Genomic and molecular maps of stemness and malignant regulatory signatures.

Gennadi V. Glinsky[1,2]

[1] Institute of Engineering in Medicine

University of California, San Diego

9500 Gilman Dr. MC 0435

La Jolla, CA 92093-0435, USA

[2]OncoSCAR, Inc.

Portland, OR 97207

Correspondence: gglinskii@ucsd.edu

Web: http://iem.ucsd.edu/people/profiles/guennadi-v-glinskii.html

Key words: malignant regulatory signatures; stem cell-associated retroviral sequences; retrotransposition; human embryogenesis; cancer survival genes; cancer driver genes; multi-lineage markers expressing human embryonic cells.

Abstract

Repetitive DNA sequences (repeats) colonized two-third of human genomes and a majority of repeats comprised of transposable genetic elements (TE). Evolutionary distinct categories of TE represent nucleic acid sequences that are repeatedly copied from and pasted into chromosomes at multiple genomic locations and acquired a multitude of regulatory functions. Here, genomics-guided maps of stemness regulatory signatures were drawn to dissect the contribution of TE to clinical manifestations of malignant phenotypes of human cancers. From patients' and physicians' perspectives, the clinical definition of a tumor's malignant phenotype could be restricted to the early diagnosis of sub-types of malignancies with the increased risk of existing therapy failure and high likelihood of death from cancer. It is the viewpoint from which the understanding of stemness and malignant regulatory signatures is considered in this contribution. Analyses from this perspective of experimental and clinical observations revealed the pivotal role of human stem cell-associated retroviral sequences (SCARS) in the origin and pathophysiology of clinically-lethal malignancies. SCARS were defined as the evolutionary- and biologically-related family of genomic regulatory sequences, the principal physiological function of which is to create and maintain the stemness phenotype during human preimplantation embryogenesis. For cell differentiation to occur, SCARS expression must be silenced and SCARS activity remains repressed in most terminally-differentiated human cells performing specialized functions in the human body. De-repression and sustained activity of SCARS results in various differentiation-defective phenotypes. One of the most prominent tissue- and organ-specific clinical manifestations of sustained SCARS activities is diagnosed as a pathological condition defined by a consensus of morphological, molecular, and genetic examinations as the malignant growth. Contemporary evidence are acquired, analyzed, and reported defining both reliable diagnostic tools and druggable molecular targets readily amenable for diagnosis and efficient therapeutic management of clinically-lethal malignancies: detection of high-fidelity

molecular signals of continuing SCARS activities in association with genomic regulatory

networks of thousands' functionally-active enhancers triggering engagements of down-stream

phenotype-altering genetic loci.

Introduction

The idea that malignant growth originates from stem cells is more than a quarter century old [1]. It was revived at the beginning of 21st century as the cancer stem cell theory [2; 3], which became one of dominant concepts of the contemporary cancer research. One of the key principles of the cancer stem cell theory is that a single cancer stem cell is sufficient to regrow a malignant tumor fully recapitulating morphological, molecular, genomic, and biological features of the parental tumor. Consequently, the theory predicts that cancer cannot be eradicated unless cancer stem cell-targeting therapies [4] will eliminate all cancer stem cells. This postulate is believe to be true because if even a single cancer stem cell would escape the therapeutic assault, it will continue to fuel the malignant growth. However, some fundamental clinical realities seem not necessarily fully compatible with the uniformly simplistic view of the human cancer origin and pathogenesis. First, tumors arising in the same organ are not equivalent in clinical responses to therapies, which could be correlated to their genetic and molecular features. Second, the clinical prognosis related to the organ of cancer origin is markedly different for cancers diagnosed in different organs even at the early stages. Third, in many instances, the clinical cure of malignant tumors has been achieved by the first-line cancer therapies, which are not specifically designed to target cancer stem cells.

On a parallel track, technological advances enabled genome-wide gene expression profiling analyses of human malignancies making a reality the search for gene expression signatures of clinically-lethal malignancies, thus, looking for statistically-significant gene expression correlates of increased likelihood of existing therapy failure and death from cancer. Historically, the theory defining a genomic link between degrees to which a malignancy recapitulates gene expression profiles of stem cells and clinical phenotypes of increased likelihood of therapy failure and death from cancer is originated from the discovery of the death-from-cancer gene expression signature [5]. This genomic connectivity between the phenotypes of resemblance to stemness

4

and high likelihood of death from cancer was initially documented for cancer patients diagnosed with 12 distinct types of human malignances [5]. Observations reported in the original contributions [5; 6] and follow-up studies (reviewed in 7) directly implicated sustained activation of the Polycomb Group (PcG) Proteins chromatin silencing pathway [8], specifically, the *BMI1* genes, as the principal genomic contributor defining these associations [5; 7]. Collectively, these observations formed the foundation for a concept stating that malignant clinical behaviors of human cancers are governed by stemness genomic laws [5-7; 9-12]. The universal nature of the genomic connectivity between the degree of resemblance to stemness and the extent of malignant behavior of a tumor was validated in numerous experimental cancer models, including transgenic mouse models facilitating implementation of the mouse/human translational genomics approach [5; 13; 14]; clinically-relevant orthotopic xenograft models of human cancers and xenograft-derived cancer cell lines, including blood-borne metastasis precursor cells [6; 15-19]. Mechanistic roles of genes essential for functional integrity of PcG chromatin silencing pathway were demonstrated using targeted genetic interference approaches [13; 20] and gene-specific small molecule therapeutics [21]. Overall, multiple studies have shown that BMI1 inhibition confer therapeutic effects on glioblastoma multioforme, colorectal and breast cancers, as well as chemoresistant ovarian, prostate, pancreatic, and skin cancers [21-24].

However, the major limitation of these and many other early studies was the lack of sufficient understanding of the genomic and molecular underpinning of the stemness phenotype as it emerges during human preimplantation embryogenesis. Remarkable advances in single cell expression profiling analyses of human preimplantation embryos closed this knowledge gap and provided the opportunity to address this limitation. Collectively, these advances facilitated the discovery of stem cell-associated retroviral sequences, which act as the master genomic regulatory elements driving the creation of stemness phenotype in human embryos and may be responsible for its emergence in sub-sets of human malignancies diagnosed in multiple organs.

One of the most promising avenues toward understanding of genomic and molecular underpinning of malignant regulatory signatures has its origin in the fundamental advances revealing principal regulatory elements of genomic and molecular pathways of the stemness phenotype creation and maintenance during human embryonic development. Remarkable achievements of single cell genomics of human preimplantation embryogenesis facilitated the emergence of the concept of SCARS as both intrinsic and integral components of human-specific genomic regulatory networks (GRNs), the main biological function of which is to enable the creation and maintenance of stemness features in human embryonic cells.

Several independent yet complementary approaches were utilized to discern the potential impacts of SCARS, other families of HSRS, and functionally-active hESC enhancers on physiological and pathological phenotypes of Modern Humans.

First, comprehensive lists of genes comprising down-stream targets of corresponding regulatory loci of interest have been identified.

Second, multiple gene expression signatures (GES) linked to regulatory loci of interest were deconvoluted from large sets of down-stream target genes.

Third, GSEA using an extensive collection of genomic databases have been carried out to statistically link down-stream target genes with phenotypic traits, morphological features, and physiological and pathological conditions.

Fourth, disease type-specific sets of genes were identified and assembled into panels of GES for follow-up interrogations of their potential pathophysiological impact and translational utilities.

