# D1.2

## Collection of high-quality clinical and molecular paediatric cancer datasets as well as other tumour types

| | |
|---|---|
| **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** | 53 months |
| **Programme** | H2020-SC1-DTH-2018-1 |

| | |
|---|---|
| **Deliverable type** | R |
| **Deliverable reference number** | SC1-DTH-07-826121 / D1.2 / 1.0 |
| **Work package contributing to the deliverable** | WP1 |
| **Due date** | May 31st, 2021 (M29) |
| **Actual submission date** | May 28th, 2021 |

| | |
|---|---|
| **Responsible organisation** | BCM |
| **Editor** | Pavel Sumazin |
| **Dissemination level** | PU |
| **Revision** | V1.0 |

| | |
|---|---|
| **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)

## Contributors (ordered according to beneficiary numbers)

All iPC partners contributed to this deliverable

## 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 collected demographic, clinical, and molecular profiles for multiple paediatric and adult tumors. Paediatric tumors include hepatoblastomas, leukemias, sarcomas, neuroblastomas, and medulloblastomas. Other paediatric tumor types for which data was collected include gliomas, osteosarcomas, and ependymomas. In addition, D1.2 is emphasized collections of single-cell profiles from high-risk cancers, including neuroblastomas, leukemias, sarcomas, and hepatoblastomas. These datasets will be used by WP3-8 to evaluate the effects of treatments and perturbations targeting cancer cells and 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 cells. These data are particularly interesting because it will allow to characterize cancer cell types that are predictive of outcome and cell types that are resistant to therapies. D1.2 efforts also included a collection of data to molecularly profile cancer cell lines and evaluate cancer-cell responses to drugs and molecular perturbations. All of these data are being used in iPC to construct models of regulation and responses to perturbations by gene-therapy, immunotherapy, small molecules, and radiotherapy.

# 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 paediatric cancers [1-5]. The creation of these models depends on our ability to model cancer and normal cell responses to perturbations and potential treatments. Deliverable 1.2 is a collection of demographic, clinical, and molecular profiles for multiple paediatric and adult tumors 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. Demographic and clinical data include age, sex, treatments, outcomes, and clinical features that are unique to each tumor type, including AFP levels for hepatoblastomas and MYCN amplification status for neuroblastoma. 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 patients. Paediatric cancer-specific data include clinical and molecular profiles for our cancers of interest, including models of these cancers. 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. We note that paediatric cancers are relatively rare, and the collection and aggregation of these data are necessary for powering computational models and is therefore of great value on its own accord.

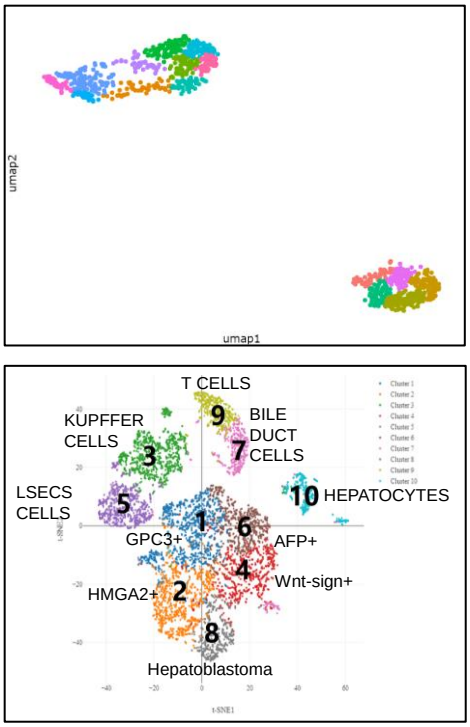


Figure 2. (Top) Single-cell copy number data revealed multiple genetic HB subclones and pre-cancer states. (Bottom) Single cell RNA profiles for the same HB revealed unique expression programs for multiple tumor subclones (subclones 1, 2, 4, 6, 8).

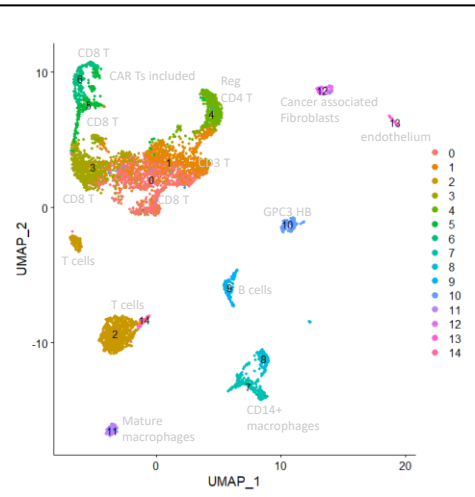


Figure 1. Single-cell RNA profiles of an HB 4 weeks following targeting by GPC3 CAR T cells revealed active CAR T cells and no GPC3-negative HBs. We enriched for immune cells to allow for CAR T discovery.

## 1.1 Data specific to our cancers of interest

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

**Ewing Sarcoma.** In total, we collected demographic, clinical, and molecular profiles for 319 Ewing sarcoma patients. In addition, single-cell and treatment data were collected from PDX models for 8 of these patients.

**Hepatoblastoma**. We collected demographic, clinical, and molecular profiles for over 400 hepatoblastoma (HB) and paediatric hepatocellular carcinoma patients. These include patients from 5 large studies, which include an ongoing international clinical trial (PHITT, clinicaltrialgov ref NCT03017326) 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. Single-cell DNA and RNA was

collected for 7 patients. These data demonstrated that high-risk hepatoblastomas contain multiple, genetically distinct tumor subclones; see figure 1 for an example. In addition, single-cell RNA was collected for 6 hepatoblastoma patients undergoing immunotherapy. For these patients, both product and post-infusion tumor samples were profiled (Figure 2).

**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. In addition, we collected single-cell protein and/or RNA data from bone marrows of 28 patients at both diagnosis and relapse. For 5 patients, samples were taken at remission when few cancer cells are expected, and profiled for single-cell protein by flow cytometry.

**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. In addition, single-cell RNA samples were collected for 37 patients.

## 1.2 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, including predominantly RNA expression and DNA alteration profiles. 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 molecular, demographic, and clinical 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 both publicly available and unpublished cancer datasets from iPC partners 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.

## 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 44 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 datsets have been collected by research centres around the world, including data from 3 iPC partners. 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, 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 short-RNA library 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 annotation, promoter methylation, and DNA alteration profiles including 122 patients with whole-genome profiles for both tumor and non-tumor 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 across the USA, 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 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 Predicted interactions from ENCODE data

We used ENCODE [17] data to predict TF and RBP targets based on ChIP-Seq and eCLIP, including 108 TFs that were profiled in 37 cell lines—including lines derived from a variety of adult and paediatric tumors—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. To improve our ability to map ENCODE perturbations to single-cell profiles, we profiled 32 cell lines for single-cell RNA expression. These data revealed that cell types identified by single-cell expression are valuable for predicting drug efficacy. For example, the proportion of HER2+ cell on the plate was predictive of ICU 50 values of cell line response to HER2 inhibitors.

# 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. In D1.2, we placed special emphasis on collecting single-cell RNA data to characterize the composition of cancers, and particularly high-risk tumors. We believe that decomposing tumors into the cancer-cell components will dramatically improve our ability to predict the efficacy of treatments on cancer models and in the clinic. 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. All of these data is stored and will be made available through the iPC portal, which is being produced in WP2. In addition, in future deliverables, we will report on data harmonization efforts that are focused on integrating molecular and clinical data across the datasets reported for each tumor type.

## 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. Encode: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.