Identification and characterisation of long non-coding RNAs expressed downstream of EGF-induced signalling programme

A thesis submitted to The University of Manchester for the degree of PhD

in the Faculty of Life Sciences

2015

Karol Piotr Nowicki-Osuch

Faculty of Life Sciences
# Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>2</td>
</tr>
<tr>
<td>List of figures</td>
<td>6</td>
</tr>
<tr>
<td>List of tables</td>
<td>9</td>
</tr>
<tr>
<td>List of supplementary materials</td>
<td>10</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>11</td>
</tr>
<tr>
<td>Abstract</td>
<td>13</td>
</tr>
<tr>
<td>Declaration</td>
<td>14</td>
</tr>
<tr>
<td>Copyright Statement</td>
<td>15</td>
</tr>
<tr>
<td>Contribution from other co-workers</td>
<td>16</td>
</tr>
<tr>
<td>Acknowledgment</td>
<td>17</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>18</td>
</tr>
<tr>
<td>1.1. Structure of the eukaryotic genome</td>
<td>18</td>
</tr>
<tr>
<td>1.1.1. The basic structure and diversity of eukaryotic genes</td>
<td>20</td>
</tr>
<tr>
<td>1.1.2. Chromatin</td>
<td>25</td>
</tr>
<tr>
<td>1.2. Regulation of the expression of eukaryotic genes</td>
<td>31</td>
</tr>
<tr>
<td>1.2.1. Structure of eukaryotic promoter</td>
<td>32</td>
</tr>
<tr>
<td>1.2.2. Transcription initiation</td>
<td>34</td>
</tr>
<tr>
<td>1.2.3. Sequence-specific transcription factors</td>
<td>35</td>
</tr>
<tr>
<td>1.3. The role of the regulation of gene expression in cell growth, migration and cancer</td>
<td>37</td>
</tr>
<tr>
<td>1.3.1. The relationship between cell growth and cancer</td>
<td>37</td>
</tr>
<tr>
<td>1.3.2. Mitogen-activated protein kinase pathway</td>
<td>40</td>
</tr>
<tr>
<td>1.3.3. Regulation of gene expression by the ERK pathway</td>
<td>43</td>
</tr>
<tr>
<td>1.3.4. The expression programme downstream of ERK signalling pathway</td>
<td>44</td>
</tr>
<tr>
<td>1.3.5. The ERK signalling pathway and cancer</td>
<td>49</td>
</tr>
<tr>
<td>1.3.6. Immortalised breast cell lines as models of cell growth and migration</td>
<td>50</td>
</tr>
</tbody>
</table>
3.2. EGF-dependant expression of long noncoding RNAs

3.2.1. Identification of long noncoding RNAs expressed in EGF-dependant manner

3.2.2. Long intergenic non-coding RNAs show characteristics of IE genes

3.2.3. Characterisation of the expression profile of EGF-regulated genes in the primary cancer samples

3.3. Chapter summary

4. Characterisation of EIN1 – EGF-inducible non-coding RNA

4.1. Genomic locus of EIN1

4.1.1. Identification of genomic features associated with EIN1

4.1.2. Investigation of the structure of EIN1 transcripts

4.1.3. Expression profile of EIN1 in human tissues

4.2. Characterisation of the protein-coding potential and localisation of EIN1

4.2.1. Computational investigation of coding potential of EIN1

4.2.2. Biochemical investigation of the coding potential of EIN1

4.2.3. The cellular localisation of EIN1

4.3. Investigation EIN1 transcription regulation by ERK1/2 signalling pathway

4.3.1. Transcription of EIN1 in mammary gland cell lines and in cancer data

4.3.2. EIN1 transcription is MEK1/2-dependant and independent of new protein synthesis

4.4. Chapter summary

5. The effect of EIN1 depletion on the EGF-inducible gene expression profile

5.1. Establishment of CRISPRi expression regulation system

5.1.1. The outline CRISPRi approach

5.1.2. Validation of the efficiency of EIN1 knockdown

5.2. Investigation of EGF-regulated expression profile after EIN1 downregulation

5.2.1. Validation of RNA-seq

5.2.2. Expression of protein-coding genes after EGF-stimulation
5.2.3. Effect of $EIN1$ downregulation on the expression of EGF-regulated genes . 154
5.3. Chapter summary ........................................................................................................ 166
6. Discussion .................................................................................................................... 168
7. References .................................................................................................................. 178

Word count: 49,339
List of figures

Figure 1.1.1. RNApolII-dependent transcription and maturation......................................................21
Figure 1.1.2-1. Basic architecture of nucleosome..............................................................................27
Figure 1.2.1-1. Early events during the assembly of preinitiation complex (PIC). ..................33
Figure 1.3.2-1. An overview of the ERK1/2 signaling cascade......................................................42
Figure 1.3.4-1. Pattern of gene expression downstream of ERK pathway. .................................46
Figure 1.4.3-1. Regulation of expression by lncRNAs.................................................................57
Figure 3.1.1-1. Quality control of the sequencing data from the Nugen and Truseq library preparation kits........................................................................................................71
Figure 3.1.1-2. Comparison of divergence between individual repeats of the Truseq libraries..........................................................................................................................72
Figure 3.1.2-1. Quality control of the high-depth RNA-seq.........................................................74
Figure 3.1.2-2. Workflow of the RNA-seq data analysis.................................................................77
Figure 3.1.3-1. Statistical analysis of the expression of EGF-regulated genes............................79
Figure 3.1.3-2. Examples of EGF-regulated IE genes.................................................................80
Figure 3.1.3–3. The analysis of functional correlation of EGF-regulated genes in MCF10A cell line......................................................................................................................83
Figure 3.1.3-4. Expression of EGF-regulated genes over time......................................................85
Figure 3.2.1-1. Distribution of differentially expressed genes......................................................88
Figure 3.2.1-2. Average profiles of RNA-seq read density around gene TSS...........................90
Figure 3.2.1-3. RNA-seq read density around TSS of individual genes......................................91
Figure 3.2.1-4. Average profiles of RNApolIII ChIP read density around the TSS of EGF-regulated genes...........................................................................................................94
Figure 3.2.2-1. Average profiles of lincRNA expression over an extended EGF time course .................................................................................................................................97
Figure 3.2.2-2. Distribution of ELK1 binding sites in the proximity of EGF-regulated genes...............................................................................................................................99
Figure 3.2.2-3. Cumulative distribution of the distance between ELK1 binding sites and selected classes of genes. ................................................................. 100

Figure 3.2.3-1. Expression of EGF-regulated genes in the primary tumour samples – EBI data ........................................................................................................ 103

Figure 3.2.3-2. Expression of selected IE genes in the primary tumour samples - EBI data ........................................................................................................ 104

Figure 3.2.3-3. Expression of EGF-regulated genes in the primary tumour samples ... 106

Figure 4.1.1-1. Sashimi plot of EIN1 structure and expression ......................... 110

Figure 4.1.1-2 Sashimi plot of EIN1 structure and expression in total RNA-seq .... 112

Figure 4.1.1-3. Genomic features of EIN1 locus ............................................. 113

Figure 4.1.2-1. Identification of EIN1 5’ ends .................................................. 116

Figure 4.1.2-2. Identification of EIN1 5’ ends – 5’ RACE ............................... 118

Figure 4.1.3-1. RP11-7F17.7 expression in normal tissues ............................ 120

Figure 4.1.3-2. SNPs associated with EIN1 expression – cis-eQTLs ............... 123

Figure 4.2.1-1. Computational analysis of EIN1 coding potential .................. 126

Figure 4.2.2-1. Sucrose cushion density centrifugation analysis of EIN1 interaction with ribosomes ................................................................. 128

Figure 4.2.3-1. EIN1 preferentially localises to the nucleus in MCF10A and MCF7 cells lines ................................................................................................. 130

Figure 4.3.1-1. EGF-dependant expression of FOS and EIN1 in MCF10A, MCF7 and MDA-MB-231 cell lines and in primary cancer cells ....................... 132

Figure 4.3.2-1. The expression of EIN1 in the presence of MEK inhibitor .......... 135

Figure 4.3.2-2. The expression of EIN1 is independent of protein synthesis ....... 136

Figure 5.1.1-1. CRISPRi-driven regulation of gene expression ....................... 142

Figure 5.1.2-1. Validation of CRISPRi-driven knockdown of EIN1 ................. 144

Figure 5.2.1-1. Quality control of RNA-seq .................................................. 148

Figure 5.2.1-2. Validation of EIN1 downregulation ..................................... 150

Figure 5.2.2-1. EGF-regulated expression of genes ........................................ 153

7 | P a g e
Figure 5.2.3-1. Scatter plots of the effect of EIN1 downregulation on the expression of EGF-regulated genes ................................................................. 155

Figure 5.2.3-3. EGF-dependant expression of genes that are affected by EIN1 downregulation at 180 minutes after EGF stimulation. ............................................. 159

Figure 5.2.3-4. Normalized expression of genes affected by EIN1 downregulation at 180 minutes after EGF stimulation. ............................................................. 160

Figure 5.2.3-5. Cellular processes associated with the genes regulated by EIN1 ........ 162

Figure 5.2.3-6. Diseases associated with the genes regulated by EIN1 .................. 163

Figure 5.2.3-7. The effect of EIN1 on the EGF-regulated cellular movement programme. ..................................................................................... 165

Figure 6.1.1-1. Proposed mechanism of EIN1 action ............................................. 175
List of tables

Table 1.1.1-1: Summary of the classes of known and putative genes present in human genome. .................................................................23

Table 1.1.2-1: Summary of the major histone modification marks for which associated function is known.........................................................29

Table 1.3.1-1: Summary of the cell signalling pathways associated with carcinogenic transformation .............................................................39

Table 2.4.3-1: Lentiviral plasmids used in the study. .........................................................65

Table 3.1.1-1: Summary of the mapping statistics for Truseq and Nugen sequencing libraries........................................................................69

Table 3.1.2-1 Summary of the mapping statistics for high-depth RNA-seq................75

Table 3.1.3-1: Summary of the upstream regulator analysis of genes affected by stimulation of MCF10A cells with EGF........................................82

Table 4.1.3-1: Summary of the SNPs significantly associated with EIN1 expression – cis eQTLs ........................................................................122

Table 5.2.1-1: Summary of the mapping statistics for RNA-seq after EIN1 downregulation........................................................................147
List of supplementary materials

Supplementary file 1. The summary of the differential expression analysis of the high-depth nuclear RNA-seq.

Supplementary file 2. The Ingenuity Pathway analysis output for the upstream regulators of EGF-regulated genes.

Supplementary file 3. The average distribution of sequencing reads around TSSs of differentially expressed genes.

Supplementary file 4. The expression data for the breast cancer samples (Horvath et al., 2013).

Supplementary file 5. The expression data for the lung cancer samples (Seo et al., 2012).

Supplementary file 6. The summary of the differential expression analysis of the total cell RNA-seq after EIN1 depletion.

Supplementary files 7 and 8. Line plot of expression of genes regulated by EGF stimulation.


Supplementary files 10-12 Ingenuity Pathways Analysis summary for analysis of EGF induced data.

Supplementary files 13-16 Ingenuity Pathways Analysis summary for analysis of EIN1-regulated genes.

Supplementary Materials file 1. List of primers used.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>Angstrom (1 Å = 10-10 metre)</td>
</tr>
<tr>
<td>aa</td>
<td>amino acid</td>
</tr>
<tr>
<td>AP-1</td>
<td>Activation Protein 1</td>
</tr>
<tr>
<td>BFP</td>
<td>Blue Fluorescent Protein</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>BRE</td>
<td>TFIIB Recognition Element</td>
</tr>
<tr>
<td>CAGE</td>
<td>Cap Analysis Gene Expression</td>
</tr>
<tr>
<td>Cas9</td>
<td>RISPR Associated Protein 9</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CHART</td>
<td>Capture Hybridization Analysis of NRA Targets</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin immunoprecipitation followed by next generation sequencing</td>
</tr>
<tr>
<td>ChIRP-seq</td>
<td>Chromatin Isolation by RNA Purification</td>
</tr>
<tr>
<td>circRNA</td>
<td>Circular RNA</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered Regularly Interspaced Short Palindromic Repeats</td>
</tr>
<tr>
<td>CRISPRi</td>
<td>Interference CRISPR</td>
</tr>
<tr>
<td>dCas9</td>
<td>catalytically Dead Cas9</td>
</tr>
<tr>
<td>DE gene</td>
<td>Delayed-Early gene</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>dsDNA</td>
<td>double stranded DNA</td>
</tr>
<tr>
<td>DUSP</td>
<td>Dual Specificity Phosphatase</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein-Barr virus</td>
</tr>
<tr>
<td>EGF</td>
<td>Epidermal Growth Factor</td>
</tr>
<tr>
<td>EIN1</td>
<td>EGF-induced non-coding RNA 1</td>
</tr>
<tr>
<td>ERK</td>
<td>Extracellular signal-Regulated protein Kinase</td>
</tr>
<tr>
<td>eRNA</td>
<td>Enhancer RNA</td>
</tr>
<tr>
<td>ETS</td>
<td>E26 Transformation-Specific or E-Twenty-Six</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments per kilobase per million reads</td>
</tr>
<tr>
<td>GEF</td>
<td>Guanine nucleotide Exchange Factor</td>
</tr>
<tr>
<td>Gro-seq</td>
<td>Global Run-On Sequencing</td>
</tr>
<tr>
<td>ID-miR</td>
<td>Immediate-Downregulated miRNA</td>
</tr>
<tr>
<td>IE gene</td>
<td>Immediate-Early gene</td>
</tr>
<tr>
<td>Inr</td>
<td>Initiator consensus sequence</td>
</tr>
<tr>
<td>kb(p)</td>
<td>Kilo base pair</td>
</tr>
<tr>
<td>KRAB</td>
<td>Krüppel Associated Box</td>
</tr>
<tr>
<td>lincRNA</td>
<td>Long Intergenic Non-Coding RNA</td>
</tr>
<tr>
<td>IncRNA</td>
<td>Long Non-Coding RNA</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>MAP2K</td>
<td>Mitogen-Activated Protein Kinase Kinase Kinase Kinase Kinase</td>
</tr>
<tr>
<td>MAP3K</td>
<td>Mitogen-Activated Protein Kinase Kinase Kinase Kinase Kinase Kinase</td>
</tr>
<tr>
<td>MAPK</td>
<td>Mitogen-Activated Protein Kinase Kinase Kinase Kinase Kinase Kinase Kinase</td>
</tr>
<tr>
<td>miRNA</td>
<td>micro RNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>Non-Coding RNA</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>PIC</td>
<td>Pre-Initiation Complex</td>
</tr>
<tr>
<td>RAP</td>
<td>RNA Antisense Purification</td>
</tr>
<tr>
<td>RIP</td>
<td>RNA Precipitation</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RNA FISH</td>
<td>RNA Fluorescence in situ Hybridisation</td>
</tr>
<tr>
<td>RNApolI</td>
<td>RNA Polymerase I</td>
</tr>
<tr>
<td>RNApolII</td>
<td>RNA Polymerase II</td>
</tr>
<tr>
<td>RNApolIII</td>
<td>RNA Polymerase III</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SOS</td>
<td>Son of Sevenless</td>
</tr>
<tr>
<td>SR gene</td>
<td>Secondary Response gene</td>
</tr>
<tr>
<td>SRE</td>
<td>Serum Response Element</td>
</tr>
<tr>
<td>ssDNA</td>
<td>single stranded DNA</td>
</tr>
<tr>
<td>TBP</td>
<td>TATA Binding Protein</td>
</tr>
<tr>
<td>TCF</td>
<td>Ternary Complex Factor</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TFIID</td>
<td>general Transcription Factor IID</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription Start Site</td>
</tr>
<tr>
<td>TTS</td>
<td>Transcription Termination Site</td>
</tr>
<tr>
<td>uaRNA</td>
<td>Upstream Antisense RNA</td>
</tr>
</tbody>
</table>
Abstract

The University of Manchester
Karol Piotr Nowicki-Osuch
PhD
Identification and characterisation of long non-coding RNAs expressed downstream of EGF-induced signalling programme.
27/09/2015

It has recently become apparent that cells encode a large number of novel non-protein-coding genes called long non-coding RNAs (lncRNAs). Whilst the biological function of many lncRNAs remains unknown, recent evidence has suggested that lncRNAs may be important regulators of cellular growth, differentiation and may play a significant role in cancer. Epidermal growth factor (EGF) – an activator of the ERK1/2 signalling cascade – is an important spatio-temporal regulator of transcription and, ultimately, of cellular growth and movement. EGF stimulation triggers a wave-like expression of immediate-early genes (IE genes), followed by delayed-early genes (DE genes) and secondary-response genes (SR genes). Over the years, considerable effort has been made to unravel the regulatory loops downstream of EGF signalling. This study investigated whether lncRNAs are sensitive to EGF signalling and whether they play a role in the transcriptional programme associated with EGF signalling.

