SHORT 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

Development and characterization of experimental tools for functional genomics research

by **Lilla Ozgyin, MSc**

Supervisor: Dr. Bálint László Bálint, MD, PhD

Doctoral School of Molecular Cell and Immune Biology, University of Debrecen

Head of the **Examination Committee**:	Dr. Gábor Szabó, MD, PhD, DSc
Members of the Examination Committee:	Dr. Péter Bay, PhD, DSc
	Dr. Zoltán Wiener, PhD

The Examination takes place at the Discussion room of the Genomic Medicine and Bioinformatic Core Facility, In Vitro Diagnostic Building 3rd floor, Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen, at 11 a.m., 9th of December, 2019.

Head of the **Defense Committee**:	Dr. Gábor Szabó, MD, PhD, DSc
Reviewers:	Dr. Margit Balázs, PhD, DSc
	Dr. László Bodai, PhD
Members of the Defense Committee:	Dr. Péter Bay, PhD, DSc
	Dr. Zoltán Wiener, PhD

The PhD Defense takes place at the Lecture Hall of Building A of the Department of Internal Medicine, Faculty of Medicine, University of Debrecen, at 1 p.m., 9th of December, 2019.

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, providing a genome-wide view on the dynamic aspects of gene regulation in pathological contexts. Moreover, public repositories of well-annotated biomaterials and high-throughput sequencing data provide an unparalleled opportunity to utilize available resources to enhance the accumulation of 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 efforts, including the landmark International HapMap Project, 1000 Genomes Project and ENCODE Project, used LCLs. The accessibility of thousands of LCLs together with well-annotated sequencing data in data 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 for RNA-Seq. Third, by uncovering, for the first time, the extent and nature of non-DNA-driven functional genomic variability of the LCL model.

THEORETICAL BACKGROUND

Normalizers in chromatin immunoprecipitation experiments

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*. 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 seriously limit their use in clinical settings.

ChIP protocols involve controls and normalizers, which ensure the elimination or identification of confounding variables, such as varying starting cell numbers and aspecific capture. 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, a number of negative controls can be used, for instance, negative control genomic loci. Another commonly used negative control experiment is IP with aspecific isotype-matched antibodies. 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, a representative chromatin aliquot. The per cent input method is particularly useful for controlling chromatin input variability.

There is no generally accepted internal control to be used in ChIP-qPCR experiments. Defining appropriate internal controls for a given ChIP experimental setup may be cumbersome. Without appropriate internal or external controls, experimental sample loss after the chromatin fragmentation step cannot be accounted for. This is especially problematic given the fact that the majority of sample handling, including extensive washing steps (~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 post-fragmentation procedural controls, as substitutes to 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 (cells, fragmented chromatin) to the experimental sample. The spiked material was precipitated in the same experiment using one

pan-specific antibody or two different antibodies. The so-called ICeChIP (Internal Standard Calibrated ChIP) enables the calculation of histone modification density, and direct comparison of experiments, by adding reconstituted and semisynthetic nucleosomes with barcoded DNA to the sample in concentration series. One serious limitation of these reagents is the lack of consistent quality. Despite these recent developments, the use of spike-in controls for comparative ChIP has not yet become widespread.

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). 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 (encoding e.g. 6-mer or 12-mer peptides) into one of the phage (e.g. M13) coat protein-encoding genes (e.g. pIV or pIII). Assembled phage vectors are then transformed into an E. coli strain, which produces fusion phages. Different vendors offer various libraries fitted to different research needs, and custom libraries can be prepared using commercially available kits.

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 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. 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. 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 in consistent quality by developing mocónoclonal reagents, and by simply reinfecting a host E .coli strain. In theory, if the phage genome-containing ChIP DNA is later decided to be sequenced, 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.

RNA sequencing in clinical research

RNA-Seq is a transcriptome profiling method applying high-throughput sequencing for the detection and quantification of cDNA. With falling sequencing costs and expanding bioinformatic toolsets, RNA-Seq has also become the leading technology of high-throughput RNA biomarker research. 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 has been on the rise in the past few years, which may pave the way for the development of RNA-based biomarker panels.

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 through biopsies is especially prone to degradation, especially 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. Therefore it is of substantial interest to preserve tissue/cellular samples in a way that enables the extraction of highly intact RNA.

Tissue banks and biosample storage at ambient temperatures

Given that native research biomaterials are sensitive to a range of chemical and physical exposures, various methods have been utilized to minimize degradation and the consequential change in features of interest. 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).

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. 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 affects 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 serious environmental burden, raising concerns over the financial sustainability of tissue banks in parallel with shrinking funding resources worldwide. Moreover, temperature fluctuations during 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.

Lyophilization

Lyophilization may enable the ambient storage and transportation of cells and tissues. Lyophilization 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. At the end of the drying process, samples are generally sealed and stored in the dark above freezing point (e.g. at room temperature). Lyophilization might serve as a suitable alternative to storage at subzero temperature, due to the following advantages: (1) low residual water content in the dried product essentially stops molecular motion and water-mediated reactions, inhibiting molecular degradation pathways; (2) it prevents heat-induced deterioration of heat-labile substances during the drying process; (3) the method requires minimal hands-on time, and (4) the cost of long term storage of freeze-dried samples may be lower than that of cryopreservation, as these samples do not require low storage temperatures; (5) temperature fluctuations (e.g. during transportation) may not have such detrimental effects on lyophilized samples. Other room temperature storage methods include formalin fixation and paraffin embedding (FFPE) and non-crosslinking fixatives. FFPE is known to lead to pervasive RNA degradation and nucleobase modification, while non-crosslinking fixatives, e.g. RNAlater, may trigger transcriptomic changes.

