D1.1

Collection of public molecular and clinical data

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	48 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	SC1-DTH-07-826121 / D1.1
Work package contributing to the deliverable	WP1
Due date	30th June, 2020
Actual submission date	1st July, 2020

Responsible organisation	BCM
Editor	Pavel Sumazin
Dissemination level	PU
Revision	V1

Abstract	The data collection will enable model development throughout the work packages in the project
Keywords	Public, molecular data, clinical data

The project iPC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

Pavel Sumazin (BCM)

Internal Reviewer

Pieter Mestdagh (UGENT)

Disclaimer

Executive Summary

The development of iPC predictive models for paediatric cancer genesis, progression, and response to therapies, as well as patient response to therapy, requires a vast quantity of molecular and clinical training data. In this deliverable, we have assembled a collection of these data to enable model construction and testing. These datasets include cancer-specific data that could be used to evaluate the effects of treatments and perturbations targeting cancer and other data that could be used to construct models and inform efforts to deconvolve regulatory interactions that are a key to understanding how the effects of alterations are propagated and affect tumor and non-tumor cells. Cancer-specific data were generated by multiple international efforts to molecularly profile primary tumors on a large-scale by multi-omics. They also include efforts to molecularly profile cancer cell lines and evaluate cancer-cell responses to perturbations, including common and potential therapies as well as targeted up- or down-regulation. Non-cancer specific data include molecular and phenotypic data from healthy tissues and non-cancer cells, as well as large-scale multi-omics profiling of individual tissues and people. All of these data are being used in iPC to construct models of regulation and responses to perturbations by gene-therapy, immunotherapy, small molecules, radiotherapy, and other potential treatments.

Table of Content

Chapter 1 Introduction

iPC set out to model patient disease and response to therapies based on molecular and clinical variants that are predictive of survival, stage, molecular classification, and response to therapies—including cure rates and toxicities—for patients with multiple types of pediatric cancers [1-5]. The creating of these models depends on our ability to model cancer and normal cell responses to perturbations and potential treatments. Deliverable 1.1 is a collection of data that will enable model development, including data that is specific to tumors of interest and data that would aid efforts to model tumors of interest. We classify these data into (1) data that are specific to our cancers of interest and (2) data to help construct models of cancer cells and tissues as well as normal patient tissues. Pediatric cancer-specific data include clinical and molecular profiles for our cancers of interest, including models of these cancers. It also includes perturbations, omics, and phenotypic assays performed on these cancer models. Data that will help to construct models include molecular and clinical data for other cancers and models, including model profiling, treatment-response data, and phenotype and genotype profiles following perturbation by chemical and biological agents, including potential treatments and treatment combinations.

Data specific to our cancers of interest

Ewing Sarcoma. In total, we collected demographic, clinical, and molecular profiles for 319 Ewing sarcoma patients.

Hepatoblastoma. We collected demographic, clinical, and molecular profiles for 352 hepatoblastoma and paediatric hepatocellular carcinoma patients. These include patients from 4 large studies, amongst others an ongoing international collaboration that continues enrolling patients. We include paediatric hepatocellular carcinoma patients in this study because recent research suggests that hepatoblastoma and paediatric hepatocellular carcinoma are linked. We collected biopsies from over 40 patients that appear to have mixed biology and phenotypes of both hepatoblastoma and paediatric hepatocellular carcinoma.

Paediatric leukaemia. We collected multiple datasets with clinical annotation and molecular profiling of Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) paediatric patients. In total, the datasets include over clinical, demographic, and molecular data for over 2,000 leukemia patients.

Medulloblastoma. iPC partners and collaborators help lead medulloblastoma research. In total, we collected demographic, clinical, and molecular profiles for 2462 medulloblastoma patients.

Neuroblastoma. Neuroblastoma has been a focus of research for multiple iPC investigators. In total, we collected demographic, clinical, and molecular profiles for 5115 neuroblastoma patients. These include patient data for 17 neuroblastoma cohorts with at least 100 patients.

