THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Development and characterization of experimental tools for functional genomics research

by **Lilla Ozgyin**

Supervisor: Dr. Bálint László Bálint, MD, PhD

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY
DEBRECEN, 2019

Table of contents

List of abbreviations

3C	chromatin conformation capture
5-FU	5-fluorouracil
ACTB	beta-actin
AMEX	X-linked amelogenin
AMELY	Y-linked amelogenin
AR	androgen receptor
ARE	AU-rich sequence element
ATAC-Seq	assay for transposase accessible chromatin with high-throughput sequencing
BSA	bovine serum albumin
BWA	Burrows-Wheeler Aligner
CDS	coding DNA sequence
CEPH	Centre d'Etude du Polymorphism Humain
ChIA-PET	chromatin interaction analysis by paired-end tag sequencing
ChIP	chromatin immunoprecipitation
ChIP-qPCR	chromatin immunoprecipitation coupled with quantitative real-time polymerase chain reaction
ChIP-Seq	chromatin immunoprecipitation sequencing
CPM	counts per million
CV	coefficient of variance
DEG	differentially expressed gene
DNase	deoxyribonuclease
dNTP	deoxynucleotide triphosphate
DPYD	dihydropyrimidine dehydrogenase
dsQTL	DNase hypersensitivity quantitative trait locus
DTT	dithiothreitol
EBV	Epstein-Barr virus

EDTA	ethylenediaminetetraacetic acid
EGCG	(-)-epigallocatechin gallate
eQTL	RNA expression quantitative trait locus
eRNA	enhancer-associated RNA
FCS	foetal calf serum
FDR	false discovery rate
FS	first strand
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GC	guanosine-cytidine dinucleotide
GEO	Gene Expression Omnibus
GFP	green fluorescent protein
H3K27ac	H3 histone acetylated at the 27th lysine residue
HEK	human embryonic kidney
HGNC	HUGO Gene Nomenclature Committee
Hi-C	chromosome conformation capture of all chromatin contacts coupled with sequencing
HLA	human leukocyte antigen
hmQTL	histone modification quantitative trait locus
IC_{50}	half-maximal inhibitory concentration
IgG	immunoglobulin G
IMT-2	acronym for a reagent containing 0.1 M D-trehalose and 0.945 mg/ml (-)-epigallocatechin gallate in phosphate buffered saline
IP	immunoprecipitation
IP-qPCR	immunoprecipitation coupled with quantitative polymerase chain reaction
LCL	human B-lymphoblastoid cell line
lncRNA	long non-coding RNA
Lyo	lyophilization; lyophilized
meQTL	DNA methylation quantitative trait locus

MTT	(3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyltetrazolium Bromide)
nTdT	nuclear terminal deoxynucleotidyl transferase
PBL	peripheral blood lymphocyte(s)
PBMC	peripheral blood mononuclear cell(s)
PBS	phosphate buffered saline
PCG	protein-coding gene
PCR	polymerase chain reaction
QC	quality control
QTL	quantitative trait locus
RIN	RNA Integrity Number
RNA-Seq	RNA sequencing
RPKM	reads per kilobase per million mapped reads
RT	reverse transcription
RT-PCR	reverse transcription polymerase chain reaction
RT-qPCR	quantitative reverse transcription polymerase chain reaction
SD	standard deviation
SDS	sodium dodecyl sulfate
SEM	standard error of the mean
sGT_1	“same genotype” (isogenic) LCL number 1 (GM22647)
sGT_2	“same genotype” (isogenic) LCL number 2 (GM22648)
sGT_3	“same genotype” (isogenic) LCL number 3 (GM22649)
sGT_4	“same genotype” (isogenic) LCL number 4 (GM22650)
sGT_5	“same genotype” (isogenic) LCL number 5 (GM22651)
SNV	single nucleotide variant
tfQTL	transcription factor binding quantitative trait locus
UTR	untranslated region

1 INTRODUCTION

The past decade has seen the emergence of the field of functional genomics and novel approaches to scientific cooperation. Clinical studies benefit from the rapid development of omics methods, which provide a genome-wide view on dynamic aspects of gene regulation in pathological contexts. Moreover, public repositories of well-annotated biomaterials and high-throughput sequencing data offer an unparalleled opportunity to utilize available resources to accelerate the accumulation of valuable scientific information. Therefore characterization of emerging model systems, development of key functional genomic methodologies and cost rationalization are key aspects of biomedical research in the post-genomic era.

The fields of transcriptomics and epigenomics have evolved in parallel with high-throughput nucleic acid profiling methods, such as RNA sequencing (RNA-Seq) and chromatin immunoprecipitation sequencing (ChIP-Seq). As for RNA-Seq, a major limitation is its sensitivity to low initial sample quality, which mostly emerges due to inappropriate sample handling, storage and transportation. In contrast, chromatin immunoprecipitation suffers from the relative lack of standardization and long hands-on time, which present major obstacles to more widespread use.

The human B-lymphoblastoid cell line (LCL) model system is widely used for uncovering general rules and genomic context of gene regulation. Many genomics and functional genomics consortial projects, including the landmark studies of International HapMap Project, 1000 Genomes Project and ENCODE Project, used LCLs as sources of biomaterial. The accessibility of thousands of LCLs together with well-annotated sequencing data in public repositories empowers research on the genetic basis of quantitative cellular phenotypes. However, LCL-to-LCL functional genomic variability in the same genotype context, which has not yet been previously evaluated, may complicate the above efforts.

With our work, we aimed at contributing to the field of functional genomics in three ways: first, by developing a novel spike-in control system to account for experimental sample loss during ChIP experiments; second, by evaluating the utility of a cost-effective cell archivation method to preserve cellular RNA; and third, by uncovering, for the first time, the extent and nature of non-DNA-driven functional genomic variability of the LCL model.

2 THEORETICAL BACKGROUND

2.1 Phage display-based chromatin immunoprecipitation procedure controls

2.1.1 Chromatin immunoprecipitation in functional genomics

Chromatin immunoprecipitation (ChIP) is a laboratory method to study interactions between genomic DNA and proteins, such as transcriptional regulators and post-transcriptionally modified histones *ex vivo*. Since its first use in the 1980s (Gilmour & Lis, 1984), numerous variations of the procedure addressing specific needs have emerged, and they have been successfully used in multiple eukaryotic model organisms. When coupled with deep sequencing, the method allows for the profiling of epigenetic marks genome-wide. Chromatin immunoprecipitation sequencing (ChIP-Seq) has also been extensively applied in clinical research for mapping disease-associated epigenetic signatures. However, multiple procedural factors, including large starting cell numbers, long hands-on time and the scarcity of standardized controls severely limit its use in clinical settings.

A typical ChIP experiment starts with stabilizing chromatin interactions by reversibly crosslinking DNA and proteins with formaldehyde (and, in some cases, together with a longer-arm crosslinker (Di(N-succinimidyl) glutarate) *in situ.* The reaction is stopped by adding glycine, then whole cells or isolated nuclei are fragmented using sonication or enzymatic treatment. Fragmented chromatin is immunoprecipitated with a ChIP-grade antibody coupled to paramagnetic beads. Bead-coupled immunocomplexes are washed several times to remove the aspecific binding, followed by immunocomplex elution from the beads. During the immunoprecipitation and washing steps, samples are kept on ice or at 4°C, and buffers are supplemented with proteinase inhibitors in order to prevent protein degradation. ChIP eluates are treated with RNase and Proteinase and purified using DNA-binding columns or by precipitation. ChIP DNA is then subjected to either qPCR (ChIP-qPCR) to quantify selected genomic regions or high-throughput sequencing (ChIP-Seq). Many laboratories use in-house protocols and buffers, but ChIP kits are also available from different vendors.

2.1.2 Controls and normalizers in chromatin immunoprecipitation experiments

ChIP protocols involve controls and normalizers, which ensure the elimination or identification of confounding variables, such as varying starting cell numbers and aspecific capture. When setting up a ChIP experiment, it may be useful to test the protocol on the target cell type, especially when one plans to work with an antibody that has not been previously confirmed ChIP-grade. Also, a commonly used positive control experiment is when a genomic region expected to be occupied by the target protein is amplified along with regions of interest by qPCR. Besides positive controls, several negative controls can be used; for instance, negative control genomic loci, which are expected to be under-represented in the sample. Another commonly used negative control experiment is immunoprecipitation (IP) with aspecific isotype-matched antibodies (IgG; purified from the serum of a non-immunized animal, or an antibody against an irrelevant epitope, such as GFP or FLAG-tag). ChIP-qPCR enrichments are almost exclusively reported using the "per cent input method" when the qPCR signal of the target region in the IP sample is divided by that of the so-called input sample. The input sample is a representative aliquot of the crosslinked and fragmented chromatin, which is set aside before the IP step, joining the protocol at the crosslink reversal step. The per cent input method is particularly useful for controlling chromatin input variability. ChIP-Seq libraries are prepared from ChIP DNA, but isotype control-precipitated samples and input samples may also be sequenced. However, there are certain drawbacks of using these controls in ChIP-Seq, and the majority of technical artifacts can be accounted for using bioinformatic methods only.

There is no generally accepted internal control to be used in ChIP-qPCR experiments. Defining appropriate internal controls for a given ChIP experimental setup – taking into account the cell type, treatment, target protein, or other variables emerging during the project – may be cumbersome. Of note, a study claims that occupancy quantitation is possible in ChIP-Seq by using an internal standard of unchanged CTCF peaks (peak-like enrichments as seen using genomic viewers). In their proposed method, a second antibody against the CTCF transcription factor (TF) is added to the reactions, and CTCF enrichments are quantified at specific loci. However, they assumed that CTCF has the same distribution and occupancy across multiple species and that CTCF does not overlap with the experimental target; this, of course, may not hold (Guertin, Cullen, Markowetz, & Holding, 2018). Without appropriate internal or external controls experimental sample loss cannot be accounted for after the chromatin fragmentation step. The lack of such procedural control is especially problematic given the fact that the majority of sample handling steps, including extensive washing (~2/3 of the experimental time), takes place after fragmentation. In the case of precious clinical samples, such as biopsies, the reality is to have only a few ChIP replicates to balance such sample loss statistically. Spike-in controls might be considered as alternative post-fragmentation procedural controls, as substitutes to missing internal controls.

A few spike-in controls have been developed for ChIP-Seq in the past few years. All protocols include the addition of xenogenic material to the experimental sample and quantification thereof during sequencing analysis. Orlando *et al.* spiked Drosophila melanogaster S2 cells to human cells (Orlando et al., 2014), while Bonhoure *et al.* and Egan *et al.* added xenogenic fragmented chromatin to their chromatin samples (Bonhoure et al., 2014; Egan et al., 2016). The first two approaches used one pan-specific antibody capturing the target from both species, while Egan *et al.* used an additional, fly-specific antibody (anti-H2Av) for capturing the exogenous target. Grzybowski *et al.* developed ICeChIP (Internal Standard Calibrated ChIP), which enable the calculation of histone modification density, and the direct comparison of experiments, by adding reconstituted and semisynthetic nucleosomes including barcoded DNA to the sample in concentration series (Grzybowski, Chen, & Ruthenburg, 2015). Despite these recent developments, the use of spike-in controls for comparative ChIP has not yet become widespread.

2.1.3 Bacteriophages as potential ChIP spike-in controls

Phage display is a widely used laboratory method for studying the interaction of proteins with certain substances (e.g. other proteins, peptides, DNA and ligands). The method is commonly used in basic research for epitope analysis of monoclonal antibodies and polyclonal sera (Moreira, Fühner, & Hust, 2018), for computational prediction of epitope structures (Halperin, Wolfson, & Nussinov, 2003), for monoclonal antibody generation (Hammers & Stanley, 2014), and in pharmaceutical biotechnology for determining drug targets, improving existing biomolecule drugs, engineering therapeutic antibodies, and mapping epitopes of clinically relevant antibodies.

A phage-based random peptide library is a pool of up to billions of different peptides displayed on the surface of bacteriophages. Libraries are generated by in-frame insertion of random DNA sequences of fixed length into one of the phage coat protein-encoding genes. Assembled phage vectors are then transformed into an E. coli strain, which produces phages with the modified coat proteins. The most commonly used bacteriophages are M13, T4, and T7. The displayed peptides are fused to major or minor coat proteins (pIV or pIII, respectively), can be of various length (e.g. 6-mer, 9-mer or 12-mer), and may be linear or form secondary structures (by incorporating cysteines). Vendors offer various libraries fitted to different research needs, and custom libraries can be prepared using commercially available kits.

Various approaches can be used for developing phage-based reagents with desired properties. In a typical experimental setup, a phage-based random peptide library is subjected to *in vitro* selection (coined “biopanning”), which comprises the affinity selection of phage-displayed peptides with a specific, immobilized substance (e.g. an antibody), followed by washing steps and phage elution. The eluate contains a fraction of the original library, which can be amplified in E.coli. Additional selection rounds can be carried out using harsher washing conditions to ensure high target affinity and specificity. Individual phages can also be cloned out from these polyclonal libraries for further development and study. Moreover, custom libraries can be designed for different purposes, such as for improving target specificity, by cloning DNA sequences for a relatively few pre-defined peptide variants into phagemid vectors and comparing the affinity of the recombinant phages to the target substance.

Phage display libraries have numerous properties that make them appealing as ChIP spike-in controls. First, phages can be developed for binding ChIP-grade antibodies through their displayed peptides. Second, the physical connection between the displayed peptide and the genotype (i.e. being part of the same phage particle) allows for the DNA-based relative quantification of affinity-selected phages (Vodnik, Zager, Strukelj, & Lunder, 2011) (Figure 1A). Phages developed through multiple rounds of *in vitro* evolution and optional cloning steps may be added to all ChIP reactions at the immunoprecipitation step, and phage DNA can be isolated together with chromatin fragments, followed by qPCR amplification (Figure 1B). Thus, the value representing qPCR-based phage recovery can be used as a post-fragmentation normalizer. Third, phages and their displayed peptides may be less sensitive to denaturation, which might support long-term reagent stability. And fourth, phages can be reconstituted at consistent quality by developing monoclonal reagents and reinfecting a host E. coli strain. In theory, if the phage genome-containing ChIP DNA is later subjected to sequencing, phage DNA will not interfere with ChIP-Seq, given that the circular ssDNA genome of the filamentous phage is incompatible with library preparation and NGS.

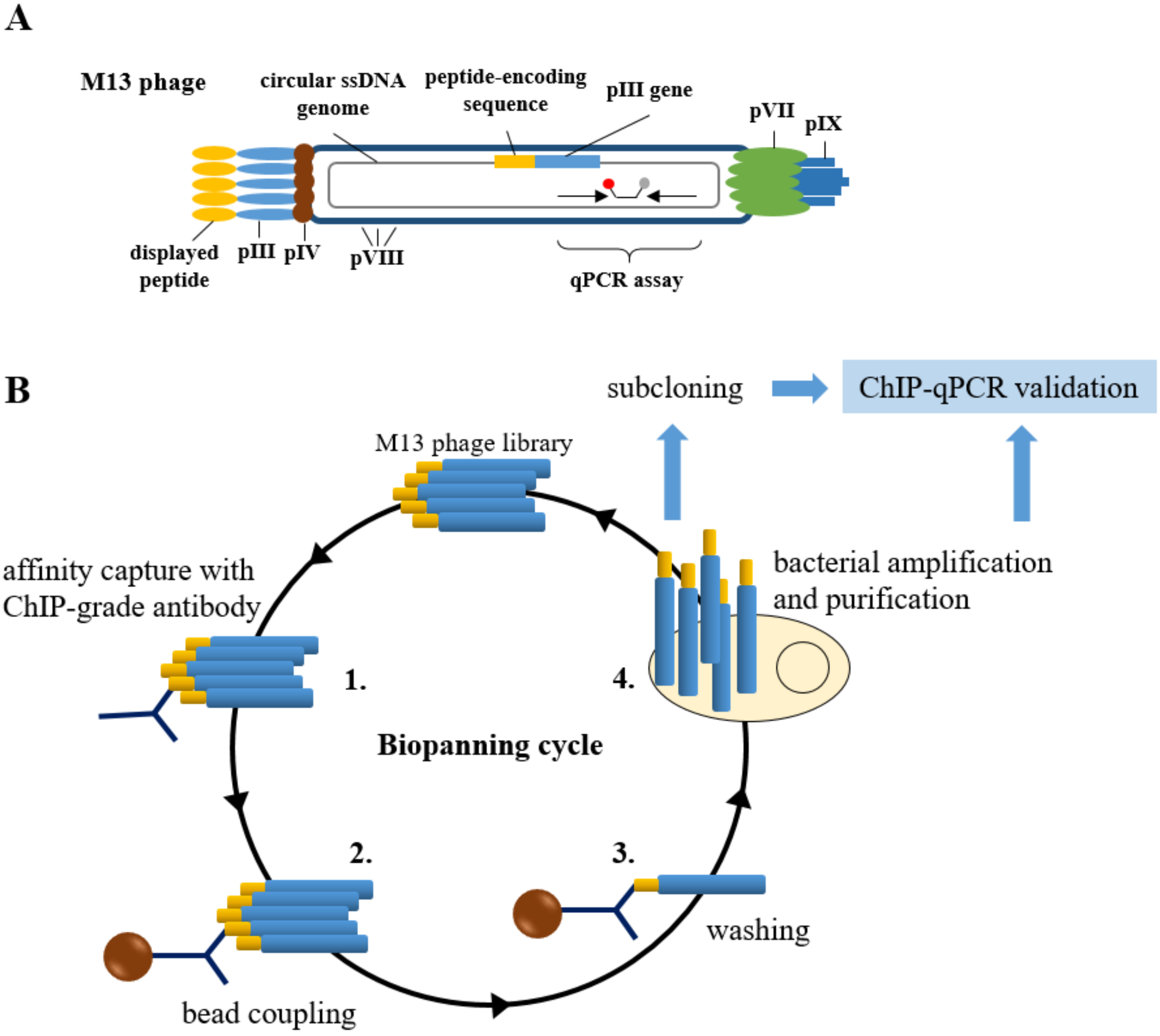


Figure 1 | M13-based phage display. A) Structural components of the M13 phage engineered for phage display. The indicated qPCR assay designed to a non-variable genomic site can be used to detect phages by qPCR. **B)** Experimental design for developing and testing phage display-based ChIP controls using multiple rounds of biopanning with ChIP-grade antibodies.

2.2 Lyophilization as an alternative method of cell stabilization for RNA-based studies

2.2.1 RNA sequencing in clinical research

RNA-Seq is a transcriptome profiling method applying high-throughput sequencing for the detection and quantification of cDNA. The technology allows for the examination of various aspects of RNA biology, such as differential expression of RNA subtypes and transcript variants, allele-specific gene expression and single-cell transcriptomics. In the post-genomic era, RNA-Seq presents as a relatively simple yet powerful tool to explore gene regulatory processes.

With falling sequencing costs and expanding bioinformatic toolsets, RNA-Seq has also become the leading technology of high-throughput RNA biomarker research. RNA levels reflect the functional state of the cells, which cannot be captured by DNA-based assays. Moreover, RNA quantitation methods are relatively mature, sensitive and highly specific. In line with the above, the number of published intra- and extracellular RNA biomarker candidates have been on the rise in the past few years (Deng et al., 2018; Ma et al., 2015; Santiago, Bottero, & Potashkin, 2018; F. Zhang, Ding, Cui, Barber, & Deng, 2019), which may pave the way for the development of RNA-based biomarker panels similar to the PAM50 breast cancer subtype predictor panel (Parker et al., 2009).

Bulk RNA sequencing methods are relatively standardized, with initial sample quality being the most critical factor for success. RNA in clinical samples collected during surgeries or as biopsies are especially prone to degradation, mostly because sample preservation is not a priority in such settings. While RT-(q)PCR-based methods have shown to be relatively insensitive to sample degradation, RNA fragmentation has serious impact on RNA-Seq analysis (Baechler et al., 2004; Bray et al., 2010; Catts et al., 2005; Gallego Romero et al., 2014; Ibberson, Benes, Muckenthaler, & Castoldi, 2009). Therefore it is of substantial interest to preserve tissue/cellular samples in a way that enables the extraction of highly intact RNA.

2.2.2 Tissue banks and different approaches to sample preservation

Given that native biomaterials are sensitive to a range of chemical and physical exposures, various methods have been utilized to minimize degradation and the change in features of interest. Such features include the antigenicity of various macromolecules and the intactness of nucleic acids during sample preparation, handling and storage. These methods are mostly based on keeping the samples in chemically inert solutions, chemical fixation, (ultra-)deep freezing, or combinations of the above. The method of choice depends on the sample type (e.g. whole tissue vs molecular preparations), source (e.g. biopsies vs cell lines) and intended storage length (short- or long-term). Keeping sample quality consistent long-term enables comparability of samples collected over a long period, such as in the case of longitudinal studies and scarce samples (e.g. rare diseases) (Monaco, Crimi, & Wang, 2014).

Clinical samples stored in tissue banks provide the starting material for basic and clinical research. These samples can be used in studies aiming to uncover molecular pathways associated with certain human conditions or identify and validate clinical biomarkers (Branković, Malogajski, & Morré, 2014; Vora & Thacker, 2015; Zatloukal & Hainaut, 2010). Additionally, clinical sample archives provide the opportunity to run diagnostic and follow-up tests weeks or months, or even years after sample collection. Biobank facilities conventionally operate ultra-deep freezers and liquid nitrogen tanks in order to prevent sample deterioration. When appropriate freezing methods and storage conditions are used, the main factor that determines the quality of nucleic acid or protein samples is sample quality before freezing. However, storage at ultra-deep temperatures entails substantial operating costs and a severe environmental burden (Zatloukal & Hainaut, 2010). Funding resources are shrinking worldwide, raising concerns over the financial sustainability of tissue banks. Moreover, temperature fluctuations due to possible power outages, as well as transportation delays due to logistical barriers may lead to transient warming cycles, increasing the risk of sample degradation. Sample storage and transportation at ambient temperatures without affecting sample quality would alleviate the above concerns.

Formalin-fixed and paraffin-embedded (FFPE) archival tissues stored at room temperature have been identified as a potentially rich source of molecular information for RNA-, DNA-, and protein-based studies. Although FFPE tissues retain their morphology for decades, extensive nucleic acid and protein degradation may occur during formaldehyde crosslinking, paraffin embedding, room temperature storage and isolation (L. N. Bell et al., 2011; Crockett, Lin, Vaughn, Lim, & Elenitoba-Johnson, 2005; Hedegaard et al., 2014; Scicchitano, Dalmas, Boyce, Thomas, & Frazier, 2009). The chemical processes seriously impairing downstream PCR- and sequencing-based analyses include methylene bridge formation between amino groups followed by hydrolysis (Chung et al., 2008; Esteve-Codina et al., 2017; Masuda, Ohnishi, Kawamoto, Monden, & Okubo, 1999; von Ahlfen, Missel, Bendrat, & Schlumpberger, 2007), as well as G>A and C>T substitutions which may lead to read mapping bias and unreliable SNV calling from sequencing data (Esteve-Codina et al., 2017; Graw et al., 2015; Hedegaard et al., 2014; Kruse, Basu, Luesse, & Wyatt, 2017).

Cell-penetrable, non-crosslinking fixatives may be used for the short-term stabilization of nucleic acids and proteins. Some of the best-known such fixatives are RNAlater (Thermo Fisher Scientific), PAXgene Tissue STABILIZER (BD), and Allprotect Tissue Reagent (Qiagen). These solutions are especially useful for "field" collection, such as during surgical tissue harvesting, as no immediate sample processing or snap-freezing is required (Drakulovski et al., 2013; Paul et al., 2005). Vendors recommend storing samples submerged in the reagent at room temperature for up to one week, between 4°C and 8°C for up to one month (up to one year for Allprotect), or in a frozen state for more extended periods. Numerous research groups have tested these reagents for non-indicated usage and use with high-throughput technologies, comparing their effectiveness to standard methods such as FFPE or snap-freezing. RNAlater, for example, is claimed to denature cellular RNases upon cell penetration rapidly, and has been successfully used for tissue preservation prior to RNA microarray analyses, histological examinations and proteomics studies (Florell et al., 2001; Mutter et al., 2004; Saito, Bulygin, Moran, Taylor, & Scholin, 2011; M. Wang, Ji, Wang, Li, & Zhou, 2018). Of note, however, high-throughput studies have reported RNAlater-specific changes in transcript and protein abundance, which warrant caution in the use of RNAlater for transcriptomic and proteomic studies (Kruse et al., 2017; Passow et al., 2019).

2.2.3 Lyophilization

Lyophilization (freeze-drying) is a drying technique which uses vacuum and a moderate amount of heat to remove water molecules from frozen (bio)materials by sublimation and desorption. The technique was first described in 1980, although its wider adoption dates back to the middle of the 20th century. Nowadays, lyophilization is widely used by the food, pharmaceutical, and technological industries to dry food products (e.g. seasonal fruits, vegetables, meats and coffee), macromolecular preparations (e.g. vaccines, antibodies and enzymes), and certain inorganic materials, respectively. Lyophilization might also serve as a suitable alternative to conventional tissue storage methods. The low residual water content in the dried products essentially stops molecular motion and inhibits intrinsic molecular degradation pathways. Moreover, tissue lyophilization requires minimal hands-on time and may preserve multiple types of heat-labile analytes in complex samples.

Lyophilization generally involves freezing samples in a so-called “lyoprotectant” solution and subjecting these pre-treated samples to conditions allowing for the frozen water molecules to directly enter the vapour phase, without passing through liquid phase. Although the main steps are common, the exact protocol and type of equipment used may substantially vary depending on biomaterial type, sample size, intended downstream use and target shelf life. Common lyoprotectants include trehalose, skimmed milk, and polyvinylpyrrolidone. The freezing step can be carried out slowly (e.g. by placing the samples into a deep freezer) or quickly (e.g. by immersing the samples into liquid nitrogen). The drying step consists of primary drying when free water is sublimed under vacuum and condensed on the surface of the cold condenser, and secondary drying, when bound water is removed via desorption. During the drying process, heat must be added to the system to enable vapour transfer to the condenser. Elaborate shelf freeze dryers are equipped with heatable shelves, while samples dried in manifold laboratory freeze dryers receive heat from the environment. At the end of the drying process, samples are generally sealed and stored in the dark, above the freezing point (at 4°C or room temperature).

Lyopreservation has numerous benefits over conventional biosample storage methods; therefore, lyopreservation has long been considered as a possible alternative to conventional approaches. In contrast to drying through evaporation (heat-drying), freeze-drying prevents heat-induced deterioration of heat-labile substances during the drying process, while low residual water activity hinders water-mediated degradation processes. Lyopreservation has the advantage of room temperature storage and convenient transportation, non-degradative pretreatment and even the possibility to recover intact and functional cellular structures, such as platelets (Wolkers, Walker, Tablin, & Crowe, 2001; X.-L. Zhou et al., 2007). Moreover, the cost of long-term storage of freeze-dried samples may be lower than that of storage at subzero temperatures, as the operating costs of a lyobank are minimal. Of note, long drying cycles (i.e. days) in industrial shelf freeze-dryers, due to their high power demand, may offset the financial benefits of ambient storage. Additionally, temperature fluctuations (e.g. during transportation) may not have such detrimental effects on lyophilized samples as on frozen samples. Despite the above benefits, the technology is not yet mature enough to become standard practice for cell and tissue preservation.

2.2.4 Lyophilization of cells and tissues

As for other biosample preservation methods, sample type, the intended use of the final product and cost considerations affect the method of choice for lyophilization. In theory, preserving the primary structure of macromolecules for low-throughput quantitative studies (e.g. qPCR) is the easiest to achieve, while fully recovering molecular activities and immunogenicity (e.g. enzymes) or even viable and functional cells might need considerable optimization efforts.

Lyophilization is considered a biomimetic strategy based on the natural phenomenon of intrinsic desiccation tolerance, called anhydrobiosis. Anhydrobiotes, including certain bacteria, yeasts, plants, and invertebrates can switch to a metabolic a state of suspended animation in case of water deficiency and regain normal metabolism when enough water is present in the environment. These organisms utilize various mechanisms to stabilize a life-compatible inner state, including sugar accumulation (such as trehalose), heat shock protein expression and the synthesis of polyunsaturated fatty acids (Watanabe, 2006).

Mammalian cell lyophilization which preserves cell functions has long been a desirable alternative to preserve cells for clinical applications, such as blood transfusion, tissue engineering and regenerative medicine. Although prokaryotes and yeasts have been extensively reported to survive lyophilization, and are routinely preserved in a dried state for laboratory and pharmaceutical use, only a few studies have so far reported successful revival and long-term viability of nucleated mammalian cells after freeze-drying. However, there has been notable success in lyophilizing and recovering functional platelets (Wolkers, Walker, Tamari, Tablin, & Crowe, 2002; X.-L. Zhou et al., 2007). Moreover, freeze-drying platelet-rich plasma has also recently delivered encouraging results (Deprés-Tremblay et al., 2018; Shiga et al., 2016, 2017). For red blood cells, however, retaining membrane intactness and preventing hemolysis has remained a challenge (Arav & Natan, 2012). The majority of lyophilization studies on nucleated cells used various stem cells and utilized trehalose as a lyoprotectant, with limited indications to retained viability and functionality (Buchanan, Pyatt, & Carpenter, 2010b; Puhlev, Guo, Brown, & Levine, 2001; S. Zhang et al., 2010). Natan *et al.* found that adding epigallocatechin gallate, a common green tea antioxidant along with trehalose lead to high membrane integrity after lyophilization, and mesenchymal stem cells retained their clonogenic potential (Natan, Nagler, & Arav, 2009). Puhlev *et al.* reported that the rate of drying, storage under vacuum and storage temperature, as well as light exposure affects viable cell recovery (Puhlev et al., 2001). These studies may pave the way for cell lyophilization for clinical use, but as relatively complicated pre- and post-lyophilization treatments are required, there is probably a long way ahead until cell lyophilization becomes widespread.

Rehydrated non-viable lyophilized cells may be able to exert certain complex functions. A series of studies using various models confirmed that freeze-dried, non-motile spermatozoa direct normal embryogenesis and development of fertile offspring after injection into oocytes (Gianaroli et al., 2012; M.-W. Li, Willis, Griffey, Spearow, & Lloyd, 2009; Liu et al., 2004; Wakayama & Yanagimachi, 1998). Moreover, freeze-dried somatic cells have been shown to direct embryonic development after nuclear transfer (Das, Kumar Gupta, Uhm, & Lee, 2010; Loi et al., 2008a). Loi *et al.* freeze-dried sheep granulosa cells in a trehalose-containing medium and found that after one day and one month of room temperature storage, 40% of the cells retained their genome integrity. Strikingly, freeze-dried granulosa cells stored for three years at room temperature, despite evident DNA fragmentation, could direct embryonic development after injection into enucleated oocytes, probably due to activated DNA repair mechanisms (Loi et al., 2008b). The above finding suggests that lyophilized cells retaining certain complex functions may not be feasible for downstream uses involving molecular biology assays.

2.2.5 Preservation of nucleic acids in lyophilized cultured cells and tissues

Several research studies have focused on evaluating the feasibility of freeze-drying as a preservation method for nucleic acids in the cellular context. Most of such studies used low-throughput assays, such as (RT-)PCR and gel electrophoresis, to assess the integrity of nucleic acids isolated from lyophilized cells and tissues. As genomic DNA-based analytical methods are qualitative, slightly degraded samples may be sufficient for analysis. However, some RNA-based methods generally used for comparing control and pathological samples can be sensitive to even partial degradation of the target transcript.

Although purified DNA is commonly stored in a dried form, DNA in a dehydrated cellular context may be prone to destabilization due to cell-intrinsic factors, such as residual DNase activity. Zhang *et al.* assessed DNA integrity in freeze-dried fibroblasts using comet assay, and found that the extent of DNA fragmentation is less pronounced when trehalose was loaded into the cells and when the samples were stored at 4°C instead of room temperature for up to one month (M. Zhang et al., 2017). Genomic DNA extracted from freeze-dried blood was successfully used for HLA-typing (Weisberg, Giorda, Trucco, & Lampasona, 1993). These lines of evidence indicate that freeze-dried cells can efficiently store highly intact DNA for analytical applications.

Lyophilization is a promising yet underexploited method of sustainable RNA preservation in the cellular context. The available research on RNA integrity in lyophilized tissues is relatively scarce, and the analytical methods used were restricted to gel electrophoretic fragment length inspection, and RT-(q)PCR or northern blot of at most a few selected mRNAs. Two early studies from the same research group concluded that RNA samples isolated from lyophilized rat liver were stable for up to a few weeks at room temperature (as assessed by gel electrophoresis), but the detectability of the beta-actin mRNA by northern blot decreased from day 5. However, GAPDH RT-PCR showed similar results to control even after four years of room temperature storage. In general, RNA stability was lower than DNA and protein stability (Matsuo et al., 1999; Takahashi, Matsuo, Okuyama, & Sugiyama, 1995). Mareninov *et al.* found that RIN values were on average ~27% lower in lyophilized brain tissues than in controls, after one year of room temperature storage in the presence of a desiccant, while a few selected mRNAs were found intact using RT-PCR (Mareninov et al., 2013). Leboeuf *et al.* tested the effect of heat, light exposure and moisture on RIN and RT-PCR Cp values after one year of lyophilized storage, and concluded that storage in the dark and in the presence of desiccants is critical for RNA stability. Surprisingly, storage at room temperature preserved RNA better than storage at 4°C (Leboeuf et al., 2008). In addition, two studies pointed out that nucleic acid integrity in lyophilized tissues might depend on tissue lipid content and oxygen-driven peroxidation (Damsteegt, McHugh, & Lokman, 2016; Matsuo, Toyokuni, Osaka, Hamazaki, & Sugiyama, 1995). The effect of lyophilization at the whole transcriptome level, however, has not yet been previously evaluated.

2.3 Human B-lymphoblastoid cell lines – a model of the omics era

2.3.1 Cell line models in biomedical research

Cell lines can be obtained from various tissues of multicellular organisms. Cancer cell lines are derived from solid tumours or leukemic cells and have the intrinsic ability to divide in culture without any prior experimental interference. Non-cancerous cell lines may have extended (e.g. embryonic stem cells) or limited life span (e.g. embryonic fibroblasts), or can be qazi-immortalized by various laboratory techniques, including viral infection (lymphoblastoid cell lines) and artificial gene expression (HEK 293T).

Cell lines, due to their advantages, have long been used to model a plethora of structural and functional properties of cells and tissues in biological sciences. Their main general advantages are the following: 1) they can substitute experiments on primary cells or sentient animals; 2) they can be easily cultured for extended periods, enabling repeated sampling of millions of cells; 3) standardized culturing protocols enable a relatively stable extracellular environment, minimizing uncontrollable effects such as different hormonal status, nutrition or drug use, that may be present in particular *in vivo* models; 4) they can be easily engineered genetically, e.g. loss of function experiments can be carried out, including gene knock-outs or knock-downs; and 5) they can be used to produce certain biomolecules, such as antibodies (hybridoma) and therapeutics (insulin). Multiple factors should be considered when choosing the appropriate cell line for a specific project, including the organism and tissue of origin and the availability of complementary data sets in the literature.

Despite the acknowledged advantages, there is a long-standing debate over the feasibility of cell line models as substitutes of primary tissues. The basis of concern includes the lack of tissue-specific interactions and molecular evolution of cell lines over long-term culturing, which may lead to significant functional changes, ultimately resulting in cellular behaviour changes. However, due to the major advantages of cell lines over primary tissues, they will remain essential tools in a researcher’s hands to model complex cellular processes.

2.3.2 Instability of cellular phenotypes in cell line models

In nature, cellular phenotypic heterogeneity may provide an evolutionary advantage by armouring the species/organism/cell population against extremes of environmental challenges.

Well-known examples of cellular phenotypic heterogeneity include: 1) phenotypic resistance of bacteria against antibiotics and bacteriophage infection, 2) tumour cell heterogeneity and therapeutic resistance, and 3) B cell receptor (BCR) rearrangement as part of the adaptive immune system. Heterogeneity of cell populations derived from multicellular eukaryotes is expected to be reflected in newly established polyclonal cell lines.

Genomic alterations in cell culture is a major contributor to changes in cellular phenotypes over time. Genomic instability refers to the increased rate of genomic changes, including aneuploidy, large chromosomal aberrations, and small-scale mutations compared to the frequency of such changes inside the body. Genomic vulnerability is especially characteristic to cancer cell lines, in which cell cycle checkpoint mechanisms and DNA repair may already be disrupted before *in vitro* culturing (Yao & Dai, 2014). It is widely accepted that overly passaged cell lines accumulate genomic changes (M. Kim et al., 2017), although the timeline and nature of changes may differ from one cell line to another. As yet, however, there is no consensus on the maximum number of acceptable cell passages. A recent study by Ben-David *et al.* provided an invaluable insight into the extent, timeline and nature of genetic variability of MCF-7, a commonly used breast cancer cell line (Ben-David et al., 2018). Using different MCF-7 strains and assessing genotypic variability of subclones, they uncovered a rapid genetic diversification as a result of culturing condition-dependent positive selection, as well as significant differences between MCF-7 batches provided by different vendors. Therefore genetic evolution of cell lines should not be overlooked during study design and data interpretation.