In order to identify lncRNAs regulated by EGF signalling, I sequenced nuclear RNA in the presence or absence of EGF stimulation. RNA-seq data showed that 173 lncRNAs are upregulated by EGF, of which 89 were intergenic lncRNAs (lincRNAs). The time-dependent expression profile of EGF-upregulated lincRNAs followed the well-established expression pattern of IE genes. Finally, investigation of the expression of lincRNAs in primary breast and lung cancer cells showed that EGF-upregulated lincRNAs were differentially expressed in cancer. The EGF-dependent induction profile and cancer enrichment were particularly strong for one of the transcripts – EGF-induced lncRNA 1 (EIN1) – and I selected it for further studies.

Firstly, using bioinformatics and biochemical approaches, I confirmed the non-coding status of the EIN1 transcript. Secondly, I confirmed that EIN1 transcription is ERK1/2-dependent and is independent of protein synthesis. Investigation of EIN1 expression in normal tissues showed its high enrichment in the human cardiovascular system. At the cellular level, the EIN1 transcript was predominantly found in the nucleus. Functionally, the depletion of endogenous EIN1 transcripts (using the newly developed CRISPRi approach) led to changes in the EGF-dependent transcription programme. EIN1 downregulation resulted in the addition of normally EGF-independent genes into the EGF-dependent expression programme.

Collectively, these results show that EGF (via the ERK1/2 pathway) can regulate transcription of lincRNAs. The EIN1 example suggests that lincRNAs may play a crucial role in the modulation of the EGF-dependent expression programme by limiting of the scope of the programme.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy.
Contribution from other co-workers

The RNA-seq library preparations were performed by Stacey Holden at the Genomic Core Facility at the University of Manchester. The initial processing (Casava software) of the raw data was performed by Peter Briggs at the Bioinformatics Core Facility at the University of Manchester.

The cDNA preparation and the use of the BioMark HD System (Fluidigm) were performed by Claire Morrisroe at the Genomic Core Facility at the University of Manchester.

The investigation of the effect of MEK1/2 inhibition on the expression of \textit{EINI} was performed by Megan Moruzzi.

The computation analysis of the data was performed on the Computational Shared Facility at The University of Manchester.

The access to the IPA software (Qiagen) was provided by the funding provided by a Wellcome Trust Institutional Strategic Support Fund (ISSF) award (097820) to the University of Manchester.
Acknowledgment

First of all, I would like to thank my supervisor, Andy Sharrocks, for the guidance, help and understanding during this PhD. I would also like to thank all members of the Sharrocks lab for a great working environment and help with many aspects of the entertaining life of a PhD student. Special thanks go to Elisa Aguilar-Martinez, Karren Palmer, Shen-Hsi Yang and Ed Britton for great discussions, patience and help with the day-to-day life of a scientist. I would like to thank everyone in the faculty and beyond (Larry Page in particular) who helped me with the technical aspects of the PhD.

I would also like to thank my friends for making the last four years full of fun and not so full of normality. After all, when you are ‘changing the world one protein at the time’, it is always good to do it in a great company, regardless if the articles are in the correct places or not. Special thanks go to Pari and Alan, a couple of crazy friends of mine who have been there for me in good and bad times and without their help this work would not be accomplished. I would also like to thank Pablo for being a great friend and flatmate and the first person that ‘changes the world one protein at the time’.

Most importantly, I would like to thank my parents for supporting me in all of my decision. Without your support I would not be where I am now. Dziękuję.

Finally, I would like to thank Wellcome Trust for generous finding.
1. Introduction

Nearly 60 years ago Francis Crick defined the principle hypothesis of molecular biology – the Central Dogma of Molecular Biology. In the original argument behind the central dogma, Crick postulated that: (1) there are nine possible directions of information transfer between DNA, RNA and Protein, (2) only three of them (DNA to DNA, DNA to RNA and RNA to Protein are universal for all of the organisms, (3) another three (RNA to RNA, RNA to DNA and DNA to Protein) are present only in selected organisms and systems, and (4) the final three scenarios (Protein to Protein, Protein to DNA and Protein to RNA) are to be never observed. The idea neatly explained the flow of information in the biological systems and the general principle behind it has held true ever since. The hypothesis, however, stipulated that proteins are the main functional element of molecular systems. Ever since the description of the central dogma, a growing body of evidence has identified that in addition to protein, RNA can have a functional role in the cellular process. Recently, the advent of next-generation sequencing technologies led to observation that over 80% of the human genome is transcribed into long non-coding RNA – RNA products with ‘functional potential’ (Djebali et al., 2012). The following sections will explore the complexity of human genome and regulatory network behind it paying particular attention to long non-coding RNAs.

1.1. Structure of the eukaryotic genome

Since the 19th century discoveries of Gregor Mendel, genes were associated with heredity and, from the evolutionary point of view, genes are the basic unit of hereditary information. However, it took nearly a century to progress our understanding of the gene from a philosophical ‘unit’ to a physical entity encoded in DNA. The current, molecular biology-based definition of a gene states that it is ‘a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions’ (Pearson, 2006). The genome is the entirety of all of the hereditable information present in each organism. The genome of the majority of living organisms is encoded in DNA (deoxyribonucleic acid); the exceptions are certain viruses where their genome is encoded in RNA (ribonucleic acid). DNA
consists of two long polymers of four nucleotides (guanine, cytosine, adenine and thymine) which form a double helical structure. The basic properties of DNA structure – its linearity and sequential structure – dictate how the genetic information is encoded within DNA, i.e. genetic information is stored in the sequence of nucleotides. However, not all of the nucleotides in the DNA encode genes.

The genomes of all of the organisms are not continuous stretches of genes; rather genes are separated by non-coding fragments of DNA. RNA polymerases are enzymes which transcribe information from DNA to RNA and ribosomes are large molecular machineries which translate RNA to protein. In many unicellular organisms, such as bacteria and archaea, and in addition to many simple eukaryotes (e.g. yeast), genes are tightly packed within the DNA. The amount of spacer non-coding DNA is usually limited, and it mainly carries information responsible for the regulation of gene expression (Gregory, 2005). On the other hand, the DNA of many evolutionary advanced eukaryotes contains vast stretches of nucleotides which do not encode for genes.

Over a decade ago the first draft of the human genome was published (International Human Genome Sequencing Consortium, 2004). It was a grand collaborative achievement of biologists, which gave scientists an excellent tool to understand the complexity of the genome. The first unexpected discovery of the Human Genome Sequencing Consortium was that the human genome encodes only ~20,000 protein-coding genes. This stands in stark opposition to the original estimate of the number of protein-coding genes; just before the human genome draft was published, scientists estimated that around 50,000–100,000 genes are needed to explain the complexity of our organism (Schuler et al., 1996). Human DNA is 2.85 billion base pairs (bp) in length and only ~1.5% of it encodes protein-coding genes (International Human Genome Sequencing Consortium, 2004). This raises the question about the role of the remaining ~98% of the DNA. For years, non-coding DNA was termed ‘junk DNA’. It, apparently, did not have any function and was an evolutionary relic. Recent evidence, however, put that hypothesis in question. Development of novel sequencing technologies showed that the majority of human DNA is actively transcribed (Djebali et al., 2012). However, before we explore the evidence behind the function of non-coding DNA, I will describe the basic structural properties of the genome. The following chapters outline the properties of eukaryotic genes (chapter
1.1.1) and the bottom-up structure of the genome (chapter 1.1.2), two features of which exploration is essential for understanding of the non-coding DNA.

1.1.1. The basic structure and diversity of eukaryotic genes

The structure of genomes varies across different organisms. The prokaryotic genome is the most basic one. Bacteria normally have only one circular chromosome that exhibits few modifications. On the other hand, the eukaryotic genome with its variety of histones, modifications and regulation is the most complex genome (described in chapter 1.1.2). Similarly to the variety of the genomic structures, a variety of gene structures exists across domains. In the case of eukaryotic genes, the most striking feature is their non-continuous nature. The vast majority of genes of higher eukaryotes are made of segments – exons – which are separated by introns (Figure 1.1.1-1A). When eukaryotic genes are transcribed from the DNA sequence, intronic sequences are removed from the RNA, through the process of splicing, to form mature RNA (Figure 1.1.1-1A). Recent evidence indicate that splicing can be a co-transcriptional processes, i.e. introns are removed from the growing RNA molecule before the whole RNA molecule is synthesised (Tilgner et al., 2012) or can take place after the mature RNA had been synthesised (Boutz et al., 2014). In the latter, transcripts with ‘detained introns’ are formed before intron removal. This suggests that, at least in some cases, splicing occurs after termination of transcription.

The number of exons in human genes varies between one per gene and 363 per gene (for Titin gene) with an average of over 10 exons per gene (Harrow et al., 2012). Additionally, various exons of the same gene can be connected in different ways and as a result they might form different mature transcripts – isoforms. Different isoforms of the same gene are formed during alternative splicing. The latest release of the human genome annotation compiled by GENCODE lists 19,814 protein-coding genes and 79,712 protein-coding isoform (Harrow et al., 2012). On average each gene has four different isoforms, however the function and significance of this complexity is largely unknown.
Figure 1.1.1. RNAPolII-dependent transcription and maturation.

A) A simplified outline of the maturation of mRNA. B) Structure of a mature protein-coding mRNA. ORF – Open Reading Frame, UTR – UnTranslated Region. Codon are in the RNA sequence.
Classically, the final, functional products of genes are proteins. Each spliced protein-coding gene contains an open reading frame (ORF, Figure 1.1.1-1B). The ORF is the part of the gene sequence that is translated into an amino acid sequence. In eukaryotes, the majority of ORFs begins with an AUG codon for methionine (with few exceptions (Ivanov et al., 2011)) and is finished with one of the stop codons (UAA, UGA, UAG). Sequences upstream and downstream of ORFs are termed untranslated regions (UTRs). The presence of an ORF is a defining element of gene structure, however recent evidence indicates that the human genome contains over 35,000 genes which are transcribed from the DNA to RNA sequence but are not translated into protein sequences (Harrow et al., 2012). These novel transcripts are termed non-coding RNAs. Ribosomal RNAs (rRNA) and transfer RNA (tRNA) are the earliest examples of the non-coding RNAs. Recently, the ENCODE (Encyclopedia of DNA Elements) data suggest that over 80% of human DNA is transcribed into some form of RNA and only 1.5% of human DNA encodes protein-coding sequences (Dunham et al., 2012). The vast majority of non-coding RNAs are long non-coding RNAs (lncRNAs, see Table 1.1.1-1 for summary). However, the ENCODE conclusions about the functionality of the lncRNAs have been fiercely debated (Doolittle, 2013; Doolittle et al., 2014; Graur et al., 2013). From an evolutionary point of view, it is difficult to defend the ENCODE position of the high functionality of lncRNAs but low conservation across organisms. Nevertheless, many examples of lncRNAs are present in the literature and new roles are constantly being assigned to them, which indicates that at least a minority are indeed functional (see section 1.4 for further details of the role and function of lncRNAs).
Table 1.1.1-1: Summary of the classes of known and putative genes present in human genome.

Main classes of eukaryotic RNA species. Subclasses for the long non-coding RNAs are compiled from gencode (Harrow et al., 2012).

<table>
<thead>
<tr>
<th>Class</th>
<th>Subclass</th>
<th>Length of the mature transcript</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Coding</td>
<td>-</td>
<td>&gt;300 nt</td>
<td>Transcripts contains ORF, encode proteins</td>
</tr>
<tr>
<td>rRNA</td>
<td>-</td>
<td>121 nt (5S) – 5070nt (28S)</td>
<td>ribosomal RNA, structural elements of ribosomes, take part in translation</td>
</tr>
<tr>
<td>Small non-coding RNA</td>
<td>miRNA</td>
<td>~22 nt</td>
<td>microRNA, regulation of mRNA level (Bartel, 2004)</td>
</tr>
<tr>
<td></td>
<td>piRNA</td>
<td>26–31 nt</td>
<td>Piwi-interacting RNA, associated with gene silencing (Meister, 2013)</td>
</tr>
<tr>
<td></td>
<td>snoRNA</td>
<td>&lt;300 nt</td>
<td>small nucleolar RNA, located in the nucleolus, required for processing of rRNA, tRNA and snRNAs (Matera et al., 2007)</td>
</tr>
<tr>
<td></td>
<td>snRNA</td>
<td>&lt;300 nt</td>
<td>small nuclear RNA, take part in mRNA splicing (Matera et al., 2007)</td>
</tr>
<tr>
<td></td>
<td>3' overlapping ncRNA</td>
<td>&lt;200 nt</td>
<td>short transcripts transcribed from the 3' UTR, unknown function (Harrow et al., 2012)</td>
</tr>
<tr>
<td></td>
<td>tRNA</td>
<td>75–95 nt</td>
<td>transfer RNA, function as carriers of amino acids and participate in protein synthesis (Phizicky and Hopper, 2010)</td>
</tr>
<tr>
<td>Long non-coding RNA</td>
<td>lincRNA</td>
<td>&gt;200 nt</td>
<td>long intergenic RNA, diverse functions</td>
</tr>
<tr>
<td></td>
<td>antisense</td>
<td>&gt;200 nt</td>
<td>Transcripts that over overlap any coding exon of a locus on the opposite strand</td>
</tr>
<tr>
<td></td>
<td>pseudogene</td>
<td>&gt;200 nt</td>
<td>Similar to known proteins but contain a frameshift and/or stop codon(s) which disrupts the ORF</td>
</tr>
<tr>
<td></td>
<td>sense intronic</td>
<td>&gt;200 nt</td>
<td>A long non-coding transcript in introns of a coding gene that does not overlap any exons</td>
</tr>
<tr>
<td></td>
<td>sense overlapping</td>
<td>&gt;200 nt</td>
<td>A long non-coding transcript that contains a coding gene in its intron on the same strand</td>
</tr>
</tbody>
</table>
Genomic loci of eukaryotic genes have multiple structural features which identify their presence. First of all, each gene has a defined transcription start site (TSS). TSS is a specific point in the genome where the transcription of a given transcript can be initiated by RNA polymerase enzymes. Global cap analysis of gene expression (CAGE) showed that there are different classes of TSS (Carninci et al., 2006; Forrest et al., 2014). Broad promoters (TSS spans up to 100 nucleotides) are the dominant form and ‘single dominant peak’ promoters are in the minority (Carninci et al., 2006). Recently, however, nascent RNA-seq experiments showed that the majority of ‘single dominant peak’ promoters may in fact be the dominant form (Scruggs et al., 2015). Immediately after the initiation of transcription, a 5’ cap is added to the 5’ end of the new mRNA sequence. The 5’ cap consist of a guanine nucleotide connected to the first nucleotide of mRNA via unusual 5’ to 5’ triphosphate linkage (Shatkin, 1976). The 3’ end of the mature mRNA is also modified through the addition of a chain of adenine nucleotides forming the polyA tail. Polyadenylation is performed by large machinery that first recognises the polyadenylation signal on the mRNA (AAUAAA), cleaves the growing mRNA and finally adds 100–150 adenine nucleotides (reviewed in Zhao et al., 1999).