The integrity of RNA samples isolated from lyophilized tissues has been assessed by multiple groups, using various methods and storage conditions. In general, lower RIN values were detected, while RT-(q)PCR results were consistent for more extended periods. Storage in the dark and in the presence of desiccants have been shown critical for RNA stability, while storage at RT preserved RNA better than storage at 4°C. In addition, nucleic acid integrity in lyophilized tissues might depend on tissue lipid content and oxygen-driven peroxidation Although RNA has been consistently found less stable than DNA or proteins,

tissue lyophilization is a promising, yet underexploited method of sustainable RNA preservation.

Cellular phenotypic instability in cell line models

Genomic alterations in cell culture is a major contributor to changes in cellular phenotypes. 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. It is widely accepted that overly passaged cell lines accumulate genomic changes, although the timeline and nature of changes may differ from one cell line to another. As yet, however, there is no general consensus on the maximum number of acceptable cell passages.

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 certain stimulus. The variability in the composition of the least controlled culturing component, foetal bovine serum, may also guide cell populations to different cell line evolutionary paths. In genetically stable cell lines, such as ESCs, environmental factors (such as culture conditions) are expected to play the most prominent role in altering molecular phenotype differences, which may lead to irreversible changes in the single-cell level. 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 and show significant geno- and phenotypic divergence. In order to minimize the effects of cell line evolution, biobanks with cell line batches of same passage numbers can be created.

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*. EBV infection leads to characteristic phenotypic

changes, generally leaving genotypes and inter-individual gene expression patterns unaffected. LCLs have been found instrumental in a variety of studies including, but not limited to, research on EBV life cycle, DNA repair, immune response, as well as research on EBV-related and non-EBV-related disorders. Moreover, LCLs were models for large omics projects, such as HapMap, 1000 Genomes, ENCODE and FANTOM5, and were the model for first kilobase-resolution genomic 3D map. Their main advantages include: 1) the easiest way to obtain renewable biosample from healthy individuals, 2) easy culturing to large quantities, 3) pre-immortal state, 4) extensively reported genetic and karyotypic stability, and 5) publicly available whole-genome sequences. Due to the latter, LCLs are widely used for molecular and drug QTL studies.

LCLs in genetic association studies

LCLs have become instrumental in QTL studies for the past decade. Quantitative trait loci (QTLs) are genomic loci which correlate with quantitative traits, such as histone modification levels (hmQTLs), gene expression levels (eQTLs) and drug response. The availability of thousands of lines with free to use genotype data in public repositories is appealing for studying genome-wide phenomena in heterogenous samples requiring large sample sizes. Moreover, in online data repositories (e.g. GEO and ArrayExpress), relevant functional genomic data are freely available for reanalysis (e.g. ENCODE ChIP-Seq datasets).

LCLs have been increasingly used as models for pharmacogenomics research. Adding to the above benefits of LCLs, a hight number of cell lines can be selected with similar age, and race; moreover, LCLs are generally free from *in vivo* confounders. LCLs are used either as primary models or as validation tools during clinical trial follow-ups. The current predominant approach to assess genotype-dependent drug response in LCLs is the use of large panels of cell lines from age- and sex-matched healthy individuals. Another way to use LCLs is functional follow-up studies. The so-called ’triangle approach’ may reduce false positive hits by incorporating gene expression data into pharmacogenomic studies. LCLs have been the models for studies assessing sensitivity to various chemotherapeutic agents, lipid-lowering-drugs and antidepressants, among others.

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 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 27^{th} 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

MATERIALS AND METHODS

Cell culture

We prepared a three-tiered biobank of LCLs purchased from Coriell Cell Repositories (GM22647-GM22651, GM12864, GM12872, GM12873). Tier 3 cells were used for all experiments. Cells were cultured in RPMI-1640 supplemented with 15% heat-inactivated FCS, 2 mM L-glutamine, and 1% penicillin-streptomycin, and were incubated at 37°C (5% CO_2) in T25 or T75 flasks in an upright position.

STR analysis

DNA was isolated from 10^6 PBS-washed cells using Roche's High Pure PCR Template Preparation Kit. Five STR regions (AMELY/AMELX, D18S51, D8S1179, TH01 and FGA) were amplified (PowerPlex S5 System, Promega) and run using the ABI PRISM 3100-Avant Genetic Analyzer. Results were analyzed using the GeneMapper ID software v4.1 at the Department of Laboratory Medicine, Faculty of Medicine, University of Debrecen.