Fifth, multiple human-specific genomic regulatory networks (GRNs) have been identified operating in developmentally and physiologically distinct human tissues and cells to dissect associations of down-stream target genes with defined human-specific GRNs.

The task of identification of down-stream-target genes was achieved using either overexpression of regulatory loci or genetic interference approaches, including shRNA-mediated interference and CRISPR/Cas9-guided epigenetic silencing (Wang et al., 2014; Theunissen et al., 2016; Pontis et al., 2019; Glinsky et al., 2018; Durruthy-Durruthy et al., 2016 a; b). Alternatively, proximity placement analyses of regulatory elements and down-stream targets were performed employing the GREAT algorithm (McLean et al., 2010, 2011).

Examples of the interrogated human-specific GRNs include the following data sets:

i)   Great Apes' whole-genome sequencing-guided human-specific insertions and deletions (Kronenberg et al., 2018);

ii)   Genome-wide analysis of retrotransposon's transcriptome in postmortem samples of human dorsolateral prefrontal cortex (Guffanti et al., 2018);

iii)   shRNA-mediated silencing of LTR7/HERVH retrovirus-derived long non-coding RNAs in hESC (Wang et al., 2014);

iv)   Single-cell expression profiling analyses of human preimplantation embryos (Glinsky et al., 2018);

v)   Network of genes associated with regulatory transposable elements (TE) operating in naïve and primed hESC (Theunissen et al., 2016; Pontis et al., 2019);

vi)   Pluripotency-related network of genes manifesting concordant expression changes in human fetal brain and adult neocortex (Glinsky, 2017);

vii)   Network of genes governing human neurogenesis in vivo (Nowakowski et al., 2017);

viii)   Network of genes differentially expressed during human corticogenesis in vitro (van de Leemput et al., 2014).

Thus, selected for these analyses human-specific GRNs appear to function in a developmentally and physiologically diverse spectrum of human cells that are biologically and anatomically highly relevant to manifestations of human-specific phenotypes ranging from preimplantation embryos to adult dorsolateral prefrontal cortex (Glinsky, 2019; 2020).

Results

Functional, structural, genetic, and molecular definitions of stem cell-associated retroviral sequences (SCARS).

The term stem cell-associated retroviral sequences (SCARS) refers to the defined set of genomic regulatory sequences sustained expression of which is essential for acquisition and maintenance of stemness phenotype [7; 25; 26]. The canonical definition of "stemness" in reference to human Embryonic Stem Cells (hESC) implies a combination of the phenotypic features of immortality/self-renewal/asymmetrical division/pluripotency. Single cell expression profiling-guided deconvolution of a developmental timeline of human preimplantation embryos enabled the discovery of human embryonic Multi-Lineage Markers Expressing cells (MLME cells), emergence of which during human embryogenesis precedes lineage segregation events and subsequent creation of hESC [27]. Specific members of SCARS termed human pluripotency-associated transcripts (HPATs) have been implicated in the creation of the MLME cells [27]. It has been hypothesized that definition of the "stemness" phenotype for the human MLME cells should be expanded to include the totipotency feature and the human MLME cells could be defined biologically as the pan-lineage precursor cells [27].

For cell differentiation to occur, the expression of SCARS must be silenced: hESC fails to properly differentiate in response to differentiation-inducing cues if SCARS expression is maintained and resulting cells display differentiation-defective phenotypes [28; 29]. It has been

8

suggested that de-repression and sustained re-activation of SCARS expression in association with continuous activation of down-stream genomic regulatory targets (collectively defined as activation of SCARS-associated genomic regulatory networks) is the hallmark of therapy-resistant clinically-lethal malignancies with clinical phenotypes of increased risk of therapy failure and high likelihood of death from cancer [7; 25; 26]. Evolutionary, SCARS are belong to the exceedingly large class of genomic sequences originated from transposable elements (TE) and comprising nearly two-third of human genome. Specifically, in hESC and human preimplantation embryos SCARS represent a functionally-related and structurally-defined sub-set of TE-derived regulatory sequences originated from LTR7/HERV-H, LTR5_Hs/HERV-K, and recently implicated SVA-D retrotransposons [30-32], the set of which was further narrowed by restrictions to human-specific (unique-to-humans) genomic regulatory sequences [7; 25; 26; 33-39].

A range of genetic and molecular definitions of SCARS directly linked to a stemness state extends to different classes of regulatory DNA sequences (transcription factor-binding sites; functional enhancer elements; alternative promoters), donors of splicing sites, and non-coding RNA molecules. Precise mapping of individual transcriptionally-active genomic loci which generated RNA molecules from repetitive sequences (repeats), including highly diverse families of transposable elements (TE) and human endogenous retroviruses (HERV) - encoded sequences, became possible only recently. Advances in RNAseq technology and bioinformatics approaches to data retrieval, processing, and analyses, including implementation of *de novo* transcriptome assembly protocols, facilitated identification of hundreds thousands of TE-encoded RNA molecules precisely mapped to corresponding transcriptionally active genomic loci in human dorsolateral prefrontal cortex [40] and across the spectrum of all major human cancer types [41]. Using the pan-cancer *de novo* transcript assembly approach, the remarkable complexity and ubiquitous nature of transcripts encoded by endogenous retroviral elements

(EREs) were uncovered in human malignancies of distinct origins and diverse spectrum of anatomical locations [41]. It has been reported that thousands of transcripts overlapping with regulatory long terminal repeats (LTRs) derived from endogenous retroviruses were expressed in a cancer-specific manner in at least one or several related cancer types [41]. Several of these cancer-specific LTR-harboring transcripts represent relatively large RNA molecules exceeding 50K nucleotides, perhaps, reflecting the read-through transcriptional activity in cancer cells due to the extensive chromatin reprogramming. Notably, cancer-specific RNA molecules derived from individual SCARS loci representing LTR7/HERV-H and LTR5_Hs/HERV-K families accounted for 31% of all reported cancer-specific LTR element-overlapping transcripts that are expressed in more than one cancer type. These cancer-specific LTR-harboring RNA molecules appear to affect the expression of disease-relevant genes and to produce previously unknown cancer-specific antigenic peptides [41]. Therefore, it is now feasible to unequivocally map SCARS-harboring RNA molecules to specific transcriptionally-active genetic loci encoding these transcripts.

Global DNA methylation reprogramming and SCARS activity contribute to creation of telomerase-positive MLME cells during human preimplantation embryogenesis.

One of the principal molecular functions of activated SCARS is illustrated by their biological activities attributed to non-coding RNA (ncRNA) molecules transcribed from regulatory DNA segments harboring SCARS. Importantly, manifestations SCARS biological activities have been demonstrated for ncRNAs derived from individual genomic loci [27; 42; 43] and in human embryos SCARS activity has been associated with the creation of telomerase-positive cells co-expressing genetic markers of all embryonic lineages [27]. These telomerase-positive Multi-Lineage Markers Expressing (MLME) cells have been identified employing single cell expression profiling analyses of viable human blastocysts and hundreds of individual cells recovered from preimplantation human embryos [27; 42; 43]. Creation of cells in part resembling

gene expression features of MLME cells was recapitulated in genetic engineering experiments, in which individual SCARS-encoded RNAs termed Human Pluripotency-Associated Transcripts (HPATs) were over-expressed in human cells [27; 42; 43]. These observations support the hypothesis that SCARS activation in human embryos may have contributed to the creation of MLME cells.