Immediately upstream of the TSS lays the promoter of the genes where the transcription initiation machinery is constructed. Classically, promoters are defined by the presence of a highly conserved TATA box (sequence TATAWAWR) located 25–35 nt upstream of the TSS (Basehoar et al., 2004). However, recent evidence indicate that the TATA box is located in only around 10% (Carninci et al., 2006) to 30 % (Suzuki et al., 2001) of promoters and that TATA box promoters are associated with tissue-specific genes highly conserved across organisms (Carninci et al., 2006). Currently, large projects are underway in order to identify TSS sites (and by association, promoters) using genome-wide CAGE. FANTOM consortium has identified over 200,000 core promoters (Forrest et al., 2014) and over 60,000 transcribed enhancer (Andersson et al., 2014; Arner et al., 2015). This is a significant increase in the number of potential transcripts in the genome. Nevertheless, similarly to the ENCODE data, the FANTOM approaches suffer from lack of functional association between the promoter and the phenotypic observation and further studies are needed in order to identify whether such high numbers of transcripts are functional.
In the case of many lncRNAs it is difficult to distinguish whether the functional effect originates from the transcript or the process of transcription itself triggers a downstream effect. The lncRNAs can work in trans (exercise their function away from the site of transcription) or in cis (exercise their function near the site of transcription) (Pelechano and Steinmetz, 2013). The distinction between the process of transcription and transcript itself is the most noticeable in the case of cis-acting lncRNAs. For example, the transcription of antisense transcripts might lead to a collision of sense- and antisense-transcribing RNApolIII enzymes what results in termination of transcription of the sense (protein-coding) transcript (Pelechano and Steinmetz, 2013). Additionally, the process of transcription and transcripts themselves might trigger changes in the DNA itself. For example, the R-loop (RNA:DNA hybrids formed during transcription) might lead to DNA damage (Hamperl and Cimprich, 2014).

Over the recent years, it has become evident that the number of genes in the human genome has been underestimated. Although, since the publishing of the first draft of the human genome sequence, the number of protein-coding genes has remained relatively stable (at around 20,000), recent discoveries have shown that ‘junk DNA’ might actually be harbouring a large number of genes that are not translated into conventional protein sequence. At least on the surface, the lncRNAs seem to share structural features with the protein-coding genes. However, further investigation is required in order to understand their biological importance. Importantly, the regulation of lncRNAs expression is a fairly unexplored area of research. In order to understand the regulation of lncRNAs, in the next section I will describe how the expression of eukaryotic genes is regulated.

1.1.2. Chromatin

The physical length of human DNA can be estimated to be around 2 metres. This creates a significant challenge in order for the cells to be able to store the DNA in an elongated form. Hence, DNA has a number of helper proteins – histones – which compact DNA to the size which is manageable for cells. Polymeric complexes of histones and DNA form chromatin. On the most simplistic, global view, eukaryotic chromatin can be divided into: 1) euchromatin – a loosely packed form of chromatin which is rich in actively transcribed genes and 2) heterochromatin – a tightly packed
form of chromatin which is usually transcriptionally inactive. Around 90% of chromatin in a given cell is euchromatin (International Human Genome Sequencing Consortium, 2004). Histones are highly conserved, small, positively charged proteins which form octamers around which DNA is wrapped; the whole DNA-histone octamer complex is called a nucleosome. Each nucleosome is usually built of two molecules of each of four different histones: H2A, H2B, H3 and H4, a fifth histone – the H1, linking histone – binds DNA between nucleosomes (reviewed in Cutter & Hayes, 2015; and Jiang & Pugh, 2009). Structurally, the nucleosome is formed from the \((H3)_{2}(H4)_{2}\) tetramer and two dimers of H2A-H2B histones around which a 147 nucleotide long stretch of DNA is wrapped (Figure 1.1.2-1) (X-ray structure: 2.8 Å resolution – Luger et al., 1997; 1.9 Å resolution – Richmond and Davey, 2003).
Figure 1.1.2-1. Basic architecture of nucleosome.
A simplified depiction of a nucleosome. Only the core histone (H2A, H2B, H3 and H4) are shown.
In addition to the linear topography of the histone-DNA interaction, nucleosomes form highly-structured 3-dimensional structures in the nucleus of the cell. Low resolution X-ray diffraction of DNA structures, where previously visualised nucleosome structure was superimposed, show that nucleosomes are tightly packed on one another to form chromatin fibres (Schalch et al., 2005). These structural features explain how it is possible for cell to compact nearly 2 metres of DNA into the nucleus which has a diameter of around 20 µm.

In addition to the compacting role, the presence of histones must also be considered during transcription and DNA replication processes. In order for RNA polymerases to access DNA, DNA must be released from nucleosomes. It is possible because chromatin is a highly flexible entity and access to the DNA is regulated by the ‘histone code’ (Jenuwein and Allis, 2001). Histones are proteins which can be highly regulated by covalent modification such as acetylation, phosphorylation, methylation, ubiquitination and sumoylation (for reviews see: Hawkins et al., 2010; Nathan et al., 2003; Weake and Workman, 2008; Yuan, 2012; Zhou et al., 2011). Global, systematic studies of the genomic location of H3K4me3 (trimethylated lysine 4 on histone H3), H3K9me3, H3K27me3 and H3K36me3 in human CD4+ T cells showed that these histone marks are distributed in a non-random fashion throughout the human genome and are associated with distinct genomic units (Wang et al., 2008; Wei et al., 2009). These studies employed chromatin immunoprecipitation combined with next-generation sequencing (ChIP-seq), a technique that allows for identification of protein-DNA interactions (at ~200-500 bp resolution) (reviewed in Landt et al., 2012). The H3K4me3 histone mark is associated with promoter/TSS of actively transcribed genes and the H3K36me3 mark is associated with actively transcribed genes, whereas H3K27me3 and H3K9me3 are repressive marks associated with inactive genes (Landt et al., 2012). Subsequent studies on a much higher number of cell lines and on various other histone marks performed by the ENCODE consortium confirmed these findings (Arvey et al., 2012; Djebali et al., 2012; Dunham et al., 2012). The modifications of histones are summarised in Table 1.1.2-1. Furthermore, ChIP-seq experiments showed that histone modifications are dynamic and that they undergo changes during cellular development. For example,
Table 1.1.2-1: Summary of the major histone modification marks for which associated function is known

<table>
<thead>
<tr>
<th>Mark</th>
<th>Signal Characteristics</th>
<th>Example of Modifying Enzyme</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>Peak</td>
<td></td>
<td>Marks elements associated with enhancer and other distal regulatory elements</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Peak</td>
<td>MLL, SET1</td>
<td>Marks elements associated with enhancer and promoter regions</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Peak</td>
<td></td>
<td>Marks elements associated with promoter and transcription start site, together with H3K36me3 marks actively transcribed genes</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>Region</td>
<td>PCAF/GCN5</td>
<td>Marks transcriptionally active elements, primary associated with promoters</td>
</tr>
<tr>
<td>H3K9me1</td>
<td>Peak/region</td>
<td></td>
<td>Marks 5' ends of genes</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>Peak</td>
<td>ESET/SETDB1</td>
<td>Marks transcriptionally inactive elements, associated with heterochromatin and repetitive elements</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Region</td>
<td>GCN5</td>
<td>Marks transcriptionally active promoters/enhancers</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>Region</td>
<td>EZH2</td>
<td>Marks elements associated with repressive domains and silent developmental genes, introduced by polycomb complex</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>Region</td>
<td>SET2</td>
<td>Marks actively transcribed regions of DNA</td>
</tr>
</tbody>
</table>

Notes: For signal characteristics, peaks denotes narrow modification region of up to 500 bp, region denotes continuous modification region up to several thousand base pairs.
Information was compiled from: (Arvey et al., 2012; Hawkins et al., 2010; Nathan et al., 2003; Weake and Workman, 2008; Yuan, 2012; Zhou et al., 2011)
the meta-analysis of the level of H3K27me3 around the protein-coding genes during development of the T cells showed that in mature T cells, the level of H3K27me3 is lower than in naïve T cells (Wei et al., 2009). Simultaneously, at the level of individual genes, the change of histone marks from H3K27me3 to H3K4me3 correlates with the change (increase) of associated transcript levels (Wei et al., 2009).

The modification of chromatin states is undertaken by specialised enzymes. Modifications of the histone code are performed by specific acetyltransferases, deacetylases, methyltransferases, demethylases, serine/threonine kinases, ubiquitinases and deubiquitinases (reviewed in Kouzarides, 2007). The majority of these enzymes show high specificity towards unique sites on histones. For example, a family of MLL (MLL1–MLL5) methyltransferases is responsible for the methylation of lysine 4 on histone H3 (Bannister and Kouzarides, 2005; Kouzarides, 2007). Finally, changes of the histone marks have been associated with cancer (reviewed in Rodríguez-Paredes and Esteller, 2011). For example, loss of function of p300 and CBP (histone acetyltransferases) is associated with leukaemia and cancers of epithelial origin (Sawan and Herceg, 2010; Yang, 2004). Similarly, gain-of-function mutations of the MLL methyltransferases are associated with leukaemia (Hess, 2004).

Nucleosomes are the smallest domains of the chromatin structure. Similarly to proteins, chromatin forms domains of a higher order than nucleosomes. As a result, one might expect that chromatin forms a complex network of interacting domains which are difficult to investigate. Currently, the most widespread approaches to study the 3D structure of chromatin are variations of chromosome conformation capture (3C) (Dekker et al., 2002). 3C technique is conceptually straightforward technique that allows one to identify regions which are close (in 3D space of chromatin, not necessarily in 1D space of DNA chain) to the known DNA region using a proximity ligation approach. With the advent of next-generation sequencing technologies, the 3C technique was expanded to whole genome studies and 4C (chromosome conformation capture-on-chip), 5C (chromosome conformation capture carbon copy), HiC, and Chia-PET technique were developed (review in de Wit and de Laat, 2012). Early HiC studies led to the discovery of topologically associating domains (TADs) – regions of DNA that tend to interact with each other in a pattern conserved between different human-derived cell lines (Dixon et al., 2012; Pope et al., 2014).
Recent developments in the 3D genome studies allow for targeted enrichment of samples for multiple regions of interest. Capture-C (Hughes et al., 2014) and Capture-HiC (Mifsud et al., 2015) were used to study interactions with focus on the promoter regions. The authors were able to significantly enrich samples for the interaction with promoter regions and they were able to annotate vast number of enhancers to their respective genes. Interestingly, the data showed that promoters preferentially interact with intergenic regions within TADs. With the advent of the next-generation techniques we are now able to measure the structure of the genomic interaction in a genome-wide manner. Nevertheless, current techniques are still limited in their resolution (limited by the genomic density of the restriction enzyme site used in the technique). Further developments are needed in order to increase the efficiency, accuracy and resolution.

Chromatin is a highly complex molecular structure which we are only now starting to understand. Currently available data show that chromatin is flexible and can undergo many changes. Different histone modifications define what happens to the DNA sequence hidden within genes. Ultimately, chromatin structure and dynamics dictate how the genes are transcribed. However, chromatin structure is at the top level of gene transcription regulation. Opening of chromatin is the first step in the initiation of transcription and the following sections will describe the details of the regulation of gene transcription.

1.2. Regulation of the expression of eukaryotic genes

Human genome contains around 20,000 protein-coding genes and tens of thousands of lncRNAs. However, not all of the genes are expressed in every cell type. Only around 50% of protein-coding genes are expressed in the given tissue at a given time (Dunham et al., 2012). Furthermore, their expression varies between cells, within the cell, and it varies over time. The key to understand the complexity of eukaryotes lies in the understanding of the regulation of expression. Eukaryotic expression is regulated at almost every stage of gene maturation, including transcription, RNA processing and localisation, and in the case of protein-coding genes – translation. Due to its significance for lncRNAs, the following sections will focus on co-transcriptional regulation of gene expression.
1.2.1. Structure of eukaryotic promoter

The simplified image of a promoter presented in section 1.1.1 can summarise the promoter as a genomic region where transcription of the gene begins. In the eukaryotic genome, the core promoter spans a region within 50 bp from the TSS. Metazoan promoters can be classified into three major types: Type I (‘adult’ promoter, associated with tissue-specific expression in the peripheral tissues), Type II (‘ubiquitous’, associated with housekeeping genes) and Type III (‘developmentally regulated’, associated with genes required for multicellular development and cellular differentiation) (reviewed in Lenhard et al., 2012). The Type I promoter is the most studied promoter of eukaryotic genes. Structurally, type I promoters are the most conserved and they contain the TATA box. TATA-containing promoters can also be characterised by two additional sequence-specific features – initiator sequence (Inr) and TFIIB Recognition Element (BRE, Figure 1.2.1-1). The original Inr sequence was defined as YY(-1)A(+1)NWYY (Smale and Kadonaga, 2003). However, data from CAGE experiments showed preference for CG, CA and TG dinucleotides at the initiator (Carninci et al., 2006; Sandelin et al., 2007). Over the years, two BRE elements have been identified, both positioned next to the TATA box. The first is an upstream BRE element, defined as SSRCGCC (Lagrange et al., 1998) and the second is the downstream BRE element RTDKKKK (Deng and Roberts, 2005).

Historically, the TATA-containing promoters were the main focus of studies. However, a sequence-based search showed that the majority (around 70%) of promoter are enriched for CpG islands and that they do not contain a TATA consensus sequence (Lenhard et al., 2012; Saxonov et al., 2006). CpG islands are defined as regions of the DNA (>200 nt long) where the frequency of CG dinucleotide is high (Deaton and Bird, 2011).

Finally, promoters of actively transcribed gene are strongly structured from the chromatin point of view. It has been observed that promoters are enriched for H2A.Z and H3.3 variant of histones (Ku et al., 2012; Sarma and Reinberg, 2005), the H3K4me3 histone mark is strongly associated with promoters and nucleosomes are strongly ‘phased’ around the promoter with the core promoter being free of nucleosomes (Jiang and Pugh, 2009; Mavrich et al., 2008).
Figure 1.2.1-1. Early events during the assembly of preinitiation complex (PIC).

Seemingly, promoters are a ‘simple’ genomic element where transcriptional machinery is assembled. However, since the gene expression profiles do not follow a single one-fits-all model, it is logical to observe variation in the structure of promoters, a variation that leads to an important regulatory step of transcription – initiation.

1.2.2. Transcription initiation

The RNA polymerases are enzymes that synthesise RNA from DNA templates. The human genome encodes three closely related RNA polymerases: RNA polymerase I (RNApolI, associated with expression of 28S, 18S and 5.8S rRNA, reviewed in Schneider, 2012), RNA polymerase II (RNApolII, associated with expression of mRNA, miRNA, lncRNA, reviewed in Shilatifard et al., 2003) and RNA polymerase III (RNApolIII, associated with expression of 5S rRNA, tRNA, snRNA and snoRNA, reviewed in Haeusler and Engelke, 2006). RNApolII is the most highly studied of these enzymes due to its role in the transcription of mRNA. The recruitment of RNApolII to the promoter region is a complex process which requires the addition of many proteins – transcription factors.

RNApolII is the main RNA polymerase associated with the expression of the protein-coding genes, as well as a vast majority of other genes present in the eukaryotic genome. However, for adequate functioning, RNApolII requires a high number of additional proteins. These helper proteins are called transcription factors (TFs). On the most superficial level, TFs can be split into two groups: sequence-specific TFs and general TFs. The sequence-specific TFs are normally modulators of transcription and the general TFs play a crucial role in the recruitment of the RNApolII to promoters (recently reviewed in Sainsbury et al., 2015).

Initiation of transcription of eukaryotic genes begins with the assembly of the pre-initiation complex (PIC, Figure 1.2.1-1). Classically, in the case of TATA-containing promoters, the assembly of the PIC begins with the binding of the TATA box-binding protein (TBP, subunit of general transcription factor IID (TFIID)) to the TATA box (recently reviewed in Sainsbury et al., 2015). In the case of TATA-less promoters, ChIP-exo experiment showed that TBP occupies the vast majority of actively transcribed genes, including TATA-less promoters (Venters and Pugh, 2013). This suggests a role of TBP in both classes of promoters. Additionally, studies
in yeast showed that mutations of the DNA-binding region of TBP do not affect the transcription of the associated gene (Kamenova et al., 2014). This further complicates the picture and more studies are needed to completely understand the role of TBP on the TATA-less promoters.