Immunophenotyping and cell cycle analysis

Combinations of antibodies were added to 10^6 cells and incubated for 15 minutes at RT in the dark. Samples were washed and fixed using 1% paraformaldehyde in PBS. Intracellular staining was carried out following the procedure described for Intrastain (Dako Glostrup). 100,000 events were acquired with a FACS Canto II flow cytometer (BD). Data were analyzed using FACS Diva (BD) and Kaluza Software version 1.2 (Beckman Coulter). Sources of antibodies: CD5, CD10, CD20, CD22, CD24, CD34, CD38, CD79b, CD81, FMC7, HLA-DR, kappa and lambda markers (BD); CD19, CD23, and CD43 (Beckman Coulter); CD21 and anti-CD45 (Exbio); nTdT and IgM (Dako). For cell cycle analysis, $2*10^6$ PBS-washed cells were fixed with 70% ethanol at 4°C, pelleted, and were incubated with RNase and Propidium-iodide for 30 minutes at RT, in the dark. 20,000 events were acquired with a FACS Calibur II flow cytometer (BD). Data were analyzed with ModFit LT (BD).

Phage biopanning

The Ph.D.™-7 random heptapeptide library (NEB) was panned in four consecutive rounds against magnetic bead-coupled AR N-20 antibody (Santa Cruz Biotechnology). 10^{11} PFUs blocked with TBST/BSA were immunoprecipitated with 10 μg AR antibody (10-60 minutes, RT). A Protein A:Protein G paramagnetic bead mix (1:1) pre-washed with TBST/BSA was incubated with the antibody-phage reactions (20 minutes, RT). Beads were

washed ten times with TBST/BSA, eluted, and the elution buffer was neutralized. Early-log F+ ER2738 bacteria were infected with the phage eluate and were grown in non-selective LB. Bacterial supernatants were precipitated with 1/6 volumes of 20% PEG-8000/2.5 M NaCl overnight (4°C). Phage pellets were dissolved in TBS and precipitated again. TBS-dissolved phages were used as starting material for three additional rounds of biopanning (10^{10} PFUs). SmartSpec Plus (Bio-Rad) was used to measure phage concentrations. Glycerol stocks (50%) were prepared and stored at -20°C.

Phage subcloning

NEB's phage titering and plaque amplification protocol was used for generating monoclonal phage reagents. The fourth round of polyclonal phage stock was diluted in LB and was used to infect mid-log ER2738 cells. Infected cells were added to 45°C "Top Agar", poured onto LB/IPTG/X-gal plates, and were incubated at 37°C overnight. The next day dilute ER2738 cultures were infected with individual blue plaques picked from the plates. Monoclonal phages were precipitated and handled the same way as polyclonal stocks.

Chromatin immunoprecipitation sequencing (ChIP-Seq)

For the sGT study, PBS-washed cells were fixed at RT either with 1 % methanol-free formaldehyde (FA; Thermo) for 10 min (sGT) or with 2 mM di(N-succinimidyl) glutarate (Sigma) for 45 minutes followed by 1 % FA for 10 minutes (trio). Reactions were stopped using 0.125 M glycine. For the lyo study, we used fresh, IMT-2 lyophilized, IMT-2 lyophilized and 1% FA fixed, trehalose preloaded and IMT-2 lyophilized (0.1 M trehalose overnight at 37°C prior to lyo), and FA-prefixed and IMT-2 lyophilized cells. For EGCG treatment, we cultured LCL cells in medium containing 2 mM EGCG 1 hour prior to cell harvesting. Nuclei were isolated and were either sonicated (Bioruptor Plus) or MNase-treated and mildly sonicated. 90% of ten-times diluted, cleared chromatins corresponding to $5*10^6$ cells were immunoprecipitated with either 2.5 μg anti-histone H3 (acetyl K27) antibody or isotype control antibody overnight at 4°C. To the top 90% of the centrifuged IPed samples BSA-blocked Protein A:Protein G (1:1) magnetic bead mixtures were added, and samples were incubated for 6 hours at 4°C. Beads were washed six times using four different wash buffers, and antibody-antigen complexes were eluted. Crosslinks were reversed, samples were treated with RNase A and Proteinase K, and DNA fragments were isolated using Qiagen's MinElute PCR purification kit. DNA concentrations were determined using the Qubit dsDNA

HS Assay Kit (Thermo). Most buffers were supplemented with Roche's cOmplete Mini proteinase inhibitor tablets. Original chromatins without enrichment were used as "input" for ChIP-qPCR experiments; isotype control and an H3K27ac negative genomic region were used as negative controls.

For phage ChIP, we used Diagenode's IP-Star Automated System until the purification step. Sonicated, 1% FA fixed and sonicated chromatins were prepared from HEK293T cells. IP plates were prepared by adding buffers and reagents into different wells: 1 μg anti-AR antibody (or isotype-matched IgG) in IP buffer, HEK293T chromatin (~100,000 cells), 10^6 phage particles in IP buffer, 100 μl Wash Buffer A, B or C, 100 μl TE buffer, 100 μl Elution Buffer, and 10 μl of Protein A:Protein G bead mix (1:1). Following the programmed IP, bead coupling, washing and elution steps, samples were de-crosslinked, treated with RNase and Proteinase, and isolated with High Pure PCR Template Preparation Kit (Roche). M13 universal primers and UPL probe 48 (Roche) was used for qPCR measurement of phage genomes on a LightCycler 480 instrument (Roche). Input samples were used to normalize data ($2^{-\Delta Cp}$ method).

