using the RF algorithm.

Availability of Source Code and Requirements

- Project name: NEN organoids project, lungNENomics
- Project homepages: <https://www.embl.org/groups/dayton/>, <http://rarecancersgenomics.com/lungnlenomics/>
- Operating system(s): Platform independent
- Programming language: Nextflow, R
- Other requirements: R packages *caret*, *randomForest*
- License: GNU GPL

All nextflow command lines for data processing are available at [51] in the readme. All R scripts for the analysis are available in the subfolder Rscripts [51].

Additional Files

Supplementary Fig. S1. Random forest (RF) classification of variants in genes not reported as driver in neuroendocrine neoplasms. (A) ROC curve. (B–D) Confusion matrix for different levels of sensitivity and specificity. Reference: somatic status assessed from whole-genome sequencing data. Prediction: somatic status predicted from RNA-seq data using the RF algorithm.

Supplementary File 1. MultiQC report for raw whole-genome-sequencing data.

Supplementary File 2. MultiQC report for raw RNA-sequencing data.

Supplementary File 3. MultiQC report for whole-genome-sequencing alignments.

Supplementary File 4. MultiQC report for RNA-sequencing alignments.

Abbreviations

AUC: area under the curve; bp: base pairs; LCNEC: large-cell neuroendocrine carcinoma; NEC: neuroendocrine carcinoma; NET: neuroendocrine tumor; PDTO: patient-derived tumor organoid; QC: quality control; RF: random forest; RNA-seq: RNA sequencing; ROC: receiver operating characteristic; VCF: variant calling format; WGS: whole-genome sequencing.

Consent for Publication

All patients signed informed consent forms for molecular analyses and to the publishing of the data.

Acknowledgments

We thank the patients for participating to the study, Utrecht Sequencing for RNA sequencing services, and the editor and the reviewers for their useful suggestions. The results shown here are in part based upon data generated by the Rare Cancers Genomics initiative (www.rarecancersgenomics.com).

Authors' Contributions

Nicolas Alcala, Catherine Voegelé (Data curation [equal], Project administration [supporting], Resources [equal], Software [supporting], Writing – review & editing [supporting]), Lise Mangiante (Conceptualization [supporting], Funding acquisition [supporting], Methodology [supporting], Writing – review & editing [supporting]), Alexandra Sexton-Oates (Methodology [supporting], Writing – review & editing [equal]), Hans Clevers (Conceptualization [supporting], Funding acquisition [supporting], Supervision [supporting]), Lynnette Fernandez-Cuesta (Conceptualization [equal], Funding acquisition [lead], Investigation [supporting], Project administration [supporting], Supervision [supporting], Writing – review & editing [supporting]), Talya L. Dayton (Conceptualization [equal], Data curation [supporting], Funding acquisition [supporting], Investigation [supporting], Methodology [supporting], Project administration [lead], Resources [lead], Supervision [supporting], Writing – review & editing [supporting]), and Matthieu Foll (Conceptualization [equal], Funding acquisition [supporting], Investigation [supporting], Methodology [equal], Software [supporting], Supervision [equal], Validation [supporting], Writing – review & editing [lead])

Funding

The study was funded by the NET Research Foundation (2017 Petersen Accelerator Award to H.C.), Worldwide Cancer Research (2020 Grant Round to L.F.-C.), NET Research Foundation (2019 Investigator Award to L.F.-C.), French National Cancer Institute (INCa, PRT-K 2017 to L.F.-C. and M.F.), and Ligue Nationale contre le Cancer (fellowship to L.M.). T.L.D. was supported by an EMBO long-term fellowship (ALTF-21-2017) and a Marie Skłodowska-Curie IF grant 797966– PNECtumor. The Oncode Institute is supported by the Dutch Cancer Society.

Data Availability

The dataset supporting the results of this article is available in the European Genome-Phenome archive repository, study EGAS00001005752. The study consists of 7 datasets: EGAD00001009988, with WGS CRAM files for 2 experiments; EGAD00001009989 with WGS CRAM files for 6 experiments; EGAD00001009990, with WGS CRAM files for 2 experiments; EGAD00001009991, with RNA-seq fastq files from 4 experiments; EGAD00001009992, with RNA-seq fastq files for 15 experiments; EGAD00001009993, with RNA-seq fastq files for 2 experiments; and EGAD00001009994, with gene expression in multiple formats (R data, tab-separated text files) and multiple units (raw counts, TPM, FPKM) for 21 samples. Because of the sensitivity of the data and the patient consent, to get access to the data, please contact the data access committee of the Division of Biomedical Genetics from UMC Utrecht at dacdbg@umcutrecht.nl. Once a data access agreement has been signed and access granted, data can be downloaded using the EGA Python client (see detailed instructions [67], and video tutorial [68]). Expression matrices in raw counts format

and small variants are also publicly available on the GitHub repository under the data folder [51].