In the second step of PIC assembly on the TATA-containing promoters (Figure 1.2.1-1), TBP binding to the TATA-like element causes recruitment of TFIID, followed by the binding of TFIIB and TFIIA general transcription factors. TFIIB binds two motifs downstream and upstream of the TATA-like elements (TFIIB recognition elements – BREs) (Deng and Roberts, 2005; Lagrange et al., 1998). Subsequently, the complex is bound by the TFIIF-RNApolII complex. In the next step, TFIIE and TFIIH are recruited to the PIC and the transcription initiation complex is formed. TFIIE and TFIIH are required for promoter DNA opening (Holstege et al., 1996). The final step of transcription initiation is the phosphorylation of RNApolIII, which triggers the release of RNApolIII from the initiation complex and allows the subsequent synthesis of a new RNA product. In the case of many promoters, RNApolIII tends to leave the core promoter, however it does not synthesise complete mRNA, but rather pauses around 50 nucleotides downstream of the TSS (Venters and Pugh, 2013).

1.2.3. Sequence-specific transcription factors

In addition to the general TFs, the genome of eukaryotic organisms encodes for a vast number of sequence-specific TFs. It has been estimated that around 10% of all protein-coding genes contain a DNA-binding domain (Babu et al., 2004). Sequence-specific TFs are important spatial and temporal regulators of gene expression (reviewed in Spitz and Furlong, 2012). Binding of TFs to the enhancer and/or promoter regions of DNA is undertaken by a specialised protein domain (the DNA-binding domain) that recognises specific 6–12 bp long DNA sequence and can lead to activation or repression of transcription. One of the possible mechanisms used for activation of transcription includes recruitment of histone modifiers or nucleosome remodelling complexes, which destabilise nucleosomes in the vicinity of the TSS (Soufi et al., 2015). Secondly, TFs recruit mediator complex to the promoter region of gene, which in turn interacts with the PIC (Malik and Roeder, 2010). Mediator is a multiprotein complex with four distinguished features termed: head, middle, tail and
the kinase domains. The head and middle domains interact with the PIC, whereas the tail domain interacts with sequence-specific TFs (Malik and Roeder, 2010).

The activity of TFs is regulated in multiple ways. The basic mode of TF regulation is a change of its cellular concentration. An example of such regulation can be seen in transcription factors that are expressed as immediate-early genes (IE genes). The AP-1 transcription factor is usually a heterodimer of products of \textit{JUN} and \textit{FOS} genes. Levels of both of them are relatively low in unstimulated cells, but when cells are stimulated with mitogenic factor, their expression and protein levels increase, which results in the subsequent increase in their activity (Karin et al., 1997). Another powerful way of regulating the activity of TFs is through the post-translational modification of TFs. The most common are: phosphorylation, oligomerisation, cofactor binding, stability modification and subcellular localisation. An example of post-translational modification is phosphorylation of the ELK1 transcription factor by the downstream kinase of the MAPK pathway (reviewed in Sharrocks, 2002; Yang et al., 2013b, also see section 1.3.3). Finally, binding of TFs to DNA is a dynamic and complex process, which depends not only on the specificity of the given TF for its DNA-binding sequence, but is also dependent on the accessibility of the chromatin, in competition with other transcription factors for the same genomic region and on the activity state of the transcription factor. For example, TF competition can be observed between STAT1 and AP1 transcription factors, which tend to occupy the same genomic regions (Whiteld et al., 2012).

The regulation of transcription is a multidimensional process. The initiation of transcription is the first of many steps required for the assembly of a functional product of a gene. Without TF-dependent assembly of the transcriptional machinery on the promoter of a gene, every other regulatory step is irrelevant. Additionally, it can be suggested that initiation of transcription is intrinsically built into a modern definition of the gene. As a result, features such as promoter structure, RNApolII regulation and TFs binding should be included during discoveries of novel genes. After all, an lncRNA can be a gene only if it is transcribed and regulated transcription cannot happen without the formerly mentioned elements.
1.3. The role of the regulation of gene expression in cell growth, migration and cancer

For the multicellular organism to function correctly, each cell must exhibit its function at precise location and time. As a result of that requirement, a multitude of systems have evolved to control every function of the cell. Cell growth (defined as cell division and an increase in size) and migration are arguably two of the most fundamental cellular processes that must be controlled. If a cell obtains an ability to grow at an uncontrollable rate, it threatens survival of the whole organism. If such a cell is not stopped in time by an organism, cancer can develop.

1.3.1. The relationship between cell growth and cancer

Cancer (medically known as malignant neoplasia) is a diverse group of diseases whose common characteristic is uncontrollable growth of cells. Cancer is caused by an introduction of abnormalities (mutations) to the genetic material of cells. Epidemiological studies have shown that the majority of clinical cases of cancer (90–95%) are caused by environmental and lifestyle factors and not genetic factors (Anand et al., 2008; Hamilton and Mack, 2003). A notable example of heritable cancer is retinoblastoma, cancer of the retina, which is caused by a mutation in the Rb gene (Du and Pogoriler, 2006). From an evolutionary point view, mutations are crucial for survival of species, as they allow them to adapt to the ever-changing environment. On the other hand, mutations can be detrimental for the survival of an individual and organisms have developed a number of checkpoints to ensure that the growth of each cell of the multicellular organism is controlled. The cell cycle, a process of growth and division of a single cell into two daughter cells, is the main cellular process which must be controlled. When cells gain the ability to grow at an uncontrollable rate, they threaten the survival of the whole organism. Since the cell cycle is such a highly regulated process, one might ask: how are cells are able to become cancer cells in first place? Carcinogenesis is possible because many growth-related cellular processes are controlled by a small number of proteins that are master regulators of those processes. The previously mentioned Rb gene is one of the tumour suppressor genes – master regulators of cell growth and differentiation. Other master regulators of cell growth and division, which might cause cancer, are proto-
oncogenes. Loss-of-function mutations in the tumour suppressor genes and gain-of-function mutations in the proto-oncogenes are essential but not sufficient steps required for development of cancer (Hanahan and Weinberg, 2011, 2000). In addition to the 1) sustained growth (caused by hyperactivity of oncogenes) and 2) evasion of growth suppression (caused by loss of tumour suppressor genes), Hanahan and Weinberg (2000) proposed four additional hallmarks of cancer – cell must: 3) resist cell death, 4) activate invasion and metastasis, 5) gain replicative immortality and 6) must be able to induce angiogenesis. Subsequently, four other features have been added: two ‘emerging hallmarks’ – 1) deregulation of cellular energetics and 2) evasion of immune destruction, and two ‘cancer enabling characteristics’ – 1) genome instability and 2) tumour-promoting inflammation (Hanahan and Weinberg, 2011).

Each cell of the multicellular organism is surrounded by many other cells and in order to maintain homeostasis, these cells must efficiently communicate with each other; this is achieved by extracellular signalling molecules. These molecules can trigger movement, growth, division or death (apoptosis) of cells. Each cell has a number of cell surface receptors that recognise these signalling molecules and subsequently pass the signal to the intracellular signalling cascades. Failure of these signalling cascades might lead to constitutive activation of the cell growth, thus leading to cancer. There are seven main cell signalling cascades associated with cancer: 1) JAK/STAT, 2) NOTCH, 3) MAPK/ERK, 4) NF-κB, 5) WNT, 6) PI3/AKT and 7) TGF-β (for reviews, see Christofori, 2006; Dreesen and Brivanlou, 2007; Karin, 2006; Shaw and Cantley, 2006). Each signalling pathway is associated with a particular signalling event (for overview see table 1.3.1 1).

The MAPK (mitogen-activated protein kinase) signalling pathway plays a central role in cell proliferation and, unsurprisingly, it is dysregulated in a broad spectrum of human cancers. A large number of detrimental changes in the MAPK pathway are due to mutations of Ras and Raf – two central proteins essential for the function of the pathway. Cancer-associated mutations of Ras and Raf cause constitutive activation of the MAPK pathway, which in turn causes uncontrolled growth of the cells.
Table 1.3.1-1: Summary of the cell signalling pathways associated with carcinogenic transformation

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Stimulators</th>
<th>Role</th>
<th>Downstream effectors</th>
<th>Cancer association</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAK/STAT</td>
<td>Cytokines, Growth Factors (EGF, PDGF, LIF)</td>
<td>Regulation of apoptosis, differentiation and proliferation</td>
<td>Transcription factors</td>
<td>Lung, breast, head and neck, brain and stomach tumours</td>
</tr>
<tr>
<td>NOTCH</td>
<td>Delta, Serrata, Lag2</td>
<td>Cell fate decision</td>
<td>Transcription factors</td>
<td>T-cell acute lymphoblastic leukaemia, also associated with breast tumours, melanoma, medullablastoma and ovarian cancers</td>
</tr>
<tr>
<td>MAPK/ERK</td>
<td>Cytokines, Growth factors</td>
<td>Regulation of cell growth, division, adhesion, migration and survival</td>
<td>Transcription factors</td>
<td>Constitutively activated in wide spectrum of human tumours</td>
</tr>
<tr>
<td>NF-κB</td>
<td>Tumours necrosis factor – α (TNF-α), interleukin-1, growth factors, bacterial or viral infection, oxidative stress</td>
<td>Coordination of innate and adaptive immune responses</td>
<td>Transcription factors</td>
<td>B-cell malignancies</td>
</tr>
<tr>
<td>WNT</td>
<td>Over 30 extracellular Wnt-ligands</td>
<td>Regulation of cell-cell communication</td>
<td>Transcription factors</td>
<td>Colon cancer, breast cancer, adenocarcinomas, gastric polyps and acute myeloid leukaemia</td>
</tr>
<tr>
<td>PI3/AKT</td>
<td>Cytokines, Growth factors, extracellular matrix molecules</td>
<td>Regulation of cellular proliferation, cell death and cytoskeletal rearrangements</td>
<td>Transcription factors, activates MDM2 (negative regulator of p53) and GSK-3 (negative regulator of Wnt signalling)</td>
<td>Glioblastomas, lung carcinomas, melanomas, breast, ovarian and thyroid cancer</td>
</tr>
<tr>
<td>TGF-β</td>
<td>Members of TGFβ superfamily</td>
<td>Negative regulation of cell proliferation</td>
<td>Transcription factors</td>
<td>Pancreatic ductal adenocarcinomas, lung carcinomas</td>
</tr>
</tbody>
</table>
1.3.2. Mitogen-activated protein kinase pathway

The MAPK pathway plays a central role in transmission of the growth factor signal from the surface of the cell to nuclear effector proteins (Turjanski et al., 2007; Yang et al., 2013b). The MAPK pathway consists of a series of protein kinases, which activate downstream elements of the cascade by phosphorylation. In humans, there are at least 11 members of the MAPK superfamily and they can be divided into six groups based on their sequence similarity: the extracellular signal-regulated protein kinases (ERK1 and ERK2); Jun N-terminal kinases (JNK1, JNK2, JNK3); p38s (p38a, p38b, p38g, p38d); ERK5 (ERK5); ERK3s (ERK3, p97 MAPK, ERK4) and ERK7s (ERK7, ERK8) (Turjanski et al., 2007; Yang et al., 2013b). Each subclass of the MAPK superfamily can be stimulated by a separate protein kinase cascade. All of the stimulation cascades follow a general path, where an extracellular signal is detected at the cell surface and it is transmitted to the downstream kinase – MAPK kinase kinase (MAP3K or MAP2K). When MAP3K is activated, it in turn phosphorylates MAPK kinase (MAPKK or MAPKK), which then activates the final effector – MAPK – which is responsible for the activation or deactivation of the downstream proteins (Osborne et al., 2012). This seemingly simple pathway is responsible for the great amplification of the signal received from the extracellular stimulus and it allows for rapid change in the activity of downstream effector proteins in a matter of minutes.

The MAPK pathway is well-conserved signalling pathway present in diverse groups of organisms including plants, fungi, nematodes, insects and mammals (Widmann et al., 1999). The ERK pathway is one of the most studied pathway of the mammalian MAPK pathways due to its role in cancer (Dhillon et al., 2007). The first step in the activation of the ERK pathway is the exposure of the cell to extracellular stimuli such as cytokines or growth factors (Figure 1.3.2-1). Epidermal growth factor (EGF) is one of activators of the ERK pathway. This extracellular signal is detected by members of the tyrosine kinase family of cell receptors (Davis, 1995; Yarden and Sliwkowski, 2001) called EGF receptors (known as ErbB1, HER or EGFR). Binding of EGF to the receptor stabilises the dimeric form of the EGF receptor and triggers a conformational change in the intracellular domains of the receptor, which activates the kinase activity of one domain of the receptor, causing phosphorylation of the other domain (Hubbard and Miller, 2007; Lemmon and Schlessinger, 2010; Lodish
H, Berk A, Zipursky SL, 2000). When fully phosphorylated, the EGF receptor is active and transmits signal downstream. The next step in the ERK pathway is the activation of the GTP-binding protein of the Ras family (Avruch, 2007). The Ras proteins are active only when they are bound to GTP and they are inactive when they bind GDP. When the EGF receptor is activated, it recruits an adaptor protein (e.g. GRB2) which in turn recruits guanine nucleotide exchange factor (GEF) – Son of Sevenless (Sos) (Shaw and Cantley, 2006). When Sos is engaged, it facilitates exchange of the GDP to GTP in the plasma membrane co-localised Ras protein (Pearson et al., 2001). In the next step, the active form of Ras (bound with GTP) activates Raf – MAP3K of the ERK cascade. The exact nature of the Raf activation by Ras is not known, but it was proposed that: Raf might induce conformation change in Raf that induces its kinase activity; that Ras might provide the ‘proper’ environment for the activity of Raf (Pearson et al., 2001); or that Raf is activated by phosphorylation by another kinase in the presence of Ras (Balan et al., 2006; recently reviewed in Plotnikov et al., 2011). In addition to these processes, Raf activity is also regulated by other proteins including 14-3-3 proteins and heat-shock protein 90 (Roberts and Der, 2007). When Raf is activated it phosphorylates the next member of the ERK pathway – MEK (MAP2K). There are two closely related MEK kinases – MEK1 and MEK2. The phosphorylation of MEKs take place on two serines located in the Ser-X-Ala-X-Ser motif (Ser218 and Ser222 in MEK1) (Alessi et al., 1994; Zheng and Guan, 1994). Signal transition from Raf to MEKs is the main signal amplification step because MEK1/2 is much more abundant in the cell then Raf (Pearson et al., 2001). When MEKs are activated, they in turn phosphorylate ERK1 and ERK2. In contrast to the Raf-MEK step, MEK1/2-ERK1/2 step does not amplify the signal because proteins are present in the cell at similar levels (Pearson et al., 2001). The phosphorylation of ERKs takes place on the conserved tyrosine and threonine within the Thr-X-Try motif ((Payne et al., 1991). Phosphorylated ERK1/2 are the final kinases in the ERK pathway and when they are activated they modulate activity of the downstream effectors including transcription factors such as ELK1, JUN, FOS, CREB and SRF (Kyosseva, 2004) and modulate expression of genes in the nucleus (downstream effects of ERK1/2 are described in section 1.3.3).
Figure 1.3.2-1. An overview of the ERK1/2 signaling cascade.

A simplified overview of the signalling cascade from Epidermal Growth Factor (EGF) to transcription factors. See text for details of the ERK1/2 pathway. Sos – Son of Sevenless, GDP – guanine diphosphate, GTP – guanine triphosphate.
All of the above processes take place in the cytoplasm and, as a result, each of the kinases involved in the ERK pathways are exposed to a number of other members of the MAPK pathway, thus a possibility of cross-interaction with members of other pathways exist. A range of evidence indicates that MAPKs have overlapping substrate specificity (Pearson et al., 2001; Waskiewicz et al., 1997). However, the specificity of the signalling pathways is increased by the presence of scaffold proteins such as KSR, which bring together members of a given pathway into single module (Dhanasekaran et al., 2007; Whitmarsh, 1998, 2006; Whitmarsh and Davis, 1998). In addition to the increase in specificity, the scaffold proteins increase the speed of signal transmission within each module and they insulate MAPKs in the module from other kinases.

