Određivanje ishodišne stanice tumora na temelju raspodjele različitih tipova mutacija

Bakšić, Ivan

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Ivan Bakšić

Određivanje ishodišne stanice tumora na temelju raspodjele različitih tipova mutacija

Diplomski rad

Zagreb, 2022.

University of Zagreb
Faculty of Science
Department of Biology

Ivan Bakšić

Cell-of-origin determination based on the distribution of different mutation types in cancer

Master thesis

Zagreb, 2022

Ovaj rad je izrađen u Grupi za bioinformatiku na Zavodu za molekularnu biologiju Prirodoslovno-matematičkog fakulteta u Zagrebu, pod mentorstvom doc. dr. sc. Rose Karlić. Rad je predan na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja magistra molekularne biologije.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Diplomski rad

Određivanje ishodišne stanice tumora na temelju raspodjele različitih tipova mutacija

Ivan Bakšić

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Usporedbom tumorskih obrazaca mutacija s kromatinskim obilježjima stanica pojedinih tipova tkiva moguće je tumor povezati s izvornim tipom zdrave stanice iz koje je tumor nastao. U ovome radu proveo sam istraživanje na 722 uzorka melanoma i tumora jetre te na integracijskim mjestima hepatitis B virusa. Cilj je bio pomoću Random Forest regresijske analize odrediti ishodišnu stanicu tumora na temelju korelacije između kromatinskih obilježja zdravih tkiva i varijanti jednog nukleotida, insercija i delecija te virsunih integracijskih mjesta. Nadalje, u radu su provedena metodološka istraživanja kojima se htjelo postići poboljšanje u određivanju ihodišnih stanica i istražiti kako broj mutacija utječe na kvalitetu predviđanja. Dokazano je kako je frekvencija varijanti jednog nukleotida povećana u regijama otvorenog kromatina kod tumorskih stanica melanoma te da je moguće postići značajno povećanje snage predviđanja isključenjem stršećih vrijednosti. Također je utvrđeno da s podacima varijanti jednog nukleotida tumora jetre, insercija i delecija te virusnih integracijskih mjesta korištenim u ovome radu nije moguće pouzdano i točno odrediti ishodišnu stanicu te da značajno smanjenje broja mutacija neće nužno dovesti do smanjenja kvalitete predviđanja ishodišnje stanice.

Ključne riječi: ishodišna-stanica, tumor, melanom, HBV, VIS, Random-Forest
(46 stranica, 12 slika, 1 tablica, 42 literaturnih navoda, jezik izvornika: engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici

Mentor: doc. dr. sc. Rosa Karlić

Ocjenitelji:
prof. dr. sc. Dunja Leljak-Levanić
izv. prof. dr. sc. Inga Urlić
prof. dr. sc. Kristian Vlahoviček

Rad prihvaćen: 31. ožujka 2022.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology

Master thesis

Cell-of-origin determination based on the distribution of different mutation types in cancer

Ivan Bakšić

Rooseveltov trg 6, 10000 Zagreb, Croatia

It is possible to corelate tumor with the healthy tissue it originated by using comparison of mutational landscape with chromatin marks of the normal cells. In this work I carried out research on 722 melanoma and liver carcinoma samples and on hepatitis B virus integration sites. The goal of this research was to determine COO correlations between chromatin marks and single nucleotide variance, insertions and deletions and hepatitis B integration sites using the Random Forest regression analysis. Moreover, this work provides methodological research with the goal of improving cell-of-origin determination and observing the effect of different mutation number on the quality of the prediction. It was proven that the melanoma single nucleotide variance frequency is higher in open chromatin regions and that it was possible to considerably increase prediction power by outlier exclusion. It was also determined that liver carcinoma single nucleotide variance data, insertion and deletion data and virus integration site data used in this work do not provide accurate nor reliable cell-of-origin prediction. Finally, it was determined how usage of considerably lower number of mutations does not necessarily lower the prediction power of the cell-of origin.

Keywords: Cell-of-origin, Cancer, Melanoma, HBV, VIS, Random-Forest
(46 pages, 12 figures, 1 table, 42 references, original in: english)
Thesis is deposited in Central Biological Library.

Mentor: Asst. Prof. Rosa Karlić, PhD

Reviewers:
Prof. Dunja Leljak-Levanić, PhD
Assoc. Prof. Inga Urlić, PhD
Prof. Kristian Vlahoviček, PhD

Thesis accepted: March 31, 2022

TABLE OF CONTENTS

1. INTRODUCTION **1**

1.1 Cancer and cell-of-origin 1

1.2 Mutations 2

1.3 Cancer genomics 3

1.4 Random Forest 4

2. RESEARCH GOALS **5**

3. MATERIALS AND METHODS **6**

3.1 Reference genome 6

3.2 Tumor data 6

3.3 HBV integration sites data 6

3.4 Chromatin data 7

3.5 Random Forest regression analysis 10

3.6 Subset generation 11

4. RESULTS **12**

4.1 Using chromatin mark peak data results in accurate predictions of COO in control datasets 12

4.2 Analysis based on melanoma SNV mutations results in accurate COO prediction for certain datasets 17

4.3 Analysis based on liver carcinoma SNV mutations results in inaccurate COO prediction for all datasets 21

4.4 Analysis based on melanoma and liver carcinoma indel mutations results in inaccurate COO prediction for all datasets 26

4.5 Analysis based on HBV integration sites results in inaccurate COO prediction for all datasets 34

5. DISCUSSION **37**

6. CONCLUSIONS **40**

7. BIBLIOGRAPHY **41**

ABBREVIATIONS

CI – confidence interval

COO – cell-of-origin

HBV – hepatitis B virus

HCC – hepatocellular carcinoma

Indel – insertion and deletion

LICA-CN – Liver Cancer – CN

LIHC-US – Liver Hepatocellular carcinoma - TCGA, US

LIRI-JP – Liver Cancer - RIKEN, JP

Mb – megabase

MELA-AU – Skin Cancer – AU

NGS – next generation sequencing

SKCA-BR – Skin Adenocarcinoma – BR

SKCM-US – Skin Cutaneous melanoma - TCGA, US

SNV – single nucleotide variation

TCGA – The Cancer Genome Atlas

WMW – Wilcoxson-Mann-Whitney

VIS – virus integration site

1. INTRODUCTION

1.1 Cancer and cell-of-origin

Cancer incidence and mortality are rapidly growing worldwide; therefore, cancer research and treatment is emphasized more than ever. Sixth most commonly diagnosed cancer and the third leading cause of cancer death worldwide in the year 2020 was primary liver cancer, with approximately 906,000 new cases and 830,000 deaths (Sung et al., 2021). Primary liver cancer includes hepatocellular carcinoma (HCC) (comprising 75%-85% of cases) and intrahepatic cholangiocarcinoma (10%-15%), as well as other rare types. The main risk factors for HCC are chronic infection with hepatitis B virus or hepatitis C virus, aflatoxin-contaminated foods, heavy alcohol intake, excess body weight, type 2 diabetes, and smoking (London and Mcglynn, 2009). Similarly, skin melanoma was responsible for approximately 325,000 new cases and 57,000 deaths (Sung et al., 2021) with several risk factors, such as: ultraviolet radiation exposure, melanocytic nevi, family history and sun sensitivity (Gruber and Armstrong, 2009). Consequently, there have been numerous approaches in cancer research to better understand the disease and to ultimately cure it, identifying origin cell type being one of them.

Need for distinguishing the normal cell from which the tumor has derived (i.e., cell-of-origin) rose from the fact that despite the advancements in imaging and histology to segregate cancer, there has been slow improvements in determining clinically and molecularly distinct cancers (Gilbertson, 2011). Advancements in this field of study would therefore provide better prediction for different cancer treatment responses and prognoses. For example, basket trial study where inhibitor drug was used on mutated proteins present in various types of cancer showed different response rates between the cancer types (Hyman et al., 2018). This emphasizes the importance of accurate tumor origin determination in cancer therapy. Majority of cancer COO research on record is based on mouse models (Gilbertson, 2011; Köhler et al., 2017; Mu et al., 2015) with immunohistochemcal staining and imaging as a primary method for COO determination. However, in recent works notable success in cancer COO prediction was achieved with the use of bioinformatics tools and analysis. Namely, in the papers Polak et al. (2015), Kübler et al. (2019) and Ha et al. (2020) authors have demonstrated that for certain cancers COO can be accurately predicted based on distribution of mutations and epigenomic marks along the cancer genome. In Polak et al. (2015) paper the authors made proof-of-concept study in order to understand association between different

epigenomic features and mutagenesis in a cell type-specific manner. They compared mutation distribution of eight cancer genome types to 424 epigenetic features deriving from 106 different cell types from 45 different tissue types and determined the best predictors of local somatic mutation density, amongst other. Later in the paper Kubler et al. (2019) this method was further elaborated by increasing the number of cancer types from 8 to 32 and increasing the number of individual samples analysed from 173 to 2,550. Authors also studied cancers that arise in the same organ but manifest as distinct subtypes, in particular breast cancer group. Cancer subtype study was also performed in Ha et al. (2020), this time for primary liver cancer. This study confirmed role of chromatin marks associated with possible COOs in shaping mutational landscape of primary liver cancer and detected distinctive contribution of each COO in different subtypes of primary liver cancer.

1.2 Mutations

Mutations can be divided in germline mutations and somatic mutations. While germline mutations contribute to inherited genetic disease, somatic mutations do not contribute to future generations. Instead, they are one of the main causes of human disease, cancer amongst other. Some of the most common types of mutations are insertions and deletions, single nucleotide variants and transposable elements. Frequency of these mutations is not uniform throughout the human genome and genomic regions may differ in their mutation frequency between mutation types and cancers types (Lawrence et al., 2013). More precisely, there is substantial variation in the density of somatic mutations along the human genome at the scale of one megabase (Hodgkinson et al., 2012). Dominant influence on regional mutation rate variation in human somatic cells is chromatin organization in different cell types and in different time (Schuster-Böckler and Lehner, 2012). Chromatin structure and DNA accessibility is regulated on epigenomic level by the processes of DNA methylation and histone modification (Dunham et al., 2012). Histone remodelling can lead to nucleosome repositioning and decrease in access to DNA sequence newly bound to the histone (Cairns, 2007). On the other hand, open chromatin regions distinguished by DNA readily accessible to enzymes in the nucleus are characterised by increase in local DNA accessibility using nucleosome ejection, sliding or restructuring. Modifications responsible for increase in DNA access include methylations: H3K4me1 (associated with enhancers), H3K4me3 (associated with

promoters), H3K36me3 (associated with transcribed chromatin); and acetylations: H3K27ac and H4ac (both associated with enhancers and promoters). By contrast, histone modifications responsible for DNA repression include H3K27me3, H3K9me3 and significant decrease in histone modifications (Beisel and Paro, 2011). As a result, histone modifications affect the accumulation rate of different types of mutations (Don et al., 2013). For example, open chromatin regions are associated with higher rates of insertion, deletion and substitution, while mildly elevated deletion and substitution rates are located in closed chromatin regions, along with lower rates of insertion, deletion and substitution. Apart from mutation rates, chromatin organization also influences the density of viral integration sites (Mitchell et al., 2004). For instance, hepatitis B virus integration sites in the liver cancer cells preferentially occur in the regions of closed chromatin, contrary to normal hepatocyte genome where their frequency is higher in active chromatin areas (Hama et al., 2018). This suggests that cells in which vital genes are disrupted by viral integration may be subject to negative selection.

1.3 Cancer genomics

Cancer genomics is the study of the totality of DNA sequence and gene expression differences between tumour cells and normal host cells. It aims to understand the genetic basis of tumour cell proliferation and the evolution of the cancer genome under mutation and selection by the body environment, the immune system and therapeutic interventions. In order to commercially afford totality of DNA sequence and use it as a clinical tool, massive parallel sequencing technologies, also named next generation sequencing, had to be developed (Goodwin et al., 2016). NGS refers to any of several high-throughput approaches to DNA sequencing using the concept of massively parallel sequencing of clonally amplified or single DNA molecules that are spatially separated in a flow cell (Voelkerding et al., 2009). Depending on the platform, NGS yields hundreds of megabases to gigabases of nucleotide sequence output in a single instrument run and subsequent data processing is performed to obtain consensus nucleotide sequence and to detect nucleotide variants. One of the many uses of NGS is in the research of DNA and protein interaction, named ChIP-sequencing or ChIP-seq. This method consists of chromatin immunoprecipitation where DNA and protein complexes are immunoprecipitated and disassembled, followed by sequencing of resulting DNA fragments using NGS. ChIP-seq method is used in determining genome locations

that various histone modifications are associated with by utilising antibodies that specifically bind to certain types of histone modifications during immunoprecipitation process (Collas, 2010).

Few of the biggest cancer genomic projects are the following: The Cancer Genome Atlas project, The International Cancer Genome Consortium and The Pan-Cancer Analysis of Whole Genomes. The Cancer Genome Atlas is a project to analyse human tumors with the goal to catalogue molecular aberrations responsible for the carcinogenesis at the DNA, RNA, protein and epigenetic levels (Weinstein et al., 2013). The International Cancer Genome Consortium is a global initiative to build a comprehensive catalogue of mutational abnormalities in the major tumor types (Hudson et al., 2010). It incorporates data from The Cancer Genome Atlas and the Sanger Cancer Genome Project. The Pan-Cancer Analysis of Whole Genomes study is an international collaboration to identify common patterns of mutation in cancer whole genomes from the International Cancer Genome Consortium (Campbell et al., 2020).

1.4 Random Forest

Random Forest is non-parametric supervised machine learning method. It can be used for regression analysis when response variable is numeric or classification analysis when response variable is categorical. Basic steps in the regression method include generating multiple datasets by randomly selecting *n* observations with replacement and building regression tree for each of the dataset. Prediction value of a numeric response variable is obtained by passing the explanatory variables through each tree model and combining model results (Breiman, 2001). The use of multiple regression trees reduces the risk of over-fitting and makes the method robust to outliers and noise in the input data.

Researched correlation between epigenetic marks and different types of mutation and VIS presents potential for COO determination. Therefore, in this work I explored indel and SNV mutations in melanoma and liver cancer sets, alongside HBV integration sites and determined their connection with epigenomic marks of normal cells which resulted in COO prediction for different cancer types.

2. RESEARCH GOALS

Using Random Forest regression, I performed various analysis on SNV, indel and VIS datasets founded on the hypothesis that mutations and VIS can be used for COO prediction based on correlation with epigenomic markers with the goal of accurate COO prediction for each tumor type. My next goal was to compare the prediction results with each other in order to conclude which mutation type provides the highest quality of the mutation density prediction. I also hypothesised that SNV mutations positively correlate with chromatin markers characteristic for closed chromatin and negatively correlate with chromatin marks typical for open chromatin. I presumed opposite for indel and VIS. Therefore, I used Spearman's correlation coefficient measure to determine the correlation between mutations, that is VIS, and the epigenomic marks. Another premise was that of the influence of outlier exclusion on regression model prediction power. The goal was to increase the prediction power of datasets that did not provide accurate COO by excluding certain and all outliers from the datasets. Finally, I hypothesised that the low prediction power might be the result of low mutation number of the researched dataset, hence I performed probability proportional to size sampling on the datasets that made accurate COO predictions in order to observe how low number of the mutations affects the COO prediction.

3. MATERIALS AND METHODS

3.1 Reference genome

I downloaded human genome (hg19) provided by Karimzadeh et al. (2018) and excluded regions with fraction of uniquely mappable 36-mers lower than 0.92. I also excluded centromeric and telomeric regions downloaded using UCSC Table Browser (Karolchik et al., 2004) and divided resulting regions into 1 Mb widows. Produced reference genome had 2,120 Mb windows.

3.2 Tumor data

I obtained whole genome mutation data from ICGC DCC Data Portal Release 28 (Zhang et al., 2019) for 722 cancer genome samples belonging to two cancer types: melanoma and hepatocarcinoma. Melanoma data contained three separate datasets: SKCM-US, MELA-AU and SKCA-BR, while hepatocarcinoma contained several different datasets of which I selected TCGA dataset LIHC-US and two additional datasets with the most amount of mutation, namely Liver LICA-CN and LIRI-JP. From collected data I filtered out single based substitution mutation type (SNV) and small indel (<= 200 bp) mutation type separately. I additionally filtered data belonging to whole genome sequencing strategy. I counted the number of mutations in each reference genome window for each dataset and mutation type using “any” condition in overlap count.