The summary of the multi-step validation protocol of human embryonic Multi-Lineage Markers Expressing (MLME) cells is shown in the Table 1. The MLME phenotype was assigned to individual telomerase-positive cells that co-expressed at least 6 genetic markers of the Epiblast (EPI) lineage; 7 genetic markers of the Trophectoderm (TE) lineage; and 4 genetic markers of the Primitive endoderm (PE) lineage; and cells must express all three main master pluripotency transcription factors (*OCT4, NANOG, SOX2*). First, the expression levels of 58 genetic markers of human embryonic lineages were considered individually in a particular single cell by comparing the expression values of the markers in a given cell and the median expression value of the marker in the population of single cells of human embryos as previously reported [27; 44]. The marker was considered expressed when the expression value in a cell exceeds the median expression value. The discovery set of 58 genetic markers of human embryonic lineages was utilized in these experiments and based on the above criteria a total of 135 MLME cells were selected from 839 telomerase-positive human embryonic cells. The discovery set of 58 genetic markers of human embryonic lineages was reported elsewhere [27; 45; 46]. Next, independent sets of lineage-specific markers comprising of top 100 individual genetic markers for each embryonic lineage that were utilized for validation of the MLME phenotype in each individually-selected cell. The validation sets of lineage-specific genetic markers of human embryonic lineages were reported elsewhere [44]. To assess the statistical significance of the enrichment of the lineage-specific genetic markers in the MLME cells, p values were estimated using the hypergeometric distribution test. Results of these analyses revealed statistically

significant enrichment of genes representing genetic markers of three main embryonic lineages among genes up-regulated in human embryonic MLME cells (Table 1). In agreement with the hypothesis that activities of SCARS contribute to creation of MLME cells, SCARS appear to affect expression of two-third of genes (8,374 of 12,735 genes; 66%) expression of which distinguishes MLME cells from other cells in preimplantation human embryos. Notably, SCARS activity affects expression of a dominant majority (84.1%) of genes up-regulated in human embryonic MLME cells, while expression of only a minor fraction of genes down-regulated in MLME cells (13.4%) appears affected by SCARS.

Zygote-to-embryo transition is accompanied by dramatic DNA methylation reprogramming which is governed by the placeholder nucleosome positioning [47]. Newly established genome-wide dynamics of the chromatin accessibility landscape and concurrent changes of promoter methylation states affect expression of thousands genes and results in embryonic genome activation [31; 48]. Importantly, DNase I hypersensitive site (DHS) sequencing revealed that human transposons SVA and HERV-K harbor DHSs and are highly expressed in early human embryos, but not in differentiated tissues [31]. Analyses of genes comprising GES of human embryonic MLME cells revealed that DNA methylation reprogramming may have contributed to the creation and maintenance of the MLME phenotype in human preimplantation embryos (Figure 1). Collectively, observed in MLME cells gene expression changes of methyltransferases would cause marked reprogramming of genome-wide DNA methylation profiles by erasing the pre-existing cytosine methyl marks and establishing *de novo* methylation patterns (Figure 1A). Concurrently diminished expression of genes encoding primate-specific zinc finger proteins, in particular, *ZNF534* and *ZNF91* genes, would relieve the repressive chromatin from SCARS loci and facilitate activation of SCARS expression (Figure 1B). Consistently, during transition from the oocyte to the morula stage of human preimplantation embryogenesis, promoters of genes comprising the MLME GES shift from nearly exclusively

homogenously closed (silenced) states to predominantly homogenously open (active) states (Figure 2). The predominantly homogenously open promoter states of genes comprising the MLME signature are maintained in human embryonic cells of the ICM, TE, and hESC (Figure 2).

SCARS represent both intrinsic and integral components of human-specific genomic regulatory networks.

SCARS are predominantly primate-specific regulatory sequences that is they are common for Modern Humans and non-human primates [7]. However, sizable fractions of different SCARS families are represented by human-specific (unique-to-human) regulatory sequences. For example, 302 of 1222 (24.7%) full-length LTR7/HERV-H elements have been identified as candidate human-specific regulatory sequences, HSRS [7]. Species-specificity of SCARS is defined by the unique genomic coordinates of the insertions of corresponding parent transposons, which appear as segments of DNA present on human chromosomes and absent on chromosomes of non-human primates. Interestingly, 37.6% of highly active in hESC LTR7/HERV-H elements have been classified as HSRS [7]. This is contrast to only 19.8% LTR7/HERV-H that are inactive in hESC being identified as candidate HSRS (p < 0.0001). Therefore, globally SCARS should be viewed within the genomic regulatory context of other classes of HSRS [38].

Candidate HSRS comprise a coherent compendium of nearly one hundred thousand genomic regulatory elements, including 59,732 HSRS which are markedly distinct in their structure, function, and evolutionary origin [38] as well as 35,074 human-specific neuro-regulatory single nucleotide changes (hsSNCs) located in differentially-accessible (DA) chromatin regions during human brain development [39; 49]. Unified activities of HSRS may have contributed to development and manifestation of thousands human-specific phenotypic traits [39]. SCARS encoded by human endogenous retroviruses LTR7/HERV-H and LTR5_Hs/HERV-K as one of the significant sources of the evolutionary origin of HSRS [7; 25-27; 33-40], including human-

specific transcription factor binding sites (TFBS) for NANOG, OCT4, and CTCF [33; 37]. It was

interest to determine whether genes previously linked to multiple classes of HSRS, which were

identified without considerations of genes expression of which is regulated by SCARS, overlap

with SCARS-regulated genes. To this end, 13,824 genes associated with different classes of

HSRS were identified using the GREAT algorithm [38; 39], subjected to the GSEA, and

compared with the sets of SCARS-regulated genes (Figure 3) identified by shRNA interference

[50] and CRISR/Cas-guided epigenetic silencing experiments comparing regulatory networks of

naïve and primed hESC [30; 32]. These analyses revealed that SCARS appear to affect

expression of a majority (8,384 genes; 61%) of genes associated with different classes of HSRS

(Table 2; Supplemental Table S1), in agreement with the hypothesis that a large fraction of

SCAS-regulated genes represents an intrinsic component of human-specific genomic regulatory

networks. Consistently, SCARS affect expression of a majority of genes (5,389 of 8,405 genes;

64%) associated with neuro-regulatory hsSNCs [39]. Overall, the common gene set of

regulatory targets independently defined for HSRS, SCARS, and neuro-regulatory hsSNCs

comprises of 7,990 coding genes or 95% of all genes associated with neuro-regulatory hsSNCs

located in DA chromatin regions during human brain development [39].

Genes associated with HSRS and neuro-regulatory hsSNCs manifest a staggering breadth of

significant associations with morphological structures, physiological processes, and pathological

conditions of Modern Humans [39], indicating that a preponderance of human-specific traits

evolved under a combinatorial regulatory control of HSRS and neuro-regulatory loci harboring

hsSNCs. SCARS-regulated genes comprise a large fraction of these human-specific genomic

regulatory networks and represent an integral component of genomic regulatory wiring

governing human-specific features of early embryonic development.