Environmental factors add up to genomic alterations in modulating phenotypes of cell lines. The so-called ‘network state’ of each cell determines the extent and nature of cellular phenotype perturbations upon a particular stimulus. The variability in the composition of the least controlled culturing component, foetal calf serum, may also guide cell populations to different cell line evolutionary paths. In genetically stable cell lines, such as embryonic stem cells (ESCs), environmental factors (such as culture conditions) are expected to play the most prominent role in altering molecular phenotypes, which may lead to irreversible changes in the single-cell level. Hastreiter et al. have shown that two different culturing conditions used to maintain ESCs in the pluripotent state, namely serum+LIF (leukaemia inhibitory factor) and 2i (a combination of a GSK and a MEK inhibitor), have different effects on the expression of Nanog, a key pluripotency factor. Replacing the serum+LIF medium to 2i on ESCs leads to the stabilization of Nanog expression across the cell population by 1) inducing Nanog expression and 2) selecting against Nanog-low cells (Hastreiter et al., 2018). Therefore cellular environment should be as standardized as possible to ensure the comparability of cellular phenotype measurements.

The baseline heterogeneity of cell lines is known to decrease over long-term cell culture. The phenomenon is generally referred to as ‘clonal evolution’ when a cell or few cells with growth advantage start to dominate the culture. These changes undoubtedly affect the results of experiments carried out on younger vs older cultures. Moreover, care should be taken when comparing results from the same cell line obtained from different sources, as they have possibly gone through multiple bottlenecks individually, and show significant geno- and phenotypic divergence (Mouriaux et al., 2016; Spetzler et al., 2010). In order to minimize the effects of cell line evolution, biobanks with cell line batches of same passage numbers can be created. Moreover, monoclonal cell lines with temporally more stable cellular phenotypes may also be used for a few types of experimental applications.

2.3.3 Human B-lymphoblastoid cell lines as a model system of the omics era

Human B-lymphoblastoid cell lines (LCLs) are derived from resting B cells by Epstein-Barr virus (EBV) infection *in vitro*. The laboratory method has been in use for decades, and most commonly involves the isolation of peripheral blood mononuclear cells (PBMCs) or peripheral blood lymphocytes (PBLs) from blood, followed by treatment with EBV in the presence of immunosuppressive agents (e.g. cyclosporin A) to prevent T cell-mediated elimination of the newly infected B cells (Neitzel, 1986). Once inside the cell, EBV uses the cellular transcriptional machinery to drive the expression of genes from its dsDNA genome, encoding proteins inevitable for growth transformation (Zhao et al., 2011). As EBV predominantly targets B cells from the PBMC/PBL pool, an actively proliferating B cell-derived LCL population emerge, comprising multiple loose cell clumps with rosette morphology. LCLs then repopulate the culture by overgrowing non-infected leukocytes in a matter of weeks (Hui-Yuen, McAllister, Koganti, Hill, & Bhaduri-McIntosh, 2011).

Studies with highly diverse research aims have taken advantage of the benefits of the LCL model over the past few decades. LCLs are relatively easy to prepare, and they can be cultured in conventional media such as fetal calf serum-supplemented MEM (Minimal Essential Medium), DMEM (Dulbecco's Modified Eagle's Medium) or RPMI (Roswell Park Memorial Institute Medium), without the need for special supplements (e.g. cytokines). Using LCLs as surrogates for primary human material eliminates the need for repeated sample collection from the same individual, allaying concerns related to the loss of subjects to follow-up. Although LCLs can be maintained in culture for a large number of population doublings (~160 passages) (Oh et al., 2013), most lines remain "pre-immortal", characterized by stable diploid karyotype, low telomerase activity and the lack of tumorigenicity (Hussain & Mulherkar, 2012). Due to the generally low somatic mutation rates (0.3%) and karyotypic stability, LCLs have been historically used as an unlimited source of DNA for clinical and experimental genetics. LCLs have also provided invaluable insights into molecular events associated with EBV infection, including B cell transformation and immunological response (Bhaduri-McIntosh, Rotenberg, Gardner, Robert, & Miller, 2008; Long et al., 2005; Pokrovskaja et al., 2002; Styles et al., 2017). LCLs have also been instrumental in studying carcinogen sensitivity and DNA repair (Mazzei et al., 2011; Žegura, Volčič, Lah, & Filipič, 2008). Moreover, LCLs with certain characteristics provided useful models for studying non-EBV-related diseases, including but not limited to, neurological (Pansarasa et al., 2018), metabolic (Grassi et al., 2016), and psychiatric conditions (Milanesi et al., 2017).

The LCL model is one of the dominant contributors to the advancement of omics research. Thousands of LCLs prepared from healthy and diseased individuals with diverse ancestries have been made available by public biobanks, including Coriell Cell Repositories, for the benefit of the scientific community. Moreover, high-quality whole genome sequencing data is available for thousands of LCLs, including families, through the 1000 Genomes Project's website, providing genotype information for high throughput molecular association studies. With an average population doubling time of 24 hours and growth in suspension, high cell densities can be obtained which facilitates reaching appropriate cell numbers for high throughput sequencing-based methods requiring tens of millions of cells (e.g. ChIP-Seq). These unique advantages rendered LCLs a reasonable choice for studying chromatin structure and gene expression regulation (Grubert et al., 2015; Maya Kasowski et al., 2013; Rao et al., 2014; Reddy et al., 2012; Waszak et al., 2015), as well as drug response (Wheeler & Dolan, 2012), most in association with the cells' genome. Furthermore, an LCL named GM12878 with publicly available high-coverage genotype data, has been one of the most widely used cell lines used to uncover fundamental functional genomic processes using high throughput sequencing-coupled technologies: (1) it was the cell line for which the first kilobase-resolution genomic 3D map; (2) a model cell line of the ENCODE Project mapping transcription factor binding sites and histone modifications genome-wide and (3) the FANTOM5 Project profiling transcription start sites, (4) and other genome-wide datasets are available for this cell line including GRO-Seq (global run-on sequencing), ATAC-Seq (assay for transposase-accessible chromatin using sequencing) and DRIP-Seq (DNA-RNA hybrid immunoprecipitation sequencing), among others.

2.3.4 LCLs as surrogates for parental B cells

EBV-linked changes to B cell biology have become a field of intensive research in the past decades. The basis of interest includes the extremely high prevalence of latent EBV infection in the human population, the relatively large viral "toolset" to interfere with the host cell's normal signaling, as well as the widespread use of LCLs in biomedical research. With the emergence of high-throughput nucleic acid-profiling methods, including ChIP-Seq, Bisulfite-Seq and RNA-Seq, one can gain a global perspective on the epigenetic and transcriptomic changes triggered by EBV infection. Research efforts have been primarily focused on comparing the genomic, epigenomic and transcriptomic signatures of parental B cells/PBMCs/PBLs to their derivative LCLs. These studies provide us with an important resource to indirectly assess as to whether potential EBV infection-triggered genomic changes likely contribute to the overall functional genomic variability among LCLs.

A key aspect of the feasibility of LCLs as models for genetic association studies is genomic integrity. Genomic instability may result from EBV infection *per se* and extensive culturing, although a rapid genomic evolution comparable to that of cancerous cell lines (Ben-David *et al*., 2018) is not expected. Based on early observations of their relative genetic stability, LCLs have been historically used as continuous sources of DNA for cytogenetics and genotyping (e.g. HapMap Project, 1000 Genomes Project and Genome in a Bottle Consortium) (Gibbs et al., 2015; Jarvis, Ball, Rickinson, & Epstein, 1974; Zook et al., 2016). Series of genome-wide studies conducted by different labs have since confirmed the above assumption, comparing parental cells and transformed culture, across cell passages and repeated freeze-thaw cycles. McCarthy *et al.* found a 99.6%< concordance between 24 low passage PBMC-LCL pairs, as well as identical Mendelian error rates and levels of heterozygosity (McCarthy, Allan, Chandler, Jablensky, & Morar, 2016). The whole-genome sequencing of an individual's PBMCs and LCLs revealed that only 0.4% and 0.3% of SNVs and indels were unique to the PBMC and LCL DNA, respectively, which is within the error rate of the used technology (Nickles et al., 2012). Another study found an even lower mismatch rate (0.03-0.12%) by comparing 19 PBMC-LCL pairs, which was not significantly higher than that between controls (Herbeck et al., 2009), and the concordance level was maintained even after 170-180 passages (Oh et al., 2013). Londin *et al.* estimated an up to 99.8% genomic concordance using exome sequencing (Londin et al., 2011). Mohyuddin *et al.* estimated the somatic mutation rate to be 0.3% after 90 generations by short tandem repeat (STR) analysis (Mohyuddin et al., 2004). Studies have also shown that copy number variations are not typical to LCLs (Jeon et al., 2007; Nickles et al., 2012). In contrast, probably due to the higher energy demand of the continuously proliferating cells, the mitochondrial genome is multiplicated in LCLs (Chakrabarty et al., 2014; Jeon et al., 2007; Nickles et al., 2012). Late-passage cells (>50), however, may show loss of heterozygosity at several regions (Oh et al., 2013). A study reported mosaic loss of regions spanning > 20 Mb in 4 out of 29 studied LCLs (Shirley et al., 2012). Even though EBV exists as multiple circular episomal copies in the nucleus, there have been reports of genomic integration in LCLs derived from diseased individuals (e.g. Burkitt Lymphoma patients) (Lestou et al., 1993; Takakuwa et al., 2004; Xiao et al., 2016). Overall, the existing literature on the genomic stability of LCLs derived from healthy individuals highlight that the rate of *de novo* genomic changes associated with EBV infection *per se* is negligible.

As EBV proteins and ncRNAs have long been known to affect multiple cellular pathways leading to transcriptional changes, the question arises to what extent the LCL epigenome reflects that of its parental B cells. Due to experimental considerations, most studies used genotype-matched whole blood/PBMCs/PBLs, instead of B cells, as a baseline for comparative DNA methylation analyses. This approach has a serious limitation, namely that the proportion of CD19+ B cells among white blood cells in blood is very low (~3% in adulthood) (Morbach, Eichhorn, Liese, & Girschick, 2010), therefore it is expected that the majority of methylation differences between PBMCs and LCLs will reflect *bona fide* differences between leukocyte types, rather than changes triggered by EBV infection itself. Notable inconsistency exists in the literature regarding the extent and direction of EBV-induced methylation changes, as well as the retention of inter-individual methylation patterns; also, the mechanism of the observed methylation changes is also far from being clarified. The majority of the genome-wide methylation studies found hypomethylation at the level of both CpG islands and genome-wide after EBV infection (Hansen et al., 2014; Sugawara et al., 2011; Taniguchi, Iwaya, Ohnaka, Shibata, & Yamamoto, 2017). Differences were also found to be associated to some extent with EBV copy number (Caliskan, Cusanovich, Ober, & Gilad, 2011). Caliskan *et al.* and Sun *et al*. found that although a large-scale demethylation occurs in case of promoter CpGs and CG nucleotides genome-wide, respectively, most EBV-induced changes between B cells and their derivative LCLs are systemic and reproducible, and inter-individual variability is largely captured by the LCL model (Caliskan et al., 2011; Sun et al., 2010). However, Grafodatskaya *et al.* highlighted that methylation variability at both individual CpGs and CpG islands is slightly higher among LCLs than among their parental B cells, and the level of variability increases in high passage LCLs, indicating that random (i.e. not reproducible) changes may also emerge as a result of EBV transformation and long-term culturing (Grafodatskaya et al., 2010).

2.3.5 LCLs in genetic association studies

Quantitative trait loci (QTLs) are genomic loci which correlate with quantitative traits. Molecular QTLs are genomic variants which are associated with quantitative molecular traits, including DNA methylation levels (meQTLs), histone modification levels (hmQTLs), transcription factor binding strength (tfQTLs), DNase sensitivity (dsQTL), and gene expression levels (eQTLs). As these traits are often interrelated, one genomic region can be identified as a QTL for multiple molecular traits (e.g. tfQTL, hmQTL and eQTL). Finding genomic regions affecting molecular phenotypes may lead to a better understanding of the nature of genotype-phenotype and phenotype-phenotype interactions, and may serve as a catalogue for forming hypotheses regarding disease mechanisms.

LCLs have been used in QTL studies for the past decade (Banovich et al., 2014; J. T. Bell et al., 2011; Grubert et al., 2015; M. Kasowski et al., 2010; Odhams et al., 2017; Waszak et al., 2015). The primary reasons behind the extensive use of LCLs include: 1) obtaining cell samples for experimental use – especially from healthy control individuals – is easiest by blood drawing; 2) by infecting B cells with EBV, the resulting LCLs serve as renewable materials for repeated experiments; 3) a high number of LCLs with genotype data are available from public repositories, and 4) their karyotypes and genotypes are generally stable. LCLs are used either as primary models or as validation tools during clinical trial follow-ups. Due to their genotype stability, public genotype data is a reliable resource for association studies without the need for resequencing, and with the improvement of NGS-based methods and decreasing sequencing prices, it is easier to obtain high-quality data. Moreover, multiple consortia and most laboratories publish their functional genomic datasets in data repositories (e.g. GEO and ArrayExpress), enabling the reanalysis of relevant data from multiple sources. High-throughput chromatin conformation studies (e.g. ChIA-PET and Hi-C) enable some form of validation through uncovering chromatin interactions. One of the most widely used and thoroughly characterized cell lines is the LCL named GM12878, from which hundreds of datasets including ATAC-Seq, RNA-Seq, Bisulfite-Seq, various transcription factor and histone ChIP-Seqs and Hi-C have been generated.

Besides seeking associations between the genotype and specific molecular phenotypes, LCLs have been increasingly used as models for pharmacogenomics research. Adding to the above benefits of LCLs, a high number of cell lines can be selected with similar age, gender and race; moreover, LCLs are generally free from *in vivo* confounders. The current predominant approach to assess genotype-dependent drug response in LCLs is the use of large panels of cell lines from age- sex- and race-matched healthy individuals. Another way to use LCLs is functional follow-up studies, where LCLs carrying the pre-selected variants are used to validate genotype-drug associations predicted in clinical samples or cell line-based screening studies. The so-called triangle approach described by Huang *et al.* (Huang et al., 2007) may reduce false positive hits by incorporating gene expression data into pharmacogenomic studies. The model assumes that RNA expression is the dominant mediator between genotype and drug response phenotypes. Therefore variants with high significance in all possible associations are the strongest candidates. LCLs have been the models for studies assessing sensitivity to various chemotherapeutic agents (Hartford et al., 2009; Huang et al., 2007; O’Donnell et al., 2010), lipid-lowering-drugs (Mangravite et al., 2010), and antidepressants (Morag et al., 2011), among others.

3 AIMS OF OUR STUDIES

Aim 1. Development and characterization of phage display-based procedure controls for ChIP experiments

ChIP protocols generally lack procedure controls allowing for normalization of uneven sample loss, possibly leading to experimental bias in comparative ChIP experiments. In our study, we aimed to

- Develop a phage display-based spike-in procedure control system to track and normalize for uneven sample loss during ChIP
- Select androgen receptor (AR)-mimicking phages through multiple rounds of *in vitro* evolution, followed by diversity assessment and ChIP-based AR antibody affinity measurement of polyclonal stocks
- Prepare monoclonal stocks from the highest affinity polyclonal batch, followed by ChIP-based AR antibody affinity measurement of resulting stocks

Aim 2. Assessing lyophilization as an alternative means to archive human cells for RNA-based studies

Lyophilization and room temperature storage has emerged as an alternative and cost-effective method of biosample stabilization for storage and transport. However, its widespread adoption for tissue handling has not yet been taken place. In this work, our purpose is to

- Test whether a cell membrane stabilizer, epigallocatechin-gallate supports cell lyophilization and subsequent RNA- and ChIP-based measurements
- Assess RNA integrity and RNA yield using low-scale methods, and measure multiple gene types at various expression levels by RT-QPCR from samples isolated from LCL cells right after lyophilization in 0.1 M trehalose/PBS, or after two or eight weeks of room temperature storage

- Profile the transcriptome of paired fresh and lyophilized LCLs stored at room temperature for two weeks by mRNA-Seq, and apply QC measures to compare library features, e.g. library complexity, read GC content and read mismatch rate

- Perform various function- and sequence-based analyses to uncover the characteristics of genes downsampled in lyophilized cells

Aim 3. Evaluation of the genotype-independent variability of LCLs at multiple cellular phenotype levels related to gene regulation and response to an external stimulus

So far, little has been known about the extent of functional genomic variability of non-genetic origin among LCLs, a widely used model for genetic association studies. Utilizing isogenic LCLs derived from the same individual, we aimed to:

- Perform basic cell line characterizations, including short tandem repeat analysis, cell cycle stage assessment and immunophenotyping using flow cytometry in order to exclude major differences that might bias our results with functional genomic assays

- Perform ChIP-Seq experiments to map H3 histones acetylated at the 27th lysine residue (H3K27ac) in order to map and compare active gene regulatory element activities genome-wide in isogenic LCLs

- Profile the transcriptome of isogenic LCLs by mRNA-Seq and analyze affected genes and the relationship between chromatin profiles and RNA levels

- Assess the biological functions and other characteristics of variable genes, and to examine a possible relationship between variable pharmacogene mRNA expression and chemotherapeutic drug response on the example of the *DPYD* gene and 5-fluorouracil

4 MATERIALS AND METHODS

4.1 Cell culture

Human B-lymphoblastoid cell lines were obtained from Coriell Cell Repositories as actively proliferating live cultures. Before experiments, three-tiered biobanks were created for each LCL following Sigma-Aldrich's guideline (https://www.sigmaaldrich.com/technical-documents/protocols/biology/good-cell-banking.html) in order to provide a continuous source of cells with the same number of freeze-thaw cycles and passages. Cells were cultured in RPMI-1640 medium (Sigma-Aldrich, cat. R0883) supplemented with 15 v/v% heat-inactivated FCS (Thermo Fisher Scientific, cat. 10270-106), 2 mM L-glutamine (Sigma-Aldrich, cat. G7513) and 1 v/v% penicillin-streptomycin (Sigma-Aldrich, cat. P4333), and were kept at 37°C in T25 or T75 cell culturing flasks in an upright position. Tier 3 cells (working biobank) were used for all experiments. Table 1 shows the main characteristics of LCLs used in our studies.

Table 1 | The main characteristics of the LCL cells used in our study.

Cell line	Alt. name	Pedigree	Age of source	Gender of source	Race of source	Disease state of the source	Relationship with other LCLs
GM22647	sGT_1	-	26	male	CEPH/UTAH	Healthy	These five cell lines were derived from the same individual (five vials of peripheral blood).
GM22648	sGT_2	-					
GM22649	sGT_3	-					
GM22650	sGT_4	-					
GM22651	sGT_5	-					
GM12864	Trio_S	CEPH/UTAH 1459, son	N/A	male	CEPH/UTAH	Healthy	Son of GM12872 and GM12873
GM12872	Trio_F	CEPH/UTAH 1459, father	N/A	male	CEPH/UTAH	Healthy	Father of GM12864
GM12873	Trio_M	CEPH/UTAH 1459, mother	N/A	female	CEPH/UTAH	Healthy	Mother of GM12864

N/A = information not available

4.2 Short tandem repeat analysis

10^6 cells were washed once with PBS (1 ml), pelleted (500g, RT) and genomic DNA was isolated using Roche's High Pure PCR Template Preparation Kit (Roche Life Science, cat. 11796828001). Five short tandem repeat (STR) regions (AMELY/AMELX, D18S51, D8S1179, TH01 and FGA) were amplified with the PowerPlex S5 System (Promega, cat. TMD021). We used the ABI PRISM 3100-Avant Genetic Analyzer and the GeneMapper ID software (version 4.1) to detect PCR products and fragment analysis, respectively (Department of Laboratory Medicine, Faculty of Medicine, University of Debrecen).

4.3 Immunophenotyping and cell cycle analysis

For immunophenotyping, we used flow cytometry and eight-colour labeling. PBS-washed LCLs were stained with combinations of fluorescently labeled antibodies in three separate tubes. Table 2 shows the antibodies, clones and fluorochromes used in our studies.

Table 2 | Antibodies, clones and fluorochromes used in our studies.

	FITC	**PE**	**PerCP-Cy5.5/ PC5.5**	**PC7**	**APC**	**APC-AF750**	**PB**	**PO**
Tube 1.	CD23 *(9P25)*	CD22 *(S-HCL-1)*	CD5 *(L17F12)*	CD19 *(J3-119)*	CD38 *(HB-7)*	CD81 *(JS-81)*	CD20 *(L27)*	CD45 *(HI30)*
Tube 2.	Kappa *(TB28-2)*	Lambda *(1-155-2)*	HLA-DR *(L243)*	CD19 *(J3-119)*	CD79b *(SN8)*	CD43 *(DFT1)*	FMC7 *(FMC7)*	CD45 *(HI30)*
Tube 3.	nTdT *(HT-6)*	cyIgM *(polyclonal)*	CD34 *(8G12)*	CD19 *(J3-119)*	CD10 *(HI10a)*	CD24 *(ML5)*	CD21 *(LT21)*	CD45 *(HI30)*

Abbreviations: APC, allophycocyanin; APC-AF750, conjugated allophycocyanin-Alexa fluor 750; cyIgM, cytoplasmic immunoglobulin M; FITC, fluorescein isothiocyanate; nTdT, nuclear terminal deoxynucleotidyl transferase; PB, pacific blue; PC5.5, phycoerythrin cyanin 5; PC7, phycoerythrin cyanin 7; PE, phycoerythrin; PerCP-Cy5.5, peridinin chlorophyll protein 5.5; PO, pacific orange.

Antibodies against CD5, CD10, CD20, CD22, CD24, CD34, CD38, CD79b, CD81, FMC7, HLA-DR, kappa and lambda markers were purchased from Becton Dickinson Biosciences (San Jose, CA); antibodies against CD19, CD23, and CD43 were purchased from Beckman Coulter (Brea, CA); anti-CD21 and anti-CD45 were purchased from Exbio (Prague, Czech Republic);

nTdT and IgM were purchased from Dako (Glostrup, Denmark). Standard procedures were used for surface staining: combinations of labeled antibodies, each at saturating concentration, were added to 10^6 cells (in 50 µl) and incubated for 15 minutes at room temperature (RT) in the dark. Samples were washed once with PBS and resuspended in 500 µl PBS containing 1% paraformaldehyde. Intracellular staining was carried out after surface staining, strictly following the procedure described for Intrastain (Dako Glostrup, Denmark). 10^5 events were acquired with a FACS Canto II flow cytometer (Becton Dickinson, San Jose, CA). Data were analyzed by FACS Diva (Becton Dickinson Biosciences, San Jose, CA) and Kaluza Software version 1.2 (Beckman Coulter, Brea, CA). The flow cytometer was calibrated daily, using Tracking fluorescent microbeads (cat. 641319, Becton Dickinson, San Jose, CA) and Autocomp software.

For cell cycle analysis, $2*10^6$ cells were washed with PBS at room temperature and fixed with 70 v/v% ethanol at 4°C. The cells were pelleted by centrifugation and were incubated with RNase (0.5 ml of 2 mg/ml RNase in PBS) and Propidium-iodide (0.5 ml working solution of 100 µg/ml Propidium-iodide, 1% Triton-X-100, 12.7 µM EDTA in PBS) for 30 minutes at room temperature, in the dark. 20,000 events were acquired using a FACS Calibur II flow cytometer (Becton Dickinson, San Jose, CA). Data were analyzed with ModFit LT for Mac 2.0 (Becton Dickinson, San Jose, CA).

4.4 In vitro evolution of transcription factor-mimicking phages

We screened a premade random heptapeptide library (Ph.D™-7 Phage Display Peptide Library Kit, New England Biolabs) against magnetic bead-coupled AR antibody (N-20) (Santa Cruz Biotechnology, cat. sc-816) in Eppendorf LoBind microcentrifuge tubes (cat. Z666548), with four consecutive rounds of biopanning. For each biopanning round, phage peptide library (equivalent to 10^{10} or 10^{11} plaque-forming units (PFUs)) was blocked with 1 ml TBST/BSA (0.1% Tween-20, 1 v/w% BSA in TBS) for 30-60 minutes at room temperature. 10 µg of the AR antibody was added to the phages, and incubated for 10-60 minutes at room temperature, using a rotating rack. A Protein A:Protein G paramagnetic bead mix (50 µl, 1:1 ratio; Life Technologies, 10002D and 10004D) was washed twice with 500 µl TBST/BSA, added to the antibody-phage reactions and incubated for 20 minutes at room temperature on a rotating rack. Beads-antibody-phage complexes were washed ten times with 500 µl TBST/BSA and eluted two times (0.2 M

Glycine-HCl, pH 2.2, with 1 mg/ml BSA) by rotating for 20 minutes at room temperature. We used 150 µl 1M TRIS-HCl, pH 9.1 for neutralizing the acidic elution buffer per ml eluate. The neutralized phage eluate was used to infect 25 ml F+ ER2738 bacterial strain (early-log growth phase), and infected bacteria were grown using vigorous shaking in non-selective LB. Bacterial cultures were centrifuged at high speed twice, and 80% of the supernatant was precipitated at 4°C with 1/6 volumes of 20% (v/w) PEG-8000/2.5M NaCl overnight. Precipitated samples were centrifuged, and pellets were dissolved in TBS. The second round of precipitation was carried out using the same precipitating agent for 1-2 hours on ice and high-speed centrifugation. Phage pellets were redissolved in 200 µl TBS. We used the SmartSpec Plus spectrophotometer from Bio-Rad to measure phage concentration (260 nm) and calculated phage concentration as per the manufacturer's recommendations. Glycerol was added to the phage stocks at 50% final concentration, and stocks were stored at -20°C. The above stock was used as starting material for three additional rounds of biopanning (10^{10} PFUs).

4.5 Phage subcloning

NEB's phage titering and plaque amplification protocol was used for generating monoclonal phage reagents, with slight modifications. The polyclonal phage stock prepared during biopanning (fourth round) was diluted in LB, and 10 µl was added to 200 µl ER2738 cells (at mid-log growth phase). The infected cells were added to 3 ml 45°C "Top Agar" and, poured onto warmed LB/IPTG/X-gal plates and were incubated at 37°C overnight. The upcoming day, an overnight culture of ER2738 was diluted 1:100 and were grown to OD_{600}=0.3-0.5, and individual blue plaques from the LB/IPTG/X-gal plates were resuspended in 1 ml of the ER2738 cells. The cultures were shaken for 4.5 hours at 37°C, and the tubes were centrifuged at 4°C. One-sixth volume of 20% (v/w) PEG-8000/2.5M NaCl was added to 80% of the supernatant and incubated overnight at 4°C. The next day, pellets were redissolved in 200 TBS and precipitated again with 1/6 volume of 20% (v/w) PEG-8000/2.5M NaCl at 4°C, resuspended in 200 µl TBS and precipitated on ice for 2 hours with 1/6 volume 20% (v/w) PEG-8000/2.5M NaCl. The pellet was resuspended in TBS. Concentrations were calculated using NEB's formula and OD_{260} values (SmartSpec Plus, Bio-Rad). Monoclonal samples were stored at -20°C with 50% glycerol.

4.6 Chromatin preparation and robotic phage ChIP-qPCR

List of buffers:

- Sonication Buffer: 1% SDS; 10 mM EDTA; 50 mM Tris-HCl (pH 8.1)
- IP Buffer: 0.01 v/v% SDS; 1.1 v/v% Triton X-100; 1.2 mM EDTA; 16.7 mM Tris-HCl (pH 8.1); 167 mM NaCl
- Wash Buffer A: 0.1 v/v% SDS; 1v/v% Triton X-100; 2 mM EDTA; 20 mM Tris-HCl (pH 8.1); 0.15 M NaCl
- Wash Buffer B: 0.1 v/v% SDS; 1 v/v% Triton X-100; 2 mM EDTA; 20 mM Tris-HCl (pH 8.1); 0.5 M NaCl
- Wash Buffer C: 1 v/v% NP-40, 1 v/v% Na-deoxycholate; 1 mM EDTA, 20 mM Tris-HCl, 0.25 M LiCl (pH 8.1)
- TE buffer: 1 mM Tris-HCl, 0.1 mM EDTA (pH 8.0)
- Elution Buffer: 0.1 M NaHCO3, 1% SDS

HEK293T cells were cultured in DMEM (Sigma-Aldrich, D5671), supplemented with 10 v/v% FCS (Life Technologies, cat. 10270-106), 2 mM L-Glutamine (Sigma-Aldrich, G7513) and 1 v/v% Penicillin-Streptomycin (Sigma-Aldrich, cat. P4333-100ML). “Buffer chromatin” was prepared from 1*10^7 HEK293T cells. Cells were fixed with 1 v/v% formaldehyde (Sigma-Aldrich, cat. F8775) for 10 minutes at room temperature, and formaldehyde was quenched using 0.125 M glycine (final concentration) at room temperature for 5 minutes. Cells were washed twice with ice-cold PBS, scraped up from the flask and resuspended in 1 ml PBS. 10^7 cells were resuspended in Sonication Buffer and sonicated using Bioruptor Plus (Diagenode). The sonicated sample was centrifuged, and the supernatant was frozen at -80°C in small aliquots until use. Diagenode’s IP-Star Automated System was used to carry out IP reactions until the bead elution step. IP reactions were prepared in plates by adding buffers and reagents into different wells of each row: 1 µg anti-AR antibody (or isotype-matched IgG) diluted in 100 µl IP buffer; chromatin diluted to 10^5 cell-equivalent/100 µl, 10^6 phage particles, and 1 mM DTT; 100 µl of Wash Buffer A; 100 µl of Wash Buffer B; 100 µl of Wash Buffer C; 100 µl of TE buffer; 100 µl of Elution Buffer and 10 µl of Protein A:Protein G bead mix (1:1 ratio, Life Technologies, 10002D and 10004D). The IP-Star was programmed so that the following IP protocol was performed per reaction: (1) the magnetic bead mix was incubated with AR antibody in IP buffer at 4°C; (2) the chromatin-phage mix was slowly suspended with the bead-antibody mix for 6 hours at 4°C; (3) complexes were washed at 4°C once with Wash Buffer A; (4) once with Wash Buffer B; (5) once with Wash Buffer C; and (6) once with TE buffer; (7) samples were eluted using Elution Buffer at room temperature. Sonication buffer, IP Buffer and Wash buffers contained 100x Proteinase inhibitor cocktail (Sigma-Aldrich, cat. P8340). 8 µl of 5 M NaCl and 2 µl of 0.5 M EDTA (pH 8.0) was added to the eluates and were incubated at 65°C for 4 hours to overnight. 10 µg RNAse A (Sigma-Aldrich, cat. R5503) was added, and samples were incubated for 30 minutes at 37°C, followed by 2 hours of incubation with 10 µg Proteinase K (Thermo Fisher Scientific, cat. EO0491) at 45°C (in a thermoshaker, at 1000 rpm). The High Pure PCR Template Preparation Kit (Roche, cat. 11796828001) was used to purify DNA from eluates and IP Buffer as per the manufacturer's instructions. M13 universal primers designed by Roche's UPL Assay Design Center, and UPL probe 48 (human Universal ProbeLibrary Set, 04683633001) was used for qPCR measurement of phage genomes. The LightCycler 480 instrument (Roche) was used to measure amplification signals, and input samples were used to normalize data ($2^{-\Delta Cp}$ method).

Sequence of qPCR primers:

M13 forward: 5' ATTCACTGGCCGTCGTTTTA 3'
M13 reverse: 5' GGCGATTAAGTTGGGTAACG 3'

4.7 Capillary sequencing

The phage genome was isolated using Roche's High Pure PCR Template Preparation kit (cat. 11796828001) following the manufacturer's instructions. We used the -96 gIII primer binding to the M13KE genome at a constant region downstream of the ROI. For details, please refer to NEB's "Ph.D. Phage Display Libraries" Manual. We used the ABI 310 Avant sequencer to run the sequencing reactions and detect fluorescent fragments. The sequencing reactions were run at the Genomic Medicine and Bioinformatic Core Facility of the University of Debrecen.

4.8 Chromatin immunoprecipitation

The presented method describes the ChIP protocol used for LCL ChIP-Seq.

List of ChIP-Seq buffers:

- ChIP Lysis Buffer: 1 v/v% Triton X-100, 0.1 w/v% SDS, 150 mM NaCl, 1 mM EDTA pH 8.0, 20 mM Tris pH 8.0

- IP Wash Buffer I: 1 v/v% Triton X-100, 0.1 w/v% SDS, 0.1 v/v% sodium-deoxycholate, 150 mM NaCl, 1 mM EDTA pH 8.0, 20 mM Tris pH 8.0
- IP Wash Buffer II: 1 v/v% Triton X-100, 0.1 w/v% SDS, 0.1 v/v% sodium-deoxycholate, 500 mM NaCl, 1 mM EDTA pH 8.0, 20 mM Tris pH 8.0
- IP Wash Buffer III: 0.5 v/v% NP-40, 0.5 v/v% sodium-deoxycholate, 0.25 M LiCl, 1 mM EDTA pH 8.0, 20 mM Tris pH 8.0
- IP Wash Buffer IV: 10 mM EDTA pH 8.0, 200 mM Tris pH 8.0
- Bead Elution Buffer: 100 mM NaHCO3, 1 w/v% SDS

For isogenic LCL experiments, we carried out experiments in biological duplicates, i.e., cells were originating from different vials of the same freezing batch (biobank tier 3; an equal number of passages and freeze-thaw cycles), but the resuscitation of cells for replicate one and two experiments was carried out on separate days. After resuscitation, cells were passaged once and expanded using standard procedures and reagents. Cell numbers were set to 800,000/ml 12 hours before experiments. Cell harvesting was carried out at the same time-point of the day for all experiments. $2*10^7$ cells were washed with PBS and crosslinked using either single-reagent crosslinking (1 % methanol-free formaldehyde (Thermo Fisher Scientific, cat. 28908) for 10 minutes at room temperature; for isogenic LCLs) or two-reagent crosslinking (2 mM di(N-succinimidyl) glutarate (Sigma-Aldrich, cat. 80424-5MG-F) for 45 minutes at room temperature and then with 1 % methanol-free formaldehyde at room temperature; for trio LCLs). Cell suspensions were then spiked with glycine (0.125 M final concentration), and incubated with for 5 minutes at room temperature to quench formaldehyde. Fixed cells were washed with ice-cold PBS, and nuclei were isolated by resuspension in ice-cold ChIP Lysis Buffer and centrifugation at 4°C (repeated three times). Nuclei were sonicated in 1 ml cold ChIP Lysis Buffer (Bioruptor Plus, Diagenode; low frequency, five cycles, 30 minutes ON and 30 minutes OFF). Nuclear debris was sedimented by high-speed centrifugation at 4°C, and the top 90% of the supernatant was carried over for further experimental steps. At this step, 50 µl chromatin was set aside and stored at -20°C until use (as “input” for ChIP-qPCR experiments). Prior to immunoprecipitation (IP), sonicated chromatin was diluted ten times with cold ChIP Lysis Buffer, and chromatin equivalent to $5*10^6$ cells were immunoprecipitated with 2.5 µg anti-histone H3 (acetyl K27) antibody (Abcam, cat. ab4729) or isotype control antibody (Santa Cruz Biotechnology, cat. sc-

2027 X) on a rotating rack, overnight, at 4°C. 47.5 µl Protein A-Protein G paramagnetic bead mix (1:1 ratio) (Thermo Fisher Scientific, cat. 10002D and 10004D) per IP reaction was blocked with 0.5 w/v% BSA/PBS overnight, at 4°C. The following day, IP reactions were centrifuged, and the top 90% was incubated with blocked beads for 6 hours on a rotating rack, at 4°C. Capture beads were washed once IP Wash Buffer I, twice with IP Wash Buffer II, once with IP Wash Buffer III and twice with IP Wash Buffer IV for 3 minutes on a rotating rack, collecting beads using a magnetic rack. All buffers were ice-cold and were supplemented with cOmplete Mini proteinase inhibitor tablets (Roche, cat. 11697498001). Beads were eluted twice using Bead Elution Buffer (2x100 µl, 2x15 minutes at room temperature, 1,000 rpm). Input samples were set to 200 µl with Bead Elution Buffer and treated the same way as IP samples during all subsequent steps. Crosslinks were reversed at 65°C overnight, in the presence of 0.4 M NaCl. Eluates were treated with 20 µg RNase A and 40 µg Proteinase K, and DNA fragments were isolated using Qiagen's MinElute PCR purification kit (cat. 28006). DNA concentrations were determined using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, cat. Q32851). Samples were tested before and after library preparation for the presence of selected positive and negative control regions by qPCR. For pre-library qPCRs, IgG-precipitated samples were used as additional controls of IP specificity, and input samples were used as normalizers. The UPL Assay Design Center was used to design primer pairs and appropriate Universal ProbeLibrary probes (Roche Applied Science, Germany). Reactions were run on a LightCycler 480 instrument.

