MEDICAL UNIVERSITY OF VIENNA

Ce–M–M–
Research Center for Molecular Medicine of the Austrian Academy of Sciences

Epigenomic and transcriptional determination of cellular identity

Doctoral thesis at the Medical University of Vienna
for obtaining the academic degree

Doctor of Philosophy

Submitted by

Johanna Klughammer

Supervisor:
Christoph Bock, PhD

CeMM Research Center for Molecular Medicine
of the Austrian Academy of Sciences

Lazarettgasse 14, AKH BT25.3
1090 Vienna, Austria

Vienna, 06/2017

I. Declaration

This thesis is written in a cumulative format and contains two published manuscripts and one that has been submitted (listed below). The author of this thesis was crucially involved in the presented research as well as the preparation of the manuscripts. Detailed author contribution statements approved by all authors are also specified in each manuscript in the section “Author contributions”.

The author of this thesis, Johanna Klughammer, wrote the remainder of this thesis with input from Christoph Bock.

List of first-author publications:

Klughammer J, Datlinger P, Printz D, Sheffield NC, Farlik M, Hadler J, Fritsch G, Bock C. **Differential DNA Methylation Analysis without a Reference Genome.** Cell Rep. 2015 Dec;13(11):2621-33. doi:10.1016/j.celrep.2015.11.024.

Contribution: The author of this thesis designed the study, acquired and prepared the samples, performed FACS sorting, developed RefFreeDMA and performed the computational analysis, and wrote the manuscript.

Klughammer J*, Kiesel B*, Roetzer T, Fortelny N, Kuchler A, Datlinger P, Peter N, Nenning K, Furtner J, Nowosielski M, Augustin M, Mischkulnig M, Ströbel T, Moser P, Freyschlag CF, Kerschbaumer J, Thomé C, Grams AE, Stockhammer G, Kitzwoegerer M, Oberndorfer S, Marhold F, Weis S, Trenkler J, Buchroithner J, Pichler J, Haybaeck J, Krassnig S, Ali KM, von Campe G, Payer F, Sherif C, Preiser J, Hauser T, Winkler PA, Kleindienst W, Würtz F, Brandner-Kokalj T, Stultschnig M, Schweiger S, Dieckmann K, Preusser M, Langs G, Baumann B, Knosp E, Widhalm G, Marosi C, Hainfellner JA, Woehrer A, Bock C. **The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space.** Submitted.

Contribution: The author of this thesis designed the study, performed the data analysis, and wrote the manuscript.

Li J*, Klughammer J*, Farlik M*, Penz T*, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S. **Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types.** EMBO Rep. 2016 Feb;17(2):178-87. doi:10.15252/embr.201540946.

Contribution: The author of this thesis processed the raw data, performed the bioinformatic analysis, and wrote the manuscript. This publication is also include the PhD thesis of Jin Li (Verbund-ID-Nr.: AC10778692; http://permalink.obvsg.at/AC10778692).

*Shared first authorship

II. Table of contents

III. List of figures

IV. Abstract

Cells are the basic units of life. The ability to maintain functionally specialized and cooperating cell types provides the foundation for multicellular organisms. One of the most fascinating aspects of multicellular life is how a shared genomic sequence supports morphologically and functionally diverse cell types. Understanding the regulatory processes and molecular mechanisms underlying cellular identity is therefore important to understanding multicellular life. With the advances of modern molecular biology, the interplay between genome, epigenome, and transcriptome has emerged as the leading determinant of cellular identity. Technological advances in high throughput sequencing have further provided the necessary tools to study these molecular determinants of cellular identity in detail.

Despite substantial scientific progress, many fundamental questions remain to unanswered. For example, studies conducted mainly in human and mouse have shown that DNA methylation is crucially involved in the determination of cellular identity, which appears to be conserved across vertebrates. Invertebrates, however, display fundamentally different DNA methylation patterns and in some species, DNA methylation cannot be detected at all. This raises the question of how and why DNA methylation acquired its defining role for cellular identity in vertebrates, what the role of DNA methylation in invertebrates might be, and how some higher organisms can exist without detectable levels of DNA methylation. Answering these evolutionary questions might identify yet unknown mechanisms involved in the determination of cellular identity or new functions of DNA methylation that may be relevant to human physiology. To advance research in this direction, we have developed a computational framework called RefFreeDMA that allows differential DNA methylation analysis without the need of a reference genome, enabling the assessment of DNA methylation in virtually any species. We successfully validated our approach in three species (human, cow, and carp) and plan to apply it next in the assessment of tissue specific DNA methylation across many more vertebrate and invertebrate species.

In the context of human diseases, the clinical relevance of changes in cellular identity is widely accepted. Aberration of cellular identity has been recognized as a fundamental factor in cancerogenesis. Focusing on glioblastoma, the most common malignant tumour of the adult central nervous system, we assessed the processes involved in tumour progression by comparing the DNA methylation profiles of primary and recurring tumours in 112 patients. We found that primary and recurring tumours display considerable variability in their subtype compositions and identified subtype specific epigenome regulatory patterns and a loss of DNA methylation in Wnt signalling genes during progression. Our results chart the dynamics of DNA methylation in the progression of glioblastoma and establish the feasibility of conducting a DNA methylation study on samples collected in a routine clinical setting.

Finally, the field of regenerative medicine has been actively researching the possibility to replenish lost pancreatic beta cell mass in diabetic patients by reprogramming the identity of other, more abundant cell types to beta cells. To support these efforts, we generated and analysed transcriptomes of 64 healthy human pancreatic islet cells by single-cell sequencing. We were able to identify four endocrine and two exocrine human pancreatic islet cell types, which allowed the generation of accurate, cell type specific expression profiles. In a subsequent study assessing the transdifferentiating potential of small molecule drugs, these expression profiles then served as reference to identify characteristic changes in the expression profiles of alpha cells upon treatment with Artemisinins.

Taken together, the work presented in this thesis contributes biologically and medically relevant results to the understanding of cellular identity. Moreover, it emphasizes and promotes the promising potential of recent high throughput sequencing technology for the advancement of this field.

V. Zusammenfassung

Zellen sind die elementaren Einheiten des Lebens. Die Fähigkeit, funktionell spezialisierte und kooperierende Zelltypen aufrecht zu erhalten, bildet die Grundlage für die Existenz multizelluläre Organismen. Einer der faszinierendsten Aspekte multizellulären Lebens ist die Ausbildung morphologisch und funktionell unterschiedlicher Zelltypen auf Basis derselben genomischen DNA-Sequenz. Molekularbiologische Forschungsbemühungen haben das Zusammenspiel zwischen Genom, Epigenom und Transkriptom als zentral für die Regulation der zellulären Identität identifiziert. Technologische Fortschritte in der Hochdurchsatz-Sequenzierung bieten die Werkzeuge, diese Mechanismen im Detail zu untersuchen.

Trotz des wissenschaftlichen Fortschritts sind viele fundamentale Fragen bisher unbeantwortet geblieben. Zum Beispiel haben Untersuchungen in Mensch und Maus gezeigt, dass die DNA-Methylierung in Säugetieren wichtig für die Festlegung zellulärer Identität ist, was über alle Vertebraten hinweg konserviert zu seinen scheint. Invertebraten jedoch zeigen grundlegend andere DNA-Methylierungsmuster, und in manchen Spezies ist überhaupt keine DNA-Methylierung messbar. Dies wirft die Fragen auf, wie und warum DNA-Methylierung in Vertebraten seine zelluläre Identität definierende Funktion erlangt hat, was genau die Rolle der DNA-Methylierung in Invertebraten ist, und wie manche höheren Organismen ohne detektierbare DNA-Methylierung existieren können. Diese evolutionären Fragen zu beantworten könnte bisher unbekannte Mechanismen identifizieren, die in der Festlegung zellulärer Identität eine Rolle spielen, oder zur Entdeckung neuer Funktionen der DNA-Methylierung führen, die möglicherweise auch für die humane Physiologie relevant sind. Um die Forschung in dieser Richtung voranzubringen haben wir eine computerbasierte Analysemethode (RefFreeDMA) entwickelt. Diese Methode ermöglicht differentielle DNA-Methylierungsanalysen auch ohne Referenzgenome, womit die Untersuchung von DNA-Methylierung in praktisch jeder Spezies möglich wird. Wir haben unseren Ansatz erfolgreich in drei Spezies (Mensch, Rind und Karpfen) validiert und planen diese Analysemethode zur Untersuchung gewebespezifischer DNA Methylierung in vielen weiteren Vertebraten- und Invertebraten-Spezies zu verwenden.

Im Kontext humaner Erkrankungen ist die klinische Relevanz von Veränderungen der zellulären Identität gut etabliert. Fehlentwicklungen zellulärer Identität wurden als fundamentaler Faktor in der Entstehung von Krebs erkannt. Mit Fokus auf dem Glioblastom, dem häufigsten bösartigsten Tumor des adulten zentralen Nervensystems, haben wir in der Tumorprogression involvierte Prozesse untersucht, indem wir DNA-Methylierungsprofile von primären und rezidivierten Tumoren in 112 Patienten vergleichen haben. Diese Untersuchung hat ergeben, dass sowohl primäre als auch rezidivierte Tumore eine erhebliche Heterogenität in ihrer Subtyp-Zusammensetzung zeigen, dass sich die Glioblastom-Subtypen durch distinkte Epigenom-regulatorische Signaturen unterscheiden lassen und dass im Laufe der Tumorprogression eine Reduktion der DNA-Methylierung in Genen der Wnt-Signalkaskade auftritt. Unsere Ergebnisse beschreiben die Dynamik der DNA-Methylierung in der Progression von Glioblastomen und etablieren die Machbarkeit von DNA-Methylierungs-Studien, die auf im klinischen Routinebetrieb gesammelten Proben basieren.

Im Bereich der regenerativen Medizin hat sich die Forschung intensiv mit der Möglichkeit beschäftigt, verlorengegangene Betazellen in Diabetes-Patienten durch die Umwandlung von anderen pankreatischen Inselzellen zu ersetzen. Um diese Bemühungen zu unterstützen, haben wir Transkriptome von 64 gesunden, humanen, einzelnen, pankreatischen Inselzellen produziert und ausgewertet. Wir konnten vier endokrine und zwei exokrine pankreatische Typen von Inselzellen identifizieren, was das Erstellen von akkuraten, zelltypspezifischen Expressionsprofilen ermöglichte. In einer darauf aufbauenden Studie, die das Transdifferentierungspotential von Wirkstoffen kleiner molekularer Größe untersuchte, dienten diese Expressionsprofile dann als Referenz, um charakteristische Veränderungen in den Expressionsprofilen von mit Artemisininen behandelten Alphazellen festzustellen.

Zusammen genommen trägt die hier präsentierte Arbeit biologisch und medizinisch relevante Ergebnisse zum Verständnis der zellulären Identität bei. Sie betont und fördert außerdem das Potential neuartiger Hochdurchsatz-Sequenzierungstechnologien für die biomedizinische Forschung.

VI. Publications arising from this thesis

*Shared first authorship

Sheffield NC, Pierron G, Klughammer J, Datlinger P, Schönegger A, Schuster M, Hadler J, Surdez D, Guillemot D, Lapouble E, Freneaux P, Champigneulle J, Bouvier R, Walder D, Ambros IM, Hutter C, Sorz E, Amaral AT, de Álava E, Schallmoser K, Strunk D, Rinner B, Liegl-Atzwanger B, Huppertz B, Leithner A, de Pinieux G, Terrier P, Laurence V, Michon J, Ladenstein R, Holter W, Windhager R, Dirksen U, Ambros PF, Delattre O, Kovar H, Bock C, Tomazou EM. **DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma.** Nat Med. 2017 Mar;23(3):386-395. doi:10.1038/nm.4273.

Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. **Pooled CRISPR screening with single-cell transcriptome readout.** Nat Methods. 2017 Mar;14(3):297-301. doi:10.1038/nmeth.4177.

Li J, Casteels T, Frogne T, Ingvorsen C, Honoré C, Courtney M, Huber KV,Schmitner N, Kimmel RA, Romanov RA, Sturtzel C, Lardeau CH, Klughammer J, Farlik M, Sdelci S, Vieira A, Avolio F, Briand F, Baburin I, Májek P, Pauler FM, Penz T, Stukalov A, Gridling M, Parapatics K, Barbieux C, Berishvili E, Spittler A, Colinge J, Bennett KL, Hering S, Sulpice T, Bock C, Distel M, Harkany T, Meyer D, Superti-Furga G, Collombat P, Hecksher-Sørensen J, Kubicek S. **Artemisinins Target $GABA_A$ Receptor Signaling and Impair α Cell Identity**. Cell. 2016 Jan 12;168(1-2):86-100.e15. doi:10.1016/j.cell.2016.11.010.

Farlik M*, Halbritter F*, Müller F*, Choudry FA, Ebert P, Klughammer J, Farrow S, Santoro A, Ciaurro V, Mathur A, Uppal R, Stunnenberg HG, Ouwehand WH, Laurenti E, Lengauer T, Frontini M, Bock C. **DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation**. Cell Stem Cell. 2016 Dec 1;19(6):808-822. doi:10.1016/j.stem.2016.10.019.

Tschurtschenthaler M*, Kachroo P*, Heinsen FA, Adolph TE, Rühlemann MC, Klughammer J, Offner FA, Ammerpohl O, Krueger F, Smallwood S, Szymczak S, Kaser A, Franke A. **Paternal chronic colitis causes epigenetic inheritance of susceptibility to colitis.** Sci Rep. 2016 Aug 19;6:31640. doi:10.1038/srep31640.

Mass E*, Ballesteros I*, Farlik M*, Halbritter F*, Günther P, Crozet L, Jacome-Galarza CE, Händler K, Klughammer J, Kobayashi Y, Gomez-Perdiguero E, Schultze JL, Beyer M, Bock C, Geissmann F. **Specification of tissue-resident macrophages during organogenesis.** Science. 2016 Sep 9;353(6304). pii: aaf4238. doi:10.1126/science.aaf4238.

Li J*, Klughammer J*, Farlik M*, Penz T*, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S. **Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types.** EMBO Rep. 2016 Feb;17(2):178-87. doi:10.15252/embr.201540946.

Klughammer J, Datlinger P, Printz D, Sheffield NC, Farlik M, Hadler J, Fritsch G, Bock C. **Differential DNA Methylation Analysis without a Reference Genome.** Cell Rep. 2015 Dec 22;13(11):2621-33. doi:10.1016/j.celrep.2015.11.024.

Farlik M*, Sheffield NC*, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, Bock C. **Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics.** Cell Rep. 2015 Mar 3;10(8):1386-97. doi:10.1016/j.celrep.2015.02.001.

VII. Acknowledgements

I wish to express my gratitude to all the many people and institutions that have supported me in various ways on my scientific journey. I am aware of how lucky I am to pursue my PhD studies in the incredibly friendly, supportive, and scientifically outstanding environment of CeMM.

In particular, I thank:

Christoph for being my PhD supervisor – I could not have wished for a better one!

All current and past members of the Bock lab, especially Nathan, for interesting discussions, valuable advice, and fruitful collaborations.

Andreas, Cecilia, Fio, Kseniya, and Michael – dear friends who have shared this journey with me and without whom it would not have been as enjoyable.

Simon for his endless patience and understanding.

My family, especially my parents and my grandma, for encouragement when needed and not taking offense in my limited presence.

This PhD thesis was supported by a DOC fellowship of the Austrian Academy of Sciences and a PhD fellowship of the German National Academic Foundation (Studienstiftung des deutschen Volkes).

VIII. Abbreviations

2-HG: 2-hydroxyglutarate

5mC: 5-methylcytosine

ATAC-seq: Assay for transposase-accessible chromatin sequencing

BCR: B cell receptor

Bio-CAP-seq: Biotinylated CxxC affinity purification followed by sequencing

bp: base pair

cDNA: Complementary DNA

CGI: CpG island

ChIP-seq: Chromatin immune precipitation followed by sequencing

CIMP: CpG island hypermethylator phenotype

CLP: Common lymphoid progenitor

CMP: Common myeloid progenitor

CNV: Copy number variation

CRISPR: Clustered regularly interspaced short palindromic repeats

DKK: Dickkopf (gene)

DNase-seq: DNase hypersensitivity sequencing

FFPE: Formalin fixed, paraffin embedded

FPKM: Fragments per kilobase per million (reads)

HSC: Hematopoietic stem cell

IDH1: Isocitrate dehydrogenase 1

Indel: Short insertion or deletion

LRR: Leucine-rich repeat

MBD: Methyl-CpG Binding Domain

MeDIP-seq: Methylated DNA immunoprecipitation followed by sequencing

MHC: Major histocompatibility complex

MPP: Multipotent progenitor

mRNA: Messenger RNA

ncRNA: Non-coding RNA

NMI: Non-methylated island

PCR: Polymerase chain reaction

Poly-A: Poly-adenylated

RAG1/2: Recombination Activating Gene 1/2

RPKM: Reads per kilobase per million

RRBS: Reduced representation bisulfite sequencing

rRNA: Ribosomal RNA

SNV: Single nucleotide variant

TCGA: The Cancer Genome Atlas (project)

TCR: T cell receptor

TET: ten-eleven translocation (proteins)

TFBS: transcription factor binding site

TLR: Toll like receptor

TPM: Transcripts per million

UMI: Unique molecular identifier

V(D)J: variable (V), diversity (D) and joining (J) gene segments

VLR: variable lymphocyte receptor

WGBS: Whole genome bisulfite sequencing

WGS: Whole genome sequencing

Wnt: Wingless-Type (gene)

1 Introduction

The research presented in this thesis spans several topics of biology and biomedicine, including brain cancer, pancreatic islets, and vertebrate evolution. As a common theme, all these lines of research contribute to the better understanding of cellular identity as a fundamental biological principle through molecular characterisation on the genomic, epigenomic, and transcriptional level. This chapter introduces relevant evolutionary, developmental, and regulatory aspects of cellular identity, as well as the state-of-the-art sequencing based methods that enable profound molecular characterisation of cells and tissues.

1.1 What is cellular identity?

The logical prerequisite to the concept of "cellular identity" was the realisation that all living organisms are composed of cells. "Cell theory" was first formulated in the years 1838/1839 by Schleiden and Schwann about 180 years after Hooke first coined the term "cell" in 1665 when describing the microscopic structure of a slice of cork (Mazzarello, 1999). With improvements in microscopy, naturalists soon observed cellular substructures and began to morphologically characterize different types of cells. By the end of the 19th century the concept of cellular hierarchies and identities emerged and was consolidated through experiments on hematopoietic differentiation by Till and McCulloch in the 1960s (Daley, 2015).

While the "cell theory" is no longer considered a theory but has rather acquired the status of an established fact, the concept of cellular identity remains vague in its evolutionary, developmental, and regulatory details. Through modern molecular biology technique, including high-throughput sequencing, we now have the necessary tools at hand for broad as well as deep assessment of key molecular determinants of cellular identity, including the genome, epigenome, and transcriptome. However, in the light of assessing millions of different cells in millions of conditions at high resolution, trying to elucidate cellular identity in detail prompts us to consider the closely related concept of "cell state". The terms "cell state" and "cell(ular) identity" are often used interchangeably and biologically the concepts are not well discriminated. Therefore, the next section clarifies the use and meaning of these concepts at least for the scope of this thesis.

1.1.1 Cell state versus cellular identity

Differentiating between cellular identity and cell state is more than a semantic problem. In fact, establishing a useful and meaningful differentiation is necessary and achievable. Intuitively one might define "cell state" as transient and "cellular identity" as long-lasting properties of a cell. In this logic, for a given cell, its cellular identity comprises all its possible cell states, which manifest as cellular phenotypes, and the cell can reversibly change its cell state without changing its identity. While a cell can only have one cellular identity at a time, cell states are not necessarily mutually exclusive. This relatively simple definition, which emphasises the temporal dynamics of properties also accommodates the fact that large and persistent changes in cell state can lead to changes in cellular identity, as is observed in de-, re-, and trans-differentiation (Efroni *et al*, 2015). During the process of differentiation a cell acquires its identity while potentially running through a series of identity changes at each asymmetric stem-cell division (Knoblich, 2008). As a simplified example, a neural progenitor cell divides asymmetrically and thereby produces one cell that retains the identity of a neural progenitor cell as well as another cell with a new identity that develops into a functional neuron. Both cellular identities (progenitor and neuron) comprise a distinct set of states, the manifestation of which does not change the identity. For example, the progenitor cell can be in a dividing or quiescent state, while the neuron never divides but can be in an excited or resting state [Fig. 1]. This concept of cellular identity classifies cells into biologically meaningful groups, also referred to as cell types, that are not confounded by

transient effects and distinguishes common long-term properties. However, the definition of a cell's identity might go further and even comprise a track record of a cell's origin, including cellular ancestors, place and time of appearance as well as uniquely identifiable marks such as distinct sets of (somatic) mutations. Depending on the properties one includes in the definition of cellular identity, the groups of cells perceived as "identical" change in inclusiveness, size, and heterogeneity. This flexibility allows to adjust the exact definition to the studied system and research question. For example, defining appropriate sets of common properties is particularly important in the study of diseases, when the responsible cells need to be held accountable in order to understand and appropriately treat the disease. A prominent example is the characterisation of cells of origin in malignant diseases (Visvader, 2011), where a more narrow definition appears appropriate. On the other hand, a classification at the level of cell type seems sufficient for the identification of non- or dysfunctional cells in diseases with a cellular cause like type I diabetes where the symptoms are caused by the mostly immune-mediated death of beta-cells (Atkinson *et al*, 2014).

This flexible definition of cellular identity together with the distinction of cellular identity and cellular state is in line with current opinion (Wagner *et al*, 2016) and provides biologically and medically meaningful classifications and is therefore sufficient for the scope of this thesis, although it is not all-encompassing. For example in the case of mono-cellular organisms (bacteria and protozoa) acquiring cellular identity through differentiation does not generally apply since each single cell represents an entire viable organism. For many mono-cellular organisms, cellular identity is therefore rather associated with speciation than specialisation and lineage hierarchies should be seen rather in a phylogenetic than ontogenetic sense (Stoeck, 2005; Yubuki *et al*, 2009).

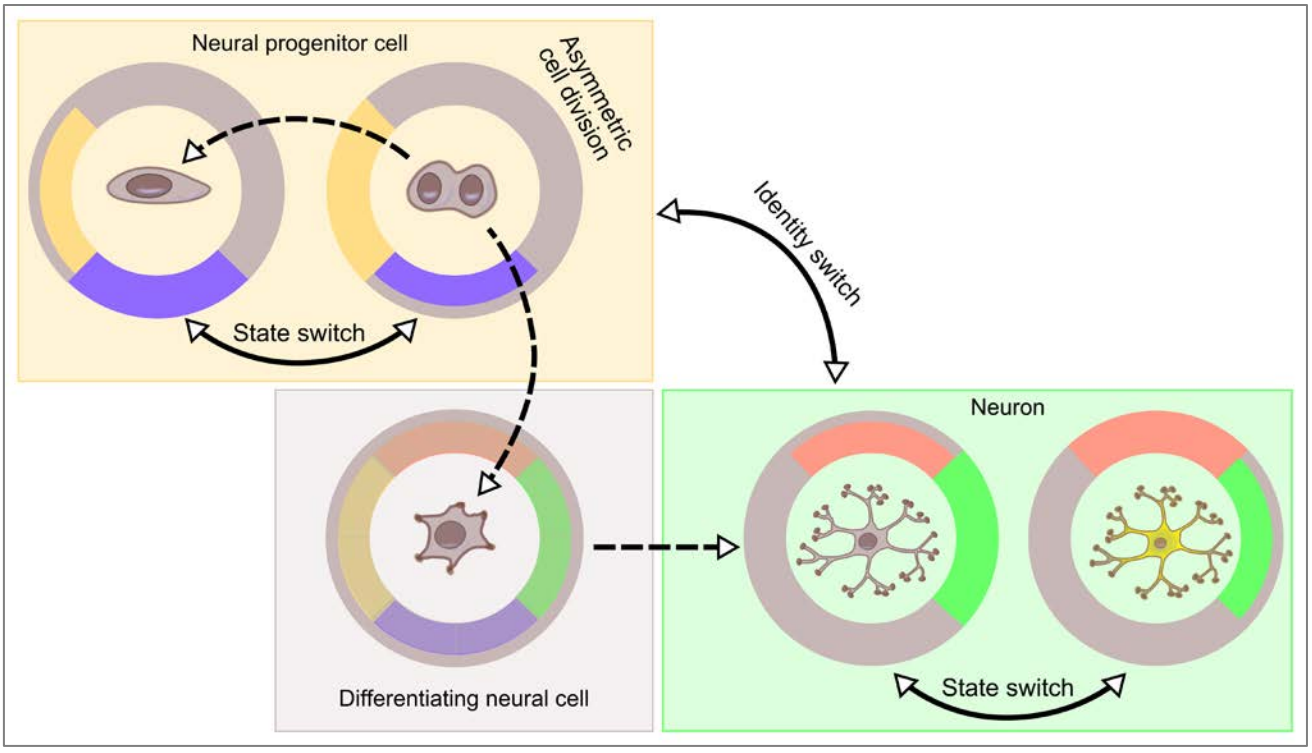


Figure 1 Schematic to illustrate the distinction between cell state and cellular identity as used in this thesis. The different colors of the inner circle represent the different states a cell can engage, with grey indicating that the state is not available. The outer circle signifies whether the respective state is active. Colors: blue: quiescent state; orange: dividing state; green: resting state; red: excited state.

1.1.2 Evolutionary aspects of cellular identity

The evolution of cellular identity is closely related to the evolution of the cell itself (Cooper, 2000) and involves three major transitions: The emergence of eukaryotic cells (Woese & Fox, 1977; Woese, 1998), the emergence of multicellular organisms (Hedges *et al*, 2004), and the emergence of vertebrates (Holland *et al*, 2015). Each of these evolutionary milestones goes along with a marked increase in complexity (Bird, 1995; Cooper, 2000). Eukaryotic cells display higher cellular complexity than prokaryotic cells (e.g. organelles), multicellular organisms display higher organisational complexity than unicellular organisms (e.g. cell signalling and specialisation), and vertebrates display higher organismal complexity than invertebrates (e.g. immune and nervous system). This view of discrete increases in complexity is useful to emphasize evolutionarily important common features in groups of organisms, but to some degree violates the mostly continuous nature of evolution. Therefore, it is not surprising that one can find examples that do not quite fit into this discrete scheme. These examples include simple yet multicellular prokaryotes such as some cyanobacteria (Flores & Herrero, 2010), signalling and interaction between unicellular organisms (e.g. biofilms and quorum sensing), differentiation in unicellular organisms such as yeasts (Herskowitz, 1989), and invertebrates with vertebrate like intelligence such as cephalopods (Kröger *et al*, 2011; Roth, 2015).

The definition and evolution of biological complexity is a difficult topic, but a trend towards higher complexity even in the absence of positive selection seems to widely accepted (Adami, 2002; Lukeš *et al*, 2011). Also, a connection between an organism's complexity and its number of different cell types is rather evident (Arendt, 2008). As a molecular measure of biological complexity, the number of distinct genes rather than the size of an organism's genome has been proposed in a time when genome sequencing was still in its infancy and the number of genes could only be estimated (Bird, 1995). This early hypothesis has since been challenged: An ever increasing number of sequenced genomes showed that organisms with similar levels of complexity display highly variable numbers of genes, which led to the proposal of alternative measures of biological complexity, stressing in particular the importance of non-protein coding DNA (Taft *et al*, 2007; Jiang & Xu, 2010). While a strictly linear relationship between complexity and gene number is indeed very unlikely, recent genomic data do support the originally proposed increases in gene number at the prokaryote/eukaryote and invertebrate/vertebrate boundary (Bird, 1995) [Fig. 2]. The emergence of novel epigenomic mechanisms to control gene expression noise (nucleus/chromatin and histones in eukaryotes and DNA methylation in vertebrates) (Bird, 1995) as well as increased energy supply through endosymbionts (mitochondria and chloroplasts) (Lane & Martin, 2010) are thought of as crucial for these increases in gene number. The large amount of genome sequencing data that is available today confirms the trend toward larger genomes and greater numbers of genes in increasingly complex groups of organisms, with increases in genome size at complexity boundaries being more clear-cut than the increase in gene number [Fig. 2]. As a consensus and with respect to the recently established regulatory importance of non-coding genomic regions (Dunham *et al*, 2012), it seems plausible that a certain genome size together with a certain number of distinct genes only enables but not enforces a certain level of cellular, organisational, or organismal complexity. This concept accommodates the fact that organisms of similar complexity often display large differences in genome size and gene number, while on average the genome size and gene number do increase with increasing complexity [Fig. 2].

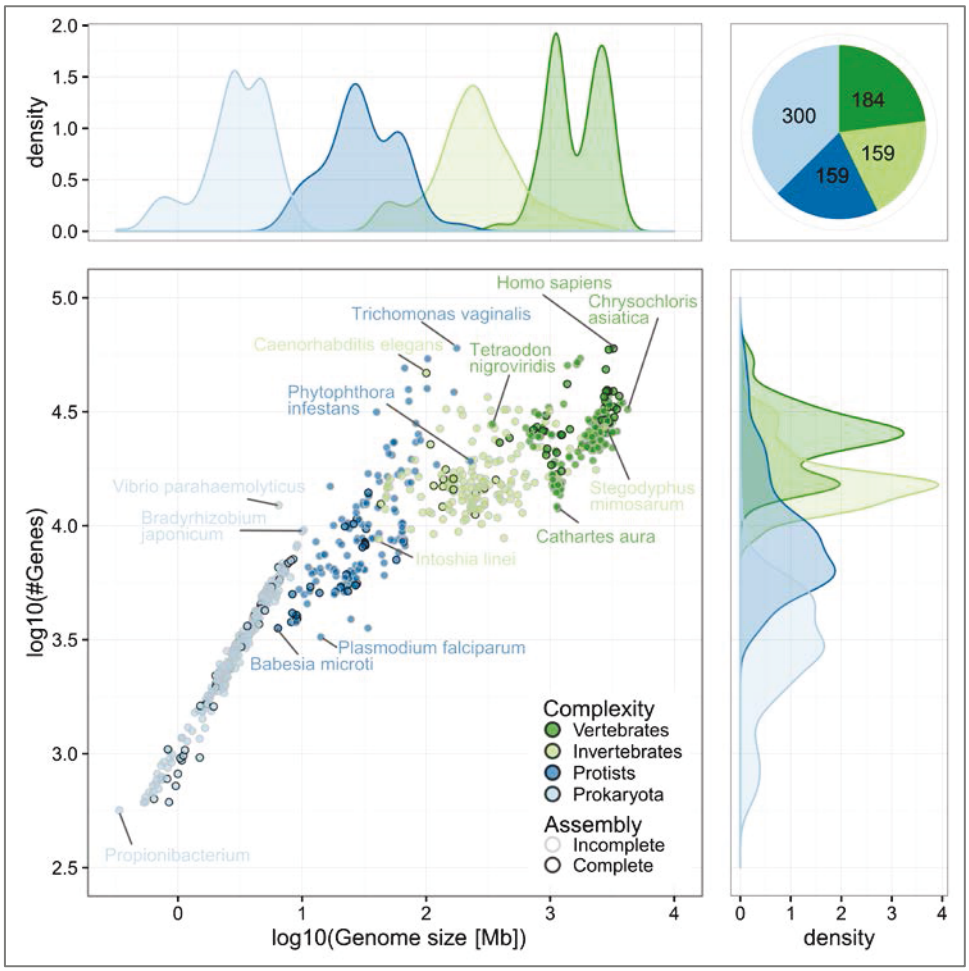


Figure 2 Relationship between genome size and number of distinct genes to illustrate the increase in average genome size as well as average gene number in groups of increasing complexity level. For each complexity level the most extreme species in genome size and gene number are annotated. The data were obtained from https://www.ncbi.nlm.nih.gov/genome/browse/ state 11.11.2016. Of the large number of prokaryote genomes available 300 were randomly selected. The pie chart shows the number of species included for each complexity level.

The ability to control the expression of a large number of genes is necessary for the production of different cellular phenotypes because each cellular phenotype is characterized by the expression of a distinct set of protein complexes that enable particular cellular functions (Arendt *et al*, 2016). The ability to produce and maintain different cellular identities in turn is a prerequisite for specialization among the cells of a multicellular organism. Therefore, the evolution of genomes, gene regulatory mechanisms and cellular identity are tightly interrelated and the emergence of new cell types and cellular systems is accompanied by novel regulatory signatures (Arendt *et al*, 2016). To illustrate, let us take a closer look at the emergence of the adaptive immune system in vertebrates, which entailed the introduction of novel specialized cell types. Both invertebrates and vertebrates possess innate immune systems with striking molecular (e.g. TLRs) and cellular (e.g. phagocytes) similarities, which detect and eliminate pathogens in a “hard-coded” and therefore fast but less flexible manner (Loker *et al*, 2004; Ottaviani, 2011). In addition to and in close interaction with innate immunity, vertebrates have developed two convergent adaptive immune systems as a means of more flexible pathogen defence. Both extant forms of adaptive immunity (VDJ-based in jawed and LRR-based in jawless vertebrates) involve new sets of cells (T/B lymphocytes in jawed and VLRA/VLRB lymphocytes in jawless vertebrates) as well as new sets of genes (RAG1/2, BCR/TCR, MHC in jawed and LRR, VLR in jawless vertebrates) that have no evident counterparts in invertebrates (Cooper & Alder, 2006; Litman *et al*, 2010) [Fig. 3]. In jawed vertebrates, lymphoid cells (the cells of the adaptive immune system) and myeloid cells (the cells of the innate immune system) arise from the same hematopoietic stem cells (HSCs) (Kondo *et al*, 2001). This implies the emergence of a novel regulatory mechanism in vertebrates that enhances differentiation along the lymphoid path and prevents differentiation along the evolutionarily older myeloid path. Interestingly, experiments in mice showed that without appropriate DNA methylation HSCs can give rise to myeloid but not lymphoid cells (Bröske *et al*, 2009), indicating that DNA methylation might be part of the novel regulatory mechanism in vertebrates that enables lymphoid differentiation.

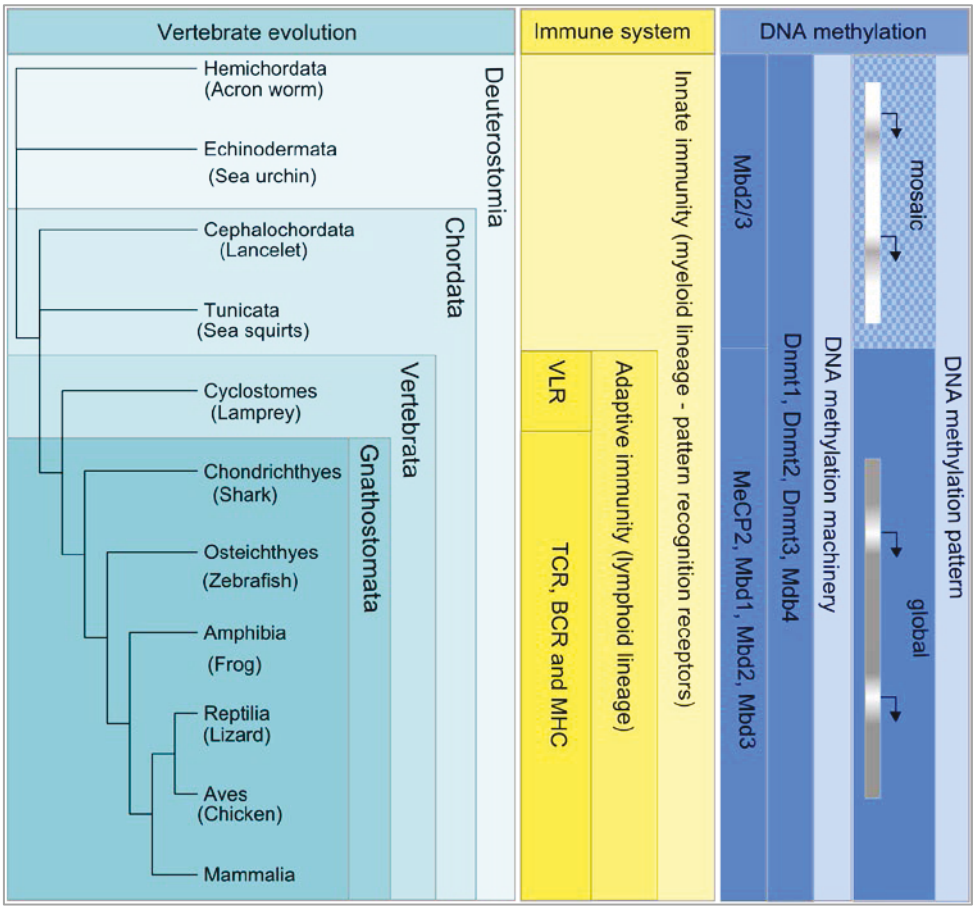


Figure 3 Schematic summarizing the evolutionary relationship between innate immunity, adaptive immunity, the DNA methylation machinery, and genome-wide DNA methylation patterns (grey: methylated, white: unmethylated, arrow: transcription start site/promoter).

1.1.3 Developmental aspects of cellular identity

Although multicellularity has evolved independently in most eukaryotic lineages (e.g. animals, plants, slime molds), all multicellular organisms start out as only one cell and develop through many rounds of mitosis while cells stay associated after division (Miller, 2010). Animals can consist of as little as 3 cell types (*Microhydra rideri*) and 20 cells (*Dicyemmenea abelis*) or as many as 122 cell types (*Morone saxatilis*) and $10^{13.7}$ cells (*Canis familiaris*) (Bell & Mooers, 1997). For humans, even as many as 411 different cell types (including 145 types of neurons) have been described (Vickaryous & Hall, 2006), but of course this number is highly dependent on how different cell types are defined. In humans (and many other sexually reproducing metazoans), organismal development begins with the totipotent zygote (fertilized egg), which gives rise to embryonic as well as extraembryonic structures. In the first stages of embryogenesis during gastrulation three germ layers (ectoderm, mesoderm, and endoderm) form (Solnica-Krezel & Sepich, 2012) that each give rise to distinct lineages of somatic cell types. As the result, each adult cell type can be assigned to its germ layer of origin (Vickaryous & Hall, 2006). Furthermore, lineage-tracing experiments based on direct observation in *Caenorhabditis elegans* first allowed researchers to document the exact fate of each cell in an entire yet simple organism (Sulston & Horvitz, 1977). Later, labelling-based approaches were used to attempt the same in more complex organisms, successfully charting complex organs such as the brain (Kretzschmar & Watt, 2012). Only recently, a proof-of-concept study in a zebrafish embryo showed that labelling through genome editing might in future allow the reconstruction of lineage relationships even throughout an entire vertebrate organism (McKenna *et al*, 2016). These experiments reveal the hierarchical organisation of ontogenesis with increasing specialisation and determination of cellular identity from root to leaves. Complementing single-cell RNA sequencing experiments can catch cells in temporary, intermediate transcriptional states and thereby unveil the trajectory of differentiation as exemplified in developing lung epithelium (Treutlein *et al*, 2014).

Cellular differentiation (i.e., the process of establishing cellular identity) is a central process in the developing as well as the adult organism. In both cases, stem cells are the point of origin. Stem cells usually do not differentiate themselves, but they can divide asymmetrically giving rise to cells retaining stem cell identity (self-renewal) and to more lineage-committed daughter cells, which follow certain differentiation trajectories (Knoblich, 2008). One of the best-understood adult differentiation systems is that of haematopoiesis. Haematopoietic stem cells (HSCs) have long been known to replenish blood cells after transplantation but only recent lineage tracing experiments in mice show that HSCs indeed give rise to all major hematopoietic lineages under homeostatic conditions (Sawai *et al*, 2016). Interestingly, HSCs themselves are not the most proliferative cells. Instead it has been shown that the most potent HSCs actually divide the least (only 5 times in a mouse life), while the majority of dividing multipotent hematopoietic cells are the more lineage committed multipotent progenitors (MPPs) (Wilson *et al*, 2008) and the lineage restricted common myeloid progenitors (CMPs) and common lymphoid progenitors (CLPs) (Höfer & Rodewald, 2016). CMPs and CLPs give rise to cells that then finally differentiate into functional myeloid (granulocytes, monocytes, megakaryocytes, and erythrocytes) and lymphoid (B cells, T cells, and NK cells) cells respectively. This principle of separation between potency and amplification is also found in other stem/progenitor cell compartments that frequently need to replenish lost cells (e.g. the gut and the skin) and is thought of as a means to protect the most valuable “backup” cells from mutational damage.

The ability to regenerate after injury varies widely between different tissues and species, depends on the age of the organism, and crucially involves immune cells - in particular tissue resident macrophages (Forbes & Rosenthal, 2014). Generally, the ability to regenerate seems to decline with age and with increasing complexity. While for example the neonatal mammalian heart and the adult fish heart efficiently replenish lost cardiomyocytes, the adult mammalian heart fails to do so. In contrast, the adult mammalian liver can easily tolerate a 70% loss and reconstitute the lost mass within weeks. Mostly, tissue specific stem cells are attributed a major role in tissue regeneration, and their limited occurrence entails limited regenerative potential, as seems to be the case in the adult mammalian heart (Forbes & Rosenthal, 2014). Interestingly, liver regeneration does not rely on stem cells, but is achieved through terminally differentiated, functional and primarily quiescent hepatocytes that re-enter the cell cycle on demand (Michalopoulos, 1997). Similarly, in zebrafish (Poss, 2002) and neonatal mice (Porrello *et al*, 2011) heart regeneration is achieved through proliferating pre-existing cardiomyocytes. These findings imply that the cellular identity of neonatal mammalian cardiomyocytes still comprises the proliferative state, while the ability to enter this state is lost in adult cardiomyocytes.

Terminally differentiated cells usually execute their functions as defined by their identity and die after a cell-type specific time-period. Nevertheless, under certain conditions cells can actually change their identity, either through dedifferentiation followed by redifferentiation or through direct transdifferentiation. In vitro, it is for example possible to dedifferentiate human fibroblast to a pluripotent state through the induced expression of only three transcription factors (Oct3/4, Sox2, and Klf4) (Nakagawa *et al*, 2007). In vivo, physiological changes in cellular identity mostly happen in the context of regeneration. Complete limb and tail regeneration in the adult Mexican axolotl has been proposed to include dedifferentiation of muscle and dermis cells to form the blastem, an undifferentiated cell mass from which cells re-differentiate to form all missing structures (skin, muscle, cartilage). Although, this regenerative process even appears to includes ectoderm to mesoderm lineage switching (Echeverri, 2002), a later study from the same research group showed that blastem forming cells do keep a memory of their tissue of origin and that complete dedifferentiation is not necessary for regeneration (Kragl *et al*, 2009). In human, plasticity of differentiated pancreatic islet cells is highly studied in the context of diabetes and the regeneration of lost insulin producing cells (beta cells). Indeed, the cells of the mammalian pancreatic islets display astonishing plasticity in that especially alpha (Thorel *et al*, 2010) but also delta and even ductal cells (Romer & Sussel, 2015) can transdifferentiate to beta cells in the case of severe beta cell loss. For humans, beta cell regeneration has been reported but seems to be less efficient compared to mice (Thorel *et al*, 2010; Li *et al*, 2017).

Given the obvious benefits of cellular plasticity for regeneration, why do not all differentiated cells retain the ability to re-enter cell cycle, dedifferentiate, or transdifferentiate? Apparently, there must be costs attached to cellular plasticity that make the stable maintenance of cellular identity more favourable for most cell types of the adult mammalian organism. One potential cost might be the uncontrolled loss and aberration of cellular identity as is observed in malignant transformation (Roy & Hebrok, 2015). In fact, the uncontrolled proliferative state is what all malignant diseases have in common and what makes them incompatible with homeostasis and ultimately life (Hanahan & Weinberg, 2011). However, similar to non-malignant tissues, malignant neoplasms do not entirely consist of highly proliferative cells. Often they display clonal hierarchies with few self-renewing stem cell like cells at the root and proliferative, phenotypic, and (epi-) genomic heterogeneity between the cells that make up the vast majority of the neoplastic cell population (Nguyen *et al*, 2012). Therapies are directed against proliferating cells in general (e.g. Antimetabolites and DNA alkylating agents) or against specific targets such as BCR-ABL fusion protein in chronic myelogenous leukemia. Unfortunately, therapy induced selective pressure applied on a heterogeneous population of malignant cells more often than not leads to the emergence of therapy resistant cells and therefore relapse of the disease. This phenomenon has prompted the idea to use combination therapies, similar to the treatment of HIV, in order to attack and control the most dangerous sub-populations of malignant cells in each treatment round (Bock & Lengauer, 2012).

Glioblastoma is an example for a malignant neoplasm that rapidly develops therapy resistance in most patients. Glioblastomas are highly invasive brain tumours of astrocytic origin with a 5-year progression free survival rate of just 4.1% (overall survival rate 9.8%) when treated with standard of care (maximal safe resection, radiotherapy, temozolomide) (Stupp *et al*, 2009). This extremely low progression free survival rate is only partially due to the inherent difficulty of completely resecting an invasive tumour from the brain. Additionally, glioblastomas display intra-tumour heterogeneity in the three major molecular determinants of cellular identity (genome, epigenome, and transcriptome) leading to the observed variable and mostly unfavourable treatment result (Sottoriva *et al*, 2013; Parker *et al*, 2016). Understanding and characterizing the deregulation and subsequent decay of cellular identity in glioblastoma on a molecular level might lead to more targeted treatment approaches with the ultimate goal to prevent tumour progression.

1.2 Regulation and determination of cellular identity

1.2.1 Genome

Given that genomes are largely identical between cells within an organism, and even between individuals of the same species, the importance of the genome in the regulation and determination of cellular identity might not be evident. However, there are plenty of rather subtle differences between genomes, which critically shape and reveal individual as well as cellular identity. These genomic variants can be broadly grouped into three classes: Single nucleotide variants (SNVs), short insertions and deletions (indels), and structural variants (SVs) including copy number variation (CNV). Genomic variation within the human population has been mapped by the 1000 Genomes Project, which analysed genomes of more than 2500 individuals and revealed that a typical genome differs from the reference genome in about 4.5 million sites (0.1% of the genome) (Auton *et al*, 2015). While most of these variants are common in the human population, some are private to subpopulations, families, or even individuals and can thus be used to infer identity (Erlich & Narayanan, 2014).

On a cellular level, somatic variants (i.e. variants that are not inherited but arise spontaneously during the course of a lifetime) can be used to infer cellular identity. For example, immune receptor diversity of jawed vertebrates is produced through genomic rearrangements in the BCR and TCR genes. These genomic rearrangements randomly select certain segments to be part of the immune receptor and remove others from the genome, which makes the selection irreversible and leaves each cell with

an individual combination of segments. Furthermore, the joining of the different segments involves random mutations, and B cells additionally undergo somatic hypermutation to further increase BCR diversity (Hood *et al*, 1985; Teng & Papavasiliou, 2007). Through the many random elements in the generation of BCRs and TCRs, each B- and T-cell is equipped with an individual version of the respective immune receptor. The specific configuration of an immune cell's immune receptor is an important part of its identity because only cells that express receptors that do not recognize self-antigens will be allowed to survive, and only those that recognize danger antigens will receive proliferative signals in order to elicit the adaptive immune response. Using the immune receptor as identity markers has also proven useful for detecting and monitoring many B- or T-cell malignancies. These malignancies usually arise from only one malignantly transformed cell resulting in the expression of the same TCR or BCR version in all descendant malignant cells, which in turn makes clonal BCR/TCR expression an important diagnostic marker (Boyd *et al*, 2009). Interestingly, some unicellular organisms (prokaryotic as well as eukaryotic) such as Neisseria (Rotman & Seifert, 2014) and Trypanosoma (Horn, 2014) use similar mechanisms of controlled enhanced genomic mutation to change their surface proteins and thereby produce antigenic variation within the population. This variation allows immune evasion or other rapid adaptations to changes in the environment by providing a multiplicity of phenotypes of which the fittest in a certain environment gets to represents the majority but not the entirety of the population.

While the generation of genomic mutations during immune receptor diversification or antigenic variation are physiological, controlled, and localized processes, genomic mutations can also occur spontaneously throughout the genome. Depending on type and location, the effects of somatic mutations range from no phenotypic changes (e.g. silent mutations) to changing the identity of a cell (e.g. deleterious mutations in crucial transcription factors). Cancer genomes are known to be especially ridden with somatic mutations (Roberts & Gordenin, 2014), and large studies have identified mutational signatures that are specific for certain types of cancers and certain mutational processes (Alexandrov *et al*, 2013). In cancer genomes, relevant mutations can be classified into two major categories: Mutations that confer growth advantage (driver mutations) and mutations that coincide with driver mutations but do not confer growth advantage themselves (passenger mutations). Driver mutations are rare events and usually affect oncogenes and tumour suppressor genes, while passenger genes occur more frequently (Greenman *et al*, 2007). Because it is very unlikely that mutations revert to the original sequence, cells tend to accumulate mutations during tumour progression. Therefore, the mutational signature of an individual cell represents a logbook of its genesis that allows to infer its origin (founder cells) and position in the clonal hierarchy as well as the timing of mutational events in growing tumours (Bozic *et al*, 2010). Mutational heterogeneity (i.e. cells with different mutational profiles within the population of malignant cells) is common in most malignancies. Similar to antigenic variation in unicellular organisms, tumours often display one major clone that makes up most of the malignant population and several minor clones that might only consist of few cells. When the environment changes and favours a different set of mutations (e.g. through therapy, immune response or metastasis) minor clones often rescue the malignant neoplasm by replacing the vulnerable major clone with the effect of therapy resistance, immune evasion or spreading metastasis (Vogelstein *et al*, 2013). The knowledge of the different cellular identities within the population of malignant cells is therefore crucial for the understanding and treatment of malignant diseases.

As illustrated in the physiological as well as pathological examples above, genomic variation is majorly involved in the determination and inference of cellular identity. However, regulation of cellular identity cannot be understood by focusing on the genomic level alone. For example, genomic regulatory mutations do not always per se affect cellular identity, but manifest through their effect on transcription factor binding, enhancers, or chromatin structure (Melton *et al*, 2015; Lappalainen *et al*, 2013). Therefore, crucial aspects in the regulation of cellular identity happen at the level of and through interaction with epigenome and transcriptome, which are discussed in the following sections.

1.2.2 Epigenome

The epigenome comprises regulatory, potentially heritable modifications of the genome that do not affect the genomic sequence but rather the interpretation thereof (Bernstein *et al*, 2007). If the genome was a cookbook, the epigenome would consist of post-it notes with instructions like “For 8 people use 1.5 times the ingredients.” to scale up a recipe or “Never do that again!” to mark a particularly horrible recipe. Inherited cookbooks, might even still contain notes from earlier generations. Although the text in all copies of a certain cookbook is the same, different households will select different favourite dishes and preparation of the dishes will vary slightly. The comments in the cookbook help to ensure continuity within one household and thereby contribute to culinary identity. However, because the actual text is not altered, changes in the menu and corrections remain possible. In this analogy, different households with the same copy of a cookbook represent the different cells in an organism that are equipped with essentially the same genome but display phenotypic differences due to the way the genome is interpreted. In each cell, the epigenome specifically guides the interpretation of the genome (with gene expression as a crucial step), to enable stable cellular identity without changing the genomic sequence itself. Thereby each cell in principle retains the potential to produce any of the gene products encoded in the genome and thus display any phenotype according to the configuration of the epigenome. This potential is for example exploited *in vivo* during de-, re-, and transdifferentiation in the context of regeneration and *in vitro* for the generation of induced pluripotent stem cells.

During normal cellular differentiation, the epigenome orchestrates the shift from the expression of pluripotency genes in stem and progenitor cells to the expression of lineage specific genes in differentiating cells. By tagging the respective genes with activating or repressive marks, the epigenome plays a crucial role for the controlled expression of cell type specific genes – if disrupted inappropriately, cells fail to maintain or establish their identity. In addition, pathogenic changes in the epigenome, especially those that lead to silencing of tumour-suppressor genes or activation of oncogenes, are frequently involved in malignant transformation (Berdasco & Esteller, 2010). Similar to genetic heterogeneity, epigenetic heterogeneity is often observed in malignant neoplasms and poses similar medical challenges (Mazor *et al*, 2016). The aforementioned reversibility of changes in the epigenome bears the potential to correct pathogenic aberrations but also the risk to cause even more. Therefore, a profound understanding of the molecular mechanisms behind the epigenome is important for cancer therapy.

The epigenome is shaped by DNA methylation and histone modifications (Bernstein *et al*, 2007), but also comprises non-coding RNAs (Koerner & Barlow, 2010; Lee, 2012) and chromatin remodelling complexes. While DNA methylation directly modifies the DNA, histone modifications alter the way the DNA is packed. There are several different types of histone modifications (different histones, positions, and marks) with effects ranging from transcriptional silencing over poised transcription to active transcription initiation and/or elongation. For example, histone acetylation is generally associated with transcriptional activation, as its negative charge opens up the chromatin and makes the DNA accessible for transcription (Strahl & Allis, 2000). In contrast, DNA methylation (in eukaryotes always at the C5 position of cytosines), is predominantly associated with transcriptional repression (Schübeler, 2015). Both epigenetic modifications (DNA methylation and histone modifications) can be maintained through cell divisions, enabling propagation of cellular identity to the next cell generations. While the mechanisms behind the heritability of histone modifications are not well understood, propagation of DNA methylation has been beautifully explained: In vertebrates, DNA methylation mostly occurs at cytosines in CpG motifs. Directly after replication, a methylated CpG motif displays a methylated cytosine on the pre-existing strand, while the corresponding cytosine on the newly synthesized strand is unmethylated. Due to the sequence symmetry of CpG motives, the maintenance DNA methyltransferase DNMT1 can specifically detect such hemi methylation CpG motifs and ad the missing methylgroup to the cytosine on the newly synthesized strand (Song *et al*, 2011). Demethylation is achieved passively during replication though reduced DNMT1 activity, or actively through ten-eleven translocation (TET) family of proteins (Tahiliani *et al*, 2009).

The importance of cell type specific DNA methylation for establishing and maintaining cellular identity has been demonstrated by assessing and comparing DNA methylation in various differentiated cell types and tissues (Ziller *et al*, 2013) as well as stem and progenitor cells (Bock *et al*, 2012; Farlik *et al*, 2016). These studies revealed that cell type specific DNA methylation is highly associated with regulatory genomic elements such as promoters, enhancers, and transcription factor binding sites (TFBSs), suggesting its involvement in transcriptional regulation. Although most studies are focused on human and mouse, comparative profiling of non-methylated DNA provides evidence that the regulatory role of DNA methylation is conserved across vertebrates (Long *et al*, 2013). In the contexts of hematopoietic differentiation it was shown that hypermethylated regions in lymphoid cells (in comparison to myeloid cells) are enriched in myeloid transcription factor binding sites and vice versa. This phenomenon was observed in mammalian species as well as in fish (Klughammer *et al*, 2015), but whether the lack of transcription factor binding elicits methylation of the respective binding sites, or whether DNA methylation prevents transcription factor binding is not clear to date. Additionally, lymphoid cells have been found to display more hypermethylated regions compared to myeloid cells (Klughammer *et al*, 2015; Farlik *et al*, 2016), which complements the finding that proper DNA methylation is crucial for lymphoid differentiation but dispensable for myeloid differentiation (Bröske *et al*, 2009).

From an evolutionary prospective, the emergence of lymphoid cells in vertebrates coincides with a drastic change in genome-wide DNA methylation patterns at the invertebrate-vertebrate boundary. Although the DNA methyltransferases and some methyl-CpGbinding domain (MBD) proteins are conserved between vertebrates and invertebrates (Albalat, 2008), vertebrates display globally methylated genomes with few unmethylated regulatory sites mostly associated with CpG islands (CGIs), while invertebrates display vastly unmethylated genomes where DNA methylation is mostly found in gene bodies (Suzuki & Bird, 2008; Zemach *et al*, 2010) [Fig. 3]. This difference in genome-wide DNA methylation patterns between invertebrates and vertebrates might indicate a new regulatory role of DNA methylation in vertebrates tightly associated with cell type specific transcription and regulation of cellular identity. One approach to further investigate this hypothesis is to assess cell type (or tissue) specific DNA methylation in a wide range of invertebrates and vertebrates, which in combination with transcriptional analysis would allow to draw conclusions regarding the evolutionary role of DNA methylation in the regulation of cellular identity.

1.2.3 Transcriptome

The transcriptome is the entirety of different transcripts (i.e. RNAs) that are produced in a cell, a tissue, or an organism, depending on the level of investigation. A cell's transcriptome is the result of its interpretation of the genome and the primary display of its molecular phenotype. To create functional phenotypes, the coding part of the transcriptome can be translated into proteins, which in turn catalyse most cellular processes (CRICK, 1970). Ultimately, all regulatory or pathogenic changes to the genome and epigenome directly or indirectly affect and act through the transcriptome, making the transcriptome (together with the proteome) the instance that actually implements and manifests cellular identity. However, in contrast to the genome and epigenome, the transcriptome is highly susceptible to short-term endogenous or exogenous regulatory changes that affect the state of a cell rather than its identity. One prominent example for short-term changes in cell state that involve distinct transcriptional signatures is the cell cycle in which a cell passes through four stages (G1, S, G2, and M) until it divides (Scialdone *et al*, 2015). However, it is possible to account for the transcriptional changes that come with the cell cycle or other potentially unknown perturbances and thereby discriminate between cell state and identity, which in turn allows the characterisation of otherwise hidden subpopulations (Buettner *et al*, 2015). When cell types display very strong and specific transcriptional profiles as is the case for the cells of the pancreatic islets (Li *et al*, 2016a) the transcriptome can be informative about cellular identity and cell state at the same time. This was demonstrated by studies comparing the different pancreatic islet cells from healthy and diabetic individuals (Segerstolpe *et al*,

2016; Wang *et al*, 2016; Xin *et al*, 2016). In these studies, the strong expression of cell type specific marker genes (e.g. glucagon for alpha cells, insulin for beta cells, somatostatin for delta cells, and pancreatic polypeptide for PP cells) allowed to determine the identity of a cell, while the broader transcriptional profile was used to compare healthy and diseased (diabetic) cell states. The cell type resolved comparison of cell state (healthy vs. diabetic) led to the discovery of novel disease associate genes and pathways (Segerstolpe *et al*, 2016; Xin *et al*, 2016) as well as signs of dedifferentiation of alpha and beta cells during type 2 diabetes (Wang *et al*, 2016). The transcriptome also allows to precisely track directed changes in cellular identity and for example place transdifferentiating cells on an "identity continuum" between the original cell type and the completely transdifferentiated cell type (Treutlein *et al*, 2016). These and other studies (Treutlein *et al*, 2014; Grün *et al*, 2015; Zeisel *et al*, 2015), demonstrate that the transcriptome can be a suitable readout for cellular identity and its dynamic changes.

In order to understand the regulatory contribution of the transcriptome to creating, changing, and maintaining cellular identity, one component is of particular importance: Transcription factors and their functioning in gene regulatory networks. Transcription factors are a diverse group of proteins that have the ability to regulate the rate of transcription of their target genes. Thereby transcription factors crucially contribute to the implementation of cell type specific gene expression, which in turn is crucial to establish and maintain cellular identity. Transcription factors act through their specific binding to the DNA and their interaction with the transcription machinery, which can be activating or repressive. The specificity of transcription factors is provided through their DNA binding preferences at certain transcription factor specific sequence motifs as well as co-binding as part of protein complexes (Latchman, 1997). The activity of transcription factors themselves is regulated through the rate of their transcription, accessibility of their binding sites, interaction with other transcription factors, and post-translational modifications (e.g. phosphorylation). The processes that regulate transcription factors in turn involve proteins (e.g. kinases, other transcription factors, and chromatin modifiers) the expression of which is again regulated by transcription factors. To disentangle these gene regulatory networks and identify master transcription factors (i.e. transcription factors that alone or in small groups can induce or abolish complete transcriptional programs) has greatly aided the understanding of the regulatory processes underlying cellular identity. For example, evidence has been found that in order to confer cellular identity, the master transcription factors for ESCs (Oct4, Sox2, Nanog, Klf4 and Esrrb), B cells (PU.1, Ebf1, E2A and Foxo1), T helper cells (T-Bet), macrophages (C/EBPα), and myotubes (MyoD) occupy large enhancer regions (super-enhancers) that are exceptionally dense in transcription factor binding sites and associated with cell type specific genes (Whyte *et al*, 2013). Furthermore, expression levels of just 18 key hematopoietic transcription factors were sufficient to distinguish five types of hematopoietic stem and progenitor cell types, hinting at the importance of controlled combinatorial expression of key transcription factors (Moignard *et al*, 2013). Furthermore, in T cell maturation the temporally coordinated expression of key transcription factors and the transitions between three distinct gene network phases has been identified as crucial for acquiring cellular identity and preventing leukemic transformation (Yui & Rothenberg, 2014).

Despite the seemingly unlimited regulatory potential of transcription factor combinations, the imperfect specificity of transcription factor binding leading to global crosstalk has been suggested as the limiting factor of large gene regulatory networks (Friedlander *et al*, 2016). This limitation implies that there is only a finite number of different transcription factors and transcription factor binding sites a cellular system can control, which in turn sets limits to genome size, gene number, and ultimately the number of different cell types (i.e. cellular identities) within an organism. In fact, comparative analysis have provided evidence that transcription factors and transcription factor binding motifs are highly conserved among metazoans and that evolutionarily novel transcription factors emerge together with new cell types (Nitta *et al*, 2015). Despite the conservation of transcription factors and their binding motifs, aligned binding events in orthologous genomic regions across different species appear to be rare, in-

dicating species specificity of transcriptional regulation, after all (Schmidt *et al*, 2010). From an evolutionary perspective, given these transcription factor associated constraints, it is interesting to speculate on the maximally possible number of different cell types within one organism and whether this number might be already reached in vertebrates. Non-coding RNAs have been suggested to create a large additional regulatory space in higher organisms (Mattick, 2001) and another way to control gene expression without the necessity of transcription factors and thereby exceed this limit would be to remove genes from the genome in a cell type specific manner. In principle, this mechanism would be very similar to the processes involved in B and T cell receptor generation in vertebrates, where genomic segments are excised from the genome to ensure the exclusive expression of a particular receptor composition.