TruSeq ChIP Sample Preparation Guide 15023092 B (Illumina) was used for library preparation from 10 ng ChIP DNA (sGT study, H3K27ac). Libraries were sequenced at the Genomic Medicine and Bioinformatics Core Facility (University of Debrecen; NextSeq 500 system) and at the EMBL Genomics Core Facility (Heidelberg, Germany; HiSeq2000 system) (50-bp, single-end). BWA 0.7.10 was used to align reads to the hg19 reference genome. HOMER 4.9.1 was used for predicting enriched regions, for preparing reads for super-enhancer (SE) prediction and read distribution heatmap, for predicting super-enhancers and for predicting nucleosome-free regions from pooled bam files. Pheatmap (R) was used to cluster differentially acetylated regions. Java TreeView was used to plot tag densities. DiffBind was used for defining consensus regions, for calculating RPKM values, and for creating a correlation heatmap visualized with Plotly 3.0.0. We used two-way ANOVA combined with Tukey's post hoc test with functions aov() and TukeyHSD() (MASS, R) to define differentially enriched regions ($P < 0.05$, $FC > 2$). Closest genes were assigned using bedtools. Integrative Genomics Viewer (Broad Institute)was used to visualise bedgraphs, and The 3D Genome Browser (Yue Lab) was used to visualize Hi-C data from the GM12878 cell line (40 Kb resolution). PhenoGram was used to visualise differentially acetylated regions over chromosome 3.

MNase profiling

Nuclei were washed and were treated with MNase at 37°C for 30 minutes at 66.6 GU, 22.2 GU or 7.4 GU per $1.5*10^6$ nuclei. Stopped reactions were mildly sonicated (Bioruptor Plus, Diagenode), and cell debris was removed. 80% of the supernatant was precipitated with 3 volumes of absolute ethanol at -20°C overnight. The next day nucleic acids were pelleted and desiccated. Fragmented DNA was recovered essentially the same way as during the ChIP protocol described above. Samples were run on a 1% agarose gel to analyze fragment lengths.

Chromatin conformation capture sequencing (3C-Seq)

We followed the protocol for multiplexed 3C-sequencing described previously. Cells were crosslinked in 10% FCS/PBS with 1% FA (Thermo) for 10 minutes at RT. Reactions were quenched with 0.125 M Glycine, and nuclei were isolated from washed cell pellets. Washed nuclei were digested using 400 U EcoRI-HF (NEB) at 37°C. Digested samples were ligated with 100 U T4 ligase (NEB) under dilute conditions at 16°C. After reversing crosslinks and RNase/Proteinase treatment, DNA was isolated using the phenol-chloroform-isoamyl alcohol (PCI) method. 10 μg of DNA was digested with 10 U MseI (NEB) at 37°C. Samples were isolated with the PCI method and ligated using 100 U of T4 ligase at 16°C overnight. Ligation samples were isolated with the PCI method, and DNA samples were amplified using Expand Long Range, dNTPack (Sigma) using primers specific to a region in the P3H2 gene body (inverse PCR). PCR reactions were cleaned up using Qiagen's MinElute PCR purification kit. Libraries were prepared using TruSeq ChIP Sample Preparation Guide 15023092 B (Illumina) with minor modifications. Samples were sequenced (75-bp) at the EMBL Genomics Core Facility (Heidelberg, Germany; HiSeq2000 system).

Lyophilization

$3*10^6$ PBS-washed LCL cells were resuspended in 0.5 ml lyophilization solution containing 0.1 M D-(+)-trehalose dihydrate in PBS or in IMT-2 solution containing 0.1 M D-(+)-trehalose dihydrate and 0.945 mg/ml (-)-epigallocatechin gallate in PBS, in microcentrifuge tubes. Cell suspensions were then snap-frozen in liquid nitrogen, a parafilm with seven 1-mm holes was placed on top of the tubes'opening, and samples were loaded into the freeze dryer (CoolSafe 110, ScanVac), at a condenser temperature of -110°C (Proteomics Core Facility, Faculty of Medicine, University of Debrecen). Lyophilization was carried out at 0.004 mBar for six hours. After finishing the lyophilization cycle, tubes were closed, and

lyophilized cell powders were processed immediately, or stored for 2 weeks or 2 months at room temperature (23–25°C) in the dark, in the presence of $CaCl_2$ dihydrate.

RNA isolation

Total RNA was isolated using the TRIzolate method. In short, 2-3*10^6 LCL cells (fresh or lyophilized) were washed with PBS, pelleted, and vortexed for 5 minutes in 1 ml TRIzolate (UD-Genomed Ltd). Phases were separated by adding chloroform (1:5) and high-speed centrifugation. Nucleic acids were precipitated from the aqueous phase using 1:1 isopropanol. Pellets were washed twice with chilled 75% ethanol and vacuum-desiccated. RNA pellets were redissolved in nuclease-free water. Sample purity was assessed using a NanoDrop 1000 instrument (Thermo), and concentrations were measured using Qubit fluorometer (RNA HS Assay Kit, Thermo). Agilent RNA 6000 Nano microchips were used to analyze fragment distributions and to determine RIN values.