For the lyophilization study, we used the same protocol, but used $5*10^6$ (1) fresh cells, (2) cells lyophilized in IMT-2 (0.1 M trehalose, 0.945 mg/ml (-)-epigallocatechin gallate (EGCG) in PBS), (3) cell lyophilized in IMT-2 and fixed with 0.1% formaldehyde (FA), (4) cells incubated with 0.1 M trehalose overnight at 37°C prior to lyophilization in IMT-2, and (5) FA-prefixed cells lyophilized in IMT-2 . For assessing the effect of pure EGCG on ChIP efficiency of SPI1 and CT64, we added 2 mM EGCG to the culturing medium 1 hour before cell harvesting. Otherwise, the above protocol was followed.

Sequence of ChIP oligos:

	Forward (5’ > 3’)	Reverse (5’ > 3’)
SPI1	GATGGGAGGGAGAACGTGT	GCATTTGTTGGGTTAGAGCAA
CT64	CAGCAATTGTGAGGCTCTGA	GCACCTGTTGAGTTTGGTCTG

4.9 MNase profiling

For MNase profiling, nuclei were washed in MNase buffer containing BSA (1x concentration, NEB's kit, cat. M0247) and proteinase inhibitors, and were treated at 37°C for 30 minutes with MNase enzyme at ratios of 66.6 Gel Units, 22.2 Gel Units and 7.4 Gel Units per $1.5*10^6$ nuclei (in 200 µl volume). Reactions were stopped with 20 µl MNase stop solution (5 v/w% SDS, 250 mM EDTA), were supplemented with 280 µl ChIP Lysis Buffer (containing proteinase inhibitors) and sonicated using a Bioruptor Plus sonicator (Diagenode) for 3 minutes at low setting. Cell debris was pelleted at 4°C, and 80% of the supernatant was precipitated with three volumes of absolute ethanol at -20°C overnight. The next day nucleic acids were pelleted at 4°C and were desiccated at room temperature. Fragmented DNA was recovered substantially the same way as during the ChIP protocol: reverse crosslinked (here for 4 hours), treated with RNase A and Proteinase K, and isolated using Quiagen's MinElute kit. Sample concentrations were approximated using NanoDrop and were run on a 1% agarose gel stained with ethidium bromide, using 1 kb DNA ladder (Thermo Fisher Scientific, cat. 15615-024) as DNA fragment length markers. For ChIP with MNase digestion, we chose and used the 22.2 Gel Units/$1.5*10^6$ cells setup, and after the mild sonication, chromatins were subjected to the ChIP protocol described in the previous paragraph.

4.10 ChIP-Seq library preparation and sequencing

We followed the TruSeq ChIP Sample Preparation Guide 15023092 B (Illumina, San Diego, CA, USA) with minor modifications to prepare indexed ChIP-Seq libraries, using 10 ng ChIP DNA as starting material. Libraries were pooled and sequenced to a 50-bp read length (single-end) at the Genomic Medicine and Bioinformatics Core Facility (University of Debrecen; NextSeq 500 system) or at the EMBL Genomics Core Facility (Heidelberg, Germany; HiSeq2000 system). We used the bcl2fastq software for demultiplexing sequencing data.

4.11 ChIP-Seq data analysis

Sequencing reads were aligned to the human genome (hg19; GRCh37) using BWA 0.7.10. H3K27ac-enriched genomic regions were predicted using HOMER 4.9.1, and 'blacklisted' genomic regions identified by the ENCODE Project were removed using bedtools (subtract). We used DiffBind (Bioconductor) to define the consensus (merged) region set for further analyses.

RPKM values were calculated, replicate samples were clustered, and a correlation matrix mas generated using DiffBind. Plotly 3.0.0. (Python) was used for creating the correlation heatmap. We used two-way ANOVA combined with Tukey's post hoc test (functions: aov() and TukeyHSD()) from the MASS R package to define differentially enriched regions (P value < 0.05, fold-difference > 2). Super-enhancers were predicted from bam files pooled from all replicates (SAMtools). Tag directory was created using HOMER's makeTagDirectory, and super-enhancers were predicted using findPeaks (HOMER). We used the R package pheatmap to cluster differentially acetylated regions, and data to the read distribution heatmap was generated by annotatePeaks (-hist function; HOMER). Regions were centered to the most central nucleosome-free region (predicted using HOMER's getPeakTags with -nfr function), and tag densities were visualised using Java TreeView (+/- 1 kb). Closest genes were assigned using bedtools (closest). BedGraphs were visualized using Integrative Genomics Viewer (IGV, Broad Institute). The 3D Genome Browser (Yue Lab, http://promoter.bx.psu.edu/hi-c/index.html) was used to visualize Hi-C data from the GM12878 cell line with 40 kb resolution. PhenoGram was used to visualise differentially acetylated regions over chromosome (sGT_1 vs sGT_2).

4.12 Multiplexed 3C-Seq

We followed the protocol for multiplexed 3C-sequencing described by Stadhouders *et al.* (Stadhouders et al., 2013) with minor modifications. We washed 10^6 untreated GM12872, GM12873 and GM12864 cells in 10% FCS/PBS, and fixed in 10% FCS/PBS containing 1% ultrapure formaldehyde (Thermo Fisher Scientific, cat. 28908) for 10 minutes at room temperature, gently rotating the tubes. Reactions were quenched with 0.125 M Glycine/PBS, which was followed by centrifugation at room temperature. Cells were washed twice with ice-cold PBS, resuspended in 3C Lysis Buffer (10 mM Tris-HCl pH 8.0, ten mM NaCl, 0.2% NP-40) containing proteinase inhibitors (Roche, cat. 11697498001), and incubated on ice for 7 minutes. The nuclear prep was washed with ice-cold PBS. Nuclei were pretreated with 0.3% SDS for 1 hour (37°C, 900 rpm) and another hour with 2% Triton-X-100, followed by digestion with 400 U HindIII-HF (NEB, cat. R3104S) overnight at 37°C, followed by another round of digestion with 400 U HindIII-HF for 6 hours. We added 7% SDS, and heat-inactivated HindIII at 65°C for 25 minutes. Nuclei were incubated for 1 hour at 37°C with 5% Triton-X-100 in ligation buffer. Digested samples were ligated with 100 U T4 ligase (NEB, cat. M0202S) in 7 ml final volume at

16°C for 4 hours, followed by de-crosslinking at 65°C overnight in the presence of 3 μg Proteinase K (Thermo Fisher Scientific, cat. EO0491). The next day, samples were incubated with 10 μg RNase A ((Sigma-Aldrich, cat. R5503) for 1 hour at 37°C. DNA was isolated using the phenol-chloroform-isoamyl alcohol (PCI) method (Sigma-Aldrich, cat. P3803I), and stored in TRIS-HCl (pH 7.5) at -20°C (3C library) until the second digestion-ligation round. 10 μg of DNA was digested overnight with 10 units of MseI (NEB, cat. R0525S) at 37°C. Samples were isolated with phenol-chloroform-isoamyl alcohol and ligated using 100 units of T4 ligase at 16°C overnight. Ligation samples were isolated with phenol-chloroform-isoamyl alcohol and resuspended in Tris-HCl (pH 7.5). DNA samples were amplified using Expand Long Range, dNTPack (Sigma-Aldrich, cat. 4829034001) using primers specific to the so-called bait sequence, a region in the P3H2 gene body (inverse PCR). PCR reactions were cleaned up using Qiagen's MinElute PCR purification kit (cat. 28006). We followed the TruSeq ChIP Sample Preparation Guide 15023092 B (Illumina, San Diego, CA, USA) with minor modifications to prepare indexed 3C-Seq libraries. Digestion and ligation efficiencies were checked against appropriate controls using agarose gels. Sample concentrations were measured using the Qubit HS dsDNA kit throughout the protocol, and agarose gels were used to assess digestion and ligation efficiencies. Samples were sequenced (75-bp) at the EMBL Genomics Core Facility (Heidelberg, Germany; HiSeq2000 system).

Sequence of inverse PCR primers:

P3H2_left: CAGTGGTGGAGTGCTGTAAAG
P3H2_right: CCCACACTACTAAGGAAAGCTC

4.13 Lyophilization

$3*10^6$ cells at log-growth phase were washed with PBS and resuspended in 0.5 ml lyophilization solution containing 0.1 M D-(+)-Trehalose dihydrate (Sigma-Aldrich, cat. T9531) in PBS, in polypropylene microcentrifuge tubes. Cell suspensions were then snap-frozen by immersing in liquid nitrogen and kept on dry ice. Before loading into the freeze dryer, each tube was opened, and a piece of parafilm with seven holes (1 mm diameter each) was placed over the tube’s opening. We loaded the samples into a pre-cooled CoolSafe 110 freeze dryer (ScanVac, LaboGene, Denmark), at a condenser temperature of -110°C (Proteomics Core Facility, University of Debrecen). Lyophilization was carried out at 0.004 mBar for six hours, at an environmental temperature set to 22°C. After finishing the lyophilization cycle, tubes were closed, and lyophilized cell powders were processed immediately, or stored for two weeks or two months at room temperature (23–25°C) in a tightly sealed, non-transparent box, in the presence of $CaCl_2$ dihydrate.

4.14 RNA isolation

Total RNA was isolated using the TRIzolate method. In the lyophilization project, $3*10^6$ PBS-washed, pelleted cells (control) or lyophilized cell powders (in trehalose matrix or in IMT-2 matrix where indicated in the Results section) or washed lyophilized cells (in IMT-2 matrix; after resuspension in 0.1 ml nuclease-free water and two rounds of PBS washes at room temperature) were resuspended in 1 ml TRIzolate (UD-Genomed Medical Genomic Technologies Ltd., cat. URN0103), and vortexed for 5 minutes at room temperature. In the project involving isogenic LCLs, $2*10^6$ cells (biological duplicates, see Chromatin immunoprecipitation section) were washed twice with PBS, resuspended in 1 ml TRIzolate, and vortexed for 5 min at room temperature. Chloroform (Sigma-Aldrich, cat. C2432) was added (1:5 ratio), and samples were vortexed for 5 minutes at room temperature, followed by high-speed centrifugation. The aqueous phase was collected and precipitated using isopropanol (1:1, Sigma-Aldrich, cat. I9516), for 10 minutes at room temperature. Pellets were washed twice with chilled 75% ethanol (absolute ethanol: VWR International, cat. 20821.296), and vacuum-desiccated. RNA pellets were redissolved in nuclease-free water (AccuGENE, Lonza, cat. 51200) and incubated at 65°C for 10 minutes. A NanoDrop 1000 instrument (Thermo Fisher Scientific, Waltham, MA, USA) was used to assess sample purity, and concentrations were determined using the Qubit RNA HS Assay Kit (Thermo Fisher Scientific, cat. Q32855). Agilent RNA 6000 Nano microchips (Agilent, Santa Clara, CA, USA) were used to analyze fragment distributions and to determine RIN values according to the manufacturer’s instructions.

4.15 mRNA-Seq library preparation and sequencing

RNA-Seq libraries were prepared from 1 μg total RNA following Illumina’s TruSeq RNA Sample Preparation v2 Guide (with poly(A) selection). Indexed and pooled libraries were sequenced on the NextSeq 500 system using 50-bp (isogenic LCL study) or 75-bp (lyophilization study) read length (single-end). We used the bcl2fastq Conversion Software (Illumina, San

Diego, CA, USA) for demultiplexing the sequencing data based on sample-specific indices. For the isogenic LCL study, all steps related to library preparation, cluster generation, sequencing and base calling were carried out at the Genomic Medicine and Bioinformatic Core Facility (University of Debrecen) using a NextSeq 500 sequencer (Illumina, San Diego, CA, USA). For the lyophilization study, library preparation was performed at the Genomic Medicine and Bioinformatic Core Facility (University of Debrecen), while cluster generation, sequencing and base calling were performed at the 2nd Department of Pediatrics (Semmelweis University) using a NextSeq 500 sequencer (Illumina, San Diego, CA, USA).

4.16 mRNA-Seq data analysis

In the lyophilization project, reads were aligned to hg19 (GRCh37) with TopHat v2.0.7 (--max-multihits 1). Transcript abundances were calculated using Cufflinks and UCSC's gene annotation track. All genes below the FPKM threshold of 1 in all samples were discarded, as were poly(A)-free small RNAs (due to ambiguous capture during library preparation). We used Cuffdiff to find differentially sampled RNAs between control and lyophilized samples (FDR = 0.05). We used the QoRTs package to obtain metadata regarding GC content, per-base mismatch profile, chromosome distribution, gene body coverage and absolute read count per gene. We assessed cumulative gene diversity by plotting the fraction of reads mapping to the top 10, 100, 1,000 and 10,000 genes. RNA biotypes were assigned to genes with available HGNC IDs (retrieved from https://www.genenames.org/) using ENSEMBL v91 annotation. The DAVID Bioinformatics Resources 6.8 tool was used for functional annotation (https://david.ncifcrf.gov/). We downloaded human lncRNAs (most extended variant) and protein-coding genes (most extended variant with available coding sequence (CDS)) from the HGNC database. DNA sequences were retrieved from ENSEMBL v91 genes using BioMart, and a custom bash script was used to calculate sequence lengths and GC contents. We used the two-tailed non-parametric Wilcoxon rank-sum test (Mann-Whitney U test), which also accounts for the different sample sizes to compare features of control and differentially expressed transcripts. We acquired ARE data from the ARED-Plus database, and draw the Venn diagram using BioVenn.

In the isogenic LCL project, sequencing reads were aligned to the human reference genome hg19 (GRCh37) with TopHat v2.0.7 (--max-multihits 1), we calculated transcript abundances and filtered bath effects using edgeR and UCSC's gene annotation track (hg19, Illumina's iGenomes database, version 07/17/2015). Genes below the CPM threshold of 5 across all samples were discarded. Expression values were represented as RPKM values. EdgeR (ANOVA) was used to identify differentially expressed genes (FDR = 0.05, fold-difference > 2). Z-scores were calculated for each gene per sample using the following formula: z-score$_{\text{sample1\_gene1}}$ = (RPKM$_{\text{sample1\_gene1}}$ - mean RPKM$_{\text{gene1}}$)/SD$_{\text{gene1}}$. The mRNA-Seq heatmap was created using heatmapper (http://www.heatmapper.ca). The DAVID Bioinformatics Resources 6.8 tool was used for functional annotations (https://david.ncifcrf.gov/).

4.17 RT-qPCR

qPCR primers were designed using either Roche's UPL Assay Design Center (UBR2, TRERF1, PTPRJ, SLC6A4, RXRA and TCL1A assays) or the Primer 3 Plus software (DPYD; ACTB; lncRNAs: MALAT1, GAS5, TUG1; eRNAs: eIRF4_-1.9kb, eSPI1_-16kb and eMYC_-170kb), and were analyzed using the OligoAnalyzer Tool (Integrated DNA Technologies). GAPDH primers were derived from Sigma-Aldrich. Enhancer-associated RNAs were designed based on in-house LCL H3K27ac ChIP-seq and mRNA-seq data, as well as public polII ChIA-PET (GSM1872887) and GRO-cap (global nuclear run-on sequencing capturing transcription start sites; GSM1480323) data from GM12878 cells. Primer sequences can be found in Table 3.

We treated total RNA samples with RQ1 DNase according to the manufacturer's recommendations (Promega, cat. M6101), and were reversely transcribed using the SuperScript II system (Thermo Fisher Scientific, cat. 18064014). Each reverse transcription (RT) reaction contained 1x FS buffer, 10 mM DTT, 0.5 mM of dNTPs, 0.8 units of the SSII enzyme, and either 0.012 µg random hexamer primers (mRNAs) or 100 nM gene-specific RT primers (lncRNAs and eRNAs) or 0.4 µg oligo-p(dT)15 primers (GAPDH). The thermal profiles were as follows: for mRNAs, 25°C for 10 minutes, 42°C for 50 minutes, 70°C for 15 minutes; for lncRNAs and eRNAs, 42°C for 50 minutes, 70°C for 15 minutes; for GAPDH, 42°C for 2 hours and 70°C for 15 minutes. Reverse transcription negative control reactions lacking the SSII enzyme were prepared for each sample. RT reactions were diluted five-fold with nuclease-free water prior to qPCR. For the assessment of RT inhibition by EGCG, we prepared 20-µl RT reactions containing the above components for mRNA reverse transcription, as well as 2 µl of 10-fold concentrated EGCG stock to obtain final concentrations between 10^6 and 10^7 M.

We amplified target regions using the LightCycler 480 SYBR Green I Master (Roche Applied Science, cat. 04887352001) with 0.375 µM of each primer. The qPCR reactions were prepared in triplicates. The cycling conditions were as follows: 95°C for 10 minutes, 50 cycles of 95°C for 10 s and 60°C for 30 s (mRNAs, lncRNAs and eRNAs, or 95°C for 10 minutes, 50 cycles of 95°C for 5 s, 55°C for 15 s and 72°C for 10 s (GAPDH). Expression levels were quantified using the $2^{-\Delta Cp}$ method (normalized to ACTB).

Table 3 | Primer sequences used for (RT)-qPCR in our studies.

Assay	Forward primer (5'→3')	Reverse primer (5'→3')
DPYD	CGTGTCAGAAGAGCTGTCCA	GGTCCCTCTTCAGTGGCATA
ACTB	CCCTGGCACCCAGCAC	GCCGATCCACACGGAGTAC
GAPDH	AGTCCCTGCCACACTCAG	ACTTTATTGATGGTACATGACAAGG
TCL1A	GGGAGGAATGGACAGACAGA	AGTGGGTGTGCAACATGAAA
RXRA	CCAGTACTGCCGCTACCAG	CATTCTCGTTCCGGTCCTT
TRERF1	AGGGTGAACCTCAGGAGACC	CAGGATTGCCAGTGACCAG
UBR2	CCTTCCTCTTTCCCTCCATC	CCAGGAGGCAGAGGTTGTAG
PTPRJ	GTCCTGTCCTAGGTGACATCG	GGAAGTCAGAAACTGGAACAGG
SLC6A4	CATTCTCGTTCCGGTCCTT	GTTGGCTATCGCTTCTGCAT
MALAT1	AAAAAGCTACTAAAAGGACTGGTGTAA	*TCCAAATTCTTCTAACTCTTCCAAA*
GAS5	*GCCATGAGACTCCATCAGGC*	CCTCACCCAAGCTAGAGTGC
TUG1	CTGACGAAGACACCCATTCC	*GTGGAGGTAAAGGCCACATC*
eIRF4_-1.9kb	*TGGCAAATGAGTAAACCAGAAG*	TCAACATACCCTCCCCTCAC
eSPI1_-16kb	CTCTGGGCAGGGTCACAG	*GGGCGCTTCCTGTTTTCT*
eMYC_-170kb	*ACTCCAAAGTTCAAGCCCTCT*	GCACACCCGCTGTAACATT

Italicized oligos were used for gene-specific priming.

4.18 5-fluorouracil treatment and MTT assay

The optimal seeding number for 5-fluorouracil (5-FU) treatment was determined to be 20,000 cells per one well of a 96-well plate (for 100 µl final volume), using untreated sGT_1 cells. For the actual assay, sGT_1 or sGT_2 cells were grown to log-growth phase and were washed with indicator-free RPMI (Sigma-Aldrich, cat. R7509) containing 15 v/v% heat-inactivated FCS, 2 mM L-glutamine and 1 v/v% penicillin-streptomycin. 200,000 cells were seeded in quadruplicates per treatment type in wells of a 96 U-well plate (Sigma-Aldrich, M2186). The 5-FU stock solution (TEVA Pharmaceutical Industries; registration number given by the Hungarian National Institute of Pharmacy and Nutrition: OGYI-T-4272/07) was diluted using sterile ultrapure water, and we used two-fold serial dilution in indicator-free RPMI to prepare two-fold concentrated working solutions. 50 µl of 5-FU working solutions were added to the designated wells in the indicated concentrations. Ultrapure water was added to designated wells as a vehicle, and medium-only wells were used as background. Cells were incubated at 37°C and 5% CO_2 for 72 hours. Ten microliters of the MTT stock solution (4.5 mg/ml in PBS) (Sigma-Aldrich, cat. M5655) was added to the wells, and the plate was incubated at 37°C for 6 hours tightly sealed and wrapped in a non-transparent foil. Well, contents were resuspended with 100 µl Lysis Solution (20 w/v% SDS, 20 mM HCl in PBS). The plate was resealed and wrapped, followed by incubation for 1 hour. A VICTOR3 Multilabel Plate Reader (PerkinElmer, MA, USA) was used to measure absorbances at 595 nm. Data were normalized to medium-only wells and was visualized as mean OD_{595} values of the four wells per treatment (+/- SD) divided by the OD_{595} values of vehicle-treated wells.

4.19 Data visualization

We used GraphPad Prism version 6.01 and 7.04 (La Jolla California, USA; www.graphpad.com) or Microsoft Professional Plus 2013 (Excel) (Microsoft, Redmond, Washington, USA) for most visualizations and statistical analyses. All other visualization tools used during the studies were clearly indicated in the relevant Materials and Methods section.

4.20 Data availability

RNA-Seq and ChIP-Seq data related to the studies that form the basis of the present Doctoral Thesis has been made available in the GEO database under accessions GSE106344 (lyophilization study) and GSE121926 (isogenic LCL study).

5 RESULTS

5.1 Development and characterization of phage display-based spike-in procedure controls for ChIP

Here we describe the development and characterization of an AR-mimicking phage control system that may later be used as a spike-in control reagent for AR ChIP experiments, or as an indifferent control in ChIP experiments targeting chromatin-associated proteins other than AR.

5.1.1 Development and characterization of a polyclonal AR-mimicking phage control reagent

We subjected NEB's Ph.D.™-7 random heptapeptide library to four consecutive rounds of affinity selection with a ChIP-grade anti-AR antibody (cat. sc-816; with 309 citing articles thus far). In theory, phage affinity to (i.e. the fraction of input phages captured by) AR antibody increases, while phage mixture complexity decreases with the number of selection rounds applied. In order to get a general view of the enrichment of certain bases at the 21 variable genomic positions across the selection rounds, we performed capillary sequencing for each of the four polyclonal phage reagents generated (Figure 2). We observed that the number of positions dominated by one type of nucleobase increases, and the guanidine dinucleotide background becomes less dominant with increasing round numbers. It should be noted that obtaining an even less genetically complex polyclonal population by biopanning may be restricted by the facts that various peptide sequences can mimic the AR epitope, multiple codons may code for a single amino acid (degenerated genetic code), and phages affine to the plasticware used during selection may also be present. Moreover, drastically decreasing the complexity of phage libraries may result in the overrepresentation of phages which infect more rapidly growing ER2738 cells, which may be unrelated to the anti-AR-affinity.

We next tested whether the generated polyclonal libraries can be efficiently captured in an IP reaction. We set up an immunoprecipitation (IP) experiment largely carried out by an automated (robotic) system, which recapitulated the steps of a ChIP protocol: each reaction contained crosslinked and sonicated chromatin (from HEK293T cells), anti-AR or isotype control antibody, and 10^6 phage particles. The ChIPped DNA was subjected to qPCR with primers complementary with non-variable phage genomic regions. We found that the fraction of phages that could be recovered after the simulated ChIP reaction increased with the number of selection rounds. The isotype control IgG signals, which indicate all non-specific binding (e.g. to plasticware, beads, antibody constant regions, and so forth), remained relatively stable across the four phage batches. Approximately 50% of spiked phages could be recovered using the ‘round 4’ batch (Figure 3).

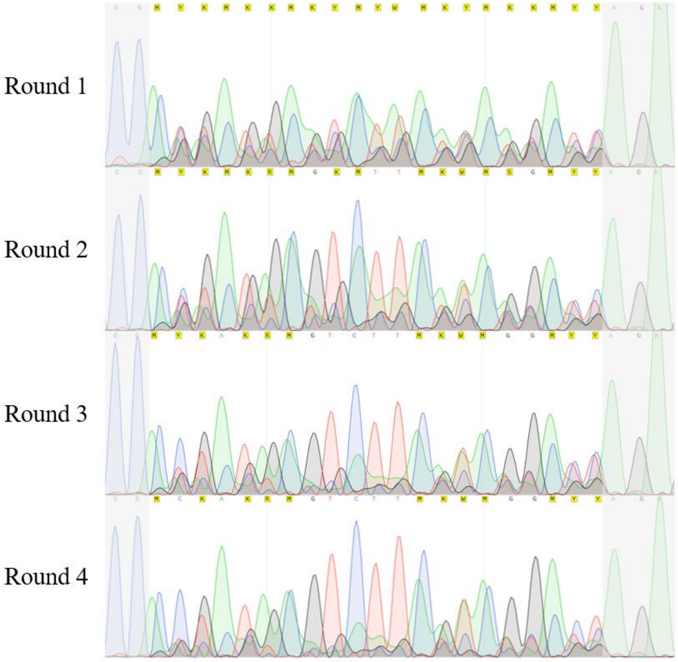


Figure 2 | Capillary sequencing of polyclonal AR phage batches produced using four consecutive rounds of biopanning with a ChIP-grade anti-AR antibody.

5.1.2 Development and characterization of monoclonal AR-mimicking phage controls

Due to the fact that affinity-selected polyclonal phage mixtures may contain phage clones that are specific to components of the selection environment other than the variable region of the antibody, as well as the general phenomenon that during the regeneration of polyclonal batches there may be a shift in clone distribution (due to variable infectivity and ER2738 growth), we decided to select and test individual AR phage clones for anti-AR-affinity in simulated ChIP experiments. We infected ER2738 cells with highly diluted ‘round 4’ phages, picked and grown individual bacterial colonies from IPTG/X-gal plates, and purified phage clones. We selected the clones that had a high enough yield for subsequent experiments. We used the semi-robotic IP-qPCR method described for polyclonal reagents (for details please refer to the Materials and

Methods section) to measure the percentage of monoclonal phages isolated from the elution buffer. Moreover, we purified phages from the IP buffer as well, which contained the phages not bound to the antibody-bead complex during the immunoprecipitation step. Figure 4 shows the ChIP results of seven AR phage clones. The elution buffer:IP buffer ratio was 50%< for all clones, with one clone with exceptionally high (90%<) affinity to the anti-AR antibody. High affinity is preferred for a phage clone to be used as a spike control for real-world ChIP experiments.

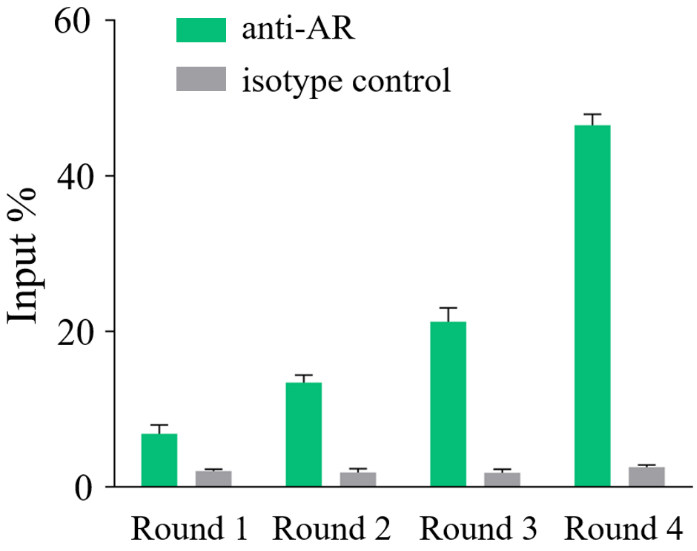


Figure 3 | Assessment of phage recovery from round 1-4 polyclonal phage stocks, using anti-AR robotic IP followed by qPCR.

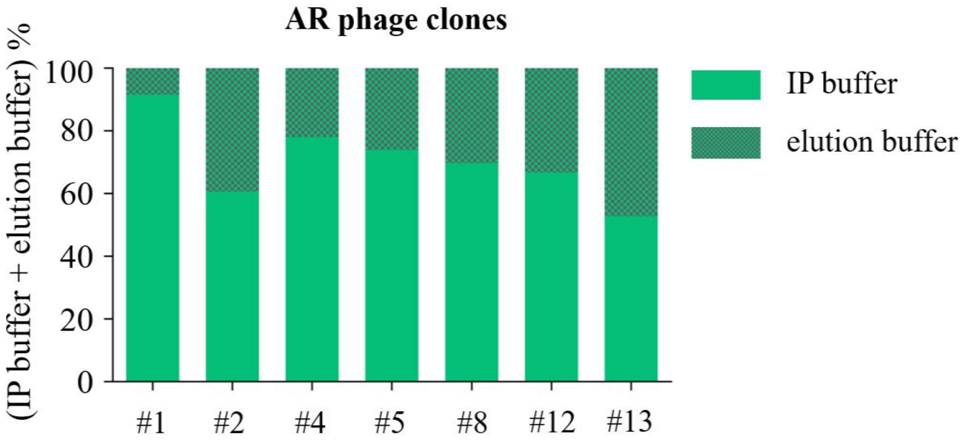


Figure 4 | Assessment of monoclonal AR phage recovery using anti-AR robotic IP followed by qPCR. Phage genomic DNA was quantified from IP buffer and elution buffer.

5.2 Lyophilization as a means to preserve cellular RNA in human cells

We used low- and high-throughput methods to investigate the effect of whole-cell lyophilization *per se*, as well as weeks of room temperature storage on the quality and quantity of extracted RNA molecules.

5.2.1 RT-qPCR and ChIP-qPCR experiments with LCL cells lyophilized in IMT-2

Based on Arav *et al.*'s and Natan *et al.*'s success with mammalian cell lyophilization and recovery of live cells (Arav & Natan, 2012; Natan et al., 2009), we decided to apply IMT-2 solution as lyoprotectant in our pilot experiments. IMT-2 contains 0.1 M trehalose, and 0.945 mg/ml (-)-epigallocatechin gallate (EGCG) in PBS, as both components have been shown to have various properties making them good candidates as cell lyoprotectants, which may facilitate multiple downstream applications including ones requiring intact cells, such as ChIP.

Three batches of human cells were lyophilized in 0.5 ml IMT-2 and were rehydrated immediately after the lyophilization cycle with 0.5 ml room-temperature ultrapure water. We examined recovered cells under a standard inverse light microscope and found that although cell size and morphology were identical to fresh cells, trypan blue penetrated all lyophilized cells during 5 minutes of incubation. This suggests the presence of membrane discontinuities enabling the rapid penetration of the dye. In line with that, there was no sign of cell proliferation for up to one week of culturing under standard cell culture conditions. Although the used lyophilization method did not provide sufficient protection against membrane damage, our primary aim was to extract intact biomolecules from lyophilized cells. Therefore we decided to assess RNA quality and ChIP DNA complexity from LCL cells lyophilized in IMT-2.

We next lyophilized three LCL batches on three different days in IMT-2, and resuspended lyophilized powders in 1 ml trizol. Total RNA was isolated, and as a first quality control step, sample absorbance spectra were taken. Surprisingly, we found that lyophilized samples had lower $OD_{260/230}$ ratios (P value = 0.03, paired t-test), suggestive of sample contamination with substances with high absorbance around 230 nm. Also, during RNA isolation, we observed that RNA pellets from lyophilized samples, even after ethanol washes, had brownish-grey colour. Running Agilent chip-based electrophoresis to examine fragment distributions, we found relatively high RIN values for lyophilized samples (mean = 7.9), although it was lower than those for paired controls (P value = 0.03, paired t-test). Although successful RT-qPCR experiments were expected based on RINs, lyophilized samples showed significantly elevated Cp values for a highly expressed housekeeping gene, GAPDH, compared to controls (Figure 5A), which could be reversed by diluting total RNA samples 100-fold prior to reverse transcription (RT) (Figure 5B). This suggests that the RT-qPCR reaction was inhibited by an experimental reagent, possibly EGCG. To prove the inhibitory effect of EGCG, we spiked RT reactions with EGCG at different final concentrations, and demonstrated that the presence of this polyphenol (even at concentrations comparable to those achievable in the blood plasma after green tea consumption (C. S. Yang et al., 1998)), inhibits RT-qPCR in a concentration-dependent manner (Figure 5C). We also found that washing the cells prior to trizol isolation did not lead to a change in RT-qPCR-based measurability (Figure 5D). Noteworthy, it has been reported that EGCG inhibits the reverse transcription step of HIV infection (S. Li, Hattori, & Kodama, 2011). Collectively, when EGCG is used as a lyoprotectant, total RNA isolated with trizol is contaminated with EGCG, which interferes with RT-qPCR reactions.

Although lyophilization in IMT-2 was not proven feasible for RT-qPCR-based RNA quantitation, we decided to check whether lyophilized cells can be used for ChIP-based studies. We first compared four sample types for H3K27ac ChIP efficiency: control cells (not lyophilized), cells lyophilized in IMT-2 after fixation in 1% formaldehyde (FA), cells lyophilized in IMT-2, and cells preloaded with 0.1 M trehalose overnight before IMT-2 lyophilization. Of note, ChIP DNA isolated from cells lyophilized without prior fixation showed a mild brownish discolouration, similarly to RNA pellets (see the previous paragraph). We observed high ChIP efficiency for cells prefixed with 1% formaldehyde prior to IMT-2 lyophilization. However, cells lyophilized without prior fixation showed lower IP efficiency (P value < 0.01) (Figure 6A). All input (unprecipitated chromatin) measurements resulted in similar Cp values, suggesting that qPCR inhibition is less likely the cause for the above phenomenon. Nevertheless, we prepared 10x and 100x dilutions from the IMT-2-lyophilized, H3K27ac-immunoprecipitated samples and measured positive and negative control regions by qPCR. Not surprisingly, relative expression values normalized to the undiluted sample were highly similar for the different dilutions (Figure 6B). As we observed the leakiness of the cell membranes after lyophilization, we hypothesized that FA fixation taken place after lyophilization might overfix cellular structures, leading to less available chromatin epitopes. Therefore we included one more sample type which was first lyophilized in IMT-2 and were subsequently fixed with 0.1% FA instead of the standard 1%. We found that 0.1% FA fixation or no fixation at all did not improve ChIP efficiency (Figure 6C). We next decided to perform micrococcal nuclease (MNase) digestion of our chromatins. MNase is an exo-endonuclease that digests internucleosomal regions, resulting in characteristic electrophoretic patterns; therefore MNase treatment may be useful to assess the intactness of the beads-on-a-string-level chromatin structure which, in theory, should enable H3K27ac ChIP experiments. Besides control and FA-prefixed, IMT-2-lyophilized cells, chromatin from non-prefixed cells also show the definitive nucleosomal pattern, indicating that lyophilization did not lead to the displacement of histones from genomic DNA (Figure 6D). Our next hypothesis was that the presence of EGCG inhibits the ChIP reaction, regardless of lyophilization. We treated LCL cells for 1 hour with 2 mM EGCG in duplicates and performed H3K27ac ChIP-qPCR. We found that the addition of EGCG resulted in lower IP efficiency of the positive control region (Figure 6E). This might be the result of either biological response to EGCG, or the interference of this substance with the ChIP experiment. The former explanation is supported by evidence on the downregulation of *SPI1*, the promoter region of which was used in our experiments as a positive control region, by EGCG in differentiating naïve CD4+ T helper cells (J. Wang, Pae, Meydani, & Wu, 2013). As it has been recognized that EGCG affects gene regulation (H.-S. Kim, Quon, & Kim, 2014), and it seemingly penetrate rapidly into the cells (discolouration of precipitated nucleic acids in ~ 5 minutes), we concluded that EGCG should not be used as a cellular lyoprotectant when the downstream applications involve RNA- or chromatin-based studies.