1.2.4 Interaction between genome, epigenome, and transcriptome

Although introduced separately in the previous sections, genome, epigenome, and transcriptome are highly interdependent and changes on one level often also affect the others. Together with the above described individual effects, these interactions are critically involved in the regulation, determination, and aberration of cellular identity. While some interactions between genome, epigenome, and transcriptome are self-evident, others are rather unexpected and, amongst other insights, their recognition allows a better understanding of evolutionary as well as pathogenic processes.

Although originally perceived as a disease caused by aberrations of the genome, for many cancers a far-reaching interplay between genomic, epigenomic, and transcriptomic changes has been identified. While the interplay between genomic and transcriptomic changes mostly involves mutations in genes coding for signalling proteins, transcription-factor binding sites, and enhancers (Sur & Taipale, 2016), the interplay between genomic and epigenomic changes mostly involves genomic mutations altering the function of readers, writers, and erasers of the epigenome (Shen & Laird, 2013). The resulting changes to the epigenome can in turn contribute to realizing the hallmarks of cancer, for example by aberrantly silencing tumour suppressor genes or activating oncogenes. A prominent example is the CpG island hypermethylator phenotype (CIMP) that has been traced back to a certain Isocitrate dehydrogenase 1 (IDH1) mutation in gliomas. The mutated IDH1 produces 2-hydroxyglutarate (2-HG) instead of α-ketoglutarate which inhibits DNA demethylases (the TET enzymes) and thereby causes the hypermethylator phenotype (Turcan *et al*, 2012). Further, mutated IDH1 has been found to block normal differentiation and increase expression of stem cell identity markers, which may explain part of its oncogenic potential. However, although 2-HG (the product of mutated IDH1) is generally considered an oncometabolite, glioma patients with mutated IDH1 have a better prognosis than those with wildtype IDH1, possibly due to increased chemosensitivity (Waitkus *et al*, 2016). Interestingly, promoter hypermethylation induced reduction in the expression of O-6-methylguanine-DNA methyltransferase (MGMT) has been described as a predictor of chemosensitivity (Weller *et al*, 2010). Mechanistically, MGMT is an enzyme that repairs mutagenic DNA lesions and thereby counteracts the effect of alkylating chemotherapeutics. Although MGMT promoter hypermethylation is also observed in IDH1 wildtype tumours, it is more likely to occur in the context of a hypermethylator phenotype. This interplay between IDH1 mutation, MGMT promoter hypermethylation, and disease phenotype is a prime example for the interdependency between genome, epigenome, and transcriptome in the context of malignant disease.

Apart from modulating expression levels of DNA repair enzymes such as MGMT, the epigenome is critically involved in further aspects of maintaining genome integrity. For example, chromatin structure is fundamentally shaped by the epigenome, and especially the densely packed heterochromatin appears to protect the DNA from double-strand breaks as caused for example by ionizing radiation. Additionally, protective changes in chromatin structure have been observed when exposing cells to DNA damaging agents (Lukas *et al*, 2011; Venkatesh *et al*, 2016). Consequently, changes to the epige-

nome that impair a cell's capability to establish and maintain certain DNA protective chromatin conformations will lead to increased rates of DNA damage and it has been shown that mutation rates and types throughout cancer genomes correlate with chromatin structure (Schuster-Böckler & Lehner, 2012). A particular role for the protection of the genome from endogenous damage caused by mobile genomic elements (i.e. transposons) has been attributed to DNA methylation. The translocation of transposons can cause major genomic damage for example by inserting into and disrupting essential genes or by destabilizing the DNA through gaps that remain when a transposon leaves its locus. DNA methylation of transposable elements, however, appears to drastically reduce transposon mobility and thus prevent such damages, while impairment of DNA methylation has been shown to increase transposon mobility (Miura *et al*, 2001). Although increased transposon mobility does not seem to be involved, also hypomethylation of classical satellite DNA has been linked to chromosomal instability in the Immunodeficiency–centromeric instability–facial anomalies (ICF) syndrome especially hampering the late phase of B cell development (Jeanpierre *et al*, 1993; Ehrlich, 2003). Considering that a stable genome forms the fundament to stable cellular identity, this genome-stabilizing role of the epigenome gains even more importance.

Given the important role of DNA methylation for protecting genome integrity on the one hand and its regulatory functions on the other, both of which are known to be associated with certain DNA sequence properties (e.g. repeats) and motifs (e.g.TF binding motifs), it is not too surprising, that DNA methylation status can be predicted based on certain genomic features. Several studies have demonstrated this predictability of DNA methylation status in human, identifying DNA sequence (Bock *et al*, 2006; Das *et al*, 2006) as well as transcription factor binding sites (TFBSs) (Fang *et al*, 2006) as predictive genomic features. The interaction between DNA methylation and transcription factor binding appears to go in both directions: Some transcription factors (e.g. SP1) prevent DNA methylation at adjacent CpGs, while others (e.g. NR6A1) attract DNA methyltransferases and thereby enhance surrounding DNA methylation (Blattler & Farnham, 2013). Conversely, DNA methylation also appears to regulate transcription factor binding as has been demonstrated through the example of the transcription factor CTCF, where methylation of a certain CpG within the TFBS drastically reduces binding affinity (Bell & Felsenfeld, 2000). In line with the described crosstalk between DNA methylation and transcription factor binding, DNA methylation based prediction of transcription factor binding has been demonstrated using a computational supervised learning approach and appears to be possible across different cell types and even across species (human and mouse) (Xu *et al*, 2015). If this transferability of predictive models also holds true across greater evolutionary distances, it would enable the DNA methylation based assessment of transcription factor binding in a wide range of non-model organisms and thereby contribute to a better understanding of the evolutionary processes shaping the transcription factor mediated regulation of cellular identity.

Comparative analyses between hominids (human, chimpanzee, gorilla, and orangutan) have further investigated the interplay between genome and epigenome evolution. These analyses demonstrated a close interplay between changes in DNA sequence and DNA methylation state during hominid evolution (assessed in regions of incomplete lineage sorting) (Hernando-Herraez *et al*, 2015) as well as increased nucleotide evolution in the neighborhood of differentially methylated regions (Hernando-Herraez *et al*, 2013). Similar analysis across other vertebrate groups or even across the entire vertebrate tree would allow to determine the generalizability of these findings. However, such endeavors are analytically very demanding due to larger evolutionary distances and poorer quality or absence of reference genomes.

1.3 High-throughput sequencing enables profound cellular characterization

As described in the previous chapter, nucleic acid (DNA and RNA) sequences form the molecular basis of cellular identity. Assessing the sequence of nucleic acids in order to characterize cells has been scientific practice ever since it became possible to analyse nucleic acid sequences in the late 60s

(Sanger, 2001). The methods used for reading nucleic acid sequences (i.e. sequencing in the broadest sense) have been greatly improved especially in the past 10 years. Starting from low resolution restriction fragment analysis over low throughput chain-termination (Sanger) sequencing to modern high throughput reversible dye-terminator (Illumina) sequencing, sequencing throughput has drastically increased and costs have severely dropped. Today, high throughput sequencing is a standard method in nearly all bio-medical research fields. Accordingly, in the past few years a multitude of specialized sequencing protocols for the assessment of genome, epigenome, and transcriptome have been devised. In particular, single-cell sequencing approaches have opened the door for cellular characterisation in unprecedented detail and resolution. Generally, sequencing strategies fall into two groups: Those that are designed to broadly capture “everything” (e.g. whole genome sequencing) and those that are optimized to capture in depth only regions of interest (e.g. exome sequencing). Therefore, when selecting a strategy for a project, one needs to weigh sequencing breadth against sequencing depth. For most research questions, sequencing depth will be more beneficial, but for some applications, such as genome assembly, sequencing breadth is indispensable. However, if sequencing cost is not an issue, breadth and depth can be achieved at the same time. This chapter presents a selection of commonly used high-throughput sequencing approaches as well as data-analytical strategies.

1.3.1 Genome profiling

Genome profiling in the context of cellular identity serves two major purposes: Detection of genomic variants and genome (de novo) assembly of yet unsequenced species. Detection of genomic variants is usually carried out relative to a species-specific reference genome. In order to generate a reference genome, the entire genome needs to be sequenced and the resulting sequencing reads need to be re-assembled in correct order. The assembly is achieved through detecting unambiguous overlaps between reads and therefore significantly facilitated by longer read lengths, especially in repetitive regions. Although one of the most recent de novo assembled vertebrate genome (spotted gar) was still sequenced using Illumina technology (Braasch *et al*, 2016) with a maximal read length of 300 bases, recent single-molecule sequencing technologies which can achieve read lengths > 10000 bases are increasingly used for de-novo genome assembly (Loman *et al*, 2015) and also to refine existing assemblies (Chaisson *et al*, 2015).

For the detection of genomic variants, read lengths between 50 and 100 bases are mostly sufficient and sequencing depth, as opposed to breadth, is usually paramount. This means that given a certain budget, one would rather invest in sequencing depth to reach at least 30-40x coverage in the loci of interest instead of aspiring to cover the entire genome. Especially in genetically heterogeneous cancer samples, where a mixture of different cells is assessed, the variant frequency detection limit directly depends on sequencing depth. A popular method of choice, especially in the clinical setting, is exome sequencing, where only the coding and thus easily interpretable part of the genome is sequenced (Clark *et al*, 2011). However, exome enrichment strategies tend to display increased sequence coverage bias (i.e. uneven coverage of protein coding regions) and do not allow for the detection of structural or non-coding variants. Comparable whole genome sequencing (WGS) on the other hand is currently two to four times more expensive and computationally more intense (Lelieveld *et al*, 2015). Due to sequencing noise, the detection of genomic variants (i.e. variant calling) from high-throughput sequencing data is intrinsically difficult. In the past years, researchers have created a multitude of different computational tools. The genome analysis tool kit (GATK), due to its comprehensiveness and extensive documentation, represents a widely used and well-validated method (Van der Auwera *et al*, 2013). However, concordance between different variant-calling approaches appears to be only around 50% with many exclusively detected true positives. It has therefore been proposed to use the union of all variants detected by different callers in the discovery phase when sensitivity is more important than specificity (O’Rawe *et al*, 2013). In order to interpret discovered variants in a biologically or medically meaningful way, variants need to be annotated according to their presumed

phenotypic effect. The ANNOVAR tool for example takes into account a variants disruptive effect on protein expression, degree of conservation, and frequency in the general population in order to estimate its phenotypic effect (Wang *et al*, 2010).

Although high coverage (>100x) together with dedicated protocols and suitable bioinformatics make it possible to detect variants that occur in only 1% of the cells, standard variant detection at 40x coverage in bulk samples has an intrinsic blind spot for variants that occur in less than about 20 % of the cells. Furthermore, bulk DNA sequencing does not allow to decipher which combinations of variants come from the same cell and therefore cannot appropriately resolve the cellular components within the population of sequenced cells. While for many applications such as the detection of germline variants, which should be present in all cells of a sample, these limitations are tolerable, they are detrimental when for example the clonal architecture of a tumour is to be assessed. For such cases, single-cell genome sequencing will offer a near to ideal solution once the still high false positive and false negative rates can be controlled (Gawad *et al*, 2016). Technical challenges arise mainly from the extremely low amount of input DNA (exactly two copy of the genome in diploid cells), which requires whole genome amplification, which in turn is prone to biases and can still not replace parts of the genome that may have been lost in the process. Nevertheless, variant analysis in single cells has proven its usefulness, for example in the assessment of clonal evolution in myeloproliferative neoplasms (Hou *et al*, 2012).

1.3.2 Epigenome profiling

Profiling of the epigenome can be carried out on several different levels, ranging from the broad assessment of chromatin structure to the precise measurement of epigenetic modifications. Methods that assess the DNA accessibility, such as “DNase hypersensitivity sequencing” (DNase-seq) or “Assay for Transposase-Accessible Chromatin sequencing” (ATAC-seq), allow genome-wide profiling of chromatin openness, which is considered a proxy for regulatory activity. The exact genome-wide distribution of histone modifications can be determined through “chromatin immunoprecipitation followed by sequencing” (ChIP-seq) experiments, where histone-modification specific antibodies are used to precipitate DNA fragments crosslinked to the respective histones. Similarly, precipitation-based methods have also been used for the genome-wide assessment of DNA methylation. For example, “methylated DNA immunoprecipitation followed by sequencing” (MeDIP-seq), precipitates only methylated DNA fragments using a 5mC-specific antibody (Down *et al*, 2008), while “biotinylated CxxC affinity purification followed by sequencing” (Bio-CAP-seq) precipitates only fragments containing unmethylated CpG motifs (Blackledge *et al*, 2012). However, precipitation-based methods bear the intrinsic shortcoming that absence of signal can signify either a true negative or a false negative call with no possibility for disambiguation. This ambiguity is omitted in bisulfite conversion based approaches, where methylated as well as unmethylated sequences are captured. To discriminate between methylated and unmethylated CpGs, the DNA is treated with sodium bisulfite, which converts unmethylated cytosines to uracil, while methylated cytosines remain cytosines (Frommer *et al*, 1992). Due to their relative robustness, bisulfite based approaches are widely used in basic research as well as in the clinical setting (Bock *et al*, 2016b). Readout of bisulfite converted sequences is possible either through microarrays or high-throughput sequencing (bisulfite sequencing). While microarrays need to be designed separately for each genome/application and only capture predefined sets of regions, high-throughput sequencing allows species-independent, versatile, genome-wide analysis. Therefore, microarrays such as the Infinium 450k array are often used in biomedical research where mainly human samples are to be assessed, whereas sequencing is preferred in basic research offering more room for discovery.

Like genome sequencing, bisulfite sequencing can be performed in a global or targeted manner. In whole genome bisulfite sequencing (WGBS) the entire genome is sequenced regardless of whether or not a sequence actually contains relevant cytosines (i.e. CpG motifs in vertebrates). On the one

hand, this makes sure that every potentially interesting region is captured, but on the other hand, it increases, sequencing cost and compute time. In contrast, reduced representation bisulfite sequencing (RRBS) is designed in a way that each sequenced DNA fragment should contain at least one CpG (Meissner, 2005). This enrichment of relevant sequences is achieved by digesting the DNA with CpG methylation insensitive restriction enzymes (e.g. MspI, TaqI) the restriction motif of which contains a CpG, which makes sure that also the resulting read (which originates from the ends of the fragment) contains at least this restriction CpG. Furthermore, subsequent fragment size-selection enriches for fragments of less than 500 bp length, making sure that fragments originate predominantly from CpG-rich regions including CpG islands (Veillard *et al*, 2016). In human, a typical RRBS library covers 70-80% of all promoters (Bock *et al*, 2010), about 50% of all CpGs that fall into promoter regions and 25-30% of CpGs that fall into further regulatory regions [Fig. 4]. However, pre-fragmentation of the DNA, as it happens spontaneously in DNA stored at room temperature, during apoptosis, or harsh treatments (e.g. formalin fixation), counteracts the enrichment procedure and thereby reduces coverage in the desired genomic regions. Nevertheless, the presented characteristics make RRBS a viable solution for projects that focus on the regulatory functions of DNA methylation in many different samples.

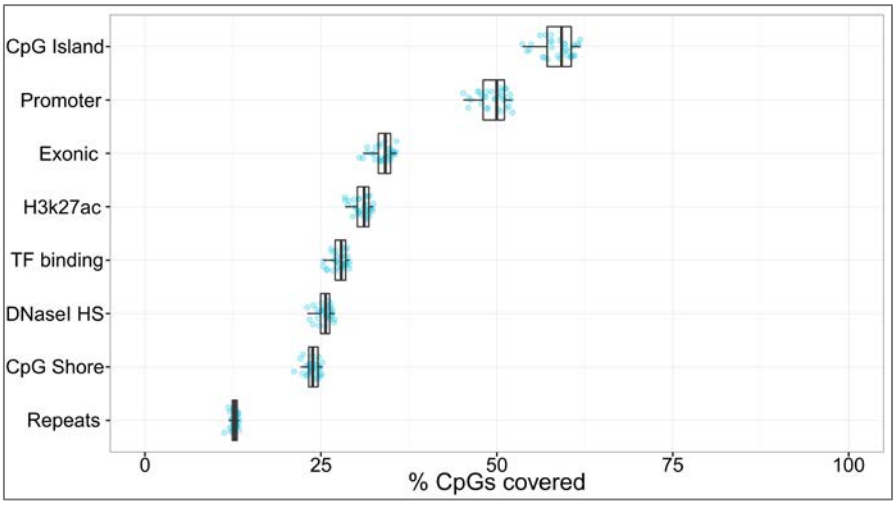


Figure 4 Fraction of CpGs captured in different genomic regions (y-axis). The data represent 32 typical human RRBS libraries prepared as described in (Klughammer *et al*, 2015) and mapped to the human genome assembly hg19. Genomic annotations were retrieved from RefSeq (Promoter, Exonic), UCSC-GB (CpG Island, CpG Shore, Repeats) and ENCODE (H3k27ac, TF binding, DNase HS). TF: transcription factor; HS: hypersensitive

The analysis of bisulfite sequencing data is generally performed by aligning the sequencing reads to a reference genome, taking into account the bisulfite induced cytosine to thymine conversions. Following alignment, for each cytosine the percentage of methylated reads covering that cytosine is calculated, which for example allows the comparison of methylation levels between samples (Bock, 2012). In order to omit the necessity of a reference genome for differential DNA methylation analysis, we (Klughammer *et al*, 2015) and others (van Gurp *et al*, 2016) have developed computational approaches, that deduce a reference directly from the sequencing reads. These advances now allow the assessment of DNA methylation any species independent of whether or not a reference genome is available, offering great opportunities for the comparative investigation of DNA methylation.

An additional advantage of bisulfite sequencing based approaches is that inference of DNA methylation heterogeneity within the population of sequenced cells is to some degree possible. The assessment of methylation patterns within reads that contain several CpG motifs allows to measure local epiallele composition (Li *et al*, 2014) and calculate local heterogeneity scores such as the proportion of discordant reads (PDR) (Landau *et al*, 2014). Clinical relevance of DNA methylation heterogeneity has for example been demonstrated in leukaemias where increased DNA methylation heterogeneity was significantly linked to adverse clinical outcome (Landau *et al*, 2014; Li *et al*, 2016b). Although these computational approaches allow to infer DNA methylation heterogeneity from bulk samples, single-cell bisulfite sequencing is needed to resolve the true clonal composition of a population of cells and assess DNA methylation patterns individually in each cell. There are single-cell versions of

WGBS (Farlik *et al*, 2015; Smallwood *et al*, 2014) as well as RRBS (Guo *et al*, 2013) but they all significantly suffer from low coverage especially owing to loss of DNA during bisulfite treatment. This lack in coverage results in a reduced number of commonly covered CpGs and thereby reduces comparability between different samples. However, because related regulatory elements all over the genome tend to display similar changes in DNA methylation, it is possible to combine DNA methylation measurements across these regions without losing too much information but significantly increasing comparability (Farlik *et al*, 2015; Sheffield & Bock, 2015).

1.3.3 Transcriptional profiling

In contrast to profiling the genome or epigenome, profiling the transcriptome is performed by assessing the RNA not the DNA. Because RNA is biochemically much more instable than DNA and additionally readily digested by omnipresent RNases, RNA-sequencing experiments are intrinsically more prone to batch effects than DNA-sequencing experiments. Accordingly, experiments need to be planned in a way that technical and biological variability is as little as possible confounded (e.g., case and control samples should not be processed in different batches), which allows for computation correction of technical artefact (batch-effect correction) (Leek *et al*, 2012). Also, due to the intrinsic instability of RNA, sequencing is rarely actually performed on RNA, but on DNA copies of the RNA, the cDNA. Although conducted by different means, this reverse transcription of RNA into DNA is one of the first steps in the process of preparing an RNA-sequencing library. Further, many protocols commonly used for transcriptome profiling include steps to prevent inclusion of ribosomal RNA (rRNA), which represents the vast majority of RNA, into the sequencing library. Inclusion of rRNA can be prevented through depletion of rRNA or enrichment of poly-adenylated (poly-A) RNAs that include all messenger RNAs (mRNA) and some but far from all non-coding RNAs (ncRNA). Poly-A enrichment is either achieved by poly-T-bead pull-down or by using poly-T primers to initiate reverse transcription, which integrates poly-A enrichment into the workflow and thereby prevents additional steps with potential loss of non-rRNA. Especially single-cell RNA sequencing protocols, where the amount of input RNA is extremely low (Kolodziejczyk *et al*, 2015), but also high-efficiency bulk RNA sequencing protocols such as QuantSeq (Moll *et al*, 2014) make use of the latter, more economical method. Especially when working with low amounts of starting material, libraries need to be amplified by PCR (>15 cycles) in order to increase the concentration of the sequencing library to a level at which it can be handled. To be able to analytically correct for PCR duplicates, short (5 - 6 bp) random DNA sequences called unique molecular identifiers (UMIs) can be attached to cDNA fragments before amplification, which then after amplification allows to determine reads, that originate from the same cDNA fragment (Islam *et al*, 2014). Removing the PCR duplicates and keeping only one copy each, prevents PCR-amplification biases and allows more accurate determination of expression levels.

Transcript expression levels are calculated by counting the number of reads (or UMIs) that map to a certain transcript and normalising this number to the total number of reads in the sequencing library. This library-size normalisation makes transcript expression levels comparable across different libraries (i.e. samples) and is therefore crucial for differential expression analysis. In order to compare the expression levels of different transcripts, for example to determine the most highly expressed gene or transcript, it is necessary to also normalise the read count to transcript length, because at the same expression level, longer transcripts generally collect more reads than shorter transcripts. Transcript expression levels are therefore often presented as reads per kilobase (transcript length) per million reads (library size) (RPKM) or related measures (FPKM, TPM) (Conesa *et al*, 2016). While library-size normalisation is largely unaffected by any kind of bias there might occur during library preparation or sequencing, transcript-length normalisation is affected by systematic unevenness in transcript coverage, which if neglected leads to systematically underestimating the expression levels of longer reads. Therefore, often an effective transcript-length is calculated based on the observed read distribution and used for normalisation instead of simply using the annotated transcript-length.

A commonly observed transcript-coverage bias in poly-A enriched libraries is the trend to higher coverage of a transcript's 3' end, where the poly-A is situated. This is because selection as well as reverse transcription originates from the 3' end of a transcript and there is a certain propensity for RNA breaks as well as interruptions of the reverse transcription, which decreases the probability for a sequence to be captured with increasing distance to the 3' end of the transcript. In poly-A enriched libraries, in which reverse transcription is initiated by random priming (instead of poly-T priming) such as the Illumina TruSeq protocol (Illumina, 2011) the 3' bias is negligible, as long as the library is constructed from high-quality (not fragmented) RNA. However, protocols that take advantage of combining the steps of poly-A enrichment and reverse transcription by using poly-T primed reverse transcription, intrinsically display a slight 3-prime bias regardless of RNA quality (Adiconis *et al*, 2013).

QuantSeq, a protocol for high efficiency bulk RNA sequencing, elegantly turns the 3' bias into a virtue, by deliberately aiming to incorporate only the 3' ends of a transcript (Moll *et al*, 2014). Although these kind of data do not allow assessment of the entire transcript, they do allow robust quantification of transcript expression levels at extremely low sequencing coverage, which is especially useful for large-scale screening assays with transcriptomic readout, where with replicates easily hundreds of samples need to be assessed. This has been demonstrated for example in a successfully conducted reverse genetic screen where pools of 48 samples were sequenced in 50-bp single-read on an Illumina HiSeq 2000 machine yielding only 2-4 million reads per sample (Gapp *et al*, 2016).

However, the greatest recent advances in RNA sequencing have been made in the area of single-cell RNA sequencing with an ever-increasing number of published studies and methods. For example Smart-seq2 is a popular single-cell RNA sequencing protocol that, like its predecessor Smart-seq, attempts to capture the entire transcript at a reasonably homogeneous coverage through template-switching in the process of cDNA synthesis (Picelli *et al*, 2014). Having overcome the technical challenges of producing effective sequencing libraries from minuscule amounts of RNA, the next milestone was to drastically scale up the efficiency of library preparation, so that thousands of single cells could be analysed. A first step in that direction was the application of microfluidics as demonstrated for example trough the transcriptional profiling and characterisation of 3005 single mouse brain cells, discovering so far unknown cell subtypes (Zeisel *et al*, 2015). This already impressive number of assessed cells can now easily be surpassed by an order of magnitude through a technology that uses nanoliter droplets as reaction chambers (Macosko *et al*; Klein *et al*, 2015). Each droplet contains one cell as well as barcoded reverse transcription primers, through which, during cDNA synthesis, each transcript within the droplet is labelled with the same unique cell barcode. Barcoded cDNAs can then be released from the droplets and the following steps of the library preparation as well as sequencing can be performed in a pooled manner. Through the incorporated barcode, each cDNA can be uniquely reassigned to its cell of origin. The scalability of this approach is rapidly sparking the development of new single-cell screening approaches such as for example pooled genome-editing (CRISPR) screens with transcriptomic readout (Datlinger *et al*, 2016).

1.3.4 Complementarity of single-cell and bulk sequencing

Despite the impressive results and new insights that can be obtained through single-cell sequencing, be it genome, epigenome, or transcriptome sequencing, single-cell sequencing is not expected to replace sequencing of bulk samples. Rather than compete, these two approaches complement each other in a powerful way especially for research that involves the profound characterisation of cell populations as depicted in Figure 5. Bulk sequencing approaches do not allow the discovery of unknown cell types, barely allow the assessment of population heterogeneity, and are not well suited for the characterisation of rare cell types (unless they can be purified beforehand). Single-cell sequencing in contrast makes all these endeavours possible, however with one crucial drawback: Once a cell has been subjected to sequencing, it is destroyed and can no longer be assessed by other means such as functional assays, although there are solutions to simultaneously obtain different molecular readouts

from the same cell (e.g. DNA and RNA, RNA and Protein, DNA methylation and sequence) (Bock *et al*, 2016a). The combination of several assays, however, is easily possible with bulk samples, where the sample can be split and only part of the cells can be subjected to sequencing, while the rest remains available for further assessments. Yet, in order to produce pure bulk samples consisting only of the cell (sub) type of interest, cellular characteristics to select by, such as for example marker gene expression, need to be available. These cellular characteristics, in turn can be identified through single-cell analysis, which closes the circle of discovery [Fig. 5].

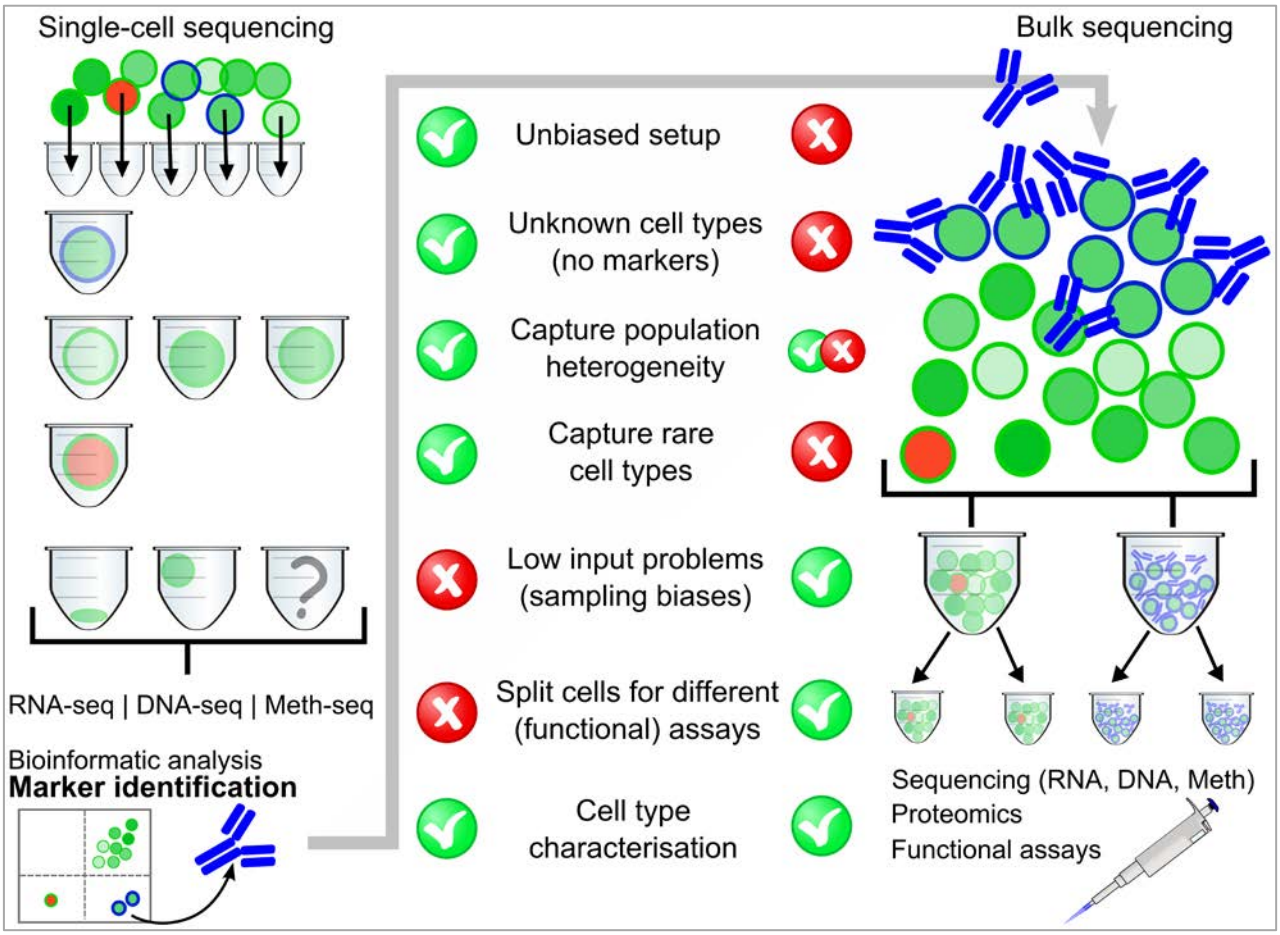


Figure 5 Summary of the strengths and weaknesses of single-cell sequencing and bulk sequencing, emphasizing the complementary nature of these two approaches. Depicted schematically is the molecular characterisation of a heterogeneous population of cells by either single-cell (left) or bulk (right) sequencing. The grey arrow signifies the application of cell (sub) type markers (represented by an antibody) identified through single-cell sequencing, for the purification of cell sub populations for bulk sequencing experiments.

2 Aim

This thesis aims at providing new insights into the complex and fundamental phenomenon of cellular identity by integrating physiological as well as pathological states, evolutionary relationships, and molecular data.

First, the work presented in this thesis builds a foundation to uncover the evolutionary role of DNA methylation for establishing and maintaining cellular identity across invertebrate and vertebrate species. To this end, an RRBS-based computational method for reference genome independent differential DNA methylation analysis (RefFreeDMA) was developed, validated, and applied.

Second, this thesis assesses cellular identity and its alterations in the pathological condition of a malignant disease. To that end, DNA methylation and also chromosomal aberrations, and genomic variants, were measured and analysed in matched (diagnosis, progression) glioblastoma samples from a cohort of 112 patients (GBMatch).

Third, transcriptional aspects of cellular identity under physiological conditions are investigated and revealed by identifying and characterizing single human pancreatic islet cells based on their transcriptional profiles (HumanIslet).

3 Results

3.1 RefFreeDMA

Differential DNA Methylation Analysis without a Reference Genome.

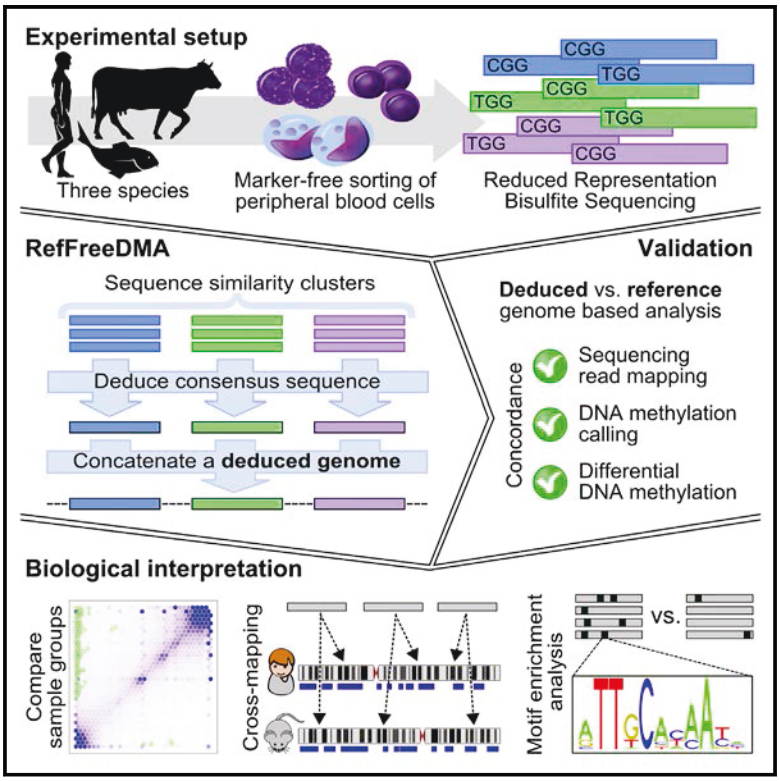
Klughammer J, Datlinger P, Printz D, Sheffield NC, Farlik M, Hadler J, Fritsch G, Bock C.
Cell Rep. 2015 Dec;13(11):2621-33. doi:10.1016/j.celrep.2015.11.024.

Resource

Cell Reports

Differential DNA Methylation Analysis without a Reference Genome

Graphical Abstract



Highlights

- Bioinformatic method for DNA methylation analysis without a reference genome
- Coverage-optimized high-throughput RRBS protocol validated in nine species
- Antibody-free FACS purification of blood cell types validated in three species
- Analysis of blood cell-type-specific DNA methylation in human, cow, and carp

Authors

Johanna Klughammer, Paul Datlinger, Dieter Printz, ..., Johanna Hadler, Gerhard Fritsch, Christoph Bock

Correspondence

cbock@cemm.oeaw.ac.at

In Brief

Klughammer et al. describe a method for reference-genome-independent analysis and interpretation of DNA methylation patterns. A combination of experimental and computational advances enables cost-effective DNA methylation analysis in natural populations and species without a reference genome, thus facilitating epigenome-wide association studies in the context of ecology and evolution.

Accession Numbers

GSE74026

Klughammer et al., 2015, Cell Reports *13*, 1–13
December 22, 2015 ©2015 The Authors
http://dx.doi.org/10.1016/j.celrep.2015.11.024

CellPress

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

Cell Reports
Resource

Differential DNA Methylation Analysis without a Reference Genome

Johanna Klughammer,[1] Paul Datlinger,[1] Dieter Printz,[2] Nathan C. Sheffield,[1] Matthias Farlik,[1] Johanna Hadler,[1] Gerhard Fritsch,[2] and Christoph Bock[1,3,4,*]
[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria
[2]Children's Cancer Research Institute, St. Anna Kinderkrebsforschung, 1090 Vienna, Austria
[3]Department of Laboratory Medicine, Medical University of Vienna, 1090 Vienna, Austria
[4]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany
*Correspondence: cbock@cemm.oeaw.ac.at
http://dx.doi.org/10.1016/j.celrep.2015.11.024

SUMMARY

Genome-wide DNA methylation mapping uncovers epigenetic changes associated with animal development, environmental adaptation, and species evolution. To address the lack of high-throughput methods for DNA methylation analysis in non-model organisms, we developed an integrated approach for studying DNA methylation differences independent of a reference genome. Experimentally, our method relies on an optimized 96-well protocol for reduced representation bisulfite sequencing (RRBS), which we have validated in nine species (human, mouse, rat, cow, dog, chicken, carp, sea bass, and zebrafish). Bioinformatically, we developed the RefFreeDMA software to deduce ad hoc genomes directly from RRBS reads and to pinpoint differentially methylated regions between samples or groups of individuals (http://RefFreeDMA.computational-epigenetics.org). The identified regions are interpreted using motif enrichment analysis and/or cross-mapping to annotated genomes. We validated our method by reference-free analysis of cell-type-specific DNA methylation in the blood of human, cow, and carp. In summary, we present a cost-effective method for epigenome analysis in ecology and evolution, which enables epigenome-wide association studies in natural populations and species without a reference genome.

BACKGROUND

DNA methylation is an epigenetic mechanism that is indispensable for animal development (Reik, 2007) and also broadly relevant for plant biology (Law and Jacobsen, 2010). Defects in the DNA methylation machinery are associated with widespread changes in cellular identity and interfere with the developmental potential of stem cells (Jones, 2012). Altered DNA methylation patterns are ubiquitous in cancer (Baylin and Jones, 2011; Feinberg and Tycko, 2004), and they have been observed in numerous other diseases (Portela and Esteller, 2010; Robertson, 2005). Moreover, there is mounting evidence for associations between DNA methylation patterns and environmental factors such as stress, nutrition, toxic exposures, and substance abuse (Foley et al., 2009; Mill and Heijmans, 2013).

In humans, epigenome-wide association studies (EWASs) have emerged as a widely used paradigm for linking DNA methylation to environmental exposures and to diseases (Michels et al., 2013; Rakyan et al., 2011). A small number of associations between the epigenome and the environment have also been validated in inbred mouse and rat models, for example, identifying connections between early life exposures and the propensity to subsequently develop certain diseases and behavioral phenotypes. A widely discussed hypothesis posits that epigenetic mechanisms provide a mechanistic link between exposures and diseases, thus contributing to the developmental origins of health and disease in humans (Gillman, 2005; Waterland and Michels, 2007). Furthermore, DNA methylation can be transgenerationally inherited at certain genomic loci (Feil and Fraga, 2011) and may contribute to species evolution (Jablonka and Raz, 2009).

There is tremendous potential in studying environmental influences and epigenetic inheritance not only in laboratory animals, but also in natural populations and non-model organisms. For example, animals in the wild are often exposed to complex evolutionary pressures and ecological interactions that cannot be modeled in the laboratory. Initial studies along these lines have suggested a role of epigenetics in the evolution of Darwin's finches (Skinner et al., 2014) and in speciation among marsupials (O'Neill et al., 1998), and they identified DNA methylation as a potential source of random variation in natural populations of fish (Massicotte et al., 2011) and songbirds (Liebl et al., 2013; Schrey et al., 2012).

However, systematic epigenetic studies in natural populations and non-model organisms have been hampered by the lack of methods for high-resolution and high-throughput DNA methylation analysis that work well across a broad range of species. To date, most studies of DNA methylation in ecology and

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

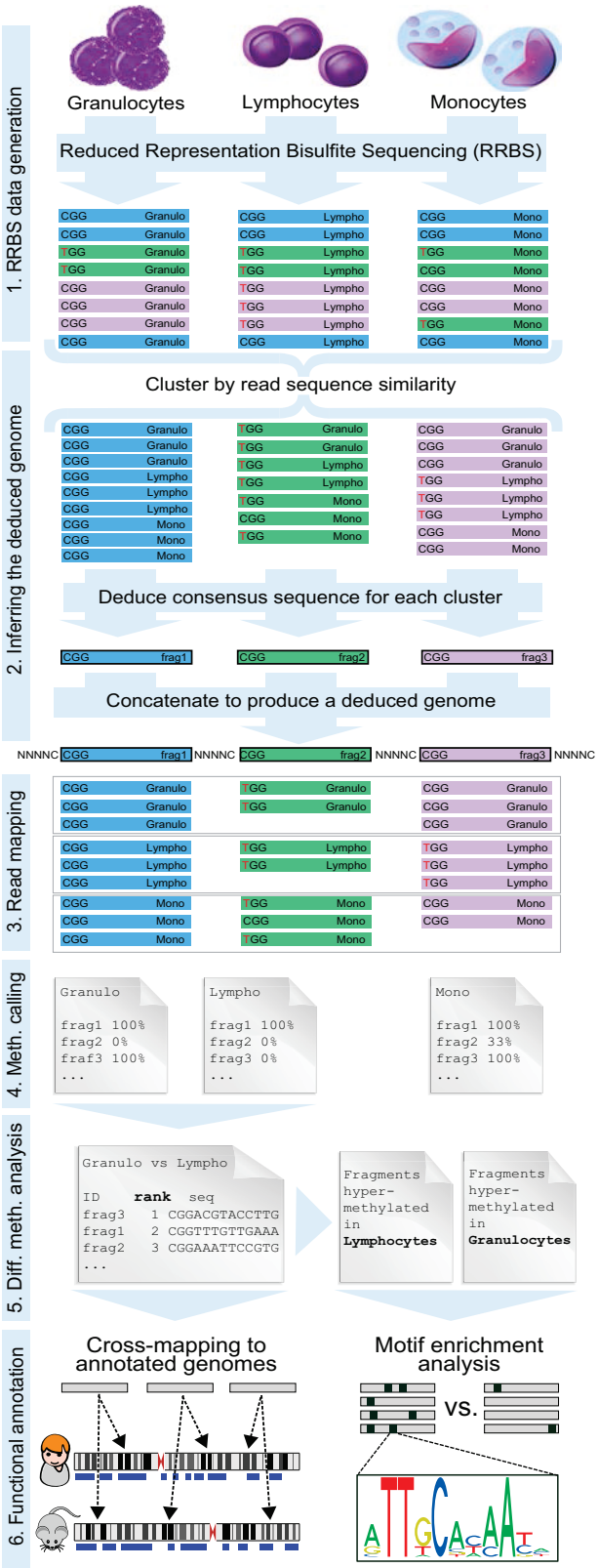


Figure 1. DNA Methylation Analysis without a Reference Genome

Workflow for reference-genome-independent analysis of differential DNA methylation using an optimized RRBS protocol and the RefFreeDMA software. Colored bars represent RRBS sequencing reads, and identical colors indicate high sequence similarity. Bisulfite-converted MspI restriction sites are shown at the beginning of each read (CGG for methylated sites and TGG for unmethylated sites). To derive a deduced genome, reads from all samples are clustered by sequence similarity, and a consensus sequence is determined. These deduced genome fragments (black-edged bars) are concatenated into one deduced genome, to which the RRBS reads for each sample are mapped. DNA methylation levels are obtained by counting the number of Cs versus Ts for individual cytosines in the deduced genome (this step typically focuses on CpG sites, but the method also supports the analysis of non-CpG methylation). Differential methylation analysis is performed by comparing site-specific and fragment-specific DNA methylation levels between sample groups. Finally, the identified differentially methylated fragments are analyzed by cross-mapping to well-annotated genomes of other species (e.g., mouse or human) and by motif enrichment analysis (e.g., for identifying enriched transcription factor binding sites).

evolution have relied on low-throughput, gel-based assays such as MS-AFLP (Schrey et al., 2013). Much more powerful assays are being used for DNA methylation analysis in human, including the Infinium microarray, whole-genome bisulfite sequencing (WGBS), and reduced representation bisulfite sequencing (RRBS). However, none of these assays is directly applicable for studying DNA methylation in natural populations and non-model organisms: The Infinium assay requires a commercial microarray that is only available for the human genome (Bibikova et al., 2011); WGBS is excessively expensive when studying more than a handful of samples (Beck, 2010), and RRBS suffers from the technical complexity of the original protocol (Gu et al., 2011) and from concerns that the restriction enzyme MspI may not provide good genome coverage in other species. Furthermore, there is a general lack of bioinformatic methods for analyzing sequencing-based DNA methylation data in the absence of a high-quality reference genome and in genetically diverse populations for which existing reference genomes would unduly bias the analysis.

Here, we describe an integrated approach for analyzing DNA methylation at single-base-pair resolution in a broad range of species. We combine an optimized high-throughput RRBS protocol with a tailored computational method called RefFreeDMA in order to detect differential DNA methylation without a reference genome. RefFreeDMA constructs a deduced genome directly from RRBS sequencing reads, it maps the sequencing reads to the deduced genome, performs DNA methylation calling, and identifies differentially methylated cytosines and DNA fragments (Figure 1). We validated our method by studying blood cell-type-specific DNA methylation in three species (human, cow, and carp), benchmarking the reference-free analysis against a reference-based analysis using the existing reference genomes. The experimental protocol was also validated in six additional vertebrate species (rat, mouse, dog, chicken, sea bass, and zebrafish). We expect that the described method will be broadly useful for DNA methylation analysis in non-model organisms, for example, to identify and interpret DNA methylation differences between samples (e.g., different cell types) or groups of individuals (e.g., animals that have been exposed to different environments).

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

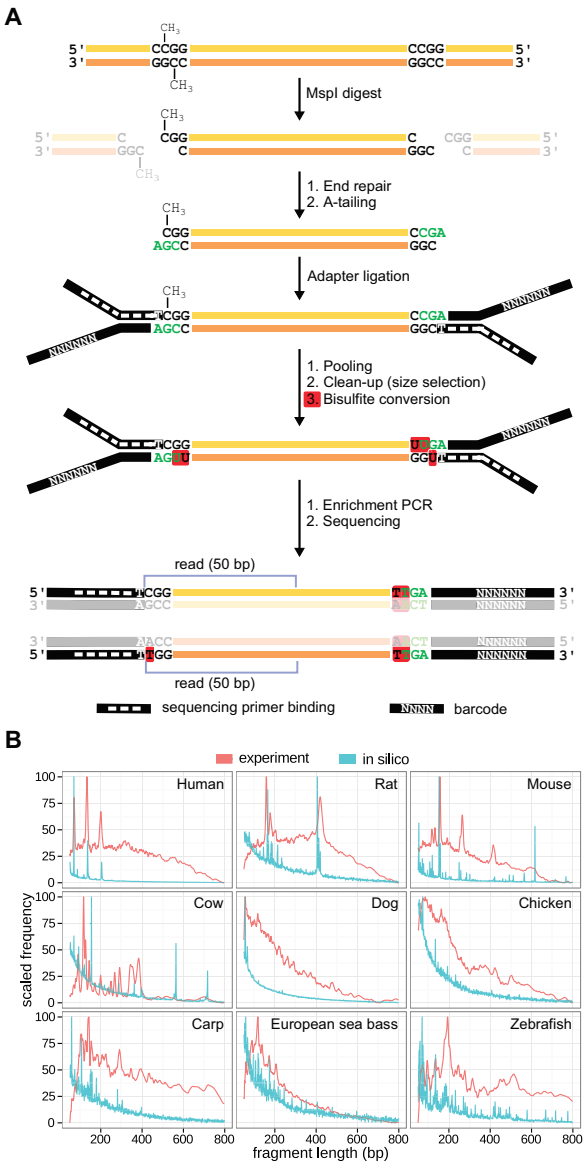


Figure 2. An Optimized RRBS Protocol Validated in Nine Species

(A) Schematic outline of RRBS library preparation and the corresponding sequencing reads.

(B) Computationally predicted (blue) and experimentally measured (red) fragment length distribution of RRBS libraries in nine vertebrate species. Predictions were based on in silico MspI restriction digests of the reference genomes using the BSgenome R package. Experimental results were obtained by electrophoresis (Experion DNA 1k chip). In species with a reference genome, concordance between predicted and experimentally measured peaks can be used to confirm successful RRBS library preparation.

RESULTS

High-Throughput DNA Methylation Mapping in Diverse Animal Species Using RRBS

RRBS enables genome-scale DNA methylation mapping at single-base-pair resolution for a fraction of the cost of WGBS (Meissner et al., 2005). It exploits the highly characteristic distribution of DNA methylation in vertebrate genomes, which occurs mainly at CpG dinucleotides. DNA is digested with the restriction enzymes MspI (restriction site: C^CGG) and/or TaqI (restriction site: T^CGA), which are insensitive to DNA methylation at the central CpG, and short size-selected restriction fragments are subjected to bisulfite sequencing (Figure 2A).

We adapted an existing RRBS protocol (Boyle et al., 2012) and optimized it for genome coverage and sample throughput (see Experimental Procedures for details). The optimized protocol increases the number of covered CpG sites from ~2.5M to ~4M (human genome, using the MspI enzyme), and it allows a single person to process up to 192 samples per week. For most vertebrates, good sequencing coverage can be obtained when 6–12 barcoded samples are sequenced on a single lane of Illumina HiSeq, which makes the protocol approximately 10-fold cheaper than WGBS. To validate the assay, we generated RRBS libraries for nine species (human, rat, mouse, cow, dog, chicken, carp, sea bass, and zebrafish). These libraries showed characteristic fragment length distributions, which reflect the distribution of CpG-rich repetitive elements in these species and which provide a convenient metric for assessing the quality of RRBS libraries prior to sequencing (Figure 2B).

Using our optimized RRBS protocol, we established a DNA methylation dataset for the major nucleated cell populations in peripheral blood of three species (human, cow, and carp), with four biological replicates per cell type and species. The human and cow datasets comprise granulocytes, monocytes, and lymphocytes, whereas the carp dataset also includes nucleated erythrocytes and one additional leukocyte population that morphologically resembles granulocytes and monocytes (Figure 3A). In total, the dataset comprises 44 blood cell samples from three species and 789 million sequencing reads (Table S1). All cell types were fluorescence-activated cell sorting (FACS) purified based on forward and side scatter alone, demonstrating the feasibility of separating blood cell types in species that lack suitable FACS antibodies. The purity of the sorted cell populations was assessed visually through cytospins, and it exceeded 95% in all samples. Here, our analysis focuses on DNA methylation differences between these cell populations, but the same sorting strategy can also be used for minimizing the impact of differences in cell composition between individuals, which is a major confounder in human EWAS (Houseman et al., 2012; Jaffe and Irizarry, 2014).

RefFreeDMA: Analyzing Differential DNA Methylation without a Reference Genome

We devised a workflow for reference-free DNA methylation analysis consisting of six main steps (Figure 1): (1) preparation and sequencing of RRBS libraries, (2) inference of a deduced genome from the RRBS sequencing reads, (3) read alignment to the deduced genome, (4) DNA methylation calling, (5) identification and ranking of differentially methylated CpGs and deduced genome fragments, and (6) functional annotation of differential DNA methylation. RefFreeDMA is implemented as a Linux-based software pipeline, supporting small to moderately sized analyses on a desktop computer (e.g., 40-hr total runtime for 20 samples), whereas large analyses are efficiently parallelized on a computing cluster. A detailed overview of the

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

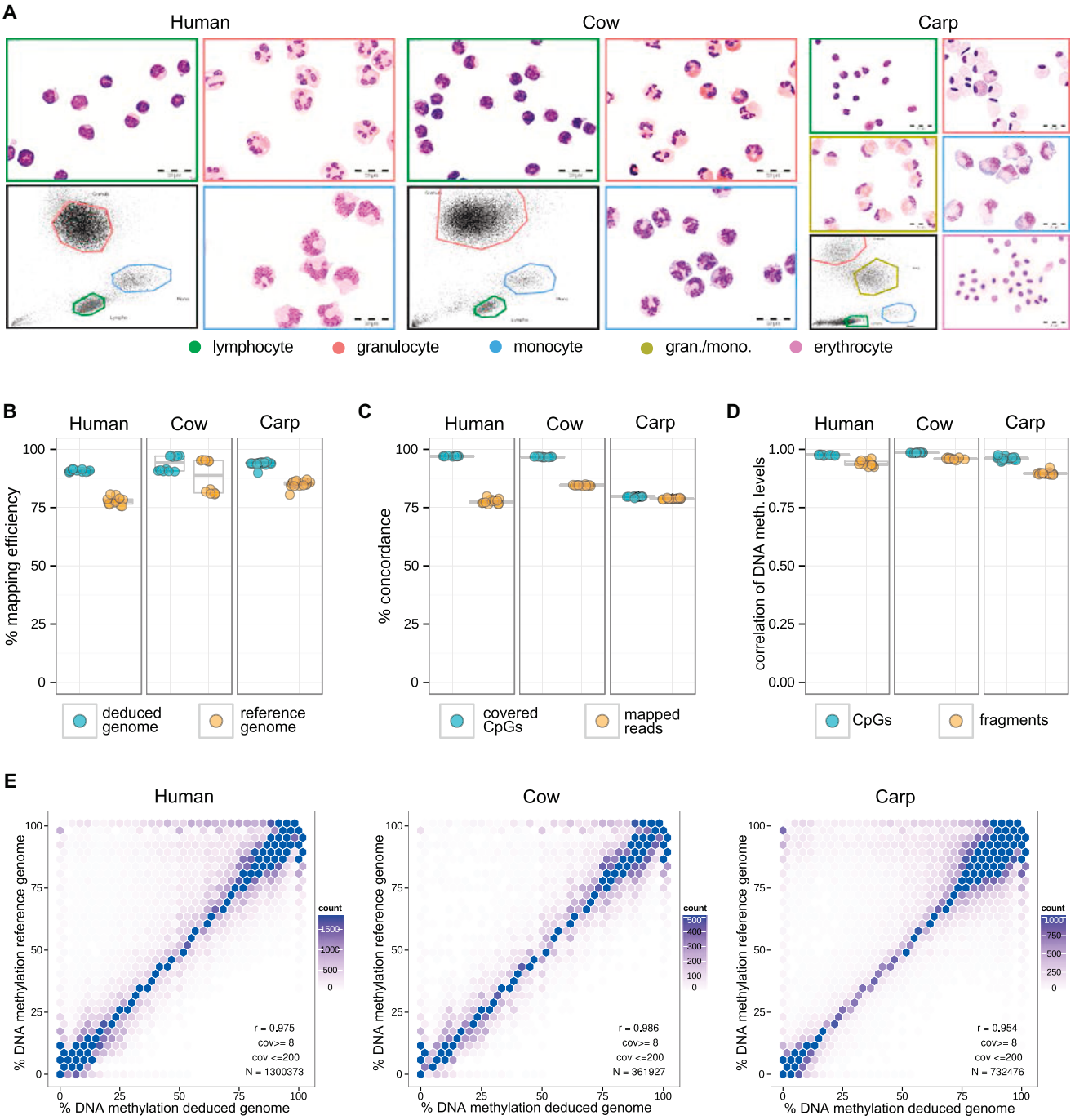


Figure 3. Validation of Reference-Free DNA Methylation Mapping

(A) Representative images (Giemsa-stained cytospins at 100× magnification) of blood cell populations that were purified by FACS using an antibody-independent protocol based on forward scatter (x axis) and side scatter (y axis). Gated cell populations are highlighted in different colors, and their DNA was used for RRBS library preparation.

(B) Percent mapping efficiency (alignment rate) for RRBS reads using the deduced genome versus the reference genome. Mapping rates are expectedly lower than 100% for the reference-free method because low-confidence reads are used during alignment but not for building the deduced genome.

(C) Percentage of CpGs and sequencing reads with concordant mapping between the two approaches in non-repetitive genomic regions (see Figure S3A for details).

(D) Pearson correlation of DNA methylation levels for the two approaches, compared at the level of CpG sites and deduced genome fragments using RefFreeDMA's standard filtering criteria (coverage of least eight and not more than 200 mapped reads).

(E) DNA methylation scatterplots at the level of CpG sites (r, Pearson correlation; N, number of CpGs; cov, minimum and maximum read coverage used for filtering).

RefFreeDMA pipeline is provided as a Unified Modeling Language (UML) diagram in Figure S1.

A key aspect of RefFreeDMA is the construction of a deduced genome directly from the RRBS reads. This deduced genome is not based on classical de novo assembly of bisulfite sequencing reads, which is computationally expensive and would require very deep sequencing. Rather, we exploit a specific characteristic of RRBS with its defined fragment start and end positions at MspI restriction sites to simplify the problem. RefFreeDMA constructs the deduced genome by clustering the RRBS reads from all samples in a given species according to their sequence similarity, followed by inference of the consensus sequence for each read cluster. In the consensus sequence, positions with both cytosines (Cs) and thymines (Ts) among the clustered reads are retained as Cs (Figure 1), given that they are likely to reflect genomic cytosines that are methylated and protected from bisulfite sequencing in some but not all samples. We developed an efficient two-step approach in which all quality-filtered, non-duplicate sequencing reads are initially clustered in an approximate and computationally efficient manner, followed by a more precise and computationally demanding finalization step (see Experimental Procedures for details). Finally, all consensus sequences are concatenated with spacer sequences (i.e., stretches of Ns) to facilitate computational processing, resulting in a deduced genome that is specific for a given species and analysis but shared among all samples contributing to the analysis.

The subsequent steps of read alignment, DNA methylation calling, and differential methylation analysis are performed in much the same way as for DNA methylation analysis with a reference genome (Bock, 2012). Specifically, we use BSMAP/RRBSMAP (Xi et al., 2012; Xi and Li, 2009) for read alignment and a custom DNA methylation calling script (Bock et al., 2010) for calculating the fraction of methylated reads at each CpG position in the deduced genome. Differentially methylated CpGs and deduced genome fragments between sample groups are then identified using a modified t test statistic as described for the RnBeads software (Assenov et al., 2014). The analysis gives rise to lists with individual CpGs as well as deduced genome fragments ranked by their degree of differential methylation.

In a final step, the top-ranking differentially methylated fragments are exported as FASTA/FASTQ files, which provide the basis for biological interpretation by cross-mapping to well-annotated genomes and by reference-free motif enrichment analysis. The principle behind cross-mapping is to link deduced genome fragments in the analyzed species to orthologous regions in well-annotated genomes of other vertebrate species and to use the genome annotations that are available in the latter species (e.g., genes, transcription factor binding sites, histone modifications, and DNase hypersensitivity sites) for cross-species enrichment analysis. This approach is of course limited to genomic regions that are conserved across species; hence, it is most powerful for species that are closely related to well-characterized model organisms.

Motif enrichment analysis provides an alternative approach to biological interpretation that is independent of any reference genomes. It is based on the observations that transcription factor binding motifs are highly conserved across all vertebrates (Nitta et al., 2015) and that DNA methylation levels at motif sequences have been shown to correlate with cell-type-specific transcription factor binding (Bock et al., 2012; Feldmann et al., 2013; Stadler et al., 2011). By analyzing motif enrichment among differentially methylated DNA fragments using existing databases (such as JASPAR; Mathelier et al., 2014) and software tools (such as AME; McLeay and Bailey, 2010), it is possible to gain insight into the regulatory mechanisms that distinguish the studied cell types and sample groups.

Validating Reference-Free DNA Methylation Analysis across Three Species and 44 Samples

To validate our approach, we performed reference-free analysis of the RRBS blood cell dataset (Figure 3A) and compared the results to those obtained by reference-based analysis of the same data (see Experimental Procedures for details). The fraction of aligned reads was in the range of 90% to 98% for the deduced genomes and slightly lower (75% to 95%) for the published reference genome of each species (Figure 3B; Table S1). The number of covered CpGs was predominantly species specific (3–4 million for human, ~3 million for cow, and 1.5–2 million for carp) and broadly similar between the reference-based and reference-free analysis. Average DNA methylation levels at CpG sites were also similar for both approaches, whereas the observed C-to-T conversion rates at non-CpG sites were substantially lower in the reference-free analysis (Table S1). This is because ubiquitously unmethylated Cs—which in vertebrates are mostly found in non-CpG context—are counted as Ts by the reference-free analysis (case 4 in Figure S2) and therefore do not contribute to high non-CpG conversion rates. To circumvent this potential problem our RRBS protocol uses methylated and unmethylated spike-in controls to monitor bisulfite conversion rates (Table S1), rather than relying on non-CpG conversion rates. The issue can also be avoided altogether by sequencing a single RRBS sample without bisulfite conversion and including it in the analysis. Finally, to assess the comparative performance of our reference-free method, we benchmarked it against simply cross-mapping the RRBS reads for carp to the well-annotated genomes of human, mouse, and zebrafish. The results showed a one to two orders of magnitude higher genome-wide CpG coverage using RefFreeDMA than observed for the basic cross-mapping approach (Table S2).

We also compared the alignment of individual reads, the coverage of individual CpGs, and the DNA methylation levels of single CpGs and deduced genome fragments between the two approaches. To that end, the deduced genome fragments were aligned to the corresponding reference genome, allowing us to link most RRBS fragments (human: 1,254,324 out of 1,522,786; cow: 1,276,537 out of 1,521,946; and carp: 455,821 out of 780,757) to their putative position in the reference genome. More than 75% of reads and CpGs in non-repetitive regions where concordantly mapped by both approaches (Figure 3C), whereas the agreement was much lower for repetitive regions and reads that map to multiple positions in the genome (Figure S3A). We investigated these discrepancies and identified four scenarios in which there may be deviations between the reference-free method and the reference-based method (Figure S2). Most frequently, a sequencing read maps to multiple positions throughout the reference genome, and the aligner

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

randomly assigns it to one of these positions. We indeed observed similarly low concordance rates in repetitive regions when running the reference-based method twice with different random seed parameters (Figure S3A). Based on these results, it might even be argued that the clustering and combining of highly similar repetitive reads into a single consensus provide a more appropriate way of handling multimapping reads than their random assignment in the reference-based analysis, and similar approaches have successfully been used for studying epigenetic marks in repetitive regions of the genome (Bock et al., 2010; Day et al., 2010). Finally, despite these special cases, we observed excellent agreement between the two approaches when plotting alignment positions across a representative chromosome (Figure S3B), and the DNA methylation values obtained with the two approaches were highly correlated in all samples and all species—with Pearson correlation coefficients above 0.9 across all CpGs and fragments and above 0.95 for those CpGs and fragments that have good sequencing coverage (Figures 3D, 3E, and S3C).