RT-qPCR

qPCR primers were designed using either Roche's UPL Assay Design Center (UBR2, TRERF1, PTPRJ, SLC6A4, RXRA, TCL1A, SPI1 and CT64 assays) or Primer 3 Plus (DPYD; ACTB; lncRNAs: MALAT1, GAS5, TUG1; eRNAs: eIRF4_-1.9kb, eSPI1_-16kb and eMYC_-170kb). GAPDH primers were derived from Sigma. Total RNA samples were treated with RQ1 DNase (Promega) and were reversely transcribed (SuperScript II, Thermo). RT reactions were diluted five-fold with nuclease-free water prior to qPCR. Prior to the EGCG inhibition assessment, 20-µl RT reactions were prepared with final concentrations of EGCG between 10^6 and 10^7 M. We amplified target regions using the LightCycler 480 SYBR Green I Master (Roche) with 0.375 µM of each primer. The qPCR reactions were prepared in triplicates, Reverse transcription negative control reactions lacking the SSII enzyme were prepared for each sample. The $2^{-\Delta Cp}$ method was used for quantification against the ACTB normalizer, where applicable.

RNA-Sequencing

RNA-Seq libraries were prepared from 1 µg total RNA following Illumina's TruSeq RNA Sample Preparation v2 Guide. Libraries were sequenced on the NextSeq 500 system using 50-bp (sGT study) or 75-bp (lyophilization study) read length (single-end). For the sGT study, all steps were carried out at the Genomic Medicine and Bioinformatic Core Facility (University of Debrecen). 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). Reads were aligned to the human reference genome hg19 with TopHat v2.0.7. For the sGT study, transcript abundances were calculated using edgeR and genes with CPM<5 were discarded. EdgeR was used to identify differentially expressed genes (FDR = 0.05, FC > 2). For the lyo study, transcript quantitation was carried out with Cufflinks genes with RPKM<1, as well as small RNAs were discarded. We used Cuffdiff to find differentially sampled RNAs (FDR = 0.05). DAVID Bioinformatics Resources 6.8 tool was used functional annotations. The QoRTs package was used to obtain metadata regarding the sequencing library, such as read GC content, mismatch profile, and gene body coverage. Transcript information was derived from the HGNC database using BioMart and custom scripts, and we used the Mann-Whitney U test for statistical analysis. ARE data was derived from the ARED-Plus database.

5-FU treatment and cytotoxicity assay

20,000 cells were seeded in ’FU medium’ (indicator-free RPMI, 15% heat-inactivated FCS, 2 mM L-glutamine, 1 % penicillin-streptomycin) to a 96-well U-bottom plate in quadruplicates. Serial dilutions were prepared from a 5-FU stock (TEVA) in ’FU medium’ and were added to the cells at indicated concentrations. Ultrapure water was used as vehicle and medium-only wells as background. Cells were incubated with 5-FU at 37°C (5% CO_2) for 72 hours. MTT stock (4.5 mg/ml in PBS) was added to the wells, and the plate was incubated at 37°C for 6 hours in a non-transparent foil. Cells were lysed and incubated for one additional hour. A VICTOR3 Multilabel Plate Reader (PerkinElmer) was used to measure absorbances at 595 nm.

Capillary sequencing

The phage genome was isolated using Roche’s High Pure PCR Template Preparation kit following the manufacturer’s instructions. Samples were amplified using the -96 gIII primer, and the ABI 310 Avant sequencer was used to detect fragments (Genomic Medicine and Bioinformatic Core Facility, University of Debrecen).

RESULTS

DEVELOPMENT AND CHARACTERIZATION OF A POLYCLONAL AR-MIMICKING PHAGE SPIKE-IN CONTROL REAGENT

We selected polyclonal AR-mimicking phages using NEB's Ph.D.™-7 random heptapeptide library by four consecutive rounds of affinity selection with a ChIP-grade anti-AR antibody. In order to get a general view of the enrichment of certain bases at the 21 variable genomic positions over the selection rounds, we performed capillary sequencing for each polyclonal batch. 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. We next tested whether the generated polyclonal libraries can be efficiently captured in a ChIP reaction. We set up an IP experiment that 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, etc.), remained relatively stable across the four phage batches. Approximately 50% of spiked phages could be recovered using the 'round 4' batch.

DEVELOPMENT AND CHARACTERIZATION OF MONOCLONAL AR-MIMICKING PHAGE SPIKE-IN 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 an IP experiment. We infected ER2738 cells with highly diluted 'round 4' phages and amplified/purified individual clones, and using the semi-robotic IP-qPCR method, we found that the elution buffer:IP buffer ratio was 50%< for all clones, with one clone with exceptionally high (90%<) affinity to the anti-AR antibody.

THE EFFECT OF EPIGALLOCATECHIN GALLATE LYOPROTECTANT ON RT-QPCR AND CHIP-QPCR

Based on previous success with mammalian cell lyophilization and recovery of live cells, we attempted to perform RT-qPCR and ChIP-qPCR experiments from RNA and DNA samples isolated from LCLs lyophilized in IMT-2, containing epigallocatechin gallate, in our pilot experiments.