3.3 HBV integration sites data

I downloaded HBV integration sites from VISDB database (Tang et al., 2020) and ViMIC database (Wang et al., 2020). I filtered sites obtained from VISDB database by “Tumor” Sample type and “GRCh37,” “GRCh37.55” and “GRCh37/hg19” Reference Human Genome type. I used liftover from hg38 to hg19 on “Tumor” Sample integration sites that belonged to “GRCh38” Reference Human Genome type. Next, I assigned Begin Location as End Location to all the integration sites that did not contain End Location and vice versa. This resulted in 9,799 integration sites. In addition, I filtered data from ViMIC database by “Tumor”, “Tumor ” and “tumor” Sample type. I applied liftover from hg38 to hg19 to the filtered data. This resulted in 14,588 integration sites. I combined generated integration sites from VISDB and ViMIC databases and counted the number

of integration sites in reach reference genome window using "any" condition in overlap count which resulted in the sum of 16,905 integration sites when aligned to reference genome.

3.4 Chromatin data

I gathered chromatin data from DeepBlue epigenomic data server (Albrecht et al., 2016) in a form of peaks using DeepBlueR package (Albrecht et al., 2017). I downloaded data for six active histone modifications (H3K27ac, histone H3 lysine 27 acetylation; H3K27me3, histone H3 lysine 27 trimethylation; H3K36me3, histone H3 lysine 36 trimethylation; H3K4me1, histone H3 lysine 4 monomethylation; H3K4me3, histone H3 lysine 4 trimethylation; H3K9ac, histone H3 lysine 9 acetylation) and one repressive histone modification (H3K9me3, histone H3 lysine 9 trimethylation). Projects with provided peak data were ENCODE and Roadmap Epigenomics. There were no available DEEP (IHEC) project peak data on DeepBlue epigenomic data server. After filtering only primary cell, primary tissue, primary cell tissue cell types and removing all the cell types containing data for only one histone modification, I was left with 15 ENCODE cell types and 89 Roadmap Epigenomics cell types. For each histone modification I counted the number of peaks in each reference genome window, considering any type of overlap between two ranges. Certain cell types had biological replicates available and I considered them as one: fibroblast primary cells, keratinocytes, T helper memory cells, keratinocytes primary cells, melanocyte primary cells, T helper naive cells and rectal mucosa. Furthermore, I formed 11 groups based on histological relationship between certain cell types: 'Blood – other', consisting of peripheral blood mononuclear cells excluding T cells; 'Bone/Soft tissue', consisting of adipose, bone, mesenchymal and muscle cells; 'Brain', consisting of cells originated from different parts of the brain and neurospheres; 'Breast', consisting of myoepithelial and mammary epithelial cells; 'CD19', consisting of B cells; 'CD34', consisting of hematopoietic stem cells; 'Colorectal Mucosa', consisting of colon and rectal mucosa cells; 'Gastrointestinal', consisting of stomach, large intestine and small intestine cells; 'Heart', consisting of cells originated from different parts of the heart, 'Squamous', consisting of keratinocytes and epithelial cells; 'T cells', consisting of all the T cells. I also grouped cell types with their fetal counterparts if there were any. This includes: Muscle cells, Brain cells, Intestine cells, Heart cells, Lung cells and Thymus cells. Finally, I grouped

Stomach Mucosa and Gastric cell types into Stomach Mucosa group. This resulted in 29 distinct cell groups (**Table 1**).

Table 1. List of chromatin marks originated from normal cell types. Chromatin mark peak data was obtained for each cell type using DeepBlue database id. Cell types were grouped by their histological relationship. Abbreviated cell type names are used in the work.

DeepBlue id	Project	Cell type name	Cell type abbreviated name	Cell type group
s2802	Roadmap	Fetal Adrenal Gland	Fetal Adrenal Gland	Adrenal
s7317	ENCODE	astrocyte	Astrocyte	Astrocyte
s2694	Roadmap	Primary monocytes from peripheral blood	Primary monocytes (PB)	Blood - other
s2695	Roadmap	Primary neutrophils from peripheral blood	Primary neutrophils (PB)	Blood - other
s2711	Roadmap	Primary Natural Killer cells from peripheral blood	Primary Natural Killer cells (PB)	Blood - other
s2726	Roadmap	Primary mononuclear cells from peripheral blood	Primary mononuclear cells (PB)	Blood - other
s2785	Roadmap	Monocytes CD14+ Primary Cells	Monocytes CD14+ Primary Cells	Blood - other
s6987	ENCODE	mononuclear cell	Mononuclear Cell	Blood - other
s7029	ENCODE	CD14+ monocyte	CD14+ Monocyte	Blood - other
s2691	Roadmap	Bone Marrow Derived Cultured Mesenchymal Stem Cells	Bone Marrow Derived Cultured MSCs	Bone/Soft tissue
s2714	Roadmap	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells	Stem Cell Derived Chondrocytes	Bone/Soft tissue
s2717	Roadmap	Muscle Satellite Cultured Cells	Muscle Satellite Cultured Cells	Bone/Soft tissue
s2790	Roadmap	Osteoblast Primary Cells	Osteoblast Primary Cells	Bone/Soft tissue
s2800	Roadmap	Adipose Derived Mesenchymal Stem Cell Cultured Cells	Adipose Derived MSC Cultured Cells	Bone/Soft tissue
s2738	Roadmap	Colon Smooth Muscle	Colon Smooth Muscle	Bone/Soft tissue
s2740	Roadmap	Duodenum Smooth Muscle	Duodenum Smooth Muscle	Bone/Soft tissue
s2750	Roadmap	Fetal Muscle Trunk	Fetal Muscle Trunk	Bone/Soft tissue
s2751	Roadmap	Fetal Muscle Leg	Fetal Muscle Leg	Bone/Soft tissue
s2761	Roadmap	Psoas Muscle	Psoas Muscle	Bone/Soft tissue
s2764	Roadmap	Rectal Smooth Muscle	Rectal Smooth Muscle	Bone/Soft tissue
s2768	Roadmap	Skeletal Muscle Male	Skeletal Muscle Male	Bone/Soft tissue
s2769	Roadmap	Skeletal Muscle Female	Skeletal Muscle Female	Bone/Soft tissue
s2772	Roadmap	Stomach Smooth Muscle	Stomach Smooth Muscle	Bone/Soft tissue
s2801	Roadmap	Adipose Nuclei	Adipose Nuclei	Bone/Soft tissue
s7314	ENCODE	osteoblast	Osteoblast	Bone/Soft tissue
s7322	ENCODE	skeletal muscle myoblast	Skeletal Muscle Myoblast	Bone/Soft tissue
s2718	Roadmap	Cortex derived primary cultured neurospheres	Cortex derived neurospheres	Brain
s2719	Roadmap	Ganglion Eminence derived primary cultured neurospheres	Ganglion Eminence derived neurospheres	Brain
s2729	Roadmap	Brain Angular Gyrus	Brain Angular Gyrus	Brain
s2730	Roadmap	Brain Anterior Caudate	Brain Anterior Caudate	Brain
s2731	Roadmap	Brain Cingulate Gyrus	Brain Cingulate Gyrus	Brain
s2732	Roadmap	Brain Germinal Matrix	Brain Germinal Matrix	Brain
s2733	Roadmap	Brain Hippocampus Middle	Brain Hippocampus Middle	Brain
s2734	Roadmap	Brain Inferior Temporal Lobe	Brain Inferior Temporal Lobe	Brain
s2735	Roadmap	Brain_Dorsolateral_Prefrontal_Cortex	Brain Dorsolateral Prefrontal Cortex	Brain
s2736	Roadmap	Brain Substantia Nigra	Brain Substantia Nigra	Brain
s2742	Roadmap	Fetal Brain Male	Fetal Brain Male	Brain
s2743	Roadmap	Fetal Brain Female	Fetal Brain Female	Brain
s2692	Roadmap	Breast Myoepithelial Primary Cells	Breast Myoepithelial Primary Cells	Breast
s2693	Roadmap	Breast variant Human Mammary Epithelial Cells (vHMEC)	Breast vHMEC Mammary Epithelial	Breast
s7310	ENCODE	mammary epithelial cell 01	Mammary Epithelial Cell 1	Breast
s6788	ENCODE	mammary epithelial cell 02	Mammary Epithelial Cell 2	Breast
s2696	Roadmap	Primary B cells from cord blood	CD19 - Primary B cells (CB)	CD19
s2697	Roadmap	Primary B cells from peripheral blood	CD19 - Primary B cells (PB)	CD19
s6684	ENCODE	B cell 01	B cell 1	CD19
s2700	Roadmap	Primary hematopoietic stem cells	CD34 - Primary HSC	CD34
s2701	Roadmap	Primary hematopoietic stem cells short term culture	CD34 - Primary HSC short term culture	CD34
s2715	Roadmap	Primary hematopoietic stem cells G-CSF-mobilized Female	CD34 - Primary HSC G-CSF-mobilized Female	CD34
s2716	Roadmap	Primary hematopoietic stem cells G-CSF-mobilized Male	CD34 - Primary HSC G-CSF-mobilized Male	CD34
s2737	Roadmap	Colonic Mucosa	Colonic Mucosa	Colorectal Mucosa
s2762	Roadmap	Rectal Mucosa Donor 29	Rectal Mucosa 29	Colorectal Mucosa

s2763	Roadmap	Rectal Mucosa Donor 31	Rectal Mucosa 31	Colorectal Mucosa
s2739	Roadmap	Duodenum Mucosa	Duodenum Mucosa	Duodenum Mucosa
s2745	Roadmap	Fetal Intestine Large	Fetal Intestine Large	Gastrointestinal
s2746	Roadmap	Fetal Intestine Small	Fetal Intestine Small	Gastrointestinal
s2753	Roadmap	Fetal Stomach	Fetal Stomach	Gastrointestinal
s2767	Roadmap	Sigmoid Colon	Sigmoid Colon	Gastrointestinal
s2770	Roadmap	Small Intestine	Small Intestine	Gastrointestinal
s2744	Roadmap	Fetal Heart	Fetal Heart	Heart
s2756	Roadmap	Left Ventricle	Left Ventricle	Heart
s2765	Roadmap	Right Atrium	Right Atrium	Heart
s2766	Roadmap	Right Ventricle	Right Ventricle	Heart
s2747	Roadmap	Fetal Kidney	Fetal Kidney	Kidney
s2728	Roadmap	Liver	Liver	Liver
s2749	Roadmap	Fetal Lung	Fetal Lung	Lung
s2757	Roadmap	Lung	Lung	Lung
s7319	ENCODE	fibroblast of lung 01	Lung Fibroblast 1	Lung Fibroblast
s6726	ENCODE	fibroblast of lung 04	Lung Fibroblast 4	Lung Fibroblast
s2724	Roadmap	Foreskin Melanocyte Primary Cells skin 01	Melanocyte 01	Melanocyte
s2725	Roadmap	Foreskin Melanocyte Primary Cells skin 03	Melanocyte 03	Melanocyte
s2758	Roadmap	Ovary	Ovary	Ovary
s2759	Roadmap	Pancreas	Pancreas	Pancreas
s2748	Roadmap	Pancreatic Islets	Pancreatic Islets	Pancreatic Islets
s2752	Roadmap	Placenta	Placenta	Placenta
s2760	Roadmap	Placenta Amnion	Placenta Amnion	Placenta Amnion
s2720	Roadmap	Foreskin Fibroblast Primary Cells skin 01	Skin Fibroblast 01	Skin Fibroblast
s2721	Roadmap	Foreskin Fibroblast Primary Cells skin 02	Skin Fibroblast 02	Skin Fibroblast
s7316	ENCODE	fibroblast of dermis	Dermis Fibroblast	Skin Fibroblast
s2774	Roadmap	Spleen	Spleen	Spleen
s2722	Roadmap	Foreskin Keratinocyte Primary Cells skin 02	Skin Keratinocyte 02	Squamous
s2723	Roadmap	Foreskin Keratinocyte Primary Cells skin 03	Skin Keratinocyte 03	Squamous
s2741	Roadmap	Esophagus	Esophagus	Squamous
s7323	ENCODE	keratinocyte 01	Keratinocyte 1	Squamous
s6963	ENCODE	bronchial epithelial cell	Bronchial Epithelial Cell	Squamous
s6779	ENCODE	kidney epithelial cell	Kidney Epithelial Cell	Squamous
s6805	ENCODE	keratinocyte 02	Keratinocyte 2	Squamous
s2755	Roadmap	Gastric	Gastric	Stomach Mucosa
s2771	Roadmap	Stomach Mucosa	Stomach Mucosa	Stomach Mucosa
s2698	Roadmap	Primary T cells from cord blood	Primary T cells (CB)	T cells
s2699	Roadmap	Primary T cells fromperipheralblood	Primary T cells (PB)	T cells
s2702	Roadmap	Primary T helper memory cells from peripheral blood 2	Primary Th memory cells (PB) 2	T cells
s2703	Roadmap	Primary T helper naive cells from peripheral blood	Primary Th naive cells (PB) 1	T cells
s2704	Roadmap	Primary T helper naive cells from peripheral blood	Primary Th naive cells (PB) 2	T cells
s2705	Roadmap	Primary T helper memory cells from peripheral blood 1	Primary Th memory cells (PB) 1	T cells
s2706	Roadmap	Primary T helper cells PMA-I stimulated	Primary Th cells PMA-I stimulated	T cells
s2707	Roadmap	Primary T helper 17 cells PMA-I stimulated	Primary Th 17 cells PMA-I stimulated	T cells
s2708	Roadmap	Primary T helper cells from peripheral blood	Primary Th cells (PB)	T cells
s2709	Roadmap	Primary T regulatory cells from peripheral blood	Primary Treg cells (PB)	T cells
s2710	Roadmap	Primary T cells effector/memory enriched from peripheral blood	Primary T cells effector/memory enriched (PB)	T cells
s2712	Roadmap	Primary T CD8+ naive cells from peripheral blood	Primary T killer naive cells (PB)	T cells
s2713	Roadmap	Primary T CD8+ memory cells from peripheral blood	Primary T killer memory cells (PB)	T cells
s2754	Roadmap	Fetal Thymus	Fetal Thymus	Thymus
s2773	Roadmap	Thymus	Thymus	Thymus
s2727	Roadmap	Aorta	Aorta	Vascular

3.5 Random Forest regression analysis

I used Random Forest regression to predict mutational density and VIS density based on chromatin profiles for every data set and subset. I created prediction using every chromatin profile of every type of cell separately, resulting in 104 models for each run. Every Random Forest regression was computed with 1,000 trees. I reported performance using 10-fold cross-validation. That is, I divided all windows into ten sets, trained the model on nine and made the prediction of the remaining set. I test the accuracy of the prediction by comparing it with the remaining set and calculating R^2. Finally, I determined average R^2 of all the predicted and compared sets and used it as a model performance measure. I compared model results between cell types by using paired Wilcoxon-Mann-Whitney test between the R^2 values from ten-fold cross-validation tests. Additionally, to compare models trained on histone marks from different cell types, I calculated the robust $\hat{z}$-score. I presented cell type with the highest average R^2 and significant difference in R^2 set values compared to the other cell types belonging to different cell type groups as supposed COO. I performed Random Forest regression analysis using caret (Kuhn, 2008) and ranger (Wright and Ziegler, 2017) packages.

I obtained control data from Kübler et al. (2019) by performing Random Forest regression on provided "Liver-HCC" and "Skin-Melanoma" SNV datasets. To prove the validity of chromatin mark peak use instead of the number of reads utilised in the work and to enable comparison between the results presented in this work and control results, I performed Random Forest regression based on control mutation window count data for each dataset, control reference genome and chromatin mark peak data.

For certain cancer types I visualised mutation density by plotting the predicted number of mutations in each 1Mb window against observed number of mutations in each 1Mb window. I additionally analysed correlations between chromatin mark density and mutational density and between chromatin mark density and VIS density using Spearman's rank correlation coefficient and visualised it as correlation coefficient versus cell type and as a mutation density versus each chromatin mark density.

3.6 Subset generation

In order to observe the effect how outlier influences the results, I filtered out the biggest outlier in SKCM-US SNV, SKCA-BR SNV, MELA-AU SNV, LIHC-US SNV, LICA-CN SNV, LIRI-JP SNV, LICA-CN indel and LIRI-JP indel data, top two outliers in SKCA-BR indel data and top 5 outliers in MELA-AU indel data. In addition, I made data subsets with all of the outliers excluded for each dataset and performed Random Forest regression analysis on all of the listed subsets.