Reference-Free Analysis of Differential DNA Methylation between Cell Types of the Blood

Importantly, the reference-free method was able to recapitulate the known biological similarities and differences among the different blood cell types in almost perfect concordance with the reference-based method (Figure 4A). Many genes with a known role in hematopoietic cells were identified by both methods, as illustrated by the myeloid-specific MPO gene and the lymphoid-specific LAX1 gene (Figure 4B). There was also strong correlation ($r \geq 0.95$) between the differential DNA methylation ranks obtained with the two methods in all three species (Figure S4A). Furthermore, the vast majority of the top-1,000 differentially methylated fragments identified by the reference-free method were also among the top-1,000 or top-5,000 differentially methylated regions based on the reference-based method (Figure S4B). The magnitude of the DNA methylation differences calculated by either method were also highly correlated (Figure S4C). Furthermore, both methods identified a consistent and biologically interesting trend toward increased DNA methylation levels in lymphoid as opposed to myeloid cells, which was very prominent in human, weaker in cow, and essentially absent in carp (Figures 4C and S4D), suggesting species-specific differences in the genome-wide regulation of DNA methylation in the hematopoietic system.

We pursued two complementary approaches for interpreting the identified DNA methylation differences without a reference genome for the target species. First, we cross-mapped the deduced genome fragments obtained in each species to the human and mouse genome, for which extensive functional genomics data exist from projects such as ENCODE (ENCODE Project Consortium, 2004), IHEC (http://www.ihec-epigenomes.org/), and BLUEPRINT (Adams et al., 2012). Cross-species mapping rates were expectedly low, amounting to ~20% for human and cow and ~10% for carp at a maximum mismatch rate of 20%. (Figure S5A). Nevertheless, for those deduced reference fragments that did map, we were able to perform enrichment analysis relative to the extensive biological annotations of the human and mouse genomes. Fragments that were less methylated in lymphocytes as compared with granulocytes (hypermethylated in granulocytes) were often associated with lymphoid-specific regulatory elements and transcription factor binding mapped by ChIP-seq and similar technologies (Figures 5A and S5B). The enrichment was not always consistent between species, but we found recurrent and biologically meaningful associations. Most notably, the binding sites of two key myeloid transcription factors, CEBPA and CEBPB (Akagi et al., 2010; Rosenbauer and Tenen, 2007), were hypermethylated in both human and cow lymphocytes, and binding sites of MYB, a transcription factor implicated in lymphocyte and erythrocyte development (Greig et al., 2008), were hypermethylated in human and cow granulocytes. In contrast, carp appears to be too evolutionary distant to obtain interesting results by cross-mapping to mammalian genomes (Figure S5B).

Second, we exploited the fact that transcription factor binding motifs are much more conserved than most regulatory elements (Nitta et al., 2015) and performed alignment-free motif enrichment analysis for those deduced reference fragments that were most differentially methylated between lymphocytes and granulocytes. In all three species, there was a higher ratio of GC-rich and CpG-rich motifs among fragments that are hypermethylated in granulocytes (Figures 5B and S5C), which we corrected for in the motif analysis by using random sequences with matched base composition as controls (see Experimental Procedures for details). Those fragments that were less methylated in lymphocytes (hypermethylated in granulocytes) were enriched for 29 sequence motifs, of which four were shared across two species (EGR2, KLF5, KLF1, and RREB1; shown in Figure S5D). Those fragments that were less methylated in granulocytes (hypermethylated in lymphocytes) were enriched for 40 sequence motifs, and four motifs were shared between all three species (CEBPA, CEBPB, HLF, and JUN) (Figures 5C and S5D). Three of these transcription factors are well-established regulators of myeloid cell differentiation (Akagi et al., 2010; Orkin, 1995; Rosenbauer and Tenen, 2007), whereas HLF is associated with hematopoietic stem cells (Gazit et al., 2013). Finally, we also searched for motifs that were enriched in lymphocyte-specific as well as in granulocyte-specific differentially methylated fragments (Figures 5C and S5E), and a total of 27 sequence motifs were identified, of which six were shared across all three species (BRCA1, FOXL1, PAX4, RREB1, RUNX1, and RUNX2). Of these, RUNX1 and RUNX2 in particular are known to play a role in both lymphoid and myeloid cell differentiation and function (Klunker et al., 2009; Liebermann and Hoffman, 2002; Tenen et al., 1997).

DISCUSSION

We present an integrated experimental and computational method for DNA methylation analysis and interpretation in non-model organisms, unsequenced species, and natural populations. Our method addresses a major bottleneck for epigenome studies in the context of comparative genomics, ecology, and evolution, where whole genome bisulfite sequencing is rarely affordable for sufficiently large cohorts and other widely used methods such as MS-AFLP are strongly limited in the information they can provide.

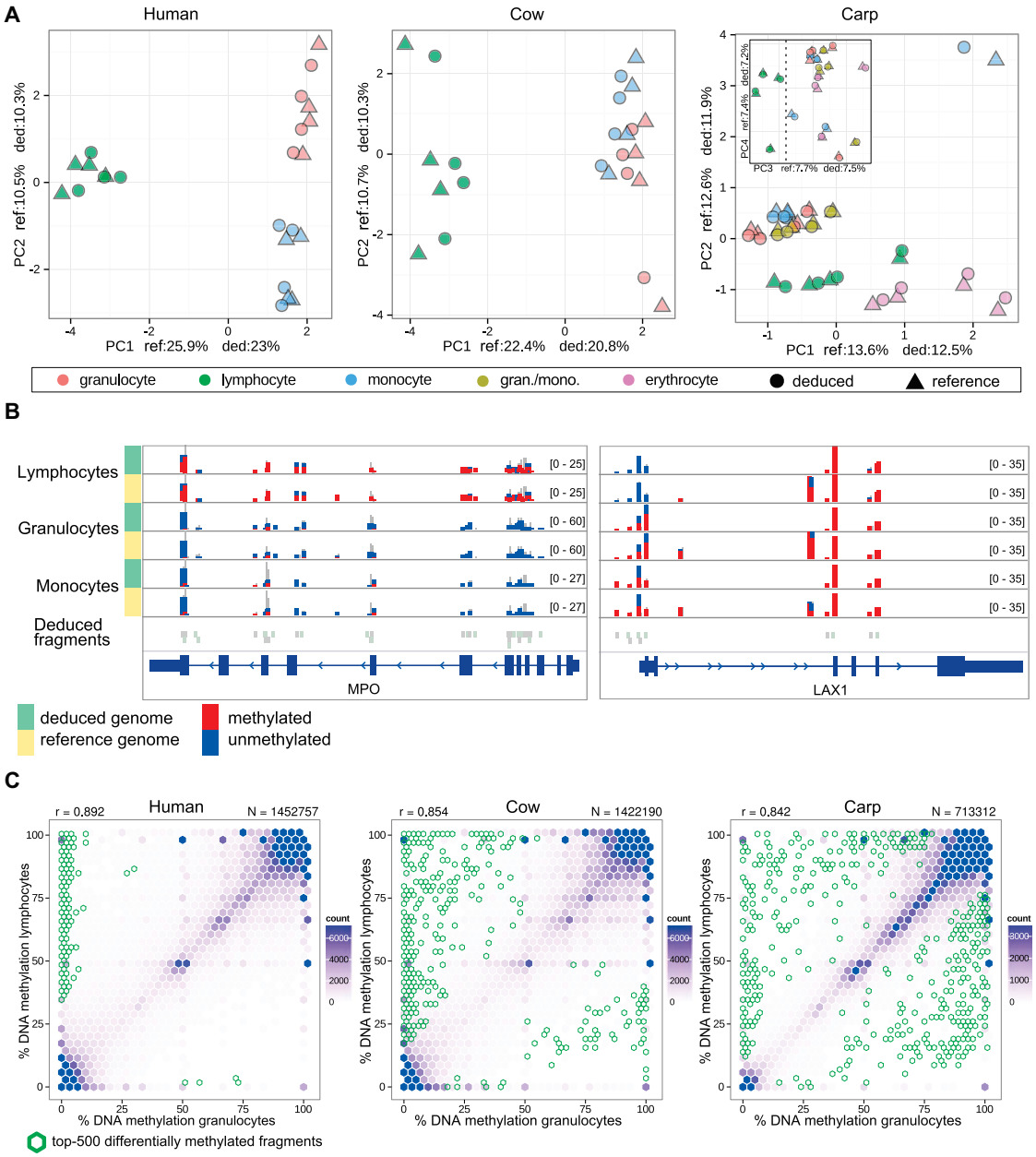


Figure 4. Differential DNA Methylation Analysis without a Reference Genome

(A) Global concordance between reference-free and reference-based DNA methylation analysis illustrated by principal component analysis. Shown are the first two principal components (x axis and y axis) for the reference-free (circles) and reference-based (triangles) approaches as well as the percentage of variance explained by these principal components. The inset for carp shows the third and fourth principal components, which provides clearer separation of lymphoid versus myeloid cell types.

(B) Representative genome browser tracks displaying DNA methylation levels at single CpG sites as determined by the reference-free and reference-based approach, focusing on genes with known myeloid (MPO) and lymphoid (LAX1) function. The "Deduced fragments" track depicts the mapping between deduced genome fragments (gray boxes) and the reference genome.

(C) DNA methylation scatterplots showing differential DNA methylation in granulocytes (x axis) versus lymphocytes (y axis) based on the reference-free approach. Means across four biological replicates per cell type are shown, and the green hexagons indicate the top-500 most differentially methylated fragments (r, Pearson correlation; N, number of deduced genome fragments). Matched scatterplots for the reference-based analysis are shown in Figure S4D.

On the experimental side, our method uses an optimized 96-well RRBS protocol, which provides an excellent trade-off between single-base-pair resolution, affordable cost, and practical feasibility for studies with hundreds (or even thousands) of individuals. Building upon the track record of RRBS in mouse and human and the popularity of reduced representation genome

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

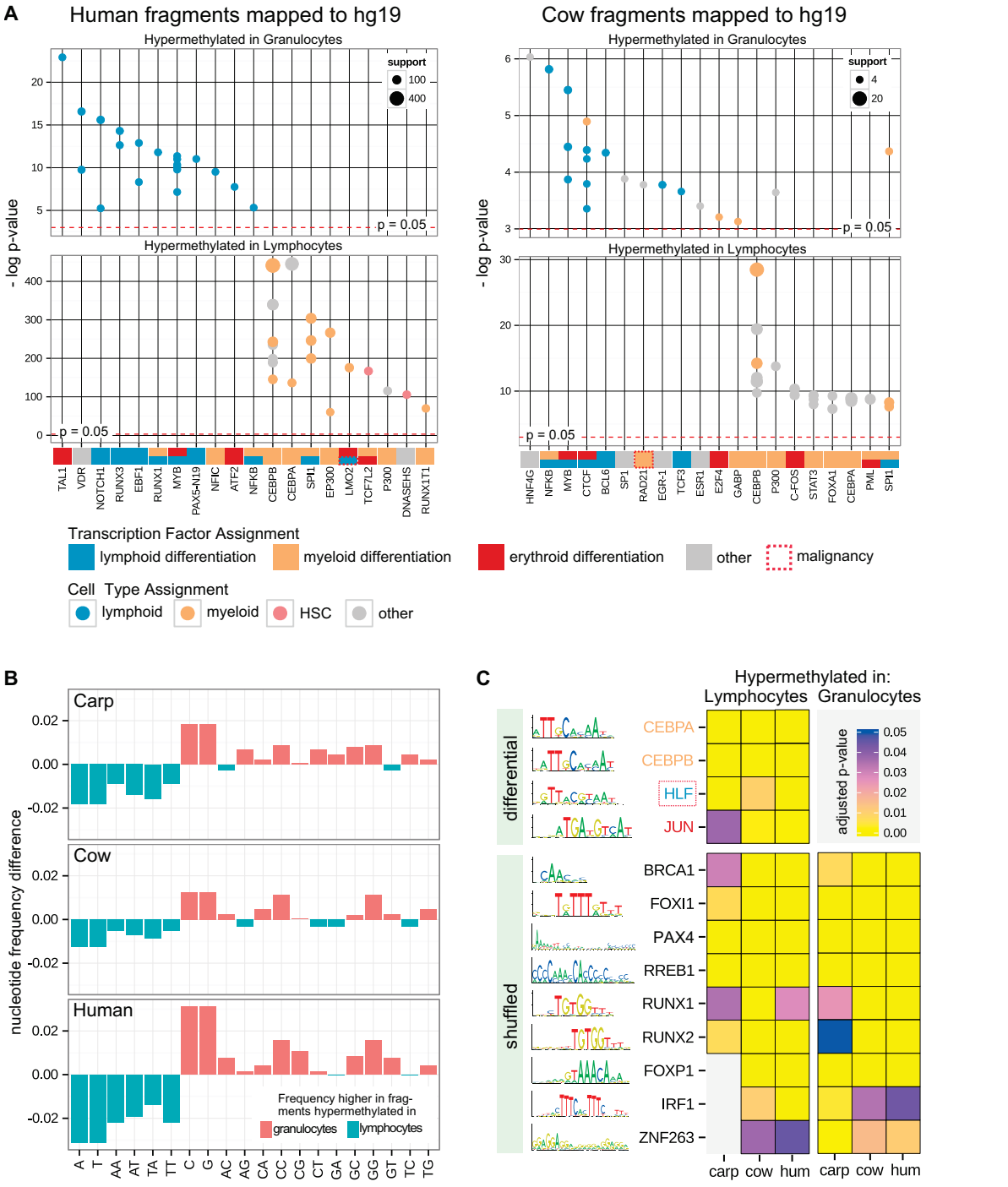


Figure 5. Biological Interpretation of DNA Methylation Differences

(A) Region enrichment analysis for differentially methylated deduced genome fragments that have been cross-mapped to the human genome (hg19). The top-20 enriched region sets obtained by LOLA analysis are shown. Uncorrected p values are plotted on the y axis, and the number of overlapping regions is indicated by bubble size. Each dot represents a region set in the database, and the red dashed line indicates p values of 0.05. Similar plots for carp and for cross-mapping to the mouse genome (mm10) are shown in Figure S5B. Cell-type-specific gene functions are based on literature search and indicated through colored boxes on the x axis.

(B) Nucleotide frequency differences between the top-500 deduced genome fragments with increased DNA methylation in granulocytes versus lymphocytes (red) and vice versa (blue).

(legend continued on next page)

sequencing assays such as RAD-seq (Baird et al., 2008) and GBS (Elshire et al., 2011) for research in natural populations and non-model organisms, we expect our method to be broadly useful for EWASs in the context of ecology and evolution.

The described method should be applicable to any animal and plant species with appreciable levels of DNA methylation, and it is readily adapted to different genome compositions and sequencing depths by selecting an appropriate restriction enzyme (or enzyme combinations). Here we focused on vertebrates, where DNA methylation is largely restricted to CpG dinucleotides and the MspI restriction enzyme is an ideal choice. MspI enriches for CpG islands and gene promoters, while also providing a broad sampling of other genomic regions such as enhancers, gene bodies, CpG island shores, and repetitive elements. Furthermore, every read contains at least one CpG (at the MspI restriction site), which increases cost-effectiveness for vertebrate genomes. Importantly, our method can be used to map not only CpG methylation, as we demonstrate here, but also non-CpG methylation (Ziller et al., 2011), which is widespread among non-vertebrate species and also present in certain vertebrate cell types.

On the computational side, we developed the RefFreeDMA method and software to build a deduced genome directly from the bisulfite sequencing reads, to quantify DNA methylation at the level of single CpG sites and deduced fragments, and to detect and rank DNA methylation differences between samples and sample groups. RefFreeDMA overcomes relevant limitations of an existing method that uses de novo assembly of MeDIP-seq reads (Kaspi et al., 2014), namely low resolution, susceptibility to biases, and lack of quantification, and it is more powerful and more widely applicable than read mapping to the genome of a related species (Weyrich et al., 2014), which requires a closely matched genome and a second, unconverted library. Furthermore, we present two approaches (cross-mapping and motif enrichment analysis) for interpreting the identified differentially methylated regions in the absence of a reference genome.

To validate our method, we established and analyzed a cross-species DNA methylation dataset comprising multiple blood cell types in two mammalian species (human and cow) and one fish (carp). All cell types were enriched to >95% purity by a sorting strategy that is particularly useful for working with non-model organisms because it does not require any species-specific antibodies. Bioinformatic analysis in the three species with and without the respective reference genomes gave rise to consistent and informative results. For example, we observed that the most differentially methylated fragments in the two mammalian species were predominantly hypermethylated in lymphocytes, whereas no such bias was present in carp (Figures 4C and S4D). We also identified characteristic binding motifs of lineage-specific transcription factors that were consistently enriched among differentially methylated fragments of all three species (Figure 5C).

Despite the good results that we obtained in our validation of RefFreeDMA, there are several inherent limitations of reference-free DNA methylation analysis that potential users of our method should keep in mind. First, repetitive elements with high sequence similarity can get merged into a single deduced genome fragment, which is why RefFreeDMA tends to report moderately fewer covered CpGs than we obtained using reference-based analysis. Second, cytosines that are unmethylated in all samples of one species will not be represented in the deduced genome (case 4 in Figure S2), unless one RRBS sample is sequenced without bisulfite conversion and added to the analysis. Third, our method does not perform de novo assembly of deduced genome fragments, which would require substantially deeper and broader sequencing coverage than is typically affordable. It can therefore happen that the same CpG is included twice in two partially overlapping fragments (case 2 in Figure S2). However, based on our analysis of the validation dataset, this type of bias appears to be negligible (Figure S4C).

In summary, we expect that RefFreeDMA in combination with our optimized RRBS protocol will be useful for researchers who are interested in analyzing DNA methylation in non-model organisms without the need of a reference genome. Apart from assessing cell-type-specific DNA methylation as demonstrated here, other applications of RefFreeDMA may include EWASs for phenotypic differences in natural populations, agricultural research on the epigenetic effect of different feeds, drugs, and rearing conditions, and meta-epigenome studies of DNA methylation in entire ecosystems.

EXPERIMENTAL PROCEDURES

Sample Acquisition

For human, cow, and carp, 5–10 ml of peripheral blood was obtained from two male and two female individuals, anti-coagulated by 2 mg/ml K_2EDTA and processed within 1 hr after collection. Human blood samples were obtained by venipuncture from healthy donors by a qualified physician. All donors provided informed consent. The study was conducted in accordance with the principles laid down in the Declaration of Helsinki, overseen by the ethics commission of the Medical University of Vienna. Cow blood samples were obtained post-mortem from a slaughterhouse. Carp blood samples were obtained post-mortem from a fish vendor. For the other species (mouse, rat, dog, chicken, sea bass, and zebrafish), purified DNA was provided by the collaborators listed in the Acknowledgments.

Cell Purification

Leukocytes were isolated from whole blood by removing the erythrocytes through hypotonic lysis. Specifically, 5 ml of whole blood was incubated with 9 ml ddH_2O for 1 min. The lysis was stopped by adding 1 ml of 10× PBS to the sample. Leukocytes were pelleted by centrifuging for 5 min at 550 *g*. If the pellet was still red, a second round of lysis was initiated by resuspending the pellet in 1 ml 1× PBS. Subsequently, 4.5 ml of ddH_2O was added and after 30 s the lysis reaction was stopped by adding 0.5 ml 10× PBS. Leukocytes were pelleted by centrifuging for 3 min at 550 *g*. Finally, the pellet was washed in 1 ml 1× PBS and then resuspended in 500–800 μl RPMI-1640 medium supplemented with 10% fetal calf serum (FCS). The cell suspension was then filtered into a FACS tube, and cell populations were sorted by FACS based on their forward and side scatter properties. Sorting was performed on a BD FACS Aria 1 with a 70-μm nozzle, which allowed for a maximum sorting speed of 30,000 events per second. For each population,

(C) Enrichment of known sequence motifs associated with transcription factor binding sites among the top-500 deduced genome fragments with increased DNA methylation in granulocytes versus lymphocytes (right) and vice versa (left). The motif analysis used either the opposing group ("differential") or randomly shuffled sequences with the same mono- and dinucleotide composition ("shuffled") as background. The diagram only shows motifs that were enriched in all three species; the complete sets of enriched transcription factor binding motifs are shown in Figures S5D and S5E.

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

between 500,000 and 3 million cells were obtained. Giemsa stained cytospins were produced for each sorted cell population, and the purity was assessed at 100× magnification.

DNA Isolation

The Allprep DNA/RNA Mini kit (QIAGEN) was used for DNA isolation. Cells were lysed in 600 μl Buffer RLT Plus supplemented with 1% β-Mercaptoethanol and vortexed thoroughly for at least 5 min. The procedure of isolating DNA and RNA was performed according to protocol. DNA was stored at −20°C.

RRBS Library Preparation

For RRBS, 100 ng of genomic DNA was digested for 12 hr at 37°C with 20 units of MspI (New England Biolabs, R0106L) in 30 μl of 1× NEB buffer 2. To retain even the smallest fragments and to minimize the loss of material, end preparation and adaptor ligation were performed in a single-tube setup. End fill-in and A-tailing were performed by addition of Klenow Fragment 3′ > 5′ exo- (New England Biolabs, M0212L) and dNTP mix (10 mM dATP, 1 mM dCTP, 1 mM dGTP). After ligation to methylated Illumina TruSeq LT v2 adaptors using Quick Ligase (New England Biolabs, M2200L), the libraries were size selected by performing a 0.75× cleanup with AMPure XP beads (Beckman Coulter, A63881). The libraries were pooled in combinations of six based on qPCR data and subjected to bisulfite conversion using the EZ DNA Methylation Direct Kit (Zymo Research, D5020) with the following changes to the manufacturer's protocol: conversion reagent was used at 0.9× concentration, incubation performed for 20 cycles of 1 min at 95°C, 10 min at 60°C, and the desulphonation time was extended to 30 min. These changes increase the number of CpG dinucleotides covered by reducing double-strand break formation in larger library fragments. Bisulfite-converted libraries were enriched using PfuTurbo Cx Hotstart DNA Polymerase (Agilent, 600412). The minimum number of enrichment cycles was estimated by qPCR. After a 2× AMPure XP cleanup, quality control was performed using the Qubit dsDNA HS (Life Technologies, Q32854) and Experion DNA 1k assays (BioRad, 700-7107). RRBS libraries were sequenced on the Illumina HiSeq 2000 platform in 50-bp single-read mode.

Bisulfite Conversion Controls

In order to monitor the efficiency of the bisulfite conversion and to check for underconversion of unmethylated cytosines as well as overconversion of methylated cytosines, custom-designed and synthesized methylated and unmethylated oligonucleotides were spiked into each sample at a concentration of 0.1% of the genomic DNA. For each sample, sequencing reads were aligned to the control sequences using Bismark with default settings (Krueger and Andrews, 2011). Conversion metrics are reported in Table S1.

RRBS Data Preprocessing

Sequencing data were processed with illumina2bam-tools v.1.12, and the resulting BAM files were converted to fastq format using SamToFastq.jar (picard-tools v.1.100) with the INCLUDE_NON_PF_READS parameter set to FALSE. All reads were trimmed for adaptor sequences and low-quality sequences using trimgalore v.0.3.3 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the following command: *trim_galore -q 20–phred33 -a "AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC"–stringency 1 -e 0.1–length 16–output_dir $output_dir $input_fastq.*

Derivation of a Deduced Genome

Based on the trimmed RRBS reads for a given species and analysis, a deduced genome is constructed in six steps: (1) *Pre-filtering.* To reduce the number of reads that need to be processed, one representative read is kept for each read sequence and sample. Furthermore, reads that stand a high chance of arising from sequencing errors are discarded by requiring that each read occurs at least twice among four samples after converting all Cs to Ts. (2) *Preliminary read grouping.* To be computationally effective, we perform read grouping initially by exact string matching. Reads that share the same sequence in their fully converted form (all Cs replaced by Ts) are combined into one pre-consensus sequence by assigning a C to each position at which at least 5% of the reads contain a C in their unconverted form. (3) *Consensus building.* To combine highly similar but not identical fragments into one consensus, the pre-consensus fragments are grouped by sequence similarity using an all-against-all alignment of the C to T converted fragments with Bowtie2 v.2.2.3 (Langmead and Salzberg, 2012) using the following command: *bowtie2 -t -q–phred33–end-to-end -N 1 -L 22–norc–n-ceil "L,0,0.2"–mp 3–np 0–score-min "L,-0.6,-0.6" -k 300 -D 3–rdg "20,20"–rfg "20,20" -p 4 -x $reference -U $fastq -S $out_sam*. Fragments that match with less than 8% maximum mismatch ratio are merged by assigning them to the largest available group. For each group, a consensus sequence is deduced by assigning the majority base to each position, while assigning Cs to all positions at which at least 5% of the fragments contain a C. (4) *Consensus refinement.* For those groups in which some fragments exhibit more than 5% mismatches relative to the consensus, the diverging reads are assigned to separate groups, and a new consensus is built for the respective groups. This procedure is repeated until no fragment-to-consensus mismatch rate exceeds 5%. (5) *Merging of reverse complements.* After bisulfite conversion, reads originating from the two strands of the same DNA fragment are often not identified as reverse complements during the Bowtie2 alignment and are therefore not automatically merged into one consensus. To overcome this problem, all reads that start and end with the RRBS restriction site (MspI: 5′ [CT]GG – [CT][CT]G 3′) are tested for whether they become perfect reverse complements of each other when all Cs are replaced by Ts and all Gs are replaced by As. For each pair to be merged, a consensus is formed by assigning a C to all T positions in the sequence of the forward partner at which the reverse-complement partner shows a C. (6) *Concatenation into one deduced genome.* In the final step, the merged deduced genome fragments are concatenated into one deduced genome that can be used for alignment, DNA methylation calling, and differential methylation analysis in the same way as a regular reference genome. To avoid creating artificial sequences at the concatenation sites, spacer sequences consisting of 50 Ns (equaling the read length) are added between the deduced genome fragments. Of note, all key parameters in RefFreeDMA have been empirically optimized and can be changed by the user of the software.

Mapping and DNA Methylation Calling

Bisulfite alignment of the RRBS reads to the deduced genomes and to the reference genomes, as well as the mapping of the deduced genome fragments to the reference genomes was performed using BSMAP v2.74 (Xi and Li, 2009) with the following command line: *bsmap -a $input_fastq -d $ref_genome_fasta -o $output_bam -D C-CGG -w 100 -v 0.08 -r 1 -p 4 -n 0 -S 1 -f 5 –u*. For cross-mapping and alignment to the deduced genomes, the -D parameter was not set, disabling the RRBS mode to allow mapping of reads independently of restriction sites. Also, for cross-mapping, the maximum allowed error rate (*-v*) was set to 0.2. The human (hg19) and cow (bosTau6) reference genomes were downloaded from the UCSC Genome Browser, and the carp reference genome was downloaded from the European Nucleotide Archive (ENA) project PRJEB7241 assembly GCA_000951615.1. For better handling, the 9,377 scaffolds of the carp genome were concatenated into ten artificial chromosomes using stretches of Ns as separators. DNA methylation calling was performed using the *biseqMethCalling.py* software (Bock et al., 2010).

Differential Methylation Analysis

CpG sites exhibiting differential DNA methylation between predefined groups of samples were identified using hierarchical linear models as implemented in the *limma* R package. Multiple testing correction was performed for CpG sites using the false discovery rate method implemented in R's *p.adjust()* function. To assess the significance of differential DNA methylation for entire fragments, multiple testing corrected p values for all CpG sites contained in a fragment were combined using an extension of Fisher's method (Makambi, 2003) as implemented in RnBeads (Assenov et al., 2014). Differentially methylated fragments were priority ranked based on statistical significance as well as effect size, calculating ranks individually for p value, log fold change, and absolute difference in DNA methylation levels and then selecting the worst of the three ranks as representative for the fragment. This way, fragments that achieve top ranks in all of the measures are favored, whereas fragments that are assigned a bad rank in one or more of the measures are penalized.

Software Properties

RefFreeDMA is a Linux-based software pipeline that supports the various steps of reference genome independent analysis of differential DNA methylation based on RRBS data. External software requirements are limited to standard command line tools for next generation sequencing analysis, including picardtools, samtools, trimgalore, bowtie2, and bsmap. Runtime and memory usage depend on the number of samples, the number of reads per sample, the RRBS library complexity, and whether RefFreeDMA's support for parallelization is used. For the presented datasets, which comprise 12 to 20 samples per species with ~18 million 50-bp single-end reads per sample, one complete run using four cores (Intel Xeon E5-2650 processor) takes about 9 hr (wall-clock time) with parallelization and 40 hr (wall-clock time) without. The peak memory usage is 15 GB during consensus building. Although this study focuses on CpG methylation, our software also supports non-CpG methylation (when the *nonCpG* parameter is set to TRUE). RefFreeDMA is available as open source under the GPLv3 license: http://RefFreeDMA.computational-epigenetics.org.

Comparison between Reference-Free and Reference-Based Analysis

Correspondence between the published reference genomes and the deduced genomes is determined by mapping the deduced genome fragments to the corresponding reference genome. The resulting associations between CpG sites in the deduced genome and the reference genome serve as the basis for the validations. Figure S2 depicts the correct match between the two approaches (case 1) as well as four scenarios in which discrepancies between reference-free and reference-based analysis are expected (cases 2 to 5). Comparisons between the reference-free and reference-based approaches are performed at the level of individual CpGs and at the level of deduced genome fragments.

Cross-Mapping Analysis

In order to establish a connection between deduced genome fragments identified by RefFreeDMA in one species and well-annotated genomes of other species, deduced fragments were mapped to the human genome (hg19) and the mouse genome (mm10) using BSMAP/RRBSMAP with a maximum allowed mismatch rate of 20% as described in Mapping and DNA Methylation Calling. Overlaps between the genomic positions of mapped deduced genome fragments and annotations on the respective genome can then be used to perform enrichment analysis for the deduced fragments. We assessed differentially methylated fragments for enrichment of genomic annotations using LOLA (Sheffield and Bock, 2015). LOLA tests for significant enrichment of overlap between user-defined genomic regions of interest (i.e., the fragment mapping positions) and experimentally annotated genomic regions, which are provided as a database. The matched genomic regions for the differentially methylated fragments (mean coverage > 2 and adjusted $p < 0.05$) of granulocytes or lymphocytes were used as primary input regions (user set), while the genomic regions of all mapped deduced genome fragments were used as background (universe). The regions database for human (hg19) consisted of region sets downloaded from Cistrome, CODEX, ENCODE, and the UCSC Genome Browser as well as custom sets for DNase hypersensitivity sites (Sheffield et al., 2013). The region database for mouse (mm10) consisted of region sets downloaded from CODEX and ENCODE.

Motif Enrichment Analysis

Motif enrichment analysis was performed using the command-line version of the AME tool (McLeay and Bailey, 2010) from the MEME package. We used the average odds score as sequence scoring method and the rank-sum test as motif enrichment test. All motifs were obtained from the JASPAR CORE (2014) Vertebrates database (Mathelier et al., 2014). Only enrichments with an adjusted p value lower than 0.05 were reported. In order to find motifs that are differentially enriched among differentially methylated fragments, the top-500 differentially methylated fragments (mean coverage > 2 and adjusted $p < 0.05$) of one sample group were used as primary input sequences, while the top-500 differentially methylated fragments of the other group were used as background (control sequences). To correct for motif enrichment due to base composition bias (Figures 5B and S5C), we performed the same analysis on random sequences that were constructed to reflect the base compositions of both groups on single nucleotide and dinucleotide level in 50 iterations each. To this end, the base compositions of the original sequences were determined using the fasta-get-markov tool from the MEME package. The 0^{th}- and 1^{st}-order Markov models for each group were then used as input for the gendb tool, which constructed 500 random sequences (length ~50 bases) according to the models. This process was repeated 50 times with different random seeds. Finally, for each iteration AME was run on the shuffled sequences of one group as input and the shuffled sequences of the other group as background. All motifs that were detected as significantly enriched in more than 60% of all iterations were identified as false positives due to base composition bias and removed from the list of differentially enriched motifs identified for the original sequences. Furthermore, to identify motifs that might be enriched in differentially methylated fragments of both groups, we ran AME using the original sequences as input and the respective shuffled sequences as background. Only motifs that were found to be enriched in at least 95% of the iterations were reported as truly enriched in the differentially methylated fragments compared with the randomly shuffled sequences. For each enriched motif, the least significant p value was reported.

ACCESSION NUMBERS

The DNA methylation data reported in this paper have been submitted to the NCBI GEO and are available under accession number GEO: GSE74026.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.11.024.

AUTHOR CONTRIBUTIONS

J.K. and C.B. designed the study. P.D., M.F., and C.B. optimized the RRBS protocol. J.K. acquired and prepared the samples. J.K., D.P., and G.F. performed FACS sorting. P.D. and J.H. made the RRBS sequencing libraries. J.K. developed RefFreeDMA and performed the computational analysis with input from N.C.S. and C.B. J.K. and C.B. wrote the manuscript with input from all co-authors.

ACKNOWLEDGMENTS

We thank the Biomedical Sequencing Facility at CeMM for assistance with next generation sequencing, Fabian Müller for providing the *biseqMethCalling.py* software, and all members of the Bock lab for their help and advice. We also thank Sylvia Knapp, Denise Barlow, Thomas van Gurp, and Christian Remmele for comments and suggestions, Marc Mößmer (Biofisch GmbH) for providing carp blood, Fleischerei Leopold Hödl for providing cow blood, and the following researchers for providing DNA from additional species: Clarissa Gerhäuser (rat), Vardhman Rakyan (dog), Marcela Hermann (chicken), Kaja H. Skjærven (zebrafish), and Francesc Piferrer (sea bass). This work was performed in the context of the BLUEPRINT project (European Union's Seventh Framework Programme grant agreement No. 282510) and the ERA-NET projects EpiMark (FWF grant agreement no. I 1575-B19) and CINOCA (FWF grant agreement no. I 1626-B22). It was co-funded by a Marie Curie Career Integration Grant (European Union's Seventh Framework Programme grant agreement No. PCIG12-GA-2012-333595). J.K. was supported by a DOC Fellowship of the Austrian Academy of Sciences. N.C.S. was supported by a Human Frontier Science Program long-term fellowship (LT000211/2014). C.B. was supported by a New Frontiers Group award of the Austrian Academy of Sciences.

Received: September 17, 2015
Revised: October 12, 2015
Accepted: November 4, 2015
Published: December 3, 2015

Please cite this article in press as: Klughammer et al., Differential DNA Methylation Analysis without a Reference Genome, Cell Reports (2015), http://dx.doi.org/10.1016/j.celrep.2015.11.024

REFERENCES

Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. Nat. Biotechnol. *30*, 224–226.

Akagi, T., Thoennissen, N.H., George, A., Crooks, G., Song, J.H., Okamoto, R., Nowak, D., Gombart, A.F., and Koeffler, H.P. (2010). In vivo deficiency of both C/EBPβ and C/EBPε results in highly defective myeloid differentiation and lack of cytokine response. PLoS ONE *5*, e15419.

Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. Nat. Methods *11*, 1138–1140.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE *3*, e3376.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nat. Rev. Cancer *11*, 726–734.

Beck, S. (2010). Taking the measure of the methylome. Nat. Biotechnol. *28*, 1026–1028.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288–295.

Bock, C. (2012). Analysing and interpreting DNA methylation data. Nat. Rev. Genet. *13*, 705–719.

Bock, C., Tomazou, E.M., Brinkman, A.B., Müller, F., Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H.G., and Meissner, A. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat. Biotechnol. *28*, 1106–1114.

Bock, C., Beerman, I., Lien, W.H., Smith, Z.D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D.J., and Meissner, A. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol. Cell *47*, 633–647.

Boyle, P., Clement, K., Gu, H., Smith, Z.D., Ziller, M., Fostel, J.L., Holmes, L., Meldrim, J., Kelley, F., Gnirke, A., and Meissner, A. (2012). Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol. *13*, R92.

Day, D.S., Luquette, L.J., Park, P.J., and Kharchenko, P.V. (2010). Estimating enrichment of repetitive elements from high-throughput sequence data. Genome Biol. *11*, R69.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE *6*, e19379.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Feil, R., and Fraga, M.F. (2011). Epigenetics and the environment: emerging patterns and implications. Nat. Rev. Genet. *13*, 97–109.

Feinberg, A.P., and Tycko, B. (2004). The history of cancer epigenetics. Nat. Rev. Cancer *4*, 143–153.

Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. *9*, e1003994.

Foley, D.L., Craig, J.M., Morley, R., Olsson, C.A., Dwyer, T., Smith, K., and Saffery, R. (2009). Prospects for epigenetic epidemiology. Am. J. Epidemiol. *169*, 389–400.

Gazit, R., Garrison, B.S., Rao, T.N., Shay, T., Costello, J., Ericson, J., Kim, F., Collins, J.J., Regev, A., Wagers, A.J., and Rossi, D.J.; Immunological Genome Project Consortium (2013). Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. Stem Cell Reports *1*, 266–280.

Gillman, M.W. (2005). Developmental origins of health and disease. N. Engl. J. Med. *353*, 1848–1850.

Greig, K.T., Carotta, S., and Nutt, S.L. (2008). Critical roles for c-Myb in hematopoietic progenitor cells. Semin. Immunol. *20*, 247–256.

Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat. Protoc. *6*, 468–481.

Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 86.

Jablonka, E., and Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. Q. Rev. Biol. *84*, 131–176.

Jaffe, A.E., and Irizarry, R.A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. *15*, R31.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Kaspi, A., Ziemann, M., Keating, S.T., Khurana, I., Connor, T., Spolding, B., Cooper, A., Lazarus, R., Walder, K., Zimmet, P., and El-Osta, A. (2014). Non-referenced genome assembly from epigenomic short-read data. Epigenetics *9*, 1329–1338.

Klunker, S., Chong, M.M.W., Mantel, P.Y., Palomares, O., Bassin, C., Ziegler, M., Rückert, B., Meiler, F., Akdis, M., Littman, D.R., and Akdis, C.A. (2009). Transcription factors RUNX1 and RUNX3 in the induction and suppressive function of Foxp3+ inducible regulatory T cells. J. Exp. Med. *206*, 2701–2715.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. *11*, 204–220.

Liebermann, D.A., and Hoffman, B. (2002). Myeloid differentiation (MyD) primary response genes in hematopoiesis. Oncogene *21*, 3391–3402.

Liebl, A.L., Schrey, A.W., Richards, C.L., and Martin, L.B. (2013). Patterns of DNA methylation throughout a range expansion of an introduced songbird. Integr. Comp. Biol. *53*, 351–358.

Makambi, K. (2003). Weighted inverse chi-square method for correlated significance tests. J. Appl. Stat. *30*, 225–234.

Massicotte, R., Whitelaw, E., and Angers, B. (2011). DNA methylation: A source of random variation in natural populations. Epigenetics *6*, 421–427.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. *42*, D142–D147.

McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. *33*, 5868–5877.

Michels, K.B., Binder, A.M., Dedeurwaerder, S., Epstein, C.B., Greally, J.M., Gut, I., Houseman, E.A., Izzi, B., Kelsey, K.T., Meissner, A., et al. (2013). Recommendations for the design and analysis of epigenome-wide association studies. Nat. Methods *10*, 949–955.

Mill, J., and Heijmans, B.T. (2013). From promises to practical strategies in epigenetic epidemiology. Nat. Rev. Genet. *14*, 585–594.

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.M., and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife *4*, 1–20.

O'Neill, R.J., O'Neill, M.J., and Graves, J.A. (1998). Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature *393*, 68–72.

Orkin, S.H. (1995). Transcription factors and hematopoietic development. J. Biol. Chem. *270*, 4955–4958.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nat. Biotechnol. *28*, 1057–1068.

Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. Nat. Rev. Genet. *12*, 529–541.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature *447*, 425–432.

Robertson, K.D. (2005). DNA methylation and human disease. Nat. Rev. Genet. *6*, 597–610.

Rosenbauer, F., and Tenen, D.G. (2007). Transcription factors in myeloid development: balancing differentiation with transformation. Nat. Rev. Immunol. *7*, 105–117.

Schrey, A.W., Coon, C.A.C., Grispo, M.T., Awad, M., Imboma, T., McCoy, E.D., Mushinsky, H.R., Richards, C.L., and Martin, L.B. (2012). Epigenetic variation may compensate for decreased genetic variation with introductions: A case study using house sparrows (Passer domesticus) on two continents. Genet. Res. Int. *2012*, 1–7.

Schrey, A.W., Alvarez, M., Foust, C.M., Kilvitis, H.J., Lee, J.D., Liebl, A.L., Martin, L.B., Richards, C.L., and Robertson, M. (2013). Ecological Epigenetics: Beyond MS-AFLP. Integr. Comp. Biol. *53*, 340–350.

Sheffield, N.C., and Bock, C. (2015). LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics, Published online October 27, 2015. http://dx.doi.org/10.1093/bioinformatics/btv612.

Sheffield, N.C., Thurman, R.E., Song, L., Safi, A., Stamatoyannopoulos, J.A., Lenhard, B., Crawford, G.E., and Furey, T.S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. *23*, 777–788.

Skinner, M.K., Gurerrero-Bosagna, C., Haque, M.M., Nilsson, E.E., Koop, J.A.H., Knutie, S.A., and Clayton, D.H. (2014). Epigenetics and the evolution of Darwin's Finches. Genome Biol. Evol. *6*, 1972–1989.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Tenen, D.G., Hromas, R., Licht, J.D., and Zhang, D.E. (1997). Transcription factors, normal myeloid development, and leukemia. Blood *90*, 489–519.

Waterland, R.A., and Michels, K.B. (2007). Epigenetic epidemiology of the developmental origins hypothesis. Annu. Rev. Nutr. *27*, 363–388.

Weyrich, A., Schüllermann, T., Heeger, F., Jeschek, M., Mazzoni, C.J., Chen, W., Schumann, K., and Fickel, J. (2014). Whole genome sequencing and methylome analysis of the wild guinea pig. BMC Genomics *15*, 1036.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Xi, Y., Bock, C., Müller, F., Sun, D., Meissner, A., and Li, W. (2012). RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. Bioinformatics *28*, 430–432.

Ziller, M.J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C.B., Bernstein, B.E., Lengauer, T., et al. (2011). Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLoS Genet. *7*, e1002389.

Cell Reports

Supplemental Information

Differential DNA Methylation Analysis without a Reference Genome

Johanna Klughammer, Paul Datlinger, Dieter Printz, Nathan C. Sheffield, Matthias Farlik, Johanna Hadler, Gerhard Fritsch, and Christoph Bock

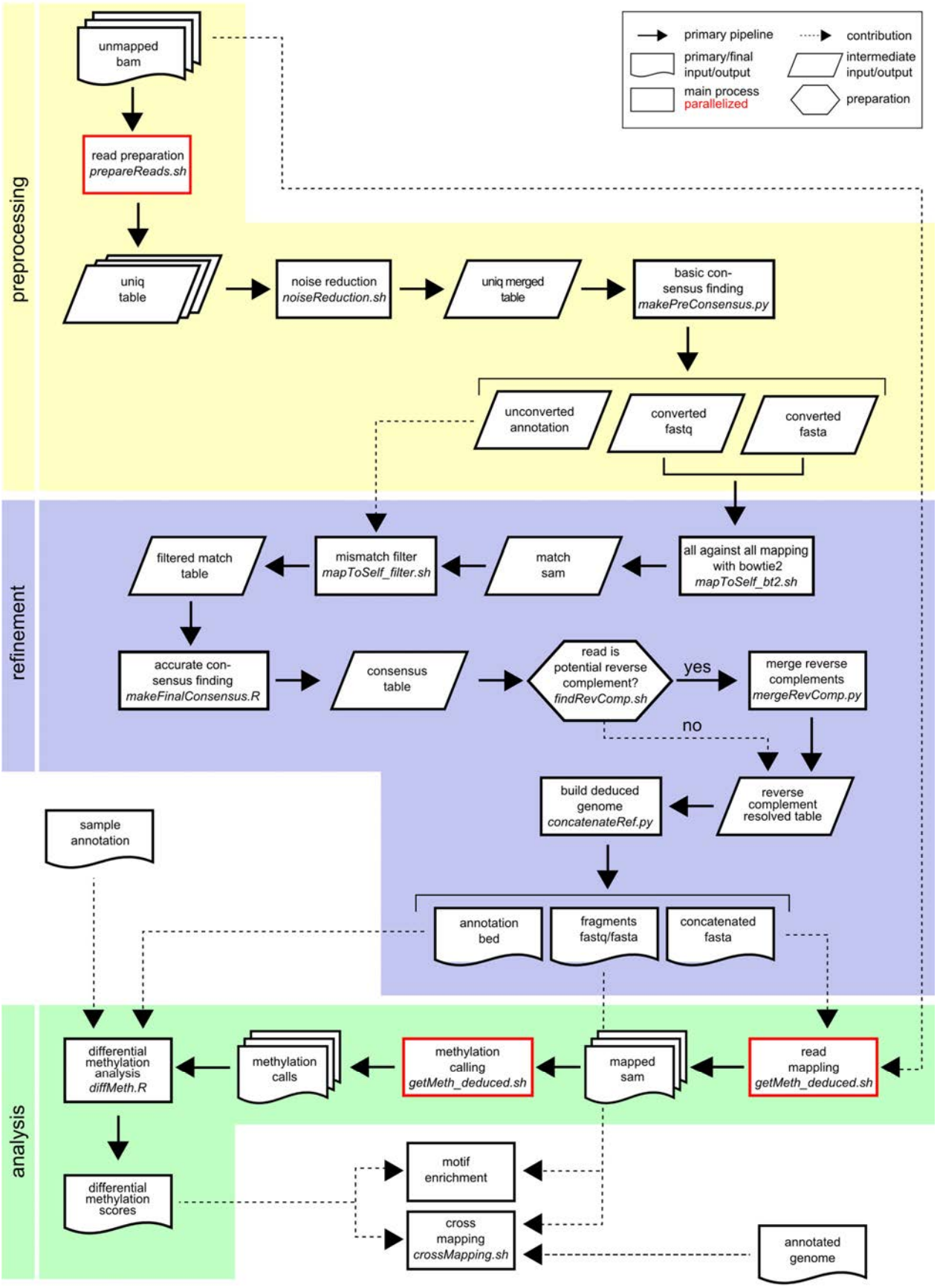


Figure S1. UML diagram outlining the RefFreeDMA software and analysis workflow, Related to Figure 1

The diagram illustrates the RefFreeDMA software and its key computational steps for performing reference-free analysis of differential DNA methylation, starting from raw RRBS reads and resulting in a ranked list of differentially methylated sites and fragments.

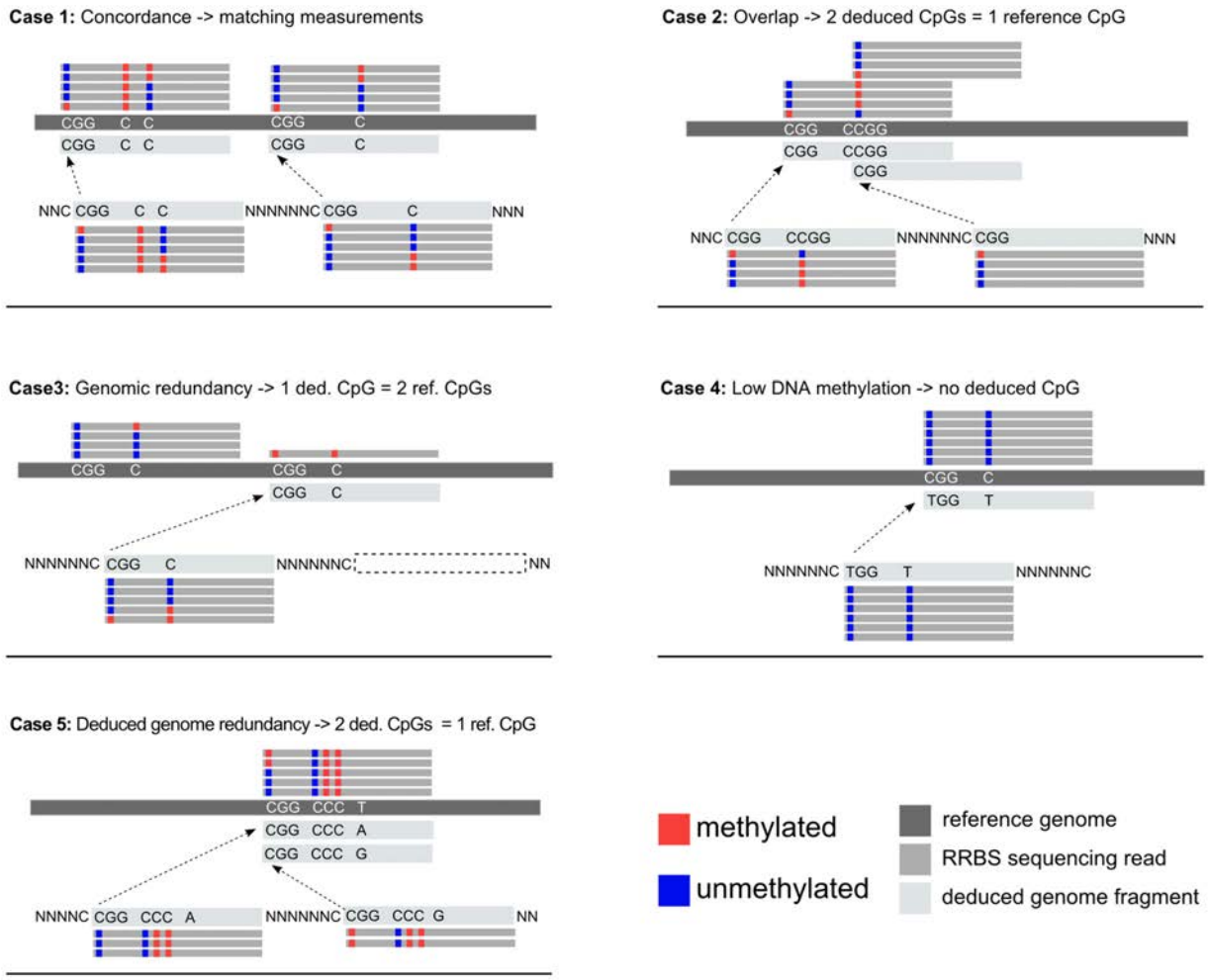


Figure S2. Sources of discrepancy between reference-free and reference-based analysis, Related to Figure 1

Case 1 depicts concordance between the two approaches, which applies to the vast majority of non-repetitive fragments that are not entirely unmethylated in all samples. All matching CpGs are uniquely assigned to each other when aligning the deduced genome fragments to the reference genome. *Case 2* depicts a scenario in which two deduced genome fragments overlap when aligned to the reference genome. Here, two measurements in the deduced genome are represented by only one measurement in the reference genome. *Case 3* depicts genomic redundancy caused by repetitive sequences in the reference genome. In the deduced genome, these similar or sequence-identical regions are represented by one deduced genome fragment. Multiple CpG sites in the reference genome are thus represented by only one site in the deduced genome. *Case 4* depicts the scenario where all reads are completely unmethylated for a given set of CpG sites. Deduced genome fragments covering these sites will contain a T instead of a C at the respective position, thereby reducing the number of CpG sites in the deduced genome. *Case 5* depicts the effect of deduced genome redundancy, which can occur when fragments contain sequencing errors that make them too dissimilar to be merged into one consensus.

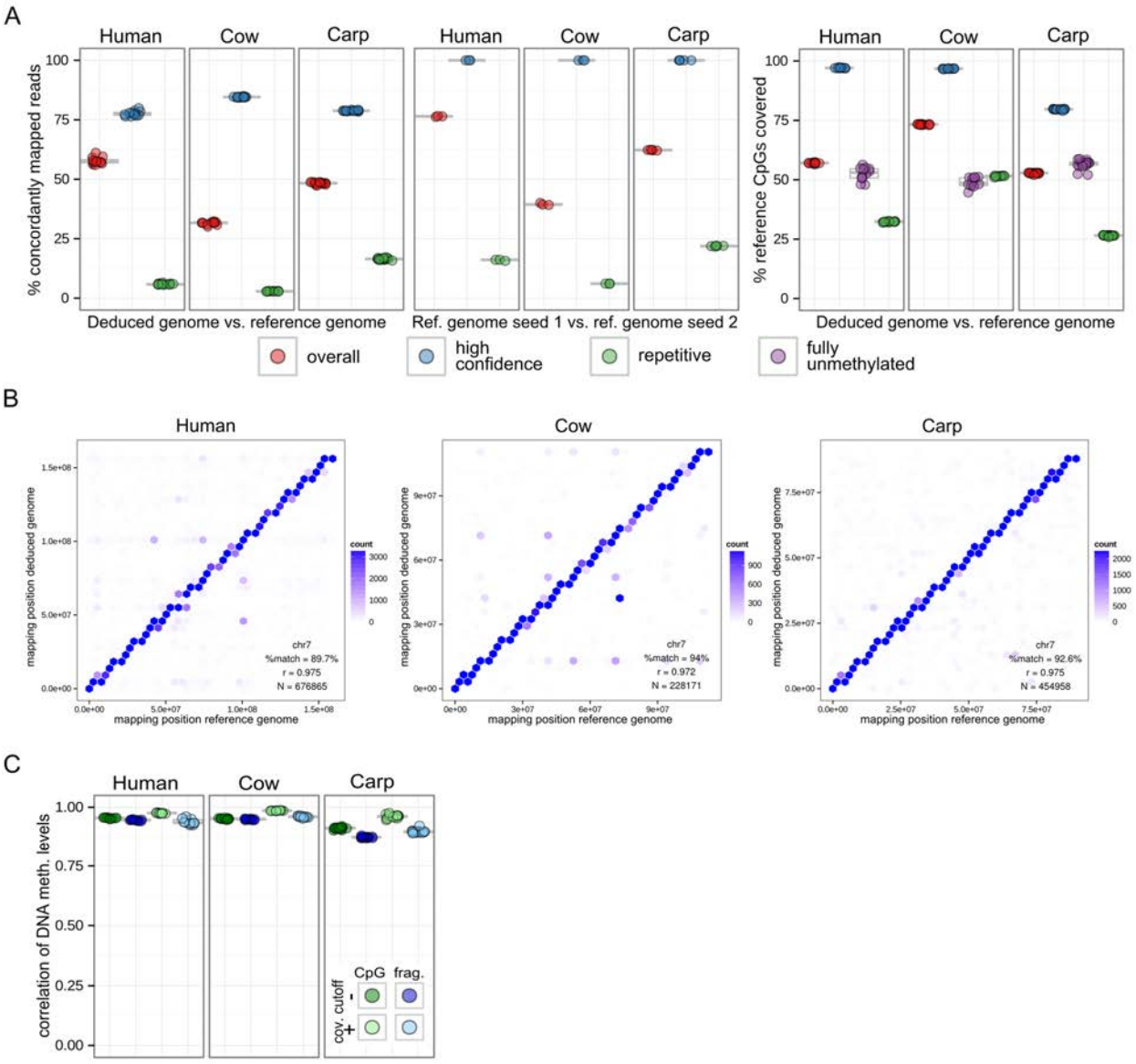


Figure S3. Comparison of reference-free & reference-based DNA methylation analysis, Related to Figure 3

(**A**) Concordance of mapped read positions (left) and covered CpG sites (right) between the reference-based and reference-free methods. For comparison, the concordance is also shown for the case of aligning the reads twice to the reference genome using different seeds for random assignment of reads that map to multiple positions (middle). "High confidence" fragments are those that are neither repetitive nor unmethylated in all samples. (**B**) Scatterplots illustrating the concordance of read mapping positions between the reference-free (y-axis) and reference-based (x-axis) methods. Representative plots of chromosome 7 are shown for each species (r: Pearson correlation; N: number of RRBS reads). (**C**) Pearson correlation of DNA methylation levels obtained with the two approaches, calculated for CpG sites as well as deduced genome fragments (frag.) with (+) and without (-) coverage filtering (requiring at least eight and not more than 200 mapped reads per CpG site or fragment).

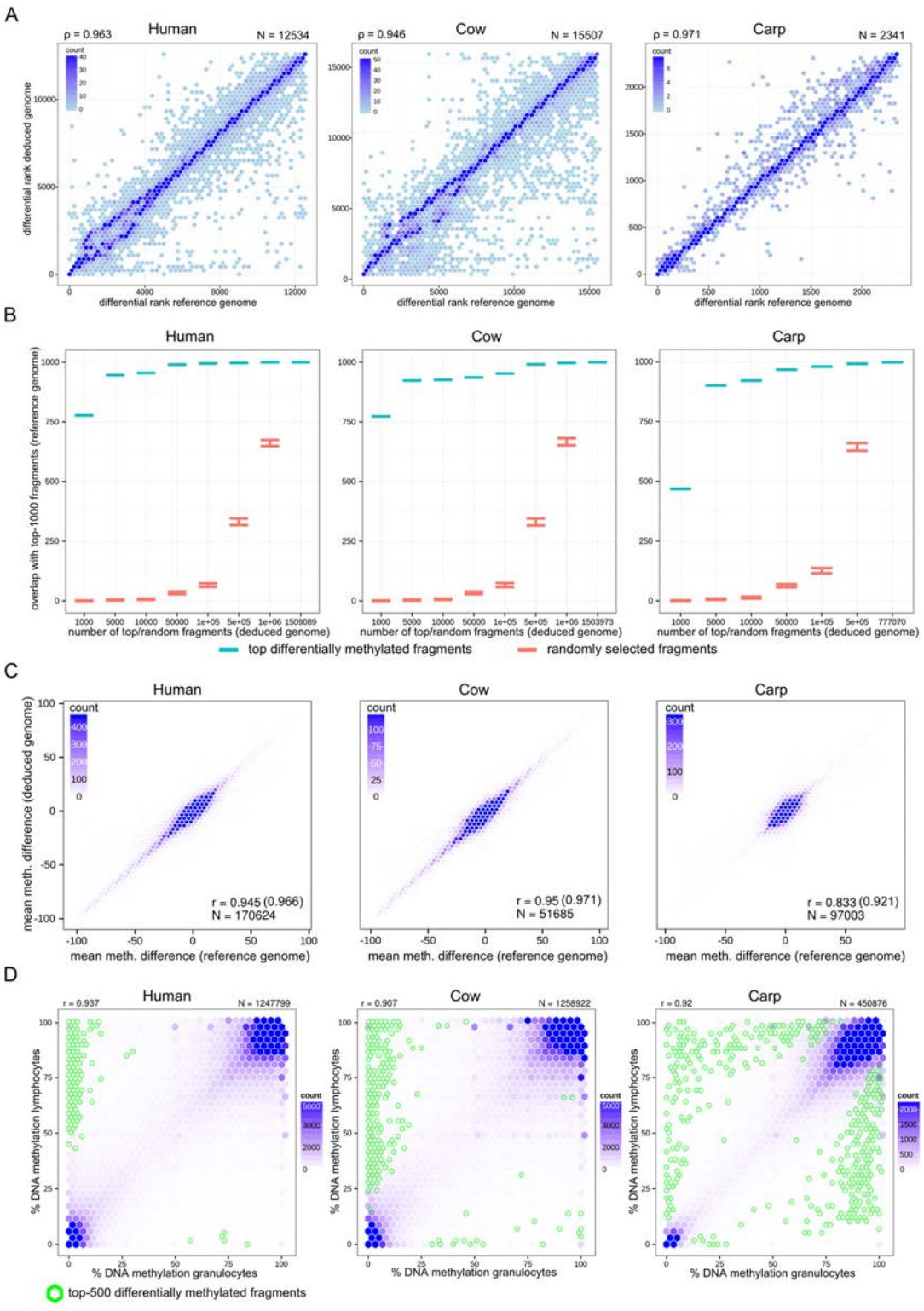


Figure S4. Validation of reference-free analysis of differential DNA methylation, Related to Figure 4

(**A**) Scatterplots displaying the agreement between differential methylation ranks for differentially methylated fragments (p-value < 0.05) using the two approaches (ρ: Spearman correlation coefficient; N: number of deduced genome fragments). (**B**) Recovery of the top-1000 differentially methylated deduced genome fragments (p-value < 0.05, coverage ≥ 8, non-overlapping) determined by the reference-based approach in a gradually increasing number of top differentially methylated deduced genome fragments using the reference-free approach (blue). The recovery within an equal number of randomly selected deduced genome fragments is shown for comparison (red). (**C**) Scatterplots showing the difference in mean fragment methylation between granulocytes and lymphocytes as determined by the reference-based (x-axis) vs. the reference-free (y-axis) approach for fragments that overlap with each other when mapped to the reference genome. Pearson correlations (r) for non-overlapping fragments are indicated in brackets. This plot shows that differential DNA methylation values are not strongly affected by overlapping fragments (Case 2 in Figure S2). All fragments were coverage-filtered for at least eight and not more than 200 mapped reads. (**D**) DNA methylation scatterplots demonstrating differential DNA methylation in granulocytes (x-axis) vs. lymphocytes (y-axis) using the reference-based approach. Means across four biological replicates are shown for each cell type, and the green hexagons indicate the top-500 most differentially methylated fragments. Matched scatterplots for the reference-free analysis are shown in Figure 4C.