The ERK signalling pathway reveals that a single cell surface event, such as binding of the EGF to the EGF receptor, triggers complex changes in intracellular proteins, which in turn exerts various functions within cell. Furthermore, it demonstrates that signalling pathways are highly regulated processes where the specific signal reaches only its target effectors, and that the signal is not lost or diverted to other effectors. Additionally, the structure of the cascade shows the importance of the early steps in the cascade (where a small number of molecules can trigger a significant cellular response through amplification of the signal) and explains why Ras and Raf are often the dominant proteins mutated in the cancer cells.

1.3.3. Regulation of gene expression by the ERK pathway

In order to fully appreciate the role of the ERK pathway in the regulation of cell proliferation, one should investigate what are the immediate effects of the ERK activation. ERK is a serine/threonine kinase and in its active form it phosphorylates proteins that contain the Pro-X-Ser-/Thr-Pro consensus sequence (Avruch, 2007; Yoon and Seger, 2006). Over 160 proteins are known to be phosphorylated by ERK and the majority of them are transcription factors (Yang et al., 2013b; Yoon and Seger, 2006). The first transcription factor identified as a target of ERK was ELK1 (Gille et al., 1992). ELK1 is a member of the TCF (ternary complex factor) subfamily of ETS (E26 transformation-specific or E-twenty-six) family of transcription factors and its activity is associated with the regulation of the expression of immediate-early (IE) genes (Sharrocks, 2002).
The transcription of IE genes occurs within minutes after stimulation, with rapid expression kinetics and is independent of protein synthesis (see section 1.3.4 for further description). The phosphorylation of target transcription factors triggers changes to various properties of these proteins such as 1) cellular localisation (transcription factors can be excluded from the nucleus or targeted to the nucleus), 2) expression levels and the stability and 3) affinity to cofactors and DNA (reviewed in Yang et al., 2003, 2013b). In addition to transcription factors, ERK also phosphorylates other proteins such as coregulatory proteins (e.g. CBP, p300), kinases (e.g. DAPK) and phosphatases, cytoskeletal proteins (e.g. Calnexin), signalling proteins (e.g. EGFR) and various other proteins (reviewed in Yang et al., 2003, 2013b; Yoon and Seger, 2006).

Finally, ERK activity is not always restricted to the nucleus. For example, it has been shown that activation of ERK is associated with a rapid decrease in the level of some miRNAs (Avraham et al., 2010). The putative targets of these miRNAs, termed immediately down-regulated miRNAs (ID-miRs), are some of the IE genes (see section 1.3.4 for more details). By downregulating levels of the miRNAs that target IE genes and simultaneously upregulate transcription levels by activation of transcription factors, the regulation of the downstream effectors by ERK seems to be a multi-level process where many pathways converge on the same target. This observation combined with cases of self-regulation of ERK activity by feedback and feedforward loops, shows that the downstream targets of the ERK pathway form a highly complicated and well-organised network (Avraham and Yarden, 2012; Yang et al., 2013b).

1.3.4. The expression programme downstream of ERK signalling pathway

Stimulation of cells with EGF triggers the expression of genes. However, the expression patterns of EGF-regulated genes are not random. They follow a wave-like pattern where different groups of genes are expressed at different times. Immediately after the stimulation of cells with EGF, the expression of IE genes occurs. The maximum level of IE genes is usually detected within one hour after stimulation of the cells (O’Donnell et al., 2012). The expression of IE genes is followed by the
expression of delayed-early genes (DE genes) and finally, secondary-response genes (SR genes) are produced (Figure 1.3.4-1A) (Amit et al., 2007).

The FOS proto-oncogene was identified as the first IE gene regulated by growth factors (Greenberg and Ziff, 1984). Subsequently, a large group of genes with expression kinetics similar to FOS were identified (Lau and Nathans, 1987). IE genes are a diverse group of genes whose common feature is their expression pattern, which is rapid and independent of protein synthesis (Cochran et al., 1984). The lack of the requirement for protein synthesis stems directly from the MAPK signalling pathway, where elements of the pathway are activated by phosphorylation of already synthesised proteins. Since many of the IE genes are transcription factors (Amit et al., 2007; Odrowaz and Sharrocks, 2012), their expression is an intermediate step in the response to the extracellular stimuli that triggers the expression of DE genes.

The majority of what is known about the regulation of the expression of IE genes comes from studies of the proto-oncogene FOS (review in O’Donnell et al. 2012). The promoter region of FOS contains a serum response element (SRE) which is recognised by a complex of serum response factor (SRF) and a member of the TCF transcription factor family. TCFs are a subfamily of ETS-domain transcription factors (review in Sharrocks 2001). ELK1 is the most studied member of the TCF family and it can be directly activated by the ERK protein kinases upon growth factor stimulation. When ELK1 is in a complex with SRF it binds ETS-binding sites present in the promoter regions of many IE genes including FOS and EGR1 (Sharrocks, 2001; Stevens et al., 2002). In the absence of the growth factor signal, ELK1 is SUMO (small ubiquitin-related modifier)-modified. In this state, ELK1 recruits histone deacetylases to the FOS promoter, and this allows for maintenance of low basal expression level. When ELK1 is activated by ERK-dependant phosphorylation, it triggers, in the vicinity of the FOS promoter, an increase of histone acetylation levels through p300 acetyltransferase and promotes phosphorylation of the H3S10 (review in O’Donnell et al. 2012). In the next step, the mediator complex is recruited to the FOS promoter site and subsequently RNApolIII is recruited and transcription is initiated.
Figure 1.3.4-1. Pattern of gene expression downstream of ERK pathway.

A) A stylised representation of the pattern of gene expression downstream of ERK signalling pathway. ERK activation by phosphorylation triggers a wave-like expression of Immediate-early genes (IEG), followed by expression of delayed-early genes (DEG) and secondary response genes (SRG).

B) Depiction of interactions between genes expressed downstream of ERK pathway. ERK activation triggers expression of IEGs, which can activate transcription of DEG (1) – feedforward, inhibit ERK activity (2) transcription of IEG (3) via negative feedback loops or activate transcription of IEG (4) via positive feedback loops. DEGs can trigger expression of SRG or affected transcription of IEG via positive and negative (5) feedback loops. Finally, ERK activity causes inactivation of immediately-downregulated microRNAs (ID-miRs) which downregulate IEGs (6). See text for further details and examples of gene interactions.
IE genes are not only characterised by a rapid increase of transcription but also by rapid inactivation of transcription and degradation of the product mRNA. The FOS mRNA level peaks 30–60 minutes after cell stimulation with EGF and it returns to basal levels within two hours. It has been shown that the RNA-binding protein ZFP36 promotes degradation of IE genes transcripts (Amit et al., 2007). ZFP36 itself is a product of an IE gene, hence the expression of IE genes is regulated by a negative feedback loop through ZFP36 (Figure 1.3.4-1B). Additionally, one of the early studies of FOS expression regulation indicated that AP-1 (activation protein 1) might down-regulate expression of FOS (Sassone-Corsi et al., 1988), hence down-regulating its own activity via a negative feedback loop (Figure 1.3.4-1B). The FOS paradigm shows that there are multiple levels of IE genes transcription regulation and these various regulatory systems must cooperate in order to produce rapid expression of the genes.

FOS is a component of the AP-1 transcription factor. AP-1 is a homodimer of JUN family members or a heterodimer composed of JUN, FOS, or ATF (activating transcription factor) subunits (Karin, 1995). The activity of AP-1 is regulated at two major levels: by regulation of activity of the AP-1 proteins and by regulation of AP-1 proteins abundance (review in Karin & Liu 1997). The protein levels of AP-1 components are regulated by changes in transcription levels, control of mRNA stability and by turnover of pre-existing or newly synthesised proteins (Hess et al., 2004). Similarly to FOS, members of JUN family are also IE genes and their transcription is regulated by the MAPK pathway. Expression of JUN and FOS after EGF stimulation triggers the formation of AP-1, which in turn regulates subsequent expression of DE genes required for the cellular response to EGF stimulation (Figure 1.3.4-1A).

The transcription of DE genes is dependent on the production of the protein product of the IE genes. As a result, inhibition of translation of IE genes using cycloheximide (CHX) stops transcription of DE genes (Lanahan et al., 1992; Tullai et al., 2007). The maximum level of DE genes transcripts takes place around two hours after EGF stimulation and, by definition, it is after the peak of IE genes transcriptions. IE genes can both down- and up-regulate expression of many genes (including IE genes and DE genes). DE genes, similarly to the IE genes, are involved in multiple regulatory loops (Figure 1.3.4-1B). An example of negative feedback involving DE genes may
be seen by members of the dual specificity phosphatases family (DUSP genes). As direct targets of the members of DUSP family are MAPKs, one can see that ERK activity increases levels of DUSPs (via IE genes), which in turn deactivates MAPK signalling and eventually leads to deactivation of DUSPs (Avraham and Yarden, 2011; Blüthgen et al., 2009).

Finally, ERK signalling does not only regulate the transcription of protein-coding genes, as it can also have an effect on non-coding RNAs. Previously mentioned, (section 1.3.4) ID-miRs seem to be a subclass of micro RNAs (miRNAs) that targets IE genes (Avraham et al., 2010). For example, the level of mir-155 decreases within minutes after EGF stimulation. One of the directed targets of mir-155 is the mature form of FOS mRNA, thus a conclusion can be drawn in which ID-miRs are gatekeepers of the level of IE genes. When ERK activity is triggered by EGF signal, the level of ID-miRs decreases, resulting in an increase in the level of IE genes (Figure 1.3.4-1B) (Avraham et al., 2010).

Eventually, the interconnected network of IE genes, DE genes and ID-miRs leads to expression of SR genes. In contrast to the IE and DE genes, the level of SR genes is much more stable. SR genes are the final effector genes associated with apoptosis, cell cycle progression and various other growth and migration-associated functions (Avraham and Yarden, 2011; Ozanne et al., 2007; Shaulian and Karin, 2002).

At first sight, the expression programme downstream of the ERK pathway, with a multitude of feedback and feedforward loops, is overly complicated. However, a high level of interconnection between elements of the programme allows for greater control of the system. For example, it can be suggested that ID-miRs provide a fail-safe mechanism; uncontrollable activation of IE genes will be dampened if the ID-miRs are not inactivated. Similar action can be expected of the DUSP family phosphatases. It is safe to assume that ERK signalling should be short lived because of these fail-safe mechanisms. However, as the next section will show, the highly regulated programme can be disrupted. Examples of the gain-of-function mutations upstream of the ERK, mutations that are associated with cancer, are one of the most important examples.
1.3.5. The ERK signalling pathway and cancer

The cellular response to mitogens is a well-orchestrated event. Starting with the reception of the signal at the cell surface, going through the signal transduction cascade (section 1.3.2 and 1.3.3) and finishing with the changes in the gene expression (section 1.3.4), every reaction is well timed and controlled. The final outcome of the signalling may be (depending on the message) cell division, increased survival or a change in cellular movement. Disruption in the control or activity of any of the elements of the pathway could lead to cancer. For example, around 30% of breast cancer patients harbour an overexpression of the HER2 receptor (a member of the EGF receptor family), which is associated with a more aggressive form of the disease, with a lower survival rate and higher recurrence rate (Koboldt et al., 2012; Mitri et al., 2012; Slamon et al., 1987). Similarly, ERK is hyper-activated in up to 90% of human melanoma cases, mainly due to gain-of-function mutations of Ras and Raf kinases (Gray-Schopfer et al., 2007). Importantly, the gain-of-function of BRAF is mainly caused by a single mutation of valine at position 600 into glutamic acid (V600E BRAF).

As the signalling pathways contain a small number of key elements, these proteins became an interesting target for cancer treatments. For example Trastuzumab – targeting HER2 positive breast cancer – is one of the first effective monoclonal antibody based anti-cancer therapies (Mitri et al., 2012; Piccart-Gebhart et al., 2005) and V600E BRAF was selectively targeted by Vemurafenib (Bollag et al., 2012; Chapman et al., 2011). Nevertheless, the mechanism of action of some of the available anti-cancer drugs is still not fully understood. For example, the overexpression of HER2 is a known driver of breast cancer, yet its cellular role is not fully understood; recently it has been shown that the overexpression of HER2 leads not only to changes in the expression of protein-coding but also affects lincRNAs expression (Merry et al., 2015).

Additionally, the role of IE and DE genes (downstream targets) in cancer is not fully understood. Correlation studies of cancer patient survival and FOS expression showed that overexpression of FOS correlates both with a decrease in patient survival (Bland et al., 1995) and an increase in patient survival (Mahner et al., 2008). This shows that even for a highly-studied IE gene, such as FOS, the function in cancer is not clear. It appears that FOS might function both as a cancer repressor and
as an oncogene. The image for other IE genes is similarly confusing (Healy et al., 2012; Murphy and Blenis, 2006). Often it is assumed that IE genes are, similarly to the elements of MAPK pathways, drivers of cancer; however, to the best of my knowledge, a global overview of the expression of IE genes in cancer has not been investigated.

More effort is required in order to understand the cellular mechanisms and clinical significance of the MAPK pathways on the expression of downstream IE genes. Most of the studies of IE genes are performed in cellular systems where IE gene expression is induced transiently. However, in cancer cells, the signalling pathways are constitutively hyper-activated. It is safe to assume that this pushes the crosstalk between elements of the pathway out of the normal equilibrium and causes consequence that are difficult to predict. In addition to that, recent discoveries in the field of lncRNAs (see section 1.4) show that there still might be many elements of the signalling pathway that are unknown and further effort is required to understand the molecular mechanism of the connection between MAPK and cancer.

1.3.6. Immortalised breast cell lines as models of cell growth and migration

Our understanding of many of the regulation processes associated with EGF-dependent growth comes from studies on immortalised breast cell lines including many breast cancer cell lines.

MCF10A cells are an immortalised, non-transformed epithelial cell line derived from human fibrocystic mammary tissue. These cells are defined as ‘normal’ breast epithelial cells as they have a near diploid karyotype and are dependent on exogenous growth factors for proliferation. They also lack the ability to form tumours in nude mice and lack the ability to grow in anchorage-independent assays (Soule et al., 1990). Additionally, these cells harbour a deletion of the locus containing p16 and p14ARF, as well as amplification of MYC. Furthermore, MCF10A cells express wild-type p53 and are negative for oestrogen receptor (ER). MCF10A cells are an excellent model system for understanding epithelial cell biology. When plated in a mixture of collagen and laminin they form 3D structures that resemble acini structures of the human breast (Debnath et al., 2003). Finally, it has been shown that ELK1 – a direct target of ERK pathway – is an essential
transcription factor required for survival of MCF10A cell line (Silva et al., 2008). A number of studies showed that the MCF10A cell line is an excellent system used to interrogate how the cancer genes influence carcinogenic transformation (Debnath et al., 2002; Herr et al., 2011; Muthuswamy et al., 2001). Importantly, it has been shown that the cellular response to epidermal growth factor is intact in a MCF10A cell (Amit et al., 2007; Odrowaz and Sharrocks, 2012). In addition to MCF10A, a large panel of breast cancer cell lines have been used as models of breast cancer (Lacroix and Leclercq, 2004). MCF7 cell lines is one of the most widely used cell lines and it has been used in multiple studies involving IncRNAs (Djebali et al., 2012). In contrast to MCF10A, MCF7 cells are positive for oestrogen receptor (ER). However, similarly to MCF10A cell line, MCF7 cells are highly responsive to EGF stimulation (Davidson et al., 1987). Another widely used cell line is MDA-MB-231 – a mammary gland cell line – that is derived from a metastatic site (Lacroix and Leclercq, 2004). In contrast to MCF7 and MCF10A cells, the EGF receptor is strongly overexpressed in the MDA-MB-231 (Davidson et al., 1987) and they are resistant to EGF stimulation. Both MCF7 and MDA-MB-231 do not require EGF for normal proliferation.

Out of the above cell lines, MCF10A seems to be the most suited for studies of EGF-dependant signalling. Studies from our laboratory indicate that the EGF stimulation triggers wave-like expression of immediate-early genes (see section 1.3.4) and that ELK1 is an important regulator of cellular movement and function, which is dependent on the expression of immediate-early genes (Odrowaz and Sharrocks, 2012). In addition, both MCF10A and MCF7 cells are some of the key cell lines used in the ENCODE project (Djebali et al., 2012) and MCF7 cells have been used in the FAMTOM5 project (Forrest et al., 2014). Use of these cell lines in global projects provides the research community with additional easily available genomic information and it makes MCF10A an excellent model to study EGF-dependant expression programme.