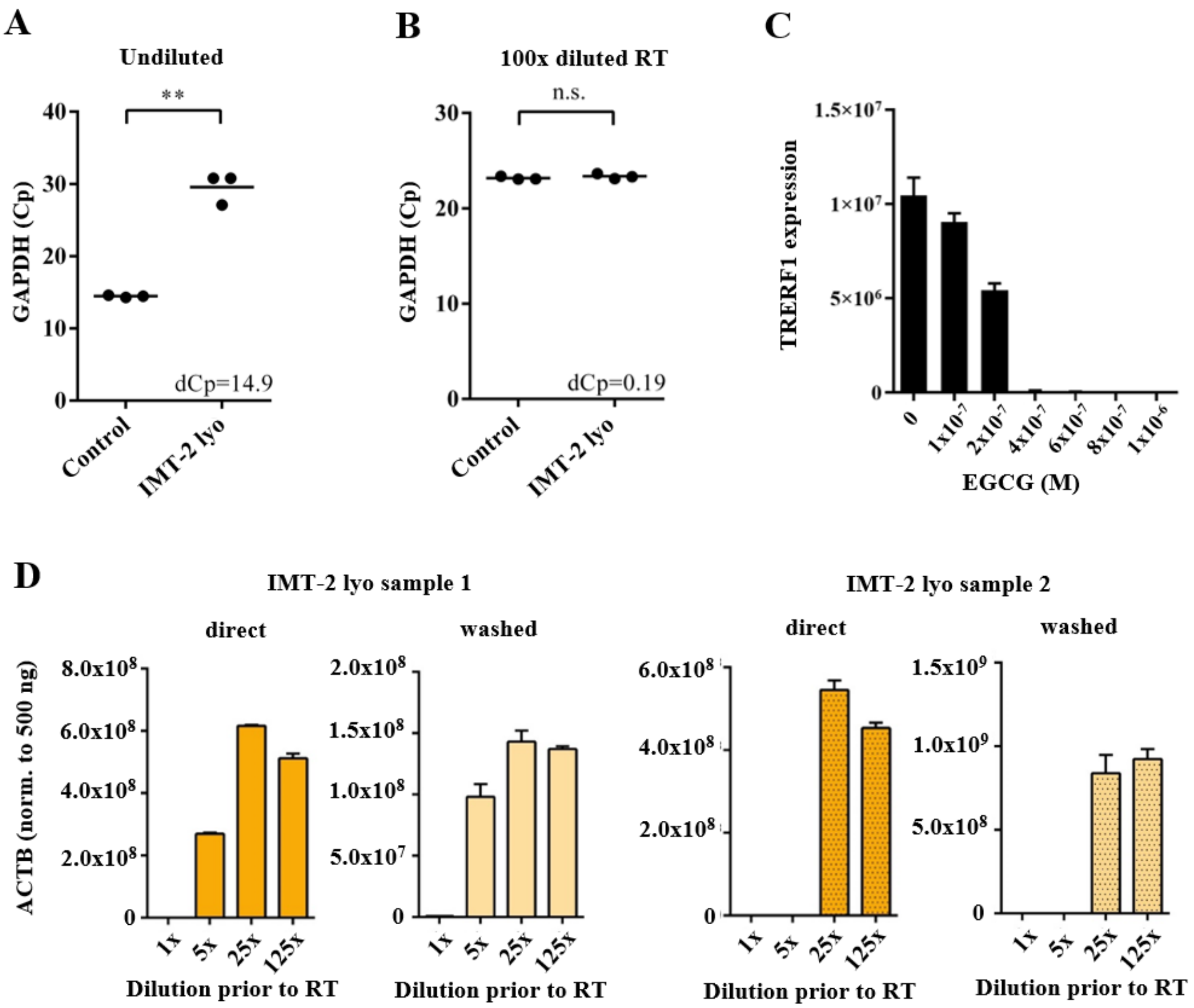


Figure 5 | The effect of (-)-epigallocatechin-gallate as a cell lyoprotectant on RT-qPCR of isolated total RNA. Cp values of the GAPDH mRNA in paired control cells and cells lyophilized in IMT-2 as measured using the same amount of total RNA for reverse transcription, using **A)** undiluted total RNA and **B)** diluted total RNA. Vertical lines represent the means. The Cp value differences are indicated over the y-axis (paired t-test, **P value < 0.01). **C)** RT-qPCR of pure, highly intact (RIN = 10) total RNA spiked with different concentrations of epigallocatechin-gallate. Error bars represent SD values of three replicate qPCR measurements. **D)** RT-qPCR measurements of total RNA isolated from IMT-2-lyophilized and directly isolated, and paired IMT-2-lyophilized and washed cells. Five-fold dilution series of total RNA samples were used for reverse transcription reactions, and presented qPCR values were normalized to 500 ng input RNA. Error bars represent SD values of triplicate qPCR measurements. RT = reverse transcription; EGCG = epigallocatechin-gallate; n.s. = not significant.

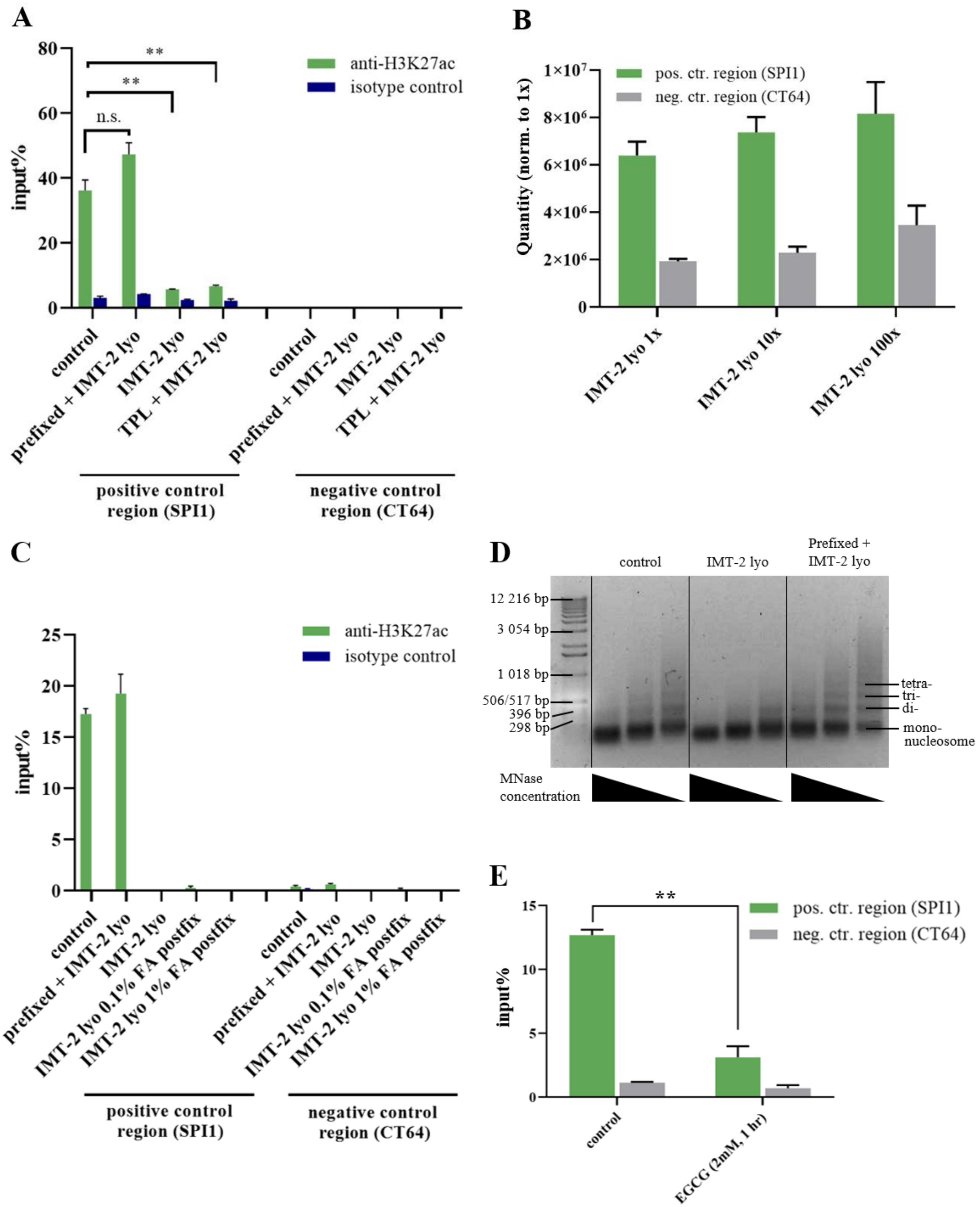


Figure 6. *Figure caption can be found on page 54.*

Figure 6 *(page 53)* **| The effect of epigallocatechin gallate (EGCG) on chromatin immunoprecipitation efficiency. A)** H3K27ac ChIP-qPCR from LCL cells under the following conditions: fresh (control), fixed in 1% FA and lyophilized in IMT-2, lyophilized in IMT-2, and trehalose-preloaded and lyophilized in IMT-2, using sonication as the fragmentation method. ChIP experiments were performed in two biological replicates, positive and negative control genomic regions from IP samples were normalized to chromatin input (+/-SD). **B)** QPCR measurements of diluted H3K27ac ChIP samples, normalized to undiluted sample (three replicate qPCR measurements). Positive and negative control genomic regions are indicated. **C)** H3K27ac ChIP-qPCR from LCL cells under the following conditions: fresh (control), fixed in 1% FA and lyophilized in IMT-2, lyophilized in IMT-2, lyophilized in IMT-2 and fixed with 0.1% FA, lyophilized in IMT-2 and fixed with 1% FA, using MNase treatment as the fragmentation method. Bar plots represent means of triplicate qPCR measurements. Positive and negative control genomic regions from IP samples were normalized to chromatin input. **D)** DNA fragment distribution of MNase-treated (66.6 Gel Units, 22.2 Gel Units, and 7.4 Gel Units MNase/$1.5*10^6$ cells) chromatin isolated from the following cells: control, lyophilized in IMT-2 prefixed and lyophilized in IMT-2. **E)** H3K27ac ChIP-qPCR on positive and negative control regions from control cells and cells treated for 1 hr with 2 mM EGCG (N = 2, +/-SD). Where biological duplicates were used, we applied Student's t-test. **P value < 0.01. TPL = trehalose-preloaded; FA = formaldehyde.

5.2.2 Cellular RNA quality and quantity remains stable during lyophilization and weeks of room temperature storage as measured by standard methods

First, we decided to examine whether the lyophilization cycle itself affects RNA quality and quantity, using low-throughput methods. LCL cells were lyophilized in 0.5 ml 0.1 M trehalose/PBS for 6 hours in a state-of-the-art manifold freeze dryer (for details, please refer to the Materials and methods section). RNA was isolated immediately and was quantified using fluorometry followed by fragment distribution-based quality assessment by a capillary electrophoresis-based method (Agilent microchip). RNA from lyophilized cells showed no sign of degradation, with similar yields and RIN values compared to RNA isolated from paired fresh cells (N = 6) (Figure 7A-D). RIN (RNA integrity number) values calculated from ribosomal peaks of electropherograms were remarkably high (mean = 9.8), and were not significantly different from paired fresh cells (P values > 0.01). RIN values are a good indicator of overall RNA quality and are therefore routinely used before sensitive high-throughput methods (microarrays and RNA-Seq). The RIN cutoff used in such cases is somehow arbitrary, but samples with RIN > 7 are generally considered eligible for use in both RT-qPCR and high-throughput applications. Notably, however, it has been argued that the integrity of the mRNA fraction, commonly assayed in research, may not be reliably reflected by the RIN, probably due to structural differences between rRNAs and mRNAs (Fleige & Pfaffl, 2006; Mayne, Shepel, & Geiger, 1999; Vermeulen et al., 2011).

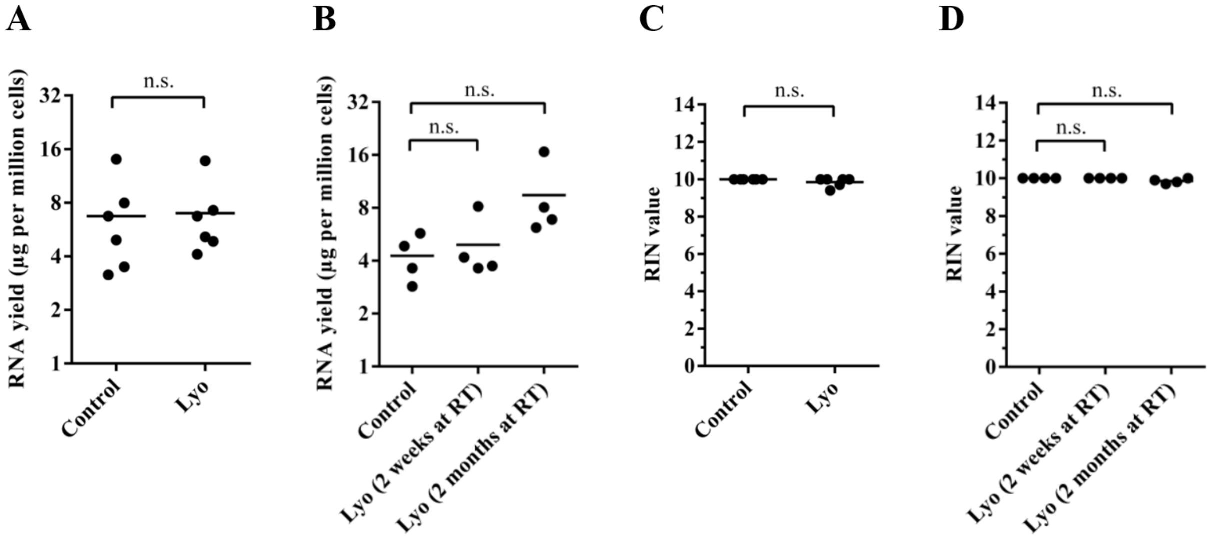


Figure 7 | Quality and quantity of total RNA samples isolated from lyophilized LCLs right after, or two or eight weeks of room temperature storage after lyophilization. Yields of RNA samples isolated **A)** right after lyophilization and **B)** after two or eight weeks of room temperature storage; RIN values of RNA samples isolated **A)** right after lyophilization and **B)** after two or eight weeks of room temperature storage. P values were over 0.01 in all cases (n.s., not significant). RT = room temperature.

We next assessed whether the quantitation of variably abundant mRNAs, long non-coding RNAs (lncRNAs) and the recently described class of enhancer-associated RNAs (eRNAs) shows differences between fresh and lyophilized cells. We chose six genes which span the expression levels of extremely low (RPKM < 1), low (RPKM = 1-10), moderate (RPKM = 10-100), and high (RPKM > 100), based on RNA-Seq data in two LCLs (GM12872 and GM12873). Of note, 65% of detected genes fall into the low or extremely low categories (Figure 8A). We found that all measured genes were amplified to a similar degree in control vs lyophilized cells, regardless of their abundance (Figure 8B). We also measured representative genes from two other RNA classes. Namely, lncRNAs, which have been described as potential disease biomarkers (Fu et al., 2016; Prensner et al., 2014; M.-H. Yang et al., 2015) and eRNAs, potent indicators of enhancer-associated pathological conditions (Jiao et al., 2018; P. Li et al., 2019). Three lncRNAs with clinical relevance (MALAT1, GAS5 and TUG1), as well as three eRNAs at super-enhancers of highly expressed LCL genes (IRF4, SPI1, and MYC) were selected for RT-qPCR. Expression levels between paired fresh and lyophilized GM12873 cells were found to be highly concordant (Figure 9A-B). Collectively, these results suggest that lyophilized cellular samples enable accurate gene expression quantitation by RT-qPCR.

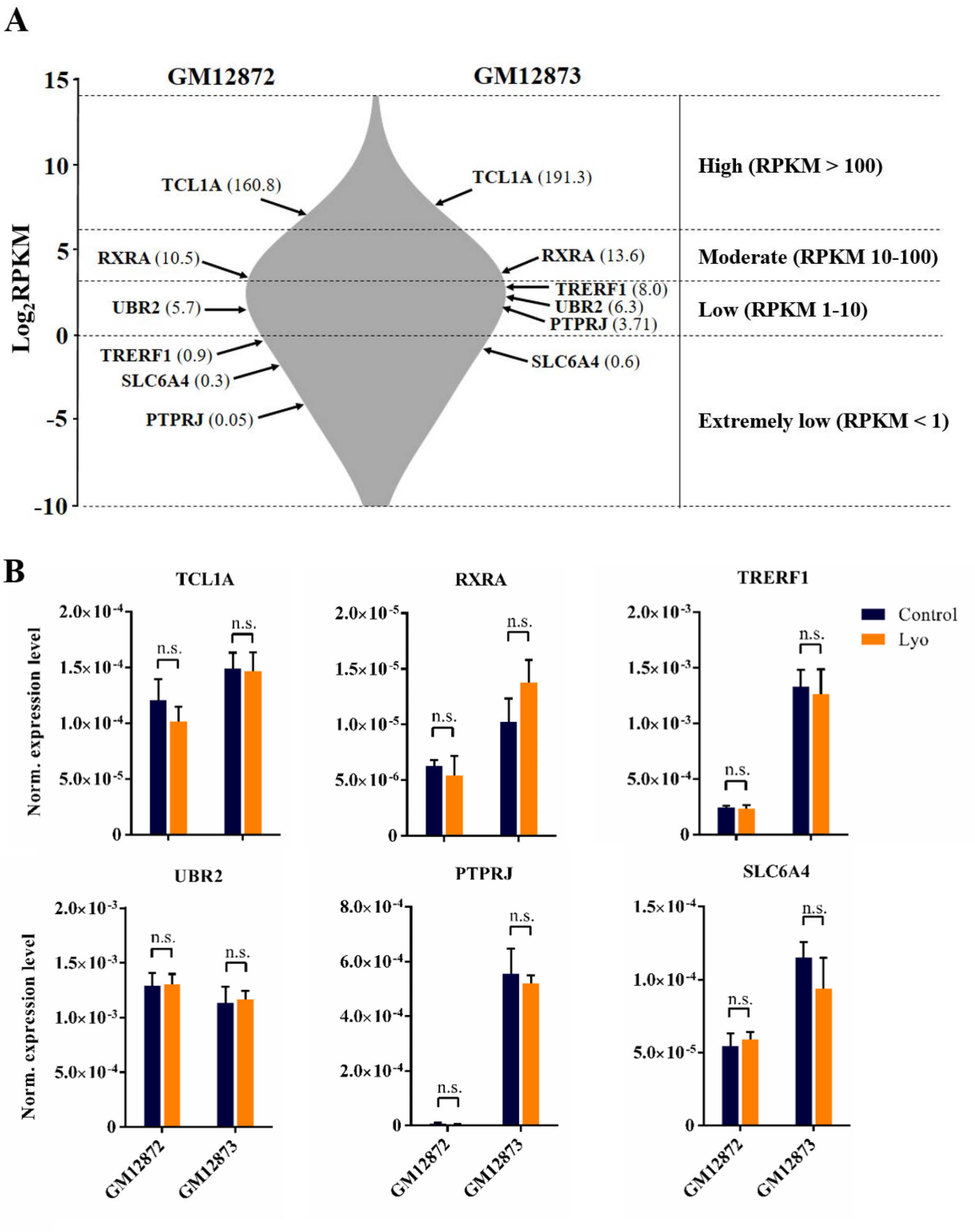


Figure 8 | RT-qPCR measurement of mRNAs in paired control and lyophilized cells. A) Violin plot of log$_2$RPKM values of genes selected for qPCR quantitation based on previous RNA-Seq data of fresh LCL cells (GM12872, GM12873). The numbers in brackets represent RPKM values. **B)** ACTB-normalized RT-qPCR expression values per gene, for each cell line (P values > 0.01; n.s. = not significant).

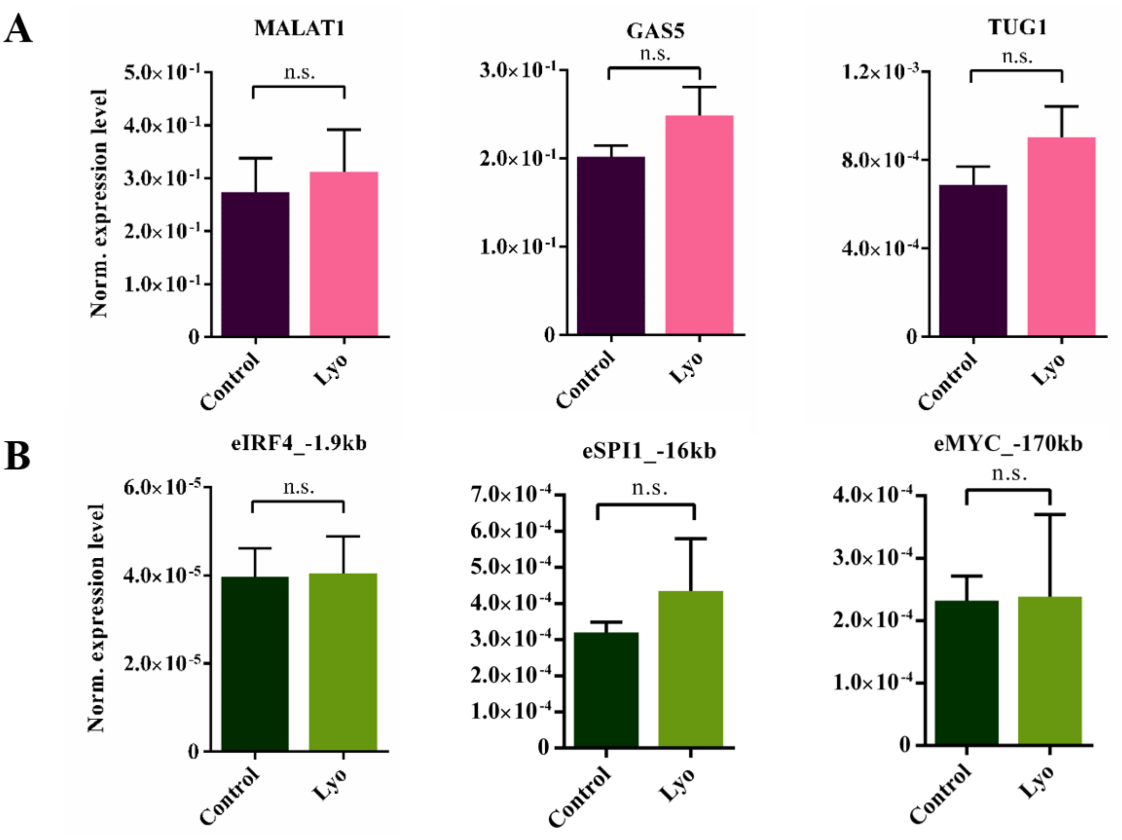


Figure 9 | RT-qPCR measurement of lncRNAs and eRNAs in paired control and lyophilized cells. A) ACTB-normalized expression of three lncRNAs **and B)** three eRNAs in GM12873 cells between paired control and lyophilized cells (P values > 0.01; n.s., not significant).

5.2.3 RNA-Seq reveals a highly similar transcriptome profile between control and lyophilized cells

Three total RNA samples isolated from lyophilizates stored for two weeks at room temperature, together with their paired control samples, were selected for transcriptome-wide analysis by RNA-Seq. Libraries were prepared using poly(A) selection and were sequenced using Illumina's NextSeq500 platform to read numbers over 1.5*10^7 (Table 1). Overall library qualities can be inferred by calculating standard quality metrics, including the percentage of uniquely mapping reads and duplicated reads. These metrics may vary from experiment to experiment, affected by various factors such as cell type, library type and sequencing conditions. These quality metrics, therefore, should be similar within an experiment. Table 4 summarizes basic metrics, as well as raw read numbers and the number of predicted genes for each sample. The fraction of uniquely mapping reads were over 90% for all samples, with -2.6% to 0.4% deviation from the median. Reads with identical starting positions occur in libraries predominantly due to sampling and library fragmentation bias. However, as these may also represent PCR duplicates or signal sequencing artifacts in case of low quality or quantity samples, duplication rate is commonly assessed before RNA-Seq analysis (Parekh, Ziegenhain, Vieth, Enard, & Hellmann, 2016). Duplication rates ranged from -13.6% to 6.7% compared to the median. Based on the best practices described by Conesa *et al.*, quality metrics should show less than 30% disagreement so as not to consider a sample an outlier (Conesa et al., 2016); thus, all of our samples were included in further analyses. The above observations suggest that libraries of lyophilized samples were of high quality, showed sufficiently high complexity and that none of the experimental steps, including lyophilization and room temperature storage, induced base modifications affecting read mappability, which contrasts with findings in formalin-fixed and paraffin-embedded samples where mapping quality decreased due to formalin-induced base changes (Esteve-Codina et al., 2017; Graw et al., 2015).

Table 4 | Per sample RNA-Seq library information.

Sample pair	**Sample name**	**# raw reads**	**% deviation of UMRs from the median**	**% deviation of DR from the median**	**# expressed genes**
Pair 1	**Control 1**	15 173 243	-2.6	+7.5	11 017
	Lyo 1	27 984 754	0.1	-12.9	10 918
Pair 2	**Control 2**	20 352 186	-0.1	+5.5	10 979
	Lyo 2	27 149 029	0.1	-6.8	10 881
Pair 3	**Control 3**	24 497 577	-0.2	+0.8	11 035
	Lyo 3	22 484 771	0.4	-0.8	10 970

UMRs = uniquely mapping reads; DR = duplication rate.

In order to obtain a deeper insight into library quality, we calculated read GC content, chromosomal distribution, biotype distribution, gene body coverage, cumulative gene diversity and per base mismatch rate (Figure 10). It has been described that GC bias in RNA sequencing reads, which tend to be sequencing lane-specific, may affect gene expression measures and other downstream analyses (Risso, Schwartz, Sherlock, & Dudoit, 2011). Therefore we calculated the fraction of purine bases for each sequencing read and mapped the percentage of reads with certain GC% values. The resulting GC plots showed that most reads fell to the 37-39% region of an approximately normally distributed data, with no GC shifts or peaks observed at the tail regions (Figure 10A). The majority of reads mapped to the autosomes, and although slight differences occurred between samples, differences were consistent within cell batches (control-lyo pair), suggesting that cell states at harvest may have been responsible for the observed differences (Figure 10B). Most reads mapped to protein-coding genes in all samples, and there was no statistically significant difference between control and lyophilized samples (Figure 10C). Differences between input RNA sample integrity may be reflected by gene body coverage of genes. Figure 10D suggests that there was no pronounced mapping bias towards each end of genes at the upper middle quartile read count range, indicating that there was no 5’ or 3’ bias suggestive of degradation or strand cleavage. Figure 10E shows the fraction of reads mapping to the top 10, 100, 1 000, and 10,000 genes, denoted as „cumulative gene diversity”. This metric is aimed at showing the domination of reads mapping to a few highly expressed genes, characteristic to inferior library complexity. The plots show highly similar library complexities (19-22% falling to the top 100 genes; 50-53% falling to the top 1000 genes). The first pair shows a slight left shift, which may reflect a phenomenon of biological origin rather than experimental inconsistency. We next assessed whether lyophilization and/or room temperature storage lead to base changes,   leading to increased mapping mismatch rate. Of note, mismatches may represent natural variation compared to the reference genome in use, as well as results of various physical and chemical exposures. In formalin-fixed and paraffin.embedded samples, for example, G>A and C>T transitions occur relatively frequently. We found no statistical difference across the read length, as represented by Figure 10F (as exemplified by the C>G mismatch ratio, compared to the hg19 reference genome).

After assessing overall library quality for each sample, we performed differential gene expression analysis. We found very high correlation between control and lyophilized datasets ($r^2 = 0.99$), and identified 28 genes significantly downsampled in lyophilized samples (differentially expressed genes, DEGs; FDR = 0.05) (Figure 11A). The fold-difference between controls and lyophilized samples ranged from 1.94 to 4.25 (median = 2.31), with lowly expressed genes showing higher fold-difference (Figure 11B). Among the DEGs were 21 protein-coding genes (PCGs), six lncRNAs and one pseudogene. Gene ontology analysis uncovered the enrichment of the term 'DNA-templated transcription' (GO:0006351; P value = $2.0*10^{-5}$; child term: 'Transcription by RNA polymerase II', GO:0006366, P value = $2.2*10^{-5}$). Such transcriptional regulators included POLR2A, CIC, INTS1, KDM6B and KMT2D. This finding is consistent with previous studies describing higher degradation rates of transcriptional regulators under standard culturing conditions and after room temperature storage of cells in liquid media; short half-lives of transcription factors in cells have been shown in both physiological and non-physiological conditions (Baechler et al., 2004; Rabani et al., 2011; Schwanhäusser et al., 2011; Sharova et al., 2009; E. Yang et al., 2003). Our result suggests the presence of some residual cellular decay activity in dried cells, affecting genes with special transcript features.

Figure 10 *(page 61)* **| RNA-Seq quality metrics for paired control and lyophilized samples. A)** The fraction of reads with certain GC%. **B)** Chromosome distribution of sequencing reads (GM12873 cells are derived from a female donor). **C)** The fraction of reads mapping to certain RNA classes reads (P values > 0.01, all comparisons; paired t-test). **D)** Gene body coverage profile of genes in the upper middle quartile expression range; exonic regions had been previously divided in percentiles (bins of 2.5% coding length) (P values > 0.01, all comparisons; paired t-test). **E)** Cumulative gene diversity as assessed by plotting the fraction of reads mapping to the top 10, 100, 1 000, and 10 000 genes (by read counts) (P values > 0.01, all comparisons; paired t-test). **F)** A representative mismatch profile (C>G mismatch rate), plotting mean +/- SEM; corrected for multiple testing, P values > 0.01).

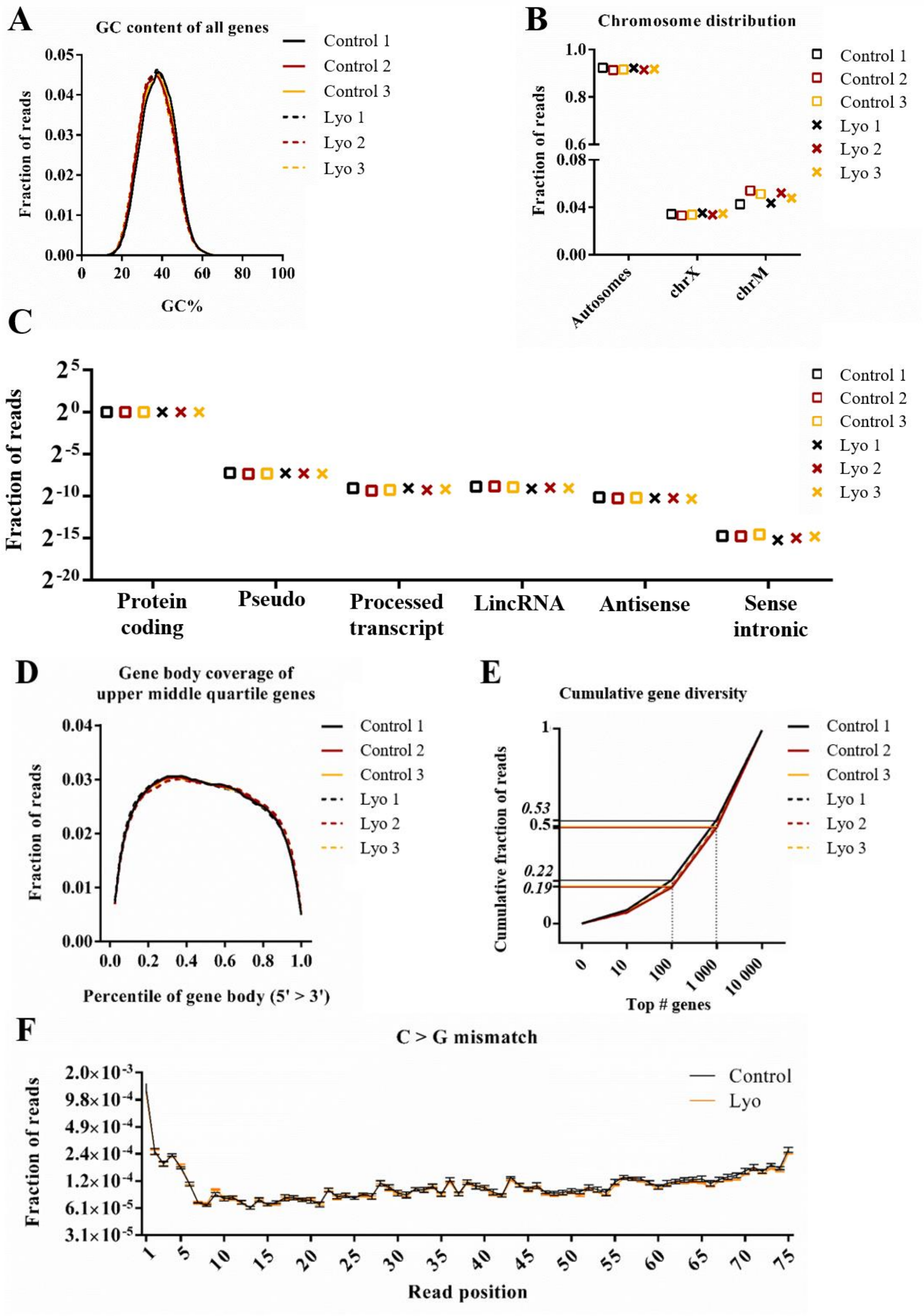


Figure 10. *Figure caption can be found on page 60.*

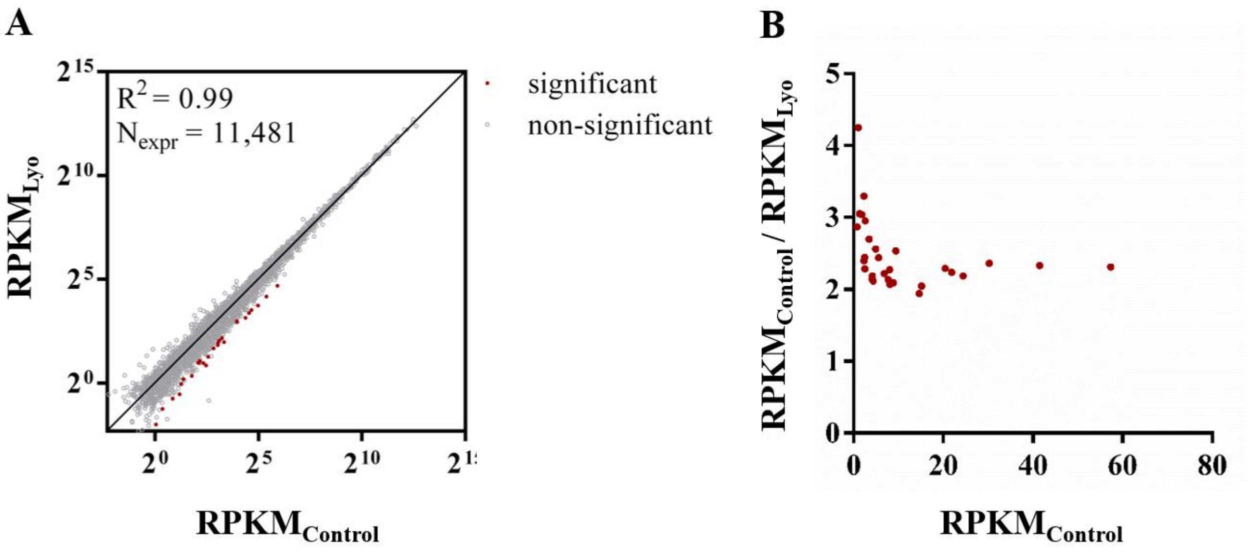


Figure 11 | Characteristics of genes downsampled in lyophilized samples. A) Mean RPKM values from lyophilized samples plotted against those of the controls (RPKM>1 in at least one sample). **B)** Fold distribution of DEGs plotted against RPKM values of controls (mean values of three replicates).

5.2.4 Characteristics of downsampled RNAs

Next, we decided to uncover the distinctive features of DEGs. First, we analyzed the read coverage of DEGs across the gene length in control any lyophilized samples, using meta-transcripts (40 bins, containing only exonic regions). Such analysis would shed light on 3' mapping bias, which may result from the wash-off of the 5' ends of cleaved RNA strands during poly(A)-selection. We calculated read coverage ratios for each bin per gene, and we found that most DEGs did not show 3' bias (P value > 0.01; N = 16). However, 8 DEGs showed significant positive, and 1 DEG (the lowly expressed *LINC01374*) showed a significant negative correlation between 3' distance and read count ratio (P value < 0.01) (Figure 12A). Figure 12B shows mapped read counts and read count ratios across the length of the *AGRN* gene, as an example. Overall, although most DEGs did not show different read distributions between control and lyophilized samples (i.e. uniform reduction across the transcript), for some transcripts, strand cleavage might have contributed to the observed decreased expression.