One of the important questions is whether the patterns of significant associations with

physiological and pathological phenotypes observed for genes linked with HSRS, hsSNCs, and

SCARS are specific and not related to the size effects of relatively large gene sets subjected to the GSEA [39]. To address this questions, 42,847 human genes not linked by the GREAT algorithm with HSRS were randomly split into 21 control gene sets of various sizes ranging from 2,847 to 6,847 genes and subjected to the GSEA [39]. Importantly, no significant phenotypic associations were observed for 21 control gene sets, indicating phenotypic associations attributed to genes linked with HSRS, hsSNCs, and SCARS are not likely due to non-specific gene set size effects captured by the GSEA. These observations are highly consistent with the conclusion that a broad spectrum of significant phenotypic associations documented for genes linked with HSRS, neuro-regulatory hsSNCs, and SCARS reflects their bona fide impacts on physiological and pathological phenotypes of Modern Humans. It should be underscored that the efficient execution of these analytical experiments was greatly facilitated by the web-based utilities provided by the Enrichr Bioinformatics System Biology platform [51; 52].

Gene set enrichment analyses (GSEA) of 8,384 genes associated with HSRS, expression of which is regulated by LTR7Y/B and LTR5_Hs/SVA_D enhancers and HERVH lncRNAs.

GSEA on multiple genomics databases revealed remarkable breadth and depth of significant associations with physiological and pathological phenotypes of Modern Humans of 8,834 SCARS-regulated genes linked with multiple families of HSRS (Supplemental Text S1). Consistent with the established role of SCARS in human embryogenesis, SCARS-regulated genes are significantly enriched in human embryo and neuronal epithelium according to GSEA of the ARCHS4 Human Tissues database. Consistently, POU5F1 and PRDM14 master stem cell regulators were identified by GSEA of the ESCAPE stem cell-focused database as top up-stream regulators, while pathways in Cancer (KEGG 2019 Human database) and Axon Guidance (KEGG 2019 Mouse database) were scored as top significantly-enriched pathways.

GSEA of the Allan Brain Atlas database focused on up-regulated genes identified 590 human brain regions among significantly enriched records, while GSEA of the Allen Brain Atlas of

15

down-regulated genes identified 847 significant records (adjusted p-value < 0.05). Notably, seven of the top ten significantly enriched records among up-regulated genes identified the Dentate Gyrus, while remaining 3 of the top 10 records identified the Fields CA3 of stratum pyramidale and stratum lucidum of the hippocampus (Supplemental Text S1; Allan Brain Atlas database; up-regulated genes).

GSEA of the Virus MINT database comprising of human genes that encode proteins known to physically interact with viruses and viral proteins identified the Epstein-Barr virus as the top-scoring record, indicating that upon entry in human cells the Epstein-Barr virus-encoded proteins target proteins encoded by SCARS-regulated genes. Overall, expression of nearly 60% of all human genes encoding virus-interacting proteins (2,574 of 4,433 VIP-encoding genes; 58%) is regulated by SCARS.

GSEA of 2,846 genes associated with created de novo HSRS, expression of which is regulated by LTR7Y/B and LTR5_Hs/SVA_D enhancers and HERVH lncRNAs.

In human genome, there are 4,528 genes comprising putative regulatory targets of ~12,000 created *de novo* HSRS [38; 39]. Notably, SCARS regulate expression of 2,846 genes (63%) of all genes identified as candidate regulatory targets of created *de novo* HSRS. GSEA of genomics databases revealed numerous significant enrichment records linked with 2,846 SCARS-regulated genes, thus highlighting their potential impacts on human physiology and pathology (Supplemental Text S2).

Unexpectedly, GSEA of the ENCODE and ChEA Consensus transcription factors (TFs) from ChIP-X database identified androgen receptor (AR) as a top-scoring candidate upstream regulator. In agreement with the above observations, GSEA of the ARCHS4 Human Tissues database identified Neuronal epithelium, Human embryo, and Prefrontal cortex as top significantly-enriched records (Supplemental Text S2). Pathways in Cancer (KEGG 2019

Human database) and Axon Guidance (KEGG 2019 Mouse database) were identified as top significantly enriched pathways. Additionally, pathways of Integrins in angiogenesis (NCI-Nature 2016 database) and Integrin signaling (Panther 2016 database) were identified as top-scoring significantly-enriched pathways (Supplemental Text S2).

GSEA of the Jensen Tissues database identified 134 significantly enriched records indicating that SCARS-regulated genes associated with created *de novo* HSRS have been previously identified among genes comprising expression signatures of many human tissues. Other notable findings were revealed by the GSEA of the Human Phenotype Ontology database (81 significant records); the MGI Mammalian Phenotype 2017 database (309 significant records); the Allen Brain Atlas databases of up-regulated genes (284 significantly-enriched brain regions) and down-regulated genes (408 significantly-enriched brain regions).

SCARS-regulated genes appear significantly enriched among genes implicated in a broad spectrum of human common and rare diseases. GSEA of the Rare Diseases AutoRIF ARCHS4 Predictions database captured 353 significantly-enriched records of human rare disorders (Supplemental Text S2). GSEA of the Disease Perturbations from Gene Expression Omnibus (GEO) database of up-regulated genes identified 246 significant records, while interrogation of the Disease Perturbations from GEO database of down-regulated genes revealed 203 significantly-enriched records (Supplemental Text S2).

Lastly, according to GSEA of the Jensen Diseases database, a significant majority of SCARS-regulated genes associated with created *de novo* HSGRS (2008 of 2846 genes; 71%) have been implicated in development and clinical manifestations of multiple types of human cancers (Supplemental Text S2).

Inference of potential impacts of SCARS on development and clinical behavior of human malignancies.

SCARS activation hypothesis postulates the central role of a sustained activity of SCARS in acquisition and maintenance of stemness features in human cancer cells, clinical manifestations of which are reflected in high likelihood of therapy failure and death from cancer [7; 25; 26]. This intrinsic propensity to evade the malignancy eradication therapies is proposed to exist even if SCARS-activation driven cancer is diagnosed as the early stage disease based on established pathomorphological and molecular criteria.

Observations capturing the principal molecular, genetic, and biological features attributed to regulatory impacts of SCARS were made in experimental models of naïve and primed hESC, human induced pluripotent stem cells (iPSC), and human preimplantation embryogenesis. These experiments identified genes expression of which is significantly altered in human cells subjected to targeted genetic manipulations to achieve SCARS over-expression [27; 42; 43} and/or silencing using shRNA interference [42; 43; 50], CRISPR/Cas gene knockout technology [42] as well as CRISPR/Cas-guided epigenetic silencing of SCARS [32], thus facilitating identification of multiple gene expression signatures (GES) reflecting fine details of experimentally-defined SCARS-associated genomic regulatory networks.

Impacts of genes comprising distinct GES regulated by LTR7Y/B and LTR5_Hs/SVA_D enhancers and HERVH lncRNAs.

Potential biological relevance of several experimentally-defined GES comprising distinct panels of SCARS-regulated genes have been evaluated using Gene Set Enrichment Analyses (GSEA) across multiple genomic databases as previously described [38; 39]. These analytical experiments were executed using the web-based tools of the Enrichr Bioinformatics System Biology platform [51; 52]. To date, the following GES of SCARS-regulated networks in hESC are available for follow-up interrogations of their biological impacts and potential translational significance:

1. GES comprising a set of 1,141 genes that are regulated by both HERVH lncRNA and LTR5_Hs/SVA_D enhancers;

2. GES comprising a set of 3,063 genes regulated by both LTR7Y/B enhancers and HERVH lncRNA;

3. GES comprising a set of 1,477 genes regulated by both LTR7Y/B enhancers and HERVH lncRNA and manifesting concordant expression profiles;

4. GES comprising a set of 1,586 genes regulated by both LTR7Y/B enhancers and HERVH lncRNA and manifesting discordant expression profiles;

The up to date summary of the key findings for each of these four SCARS GES is reported in the Supplemental Text S3. Notably, GSEA of 1,141 genes that are regulated by both LTR5_Hs/SVA_D enhancers and HERV-H lncRNA facilitated identification and characterization of sub-sets of SCARS-regulated genes implicated in Parkinson's, autism, multiple types of cancer, and human embryonic development (Supplemental Text S3).