I created subsets by randomly sampling tumor genome types of the datasets that accurately predicted COO and performed Random Forest regression in order to observe if the mutation number is responsible for inaccurate predictions of the other datasets. I sampled 10, 30 and 50 samples from SKCA-BR SNV dataset resulting in 2,048,075, 3,919,536 and 8,076,703 mutation sums and 5, 10 and 30 samples from MELA-AU SNV dataset resulting in 2,536,104, 4,324,187 and 12,369,354 mutation sums respectively. For the same reason I performed sampling and analysis with probability proportional to size of each window on the mentioned datasets where maximum number of samples was equal to the dataset that did not accurately predict COO.

All of the data processing, model building and graph displaying I performed using R program (R Core Team, 2021).

4. RESULTS

4.1 Using chromatin mark peak data results in accurate predictions of COO in control datasets

Random Forest regression based on control data (prediction based on number of reads in 1Mb windows) resulted in Melanocyte 03 COO prediction for Skin-Melanoma dataset with 81% variance explained and Liver COO prediction for Liver-HCC dataset with 87% variance explained (**Figure 1**). The cell type that fitted the model with the highest accuracy and significantly differed from the next-best histologically unrelated cell type was determined to be the COO. In both Skin-Melanoma and Liver-HCC data analysis Spearman's rank correlation coefficient was negative for H3K36me3, H3K4me1, H3K4me3 and Input chromatin marks and positive for H3K27me3 and H3K9me3 marks. Random Forest regression on a dataset consisting of control mutational density and chromatin mark peaks window count based on control reference genome, resulted in accurate predictions of COO (**Figure 2a**, **Figure 2c**), similar to those of control. The COO predictions were the same and there was no significant difference in COO variance explained when compared to the control results (p-value (Melanocyte 03) = 0.695, p-value (Liver) = 0.492, p-value < 0.03, WMW test). Spearman's rank correlation coefficient was positive for H3K27me3 and H3K9me3 marks and negative for H3K27ac, H3K36me3, H3K4me1 and H3K4me3 chromatin marks (**Figure 2b**). This proved the validity of chromatin mark peak use in COO determination and enabled me to compare the results presented in this work to the control results.

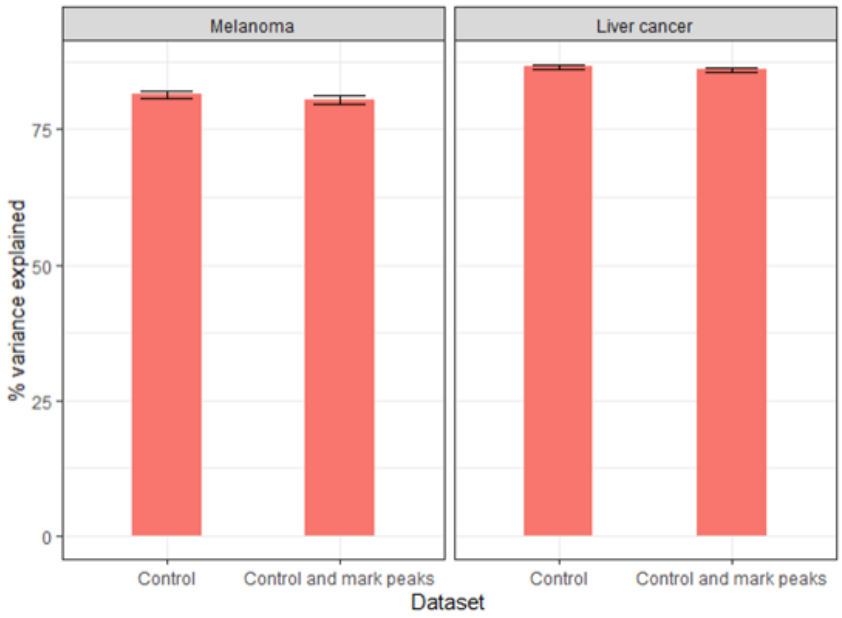

Figure 1. Prediction accuracy comparison between control models and models based upon control data mutations and peak data for melanoma and liver cancer. Models in "Control" group were built on control SNV mutation dataset and chromatin read data. Models in "Control and mark peaks" group were built on control SNV mutation dataset and chromatin peak data. Error bars represent standard errors of the mean prediction accuracy estimated using 10-fold cross-validation.

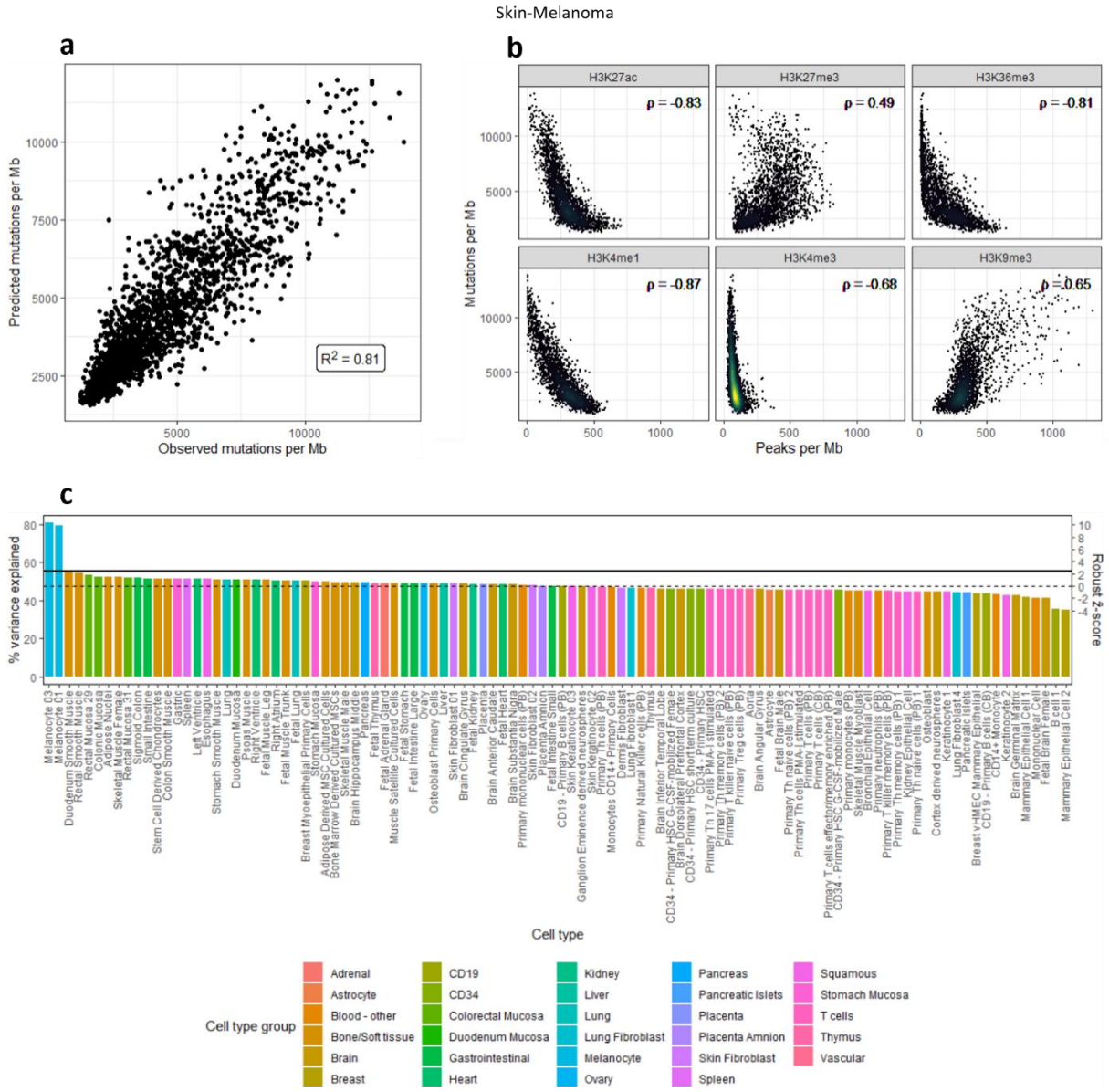

Figure 2. Analysis based on melanoma single nucleotide variation and chromatin peak data results in accurate cell-of-origin prediction. (**a, d, g**) Observed number of melanoma single nucleotide variations (SNV) in one megabase (Mb) windows versus the number of variations in one Mb windows predicted by 10-fold cross-validation Random Forest regression analysis based on chromatin peak data. Prediction accuracy is reported as R^2 value between the predicted and observed mutation numbers in reference windows across the 10-fold cross-validation. (**b, e, h**) Melanoma SNV per 1 Mb versus the number of peaks for each type of chromatin mark and their correlation described by the Spearman's rank correlation coefficient. (**c, f, i**) Model accuracy based on chromatin marks of each normal cell type. The cell type that fits the model with the highest prediction accuracy and significantly differs from the next-best histologically unrelated cell type is determined to be the cell-of-origin. Solid horizontal line describes variance explained reported by the next-best cell type model that belong to the different cell type group. Dotted horizontal line indicates median variance explained value of all the cell type models. SKCA – outliers excluded section additionally displays Spearman's rank correlation coefficient between mutation and chromatin marks for every cell type individually. Right-hand (secondary) y-axis shows robust ẑ-score. Bars are coloured by histological group.

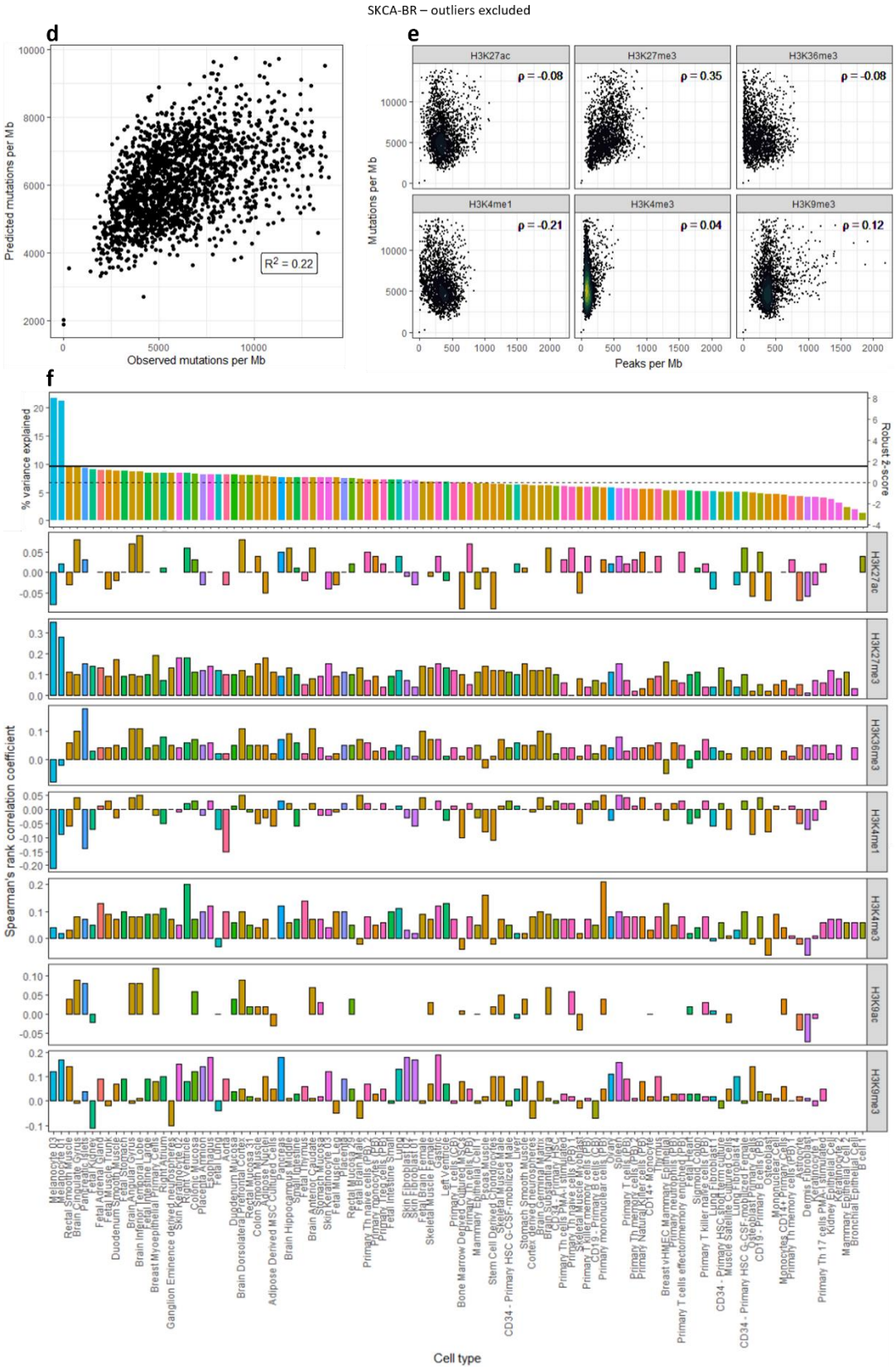

Figure 2. Continued

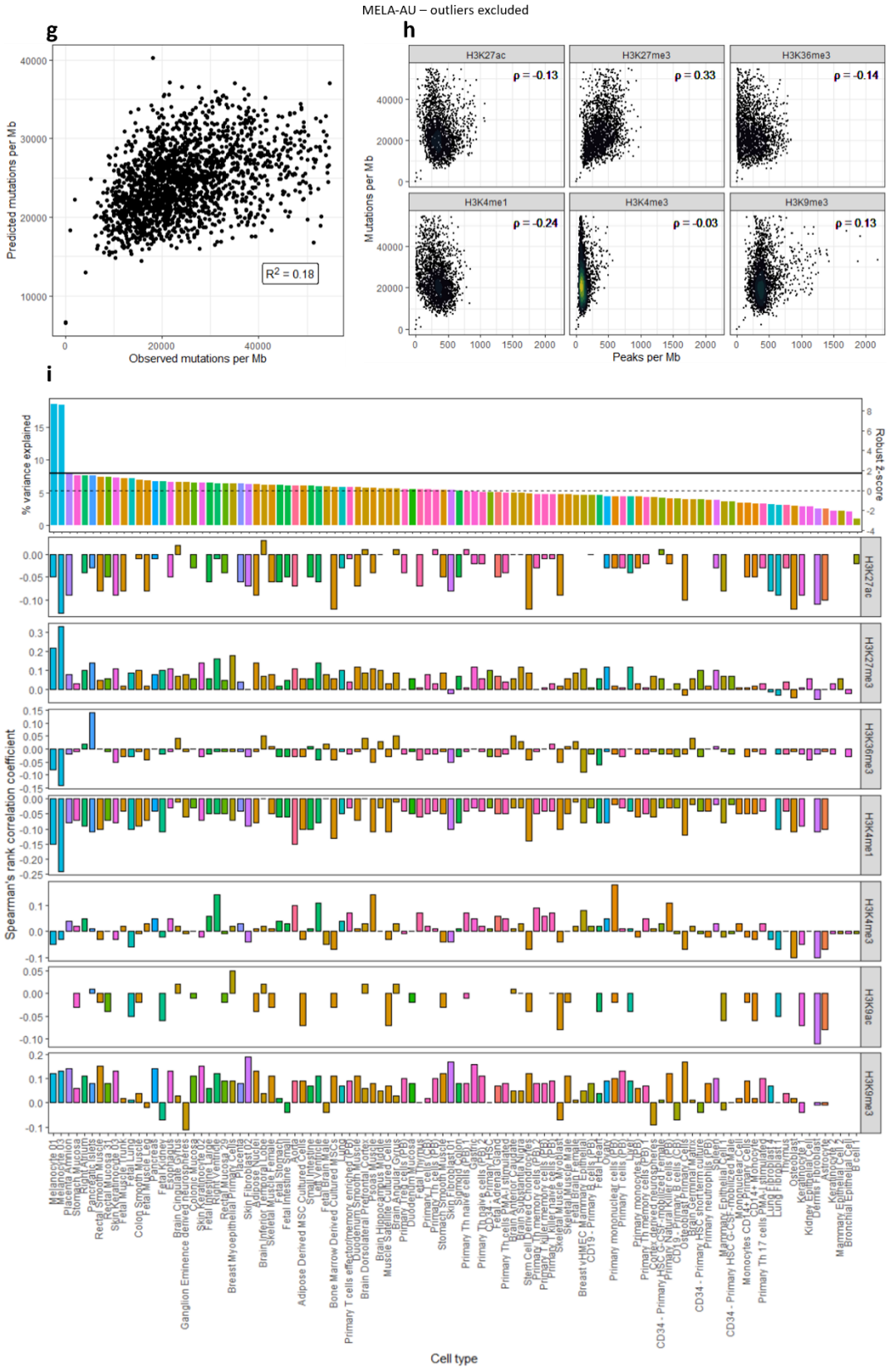

Figure 2. Continued

4.2 Analysis based on melanoma SNV mutations results in accurate COO prediction for certain datasets

Predicted COO based on SNV mutation type obtained from SKCA-BR and MELA-AU datasets was expected normal cellular counterpart. That is, Melanocyte 01 cell type explained significantly more variance than the next best histologically unrelated cell type (p-value = 0.0136, WMW test). By removing top outliers from datasets, there was no significant increase in model accuracy (p-value > 0.03 for all the cell types, WMW test) and by removing all of the outliers there was a significant increase in variance explained in 13 cell types for SKCA-BR dataset and in five cell types for MELA-AU dataset, especially in Melanocyte 03 (p-value = 0.00195, WMW test) cell type that became newly predicted COO in SKCA-BR dataset (**Figure 2d**, **Figure 2f**). Predicted COO for MELA-AU dataset did not change. Furthermore, predictive power in SKCA-BR and MELA-AU sets and outlier subsets ranges from 15.6% to 21.7% for top predictions, which differs from control case predictive power where the best model had 81% variance explained (**Figure 1**).