The multiQC report for WGS raw reads is available in [Supplementary File S1](#), the multiQC report for RNA-seq raw reads is available in [Supplementary File S2](#), the multiQC report for WGS alignments is available in [Supplementary File S3](#), and the multiQC report for RNA-seq alignments is available in [Supplementary File S4](#).

Snapshots of our code and other data further supporting this work are openly available in the GigaScience repository, GigaDB [69].

Organoid lines mentioned in this article can be requested from Hans Clevers (h.clevers@hubrecht.eu) or Talya Dayton (talya.dayton@embl.es). Distribution of organoids to third parties will have to be authorized by the relevant ethical committee, and a complete material transfer agreement will be required to ensure compliance with the Dutch “Medical Research Involving Human Subjects” act. Use of organoids is subjected to patient consent; note that upon consent withdrawal, distributed organoid lines and any derived material will have to be promptly disposed of.

Competing Interests

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

H.C.'s full disclosure is given at [70]. H.C. is inventor of several patents related to organoid technology, cofounder of Xilis, and currently an employee of Roche, Basel.

Ethical Approval

This study was approved by the medical ethical committee of each respective hospital of the patients: Verenigde Commissies Mensgebonden Onderzoek of the St. Antonius Hospital Nieuwegein, Z-12.55; UMC Utrecht, METC 12-093 HUB-Cancer; NKI Institutional Review Board (IRB), M18ORG/CFMPB582; and Maastricht University Medical Center, METC 2019-1061 and 2019-1039.

References

1. Clevers H. Modeling development and disease with organoids. *Cell* 2016;165(7):1586–97. <https://doi.org/10.1016/j.cell.2016.05.082>.
2. Kim J, Koo BK, Knoblich JA. Human organoids: model systems for human biology and medicine. *Nat Rev Mol Cell Biol* 2020;21(10):571–84. <https://doi.org/10.1038/s41580-020-0259-3>.
3. Drost J, Clevers H. Organoids in cancer research. *Nat Rev Cancer* 2018;18(7):407. <https://doi.org/10.1038/s41568-018-0007-6>.
4. Tuveson D, Clevers H. Cancer modeling meets human organoid technology. *Science* 2019;364(6444):952–5. <https://doi.org/10.1126/science.aaw6985>.
5. LeSavage BL, Suh RA, Broguiere N, et al. Next-generation cancer organoids. *Nat Mater* 2022;21(2):143–59. <https://doi.org/10.1038/s41563-021-01057-5>.
6. Dayton TL, Alcala N, Moonen L, et al. Druggable growth dependencies and tumor evolution analysis in patient-derived organoids of neuroendocrine neoplasms from multiple body sites. *Cancer Cell* 2023;41(12):2083–99. <https://doi.org/10.1016/j.ccell.2023.11.007>.