Total RNA samples isolated from IMT-2 lyophilized LCLs showed lower $OD_{260/230}$ ratios compared to controls (P value = 0.03, paired t-test). In line with that, RNA pellets had brownish-grey colour during isolation, which indicated EGCG contamination. RIN values were generally high, though lower than those for paired controls (P value = 0.03, paired t-test). Moreover, lyophilized samples showed significantly elevated Cp values for a highly expressed housekeeping gene, GAPDH, compared to controls, which could be reversed by diluting total RNA samples 100-fold prior to RT. EGCG spike experiments proved a concentration-dependent, lyophilization-independent inhibitory effect of EGCG on RT-qPCR.

Although MNase profiling revealed a normal beads-on-a-string nucleosome profile of chromatin isolated from IMT-2 lyophilized cells, only cells fixed with 1% FA prior to lyophilization showed sufficiently high IP efficiencies (H3K27ac), regardless of fragmentation type, fixation strength and trehalose preloading. Of note, ChIP DNA isolated from cells lyophilized without prior fixation showed a mild brownish discolouration, similarly to RNA pellets. All input (unprecipitated chromatin) measurements resulted in similar Cp values, suggesting that qPCR inhibition is less likely the cause for the above phenomenon. To test our hypothesis that EGCG itself leads to low IP efficiency, we treated LCL cells for 1 hour with 2 mM EGCG and performed H3K27ac ChIP-qPCR. We found that the addition of EGCG resulted in lower IP efficiency of the positive control region. We concluded that EGCG should not be used as a cellular lyoprotectant when the downstream applications involve RNA- or chromatin-based studies.

CELLULAR RNA QUALITY AND QUANTITY AFTER LYOPHILIZATION WITH TREHALOSE AND WEEKS OF ROOM TEMPERATURE STORAGE AS MEASURED BY STANDARD METHODS

Total RNA isolated immediately after lyophilization in a trehalose-containing lyoprotectant solution were characterized by remarkably high RIN values (mean = 9.8), as

well as yields not significantly different from those of fresh cells. We next assessed whether the quantitation of variably abundant mRNAs (extremely low, RPKM < 1; low, RPKM = 1-10; moderate, RPKM = 10-100; and high, RPKM > 100), long non-coding RNAs (lncRNAs) and the recently described class of enhancer-associated RNAs (eRNAs) shows differences between fresh and lyophilized cells. The expression of the selected genes in two LCLs was not significantly different between control and lyophilized samples. These results suggest that lyophilized cell-based samples enable accurate gene expression quantitation by RT-qPCR.

TRANSCRIPTOMIC EFFECT OF TREHALOSE LYOPHILIZATION OF LCLs

Three total RNA samples isolated from lyophilizates stored for two weeks at RT, together with their paired control samples were selected for transcriptome-wide analysis by mRNA-Seq. First, overall library qualities were determined using standard quality metrics. The percentage of uniquely mapping reads (over 90%) and duplicated reads were similar for all samples, and no outlier library was detected. Next, we calculated read GC content, chromosomal distribution, biotype distribution, gene body coverage, cumulative gene diversity and per base mismatch rate; none of these metrics showed a significant difference between control and lyophilized samples. Using differential gene expression analysis, we found a high correlation between control and lyophilized datasets ($r^2 = 0.99$) and identified 28 genes significantly downsampled in lyophilized samples (FDR = 0.05; 21 protein-coding genes (PCGs), 6 lncRNAs and 1 pseudogene). Lowly expressed genes showed higher fold-difference. GO analysis uncovered the enrichment of the term ‘DNA-templated transcription’ ($P = 2.0*10^{-5}$), including the transcriptional regulators POLR2A, INTS1, and KMT2D genes.

CHARACTERISTICS OF RNAS DOWNSAMPLED IN TREHALOSE LYOPHILIZED LCLs

Next, we decided to uncover the distinctive features of DEGs. We calculated read coverage ratios for metatranscripts of each gene, and we found that most differentially expressed genes (DEGs) did not show 3’ bias ($P > 0.01$; $N = 16$). 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 ($P < 0.01$). Downsampled lncRNAs and protein-coding RNAs (5’UTR+CDS+3’UTR) were shown to have significantly higher transcript length and GC fraction than all corresponding human transcripts ($P < 0.0001$, Mann-Whitney test). 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 <

0.001; Mann-Whitney test). Also, twelve protein-coding DEGs (57%) were listed as containing at least one 3'UTR or intronic ARE in the ARED-Plus database.

STR-, PLOIDY- AND CELL CYCLE STAGE ANALYSIS OF ISOGENIC LCLs

We found that the five reportedly isogenic LCLs derived from the same CEPH/UTAH 26-year-old healthy male have diploid genotypes, a similar cell cycle progression, and using STR markers, the cells were confirmed to be indeed genetically identical and originating from a male donor, with no sign of contamination with genetically distinct cells. These results suggested that the main characteristics of the selected five isogenic LCLs were suitable for the purposes of our study.

IMMUNOPHENOTYPING OF ISOGENIC LCLs

The method used allowed us to identify the source cell type, assess cell line clonality and surface marker expression heterogeneity within cell lines. Flow cytometric analysis of sGT 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 represent a lower level of complexity, they were also derived from multiple B cell clones. The cell lines showed expression patterns characteristic to mature B cells (CD19+ with low side scatter, 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), which was possibly the result of EBV infection. Moreover, all sGT 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 (54-76%), CD38: 20-87%, cIgM: 1-84%.