Analysis based on SKCM-US dataset did not result in accurate nor significant COO prediction, regardless of the outlier removal (**Figure 3a, Figure 3c**). Removing the top outlier resulted in significant increase of the variance explained in two cell types (p-value < 0.03 WMW test) and removing all outliers resulted in significant decrease in in nine cell types (p-value < 0.03 WMW test). Since SKCM-US dataset had significantly less mutations than the other two (**Figure 4**), I performed the analysis on SKCA-BR and MELA-AU sample subsets to determine whether the lower number of mutations was the reason for inaccurate COO prediction.

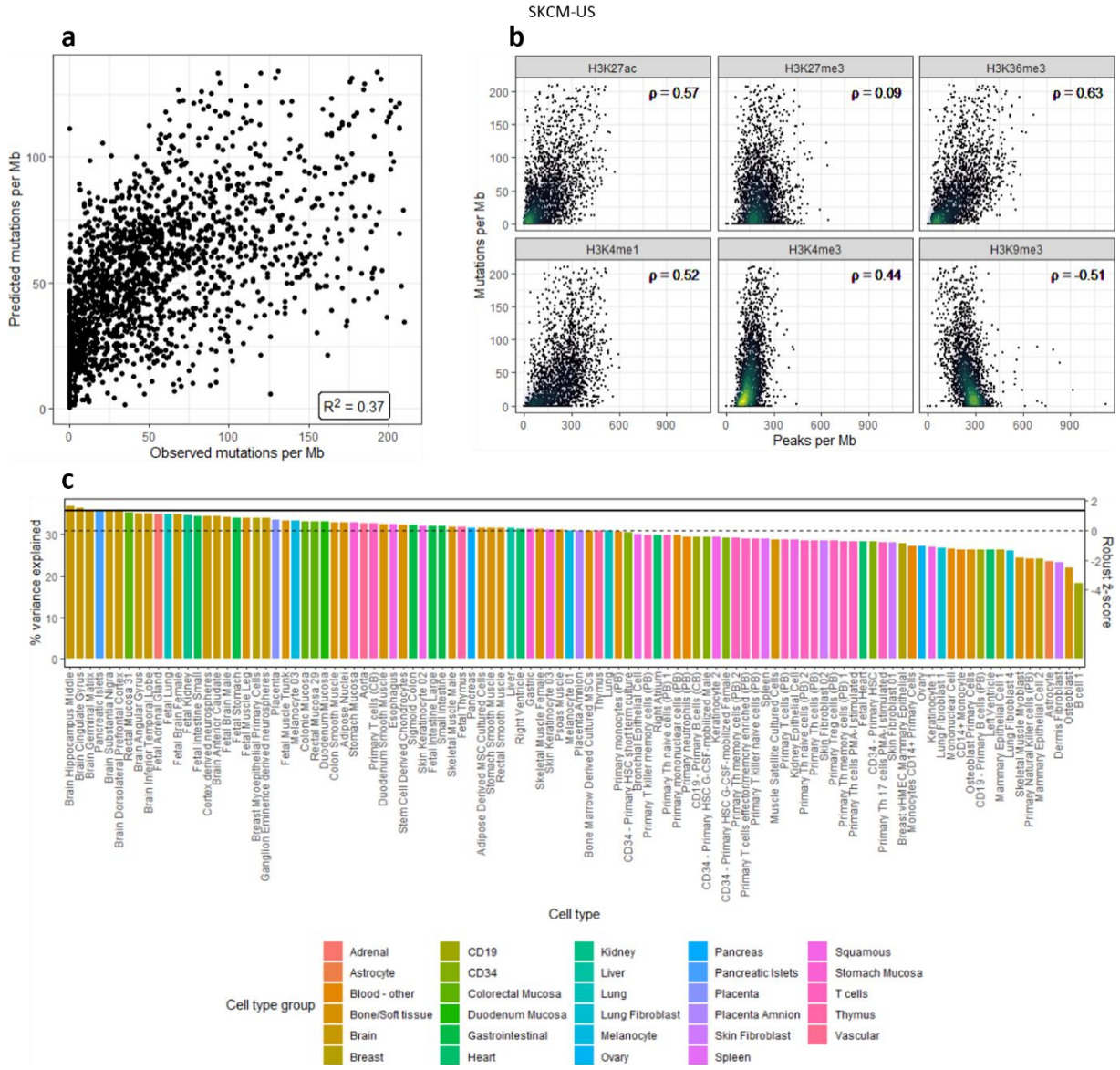

Figure 3. Analysis based on melanoma single nucleotide variation and chromatin peak data results in inaccurate cell-of-origin prediction. (**a**) Observed number of melanoma single nucleotide variations (SNV) in one megabase (Mb) windows versus the number of variations in one Mb windows predicted by 10-fold cross-validation Random Forest regression analysis based on chromatin peak data. Prediction accuracy is reported as R^2 value between the predicted and observed mutation numbers in reference windows across the 10-fold cross-validation. (**b**) Melanoma SNV per 1 Mb versus the number of peaks for each type of chromatin mark and their correlation described by the Spearman's rank correlation coefficient. (**c**) Model accuracy based on chromatin marks of each normal cell type. The cell type that fits the model with the highest prediction accuracy and significantly differs from the next-best histologically unrelated cell type is determined to be the cell-of-origin. Solid horizontal line describes variance explained reported by the next-best cell type model that belong to the different cell type group. Dotted horizontal line indicates median variance explained value of all the cell type models. Right-hand (secondary) y-axis shows robust $\hat{z}$-score. Bars are coloured by histological group.

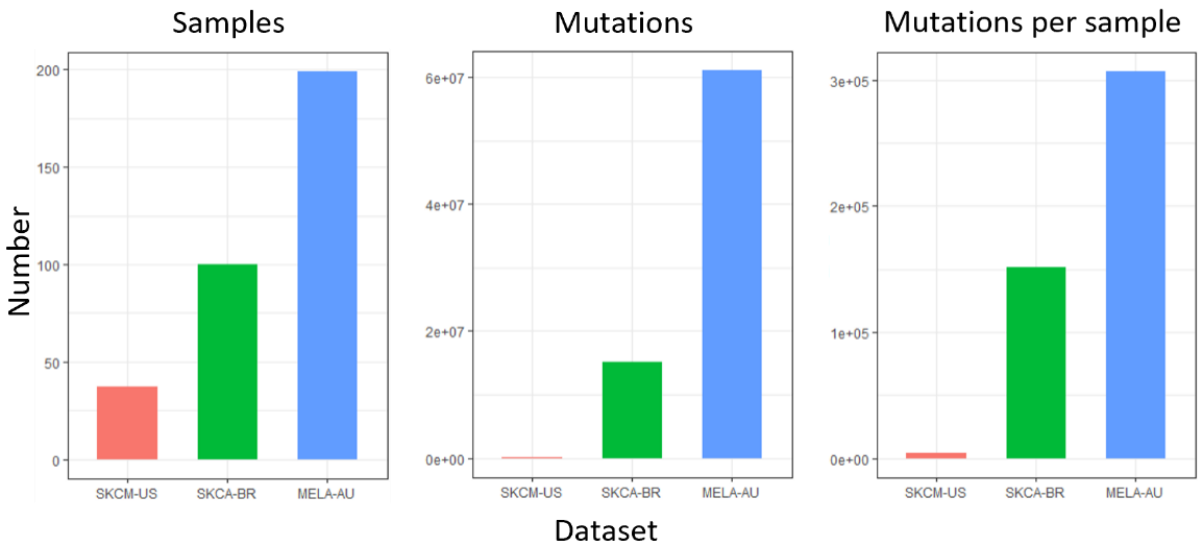

Figure 4. Melanoma single nucleotide variation datasets vary in sample and mutation numbers. "SKCM-US" stands for Skin Cutaneous melanoma - TCGA, US dataset, "SKCA-BR" stands for Skin Adenocarcinoma – BR dataset and MELA-AU stands for Skin Cancer – AU dataset.

All of the subset analysis resulted in accurate prediction, with no significant change in the variance explained for the top cell type group (*p*-value > 0.03, WMW test). Likewise, SKCA-BR and MELA-AU subsets scaled to mutational sum equal to the SKCM-US dataset also resulted in accurate prediction of COO when analysed, with no significant change in the variance explained for the top cell type group (*p*-value > 0.03, WMW test) (**Figure 5**).

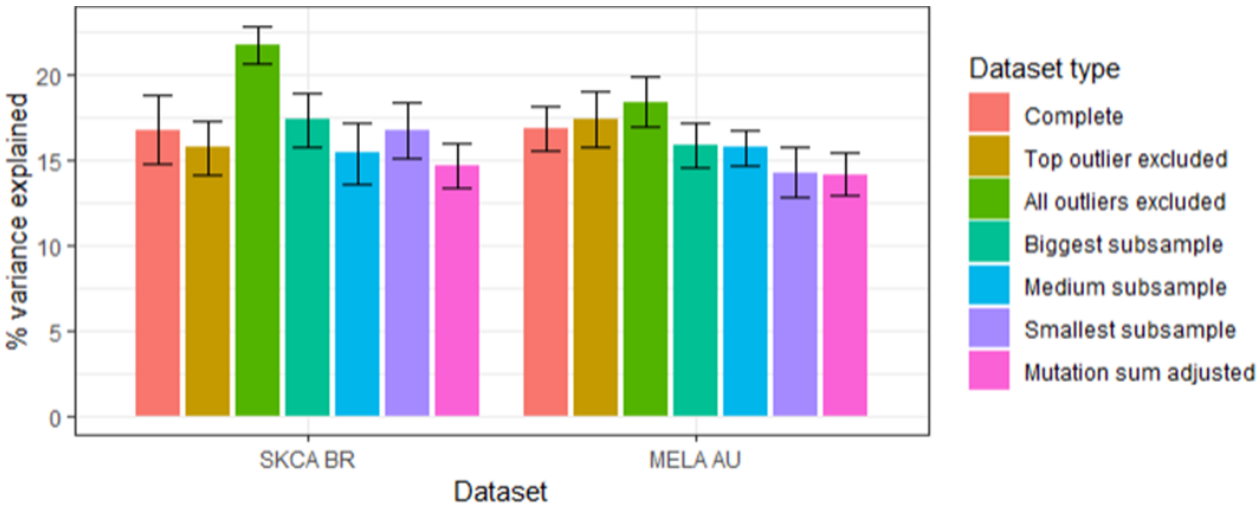

Figure 5. Prediction accuracy comparison between melanoma models that accurately predicted cell-of-origin. "Complete" dataset type refers to mutation datasets that did not had had any window count observation removed. "Biggest subsample", "Medium subsample" and "Smallest subsample" dataset types refer to Skin Adenocarcinoma – BR dataset build upon 50, 30 and 10 samples respectively and Skin Cancer – AU dataset build upon 30, 10 and 5 samples respectively. "Mutation sum adjusted" dataset type consists of datasets build by sampling mutations with probability proportional to size where mutation sum was equal to the dataset which model did not accurately predict COO. Error bars represent standard errors of the mean prediction accuracy estimated using 10-fold cross-validation.

By comparing Spearman's rank correlation coefficients between the control COO and the SKCM-US dataset, there was a discrepancy in correlation coefficients. SKCM-US had the same direction in relationship only between mutations and H3K27me3 epigenomic marker for the majority of the cell types (97/103). All the other mutual epigenomic markers had different correlation sign, including all of the H3K36me3 and H3K4me1 chromatin mark cell types and the majority of H3K4me3 (101/104) and H3K9me3 (85/99) chromatin mark cell types (**Figure 3b**). On the other hand, both SKCA-BR and MELA-AU datasets shared the same direction in relationship for majority (more than 50%) of cell types in all of the mutual chromatin marks, especially in the top two cell types: Melanocyte 01 and Melanocyte 03 (**Figure 2f**).

4.3 Analysis based on liver carcinoma SNV mutations results in inaccurate COO prediction for all datasets

LIHC-US, LICA-CN and LIRI-JP datasets and subsets all provided inaccurate and unsignificant (*p*-value > 0.03, WMW test) results when predicting COO (**Figure 6a**, **Figure 6b**, **Figure 6c**) regardless of the outlier removal, except for LICA-CN dataset with all of the outliers removed that had accurate but insignificant COO prediction (*p*-value = 0.846, WMW test). Removing the top outliers resulted in no significant change of variance explained for all the cell types in all of the datasets, while removing all of the outliers resulted in significant change of variance for 54 cell types of LIHC-US dataset, 77 cell types for LICA-CN dataset and 51 cell types for LIRI-JP dataset. Median variance explained decreased from 31.8% to 21.8% for LIHC-US dataset and increased from 2.8% to 7.5% for LICA-CN dataset and from 2.9% to 6.7% for LIRI-JP dataset (**Figure 7**).