Data to help construct models of cancer and normal cell response to therapies

We collected demographic, clinical, treatment, and molecular profiles for over 350,000 tumor biopsies. These data—with The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) accounting for the largest datasets - will be used to build models that predict outcome and responses for therapy by cancer cells. In addition, we have assembled multiple large paediatric cancer datasets, including clinical and molecular data from an international consortium that is focused on paediatric glioma. While these cancers were not originally selected as areas of focus for iPC, they are closely related to our mission to study and build methods to design new and personalized diagnostic and therapeutic strategies for paediatric cancers. In addition to data associated with tumor biopsies, we have collected interaction data, regulatory region data, perturbation data, therapy-response data, and genome-wide essentiality data for cancer cells, non-cancer cells, cancer tissues, and non-cancer tissues. All these data are intended to help build models for regulation and perturbation-responses on the cellular, tissue, and patient levels.

Chapter 2 Data specific to our cancer types

We aim to study 5 types of paediatric cancers. To prepare for this effort, we collected publicly available cancer datasets for each of the 5 cancer types. Datasets include demographic and clinical data for cancer patients, as well as molecular profiles for primary and metastatic tumors at diagnosis, resection, or relapse. While multiple small datasets are available for each tumor, in the following we list the largest patient datasets collected for each tumor type.

2.1 Ewing sarcoma

Our Ewing sarcoma datasets are the result of efforts by multiple groups over the last 15 years. The largest of these efforts are outlined below.

- The dataset by Postel-Vinay et al. [6] includes demographic information and molecular profiles for 117 Ewing sarcoma patients
- The dataset by Savola et al. [7] includes demographic and RNA expression profiles for 117 Ewing sarcoma patients.
- The dataset by Volchenboum et al. [8] includes rich clinical characterization, demographic, and RNA expression profiles for 85 Ewing sarcoma patients. Clinical data includes detailed descripting on tumor sites, morphology, and immunohistochemistry, in addition to survival, EFS, and any event data.
- The dataset by Tirode et al. [9] includes molecular profiles for 39 Ewing sarcoma patients with clinical, demographic, and molecular profiling includes RNA expression and phenotypic assays include differentiation assays and cell-cycle analyses of Ewing sarcoma models.

2.2 Hepatoblastoma

Our hepatoblastoma datasets have been collected by research centres around the world. The largest datasets are given below, and current efforts are geared to collect and profile samples with rich clinical data within cooperative efforts across the globe. These efforts are producing data for analysis by iPC while trials are ongoing.

- The dataset by Carrillo-Reixach et al. [10] includes rich clinical characterization, demographic, and molecular profiles for 159 tumor samples from 113 patients. Molecular profiles include DNA alterations, RNA expression, and promoter methylation profiles.
- The dataset by Sumazin et al. [11] includes rich clinical characterization, demographic, proteomic, DNA alteration, and RNA expression profiles for 82 hepatoblastoma patients.
- We entered in an agreement for data acquisition from the Fibrolamellar Registry, which includes an extremely rich clinical and demographic annotation as well as molecular profiles for over 250 patients. Clinical and demographic annotation includes a 600-item questioner that is required for each enrolled patient. Molecular profiles include expression and methylation data.

2.3 Paediatric leukaemia

Our leukemia datasets include large scale profiles of samples and data collected within national and international consortiums. These include the datasets below. Multiple efforts to produce single-cell level profiles are ongoing by iPC collaborators and are expected to generate validation data for the project.

- The Therapeutically Applicable Research To Generate Effective Treatments (TARGET) ALL data set includes data for 819 ALL patients with tumor and matched normal DNA profiles—

including targeted sequencing and WES, WGS, or both—as well as copy number analyses, RNA-expression profiles by RNA-Seq, and some protein expression profiles. Clinical and demographic data include patient age at diagnosis, gender, event free survival, vital status, dates of follow-ups and events, treatment protocols, trial protocols, therapeutic protocols, clinical tests that include pathology and genetic testing, and free comments.

- The Therapeutically Applicable Research To Generate Effective Treatments (TARGET) AML data set includes data for 988 AML patients. Molecular data include tumor and matched normal DNA profiles, microRNA profiling using smallRNA sequencing, RNA-expression profiles by RNA-Seq. Clinical and demographic data include patient age at diagnosis, gender, event free survival, vital status, dates of follow-ups and events, treatment protocols, trial protocols, therapeutic protocols, clinical tests that include pathology and genetic testing, and free comments. Clinical tests also include data on white blood cell counts at diagnosis, bone marrow leukemic blast percentage, detection of CNS disease, chloroma, the presence of common genetic alterations including translocations, deletions, and trisomies, and cytogenetic complexities.
- The dataset by Ploak et al. [12] includes outcome, subtype, and some RNA and protein expression profiles of biopsies from 654 ALL patients.