REGULATORY ELEMENT-LEVEL VARIABILITY OF ISOGENIC LCLs

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. We compared all regulatory elements marked with H3K27ac, to get an overall view of the level of similarity between LCLs in the same genotype context, an we found that although the correlation coefficients were remarkably high across

the sGT LCL dataset (between 0.9 and 0.97), biological replicates clustered together, indicating that sGT LCLs have their unique epigenetic profile. Of the 42,923 consensus regions, we found that almost one-fourth (9,685 sites) had variable H3K27ac enrichments across sGT LCLs (RPKM fold-difference > 2, P value < 0.05, between at least two cell lines). When we compared each pair of cell lines, we found 1,056 to 4,174 variable regulatory elements. We also found that intergenic enhancers were highly affected, while acetylation levels are relatively stable at promoter elements across the cell lines.

COORDINATED CHANGES IN REGULATORY ELEMENT ACTIVITIES OVER EXTENDED GENOMIC REGIONS IN ISOGENIC LCLs

Super-enhancers (SEs), linearly clustered gene regulatory elements < 12.5 Kb apart from each other and spanning several kilobases, were predicted from H3K27ac ChIP-Seq tags pooled across all samples. The predicted SEs were located in the proximity of genes involved in B cell and immune functions (e.g. PAX5 and IRF2), are related to EBV infection ('EBV super-enhancers', e.g. BCL2, MIR155 and MYC). Of the 1,058 predicted putative SEs, 49% contained at least one variable constituent enhancer, but only 31 (2.9%) SE were found variable as a whole entity (FC > 2, P < 0.05) across sGT LCLs. Genes closest to variable SEs were associated with immune functions such as leukocyte activation ($P = 5.5*10^{-4}$) and leukocyte cell-cell adhesion ($P = 5.6*10^{-4}$). Transcription factor activity ($P = 1.2*10^{-2}$) and LPS binding ($P = 3.2*10^{-2}$) were among the most enriched molecular functions.

We took an alternative approach to assess whether coordinated loss or gain of enhancer activity extends to non-super-enhancer regions as well. We found that variable enhancers cluster based on the direction of change over long genomic regions not previously classified as super-enhancers. Using a publicly available LCL Hi-C dataset, we found that coordination may extend through multiple topologically associated domains (TADs), and signal direction may also switch from one TAD to another. Using the P3H2 gene as a bait, we performed multiplexed 3C-Seq and found that a decrease in chromatin contact frequencies was associated with a coordinated decrease in H3K27ac signal over megabases of the genome.

H3K27AC VARIABILITY IS LINKED TO TRANSCRIPTOMIC VARIABILITY IN AND AFFECT CLINICALLY RELEVANT PATHWAYS ISOGENIC LCLs

When assessing whether and to what extent chromatin activity differences were mirrored by transcriptome-level differences, we observed that gene variation patterns

generally follow that of the H3K27ac signal. However, only 525 (4.6%) of genes were found to be significantly differentially expressed (differentially expressed genes; DEGs) across sGT LCLs (FDR = 0.05, FC > 2). Pairwise comparison of cell lines resulted in 25 to 229 variable genes (mean = 119.8; median = 107.5). Notably, none of the previously reported EBV copy number-related genes, CXCL16, AGL, and ADARB2, were found to be differentially expressed in our dataset, suggesting that gene expression changes had not been induced by EBV infection differences.

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. Surprisingly, 121 of DEGs had been previously categorized as being pharmacogenes (genes associated with response to pharmaceuticals) based on the Genetic Association Database (GAD). 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. 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.

NON-GENETIC VARIABILITY MIGHT LEAD TO ALTERNATIVE CELLULAR RESPONSE TO DRUG TREATMENT IN ISOGENIC LCLs

We decided to asses whether drug response phenotype level is also affected, which may have implications in LCL-based pharmacogenomic research. 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), and as coupled 5-fluorouracil (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. 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). Our result suggests that non-genetic factors might affect LCL cells' response to drugs in pharmacogenomic screenings.

DISCUSSION

Spike-in phages as novel controls for chromatin immunoprecipitation experiments

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.

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, the requirement of which may be addressed by capturing bead-antibody complexes in a separate experiment, and adding these beads to the reactions at an appropriate time point, or by using a separate antibody and phage reagent to exclude cross-reaction . 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 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.

Lyophilization and ambient storage of human cells to preserve RNA

Operating conventional biobanks, especially ultra-low temperature freezers, have high associated costs and considerable environmental impact. 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. Studies are underway to extend the utility of lyophilization to stabilizing mammalian cells and tissues. However, lyophilization of mammalian cells for downstream RNA studies has not become standard, despite a few studies reporting only minimal RNA degradation using RT-qPCR of certain genes and gel

electrophoresis. In our research, we aimed at extending our knowledge on lyophilizing mammalian cells for RNA-based applications, including novel insights by measuring low abundance genes, lncRNAs and enhancer RNAs, and profiling the transcriptome.