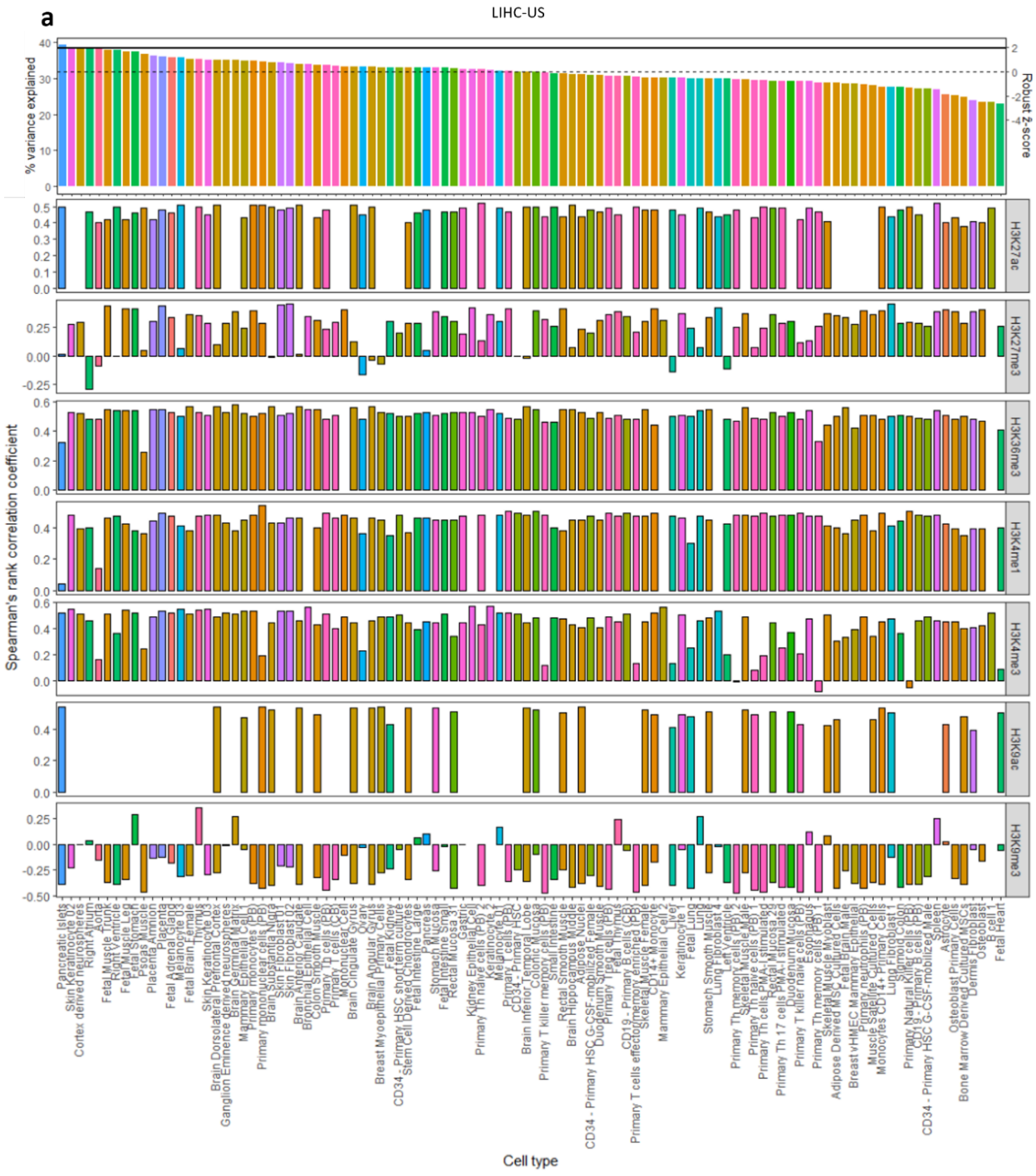

Figure 6. Analysis based on liver cancer single nucleotide variation and chromatin peak data results in inaccurate cell-of-origin prediction. (**a**, **b**, **c**) Model accuracy based on chromatin marks of each normal cell type. The cell type that fits the model with the highest prediction accuracy and significantly differs from the next-best histologically unrelated cell type is determined to be the cell-of-origin. Solid horizontal line describes variance explained reported by the next-best cell type model that belong to the different cell type group. Dotted horizontal line indicates median variance. Spearman's rank correlation coefficient between mutation and chromatin marks is reported for every cell type individually. Right-hand (secondary) y-axis shows robust $\hat{z}$-score. Bars are coloured by histological group.

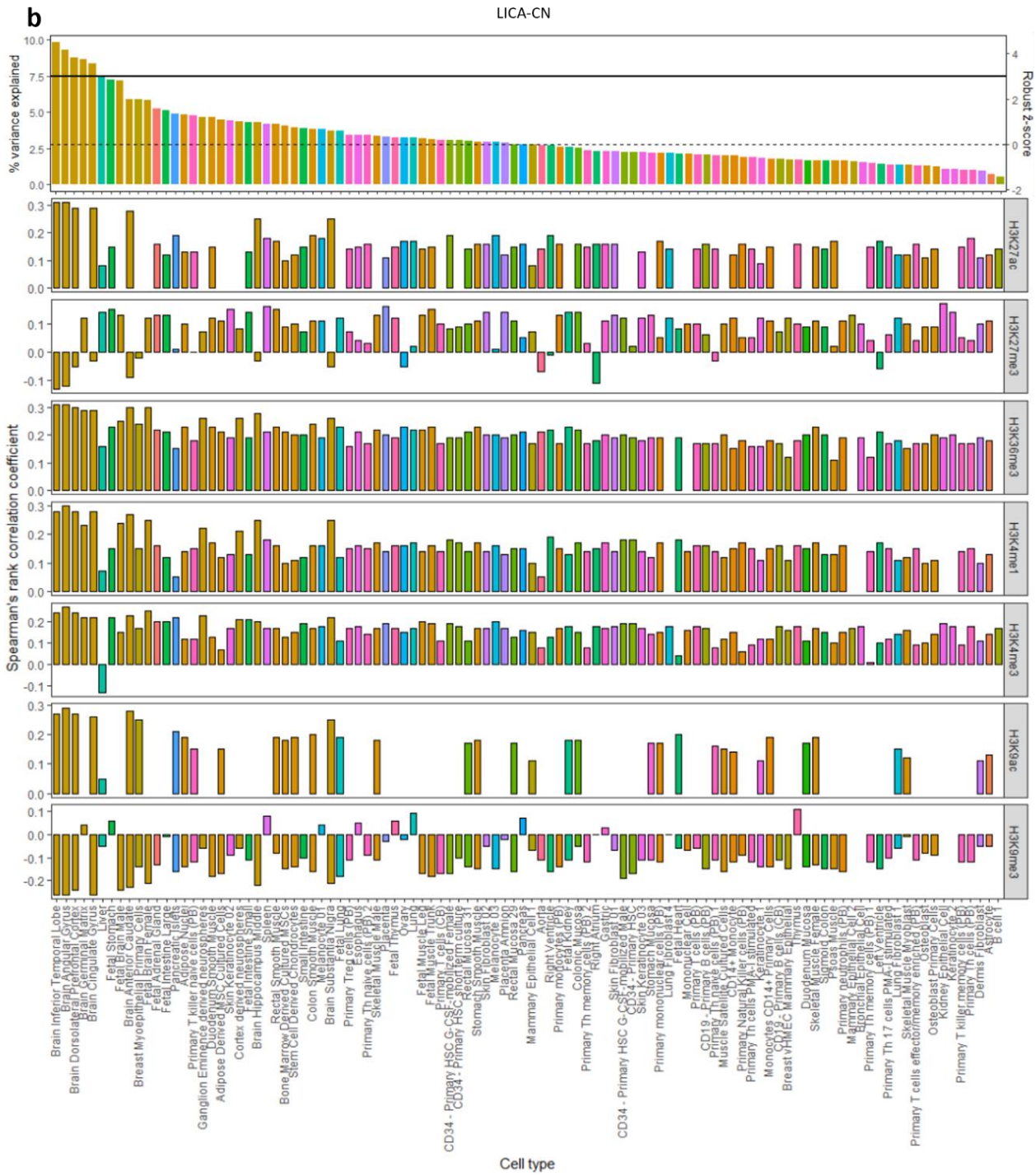

Figure 6. Continued

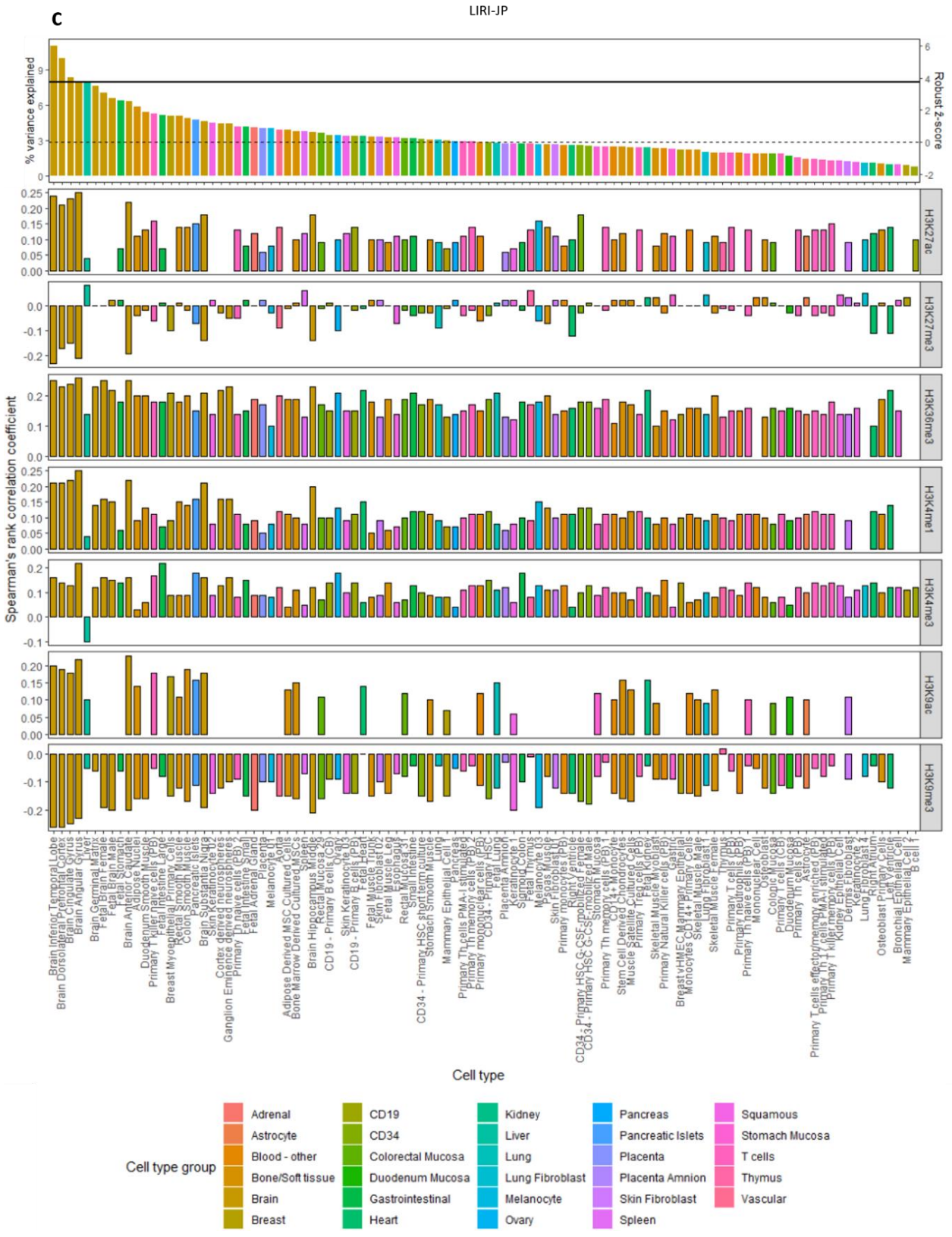

Figure 6. Continued

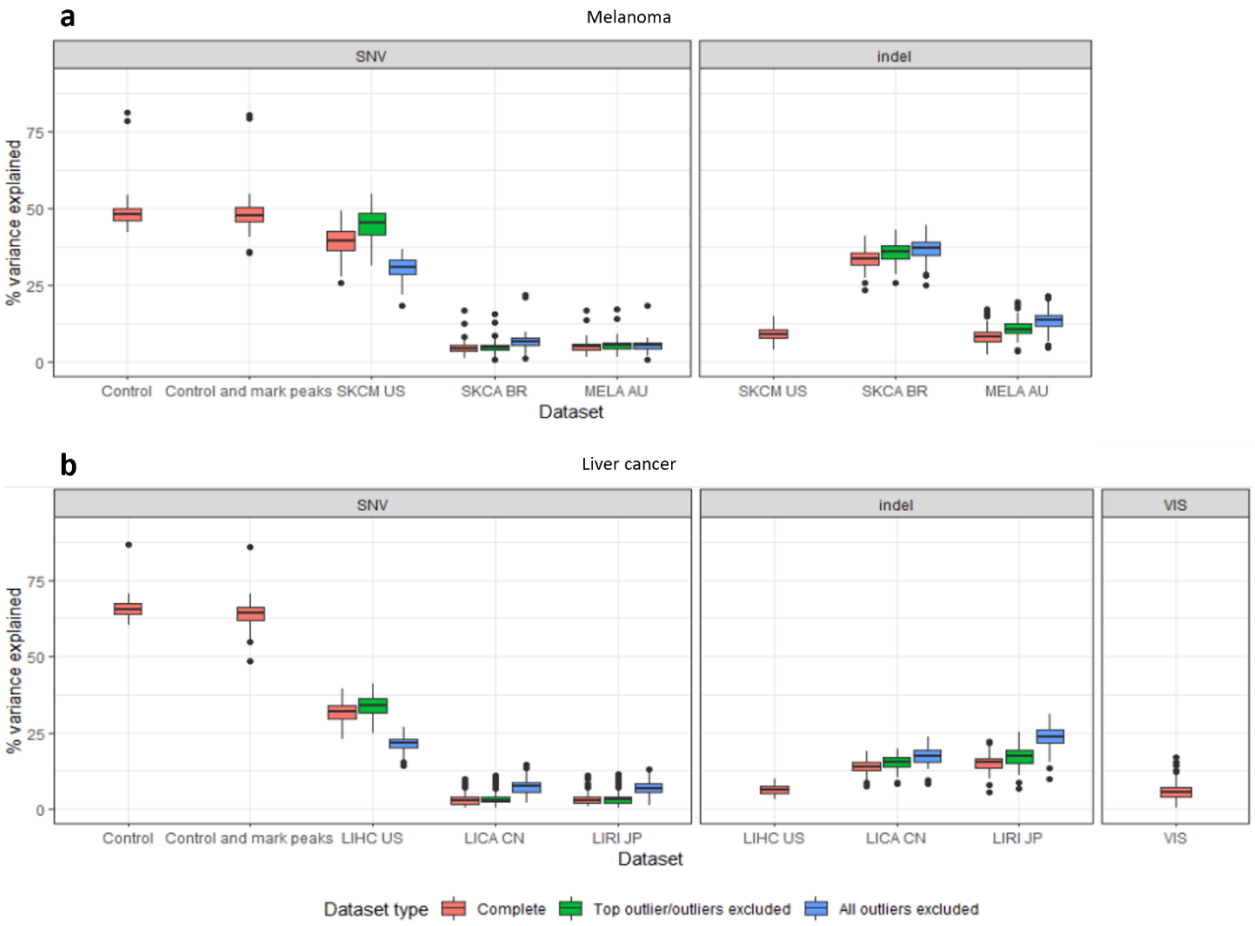

Figure 7. Distribution of variance explained of cell type models for every researched dataset Distribution of accuracy of 104 cell type models trained on various melanoma or liver carcinoma single nucleotide variation and insertion and deletion datasets. Models labelled "Control and mark peaks" were built on control SNV mutation dataset and chromatin peak data and "Complete" dataset type refers to datasets that did not had had any window count observation removed. (**a**) Distribution of variance explained based on models built on different melanoma datasets. (**b**) Distribution of variance explained based on models built on different liver cancer datasets. Box plots, band inside the box, median; box, first and third quartiles; whiskers, most extreme values within 1.5 × inter-quartile range from the box; points, outliers.

Spearman's rank correlation coefficient direction of LIHC-US dataset was positive for all of the H3K27ac, H3K36me3, H3K4me1 and H3K9ac marks (**Figure 6a**) regardless of the outlier removal. LICA-CN and LIRI-JP datasets and subsets provided the same results with the exception of H3K4me3 chromatin mark where Liver cell type had different correlation coefficient when compared to the rest of the cell types (**Figure 6b, Figure 6c**). Majority (more than 85%) of the H3K9me3 cell types in all datasets and subsets had negative correlation coefficient direction with Liver cell type being one of the exceptions in LICH-US dataset, while H3K27me3 had positive

correlation coefficient in majority (more than 84%) of the cell types when considering LIHC-US and LICA-CN datasets and subsets. LIRI-JP dataset had correlation coefficient close to zero regarding the H3K27me3 chromatin mark (median = -0.01, 95% CI = -0.02, 0.00), which also applied to its top outlier removed subset (median = -0.01, 95% CI = -0.02, 0.00) and all outliers removed subset (median = 0.02, CI = 0.00, 0.04). These results show contrast when compared to the control Liver-HCC results where positive Spearman's rank correlation coefficient was observed in H3K9me3 mark and negative in H3K36me3, H3K4me1 and H3K4me3 marks. The only match between tested and control result was positive correlation coefficient of the H3K27me3 mark.

4.4 Analysis based on melanoma and liver carcinoma indel mutations results in inaccurate COO prediction for all datasets

All of the melanoma and liver carcinoma datasets and outlier subsets had inaccurate and unsignificant prediction ((*p*-value > 0.03, WHW test) of either melanoma or liver carcinoma COO (**Figure 8a, Figure 8b**, **Figure 8c**, **Figure 9a**, **Figure 9b**, **Figure 9c**).