7. Rindi G, Klimstra DS, Abedi-Ardekani B, et al. A common classification framework for neuroendocrine neoplasms: an International Agency for Research on Cancer (IARC) and World Health Organization (WHO) expert consensus proposal. *Mod Pathol* 2018;31(12):1770–86. <https://doi.org/10.1038/s41379-018-0110-y>.
8. Travis W, Beasley M, Cree I, et al. Lung neuroendocrine neoplasms. In: WHO Classification of Tumours: Thoracic Tumours, 5th ed. Lyon, France: International Agency for Research on Cancer; 2021:127–149.
9. Klimstra D, Klöppel G, La Rosa S, et al. Classification of neuroendocrine neoplasms of the digestive system. In: WHO Classification of Tumours: Digestive System Tumours, 5th ed. Lyon, France: International Agency for Research on Cancer; 2019:16–19.
10. Rudin CM, Poirier JT, Byers LA, et al. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nat Rev Cancer* 2019;19(5):289–97. <https://doi.org/10.1038/s41568-019-0133-9>.
11. Derks JL, Leblay N, Lantuejoul S, et al. New insights into the molecular characteristics of pulmonary carcinoids and large cell neuroendocrine carcinomas, and the impact on their clinical management. *J Thorac Oncol* 2018;13(6):752–66. <https://doi.org/10.1016/j.jtho.2018.02.002>.
12. Fernandez-Cuesta L, Foll M. Molecular studies of lung neuroendocrine neoplasms uncover new concepts and entities. *Transl Lung Cancer Res* 2019;8(Suppl 4):S430. <https://doi.org/10.21037/tlcr.2019.11.08>.
13. Zhao Z, Chen X, Dowbaj AM, et al. Organoids. *Nat Rev Methods Primers* 2022;2(1):94. <https://doi.org/10.1038/s43586-022-00174-y>.
14. IARC bioinformatics platform GitHub repository. <https://github.com/IARCBioinfo/>. Accessed 8 February 2024.
15. Alcala N, Leblay N, Gabriel A, et al. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. *Nat Commun* 2019;10, article number 3407. <https://doi.org/10.1038/s41467-019-11276-9>.
16. Gabriel AA, Mathian E, Mangiante L, et al. A molecular map of lung neuroendocrine neoplasms. *Gigascience* 2020;9(11):giaa112. <https://doi.org/10.1093/gigascience/giaa112>.
17. Di Tommaso P, Chatzou M, Floden EW, et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35(4):316. <https://doi.org/10.1038/nbt.3820>.
18. Dockerhub home page. <https://hub.docker.com/>. Accessed 8 February 2024.
19. Singularity hub home page. <https://singularity-hub.org/>. Accessed 8 February 2024.
20. IARCBioinfo whole-genome sequencing alignment pipeline. <https://github.com/IARCBioinfo/alignment-nf>. Accessed 8 February 2024.
21. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010;26(5):589–95. <https://doi.org/10.1093/bioinformatics/btp698>.
22. Vasmuddin M, Misra S, Li H, et al. Efficient architecture-aware acceleration of bwa-mem for multicore systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Rio de Janeiro, Brazil: IEEE; 2019: 314–24.
23. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 2014;30(17):2503–5. <https://doi.org/10.1093/bioinformatics/btu314>.
24. Tarasov A, Vilella AJ, Cuppen E, et al. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 2015;31(12):2032–4. <https://doi.org/10.1093/bioinformatics/btv098>.
25. IARCBioinfo RNA sequencing alignment pipeline. <https://github.com/IARCBioinfo/RNAseq-nf>. Accessed 8 February 2024.
26. Krueger F. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisulfite-Seq) libraries. 2012. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed 28 June 2019.
27. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;17(1):10–12. <https://doi.org/10.14806/ej.17.1.200>.
28. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
29. IARCBioinfo local re-alignment pipeline. <https://github.com/IARCBioinfo/abra-nf>. Accessed 8 February 2024.
30. Mose LE, Wilkerson MD, Hayes DN, et al. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* 2014;30(19):2813–5. <https://doi.org/10.1093/bioinformatics/btu376>.
31. IARCBioinfo base quality score recalibration pipeline. <https://github.com/IARCBioinfo/BQSR-nf>. Accessed 8 February 2024.
32. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinform* 2013;43(1):11–10.
33. Benjamin D, Sato T, Cibulskis K, et al. Calling somatic SNVs and indels with Mutect2. *BioRxiv* 2019; 861054. <https://doi.org/10.1101/861054>.
34. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. Sebastopol, USA: O'Reilly Media; 2020.
35. IARCBioinfo GATK mutect2 variant calling pipeline. <https://github.com/IARCBioinfo/mutect-nf>. Accessed 8 February 2024.
36. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10(2):giab008. <https://doi.org/10.1093/gigascience/giab008>.
37. IARCBioinfo variant calling format files normalization pipeline. https://github.com/IARCBioinfo/vcf_normalization-nf. Accessed 8 February 2024.
38. IARCBioinfo variant calling format files annotation with ANNOVAR pipeline. https://github.com/IARCBioinfo/table_annovar-nf. Accessed 8 February 2024.
39. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;15(8):591–4. <https://doi.org/10.1038/s41592-018-0051-x>.
40. IARCBioinfo strelka2 variant calling pipeline. <https://github.com/IARCBioinfo/strelka2-nf>. Accessed 8 February 2024.
41. Andrews S, Krueger F, Segonds-Pichon A, et al. FastQC. Babraham, UK: Babraham Institute; 2012.
42. Ewels P, Magnusson M, Lundin S, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;32(19):3047. <https://doi.org/10.1093/bioinformatics/btw354>.
43. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 2015;32(2):292–4. <https://doi.org/10.1093/bioinformatics/btv566>.
44. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;28(16):2184–5. <https://doi.org/10.1093/bioinformatics/bts356>.