In our preliminary experiments, we used IMT-2 solution containing trehalose and EGCG in PBS as lyoprotectants, which was described as an efficient membrane stabilizer, which would have facilitated applications requiring whole cells (e.g. ChIP). After cell rehydration, we found microscopically intact cells trypan blue-penetrable cell membranes, and cells could not proliferate in culture. Also, isolated RNAs showed lower RIN values, which may have resulted from endogenous RNAse-based degradation during cell washes, or might result from the activation of apoptotic pathways previously described for EGCG-treated LCLs. Moreover, reverse transcription was shown to be inhibited by the EGCG co-precipitated with total RNA, and H3K27ac ChIP was also unable to enrich the positive control region, the regulatory element of a TF gene highly expressed in the steady-state. This might be the result of the inhibition of the experiment, or more likely reflect a biological response to EGCG. We concluded that using such a potent gene regulator and RT inhibitor with rapid membrane penetration would hinder the comparability of lyophilized samples.

In the subsequent experiments, trehalose was used as lyoprotectant, which is a relatively cheap reagent and has been shown to sequester reactive oxygen species, as well as to protect cells during dehydration. Keeping in mind the importance of cost-effectiveness in biobanking, we aimed at using a short lyophilization cycle (low energy consumption). We could lyophilize samples in six hours, which is substantially less than cycle lengths used in the pharmaceutical industry. Trehalose did not allow for the recovery of intact cells, but the dried products could easily be resuspended in TRIzol and resulted in high quality (RIN) and quantity RNA isolates even after two weeks or two months of room temperature storage. Selected variably abundant mRNAs, lncRNAs and eRNAs were measured by RT-qPCR and were found to be expressed at highly similar levels to paired controls.

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. Therefore we performed RNA-Seq from RNAs isolated from lyophilized samples stored for two weeks at RT, and the generated sequencing libraries were shown to be highly comparable using multiple quality metrics, such as UMRs, 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-change. Of this gene set, lower abundance genes showed higher fold-changes, which may result from higher degradation rates or the combination of the higher stochasticity or measurement bias of low expression genes. Assessing transcript features of DEGs, we found that affected transcripts are generally longer, contain more G and C residues, and often encode transcription factors. It was previously shown that longer transcripts and TF-encoding transcripts are less stable both *in vivo* and *in vitro*. However, GC content at the third codon was found to be inversely correlated with degradation rates. 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, a study reported their annual costs of lyophilized vs -80°C vs LN_2 storage, which was 3, 24 and 31 EUR, respectively. Our lyophilization cost estimation taking into account the energy consumption of the CoolSafe freeze dryer and the price of LN_2 and trehalose resulted in 0.87 USD per sample when only one sample is lyophilized per run. 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. whole-cell lyophilization for gene regulation-related studies are yet to be established.

Genotype-independent cellular phenotypic variability of the LCL model

With the combination of classical molecular biology techniques and high throughput methods, the scientific community has got the opportunity to revisit commonly held assumptions about the most popular cell lines. Studies using LCLs commonly refer to inter-cell line differences, rather unfoundedly, as inter-individual differences. A few studies 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.

In our study, we aimed at exploring the cis regulatory element- and transcriptome-level variability of LCLs using five genetically identical cell lines. The preparation of these isogenic cells resembled that of genetically distinct, commercially available LCLs. In order to minimize variability emerging during our research, we ensured that all LCLs had the same number of freeze-thaw cycles and passages prior to initiation experiments by preparing a three-tiered biobank. We also handled all cell lines in parallel, harvested cells at the same time point of the day, and used biological replicates to exclude differences due to random fluctuations. Altogether, this model reflects variability emerging during cell line generation and short-term culture.

All cell lines were confirmed to be euploid, and have been derived from the same human male source, and showed similar cell cycle stage distribution. All cell lines showed mature B cell phenotype, with only one cell line showing evidence to monoclonality. Of note, the shrinkage of diversity has been shown to occur mostly at the early steps of cell line generation and, to a lesser extent, is affected by later culturing. We assumed that polyclonal cells better mirror the heterogeneity of the original B cell pool and that derivatives of one descendant cell of the initially diverse cell population might show divergence from the other cell lines. However, the monoclonal line did not show any outstanding features in our study.

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. Intergenic enhancer regions showed the highest fraction of variable regions, in contrast to promoters, whose activity levels were comparably stable. This is in line with evidence suggesting that promoters are more resistant to short-term and evolutionary-scale perturbations than enhancers. The origin of promoter robustness can be partly explained by enhancer redundancy (“shadow enhancers”), 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 modest differences in the levels of poly(A)+ RNAs. This is consistent with findings in genetically distinct LCLs and yeast.

Individual cells of a tissue type can be highly heterogeneous in terms of their functional genomic features. Phenotypic plasticity enabled by heterogeneity allows a more effective

response to unpredictable external exposures, promoting survival. The observed variability might be 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. 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.

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 a correlation between 5-FU sensitivity and *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. As we did not assess these factors, we cannot exclude their confounding effects. Our study suggests that using the triangle study model, i.e. including the measurement of baseline RNA levels into pharmacogenomic study design, would be beneficial. The limited number of isogenic cells limit our ability to extrapolate our findings to large panels of LCL; hence, studies on a large number of isogenic lines would be desirable.

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.

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 key aspects of biomedical research: the development of emerging 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.

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 to 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 Laboratory 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.

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

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

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

DEBRECENI EGYETEM Egyetemi és Nemzeti Könyvtár