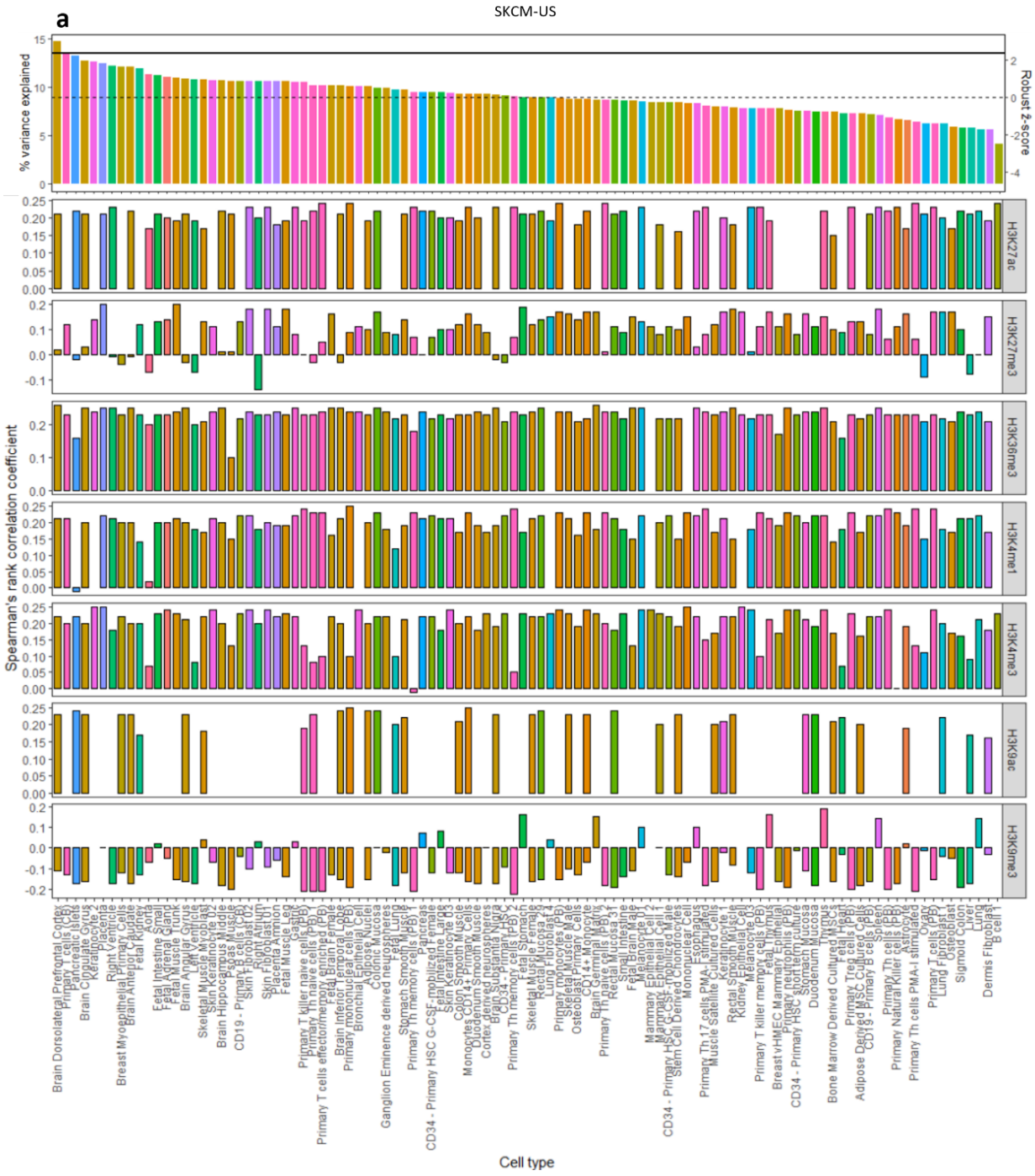

Figure 8. Analysis based on melanoma insertion and deletion and chromatin peak data results in inaccurate cell-of-origin prediction. (**a**, **b**, **c**) Model accuracy based on chromatin marks of each normal cell type. The cell type that fits the model with the highest prediction accuracy and significantly differs from the next-best histologically unrelated cell type is determined to be the cell-of-origin. Solid horizontal line describes variance explained reported by the next-best cell type model that belong to the different cell type group. Dotted horizontal line indicates median variance. Spearman's rank correlation coefficient between mutation and chromatin marks is reported for every cell type individually. Right-hand (secondary) y-axis shows robust ẑ-score. Bars are coloured by histological group.

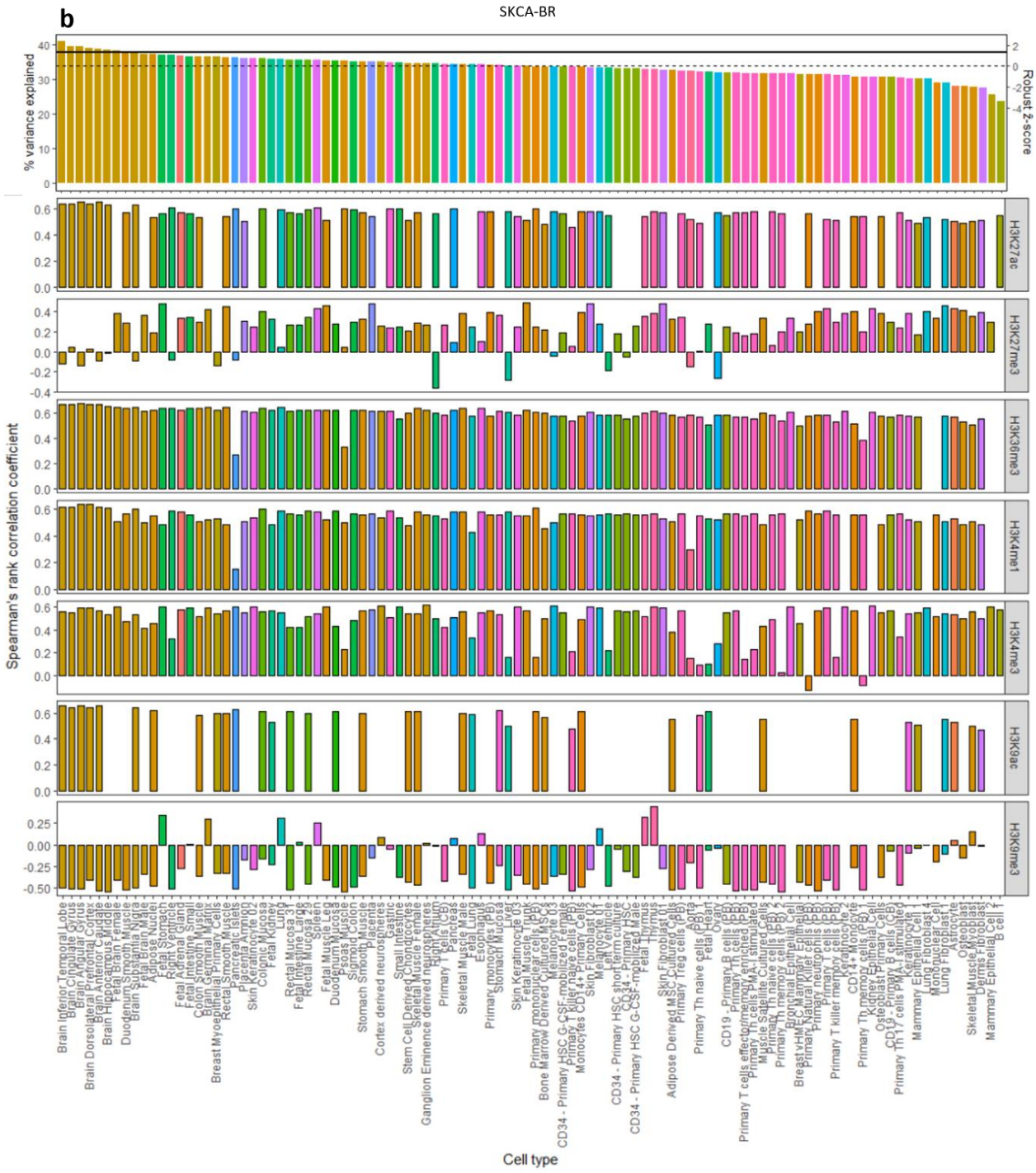

Figure 8. Continued

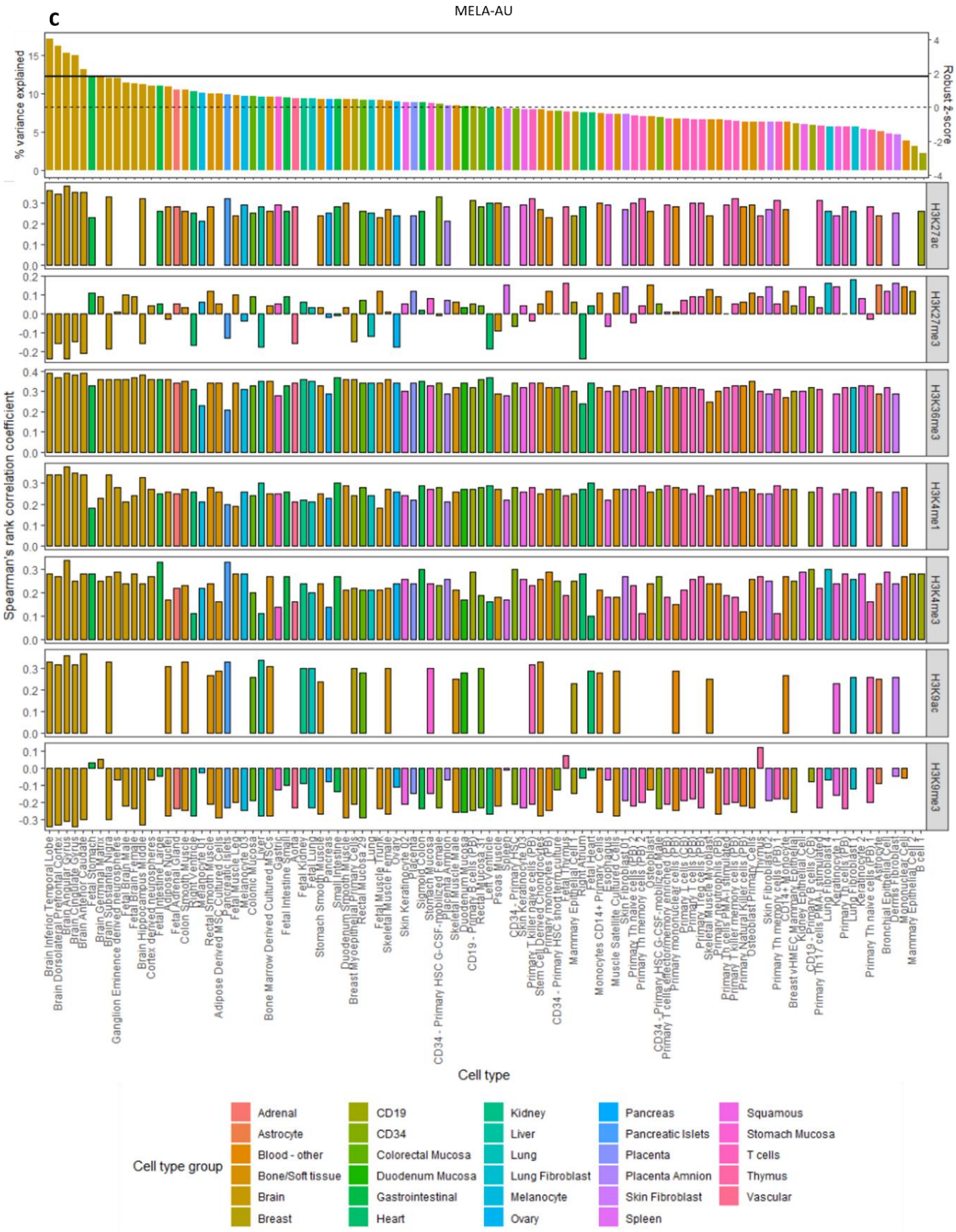

Figure 8. Continued

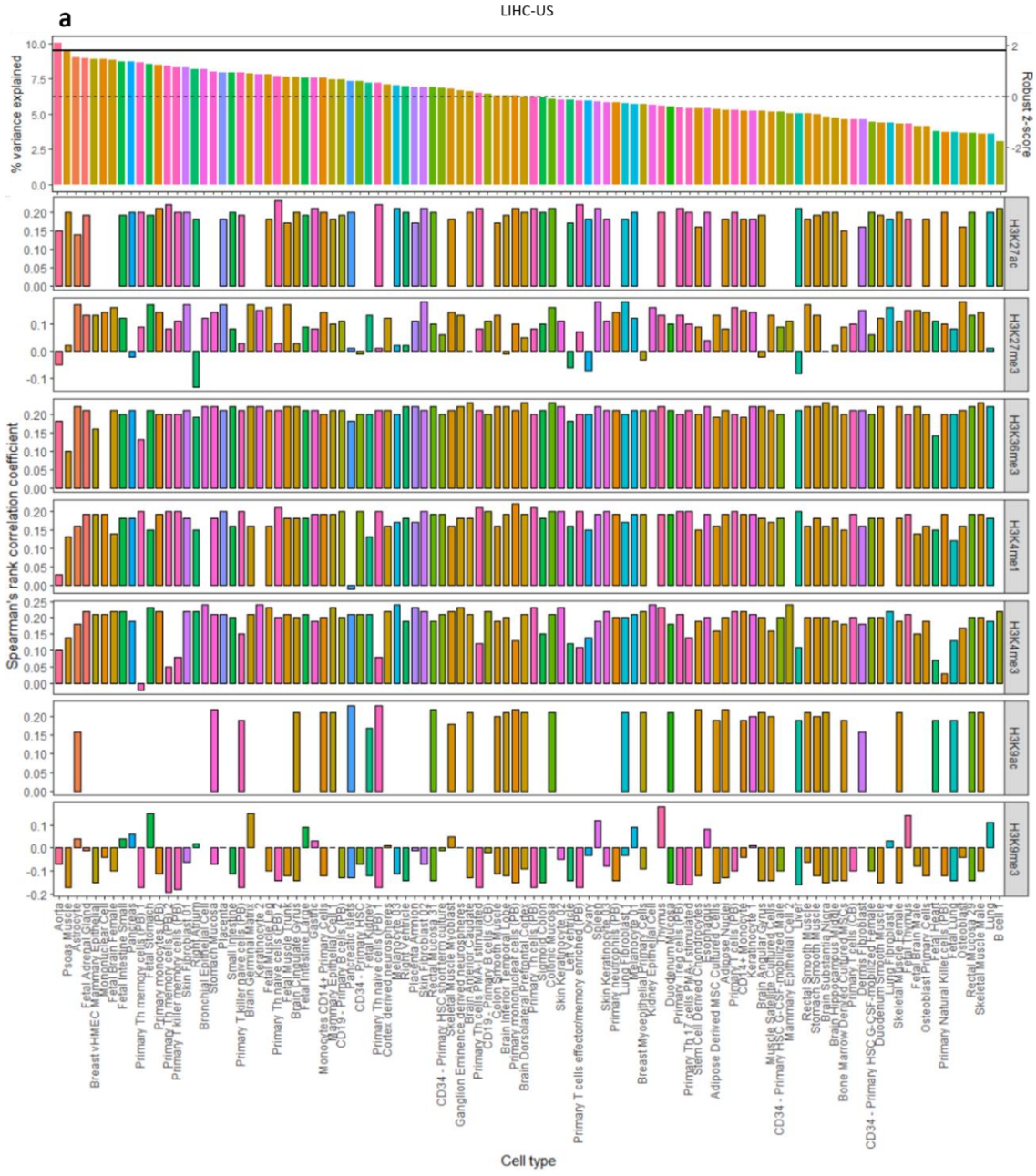

Figure 9. Analysis based on liver cancer insertion and deletion and chromatin peak data results in inaccurate cell-of-origin prediction. (**a**, **b**, **c**) Model accuracy based on chromatin marks of each normal cell type. The cell type that fits the model with the highest prediction accuracy and significantly differs from the next-best histologically unrelated cell type is determined to be the cell-of-origin. Solid horizontal line describes variance explained reported by the next-best cell type model that belong to the different cell type group. Dotted horizontal line indicates median variance. Spearman's rank correlation coefficient between mutation and chromatin marks is reported for every cell type individually. Right-hand (secondary) y-axis shows robust $\hat{z}$-score. Bars are coloured by histological group.