GSEA of the Jensen Diseases database revealed that a significant majority of genes regulated by both HERV-H lncRNA and LTR7Y/B enhancers (1905 of 3063 genes; 62%) have been implicated in development and clinical manifestations of multiple types of human cancer. Similarly, a significant majority of genes regulated by both HERV-H lncRNA and LTR7Y/B enhancers and manifesting concordant expression profiles (972 of 1477 genes; 66%) have been implicated in development and clinical manifestations of multiple types of cancer (Supplemental Text S3).

HSRS and SCARS regulate expression of a majority of cancer survival predictor genes and cancer driver genes.

One of the approaches to evaluation of potential impacts of SCARS on development and clinical manifestations of human malignancies could be the assessment of regulatory effects of SCARS

on cancer survival and cancer driver genes. To this end, analyses of 10,713 protein-coding genes expression changes of which are significantly associated with the increased likelihood of survival of cancer patients diagnosed with 17 major cancer types [53] and 460 cancer driver genes identified in 28 human cancer types [54] revealed that SCARS regulate a majority of either cancer survival genes or cancer driver genes (Tables 3; 4; Figure 4; Supplemental Text S4).

It has been observed (Table 3) that a prominent majority of human cancer survival predictor genes is regulated by HSRS (7,738 genes; 72%). As shown in Table 4, SCARS regulate expression of 7,609 of 10,713 (71%) human cancer survival predictor genes (Table 4).

SCARS regulate expression of two-third cancer driver genes (305 of 460 genes; 66%) and as many as 73-75% of high-confidence cancer driver genes (Figure 4), which were defined by either the level of peer-reviewed literature support (Figure 1A) or the statistical significance levels (Figure 4B). Notably, SCARS regulate expression of a majority of cancer driver genes regardless of their maximum mutations' frequency (Figure 4D). SCARS-regulated cancer driver genes were identified in all analyzed to date 28 types of human cancer (Table 5). From the therapeutic strategy stand point, it is important to map actionable cancer therapy-guiding nodes defined by the SCARS stemness matrix mapped to connect Cancer Driver Genes/Cancer Type/Regulatory SCARS (Table 5). Further details describing regulatory effects of HSRS and SCARS on cancer survival and cancer driver genes are reported in the Supplemental Text S4.

Implications for mechanistic studies of normal development and pathophysiology of Modern Humans.

One of the most intriguing molecular functions of SCARS is highlighted by their role as functionally active enhancers as well as the ability of SCARS to influence enhancers' activity. Enhancer elements could be divided into functionally silent and functionally active categories.

Exceedingly large set of functionally silent enhancers could be defined by characteristic chromatin marks indicating that specific DNA sequences harboring these chromatin marks may function as enhancer elements. Accurate molecular and genetic definitions of functionally active enhancers require the application of specific assays in a particular cell type as it has been reported for hESC [55]. It has been observed that SCARS are significantly enriched among regulatory DNA sequences identified in either primed or naïve hESC as functionally active enhancer elements [37; 55]. Furthermore, human embryonic MLME cells, creation of which was associated with SCARS activity [27], appear to capture GES of both Naïve and Primed hESC (Supplemental Text S5) with more significant resemblance of hESC in the Naïve state. Therefore, assessments of biological roles of functionally active enhancers in hESC may shed a light on our understanding of potential biological impacts of SCARS-associated genomic regulatory networks.

Arguably, two key biologically-distinct functions of active enhancers in hESC are the maintenance of self-renewal and pluripotency states by restricting the differentiation potential and changing on demand the expression of genes linked to major embryonic lineages. Primed hESCs, in particular, are thought to represent a state poised to differentiation in which functionally active enhancers linked to differentiation of various lineages can be quickly switched on or off in response to developmental cues (likely in response to changes in chromatin and histone modification patterns). The biological role of functionally active hESC enhancers could be inferred by evaluating the enrichment within regulatory networks governed by naïve and primed hESC enhancers of genes comprising expression signatures of different human and non-human embryonic lineages (Table 6). In these analyses gene expression signatures of major embryonic lineages of distinct species, including humans, monkeys, and mice were evaluated [27; 45; 46; 50; 56-58]. To this end, all genes comprising expression signatures of distinct embryonic lineages were assessed and genes which are located in close genomic proximity (at a distance of 10 kb or less) to naïve and primed hESC functionally active

enhancers were identified. It has been observed that in all instances a high proportion of marker genes distinguishing embryonic lineages are located in close genomic proximity to hESC functional enhancers (Table 6). Notably, proportions of genes associated with naïve and primed hESC enhancers appear similar, consistent with the hypothesis that both naive and primed hESC represent functionally distinct states with the complimentary relevance to mechanistic exploration of developmental pathways.

To assess the statistical significance of these findings, observed numbers of genes associated with hESC functional enhancers were compared to the expected values based on associations by chance alone. Expected values were estimated based on the number of genes in the human genome (63,677); number of genes associated with functional enhancers of the Naïve hESC (18,766); number of genes associated with functional enhancers of the Primed hESC (17,131); number of genes associated with functional enhancers of both Naive and Primed hESC (25,421); and numbers of genes in the corresponding expression signatures of embryonic lineages. These analyses revealed that in all instances differences between the observed and expected numbers of observations appear highly statistically significant (Table 6). These findings indicate that genomic networks governed by both naïve and primed functional enhancers in hESC may represent valuable models for follow-up mechanistic studies of regulatory mechanisms governing critical stages of the human pre-implantation embryogenesis. This line of investigations have been extended to evaluate the potential biological role of hESC functionally active enhancers by performing the proximity placement analyses of genes associated with regulatory networks of naïve and primed hESC functional enhancers and compare these with genes involved in human embryonic, neurodevelopmental, and cancer survival predictors' transcriptional networks (Supplemental Tables S9 and S10), which were previously identified in multiple independent studies {27; 30; 40; 45; 46; 50; 53; 56-60]. A comprehensive genome-wide proximity placement analyses identifies all genes associated with functional enhancers, which were defined based on the location of their genomic coordinates

22

within +/- 10 Kb windows of the corresponding enhancer's genomic coordinates [55]. All genes in common have been identified for a set of genes associated with enhancers and a set of genes comprising the expression signatures of corresponding embryonic, neurodevelopmental, and cancer survival predictors' networks. Finally, the assessment of statistical significance of observed versus expected numbers of genes in common has been performed for corresponding gene sets. Highly significant associations (Supplemental Tables S9 and S10) of genes defining human embryonic, neurodevelopmental, and cancer survival predictors' transcriptional networks with naïve (Supplemental Tables S9) and primed (Supplemental Tables S10) hESC functionally active enhancers have been observed. Genes associated with functionally active enhancers in Naïve and Primed hESC are significantly enriched for genes comprising human-specific expression signatures of excitatory neurons (Figure 5A), radial glia (Figure 5B), induced pluripotent cells (Figure 5C), and human genes encoding a majority of virus-interacting proteins (Figure 5D). It should be noted that these regulatory genomic features of functionally active hESC enhancers are markedly similar to the regulatory impacts of HSRS and SCARS on genes implicated in pathogenesis of neurodevelopmental, neuropsychiatric, and neurodegenerative disorders [35; 38; 39]. The summary of observations supporting this conclusion is reported in the Supplemental Text S6.