As a next step, we focused on transcript properties that may have been responsible for accelerated DEG decay in lyophilized samples. Downsampled lncRNA transcripts and protein-coding RNAs 5' untranslated region (5'UTR) + coding sequence (CDS) + 3' untranslated region (3'UTR) were shown to have significantly higher transcript length and GC fraction than all corresponding human transcripts (P value < 0.0001, Mann-Whitney test) (Figure 12C-D). We also found that the length of the CDS, as well as the GC fraction of the CDS and the 5'UTR and 3'UTR, were significantly higher for DEGs (P value < 0.001; Mann-Whitney test) (Figure 12E-G).

The presence of 3'UTR AU-rich elements (AREs) had been associated with a shorter half-life as a result of ARE-binding protein-mediated poly(A) degradation (C. Y. Chen & Shyu, 1995). Twelve protein-coding DEGs (57%) were listed as containing at least one 3'UTR or intronic ARE in the ARED-Plus database (Bakheet, Hitti, & Khabar, 2018). Of note, the DEGs with downsampled 5' transcript end and ARE-containing DEGs showed only a modest overlap (2 DEGs with both) (Figure 12H), suggesting that either fragmentation or ARE-mediated decay might be responsible for the downsampling of differentially expressed protein-coding transcripts.

Figure 12 *(page 64)* **| Transcript properties of DEGs. A)** Read count ratios across meta-transcripts (assembled from exons and divided to 40 bins) for the 9 DEGs with end bias (linear regression; P value < 0.01) and 19 DEGs without end bias. **B)** Mean mapped read counts (+/- SEM) for each sample group and read count ratios over the AGRN metatranscript. Box-and-whisker plots of **C)** transcript lengths and GC% of human vs. DE lncRNAs; **D)** cDNA length and cDNA GC%, **E)** CDS length and CDS GC%, **F)** 5'UTR length and 5'UTR GC%, and **G)** 3'UTR length and 3'UTR GC% of all human vs. DE protein-coding transcripts. **H)** Proportional Venn diagram showing the overlap between DE RNAs containing ARE(s) and DEGs with 5' bias. The box-and-whisker plots display medians as horizontal lines and interquartile ranges as boxes, as well as minimum to maximum values as whiskers. DE = differentially expressed; PCG = protein-coding gene; n.s. = not significant.

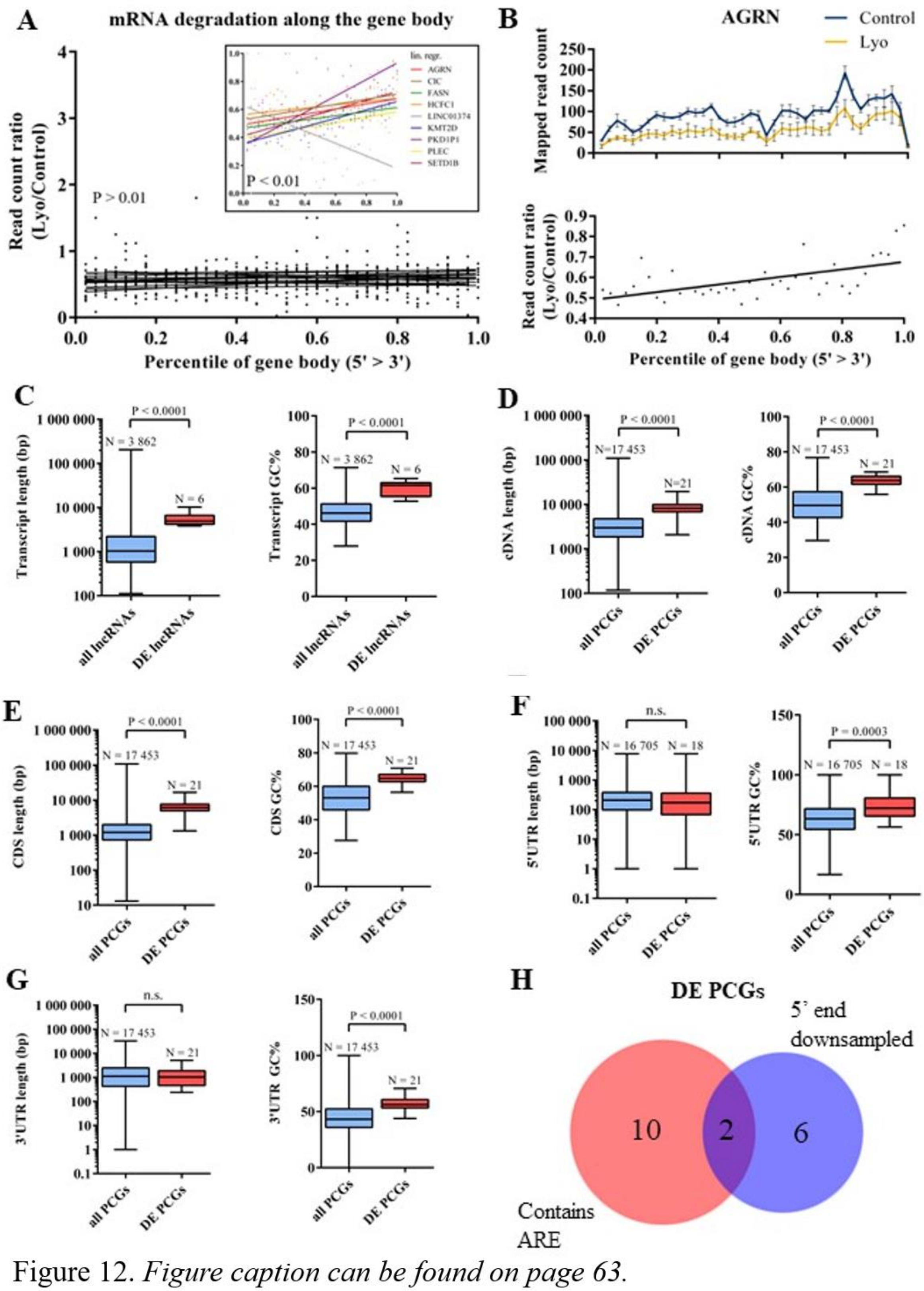


Figure 12. *Figure caption can be found on page 63.*

5.3 Characterization of isogenic LCLs at multiple cellular phenotype levels

5.3.1 Basic characterization of isogenic LCLs

A set of five LCLs were selected from Coriell Cell Repositories, a major source of LCLs for biomedical research, as a model system to characterize the extent and nature of genotype-independent variability among LCLs at the levels of selected cellular phenotypes. The five isogenic LCLs, namely GM22647, GM22648, GM22649, GM22650 and GM22651 (denoted as sGT_1, sGT_2, sGT_3, sGT_4 and sGT_5, respectively, where sGT stands for "the same GenoType"), had been derived from five collection tubes of blood drawn from the same CEPH/UTAH 26-year-old healthy male (Shirley et al., 2012). PBMCs from the five tubes had been independently infected with a common laboratory EBV strain (B95-8), expanded and archived at Coriell. Therefore the procedure of isogenic LCLs preparation was largely equivalent with the one used during LCL preparation from genetically unrelated individuals. A study from the cell lines' source laboratory compared the genotypes of the five LCLs with the genotype of the parental cells (PBMC) and found high concordance at the levels of SNVs and indels. Moreover, they found no evidence for the presence of mosaic regions and chromosomal aberrations (Shirley et al., 2012).

The following actions were taken to exclude the majority of cell line handling-related confounders: 1) isogenic LCLs were shipped together from Coriell as live cultures, screened by the company for bacterial, viral and fungal infections; 2) a three-tiered biobank was created from all cell lines to provide working batches of cells with the same number of freeze-thaw cycles and passages (N = 14) for multiple experiments; 3) the five cell lines were handled in parallel during biobanking and experiments; 4) cell harvesting for epigenomic and transcriptomic profiling were carried out at the same time point of the day (excluding circadian effects); 5) and we used biological replicates to exclude differences due to random fluctuations. Using cell lines prepared from the same individual and excluding cell culturing-related confounders, isogenic LCLs represent a suitable model for exploring molecular phenotype variability among LCLs.

Prior to experiments, we decided to verify that the cell lines are of the same genetic origin, have diploid genotypes, and have a similar cell cycle progression. We used short tandem repeat (STR) analysis to confirm the same genetic background of the cells, and that the male sex chromosome is present. We assessed four autosomal (D18S51, D8S1179, TH01 and FGA) and one allosomal (Amelogenin; AMELX and AMELY on the X and Y chromosomes, respectively) short tandem repeat by electrophoresizing fluorescently labelled PCR products of the selected regions (Figure 13). PCR product sizes were found identical between the cell lines, and both AMELX and AMELY variants were present. This result indicates that the isogenic cell lines were indeed derived from the same male individual and that there was no sign of contamination with genetically distinct cells.

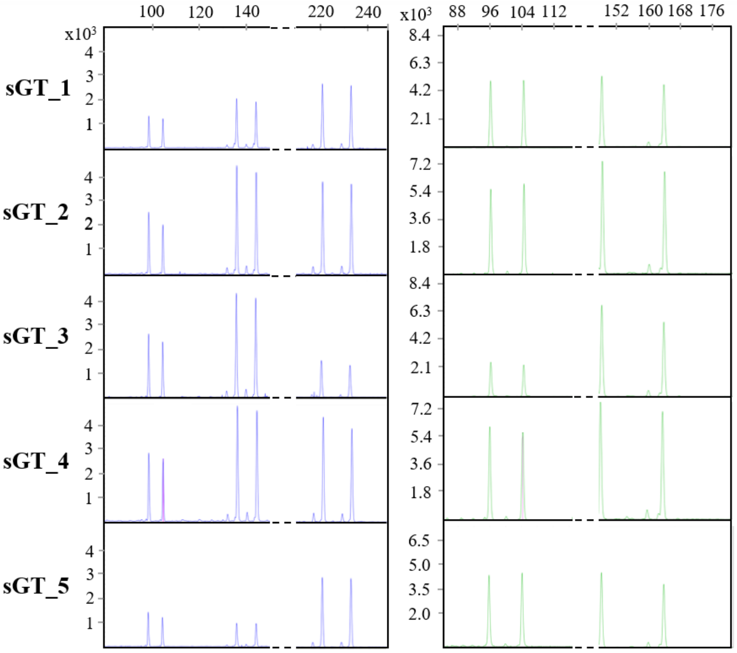


Figure 13 | Capillary electropherograms of PCR-amplified short tandem repeat regions.

The propidium-iodide intercalator was used for counting the fraction of cells at each stage of the cell cycle (G_0-G_1, S, G_2-M) and for assessing ploidy. We found that the cell lines had similar cell cycle stage distributions, and DNA indices (calculated DNA content per nucleus per haploid genome size) for all cell lines were below 1.1, indicating euploidy (Figure 14). The presented results suggested that the main characteristics of the selected five isogenic LCLs were similar, therefore are suitable for our study.

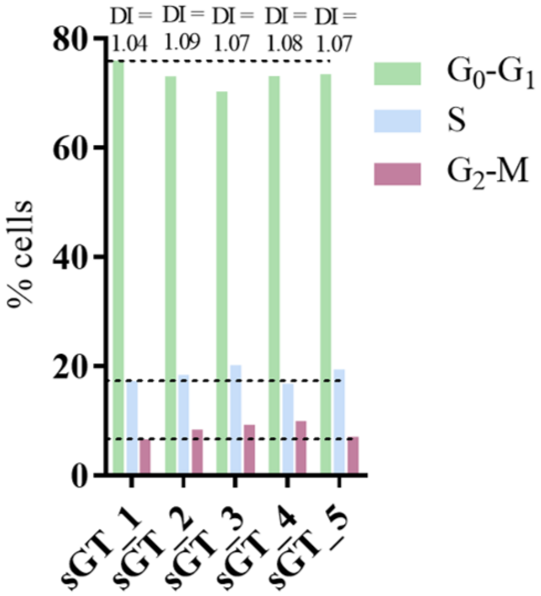


Figure 14 | The percentage of cells at each cell cycle stage and DNA indices calculated for isogenic LCLs based on propidium-iodide staining.

5.3.2 Intra- and inter-cell-line variability of isogenic LCLs at the level of protein surface markers as assessed by flow cytometry

After basic cell line characterizations, we decided to compare the expression of selected immune cell protein markers between isogenic LCLs using flow cytometry-based immunophenotyping. The method allowed us to identify the source cell type, assess surface marker expression heterogeneity within cell lines and cell line clonality.

It has been described that commercially available LCLs are commonly mono- or pauciclonal, and that cell line complexity (the number of constituent clones) quickly decreases during culturing (Plagnol et al., 2008). As LCLs are derived from B cells, clonality can be assessed based on measuring the cell surface expression of the kappa (κ) and the lambda (λ) immunoglobulin light chains; as light chain restriction is indicative for neoplasia, the method is widely used in the clinics to characterize leukaemias. In a mature B cell, as a result of DNA rearrangement during cell maturation, either the kappa or the lambda light chain is expressed (a.k.a. light chain exclusion). As the selection of a light chain is random, the ratio of kappa chain- and lambda chain-expressing B cells in the blood of a healthy individual is between 1-2 (O'donahue, Johnson, Hedley, & Vaughan, 2018). Flow cytometric analysis of isogenic LCLs double-stained with fluorescent anti-kappa and anti-lambda antibodies suggested that one cell line (sGT_4) was lambda-restricted monoclonal (with dim lambda expression), but pauciclonality cannot be excluded; while the other four cell lines expressed both light chains at various ratios: sGT_2 and sGT_3 were possibly polyclonal, and although sGT_1 and sGT_5 possibly represent a lower level of complexity, they were also derived from multiple B cell clones (Figure 15A). Assessing mRNA expression from the *IGL* and *IGK* loci by RNA-Seq supported our findings with flow cytometry (Figure 15B).

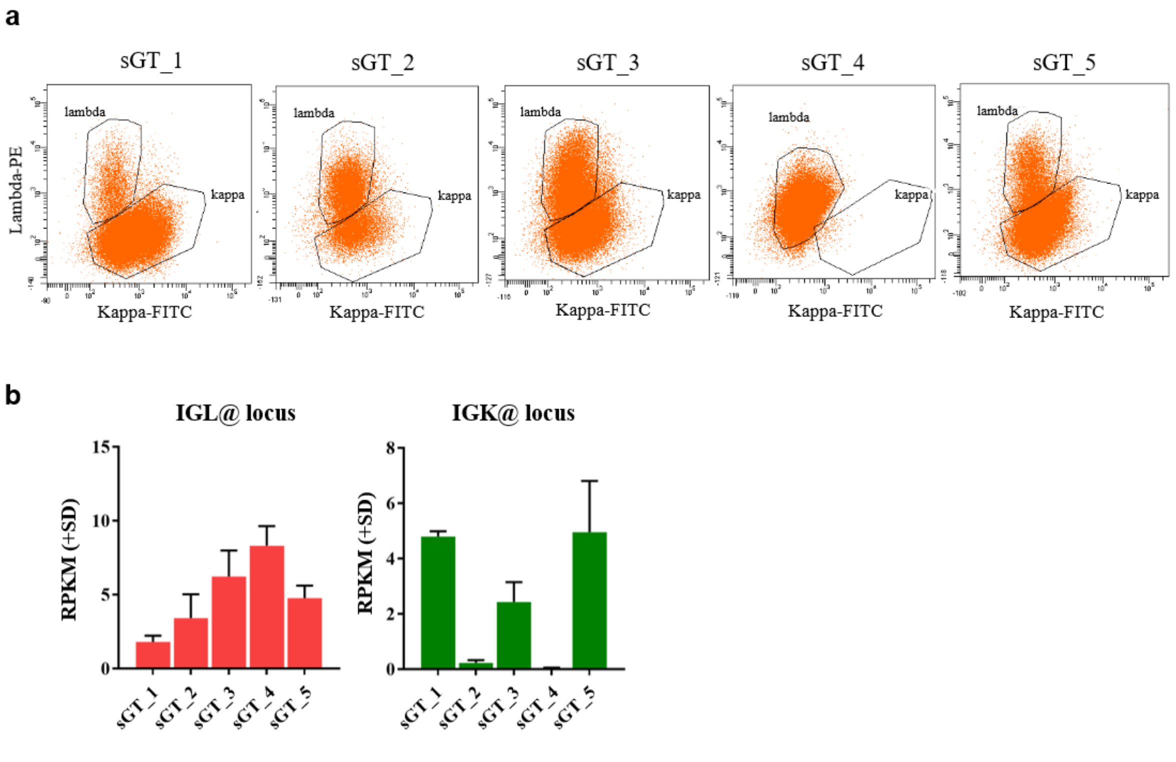


Figure 15 | Protein- and RNA-level expression of kappa and lambda immunoglobulin light chains in isogenic LCLs. a) Scatter plots display Lambda-PE and Kappa-FITC fluorescence intensities for the five isogenic LCLs, as well as the events falling within the lambda and kappa gates. **b)** RNA-level expression of the immunoglobulin lambda and immunoglobulin kappa loci (mRNA-Seq, N = 2).

Immunophenotyping is commonly used in the context of laboratory diagnostics to seek for leukaemia markers in blood samples, as well as in basic science. We used a panel of twenty antibodies in different combinations (for details, please refer to Materials and Methods) to confirm cell type of origin, clonality (see the previous section) and expression heterogeneity in isogenic LCLs. Overall, the cell lines showed expression patterns characteristic to mature B cells (CD19+, CD20+, CD22+, CD23+, CD45+, HLADR+, dim FMC7+, dim CD21+, dim CD43+, CD5-, CD10-, CD34-, and nTdT-). Interestingly, however, the pan B cell marker CD24 was not present in either of our cell lines (0.6-1.9% CD24+ cells). This might be the result of EBV infection, as EBV infection has been reported to diminish CD24 levels (Karran et al., 1995). Moreover, all isogenic LCLs were negative to CD79b encoding the beta component of the B cell receptor. The fraction of cells positive to certain markers showed high variability between the cell lines: CD81 was exposed in 54-76% of individual cells, 20-87% of cells stained positive for the activation marker CD38, and the expression of pre-B cell marker cytoplasmic immunoglobulin M (cyIgM) showed a marked difference between the cell lines (1-84%) (Table 5). In summary, isogenic LCLs were confirmed to have been derived from human B cells, and the expression of certain immune markers (CD81, CD38, cyIgM) high variability between the cell lines.

Table 5 | Protein marker profile of isogenic LCLs.

Marker	**sGT_1**	**sGT_2**	**sGT_3**	**sGT_4**	**sGT_5**
CD19	+	+	+	+	+
CD20	+	+	+	+	+
CD22	+	+	+	+	+
CD23	+	+	+	+	+
CD45	+	+	+	+	+
HLA-DR	+	+	+	+	+
CD21	dim	dim	dim	dim	dim
CD43	dim	dim	dim	dim	dim
FMC7	dim	dim	dim	dim	dim
CD10	-	-	-	-	-
CD34	-	-	-	-	-
CD5	-	-	-	-	-
CD79b	-	-	-	-	-
nTdT	-	-	-	-	-
CD24	0.6%	1.1%	1.2%	1.9%	1%
CyIgM	1%	84%	36%	15%	10%
Lambda	5%	59%	36%	dim	13%
Kappa	95%	36%	60%	-	86%
CD81	54%	54%	58%	74%	76%
CD38	87%	87%	63%	20%	64%

Dim = positive staining with low fluorescence intensity.

5.3.3 Marked differences in gene regulatory element activities among isogenic LCLs

We next decided to profile isogenic LCLs with regard to gene regulatory element activity using chromatin immunoprecipitation sequencing (ChIP-Seq) in biological duplicates. Active promoters and enhancers are known to be enriched for nucleosomes containing acetylated H3 histone at the 27th lysine residue (H3K27ac). Therefore it serves as a general gene regulatory element activity mark (a.k.a. active histone mark). The deposition of H3K27ac is guided by transcription factors and histone acetyltransferase activity.

In order to get an overall view of the level of similarity between LCLs in the same genotype context, we first derived a 'consensus' set of autosomal active regulatory regions by merging predicted H3K27ac-enriched regions (overlapping in at least two samples) across the replicate measurements of isogenic LCLs, as well as three LCLs derived from a CEPH/UTAH trio (mother, father and son), the members of which were genetically unrelated to the isogenic LCLs. We drew a heatmap of calculated pair-wise correlation scores and found that although the correlation coefficients were remarkably high across the isogenic LCL dataset (between 0.9 and 0.97), biological replicates clustered together, indicating that isogenic LCLs have their unique epigenetic profile. Not surprisingly, the correlation coefficients were generally lower between genetically distinct cell lines, clearly distinguishing the trio cell lines from isogenic LCLs (Figure 16).

For our further analyses, we defined an isogenic LCL-specific 'consensus' set of regions using only the isogenic LCL dataset (including both autosomal and allosomal sites), resulting in a working set of 42,923 regions. We calculated normalized ChIP-Seq read counts (RPKM values; reads per kilobase per million mapped reads) sample-wise over the consensus set. Strikingly, we found that almost one-fourth of consensus regions (9,685 sites) had variable H3K27ac enrichments across isogenic LCLs (RPKM fold-difference > 2, P value < 0.05, between at least two cell lines; for details, see Materials and Methods) (Figure 17). When we compared each pair of cell lines, we found 1,056 to 4,174 variable regulatory elements (fold-difference > 2, P value < 0.05) (Figure 18).

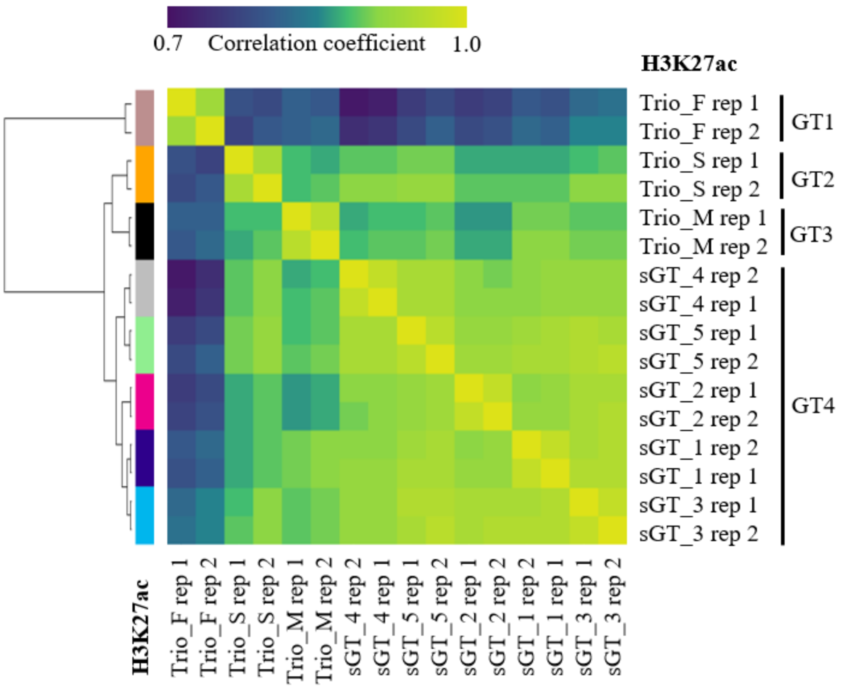


Figure 16 | Correlation heatmap of H3K27ac signals over a consensus set of autosomal genomic regions in four individuals: a CEPH/UTAH trio (Trio_F = father; Trio_M = mother; Trio_S = son) with three different genotypes marked as GT1-GT3, and the CEPH/UTAH male from whom the five isogenic LCLs of this study was prepared (GT4; sGT_1-sGT_5). GT = genotype.

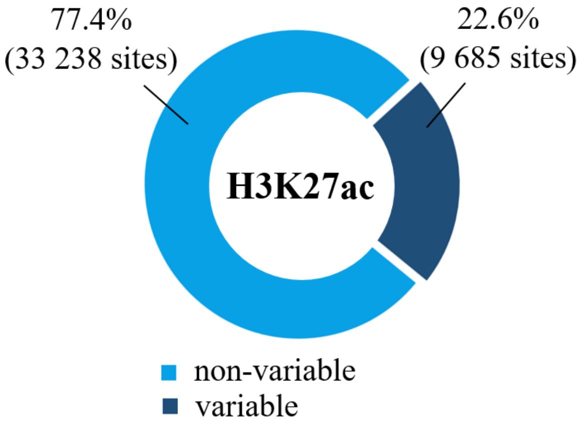


Figure 17 | Percentage of consensus regions showing variable (fold-difference > 2, P value < 0.05, between at least two cell lines) and non-variable H3K27ac enrichment across the five isogenic LCLs.

Next, we performed a genomic feature association analysis, comparing regions with variable and non-variable H3K27ac levels across all cell lines. We found that intergenic regulatory elements (enhancers) were highly affected, while promoter regions were under-represented in the variable set. These observations indicate that H3K27 acetylation levels are relatively stable at promoter elements across the cell lines and that this kind of robustness is not characteristic to intergenic enhancers (Figure 19).

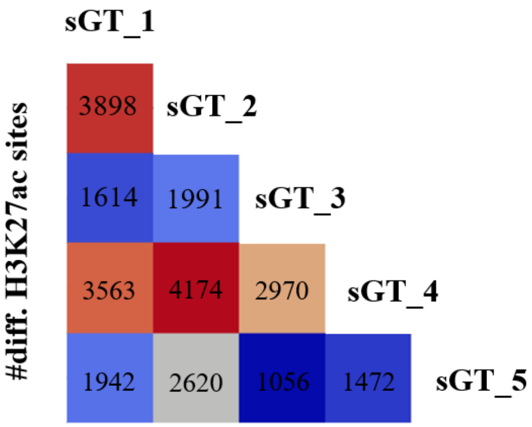


Figure 18 | Pairwise comparison of isogenic LCLs based on their H3K27ac signals over the consensus set of regions (fold-difference > 2, P value < 0.05).

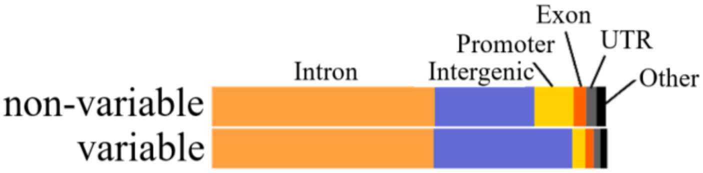


Figure 19 | The fraction of variable and non-variable regions mapping to certain Genome Ontology annotation categories.

5.3.4 Coordinated loss or gain of gene regulatory element activities over extended genomic regions

In the next step, we decided to assess whether the activity of super-enhancers (SEs), a distinct class of regulatory elements also vary between isogenic LCLs. Super-enhancers have been described as linearly clustered gene regulatory elements less than 12.5 Kb apart from each other, spanning several kilobases. SEs are characterized by distinct chromatin signatures, are highly active, and are commonly found in the close proximity of key cell identity genes and oncogenes (Lovén et al., 2013; Whyte et al., 2013). Dysregulation of SEs has been associated with several pathological conditions, including autoimmune diseases and cancer (Farh et al., 2015; Hnisz et al., 2013; Mansour et al., 2014). We predicted SEs from H3K27ac ChIP tags pooled across all samples, which resulted in 1,058 putative SEs (Figure 20). The predicted SEs were located in the proximity of genes involved in B cell and immune functions, including the B cell master regulator transcription factor PAX5 and IRF2. Moreover, several predicted SEs have recently been identified as being formed as the result of EBV infection (EBV super-enhancers, e.g. BCL2, ETS1, MIR155 and MYC) (H. Zhou et al., 2015). We found that although almost half of the SEs (518; 49%) contained at least one constituent enhancer element that had been classified as variable, changes in H3K27ac levels over the whole SE regions were modest; only 31 (or 2.9%) were found variable (P value < 0.05, fold-difference > 2) across isogenic LCLs (Figure 21). This finding suggests that individual SE elements may vary across cell lines without seriously affecting SE activity *per se*. Super-enhancer activity differences show a linear relationship with the expression of the closest gene (Figure 22). Genes closest to variable SEs were associated with immune functions such as leukocyte activation ($5.5*10^{-4}$) and leukocyte cell-cell adhesion ($5.6*10^{-4}$). Transcription factor activity ($P = 1.2*10^{-2}$) and LPS (lipopolysaccharide) binding ($P = 3.2*10^{-2}$) were among the most enriched molecular functions.

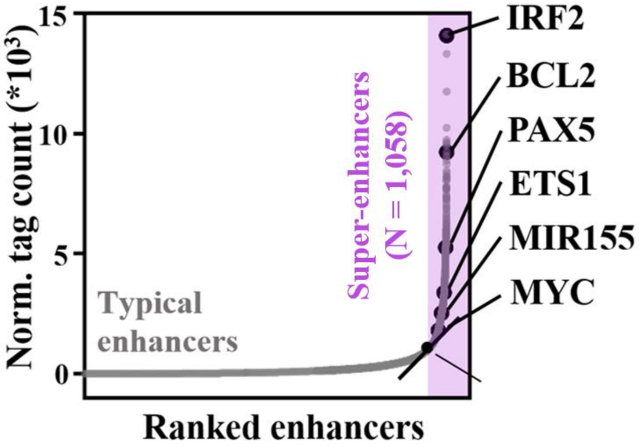


Figure 20 | H3K27ac ChIP-Seq tag counts over typical and super-enhancers (SEs), ranked by tag count; the black dots represent SEs which are associated with genes related to B cell-specific and immune functions or related to EBV infection.

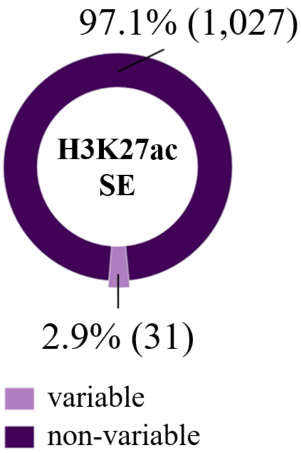


Figure 21 | The percentage of super-enhancers with variable H3K27ac enrichment across the five isogenic LCLs (fold-difference > 2, P value < 0.05).

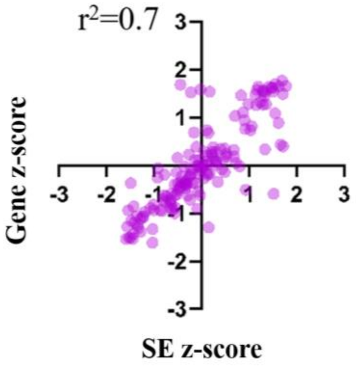


Figure 22 | Correlation between H3K27ac variability over super-enhancers and the expression of the closest expressed gene; z-scores were plotted against each other, for all possible comparisons (N=155).

As the definition and therefore, the prediction of super-enhancers is somehow arbitrary, we took an alternative approach to assess whether coordinated loss or gain of regulatory element activity extends to non-super-enhancer regions as well. We found that variable enhancers cluster based on the direction of change (e.g. sGT_1 > sGT_2 and sGT1 < sGT_2) over long genomic regions not previously classified as super-enhancers The karyogram on Figure 23 shows the relationship between H3K27ac signal intensities over variable regions of chromosome 3, between sGT_1 and sGT_2 cell lines. An approximately 4 Mb-long region is presented, including BedGraphs of isogenic cell lines and a public LCL Hi-C (chromosome conformation capture of all chromatin contacts coupled with sequencing) dataset, which shows that coordination may extend through multiple topologically associated domains (TADs), and signal direction may also switch from one TAD to another (Figure 23).

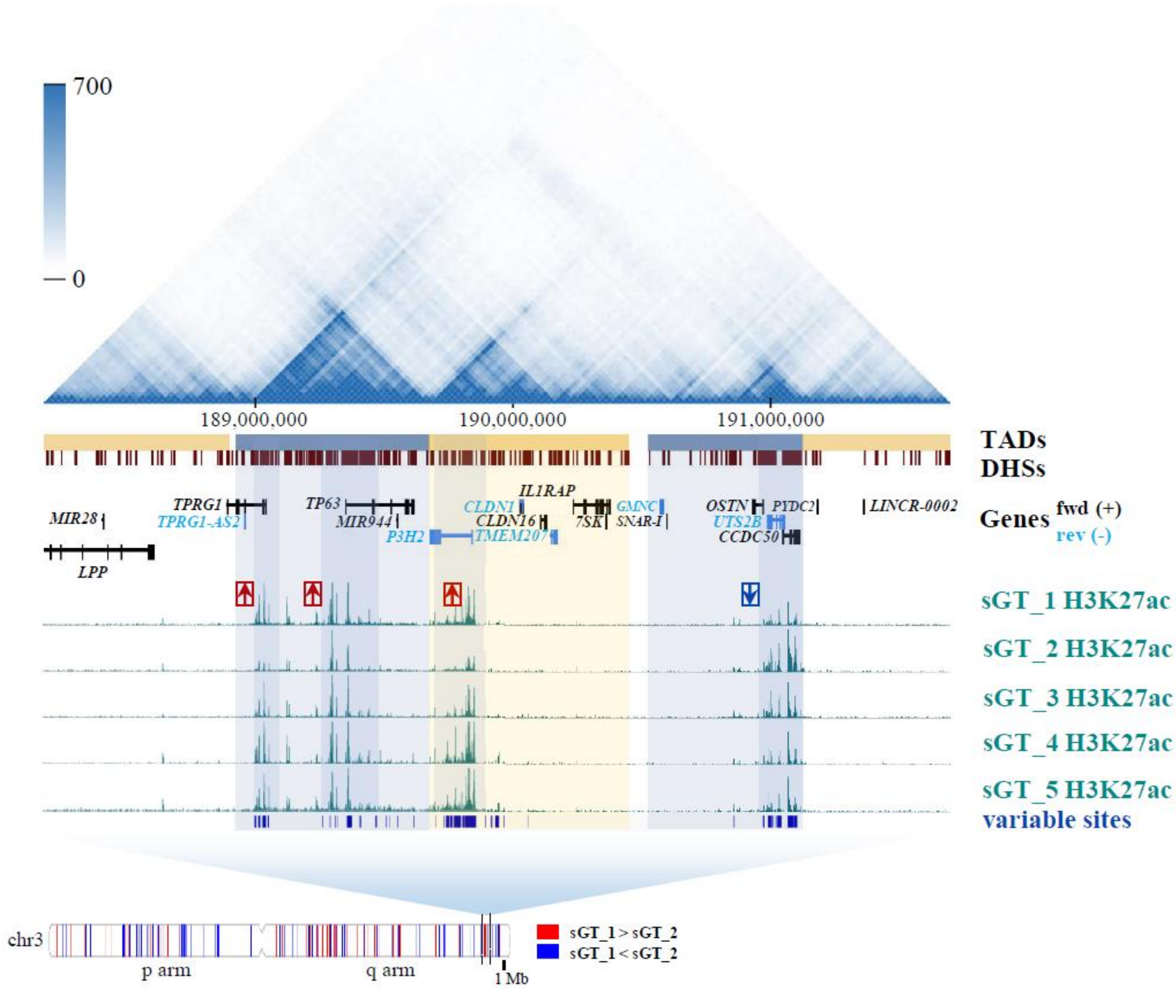


Figure 23 | Long-range coordination of enhancer activity in LCLs. The heatmap shows Hi-C contact frequencies over a ~4 Mb-long genomic region on chromosome 3 (GM12878 cell line), blue and yellow vertical lines show predicted topologically associated domains (TADs), while black horizontal lines show DNase hypersensitive sites (DHSs). Red arrows indicate the direction of change between the two cell lines. BedGraph tracks at the same genomic location from sGT_1-sGT_5 are shown. Variable sites are represented as blue vertical lines under the BedGraphs. The karyogram shows variable sites between sGT_1 and sGT_2, coloured based on the direction of change.

In order to assess whether the long-range coordinated change is associated with chromatin contact frequencies, we used the trio LCLs, which show high (mother), moderate (child) and low (father) H3K27ac signals around the P3H2 gene, which is located in the 4-Mb region described above (Figure 24). We performed multiplexed 3C-Seq in the trio cells using a P3H2 promoter-proximal region as a bait. When compared to the H3K27ac signals, we found that chromatin contact frequencies correlate with histone acetylation signal. Notably, this kind of experiment has no power to suggest causality or shed light on the temporal order of gene regulatory events.