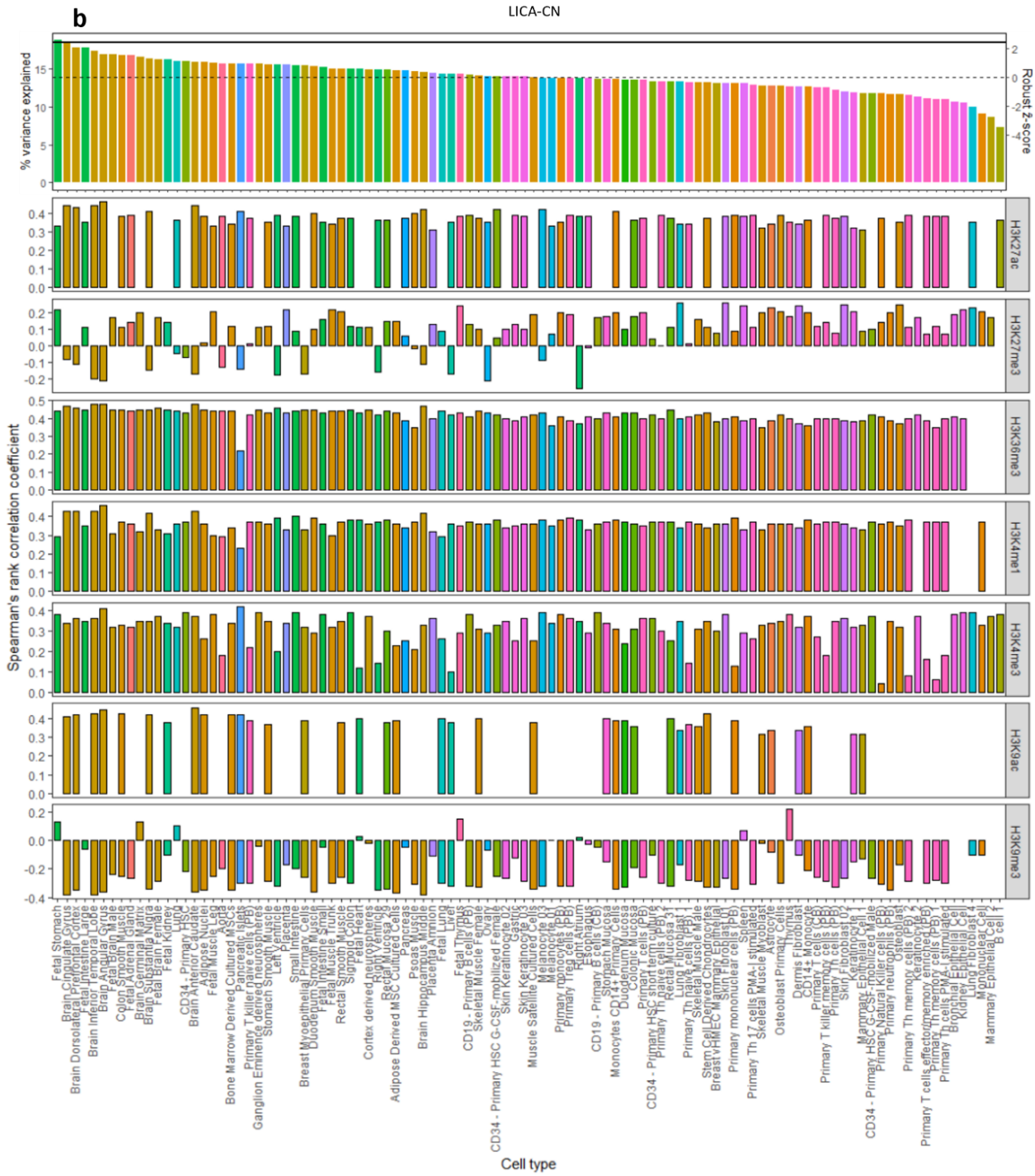

Figure 9. Continued

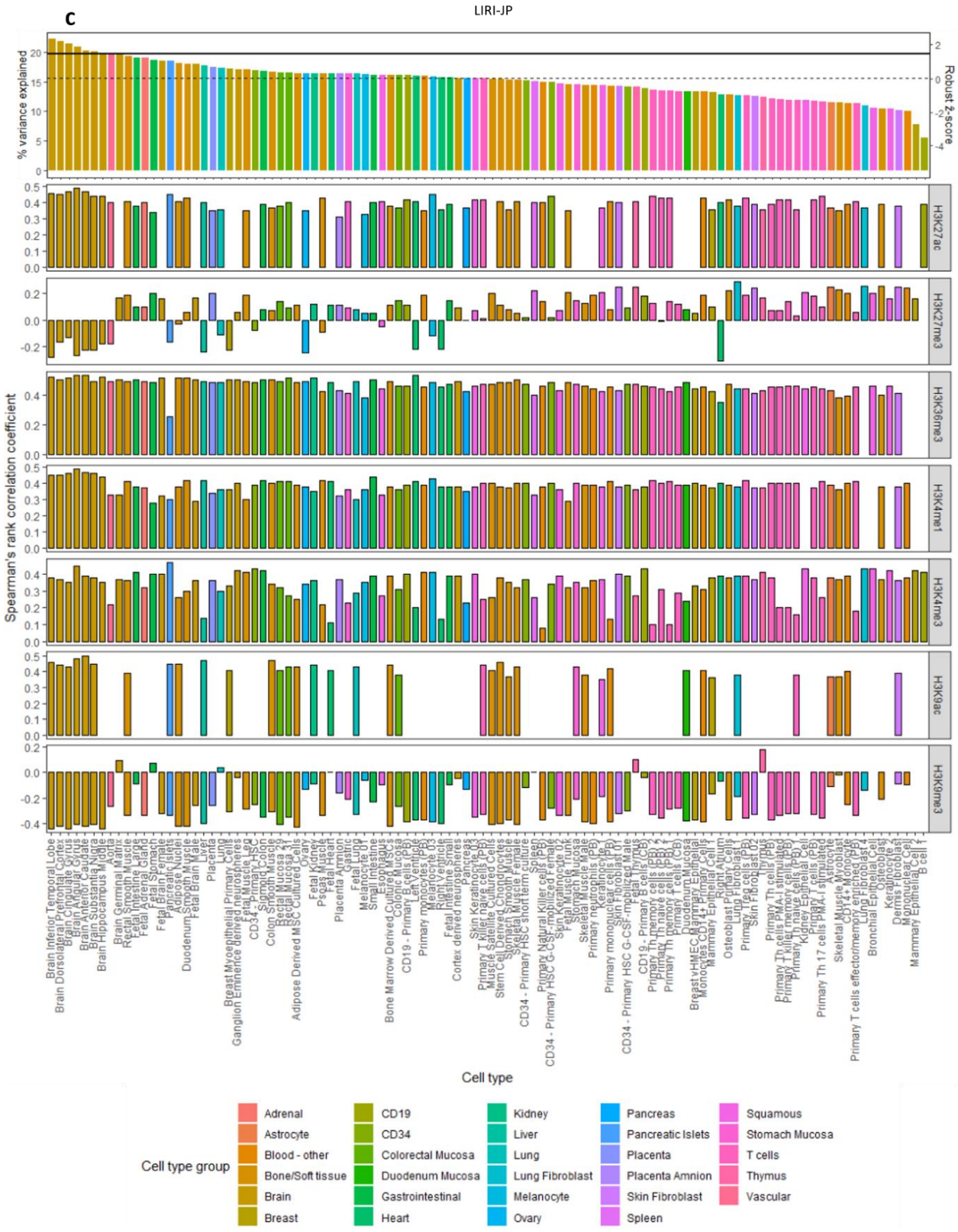

Figure 9. Continued

By removing outliers in SKCM-US and LIHC-US data there was not enough information to perform Random Forest regression analysis since SCKCM-US and LICH-US datasets contained significantly less data compared to SKCA-BR and MELA-AU datasets (**Figure 10, Figure 11**).

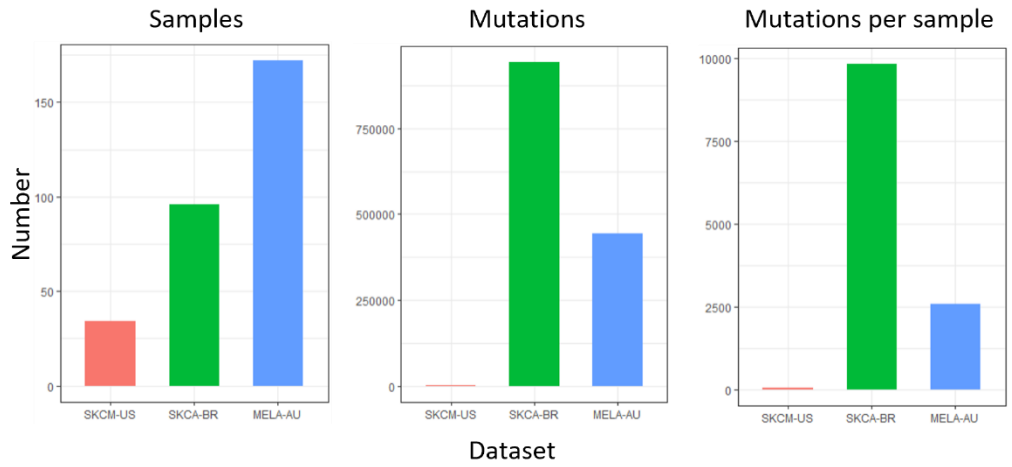

Figure 10. Melanoma insertion and deletion datasets vary in sample and mutation numbers. "SKCM-US" stands for Skin Cutaneous melanoma - TCGA, US dataset, "SKCA-BR" stands for Skin Adenocarcinoma – BR dataset and MELA-AU stands for Skin Cancer – AU dataset.

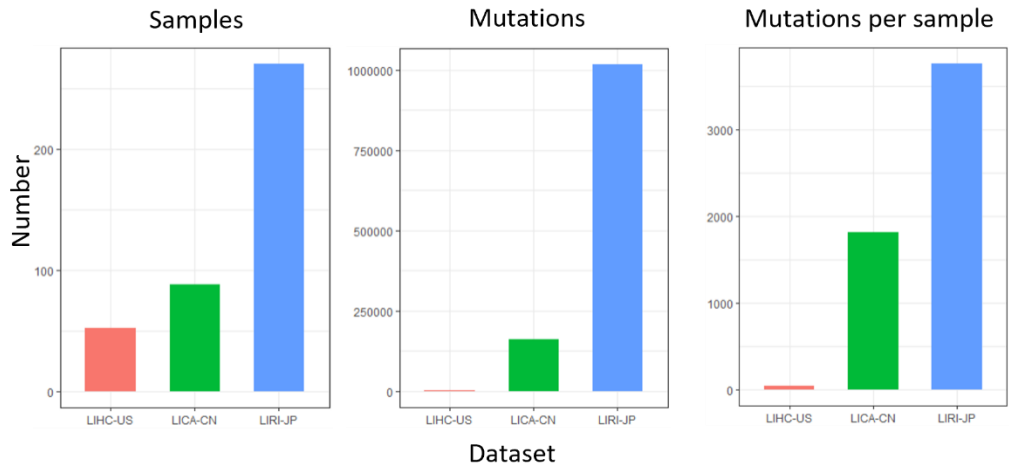

Figure 11. Liver cancer insertion and deletion datasets vary in sample and mutation numbers. "LIHC-US" stands for Liver Hepatocellular carcinoma - TCGA, US dataset, "LICA-CN" stands for Liver Cancer – CN dataset and "LIRI-JP" stands for Liver Cancer - RIKEN, JP dataset.

On the other hand, removing top outliers in SKCA-BR, MELA-AU, LICA-CN and LIRI-JP resulted in significant increase in variance explained for five cell types in MELA-AU dataset and no significant change in variance explained for SKCA-BR, LICA-CN and LIRI-JP datasets. Removing all outliers resulted in significant increase in variance explained for one cell type in SKCA-BR dataset, 52 cell types in MELA-AU dataset, five cell types in LICA-CN dataset and 79 cell types in LIRI-JP dataset. Median variance explained increased from 8.2% to 10.8% for MELA-AU dataset and from 15.5% to 23.6% for LIRI-JP dataset (**Figure 7**).

Both melanoma and liver carcinoma indel datasets and outlier subsets had positive Spearman's rank correlation coefficient for all of the cell types when comparing indels to H3K27ac, H3K36me3 and H3K9ac marks. Majority (more than 75%) of all cell types had positive H3K27me3 correlation coefficient and negative H3K9me3 correlation coefficient, with notable exceptions being Liver that had negative H3K27me3 correlation coefficient in all the liver carcinoma datasets and outlier subsets (**Figure 9a**, **Figure 9b**, **Figure 9c**), Melanoma 01 with positive H3K9me3 correlation coefficient in SKCM-US dataset (**Figure 8a**) and MELA-AU (**Figure 8c**) datasets and Melanoma 03 with negative H3K27me3 correlation coefficient for all the SKCA-BR **(Figure 8b**) and MELA-AU (**Figure 8c**) datasets and subsets.

4.5 Analysis based on HBV integration sites results in inaccurate COO prediction for all datasets

Analysis based on HBV integration sites did not accurately nor significantly predict the COO (**Figure 12a**) which was supposed to be Liver cell type (p-value = 0.695, WMW test). Median variance explained was 5.4% with 95% confidence interval ranging from 4.9% 5.8% (**Figure 3**). Since dataset only had 16,905 VIS, the number of mutations was too low to perform analysis based on the outlier removal.

Regarding the VIS and chromatin marks correlation, Spearman's rank correlation coefficient was positive for all of the cell types when referring to H3K27ac and H3K9ac marks and majority of the cell types when referring to H3K27me3 (95/103), H3K4me1 (94/97), H3K4me3 (102/104) and H3K9me3 (74/99) chromatin marks (**Figure 12b**). It is also worth noting that absolute maximal coefficient value between all the cell types and chromatin marks was 0.17 which means that overall correlation between chromatin marks and VIS was not large.

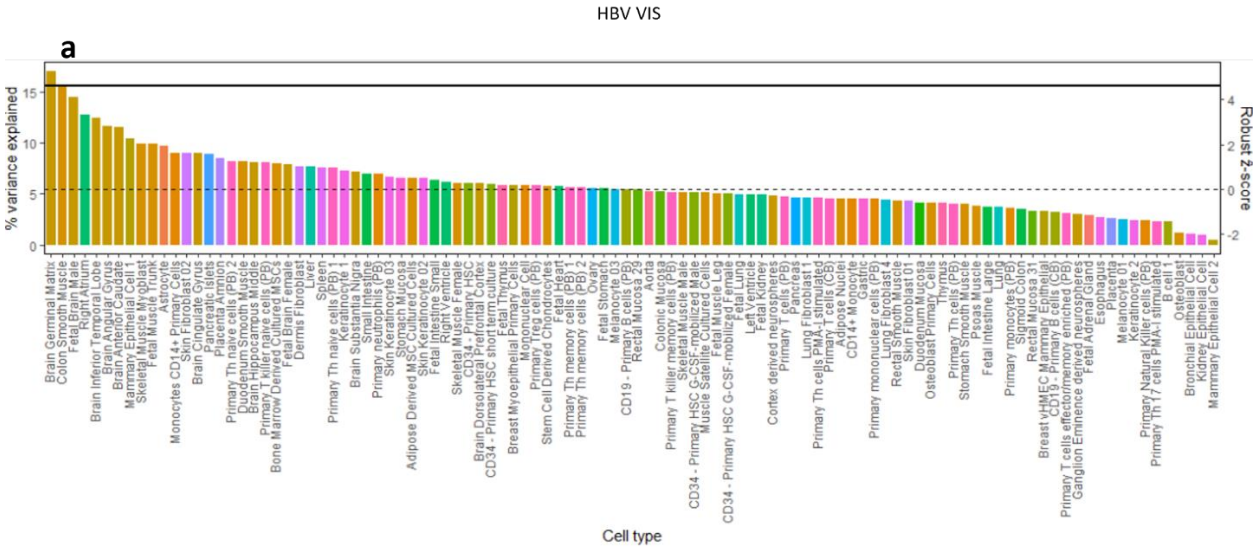

Figure 12. Analysis based hepatitis B virus integration sites and chromatin peak data results in inaccurate cell-of-origin prediction. (**a**) Model accuracy based on chromatin marks of each normal cell type. The cell type that fits the model with the highest prediction accuracy and significantly differs from the next-best histologically unrelated cell type is determined to be the cell-of-origin. Solid horizontal line describes variance explained reported by the next-best cell type model that belong to the different cell type group. Dotted horizontal line indicates median variance. (**b**) Spearman's rank correlation coefficient between mutation and chromatin marks *is reported* for every cell type individually. Right-hand (secondary) y-axis shows robust *ẑ*-score. Bars are coloured by histological group.