2.4 Medulloblastoma

International efforts to study medulloblastoma patients have already produced richly annotated datasets with clinical, demographic, and molecular profiles. The largest of these datasets are given below. Efforts to produce additional profiling types are ongoing by iPC collaborators and are expected to produce additional data in the near future.

- The dataset published by Northcott et al. [13] includes demographic and molecular data for 579 patients, with molecular profiles including whole-genome and whole-exome sequencing, RNA expression by RNA-Seq, and promoter methylation by Illumina microarrays.
- The dataset by Hovestadt et al. [14] includes demographic information and expression profiles for 284 patients.
- The dataset by Northcott et al. [15] includes demographic and molecular data for 212 patients, including focal deletions and amplification and other DNA alterations, as well as RNA expression, copy number, and methylation profiles.

2.5 Neuroblastoma

We have collected 17 datasets with biopsies from over 100 clinically and molecularly characterized neuroblastoma patients and tumors. The largest of these are outlined below. Multiple iPC partners are generating additional data, including single-cell RNA and protein profiles that are expected to be available for the project.

- The Therapeutically Applicable Research To Generate Effective Treatments (TARGET) neuroblastoma dataset includes data for 628 patients, including clinical annotations, RNA expression, copy number profiles, promoter methylation, and DNA alteration profiles including 122 patients with whole-genome profiles for both tumor and matched normal samples and an additional 222 patients with whole-exome profiles.
- The dataset generated by Kocak et al. [16] included demographic, clinical, and RNA expression profiles for tumor biopsies from 649 patients.
- The Sequencing Quality Control (SEQC) project generated expression profiles for biopsies from 498 clinically characterized patients. These data are unique because no patient-selection bias was known to be introduced. Other datasets are often enriched for poor-outcome patients or high-risk patients. Lack of bias makes this dataset ideal for studying molecular features that are common across the entire patient population.

Chapter 3 Other data to help construct models

We collected demographic, clinical, and molecular data from cancers outside of our focus group, as well as other data that is intended to help build models for regulation and perturbation-responses on the cellular, tissue, and patient levels. An outline of the datasets and types is given below. We also include an outline of first-order efforts to use these data to predict regulatory interactions.

3.1 Other paediatric tumor profiles

We collected a total of 19 large-scale paediatric cancer datasets with patient demographic and clinical data as well as molecular tumor profiles outside of our stated 5 tumor types of interest. These total data for over 4,500 paediatric cancer patients. The largest and most comprehensive of these profiles was assembled by The Pediatric Brain Tumor Consortium (PBTC). PTBC has enrolled nearly 1,900 paediatric brain-cancer patients, has rich clinical data for many of these patients including outcome, events, and treatment protocols, and over 1,000 of these patients have available profiled biopsies including DNA, RNA, and protein expression profiles. Multiple iPC groups are working on this dataset with several manuscripts nearing submission. Models based on these data predict response to therapies including chemotherapy and radiation, which is a standard high-toxicity treatment for high-grade gliomas.

3.2 Adult tumor profiles

We collected demographic, clinical, treatment, and molecular profiles for over 350,000 tumor biopsies. TCGA tumors comprise some of the largest of these datasets, and we briefly outline some of these datasets, as an example, below. We used RNA- and microRNA (miRNA)-expression and copy number profiles of TCGA tumors from 32 types. RNA, including both mRNA, long non-coding RNA (lncRNA), and microRNA (miRNA) expression was profiled using RNA-Seq and miRNA-Seq, while copy numbers were estimated using SNP Arrays. All included tumor was profiled by each of these assays. The number of profiled tumors in the 10 largest multi-omics (DNA, RNA, microRNA) collections is given below. When available, tumor subtypes were obtained from TCGA phenotype descriptions. For example, BRCA subtypes for TCGA and METABRIC (collected but not discussed in this document) were based on PAM50 inference.