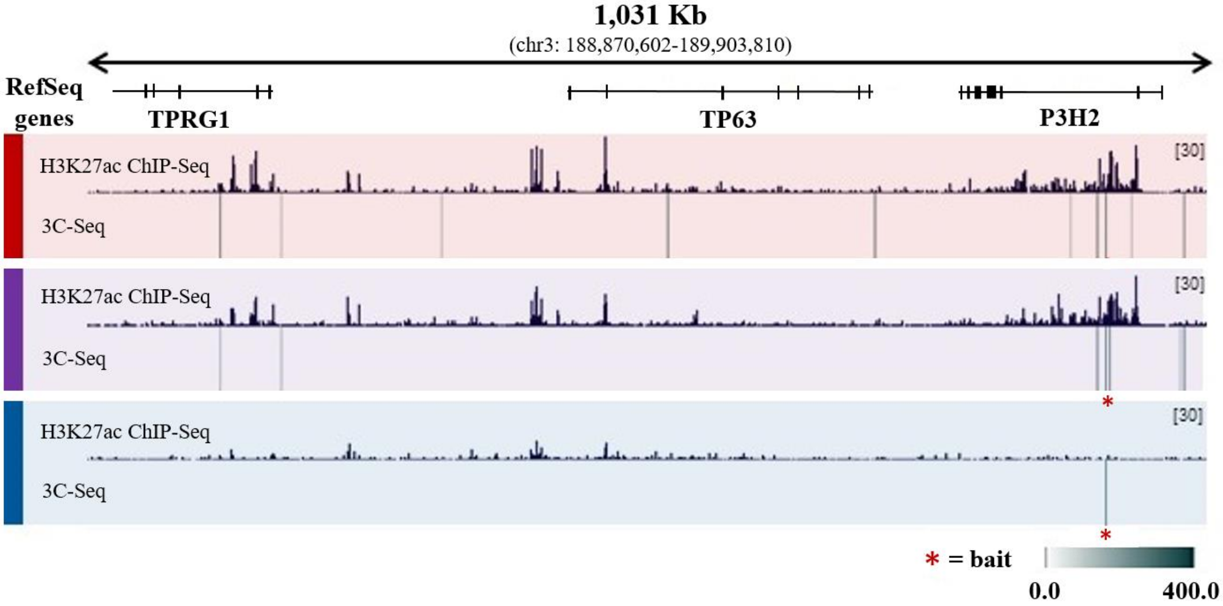


Figure 24 | 3C chromatin contacts in trio LCLs. H3K27ac BedGraph tracks are shown along with contact frequencies between the bait sequence (indicated with red asterisks) and neighbouring genomic regions demarcated by HindIII cutting sites. Red = mother (GM12873), purple = child (GM12864), blue = father (GM12872).

5.3.5 H3K27ac variability is linked to transcriptomic variability in isogenic LCLs and affect clinically relevant pathways

We next assessed whether and to what extent chromatin activity differences were mirrored by transcriptome-level differences. We performed mRNA-Seq experiments in the five isogenic cell lines in biological duplicates. In order to compare the changing patterns of H3K27ac signal and RNA levels, we plotted H3K27ac ChIP-Seq tag counts of 9,685 variable sites +/- 1kb of the cental nucleosome-free region, and the z-scores of closest genes. We performed k-means clustering of variable H3K27ac sites in order to capture the main patterns of H3K27ac variability, resulting in six clusters. We observed that gene variation patterns generally follow that of the H3K27ac signal (Figure 25). Representative examples of each cluster are shown in Figure 26. However, only 525 (4.6%) of genes were found to be significantly differentially expressed (differentially expressed genes; DEGs) across isogenic LCLs (FDR = 0.05, fold-difference > 2).

Pairwise comparison of cell lines resulted in 25 to 229 variable genes (mean = 119.8; median = 107.5) (Figure 27). Notably, none of the previously reported EBV copy number-related genes (CXCL16, AGL, ADARB2) (Houldcroft et al., 2014) were found to be differentially expressed in our dataset, suggesting that gene expression changes had not been induced by EBV infection differences.

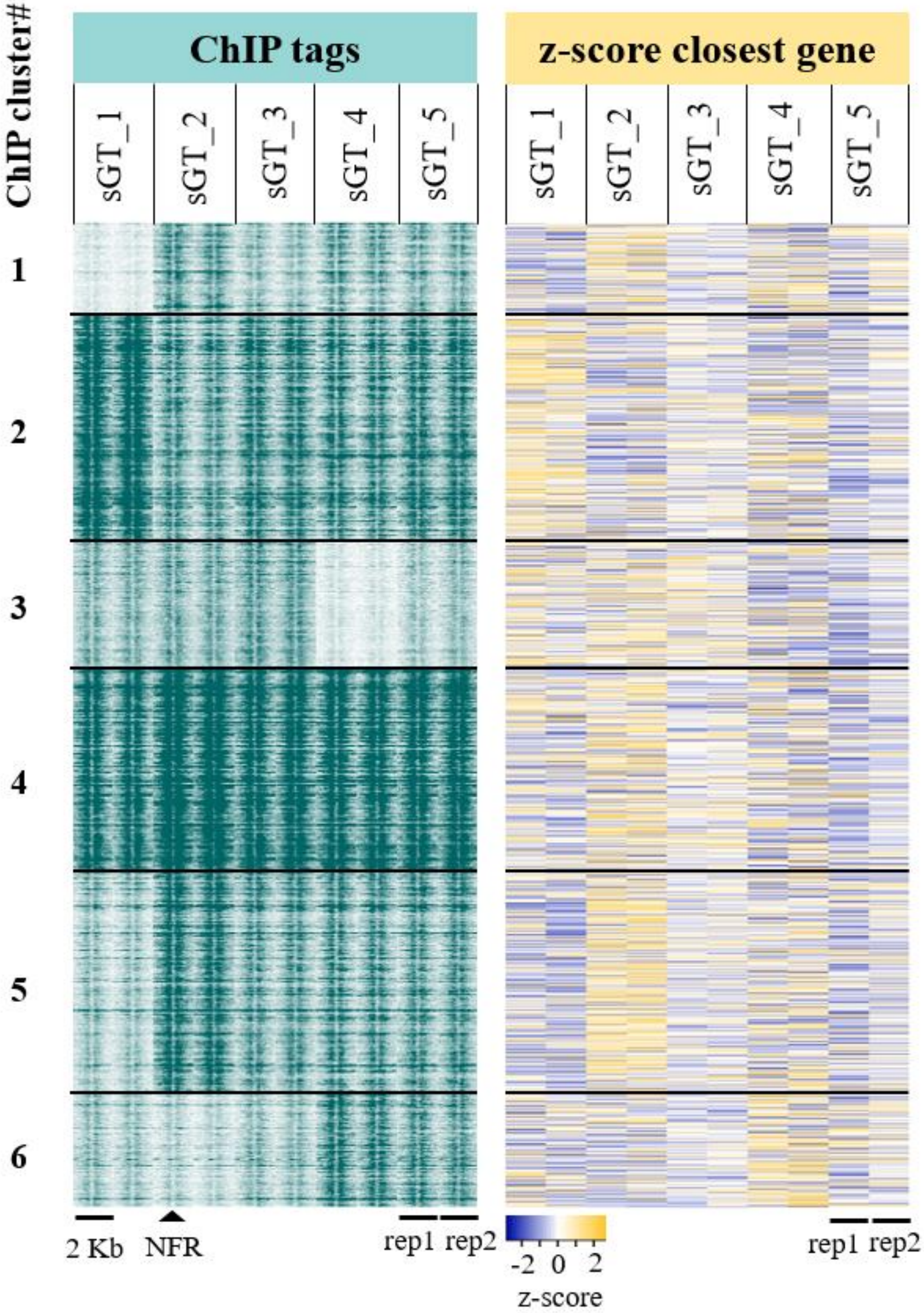


Figure 25 | Correlation between H3K27ac ChIP-Seq signal and gene expression in isogenic LCLs genome-wide. Read distribution heatmap on the left-hand side shows ChIP-Seq tag counts around the central 'valleys' variable H3K27ac-enriched regions (+/- 1 kb). The heatmap shows z-scores of closest genes. Both replicates for each sample are represented.

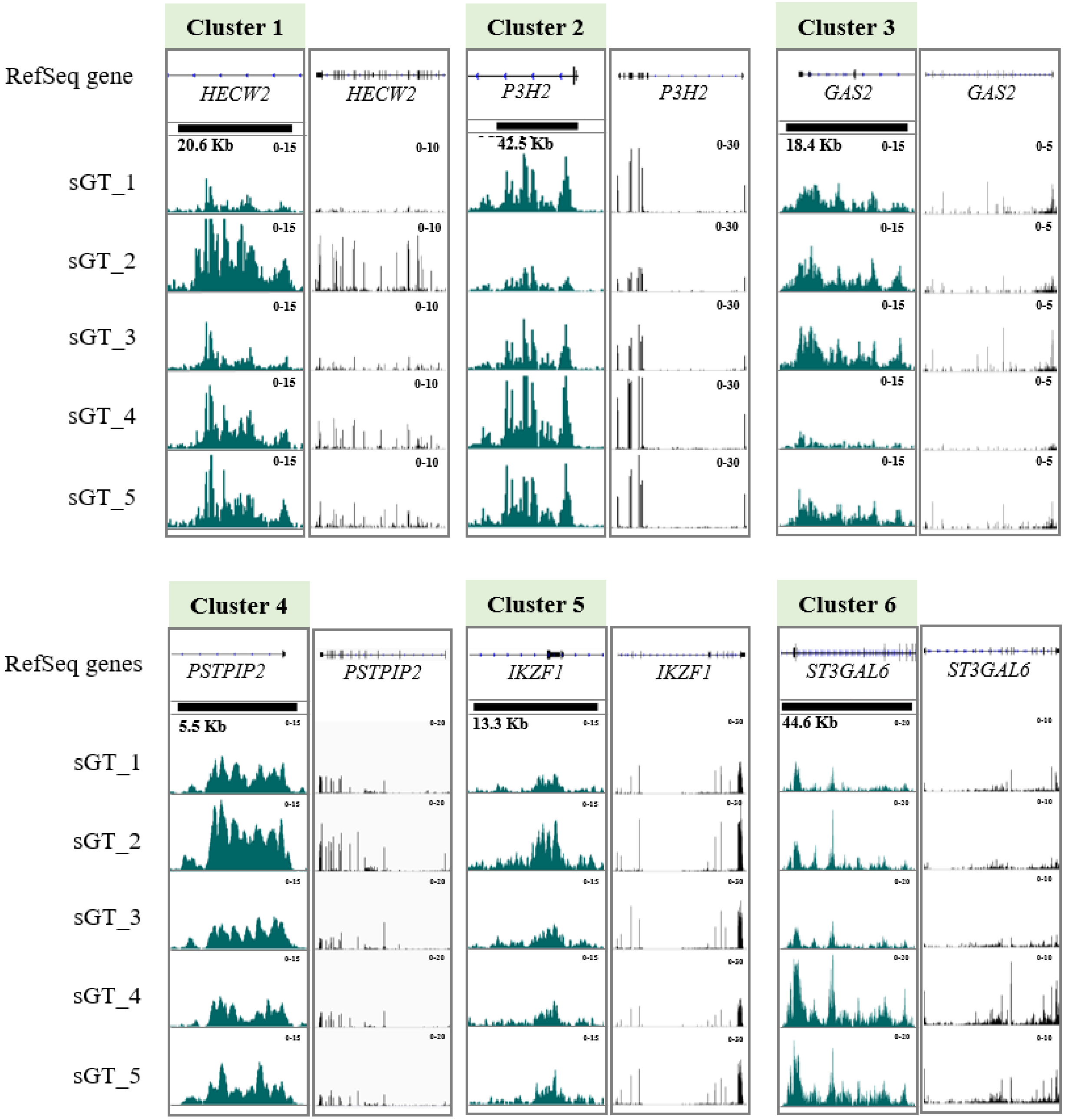


Figure 26 | Examples of correlation between H3K27ac signal and gene expression in isogenic LCLs. Integrative Genomics Viewer snapshots show scale-matched ChIP-Seq, and RNA-Seq signals (BedGraphs) generated from pooled replicates per sample. Each example represents one cluster of the six presented in Figure 10. The black rectangles over the ChIP-Seq tracks show the length of each variable region.

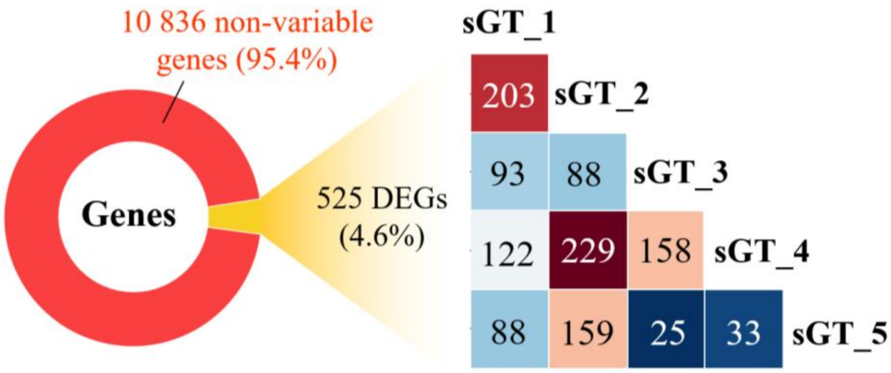


Figure 27 | Overall gene expression variability and pairwise gene expression differences in isogenic LCLs. The pie chart shows the number and percentage of differentially expressed genes across all five cell lines, while the right-hand panel shows the number of genes variable between each pair of isogenic LCL. DEG = differentially expressed gene.

Among the enriched biological process gene ontology terms related to the DEG set were cell migration (P = $2.8*10^{-12}$), intracellular signal transduction (P = $9.8*10^{-9}$), and regulation of apoptotic process (P = $3.4*10^{-8}$), and few of the most enriched molecular functions in the gene set included immune receptors and transcription factors (Figure 28A-B). Surprisingly, 121 of DEGs had been previously categorized as being pharmacogenes (genes associated with response to pharmaceuticals) based on the Genetic Association Database (GAD) (Figure 28C). By comparing the coefficient of variance (CV) and gene expression level trends, we found that the lower the mean gene expression level is, the higher the associated CV value is. This may be explained by the intrinsically higher variability of lowly expressed genes, as well as the sampling bias during the experiments. Higher CV values were associated with receptor function, cell surface localization, and play roles in signal transduction and cell motility. Genes with lower CV values were predominantly located inside the cell, related to signal transduction pathways and mediate immune and apoptotic functions (data not shown). Taken together, RNA-level variability among isogenic LCLs is far less pronounced than H3K27ac activity level variability, which is not unexpected in the light of studies describing enhancer redundancy (Osterwalder et al., 2018).

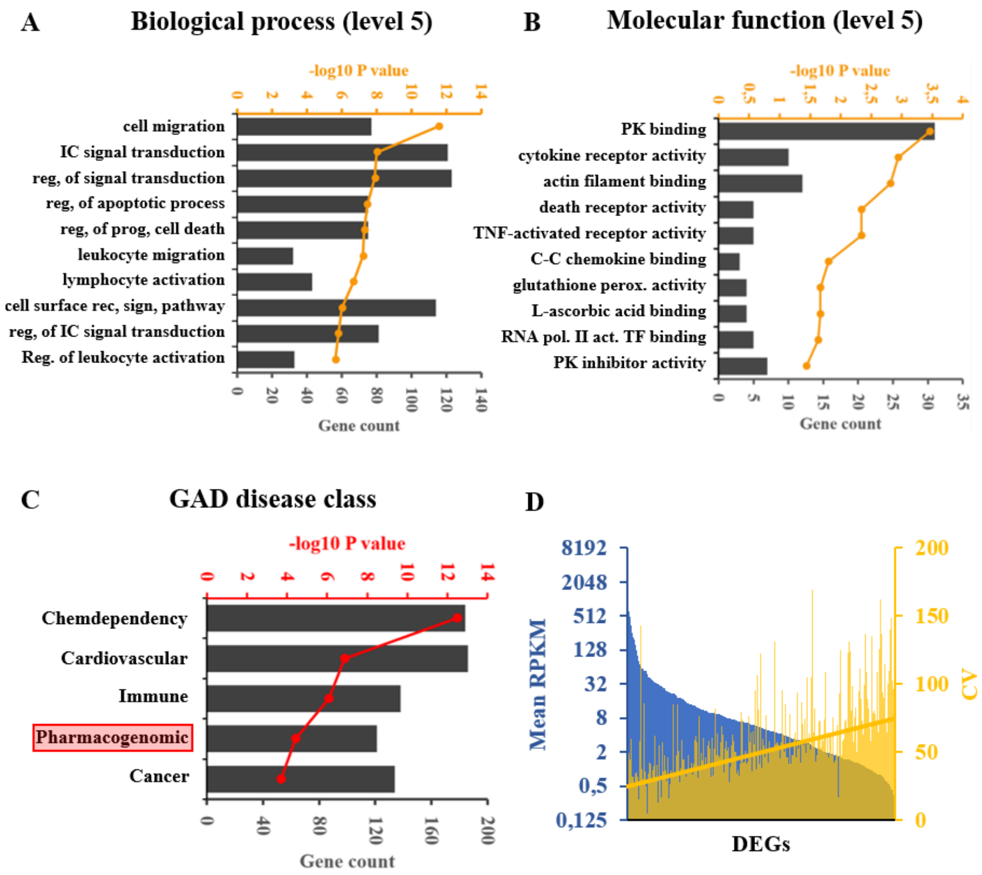


Figure 28 | Gene Ontology (GO) annotation and Coefficient of Variance (CV) of differentially expressed genes. a) Gene numbers and –log10 P values related to the most highly enriched biological processes, **B)** molecular functions, and **C)** Gene Association Database (GAD) disease classes. **D)** Mean RPKM of differentially expressed genes across isogenic LCLs, and corresponding CV values. The yellow line is the linear fit of CV values. DEG = differentially expressed gene.

5.3.6 Non-genetic variability might lead to an altered response to drug treatment

Having uncovered the extent of epigenetic and transcriptomic variability in isogenic LCLs, and as numerous DEGs were associated with drug response, we decided to asses whether this phenotype level is also affected, which may have implications in LCL-based pharmacogenomic research. Dihydropirimidine dehydrogenase (DPYD) catalyzes the rate-limiting step of pyrimidine catabolism, including the anti-cancer drug 5-fluorouracil (5-FU). Familial DPYD deficiency has been known to lead to serious, sometimes life-threatening 5-FU toxicity (Diasio, Beavers, & Carpenter, 1988). We chose DPYD as our model gene as it had been found significantly differentially acetylated and expressed between sGT_1 and sGT_2 (8.5x fold-difference) (Figure 29A), and as 5-FU toxicity can be measured by an MTT-based viability assay.

We could also validate gene expression difference by RT-qPCR using RNA samples independent of those used for RNA-Seq (Figure 29B). We treated sGT_1 and sGT_2 cells with different concentrations of 5-FU and measured their viability, as a measure of cytotoxicity, after 72 hours. We found that sGT_2 cells (higher DPYD expression) were less sensitive to 5-FU treatment compared to low DPYD expressing sGT_1, with an almost 2-fold increase in the half-maximal inhibitory concentration (IC50 sGT_1 = 0.63 µM, IC50 sGT_2 = 1.21 µM) (Figure 29C). Our result suggests that non-genetic factors might as well as affect LCL cells' response to drugs in pharmacogenomic screenings.

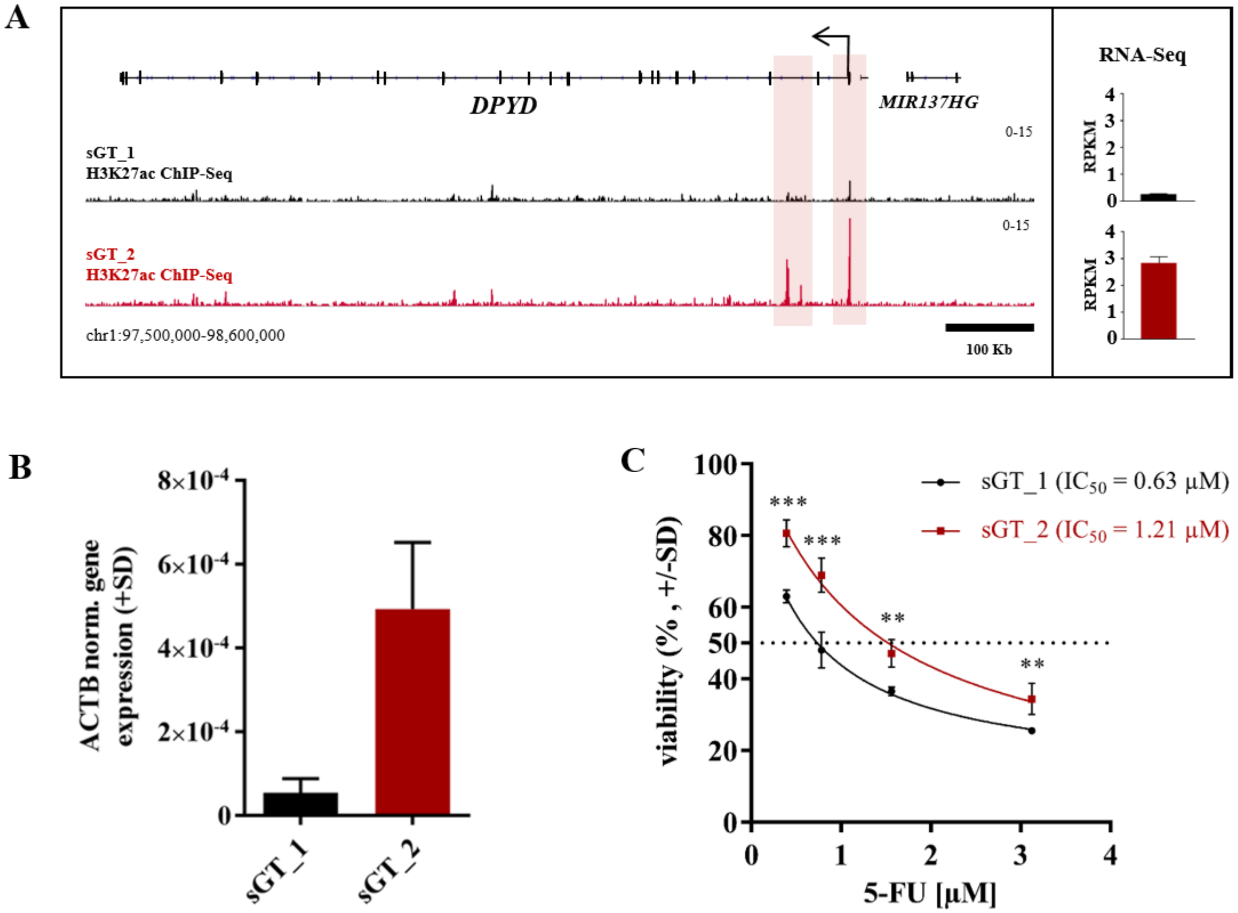


Figure 29 | H3K27ac landscape and mRNA level of DPYD and 5-fluorouracil sensitivity. A) Integrative Genomics Viewer track view of H3K27ac BedGraphs from pooled replicates, surrounding the DPYD gene in sGT_1 and sGT_2 cell lines; RNA expression of DPYD based on two replicates of mRNA-Seq data (N =2; RPKM+SD). B) DPYD expression from independent samples as measured by RT-qPCR (+SD, N = 2) in sGT_1 and sGT_2. C) Dose-response curves of sGT_1 and sGT_2 related to 5-fluorouracil (5-FU) treatment for 72 hours (MTT assay). IC_{50} values were calculated using the non-linear regression curve fit using the least-squares method. *** = P value < 0.001, ** = P value < 0.01.

6 DISCUSSION

6.1 Bacteriophages selected through in vitro evolution as novel spike-in controls for chromatin immunoprecipitation experiments

Since its development (Smith, 1985), phage display has been used for various applications including antibody epitope mapping (Moreira et al., 2018), and antibody generation (Hammers & Stanley, 2014). Multiple phage types and displaying peptides can be used depending on the aims of the study, and libraries with various displayed peptide lengths and even conformational mimotopes can be generated or purchased from different vendors. Phage display is a versatile method allowing not only for epitope mapping of clinical antibodies with unknown specificity but for the generation and production of experimental reagents (e.g. antibodies). Phages with desired specificities can be generated by selection rounds resembling natural evolution (biopanning), monoclonal reagents can be developed from pre-selected, polyclonal phage libraries, and DNA sequences coding for desired peptides, or peptide variants, can even be cloned into the phage genome.

While ChIP-qPCR and ChIP-Seq hold great promise for clinical applications, ChIP protocols are labour-intensive, various protocols and reagents are in use with a limited number of commercially available kits, and general best practices have not been laid, which may lead to limited lab-to-lab reproducibility. While initial cell number differences can be controlled for using the per cent input method, there is a source of variation which remains uncontrolled: variable sample loss due to pipetting errors through many steps of the protocols. As the majority of sample handling comes after setting aside input samples, this represents a major and unresolved problem. Although using replicates may, to some extent, cushion the effect of these errors, but limited sample availability and relatively high input cell numbers needed may not allow for using replicates. Of note, internal controls like housekeeping genes for RT-qPCR are not an option, as there is no information on ChIP signal stability in the context of various cell types, treatment types, and for the plethora of ChIP antibodies. In theory, spike-in controls may substitute for missing internal controls.

RNA spike-ins, synthetic transcripts of known sequence and quantity or a mixture such transcripts, are mixed into experimental RNA samples and are used to calibrate RT-qPCR, hybridization or RNA-Seq experiments. Using spike-in controls enable the assessment of quantitation sensitivity and accuracy, promoting comparable analysis of different platforms, protocols and samples (Jiang et al., 2011; Munro et al., 2014). DNA spike-ins can be used can also be used to check for the presence of PCR inhibitors (Buckwalter et al., 2014).

A suitable ChIP spike-in procedure control would be a reagent containing a protein-DNA complex or pools of protein-DNA complexes, which are carried over through the experiment from IP until the end of ChIP fragment isolation, and is easily quantifiable. Obviously, spike-ins should not compete with chromatin epitopes. This requirement may be addressed by coupling the spike-ins to antibodies, or capturing bead-antibody complexes in a separate experiment, and adding these beads to the reactions at an appropriate time point or by using indifferent phage-antibody pairs. This would enable a within-experiment and between-experiment control of experimental biases. In phage display, either biopanning or directional cloning would be the viable option to develop phages with high affinity to ChIP-grade antibodies. The primary advantage of phage display in this context over other proposed spike-ins (Bonhoure et al., 2014; Egan et al., 2016; Grzybowski et al., 2015; Orlando et al., 2014) is that phages can be easily regenerated in-house by re-infecting a specific strain of bacteria followed by a relatively easy phage purification protocol. In addition, monoclonal phage stocks may have a more consistent quality than xenogenic chromatin-based spikes.

As the M13KE genome is circular without free DNA ends, this genome would be selected against during ChIP-Seq library preparation. This would not be a problem when ChIP-qPCR is the selected method of quantification but hinders its use as potential ChIP-Seq normalizers. In case ChIP-Seq is the primary method of choice, the T7 phage with linear dsDNA genome would present a more suitable choice.

We presented a method to develop phage-display-based spike-in procedural controls, which may later be used for filling the controllability gap of ChIP experiments. We tested the biopanning-based method for several ChIP-grade antibodies, including anti-AR, anti-ER (estrogen receptor), anti-RXRα (retinoid X receptor alpha) and anti-CTCF (CCCTC-binding factor), all designated for human samples. Also, we generated monoclonal stocks from biopanned, mixed phage libraries, as polyclonal stocks may evolve in terms of ratios of constituent phages during library propagation in bacteria, which may hinder reproducibility. Notably, in an ideal scenario, full-length peptides used for immunization would be cloned into the phage genome and propagated, but information on the immunizing peptide(s) is not available for many of the commercially available ChIP-grade antibodies. In this case, only biopanning can be considered. The author of the present dissertation was responsible for carrying out or supervise all AR-based experiments. Therefore the results are demonstrated based on the example of AR. We showed that biopanning of a heptapeptide phage library resulted in stocks of polyclonal phages with increasing recovery in simulated ChIP experiments in each round. Also, all monoclonal phages generated from the highest affinity polyclonal stock were shown to bind to the selection antibody in the context of a ChIP experiment. An important limitation of our study was that high-affinity phage stocks for post-translationally modified histones, common targets for ChIP, could not be generated. Further characterizations, i.e. reproducibility measurements, are needed for developing a ChIP quality control system that can be used for clinical research or diagnostic procedures.

We believe that phages mimicking transcription factors epitopes as ChIP spike-in controls might be of interest for the scientific community, and might pave the way for developments aiming a better standardization of ChIP experiments.

6.2 Lyophilization and ambient storage of human cells to preserve RNA

It is imperative to consider the financial aspects of biological sample storage for research and diagnostic applications. Conventional storage systems are designed to preserve biomaterials in the long run, therefore operate at ultra-deep temperatures (from -180°C to -20°C) in order to minimize molecular motion. Operating these storage systems, especially ultra-low temperature freezers, have high associated costs and considerable environmental impact. The US and South-east Asia have to face disasters like floods and thunderstorms, which endangers sample storage systems relying on continuous electricity supply. Long-term electric outages might lead to sample thawing which is detrimental to sample quality (both structurally and at the molecular level), even if equipped with an electronic backup system. The growing number of international collaborations and the emergence of biobanks shipping globally increasingly require transport methods ensuring sample integrity during temperature fluctuations or long waiting times at border controls. Lyophilization has been proposed as a method of choice to safely dry biosamples, resulting in long shelf-lives.

Lyophilization has become a fairly common technique to stabilize food products, liquid pharmaceutical formulations, and bacterial strains. Studies are underway to extend its utility to lyophilization of mammalian cells and tissues, for various downstream applications. DNA, RNA and protein stabilities in lyophilized cells and tissues have been assessed by low-throughput methods, using highly variable protocols and equipment, with generally positive results (Leboeuf et al., 2008; Mareninov et al., 2013; Matsuo et al., 1999; Takahashi et al., 1995; Weisberg et al., 1993; Wu et al., 2012; M. Zhang et al., 2017). Lyophilized materials have been shown promising for *in vitro* fertilization (Das et al., 2010; Gianaroli et al., 2012; M.-W. Li et al., 2009; Liu et al., 2004; Loi et al., 2008b; Wakayama & Yanagimachi, 1998), blood transfusion-based uses (Arav & Natan, 2012; Deprés-Tremblay et al., 2018; Shiga et al., 2016, 2017; Wolkers et al., 2002; X.-L. Zhou et al., 2007), and regenerative medicine (Buchanan, Pyatt, & Carpenter, 2010a; Natan et al., 2009; S. Zhang et al., 2010). Lyophilization of mammalian cells for downstream RNA studies has not become standard, despite a few studies reporting only minimal RNA degradation using a limited toolset of RT-qPCR of certain genes and gel electrophoresis. In our research, we aimed at extending our knowledge on lyophilizing mammalian cells for RNA-based downstream applications, including novel insights by measuring low abundance genes, enhancer RNAs and profiling the transcriptome by RNA-Seq.

RNAs are prone to RNAse-mediated degradation, especially when tissues are removed from the *in vivo* context and subject to ischaemia. *In vitro*, RNAse contamination or release due to tissue damage (Augereau, Lemaigre, & Jacquemin, 2016), alkaline cleavage and high temperatures (Mikkola, Lönnberg, & Lönnberg, 2018), and to a lesser extent, oxidative stress (Thorp, 2000) contribute to RNA instability. In lyophilized tissues residual moisture, light exposure (Leboeuf et al., 2008) and lipid-driven peroxidation were shown to affect RNA integrity (Damsteegt et al., 2016; Matsuo et al., 1995).

In our preliminary experiments, we used IMT-2 solution containing trehalose and epigallocatechin-gallate in PBS as lyoprotectants, which was described as an efficient membrane stabilizer (Natan et al., 2009), which would facilitate applications requiring intact cells, including ChIP. Although we could recover microscopically intact cells with discontinuities allowing for trypan blue penetration, and fragility to shear force introduced by pipetting, isolated RNAs showed lower overall RNA integrity as assessed by calculating RIN values. Lower RINs may have resulted from endogenous RNAse-based degradation during cell washes or might result from the activation of apoptotic pathways previously described for epigallocatechin-gallate -treated LCLs (Noda et al., 2007). Moreover, reverse transcription was shown to be inhibited by the epigallocatechin-gallate co-precipitated with total RNA, and H3K27ac ChIP was also unable to enrich the positive control region, the regulatory element of a transcription factor gene highly expressed in the steady-state. This might be the result of the inhibition of the experiment at some step (i.e. fixation), or more likely reflect a biological response to epigallocatechin-gallate during the max. Five minutes of incubation in IMT-2 prior to snap freezing or during rehydration. After obtaining these results, we concluded that using such a potent gene regulator (H.-S. Kim et al., 2014; Noda et al., 2007) and reverse transcription inhibitor with rapid membrane penetration would hinder the comparability of lyophilized samples as lyophilized cells would not represent cells at the steady-state, especially when different cell types or cells kept in lyophilization solution for technical reason to varying times are to be compared. After considering the above concerns, we decided to carry out our experiments without epigallocatechin gallate and focused on the quality and quantity of RNA isolated from cells lyophilized in 0.1 M trehalose/PBS.

In the subsequent experiments, trehalose was used as lyoprotectant, which is a relatively cheap reagent and has been shown to sequester reactive oxygen species (Benaroudj, Lee, & Goldberg, 2001), as well as to protect cells during dehydration even when it remains extracellular (Eroglu et al., 2000). Keeping in mind the importance of cost-effectiveness in biobanking, we aimed at using a short lyophilization cycle, as energy consumption by the freeze dryer may easily become prohibitive to adopting the method. Using a manifold freeze dryer, we could dry 0.5 ml samples in six hours, which is substantially less than cycle lengths used in the pharmaceutical industry (sometimes days). Not surprisingly, trehalose did not allow for the recovery of intact cells probably due to membrane damage, but the dried products could easily be resuspended in TRIzol without previous washing, and resulted in high quality (RIN) and quantity RNA isolates even after two weeks or two months of room temperature storage in the dark, in the presence of a desiccant. As RIN values may not sufficiently represent RNA types other than rRNAs, we measured mRNAs, as well as lncRNAs and eRNAs from lyophilized samples by RT-qPCR, without significant change even for extremely low abundance genes compared to paired control samples.

PCR-based applications using selected genes may not necessarily reflect transcriptome-wide changes, as supported by studies showing the relative insensitivity of RT-qPCR to overall sample quality (Baechler et al., 2004; Bray et al., 2010; Catts et al., 2005; Gallego Romero et al., 2014; Ibberson et al., 2009). As RNA-Seq is increasingly used for biomarker studies, as well as to assess transcriptome changes due to lyophilization, we performed RNA-Seq from RNAs isolated from lyophilized samples stored for two weeks at room temperature. Overall, the generated sequencing libraries were shown to be highly comparable using multiple quality metrics, such as uniquely mapping reads, read duplication rates and GC fraction, library complexity, read coverage of genes, and read mapping to different RNA biotypes and chromosomes. There was no sign of base modifications affecting reliable read mapping. The 28 genes downsampled in lyophilized samples represent 0.4% of expressed genes, with a low median fold-difference. Of this gene set, lower abundance genes showed higher fold-differences, which may result from higher degradation rates or the combination of the higher stochasticity or measurement bias of low expression genes (Bray et al., 2010; Gallego Romero et al., 2014; Ibberson et al., 2009). Assessing transcript features of DEGs, we found that affected transcripts are generally longer, contain more G and C residues, and often encode transcription factors. In vivo, it was previously shown that longer transcripts (Feng & Niu, 2007) to degradation and transcription factor-encoding transcripts are less stable both *in vivo* and *in vitro* (Conesa et al., 2016; P. Li et al., 2019; Parekh et al., 2016; Risso et al., 2011). However, GC content at the third codon was found to be inversely correlated with degradation rates (Neymotin, Ettorre, & Gresham, 2016). These and the presence of ARE sequences in a few of DEG transcripts hint to the presence of residual, regulated decay mechanisms in lyophilized cells.

Regarding sample costs, Leboeuf *et al.* (Leboeuf et al., 2008)reported their annual costs of lyophilized vs. -80°C vs storage in liquid nitrogen, which were 3, 24 and 31 EUR, respectively, taking into account salaries, maintenance of freezers, CO_2 backup, air conditioning and consumables, without taking into account possibly decreased transport costs. Our lyophilization cost estimation, taking into account the energy consumption of the CoolSafe freeze dryer and the price of liquid nitrogen and trehalose, resulted in 0.87 USD per sample when only one sample is lyophilized per run. Given that longer periods may pass between collections of surgical and blood samples, to minimize lyophilization costs, an ideal procedure would involve the collection of samples in lyophilization solutions, followed by transient storage at ultra-low temperatures until a sufficient number of samples are gathered for a lyophilization run.

Overall, the findings of our study support the feasibility of lyophilization in trehalose and room temperature storage for human samples dedicated to RNA-based applications. Best practices and methods for problematic samples and whole-cell lyophilization for gene regulation-related studies are yet to be established.