1.4. Long non-coding RNAs

In section 1.1.1, I introduced the concept of the long non-coding RNAs. They are a novel class of transcripts that do not share many features except for the lack of an obvious open reading frame and length of over 200 bp. Based on the research
evidence from the next-generation sequencing projects (Djebali et al., 2012) as well as many independent ‘small science’ studies, the existence of some of the lncRNAs is not questioned. However, in the literature there is ongoing discussion about how widespread the expression of lncRNAs actually is. The following chapter will summarise the history of the lncRNA research and their role in signalling pathways and cancer.

1.4.1. A brief history of long non-coding RNAs

In the early days of molecular biology two types of non-coding RNAs were identified – transfer RNA (tRNA) and ribosomal RNA (rRNA). Since then most of the research focused on protein-coding genes. However, in 1993 a new class of genes were discovered. Lin-4 was identified as the first small interfering RNA (siRNAs) – a class of ~22 nucleotide long transcripts that regulate the level of the protein-coding genes (Lee et al., 1993). Identification of new classes of small ncRNAs was associated with a change in methodology. In order to discover novel genes, scientists started using biochemical screens instead of genetic screens (Mattick, 2009). This shift in the methodology eventually led to discovery of lncRNAs.

One of earliest lncRNAs identified was XIST (X-inactive specific transcript) (Brown et al., 1992), a gene responsible for random inactivation of the X-chromosome of placental mammals. Following the development of next-generation sequencing techniques, a large number (currently estimated at over 30,000 genes) of long non-coding RNAs were identified. A high number of the putative lncRNAs were discovered using chromatin-state maps in mouse embryonic stem cells (Guttman et al., 2009). In this study, the authors investigated the genomic locations that are marked by H3K4me3 and H3K36me3 (H3K4me3 and H3K36me3 marked regions are actively transcribed by polymerase II) and they identified over 1500 lncRNAs. Collaborative efforts by large international consortia have increased the number of the putative lncRNAs to over 10,000 (Djebali et al., 2012). Recently, lncRNAs have been identified in unexpected forms and location. For example, enhancer RNAs (eRNAs) are expressed from the enhancer regions (Kim et al., 2010), whereas circular RNAs (circRNAs) are formed by 3’-end to 5’-end splicing of RNA and their role has been associated with miRNA sponges (Memczak et al., 2013).
Currently, the outstanding question in the lncRNAs field is the identification of their biological role. This question will have to be addressed by ‘small science’ approaches, where the role of each of the lncRNAs is investigated in detail by separate groups. The following chapters will outline the approaches used in the process of lncRNA identification and examples of the role of lncRNAs that have recently been identified.

1.4.2. Major methods used for functional studies of lncRNAs

The arrival of next generation sequencing technique and development of RNA-seq resulted in unparalleled discoveries of lncRNAs. Over recent year, multiple RNA-seq protocols have been developed. Currently, the sequencing market is dominated by sequencing by synthesis technique developed by Solexa and owned by Illumina (Wang et al., 2009). The key step in the preparation of the sequencing libraries is removal of rRNA (>90% of total RNA). Currently, two methods are widely used for rRNA removal – enrichment for polyA-containing transcripts or depletion of rRNA and both have been successfully used for identification of lncRNAs (Djebali et al., 2012; Kretz et al., 2013). Pairwise comparison shows that for the mature RNAs of good quality, the difference between rRNA depletion methods is negligible (Chen et al., 2014; Tariq et al., 2011; Zhao et al., 2014). However, for some of the novel classes of lncRNAs, such as eRNAs and circRNAs, due to lack of polyA sequences in the mature transcripts, rRNA depletion methods are essential (Kim et al., 2010; Memczak et al., 2013).

In addition to the RNA-seq techniques that sequence the whole length of the transcripts, multiple techniques have been developed that measure the level of actively transcribed, nascent RNAs. For example, global run-on sequencing (Gro-seq (Core et al., 2008)) is used to study eRNAs (Hah et al., 2013) and Start-seq (Scruggs et al., 2015) can be used to study upstream antisense RNAs (uaRNAs) – antisense transcripts located upstream from a known gene, expressed from bidirectional promoter. However, most of the nascent RNA-seq technique cannot detect full length transcripts and they are mainly used for studies of the processes of transcription. Nonetheless, since nascent RNA-seq detects events of transcription initiation, involuntary they also lead to discoveries of lncRNAs.
RNA-seq techniques seem to be indispensable for the process of lncRNAs discovery; nevertheless they do not provide information about the function of particular lncRNAs. Functions of some of the classes have been described globally using computation approach (e.g. circRNAs (Memczak et al., 2013); however, only a handful of individual lncRNAs have a well-defined cellular role (see next section for examples). Lack of discoveries of lncRNAs in classical genetic screens indicates that they play subtle, regulatory roles rather than function as master regulators of cellular processes. As a result, functional studies are often performed using ‘guilt-by-association’ approaches where one searches for molecular partners or genes of known function whose expression correlate with the lncRNAs of interest. For example the functional association of PCAT1 (prostate cancer-associated ncRNA transcript 1) and Polycomb Repressive Complex 2 (PRC2) was deduced from the mutually exclusive expression pattern of PCAT1 and EZH2 (component of PRC2) (Prensner et al., 2012).

On the cellular level the function of lncRNAs can be investigated using multiple co-interaction methods. When a potential protein co-interactor is identified, one can immunoprecipitate the protein of interest together with lncRNA using RNA immunoprecipitation (RIP) (Kretz et al., 2013; Zhao et al., 2010). Recently, a number of techniques have been proposed where one can investigate lncRNAs interaction without prior knowledge of protein partners. Chromatin Isolation by RNA Purification (ChIRP-seq) (Chu et al., 2011), RNA Antisense Purification (RAP) (Engreitz et al., 2013) and Capture Hybridization Analysis of RNA Targets (CHART) (Simon et al., 2011) allow for identification of chromatin binding by lncRNAs. All three of the techniques use similar methodology where one designs biotinylated (or otherwise tagged) antisense oligonucleotides that target the transcripts of interest. Subsequently, the chromatin fraction associated with the transcript of interest is purified and sequenced. Additionally, one can identify proteins interacting with the transcripts of interest using ChIRP-MS (ChIRP followed by mass spectroscopy instead of sequencing) (Chu et al., 2015).

In addition to relatively new techniques which rely on next generation sequencing, more traditional techniques have been used to study lncRNAs. The function on lncRNAs can often be deduced from cellular localisation of the transcripts. RNA Fluorescence in situ Hybridisation (RNA-FISH) an biochemical fractionation have
been successfully used to identify nuclear and cytoplasmic lncRNAs (Djebali et al., 2012; Khaitan et al., 2011; Marín-Béjar et al., 2013; Sauvageau et al., 2013). Additionally, ribosome profiling has been used to investigate the noncoding status of lncRNAs (Guttman et al., 2013; van Heesch et al., 2014). Techniques mentioned above show that the lncRNA field is technologically fast-moving. Functional studies of lncRNA require novel biochemical approaches oriented towards RNA rather than traditional protein- or DNA-focus techniques. Key to many of the above techniques are next generation sequencing techniques and computation approaches developed for them. However, the ultimate goal of the field is identification of the physiological role of lncRNAs.

1.4.3. Role of lncRNA in cell growth, differentiation and cancer

The role of lncRNAs has been implicated in many cellular processes including regulation of transcription, RNA processing, siRNA attenuation and RNA stabilisation (reviewed in Cech and Steitz, 2014; Fatica and Bozzoni, 2013; Holoch and Moazed, 2015; Yang et al., 2013a). Over recent years, lncRNAs became prominent elements of regulatory networks associated with cells growth, differentiation and cancer progression. The following section will present some of the potential phenotypes and molecular mechanism associated with these processes.

One of the first examples of lncRNAs associated with cancer-related pathways is lincRNA-p21 (Huarte et al., 2010). In this study a set of over 30 lncRNAs regulated by p53 was identified in two mouse cell lines using custom tiling array. Further analysis of the data showed that lincRNA-p21 is an lncRNA residing ~15kbp upstream of the p21 gene (a canonical target of p53) and it regulates apoptosis (but not cell cycle arrest) in a p53-dependant manner. Using biotinylated in vitro-synthesised lincRNA-p21, the authors showed that lincRNA-p21 interacts with hnRNP-K, a known RNA-binding protein and transcriptional repressor. When lincRNA-p21 was down-regulated using siRNA, the binding of hnRNP-K to many of its target sites was abolished (Figure 1.4.3-1A). lincRNA-p21 is an example of a functional lncRNA that acts downstream of the DNA damage pathway to modulate gene expression. A subsequent study suggested that lincRNA-p21 may function in the cytoplasm (Yoon et al., 2012). The authors investigated the interaction of lincRNA-p21 with HuR – an RNA-binding protein (Figure 1.4.3-1A). They observed that
knockdown of HuR increases stability of \textit{lincRNA-p21} transcripts and decreases translation potential of \textit{JunB} and \textit{CTNNB1}. Their results showed that \textit{lincRNA-p21} transcript targets \textit{JUNB} and \textit{CTNNB1} transcripts to Rck – a known translation repressor.

An example of lncRNA that regulates expression of genes by directing protein cofactors to the DNA is HOTAIR. \textit{HOTAIR} has been identified in the \textit{HOXC} locus (contains genes regulating body plan in embryo along the anterior-posterior axis) and it was shown that it is required for the PRC2-dependant silencing of the \textit{HOXD} locus (Figure 1.4.3-1B) (Rinn et al., 2007). Subsequent studies showed that HOTAIR levels are increased in primary breast tumours and that HOTAIR expression increases cancer invasiveness and metastasis in a PRC2-dependant manner (Gupta et al., 2010). Furthermore, high overexpression (>125 times) of HOTAIR has been associated with the decreased survival rate of the affected patients. This shows that epigenetic changes introduced by HOTAIR are of high significance for cancer studies.

\textit{TINCR} (terminal differentiation-induced ncRNA) is an example of an lncRNA whose function is associated with cell differentiation but functions by regulating RNA stability (Kretz et al., 2013). \textit{TINCR}, similar to the \textit{lincRNA-p21} mechanism of action, interacts with mRNAs and targets them to STAU1 – a double stranded RNA-binding protein. The authors did not investigate the changes in the level of downstream protein products; hence, the effect of mRNA-TINCR-STAU1 is unknown. However, the interaction is essential for normal epidermal differentiation.
Figure 1.4.3-1. Regulation of expression by lncRNAs.

A) lincRNA-p21 regulates expression via two mechanisms. It can suppresses transcription in hnRNP-K (Heterogeneous nuclear ribonucleoprotein K) dependant manner, or in the absence of the RNA-binding protein HuR, it decreases translation potential of JunB and CTNNB1 genes in a Rck-dependant manner (forms cytoplasmic processing bodies). B) HOTAIR suppresses transcription by recruitment of Polycomb Repressive Complex (PRC) to the HOXD locus.
A number of studies of global RNA levels showed that many lncRNAs are differentially expressed in various cancers. For example RNA-seq experiments in prostate cancer cells identified 121 novel lncRNA associated with that disease (Prensner et al., 2011). Further investigation of the most upregulated lncRNA, PCAT1, indicated that it is predominantly repressive in nature and it influences expression of genes associated with mitosis and cell cycle. In addition to RNA-seq, microarrays have also been used to investigate differential expression patterns of lncRNA. The Reis group investigated expression of the lncRNAs in pancreatic cancer (Tahira et al., 2011). The authors identified 134 ncRNAs that are differentially expressed between normal and cancer pancreatic cells. Eleven of these ncRNAs were intronic transcripts mapping to the introns of MAPK pathway and/or apoptosis-related genes. Subsequent validation of these genes using RT-qPCR showed that three of these intronic lncRNAs, PPP3CB, MAP3K14, and DAPK1, are differentially expressed between primary and metastatic pancreatic cancer cells, but the mRNA for the host genes are expressed at equal levels. Additionally, four transcripts have been identified as antisense to PPP3CB, ATF2, TGFBR2 and MAPK1. These results suggests that genomic loci harbouring MAPK genes, also encode lncRNAs, but their functions are unknown.

Another example of an lncRNA associated with cancer is MALAT1 (metastasis associated lung adenocarcinoma transcript 1). Originally, MALAT1 was identified as a highly upregulated transcript in lung cancer samples and it was associated with poor prognosis (Ji et al., 2003). MALAT1 was described as a highly conserved lncRNA. A Subsequent study showed that MALAT1 plays a role in the regulation of alternative splicing in nuclear speckles (Tripathi et al., 2010). These observations, combined with the high conservation across organisms, indicated that MALAT1 may play an important role in the cell. However, subsequent knock-out of MALAT1 in mice did not show a significant phenotype (Nakagawa et al., 2012). This indicates that even the highly conserved, highly expressed lncRNAs most likely exert their function in a limited number of specific conditions.

The lncRNAs described in this section are examples of a small fraction of lncRNAs associated with cancer, differentiation and cellular growth. I focused only on some of the most comprehensively studies lncRNAs. A Pubmed search using ‘lncRNA’ and ‘cancer’ results in nearly 1700 articles with nearly 50% of them published in the last
18 months. This highlights the growing reach of the field and the potential importance of IncRNAs in the regulation of cell growth.

1.5. Aims of the project

EGF – an important regulator of cellular growth – triggers a transcriptional programme, which via series of transcriptional bursts ultimately leads to cell division. A key element of the EGF-induced response is MAPK-dependant signal transduction from the cell surface to the nucleus. MAPK activity triggers expression of multiple IE genes followed by DE genes and SR genes. The expression programme is tightly controlled via multiple feedback and feedforward loops.

Over the recent years, IncRNAs have emerged as important regulators of cellular processes. Their role has been implicated with regulation of apoptosis, cellular differentiation and their expression is associated with cancer. However, the role of IncRNAs has not been investigated in a tightly controlled EGF-induced expression programme.

In this study I hypothesise that the expression of IncRNAs is regulated by the EGF-induced signalling programme and that these IncRNAs play a role in the expression programme associated with the IE and DE response. They key aims of the project were:

- identification of EGF-regulated IncRNAs using RNA-seq
- global characterisation of the family of EGF-regulated IncRNAs including their expression pattern and association with diseases
- selection of specific IncRNAs and their further characterisation including:
  - investigation of the coding potential
  - investigation of their interactions with other elements of EGF-regulated expression programme
  - investigation of the effect of knockdown and overexpression on the EGF-regulated expression programme.

The MCF10A cell line, with its near normal karyotype, strong reliance on EGF for growth and well-defined phenotypes, seems to be an ideal in vitro system for this project.
2. Materials and Methods

2.1. Cell culture

$5 \times 10^5$ MCF10A cells were grown in the T75 flasks in DMEM/F12 (Gibco, 11320-033) containing 5% horse serum (Biosera, DH291), 20 ng/ml EGF (Sigma, E1257), 10 µg/ml insulin (Sigma, I0516), 100 ng/ml cholera toxin (Sigma, C9903) and 0.5 µg/ml hydrocortisone (Sigma, H0396) (complete medium).

MCF7, MDA-MB-231, Hela and HEK293T cells were grown in DMEM (Gibco, 22320-22) supplemented with 10% Fetal Bovine Serum (Life Technologies, 10270-098).

All of the cells were grown up to 90% confluence and were passaged every 2-3 days by brief washing of the cells with Dulbecco's phosphate-buffered saline (DPBS, Life Technologies, 14190-094), detached with Trypsin-EDTA (0.05%) (Life Technologies 25300-054) and replated at 1 in 5 dilution in full media.