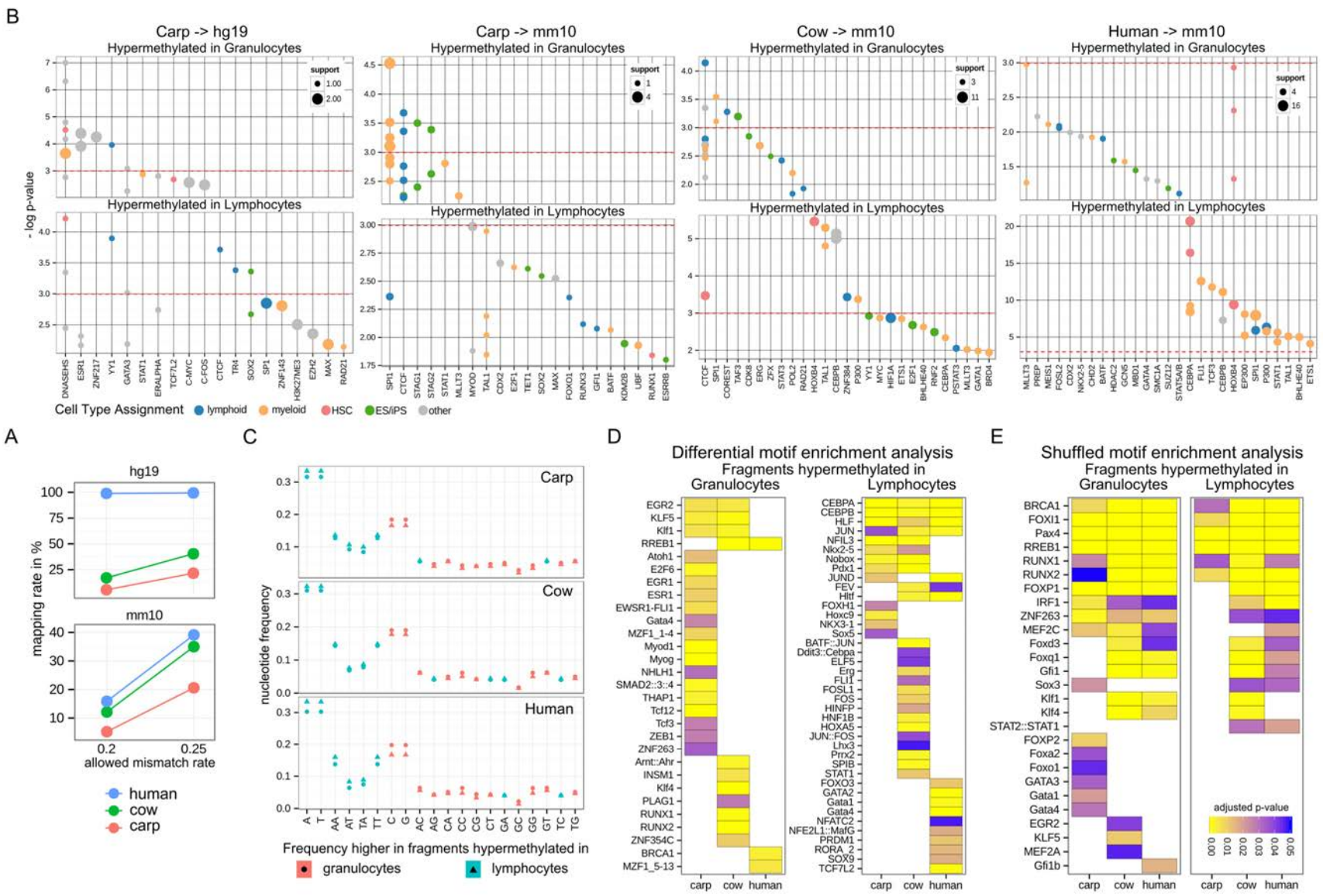


Figure S5: Interpretation of DNA methylation differences through cross-mapping to annotated genomes and motif enrichment analysis, Related to Figure 5

(**A**) Mapping of the deduced genome fragments of human, cow, and carp to the reference genomes of human (hg19) and mouse (mm10). Mapping rates are displayed for maximum mismatch rates of 20% and 25%. (**B**) Region enrichment analysis for reference-free deduced genome fragments that have been cross-mapped to the reference genomes of human (hg19) and mouse (mm10). For each group, the top-20 enrichments obtained by LOLA analysis are shown. Uncorrected p-values are plotted on the y-axis, and the number of overlapping regions is indicated by bubble size. Each dot represents an experiment listed in the database, and the red dashed lines indicate p-values of 0.05. Similar plots for human and cow cross-mapping to the human genome (hg19) are shown in Figure 5A. (**C**) Nucleotide frequency differences between the top-500 deduced genome fragments in granulocytes (dots) and lymphocytes (triangles). (**D**) Complete list of enriched sequence motifs from JASPAR CORE (2014) Vertebrates database among the top-500 deduced genome fragments with increased DNA methylation in granulocytes vs. lymphocytes (right) and vice versa (left). The motif analysis used the opposing group as background. (**E**) Same as in panel D, but using randomly shuffled sequences with the same mono- and dinucleotide composition as background. The displayed motifs were identified as significantly enriched in at least 95% of iterations.

Table S1. Summary statistics for the reference-free and reference-based analysis of DNA methylation in the blood dataset, Related to Figure 2

Table showing for each of the analyzed samples and biological replicates the number of total reads, mapped reads, and informative reads (i.e., those that give rise to at least one valid DNA methylation measurement), mean DNA methylation levels of methylated and unmethylated spike-in controls, mean DNA methylation levels across CpG sites, non-CpG conversion rates, as well as the number of CpG measurements, number of covered CpGs, and mean informative sequencing coverage per CpG site.

This table is provided as a separate Excel file.

Table S2. Summary statistics for direct cross-mapping of carp RRBS reads to the human, mouse, and zebrafish genome with various choices of alignment parameters, Related to Figure 5

Table listing for each of the carp samples the number of mapped reads, the percentage of mapped reads, and the number of CpGs covered using four different mapping approaches with different BSMAP parameters: Maximum mismatch rate of 0.08 with multi-mapping reads; maximum mismatch rate of 0.08 without multi-mapping reads; maximum mismatch rate of 0.2 with multi-mapping reads; and maximum mismatch rate of 0.2 without multi-mapping reads.

This table is provided as a separate Excel file

3.2 GBMatch

The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space.

Klughammer J*, Kiesel B*, Roetzer T, Fortelny N, Kuchler A, Datlinger P, Peter N, Nenning K, Furtner J, Nowosielski M, Augustin M, Mischkulnig M, Ströbel T, Moser P, Freyschlag CF, Kerschbaumer J, Thomé C, Grams AE, Stockhammer G, Kitzwoegerer M, Oberndorfer S, Marhold F, Weis S, Trenkler J, Buchroithner J, Pichler J, Haybaeck J, Krassnig S, Ali KM, von Campe G, Payer F, Sherif C, Preiser J, Hauser T, Winkler PA, Kleindienst W, Würtz F, Brandner-Kokalj T, Stultschnig M, Schweiger S, Dieckmann K, Preusser M, Langs G, Baumann B, Knosp E, Widhalm G, Marosi C, Hainfellner JA, Woehrer A, Bock C.

Submitted.

The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space

Johanna Klughammer[1]*, Barbara Kiesel[2,3]*, Thomas Roetzer[3,4], Nikolaus Fortelny[1], Amelie Kuchler[1], Paul Datlinger[1], Nadine Peter[3,4], Karl-Heinz Nenning[5], Julia Furtner[3,6], Martha Nowosielski[7,8], Marco Augustin[9], Mario Mischkulnig[2,3], Thomas Ströbel[3,4], Patrizia Moser[10], Christian F. Freyschlag[11], Johannes Kerschbaumer[11], Claudius Thomé[11], Astrid E. Grams[12], Günther Stockhammer[7], Melitta Kitzwoegerer[13], Stefan Oberndorfer[14], Franz Marhold[15], Serge Weis[16], Johannes Trenkler[17], Johanna Buchroithner[18], Josef Pichler[19], Johannes Haybaeck[20,21], Stefanie Krassnig[20], Kariem Madhy Ali[22], Gord von Campe[22], Franz Payer[23], Camillo Sherif[24], Julius Preiser[25], Thomas Hauser[26], Peter A. Winkler[26], Waltraud Kleindienst[27], Franz Würtz[28], Tanisa Brandner-Kokalj[28], Martin Stultschnig[29], Stefan Schweiger[30], Karin Dieckmann[3,31], Matthias Preusser[3,32], Georg Langs[5], Bernhard Baumann[9], Engelbert Knosp[2,3], Georg Widhalm[2,3], Christine Marosi[3,32], Johannes A. Hainfellner[3,4], Adelheid Woehrer[3,4]#§, Christoph Bock[1,33,34]#

1 CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria.
2 Department of Neurosurgery, Medical University of Vienna, Vienna, Austria.
3 Comprehensive Cancer Center, Central Nervous System Tumor Unit, Medical University of Vienna, Austria.
4 Institute of Neurology, Medical University of Vienna, Vienna, Austria.
5 Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab, Medical University of Vienna, Vienna, Austria.
6 Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria.
7 Department of Neurology, Medical University of Innsbruck, Innsbruck, Austria.
8 University Medical Center, Neurology, and Neurooncology, German Cancer Research Center (DKFZ) and DKTK, Heidelberg, Germany
9 Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria.
10 Department of Pathology, Medical University of Innsbruck, Innsbruck, Austria.
11 Department of Neurosurgery, Medical University of Innsbruck, Innsbruck, Austria.
12 Department of Neuroradiology, Medical University of Innsbruck, Innsbruck, Austria.
13 Department of Pathology, University Hospital of St. Poelten, Karl Landsteiner University of Health Sciences, St. Poelten, Austria.
14 Department of Neurology, University Hospital of St. Poelten, Karl Landsteiner University of Health Sciences, St. Poelten, Austria.
15 Department of Neurosurgery, University Hospital of St. Poelten, Karl Landsteiner University of Health Sciences, St. Poelten, Austria.
16 Department of Neuropathology, Neuromed Campus Wagner-Jauregg, Kepler University Hospital, Johannes Kepler University of Linz, Linz, Austria.
17 Department of Radiology, Neuromed Campus, Kepler University Hospital, Johannes Kepler University of Linz, Linz, Austria.
18 Department of Neurosurgery, Neuromed Campus Wagner-Jauregg, Kepler University Hospital, Johannes Kepler University of Linz, Linz, Austria.
19 Department of Internal Medicine, Neuromed Campus Wagner-Jauregg, Kepler University Hospital, Johannes Kepler University of Linz, Linz, Austria.
20 Department of Neuropathology, Institute of Pathology, Medical University of Graz, Graz, Austria.
21 Department of Pathology, Otto-von-Guericke University of Magdeburg, Magdeburg, Germany.
22 Department of Neurosurgery, Medical University of Graz, Graz, Austria.
23 Department of Neurology, Medical University of Graz, Graz, Austria.
24 Department of Neurosurgery, Krankenanstalt Rudolfstiftung, Vienna, Austria.
25 Department of Pathology, Krankenanstalt Rudolfstiftung, Vienna, Austria.
26 Department of Neurosurgery, Christian-Doppler-Klinik, Paracelsus Private Medical University, Salzburg, Austria.
27 Department of Neurology, Christian-Doppler-Klinik, Paracelsus Private Medical University, Salzburg, Austria.
28 Institute of Pathology, State Hospital Klagenfurt, Klagenfurt, Austria.
29 Department of Neurology, State Hospital Klagenfurt, Klagenfurt, Austria.
30 Department of Neurosurgery, General Hospital Wiener Neustadt, Wiener Neustadt, Austria.
31 Department of Radiotherapy, Medical University of Vienna, Vienna, Austria.
32 Department of Medicine I, Medical University of Vienna, Vienna, Austria.
33 Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria.
34 Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany.

* These authors contributed equally to this work
# These authors jointly directed this work
§ To whom correspondence should be addressed: adelheid.woehrer@meduniwien.ac.at (A.W.)

Keywords: Glioblastoma, epigenetic heterogeneity, DNA methylation, disease progression, tumor microenvironment, transcriptional subtypes, bioinformatics, integrative data analysis, medical epigenomics

Abstract

Glioblastoma is characterized by widespread genetic and transcriptional heterogeneity, yet little is known about the role of the epigenome in glioblastoma disease progression. Here, we present genome-scale maps of the DNA methylation dynamics in matched primary and recurring glioblastoma tumors, based on a national population registry and a comprehensively annotated clinical cohort. We demonstrate the feasibility of DNA methylation mapping in a large set of routinely collected formalin-fixed paraffin-embedded (FFPE) samples, and we validate bisulfite sequencing as a multi-purpose assay that allowed us to infer a range of different genetic, epigenetic, and transcriptional tumor characteristics. Based on these data, we identified characteristic differences between primary and recurring tumors, links between DNA methylation and the tumor microenvironment, and an association of epigenetic tumor heterogeneity with patient survival. In summary, this study provides a Resource for dissecting DNA methylation heterogeneity in genetically diverse and heterogeneous tumors, and it demonstrates the feasibility of integrating epigenomics, radiology, and digital pathology in a representative national cohort, leveraging samples and data collected as part of routine clinical practice.

Introduction

Glioblastoma is a devastating cancer with a median age at diagnosis of 64 years[1]. Even under the best available care, the median survival is little more than one year, and very few patients live for more than three years[2,3]. Despite intense efforts, limited therapeutic progress has been made over the last decade, and a series of phase III clinical trials with targeted agents have failed to improve overall survival[4-6].

Glioblastoma shows extensive temporal and spatial heterogeneity, which appears to contribute to therapeutic resistance and inevitable relapse[7-13]. Prior research on tumor heterogeneity in glioblastoma has focused mainly on the genomic and transcriptomic dimensions[7-20], while the dynamic role of the epigenome in glioblastoma disease progression is much less understood[21].

Recent data in other cancers have conclusively shown the power of DNA methylation sequencing for analyzing epigenetic heterogeneity. For example, DNA methylation heterogeneity has been linked to clonal progression in prostate cancer[22], low-grade glioma[23], esophageal squamous cell carcinoma[24], and hepatocellular carcinoma[25]; and new measures of DNA methylation heterogeneity such as epi-allele burden, proportion of discordantly methylated reads (PDR), and DNA methylation inferred regulatory activity (MIRA) have been linked to clinical variables in acute myeloid leukemia[26], chronic lymphatic leukemia[27], and Ewing sarcoma[28].

To investigate the contribution of epigenetics to the temporal and spatial heterogeneity of glioblastoma, we performed DNA methylation sequencing on a large cohort of IDH wildtype glioblastoma patients (n = 112) with matched samples from primary and recurring tumors (between 2 and 4 time points per patient), and we also included multiple subregion samples for a subset of these tumors. Importantly, by using an optimized reduced representation bisulfite sequencing (RRBS) protocol[28,29] we obtained a high success rate for archival formalin-fixed and paraffin-embedded (FFPE) samples, coverage of 3-5 fold more CpGs compared to Infinium microarrays[30-32], and single-CpG as well as single-allele resolution (in RRBS, each sequencing read captures the DNA methylation status of one or more individual CpGs in one single allele from one single cell).

The presented dataset – comprising 349 RRBS-based DNA methylation profiles, of which 320 were derived from FFPE samples – constitutes the largest cohort of FFPE samples that has yet undergone genome-scale DNA methylation sequencing, and it conclusively shows the technical feasibility of performing large multicenter DNA methylation studies based on routinely collected FFPE material. The RRBS data not only identified epigenetic disease subtypes and quantified epigenetic heterogeneity, but also allowed us to infer transcriptional subtypes, copy number aberration, single nucleotide variants (SNVs), small insertions and deletions

(indels), *MGMT* promoter methylation, and G-CIMP/ IDH mutational status. This study thus highlights the power of DNA methylation sequencing in routinely collected clinical FFPE tumor samples.

Linking DNA methylation profiles to magnetic resonance (MR) imaging, tumor morphology, tumor microenvironment, and clinical variables such as patient survival, we obtained a detailed picture of temporal and spatial heterogeneity in glioblastoma. We observed characteristic differences in DNA methylation between primary and recurring tumors, an association with the tumor microenvironment, and a link between tumor microenvironment and previously reported clinically relevant MR imaging-based progression types[33]. DNA methylation was highly predictive of the established glioblastoma transcriptional subtypes, and disease progression-associated loss of DNA methylation in the promoters of Wnt signaling genes was associated with worse prognosis. We also observed across several comparisons that the association with survival was stronger for properties of the recurring tumor than for properties of the primary tumor. In summary, our study provides a comprehensive resource of the DNA methylation dynamics and heterogeneity in glioblastoma, proof-of-concept for DNA methylation sequencing in large FFPE sample sets collected as part of routine diagnostics, and an integrative analysis of DNA methylation with various types of clinical and histopathological information.

Results

DNA methylation sequencing in a cohort of matched primary and recurring glioblastoma samples

To investigate the DNA methylation dynamics associated with disease progression in glioblastoma, we established a richly annotated dataset of patients that underwent tumor resection at diagnosis and at least once upon tumor recurrence (Figure 1a). These patients were identified through the population-based Austrian Brain Tumor Registry[34], and 112 primary glioblastoma patients (wildtype *IDH* status) with tumor samples for at least two time points (primary tumor and first recurrence) were included in the analysis (Supplementary Fig. 1a). Due to the requirement of having undergone at least two tumor resections, the selected patients were on average younger (median age at diagnosis of 58 years) and had longer overall survival (median overall survival of 22.4 months) compared to the unselected population-based cohort (median age at diagnosis 63 years, median overall survival of 8 months) (Figure 1b and Supplementary Table 1).

For each of these tumor samples, we established genome-scale DNA methylation profiles using reduced representation bisulfite sequencing (RRBS). The RRBS assay provides single-CpG and single-allele resolution with excellent quantitative accuracy even on challenging clinical samples[35,36], and it has been successfully used to dissect intra-tumor heterogeneity in several cancer types[26-28]. The RRBS-based DNA methylation data were complemented by time-matched MR imaging data as well as quantitative pathology data capturing the morphology, proliferative activity, and tumor microenvironment of the same tumors. All data are available from the Supplementary Website (http://glioblastoma-progression.computational-epigenetics.org/). We systematically integrated these datasets using statistical analysis and machine learning methods (Figure 1a).

DNA methylation profiling based on routinely collected FFPE material was successful for all samples that had adequate tumor cell content, and 95% of the resulting RRBS profiles yielded more than 500,000 covered CpGs (Supplementary Table 2). The median number of covered CpGs in FFPE samples (1,880,675) was lower than for fresh-frozen samples (4,473,349), but higher than for an alternative ethanol-based fixation method[37] (1,005,828) that was tested by one sample-providing center (Supplementary Fig. 1b and Supplementary Table 2). The measured bisulfite conversion rates were highly consistent with expectations: 99% of genomic cytosines outside of CpGs were read as thymines, the mean underconversion rate on unmethylated spike-in controls was 1%, and the mean overconversion rate on methylated spike-in controls was 2% (Supplementary Fig. 1c

and Supplementary Table 2). All DNA methylation profiles showed the expected distribution of DNA methylation levels across CpG islands, promoters, and genome-wide tiling regions (Supplementary Fig. 1d), with a slight tendency toward lower DNA methylation levels in low-quality samples (Supplementary Fig. 1e).

Comparing DNA methylation levels of 5-kilobase tiling regions between primary and recurring tumors, there was a high correlation ($r > 0.94$) across the genome (Supplementary Fig. 1f). Nevertheless, we observed widespread epigenetic heterogeneity at individual loci (Figure 1c). An example is the promoter of the *MGMT* gene, whose DNA methylation status has been shown to correlate with sensitivity to alkylating chemotherapy[38]. The *MGMT* promoter was unmethylated in the majority of samples (Supplementary Fig. 1g), and patients with a methylated *MGMT* promoter in their recurring tumors had significantly better progression-free survival (PFS) and overall survival (OS) compared to patients with unmethylated *MGMT* promoters (Supplementary Fig. 1h).

To compare our *IDH*-wildtype primary glioblastoma cohort to *IDH*-mutant brain tumor samples, we also performed RRBS on primary and recurring tumors of 13 *IDH*-mutant oligodendroglioma/astrocytoma/glioblastoma patients from the same population. The DNA methylation profiles of these tumors showed a characteristic CpG island methylator phenotype (CIMP) as expected from previous observations[39] (Figure 1d), which provides further validation of the accuracy and robustness of our DNA methylation profiling on FFPE material.

Inference of genomic information from the RRBS data

All tumor samples were obtained via a national brain tumor registry and reflect routine clinical care in Austria, which currently does not include whole genome or whole exome sequencing. We therefore evaluated whether certain types of genomic information could be inferred directly from the RRBS data.

First, we reconstructed genome-wide maps of copy number aberrations (CNAs) from the RRBS data using the CopywriteR algorithm[40]. We detected various CNAs previously described in glioblastoma, including amplifications of the *EGFR* locus and deletions of chromosome 10 (Supplementary Fig 2a). In particular, we observed that 10q deletions in the recurring tumors (which affect the *MGMT* gene and have been shown to correlate with sensitivity to alkylating chemotherapy[41]) were associated with longer survival. Based on the CNA data, we also verified that none of our primary glioblastoma samples harbored the 1p19q co-deletion, thereby excluding the presence of any misclassified cases of anaplastic oligodendroglioma (Supplementary Fig. 2c).

Second, we inferred single nucleotide variants (SNVs) and small insertions/deletions (indels) from the RRBS data using the Bis-SNP algorithm[42]. Although confident SNV and indel detection in RRBS data is limited to a relatively small subset of the genome, it allowed us to confirm that none of our samples displayed a hypermutator phenotype and that there was no strong trend toward higher mutational rates in primary or recurring tumors (Supplementary Fig. 2d). Furthermore, among the variants with high predicted impact on protein expression we found multiple genes with known relevance in glioblastoma (Supplementary Fig 2e).

Prediction of transcriptional subtypes in glioblastoma based on DNA methylation

Recent research defined three transcriptional subtypes of glioblastoma (classical, mesenchymal, and proneural)[43], while a previously included fourth (neural) subtype has been described as an artifact of contaminating non-tumor tissue. Because RNA sequencing of FFPE material is challenging and often infeasible, we tested whether these transcriptional subtypes can be inferred from DNA methylation data (Figure 2a). Machine learning classifiers using L2-regularized logistic regression were trained and evaluated on matched DNA methylation and transcriptional subtype data[44,45] from The Cancer Genome Atlas (http://cancergenome.nih.gov/). Based on these data, we obtained good prediction accuracies with cross-validated receiver operating characteristic (ROC) area under curve (AUC) values above 0.8 for most samples (Supplementary Fig 3a).

Applying the trained classifiers to our DNA methylation dataset, we assigned class probabilities to each tumor sample. While these percentage values primarily reflect the confidence with which each sample is assigned to each of the transcriptional subtypes, we can also interpret them as an estimate of the relative contribution of each of the three subtypes to individual tumor samples, thus providing an initial assessment of intra-tumor heterogeneity. We found that all three transcriptional subtypes were common in our cohort of *IDH* wildtype primary glioblastoma (Figure 2b) – in contrast to the *IDH* mutated tumors, which were as almost always assigned to the proneural subtype (Supplementary Fig 3b).

Predicted transcriptional subtypes were heterogeneous both in space and in time. Most individual tumor samples showed signatures of more than one transcriptional subtype (Figure 2b), which is consistent with recent single-cell RNA-seq data that identified similar heterogeneity within individual samples[11]. Moreover, five out of six patients with multi-sector samples displayed at least two transcriptional subtypes (Figure 2c and Supplementary Fig. 3c-e), and about half of the patients showed different predominant transcriptional subtypes between the primary and recurring tumors (Figure 2d). Predicted transcriptional subtypes in the recurring tumor (but not in the primary tumor) were also associated with patient survival (Figure 2e), the mesenchymal subtype being associated with the worst prognosis and the classical subtype with the best prognosis. Finally, patients whose tumors switched to the mesenchymal subtype displayed the worst PFS and OS (Figure 2e).

To investigate the epigenetic differences between the three transcriptional subtypes, we compared the DNA methylation profiles of those tumors that were most confidently assigned to one specific subtype (class probability > 0.8) with each other (Figure 2f). Region set analysis of differentially methylated CpGs using LOLA[46] identified a moderate enrichment of chromatin protein binding sites for regions that were specifically hypomethylated in the mesenchymal subtype, including EZH2, KDM4A, RBBP5, and SUZ12 (Figure 2g).

We also calculated 'DNA methylation inferred regulatory activity' (MIRA) scores[28] for each individual tumor sample, where high scores indicate strong local depletion of DNA methylation at specific transcription factor binding sites (Figure 2h). We observed significantly higher MIRA scores (corresponding to deeper DNA methylation dips and increased regulatory activity) for CTCF, EZH2, and KDM4A in the mesenchymal subtype (Figure 2i). In contrast, MIRA scores for key regulators of pluripotency (NANOG, SOX2, POU5F1) were reduced in the mesenchymal subtype (Figure 2i). To corroborate this observation, we plotted MIRA scores for EZH2 and NANOG against class probabilities for the different transcriptional subtypes across all samples (not only those with high class probability), and we indeed observed highly significant correlations with mesenchymal class probabilities of 0.49 (EZH2) and -0.48 (NANOG) (Supplementary Fig. 3f).

Linking DNA methylation differences to changes in the tumor microenvironment

To test whether the DNA methylation data captures relevant aspects of the tumor microenvironment, we quantified the abundance of various types of immune cells in the primary and recurring tumors. Specifically, we performed single-plex stainings for markers that identified different immune cell types (CD3, CD8, CD80, CD68) including anti-inflammatory (CD163, FOXP3) and memory T-cell (CD45ro) subpopulations. We observed significant differences in the immune cell count between the three transcriptional subtypes (Figure 3a-b), with the highest number of immune cells found in tumors of the mesenchymal subtype. The increased level of immune cell infiltration was accompanied by larger necrotic areas, fewer vital tumor areas, lower tumor cell proliferation, and lower cell density in tumor tissues (Supplementary Fig. 4a-b). Moreover, high levels of CD68 positive cells (macrophages of all types) were associated with poor prognosis in recurring tumors; and high levels of CD163 positive cells (anti-inflammatory, tumor-propagating M2 macrophages) were associated with poor prognosis in both primary and recurring tumors (Figure 3c).

Comparing immune cell infiltration between primary and recurring tumors of the same patients, we observed significantly increased levels of inflammatory infiltrates upon recurrence, while the infiltration levels of anti-inflammatory macrophages and memory T-cells did not change significantly (Figure 3d-f). The increase in inflammation in the recurring tumors was accompanied by a decrease in necrotic volume and contrast-enhancing (active) tumor mass, but an increase in edema according to matched diagnostic MR imaging data (Supplementary Fig 4c-d). Patients with less necrotic or contrast-enhancing (active) tumor mass upon recurrence presented with a more favorable clinical outcome (Supplementary Fig. 4e).

When we stratified patients according to prognostically relevant MR imaging-based progression types (i.e., classic T1, cT1 relapse / flare-up, and T2 diffuse)[33] (Supplementary Fig. 5a), we found that primary and recurring tumors from patients displaying the "cT1 relapse / flare-up" subtype had lower infiltration of pro-inflammatory immune cell types (CD3, CD8, CD68) and a lower fraction of proliferating cells (Figure 3g). In concordance with previous work[33], cT1 relapse / flare-up patients displayed a slightly more favorable prognosis compared to the other two progression types (Supplementary Fig. 5b).

Several of these characteristics of the tumor microenvironment could be predicted from the DNA methylation data using machine learning methods (Supplementary Fig. 5c-d). Differentiating between tumors with a high and low level of immune cell infiltration, we observed a high cross-validated prediction performance for CD163 (ROC AUC = 0.88), CD68 (0.79), CD45ro (0.93), CD3 (0.87) and CD8 (0.79) (Figure 3h; Supplementary Fig. 5e). For those immune cell types with high cross-validation accuracy, DNA methylation levels at the most predictive genomic tiling regions accurately grouped the samples by their immune cell infiltration levels in a hierarchical clustering analysis (Supplementary Fig 5f). As it was recently shown for RNA expression profiles[20,47], our data support the feasibility of inferring immune cell infiltration from DNA methylation data.

Linking DNA methylation differences to tumor cell-intrinsic properties

To investigate the relationship between DNA methylation and tumor cell characteristics such as proliferation and nuclear morphology, we performed detailed histopathological analysis for the majority of tumor samples.

Cell proliferation was measured by staining for the cell proliferation marker Ki67 (MIB1) followed by quantification of the relative abundance of MIB1 positive cells. The percentage of proliferating cells showed no consistent changes between primary and recurring tumors (Figure 4a). Nevertheless, high proliferation in the recurring tumors (but not in the primary tumors) was significantly associated with increased PFS (Figure 4b). DNA methylation patterns discriminated with high accuracy between tumors characterized by high versus low proliferation rates (ROC AUC = 0.89) (Figure 4c). This was not due to differences in mean DNA methylation levels (Figure 4d-e). Rather, highly proliferating tumors showed intermediate DNA methylation levels at discriminatory regions, while low proliferating tumors showed more extreme methylation patterns (Figure 4e).

Nuclear morphology of tumor cells was measured by the size and eccentricity (a measure of elongated shape) of the tumor cell nuclei, and by the variability of the two parameters. None of these measures were significantly different between primary and recurring tumors. However, tumors that shifted to a sarcoma-like phenotype (i.e., secondary gliosarcoma) upon recurrence had a significant increase in nuclear eccentricity and a decrease in its variability (Figure 4f-g), accompanied by an increase in CD8 immune cell infiltration, proliferative rates (MIB+ cells), and relative tumor mass in the recurring tumor (Figure 4h). DNA methylation patterns predicted nuclear eccentricity (ROC AUC = 0.83) and its variability (0.76) (Figure 4i), and patients with shape-shifting tumors (i.e., classic to sarcoma) displayed significantly shorter PFS and a trend towards reduced OS (Figure 4j), which might be explained by increased tissue infiltration of the spindle-shaped cells.

DNA methylation heterogeneity and dynamics between primary and recurring tumors

To quantify epigenetic tumor heterogeneity in glioblastoma progression, we used two complementary approaches (Figure 5a). Sub-clonal heterogeneity was measured by epi-allele entropy[48], and stochastic DNA methylation erosion was measured by the proportion of discordant reads (PDR)[27]. Both measures identified extensive heterogeneity between patients (Supplementary Fig. 6a-b) but no strong trend between primary and recurring tumors (Figure 5b).

Comparing the epi-allele composition of the 20% most heterogeneous and the 20% least heterogeneous samples (Figure 5b), we found extensive variability between patients and over time (Figure 5c). Samples with high epi-allele entropy and high diversity in their epi-allele composition also showed high PDR values, indicating that these samples were characterized by broadly increased epigenetic heterogeneity.

The observed differences in epigenetic heterogeneity were not a side effect of different tumor sizes, as there was little to no correlation between the two measures of tumor heterogeneity on the one hand and the tumor size as measured by MR imaging on the other hand (Figure 5d). In contrast, we did observe a significant association between PDR values and clinical outcome specifically in the primary tumors (Figure 5e and Supplementary Fig. 6c), and we also found that a longer time span between first and second surgery was weakly associated with fewer differentially methylated regions (Figure 5f) but not with the extent of epi-allelic shifting (Supplementary Fig. 6d).

To further dissect the temporal dimension of DNA methylation heterogeneity in glioblastoma, we performed differential DNA methylation analysis on all matched pairs of primary and first recurring tumors. Focusing on gene promoters, we observed high correlation of DNA methylation levels ($r = 0.86$) and a small number of promoters with strong differential methylation in multiple patients (Figure 6a). Most of these promoters showed consistent trends toward either gain or loss of DNA methylation upon tumor recurrence, although for one gene with known involvement in brain cancer (*OTX2*) we observed a progression-associated gain of DNA methylation in some patients and a loss in others (Figure 6b upper panel). When we classified the patients into those that followed the cohort-level trend in differential DNA methylation (trend patients) and those that did not (anti-trend patients) (Figure 6b, lower panel and Figure 6c), the trend patients showed worse prognosis (Figure 6e), suggesting that some of the observed differences may contribute to disease progression.

Pathway analysis identified an enrichment of genes involved in development and apoptosis among those genes whose promoters gained DNA methylation during disease progression; in contrast, genes whose promoters lost DNA methylation were enriched in the Wnt signaling pathway and T cell activation (Figure 6d). Corroborating the latter finding, when we classified all patients according to whether they on average gained or lost DNA methylation in the promoters of Wnt signaling genes, we observed a significant association between loss of DNA methylation and reduced PFS and OS (Figure 6f).

Discussion

Focusing on glioblastoma as one of the genetically most complex cancers, we sought to determine the prevalence and character of epigenetic tumor heterogeneity in time and space. To that end, we established a comprehensive set of DNA methylation profiles covering primary and recurring tumors from the same patients as well as multi-sector samples in a subset of patients. A longitudinal cohort of 112 patients with *IDH*-wildtype primary glioblastoma that had undergone at least two (and up to four) tumor resections was assembled based on the Austrian Brain Tumor Registry, thus providing a population-scale representation of glioblastoma patients. An optimized RRBS protocol allowed us to work with minute amounts of FFPE tissue, while providing single-CpG and single-allele resolution and insights into epigenetic heterogeneity at single-cell level that would be difficult or impossible to obtain using microarray-based methods for DNA methylation profiling.

Based on the RRBS dataset, we were able to infer a broad range of tumor properties – including glioblastoma transcriptional subtypes, aspects of the tumor microenvironment, and tumor cell-intrinsic attributes such as cell proliferation. Our analysis revealed changes in tumor microenvironment between primary and recurring tumors and identified the composition of the tumor microenvironment to be a major discriminatory factor between transcriptional subtypes as well as MR progression types, suggesting potential clinical applications of DNA methylation based prediction of the tumor microenvironment. In line with recent work[13,20] we observed co-occurrence of multiple transcriptional subtypes within the same tumor and frequent switching of the dominant subtype over time. Moreover, patients whose tumor switched to the mesenchymal subtype had reduced survival. Assessing the regulatory basis of the transcriptional subtypes, we identified epigenomic signatures of increased EZH2 activity in glioblastoma of the mesenchymal subtype. In light of the significantly worse prognosis of mesenchymal tumors, these results might enable new subtype-specific therapeutic approaches that exploit the observed regulatory differences, such as the use of emerging EZH2 inhibitors[49-51].

We also observed characteristic trends in DNA methylation between primary and recurring tumors, including a demethylation of Wnt signaling gene promoters that was associated with worse prognosis. Aberrant activation of the Wnt signaling pathway is observed in various cancers including glioblastoma, where hypermethylation mediated suppression of Wnt signaling inhibitors was identified as a source of aberrant activation of this pathway[52]. We quantified epigenetic tumor heterogeneity in two complementary ways, and DNA methylation erosion as measured by the PDR score was associated with survival. Patients whose primary tumors harbored higher levels of DNA methylation erosion showed longer PFS and a tendency towards longer OS. These results were surprising given that previous studies in hematopoietic malignancies (AML[26] and CLL[27]) had associated increased epigenomic heterogeneity with worse prognosis. This discrepancy might be explained by the fact that chemotherapy in leukemia is applied to the entire population of malignant cells, while the bulk of glioblastoma is surgically removed prior to chemotherapy and radiotherapy. The non-selective bottleneck of tumor resection might turn the evolutionary advantage of an heterogeneous population of tumor cells into a disadvantage, especially in the case of stochastic and therefore mostly detrimental DNA methylation erosion.

Finally, we observed that several properties of the recurring tumors were specifically associated with survival, while there was no strong association in the primary tumors. This was true for transcriptional subtypes (Figure 2e), MR-imaging derived necrotic and enhancing tumor volumes (Supplementary Fig. 4e), CD68+ macrophage infiltration (Figure 3c), and tumor cell proliferation levels (Figure 4b). These results emphasize the potential clinical relevance of repeated biopsy and detailed diagnostic work-up of recurring tumors in order to promote more personalized treatment decisions upon glioblastoma recurrence.

In summary, our study establishes a rich resource describing the DNA methylation dynamics of glioblastoma progression in a highly annotated clinical cohort with matched MR imaging and detailed histopathological analyses that included the tumor microenvironment. Importantly, all data are openly available through public repositories and a detailed Supplementary Website (http://glioblastoma-progression.computational-epigenetics.org/). This study also highlights the feasibility and potential of working with national patient registries and large patient cohorts, with FFPE samples, and with clinical data that have been collected as part of routine clinical care. Finally, in combination with research that established the accuracy and robustness of DNA methylation assays for clinical diagnostics[54], our data support that DNA methylation sequencing can make a relevant contribution to the clinical assessment of tumor heterogeneity, providing potential biomarkers for improved diagnosis, prognosis, and personalized therapy in glioblastoma and other heterogeneous cancers.

Methods

Sample acquisition via a population-based registry

All glioblastoma cases were selected from the Austrian Brain Tumor Registry[34], including only patients over 18 years of age with a first surgery at diagnosis and at least one additional surgery upon recurrence. Tumor samples and clinical data were provided by the following partner institutions: Medical University of Vienna, Kepler University Hospital Linz, Paracelsus Medical University Salzburg, Medical University of Innsbruck, Karl-Landsteiner University Hospital St. Pölten, State Hospital Klagenfurt, State Hospital Wiener Neustadt, Hospital Rudolfstiftung Vienna, and the Medical University of Graz. The resulting cohort comprised 159 patients with matched FFPE samples for the primary tumor and at least one recurring tumor, which were deposited in the neurobiobank of the Medical University of Vienna (ethics vote EK078-2004). After screening for sufficient tumor content in both the primary and the recurring tumor, 47 patients were excluded. A total of 112 patients were retained, each with at least two and up to four time points (283 tumor samples in total, including 6 patients with multi-sector sampling). The diagnosis of primary glioblastoma, *IDH*-wildtype was confirmed by central pathology review according to the 2016 update of the WHO classification[55] including targeted assessment of the IDH R132H mutational status. In addition, 13 patients (32 tumor samples) with IDH-mutant oligodendroglioma/astrocytoma/glioblastoma, and 5 patients (5 samples) who underwent temporal lobe surgery due to epilepsy (Medical University of Vienna) were included as controls. Informed consent was obtained according to the Declaration of Helsinki, and the study was approved and overseen by the ethics committee of the Medical University of Vienna (ethics votes EK550/2005, EK1412/2014, EK 27-147/2015).

DNA isolation from FFPE tumor samples

Areas of highest tumor cell content were selected based on hematoxylin-eosin stained sections, while any samples with tumor cell content below 50% in the region-of-interest were excluded from further analysis. Genomic DNA was extracted from FFPE tissues using the QIAamp DNA FFPE Tissue Kit following manufacturer's instructions.

DNA methylation profiling by RRBS

RRBS was performed as described previously[29] using 100 ng of genomic DNA for most samples, while occasionally going down to 2 ng (if not more DNA was available) and up to 200 ng (Supplementary Table 2). To assess bisulfite conversion efficiency independent of CpG context, methylated and unmethylated spike-in controls were added in a concentration of 0.1%. DNA was digested using the restriction enzymes MspI and TaqI in combination (as opposed to only MspI in the original protocol) in order to increase genome-wide coverage. Restriction enzyme digestion was followed by fragment end repair, A-tailing, and adapter ligation. The amount of effective library was determined by qPCR, and samples were multiplexed in pools of 10 with similar qPCR C_t values. The pools were then subjected to bisulfite conversion followed by library enrichment by PCR. Enrichment cycles were determined using qPCR and ranged from 12 to 21 (median: 16). After confirming adequate fragment size distributions on Bioanalyzer High Sensitivity DNA chips (Agilent), libraries were sequenced on Illumina HiSeq 3000/4000 machines in a 50 or 60 basepair single-read setup.

DNA methylation data processing

RRBS data were processed using a custom pipeline based on Pypiper (http://databio.org/pypiper) and Looper (http://databio.org/looper). Adapter sequences were trimmed, and 60 basepair reads were cropped to 50 base-

pairs using Trimmomatic[56] with the following settings: ILLUMINACLIP:RRBS_adapters.fa:2:40:7 SLIDINGWINDOW:4:15 MAXINFO:20:0.50 CROP:50 MINLEN:18. Trimmed reads were then aligned to the human genome build hg38 using BSMAP in RRBS mode[57], and DNA methylation calling was performed with a custom python script (biseqMethCalling.py) published previously[29]. To assess bisulfite conversion efficiency, unmapped reads were aligned to the spike-in reference sequences using Bismark[58], and DNA methylation calls for methylated and unmethylated controls were extracted from the alignment file. CpGs in repetitive regions according to the UCSC RepeatMasker track were excluded from further analysis. DNA methylation data were analyzed at the level of single CpGs or in a binned format with mean DNA methylation values calculated across 5-kilobase regions, CpG islands (as defined in the UCSC Genome Browser) or GENCODE promoter regions (1 kilobase upstream to 500 bases downstream of the transcription start site).

Identification of copy number aberrations from RRBS data

Copy number aberrations for each sample were identified using the R/Bioconductor package CopywriteR[40] based on the BSMAP-aligned BAM files and a bin size of 100,000. Data from five normal brain controls were merged at the level of aligned bam files to serve as the shared control for all analyses. Each individual sample was then normalized either against the merged control or against the cohort median, whichever showed the less extreme (i.e., more conservative) value for a given bin. Genomic segments identified by CopywriteR were classified as significantly amplified or deleted if their normalized absolute copy number value deviated more than one cohort standard deviation (mean standard deviation across all bins in a given segment) from 0. Significantly amplified or deleted segments for each sample were then plotted in an overview graph sorted by segment length.

Identification of single nucleotide variants from RRBS data

Single nucleotide variants and small insertions and deletions were identified using Bis-SNP[42] based on the BSMAP-aligned BAM files, the human reference genome build hg38, and dbSNP build 147. Identified variants were annotated using SnpEff v4.2.

Annotation of glioblastoma associated genes

Glioblastoma associated genes were taken from a recent publication[59] and annotated for their cancer-linked function (oncogene, tumor suppressor gene, drug resistance gene) based on a published classification of cancer genes[60]. Genes not contained in this classification were manually annotated according to their known or suspected molecular functions as described in GeneCards (http://www.genecards.org).

Patient survival analysis

Survival analysis was performed using the functions survfit() and survdiff() of the R package 'survival'. For continuous variables, patients with the 50% highest values were compared to those with the 50% lowest values (unless indicated otherwise). Survival curves were plotted with ggsurvplot() from the R package 'survminer'.

Inference of transcriptional subtypes from RRBS data

Glioblastoma transcriptional subtypes[43] were predicted from DNA methylation data at the level of single CpGs using L2-regularized logistic regression as implemented in the R package 'LiblinearR'. Classifiers were trained

and evaluated on Infinium 27k DNA methylation data for glioblastoma tumors[44] obtained from the TCGA data portal (https://portal.gdc.cancer.gov/). TCGA data were restricted to IDH wildtype, non-G-CIMP samples. Furthermore, neural subtype samples were excluded because this subtype of glioblastoma had previously been associated with tumor margin and contamination with non-tumor brain tissue[61]. For each sample in our cohort, a classifier was trained and evaluated on the TCGA data, using those CpGs that were covered also in the sample of which the transcriptional subtype was to be predicted (1,249 CpGs on average). After performance evaluation by 10-fold cross-validation and calculation of the cross-validated receiver operating characteristic (ROC) area under curve (AUC) values, a final classifier was built using all selected TCGA samples. This classifier was used to predict the transcriptional subtype including class probabilities of the respective sample.

Groupwise differential DNA methylation analysis

Differentially methylated CpGs between predefined groups of tumor samples were identified with a custom R script that uses a two-sided Wilcoxon rank-sum test. Groups containing less than five samples were excluded from the analysis, and only CpGs covered by at least five reads per sample in at least 30% of samples were included. CpGs in repetitive regions ("RepeatMasker", "Simple Repeats", and "WM + SDust" tracks from the UCSC Genome Browser, downloaded 6 September 2016) were also excluded. For the retained CpGs, differential DNA methylation between groups of samples was assessed using the Wilcoxon rank-sum test (wilcox.test() in R), and p-values were adjusted for multiple testing using the Benjamini-Hochberg method (p.adjust() in R). CpGs with multiple-testing adjusted p-values smaller than 0.05 and with a median difference of beta values larger than 0.1 were considered significant.

Region set enrichment analysis using LOLA

Enrichment of genomic region sets among the differentially methylated regions was assessed using the LOLA software[46]. To reduce potential biases from co-located CpGs, CpGs were merged into 1-kilobase tiling regions across the genome prior to LOLA analysis. In LOLA, the hypermethylated or hypomethylated regions were used as the query set, and the set of all differentially methylated tiling regions were used as the universe. Only regions from astrocytes or embryonic stem cells in the LOLA Core database were included in the analysis for better interpretability. P-values were corrected for multiple testing using the Benjamini and Yekutieli method (p.adjust() in R), and all enrichments with an adjusted p-value below 0.05 were considered significant. In a control experiment, to assess potential effects of imbalance between hypermethylated and hypomethylated region sets, the analysis was repeated using the top-N highest ranking regions from both sets.

DNA methylation inferred regulatory activity (MIRA)

MIRA scores for selected sets of transcription factor binding sites from the LOLA Core database[46] were calculated as in the original publication[28]. Briefly, aggregated DNA methylation profiles around the binding sites (2.5 kilobases upstream and downstream, split into 21 bins) were created for each sample and transcription factor. MIRA scores were calculated as the log ratio between aggregated DNA methylation values for the center bin (bin 0, reflecting the binding site) and the average of two flanking bins (bins -5 and +5).

Immunohistochemistry

The following antibodies were used for immunohistochemistry: IDH1 (1:60 Dianova #DIA-H09), CD3 (1:200 Thermo Scientific #RM-9107-S1), CD8 (1:100 Dako Cytomation #M7103), CD45Ro (1:500 Dako Cytomation #M0742), CD8 (1:100 Dako Cytomation #C8/144B), FoxP3 (1:25 BioLegend #320116), CD163 (1:1000 Novocastra #NCL-L-CD163), CD68 (1:5000 Dako Cytomation #M0814), HLA-DR (1:400 Dako Cytomation #M0775), MIB1 (1:200 Dako Cytomation #M7240), CD34 (1:100 Novocastra #NCL-l-END).

FFPE blocks were cut at a thickness of 3 µm, and sections were stained on a Dako autostainer system using the following primary antibodies MIB1, HLA-DR, CD34, CD45Ro, CD68, CD3, CD8, CD163. Antigens were retrieved by heating the sections in 10 mM sodium citrate (pH 6.0) at 95°C for 20 min, followed by incubation with primary antibodies for 30 min at room temperature. The Dako Flex+mouse detection system was used according to manufacturer´s recommendations. For primary antibodies against IDH1 and FoxP3, a Ventana BenchMark automated staining system was used, followed by visualization using the Ultra View detection kit. All sections were counterstained with hematoxylin. For each antibody used positive and negative controls were used per 30-slide-batch. Negative controls were performed by omitting the primary antibody and by using Universal Negative Control rabbit (Dako) for polyclonal rabbit antibodies or purified mouse myeloma IgG1 (Zymed Laboratories, San Francisco, CA) for monoclonal mouse antibodies. The slides were scanned using a Hamamatsu NanoZoomer 2.0 HT slide scanner, and images were analyzed using the NDPview 2 software. Whole slide scans were downsampled to 5x magnification and exported as JPEG images. Fiji was used for further image processing[62]. First, the inbuilt Color Deconvolution method was used to separate the hematoxylin from the DAB stain. The 8-bit greyscale hematoxylin image was thresholded using Phansalkar thresholding, and nuclei were counted using the built-in Analyze Particles algorithm to determine reference cell counts for each image. For each antibody stain Phansalkar thresholding parameters were manually optimized, followed by automated counting of DAB-positive cells. The inferred counts for antigen-expressing cells and corresponding total nuclei counts were saved as spreadsheet for further statistical analysis.

Histopathological analysis of whole slide scans

H&E stained slides were scanned using a Hamamatsu NanoZoomer 2.0 HT slide scanner. The Hamamatsu NDP.view2 software was used to annotate relevant regions in the slide scans. Areas including hemorrhages, necrosis, scars, squeezed tissue, and preexisting brain parenchyma were manually segmented by a specialist in neuropathology using the Freehand Region tool. The slide scans were downsampled to 10x magnification and exported as jpeg images using the NDPITools plugin of Fiji[62-64] along with xml files featuring the annotations (*.ndpa). The JPEG images were loaded into Fiji for further processing.

To obtain a cell nuclei mask for each slide scan image, the Color Deconvolution method of Fiji was used to obtain an 8-bit greyscale image of the hematoxylin stain[65]. Automated local thresholding based on Phansalkar's method segmented cell nuclei (parameters $k = 0.2$, $r = 0.5$, radius $= 8$)[66], followed by the binary Close and Open operations. The Watershed method was used to separate clustered nuclei[67]. To obtain a mask of the gross tissue on the slide scan, the original image was first converted to an 8-bit greyscale image followed by global thresholding (thresholding values 0-207). The masks were loaded into MATLAB R2014b (MathWorks) for further automated image analysis. For each image, the annotation data was fetched from the respective xml file and converted into polygons. These polygons were subsequently grouped to form binary masks based on their annotation (e.g., a necrosis mask comprising all annotated necrotic areas). The gross tissue was determined by first loading the previously obtained gross tissue mask into MATLAB, performing the binary Close operation (dilating and eroding the binary image by 30 pixels) and then removing all connected components

smaller than 1,000 pixels. Pixels that were neither included in the background nor in any of the image annotation masks were assigned to the tumor. Using the pixel-to-area conversion, the areas covered by tumor, necrosis and other annotated tissues were calculated.

Subsequently, the pre-obtained binary nucleus mask was analyzed in blocks of 160 x 160 pixels (~146 µm x 146 µm). This block size was empirically found to provide a good compromise between spatial resolution and nuclear content required for statistical analysis. The centroids of the nuclei were localized and the area and eccentricity (which is calculated as the distance between the foci of a fitted ellipse divided by its major axis length, with values close to 0 corresponding to circular nuclei and values close to 1 corresponding to spindle-shaped nuclei) of each localized nucleus was assessed. Next, the number of nuclei per block was calculated from the centroid map. From the nucleus areas in each block, the average area and eccentricity as well as the standard deviation of the nucleus areas and eccentricities were determined. The coefficient of variation was calculated by dividing the standard deviation by the average for each block. For nuclei number, average area and eccentricity, standard deviation and coefficient of variation, the data from all blocks were assembled in a matrix and saved as grey scale bitmap (.bmp) image as well as color portable network graphics (.png) image. Furthermore, using the binary annotation masks, each block was assigned to its corresponding region, given that >90% of the pixels in that block were uniformly annotated. Subsequently, for each type of region, the mean, standard deviation, coefficient of variation, median, and mode of the aforementioned nucleus characteristics were calculated (e.g., obtaining the mean average nucleus density in tumor tissue). For further statistical processing, the numerical data extracted from the slide scan images were saved into a spreadsheet.

Radiological evaluation of glioblastoma patients

MR images of glioblastomas at time of first diagnosis and recurrence in sufficient quality were available for 54 of the glioblastoma patients included in this study, which were contributed by 6 different radiology departments. Both T1-weighted images with contrast enhancement (CE) and fluid-attenuated inversion recovery (FLAIR)/T2-weighted axial images were reviewed for topographic tumor location to assess solitary versus multicentric tumors and local versus distant recurrences. Multicentric glioblastomas were defined as at least two spatially distinct lesions that are not contiguous with each other and whose surrounding abnormal FLAIR/T2 signals do not overlap[68]. Tumor segmentation was performed with BraTumIA[69,70], which uses multi-modal MRI sequences for fully automated volumetric tumor segmentation. T1, T1 contrast enhanced, T2, and FLAIR sequences were used to segment four tumor tissue types: necrotic, cystic, edema/non-enhancing, and enhancing tumor. Due to differences in MRI protocols across the study sites, the multi-modal sequences were affine registered to the T1 sequence with SPM122 and resampled to 1x1x3mm voxel size prior to segmentation. The BraTumIA-derived segmentations were reviewed by an expert radiologist, and errors in the automatic segmentation were manually corrected.

Evaluation of MR imaging-based progression phenotypes

MR imaging-based tumor progression was assessed according to the Response Assessment in Neuro-Oncology (RANO) standard[71]. Serial T1-weighted images with CE and FLAIR/T2-weighted images were available for 43 patients. Progression subtypes were classified as described previously[33,72]: (i) Classic T1 (incomplete disappearance of T1-CE during therapy followed by T1-CE increase at progression), (ii) cT1 relapse / flare-up (complete disappearance of T1-CE during therapy followed by T1-CE reoccurrence at progression), (iii) Primary non-responder (increase and/or additional T1-CE lesions at first MR imaging follow-up after start of

therapy), (iv) T2-circumscribed (bulky and inhomogeneous T2/FLAIR progression, no or single faintly speckled T1-CE lesions at progression), and (v) T2-diffuse (complete decrease in T1-CE during therapy but exclusive homogeneous T2/FLAIR signal increase with mass effect at progression).

DNA methylation based prediction of tumor properties

Tumor properties such as the immune cell infiltration and tumor cell morphology were predicted from DNA methylation data using a machine learning approach that was based on the R package 'LiblinearR'. The DNA methylation data were prepared by calculating for each sample the mean DNA methylation levels in 5-kilobase tiling regions across the genome. Tiling regions covered in less than 90% of the samples were excluded from the analysis, and the filtered data matrix (samples x tiling regions) was subjected to imputation using the function impute.knn() from the R-package 'impute', with the parameter k (i.e., the number of nearest neighbors considered) set to 5. Tumor properties represented by continuous response variables were converted into categorical variables by setting the 20% highest values to 'high', the 20% lowest values to 'low', and the remaining samples to 'NA'. Imputed beta values were used to train and evaluate the classifiers using LiblineaR(). In the confirmatory hierarchical clustering based on the most predictive features identified by the classifiers, the beta values were scaled across samples for better visualization and comparability. LiblineaR() was set to use support vector classification by Crammer and Singer as model type, and the appropriate cost parameter was estimated from the imputed data matrix using the function heuristicC() from the same package. For each tumor property, the performance of the classifiers was determined through leave-one-out cross-validation, and 10 control runs with randomly shuffled labels were included to detect potential overfitting. ROC curves and ROC AUC values were determined using the functions prediction() and performance() of the R-package 'ROCR'. Finally, we trained a classifier on the entire dataset using the selected model and cost parameter, which was then used for further analysis including the extraction of the most predictive features, hierarchical clustering, and the prediction of additional samples (for the transcriptional subtypes).

Estimation of DNA methylation heterogeneity

The epi-allele entropy as a measure of sub-clonality within a tumor was calculated using a slightly modified version of methclone[48]. We calculated epi-allele entropies separately for each of the samples and, independently, for each matched pair of primary and recurring tumors. Input files to methclone were created by aligning the trimmed RRBS reads to the human reference genome build hg38 using Bismark[58]. As in the original publication, methclone was set to require a minimum of 60 reads in order to consider a locus, and loci with a combinatorial entropy change below -80 were classified as epigenetic shift loci (eloci) between primary and recurring tumors[48]. For each pair, we then calculated epi-allele shifts per million loci (EPM), dividing the number of eloci by the total number of assed loci normalized to 1 million loci[48].

The proportion of discordant reads (PDR) as a measure of local erosion and DNA methylation disorder was calculated as described in the original publication[27]. Briefly, the number of concordantly or discordantly methylated reads with at least four valid CpG measurements was determined for each CpG using a custom python script. The PDR at each CpG was then calculated as the ratio of discordant reads compared to all valid reads covering that locus. CpGs at the end of a read were disregarded to remove potential biases due to the end-repair step of RRBS library preparation. Because the PDR and epi-allele entropy calculation is highly sensitive to differences in the read composition of the underlying RRBS library, we focused this analysis on RRBS libraries with a similar number of enrichment cycles (13-15) to ensure high consistency between samples (Supplementary Fig. 6a).

Sample-wise PDR and epi-allele entropy values were calculated by averaging across promoters that were covered in more than 75% of the samples. Promoter regions were defined as the genomic region 1 kilobase upstream to 500 basepairs downstream of a given transcription start site as annotated by GENCODE[73].

Pairwise differential DNA methylation analysis

Differentially methylated CpGs between sample pairs (i.e., primary tumor versus matched recurring tumor) were identified with a custom R script that uses Fisher's exact test. This test was applied to the methylated and unmethylated read counts derived from the BSMAP-aligned reads by the biseqMethCalling.py script. P-values were adjusted for multiple testing using the Benjamini-Hochberg method. To obtain promoter-wise differential DNA methylation calls, p-values were combined using a generalization of Fischer's method[74] as implemented in RnBeads[75]. Promoter methylation levels for each sample were calculated as the mean of all CpGs in the promoter region. Promoter regions were defined as the genomic region 1 kilobase upstream to 500 basepairs downstream of a given transcription start site as annotated by GENCODE[73].

Pathway enrichment analysis

Enrichment analysis for gene sets and pathways was performed using enrichR[76,77] through an R interface (https://github.com/definitelysean/enrichR) querying the Panther_2016 database (http://www.pantherdb.org/) for enrichments with an adjusted p-value below 0.05.

Data and code availability

All data are available through the Supplementary Website (http://glioblastoma-progression.computational-epigenetics.org/). Genome browser tracks facilitate the locus-specific inspection of the DNA methylation data, and a d3.js based graphical data explorer enables interactive analysis of associations in the annotated dataset (Supplementary Fig. 7). The Supplementary Website also hosts the raw and segmented image data from the histopathological analysis as well as the raw and segmented MR imaging data. The processed DNA methylation data will also be available for download from NCBI GEO (accession number: GSE100351, reviewer link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100351, login token: ufahcwkajrqnhef), and the raw sequencing data will be available from EBI EGA. Finally, in the spirit of reproducible research[78] the Supplementary Website makes the source code underlying the presented analyses publicly available.

Acknowledgements

We thank all patients who have donated their samples for this study. We also thank Gloria Wilk, Martina Muck, Susanne Schmid, and Ulrike Andel for technical assistance with immunohistochemical stainings, macrodissection and tumor tissue shavings, Simon Mages for contributing to the interactive data visualization, the Biomedical Sequencing Facility at CeMM for assistance with next generation sequencing, and all members of the Bock lab for their help and advice.

The study was funded in part by an Austrian Science Fund grant (FWF KLI394) to AW, a Marie Curie Career Integration Grant (European Union's Seventh Framework Programme grant agreement no. PCIG12-GA-2012-333595) to CB, and an ERA-NET project (EpiMark FWF I 1575-B19). C.B. is supported by a New Frontiers Group award of the Austrian Academy of Sciences and by an ERC Starting Grant (European Union's Horizon 2020 research and innovation programme, grant agreement no. 679146). ABTR activities are further supported by unrestricted research grants of Roche Austria to JAH and the Austrian Society of Neurology to SO.

Author contributions

JKl, AW, and CB designed the study. BK, TR, NP, KHN, JF, MN, MA, MM, TS, GL, BB, JAH, and AW established and annotated the clinical cohort. AK and PD performed DNA methylation profiling. JKl performed the data analysis with contributions from NF. PM, CFF, JKe, AEG, GS, MK, SO, FM, SW, JT, JB, JPi, JH, SK, KMA, GvC, FP, CS, JPr, PAW, WK, FW, TBK, MS, SS, KD, MP, EK, GW, and CM contributed tumor samples and clinical data. JKl, AW, and CB wrote the manuscript with contributions from all authors.

References

1. Ostrom, Q.T.*, et al.* CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2009-2013. *Neuro-Oncology* **18**, v1-v75 (2016).
2. Ferlay, J.S., I.; Ervik, M.; Forman, D.; Bray, F. GLOBOCAN 2012 v1.0 Cancer Incidence and Mortality Worldwide. IARC CancerBase No. 11. (IARC press, Lyon, 2017).
3. Woehrer, A., Bauchet, L. & Barnholtz-Sloan, J.S. Glioblastoma survival: has it improved? Evidence from population-based studies. *Current Opinion in Neurology* **27**, 666-674 (2014).
4. Chinot, O.L.*, et al.* Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *The New England Journal of Medicine* **370**, 709-722 (2014).
5. Gilbert, M.R.*, et al.* A randomized trial of bevacizumab for newly diagnosed glioblastoma. *The New England Journal of Medicine* **370**, 699-708 (2014).
6. Stupp, R.*, et al.* Cilengitide combined with standard treatment for patients with newly diagnosed glioblastoma with methylated MGMT promoter (CENTRIC EORTC 26071-22072 study): a multicentre, randomised, open-label, phase 3 trial. *The Lancet Oncology* **15**, 1100-1108 (2014).
7. Doucette, T.*, et al.* Immune heterogeneity of glioblastoma subtypes: extrapolation from the cancer genome atlas. *Cancer Immunology Research* **1**, 112-122 (2013).
8. Kim, H.*, et al.* Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Research* **25**, 316-327 (2015).
9. Kim, J.*, et al.* Spatiotemporal Evolution of the Primary Glioblastoma Genome. *Cancer Cell* **28**, 318-328 (2015).
10. Lee, J.K.*, et al.* Spatiotemporal genomic architecture informs precision oncology in glioblastoma. *Nature Genetics* **49**, 594-599 (2017).
11. Patel, A.P.*, et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
12. Sottoriva, A.*, et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 4009-4014 (2013).
13. Wang, J., Cazzato, E., Ladewig, E., Frattini, V. & Rosenbloom, D.I. Clonal evolution of glioblastoma under therapy. *Nature Genetics* **48**, 768-776 (2016).
14. Aubry, M.*, et al.* From the core to beyond the margin: a genomic picture of glioblastoma intratumor heterogeneity. *Oncotarget* **6**, 12094-12109 (2015).
15. Hu, W., Wang, T., Yang, Y. & Zheng, S. Tumor heterogeneity uncovered by dynamic expression of long noncoding RNA at single-cell resolution. *Cancer Genetics* **208**, 581-586 (2015).
16. Kumar, A.*, et al.* Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. *Genome Biology* **15**, 530 (2014).

17. Meyer, M.*, et al.* Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 851-856 (2015).

18. Snuderl, M.*, et al.* Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell* **20**, 810-817 (2011).

19. Sturm, D.*, et al.* Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425-437 (2012).

20. Wang, Q.*, et al.* Tumor evolution of glioma intrinsic gene expression subtype associates with immunological changes in the microenvironment. *bioRxiv* (2016).

21. Alentorn, A.*, et al.* Differential gene methylation in paired glioblastomas suggests a role of immune response pathways in tumor progression. *Journal of Neuro-Oncology* **124**, 385-392 (2015).

22. Brocks, D.*, et al.* Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Reports* **8**, 798-806 (2014).

23. Mazor, T.*, et al.* DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. *Cancer Cell* **28**, 307-317 (2015).

24. Hao, J.J.*, et al.* Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nature Genetics* **48**, 1500-1507 (2016).

25. Lin, D.C.*, et al.* Genomic and epigenomic heterogeneity of hepatocellular carcinoma. *Cancer Research* **77**, 2255-2265 (2017).