Collectively, these findings strongly argue that a comprehensive catalog of functionally active enhancers in hESC together with GES of SCARS-regulated genes may serve as an important previously unavailable resource for evidence-based mechanistic dissections of fine genomic regulatory architectures governing expression of genes implicated in transcriptional networks relevant to human development and diseases.

Evolutionary aspects of the emergence of overlapping genetic networks associated with cancer and other common human disorders.

Present analyses support the idea of shared genomic regulatory networks impacting human cancer survival, neuropsychiatric, neurodevelopmental, and neurodegenerative disorders. Many

genes that expressed in the human brain tend to be long because they have more introns, which is also true for genes expressed in specific cells in human preimplantation embryos because there is a large overlapping genetic networks operating in MLME cells of human embryos and fetal/adult neocortex of human brains. Overall, we have more introns in our genes than, for example mouse, and about 10% less protein coding genes. Thus, in genomes of Modern Humans high transcripts' diversity (which impacts both regulatory diversity of RNA molecules and diversity of peptides and proteins) was achieved by inserting more intronic sequences and increasingly relying on splicing. Retrotransposition is one of the major mechanistic contributors to these continuing processes with major impacts on stem cells survival and expansion to sustain the regeneration and dying cells' replenishment in various tissues and organs (Figure 6). DNA sequences of long genes that are expressed and continuously transcribed in these long living cells for many years of the individuals' lifetime have a significantly higher probability to acquire and accumulate functionally deleterious, regulatory, and disease causing mutations. Depending on when and where it happened, it would manifest as different diseases: for example, in cells of coherent peripheral tissues it would be diagnosed as malignant tumors, while in cells of central nervous system it would be diagnosed as neurodevelopmental, neuropsychiatric, or neurodegenerative disorders. It has been suggested [33] that, in addition to deamination of methyl-cytosine, one of the main mechanisms promoting the increased likelihood of mutations is the RNA-mediated formation of energetically-stable DNA:RNA triple-stranded complexes designated R-loops, specifically, R-loops formation of which is driven by SCARS-encoded RNA molecules.

Hypothesis of a singular source code for essential faithful execution of early embryogenesis programs and driving the emergence of disease states in human cells.

Precisely controlled waves of activities of distinct families of transposable elements, including SCARS, provides a genomic source code for proper execution of high-complexity developmental programs during human preimplantation embryogenesis. In human embryonic

stem cells (hESC), sustained activities of SCARS is required for maintenance of the stemness state. Conversely, failure to silence SCARS during neuronal differentiation of hESC is associated with development of differentiation-defective phenotypes, indicating that SCARS activity is not compatible with physiological functions of differentiated human cells. Consequently, aberrant activation of SCARS in long-living human cells might represent a genomic source code driving the emergence of various disease states, including cancer, neurodegeneration, neurodevelopmental and neuropsychiatric disorders (Figure 6). In this contribution, experimental evidence and theoretical considerations have been summarized supporting the model of a singular genomic source code, activation and execution of which contributes to development of multiple types of human disorders. This singular genomic source code captures the mechanistic essence of malignant regulatory signatures.

1. Removing of DNA and chromatin silencing marks at SCARS-encoding loci.
2. Activation of transcription at SCARS loci and production of SCARS RNAs.
3. Activation of a genome-wide network of functional enhancers associated with a pluripotent state.
4. Genome-scale effects of functional enhancers and SCARS-encoded RNAs on DNA conformation dynamics and chromatin states resulting in expression changes of thousands genes, including known specific disease state-causing genes.
5. Effects of SCARS-encoded RNAs on functions and maintenance of microRNAs affecting the stability, abundance, and translation of mRNAs.
6. Amplification of transcriptional activation of transposable elements, including SVA, Alu, and LINE retrotransposons.
7. Formation of energetically stable triple-stranded RNA/DNA complexes transitioning to highly prone to mutations R-loop conformations.

8. Binding of SCARS RNAs to conformation-disordered proteins, including RNA-binding proteins and virus interacting proteins (VIPs).

9. Genome-scale rewiring of and interference with liquid-liquid phase-separated condensates affecting chromatin regulatory states.

10. System-scale effects on the efficiency of compartmentalization of biochemical reactions controlled by liquid-liquid phase-separated condensates.

Conclusions

In accord with the expected in vivo regulatory role of SCARS and hESC functional enhancers during human embryonic development, a significant enrichment of genes comprising expression signatures of major embryonic lineages of distinct species, including humans, monkeys, and mice has been observed within regulatory networks of Naïve and Primed hESC functional enhancers. Results of these analyses further support the hypothesis that key regulatory features of human neurodevelopmental networks are engaged during the early-stages of human embryogenesis [33-35; Supplemental Tables S11-S12]. Analyses of regulatory networks of Naïve and Primed hESC functional enhancers revealed a highly consistent pattern of significant enrichment of genes that were previously identified as principal components of major neurodevelopmental networks (Supplemental Tables S9-S12), including GES of human neuronal and non-neuronal brain cells [59], human neurons' sub-types and neuronal diversity signatures [60], and human fetal brain/adult neocortex GES [35]. Consistent with the idea that activation of stemness genomic networks in cancer cells contributes to development of clinically-lethal death-from-cancer phenotypes, interrogation of regulatory networks of SCARS as well as Naïve and Primed hESC functional enhancers revealed a significant enrichment of cancer survival predictors' genes that were defined for 17 distinct types of human malignancies [53]. Similar regulatory connectivity has been observed for SCARS and cancer driver's genes identified for 28 human cancer types [54]. Importantly, in all instances these analyses

demonstrated that regulatory networks of SCARS and functional enhancers operating in hESC in both Naïve and Primed states appear to capture distinct arrays of genomic regulatory networks engaged in human embryogenesis, neurodevelopmental processes, and human malignancies. Consequently, collective considerations of all observations summarized in this contribution strongly argue that highly tractable experimental model systems tailored for precise structure-activity-phenotype interrogations of SCARS and functional enhancers in both Naïve and Primed hESC would represent a valuable, perhaps, indispensable, resource for dissections of principal genetic elements governing primate-specific and unique to human features of development, physiology, and pathology of Modern Humans.

Future prospects

- From the clinical perspective, perhaps, reflecting the best interest of cancer patients, the most important translational impact of malignant regulatory signatures would be the reliable early diagnosis of sub-types of malignancies with the increased risk of existing therapy failure and high likelihood of death from cancer. It is this yet unfulfilled promise of malignant regulatory signatures defining stemness of human malignancies is the main focus of this contribution.

- The predominant focus of the contemporary research effort on elucidation of molecular interconnectivity of the stemness phenotype and development of human cancers is on advancement of the cancer stem cell concept. The impact of recent remarkable advancements of single cell genomics of preimplantation human embryos, the bone fide source of the stemness phenotype creation during human development, had relatively modest influence on cancer research and, in particular, on progress in our understanding of mechanistic underpinning of malignant regulatory signatures.

- The in-depth analyses of the critically important impact of stem cell associate retroviral sequences (SCARS) as the essential elements of malignant regulatory signatures of

clinically lethal human cancers will be one of the main topic of the future research. These studies should include precise identification and detailed structure-function analyses of individual transcriptionally-active regulatory genomic loci harboring SCARS and down-stream target genes making vital contributions to pathogenesis of human malignancies and multiple other common disorders.