6.3 Genotype-independent molecular phenotypic variability of the LCL model

It is beyond dispute that cell line models will remain the workhorse of biomedical research. Beyond the well-known advantages of cell lines over primary tissues, each has its set features rendering them suitable and adequate for addressing certain scientific hypotheses. LCLs, for instance, are models of the post-genomic era in the sense that thousands of cell lines and their genomic sequences are publicly available, providing a cost-effective model for studies on genotype-cellular phenotype interactions. With the combination of classical molecular biology techniques and emerging high throughput methods the scientific community has not only become able to explore novel aspects of cellular behaviour but has also got the opportunity to revisit commonly held assumptions about the most popular cell lines. The findings of cell line characterization studies may drive more informed experimental designs. Although LCLs are widely used for mapping molecular QTLs, it is lesser-known what fraction of certain phenotypic features vary irrespective of the genetic background, potentially hindering the identification of strong QTL candidates. In our study, we aimed at exploring, for the first time, cis-regulatory element- and transcriptome-level variability of LCLs using five genetically identical cell lines.

Studies using LCLs commonly refer to inter-cell line differences, rather unfoundedly, as inter-individual differences. A few studies from two groups of researchers reported genotype-dependent quantitative chromatin features, including coordinated changes in association with chromatin folding; however, their model of genetically distinct LCLs could not be used for discriminating between genotype-independent changes and genotype-dependent changes failing to be associated with variants that reach QTL significance threshold (Grubert et al., 2015; M. Kasowski et al., 2010; H. Wang et al., 2012; Waszak et al., 2015).

The five genetically stable isogenic cell lines prepared by the Pevsner Laboratory (Shirley et al., 2012) is currently the best available set of cells for modelling LCL variability emerging from non-genetic sources. The preparation of these isogenic cells resembled that of genetically distinct, commercially available LCLs, that is, they were prepared from five different blood batches (of the same individual), and were handled separately. Here we note that numerous laboratories prepare, culture, and make LCLs available, and in those cases, lab-to-lab differences may further contribute to variability. In order to minimize variability emerging during our research, we first ensured that all LCLs had the same number of freeze-thaw cycles and passages prior to initiation experiments by preparing a three-tiered biobank and handling all cell lines in parallel. We also harvested cells at the same time point of the day in order to minimize circadian effects and used biological replicates to exclude differences sue to random fluctuations. Altogether, this model reflects variability emerging during cell line generation and short-term culture.

Although the five isogenic LCLs were selected based on their reported genomic stability (Shirley et al., 2012), and proper cell line identification from the part of the vendor could be assumed, we checked and confirmed that all cell lines were derived from the same human male source, and that all cells were euploid. Besides that, at passage 14, all lines showed rosette morphology, no single cell line was characterized with the presence of other co-cultured leukocytes besides cells with B cell phenotype, all of which are consistent with previous reports on general LCL characteristics (Hussain, Kotnis, Sarin, & Mulherkar, 2012; Joesch-Cohen & Glusman, 2017). Similar cell cycle profiles indicated that cell growth-related pathways were not markedly overactive or suppressed as a response to EBV or positive selection in any of the cell lines.

Immunophenotyping not only revealed B cell phenotype but also that only one cell line showed evidence to monoclonality. Plagnol *et al.* suggested that the shrinkage of diversity mostly occurs at the early steps of cell line generation and, to a lesser extent, is affected by later culturing (Plagnol et al., 2008). We assume that polyclonal cells better mirror the heterogeneity of the original B cell pool, as monoclonalization leads to assaying the derivative cell of only one parental cell of the originally diverse cell population, and also that long-term culturing would have eventually led to monoclonalization. However, the monoclonal line did not show any outstanding features throughout our study, showing higher cis-element activity- and RNA-level similarity with a polyclonal LCL than other polyclonals with one another. Of note, the clonality assessment by kappa/lambda staining widely used by the clinics is not able to estimate the number of constituent clones.

Reproducible H3K27ac signatures were found to discriminate the isogenic cell lines. Strikingly, almost one-fourth of assayed regions (9,685 regulatory elements) were found significantly differentially acetylated at H3K27, between at least two cell lines, using two-fold difference as the cut-off. Intergenic enhancer regions showed the highest fraction of variable regions, in contrast to promoters, whose activity levels were comparably stable. Multiple lines of evidence suggest that promoters are more resistant to short-term and evolutionary-scale perturbations than enhancers (Frankel et al., 2010; Patten et al., 2018; Villar et al., 2015), and most GWAS variants have been shown to map tp intergenic enhancers (Corradin & Scacheri, 2014). The lack of robustness in promoter activity, which is directly linked to gene expression, maybe deleterious and is therefore selected against during evolution. The origin of promoter robustness can be partly explained by enhancer redundancy ("shadow enhancers") (Frankel et al., 2010; Hong, Hendrix, & Levine, 2008), that is, multiple enhancers loop to each promoter and once an enhancer switches off, the remaining active enhancers keep promoter activity and RNA expression stable. This might also explain our finding that the variability of individual elements of super-enhancers does not lead to marked gene expression differences. Given that promoters were shown to be less affected, we were not surprised to find relatively modest differences in the levels of poly(A)+ RNAs. This is consistent with findings in genetically distinct LCLs and yeast (M. Kasowski et al., 2010; Zheng, Zhao, Mancera, Steinmetz, & Snyder, 2010).

Due to single-cell studies, it has been increasingly acknowledged that individual cells of a tissue type can be highly heterogeneous in terms of their functional genomic features, originating from early developmental steps (Rotem et al., 2015). Phenotypic plasticity enabled by heterogeneity allows a more effective response to unpredictable external exposures, promoting survival of the individual, the colony, or the species (Bódi et al., 2017; Caza & Landas, 2015). We assume that the observed variability is the result of the B cell heterogeneity in the blood samples combined with selective EBV infection of a subset of clones, and probably also growth rate differences of descendant lineages. Hence, using bulk sequencing such as ChIP-Seq and mRNA-Seq, the signatures of a limited number of parental B cells will be represented as average signals. The domination of clones with high or low lymphokine secretion, leading to differences in the composition of the culturing media, may also shift the population's phenotype. Gene ontology analysis revealed the enrichment of immune-related genes and cell surface receptors, which aligns with the results of others reporting particularly high splicing variation of B cell-specific surface receptors (Byrne et al., 2017). Notably, LCL sub-populations were identified in each cell line when probing cell surface antigens by flow cytometry.

The finding that numerous pharmacogenes are differentially expressed suggests that our study has implications not only for molecular QTL but also LCL-based pharmacogenomic QTL screenings. Our experiment with the chemotherapeutic agent 5-FU showed correlation with *DPYD* expression. Of note, it has been proposed that LCL drug response is influenced by, besides growth rate, certain other factors such as EBV copy number and baseline ATP levels, which may show genetic heritability (Choy et al., 2008; Houldcroft et al., 2014; Stark et al., 2010). As we did not assess these factors, we cannot exclude their confounding effects. The limited number of isogenic cells limits our ability to extrapolate our findings to large panels of LCL. Hence studies on a large number of isogenic lines would be desirable. A recent study using MCF-7 cells from different vendors showed inter- and intra-cell line drug response differences, though many were the result of rapid genetic evolution (Ben-David et al., 2018). Therefore genetic diversification should not be overlooked during study design and data interpretation. Our study suggests that using the triangle study model (Huang et al., 2007), i.e. including the measurement of baseline RNA levels into pharmacogenomic study design, would be beneficial.

In conclusion, our study highlights the extent and nature of LCL variability at gene regulatory element and gene expression levels, showing implications in pharmacogenomic research. Despite the above findings. We believe that LCLs will remain a powerful model for QTLs, and uncovered limitations will serve more rational experimental design.

7 SUMMARY

High-throughput functional genomics methods, such as ChIP and RNA-Seq, have revolutionized research on gene regulation. We aimed at contributing to the rapidly developing field of functional genomics considering three aspects of biomedical research: the development of key methodologies, characterization of emerging model systems, and cost rationalization.

As the ChIP method lacks well-established procedure controls, hindering the assessment of experimental sample loss, we set out to develop spike-in procedure controls based on a novel concept, using peptide-displaying bacteriophages. We could enrich phages mimicking chromatin epitopes from peptide-displaying M13 bacteriophage libraries by *in vitro* evolution. The phage control particles spiked into chromatin samples bound to the ChIP-grade antibody with high affinity and could be quantified from the ChIP eluate by qPCR. Therefore the presented concept may serve as a basis for the generation of spike-in controls for various ChIP-grade antibodies.

In the past few years, RNA-Seq has proven to be instrumental in biomedical research. However, the high operational costs of frozen tissue storage urges the scientific community to develop methods allowing for room temperature storage. Therefore we assessed the utility of lyophilization as a potentially cost-effective cell preservation method for RNA-based downstream applications. While epigallocatechin-gallate was found not to be suitable as a cellular lyoprotectant for RNA-based studies, trehalose provided sufficient RNA protection during lyophilization and weeks of room temperature storage, resulting in high yields and excellent RNA quality for both low- and high-throughput RNA studies.

The epigenomic and transcriptomic variability intrinsic to human LCL cells, which are widely used for molecular and drug QTL mapping, has not been previously elucidated. Using five LCLs from the same individual we showed that almost one-fourth of active (H3K27ac-marked) gene regulatory elements were variably acetylated, coupled to a modest transcriptomic variability. Additionally, isogenic gene expression variability may affect chemotherapeutic drug response, as shown in the example of the *DPYD* gene and 5-fluorouracil. Therefore it is suggested to consider baseline RNA levels during LCL-based QTL research design.

In summary, our results provide a baseline for more cost-effective and rational experimental design in the framework of functional genomics.

ÖSSZEFOGLALÁS

A funkcionális genomikai módszerek, például a ChIP-Seq és az RNA-Seq forradalmasították a génszabályozással kapcsolatos vizsgálatokat. Kutatásaink során a funkcionális genomika eszköztárához kapcsolódó vizsgálatokat végeztünk a következő szempontokat szem előtt tartva: kulcsmódszerek fejlesztése, modellrendszerek jellemzése és költségracionalizálás.

A ChIP módszer nem rendelkezik általánosan elfogadott, a kísérletes mintaveszteség normalizálására használható kontrollokkal. Kísérleteink során M13 bakteriofág-alapú peptidkönyvtárból kiindulva sikeresen feldúsítottunk kromatin epitópokat utánzó bakteriofágokat *in vitro* evolúció segítségével. A kromatin mintákhoz kevert fág részecskék nagy affinitással kötődtek a ChIP-minőségű antitestekhez, és kvantitálhatóak voltak ChIP eluátumokból qPCR segítségével. Mindezek alapján a bemutatott módszer kiindulópontként szolgálhat spike-in kontrollok előállításához különböző ChIP-minőségű antitestekhez.

Az RNS-szekvenálás egyre jelentősebb szerepet játszik az orvosbiológiai kutatásokban, ám a fagyasztott szövetek tárolásból eredő magas működtetési költségek miatt a szobahőmérsékleten történő mintatárolásra alkalmas módszerek fejlesztése nagy jelentőséggel bírhat. Kísérleteink során sejtek liofilezéssel történő stabilizálásának hatását vizsgáltuk RNS-alapú mérésekre, mint potenciálisan költségcsökkentő alternatíva. Míg az epigallokatekin-gallát nem bizonyult alkalmas lioprotektánsnak RNS-alapú vizsgálatokhoz, a trehalóz megfelelő védelmet nyújtott liofilezés és többhetes szobahőn történő tárolás során, magas kitermelést és kiváló RNS-minőséget eredményezve mind alacsony-, mind magas áteresztőképességű vizsgálatokhoz.

A molekuláris- és gyógyszer-QTL-ek térképezésére gyakran használt humán LCL sejtvonalak intrinzik epigenomi és transzkriptóm-szintű variabilitása korábban ismeretlen volt. Öt, azonos személyből származó LCL sejtonal vizsgálata során kimutattuk, hogy az aktív (H3K27ac-jelölt) génszabályozó elemek közel negyede mutat variábilis acetilációt, amely enyhébb transzkriptóm-variabilitással párosult. Továbbá a génexpressziós eltérések kemoterápiás szerekkel szembeni eltérő sejtválaszt okozhatnak. Emiatt javasolt az RNS-szintek beépítése az LCL-alapú QTL-vizsgálati tervekbe.

Összegezve, eredményeink kiindulási pontként szolgálhatnak költséghatékonyabb és racionálisabb kísérlettervezéshez a funkcionális genomika keretein belül.

8 KEYWORDS

phage display, spike-in control, ChIP-Seq, lyophilization, freeze-drying, epigallocatechin-gallate, trehalose, RNA-Seq, B-lymphoblastoid cell line, epigenomics, transcriptomics, pharmacogenomics

KULCSSZAVAK

fágbemutatás, spike-in kontroll, ChIP-szekvenálás, liofilezés, fagyasztva szárítás, epigallokatekin-gallát, trehalóz, RNS-szekvenálás, B-limfoblasztoid sejtvonal, epigenomika, transzkriptomika, farmakogenomika

9 AUTHOR CONTRIBUTION

The textual content and Figures 2-29 presented as part of the Doctoral Thesis have been prepared solely by Lilla Ozgyin. Figure 1A and Figure 1B were adapted from original figures by Dr. Zsolt Keresztessy and Edina Erdős, respectively, with moderate modifications. From the co-authors of the publications related to the dissertation, Dr. Attila Horváth performed the initial analyses of ChIP-Seq and RNA-Seq data (read mapping and transcript abundance calculations), and 3C-Seq analysis. Dr. Zsuzsanna Hevessy was responsible for the evaluation of immunophenotyping, clonality assessment and cell cycle analyses.

10 REFERENCES

10.1 References related to the dissertation

Arav, A., & Natan, D. (2012). Freeze Drying of Red Blood Cells: The Use of Directional Freezing and a New Radio Frequency Lyophilization Device. *Biopreservation and Biobanking*, *10*(4), 386–394. https://doi.org/10.1089/bio.2012.0021

Augereau, C., Lemaigre, F. P., & Jacquemin, P. (2016). Extraction of high-quality RNA from pancreatic tissues for gene expression studies. *Analytical Biochemistry*, *500*, 60–62. https://doi.org/10.1016/j.ab.2016.02.008

Baechler, E. C., Batliwalla, F. M., Karypis, G., Gaffney, P. M., Moser, K., Ortmann, W. A., … Behrens, T. W. (2004). Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. *Genes & Immunity*, *5*(5), 347–353. https://doi.org/10.1038/sj.gene.6364098

Bakheet, T., Hitti, E., & Khabar, K. S. A. (2018). ARED-Plus: an updated and expanded database of AU-rich element-containing mRNAs and pre-mRNAs. *Nucleic Acids Research*, *46*(D1), D218–D220. https://doi.org/10.1093/nar/gkx975

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., … Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genetics*, *10*(9), e1004663. https://doi.org/10.1371/journal.pgen.1004663

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., … Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, *12*(1), R10. https://doi.org/10.1186/gb-2011-12-1-r10

Bell, L. N., Saxena, R., Mattar, S. G., You, J., Wang, M., & Chalasani, N. (2011). Utility of formalin-fixed, paraffin-embedded liver biopsy specimens for global proteomic analysis in nonalcoholic steatohepatitis. *PROTEOMICS - Clinical Applications*, *5*(7–8), 397–404. https://doi.org/10.1002/prca.201000144

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., … Golub, T. R. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, *560*(7718), 325–330. https://doi.org/10.1038/s41586-018-0409-3

Benaroudj, N., Lee, D. H., & Goldberg, A. L. (2001). Trehalose accumulation during cellular stress protects cells and cellular proteins from damage by oxygen radicals. *The Journal of Biological Chemistry*, *276*(26), 24261–24267. https://doi.org/10.1074/jbc.M101487200

Bhaduri-McIntosh, S., Rotenberg, M. J., Gardner, B., Robert, M., & Miller, G. (2008). Repertoire and frequency of immune cells reactive to Epstein-Barr virus-derived autologous lymphoblastoid cell lines. *Blood*, *111*(3), 1334–1343. https://doi.org/10.1182/blood-2007-07-101907

Bódi, Z., Farkas, Z., Nevozhay, D., Kalapis, D., Lázár, V., Csörgő, B., … Pál, C. (2017). Phenotypic heterogeneity promotes adaptive evolution. *PLOS Biology*, *15*(5), e2000644. https://doi.org/10.1371/journal.pbio.2000644

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., … Bounova, G. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Research*, *24*(7), 1157–1168. https://doi.org/10.1101/gr.168260.113

Branković, I., Malogajski, J., & Morré, S. A. (2014). Biobanking and translation of human genetics and genomics for infectious diseases. *Applied & Translational Genomics*, *3*(2), 30–35. https://doi.org/10.1016/j.atg.2014.04.001

Bray, S. E., Paulin, F. E. M., Fong, S. C., Baker, L., Carey, F. A., Levison, D. A., … Kernohan, N. M. (2010). Gene expression in colorectal neoplasia: modifications induced by tissue ischaemic time and tissue handling protocol. *Histopathology*, *56*(2), 240–250. https://doi.org/10.1111/j.1365-2559.2009.03470.x

Buchanan, S. S., Pyatt, D. W., & Carpenter, J. F. (2010a). Preservation of differentiation and clonogenic potential of human hematopoietic stem and progenitor cells during lyophilization and ambient storage. *PloS One*, *5*(9). https://doi.org/10.1371/journal.pone.0012518

Buchanan, S. S., Pyatt, D. W., & Carpenter, J. F. (2010b). Preservation of Differentiation and Clonogenic Potential of Human Hematopoietic Stem and Progenitor Cells during Lyophilization and Ambient Storage. *PLoS ONE*, *5*(9), e12518. https://doi.org/10.1371/journal.pone.0012518

Buckwalter, S. P., Sloan, L. M., Cunningham, S. A., Espy, M. J., Uhl, J. R., Jones, M. F., …

Wengenack, N. L. (2014). Inhibition controls for qualitative real-time PCR assays: are they necessary for all specimen matrices? *Journal of Clinical Microbiology*, *52*(6), 2139–2143. https://doi.org/10.1128/JCM.03389-13

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., … Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, *8*(1), 16027. https://doi.org/10.1038/ncomms16027

Caliskan, M., Cusanovich, D. A., Ober, C., & Gilad, Y. (2011). The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics*, *20*(8), 1643–1652. https://doi.org/10.1093/hmg/ddr041

Catts, V. S., Catts, S. V., Fernandez, H. R., Taylor, J. M., Coulson, E. J., & Lutze-Mann, L. H. (2005). A microarray study of post-mortem mRNA degradation in mouse brain tissue. *Molecular Brain Research*, *138*(2), 164–177. https://doi.org/10.1016/j.molbrainres.2005.04.017

Caza, T., & Landas, S. (2015). Functional and Phenotypic Plasticity of CD4 $^{+}$ T Cell Subsets. *BioMed Research International*, *2015*, 1–13. https://doi.org/10.1155/2015/521957

Chakrabarty, S., D’Souza, R. R., Kabekkodu, S. P., Gopinath, P. M., Rossignol, R., & Satyamoorthy, K. (2014). Upregulation of TFAM and mitochondria copy number in human lymphoblastoid cells. *Mitochondrion*, *15*, 52–58. https://doi.org/10.1016/j.mito.2014.01.002

Chen, C. Y., & Shyu, A. B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends in Biochemical Sciences*, *20*(11), 465–470. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8578590

Chen, Y., Chi, P., Rockowitz, S., Iaquinta, P. J., Shamu, T., Shukla, S., … Sawyers, C. L. (2013). ETS factors reprogram the androgen receptor cistrome and prime prostate tumorigenesis in response to PTEN loss. *Nature Medicine*, *19*(8), 1023. https://doi.org/10.1038/NM.3216

Choy, E., Yelensky, R., Bonakdar, S., Plenge, R. R. M., Saxena, R., De Jager, P. L., … Team, R. (2008). Genetic Analysis of Human Traits In Vitro: Drug Response and Gene Expression in Lymphoblastoid Cell Lines. *PLoS Genetics*, *4*(11), e1000287. https://doi.org/10.1371/journal.pgen.1000287

Chung, J.-Y., Braunschweig, T., Williams, R., Guerrero, N., Hoffmann, K. M., Kwon, M., …

Hewitt, S. M. (2008). Factors in tissue handling and processing that impact RNA obtained from formalin-fixed, paraffin-embedded tissue. *The Journal of Histochemistry and Cytochemistry : Official Journal of the Histochemistry Society*, *56*(11), 1033–1042. https://doi.org/10.1369/jhc.2008.951863

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8

Corradin, O., & Scacheri, P. C. (2014). Enhancer variants: evaluating functions in common disease. *Genome Medicine*, *6*(10), 85. https://doi.org/10.1186/s13073-014-0085-3

Crockett, D. K., Lin, Z., Vaughn, C. P., Lim, M. S., & Elenitoba-Johnson, K. S. J. (2005). Identification of proteins from formalin-fixed paraffin-embedded cells by LC-MS/MS. *Laboratory Investigation*, *85*(11), 1405–1415. https://doi.org/10.1038/labinvest.3700343

Damsteegt, E. L., McHugh, N., & Lokman, P. M. (2016). Storage by lyophilization – Resulting RNA quality is tissue dependent. *Analytical Biochemistry*, *511*, 92–96. https://doi.org/10.1016/j.ab.2016.08.005

Das, Z. C., Kumar Gupta, M., Uhm, S. J., & Lee, H. T. (2010). Lyophilized somatic cells direct embryonic development after whole cell intracytoplasmic injection into pig oocytes q. https://doi.org/10.1016/j.cryobiol.2010.07.007

Deng, Q.-W., Li, S., Wang, H., Sun, H.-L., Zuo, L., Gu, Z.-T., … Yan, F.-L. (2018). Differential long noncoding RNA expressions in peripheral blood mononuclear cells for detection of acute ischemic stroke. *Clinical Science*, *132*(14), 1597–1614. https://doi.org/10.1042/CS20180411

Deprés-Tremblay, G., Chevrier, A., Hurtig, M. B., Snow, M., Rodeo, S., & Buschmann, M. D. (2018). Freeze-Dried Chitosan-Platelet-Rich Plasma Implants for Rotator Cuff Tear Repair: Pilot Ovine Studies. *ACS Biomaterials Science & Engineering*, *4*(11), 3737–3746. https://doi.org/10.1021/acsbiomaterials.7b00354

Diasio, R. B., Beavers, T. L., & Carpenter, J. T. (1988). Familial deficiency of dihydropyrimidine dehydrogenase. Biochemical basis for familial pyrimidinemia and severe 5-fluorouracil-induced toxicity. *Journal of Clinical Investigation*, *81*(1), 47–51. https://doi.org/10.1172/JCI113308

Drakulovski, P., Locatelli, S., Butel, C., Pion, S., Krasteva, D., Mougdi-Pole, E., … Mallié, M. (2013). Use of RNAlater as a preservation method for parasitic coprology studies in wild-living chimpanzees. *Experimental Parasitology*, *135*(2), 257–261. https://doi.org/10.1016/j.exppara.2013.07.002

Egan, B., Yuan, C.-C., Craske, M. L., Labhart, P., Guler, G. D., Arnott, D., … Trojer, P. (2016). An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition. *PLOS ONE*, *11*(11), e0166438. https://doi.org/10.1371/journal.pone.0166438

Eroglu, A., Russo, M. J., Bieganski, R., Fowler, A., Cheley, S., Bayley, H., & Toner, M. (2000). Intracellular trehalose improves the survival of cryopreserved mammalian cells. *Nature Biotechnology*, *18*(2), 163–167. https://doi.org/10.1038/72608

Esteve-Codina, A., Arpi, O., Martinez-García, M., Pineda, E., Mallo, M., Gut, M., … Group, on behalf of the G. (2017). A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples. *PLOS ONE*, *12*(1), e0170632. https://doi.org/10.1371/journal.pone.0170632

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. https://doi.org/10.1038/nature13835

Feng, L., & Niu, D.-K. (2007). Relationship Between mRNA Stability and Length: An Old Question with a New Twist. *Biochemical Genetics*, *45*(1–2), 131–137. https://doi.org/10.1007/s10528-006-9059-5

Fleige, S., & Pfaffl, M. W. RNA integrity and the effect on the real-time qRT-PCR performance, 27 Molecular Aspects of Medicine § (2006). https://doi.org/10.1016/j.mam.2005.12.003

Florell, S. R., Coffin, C. M., Holden, J. A., Zimmermann, J. W., Gerwels, J. W., Summers, B. K., … Leachman, S. A. (2001). Preservation of RNA for Functional Genomic Studies: A Multidisciplinary Tumor Bank Protocol. *Modern Pathology*, *14*(2), 116–128. https://doi.org/10.1038/modpathol.3880267

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., & Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, *466*(7305), 490–493. https://doi.org/10.1038/nature09158

Fu, W.-M., Lu, Y.-F., Hu, B.-G., Liang, W.-C., Zhu, X., Yang, H., … Zhang, J.-F. (2016). Long noncoding RNA Hotair mediated angiogenesis in nasopharyngeal carcinoma by direct and indirect signaling pathways. *Oncotarget*, *7*(4), 4712–4723. https://doi.org/10.18632/oncotarget.6731

Gallego Romero, I., Pai, A. A., Tung, J., Gilad, Y., Romero, I. G., Pai, A. A., … Gilad, Y. (2014). RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biology*, *12*, 42. https://doi.org/10.1186/1741-7007-12-42

Gianaroli, L., Cristina Magli, M., Stanghellini, I., Crippa, A., Maria Crivello, A., Stefano Pescatori, E., & Pia Ferraretti, A. (2012). DNA integrity is maintained after freeze-drying of human spermatozoa. *Fertility and Sterility*, *97*, 1067–1073.e1. https://doi.org/10.1016/j.fertnstert.2012.02.014

Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., … Rasheed, A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Gilmour, D. S., & Lis, J. T. (1984). Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(14), 4275–4279. https://doi.org/10.1073/pnas.81.14.4275

Grafodatskaya, D., Choufani, S., Ferreira, J. C., Butcher, D. T., Lou, Y., Zhao, C., … Weksberg, R. (2010). EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. *Genomics*, *95*(2), 73–83. https://doi.org/10.1016/J.YGENO.2009.12.001

Grassi, M. A., Rao, V. R., Chen, S., Cao, D., Gao, X., Cleary, P. A., … Group, D. R. (2016). Lymphoblastoid Cell Lines as a Tool to Study Inter-Individual Differences in the Response to Glucose. *PLOS ONE*, *11*(8), e0160504. https://doi.org/10.1371/journal.pone.0160504

Graw, S., Meier, R., Minn, K., Bloomer, C., Godwin, A. K., Fridley, B., … Chien, J. (2015). Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Scientific Reports*, *5*, 12335. https://doi.org/10.1038/srep12335

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., … Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal

Chromosomal Interactions. *Cell*, *162*(5), 1051–1065. https://doi.org/10.1016/j.cell.2015.07.048

Grzybowski, A. T., Chen, Z., & Ruthenburg, A. J. (2015). Calibrating ChIP-Seq with Nucleosomal Internal Standards to Measure Histone Modification Density Genome Wide. *Molecular Cell*, *58*(5), 886–899. https://doi.org/10.1016/j.molcel.2015.04.022

Guertin, M. J., Cullen, A. E., Markowetz, F., & Holding, A. N. (2018). Parallel factor ChIP provides essential internal control for quantitative differential ChIP-seq. *Nucleic Acids Research*, (1). https://doi.org/10.1093/nar/gky252

Halperin, I., Wolfson, H., & Nussinov, R. (2003). SiteLight: binding-site prediction using phage display libraries. *Protein Science : A Publication of the Protein Society*, *12*(7), 1344–1359. https://doi.org/10.1110/ps.0237103

Hammers, C. M., & Stanley, J. R. (2014). Antibody phage display: technique and applications. *The Journal of Investigative Dermatology*, *134*(2), e17. https://doi.org/10.1038/jid.2013.521

Hansen, K. D., Sabunciyan, S., Langmead, B., Nagy, N., Curley, R., Klein, G., … Feinberg, A. P. (2014). Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Research*, *24*(2), 177–184. https://doi.org/10.1101/gr.157743.113

Hartford, C. M., Duan, S., Delaney, S. M., Mi, S., Kistner, E. O., Lamba, J. K., … Dolan, M. E. (2009). Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. *Blood*, *113*(10), 2145–2153. https://doi.org/10.1182/blood-2008-05-154302

Hastreiter, S., Skylaki, S., Loeffler, D., Reimann, A., Hilsenbeck, O., Hoppe, P. S., … Schroeder, T. (2018). Inductive and Selective Effects of GSK3 and MEK Inhibition on Nanog Heterogeneity in Embryonic Stem Cells. *Stem Cell Reports*, *11*(1), 58–69. https://doi.org/10.1016/J.STEMCR.2018.04.019

Hedegaard, J., Thorsen, K., Lund, M. K., Hein, A.-M. K., Hamilton-Dutoit, S. J., Vang, S., … Dyrskjøt, L. (2014). Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PloS One*, *9*(5), e98187. https://doi.org/10.1371/journal.pone.0098187

Herbeck, J. T., Gottlieb, G. S., Wong, K., Detels, R., Phair, J. P., Rinaldo, C. R., … Mullins, J. I.

(2009). Fidelity of SNP array genotyping using Epstein Barr virus-transformed B-lymphocyte cell lines: implications for genome-wide association studies. *PloS One*, *4*(9), e6915. https://doi.org/10.1371/journal.pone.0006915

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., … Young, R. A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, *155*(4), 934–947. https://doi.org/10.1016/j.cell.2013.09.053

Hong, J.-W., Hendrix, D. A., & Levine, M. S. (2008). Shadow Enhancers as a Source of Evolutionary Novelty. *Science*, *321*(5894), 1314–1314. https://doi.org/10.1126/science.1160631

Houldcroft, C. J., Petrova, V., Liu, J. Z., Frampton, D., Anderson, C. A., Gall, A., & Kellam, P. (2014). Host Genetic Variants and Gene Expression Patterns Associated with Epstein-Barr Virus Copy Number in Lymphoblastoid Cell Lines. *PLoS ONE*, *9*(10), e108384. https://doi.org/10.1371/journal.pone.0108384

Huang, R. S., Duan, S., Bleibel, W. K., Kistner, E. O., Zhang, W., Clark, T. A., … Dolan, M. E. (2007). A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(23), 9758–9763. https://doi.org/10.1073/pnas.0703736104

Hui-Yuen, J., McAllister, S., Koganti, S., Hill, E., & Bhaduri-McIntosh, S. (2011). Establishment of Epstein-Barr Virus Growth-transformed Lymphoblastoid Cell Lines. *Journal of Visualized Experiments*, (57). https://doi.org/10.3791/3321

Hussain, T., Kotnis, A., Sarin, R., & Mulherkar, R. (2012). Establishment & characterization of lymphoblastoid cell lines from patients with multiple primary neoplasms in the upper aero-digestive tract & healthy individuals. *The Indian Journal of Medical Research*, *135*(6), 820–829. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22825601

Hussain, T., & Mulherkar, R. (2012). Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair. *International Journal of Molecular and Cellular Medicine*, *1*(2), 75–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24551762

Ibberson, D., Benes, V., Muckenthaler, M. U., & Castoldi, M. (2009). RNA degradation compromises the reliability of microRNA expression profiling. *BMC Biotechnology*, *9*(1),

102. https://doi.org/10.1186/1472-6750-9-102

Jarvis, J. E., Ball, G., Rickinson, A. B., & Epstein, M. A. (1974). Cytogenetic studies on human lymphoblastoid cell lines from burkitt's lymphomas and other sources. *International Journal of Cancer*, *14*(6), 716–721. https://doi.org/10.1002/ijc.2910140604

Jeon, J.-P., Shim, S.-M., Nam, H.-Y., Baik, S.-Y., Kim, J.-W., & Han, B.-G. (2007). Copy number increase of 1p36.33 and mitochondrial genome amplification in Epstein–Barr virus-transformed lymphoblastoid cell lines. *Cancer Genetics and Cytogenetics*, *173*(2), 122–130. https://doi.org/10.1016/j.cancergencyto.2006.10.010

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., … Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, *21*(9), 1543–1551. https://doi.org/10.1101/gr.121095.111

Jiao, W., Chen, Y., Song, H., Li, D., Mei, H., Yang, F., … Tong, Q. (2018). HPSE enhancer RNA promotes cancer progression through driving chromatin looping and regulating hnRNPU/p300/EGR1/HPSE axis. *Oncogene*, *37*(20), 2728–2745. https://doi.org/10.1038/s41388-018-0128-0

Joesch-Cohen, L., & Glusman, G. (2017). Differences between the genomes of lymphoblastoid cell lines and blood-derived samples. *Advances in Genomics and Genetics*, *Volume 7*, 1–9. https://doi.org/10.2147/AGG.S128824

Karran, L., Jones, M., Morley, G., Van Noorden, S., Smith, P., Lampert, I., & Griffin, B. E. (1995). Expression of a B-cell marker, CD24, on nasopharyngeal carcinoma cells. *International Journal of Cancer*, *60*(4), 562–566. https://doi.org/10.1002/ijc.2910600422

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., … Snyder, M. (2010). Variation in Transcription Factor Binding Among Humans. *Science*, *328*(5975), 232–235. https://doi.org/10.1126/science.1183621

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., … Snyder, M. (2013). Extensive variation in chromatin states across humans. *Science (New York, N.Y.)*, *342*(6159), 750–752. https://doi.org/10.1126/science.1242510

Kim, H.-S., Quon, M. J., & Kim, J. (2014). New insights into the mechanisms of polyphenols beyond antioxidant properties; lessons from the green tea polyphenol, epigallocatechin 3-gallate. *Redox Biology*, *2*, 187–195. https://doi.org/10.1016/j.redox.2013.12.022

Kim, M., Rhee, J.-K., Choi, H., Kwon, A., Kim, J., Lee, G. D., … Kim, T.-M. (2017). Passage-dependent accumulation of somatic mutations in mesenchymal stromal cells during in vitro culture revealed by whole genome sequencing. *Scientific Reports, 7*(1), 14508. https://doi.org/10.1038/S41598-017-15155-5

Kruse, C. P. S., Basu, P., Luesse, D. R., & Wyatt, S. E. (2017). Transcriptome and proteome responses in RNAlater preserved tissue of Arabidopsis thaliana. *PLOS ONE, 12*(4), e0175943. https://doi.org/10.1371/journal.pone.0175943

Leboeuf, C., Ratajczak, P., Zhao, W.-L., François Plassa, L., Court, M., Pisonero, H., … Janin, A. (2008). Long-Term Preservation at Room Temperature of Freeze-Dried Human Tumor Samples Dedicated to Nucleic Acids Analyses. *Cell Preservation Technology, 6*(3), 191–198. https://doi.org/10.1089/cpt.2008.0003

Lestou, V. S., De Braekeleer, M., Strehl, S., Ott, G., Gadner, H., & Ambros, P. F. (1993). Non-random integration of epstein-barr virus in lymphoblastoid cell lines. *Genes, Chromosomes and Cancer, 8*(1), 38–48. https://doi.org/10.1002/gcc.2870080108

Li, M.-W., Willis, B. J., Griffey, S. M., Spearow, J. L., & Lloyd, K. C. K. (2009). Assessment of three generations of mice derived by ICSI using freeze-dried sperm. *Zygote*, *17*(3), 239–251. https://doi.org/10.1017/S0967199409005292

Li, P., Marshall, L., Oh, G., Jakubowski, J. L., Groot, D., He, Y., … Labrie, V. (2019). Epigenetic dysregulation of enhancers in neurons is associated with Alzheimer’s disease pathology and cognitive symptoms. *Nature Communications*, *10*(1), 2246. https://doi.org/10.1038/s41467-019-10101-7

Li, S., Hattori, T., & Kodama, E. N. (2011). Epigallocatechin Gallate Inhibits the HIV Reverse Transcription Step. *Antiviral Chemistry and Chemotherapy*, *21*(6), 239–243. https://doi.org/10.3851/IMP1774

Liu, J.-L., Kusakabe, H., Chang, C.-C., Suzuki, H., Schmidt, D. W., Julian, M., … Yang, X. (2004). Freeze-Dried Sperm Fertilization Leads to Full-Term Development in Rabbits1. *Biology of Reproduction*, *70*(6), 1776–1781. https://doi.org/10.1095/biolreprod.103.025957

Loi, P., Matsukawa, K., Ptak, G., Clinton, M., Fulka, J., Nathan, Y., … Arav, A. (2008a). Freeze-dried somatic cells direct embryonic development after nuclear transfer. *PloS One*, *3*(8), e2978. https://doi.org/10.1371/journal.pone.0002978

Loi, P., Matsukawa, K., Ptak, G., Clinton, M., Fulka, J., Nathan, Y., … Arav, A. (2008b). Freeze-dried somatic cells direct embryonic development after nuclear transfer. *PloS One*, *3*(8), e2978. https://doi.org/10.1371/journal.pone.0002978