45. IARCbioinfo NGSCheckMate sample matching pipeline. <https://github.com/parklab/NGSCheckMate>. Accessed 8 February 2024.
46. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550. <https://doi.org/10.1186/s13059-014-0550-8>.
47. Breiman L. Random forests. *Mach Learn* 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>.
48. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002;2(3):18–22.
49. Di Genova A, Mangiante L, Sexton-Oates A, et al. A molecular phenotypic map of malignant pleural mesothelioma. *Gigascience* 2023;12:giac128. <https://doi.org/10.1093/gigascience/giac128>.
50. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;99(4):877–85. <https://doi.org/10.1016/j.ajhg.2016.08.016>.
51. IARCbioinfo NGSCheckMate sample matching pipeline. https://github.com/IARCbioinfo/MS_panNEN_organoids. Accessed 8 February 2024.
52. Ishwaran H, Kogalur UB, Gorodeski EZ, et al. High-dimensional variable selection for survival data. *J Am Statist Assoc* 2010;105(489):205–17. <https://doi.org/10.1198/jasa.2009.tm0862.2>.
53. Banerjee M, Ding Y, Noone AM. Identifying representative trees from ensembles. *Stat Med* 2012;31(15):1601–16. <https://doi.org/10.1002/sim.4492>.
54. Banck MS, Kanwar R, Kulkarni AA, et al. The genomic landscape of small intestine neuroendocrine tumors. *J Clin Invest* 2013;123(6):2502–8. <https://doi.org/10.1172/JCI67963>.
55. Sei Y, Zhao X, Forbes J, et al. A hereditary form of small intestinal carcinoid associated with a germline mutation in inositol polyphosphate multikinase. *Gastroenterology* 2015;149(1):67–78. <https://doi.org/10.1053/j.gastro.2015.04.008>.
56. Miyoshi T, Umemura S, Matsumura Y, et al. Genomic profiling of large-cell neuroendocrine carcinoma of the lung. *Clin Cancer Res* 2017;23(3):757–65. <https://doi.org/10.1158/1078-0432.CCR-16-0355>.
57. Pelosi G, Bianchi F, Dama E, et al. Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm. *Virchows Arch* 2018;472:567–77. <https://doi.org/10.1007/s00428-018-2307-3>.
58. Simbolo M, Vicentini C, Mafficini A, et al. Mutational and copy number asset of primary sporadic neuroendocrine tumors of the small intestine. *Virchows Arch* 2018;473:709–17. <https://doi.org/10.1007/s00428-018-2450-x>.
59. Walter D, Harter PN, Battke F, et al. Genetic heterogeneity of primary lesion and metastasis in small intestine neuroendocrine tumors. *Sci Rep* 2018;8(1):3811. <https://doi.org/10.1038/s41598-018-22115-0>.
60. Samsom KG, Levy S, van Veenendaal LM, et al. Driver mutations occur frequently in metastases of well-differentiated small intestine neuroendocrine tumours. *Histopathology* 2021;78(4):556–66. <https://doi.org/10.1111/his.14252>.
61. Fernandez-Cuesta L, Peifer M, Lu X, et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat Commun* 2014;5(1):3518. <https://doi.org/10.1038/ncomms4518>.
62. Mangiante L, Alcalá N, Sexton-Oates A, et al. Multiomic analysis of malignant pleural mesothelioma identifies molecular axes and specialized tumor profiles driving intertumor heterogeneity. *Nat Genet* 2023;55(4):607–18. <https://doi.org/10.1038/s41588-023-01321-1>.
63. Lee SH, Hu W, Matulay JT, et al. Tumor evolution and drug response in patient-derived organoid models of bladder cancer. *Cell* 2018;173(2):515–28. <https://doi.org/10.1016/j.cell.2018.03.017>.
64. Laddha SV, Da Silva EM, Robzyk K, et al. Integrative genomic characterization identifies molecular subtypes of lung carcinoids. *Cancer Res* 2019;79(17):4339–47. <https://doi.org/10.1158/0008-5472.CAN-19-0214>.
65. Alvarez MJ, Subramaniam PS, Tang LH, et al. A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nat Genet* 2018;50(7):979–89. <https://doi.org/10.1038/s41588-018-0138-4>.
66. Hofving T, Liang F, Karlsson J, et al. The microenvironment of small intestinal neuroendocrine tumours contains lymphocytes capable of recognition and activation after expansion. *Cancers* 2021;13(17):4305. <https://doi.org/10.3390/cancers13174305>.
67. EGA python client for data download home page. <https://github.com/EGA-archive/ega-download-client>. Accessed 8 February 2024.
68. Video tutorial for the EGA Python client. <https://embl-ebi.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=be79bb93-1737-4f95-b80f-ab4300aa6f5a>. Accessed 8 February 2024.
69. Alcalá N, Voegelé C, Mangiante L, et al. Supporting data for “Multi-Omic Dataset of Patient-Derived Tumor Organoids of Neuroendocrine Neoplasms.” *GigaScience Database*. 2024. <http://dx.doi.org/10.5524/102494>.
70. Pr. Hans Clevers competing interest disclosure. <https://www.uu.nl/staff/JCClevers/>. Accessed 8 February 2024.