- Bladder urothelial carcinoma (BLCA): 251 tumors
- Breast invasive carcinoma (BRCA): 835 tumors
- Head and neck squamous cell carcinoma (HNSC): 423 tumors
- Kidney renal clear cell carcinoma (KIRC): 437 tumors
- Brain low grade glioma (LGG): 498 tumors
- Lung adenocarcinoma (LUAD): 488 tumors
- Ovarian serous cystadenocarcinoma (OV): 261 tumors
- Prostate adenocarcinoma (PRAD): 371 tumors
- Thyroid carcinoma (THCA): 502 tumors
- Uterine corpus endometrial carcinoma (UCEC): 309 tumors

In addition, when estimating gene-expression dysregulation, we compared the expression of a gene in tumor samples to tumor-adjacent normal samples. Coding genes and lncRNAs were identified as “expressed” if they had a nonzero median absolute deviation (MAD) score. The number of profiled tumor-adjacent samples for a selection of these tumor types is given below.

- Bladder Urothelial Carcinoma (BLCA): 19 tumor adjacent samples

- Breast invasive carcinoma (BRCA): 105 tumor adjacent samples
- Head and neck squamous cell carcinoma (HNSC): 42 tumor adjacent samples
- Kidney renal clear cell carcinoma (KIRC): 67 tumor adjacent samples
- Kidney renal papillary cell carcinoma (KIRP): 30 tumor adjacent samples
- Liver hepatocellular carcinoma (LIHC): 50 tumor adjacent samples
- Lung adenocarcinoma (LUAD): 58 tumor adjacent samples
- Prostate adenocarcinoma (PRAD): 52 tumor adjacent samples
- Thyroid carcinoma (THCA): 59 tumor adjacent samples

3.3 The Connectivity Map database

The Connectivity Map database [17] includes Luminex-based multiplexed assays to measure the expression of 1171 genes (L1000) in response to a variety of perturbations—including shRNAs, 164 drugs, and 19,811 small molecule drugs—in up to 86 cell lines. Gene expression was measured using the L1000 assay at multiple time points following each perturbation. In Connectivity Map perturbation assays, while some data points are missing due to quality control metrics, the expression of most genes was profiled in triplicates after perturbations. In total, Connectivity Map includes 1,319,138 L1000 profiles from 42,080 perturbagens (19,811 small molecule compounds, 18,493 shRNAs, 3,462 cDNAs, and 314 biologics), corresponding to 25,200 biological entities (19,811 compounds, shRNA and/or cDNA against 5,075 genes, and 314 biologics) for a total of 473,647 signatures (consolidating replicates). These data can be used to model cancer and non-cancer cell responses to perturbations.

3.4 Cancer Dependency Map (Depmap) datasets

The Depmap consortium is devoted to generating data to model cancer cell responses to biochemical, physical, and biological perturbations. It includes multiple smaller consortium to investigate gene essentiality, potential cancer targets, and cancer driver genes and alterations. As of June 2020, Depmap data included molecular—DNA, RNA, protein (<15% complete), and promoter methylation—and phenotypical characterizations of 1804 cancer cell lines, essentiality data for 769 cell lines that was obtained using CIRSPR and shRNA pools, and drug sensitivity data for nearly 5,000 compounds on over 1,300 cell lines. Nearly all iPC models rely heavily on these data to predict cancer responses to perturbations and therapies.

3.5 Non-coding RNA expression

Multiple iPC groups are interested in non-coding RNAs and their influence on tumor initiation, progression, and response to therapies. We collaborated with multiple consortia to gain early access to large-scale molecular data that would allow for including non-coding RNAs in predictive model development. The largest of these, RNA-Atlas, includes nearly 900 molecular profiles of cancer and non-cancer cells and tissues. RNA-Atlas [18] includes a library of annotated and predicted non-coding RNAs from multiple orthogonal sources, and over 6,000 previously uncharacterized RNA species. These data allow for modelling new transcriptional and post-transcriptional regulatory networks and help indicate which non-coding RNAs are expressed in cancer specific manner, are dysregulated in cancer, and alter key cancer pathways in paediatric and adult cancers.