26. Li, S.*, et al.* Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nature Medicine* **22**, 792-799 (2016).

27. Landau, D.A.*, et al.* Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813-825 (2014).

28. Sheffield, N.C.*, et al.* DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. *Nature Medicine* **23**, 386-395 (2017).

29. Klughammer, J.*, et al.* Differential DNA Methylation Analysis without a Reference Genome. *Cell Reports* **13**, 2621-2633 (2015).

30. Bibikova, M.*, et al.* Genome-wide DNA methylation profiling using Infinium(R) assay. *Epigenomics* **1**, 177-200 (2009).

31. Dedeurwaerder, S.*, et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**, 771-784 (2011).

32. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8**, 389-399 (2016).

33. Nowosielski, M.*, et al.* Progression types after antiangiogenic therapy are related to outcome in recurrent glioblastoma. *Neurology* **82**, 1684-1692 (2014).

34. Woehrer, A.*, et al.* The Austrian Brain Tumour Registry: a cooperative way to establish a population-based brain tumour registry. *Journal of Neuro-Oncology* **95**, 401-411 (2009).

35. Bock, C.*, et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotechnology* **28**, 1106-1114 (2010).

36. Gu, H.*, et al.* Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature Methods* **7**, 133-136 (2010).

37. Stefanits, H.*, et al.* KINFix--A formalin-free non-commercial fixative optimized for histological, immunohistochemical and molecular analyses of neurosurgical tissue specimens. *Clinical Neuropathology* **35**, 3-12 (2016).

38. Weller, M.*, et al.* MGMT promoter methylation in malignant gliomas: ready for personalized medicine? *Nature Reviews Neurology* **6**, 39-51 (2010).

39. Turcan, S.*, et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479-483 (2012).

40. Kuilman, T.*, et al.* CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biology* **16**, 49 (2015).

41. Wemmert, S.*, et al.* Patients with high-grade gliomas harboring deletions of chromosomes 9p and 10q benefit from temozolomide treatment. *Neoplasia* **7**, 883-893 (2005).

42. Liu, Y., Siegmund, K.D., Laird, P.W. & Berman, B.P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology* **13**, R61 (2012).

43. Bowman, R.L., Wang, Q., Carro, A., Verhaak, R.G. & Squatrito, M. GlioVis data portal for visualization and analysis of brain tumor expression datasets. *Neuro-Oncology* **19**, 139-141 (2017).

44. Brennan, C.W.*, et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-477 (2013).

45. Verhaak, R.G.*, et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110 (2010).

46. Sheffield, N.C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587-589 (2016).

47. Gentles, A.J.*, et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine* **21**, 938-945 (2015).

48. Li, S.*, et al.* Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biology* **15**, 472 (2014).

49. Kim, K.H. & Roberts, C.W. Targeting EZH2 in cancer. *Nature Medicine* **22**, 128-134 (2016).

50. Mohammad, F.*, et al.* EZH2 is a potential therapeutic target for H3K27M-mutant pediatric gliomas. *Nature Medicine* **23**, 483-492 (2017).

51. Suva, M.L.*, et al.* EZH2 is essential for glioblastoma cancer stem cell maintenance. *Cancer Research* **69**, 9211-9218 (2009).

52. Lee, Y., Lee, J.K., Ahn, S.H., Lee, J. & Nam, D.H. WNT signaling in glioblastoma and therapeutic opportunities. *Laboratory Investigation* **96**, 137-150 (2016).

53. Bady, P., Delorenzi, M. & Hegi, M.E. Sensitivity Analysis of the MGMT-STP27 Model and Impact of Genetic and Epigenetic Context to Predict the MGMT Methylation Status in Gliomas and Other Tumors. *The Journal of Molecular Diagnostics* **18**, 350-361 (2016).

54. Bock, C., *et al.* Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nature Biotechnology* **34**, 726-737 (2016).

55. Louis, D.N., Ohgaki, H., Wiestler, O.D. & Cavanee, W.K. *WHO Classification of Tumours of the Central Nervous System, 4th Edition Revised*, IARC press (2016).

56. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

57. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).

58. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).

59. Sahm, F.*, et al.* Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets. *Acta Neuropathologica* **131**, 903-910 (2016).

60. Vogelstein, B.*, et al.* Cancer genome landscapes. *Science* **339**, 1546-1558 (2013).

61. Gill, B.J., *et al.* MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12550-12555 (2014).

62. Schindelin, J., *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**, 676-682 (2012).

63. Deroulers, C., *et al.* Analyzing huge pathology images with open source software. *Diagnostic Pathology* **8**, 92 (2013).

64. Schindelin, J., Rueden, C.T., Hiner, M.C. & Eliceiri, K.W. The ImageJ ecosystem: An open platform for biomedical image analysis. *Molecular Reproduction and Development* **82**, 518-529 (2015).

65. Ruifrok, A.C. & Johnston, D.A. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology* **23**, 291-299 (2001).

66. Phansalkar, N., More, S., Sabale, A. & Joshi, M. Adaptive local thresholding for detection of nuclei in diversity stained cytology images. in *Communications and Signal Processing (ICCSP), 2011 International Conference on* 218-220 (IEEE, 2011).

67. Vincent, L. & Soille, P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE transactions on pattern analysis and machine intelligence* **13**, 583-598 (1991).

68. Liu, Q., *et al.* Genetic, epigenetic, and molecular landscapes of multifocal and multicentric glioblastoma. *Acta Neuropathologica* **130**, 587-597 (2015).

69. Meier, R., *et al.* Clinical Evaluation of a Fully-automatic Segmentation Method for Longitudinal Brain Tumor Volumetry. *Scientific Reports* **6**, 23376 (2016).

70. Porz, N., *et al.* Multi-modal glioblastoma segmentation: man versus machine. *PloS One* **9**, e96873 (2014).

71. Wen, P.Y., *et al.* Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *Journal of Clinical Oncology* **28**, 1963-1972 (2010).

72. Nowosielski, M., *et al.* Radiologic progression types are treatment specific: An exploratory analysis of a phase 3 study of bevacizumab plus radiotherapy plus temozolomide for patients with newly diagnosed glioblastoma (AVAglio). *Journal of Clinical Oncology* **34**, 2048-2048 (2016).

73. Harrow, J., *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760-1774 (2012).

74. Makambi, K. Weighted inverse chi-square method for correlated significance tests. *Journal of Applied Statistics* **30**, 225-234 (2003).

75. Assenov, Y., *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nature Methods* **11**, 1138-1140 (2014).

76. Chen, E.Y., *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

77. Kuleshov, M.V., *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90-97 (2016).

78. Gentleman, R. & Temple Lang, D. Statistical analyses and reproducible research. Bioconductor project working group. Working paper 2. (2004). http://biostats.bepress.com/bioconductor/paper2

Figures

Figure 1. DNA methylation landscape of glioblastoma disease progression

A: Integrative analysis of longitudinal DNA methylation data (RRBS) with matched magnetic resonance (MR) imaging data (morphology, segmentation), clinical annotation data (e.g., treatment, progression, IDH mutation status), and histopathological data (segmentation, morphology, immunohistochemistry) using statistical methods and machine learning. TMZ: Temozolomide; RTX: Radiation therapy; PC: Palliative care.

B: Patient cohort overview summarizing the disease courses of 112 primary glioblastoma patients with IDH-wildtype status, ordered by time of first surgery.

C: DNA methylation profiles for primary and recurring tumors at three relevant gene loci (*BCL2L11*, *SFRP2*, and *MGMT*). Genes and ENCODE histone H3K27ac tracks were obtained from the UCSC Genome Browser.

D: DNA methylation levels at CpGs indicative of the CpG island methylator phenotype (CIMP), shown separately for IDH mutated control samples (which are CIMP-positive) and the IDH wildtype primary glioblastoma samples from the study cohort (which are CIMP-negative). The fold change of DNA methylation levels between IDH mutated and wildtype samples is indicated based on data from a previous study[39].

Figure 2. Glioblastoma transcriptional subtypes inferred from DNA methylation

A: Overview of the machine learning approach for classifying tumor samples by their transcriptional subtypes using DNA methylation data. Classifiers were trained on DNA methylation data (Infinium 27k assay) of TCGA glioblastoma samples with known transcriptional subtype, using only CpGs shared by RRBS. All classifiers were evaluated by tenfold cross-validation on the TCGA samples and then applied to the RRBS profiles, predicting class probabilities that indicate the relative contribution of each transcriptional subtype.

B: Transcriptional subtype heterogeneity within cohort samples, as indicated by class probabilities of the subtype classifier. Samples are grouped and ordered by their dominant subtype.

C: Distribution of class probabilities across different regions of the same tumor (indicated by Roman numbers) and across different surgeries (indicated by Arabic numbers) for two patients with multisector samples.

D: Riverplot depicting transitions in the predicted transcriptional subtype between primary and recurring tumors. The number of samples in each state is indicated. Only patients whose primary and recurring tumors were classified with high accuracy (ROC AUC > 0.8) were included in this analysis.

E: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified by predicted transcriptional subtypes (left) and switching from a non-mesenchymal to mesenchymal subtype during disease progression (right). Only tumor samples that were classified with high accuracy (ROC AUC > 0.8) were included in this analysis.

F: Heatmap displaying the DNA methylation levels of the most differential CpGs between the three transcriptional subtypes. Only tumor samples that were classified with high class probabilities (>0.8) were included in this analysis.

G: LOLA region-set enrichment analysis of differentially methylated CpGs between the different transcriptional subtypes (binned into 1-kilobase tiling regions). Adjusted p-values (Benjamini & Yekutieli method) are displayed for all significantly (adjusted p-value < 0.05) enriched region sets (binding sites, x-axis) measured in astrocytes or embryonic stem cells (ESCs).

H: Schematic depicting the calculation of 'DNA methylation inferred regulatory activity' (MIRA) scores. DNA methylation profiles are combined across centered genomic regions of interest (e.g., transcription factor binding sites) for each sample and each region set. The MIRA score is then calculated as the ratio of DNA methylation levels at the flank to DNA methylation levels at the center (binding site) of the combined DNA

methylation profile. High MIRA scores therefore reflect local demethylation at the binding site, which indicates high regulatory activity of the respective factor.

I: DNA methylation profiles (upper row) and corresponding MIRA scores (lower row) for three region sets enriched in CpGs that are hypomethylated in the mesenchymal subtype (CTCF binding in astrocytes, EZH2 binding in astrocytes, and KDMA binding in ESCs) as well as three region sets of key regulators of pluripotency measured in ESCs (POUF1, NANOG, SOX2). The significance of differences between the three transcriptional subtypes was assessed using a two-sided Wilcoxon rank sum test: * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, ns: not significant.

Figure 3. DNA methylation and the tumor microenvironment

A: Comparison of tumor-infiltrating immune cell levels between different transcriptional subtypes as measured by quantitative immunohistochemistry for the indicated marker proteins.

B. Immunohistochemical stainings for FOXP3 and CD45ro in selected samples assigned to each of the three transcriptional subtypes.

C: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified according to the level of CD163-positive and CD68-positive immune cell infiltration in their primary and recurring tumors.

D, E: Differences in the relative proportion of tumor-infiltrating pro-inflammatory (D) and anti-inflammatory or neutral (E) immune cells between tumor samples originating from the patients' first surgery (primary tumor), second surgery (recurring tumor), or third surgery.

F: Comparative immunohistochemical stainings between primary and recurring tumors for three selected markers (CD68, CD8, CD163).

G: Comparison of the levels of tumor-infiltrating immune cells (cells positive for CD3, CD8, or CD68) and proliferating cells (MIB-positive cells) between the different progression types based on magnetic resonance (MR) imaging: Classic T1 (claT1), cT1 relapse / flare-up (cT1), and T2 diffuse (T2). Primary (left panel) and recurring (right panel) tumors were analyzed separately.

H: ROC curves for the DNA methylation based prediction of immune cell infiltration levels, as determined by leave-one-out cross-validation. ROC curves and the ROC area under curve (AUC) are indicated for the actual prediction (blue) and for background predictions with randomly shuffled labels (grey).

All significance tests comparing groups of samples in this figure were performed using a two-sided Wilcoxon rank sum test: * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, ns: not significant.

Figure 4. DNA methylation and histopathological tumor characteristics

A. Comparison of the fraction of proliferating (MIB-positive) cells between first surgery (primary tumor), second surgery (recurring tumor), and third surgery.

B. Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified according to the fraction of proliferating (MIB-positive) cells in the primary and recurring tumor.

C: ROC curves for the DNA methylation based prediction of levels of proliferating (MIB positive) cells, as determined by leave-one-out cross-validation.

D: Hierarchical clustering based on column-scaled DNA methylation levels of the most predictive genomic regions from the classifier predicting the fraction of proliferating (MIB-positive) cells (5-kilobase tiling regions).

E: Distribution of DNA methylation levels of the most predictive genomic regions from the classifier predicting the fraction of proliferating (MIB-positive) cells, displayed separately for samples with high (red) or low (blue) fractions and for features with positive (top) or negative (bottom) weights assigned by the classifier.

F: Comparison of the average nuclear eccentricity (AVG) and its coefficient of variation (COV) between tumors that shift to a sarcoma-like phenotype during disease progression and those that retain a stable histological phenotype.

G: Hematoxylin and eosin stains of matched primary and recurring tumors, illustrating the morphological changes observed when tumors shift to a sarcoma-like phenotype during disease progression.

H: Comparison of additional tumor properties between tumors that shift to a sarcoma-like phenotype during disease progression and those that retain a stable histological phenotype.

I: ROC curves for the DNA methylation based prediction of average nuclear eccentricity and its coefficient of variation.

J: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified according to whether their tumors shift to a sarcoma-like phenotype during disease progression (green) or retain a stable histological phenotype (orange).

All significance tests comparing groups of samples in this figure were performed using a two-sided Wilcoxon rank sum test: * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, ns: not significant.

Figure 5. DNA methylation heterogeneity in glioblastoma disease progression

A. Illustration of epi-allele entropy (EPY) and proportion of discordant reads (PDR) as two complementary measures of epigenetic tumor heterogeneity. Individual loci can have high values for one but not for the other measure (locus 1 and 2), or the measures can agree with each other (locus 3). Loci that undergo extensive changes in their epi-allele composition (locus 2 and 3) have been termed eloci[48]. '=': no change in heterogeneity between primary and recurring tumor; '>' increased heterogeneity in the recurring tumor.

B. Comparison of mean sample-wise PDR and epi-allele entropy values between first surgery (primary tumor) and second surgery (recurring tumor). The samples with the 20% highest and lowest heterogeneity values are color-coded and form the basis for the analyses presented in panels C to E.

C: Relative epi-allele frequencies in promoter regions for each of the color-coded samples from panel B (bottom) as well as the distribution of high and low heterogeneity samples along the gradient defined by their relative epi-allele composition (top). For clearer visualization, the "0000" majority epi-allele with a frequency of 70% to 80% is not displayed. 0: unmethylated, 1: methylated

D: Correlation between intra-tumor heterogeneity and enhancing (active) tumor mass as determined by MR imaging. r: Pearson correlation.

E: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified according to their PDR and epi-allele entropy values.

F: Correlation between the number of differentially methylated promoters during progression per million assessed promoters (DPM) and the time between first and second surgery. r: Pearson correlation.

Figure 6. DNA methylation differences between primary and recurring tumors

A. Scatterplot depicting the relationship of promoter DNA methylation between primary and recurring tumors. Promoters that were differentially methylated between primary and recurring tumors in at least 5 patients are highlighted (DNA methylation difference greater than 75%, adjusted p-value below 0.001, and average RRBS read coverage greater than 20 reads). r: Pearson correlation.

B. Barplots (top) depicting the number of patients that show significant gain or loss of DNA methylation in the differentially methylated promoters highlighted in panel A; scatterplots and line plots (bottom) showing the change in DNA methylation associated with disease progression (measured as percentage points, pp) for patients following the cohort-trend (red) or not (blue). Trend lines were calculated using the *loess* method.

C. Definition of "trend" and "anti-trend" patients based on the Manhattan distance between the maximal trend at differentially methylated promoters (DNA methylation values of 0% or 100%) and the observed difference in DNA methylation for each patient. "Trend" patients are those whose DNA methylation profiles are similar to the maximal trend (low normalized Manhattan distance); "Anti-trend" patients are those whose methylation profiles are most different from the maximal trend (high normalized Manhattan distance).

D: Pathway enrichment analysis of those genes that recurrently lose DNA methylation during disease progression and those that recurrently gain DNA methylation during disease progression.

E: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified by whether they followed the cohort trend of differential promoter DNA methylation (trend patients) or not (anti-trend patients), according to the definition in panel C.

F: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified into the top-30% patients with increasing or decreasing average DNA methylation levels at the promoters of Wnt signaling genes during disease progression.

Supplementary Tables

Supplementary Table 1. Patient summary table

Supplementary Table 2. RRBS summary table

Supplementary Figures

Supplementary Figure 1. RRBS profiling in a population-based glioblastoma cohort

A: Overview of the clinical centers that contributed to this study. The numbers of IDH wildtype and IDH mutated patients are indicated for each center.

B: Scatterplots summarizing the RRBS sequencing data. The proportion of randomly fragmented reads (i.e., reads not starting with the expected RRBS restriction sites) reflect the degree of pre-fragmentation of the input DNA. The different sample types (FFPE: formalin-fixed paraffin-embedded; Frozen: fresh-frozen, RCL: ethanol-based conservation) are indicated by color.

C: DNA methylation levels of methylated and unmethylated synthetic spike-in control sequences. Dashed lines indicate DNA methylation levels of 5% and 95%.

D, E: Distribution of DNA methylation levels across different genomic regions (covered by more than 10 reads per CpG) and for the different sample types (D) and quality tiers (E) defined by the number of unique CpGs detected in each RRBS library (tier 1: more than 3 million; tier 2: between 2 and 3 million; tier 3: between 1 and 2 million; tier 4: below 1 million).

F: Scatterplots depicting the relationship of DNA methylation levels in 5-kilobase tiling regions (containing more than 25 CpGs and covered by more than 10 reads per CpG) between primary and recurring tumors for the three different sample types. r: Pearson correlation.

G: Mean *MGMT* promotor methylation levels averaged across two CpGs (cg12434587 and cg12981137)[53]. Error bars indicate the maximum and minimum detected methylation levels in each samples. The dashed line indicates the threshold (36%) below which samples are considered unmethylated.

H: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified by *MGMT* promoter methylation status as depicted in panel G.

Supplementary Figure 2. Inference of genetic information from RRBS data

A: Overview of the RRBS-based inference of copy number aberrations (CNAs) in the glioblastoma cohort. The horizontal red and green lines represent the aberrations identified in each of the samples. The genomic position of genes with reported relevance in glioblastoma are indicated.

B: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified by chromosome 10q deletion status as depicted in panel A.

C: Assessment of 1p/19q co-deletion status in the IDH wildtype primary glioblastoma samples as well as the 13 oligodendroglioma samples from 6 patients with known 1p/19q co-deletion as positive controls. The size of the bubbles represents the mean fraction of the respective chromosome arms that are affected by the indicated CNAs.

D: Scatterplot displaying the relationship of normalized RRBS-derived mutation calls (SNPs and InDels) between primary and recurring tumors.

E: Cohort-wide mutational profile of genes with known relevance in glioblastoma. Only mutations with high predicted impact on protein function are displayed. Each dot represents a tumor sample in which the indicated mutation was detected. Variant qualities above 100 were set to 100.

Supplementary Figure 3. Prediction of glioblastoma transcriptional subtypes from RRBS data

A: Sample-wise ROC AUC values as a measure for the accuracy of transcriptional subtype prediction from RRBS data. Dashed line: ROC AUC = 0.8

B: Transcriptional subtype distribution for IDH mutated samples. The number of samples assigned to each subtype is indicated.

C, D, E: Distribution of class probabilities across different regions of the same tumor (indicated by Roman numbers) and across different surgeries (indicated by Arabic numbers) for four out of six patients with multi-sector samples (C) and the remaining two patients (D,E) accompanied by matched hematoxylin and eosin stains to display the tumor regions from which the multi-sector samples originated. Section III of the primary tumor of patient 44 did not yield enough CpGs to support confident classification.

F: Scatterplots displaying the relationship between MIRA scores for the indicated factors and the class probabilities of the indicated transcriptional subtypes. r: Pearson correlation, p: p-value.

Supplementary Figure 4. Association of histopathological and MR imaging-derived tumor properties with transcriptional subtypes and disease progression

A: Segmented hematoxylin and eosin stains illustrating the quantification of the histopathological tumor properties (red: hemorrhage, green: necrosis, yellow: meningeal scarring).

B: Comparison of histopathological tumor properties between the different transcriptional subtypes (Cla: classical, Mes: mesenchymal, Pro: proneural).

C: Comparison of histopathological tumor properties between first surgery (primary tumor) and second surgery (recurring tumor).

D: Segmented MR imaging pictures illustrating the quantification of the different MR imaging-derived tumor properties. Heatmap intensity overlays indicate the extent of necrosis, contrast-enhancing (active) tumor volume, and edema in the entire cohort.

E: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified according to the level of necrotic and contrast-enhancing (active) tumor volume as derived from MR imaging in primary and recurring tumors.

All significance tests comparing groups of samples in this figure were performed using a two-sided Wilcoxon rank sum test: * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, ns: not significant.

Supplementary Figure 5. MR imaging progression types and DNA methylation based prediction of tumor properties

A: T1-contrast enhanced and T2/FLAIR MR sequences at each follow-up visit illustrating the three MR imaging progression types in this cohort (cT1 relapse / flare-up, classic T1, T2 diffuse). '*': tumor recurrence.

B: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified according to their MR imaging progression types.

C: Schematic illustrating the machine learning approach (including the data pre-processing) used to assess the predictability of various tumor properties from RRBS DNA methylation data. White squares indicate missing values; grey squares indicate imputed values.

D: ROC curves showing high prediction accuracy based on DNA methylation data for two features with high expected predictability (IDH mutation status, patient sex).

E: ROC curves evaluating the DNA methylation based prediction of several histopathologic tumor properties.

F: Hierarchical clustering based on the column-scaled DNA methylation values of the most predictive features (5-kilobase tiling regions) as identified by the machine leaning classifiers built to predict the infiltration levels of the indicated immune cell types (cells positive for CD163, CD68, CD45ro, CD3, or CD8) or the extent of indicated tumor properties (cells positive for CD34 or HLA-DR).

Supplementary Figure 6. Analysis of DNA methylation heterogeneity in primary and recurring tumors

A: Scatterplots displaying the relationship between PCR enrichment cycles during RRBS library preparation and the indicated measures of DNA methylation heterogeneity. In order to reduce the effect of technical variability, we limited the analysis of epigenomic heterogeneity to samples that fall into a defined narrow range of PCR enrichment cycles (13-15 cycles, indicated by black boxes).

B: Degree of epi-allelic shifting between normal brain control and primary or recurring tumors, as well as between primary and recurring tumors measured by the relative number of loci that show high changes in epi-allele composition (EPM: eloci per million assessed loci). ***: p-value < 0.001 (two-sided Wilcoxon rank sum test)

C: Kaplan-Meier plots displaying progression-free survival and overall survival probabilities over time for patients stratified according to their degree of epi-allelic shifting (as measured by EPM) between primary and recurring tumors.

D: Correlation between epi-allelic shifting during progression (as measured by EPM) and the time between first and second surgery. r: Pearson correlation.

Supplementary Figure 7. Illustration of the graphical data explorer on the Supplementary Website (http://glioblastoma-progression.computational-epigenetics.org/)

A: Comparison between two continuous variables.

B: Comparison between one continuous and one categorical variable.

C: Comparison between two categorical variables.

D: Hovering over an individual data point shows information about the specific data point and also highlights all matched samples from the same patient.

E, F: Clicking on a data point locks the highlighting (E) to follow the selected data point through additional analyses (F).

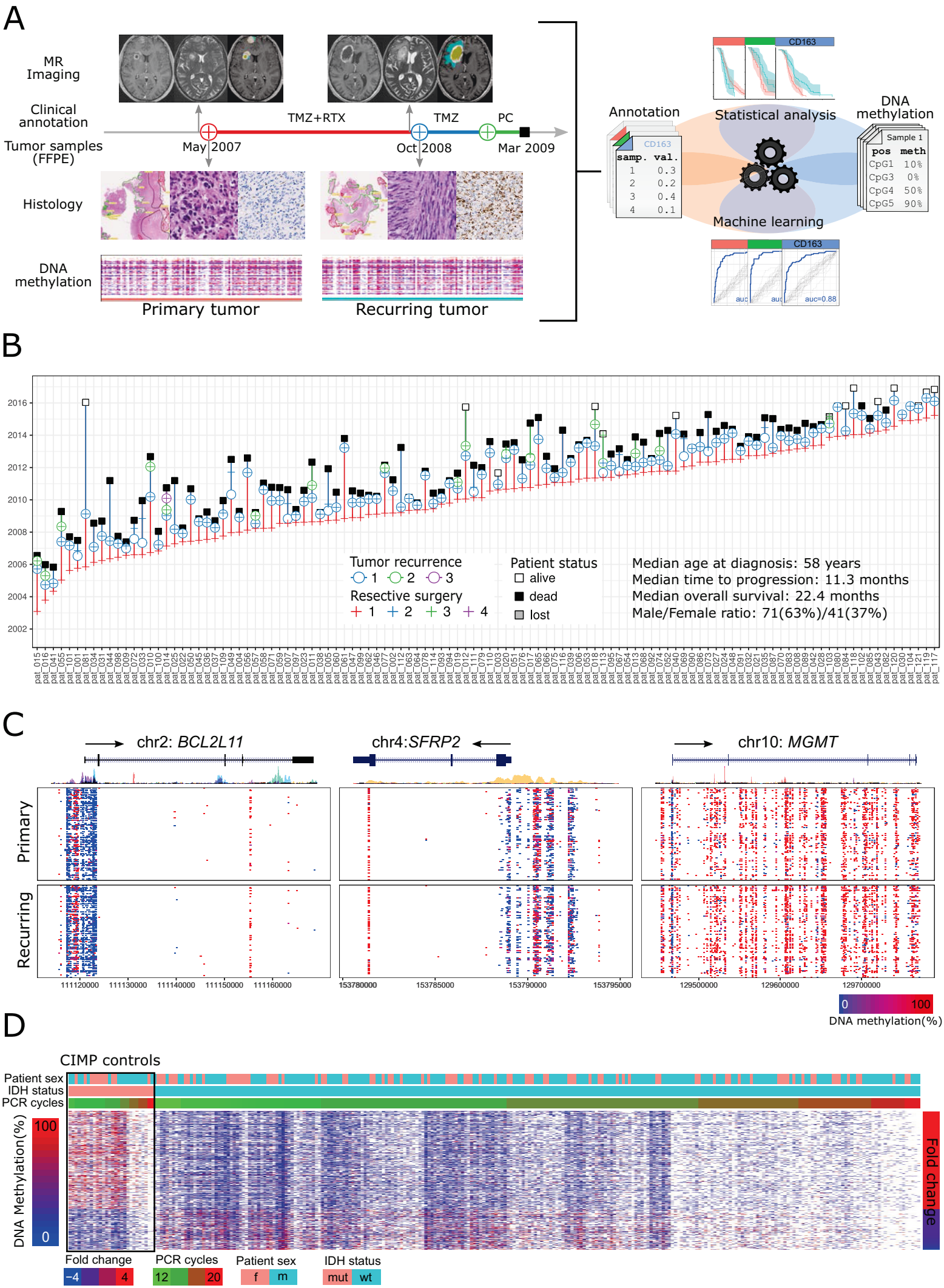

A
MR Imaging
Clinical annotation
Tumor samples (FFPE)
Histology
DNA methylation
TMZ+RTX
TMZ
PC
May 2007
Oct 2008
Mar 2009
Primary tumor
Recurring tumor
Annotation
Statistical analysis
DNA methylation
Machine learning
CD163
samp. val.
1 0.3
2 0.2
3 0.4
4 0.1
Sample 1
pos meth
CpG1 10%
CpG3 0%
CpG4 50%
CpG5 90%
auc=0.88
B
Tumor recurrence
1 2 3
Resective surgery
1 2 3 4
Patient status
alive
dead
lost
Median age at diagnosis: 58 years
Median time to progression: 11.3 months
Median overall survival: 22.4 months
Male/Female ratio: 71(63%)/41(37%)
C
chr2: BCL2L11
chr4:SFRP2
chr10: MGMT
Primary
Recurring
111120000 111130000 111140000 111150000 111160000
153780000 153785000 153790000 153795000
129500000 129600000 129700000
0 100
DNA methylation(%)
D
CIMP controls
Patient sex
IDH status
PCR cycles
DNA Methylation(%)
100
0
Fold change
Fold change
−4 4
PCR cycles
12 20
Patient sex
f m
IDH status
mut wt

Figure 1

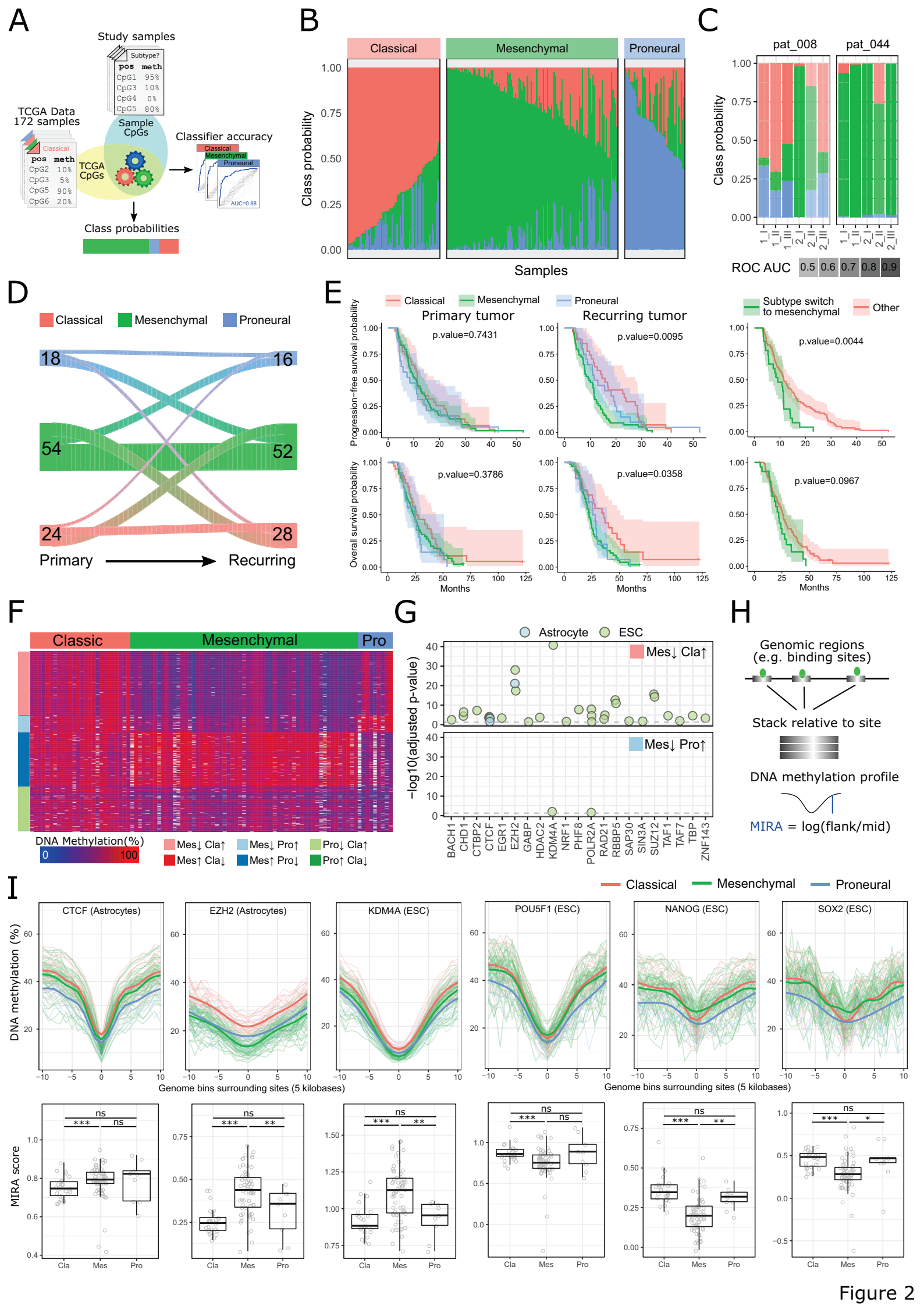

A
Study samples
Subtype?
pos meth
CpG1 95%
CpG3 10%
CpG4 0%
CpG5 80%
TCGA Data
172 samples
Classical
pos meth
CpG2 10%
CpG3 5%
CpG5 90%
CpG6 20%
Sample CpGs
TCGA CpGs
Classifier accuracy
Classical
Mesenchymal
Proneural
AUC=0.88
Class probabilities
B
Classical
Mesenchymal
Proneural
Class probability
Samples
C
pat_008
pat_044
Class probability
ROC AUC
0.5 0.6 0.7 0.8 0.9
D
Classical
Mesenchymal
Proneural
18
16
54
52
24
28
Primary
Recurring
E
Classical
Mesenchymal
Proneural
Primary tumor
Recurring tumor
Subtype switch to mesenchymal
Other
Progression−free survival probability
Overall survival probability
p.value=0.7431
p.value=0.0095
p.value=0.0044
p.value=0.3786
p.value=0.0358
p.value=0.0967
Months
F
Classic
Mesenchymal
Pro
DNA Methylation(%)
0
100
Mes↓ Cla↑
Mes↓ Pro↑
Pro↓ Cla↑
Mes↑ Cla↓
Mes↑ Pro↓
Pro↑ Cla↓
G
Astrocyte
ESC
Mes↓ Cla↑
Mes↓ Pro↑
−log10(adjusted p-value)
BACH1
CHD1
CTBP2
CTCF
EGR1
EZH2
GABP
HDAC2
KDM4A
NRF1
PHF8
POLR2A
RAD21
RBBP5
SAP30
SIN3A
SUZ12
TAF1
TAF7
TBP
ZNF143
H
Genomic regions (e.g. binding sites)
Stack relative to site
DNA methylation profile
MIRA = log(flank/mid)
I
Classical
Mesenchymal
Proneural
CTCF (Astrocytes)
EZH2 (Astrocytes)
KDM4A (ESC)
POU5F1 (ESC)
NANOG (ESC)
SOX2 (ESC)
DNA methylation (%)
Genome bins surrounding sites (5 kilobases)
MIRA score
ns
***
**
*
Cla
Mes
Pro

Figure 2

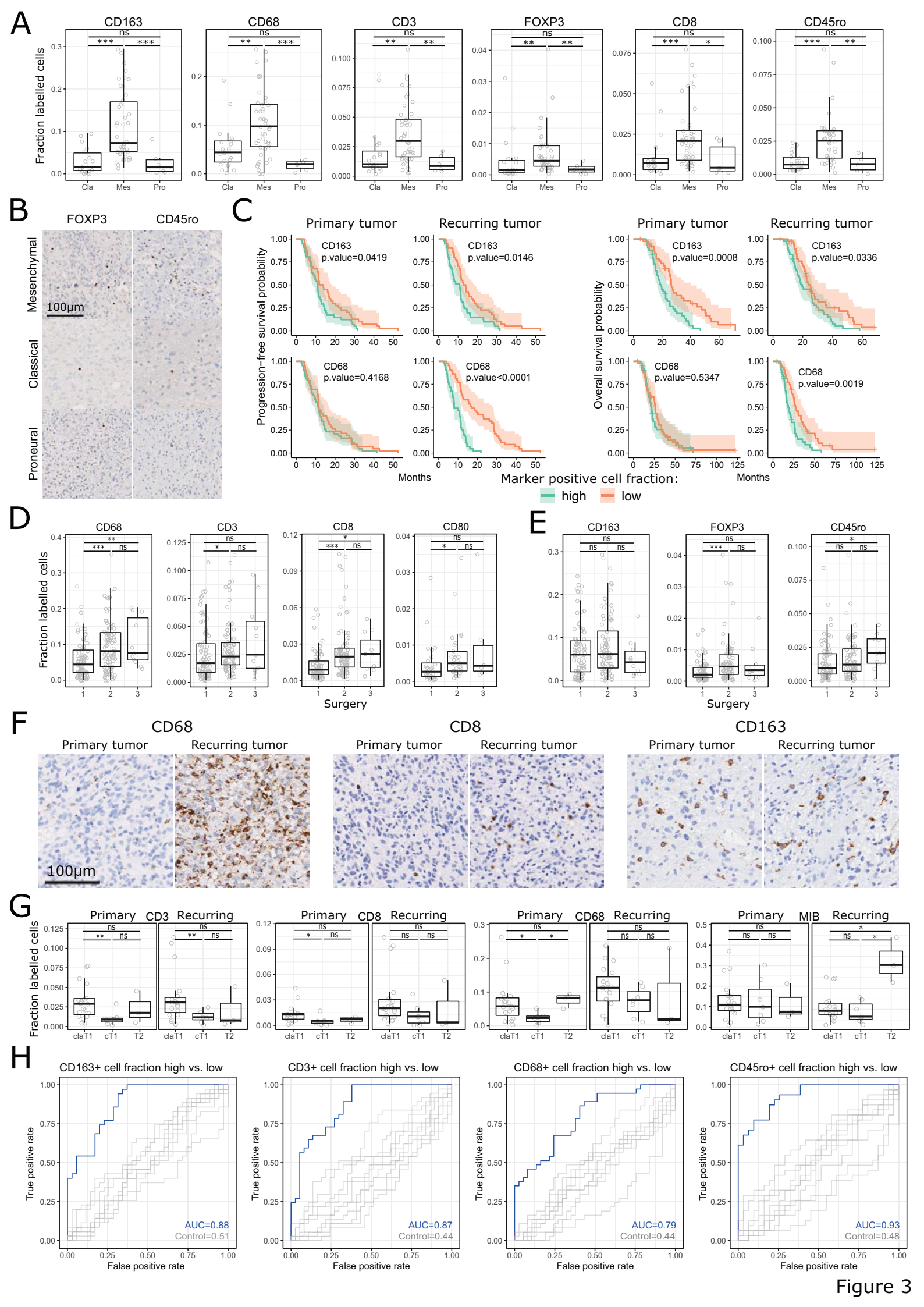

A
CD163
CD68
CD3
FOXP3
CD8
CD45ro
Fraction labelled cells
Cla
Mes
Pro
B
FOXP3
CD45ro
Mesenchymal
Classical
Proneural
100µm
C
Primary tumor
Recurring tumor
CD163 p.value=0.0419
CD163 p.value=0.0146
CD163 p.value=0.0008
CD163 p.value=0.0336
CD68 p.value=0.4168
CD68 p.value<0.0001
CD68 p.value=0.5347
CD68 p.value=0.0019
Progression−free survival probability
Overall survival probability
Months
Marker positive cell fraction:
high
low
D
CD68
CD3
CD8
CD80
Fraction labelled cells
Surgery
E
CD163
FOXP3
CD45ro
Surgery
F
CD68
CD8
CD163
Primary tumor
Recurring tumor
100µm
G
Primary CD3 Recurring
Primary CD8 Recurring
Primary CD68 Recurring
Primary MIB Recurring
Fraction labelled cells
claT1
cT1
T2
H
CD163+ cell fraction high vs. low
CD3+ cell fraction high vs. low
CD68+ cell fraction high vs. low
CD45ro+ cell fraction high vs. low
True positive rate
False positive rate
AUC=0.88
Control=0.51
AUC=0.87
Control=0.44
AUC=0.79
Control=0.44
AUC=0.93
Control=0.48

Figure 3

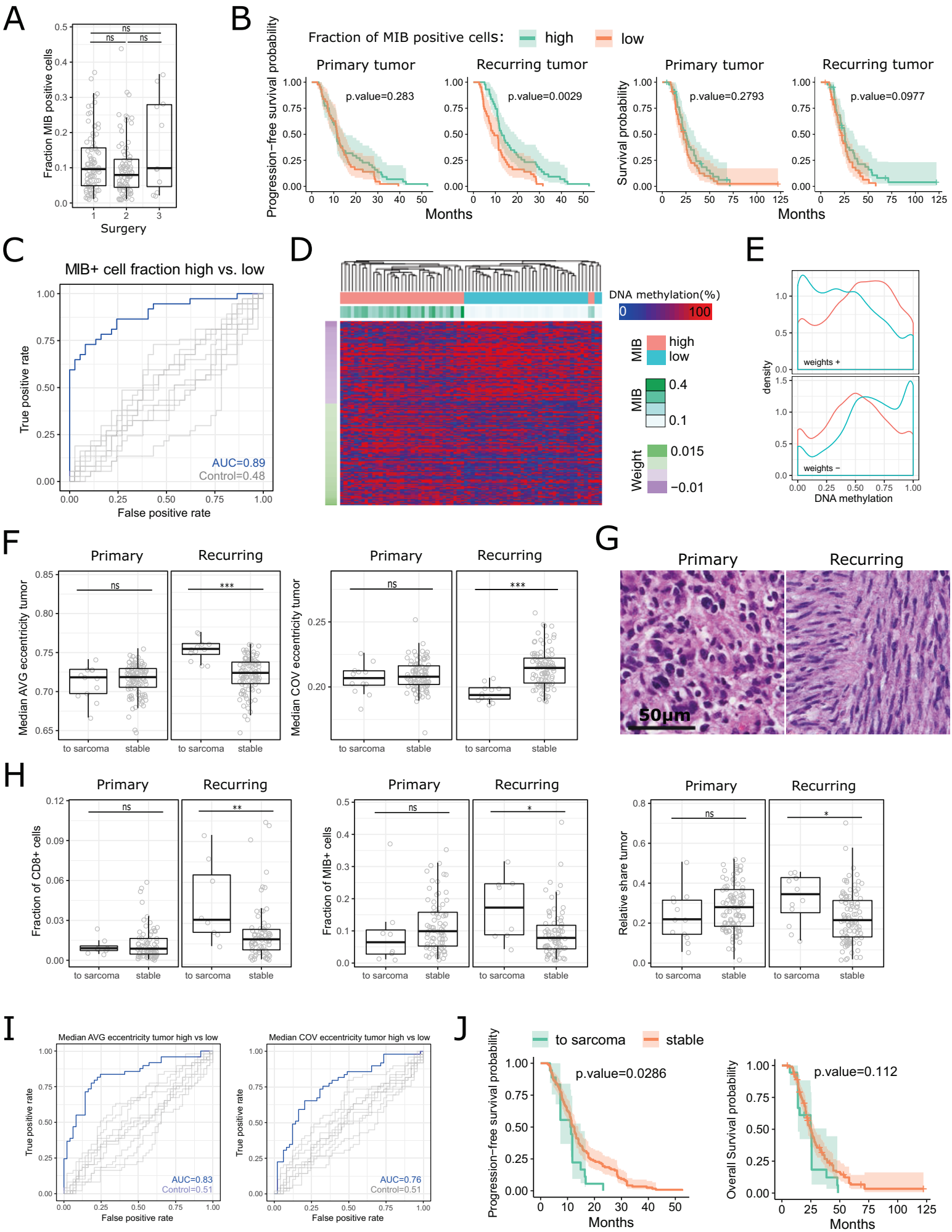

A
Fraction MIB positive cells
Surgery
ns
B
Fraction of MIB positive cells: high low
Primary tumor
Recurring tumor
p.value=0.283
p.value=0.0029
p.value=0.2793
p.value=0.0977
Progression−free survival probability
Survival probability
Months
C
MIB+ cell fraction high vs. low
True positive rate
False positive rate
AUC=0.89
Control=0.48
D
DNA methylation(%)
MIB
high
low
Weight
0.015
−0.01
E
density
weights +
weights −
DNA methylation
F
Primary
Recurring
Median AVG eccentricity tumor
Median COV eccentricity tumor
to sarcoma
stable
G
50µm
H
Fraction of CD8+ cells
Fraction of MIB+ cells
Relative share tumor
I
Median AVG eccentricity tumor high vs low
Median COV eccentricity tumor high vs low
AUC=0.83
Control=0.51
AUC=0.76
Control=0.51
J
to sarcoma
stable
p.value=0.0286
p.value=0.112
Overall Survival probability

Figure 4

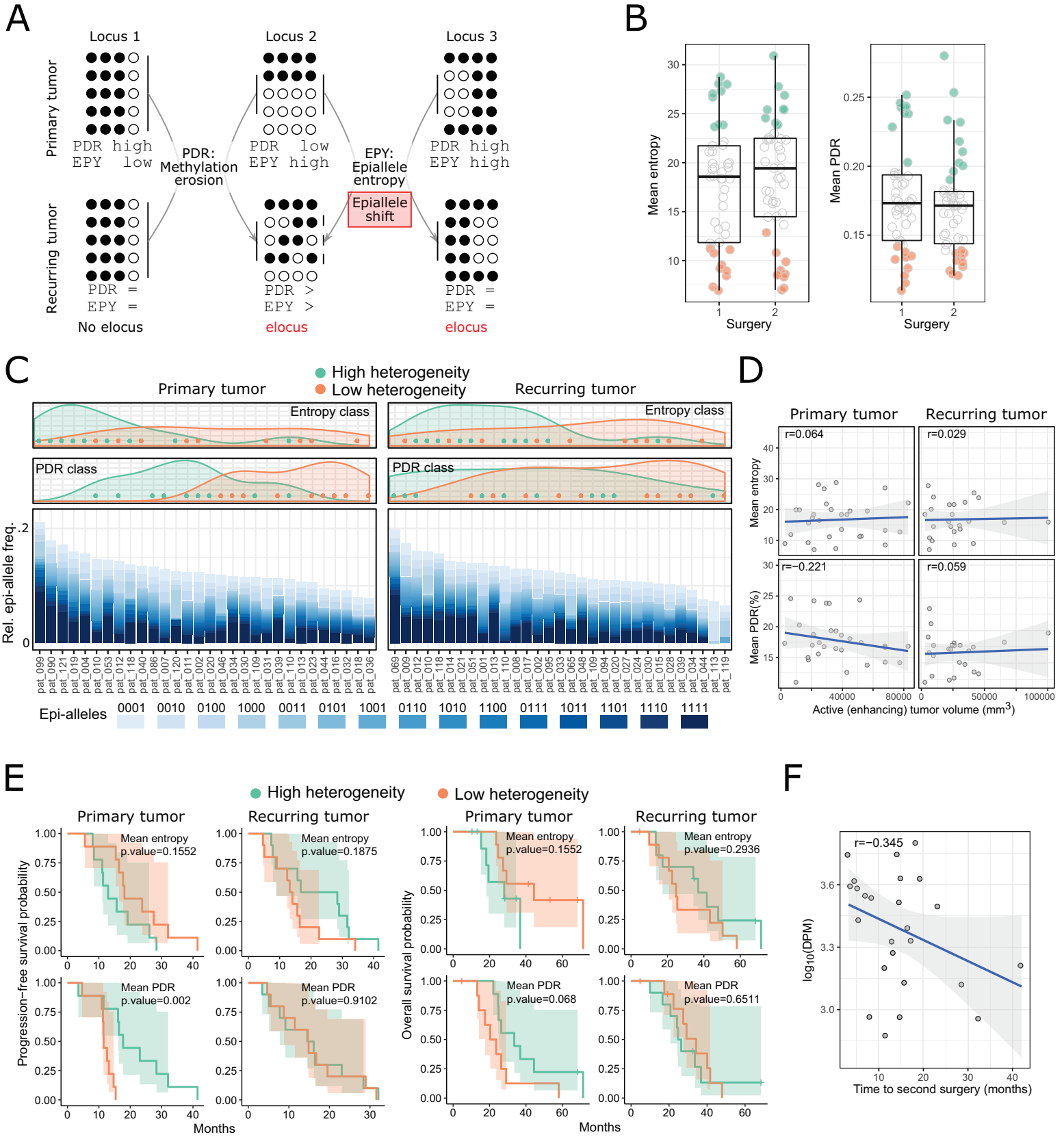

A
Locus 1
Locus 2
Locus 3
Primary tumor
PDR high
EPY low
PDR:
Methylation
erosion
PDR low
EPY high
EPY:
Epiallele
entropy
PDR high
EPY high
Epiallele
shift
Recurring tumor
PDR =
EPY =
No elocus
PDR >
EPY >
elocus
PDR =
EPY =
elocus
B
Mean entropy
Surgery
Mean PDR
Surgery
C
High heterogeneity
Low heterogeneity
Primary tumor
Recurring tumor
Entropy class
PDR class
Rel. epi-allele freq.
Epi-alleles
0001
0010
0100
1000
0011
0101
1001
0110
1010
1100
0111
1011
1101
1110
1111
D
Primary tumor
Recurring tumor
r=0.064
r=0.029
r=−0.221
r=0.059
Mean entropy
Mean PDR(%)
Active (enhancing) tumor volume (mm3)
E
High heterogeneity
Low heterogeneity
Primary tumor
Recurring tumor
Mean entropy
p.value=0.1552
Mean entropy
p.value=0.1875
Mean entropy
p.value=0.1552
Mean entropy
p.value=0.2936
Mean PDR
p.value=0.002
Mean PDR
p.value=0.9102
Mean PDR
p.value=0.068
Mean PDR
p.value=0.6511
Progression−free survival probability
Overall survival probability
Months
F
r=−0.345
log10(DPM)
Time to second surgery (months)

Figure 5

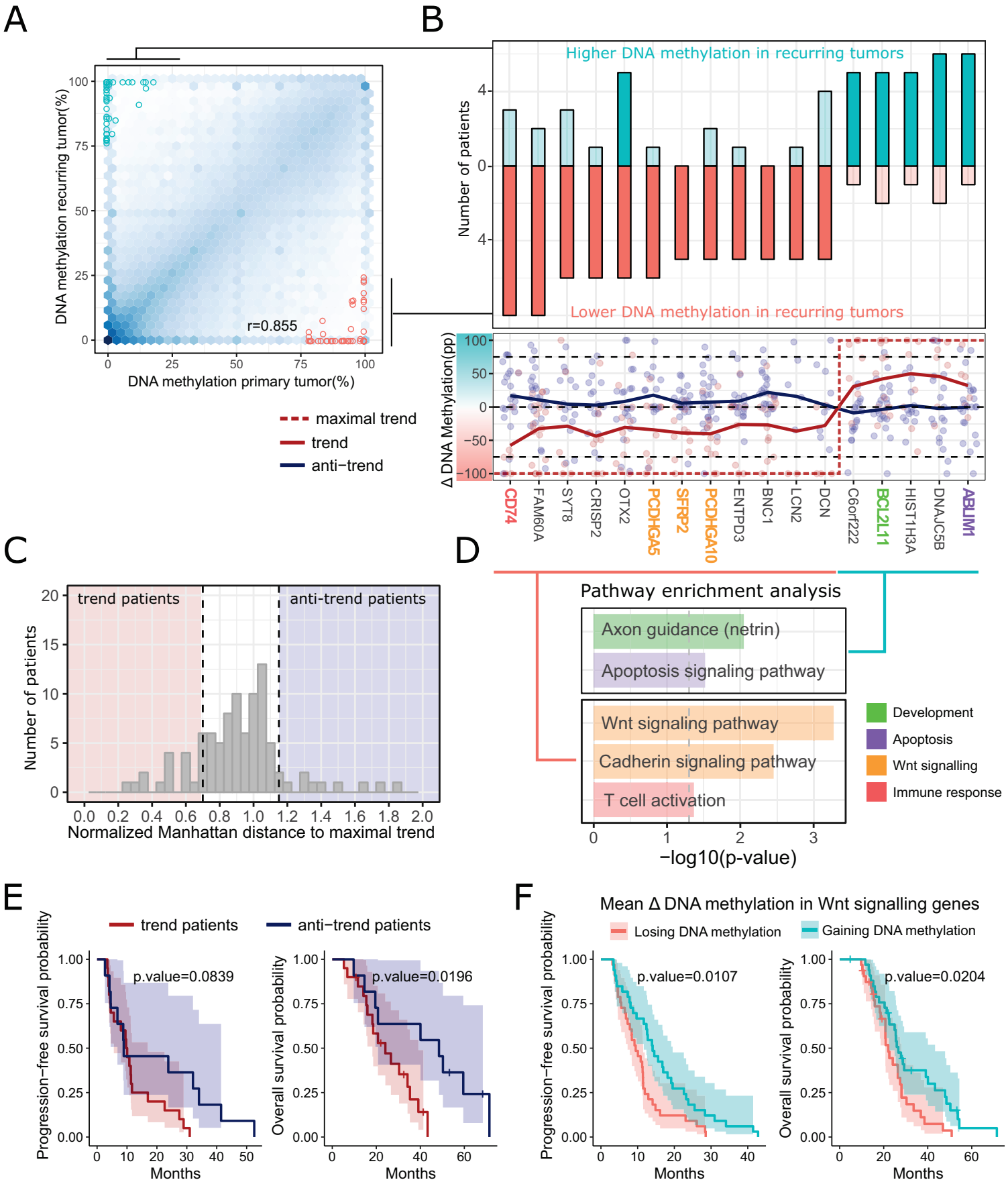
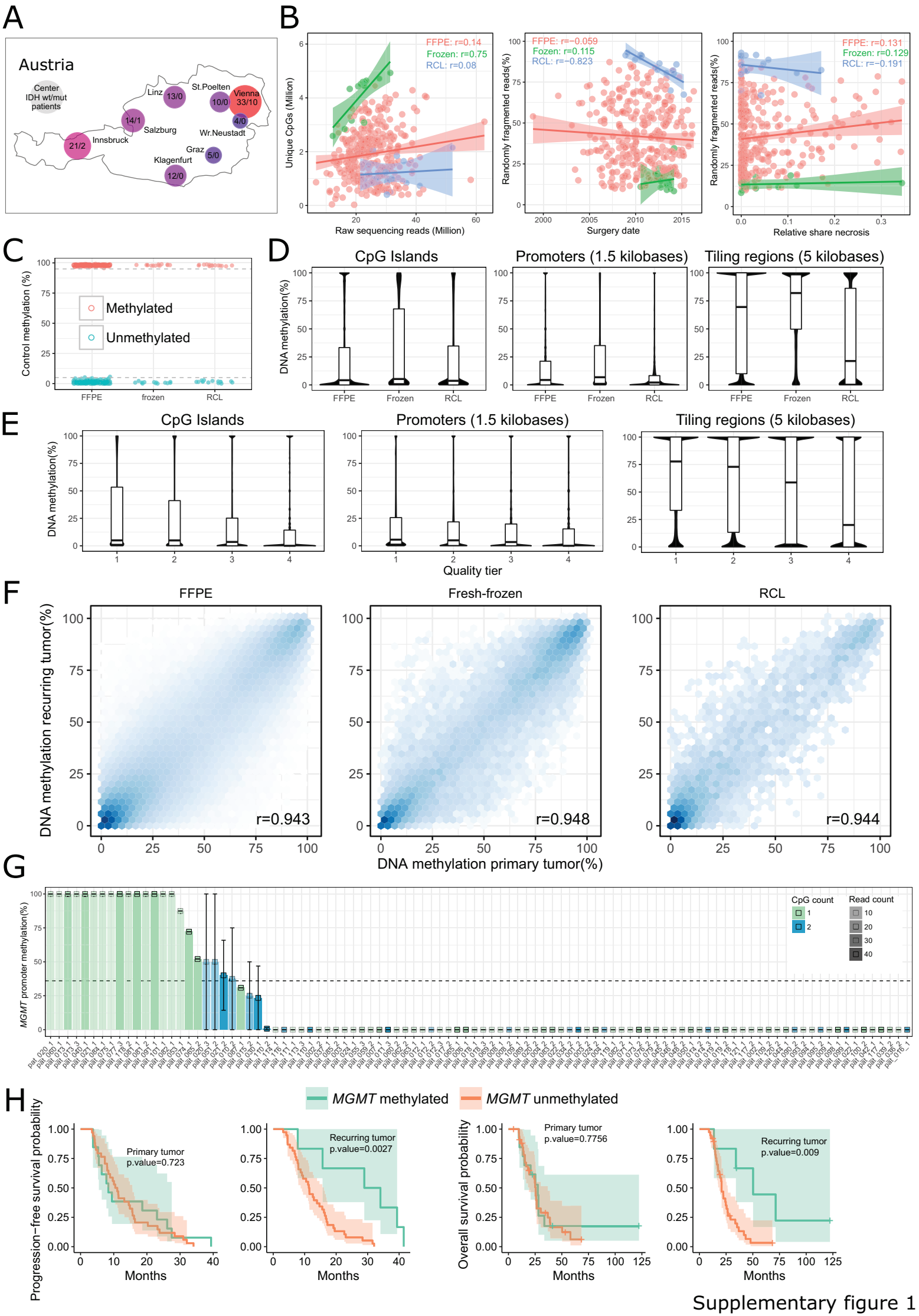
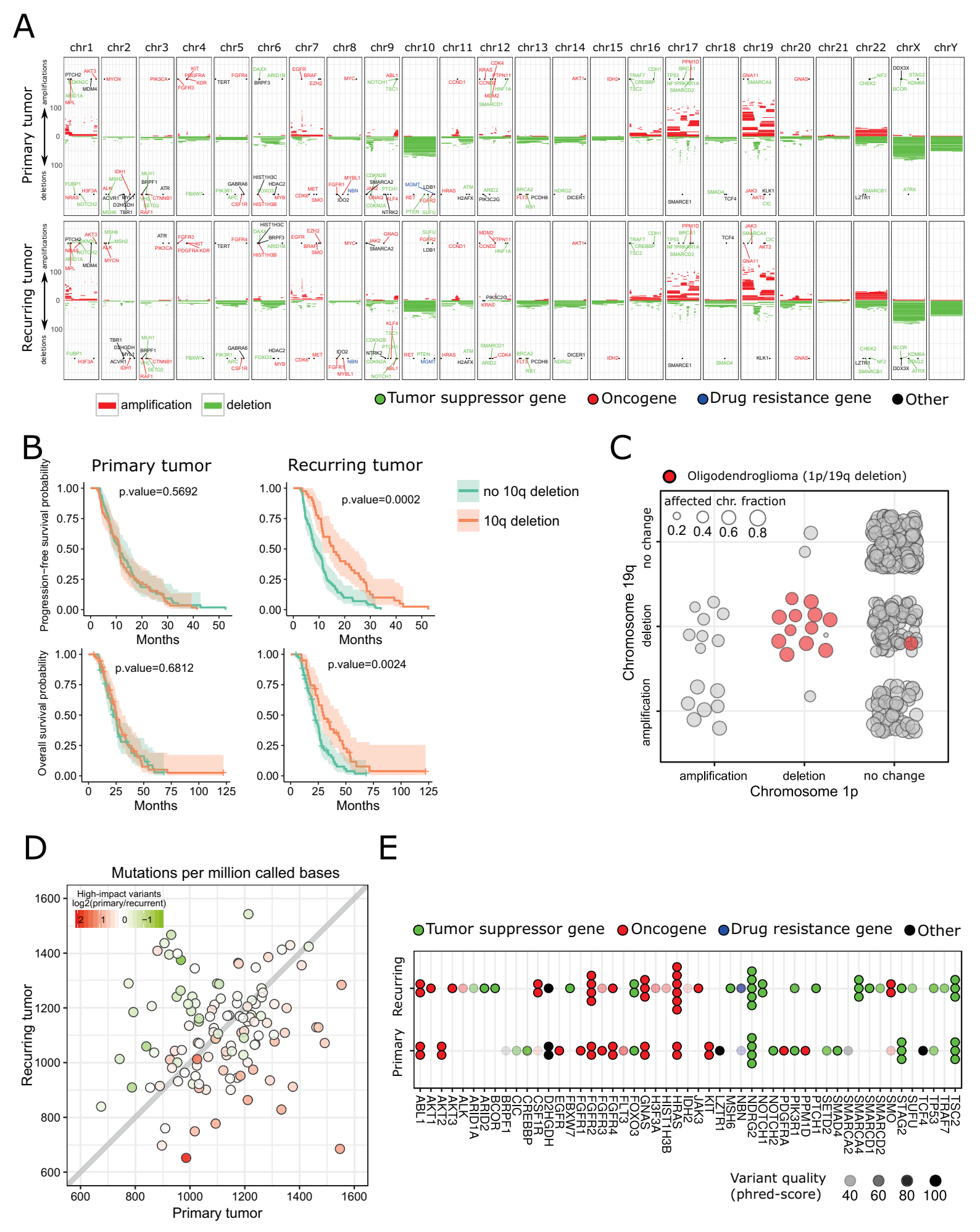

A
B
C
D
E
F
DNA methylation recurring tumor(%)
DNA methylation primary tumor(%)
r=0.855
maximal trend
trend
anti−trend
Higher DNA methylation in recurring tumors
Lower DNA methylation in recurring tumors
Number of patients
Δ DNA Methylation(pp)
CD74
FAM60A
SYT8
CRISP2
OTX2
PCDHGA5
SFRP2
PCDHGA10
ENTPD3
BNC1
LCN2
DCN
C6orf222
BCL2L11
HIST1H3A
DNAJC5B
ABLIM1
trend patients
anti-trend patients
Normalized Manhattan distance to maximal trend
Pathway enrichment analysis
Axon guidance (netrin)
Apoptosis signaling pathway
Wnt signaling pathway
Cadherin signaling pathway
T cell activation
Development
Apoptosis
Wnt signalling
Immune response
−log10(p-value)
trend patients
anti−trend patients
Progression−free survival probability
Overall survival probability
p.value=0.0839
p.value=0.0196
Months
Mean Δ DNA methylation in Wnt signalling genes
Losing DNA methylation
Gaining DNA methylation
p.value=0.0107
p.value=0.0204