Methods

Data source and analytical protocols

A total of 94,806 candidate HSRS, including 35,074 neuro-regulatory human-specific SNCs, detailed descriptions of which and corresponding references of primary original contributions are reported elsewhere (Glinsky et al., 2015-2020; Kanton et al., 2019). Solely publicly available datasets and resources were used in this contribution. The significance of the differences in the expected and observed numbers of events was calculated using two-tailed Fisher's exact test. Additional placement enrichment tests were performed for individual classes of HSRS taking into account the size in bp of corresponding genomic regions. Detailed description of methodological and analytical approaches are provided in the Supplemental Methods and previously reported contributions (Glinsky et al., 2015-2020).

Statistical Analyses of the Publicly Available Datasets

All statistical analyses of the publicly available genomic datasets, including error rate estimates, background and technical noise measurements and filtering, feature peak calling, feature selection, assignments of genomic coordinates to the corresponding builds of the reference human genome, and data visualization, were performed exactly as reported in the original publications and associated references linked to the corresponding data visualization tracks (http://genome.ucsc.edu/). Any modifications or new elements of statistical analyses are described in the corresponding sections of the Results. Statistical significance of the Pearson correlation coefficients was determined using GraphPad Prism version 6.00 software. Both nominal and Bonferroni adjusted p values were estimated. The significance of the differences in the numbers of events between the groups was calculated using two-sided Fisher's exact and Chi-square test, and the significance of the overlap between the events was determined using the hypergeometric distribution test (Tavazoie et al., 1999).

References

1. Sell S, Pierce GB. 1994. Biology of Disease: Maturation arrest of stem cell differentiation is a common pathway for the cellular origin of teratocarcinomas and epithelial cancers. Lab. Invest. 70: 6-21.

2. Reya T, Morrison SJ, Clarke MF, Weissman IL. 2001. Stem cells, cancer, and cancer stem cells. Nature 414: 105-11.

3. Pardal R, Clarke MF, Morrison SJ. 2003. Applying the principles of stem-cell biology to cancer. Nat Rev Cancer 3: 895-902.

4. Sell S, Glinsky GV. 2010. Preventive and therapeutic strategies for cancer stem cells. In: Cancer Stem Cells, Ed. William Farrar. Cambridge University Press. Pp. 68-92.

5. Glinsky GV, Berezovska O, Glinskii AB. 2005. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J Clin Invest 115:1503-1521.

6. Glinsky, GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. 2004. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest 113: 913-923.

7. Glinsky GV. 2016. Activation of endogenous human stem cell-associated retroviruses (SCARs) and therapy-resistant phenotypes of malignant tumors. Cancer Lett. 376: 347-359.

8. Grossniklaus U, Renato Paro R. 2014. Transcriptional silencing by Polycomb-Group proteins. Cold Spring Harb Perspect Biol. 6: a019331. doi: 10.1101/cshperspect.a019331.

9. Glinsky GV. 2005. Death-from-cancer signatures and stem cell contribution to metastatic cancer. Cell Cycle 4: 1171-1175.

10. Glinsky GV. 2006. Genomic models of metastatic cancer: Functional analysis of death-from-cancer signature genes reveals aneuploid, anoikis-resistant, metastasis-enabling phenotype with altered cell cycle control and activated Polycomb Group (PcG) protein chromatin silencing pathway. Cell Cycle 5: 1208-1216.

11. Glinsky GV. 2007. Stem cell origin of death-from-cancer phenotypes of human prostate and breast cancers. Stem Cells Reviews 3:79-93.

12. Glinsky GV. 2008. "Stemness" genomics law governs clinical behavior of human cancer: Implications for decision making in disease management. Journal of Clinical Oncology, 26: 2846-53.

13. Ma J, Lanza DG, Guest I, Uk-Lim C, Glinskii A, Glinsky G, Sell S. 2012. Characterization of mammary cancer stem cells in the MMTV-PyMT mouse model. Tumour Biol. 33: 1983-1996.

14. Lanza DG, Ma J, Guest I, Uk-Lim C, Glinskii A, Glinsky G, Sell S. 2012. Tumor-derived mesenchymal stem cells and orthotopic site increase the tumor initiation potential of putative mouse mammary cancer stem cells derived from MMTV-PyMT mice. Tumour Biol. 33: 1997-2005.

15. Glinsky GV, Krones-Herzig A, Glinskii AB, Gebauer G. 2003. Microarray analysis of xenograft-derived cancer cell lines representing multiple experimental models of human prostate cancer. Molecular Carcinogenesis, 37: 209-221.

16. Glinsky GV, Higashiyama T, Glinskii AB. 2004. Classification of human breast cancer using gene expression profiling as a component of the survival predictor algorithm. Clinical Cancer Res 10: 2272-2283.

17. Glinskii AB, Smith BA, Jiang P, Li X-M, Yang M, Hoffman RM, Glinsky GV. 2003. Viable circulating metastatic cells produced in orthotopic but not ectopic prostate cancer models. Cancer Res 63: 4239-4243.

18. Glinsky, GV, Glinskii, AB, Berezovskaya, O, Smith, BA, Jiang, P, Li, X-M, Yang, M, Hoffman, RM. 2006. Dual-color-coded imaging of viable circulating prostate carcinoma cells reveals genetic exchange between tumor cells in vivo, contributing to highly metastatic phenotypes. Cell Cycle, 5: 191-197.

19. Berezovskaya O, Schimmer AD, Glinskii AB, Pinilla C, Hoffman RM, Reed JC, Glinsky GV. 2005. Increased expression of apoptosis inhibitor XIAP contributes to anoikis resistance of circulating human prostate cancer metastasis precursor cells. Cancer Res 65: 2378-2386.

20. Berezovska OP, Glinskii AB, Yang Z, Li X-M, Hoffman RM, Glinsky GV. 2006. Essential role Activation of the Polycomb Group (PcG) protein chromatin silencing pathway in metastatic prostate cancer. Cell Cycle 5: 1886-1901.

21. Srinivasan M, Bharali DJ, Sudha T, Khedr M, Guest I, Sell S, Glinsky GV, Mousa SA. 2017. Downregulation of Bmi1 in breast cancer stem cells suppresses tumor growth and proliferation. Oncotarget 8: 38731-38742.

22. Abdouh M, Facchino S, Chatoo W, Balasingam V, Ferreira J, Bernier G. 2009. BMI1 sustains human glioblastoma multiforme stem cell renewal. The Journal of Neuroscience 29: 8884–96.

23. Kreso A, van Galen P, Pedley NM, Lima-Fernandes E, Frelin C, Davis T, Cao L, Baiazitov R, Du W, Sydorenko N, Moon YC, Gibson L, Wang Y, Leung C, Iscove NN, Arrowsmith CH, Szentgyorgyi E, Gallinger S, Dick JE, O'Brien CA. 2014. Self-renewal as a therapeutic target in human colorectal cancer. Nature Medicine. 20: 29–36.

24. Siddique HR, Saleem M. 2012. Role of BMI1, a stem cell factor, in cancer recurrence and chemoresistance: Preclinical and clinical evidences. Stem Cells 30: 372-378.

25. Glinsky GV. 2016. Single cell genomics reveals activation signatures of endogenous SCARS networks in aneuploid human embryos and clinically intractable malignant tumors. Cancer Lett. 381: 176-193.