2.2. RNA methods

2.2.1. Nuclear RNA-seq

$1.5 \times 10^6$ MCF10A cells were seeded onto 60 mm dish in complete media without EGF and with 0.5 % Horse serum instead of 5 % (see section 2.1). After 48 hrs incubation, EGF was added to a final concentration of 20 ng/ml. After 30 minutes, nuclear fractions were isolated according to the protocol described previously (Djebali et al., 2012). At the same time nuclear fractions were isolated from cells which were not stimulated with EGF. Briefly, cells were washed with 1xPBS and then incubated for 10 min with 1 ml of RLN buffer (50 mM TrisHCl pH=8.0, 140 mM NaCl, 1.5 mM MgCl$_2$, 0.5% (v/v) IGEPAL CA-630 and 2 U/ml superaseIn – RNase inhibitor, Invitrogen AM2696). Subsequently, the lysate was centrifuged at 1 000 g for 5 min at 4°C, the pellet was collected, resuspended in 1 ml of RLN buffer, incubated for a further 5 min on ice and centrifuged at 1 000 g for 5 min at 4°C. At this stage nuclear fractions from different experimental days were stored at -80°C for subsequent RNA extraction. The RNA was extracted from the nuclear fraction using an RNeasy plus kit (Qiagen, 74134) with DNase treatment according to the
manufacturer’s protocol. After isolation, the purity and integrity of RNA were investigated using a Bioanalyser. Only non-degraded samples with a ratio of 260/280 nm absorption above 2 were used for sequencing library preparation.

The cDNA libraries were prepared using two protocols:

- Encore® Complete RNA-Seq Library Systems (NuGEN, 0311) for total RNA sequencing
- TruSeq Stranded mRNA Sample Prep Kit (Illumina, RS-122-2101) for polyadenylated-tailed RNA sequencing.

The preparations of libraries together with subsequent sequencing reactions were performed by Stacey Holden at The Genomic Facility, University of Manchester according to the manufacturer’s recommendations. The sequencing reaction was performed on the Illumina HiSeq 2500.

2.2.2. Reverse transcription – quantitative polymerase chain reaction

RNA was extracted using the RNeasy plus kit (Qiagen, 74134) following the manufactures protocol. Subsequently, RNA samples were quantified using Nonodrop 2000 (Thermo Scientific) and concentrations were normalized to 20 ng/µl. 40 ng of each sample was used per reverse transcription – quantitative polymerase chain reaction (RT-qPCR) reaction using the QuantiTect SYBR® Green RT-PCR Kit (Qiagen, 204243) on Rotor-Gene Q (Qiagen) real-time PCR machine. Samples were quantified duplicate.

When indicated, the nanolitre volume middle throughput RT-qPCR was performed using the Fluidigm Biomark HD system using EvaGreen chemistry. The reactions were performed by Claire Morrisroe at The Genomic Facility, University of Manchester following manufacturer’s protocol. The output data were processed following the default quality protocol. Furthermore, the data points with more than one peak in the melt analysis were discarded.
The final results are normalised to the house keeping genes using the delta delta Ct method (Livak and Schmittgen, 2001). Supplementary Methods File 1 contains the list of primers used.

2.2.3. Ribosome analysis – sucrose cushion

MCF7 cells were plated onto 150 mm petri dishes at the density required for ~70% confluency on the assay day (around 5 x 10^6 cells). On the day of experiment, cells were stimulated with EGF for indicated times. Cells were incubated for 5 minutes before lysate collection with 100 µg/ml of Cycloheximide (CHX, Sigma, C7698, protein synthesis inhibitor that ‘freezes’ the actively translated mRNA with ribosomes and nascent protein). Subsequently, dishes were transferred onto ice/water bath, media was removed and cells were washed with PBS containing 100 µg/ml of CHX. Cells were gently scraped into 1 ml of PBS and centrifuged at 200 x g for 5 minute at 4°C. Subsequently, cells were resuspended in 200 µl of CSB buffer (300 mM sorbitol, 20 mM Hepes, pH 7.5, 1 mM EGTA, 5 mM MgCl_2, 10 mM KCl, 10% Glycerol, 100 µg/ml CHX and protease inhibitor cocktail) and the cell membrane was disrupted with glass beads for 45 s. Lysate was cleared by centrifugation at 9300 x g for 10 min. The Optical density (OD) at 260 nm was measured and lysates were normalized for OD. For EDTA treatment, EDTA was added to a final concentration of 50 mM and lysates were overlaid over 60% sucrose solution in CSB buffer (without Sorbitol and protease inhibitors). Samples were centrifuged for 150 min. at 4°C at 55 000 x g. RNA was collected from the pellet and supernatant fractions (supplemented with Luciferase RNA (Promega, L4561) – normalization spike-in) using Trizol LS Reagent (Life technologies, 10296). RNA levels were quantified with RT-qPCR (section 2.2.2) and normalized to the input.

2.3. Lentiviral methods

2.3.1. Lentiviral production

Second generation lentiviral particles were produced using the established lentiviral protocol with modifications (Tiscornia et al., 2006). Firstly, HEK293T cells were seeded into T75 flasks (Corning, 430641U) in full media. When 70% confluent cell were transfected with 3 µg of transfer plasmid, 2.25 µg of psPAX2 (packaging plasmid expressing gag, pol, rev and tat genes, addgene, 12260) and 1.5 µg of
pMD2.G (VSV-G expressing envelop vector, addgene, 12259) using 67.5 µl of Polyfect Transfection Reagent (Qiagen, 301107). After 8 hrs the media was changed to full media supplemented with 10 mM sodium butyrate (increases viral titer (Cribbs et al., 2013)). After overnight incubation, the media was changed to full media. Subsequently, media containing viral particles was collected after 24 and 48 hrs, combined and concentrated using PEG-it Solution (System Biosciences, LV825A-1) following the manufacturer’s protocol. Viral particles were quantified using the qPCR method (Scherr et al., 2001). Similarly to Scherr et.al., 2001, I observed that ~1% of viral particles measured using qPCR are infectious.

2.3.2. Lentiviral transductions

Transductions of the human cell lines with concentrated viral particles were performed in the 6-well plates. Firstly, cells were plated at ~40% confluency in the media appropriate for the cell line (termed full media, section 2.1). After 8 hrs, growth media was changed to the fresh growth media supplemented with 10 µg/ml of Polybrene (Millipore, TR-1003-G) and the appropriate amount of the viral particles. The following day, media was changed to the assay media required for the subsequent analysis.

The infection rate of lentiviral particles follows a Poisson distribution. The proportion of infected cells can be calculated following:

\[ P(I) = 1 - e^{-moi} \]

where moi is the proportion between number of viruses and number of cells in a well (infection chamber). For transient gene expression, we routinely infected with 95% infection rate (moi=3) and for stable constructs we used 25% infection rate (moi=0.29) to infect the majority of cells with a single virus.

2.4. CRISPRi methods

2.4.1. Guide RNA design

Guide target sequences were designed following the protocol described in (Heigwer et al., 2014). I only used guide RNA-sequences which started with G nucleotide (required for transcription initiation by RNApolIII downstream of U6 promoter),
followed by *S. pyogenes* canonical PAM sequence (NGG) and did not map to any other human genomic locus as determined by BLAT search (Kent, 2002).

### 2.4.2. CRISPRi plasmids

The dCas9-BFP-KRAB construct was expressed from pHR-SFFV-dCas9-BFP-KRAB plasmid (addgene #46911, pAS4484 (Gilbert et al., 2013). The entry plasmid for the guide RNAs (gRNAs) was lentiGuide-Puro (addgene #52963, pAS4483). Cloning of the guides into the lentiGuide-Puro was performed as described previously (Sanjana et al., 2014). Briefly, for each gRNA two DNA oligonucleotides were designed: 1) sense oligo: sequence CACC followed by 20 nt of guide sequence (see 2.4.1) and 2) antisense oligo: sequence AAAC followed by 20 nt of reverse complementary sequence to the guide sequence. DNA oligonucleotides were ordered from Eurofins Genomics (http://eurofinsgenomics.eu) at standard desalting quality. Subsequently, complementary DNA oligonucleotides were annealed in the annealing buffer (10 mM Tris pH 8, 50 mM NaCl, 1 mM EDTA) at 1 pmol/µl concentration. Annealing was performed in the PCR cycler by heating the samples to 98°C and slowly cooling down to 20°C (5°C/minute). The annealed oligos were cloned into the LentiGuide-Puro digested with the BsmBI (NEB #R0580S) restriction enzyme. Ligation was performed with T4 DNA Ligase (NEB #M0202S) in the presence of T4 Polynucleotide Kinase (NEB #M0201S – used for phosphorylation of annealed oligos). Plasmid DNA was extracted from the positive colonies and sequence. The guide RNA plasmids are listed in table 2.4.2.-1.

### 2.4.3. Total RNA-seq after CRISPRi knockdown

MCF10A-KRAB cells were created by infection of the MCF10A cells with dCas9-BFP-KRAB expressing viruses. Two weeks after infection, cells were sorted in the FACS Aria cell sorter (BD Biosciences) and cells positive for Blue Fluorescent Protein were expanded. Subsequently, the MCF10A-KRAB cells were transduced with the gRNA expressing viruses and two day after infection stable clones were selected with 1 µg/ml of Puromycin for 10 days. Subsequently, 1.5 x 10^6 MCF10A cells were seeded onto 60 mm dish in complete media without EGF and with 0.5 % Horse serum instead of 5 % (see section 2.1). After 48 hrs incubation, EGF was added to a final concentration of 20 ng/ml for 30, 90 or 180 minutes. The RNA was extracted from the total cell extract using an RNeasy plus kit (Qiagen, 74134) with
DNase treatment according to the manufacturer’s protocol. After isolation, the purity and integrity of RNA were investigated using a Bioanalyser. Only non-degraded samples with a ratio of 260/280 nm absorption above 2 were used for sequencing library preparation.

The cDNA libraries were prepared using TruSeq Stranded mRNA Sample Prep Kit (Illumina, RS-122-2101) for polyadenylated-tailed RNA sequencing. The preparations of libraries together with subsequent sequencing reactions were performed by Stacey Holden at The Genomic Facility, University of Manchester according to the manufacturer’s recommendations. The sequencing reaction was performed on the Illumina HiSeq 2500.

2.5. Computational data analysis

2.5.1. RNA-seq

The raw sequencing data were processed using the Casava software by Peter Briggs. I performed the subsequent genome alignment and testing for differential expression of RNA-seq results using Cufflinks packages (Kim et al., 2013; Trapnell et al., 2012) according to the following protocol. Firstly, raw sequencing files (fastq format) were investigated for the quality of reads, duplication level and GC content using fastqc packages (Andrews, 2013). Subsequently, raw reads were processed with the Trimmomatic tool which removed low quality reads and contamination with sequencing adapters (Bolger et al., 2014). The trimmed sequencing reads were aligned to ensembl transcription (release 72, 2013-03-06) and human genome version

Table 2.5.1-1: Lentiviral plasmids used in the study.

<table>
<thead>
<tr>
<th>Plasmid ID</th>
<th>Name</th>
<th>Guide RNA sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>4489</td>
<td>LentiGuide-Puro EIN1 -137</td>
<td>GGCTGCAGCCTGTCTTTACA</td>
</tr>
<tr>
<td>4490</td>
<td>LentiGuide-Puro EIN1 -103</td>
<td>GGGTAGTCACTAAGTGGCT</td>
</tr>
<tr>
<td>4491</td>
<td>LentiGuide-Puro EIN1 -054</td>
<td>GGTAGCAGAGGGAGGGGTG</td>
</tr>
<tr>
<td>4492</td>
<td>LentiGuide-Puro EIN1 -023</td>
<td>GACAAGAGTGTAGCTAGTTGGCT</td>
</tr>
<tr>
<td>4493</td>
<td>LentiGuide-Puro EIN1 +014</td>
<td>GTAGAAAAAGACACTTCCAA</td>
</tr>
<tr>
<td>4494</td>
<td>LentiGuide-Puro EIN1 +071</td>
<td>GTCACTTTATTTTACAACCC</td>
</tr>
<tr>
<td>4500</td>
<td>LentiGuide-Puro NT1</td>
<td>g TAAGGCTATGAAGGATAC</td>
</tr>
<tr>
<td>4501</td>
<td>LentiGuide-Puro NT2</td>
<td>g TGGTTTACATGTCGACTAA</td>
</tr>
</tbody>
</table>
19 (hg19) using *RNA-Star* aligner (version 2.3.0e) (Dobin et al., 2013). The aligner was run with default settings with option:

```
--outFilterMultimapNmax 100
```

Which allows for up to 100 genomic locations per reads. Subsequently, reads that were mapped to multiple loci and redundant reads were removed using Pickard tools (http://broadinstitute.github.io/picard/). *De novo* transcriptome assembly was performed with *Cufflinks* package version 2.2.1 and it was run with default setting. Transcripts were identified using *Cuffmerge* packages (Kim et al., 2013; Trapnell et al., 2012) with default settings and merged with the GENCODE (v. 19) transcriptome. Differential analysis was performed with *Cuffdiff* package where transcriptome assembled by *Cuffmerge* package was used as a reference. Subsequently, data quality was analysed using *CummeRbund* package version 2.0.0 (Goff et al., 2012). All of the above calculations were performed at the Computational Shared Facility at The University of Manchester.

Subsequent data analysis was performed using custom written Perl scripts and R programming language and data were also processed in the *MS Excel* software. *TiBCO Spotfire* software was used for data visualisation.

For the RNA-seq data downloaded from public databases, the same protocol was applied.

The RNA level in all of the RNA-seq based experiments is presented as fragments per kilobase per million reads (FPKM).

### 2.5.2. ChIP-seq

Raw ELK1 sequencing data published by Dr Odrowaz (Odrowaz and Sharrocks, 2012) were aligned to hg19 using *Batch Coordinate Conversion (liftOver)* tool available from UCSC website (http://genome.ucsc.edu/util.html). Subsequently, *MACS* version 1.4.2 (Feng et al., 2012) was used for identification of enriched genomic locations in the ChIP-seq data. *MACS* was executed with default setting with the exception of P-value cut-off setting which was changed to $1 \times 10^{-3}$. The *Bedtools* (Quinlan and Hall, 2010) package was used for identification of the overlapping regions (1 nucleotide overlap).
3. Identification of long non-coding RNAs regulated by EGF

3.1. The expression of immediate-early genes in MCF10A cell line

3.1.1. Establishment of strategies for global RNA level measurements

Over the last 10-15 years a plethora of techniques has been developed to investigate the global expression profiles of genes at the transcript levels. In the past, the expression of the IE genes was investigated using microarrays (Affymetrix) (e.g. Amit et al., 2007). However, the lack of an ability to detect novel transcripts is an important limitation of microarrays during studies aimed at identification of novel lncRNA. In order to identify lncRNA regulated by EGF, we decided to employ the RNA-seq technique. Since its development, RNA-seq has been used extensively for studies of lncRNAs. The results of global investigation of transcription in the cancer cells by the ENCODE project (Djebali et al., 2012) indicated that the lncRNAs localise mainly to the nucleus of cells. Secondly, the expression of the IE genes is triggered within minutes after stimulation of the cells with EGF and the maximum expression level is observed with the first 30-45 minutes after stimulation (O’Donnell et al., 2012). Since the aim of the experiment was identification of EGF-induced transcripts, I decided to analyse gene expression using only the nuclear fraction of the cells and compare transcriptome of cells stimulated with EGF for 30 minutes with non-stimulated cells. The study of the nuclear fraction will allow for identification of newly synthesised transcripts and will enrich the samples for the lncRNAs. The final element of the experimental setup was selection of the cell line. We decided to use the MCF10A cells because:

- The EGF-dependant signalling is intact and essential for cell growth (Amit et al., 2007; Odrowaz and Sharrocks, 2012)
- ELK1 transcription factor plays an important role in the regulation of expression of IE genes (Odrowaz and Sharrocks, 2012; Silva et al., 2008)
• It is a ‘normal’ immortal cell line with almost diploid karyotype and has not mutations in the MAPK pathways (Soule et al., 1990)
• It has well-defined growth and movement phenotype that can be employed to study phenotypes of potentially discovered genes (Debnath et al., 2003; Odrowaz and Sharrocks, 2012)

The above characteristics, arguable, make MCF10A cell line one of the best cell lines to study gene expression and cellular process in a near normal condition.