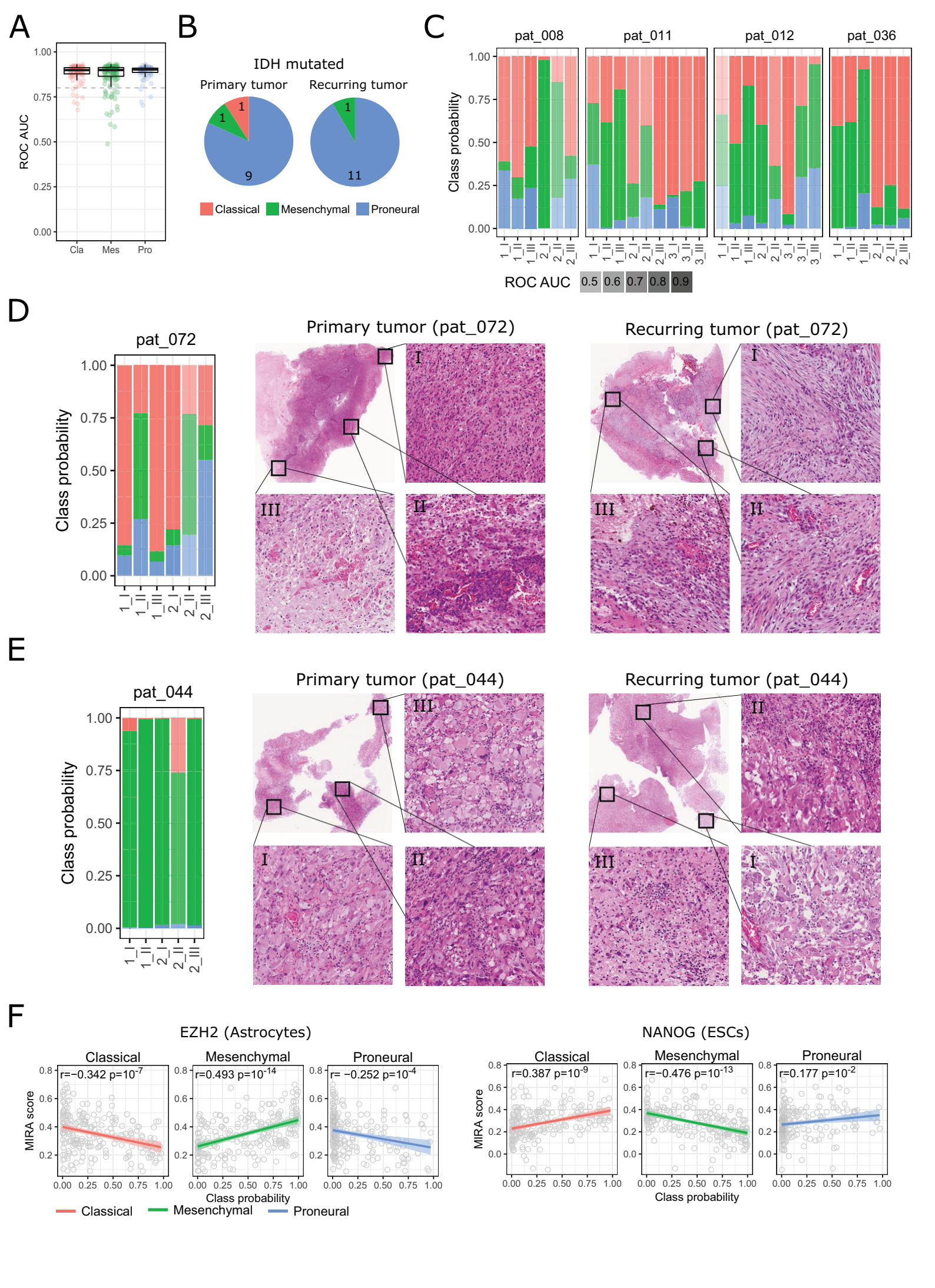
Figure 6



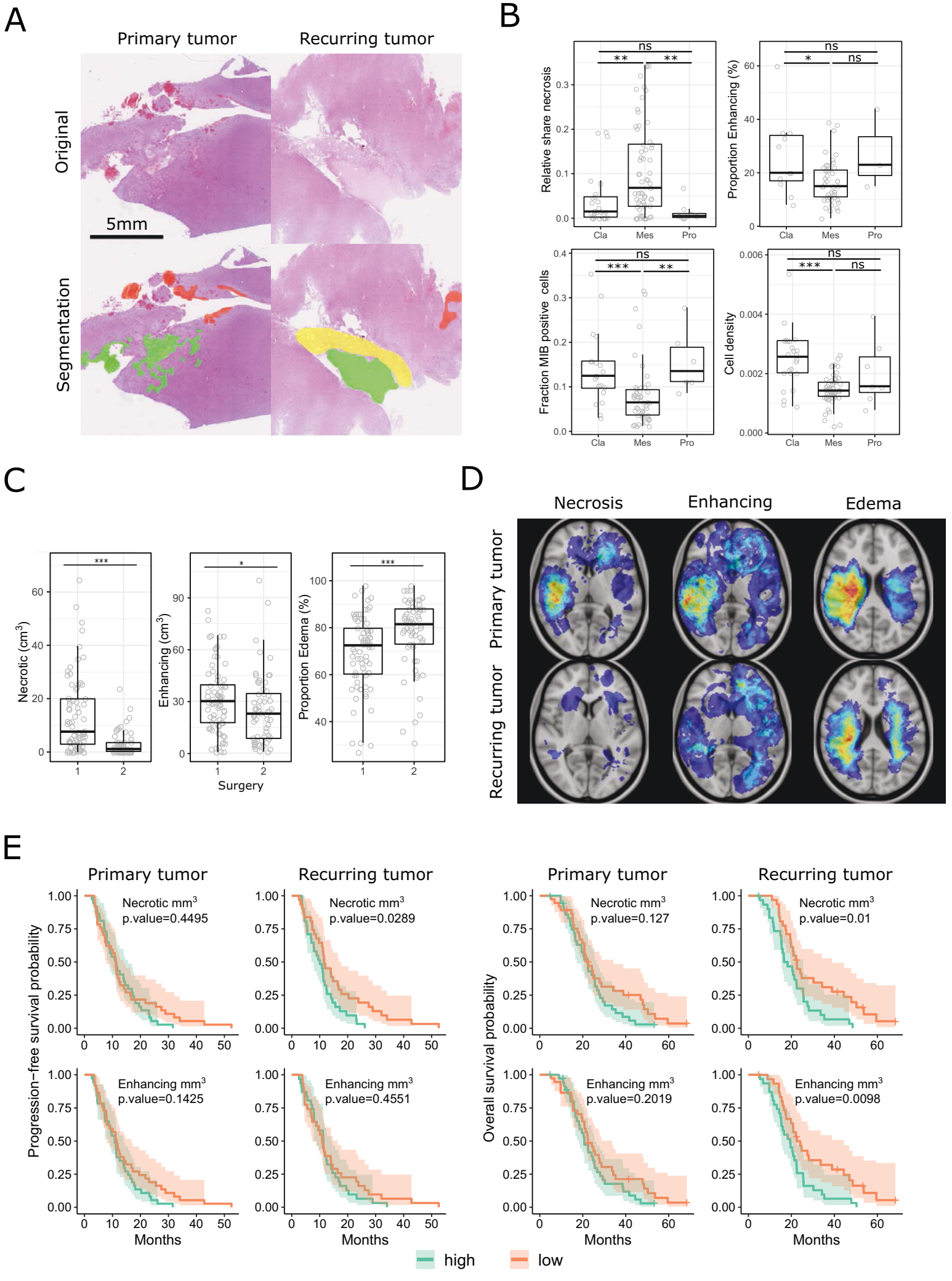
Supplementary figure 1



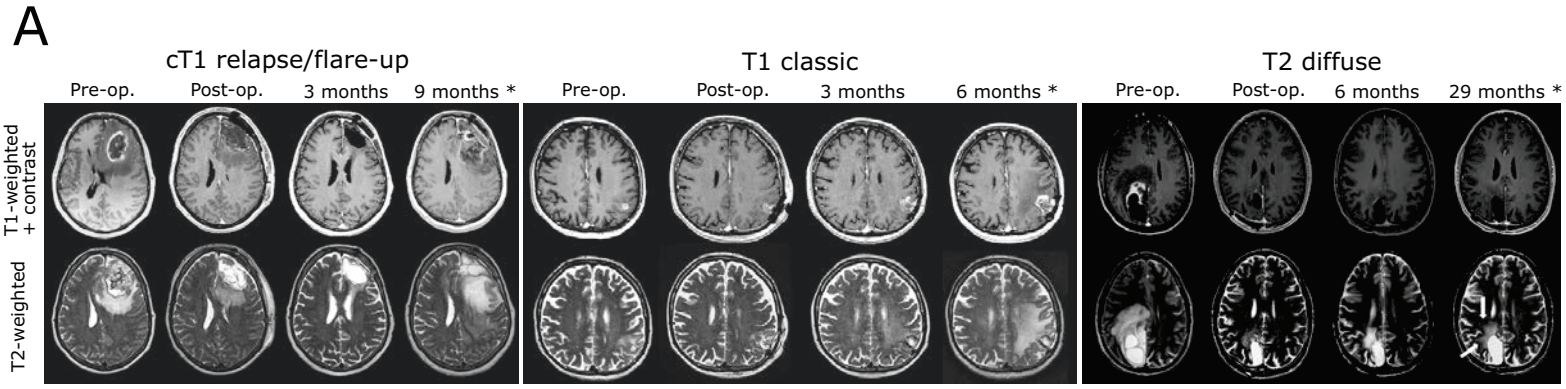
Supplementary figure 2

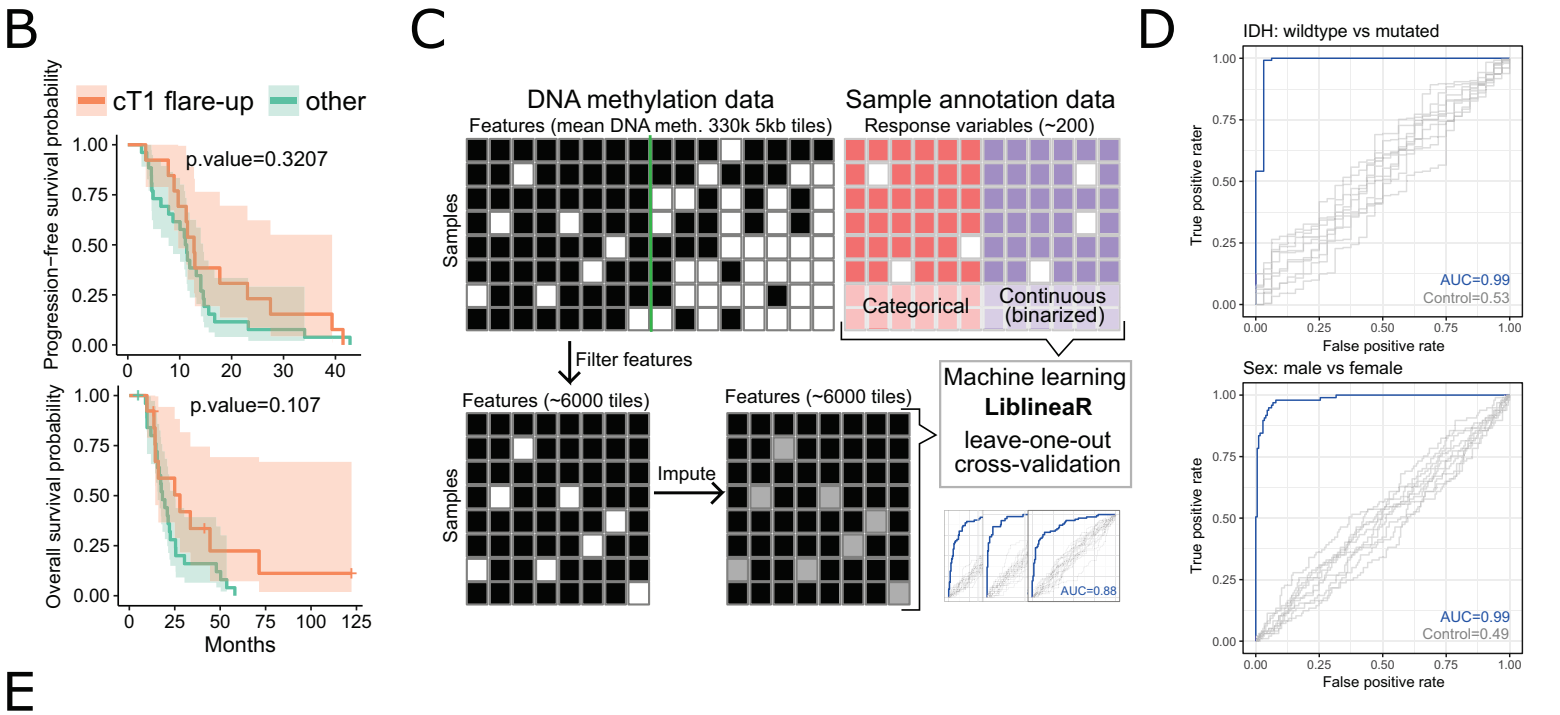


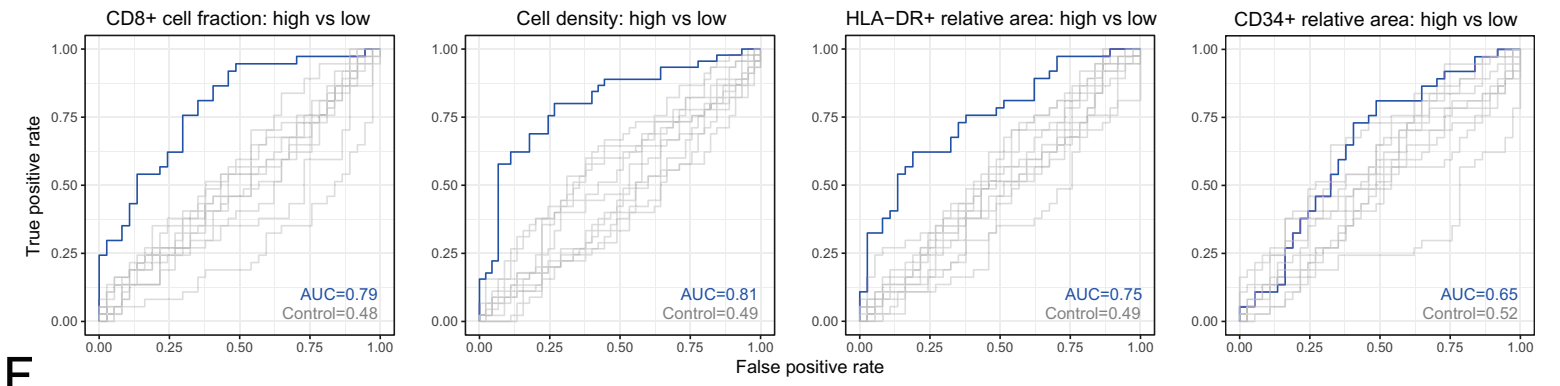
Supplementary figure 3

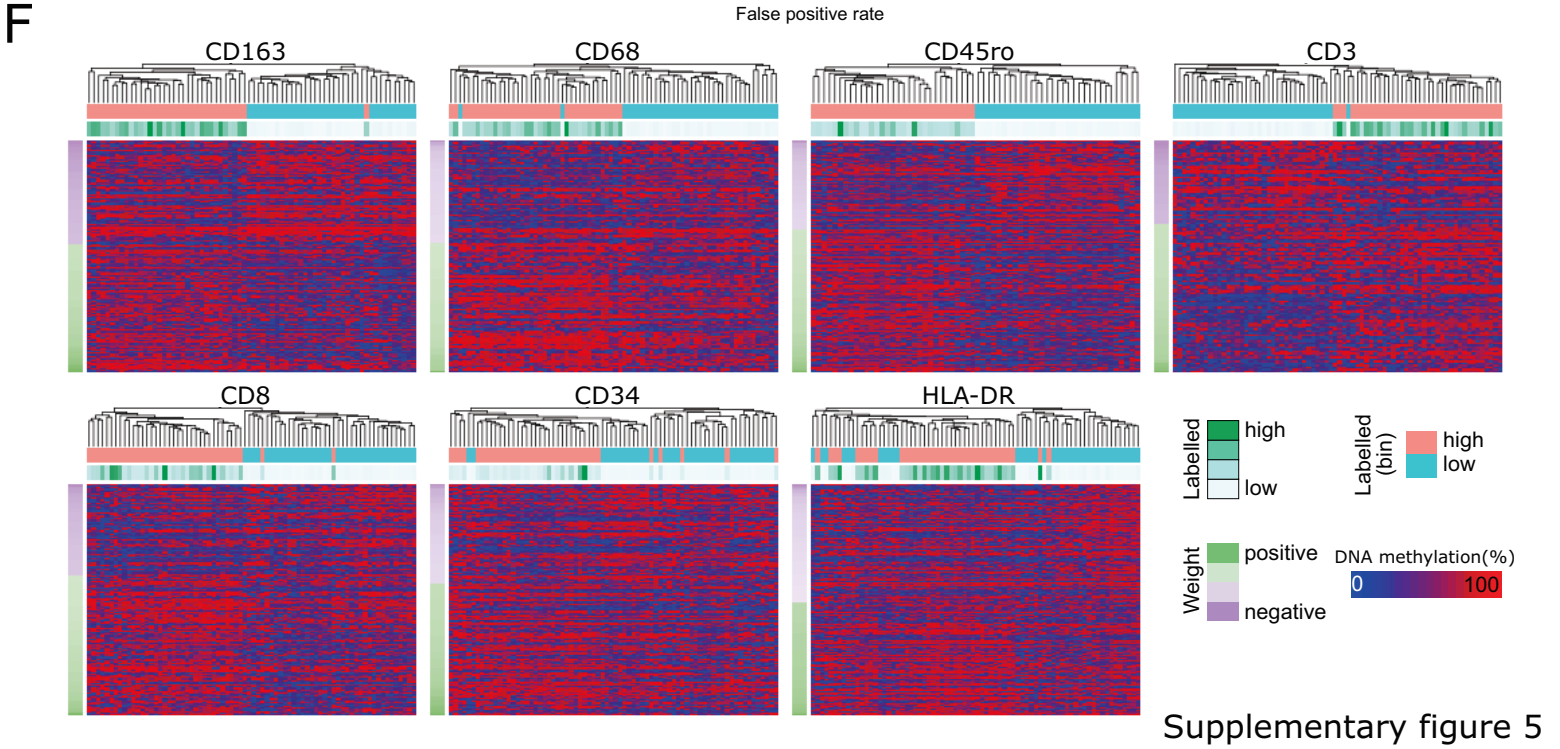


Supplementary figure 4

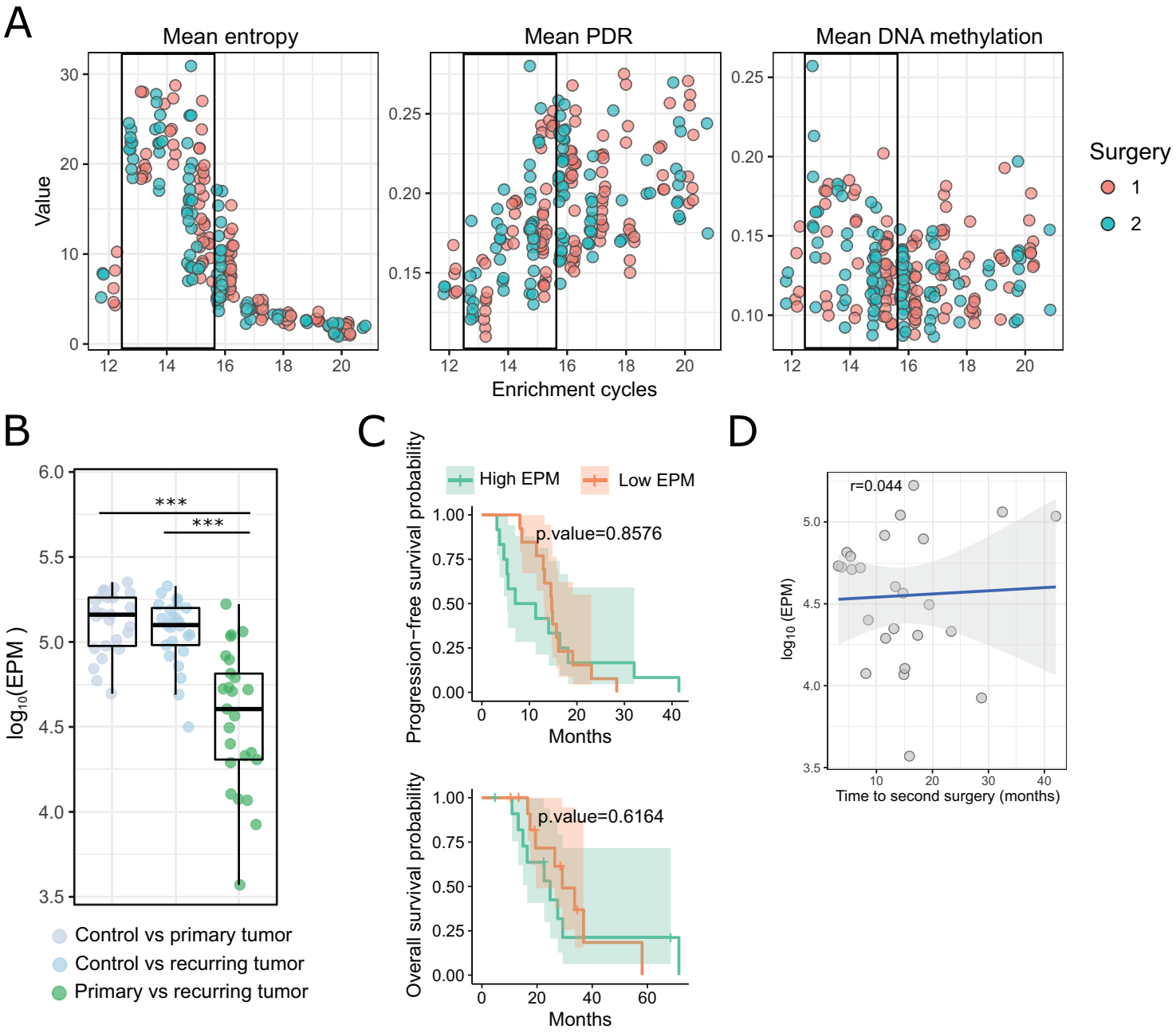




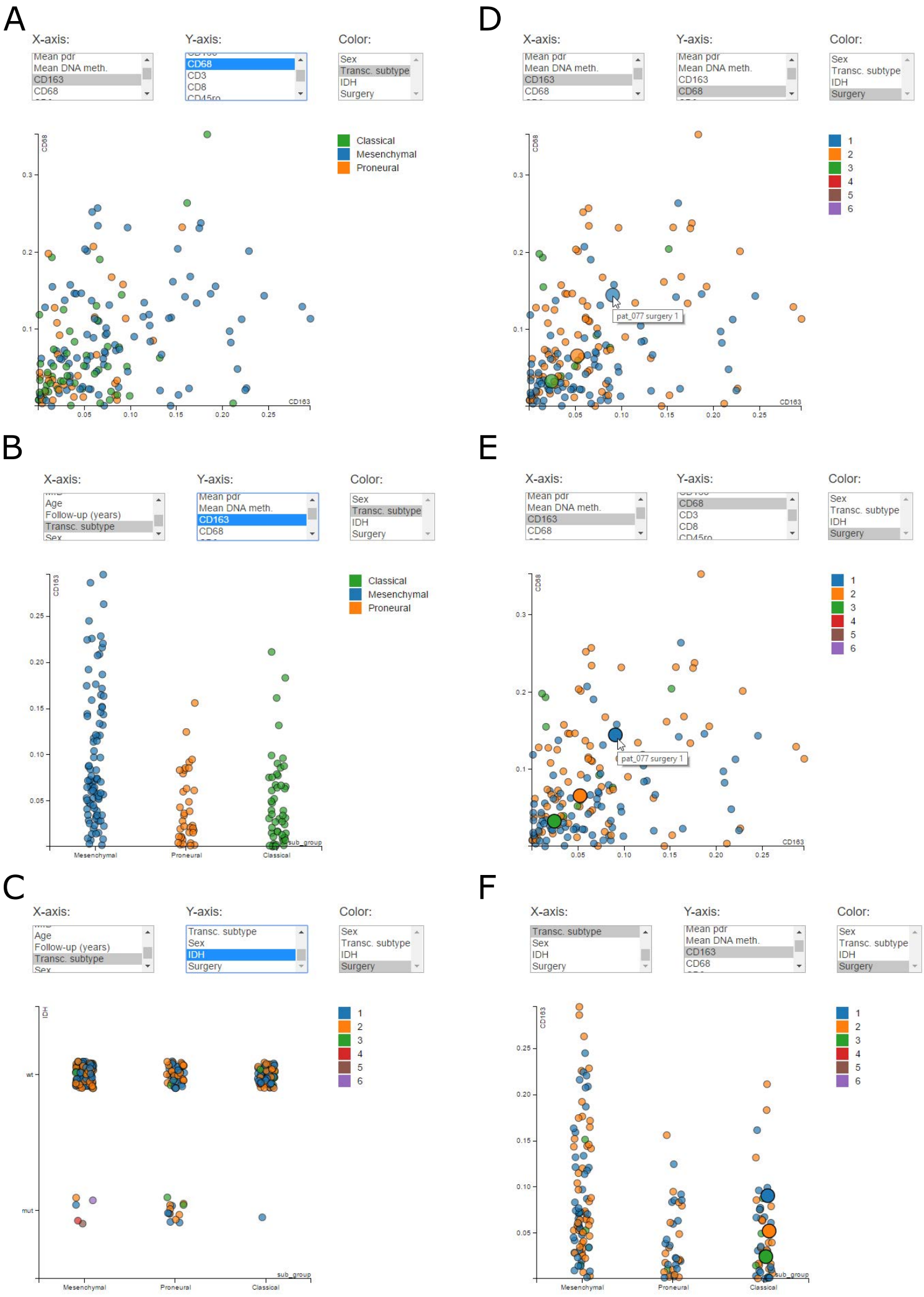




Supplementary figure 5



Supplementary figure 6



Supplementary figure 7

3.3 HumanIslet

Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types.

Li J*, Klughammer J*, Farlik M*, Penz T*, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S.

EMBO Rep. 2016 Feb;17(2):178-87. doi:10.15252/embr.201540946.

Published online: December 21, 2015

Scientific Report

TRANSPARENT PROCESS OPEN ACCESS

EMBO reports

Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types

Jin Li[1,†], Johanna Klughammer[1,†], Matthias Farlik[1,†], Thomas Penz[1,†], Andreas Spittler[2], Charlotte Barbieux[3], Ekaterine Berishvili[3], Christoph Bock[1,4,5,*] & Stefan Kubicek[1,6,**]

Abstract

Pancreatic islets of Langerhans contain several specialized endocrine cell types, which are commonly identified by the expression of single marker genes. However, the established marker genes cannot capture the complete spectrum of cellular heterogeneity in human pancreatic islets, and existing bulk transcriptome datasets provide averages across several cell populations. To dissect the cellular composition of the human pancreatic islet and to establish transcriptomes for all major cell types, we performed single-cell RNA sequencing on 70 cells sorted from human primary tissue. We used this dataset to validate previously described marker genes at the single-cell level and to identify specifically expressed transcription factors for all islet cell subtypes. All data are available for browsing and download, thus establishing a useful resource of single-cell expression profiles for endocrine cells in human pancreatic islets.

Keywords alpha cells; beta cells; diabetes; marker genes; single-cell RNA-seq
Subject Category Systems & Computational Biology
DOI 10.15252/embr.201540946 | Received 29 June 2015 | Revised 16 November 2015 | Accepted 19 November 2015

Introduction

Located within the pancreas, the islets of Langerhans are composed of endocrine cells expressing glucagon (alpha cells), insulin (beta cells), somatostatin (delta cells), pancreatic polypeptide (PP cells), and ghrelin (epsilon cells). Furthermore, they are heavily vascularized and innervated, and in contact with the surrounding acinar and ductal cells of the exocrine pancreas. Pancreatic islets function as highly specialized micro-organs that monitor and maintain blood glucose homeostasis. While damage to beta cells causes diabetes, the other pancreatic cell types may also contribute to pathogenesis in ways that are not well understood. Recent studies showed that both alpha [1] and delta cells [2] have the potential to replenish beta cell mass in animal models.

Development of diabetes correlates with global changes in the transcriptome of pancreatic islets [3]. These gene expression changes could reflect alterations in the cell subtype composition of the islet and/or changes in the transcriptomes of beta cells or other individual cell types. Analyzing islet cell-specific gene expression changes has the potential to shed light on the etiology of diabetes. Recently, alpha and beta cell purification protocols from human [4–6] and mouse islets [7,8] have yielded initial maps of cell type-specific transcriptomes. The available transcriptome datasets further comprise primary mouse and human alpha cells, beta cells, and delta cells, a number of rodent alpha and beta cell lines, and one human beta cell line [4,9–12]. Despite the rapid progress in this field, a comprehensive transcriptome database for individual human islet cell types is still missing, and no transcriptome data are currently available for PP cells.

Recent advances in next-generation sequencing and library preparation enabled for the first time the transcriptome characterization of single cells from primary tissue. For example, this approach was successfully used to establish transcriptome profiles and dissect cell type heterogeneity for primary tissue obtained from the lung [13], the spleen, and the brain [14,15].

Here, we used single-cell RNA-seq to establish a comprehensive transcriptome database for the cell types that are present in primary human pancreatic islets. Principal component analysis in combination with visualization as biplots identified alpha cells, beta cells, delta cells, PP cells, acinar cells, and pancreatic duct cells directly from the single-cell transcriptome profiles. We illustrate the utility of this resource by discovering novel cell type-specific marker genes, and we identified human-specific expression patterns in alpha and beta cells. All data are readily available for user-friendly online browsing and download to foster research on pancreatic islet biology and diabetes-related mechanisms in human.

1 CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
2 Medical University of Vienna, Anna Spiegel Forschungsgebäude, Vienna, Austria
3 Department of Surgery, Cell Isolation and Transplantation Center, Geneva University Hospitals, University of Geneva, Geneva, Switzerland
4 Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria
5 Max Planck Institute for Informatics, Saarbrücken, Germany
6 Christian Doppler Laboratory for Chemical Epigenetics and Antiinfectives, CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
*Corresponding author. Tel: +43 1 40160 70070; Fax: +43 1 40160 970 000; E-mail: cbock@cemm.oeaw.ac.at
**Corresponding author. Tel: +43 1 40160 70036; Fax: +43 1 40160 970 000; E-mail: skubicek@cemm.oeaw.ac.at
†These authors contributed equally to this work

Results and Discussion

Single-cell transcriptomes recapitulate pancreatic endocrine cell types

Primary human pancreatic islets of Langerhans were disassociated into single cells, and these cells were sorted into individual wells of a 96-well plate by FACS [16]. The Smart-Seq2 protocol [17] was then applied to obtain single-cell transcriptomes. Following the generation and amplification of cDNA, we determined the levels of beta-actin expression by qRT–PCR and selected all cell-containing wells for library preparation and next-generation sequencing (Fig 1A). Seventy cells were sequenced in total, of which 64 cells passed quality control (see Materials and Methods) and were included in the analysis (Fig EV1A and B, and Dataset EV1). We obtained an average of 12.7 million high-quality reads per single cell, of which 62.9% aligned to the human reference genome. RNA expression levels were calculated using the BitSeq software which uses RPKM normalization and corrects for non-uniform read distribution along the transcripts (e.g., 3-prime bias) [18]. Data quality was validated by assessing the relation between expression level and transcript length in native RNA (Fig EV1C) as well as ERCC spike-in controls (Fig EV1D). While transcript length and expression level were not correlated in the ERCC spike-in controls, we detected a negative correlation ($r = -0.405$) in the native RNA which was in the range of what had been previously reported as biologically significant finding [19]. However, a potential bias due to transcript length normalization cannot be completely excluded; therefore, comparing expression levels of different transcripts/genes should be performed with caution. To define global similarities among the single cells and the marker genes that drive these similarities, we performed principal component analysis (PCA) on the transcriptome dataset and displayed the results as biplots. PCA on the full dataset separates a group of 18 cells based on high *glucagon (GCG)* and *transthyretin (TTR)* expression and a group of 9 cells expressing *pancreatic polypeptide (PPY)* from a heterogeneous group of 37 cells (Fig 1B). In a second PCA on the 37 yet undefined cells, we identified a group of 12 cells with high *insulin (INS)* expression, a group of 11 cells characterized by *PRSS2*, *CTRB2*, *REG3A*, *REG1A*, and *REG1B* and a group of two *somatostatin (SST)*-expressing cells. In a third PCA on the remaining 12 undefined cells, a group of 8 cells was characterized by *keratin18 (KRT18)* and *keratin8 (KRT8)*. Based on the expression profiles of the identified marker genes, we were able to uniquely assign 60 out of 64 single-cell transcriptomes to the alpha, beta, delta, PP, acinar, or ductal cell type (Fig 1C).

As an additional validation of our cell type classification, we visualized the global transcriptional similarity of individual pancreatic cells by multidimensional scaling (MDS), where each single-cell transcriptome was colored by the cell type derived from PCA (Fig 1D). When mapped upon the MDS plot, the known cell type-specific marker genes *INS*, *GCG*, *PPY*, *SST*, *REG1A,* and *KRT8* show the expected expression patterns, with different amounts of variability within the subgroups (Fig 1E). The validity of our single-cell RNA-seq dataset was further confirmed in direct comparison to an external dataset consisting of bulk RNA-seq data for whole islet, beta, and acinar cells [20]. Using MDS, we show high transcriptional similarity between the corresponding cell types of both datasets (Fig EV1E). The expression information of individual cells and merged expression values for each cell type is available in Dataset EV2.

To rule out technical reasons as a major source of gene expression variability, we identified presumably pure alpha and beta cells among the assessed single cells (Fig EV2A). Their transcription profiles were used to simulate transcriptomes with defined percentages of alpha and beta cell contribution (Fig EV2B). Individual alpha and beta cells were then compared to these virtual transcriptomes to estimate upper limits for potential cross-contamination (Fig EV2C–E). All beta cell transcriptomes were found to be free from any alpha cell contribution, whereas beta cell profiles could explain a small proportion (< 3%) of the variance observed in 8 of the 18 alpha cells studied. However, given that these alpha cells further show higher unexplained variance, it is likely that they are characterized by high inherent variability rather than cross-contamination from beta cells. We conclude that the differences between alpha and beta cell heterogeneity are in line with biological rather than technical effects which supports the hypothesis that alpha cells might be more plastic than beta cells [4].

The heterogeneity within the different cell types was further explored by separate PCAs for each cell type (Appendix Fig S1). Particularly for endocrine cells, heterogeneity was mainly driven by expression differences of marker genes as identified in the initial cell type classification by PCA, suggesting that these cell types are characterized by a spectrum of marker gene expression levels. While this analysis provides evidence for transcriptional heterogeneity, more cells are needed to thoroughly characterize subgroups within the different cell types.

A transcriptome resource to reveal marker genes of human pancreatic cell types

To maximize the utility of our dataset for the identification of cell type-specific expression patterns, we generated a resource of genome browser tracks of all individual cells as well as cumulative tracks for the cell type clusters identified by PCA (http://islet-transcriptome.computational-epigenetics.org/). One interesting use of

Figure 1. Single-cell transcriptomes recapitulate the major pancreatic cell types.

A Workflow for obtaining and analyzing single-cell RNA-seq data from human pancreatic islets.

B Iterative PCA/biplot-based approach for the identification of cell types and cell type-defining transcripts from single-cell RNA-seq data.

C Expression (scaled RPKM values) of cell type-defining genes as identified in (B) across all single cells. Transcripts and single cells are grouped by cell type as identified in (B).

D Display of transcriptional similarity between all single cells by MDS. The coloring scheme is based on the cell types as identified in (B).

E Relative expression (scaled RPKM value) of canonical marker genes for the 6 identified pancreatic islet cell populations represented by bubble size and projected onto the MDS profile.

Published online: December 21, 2015

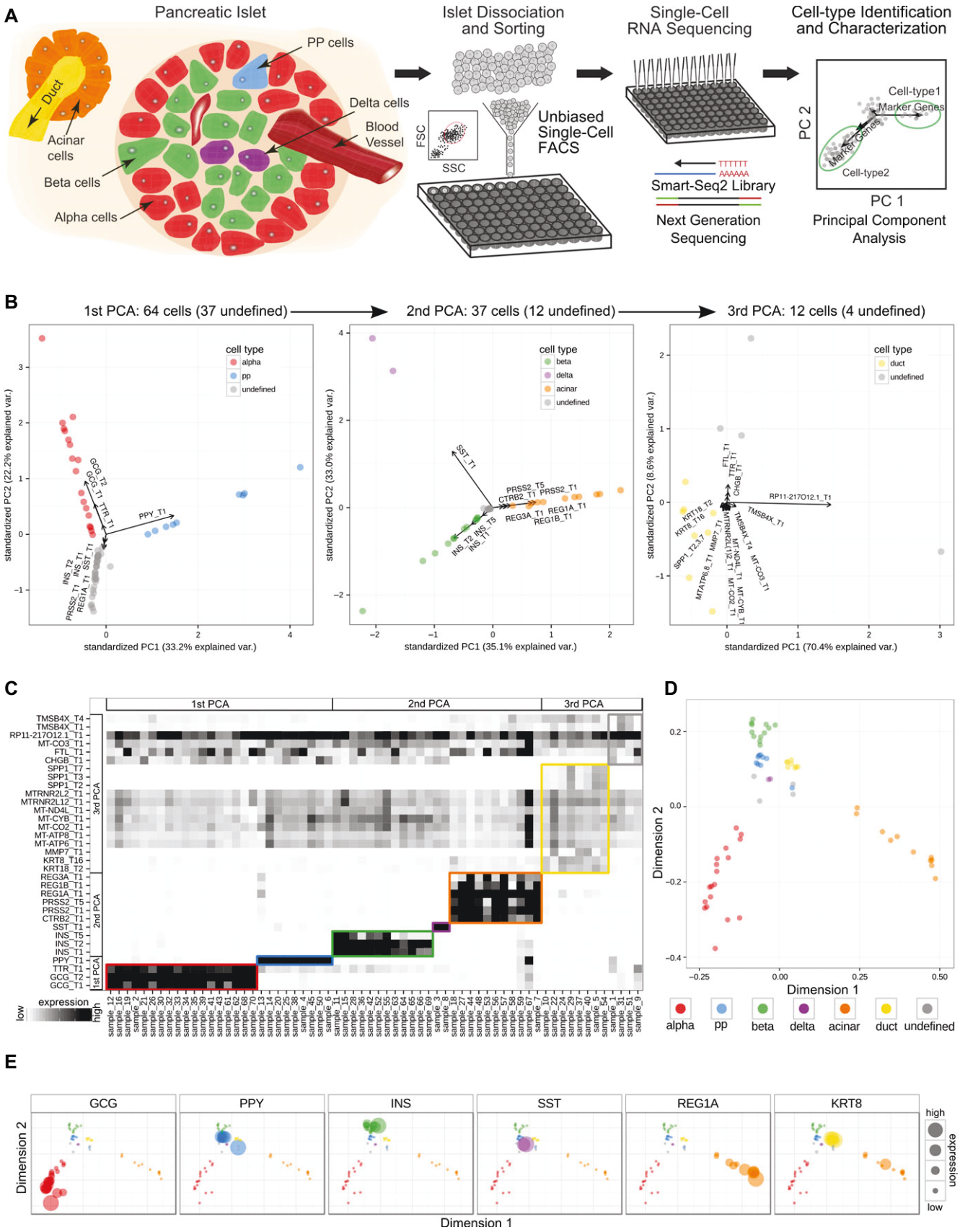


Figure 1.

this resource is the analysis of master regulatory transcription factors, which are fundamental for the development and the maintenance of different pancreatic cell types based on animal models and human genetics. The genome browser tracks illustrate the beta cell-specific expression of *PDX1*, a master regulatory transcription factor directly controlling insulin expression. In contrast, the alpha lineage-defining factor *ARX* is expressed in both alpha and some PP cells (Fig 2A and Appendix Fig S2). Other transcription factors that are important for pancreas development have different degrees of cell type-specific expression in mature human islets, including pan-endocrine (*PAX6*), beta cell-specific (*PAX4*), and duct/delta (*HHEX*) patterns (Fig 2B). While *MAFA* is transcribed in beta cells specifically, we observed robust *MAFB* expression in alpha, beta, and delta cells. Half of the beta cells studied expressed *MAFA* and *MAFB* concomitantly. In addition to these previously described factors, we also observed cell type-specific expression for transcriptional regulators, which have not yet been extensively characterized in the endocrine pancreas. For example, *MORF4L1* shows a similar pan-endocrine pattern to the canonical islet cell marker *NEUROD1* (Fig 2C). A subset of alpha cells express *IRX2* (Fig 2D), some beta cells show high expression of *polycomb ring finger oncogene (BMI1)* (Fig 2E), and PP cells can be characterized by the transcription factors *ETV1* and *MEIS1* (Fig 2F).

We further performed pairwise correlation analysis on transcript level to identify genes, of which the expression profiles correlate highly ($r > 0.9$) with those of the endocrine marker genes *INS*, *GCG*, *SST*, and *PPY* (Fig EV3). While several highly correlated genes could be identified for *INS* and *SST* (e.g., zinc transporter *SLC39A4* and Notch pathway component *DLK1* for *INS* and transcription factors *NKX6-3*, *ZNF430* for *SST*), the expression profiles of *GCG* and *PPY* did not show high correlation with any other genes.

To extend our analysis beyond transcription factors and known marker genes, we performed pairwise comparisons of cell type-specific transcriptomes by gene set enrichment analysis (Dataset EV3). Interestingly, we observed strong enrichment of a gene set containing the REST-binding motif in all endocrine cell types compared to acinar and ductal cells (Fig 3A). Most genes that contain the REST motif in their promoters are expressed in alpha, beta, delta, and PP cells, whereas they are repressed in ductal and acinar cells (Fig 3B). The transcriptional repressor REST targets the REST-binding motif. In line with the target gene expression pattern, *REST* is specifically expressed in ductal and acinar cells (Fig 3C and Appendix Fig S3).

Finally, based on pairwise differential expression analysis between the pancreatic cell types, genes with highly specific expression patterns were identified (Fig EV4 and Appendix Fig S4, Datasets EV4 and EV5). We then used these data to assess islet cell type-specific expression in two areas of high relevance for diabetes research–diabetes risk genes and mouse–human species differences.

Genomewide association studies (GWAS) have identified genomic loci conferring increased risk for the development of diabetes. We examined whether any of the diabetes-related genes predicted by GWAS were specifically expressed in one of the pancreatic islet cell types and genes differentially expressed between the endocrine and exocrine cell types (Fig EV5A). For both type 1 and type 2 diabetes, we identified GWAS genes with beta cell- and endocrine-specific expression. Other genes show broader expression patterns, emphasizing the complexity of functional annotation of diabetes GWAS results. Furthermore, key MODY (Mature Onset of Diabetes in Young) [21] genes *PDX1, PAX4, INS, HNF1A, GCK* are predominantly specific to beta cells (Fig EV5B).

To investigate species-specific differences of alpha and beta cell transcriptomes, we assessed the degree to which the previously identified differentially expressed mouse genes [7,9] are also differentially expressed in human islets and vice versa (Appendix Fig S5). We found that the human alpha cell-specific gene *group-specific component (vitamin D binding protein) GC* and the human beta cell-specific gene *DLK1* (Fig 3D) displayed opposite expression patterns as to what had been reported in mouse islet cells. To confirm the cell type-specific expression of *DLK1* and *GC*, we performed immunofluorescence staining on both human and mouse pancreatic tissue sections. In human islets, *DLK1* was specifically expressed only in insulin-positive cells (Fig 3E), whereas this protein was observed in glucagon-positive cells in mouse tissue (Fig 3F). Similarly, *GC* expression showed alpha cell specificity in human tissues (Fig 3G), whereas it was co-expressed with insulin in mouse tissues (Fig 3H). These results suggest that two of the most differentially expressed cell type-specific marker genes for human alpha and beta cells have opposite expression patterns in mouse islets.

Pancreatic islets comprise different cell types with characteristic transcriptomes, which confounds transcriptome studies that focus on whole pancreatic islets in physiological and pathological conditions. Lineage-labeled transgenic mice have made it possible to obtain transcriptomes for highly pure alpha and beta cell populations in mouse. For human islets, however, cell type-specific enrichment strategies depend on the availability of specific antibodies. Efforts have been made to measure the transcription of individual genes in single human islet cells by qRT–PCR [22], but our dataset is the first to provide genomewide transcriptional information of human islets at single-cell resolution. Using single-cell data, we also for the first time defined the transcriptomes of human delta cells and PP cells, thereby providing reference transcriptomes for all major endocrine cell types in human pancreatic islets.

We illustrated the practical utility of our resource and dataset by three case studies. First, after confirming the cell type specificity of the major transcription factors involved in pancreatic endocrine lineage determination, we identified transcripts encoding transcription factors expressed in islet cells. These include the pan-endocrine marker *MORF4L1*, alpha cell-specific *IRX2*, beta cell-specific *BMI1*, and PP cell-specific *MEIS1* and *ETV1*. These data can provide the basis for future functional studies in the roles of these transcription factors in the pancreas and in diabetes.

In a second example, we analyzed cell type-specific enrichment of previously characterized gene sets. The specific expression of REST-motif-containing genes in the endocrine cell types led us to identify the specific expression of the transcriptional repressor *REST* in the exocrine pancreas. REST recruits a large complex of chromatin regulators, including many factors that allow pharmacological modulation like histone deacetylases and the histone demethylase LSD1. REST repression in non-endocrine cells activates the promoters of important beta cell transcription factors, including PAX4 and PDX1 and is a key step in reprogramming to insulin-producing cells [23–26]. Future studies will show whether REST is critical in restricting ductal differentiation potential and may be a target for inducing beta cell neogenesis from duct cells.

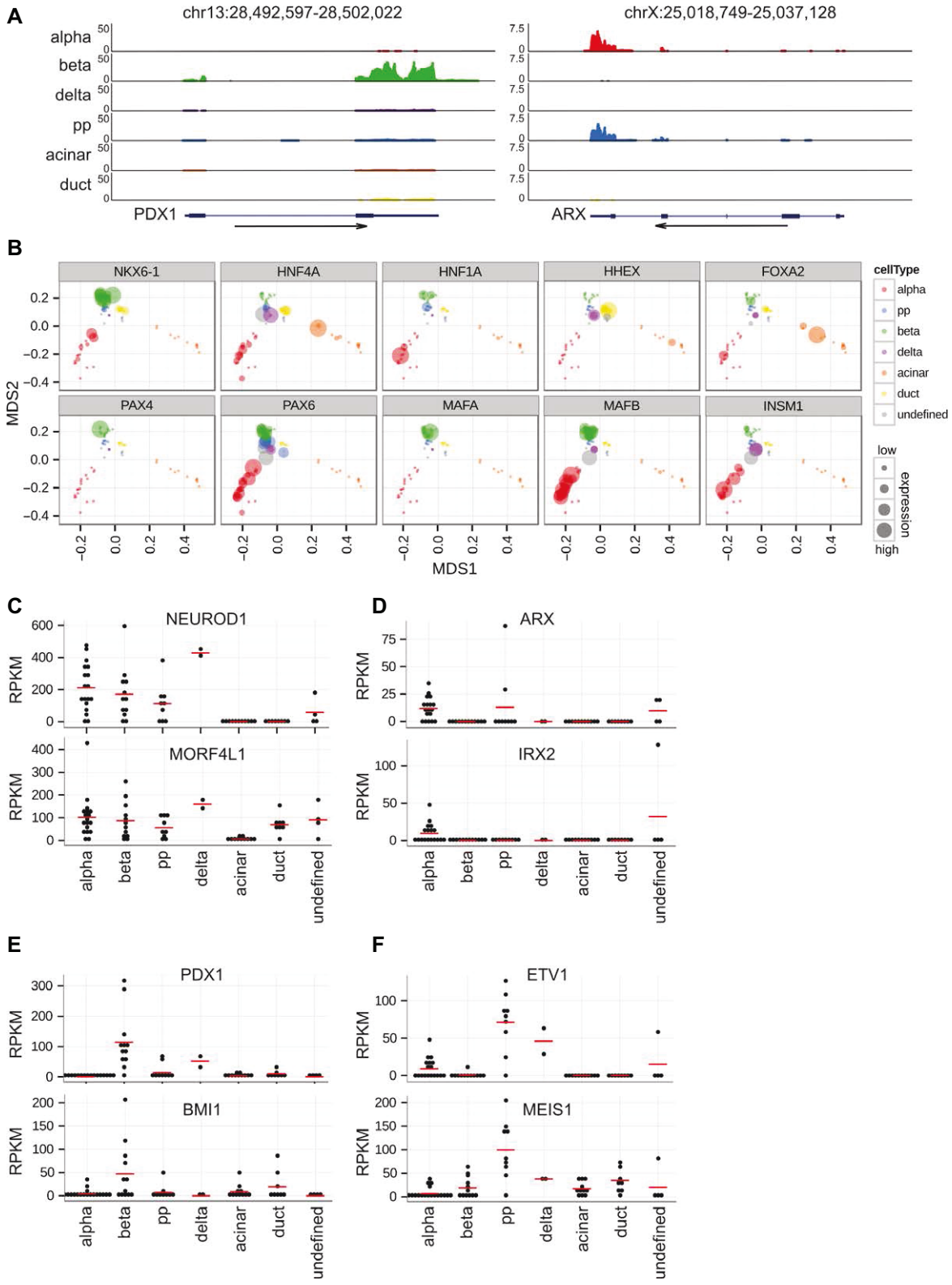


Figure 2. Expression of cell type-specific transcription factors at single-cell resolution.

A Merged UCSC Genome Browser tracks for the *PDX1* and *ARX* loci. The respective tracks for all single cells are presented in Appendix Fig S2.

B Relative expression (scaled RPKM value) of important transcription factors represented by bubble size and projected onto the MDS profile.

C–F Cell type-specific expression of pan-endocrine (C), alpha cell (D), beta cell (E), and PP cell (F) transcription factors (red bar: mean expression). The statistical significance of the differential gene expression is presented in Appendix Fig S6.

Published online: December 21, 2015

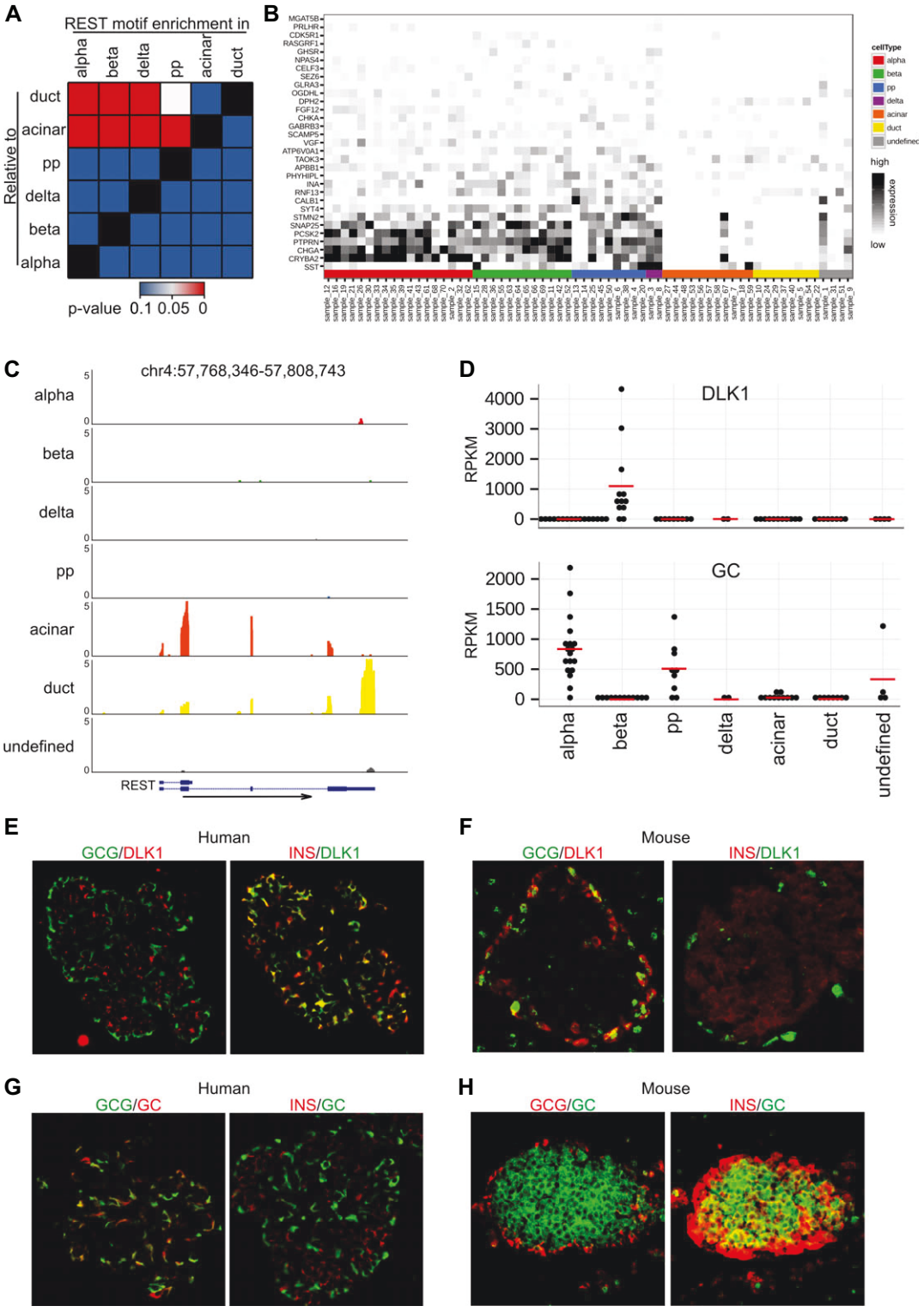


Figure 3. Single-cell transcriptomes reveal unique features of human islets.

A Heatmap displaying the *P*-values obtained by pairwise Gene Set Enrichment Analysis (GSEA) for the REST-binding motif.

B Relative expression (scaled RPKM value) of genes contained in the REST-binding motif gene set.

C Merged UCSC Genome Browser tracks for *REST*. The respective tracks for all single cells are presented in Appendix Fig S3.

D Expression of DLK1 and GC in human islet cell types (red bar: mean expression). The statistical significance of the differential gene expression is presented in Appendix Fig S6.

E–H Co-staining of *DLK1* (E, F) or *GC* (G, H) with insulin and with glucagon in representative human (E, G) and mouse (F, H) islets.

Finally, in a third example, we focused on differences between mouse and human islets. Previous studies have noticed such differences regarding the overall architecture and specific physiological properties [7,27]. Our human islet single-cell transcriptomes confirm that the expression of hormones and canonical transcription factors is conserved between human and mouse. However, two genes—*GC* and *DLK1*—that are among the most characteristic for human alpha and beta cells, respectively, are expressed in opposite patterns in the mouse. Both *DLK1* and *GC* are relevant to diabetes [5,28], and further research is necessary to dissect their roles in both human and mouse islet biology.

These examples highlight the utility of the current single-cell transcriptome database for islet biology. In addition, we expect future growth of our resource with the addition of single-cell expression data from diabetic donors and from islets treated with drugs and metabolites *ex vivo*, contributing to the utility of the presented resource for studies on all aspects of human islet biology. In summary, our study establishes a transcriptional dataset for all the cell types in human pancreatic islets with single-cell resolution and defines distinctly human features in the patterns of alpha and beta cell-expressed genes.

Materials and Methods

Reagents

Antibodies used in this project are directed against insulin (Sigma I8510), glucagon (Abcam ab92517), DLK1 (R&D MAB1144-100), and GC (Abcam ab81307). The sequences of primers for actin have been published recently [29]. All the fluorescently labeled secondary antibodies were purchased from Life Technologies Corporation. The reagents used for the Smart-seq2 protocol for cDNA synthesis, amplification, and sequencing library preparation have been published recently [17].

Cell culture

Human islets were provided through the JDRF award 31-2012-783 (ECIT: Islet for Research program). They were from a 37-year-old male donor whose BMI was 22. Islets were cultured in CMRL medium (Life Technologies) supplemented with 10% FBS, 2 mM glutamine, 100 U/ml penicillin, and 100 μg/ml streptomycin. Islets were collected following overnight culture after receiving them. To disassociate islets into single cells, islets were incubated in Accutase (Life Technologies) in 37°C for 20 min, neutralized by CMRL medium. Purification of single cells was performed by flow cytometry cell sorting on a Moflo AstriosEQ (Beckman Coulter, Miami) as previously described in [16].

Immunofluorescence

The human pancreatic histology slides were ordered from Abcam (ab4611). The mouse pancreatic histology slides from 129SV mice were gifts from Patrick Collombat. The staining followed a published protocol [30]. Briefly, the paraffin was removed from the tissues. Afterwards, rehydration and antigen retrieval was performed. The tissues were blocked by 3% BSA for half an hour and incubated overnight at 4°C with primary antibodies in 1:1,000 dilutions. After washing with PBST, tissues were incubated with secondary antibodies and Hoechst 33342 for half an hour. Finally, the slides were mounted and sealed with nail polish and images were taken with Leica CRT6000.

Single-cell RNA-seq sample and sequencing library preparation

cDNA synthesis and enrichment were performed following the Smart-seq2 protocol as described Picelli *et al* [17]. ERCC spike-in RNA (Ambion) was added to the lysis buffer in a dilution of 1:1,000,000. Library preparation was conducted on 1 ng of cDNA using the Nextera XT library preparation kit (Illumina) as described Picelli *et al* [17]. Sequencing was performed by the Biomedical Sequencing Facility at CeMM using the 50 bp single-read setup on the Illumina HiSeq 2000/2500 platform.

qRT–PCR

After the cDNA was synthesized and amplified from single cells, quantitative PCR was performed with Power SYBR Green PCR Master Mix (Applied Biosystems) on a LightCcycler 480 qPCR instrument (Roche).

Single-cell RNA-seq data processing

The raw sequencing data were processed using a custom bioinformatics pipeline which consists of the following main steps: (i) trimming of contaminating sequencing adapter sequences, (ii) alignment of the trimmed reads to the human transcriptome as well as genome, (iii) calculation of expression estimates for each transcript, differential expression analysis and visualization as genome browser tracks.

Trimming of adapter sequences was performed with trimmomatic (v 0.32). Only reads with a minimum length of 25 bp after adapter trimming were included in the downstream analysis. Alignment of the trimmed reads to the human transcriptome (hg19 GRCh37 ftp://ftp.ensembl.org/pub/release-74/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh37.74.cdna.all.fa.gz) was performed with bowtie1 (v 1.1) [31] recording up to 100 different mapping positions for each read which takes into account that one read might originate from any of the different transcripts of one gene. Alignment to the human genome (hg19/GRCh37) was performed using Tophat (v 2.0.13) [32]. These genomic alignments were purely performed for the purpose of visualization in genome browser tracks. Conversion of the alignment files to the files needed to display the data as genome browser tracks (bigWig) was performed with RSeQC (v 2.3.9) bam2wig.py followed by UCSC tools' wigToBigWig. Calculation of normalized transcript-wise expression estimates (rpkm values) as well as differential expression analysis was performed based on the transcriptome alignments using the R (v 3.1.2) package BitSeq (v 1.10.0) [18]. In order to correct for potential biases in the read distribution, the BitSeq function getExpression() was run with the "uniform" option disabled.

Quality filtering

The minimal number of reads needed to obtain reliable RPKM values as estimates of gene expression was determined by taking

advantage of a synthetic RNA mix consisting of 92 RNAs covering a 10^6-fold concentration range (ERCC spike-in controls) that had been carried along through the entire library preparation and sequencing process with each single cell. Starting from ~25 reads per transcript, we observed the expected linear relationship between ERCC transcript abundance and measured RPKM values (Fig EV1B). For the purpose of noise reduction, we defined transcripts covered by less than 25 reads as "not expressed" and set their RPKM values to a minimal value. Furthermore, 6 samples showed less than 500 (arbitrary cutoff) reliably covered transcripts and were excluded from the analysis (Fig EV1A).

Grouping the single cells based on their gene expression profiles

In order to determine groups of cells with similar expression profiles and at the same time identify the primary defining genes for each group, we performed a stepwise principal component analysis (PCA) based on the quality-filtered expression values. PCA was performed using the function prcomp() in R. The results were displayed as a biplots showing samples (cells) as dots and the most highly loaded variables (transcripts) as vectors. Biplots were constructed using a slightly modified version of the R function ggbiplot() (https://github.com/vqv/ggbiplot).

External data

External RNA-seq raw data (next-generation sequencing reads) for bulk samples of human acinar cells, beta cells, and islet cells were obtained from ArrayExpress (E-MTAB-1294: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1294/samples/) [20]. We used the samples HI10 (islet), HI25 (islet), HI32 (islet), HIE1 (beta cells), HIE2 (beta cells), and acinar tissue donor (acinar cells). External data were processed using the same pipeline as the single-cell data. For the comparison of external and single-cell as well as 500 cell data by multidimensional scaling, batch effect correction was performed using the function ComBat() of the R package sva.