26. Glinsky GV. 2015. Viruses, stemness, embryogenesis, and cancer: a miracle leap toward molecular definition of novel oncotargets for therapy-resistant malignant tumors? Oncoscience 2: 751-754.

27. Glinsky G, Durruthy-Durruthy J, Wossidlo M, Grow EJ, Weirather JL, Au KF, Wysocka J, Sebastiano V. 2018. Single cell expression analysis of primate-specific retroviruses-derived HPAT lincRNAs in viable human blastocysts identifies embryonic cells co-expressing genetic markers of multiple lineages. Heliyon 4: e00667. doi: 10.1016/j.heliyon.2018.e00667.

28. M. Koyanagi-Aoi M, M. Ohnuki M, K. Takahashi K, K. Okita K, H. Noma H, Y. Sawamura Y et al. 2013. Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proc Natl Acad Sci USA 110: 20569–20574.

29. Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, et al. 2014. Dynamic regulation of human endogenous retroviruses mediates factor induced reprogramming and differentiation potential. Proc Natl Acad Sci USA 111: 12426–12431.

30. Theunissen TW et al. 2016. Molecular criteria for defining the naive human pluripotent state. 2016. Cell Stem Cell 19: 502-515.

31. Gao L et al. 2018. Chromatin accessibility landscape in human early embryos and its association with evolution. Cell 2018; 173: 248–259.

32. Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. 2019. Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. Cell Stem Cell 24: 724-735.

33. Glinsky, G.V. 2015. Transposable Elements and DNA Methylation Create in Embryonic Stem Cells Human-Specific Regulatory Sequences Associated with Distal Enhancers and Noncoding RNAs. Genome Biology and Evolution *7*: 1432-1454.

34. Glinsky, G.V. 2016. Mechanistically Distinct Pathways of Divergent Regulatory DNA Creation Contribute to Evolution of Human-Specific Genomic Regulatory Networks Driving Phenotypic Divergence of Homo sapiens. Genome biology and evolution 8: 2774-2788.

35. Glinsky G.V. 2017. Human-specific features of pluripotency regulatory networks link NANOG with fetal and adult brain development. BioRxiv. https://www.biorxiv.org/content/10.1101/022913v3 doi: https://doi.org/10.1101/022913

36. Glinsky, G.V. 2018. Contribution of transposable elements and distal enhancers to evolution of human-specific features of interphase chromatin architecture in embryonic stem cells. Chromosome Res 26: 61-84.

37. Glinsky GV, Barakat TS. 2019. The evolution of Great Apes has shaped the functional enhancers' landscape in human embryonic stem cells. 37: 101456.

38. Glinsky GV. 2020. A catalogue of 59,732 human-specific regulatory sequences reveals unique to human regulatory patterns associated with virus-interacting proteins, pluripotency and brain development. DNA and Cell Biology 39: 126-143. doi: 10.1089/dna.2019.4988.

39. Glinsky GV. 2020. Impacts of genomic networks governed by human-specific regulatory sequences and genetic loci harboring fixed human-specific neuro-regulatory single nucleotide mutations on phenotypic traits of Modern Humans. BioRxiv. https://www.biorxiv.org/content/10.1101/848762v3 doi: https://doi.org/10.1101/848762

40. Guffanti G, Bartlett A, Klengel T, Klengel C, Hunter R, Glinsky G, Macciardi F. 2018. Novel bioinformatics approach identifies transcriptional profiles of lineage-specific transposable elements at distinct loci in the human dorsolateral prefrontal cortex. Mol Biol Evol. 35: 2435-2453.

41. Attig J, Young GR, Hosie L, Perkins D, Encheva-Yokoya V, Stoye JP, Snijders AP, Ternette N, Kassiotis G. 2019. LTR retroelement expansion of the human cancer transcriptome and immunopeptidome revealed by de novo transcript assembly. Genome Res 29: 1578-1590.

42. Durruthy-Durruthy J, Sebastiano V, Wossidlo M, Cepeda D, Cui J, Grow EJ, Davila J, Mall M, Wong WH, Wysocka J, Au KF, Reijo Pera RA, 2016. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat. Genet. 48: 44-52.

43. Durruthy-Durruthy J, Wossidlo M et al., 2016. Spatiotemporal reconstruction of the human blastocyst by single-cell gene expression analysis informs induction of naive pluripotency. Dev. Cell 38: 100-115.

44. Petropoulos S, Edsgard D, Reinius B, Deng Q, Panula SP, Codeluppi S, Plaza Reyes A, Linnarsson S, Sandberg R, Lanner F. 2016. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell 165: 1012-1026.

45. Blakeley P, Fogarty NM, del Valle I, Wamaitha SE, Hu TX, Elder K, Snell P, Christie L, Robson P, Niakan KK. 2015. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development 2015; 142: 3151-3165.

46. Bai, Q., Assou, S., Haouzi, D., Ramirez, J.M., Monzo, C., Becker, F., Gerbal-Chaloin, S., Hamamah, S., and De Vos, J. 2012. Dissecting the first transcriptional divergence during human embryonic development. Stem Cell Rev. 2012; 8: 150–162.

47. Murphy PJ et al. 2018. Placeholder nucleosomes underlie germline-to-embryo DNA methylation reprogramming. Cell 172: 993–1006.

48. Li L et al. 2018. Single-cell multi-omics sequencing of human early embryos. Nature Cell Biology 20: 847–858.

49. Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, Guijarro P, Sidow L, Fleck JS, Han D, Qian Z, Heide M, Huttner WB, Khaitovich P, Pääbo S, Treutlein B, Camp JG. 2019. Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature. 2019; 574: 418-422. PMID: 31619793; DOI: 10.1038/s41586-019-1654-9.

50. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvák Z. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature 2014; 516: 405-409.

51. Chen EY, Tan, CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinform. 2013, 14, 128.

52. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016, 44, W90–W97.

53. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K, von Feilitzen K, Oksvold P, Lundberg E, Hober S, Nilsson P, Mattsson J, Schwenk JM, Brunnström H, Glimelius B, Sjöblom T, Edqvist PH, Djureinovic D, Micke P, Lindskog C, Mardinoglu A, Ponten F. 2017. A pathology atlas of the human cancer transcriptome. Science 357(6352): eaan2507. doi: 10.1126/science.aan2507.2017.

54. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, Lander ES, Van Allen EM, Sunyaev SR. 2020. Identification of cancer driver genes based on nucleotide context. Nat Genet. 52: 208-218. doi: 10.1038/s41588-019-0572-y.

55. Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I. 2018. Functional dissection of the enhancer repertoire in human embryonic stem cells. Cell Stem Cell. 23: 276-288.

56. Nakamura T, Okamoto I, Sasaki K, Yabuta Y, Iwatani C, Tsuchiya H, Seita Y, Nakamura S, Yamamoto T, Saitou M. A developmental coordinate of pluripotency among mice, monkeys and humans. Nature 2016; 537: 57-62.

57. Boroviak T. and Nichols J. Primate embryogenesis predicts hallmarks of human naïve pluripotency. Development 2017; 144: 175-186.

58. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, Huang J, Li M, Wu X, Wen L, Lao K, Li R, Qiao J, Tang F. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 2013; 20: 1131-1139.

59. Lake BB et al. 2017. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nature Biotechnology, Advance online publication 11 December (2017). doi:10.1038/nbt.4038

60. Lake BB et al. 2016. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science 352, 1586-1590. DOI: 10.1126/science.aaf1204