3.6 Regulatory regions

Proximal Promoters and 3' UTR were used when predicting transcription factor (TF), RNA-binding protein (RBP), and miRNA binding sites. Binding site evidence across multiple promoters and 3' UTRs associated with the same gene was aggregated to produce gene-level binding evidence. We used 2kbps promoters: [-1000, 1000] relative to the transcription start sites.

When predicting TF binding sites in proximal promoters using position-weight matrices, motif scores were compared to 5'-flanking regions of length 2kbps of their cognate proximal promoters; the methodology is detailed at "TF-target prediction" section. When scoring TF binding sites in lncRNAs, comparisons were made relative to di-nucleotide preserved shuffled promoters. Binding sites for RBPs and miRNAs were identified in 3' UTRs, as evidence suggest that sites that are more likely to alter RNA stability and degradation are located in these regions [19]. Both 3' UTRs and proximal promoters were extracted based on hg19 RefSeq annotation. Note that there are 22388 proximal promoters and 38,669 3' UTRs corresponding to 17,792 PCGs. Their 3' UTR lengths were between 1 to 25,393bps with a median length of 999bps.

3.7 Verified TF-target interactions

Focusing on TFs and targets with profiles in TCGA RNASeqV2 data, we collected a total of 6,566 non-redundant and experimentally-verified human TF-target interactions for 557 TFs and 2528 targets from 3 sources; of these 388 have characterized motifs. Interactions were collected from the following sources:

- HTRIdb [20] build dating 03/20/2014: 2209 interactions involving 277 TFs and 1381 targets that were verified by small and mid-scale techniques. These excluded interactions detected by ChIP-chip or ChIP-seq due to their lower confidence.
- Table 3 of Whitfield et al. [21, 22] which included 63 interactions between 7 TFs and 54 target genes.
- TRANSFAC Professional [23] from February 2013, 4,888 interactions between 501 TFs and 1669 targets. We excluded interactions involving more than one TF per target to avoid non-specific binding by co-factors.

3.8 Verified miRNA-target interactions

miRNA-target interactions were compiled from miRecords, TarBase, TRANSFAC, and miRTarBase (v4.5 in 11/01/2013). Only human miRNA-target gene interactions with strong experimental evidence, i.e., reporter assay or western blot, were selected. In addition, we included validated targets from the Table S2 of Grosswendt et al. [24], which included interactions between 359 miRNAs and 2463 genes, where both were included in our TCGA profiles. In total, these 4,696 interactions were used to train classifiers and predict miRNA-target interactions genome wide.

3.9 Predicted interactions from ENCODE data

We used ENCODE [25] data to predict TF and RBP targets based on ChIP-Seq and eCLIP, including 108 TFs that were profiled in 37 cell lines, with the majority of assays performed in replicates. ChIP-seq data were downloaded from the UCSC genome browser, using hg19 annotation. Included eCLIP data profiled targets for 96 RBPs in 2 cell lines (HepG2 and K562), with each assay performed in duplicates. Transcription factor binding sites in proximal promoters and RBP sites in 3' UTRs were selected as sequence-based targets and used in the subsequent expression-based analysis.

3.10 Transcription factor binding motifs

In total, we collected 1634 position weight matrices (PWMs) for 642 human TFs with expression in TCGA RNASeqV2 from 5 sources. To avoid matrix entries of value 0, a pseudo count 1 was added to each entry before calculating the relative occurrence frequencies (%) of nucleotides at each position. We used this frequency table to scan TF binding sites from the proximal promoters. Sources include the following:

- JASPAR [26] version: 5.0_ALPHA: 104 PWMs for 100 TFs.
- SwissRegulon [27] downloaded on 03/18/2014: 353 PWMs for 340 TFs.
- HumanTF [28], downloaded from Table S3 in their paper: 661 PWMs for 365 TFs. Only higher-confidence motifs were included (motifs indicated in orange or green were not included).
- HOCOMOCO [29] version: 9.0: 430 PWMs for 402 TFs. Only motifs of quality A, B, C, or D were extracted.
- Factorbook [21], downloaded from Table S2 in their paper: 86 PWMs for 76 TFs. These excluded unannotated motifs in their publication.