Defining cell type-specific gene expression profiles

Cell type-specific gene expression profiles were defined by performing pairwise differential expression analysis between all previously defined groups of cells. Differential expression analysis was performed using the function estimateDE() of the R package BitSeq. For each cell type in each comparison, the specificity of the expression of each transcript was deduced under consideration of effect size (absolute difference and log2 fold change) as well as statistical significance (probability of positive log ratio, PPLR) of the measured differential expression. Technically, for each comparison, all transcripts were ranked by absolute difference in gene expression, log fold change of gene expression, and probability of positive log ratio and a combined rank for each transcripts was produced by selecting the worst (i.e., highest) of these three ranks as a representative rank. Finally, the representative ranks from all comparisons for each cell type were again combined by selecting the worst rank for each transcript (Appendix Fig S7). Therefore, the lower the combined rank, the more specific the expression of the respective transcript for the assessed cell type. To identify the cell type for which the expression of a given gene is most specific, we compared the assigned combined ranks between all cell types and selected the cell type that showed the lowest combined rank for this gene.

Assessing cross-contamination between cell types

We assessed potential cross-contamination between two cell types using a four-step approach: (i) selection of cell type-specific genes (profile genes), (ii) selection of the purest single cells for each cell type (profile cells), (iii) calculation of pure and increasingly contaminated gene expression profiles *in silico* (mix profiles), and (iv) identification of the mix profiles that best match the expression profile of each single cell.

As profile genes, we selected all genes among the top 500 cell type-specific genes for each of the two cell types that showed an absolute mean expression difference of greater than 0.5 and a relative mean expression difference of at least twofold. This selection resulted in 233 profile genes for alpha cells and 252 profile genes for beta cells.

To identify the purest cells of each cell type, we calculated a weighted mean of scaled expression values (sample-wise, scale 0 to 1; lower percentile: 0.05, upper percentile: 0.95) for both groups of profile genes for each single cell (profile scores). We used a rank-based weighting system in order to give more power to more cell type-specific profile genes. All single cells were then plotted according to their profile scores, and per cell type, the three cells with the most cell type-specific profile scores (highest distance to the diagonal) were selected as profile cells (Fig EV2A).

Pure expression profiles consisting of both groups of profile genes were calculated as the mean expression values of the three profile cells. We then used these two cell type-specific profiles to artificially construct expression profiles that represented different degrees of contamination by computationally mixing the two profiles in different ratios. Specifically, we calculated weighted means of the two pure expression values for each profile gene, with the weight increasing from 0 to 100 in steps of 1 for one of the pure profiles and at the same time decreasing from 100 to 0 for the other pure profile. This resulted in 100 profiles, two pure (cell type specific) and 98 mixed profiles (Fig EV2B).

We then calculated the Pearson correlation of each of the artificial 100 profiles with the actual expression profiles of each of the single cells (Fig EV2C) and selected the highest correlating mix profile for each single cell. These selected mix profiles represent the fraction of variance in profile gene expression that is explained by either of the two cell type-specific profiles as well as the fraction of variance that remains unexplained (Fig EV2D and E).

Gene set enrichment analysis

Binding motif analysis was done with Gene Set Enrichment Analysis (GSEA) [33,34]. For each single cell, the most highly expressed transcript was selected as representative for the respective gene. Finally, gene expression values for each cell type were found by calculating the median across all cells of a particular cell type. These median expression values were used as input for GSEA. Genes that were not found to be expressed in any of the cell types were removed from the input dataset. Pairwise comparisons were done among all six

assigned cell types except the "undefined" amounting to 30 comparisons in total. The REST-binding motif was significantly enriched ($P < 0.05$, FDR < 25%) in all of the comparisons between endocrine cell types and exocrine cell types.

GWAS analysis

GWAS results relevant for diabetes (search for "diabetes") were downloaded from the GWAS catalog (https://www.ebi.ac.uk/gwas/). We categorized the reported traits into type 1 and type 2 diabetes according to whether "1" or "2" appeared in the trait description. Each gene that was identified as significant in a GWAS (reported gene) was assigned to the cell type for which it was identified as most specific (see "Defining cell type-specific gene expression profiles"). Because in this analysis specificity among the endocrine cells (alpha cells, beta cells, delta cells, PP cells) and among the exocrine cells (acinar cells, duct cells) was not paramount, cell type specificity was determined only in comparison with cell types of the other group. This approach was chosen in order to not dismiss genes as unspecific if they are endocrine or exocrine specific but not necessarily cell type specific. The eight MODY genes were taken from [21].

Data deposition

Sequencing datasets described in this work have been deposited in the Gene Expression Omnibus (GEO) repository under accession number GSE73727.

Expanded View for this article is available online.

Acknowledgements
We would like to thank the Core Facility Flow Cytometry at the Medical University of Vienna for their expertise and assistance with FACS sorting and the Biomedical Sequencing Facility at CeMM for next-generation sequencing and single-cell technology development. We thank Patrick Collombat (INSERM, Nice) for providing mouse pancreatic tissue slides. This work was funded and supported by JDRF grants 3-SRA-2015-20-Q-R and 17-2011-258 (Generation of beta cells from alternative pancreatic subtypes). Human islets were provided through the JDRF awards 31-2012-783 and 1-RSC-2014-100-I-X (ECIT: Islet for Research program). Research in the Kubicek laboratory is supported by the Austrian Federal Ministry of Science, Research and Economy, the National Foundation for Research, Technology, and Development, and the Marie Curie Career Integration Grant EPICAL. The single-cell sequencing infrastructure at CeMM was supported by a New Frontiers Research Infrastructure grant from the Austrian Academy of Sciences. J.K. is Recipient of a DOC Fellowship of the Austrian Academy of Sciences.

Author contributions
SK, CBo, MF, JL and TP conceived and designed the study; CBa and EB provided human islets; JL, AS and MF performed the experiments; TP and CBo generated next-generation sequencing data; JK processed the raw data; JK and JL performed the bioinformatic analysis; SK, CBo, MF, JL and JK wrote the manuscript with contributions from all co-authors.

Conflict of interest
The authors declare that they have no conflict of interest.

References

1. Collombat P, Xu X, Ravassard P, Sosa-Pineda B, Dussaud S, Billestrup N, Madsen OD, Serup P, Heimberg H, Mansouri A (2009) The ectopic expression of Pax4 in the mouse pancreas converts progenitor cells into alpha and subsequently beta cells. *Cell* 138: 449 – 462
2. Chera S, Baronnier D, Ghila L, Cigliola V, Jensen JN, Gu G, Furuyama K, Thorel F, Gribble FM, Reimann F *et al* (2014) Diabetes recovery by age-dependent conversion of pancreatic delta-cells into insulin producers. *Nature* 514: 503 – 507
3. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, Storm P, Osmark P, Ladenvall C, Prasad RB *et al* (2014) Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci USA* 111: 13924 – 13929
4. Bramswig NC, Everett LJ, Schug J, Dorrell C, Liu C, Luo Y, Streeter PR, Naji A, Grompe M, Kaestner KH (2013) Epigenomic plasticity enables human pancreatic alpha to beta cell reprogramming. *J Clin Invest* 123: 1275 – 1284
5. Kameswaran V, Bramswig NC, McKenna LB, Penn M, Schug J, Hand NJ, Chen Y, Choi I, Vourekas A, Won KJ *et al* (2014) Epigenetic regulation of the DLK1-MEG3 microRNA cluster in human type 2 diabetic islets. *Cell Metab* 19: 135 – 145
6. Hrvatin S, Deng F, O'Donnell CW, Gifford DK, Melton DA (2014) MARIS: method for analyzing RNA following intracellular sorting. *PLoS One* 9: e89459
7. Benner C, van der Meulen T, Caceres E, Tigyi K, Donaldson CJ, Huising MO (2014) The transcriptional landscape of mouse beta cells compared to human beta cells reveals notable species differences in long non-coding RNA and protein-coding gene expression. *BMC Genom* 15: 620
8. Benitez CM, Qu K, Sugiyama T, Pauerstein PT, Liu Y, Tsai J, Gu X, Ghodasara A, Arda HE, Zhang J *et al* (2014) An integrated cell purification and genomics strategy reveals multiple regulators of pancreas development. *PLoS Genet* 10: e1004645
9. Kubicek S, Gilbert JC, Fomina-Yadlin D, Gitlin AD, Yuan Y, Wagner FF, Holson EB, Luo T, Lewis TA, Taylor B *et al* (2012) Chromatin-targeting small molecules cause class-specific transcriptional changes in pancreatic endocrine cells. *Proc Natl Acad Sci USA* 109: 5364 – 5369
10. Scharfmann R, Pechberty S, Hazhouz Y, von Bulow M, Bricout-Neveu E, Grenier-Godard M, Guez F, Rachdi L, Lohmann M, Czernichow P *et al* (2014) Development of a conditionally immortalized human pancreatic beta cell line. *J Clin Invest* 124: 2087 – 2098
11. Dorrell C, Schug J, Lin CF, Canaday PS, Fox AJ, Smirnova O, Bonnah R, Streeter PR, Stoeckert CJ Jr, Kaestner KH *et al* (2011) Transcriptomes of the major human pancreatic cell types. *Diabetologia* 54: 2832 – 2844
12. Blodgett DM, Nowosielska A, Afik S, Pechhold S, Cura AJ, Kennedy NJ, Kim S, Kucukural A, Davis RJ, Kent SC *et al* (2015) Novel observations from next-generation RNA sequencing of highly purified human adult and fetal islet cell subsets. *Diabetes* 64: 3172 – 3181
13. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509: 371 – 375
14. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A *et al* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343: 776 – 779

15. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C *et al* (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347: 1138 – 1142
16. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schonegger A, Klughammer J, Bock C (2015) Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* 10: 1386 – 1397
17. Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 9: 171 – 181
18. Glaus P, Honkela A, Rattray M (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28: 1721 – 1728
19. Chiaromonte F, Miller W, Bouhassira EE (2003) Gene length and proximity to neighbors affect genome-wide expression levels. *Genome Res* 13: 2602 – 2608
20. Moran I, Akerman I, van de Bunt M, Xie R, Benazra M, Nammo T, Arnes L, Nakic N, Garcia-Hurtado J, Rodriguez-Segui S *et al* (2012) Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab* 16: 435 – 448
21. Gardner DS, Tai ES (2012) Clinical features and treatment of maturity onset diabetes of the young (MODY). *Diabetes Metab Syndr Obes* 5: 101 – 108
22. Bengtsson M, Stahlberg A, Rorsman P, Kubista M (2005) Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* 15: 1388 – 1392
23. Bruce AW, Donaldson IJ, Wood IC, Yerbury SA, Sadowski MI, Chapman M, Gottgens B, Buckley NJ (2004) Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc Natl Acad Sci USA* 101: 10458 – 10463
24. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316: 1497 – 1502
25. Li HT, Jiang FX, Shi P, Zhang T, Liu XY, Lin XW, Pang XN (2012) In vitro reprogramming of rat bone marrow-derived mesenchymal stem cells into insulin-producing cells by genetically manipulating negative and positive regulators. *Biochem Biophys Res Commun* 420: 793 – 798
26. Li B, Wang S, Liu H, Liu D, Zhang J, Zhang B, Yao H, Lv Y, Wang R, Chen L *et al* (2011) Neuronal restrictive silencing factor silencing induces human amniotic fluid-derived stem cells differentiation into insulin-producing cells. *Stem Cells Dev* 20: 1223 – 1231
27. Bosco D, Armanet M, Morel P, Niclauss N, Sgroi A, Muller YD, Giovannoni L, Parnaud G, Berney T (2010) Unique arrangement of alpha- and beta-cells in human islets of Langerhans. *Diabetes* 59: 1202 – 1210
28. Wang G, Li Y, Li L, Yu F, Cui L, Ba Y, Li W, Wang C (2014) Association of the vitamin D binding protein polymorphisms with the risk of type 2 diabetes mellitus: a meta-analysis. *BMJ Open* 4: e005617
29. Fomina-Yadlin D, Kubicek S, Walpita D, Dancik V, Hecksher-Sorensen J, Bittker JA, Sharifnia T, Shamji A, Clemons PA, Wagner BK *et al* (2010) Small-molecule inducers of insulin expression in pancreatic alpha-cells. *Proc Natl Acad Sci USA* 107: 15099 – 15104
30. Li B, Tsao SW, Li YY, Wang X, Ling MT, Wong YC, He QY, Cheung AL (2009) Id-1 promotes tumorigenicity and metastasis of human esophageal cancer cells through activation of PI3K/AKT signaling pathway. *Int J Cancer* 125: 2576 – 2585
31. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25
32. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14: R36
33. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545 – 15550
34. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34: 267 – 273

Expanded View Figures

Figure EV1. Statistical analysis of single-cell RNA-seq data.

A Number of detected transcripts and total aligned reads for each single cell. The red line denotes 500 transcripts, below which samples were excluded from the analysis.

B Scatter plots displaying the correlation between the number of input ERCC RNA molecules and measured RPKM values in four representative single cells. The Pearson correlation (*r*) is noted in the upper left corner.

C Scatter plots correlating raw read counts (left) and RPKM normalized expression values (right) with transcript length. All adequately covered transcripts (> 25 reads) of all 64 samples were included in this analysis. The observed negative correlation after RPKM normalization lies in the range of what had been reported previously [19].

D Raw counts and RPKM normalized expression values for 10 groups of ERCC spike-in controls. The amount of molecules spiked into the sequencing reaction is constant within one group, whereas the length of the transcripts varies considerably. Within each group, equal expression values across different transcript lengths thereby confirm that RPKM normalization does not systematically penalize longer transcripts in the assessed ERCC transcript length range of ~200 to 2,000 bp. ERCC spike-in controls covered by more than 25 reads are indicated in blue and those with ≤ 25 reads in red.

E MDS displaying transcriptional similarity between corresponding cell types of a published dataset [20] (prefix: ext) and the current dataset.

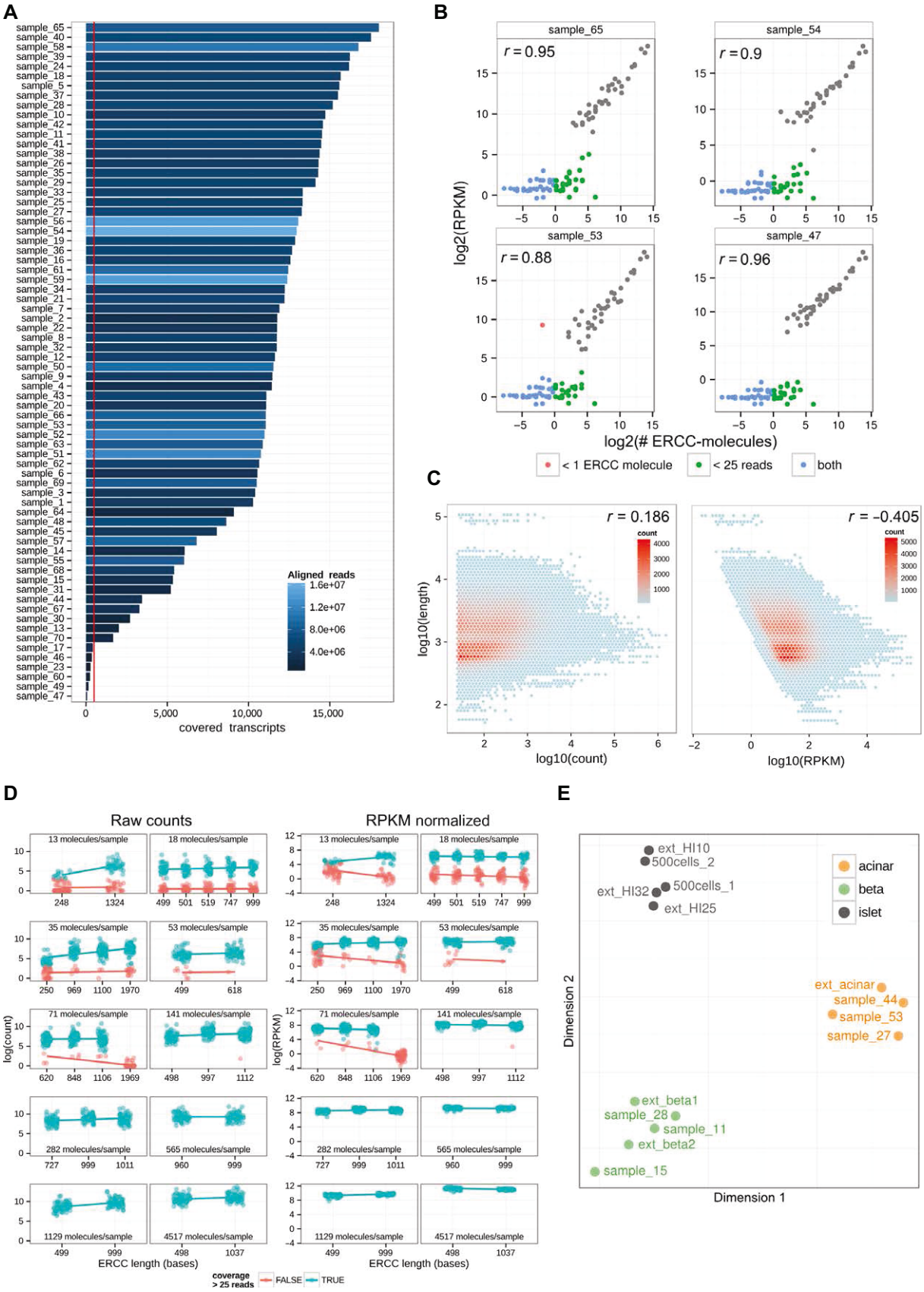


Figure EV1.

Figure EV2. Assessing cross-contamination between alpha and beta cells.

A Scatter plot displaying single alpha and beta cells, 500-cell islet samples, as well as bulk islet and beta cell samples from published datasets according to their weighted mean of scaled expression values in alpha and beta cell-specific profile genes. The three selected profile cells for each cell type are indicated by their sample ID.

B Pure and mixed expression profiles consisting of 233 alpha cell-specific genes and 252 beta cell-specific genes. Alpha and beta cell-specific profiles are calculated based on the expression values of the three selected profile cells only, while profile genes were selected based on all single cells classified as alpha or beta cells, which is why the expression gradients in the mix profiles do not always follow the same direction.

C Profile correlation curves for each individual sample. The maximum of each curve defines the maximum variance that can be explained (*y*-axis) by the corresponding mix profile (*x*-axis) providing a measure for the composition of the respective sample.

D Diagram explaining the transition from profile correlation curves to sample composition estimates. The profile composition that explains most variance is linearly scaled to the maximum variance explained.

E Sample composition estimates for each assessed sample. The differences between the 500-cell islet samples and bulk samples might be explained by technical effects that enrich for alpha cells during islet cultivation, disassociation, and FACS purification.

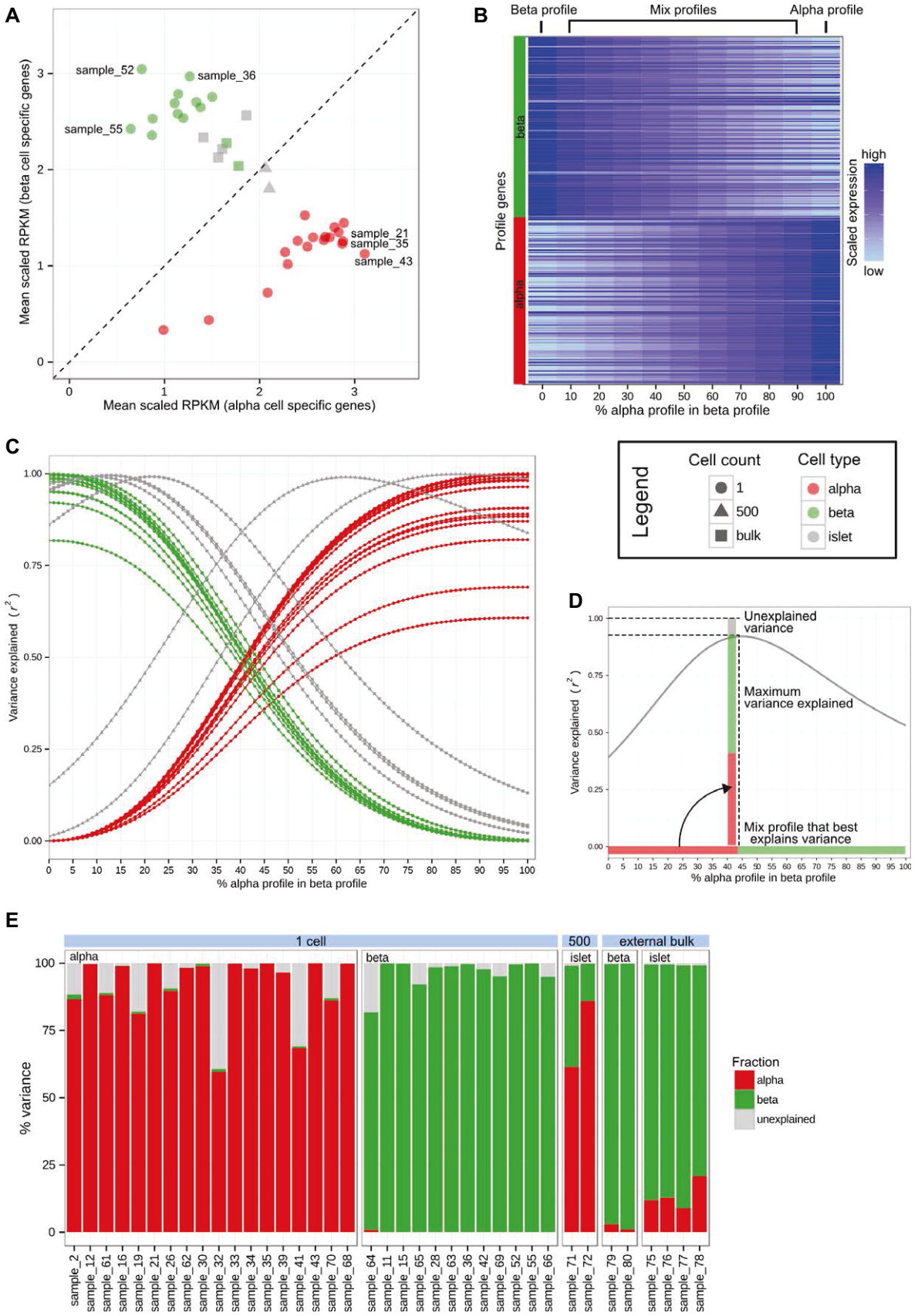


Figure EV2.

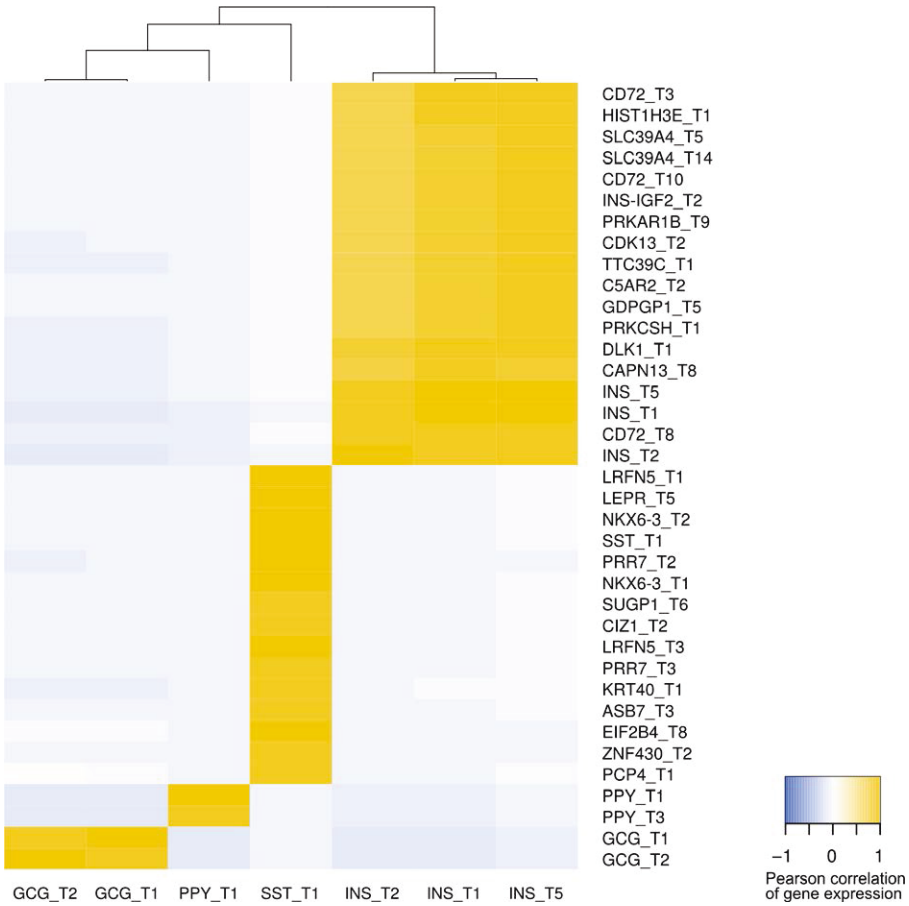


Figure EV3. Assessing pairwise correlation of endocrine marker genes.
Correlation matrix displaying all genes (*y*-axis) that are highly correlated ($r > 0.9$) with at least one of the endocrine marker genes (*x*-axis). Different transcripts of the same gene are indicated by "Tx".

Figure EV4. Specific expression of selected marker genes.
Relative expression (scaled RPKM value) of interesting genes across all single cells represented by bubble size and projected onto the MDS profile as displayed in Fig 1D.

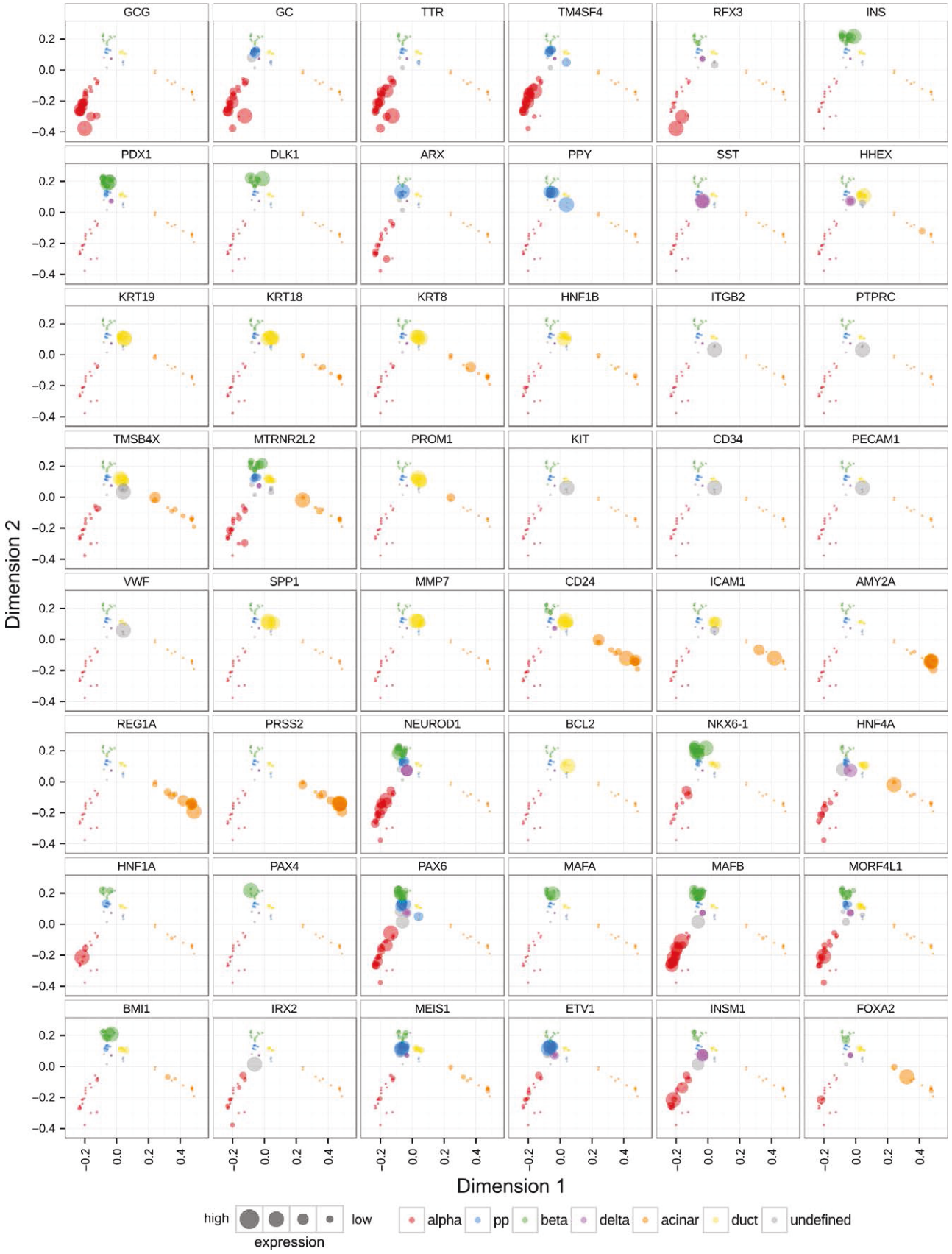


Figure EV4.

Figure EV5. Assessing cell type specificity of genes identified in diabetes-related GWAS.

A Cell type specificity for genes reported in diabetes-related GWAS. Each gene reported in a diabetes-related GWAS (search for "Diabetes" on GWAS Catalog) was assigned to the pancreatic cell type in which it was found to be most specifically expressed. Ranking was performed as described in Appendix Fig S7B and Dataset EV6.

B Heatmap showing mean expression values for the most cell type-specific diabetes-associated GWAS genes in the different here identified human islet cell types. Specifically, only genes with a specificity rank lower than 500 (dashed line in panel A) are listed, and genes with equal expression in multiple cell types are not shown. The numbers in the colored boxes indicate the number of studies in which the respective gene has been reported. Heatmaps are colored by mean ln(RPKM), the mean of the natural logarithm of the RPKM values across all cells of the respective cell type.

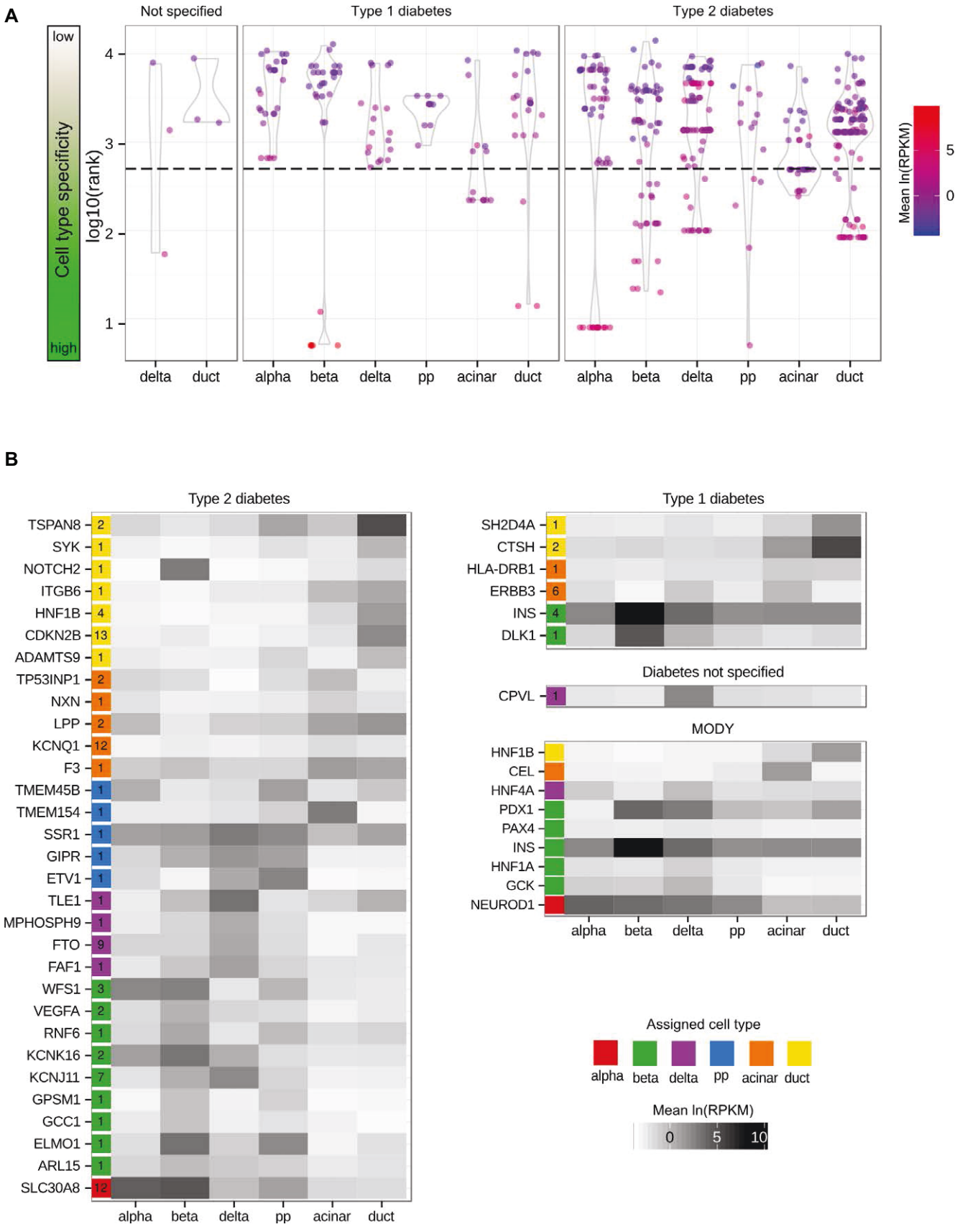


Figure EV5.

Appendix to "Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types"

Jin Li[1,7], Johanna Klughammer[1,7], Matthias Farlik[1,7], Thomas Penz[1,7], Andreas Spittler[2], Charlotte Barbieux[3], Ekaterine Berishvili[3], Christoph Bock[1,4,5,*], Stefan Kubicek[1,6,*]

[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, 1090 Vienna, Austria

[2]Medical University of Vienna, Anna Spiegel Forschungsgebäude, Lazarettgasse 14, 1090 Vienna, Austria

[3]Cell Isolation and Transplantation Center, Department of Surgery, Geneva University Hospitals and University of Geneva, Geneva, Switzerland

[4]Department of Laboratory Medicine, Medical University of Vienna, 1090 Vienna, Austria

[5]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

[6]Christian Doppler Laboratory for Chemical Epigenetics and Antiinfectives, CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria.

[7]These authors contributed equally

*To whom correspondence should be addressed:

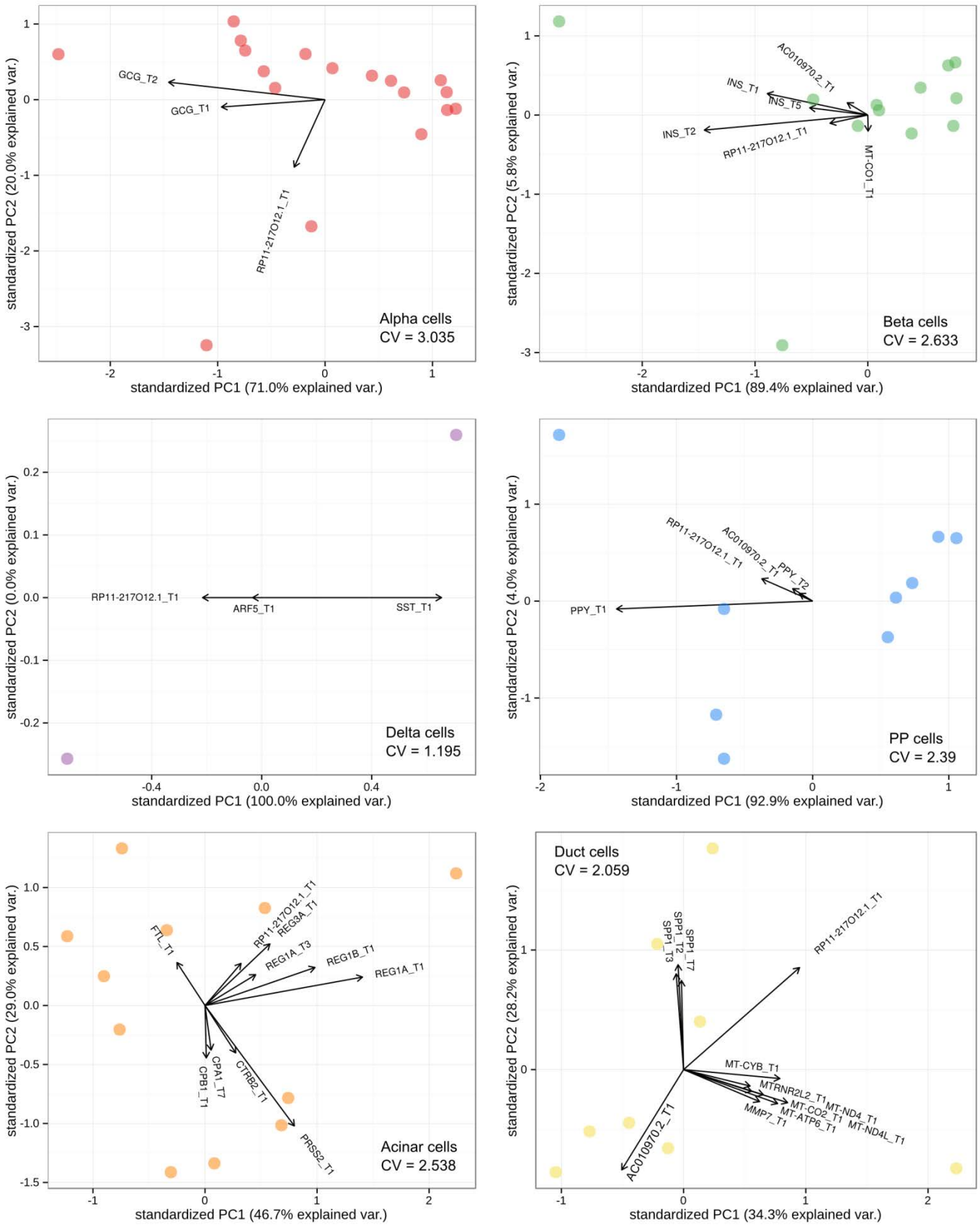
Christoph Bock
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences.
Lazarettgasse 14, 1090 Vienna, Austria
Phone : +43-1-40160-70070
Fax: +43-1-40160-970 000
Email: cbock@cemm.oeaw.ac.at

OR

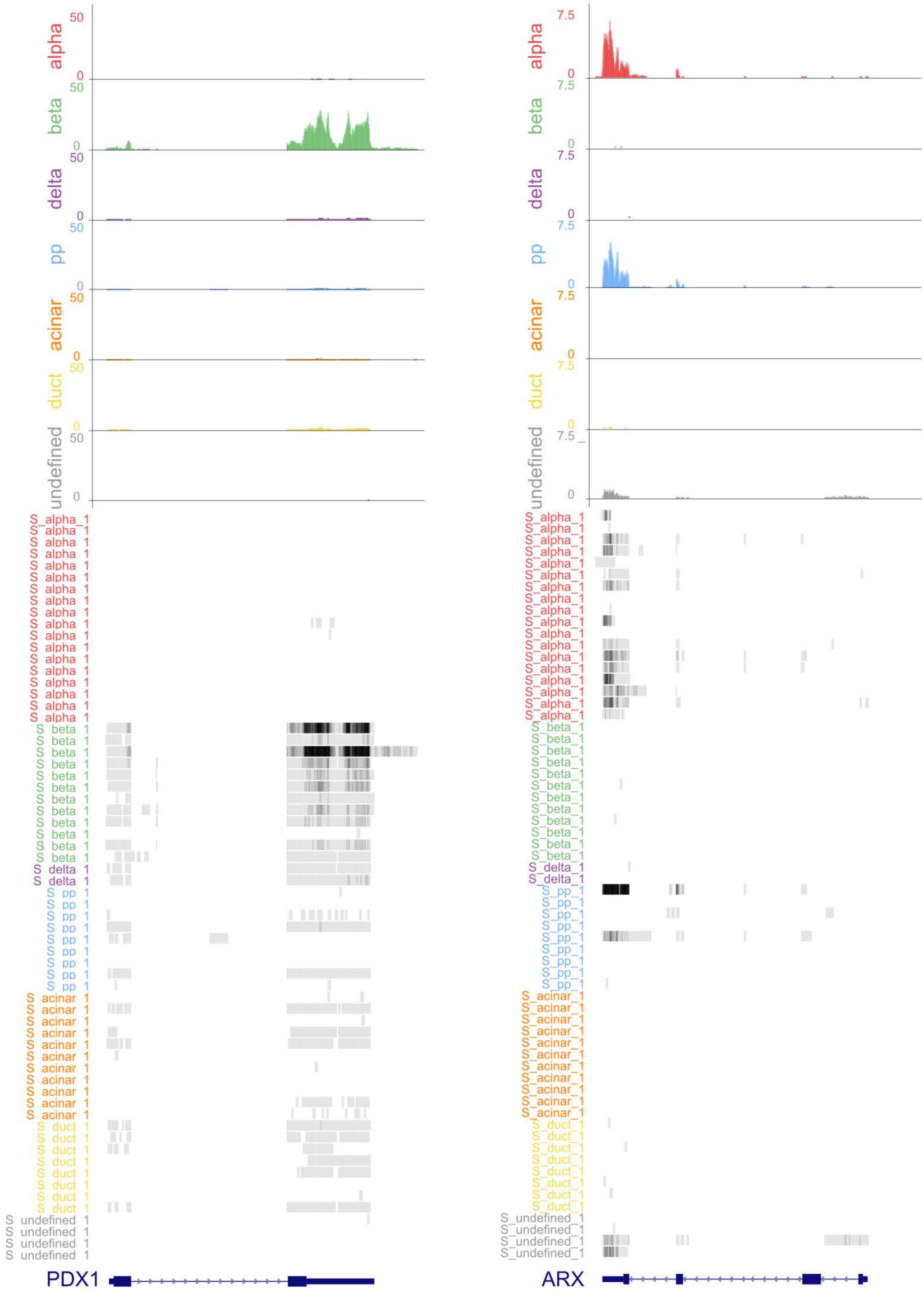
Stefan Kubicek
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences.
Lazarettgasse 14, 1090 Vienna, Austria
Phone : +43-1-40160-70036
Fax: +43-1-40160-970 000
Email: skubicek@cemm.oeaw.ac.at

Key words: Single-cell RNA-seq / human pancreatic islets / beta cells / alpha cells / marker genes

Table of contents

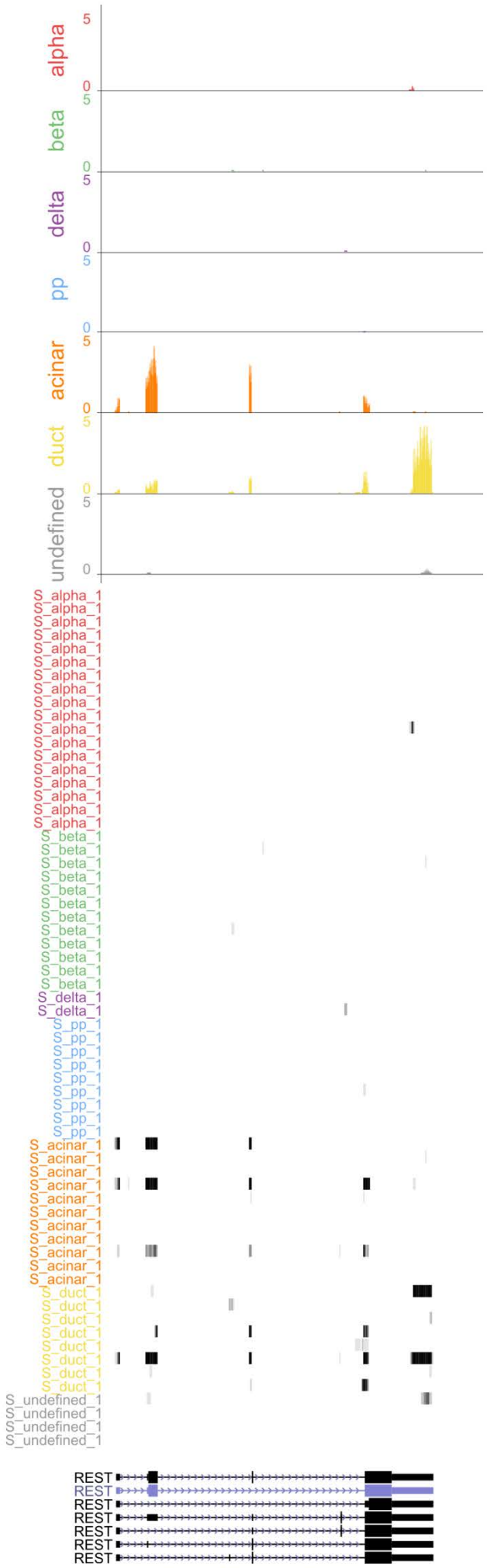


Appendix Figure S1: Assessment of heterogeneity within each identified human islet cell type
Separate biplots for each of the six identified human islet cell types show considerable variability within cell types. Only the most loaded eigenvectors (transcripts) are displayed. The coefficient of variance (CV) was calculated per cell type based on all transcripts that were expressed in at least one of the single-cells of the respective cell type.

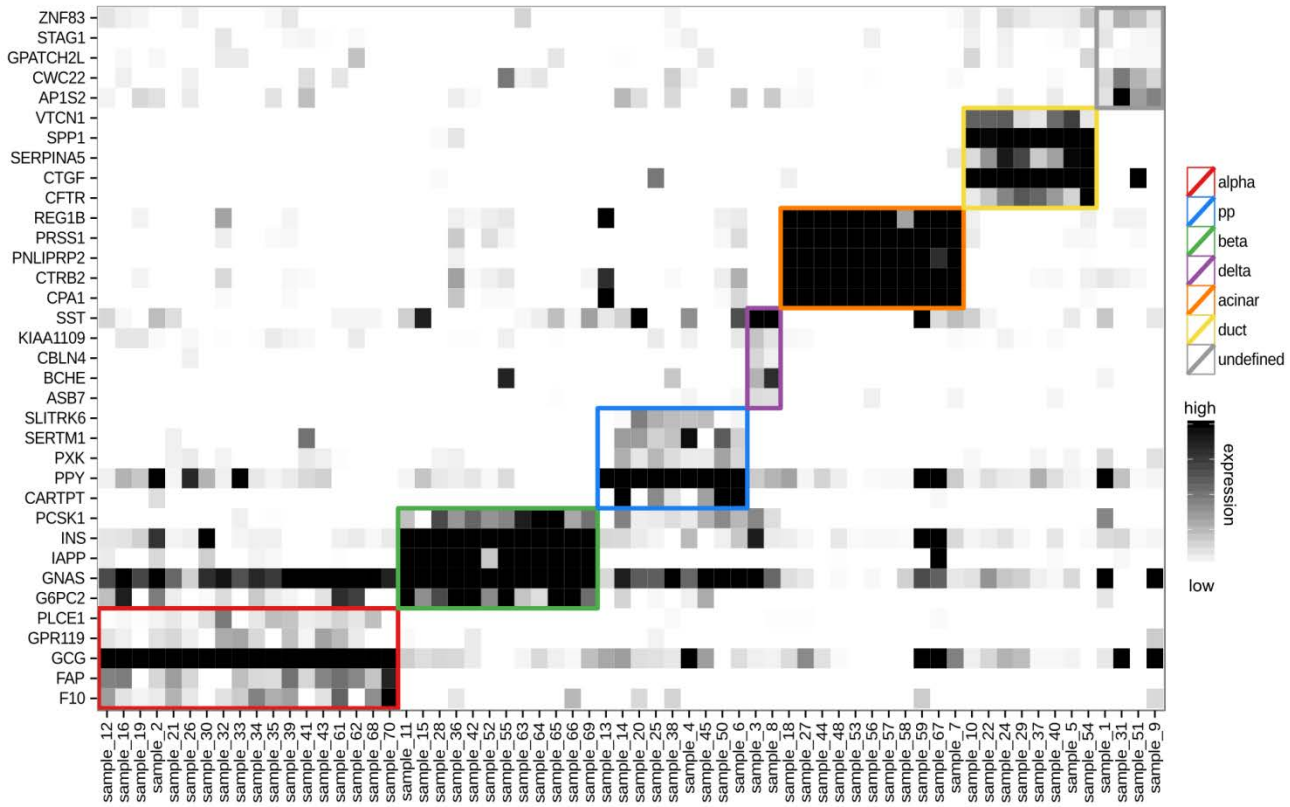


Appendix Figure S2: Single cell expression of PDX1 and ARX

UCSC Genome Browser tracks of PDX1 and ARX for all single-cells as well as merged tracks for each cell type. The scales for individual cells are 0-250 counts for PDX1 and 0-50 counts for ARX.

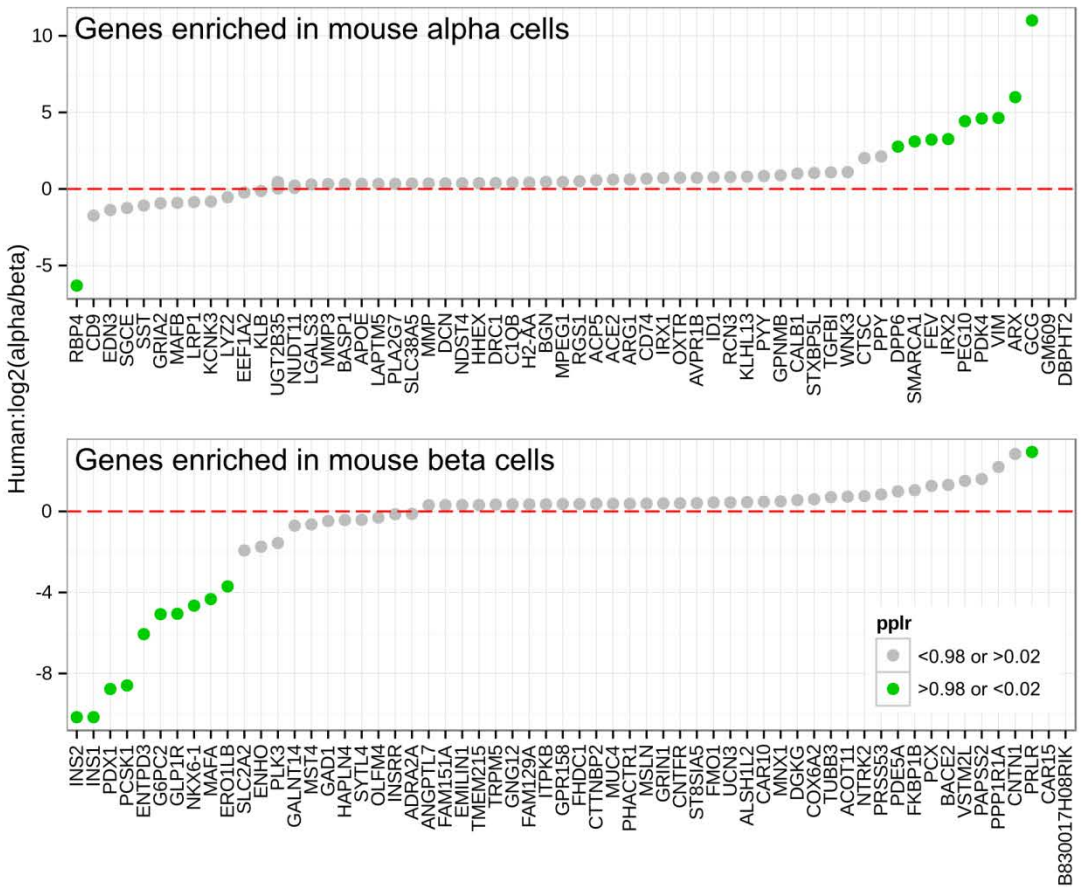


Appendix Figure S3: Single cell expression of REST
UCSC Genome Browser tracks of REST for all single-cells as well as merged tracks for each cell type. The scales for individual cells are 0-10 counts.

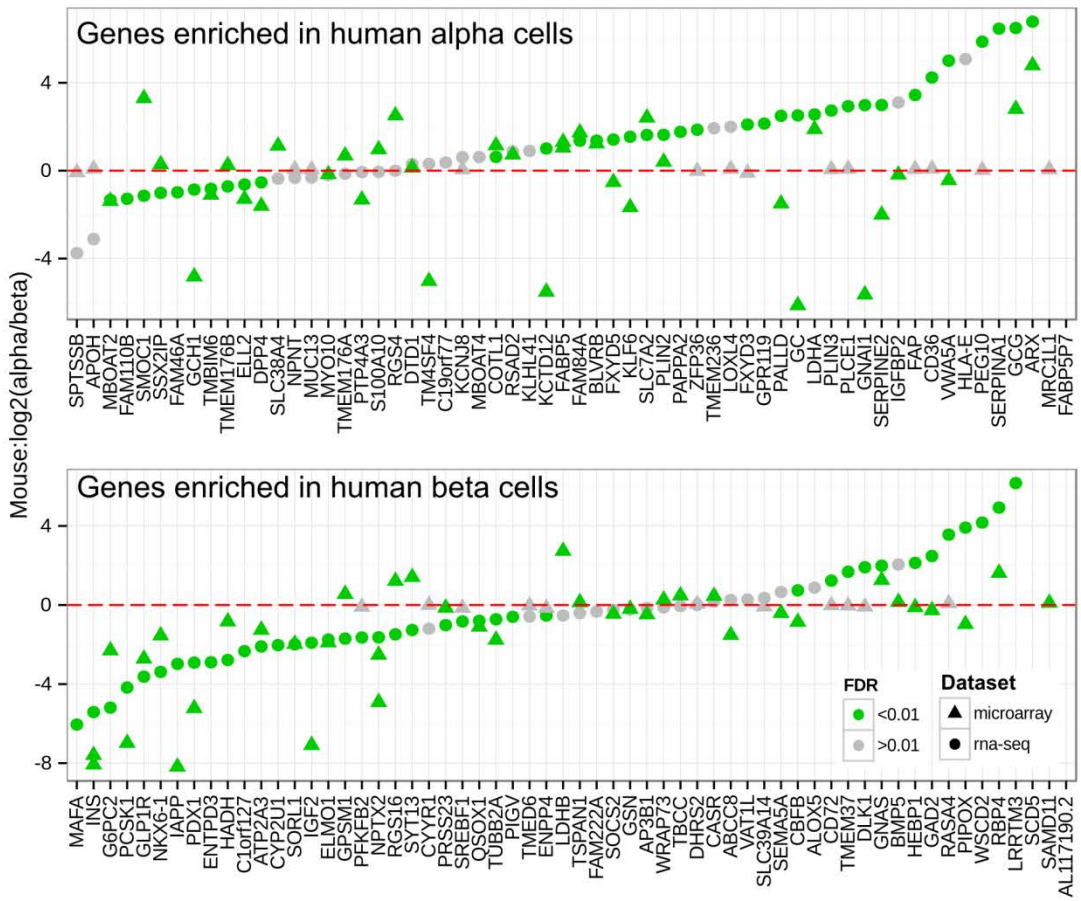


Appendix Figure S4: Identification of cell type specific transcripts
Heatmap displaying the expression patterns (scaled RPKM values) across all single cells for the top five cell type specific genes as determined by pairwise differential expression analysis.

A

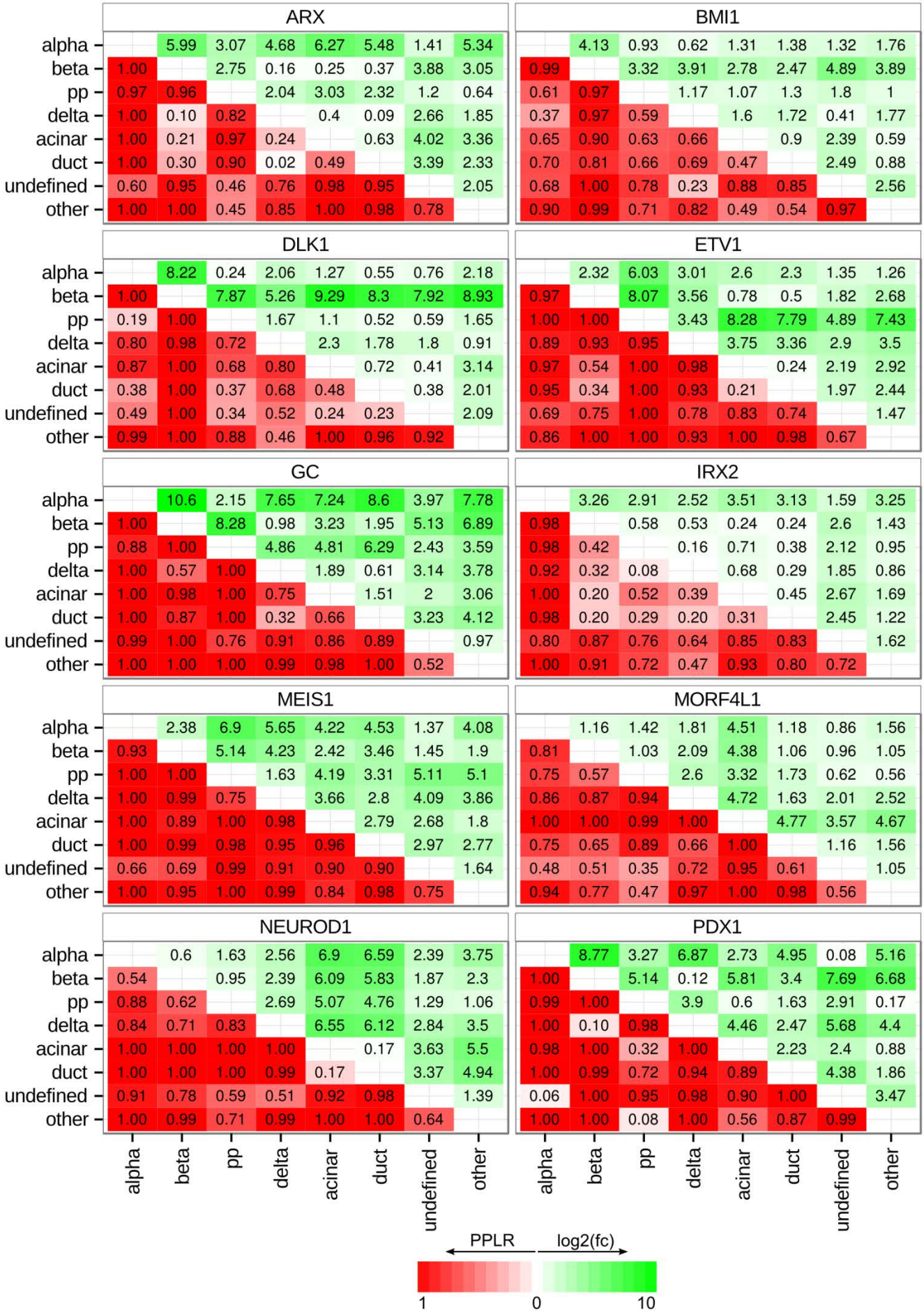


B

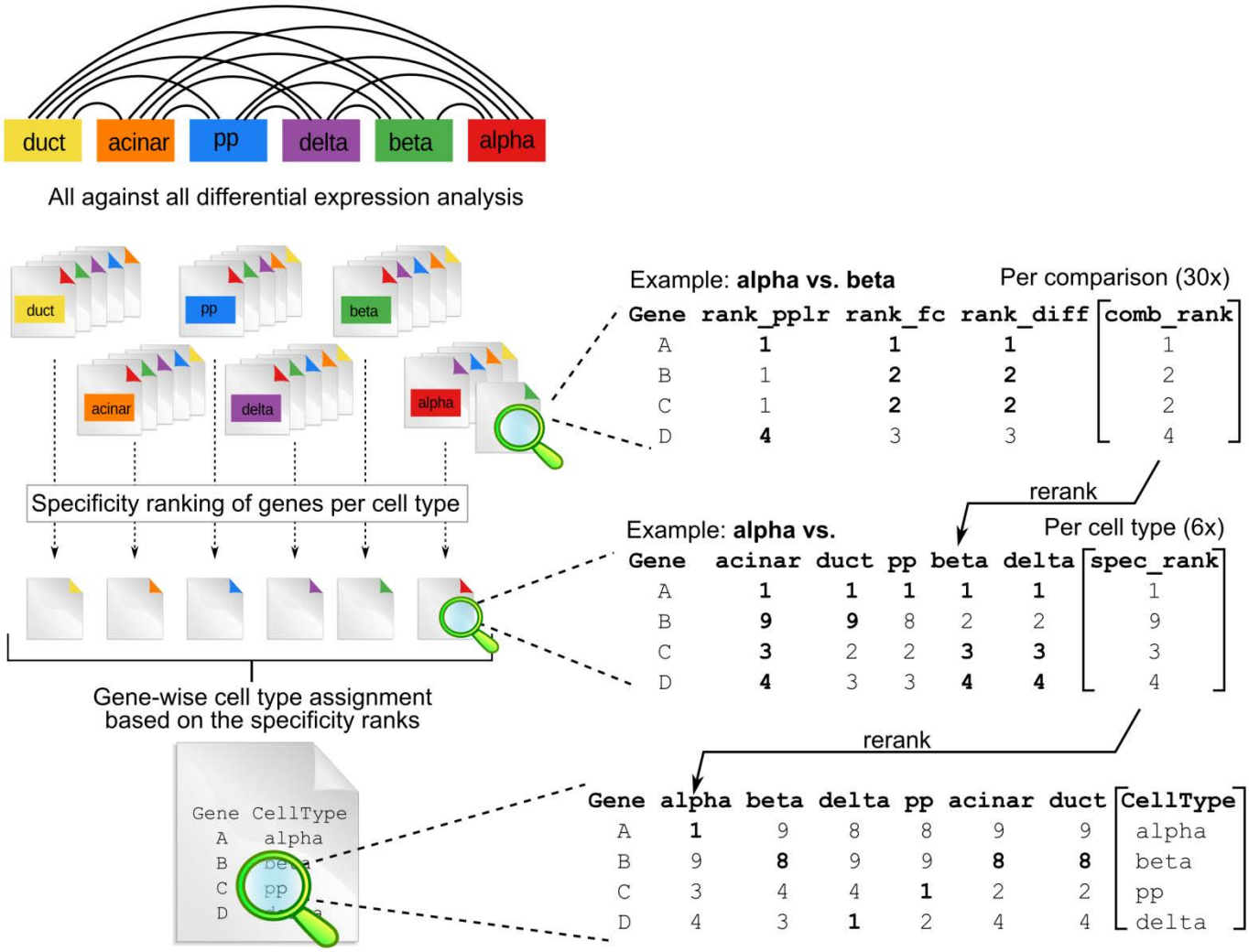


Appendix Figure S5: Comparing alpha cell and beta cell specific genes between human and mouse

(A) Log fold change in RNA expression between human alpha and beta single-cells for genes identified as the 60 most alpha or beta cell specific genes in mouse. **(B)** Log fold change in RNA expression between mouse alpha and beta cells for genes identified as the 60 most alpha or beta cell specific genes in human single-cells. PPLR: probability of positive log ratio; FDR: false discovery rate.