Figure 12. Continued

5. DISCUSSION

Connection between mutations and chromatin organization can be established through correlation between mutations and chromatin marks. In this work I presented on two independent melanoma datasets that SNV mutations negatively corelate with chromatin marks responsible for increased in DNA access and positively corelate with chromatin marks responsible for decreased DNA access. This supports the results presented in Polak et al. (2015) and Kübler et al. (2019) works where the same conclusion was made based on SNV melanoma and liver carcinoma data, amongst other. This also infers that closed chromatin regions accumulate more SNV mutations in melanoma which corresponds to the results of Polak et al. (2013) paper where lower rates of mutations are observed in accessible, regulatory DNA and explained by the higher DNA repair frequency in the open chromatin. With the mentioned datasets I confirmed previous finding that mutational landscape is significantly influenced by the normal cellular context of the COO.

Models build on liver SNV data, melanoma TCGA SNV data, indel mutations and VIS did not accurately or significantly predict COO. As it was stated before (Kübler et al., 2019), one possible reason for inaccurate prediction might be an insufficient number of mutations that makes individual cancers indistinguishable. However, the number of mutations required for the accurate prediction to occur might vary between different datasets and should be taken into consideration. Another possible explanation for the difference of prediction accuracy when comparing mentioned datasets to the datasets that accurately predicted COO could be the fact that the datasets were produced by different research groups, and the observed differences might be a consequence of differences in laboratory protocols. The same explanation can potentially elucidate the opposite correlation between certain SNV datasets and their control counterparts. On the other hand, correlation results based on the indel models with the highest percentage of variance explained indicate higher rate of indel mutation accumulation within open chromatin regions since indel mutations had positive correlation with chromatin marks associated with closed chromatin and negative correlation with chromatin marks associated with open chromatin, the only exception being H3K27me3 mark. This concurs with previous work (Don et al., 2013) where the same phenomenon was observed. Furthermore, it is possible that indel mutations occur in regions with chromatin marks shared between all cell types and therefore do not correlate with tissue specific chromatin marks. This would explain why indel models predict COO from “Brain” cell group and why there is no significant variation in variance explained between the COO and the next-best histologically

unrelated cell type. Finally, VIS data positively correlated with chromatin marks associated with both open and closed chromatin regions which varies from results presented in paper Hama et al. (2018) where VIS in liver cancer cells preferentially occur in the regions of closed chromatin, the probable reason for this difference being low number of mutations.

One of the novelties presented was the use of chromatin peak data, since works on record used read alignment information from ChIP-Seq experiments (Kübler et al., 2019; Polak et al., 2015). Although predictions based on peak data proved to be successful, one of the potential problems is missing information in peak databases, such as DeepBlue epigenome database missing DEEP (IHEC) peak data. On the other hand, in comparison with a read data counterpart, peak data does take considerably less memory space and computational power when aligning to the reference genome which can facilitate chromatin data download and processing.

In this work I also showed that outlier exclusion might lead to both significant increase and decrease in variance explained by the Random Forest regression model. Random Forest algorithm is supposed to be tolerable of the outliers since it does not matter how much a case varies from the threshold value on a selected feature variable when tree building is performed. Furthermore, output outliers will affect the estimate of the leaf node they are in, but not the values of any other leaf node. This can be demonstrated by inducing noise where Random Forest proved to be less affected by it when compare to other algorithms (Breiman, 2001). However, Random Forest robustness can be improved through various means, for example, using median instead of mean when combining the predictions from the individual trees, using least-absolute deviations from the median, instead of least-squares, as splitting criterion or building the trees using the ranks of the response instead of the original values (Roy and Larocque, 2012), to name a few.

Although analysis based on indel mutations and VIS did not provide accurate COO predictions, several approaches can be used with the goal to improve analysis results. First, with the new cancer genomic projects in development it will be possible to increase in the number of mutations through increase in sample number. For example, ICGC-ARGO is a 10-year project started in 2018 with the goal of analysing the genomes of more than 200,000 patients by detecting mutations through extensive sequencing (*ICGC ARGO - About ICGC ARGO*, n.d.). Second, change in reference window size could be a viable approach since window size of 1 Mb roughly matches the size of topologically associated domains, but they represent only a single level in hierarchical chromatin

organization (Bonev and Cavalli, 2016) and it is possible that indel mutations associate with different chromatin organizational levels. Third, various NGS platforms report different rates of success in SNV and indel detection with indels being harder to detect (Mullaney et al., 2010) and the usage of different pipelines in mutation detection may also significantly affect the SNV and indel number (Zook et al., 2014). Therefore, obtaining data generated by different pipelines might lead to improvement in indel detection. Finally, combination of different mutations and complex model building could also provide better COO prediction. All of this research approaches are worth exploring in the future since they may lead to utilization of indel and VIS in COO determination.

6. CONCLUSIONS

In this work I confirmed that COO prediction based on SNV mutations is possible using Random Forest regression analysis. The same cannot be said for indel mutations and VIS, that did not provide accurate COO prediction. SNV mutation type provided the highest quality of the mutation density prediction. Furthermore, SNV correlation with the chromatin marks resulted in confirmed hypothesis that SNV mutations positively correlate with chromatin markers characteristic for closed chromatin and negatively correlate with chromatin marks typical for open chromatin only in cases when accurate COO was predicted. Opposite was concluded for indel mutations which positively correlated with chromatin marks associated with open chromatin and negatively correlated with chromatin marks associated with closed chromatin, with one mark exception. However, I did not confirm the hypothesis that VIS preferably occur in closed chromatin regions since it correlated with marks typical for both open and closed chromatin. The cause of this was probably low number of VIS. Regarding the lower number of mutations, COO prediction power did not change when mutation count was lowered which indicates that number of mutations required for the accurate prediction to occur might vary between different datasets. In the end, outlier exclusion leads to both increase and decrease in COO prediction power and can be a viable method used in COO prediction. In the future this type of analysis can be improved by using chromatin peak data instead of read data, modifying Random Forest algorithm to increase robustness or by using additional cancer samples in the analysis.

7. BIBLIOGRAPHY

Albrecht, F., List, M., Bock, C., and Lengauer, T. (2016). DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic Acids Research*, *44*(W1), W581–W586. https://doi.org/10.1093/NAR/GKW211

Albrecht, F., List, M., Bock, C., and Lengauer, T. (2017). DeepBlueR: Large-scale epigenomic analysis in R. *Bioinformatics*, *33*(13), 2063–2064. https://doi.org/10.1093/BIOINFORMATICS/BTX099

Beisel, C., and Paro, R. (2011). Silencing chromatin: Comparing modes and mechanisms. *Nature Reviews Genetics*, *12*(2), 123–135. https://doi.org/10.1038/NRG2932

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, *17*(11), 661–678. https://doi.org/10.1038/nrg.2016.112

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cairns, B. R. (2007). Chromatin remodeling: Insights and intrigue from single-molecule studies. *Nature Structural and Molecular Biology*, *14*(11), 989–996. https://doi.org/10.1038/NSMB1333

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C. L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … Zhang, J. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/S41586-020-1969-6

Collas, P. (2010). The Current State of Chromatin Immunoprecipitation. *Molecular Biotechnology*, *45*(1), 87–100. https://doi.org/10.1007/S12033-009-9239-8

Don, P. K., Ananda, G., Chiaromonte, F., and Makova, K. D. (2013). Segmenting the human genome based on states of neutral genetic divergence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(36), 14699–14704. https://doi.org/10.1073/PNAS.1221792110/SUPPL_FILE/SAPP.PDF

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Gilbertson, R. J. (2011). Mapping Cancer Origins. *Cell*, *145*(1), 25–29. https://doi.org/10.1016/J.CELL.2011.03.019

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gruber, S. B., and Armstrong, B. K. (2009). Cutaneous and Ocular Melanoma. *Cancer Epidemiology and Prevention*. https://doi.org/10.1093/ACPROF:OSO/9780195149616.003.0063

Ha, K., Fujita, M., Karlić, R., Yang, S., Xue, R., Zhang, C., Bai, F., Zhang, N., Hoshida, Y., Polak, P., Nakagawa, H., Kim, H. G., and Lee, H. (2020). Somatic mutation landscape reveals differential variability of cell-of-origin for primary liver cancer. *Heliyon*, *6*(2), e03350. https://doi.org/10.1016/J.HELIYON.2020.E03350

Hama, N., Totoki, Y., Miura, F., Tatsuno, K., Saito-Adachi, M., Nakamura, H., Arai, Y., Hosoda, F., Urushidate, T., Ohashi, S., Mukai, W., Hiraoka, N., Aburatani, H., Ito, T., and Shibata, T. (2018). Epigenetic landscape influences the liver cancer genome architecture. *Nature Communications*, *9*(1). https://doi.org/10.1038/S41467-018-03999-Y

Hodgkinson, A., Chen, Y., and Eyre-Walker, A. (2012). The Large-Scale Distribution of Somatic Mutations in Cancer Genomes. *Hum Mutat*, *33*, 136–143. https://doi.org/10.1002/humu.21616

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusuda, J., Lane, D. P., … Wainwright, B. J. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–998. https://doi.org/10.1038/NATURE08987

Hyman, D. M., Piha-Paul, S. A., Won, H., Rodon, J., Saura, C., Shapiro, G. I., Juric, D., Quinn, D. I., Moreno, V., Doger, B., Mayer, I. A., Boni, V., Calvo, E., Loi, S., Lockhart, A. C., Erinjeri, J. P., Scaltriti, M., Ulaner, G. A., Patel, J., … Solit, D. B. (2018). HER kinase inhibition in patients with HER2- and HER3-mutant cancers. *Nature*, *554*(7691), 189–194. https://doi.org/10.1038/nature25475

ICGC ARGO - About ICGC ARGO. (n.d.). Retrieved March 10, 2022, from https://www.icgc-argo.org/page/64/about-icgc-argo

Karimzadeh, M., Ernst, C., Kundaje, A., and Hoffman, M. M. (2018). Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research, 46*(20), e120–e120. https://doi.org/10.1093/NAR/GKY677

Karolchik, D., Hinricks, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(Database issue). https://doi.org/10.1093/NAR/GKH103

Köhler, C., Nittner, D., Rambow, F., Radaelli, E., Stanchi, F., Vandamme, N., Baggiolini, A., Sommer, L., Berx, G., van den Oord, J. J., Gerhardt, H., Blanpain, C., and Marine, J. C. (2017). Mouse Cutaneous Melanoma Induced by Mutant BRaf Arises from Expansion and Dedifferentiation of Mature Pigmented Melanocytes. *Cell Stem Cell*, *21*(5), 679-693.e6. https://doi.org/10.1016/J.STEM.2017.08.003

Kübler, K., Karlić, R., Haradhvala, N. J., Ha, K., Kim, J., Kuzman, M., Jiao, W., Gakkhar, S., Mouw, K. W., Braunstein, L. Z., Elemento, O., Biankin, A. v., Rooman, I., Miller, M., Karthaus, W. R., Nogiec, C. D., Juvenson, E., Curry, E., Kenudson, M. M., … Getz, G. (2019). Tumor mutational landscape is a record of the pre-malignant state. *BioRxiv*, 517565. https://doi.org/10.1101/517565

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/JSS.V028.I05

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. v., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., R, A. H., and Getz, G. (2013). Mutational heterogeneity in cancer and

the search for new cancer genes HHS Public Access. *Nature*, *499*(7457), 214–218. https://doi.org/10.1038/nature12213

London, W. T., and Mcglynn, K. A. (2009). Liver Cancer. *Cancer Epidemiology and Prevention*. https://doi.org/10.1093/ACPROF:OSO/9780195149616.003.0039

Mitchell, R. S., Beitzel, B. F., Schroder, A. R. W., Shinn, P., Chen, H., Berry, C. C., Ecker, J. R., and Bushman, F. D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biology*, *2*(8). https://doi.org/10.1371/JOURNAL.PBIO.0020234

Mu, X., Español-Suñer, R., Mederacke, I., Affò, S., Manco, R., Sempoux, C., Lemaigre, F. P., Adili, A., Yuan, D., Weber, A., Unger, K., Heikenwälder, M., Leclercq, I. A., and Schwabe, R. F. (2015). Hepatocellular carcinoma originates from hepatocytes and not from the progenitor/biliary compartment. *The Journal of Clinical Investigation*, *125*(10), 3891–3903. https://doi.org/10.1172/JCI77995

Mullaney, J. M., Mills, R. E., Stephen Pittard, W., and Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131–R136. https://doi.org/10.1093/HMG/DDQ400

Polak, P., Karlic, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., Reynolds, A., Rynes, E., Vlahovicek, K., Stamatoyannopoulos, J. A., and Sunyaev, S. R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, *518*(7539), 360–364. https://doi.org/10.1038/nature14221

Polak, P., Lawrence, M. S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R. E., Garraway, L. A., Mirkin, S., Getz, G., Stamatoyannopoulos, J. A., and Sunyaev, S. R. (2013). Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature Biotechnology*, *32*(1), 71–75. https://doi.org/10.1038/nbt.2778

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roy, M. H., and Larocque, D. (2012). Robustness of random forests for regression. *Journal of Nonparametric Statistics*, *24*(4), 993–1006. https://doi.org/10.1080/10485252.2012.715161

Schuster-Böckler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, *488*(7412), 504–507. https://doi.org/10.1038/nature11273

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/CAAC.21660

Tang, D., Li, B., Xu, T., Hu, R., Tan, D., Song, X., Jia, P., and Zhao, Z. (2020). VISDB: A manually curated database of viral integration sites in the human genome. *Nucleic Acids Research*, *48*(D1), D633–D641. https://doi.org/10.1093/nar/gkz867

Voelkerding, K. v., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, *55*(4), 641–658. https://doi.org/10.1373/CLINCHEM.2008.112789

Wang, Y., Tong, Y., Zhang, Z., Zheng, R., Huang, D., Yang, J., Zong, H., Tan, F., and Zhang, X. (2020). ViMIC: A Database of Human Disease-related Virus Mutations, Integration Sites and Cis-effects. *BioRxiv*, 2020.10.28.359919. https://doi.org/10.1101/2020.10.28.359919

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., … Kling, T. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Wright, M. N., and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/JSS.V077.I01

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L. D., and Ferretti, V. (2019). The International Cancer Genome Consortium Data Portal. *Nature Biotechnology*, *37*(4), 367–369. https://doi.org/10.1038/S41587-019-0055-9

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology, 32*(3), 246–251. https://doi.org/10.1038/nbt.2835

BIOGRAPHY

My name is Ivan Bakšić and I was born 28th of September 1997 in Zagreb, Croatia. I attended Elementary school "Ljubo Babić" in my home town of Jastrebarsko, where I gained interest in STEM field of study, competed in various Science school contests and even reached state championship in Technical Education. This prompted me to enrol The Fifth Gymnasium high school in Zagreb, specialised in science and mathematics where I broadened my knowledge and interest in biology and informatics. In the academic year of 2016/2017, I enrolled Undergraduate study of Molecular Biology at the Faculty of Science, University of Zagreb. During my undergraduate years I participated in "Biology Night" events with the goal of promoting science and education and I widened my practical laboratory skills by volunteering at the Molecular Biomedicine Laboratory of the Centre for Research and Knowledge Transfer in Biotechnology, University of Zagreb in the year of 2018. Furthermore, I won Chancellor's Award for the work titled "Characteristics and evolution of 5S rDNA unit arrays in species of the genus *Pulsatilla* (Mill.)", part of which I presented at the 5th Students' Symposium in Biology and Life Science held in Zagreb, 2019 with the title "Evolution of 5S rDNA in selected species of genus *Pulsatilla*". I concluded my undergraduate study in the academic year 2019/2020 with the final paper titled "Biogeographical evidence of evolution" and in the next academic year I enrolled Graduate study of Molecular Biology at the Faculty of Science, University of Zagreb. In my graduate years I took interest in computational biology and specialised myself in the field of bioinformatics. I attended Copenhagen Bioinformatics Hackathon in 2021 where I competed in building machine learning model.