PWMs were used to predict TFBS in proximal promoters, 5'-flanking regions, and lncRNA transcripts.

3.11 Cross-species conservation

Cross-species conservation estimates by phastCons [30] was used for predicting miRNA binding sites. Both complete hg19 human genome and genome-wide phastCons46way conservation scores for vertebrate were downloaded from UCSC Genome Browser annotation. All scores were normalized between 0 and 1.

3.12 Prediction of transcription factor targets

We predicted targets for 636 human TFs based on both sequence and expression evidence. First, each predicted TF-target was required to have significant binding evidence from either 751 ENCODE ChIP-seq[30,78] profiles or 1,618 human TF PWMs for 108 and 636 TFs, respectively. Second, we required each TF-target pair to exhibit significant co-expression pattern across RNA Atlas profiles samples.

ENCODE ChIP-seq data sets were profiled in 37 immortal cell lines and >60% of them are in K562 (n=121 for 61 TFs), GM12878 (n=113 for 64 TFs), HepG2 (n=97 for 51 TFs), A549 (n=67 for 27 TFs), and H1-hESC (n=62 for 36 TFs). More than one-third of TFs had at least 2 replicates in the same cell line. Human TF PWMs were collected from five sources including motifs annotated in Factorbook[79] (see Table S2 in their paper; n=86 for 76 TFs), motifs of quality A-D in HOCOMOCO v9[80] (n=427 for 395 TFs), high-confidence motifs in HumanTF[81] (see Table S3 in their paper; n=651 for 357 TFs), JASPAR[82] v5_alpha (n=103 for 99 TFs), and SwissRegulon[83] downloaded on 03/18/2014 (n=351 for 331 TFs). To avoid matrix entries of value 0, a pseudo-count 1 was added to each entry before calculating the relative occurrence frequencies of nucleotides at each position.

We interrogated each of 21,550 proximal promoters to see if there is a significant ChIP-seq peak (Q-value<1E-10) or PWM-based binding site ($P < 1x10^{-5}$). The significance of motif scores on either forward or reverse strand of the proximal promoters were compared to 5'-flanking regions of length 2kbps of their cognate proximal promoters using the CREAD[84,85] package. Binding site evidence across multiple promoters associated with the same gene were aggregated to produce gene-level binding evidence. For any protein-coding gene that satisfied this sequence-based constraint, we further required significant distance correlation (dCor)[86] at $P < 1x10^{-9}$, as calculated using expression profiles of their regulating TFs and cognate protein-coding targets profiled in

RNA Atlas. Note that only TFs and target genes of non-zero median absolute deviation (MAD) score were included for analysis. We applied permutation testing to estimate the significance of dCor by shuffling TF's expression 100K times and then calculated the randomized dCor values. These values were used to fit parameters for a generalized extreme value (GEV) distribution using the MATLAB gevfit routine to obtain a non-parametric p-value lower than 1E-5 from the cumulative density of the resulting GEV distribution. For TF-targets passed both sequence and expression constraints were investigated for transcriptional lncRNA modulation. We predicted 105,029 interactions between TFs and their protein-coding targets significantly modulated by lncRNAs. Moreover, 102,338 TF-target interactions had target transcripts of adequate exonic and intronic coverage to compute m/p-ratio profiles.