Appendix Figure S6: Statistical significance of differential gene expression
Matrix showing the scaled probability of positive log-ratio (pplr: lower left) and log2 fold change (log2(fc): upper right) of pairwise differential expression analysis for individual genes displayed in Fig. 2C-F and Fig. 3D. “Other” signifies the comparison to all other cell types taken together.



Appendix Figure S7: Rank-based approach for defining cell type specific genes
Illustration of the rank based approach of defining cell type specific genes and assigning cell types to genes. PPLR: Probability of Positive Log Ratio; FC: Fold Change (expression); Diff: absolute difference (expression).

4 Discussion

The research presented in this thesis aims at advancing the understanding of the fundamental concept of cellular identity. By focusing on different aspects of cellular identity, each of the presented projects adds to an increasingly holistic perception of cellular identity, which is however far from complete. Taken together, the presented projects integrate evolutionary and molecular approaches with a focus on DNA methylation (5-methyl-cytosine) as a prominent epigenomic mechanism and RNA expression as a crucial means for realizing genomic and epigenomic information. This final chapter seeks to place the presented results into their broader context.

4.1 General discussion

The most elementary questions surrounding the concept of cellular identity are those addressing the evolutionary processes that led to the emergence of the sophisticated molecular mechanisms that we identify today as crucially involved in the determination of cellular identity. Addressing these questions is not only necessary for the understanding of the living world as it is today, but also for the anticipation of how life might develop in the future. With the recent advances in genome as well as epigenome editing (CRISPR/Cas9) (Hilton *et al*, 2015) it is just a matter of time until these tools will allow to artificially manipulate and shape the identity of cells, the building blocks of life, in an accelerated and directed fashion.

4.1.1 Evolutionary epigenomics of cellular identity

From an evolutionary perspective in the context of cellular identity, DNA methylation is a very interesting epigenomic mechanism. While DNA methylation is indispensable for establishing and maintaining cellular identity in some branches of life (e.g. vertebrates), it is absent in others (e.g. some insects), and displays fundamentally different genome-wide patterns in different phyla (especially vertebrates vs. invertebrates). Given that maintaining cellular identity in multicellular organisms is essential but DNA methylation can apparently be dispensable, the seemingly conserved and crucial role of DNA methylation across vertebrates is rather surprising. In fact, a recent study of DNA methylation across 41 insect species detected high variability of DNA methylation levels even within the same orders and the absence of DNA methylation in one order (Diptera) (Bewick *et al*, 2016), making a conserved role of DNA methylation across all vertebrates appear even more extraordinary. Furthermore, evidence for a diverging role of DNA methylation between cold- and warm-blooded vertebrates has been presented through research attributing unique transcription regulatory functions of CGIs to the evolution of warm-blooded vertebrates (Sharif *et al*, 2010). On the other hand, studying experimentally defined non methylated CpG islands (NMIs) in seven diverse vertebrate species, revealed an unexpected degree of conservation of the regulatory role of NMIs at gene promoters (Long *et al*, 2013). Acknowledging that CGIs and NMIs are functionally distinct genomic features especially in cold-blooded vertebrates, a computational study based on the same experimental data found that although DNA sequence context is highly predictive for DNA methylation status in all vertebrates, CG-rich DNA features are more predictive in warm-blooded vertebrates, while AT-rich features are more predictive in cold-blooded vertebrates (Huska & Vingron, 2016). Taken together, the evidence to date suggests a highly conserved general role of DNA methylation across all vertebrate species but subtle subgroup specific differences in the final implementation. These subtle differences might be the clue needed to understand how and why DNA methylation acquired its unique role in vertebrates. Given that studies so far have mostly only assessed one species per vertebrate class, a study assessing DNA methylation in many more species and thereby more truthfully representing vertebrate diversity seems a promising way to better understand the driving forces that apparently fixated DNA methylation as a central mechanism of stabilizing cellular identity in vertebrates.

However, genome-wide assessment of DNA methylation has been limited by the availability of reference genomes and relatively high sequencing costs. While sequencing costs have rapidly declined in

the past years, the availability of assembled reference genomes is only increasing slowly. To overcome this limitation and to enable the genome-wide profiling of DNA methylation in virtually any species, we have developed a computational approach termed RefFreeDMA that deduces an ad-hoc reference directly from RRBS sequencing reads and can thereby perform differential DNA methylation analysis without the need of a previously assembled reference genome. This approach allows to determine global DNA methylation levels, the sequence context in which certain DNA methylation patterns are found, and differences in DNA methylation between different tissues or cell types (Klughammer *et al*, 2015). Despite some inherent limitations such as the fact that the sequence context is limited to the length of a read (typically 50 bp) and that gene annotations are generally not available, we have shown that biological interpretation of the data is nevertheless possible. Without using a reference genome we have investigated differential DNA methylation between granulocytes and lymphocytes in two mammalian (human and cow) as well as one fish (carp) species and found differentially methylated fragments to be characterized by distinct sequence compositions and enriched for binding sites of lineage specific (lymphoid/myeloid) transcription factors in all three species. These confirmatory experiments encourage the undertaking of larger studies assessing more species and more different cell types or tissues. Such large scale studies assessing tissue specific DNA methylation with 100s of species will allow more confident conclusions with regards to the role of DNA methylation across vertebrates and most importantly provide the statistical power to detect also subtle differences or trends.

4.1.2 Epigenomic assessment of cellular identity in a malignant disease

Although with the availability of more and more reference genomes DNA methylation is being assessed in a broader spectrum of species, our knowledge about DNA methylation in vertebrates has been mainly gained in two mammalian species: Human and mouse. With the discovery of the biomedical relevance of DNA methylation, researchers and society have naturally focused their attention on understanding the role of DNA methylation in human physiology and pathology. Apart from investigating the transgenerational heritability of DNA methylation and assessing the role of DNA methylation in reproduction and development, studying DNA methylation in the context of cancer has been a major research focus. Interestingly, DNA methylation aberrations are found in most if not all cancers, but the uncertainty of cause or consequence is difficult to resolve. Cancerous cells are characterized by uncontrolled, invasive growth and a loss of cellular identity (Roy & Hebrok, 2015). Depending on their specific genomic and epigenomic aberrations, but also depending on their cell type of origin, cancerous cells can display different degrees of differentiation, where less differentiated cancers are usually the more aggressive ones. Apart from the degree of differentiation, also tumour heterogeneity is being more and more recognized as a crucial factor in anticipating the course of malignant diseases. Tumour heterogeneity, meaning that malignant cells from the same entity display heterogeneous (epi)genotypes and phenotypes, is particularly relevant with regards to therapy resistance and early disease progression. Tumour cells that have acquired the molecular setup to evade therapy are a major cause for the typically early relapses observed in glioblastoma, the deadliest and most common tumour of the adult central nervous system. Although progress has been made in understanding molecular processes involved in glioblastoma and the administration of targeted therapies such as for example treatment with the anti-angiogenetic drug Bevacizumab (Narita, 2015), time to progression is still disappointingly low (~ 10 months). In order to prevent relapses, understanding tumour progression on a molecular level and studying not only the primary, but also the recurring tumour, seems indispensable. However, because of the rapid and fatal course of disease, large scale progression studies assessing primary and recurring tumours in glioblastoma are difficult to conduct and to date largely missing from the research landscape. To fill this gap we assembled an Austria-wide cohort of >110 primary glioblastoma patients with formalin-fixed, paraffin embedded (FFPE) samples of the primary and at least the first recurring tumour. Reduced representation bisulfite sequencing (RRBS) yielded DNA methylation maps for all samples and allowed profound molecular characterisation of glioblastoma progression. Comparing DNA methylation profiles between primary and recurring tumours

showed that genes that recurrently lose DNA methylation during progression are enriched for Wnt signalling pathway genes and that demethylation in those genes was associated with earlier progression. Complementarity, aberrant activation of Wnt signalling has been previously described in glioblastoma and amongst others lead back to hypermethylation of Wnt signalling inhibitors such as the DKK gene (Lee *et al*, 2015).

We further assessed two measures of DNA methylation based tumour heterogeneity in primary and recurring samples: DNA methylation erosion (Landau *et al*, 2014) and epi-allelic shifting (Li *et al*, 2014). Although these two measures of tumour heterogeneity are closely related, they each yield complementary information about the epigenetic state of a tumour sample. While DNA methylation erosion has been described as a stochastic process of locally disordered DNA methylation, epi-allelic shifting directly assesses the clonal composition and changes thereof. Of note, DNA methylation erosion is a prerequisite to, but does not necessarily entail the formation of epi-alleles. This somewhat asymmetric relationship implies that although DNA methylation erosion primarily measures the cells' inability to maintain proper DNA methylation it also indirectly informs about the clonal structure of a sample. Surprisingly, we found that higher DNA methylation erosion was associated with favourable clinical outcome. These results stand in contrast to previous research assessing DNA methylation erosion in CLL (Landau *et al*, 2014) and epi-allelic shifting in AML (Li *et al*, 2016b), where both studies reported a negative association between the respective heterogeneity measures and progression-free survival. Both studies suggested increased epigenetic plasticity as explanations for their observations, hypothesising that increased epigenetic plasticity allows the tumour to adapt to environmental pressures such as (chemo) therapy. The discrepancy to our results might be explained by the fundamentally different disease courses of solid tumours such as glioblastoma and hematopoietic malignancies such as CLL and AML. While in glioblastoma the majority of the tumour cells is removed surgically and will never encounter chemotherapy, in hematopoietic malignancies the entire pool of malignant cell is exposed to chemotherapy in order to eradicate them. Furthermore, it is plausible to assume that DNA methylation erosion is generally detrimental but in rare cases can give rise to advantageous phenotypes. In light of the evolutionary bottleneck that is applied on glioblastomas through surgery, it seems likely that the few tumour cells that remain after surgery are less fit and less resistant to chemotherapy if they originate from a tumour with high levels of DNA methylation erosion. In CLL however, where no such indiscriminately externally imposed bottleneck is applied, a single cell that has acquired an advantageous phenotype through stochastic DNA methylation erosion might give rise to a therapy resistant clone, thereby leading to early relapse. These results highlight the importance of DNA methylation derived tumour heterogeneity for tumour progression, but also the necessity to evaluate these measures in an appropriate medical context.

In contrast to epigenetic heterogeneity in glioblastoma, which is a relatively new branch of research, clinically relevant transcriptional subgroups of glioblastoma (classical, mesenchymal, and proneural) have already been described nearly 10 years ago in the context of 'The Cancer Genome Atlas' (TCGA) project (Verhaak, 2009). Using machine learning, we were able to extract subtype specific DNA methylation signatures from published and annotated TCGA data and we used those signatures to predict the transcriptional subtypes of our glioblastoma samples based on the RRBS data. In concordance with a single-cell transcriptomics study in glioblastoma (Patel *et al*, 2014), we found that tumours displayed extensive heterogeneity with most tumours containing variable proportions of all three transcriptional subtypes. We went on to characterize the transcriptional subtypes by their epigenome regulatory properties and found that the mesenchymal subtype displayed DNA methylation signatures of stronger EZH2 activity and weaker activity of stemness conferring transcription factors (NANOG, SOX2, OCT4) compared to the other subtypes. In concordance with the original publication (Patel *et al*, 2014), the mesenchymal subtype also showed a high rate of immune infiltration and a lower fraction of proliferating cells. Survival analysis further revealed that patients with recurring tumours of the mesenchymal subtype had worse clinical outcome while the transcriptional subtype of the primary tumour was clinically irrelevant. Our results link the established transcriptional subtypes to

novel characteristic epigenome regulatory signatures and thereby provide evidence in support of previous efforts to investigate EZH2 inhibitors as promising new glioblastoma therapeutics (Suva *et al*, 2009) especially in the treatment of glioblastomas of the mesenchymal subtype. Furthermore, by being able to recapitulate the transcriptional subtypes on the epigenomic (DNA methylation) level we corroborate the original suggestion that the transcriptional subtypes represent distinct neural cell types (classical: astrocytes, mesenchymal: immortalized astroglia, proneural: oligodendrocytes) (Patel *et al*, 2014) instead of only transient transcriptional states.

4.1.3 Transcriptional assessment of cellular identity

DNA methylation, as a relatively stable representation of cellular identity, is a powerful mark to discriminate different cell types while preventing the confusion of persisting cell types and short-term transcriptional states. However, assessing marker gene expression, as a representation of a cell type's function, can be extremely helpful when it comes to the identification and functional characterisation of different cell types. Furthermore, transcriptomic changes usually precede epigenomic changes in short-term stimulations or changes of cellular identity (i.e. trans-differentiation) and transcriptional states within certain cell types might also be the actual focus of interest. For these reasons, certain investigations regarding cellular identity clearly prefer transcriptional profiling to epigenomic (DNA methylation) profiling.

With this in mind, we used single-cell RNA sequencing to define accurate transcriptional profiles for healthy human pancreatic islet cells as a resource for the investigation of human islet biology and diabetes research (Li *et al*, 2016a). Being primarily defined by the hormone they produce, the major cell types making up human pancreatic islets have been known for a long time (Gersell *et al*, 1978; Orci *et al*, 1976). However, only transcriptional profiling on the single-cell level has allowed a sufficiently high resolution necessary to detect subtle changes and heterogeneity within cell populations, paving the way for new insights in molecular processes relevant for diabetes and their exploitation for potential therapies. Our initial study (Li *et al*, 2016a) revealed novel cell type specific transcription factors (alpha cells: IRX2, beta cells: BMI1, PP cells: MEIS1 and ETV1) and a transcription factor (REST) expressed exclusively in cells of the exocrine pancreas that might be a target for inducing the trans-differentiation of ductal cells to insulin producing beta cells. Furthermore, comparing our human data to published mouse data, we surprisingly found that two of the genes most specifically expressed in human alpha or beta cells (GC in alpha cells and DLK1 in beta cells) showed an opposite expression pattern in mouse, highlighting the importance of acknowledging species specific differences in islet biology. Finally, our exploratory analysis of human islet single-cell transcriptomes also yielded precise transcriptional profiles for all major human pancreatic islet cell types, representing a reference point for subsequent studies such as the induction of trans-differentiation from alpha to beta cells using Artemisinins, a class of GABAa receptor agonist small molecules, commonly used for malaria treatment (Li *et al*, 2017). Although, treatment of human pancreatic islets with Artemisinins for up to 72 hours did show robust effects on alpha cells including the downregulation of alpha cell specific genes, it did not completely alter the cell type specific transcriptional profiles, and treated alpha cells were more similar to untreated alpha cells than to beta cell on a global transcriptomic scale. Thus, although treated alpha cells showed clear signs of de- or trans- differentiation they had not yet lost their alpha cell identity when treatment was stopped due to the temporal limitation of ex-vivo experiments, indicating that a complete de- or trans-differentiation, if at all possible in human, would take longer than 72 hours. In fact, long-term treatment of diabetic mice with GABA induced the conversion of alpha like cells to functional, insulin secreting beta like cells, leading to a remarkable replenishment of previously missing beta cell mass (Ben-Othman *et al*, 2017). In spite of the known species specific differences in islet cell biology, these results allow cautious optimism that drug-induced, in situ replacement of lost beta cells might also be possible in diabetic humans. Apart from the potential medical impact, trans-differentiation experiments in (human) pancreatic islet cell types also yield valuable insights in the plasticity of cellular identity (van der Meulen & Huising, 2015). Tracking cell type conversions in this relatively

well understood cellular system using single cell transcriptional and epigenomic profiling would provide a better understanding of the molecular dynamics involved in changing the identity of a cell. In particular, it would be interesting to see, if expression patterns rather change gradually or in a discrete fashion and at which point in the trans-differentiation process epigenomic changes appear and consolidate.

4.2 Conclusion & future prospects

Since the realisation that cells are the basic units of life on earth about 180 years ago, an immense amount of knowledge and insight elucidating the determination of cellular identity has been generated. Importantly, science has managed to translate many of these insights into medical advancements ranging from the treatment of diseases such as cancer or diabetes to assisted reproduction (i.e. in vitro fertilisation). The work presented in this thesis contributes to this existing pool of biological and medical knowledge but also demonstrates the power of recent technological advances in high throughput sequencing. The plethora of nucleic acid sequencing approaches available today already allow detailed and efficient molecular characterisation of all major determinants of cellular identity (genome, epigenome and transcriptome) and applied on single cells these methods already allow the identification of previously unknown cellular subtypes as demonstrated most recently for dendritic cells and monocytes (Villani *et al*, 2017). Of course, this unprecedented sensitivity poses new challenges such as to discriminate whether an observed deviation from known molecular characteristics really represents a biologically relevant new cell type or just a certain transient cell state. Here, technologies for parallel transcriptome and epigenome profiling of single cells (Angermueller *et al*, 2016; Hu *et al*, 2016) might provide the evidence needed to increase the confidence in a presumably newly discovered cell type.

Humanity has now acquired the knowledge and technical ability to identify and characterize every cell type to be found in even the most complex organisms, promising considerable advances in understanding physiology, pathology and the evolutionary basis of our existence. For human, a consortium of leading biomedical scientists has already envisioned this ambitious endeavour and named it “The Human Cell Atlas” (www.humancellatlas.org).

5 References

Adami C (2002) What is complexity? *BioEssays* **24:** 1085–1094 Available at: http://doi.wiley.com/10.1002/bies.10192

Adiconis X, Berlin AM, Borges-Rivera D, Busby MA, DeLuca DS, Fennell T, Gnirke A, Levin JZ, Pochet N, Regev A, Satija R, Sivachenko A, Thompson DA & Wysoker A (2013) Comprehensive comparative analysis of RNA sequencing methods for degraded or low input samples. *Nat. Methods* **10:** 1–20

Albalat R (2008) Evolution of DNA-methylation machinery: DNA methyltransferases and methyl-DNA binding proteins in the amphioxus Branchiostoma floridae. *Dev. Genes Evol.* **218:** 691–701 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18813943 [Accessed October 24, 2012]

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S a JR, Behjati S, Biankin A V, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens J a, Greaves M, et al (2013) Signatures of mutational processes in human cancer. *Nature* **500:** 415–421 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3776390&tool=pmcentrez&rendertype=abstract

Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, Kelsey G, Stegle O & Reik W (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13:** 229–32 Available at: http://www.nature.com/doifinder/10.1038/nmeth.3728%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26752769%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4770512

Arendt D (2008) The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9:** 868–882

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD & Wagner GP (2016) The origin and evolution of cell types. *Nat. Rev. Genet.* Available at: http://www.nature.com/doifinder/10.1038/nrg.2016.127

Atkinson MA, Eisenbarth GS & Michels AW (2014) Type 1 diabetes. *Lancet* **383:** 69–82 Available at: http://dx.doi.org/10.1016/S0140-6736(13)60591-7

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, et al (2015) A global reference for human genetic variation. *Nature* **526:** 68–74 Available at: http://www.nature.com/doifinder/10.1038/nature15393

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K V., Altshuler D, Gabriel S & DePristo MA (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline

Bell AC & Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405:** 482–5 Available at: http://www.nature.com/doifinder/10.1038/35013100

Bell G & Mooers AO (1997) Size and complexity among multicellular organisms. *Biol. J. Linn. Soc.* **60:** 345–363 Available at: http://doi.wiley.com/10.1111/j.1095-8312.1997.tb01500.x

Ben-Othman N, Vieira A, Courtney M, Record F, Gjernes E, Avolio F, Hadzic B, Druelle N, Napolitano T, Navarro-Sanz S, Silvano S, Al-Hasani K, Pfeifer A, Lacas-Gervais S, Leuckx G, Marroquí L, Thévenet J, Madsen OD, Eizirik DL, Heimberg H, et al (2017) Long-Term GABA Administration Induces Alpha Cell-Mediated Beta-like Cell Neogenesis. *Cell* **168:** 73–85.e11 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867416315239

Berdasco M & Esteller M (2010) Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev. Cell* **19:** 698–711 Available at: http://www.ncbi.nlm.nih.gov/pubmed/21074720

Bernstein BE, Meissner A & Lander ES (2007) The Mammalian Epigenome. *Cell* **128:** 669–681

Bewick AJ, Vogel KJ, Moore AJ & Schmitz RJ (2016) Evolution of DNA Methylation across Insects. *Mol. Biol. Evol.* **34:** msw264 Available at: https://academic.oup.com/mbe/article-

lookup/doi/10.1093/molbev/msw264

Bird AP (1995) Gene number, noise reduction and biological complexity. *Trends Genet.* **11:** 94–100 Available at: http://www.ncbi.nlm.nih.gov/pubmed/7732579

Blackledge NP, Long HK, Zhou JC, Kriaucionis S, Patient R & Klose RJ (2012) Bio-CAP: A versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic Acids Res.* **40:**

Blattler A & Farnham PJ (2013) Cross-talk between site-specific transcription factors and DNA methylation states. *J. Biol. Chem.* **288:** 34287–34294

Bock C (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13:** 705–719 Available at: http://www.nature.com/doifinder/10.1038/nrg3273

Bock C, Beerman I, Lien WH, Smith ZD, Gu H, Boyle P, Gnirke A, Fuchs E, Rossi DJ & Meissner A (2012) DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol. Cell* **47:** 633–47 Available at: http://linkinghub.elsevier.com/retrieve/pii/S1097276512005448 [Accessed July 14, 2015]

Bock C, Farlik M & Sheffield NC (2016a) Multi-Omics of Single Cells: Strategies and Applications. *Trends Biotechnol.* **34:** 605–608

Bock C, Halbritter F, Carmona FJ, Tierling S, Datlinger P, Assenov Y, Berdasco M, Bergmann AK, Booher K, Busato F, Campan M, Dahl C, Dahmcke CM, Diep D, Fernández AF, Gerhauser C, Haake A, Heilmann K, Holcomb T, Hussmann D, et al (2016b) Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nat. Biotechnol.* **34:** 726–737 Available at: http://www.nature.com/doifinder/10.1038/nbt.3605

Bock C & Lengauer T (2012) Managing drug resistance in cancer: lessons from HIV therapy. *Nat. Rev. Cancer* **12:** 494–501 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22673150 [Accessed July 13, 2012]

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T & Walter J (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet.* **2:** e26 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1386721&tool=pmcentrez&rendertype=abstract [Accessed November 12, 2012]

Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, Jäger N, Gnirke A, Stunnenberg HG & Meissner A (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* **28:** 1106–1114 Available at: http://www.nature.com/doifinder/10.1038/nbt.1681

Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL & Fire AZ (2009) Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Sci. Transl. Med.* **1:** 12ra23-12ra23 Available at: http://stm.sciencemag.org/cgi/doi/10.1126/scitranslmed.3000540

Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, Karchin R, Kinzler KW, Vogelstein B & Nowak MA (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci.* **107:** 18545–18550 Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1010978107

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, Berlin AM, Campbell MS, Barrell D, Martin KJ, Mulley JF, Ravi V, Lee AP, Nakamura T, Chalopin D, Fan S, et al (2016) The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* **48:** 427–437 Available at: http://dx.doi.org/10.1038/ng.3526

Bröske A-M, Vockentanz L, Kharazi S, Huska MR, Mancini E, Scheller M, Kuhl C, Enns A, Prinz M, Jaenisch R, Nerlov C, Leutz A, Andrade-Navarro M a, Jacobsen SEW & Rosenbauer F (2009) DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nat. Genet.* **41:** 1207–15 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19801979 [Accessed October 24, 2012]

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann S a, Marioni JC & Stegle O (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* Available at: http://www.nature.com/doifinder/10.1038/nbt.3102 [Accessed January 19, 2015]

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J & Eichler EE (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–11 Available at: http://www.nature.com/nature/journal/v517/n7536/full/nature13907.html%5Cnhttp://www.nature.com/nature/journal/v517/n7536/pdf/nature13907.pdf

Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, Butte AJ & Snyder M (2011) Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29:** 908–914 Available at: http://dx.doi.org/10.1038/nbt.1975%5Cnfile:///Users/rmorin/Dropbox/Papers2/2011/Clark/Nat Biotechnol 2011 Clark.pdf%5Cnpapers2://publication/doi/10.1038/nbt.1975

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X & Mortazavi A (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17:** 13 Available at: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Cooper GM (2000) The Origin and Evolution of Cells. In *The Cell - A Molecular Approach* Sunderland (MA): Sinauer Associates Available at: https://www.ncbi.nlm.nih.gov/books/NBK9841/

Cooper MD & Alder MN (2006) The Evolution of Adaptive Immune Systems. *Cell* **124:** 815–822 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867406001528

CRICK F (1970) Central Dogma of Molecular Biology. *Nature* **227:** 561–563 Available at: http://www.google.de/imgres?sa=X&espv=210&es_sm=91&biw=1568&bih=929&tbm=isch&tbnid=HN2Do5L4VxlbnM:&imgrefurl=http://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/SCBCCH&docid=UkOef-Ev24D3kM&imgurl=http://profiles.nlm.nih.gov/ps/access/SCBCCH~.png&w=17

Daley GQ (2015) Stem cells and the evolving notion of cellular identity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370:** 20140376 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26416685%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4634003

Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH & Zhang MQ (2006) Computational prediction of methylation status in human genomic sequences. **103:**

Datlinger P, Schmidl C, Rendeiro AF, Traxler P, Klughammer J, Schuster L & Bock C (2016) Pooled CRISPR screening with single-cell transcriptome read-out. *bioRxiv*

Down T a, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, et al (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* **26:** 779–85 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2644410&tool=pmcentrez&rendertype=abstract

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, et al (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74 Available at: http://www.nature.com/doifinder/10.1038/nature11247%5Cnpapers3://publication/doi/10.1038/nature11247

Echeverri K (2002) Ectoderm to Mesoderm Lineage Switching During Axolotl Tail Regeneration. *Science (80-.).* **298:** 1993–1996 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1077804

Efroni I, Ip P-L, Nawy T, Mello A & Birnbaum KD (2015) Quantification of cell identity from single-cell gene expression profiles. *Genome Biol.* **16:** 9 Available at: http://genomebiology.com/2015/16/1/9

Ehrlich M (2003) The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease. *Clin. Immunol.* **109:** 17–28 Available at: http://linkinghub.elsevier.com/retrieve/pii/S1521661603002018

Erlich Y & Narayanan A (2014) Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15:** 409–421 Available at: http://www.nature.com/nrg/journal/v15/n6/abs/nrg3723.html%5Cnhttp://www.nature.com/nrg/journal/v15/n6/pdf/nrg3723.pdf

Fang F, Fan S, Zhang X & Zhang MQ (2006) Predicting methylation status of CpG islands in the human brain. *Bioinformatics* **22:** 2204–2209 Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btl377

Farlik M, Halbritter F, Müller F, Choudry FA, Ebert P, Klughammer J, Farrow S, Santoro A, Ciaurro V, Mathur A, Uppal R, Stunnenberg HG, Ouwehand WH, Laurenti E, Lengauer T, Frontini M & Bock C (2016) DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell* **0:** 224–226 Available at: http://linkinghub.elsevier.com/retrieve/pii/S1934590916303605

Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J & Bock C (2015) Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep.***:** 1386–1397 Available at: http://linkinghub.elsevier.com/retrieve/pii/S2211124715001096 [Accessed March 2, 2015]

Flores E & Herrero A (2010) Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nat. Rev. Microbiol.* **8:** 39–50

Forbes SJ & Rosenthal N (2014) Preparing the ground for tissue regeneration: from mechanism to therapy. *Nat. Med.* **20:** 857–869 Available at: http://dx.doi.org/10.1038/nm.3653

Friedlander T, Prizak R, Guet CC, Barton NH & Tkačik G (2016) Intrinsic limits to gene regulation by global crosstalk. *Nat. Commun.* **7:** 12307 Available at: http://www.nature.com/doifinder/10.1038/ncomms12307

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL & Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89:** 1827–31 Available at: http://www.ncbi.nlm.nih.gov/pubmed/1542678%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC48546

Gapp B V, Konopka T, Penz T, Dalal V, Bürckstümmer T, Bock C & Nijman SM (2016) Parallel reverse genetic screening in mutant human cells using transcriptomics. *Mol. Syst. Biol.* **12:** 879 Available at: http://msb.embopress.org/lookup/doi/10.15252/msb.20166890

Gawad C, Koh W & Quake SR (2016) Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17:** 175–188 Available at: http://dx.doi.org/10.1038/nrg.2015.16%5Cnhttp://10.1038/nrg.2015.16

Gersell DJ, Gingerich RL & Greider MH (1978) Regional Distribution and Concentration of Pancreatic Polypeptide in the Human and Canine Pancreas. *Diabetes* **28:** 11–15 Available at: http://diabetes.diabetesjournals.org/cgi/doi/10.2337/diab.28.1.11

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, et al (2007) Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153–158 Available at: http://www.nature.com/doifinder/10.1038/nature05610

Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H & van Oudenaarden A (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* Available at: http://www.nature.com/doifinder/10.1038/nature14966

Guo H, Zhu P, Wu X, Li X, Wen L & Tang F (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23:** 2126–35 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24179143 [Accessed January 11, 2014]

van Gurp TP, Wagemaker NCAM, Wouters B, Vergeer P, Ouborg JNJ & Verhoeven KJF (2016)

epiGBS: reference-free reduced representation bisulfite sequencing. *Nat. Methods* **13:** 322–4 Available at: http://dx.doi.org/10.1038/nmeth.3763%5Cnhttp://www.nature.com/doifinder/10.1038/nmeth.3763%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26855363

Hanahan D & Weinberg RA (2011) Hallmarks of Cancer: The Next Generation. *Cell* **144:** 646–674 Available at: http://dx.doi.org/10.1016/j.cell.2011.02.013

Hedges SB, Blair JE, Venturi ML & Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* **4:** 2 Available at: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-4-2

Hernando-Herraez I, Heyn H, Fernandez-Callejo M, Vidal E, Fernandez-Bellon H, Prado-Martinez J, Sharp AJ, Esteller M & Marques-Bonet T (2015) The interplay between DNA methylation and sequence divergence in recent human evolution. *Nucleic Acids Res.* **43:** 8204–8214 Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv693

Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, Hvilsom C, Navarro A, Esteller M, Sharp AJ & Marques-Bonet T (2013) Dynamics of DNA methylation in recent human and great ape evolution. *PLoS Genet.* **9:** e1003763 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3764194&tool=pmcentrez&rendertype=abstract [Accessed January 11, 2014]

Herskowitz I (1989) A regulatory hierarchy for cell specialization in yeast. *Nature* **342:** 749–757 Available at: http://www.ncbi.nlm.nih.gov/pubmed/2507922

Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE & Gersbach C a (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33:** 510–517 Available at: http://www.nature.com/doifinder/10.1038/nbt.3199

Höfer T & Rodewald H-R (2016) Output without input: the lifelong productivity of hematopoietic stem cells. *Curr. Opin. Cell Biol.* **43:** 69–77 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0955067416301338

Holland ND, Holland LZ & Holland PW (2015) Scenarios for the making of vertebrates. *Nature* **520:** 450–455 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25903626

Hood L, Kronenberg M & Hunkapiller T (1985) T cell antigen receptors and the immunoglobulin supergene family. *Cell* **40:** 225–9 Available at: http://www.ncbi.nlm.nih.gov/pubmed/3917857

Horn D (2014) Antigenic variation in African trypanosomes. *Mol. Biochem. Parasitol.* **195:** 123–129 Available at: http://dx.doi.org/10.1016/j.molbiopara.2014.05.001

Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D, Wu H, Ye X, Ye C, Wu R, Jian M, Chen Y, Xie W, Zhang R, Chen L, Liu X, et al (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148:** 873–885 Available at: http://dx.doi.org/10.1016/j.cell.2012.02.028

Hu Y, Huang K, An Q, Du G, Hu G, Xue J, Zhu X, Wang C-Y, Xue Z & Fan G (2016) Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17:** 88 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4858893&tool=pmcentrez&rendertype=abstract

Huska M & Vingron M (2016) Improved Prediction of Non-methylated Islands in Vertebrates Highlights Different Characteristic Sequence Patterns. *PLOS Comput. Biol.* **12:** e1005249 Available at: http://dx.plos.org/10.1371/journal.pcbi.1005249

Illumina (2011) TruSeq ™ RNA and DNA Sample Preparation Kits v2. **:** 2–5

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P & Linnarsson S (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11:** 163–166 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24363023

Jeanpierre M, Turleau C, Aurias A, Prieur M, Ledeist F, Fischer A & Viegas-pequignot E (1993) An embryonic-like methylation pattern of classical satellite DNA is observed in ICF syndrome. *Hum.*

Mol. Genet. **2:** 731–735

Jiang Y & Xu C (2010) The calculation of information and organismal complexity. *Biol. Direct* **5:** 59 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2973933&tool=pmcentrez&rendertype=abstract

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA & Kirschner MW (2015) Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161:** 1187–1201 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867415005000

Klughammer J, Datlinger P, Printz D, Sheffield NC, Farlik M, Hadler J, Fritsch G & Bock C (2015) Differential DNA Methylation Analysis without a Reference Genome. *Cell Rep.* **13:** 2621–2633 Available at: http://dx.doi.org/10.1016/j.celrep.2015.11.024

Knoblich JA (2008) Mechanisms of Asymmetric Stem Cell Division. *Cell* **132:** 583–597

Koerner M V. & Barlow DP (2010) Genomic imprinting-an epigenetic gene-regulatory model. *Curr. Opin. Genet. Dev.* **20:** 164–170 Available at: http://dx.doi.org/10.1016/j.gde.2010.01.009

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC & Teichmann SA (2015) The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58:** 610–620 Available at: http://dx.doi.org/10.1016/j.molcel.2015.04.005

Kondo M, Scherer DC, King AG, Manz MG & Weissman IL (2001) Lymphocyte development from hematopoietic stem cells. *Curr. Opin. Genet. Dev.* **11:** 520–526

Kragl M, Knapp D, Nacu E, Khattak S, Maden M, Epperlein HH & Tanaka EM (2009) Cells keep a memory of their tissue origin during axolotl limb regeneration. *Nature* **460:** 60–65 Available at: http://www.nature.com/doifinder/10.1038/nature08152

Kretzschmar K & Watt FM (2012) Lineage Tracing. *Cell* **148:** 33–45 Available at: http://dx.doi.org/10.1016/j.cell.2012.01.002

Kröger B, Vinther J & Fuchs D (2011) Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules: Extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. *BioEssays* **33:** 602–613

Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, Stevenson K, Sougnez C, Wang L, Li S, Kotliar D, Zhang W, Ghandi M, Garraway L, Fernandes SM, Livak KJ, Gabriel S, Gnirke A, Lander ES, Brown JR, et al (2014) Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell* **26:** 813–825 Available at: http://linkinghub.elsevier.com/retrieve/pii/S1535610814004164

Lane N & Martin W (2010) The energetics of genome complexity. *Nature* **467:** 929–934 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20962839

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, et al (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501:** 506–511 Available at: http://www.nature.com/doifinder/10.1038/nature12531

Latchman DS (1997) Transcription factors: An overview. *Int. J. Biochem. Cell Biol.* **29:** 1305–1312 Available at: http://linkinghub.elsevier.com/retrieve/pii/S135727259700085X

Lee JT (2012) Epigenetic regulation by long noncoding RNAs. *Science (80-.).* **338:** 1435–1439 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1231776%5Cnpapers3://publication/doi/10.1126/science.1231776

Lee Y, Lee J-K, Ahn SH, Lee J & Nam D-H (2015) WNT signaling in glioblastoma and therapeutic opportunities. *Lab. Investig.* **0:** 1–14 Available at: http://www.nature.com/doifinder/10.1038/labinvest.2015.140

Leek JT, Johnson WE, Parker HS, Jaffe AE & Storey JD (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28:** 882–

883

Lelieveld SH, Spielmann M, Mundlos S, Veltman J a & Gilissen C (2015) Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum. Mutat.* **36:** 815–822 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25973577

Li J, Casteels T, Frogne T, Ingvorsen C, Honoré C, Courtney M, Huber KVM, Schmitner N, Kimmel RA, Romanov RA, Sturtzel C, Lardeau C-H, Klughammer J, Farlik M, Sdelci S, Vieira A, Avolio F, Briand F, Baburin I, Májek P, et al (2017) Artemisinins Target GABA A Receptor Signaling and Impair α Cell Identity. *Cell* **168:** 86–100.e15 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867416315318

Li J, Klughammer J, Farlik M, Penz T, Spittler A, Barbieux C, Berishvili E, Bock C & Kubicek S (2016a) Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep.* **17:** 178–187 Available at: http://embor.embopress.org/lookup/doi/10.15252/embr.201540946%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26691212%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4784001

Li S, Garrett-bakelman F, Perl AE, Luger SM, Zhang C, To BL, Lewis ID, Brown AL, Andrea RJD, Ross ME, Levine R, Carroll M, Melnick A & Mason CE (2014) Dynamic evolution of clonal epialleles revealed by methclone. **:** 1–12

Li S, Garrett-Bakelman FE, Chung SS, Sanders MA, Hricik T, Rapaport F, Patel J, Dillon R, Vijay P, Brown AL, Perl AE, Cannon J, Bullinger L, Luger S, Becker M, Lewis ID, To LB, Delwel R, Löwenberg B, Döhner H, et al (2016b) Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* **22:** 792–799 Available at: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L72171052%5Cnhttp://www.bloodjournal.org/content/126/23/306%5Cnhttp://sfx.metabib.ch/sfx_locater?sid=EMBASE&issn=00064971&id=doi:&atitle=Divergent+dynamics+of+epigenetic+and+genetic+

Litman GW, Rast JP & Fugmann SD (2010) The origins of vertebrate adaptive immunity. *Nat. Rev. Immunol.* **10:** 543–553 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2919748&tool=pmcentrez&rendertype=abstract

Loker ES, Adema CM, Zhang S-M & Kepler TB (2004) Invertebrate immune systems--not homogeneous, not simple, not well understood. *Immunol. Rev.* **198:** 10–24 Available at: http://www.ncbi.nlm.nih.gov/pubmed/15199951

Loman NJ, Quick J & Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12:** 733–735 Available at: http://www.nature.com/doifinder/10.1038/nmeth.3444

Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, Grützner F, Odom DT, Patient R, Ponting CP & Klose RJ (2013) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife* **2:** 1–19 Available at: http://elife.elifesciences.org/lookup/doi/10.7554/eLife.00348 [Accessed February 27, 2013]

Lukas J, Lukas C & Bartek J (2011) More than just a focus: The chromatin response to DNA damage and its role in genome integrity maintenance. *Nat. Cell Biol.* **13:** 1161–9 Available at: http://www.ncbi.nlm.nih.gov/pubmed/21968989

Lukeš J, Archibald JM, Keeling PJ, Doolittle WF & Gray MW (2011) How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63:** 528–37 Available at: http://doi.wiley.com/10.1002/iub.489

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A & Mccarroll SA Highly Parallel Genome - wide Expression Profiling of Individual Cells Using Nanoliter Droplets.

Mattick JS (2001) Non-coding RNAs : the architects of eukaryotic complexity. **2:** 986–991

Mazor T, Pankov A, Song JS & Costello JF (2016) Intratumoral Heterogeneity of the Epigenome. *Cancer Cell* **29:** 440–451 Available at: http://dx.doi.org/10.1016/j.ccell.2016.03.009

Mazzarello P (1999) A unifying concept: the history of cell theory. *Nat. Cell Biol.* **1:** E13-5 Available at:

http://www.nature.com/doifinder/10.1038/8964

McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF & Shendure J (2016) Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science (80-.).* **353:** aaf7907-aaf7907 Available at: http://www.ncbi.nlm.nih.gov/pubmed/27229144%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4967023%5Cnhttp://www.sciencemag.org/cgi/doi/10.1126/science.aaf7907

Meissner A (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33:** 5868–5877 Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gki901

Melton C, Reuter JA, Spacek D V & Snyder M (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47:** 710–716 Available at: http://www.nature.com/doifinder/10.1038/ng.3332

van der Meulen T & Huising MO (2015) Role of transcription factors in the transdifferentiation of pancreatic islet cells. *J. Mol. Endocrinol.* **54:** R103–R117 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25931448

Michalopoulos GK (1997) Liver Regeneration. *Science (80-.).* **276:** 60–66 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.276.5309.60

Miller SM (2010) Volvox, Chlamydomonas, and the evolution of multicellularity. *Nat. Educ.* **3:** 65 Available at: http://www.nature.com/scitable/topicpage/volvox-chlamydomonas-and-the-evolution-of-multicellularity-14433403

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H & Kakutani T (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* **411:** 212–214

Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, Kinston S, Joshi A, Hannah R, Theis FJ, Jacobsen SE, de Bruijn MF & Göttgens B (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.* **15:** 363–72 Available at: http://dx.doi.org/10.1038/ncb2709

Moll P, Ante M, Seitz A & Reda T (2014) QuantSeq 3 ' mRNA sequencing for RNA quantification. *Nat. Methods* **11:** 25 Available at: http://www.nature.com/nmeth/journal/v11/n12/full/nmeth.f.376.html

Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N & Yamanaka S (2007) Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat. Biotechnol.* **26:** 101–106 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18059259

Narita Y (2015) Bevacizumab for glioblastoma. *Ther. Clin. Risk Manag.* **11:** 1759–1765

Nguyen L V., Vanner R, Dirks P & Eaves CJ (2012) Cancer stem cells: an evolving concept. *Nat. Rev. Cancer* **12:** 133–143 Available at: http://www.nature.com/doifinder/10.1038/nrc3184

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM & Taipale J (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4:** 1–20 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4362205&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2015]

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K & Lyon GJ (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5:** 28 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3706896%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract

Orci L, Baetens D, Rufener C, Amherdt M, Ravazzola M, Studer P, Malaisse-Lagae F & Unger RH (1976) Hypertrophy and hyperplasia of somatostatin-containing D-cells in diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **73:** 1338–42 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=430269&tool=pmcentrez&rendertype=abstract

Ottaviani E (2011) Immunocyte: the invertebrate counterpart of the vertebrate macrophage. *Isj* **8:** 1–4

Parker NR, Hudson AL, Khong P, Parkinson JF, Dwight T, Ikin RJ, Zhu Y, Cheng ZJ, Vafaee F, Chen J, Wheeler HR & Howell VM (2016) Intratumoral heterogeneity identified at the epigenetic, genetic and transcriptional level in glioblastoma. *Sci. Rep.* **6:** 22477 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4778014&tool=pmcentrez&rendertype=abstract

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed B V., Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A & Bernstein BE (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.).* **344:** 1396–1401

Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S & Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9:** 171–81 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24385147 [Accessed July 16, 2014]

Porrello ER, Mahmoud AI, Simpson E, Hill JA, Richardson JA, Olson EN & Sadek HA (2011) Transient Regenerative Potential of the Neonatal Mouse Heart. *Science (80-.).* **331:** 1078–1080 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1200708

Poss KD (2002) Heart Regeneration in Zebrafish. *Science (80-.).* **298:** 2188–2190 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1077857

Roberts SA & Gordenin DA (2014) Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* **14:** 786–800 Available at: http://dx.doi.org/10.1038/nrc3816

Romer AI & Sussel L (2015) Pancreatic islet cell development and regeneration. *Curr. Opin. Endocrinol. Diabetes Obes.* **22:** 255–264 Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=01266029-201508000-00002

Roth G (2015) Convergent evolution of complex brains and high intelligence. *Philos. Trans. R. Soc. B Biol. Sci.* **370:** 20150049–20150049 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26554042

Rotman E & Seifert HS (2014) The Genetics of Neisseria Species. *Annu. Rev. Genet.* **48:** 405–431 Available at: http://www.annualreviews.org/doi/10.1146/annurev-genet-120213-092007

Roy N & Hebrok M (2015) Regulation of Cellular Identity in Cancer. *Dev. Cell* **35:** 674–84 Available at: http://dx.doi.org/10.1016/j.devcel.2015.12.001

Sanger F (2001) The early days of DNA sequences. *Nat Med* **7:** 267–268 Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11231611

Sawai CM, Babovic S, Upadhaya S, Knapp DJHF, Lavin Y, Lau CM, Goloborodko A, Feng J, Fujisaki J, Ding L, Mirny LA, Merad M, Eaves CJ & Reizis B (2016) Hematopoietic Stem Cells Are the Major Source of Multilineage Hematopoiesis in Adult Animals. *Immunity* **45:** 597–609 Available at: http://dx.doi.org/10.1016/j.immuni.2016.08.007

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P & Odom DT (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (80-.).* **328:** 1036–40 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3008766&tool=pmcentrez&rendertype=abstract [Accessed October 31, 2012]

Schübeler D (2015) Function and information content of DNA methylation. *Nature* **517:** 321–326 Available at: http://www.nature.com/doifinder/10.1038/nature14192

Schuster-Böckler B & Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488:** 504–507 Available at: http://www.nature.com/doifinder/10.1038/nature11273

Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann S a, Stegle O, Marioni JC & Buettner F (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* Available at: http://www.ncbi.nlm.nih.gov/pubmed/26142758 [Accessed July 6, 2015]

Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, Smith DM, Kasper M, Ämmälä C & Sandberg R (2016) Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24:** 593–607 Available at: http://linkinghub.elsevier.com/retrieve/pii/S1550413116304363

Sharif J, Endo TA, Toyoda T & Koseki H (2010) Divergence of CpG island promoters : A consequence or cause of evolution ? **1:** 545–554

Sheffield NC & Bock C (2015) LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics***:** btv612 Available at: http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btv612

Shen H & Laird PW (2013) Interplay between the cancer genome and epigenome. *Cell* **153:** 38–55 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23540689 [Accessed May 23, 2013]

Smallwood S a, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W & Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11:** 6–11 Available at: http://www.nature.com/doifinder/10.1038/nmeth.3035 [Accessed July 21, 2014]

Solnica-Krezel L & Sepich DS (2012) Gastrulation: Making and Shaping Germ Layers. *Annu. Rev. Cell Dev. Biol.* **28:** 687–717 Available at: http://www.annualreviews.org/doi/10.1146/annurev-cellbio-092910-154043

Song J, Rechkoblit O, Bestor TH & Patel DJ (2011) Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. *Science (80-.).* **331:** 1036–1040 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4689315&tool=pmcentrez&rendertype=abstract

Sottoriva A, Spiteri I, Piccirillo SGM, Touloumis A, Collins VP, Marioni JC, Curtis C, Watts C & Tavaré S (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **110:** 4009–14 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3593922&tool=pmcentrez&rendertype=abstract

Stoeck T (2005) Cellular identity of an 18S rRNA gene sequence clade within the class Kinetoplastea: the novel genus Actuariola gen. nov. (Neobodonida) with description of the type species Actuariola framvarensis sp. nov. *Int. J. Syst. Evol. Microbiol.* **55:** 2623–2635 Available at: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.63769-0

Strahl BD & Allis CD (2000) The language of covalent histone modifications. *Nature* **403:** 41–45

Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, et al (2009) Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* **10:** 459–466 Available at: http://dx.doi.org/10.1016/S1470-2045(09)70025-7

Sulston JE & Horvitz HR (1977) Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev. Biol.* **56:** 110–56 Available at: http://www.sciencedirect.com/science/article/pii/0012160677901580

Sur I & Taipale J (2016) The role of enhancers in cancer. *Nat. Rev. Cancer* **16:** 483–493 Available at: http://www.nature.com/doifinder/10.1038/nrc.2016.62%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27364481

Suva M-L, Riggi N, Janiszewska M, Radovanovic I, Provero P, Stehle J-C, Baumer K, Le Bitoux M-A, Marino D, Cironi L, Marquez VE, Clement V & Stamenkovic I (2009) EZH2 Is Essential for Glioblastoma Cancer Stem Cell Maintenance. *Cancer Res.* **69:** 9211–9218 Available at: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-09-1622

Suzuki MM & Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9:** 465–76 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18463664 [Accessed October 26, 2012]

Taft RJ, Pheasant M & Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays* **29:** 288–299

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L & Rao A (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (80-.).* **324:** 930–5 Available at: http://www.sciencemag.org/content/324/5929/930.short

Teng G & Papavasiliou FN (2007) Immunoglobulin Somatic Hypermutation. *Annu. Rev. Genet.* **41:** 107–120 Available at: http://www.annualreviews.org/doi/10.1146/annurev.genet.41.110306.130340

Thorel F, Népote V, Avril I, Kohno K, Desgraz R, Chera S & Herrera PL (2010) Conversion of adult pancreatic α-cells to β-cells after extreme β-cell loss. *Nature* **464:** 1149–1154 Available at: http://www.nature.com/doifinder/10.1038/nature08894

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA & Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509:** 371–375 Available at: http://dx.doi.org/10.1038/nature13173

Treutlein B, Lee QY, Camp JG, Mall M, Koh W, Shariati SAM, Sim S, Neff NF, Skotheim JM, Wernig M & Quake SR (2016) Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534:** 391–395 Available at: http://dx.doi.org/10.1038/nature18323

Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, Campos C, Fabius AWM, Lu C, Ward PS, Thompson CB, Kaufman A, Guryanova O, Levine R, Heguy A, Viale A, Morris LGT, Huse JT, Mellinghoff IK & Chan TA (2012) IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483:** 479–483 Available at: http://www.nature.com/doifinder/10.1038/nature10866

Veillard A, Datlinger P & Laczik M (2016) Diagenode ® Premium RRBS technology : cost-effective DNA methylation mapping with superior coverage. *Nat. Publ. Gr.* **13:** i–ii Available at: http://dx.doi.org/10.1038/nmeth.f.391

Venkatesh P, Panyutin I, Remeeva E, Neumann R & Panyutin I (2016) Effect of Chromatin Structure on the Extent and Distribution of DNA Double Strand Breaks Produced by Ionizing Radiation; Comparative Study of hESC and Differentiated Cells Lines. *Int. J. Mol. Sci.* **17:** 58 Available at: http://www.mdpi.com/1422-0067/17/1/58

Verhaak RGW (2009) An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17:** 98–110 Available at: http://dx.doi.org/10.1016/j.ccr.2009.12.020

Vickaryous MK & Hall BK (2006) Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev.* **81:** 425 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16790079

Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, Griesbeck M, Butler A, Zheng S, Lazo S, Jardine L, Dixon D, Stephenson E, Nilsson E, Grundberg I, McDonald D, Filby A, Li W, De Jager PL, Rozenblatt-Rosen O, et al (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356:** eaah4573 Available at: http://www.sciencemag.org/lookup/doi/10.1126/science.aah4573

Visvader JE (2011) Cells of origin in cancer. *Nature* **469:** 314–22 Available at: http://www.ncbi.nlm.nih.gov/pubmed/21248838

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA & Kinzler KW (2013) Cancer Genome Landscapes. *Science (80-.).* **339:** 1546–1558 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1235122

Wagner A, Regev A & Yosef N (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34:** 1145–1160 Available at: http://www.nature.com/doifinder/10.1038/nbt.3711

Waitkus MS, Diplas BH & Yan H (2016) Isocitrate dehydrogenase mutations in gliomas. *Neuro. Oncol.* **18:** 16–26

Wang K, Li M & Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38:** e164

Wang YJ, Schug J, Won K-J, Liu C, Naji A, Avrahami D, Golson ML & Kaestner KH (2016) Single cell transcriptomics of the human endocrine pancreas. *Diabetes***:** db160405 Available at: http://www.ncbi.nlm.nih.gov/pubmed/27364731

Weller M, Stupp R, Reifenberger G, Brandes AA, van den Bent MJ, Wick W & Hegi ME (2010) MGMT promoter methylation in malignant gliomas: ready for personalized medicine? *Nat. Rev. Neurol.* **6:** 39–51 Available at: papers3://publication/doi/10.1038/nrneurol.2009.197

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI & Young RA (2013) Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153:** 307–319 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867413003929

Wilson A, Laurenti E, Oser G, van der Wath RC, Blanco-Bose W, Jaworski M, Offner S, Dunant CF, Eshkind L, Bockamp E, Lió P, MacDonald HR & Trumpp A (2008) Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell* **135:** 1118–1129 Available at: http://linkinghub.elsevier.com/retrieve/pii/S009286740801386X

Woese C (1998) The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* **95:** 6854–9 Available at: http://www.ncbi.nlm.nih.gov/pubmed/9618502

Woese CR & Fox GE (1977) The concept of cellular evolution. *J. Mol. Evol.* **10:** 1–6 Available at: http://www.ncbi.nlm.nih.gov/pubmed/903983

Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, Murphy AJ, Yancopoulos GD, Lin C & Gromada J (2016) RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab.* **24:** 608–615 Available at: http://dx.doi.org/10.1016/j.cmet.2016.08.018

Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, Yao B, Zhang F, Jin P, Wu H & Qin ZS (2015) Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res.***:** 1–10 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25722376 [Accessed March 3, 2015]

Yubuki N, Edgcomb VP, Bernhard JM & Leander BS (2009) Ultrastructure and molecular phylogeny of Calkinsia aureus: cellular identity of a novel clade of deep-sea euglenozoans with epibiotic bacteria. *BMC Microbiol.* **9:** 16 Available at: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-9-16

Yui MA & Rothenberg E V (2014) Developmental gene networks : a triathlon on the course to T cell identity. *Nat. Publ. Gr.* **14:** 529–545 Available at: http://dx.doi.org/10.1038/nri3702

Zeisel a., Manchado a. BM, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C, Castelo-Branco G, Hjerling-Leffler J & Linnarsson S (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).* **2:** 1–33 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aaa1934 [Accessed February 21, 2015]

Zemach A, McDaniel IE, Silva P & Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (80-.).* **328:** 916–9 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20395474 [Accessed October 25, 2012]

Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A & Meissner A (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500:** 477–481 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3821869&tool=pmcentrez&rendertype=abstract

Curriculum Vitae

Personal information

Name: Johanna Klughammer
Born: 28.06.1988, Theilheim, Germany
Nationality: German
Address: Bergsteiggasse 7/1/10, 1170 Vienna
E-Mail: jklughammer@cemm.oeaw.ac.at

Education

- **PhD cand.: Computational epigenomics**, *CeMM/ Medical university Vienna, Austria* — 2012-present
  - Thesis: *Epigenomic and transcriptional determination of cellular identity.*
  - Expected defense: 2017
  - Advised by: Christoph Bock
- **MSc: Biomedicine,** *University of Würzburg, Germany* — 2010-2012, Grade:1.1 (excellent)
  - Thesis: *Genome-wide assessment of within-host evolution of Neisseria meningitidis during invasive infection.*
  - Advised by: Christoph Schoen, Tobias Müller, Marcus Dittrich
- **BSc: Biomedicine,** *University of Würzburg, Germany* — 2007-2010, Grade:1.3 (excellent)
  - Thesis: *Functional characterization of Dmrt1a and Dmrt1bY in Medaka.*
  - Advised by: Manfred Schartl, Amaury Herpin

Fellowships

- 2014-2016: DOC-fellowship of the Austrian Academy of Sciences
- 2014-2016: PhD-fellowship of the German National Academic Foundation
- 2007-2012: Student's-fellowship of the German National Academic Foundation

Presentations, Posters, and Peer review

- Poster presentation: GBMatch, MASAMB, Vienna, 2017
- Oral presentation: GBMatch, Society Of Austrian Neurooncology, Salzburg, 2017
- Reviewer: PLOS Computational Biology, 2016
- Poster presentation: RefFreeDMA, Biology of Genomes Conference, CSH/ NY, 2016
- Oral presentation: HumanIslet, Scientific PhD-conference of the German National Academic Foundation, Bonn, 2015
- Oral presentation: CompEpi, OeAW fellows meeting, Vienna, 2015
- Oral presentation: RefFreeDMA, Scientific PhD-conference of the German National Academic Foundation, Bonn, 2014
- Oral presentation: CompEpi, CeMM-scientific advisory board meeting, Vienna, 2014
- Poster presentation: Within-host evolution in Neisseria meningitidis, XVIIIth International Pathogenic Neisseria Conference, Würzburg, 2012

Publications

*Shared first authorship

1. Klughammer J*, Kiesel B*, Roetzer T, Fortelny N, Kuchler A, Datlinger P, Peter N, Nenning K, Furtner J, Nowosielski M, Augustin M, Mischkulnig M, Ströbel T, Moser P, Freyschlag CF, Kerschbaumer J, Thomé C, Grams AE, Stockhammer G, Kitzwoegerer M, Oberndorfer S, Marhold F, Weis S, Trenkler J, Buchroithner J, Pichler J, Haybaeck J, Krassnig S, Ali KM, von Campe G, Payer F, Sherif C, Preiser J, Hauser T, Winkler PA, Kleindienst W, Würtz F, Brandner-Kokalj T, Stultschnig M, Schweiger S, Dieckmann K, Preusser M, Langs G, Baumann B, Knosp E, Widhalm G, Marosi C, Hainfellner JA, Woehrer A, Bock C. **The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space.** Submitted.

2. Sheffield NC, Pierron G, Klughammer J, Datlinger P, Schönegger A, Schuster M, Hadler J, Surdez D, Guillemot D, Lapouble E, Freneaux P, Champigneulle J, Bouvier R, Walder D, Ambros IM, Hutter C, Sorz E, Amaral AT, de Álava E, Schallmoser K, Strunk D, Rinner B, Liegl-Atzwanger B, Huppertz B, Leithner A, de Pinieux G, Terrier P, Laurence V, Michon J, Ladenstein R, Holter W, Windhager R, Dirksen U, Ambros PF, Delattre O, Kovar H, Bock C, Tomazou EM. **DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma.** Nat Med. 2017 Mar;23(3):386-395. doi:10.1038/nm.4273.

3. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. **Pooled CRISPR screening with single-cell transcriptome readout.** Nat Methods. 2017 Mar;14(3):297-301. doi:10.1038/nmeth.4177.

4. Klughammer J, Dittrich M, Blom J, Mitesser V, Vogel U, Frosch M, Goesmann A, Müller T, Schoen C. **Comparative Genome Sequencing Reveals Within-Host Genetic Changes in Neisseria meningitidis during Invasive Disease.** PLoS One. 2017 Jan 12;12(1):e0169892. doi:10.1371/journal.pone.0169892.

5. Li J, Casteels T, Frogne T, Ingvorsen C, Honoré C, Courtney M, Huber KV,Schmitner N, Kimmel RA, Romanov RA, Sturtzel C, Lardeau CH, Klughammer J, Farlik M, Sdelci S, Vieira A, Avolio F, Briand F, Baburin I, Májek P, Pauler FM, Penz T, Stukalov A, Gridling M, Parapatics K, Barbieux C, Berishvili E, Spittler A, Colinge J, Bennett KL, Hering S, Sulpice T, Bock C, Distel M, Harkany T, Meyer D, Superti-Furga G, Collombat P, Hecksher-Sørensen J, Kubicek S. **Artemisinins Target GABA$_A$ Receptor Signaling and Impair α Cell Identity**. Cell. 2016 Jan 12;168(1-2):86-100.e15. doi:10.1016/j.cell.2016.11.010.

6. Farlik M*, Halbritter F*, Müller F*, Choudry FA, Ebert P, Klughammer J, Farrow S, Santoro A, Ciaurro V, Mathur A, Uppal R, Stunnenberg HG, Ouwehand WH, Laurenti E, Lengauer T, Frontini M, Bock C. **DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation**. Cell Stem Cell. 2016 Dec 1;19(6):808-822. doi:10.1016/j.stem.2016.10.019.

7. Tschurtschenthaler M*, Kachroo P*, Heinsen FA, Adolph TE, Rühlemann MC, Klughammer J, Offner FA, Ammerpohl O, Krueger F, Smallwood S, Szymczak S, Kaser A, Franke A. **Paternal chronic colitis causes epigenetic inheritance of susceptibility to colitis.** Sci Rep. 2016 Aug 19;6:31640. doi:10.1038/srep31640.

8. Mass E*, Ballesteros I*, Farlik M*, Halbritter F*, Günther P, Crozet L, Jacome-Galarza CE, Händler K, Klughammer J, Kobayashi Y, Gomez-Perdiguero E, Schultze JL, Beyer M, Bock C, Geissmann F. **Specification of tissue-resident macrophages during organogenesis.** Science. 2016 Sep 9;353(6304). pii: aaf4238. doi:10.1126/science.aaf4238.

9. Li J*, Klughammer J*, Farlik M*, Penz T*, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S. **Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types.** EMBO Rep. 2016 Feb;17(2):178-87. doi:10.15252/embr.201540946

10. Klughammer J, Datlinger P, Printz D, Sheffield NC, Farlik M, Hadler J, Fritsch G, Bock C. **Differential DNA Methylation Analysis without a Reference Genome.** Cell Rep. 2015 Dec 22;13(11):2621-33. doi:10.1016/j.celrep.2015.11.024.

11. Farlik M*, Sheffield NC*, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, Bock C. **Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics.** Cell Rep. 2015 Mar 3;10(8):1386-97. doi:10.1016/j.celrep.2015.02.001.

12. Herpin A, Adolfi MC, Nicol B, Hinzmann M, Schmidt C, Klughammer J, Engel M, Tanaka M, Guiguen Y, Schartl M. **Divergent expression regulation of gonad development genes in medaka shows incomplete conservation of the downstream regulatory network of vertebrate sex determination.** Mol Biol Evol. 2013 Oct;30(10):2328-46. doi:10.1093/molbev/mst130.