Londin, E. R., Keller, M. A., D'Andrea, M. R., Delgrosso, K., Ertel, A., Surrey, S., & Fortina, P. (2011). Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics*, *12*, 464. https://doi.org/10.1186/1471-2164-12-464

Long, H. M., Haigh, T. A., Gudgeon, N. H., Leen, A. M., Tsang, C.-W., Brooks, J., … Taylor, G. S. (2005). CD4+ T-cell responses to Epstein-Barr virus (EBV) latent-cycle antigens and the recognition of EBV-transformed lymphoblastoid cell lines. *Journal of Virology*, *79*(8), 4896–4907. https://doi.org/10.1128/JVI.79.8.4896-4907.2005

Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., … Young, R. A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, *153*(2), 320–334. https://doi.org/10.1016/j.cell.2013.03.036

Ma, J., Lin, Y., Zhan, M., Mann, D. L., Stass, S. A., & Jiang, F. (2015). Differential miRNA expressions in peripheral blood mononuclear cells for diagnosis of lung cancer. *Laboratory Investigation*, *95*(10), 1197–1206. https://doi.org/10.1038/labinvest.2015.88

Mangravite, L. M., Medina, M. W., Cui, J., Pressman, S., Smith, J. D., Rieder, M. J., … Krauss, R. M. (2010). Combined influence of LDLR and HMGCR sequence variation on lipid-lowering response to simvastatin. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *30*(7), 1485–1492. https://doi.org/10.1161/ATVBAHA.110.203273

Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., … Look, A. T. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (New York, N.Y.)*, *346*(6215), 1373–1377. https://doi.org/10.1126/science.1259037

Mareninov, S., De Jesus, J., Sanchez, D. E., Kay, A. B., Wilson, R. W., Babic, I., … Yong, W. H. (2013). Lyophilized brain tumor specimens can be used for histologic, nucleic acid, and protein analyses after 1 year of room temperature storage. *Journal of Neuro-Oncology*, *113*(3), 365–373. https://doi.org/10.1007/s11060-013-1135-1

Masuda, N., Ohnishi, T., Kawamoto, S., Monden, M., & Okubo, K. (1999). Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Research*, *27*(22), 4436–4443. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC148727/pdf/274436.pdf

Matsumoto, T., Sakari, M., Okada, M., Yokoyama, A., Takahashi, S., Kouzmenko, A., & Kato, S. (2013). The Androgen Receptor in Health and Disease. *Annual Review of Physiology*, *75*(1), 201–224. https://doi.org/10.1146/annurev-physiol-030212-183656

Matsuo, S., Sugiyama, T., Okuyama, T., Yoshikawa, K., Honda, K., Takahashi, R., & Maeda, S. (1999). Preservation of pathological tissue specimens by freeze-drying for immunohistochemical staining and various molecular biological analyses. *Pathology International*, *49*(5), 383–390. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10417680

Matsuo, S., Toyokuni, S., Osaka, M., Hamazaki, S., & Sugiyama, T. (1995). Degradation of DNA in Dried Tissues by Atmospheric Oxygen. *Biochemical and Biophysical Research Communications*, *208*(3), 1021–1027. https://doi.org/10.1006/BBRC.1995.1436

Mayne, M., Shepel, P. N., & Geiger, J. D. (1999). Recovery of high-integrity mRNA from brains of rats killed by high-energy focused microwave irradiation. *Brain Research. Brain Research Protocols*, *4*(3), 295–302. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10592338

Mazzei, F., Guarrera, S., Allione, A., Simonelli, V., Narciso, L., Barone, F., … Dogliotti, E. (2011). 8-Oxoguanine DNA-glycosylase repair activity and expression: A comparison between cryopreserved isolated lymphocytes and EBV-derived lymphoblastoid cell lines. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *718*(1–2), 62–67. https://doi.org/10.1016/j.mrgentox.2010.10.004

McCarthy, N. S., Allan, S. M., Chandler, D., Jablensky, A., & Morar, B. (2016). Integrity of genome-wide genotype data from low passage lymphoblastoid cell lines. *Genomics Data*, *9*, 18–21. https://doi.org/10.1016/j.gdata.2016.05.006

Metzger, E., Wissmann, M., Yin, N., Müller, J. M., Schneider, R., Peters, A. H. F. M., … Schüle, R. (2005). LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature*, *437*(7057), 436–439. https://doi.org/10.1038/nature04020

Mikkola, S., Lönnberg, T., & Lönnberg, H. (2018). Phosphodiester models for cleavage of nucleic acids. *Beilstein Journal of Organic Chemistry*, *14*(1), 803–837. https://doi.org/10.3762/bjoc.14.68

Milanesi, E., Voinsky, I., Hadar, A., Srouji, A., Maj, C., Shekhtman, T., … Gurwitz, D. (2017). RNA sequencing of bipolar disorder lymphoblastoid cell lines implicates the neurotrophic factor HRP-3 in lithium's clinical efficacy. *The World Journal of Biological Psychiatry*, 1–13. https://doi.org/10.1080/15622975.2017.1372629

Mohyuddin, A., Ayub, Q., Siddiqi, S., Carvalho-Silva, D. R., Mazhar, K., Rehman, S., … Qasim Mehdi, S. (2004). Genetic instability in EBV-transformed lymphoblastoid cell lines. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1670*(1), 81–83. https://doi.org/10.1016/J.BBAGEN.2003.10.014

Monaco, L., Crimi, M., & Wang, C. M. (2014). The challenge for a European network of biobanks for rare diseases taken up by RD-Connect. *Pathobiology : Journal of Immunopathology, Molecular and Cellular Biology*, *81*(5–6), 231–236. https://doi.org/10.1159/000358492

Morag, A., Pasmanik-Chor, M., Oron-Karni, V., Rehavi, M., Stingl, J. C., & Gurwitz, D. (2011). Genome-wide expression profiling of human lymphoblastoid cell lines identifies *CHL1* as a putative SSRI antidepressant response biomarker. *Pharmacogenomics*, *12*(2), 171–184. https://doi.org/10.2217/pgs.10.185

Morbach, H., Eichhorn, E. M., Liese, J. G., & Girschick, H. J. (2010). Reference values for B cell subpopulations from infancy to adulthood. *Clinical and Experimental Immunology*, *162*(2), 271–279. https://doi.org/10.1111/j.1365-2249.2010.04206.x

Moreira, G. M. S. G., Führer, V., & Hust, M. (2018). Epitope Mapping by Phage Display. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1701, pp. 497–518). https://doi.org/10.1007/978-1-4939-7447-4_28

Mouriaux, F., Zaniolo, K., Bergeron, M.-A., Weidmann, C., De La Fouchardière, A., Fournier, F., … Guérin, S. L. (2016). Effects of Long-term Serial Passaging on the Characteristics and Properties of Cell Lines Derived From Uveal Melanoma Primary Tumors. *Investigative Opthalmology & Visual Science*, *57*(13), 5288. https://doi.org/10.1167/iovs.16-19317

Munro, S. A., Lund, S. P., Pine, P. S., Binder, H., Clevert, D.-A., Conesa, A., … Salit, M. (2014). Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nature Communications*, *5*(1), 5125. https://doi.org/10.1038/ncomms6125

Mutter, G. L., Zahrieh, D., Liu, C., Neuberg, D., Finkelstein, D., Baker, H. E., & Warrington, J. A. (2004). Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. *BMC Genomics*, *5*(1), 88. https://doi.org/10.1186/1471-2164-5-88

Natan, D., Nagler, A., & Arav, A. (2009). Freeze-drying of mononuclear cells derived from umbilical cord blood followed by colony formation. *PloS One*, *4*(4), e5240. https://doi.org/10.1371/journal.pone.0005240

Neitzel, H. (1986). A routine method for the establishment of permanent growing lymphoblastoid cell lines. *Human Genetics*, *73*(4), 320–326. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3017841

Neymotin, B., Ettorre, V., & Gresham, D. (2016). Multiple Transcript Properties Related to Translation Affect mRNA Degradation Rates in Saccharomyces cerevisiae. *G3 (Bethesda, Md.)*, *6*(11), 3475–3483. https://doi.org/10.1534/g3.116.032276

Nickles, D., Madireddy, L., Yang, S., Khankhanian, P., Lincoln, S., Hauser, S. L., … Gentleman, R. (2012). In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing. *BMC Genomics*, *13*(1), 477. https://doi.org/10.1186/1471-2164-13-477

Noda, C., He, J., Takano, T., Tanaka, C., Kondo, T., Tohyama, K., … Tohyama, Y. (2007). Induction of apoptosis by epigallocatechin-3-gallate in human lymphoblastoid B cells. *Biochemical and Biophysical Research Communications*, *362*(4), 951–957. https://doi.org/10.1016/j.bbrc.2007.08.079

O'donahue, M., Johnson, L., Hedley, B., & Vaughan, E. (2018). *Title: Flow Cytometric Testing for Kappa and Lambda light chains Sponsored and reviewed by ICCS Quality and Standards Committee*. Retrieved from https://www.cytometry.org/web/modules/Module 6.pdf

O'Donnell, P. H., Gamazon, E., Zhang, W., Stark, A. L., Kistner-Griffin, E. O., Stephanie Huang, R., & Eileen Dolan, M. (2010). Population differences in platinum toxicity as a means to identify novel genetic susceptibility variants. *Pharmacogenetics and Genomics*, *20*(5), 327–

337. https://doi.org/10.1097/FPC.0b013e3283396c4e

Odhams, C. A., Cortini, A., Chen, L., Roberts, A. L., Vi ~ Nuela, A., Buil, A., … Cunninghame Graham, D. S. (2017). Mapping eQTLs with RNA-seq reveals novel suscepti-bility genes, non-coding RNAs and alternative-splicing events in systemic lupus erythematosus. *Human Molecular Genetics*, *26*(5), 1003–1017. https://doi.org/10.1093/hmg/ddw417

Oh, J. H., Kim, Y. J., Moon, S., Nam, H.-Y., Jeon, J.-P., Ho Lee, J., … Cho, Y. S. (2013). Genotype instability during long-term subculture of lymphoblastoid cell lines. *Journal of Human Genetics*, *58*(1), 16–20. https://doi.org/10.1038/jhg.2012.123

Orlando, D. A., Chen, M. W., Brown, V. E., Solanki, S., Choi, Y. J., Olson, E. R., … Guenther, M. G. (2014). Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. *Cell Reports*, *9*(3), 1163–1170. https://doi.org/10.1016/j.celrep.2014.10.018

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., … Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature, 554*(7691), 239–243. https://doi.org/10.1038/nature25461

Pansarasa, O., Bordoni, M., Drufuca, L., Diamanti, L., Sproviero, D., Trotti, R., … Cereda, C. (2018). Lymphoblastoid cell lines as a model to understand amyotrophic lateral sclerosis disease mechanisms. *Disease Models & Mechanisms*, *11*(3), dmm031625. https://doi.org/10.1242/dmm.031625

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq, *6*(1), 25533. https://doi.org/10.1038/srep25533

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., … Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, *27*(8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370

Passow, C. N., Kono, T. J. Y., Stahl, B. A., Jaggard, J. B., Keene, A. C., & McGaugh, S. E. (2019). Nonrandom RNAseq gene expression associated with RNAlater and flash freezing storage methods. *Molecular Ecology Resources*, *19*(2), 456–464. https://doi.org/10.1111/1755-0998.12965

Patten, D. K., Corleone, G., Győrffy, B., Perone, Y., Slaven, N., Barozzi, I., … Magnani, L. (2018). Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. *Nature Medicine*, *24*(9), 1469–1480. https://doi.org/10.1038/s41591-018-0091-x

Paul, A.-L., Levine, H. G., McLamb, W., Norwood, K. L., Reed, D., Stutte, G. W., … Ferl, R. J. (2005). Plant molecular biology in the space station era: utilization of KSC fixation tubes with RNAlater. *Acta Astronautica*, *56*(6), 623–628. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15736319

Plagnol, V., Uz, E., Wallace, C., Stevens, H., Clayton, D., Ozcelik, T., & Todd, J. A. (2008). Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PloS One*, *3*(8), e2966. https://doi.org/10.1371/journal.pone.0002966

Pokrovskaja, K., Ehlin-Henriksson, B., Kiss, C., Challa, A., Gordon, J., Gogolak, P., … Szekely, L. (2002). CD40 ligation downregulates EBNA-2 and LMP-1 expression in EBV-transformed lymphoblastoid cell lines. *International Journal of Cancer*, *99*(5), 705–712. https://doi.org/10.1002/ijc.10417

Prensner, J. R., Chen, W., Iyer, M. K., Cao, Q., Ma, T., Han, S., … Feng, F. Y. (2014). PCAT-1, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer. *Cancer Research*, *74*(6), 1651–1660. https://doi.org/10.1158/0008-5472.CAN-13-3159

Puhlev, I., Guo, N., Brown, D. R., & Levine, F. (2001). Desiccation Tolerance in Human Cells. *Cryobiology*, *42*(3), 207–217. https://doi.org/10.1006/cryo.2001.2324

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., … Regev, A. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology*, *29*(5), 436–442. https://doi.org/10.1038/nbt.1861

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Reddy, T. E., Gertz, J., Pauli, F., Kucera, K. S., Varley, K. E., Newberry, K. M., … Myers, R. M. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research*, *22*(5), 860–869.

https://doi.org/10.1101/gr.131201.111

Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, *12*, 480. https://doi.org/10.1186/1471-2105-12-480

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, *33*(11), 1165–1172. https://doi.org/10.1038/nbt.3383

Saito, M. A., Bulygin, V. V., Moran, D. M., Taylor, C., & Scholin, C. (2011). Examination of Microbial Proteome Preservation Techniques Applicable to Autonomous Environmental Sample Collection. *Frontiers in Microbiology*, *2*, 215. https://doi.org/10.3389/fmicb.2011.00215

Santiago, J. A., Bottero, V., & Potashkin, J. A. (2018). Evaluation of RNA Blood Biomarkers in the Parkinson's Disease Biomarkers Program. *Frontiers in Aging Neuroscience*, *10*. https://doi.org/10.3389/FNAGI.2018.00157

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., … Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342. https://doi.org/10.1038/nature10098

Scicchitano, M. S., Dalmas, D. A., Boyce, R. W., Thomas, H. C., & Frazier, K. S. (2009). Protein Extraction of Formalin-fixed, Paraffin-embedded Tissue Enables Robust Proteomic Profiles by Mass Spectrometry. *Journal of Histochemistry & Cytochemistry*, *57*(9), 849–860. https://doi.org/10.1369/jhc.2009.953497

Sharova, L. V, Sharov, A. A., Nedorezov, T., Piao, Y., Shaik, N., & Ko, M. S. H. (2009). Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*, *16*(1), 45–58. https://doi.org/10.1093/dnares/dsn030

Shiga, Y., Kubota, G., Orita, S., Inage, K., Kamoda, H., Yamashita, M., … Ohtori, S. (2017). Freeze-Dried Human Platelet-Rich Plasma Retains Activation and Growth Factor Expression after an Eight-Week Preservation Period. *Asian Spine Journal*, *11*(3), 329. https://doi.org/10.4184/ASJ.2017.11.3.329

Shiga, Y., Orita, S., Kubota, G., Kamoda, H., Yamashita, M., Matsuura, Y., … Ohtori, S. (2016).

Freeze-Dried Platelet-Rich Plasma Accelerates Bone Union with Adequate Rigidity in Posterolateral Lumbar Fusion Surgery Model in Rats. *Scientific Reports*, *6*(1), 36715. https://doi.org/10.1038/srep36715

Shirley, M. D., Baugher, J. D., Stevens, E. L., Tang, Z., Gerry, N., Beiswanger, C. M., … Pevsner, J. (2012). Chromosomal variation in lymphoblastoid cell lines. *Human Mutation*, *33*(7), 1075–1086. https://doi.org/10.1002/humu.22062

Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science (New York, N.Y.)*, *228*(4705), 1315–1317. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4001944

Spetzler, D., Pawlowski, T. L., Tinder, T., Kimbrough, J., Deng, T., Kim, J., … Kuslich, C. (2010). The molecular evolution of prostate cancer cell line exosomes with passage number. *Journal of Clinical Oncology*, *28*(15_suppl), e21071–e21071. https://doi.org/10.1200/jco.2010.28.15_suppl.e21071

Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., … Soler, E. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature Protocols*, *8*(3), 509–524. https://doi.org/10.1038/nprot.2013.018

Stark, A. L., Zhang, W., Mi, S., Duan, S., O’Donnell, P. H., Huang, R. S., & Dolan, M. E. (2010). Heritable and non-genetic factors as variables of pharmacologic phenotypes in lymphoblastoid cell lines. *The Pharmacogenomics Journal*, *10*(6), 505–512. https://doi.org/10.1038/tpj.2010.3

Styles, C. T., Bazot, Q., Parker, G. A., White, R. E., Paschos, K., & Allday, M. J. (2017). EBV epigenetically suppresses the B cell-to-plasma cell differentiation pathway while establishing long-term latency. *PLOS Biology*, *15*(8), e2001992. https://doi.org/10.1371/journal.pbio.2001992

Sugawara, H., Iwamoto, K., Bundo, M., Ueda, J., Ishigooka, J., & Kato, T. (2011). Comprehensive DNA methylation analysis of human peripheral blood leukocytes and lymphoblastoid cell lines. *Epigenetics*, *6*(4), 508–515. https://doi.org/10.4161/EPI.6.4.14876

Sun, Y. V, Turner, S. T., Smith, J. A., Hammond, P. I., Lazarus, A., Van De Rostyne, J. L., … Kardia, S. L. R. (2010). Comparison of the DNA methylation profiles of human peripheral blood cells and transformed B-lymphocytes. *Human Genetics*, *127*(6), 651–658. https://doi.org/10.1007/s00439-010-0810-y

Takahashi, R., Matsuo, S., Okuyama, T., & Sugiyama, T. (1995). Degradation of Macromolecules during Preservation of Lyophilized Pathological Tissues. *Pathology - Research and Practice*, *191*(5), 420–426. https://doi.org/10.1016/S0344-0338(11)80729-6

Takakuwa, T., Luo, W.-J., Ham, M. F., Sakane-Ishikawa, F., Wada, N., & Aozasa, K. (2004). Integration of Epstein-Barr virus into chromosome 6q15 of Burkitt lymphoma cell line (Raji) induces loss of BACH2 expression. *The American Journal of Pathology*, *164*(3), 967–974. https://doi.org/10.1016/S0002-9440(10)63184-7

Taniguchi, I., Iwaya, C., Ohnaka, K., Shibata, H., & Yamamoto, K. (2017). Genome-wide DNA methylation analysis reveals hypomethylation in the low-CpG promoter regions in lymphoblastoid cell lines. *Human Genomics*, *11*(1), 8. https://doi.org/10.1186/s40246-017-0106-6

Thorp, H. H. (2000). The importance of being r: greater oxidative stability of RNA compared with DNA. *Chemistry & Biology*, *7*(2), R33-6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10662699

Vermeulen, J., De Preter, K., Lefever, S., Nuytens, J., De Vloed, F., Derveaux, S., … Vandesompele, J. (2011). Measurable impact of RNA quality on gene expression results from quantitative PCR. *Nucleic Acids Research*, *39*(9), e63. https://doi.org/10.1093/nar/gkr065

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., … Odom, D. T. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, *160*(3), 554–566. https://doi.org/10.1016/j.cell.2015.01.006

Vodnik, M., Zager, U., Strukelj, B., & Lunder, M. (2011). Phage Display: Selecting Straws Instead of a Needle from a Haystack. *Molecules*, *16*(1), 790–817. https://doi.org/10.3390/molecules16010790

von Ahlfen, S., Missel, A., Bendrat, K., & Schlumpberger, M. (2007). Determinants of RNA quality from FFPE samples. *PloS One*, *2*(12), e1261. https://doi.org/10.1371/journal.pone.0001261

Vora, T., & Thacker, N. (2015). Impacts of a biobank: Bridging the gap in translational cancer medicine. *Indian Journal of Medical and Paediatric Oncology*, *36*(1), 17. https://doi.org/10.4103/0971-5851.151773

Wakayama, T., & Yanagimachi, R. (1998). Development of normal mice from oocytes injected with freeze-dried spermatozoa. *Nature Biotechnology*, *16*(7), 639–641. https://doi.org/10.1038/nbt0798-639

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., … Stamatoyannopoulos, J. A. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research*, *22*(9), 1680–1688. https://doi.org/10.1101/gr.136101.111

Wang, J., Pae, M., Meydani, S. N., & Wu, D. (2013). Green tea epigallocatechin-3-gallate modulates differentiation of naïve CD4+ T cells into specific lineage effector cells. *Journal of Molecular Medicine*, *91*(4), 485–495. https://doi.org/10.1007/s00109-012-0964-2

Wang, M., Ji, X., Wang, B., Li, Q., & Zhou, J. (2018). Simultaneous Evaluation of the Preservative Effect of RNAlater on Different Tissues by Biomolecular and Histological Analysis. *Biopreservation and Biobanking*, *16*(6), 426–433. https://doi.org/10.1089/bio.2018.0055

Waszak, S. M., Delaneau, O., Gschwind, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., … Dermitzakis, E. T. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, *162*(5), 1039–1050. https://doi.org/10.1016/j.cell.2015.08.001

Watanabe, M. (2006). Anhydrobiosis in invertebrates. *Applied Entomology and Zoology*, *41*(1), 15–31. https://doi.org/10.1303/aez.2006.15

Weisberg, E. P., Giorda, R., Trucco, M., & Lampasona, V. (1993). Lyophilization as a method to store samples of whole blood. *BioTechniques*, *15*(1), 64–68. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8363839

Wheeler, H. E., & Dolan, M. E. (2012). Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. *Pharmacogenomics*, *13*(1), 55–70. https://doi.org/10.2217/pgs.11.121

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., … Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. https://doi.org/10.1016/j.cell.2013.03.035

Wolkers, W. F., Walker, N. J., Tablin, F., & Crowe, J. H. (2001). Human Platelets Loaded with

Trehalose Survive Freeze-Drying. *Cryobiology*, *42*(2), 79–87. https://doi.org/10.1006/cryo.2001.2306

Wolkers, W. F., Walker, N. J., Tamari, Y., Tablin, F., & Crowe, J. H. (2002). Towards a Clinical Application of Freeze-Dried Human Platelets. *Cell Preservation Technology*, *1*(3), 175–188. https://doi.org/10.1089/153834402765035617

Wu, Y., Wu, M., Zhang, Y., Li, W., Gao, Y., Li, Z., … Zhang, C. (2012). Lyophilization is suitable for storage and shipment of fresh tissue samples without altering RNA and protein levels stored at room temperature. *Amino Acids*, *43*(3), 1383–1388. https://doi.org/10.1007/s00726-011-1212-8

Xiao, K., Yu, Z., Li, X., Li, X., Tang, K., Tu, C., … Xiong, W. (2016). Genome-wide Analysis of Epstein-Barr Virus (EBV) Integration and Strain in C666-1 and Raji Cells. *Journal of Cancer*, *7*(2), 214–224. https://doi.org/10.7150/jca.13150

Yang, C. S., Chen, L., Lee, M. J., Balentine, D., Kuo, M. C., & Schantz, S. P. (1998). Blood and urine levels of tea catechins after ingestion of different amounts of green tea by human volunteers. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *7*(4), 351–354. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9568793

Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M., & Darnell, J. E. (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Research*, *13*(8), 1863–1872. https://doi.org/10.1101/gr.1272403

Yang, M.-H., Hu, Z.-Y., Xu, C., Xie, L.-Y., Wang, X.-Y., Chen, S.-Y., & Li, Z.-G. (2015). MALAT1 promotes colorectal cancer cell proliferation/migration/invasion via PRKA kinase anchor protein 9. *Biochimica et Biophysica Acta*, *1852*(1), 166–174. https://doi.org/10.1016/j.bbadis.2014.11.013

Yao, Y., & Dai, W. (2014). Genomic Instability and Cancer. *Journal of Carcinogenesis & Mutagenesis*, *5*. https://doi.org/10.4172/2157-2518.1000165

Zatloukal, K., & Hainaut, P. (2010). Human tissue biobanks as instruments for drug discovery and development: impact on personalized medicine. *Biomarkers in Medicine*, *4*(6), 895–903.

https://doi.org/10.2217/bmm.10.104

Žegura, B., Volčič, M., Lah, T. T., & Filipič, M. (2008). Different sensitivities of human colon adenocarcinoma (CaCo-2), astrocytoma (IPDDC-A2) and lymphoblastoid (NCNC) cell lines to microcystin-LR induced reactive oxygen species and DNA damage. *Toxicon*, *52*(3), 518–525. https://doi.org/10.1016/j.toxicon.2008.06.026

Zhang, F., Ding, L., Cui, L., Barber, R., & Deng, B. (2019). Identification of long non-coding RNA-related and –coexpressed mRNA biomarkers for hepatocellular carcinoma. *BMC Medical Genomics*, *12*(S1), 25. https://doi.org/10.1186/s12920-019-0472-0

Zhang, M., Oldenhof, H., Sydykov, B., Bigalk, J., Sieme, H., & Wolkers, W. F. (2017). Freeze-drying of mammalian cells using trehalose: preservation of DNA integrity. *Scientific Reports*, *7*(1), 6198. https://doi.org/10.1038/s41598-017-06542-z

Zhang, S., Qian, H., Wang, Z., Fan, J., Zhou, Q., Chen, G., … Sun, J. (2010). Preliminary study on the freeze-drying of human bone marrow-derived mesenchymal stem cells. *Journal of Zhejiang University. Science. B*, *11*(11), 889–894. https://doi.org/10.1631/jzus.B1000184

Zhao, B., Zou, J., Wang, H., Johannsen, E., Peng, C., Quackenbush, J., … Kieff, E. (2011). Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(36), 14902–14907. https://doi.org/10.1073/pnas.1108892108

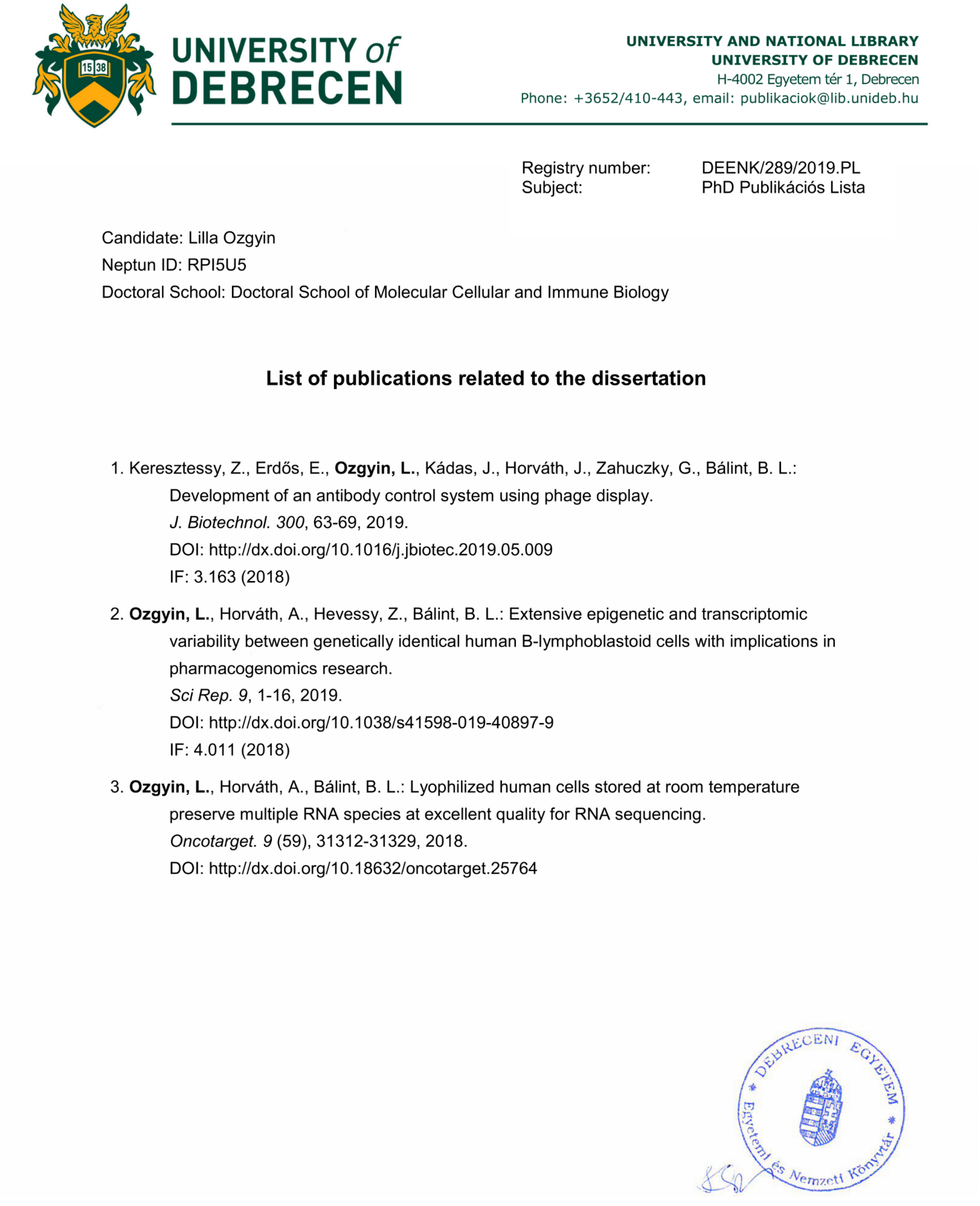
Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M., & Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. *Nature*, *464*(7292), 1187–1191. https://doi.org/10.1038/nature08934

Zhou, H., Schmidt, S. C. S., Kieff, E., Zhao Correspondence, B., Jiang, S., Willox, B., … Zhao, B. (2015). Epstein-Barr Virus Oncoprotein Super-enhancers Control B Cell Growth. *Cell Host and Microbe*, *17*, 205–216. https://doi.org/10.1016/j.chom.2014.12.013

Zhou, X.-L., Zhu, H., Zhang, S.-Z., Zhu, F.-M., Chen, G.-M., & Yan, L.-X. (2007). Freeze-drying of human platelets: influence of saccharide, freezing rate and cell concentration. *Cryo Letters*, *28*(3), 187–196. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17898906

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*, 160025. https://doi.org/10.1038/sdata.2016.25

10.2 Publication list certified by the University and National Library



UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number:	DEENK/289/2019.PL
Subject:	PhD Publikációs Lista

Candidate: Lilla Ozgyin

Neptun ID: RPI5U5

Doctoral School: Doctoral School of Molecular Cellular and Immune Biology

List of publications related to the dissertation

1. Keresztessy, Z., Erdős, E., **Ozgyin, L.**, Kádas, J., Horváth, J., Zahuczky, G., Bálint, B. L.: Development of an antibody control system using phage display.
   J. Biotechnol. 300, 63-69, 2019.
   DOI: http://dx.doi.org/10.1016/j.jbiotec.2019.05.009
   IF: 3.163 (2018)

2. **Ozgyin, L.**, Horváth, A., Hevessy, Z., Bálint, B. L.: Extensive epigenetic and transcriptomic variability between genetically identical human B-lymphoblastoid cells with implications in pharmacogenomics research.
   Sci Rep. 9, 1-16, 2019.
   DOI: http://dx.doi.org/10.1038/s41598-019-40897-9
   IF: 4.011 (2018)

3. **Ozgyin, L.**, Horváth, A., Bálint, B. L.: Lyophilized human cells stored at room temperature preserve multiple RNA species at excellent quality for RNA sequencing.
   Oncotarget. 9 (59), 31312-31329, 2018.
   DOI: http://dx.doi.org/10.18632/oncotarget.25764

List of other publications

4. Horváth, A., Dániel, B., Széles, L., Cuaranta-Monroy, I., Czimmerer, Z., **Ozgyin, L.**, Steiner, L., Kiss, M., Simándi, Z., Póliska, S., Giannakis, N., Raineri, E., Gut, I. G., Nagy, B., Nagy, L.: Labelled regulatory elements are pervasive features of the macrophage genome and are dynamically utilized by classical and alternative polarization signals.
   Nucleic Acids Res. 47 (6), 2778-2792, 2019.
   DOI: http://dx.doi.org/10.1093/nar/gkz118
   IF: 11.147 (2018)

5. **Ozgyin, L.**, Erdős, E., Bojcsuk, D., Bálint, B. L.: Nuclear receptors in transgenerational epigenetic inheritance.
   Prog. Biophys. Mol. Biol. 118 (1-2), 34-43, 2015.
   DOI: http://dx.doi.org/10.1016/j.pbiomolbio.2015.02.012
   IF: 2.581

6. Blaszczyk, K., Olejnik, A., Nowicka, H., **Ozgyin, L.**, Chen, Y. L., Chmielewski, S., Kostyrko, K., Wesoly, J., Bálint, B. L., Lee, C. K., Bluyssen, H. A.: STAT2/IRF9 directs a prolonged ISGF3-like transcriptional response and antiviral activity in the absence of STAT1.
   Biochem. J. 466 (3), 511-524, 2015.
   DOI: http://dx.doi.org/10.1042/BJ20140644
   IF: 3.562

7. Franyó, D., Boros Oláh, B., **Ozgyin, L.**, Bálint, B. L.: Befolyásolja-e az életmód génjeink működését?: az epigenetikai kutatások irányvonalai és eredményei.
   LAM KID. 2 (1), 37-42, 2012.

Total IF of journals (all publications): 24,464
Total IF of journals (publications related to the dissertation): 7,174

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

30 July, 2019

11 ACKNOWLEDGEMENTS

First I would like to express my greatest gratitude to my supervisor Dr. Bálint László Bálint for his continuous professional and personal support during both my undergraduate and graduate research studies; namely, for providing me the opportunity to learn a wide variety of experimental methods and bioinformatic analysis, for supporting my applications for travel grants, scholarships and conferences, as well as for financially supporting my English-Hungarian Medical and Health Science Translator studies during the course of my Ph.D. studies.

I would like to thank Dr. Attila Horváth for his support as a life partner throughout our Ph.D. journey, as well as for his substantial contribution to the improvement of my bioinformatics skills and both of my first-authored publications.

I would also like to thank Dr. Eric Soler, head of the Laboratory of Molecular Hematopoiesis at Inserm UMR 967, Fontenay-aux-Roses, France and Tharshana Stephen for providing me with the opportunity to obtain hands-on experience in multiplexed 3C-Seq during my Campus Hungary scholarship.

I am thankful for my fellow Ph.D. students Dóra Bojcsuk, Edina Erdős, Mária Csumita, Erik Czipa and László Halász for our inspiring and thought-provoking discussions, and for all the “fun” times spent together.

I would like to acknowledge my colleagues at the Genomic Medicine and Bioinformatic Core Facility, especially Dr. Szilárd Póliska, Ezsébet Mátyás, Ádám Pallér, Éva Nagy and Tamás Kerekes; colleagues in the Nagy Laboratory: Dr. Ixchelt Cuaranta-Monroy and Ibolya Fürtös; colleagues in the Proteomics Core Facility: Dr. Éva Csősz, Kamilla Sólyom and Lászlóné Darai (Julika); as well as Dr. Zsuzsanna Hevessy and Gombos Éva Kalmancheyne from the Dept. of Laboratoy Medicine for their experimental help.I am thankful for visiting scientists Dr. Katarzyna Blaszczyk and Dr. Daina Skiriutė for showing me their views on life and research and for becoming my friends.

Last but not least, my greatest appreciation goes to my beloved family, especially my mother, father and sister for their unconditional emotional support, and for being the kind of strong foundation for me throughout my whole life without which none of my major dreams could have come to reality.

12 SUPPLEMENTARY MATERIAL

SUPPLEMENT 1

Keresztessy, Z., Erdos, E., **Ozgyin, L.**, Kádas, J., Horváth, J., Zahuczky, G., & Balint, B. L. (2019). Development of an antibody control system using phage display. Journal of Biotechnology, 300, 63–69. https://doi.org/10.1016/J.JBIOTEC.2019.05.009 (including article supplementary material)

SUPPLEMENT 2

Ozgyin, L., Horvath, A., & Balint, B. L. (2018). Lyophilized human cells stored at room temperature preserve multiple RNA species at excellent quality for RNA sequencing. Oncotarget, 9(59), 31312–31329. https://doi.org/10.18632/oncotarget.25764 (including article supplementary material)

SUPPLEMENT 3

Ozgyin, L., Horvath, A., Hevessy, Z., & Balint, B. L. (2019). Extensive epigenetic and transcriptomic variability between genetically identical human B-lymphoblastoid cells with implications in pharmacogenomics research. Scientific Reports, 9(1), 4889. https://doi.org/10.1038/s41598-019-40897-9 (including article supplementary material)