3.13 Prediction of miRNA- and RBP-targets

We predicted targets of both types of post-transcriptional regulators through a two-step approach by requiring both sequence- and expression-based evidence. Specifically, 3'-UTRs of protein-coding transcripts and whole lncRNA transcripts were scanned for miRNA binding sites conserved across species (context score < -0.2) by TargetScan[77] v6.0 and significant RBP binding peaks at $P < 1x10^{-10}$. ENCODE eCLIP[87] datasets for 115 RBPs profiled in two human cancer cell lines, i.e., K562 and HepG2, were downloaded from UCSC Genome Browser. Among them, 66 and 49 RBPs were available in either one or two cell lines, respectively. Each RBP-cell line pair was performed in duplicates. Binding site evidence across multiple 3' UTRs associated with the same gene were aggregated to produce gene-level binding evidence. We then asked if any pair of gene, either coding or non-coding, shared a significantly large common regulator program at adjusted pFET < 0.01. For each qualified gene pair and their common regulators, we measured if correlation changes between a common miRNA/RBP and any of these two genes had evidence for being modulated by lncRNA expressions using delta dCor; see the section "lncRNA target predictions using LongHorn." below. A pair of regulator-target significantly modulated by at least one lncRNA at P < 0.05 was finally selected. miRNA/RBP-targets that passed both sequence and expression constraints were investigated for post-transcriptional lncRNA modulation. In total, 102,963 predicted interactions between miRNAs and their protein-coding targets were significantly modulated by lncRNAs and, among them, 47,999 miRNA-target transcripts had adequate exonic, intronic, and m/p-ratio reads and could be included in further analyses to compare correlations of regulator and target mRNA and pre-mRNA expression profiles. Note that, similar to experimentally verified miRNA targets, each miRNA, including both miRBase-annotated and RNA Atlas-identified miRNAs, was required to be expressed in at least 20 RNA Atlas profiled samples. RBPs were required to have a non-zero MAD score.

Chapter 4 Conclusion

The availability of large-scale data is paramount to the construction of iPC models, and paediatric cancer datasets with rich demographic, clinical, and molecular profiles remain relatively few. However, the quantity of data that can be used to construct general models to inform paediatric cancer diagnosis therapeutic efforts is vast. Our approach has been to collect as much data as possible, including data from patients and models for our cancers of interests, other paediatric cancers, other cancers, and other cells and tissues. By leveraging related data, we are building models that will be refined using data for our cancers of interests. Here, we outlined datasets that were collected over 20 years through efforts by scientists all over the world. Due to advances in profiling and accounting methods, the speed of data curation continues to increase non-linearly. We expect to continue collecting public and private data throughout the project and will make this data available to all iPC groups and to the great public as early as possible.

REFERENCES

1. Cairo S, Armengol C, De Reynies A, Wei Y, Thomas E, Renard CA, Goga A, Balakrishnan A, Semeraro M, Gresh L, et al: **Hepatic stem-like phenotype and interplay of Wnt/beta-catenin and Myc signaling in aggressive childhood liver cancer.** *Cancer Cell* 2008, **14:**471-484.
2. Chavan RS, Patel KU, Roy A, Thompson PA, Chintagumpala M, Goss JA, Nuchtern JG, Finegold MJ, Parsons DW, Lopez-Terrada DH: **Mutations of PTCH1, MLL2, and MLL3 are not frequent events in hepatoblastoma.** *Pediatr Blood Cancer* 2012, **58:**1006-1007.
3. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, Carter SL, Cibulskis K, Hanna M, Kiezun A, et al: **The genetic landscape of high-risk neuroblastoma.** *Nat Genet* 2013, **45:**279-284.
4. Viprey VF, Gregory WM, Corrias MV, Tchirkov A, Swerts K, Vicha A, Dallorso S, Brock P, Luksch R, Valteau-Couanet D, et al: **Neuroblastoma mRNAs Predict Outcome in Children With Stage 4 Neuroblastoma: A European HR-NBL1/SIOPEN Study.** *J Clin Oncol* 2014, **32:**1074-1083.
5. Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC, Boca SM, Carter H, Samayoa J, Bettegowda C, et al: **The genetic landscape of the childhood cancer medulloblastoma.** *Science* 2011, **331:**435-439.
6. Postel-Vinay S, Véron AS, Tirode F, Pierron G, Reynaud S, Kovar H, Oberlin O, Lapouble E, Ballet S, Lucchesi C: **Common variants near TARDBP and EGR2 are associated with susceptibility to Ewing sarcoma.** *Nature genetics* 2012, **44:**323-327.
7. Savola S, Klami A, Myllykangas S, Manara C, Scotlandi K, Picci P, Knuutila S, Vakkila J: **High expression of complement component 5 (C5) at tumor site associates with superior survival in Ewing's sarcoma family of tumour patients.** *ISRN oncology* 2011, **2011**.
8. Volchenboum SL, Andrade J, Huang L, Barkauskas DA, Krailo M, Womer RB, Ranft A, Potratz J, Dirksen U, Triche TJ: **Gene expression profiling of E wing sarcoma tumours reveals the prognostic importance of tumour–stromal interactions: a report from the C hildren's O ncology G roup.** *The Journal of Pathology: Clinical Research* 2015, **1:**83-94.
9. Tirode F, Laud-Duval K, Prieur A, Delorme B, Charbord P, Delattre O: **Mesenchymal stem cell features of Ewing tumors.** *Cancer cell* 2007, **11:**421-429.
10. Carrillo-Reixach J, Torrens L, Simon-Coma M, Royo L, Domingo-Sàbat M, Abril-Fornaguera J, Akers N, Sala M, Ragull S, Arnal M: **Epigenetic footprint enables molecular risk stratification of hepatoblastoma with clinical implications.** *Journal of Hepatology* 2020.
11. Sumazin P, Chen Y, Treviño LR, Sarabia SF, Hampton OA, Patel K, Mistretta TA, Zorman B, Thompson P, Heczey A: **Genomic analysis of hepatoblastoma identifies distinct molecular and prognostic subgroups.** *Hepatology* 2017, **65:**104-121.
12. Polak R, Bierings MB, van der Leije CS, Sanders MA, Roovers O, Marchante JR, Boer JM, Cornelissen JJ, Pieters R, Den Boer ML: **Autophagy inhibition as a potential future targeted therapy for ETV6-RUNX1-driven B-cell precursor acute lymphoblastic leukemia.** *haematologica* 2019, **104:**738-748.
13. Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, Ehrenberger T, Gröbner S, Segura-Wang M, Zichner T, Rudneva VA: **The whole-genome landscape of medulloblastoma subtypes.** *Nature* 2017, **547:**311-317.
14. Hovestadt V, Remke M, Kool M, Pietsch T, Northcott PA, Fischer R, Cavalli FM, Ramaswamy V, Zapatka M, Reifenberger G: **Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour**

material using high-density DNA methylation arrays. *Acta neuropathologica* 2013, **125:**913-916.

15. Northcott PA, Nakahara Y, Wu X, Feuk L, Ellison DW, Croul S, Mack S, Kongkham PN, Peacock J, Dubuc A: **Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma.** *Nature genetics* 2009, **41:**465-472.
16. Kocak H, Ackermann S, Hero B, Kahlert Y, Oberthuer A, Juraeva D, Roels F, Theissen J, Westermann F, Deubzer H: **Hox-C9 activates the intrinsic pathway of apoptosis and is associated with spontaneous regression in neuroblastoma.** *Cell death & disease* 2013, **4:**e586-e586.
17. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK: **A next generation connectivity map: L1000 platform and the first 1,000,000 profiles.** *Cell* 2017, **171:**1437-1452. e1417.
18. Lorenzi L, Chiu H-S, Cobos FA, Gross S, Volders P-J, Cannoodt R, Nuytens J, Vanderheyden K, Anckaert J, Lefever S: **The RNA Atlas, a single nucleotide resolution map of the human transcriptome.** *bioRxiv* 2019**:**807529.
19. Hausser J, Syed AP, Bilen B, Zavolan M: **Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation.** *Genome research* 2013, **23:**604-615.
20. Bovolenta LA, Acencio ML, Lemke N: **HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions.** *BMC genomics* 2012, **13:**405.
21. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.** *Genome Res* 2012, **22:**1798-1812.
22. Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, Myers RM, Weng Z: **Functional analysis of transcription factor binding sites in human promoters.** *Genome biology* 2012, **13:**R50.
23. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34:**D108-110.
24. Grosswendt S, Filipchyk A, Manzano M, Klironomos F, Schilling M, Herzog M, Gottwein E, Rajewsky N: **Unambiguous identification of miRNA: target site interactions by different types of ligation reactions.** *Molecular cell* 2014, **54:**1042-1054.
25. Encode: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
26. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32:**D91-94.
27. Pachkov M, Erb I, Molina N, Van Nimwegen E: **SwissRegulon: a database of genome-wide annotations of regulatory sites.** *Nucleic acids research* 2007, **35:**D127-D131.
28. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152:**327-339.
29. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ: **HOCOMOCO: a comprehensive collection of human transcription factor binding sites models.** *Nucleic acids research* 2013, **41:**D195-D202.
30. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:**1034-1050.