RNA-seq allows for measurement of RNA expression independently of prior knowledge about the structure and sequence of the transcripts to be investigated. A crucial element of RNA-seq experiments is construction of a library of short sequencing reads. An ideal protocol would produce a library that has

• high complexity (high number of reads with varied starting points),
• even read coverage across gene body and
• high strand specificity (Levin et al., 2010).

The ultimate purpose of the RNA-seq used in this thesis was identification of lncRNAs regulated by EGF. Hence, for the purpose of exploratory RNA-seq, high continuity in the coverage of known genes is essential. In order to identify the best library preparation method that was available at the time, I compared two stranded RNA-seq library preparation protocols – Illumina Truseq® (Illumina #RS-133-2001) and NuGEN Encore (NuGEN Inc. #0333). I will refer to former protocol as Truseq and latter as Nugen. The Truseq library preparation protocol enriches the samples for the polyA-tailed RNA and subsequently creates a stranded cDNA library by incorporation of dUTP during second strand synthesis. The Nugen protocol uses selective priming (depleted for rRNA) during first strand cDNA synthesis and subsequently a stranded library is created by incorporation a nucleotide analogue during second strand synthesis.

Firstly, in order to assess the quality of the RNA-seq libraries and compare Nugen and Truseq protocols, I performed a trial experiment using samples collected from four biological replicated of stimulated and non-stimulated MCF10A cells (eight samples for each protocol). Subsequently, all of the samples were multiplexed and sequence on a single lane of Illumina Hi-Seq sequencer. On average, the sequencing resulted in 5,768,172 reads per samples (138,436,127 reads in total, Table 3.1.1-1).
Table 3.1.1-1: Summary of the mapping statistics for Truseq and Nugen sequencing libraries.

<table>
<thead>
<tr>
<th>Library Type</th>
<th>Sample ID</th>
<th>Repeat</th>
<th>Time (min)</th>
<th>Raw Reads Count</th>
<th>Mapped Reads Count</th>
<th>Proportion of Mapped Reads</th>
<th>Uniquely mapped reads (one locus)</th>
<th>Proportion of Uniquely Mapped Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truseq 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3,612,651</td>
<td>3,499,901</td>
<td>96.88%</td>
<td>3,369,571</td>
<td>96.28%</td>
</tr>
<tr>
<td>Truseq 2</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>7,002,619</td>
<td>6,754,285</td>
<td>96.45%</td>
<td>6,482,465</td>
<td>95.98%</td>
</tr>
<tr>
<td>Truseq 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3,288,547</td>
<td>3,163,077</td>
<td>96.18%</td>
<td>3,044,473</td>
<td>96.25%</td>
</tr>
<tr>
<td>Truseq 4</td>
<td>1</td>
<td>30</td>
<td>0</td>
<td>4,769,172</td>
<td>4,586,920</td>
<td>96.18%</td>
<td>4,401,735</td>
<td>95.96%</td>
</tr>
<tr>
<td>Truseq 5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>5,378,915</td>
<td>5,177,778</td>
<td>96.26%</td>
<td>4,983,371</td>
<td>96.25%</td>
</tr>
<tr>
<td>Truseq 6</td>
<td>2</td>
<td>0</td>
<td>30</td>
<td>6,268,170</td>
<td>6,022,300</td>
<td>96.08%</td>
<td>5,783,875</td>
<td>96.04%</td>
</tr>
<tr>
<td>Truseq 7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3,826,443</td>
<td>3,662,562</td>
<td>95.72%</td>
<td>3,524,727</td>
<td>96.24%</td>
</tr>
<tr>
<td>Truseq 8</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>5,429,772</td>
<td>5,241,137</td>
<td>96.53%</td>
<td>5,029,256</td>
<td>95.96%</td>
</tr>
<tr>
<td>NuGEN 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7,096,995</td>
<td>5,607,127</td>
<td>79.01%</td>
<td>5,369,644</td>
<td>95.76%</td>
</tr>
<tr>
<td>NuGEN 2</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>6,214,588</td>
<td>5,068,282</td>
<td>81.55%</td>
<td>4,854,791</td>
<td>95.79%</td>
</tr>
<tr>
<td>NuGEN 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4,773,859</td>
<td>3,827,535</td>
<td>80.18%</td>
<td>3,651,166</td>
<td>95.39%</td>
</tr>
<tr>
<td>NuGEN 4</td>
<td>1</td>
<td>30</td>
<td>0</td>
<td>5,196,188</td>
<td>4,249,184</td>
<td>81.78%</td>
<td>4,063,389</td>
<td>95.63%</td>
</tr>
<tr>
<td>NuGEN 5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>8,648,953</td>
<td>7,070,443</td>
<td>81.75%</td>
<td>6,730,688</td>
<td>95.19%</td>
</tr>
<tr>
<td>NuGEN 6</td>
<td>2</td>
<td>0</td>
<td>30</td>
<td>3,624,942</td>
<td>3,012,358</td>
<td>83.10%</td>
<td>2,884,966</td>
<td>95.77%</td>
</tr>
<tr>
<td>NuGEN 7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6,104,846</td>
<td>4,565,410</td>
<td>74.78%</td>
<td>4,358,404</td>
<td>95.47%</td>
</tr>
<tr>
<td>NuGEN 8</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>7,769,548</td>
<td>5,843,861</td>
<td>75.21%</td>
<td>5,572,263</td>
<td>95.35%</td>
</tr>
</tbody>
</table>
The Nugen libraries were sequenced to higher depth (6,178,739 reads per sample vs. 4,947,036 reads per sample of Truseq library). However, after mapping of the samples to human genome the difference became negligible (4,905,525 (Nugen) vs 4,763,495 (Truseq)).

On average, over 96% of the Truseq reads mapped to the genome and only around 80% of Nugen reads were mapped. The evaluation of the quality of the sequencing base calling (in Phred) showed a strong decrease in the quality across the length of the reads in the Nugen library (one would expect the majority of bases to have phred > 30 (Figure 3.1.1-1A).

Basic quality control of the sequencing samples showed that the reads from the Nugen libraries were of lower quality than the reads from the Truseq library. However, the ultimate goal of the sequencing was quantification and lncRNA discovery. In order to perform it accurately, one expects equal coverage of the gene bodies with sequencing reads. In order to assess this, I aligned the sequencing reads to the Ensembl reference transcriptome and investigate the distribution of the reads in 1000 randomly selected genes. The distribution followed the expected result only in the Truseq library (Figure 3.1.1-1B). Read distribution in the Nugen library was strongly skewed towards the 3'-end of the genes. Bias toward 3'-ends in the sequencing libraries has previously been associated with low quality of the RNA (Chen et al., 2014). This, however, can be ruled out, as both Nugen and Truseq libraries were prepared the same (high quality) RNA sample. If the samples were degraded, one would expect 3’-end bias in both protocols. Regardless of the source of the uneven distribution of reads, lack of uniformity makes the Nugen libraries less optimal for studies of lncRNAs. As a result, we decided to use the Truseq library preparation protocol for all of the subsequent sequencing studies.

During the trial run, we sequenced four biological repeats. However, sequencing quality control showed that two of the samples differed from the other six samples. Both, the hierarchical clustering and principle component analysis (Figure 3.1.1-2) showed that both time point samples from the first repeat (sample 1 and 2) differ from the other six samples. In the experiment setup, I was looking for potential small changes in the expression values of a small number of genes; as a result small variation in the data quality was required. Hence, I decided to sequence only the three most similar samples, but to higher depth.
Figure 3.1.1-1. Quality control of the sequencing data from the Nugen and Truseq library preparation kits.
A) An example (sample_5, 0 minutes, repeat 3) of the quality across the length of the sequencing reads in (left) Nugen library and (right) Truseq library. B) Normalized read coverage in the gene body of 1000 randomly selected genes in (left) Nugen library and (right) Truseq library.
Figure 3.1.1-2. Comparison of divergence between individual repeats of the Truseq libraries.
A) Jensen-Shannon divergence between conditions calculated using cummerbund package (Goff et al., 2012). Sample names correspond to Sample ID column in table 3.1.1-1. B) Principle component analysis of the distance between samples. Each point represent dimensionally reduced (Principle component 1 (PC1) and 2 (PC2) gene expression values. Arrows indicate influence of the samples on the principle components. Sample names correspond to concatenated values of column ‘Time’ and ‘Repeat’ from table 3.1.1-1.
The initial assessment of the library preparation protocols showed that the choice of the library preparation system might affect the discovery process. From the pipelines available at the time, only the Truseq library was optimal for new gene discovery protocols. However, even within the same protocol and with use of tightly regulated experimental setup, I observed biological variability. As a result, I selected the three least diverse biological replicates for the sequencing using higher depth. This approach increases the chances of making the most relevant discoveries.

3.1.2. Nuclear RNA-seq – quality control and raw data processing

Based on the results presented in previous chapter, I selected six Truseq samples for subsequent sequencing to higher depth (three repeats of EGF non- and stimulated cells). The following sections summarise the results of the RNA-seq and the process of lncRNA discovery. Since a golden standard for analysis of the RNA-seq data has not yet emerged, firstly, I will outline the bioinformatics approach used in the discovery of the genes and the description of the discoveries will follow. Bioinformatically, the analysis of RNA-seq data involves three basic steps:

- assessment of the quality of the raw sequencing data,
- alignment to the transcriptome and genome of reference (with potential identification of novel transcripts),
- quantification of the expression at gene (or transcript level) and subsequent statistical interrogation for differentially expressed genes.

The assessment of the quality of raw data is performed by searching for any systemic bias in the data. Investigation of the raw sequencing reads with FastQC package (Andrews et al., 2015) showed that the quality of the reads is very good (Figure 3.1.2-1A) and it lacks of bias in the distribution of GC content (Figure 3.1.2-1B). The next step of the analysis involved removal of potential adapter contamination. In has been shown that trimming of the adapter sequences ligated during sequencing library preparation and removal of reads ends with low quality improves alignment efficiency (Bolger et al., 2014). The sequencing resulted in a total of 565,764,183 raw sequencing reads (on average, just over 94 million per samples) and the adapter trimming removed 2.7% (15,286,255 reads, Table 3.1.2-1). This further indicates very good quality of the library and low level of adapter contamination.
Figure 3.1.2.1. Quality control of the high-depth RNA-seq.

A) Quality of the sequencing across the length of the reads for (left) forward read and (right) reverse read. B) Distribution of the GC content in the sequenced reads for (left) forward read and (right) reverse read. Both panels are representative examples (repeat 2 at 0 minutes). Red line indicates observed distribution and blue line expected distribution. Data were compiled using FastQC software (Andrews et al., 2015).
Table 3.1.2-1 Summary of the mapping statistics for high-depth RNA-seq.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Time Point (min)</th>
<th>Repeat</th>
<th>Pairs of reads before trimming</th>
<th>Pairs of reads after trimming</th>
<th>Uniquely mapped pairs</th>
<th>Pairs mapped to multiple locations</th>
<th>Mapping rate</th>
<th>Non-redundant pairs</th>
<th>Redundancy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>68,134,512</td>
<td>66,626,018</td>
<td>63,048,024</td>
<td>3,200,685</td>
<td>99.43%</td>
<td>48,787,045</td>
<td>22.62%</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1</td>
<td>105,461,481</td>
<td>102,769,009</td>
<td>96,918,557</td>
<td>5,204,150</td>
<td>99.37%</td>
<td>65,302,229</td>
<td>32.62%</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>2</td>
<td>114,793,735</td>
<td>112,129,190</td>
<td>105,897,699</td>
<td>5,367,462</td>
<td>99.23%</td>
<td>71,302,496</td>
<td>32.67%</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>3</td>
<td>100,992,818</td>
<td>96,168,409</td>
<td>90,642,433</td>
<td>4,759,561</td>
<td>99.20%</td>
<td>66,709,116</td>
<td>26.40%</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>3</td>
<td>68,821,842</td>
<td>67,143,047</td>
<td>63,521,522</td>
<td>2,973,986</td>
<td>99.04%</td>
<td>43,369,937</td>
<td>31.72%</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>3</td>
<td>107,559,795</td>
<td>105,342,255</td>
<td>100,016,588</td>
<td>4,624,603</td>
<td>99.33%</td>
<td>58,264,514</td>
<td>41.75%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>-</td>
<td>-</td>
<td>565,764,183</td>
<td>550,177,928</td>
<td>520,044,823</td>
<td>26,130,447</td>
<td><strong>99.27%</strong></td>
<td>-</td>
<td><strong>31.30%</strong></td>
</tr>
</tbody>
</table>

Notes:
Sample IDs and Repeat numbers correspond to the sample IDs from table 3.1.1-1
Pairs of reads before trimming – indicates number of raw pairs of sequencing reads before any processing
Pairs of reads after trimming – indicates number of paired reads that survived Trimmomatic quality control
Uniquely mapped pairs – number of read pairs that mapped a single genomic locus
Pairs mapped to multiple locations – number of read pairs that mapped to at least two genomic loci
Mapping rate – ((Uniquely mapped pairs) + (Pairs mapped to multiple locations))/(Pairs of reads after trimming)
Non-redundant pairs – number of loci to which uniquely mapped pairs mapped i.e. if two or more pairs of reads mapped to the same locus, the locus is counted as one.
Redundancy rate – 1-((Non-redundant pair)/(Uniquely mapped pairs))
Subsequently, I aligned the trimmed reads to human genome using GENCODE transcriptome as a guide (Figure 3.1.2-2). In the course of my work I used RNA-Star aligner (Dobin et al., 2013) due to its superior runtime without compromising on accuracy (Engström et al., 2013). RNA-Star aligned over 99% of trimmed reads to the human genome (Table 3.1.2-1). The vast majority of reads aligned to a single, unique locus in the genome (94.5%) and the redundancy rate (proportion of reads mapping to the same genomic location) was on average 31.3%.
Figure 3.1.2. Workflow of the RNA-seq data analysis
The analysis of the RNA-seq data involved the following steps: 1) quality control and removal of low quality reads, 2) gencode (v. 19)-guided alignment of raw data to hg19 build of human genome, 3) removal of reads mapped to multiple genomic loci and removal of redundant (PCR duplicated) reads, 4) unguided transcriptome assembly, 5) gencode (v. 19)-guided assembly of transcriptomes identified in each sample, 6) removal of isoforms (and genes) with spliced length below 400 bp and 7) differential expression analysis with Cuffdiff (part of Tuxedo package).
3.1.3. Characterisation of protein-coding genes regulated by EGF stimulation of MCF10A cells

The next step of the analysis involved transcriptome assembly and the assessment of differential gene expression. A number of bioinformatics tools exists for these purposes, however systematic comparisons of the protocols did not show significant differences between them (Rapaport et al., 2013; Steijger et al., 2013). Hence, due to the ease of use, I used the Tuxedo suite of programs (Trapnell et al., 2012, 2013). RNA-seq data provide unparalleled opportunity for discovery of long noncoding RNA. However, the RNA-seq data are only as good as the data provided from the biological samples. Hence, before we investigate the differential expression of IncRNAs, it is highly advantageous to confirm that the known genes and mechanisms can be detected in the experimental setup.

Differential analysis of the sequencing data showed that 154 protein coding genes are regulated by EGF (p-value < 0.05 and fold change >1.5 (absolute), (Figure 3.1.3-1, Supplementary file 1). The expression of majority of these genes increased after stimulation with EGF – 144 were upregulated and 10 genes were modestly downregulated. The list of some of the most highly upregulated genes included previously described IE genes – NR4A1, EGR1, EGR2, EGR3, EGR4, FOS, FOSB, KLF2, CTGF, and DUSP5 (see Figure 3.1.3-2 for examples of expression displayed on the UCSC genome browser) (Amit et al., 2007; Odrowaz and Sharrocks, 2012; Tullai et al., 2007).
Figure 3.1.3-1. Statistical analysis of the expression of EGF-regulated genes

A) Scatterplot of average expression at 0 and 30 minutes after stimulation of MCF10A cells with EGF. B) Volcano plot of log₂ transformed ratio of expression at 30 minutes and at 0 minutes (x-axis, positive values – expression is higher at 30 minutes). In the statistical analysis, p-values were capped at 0.00005, log10=-4.3. In both panels, red points indicated genes with p-value below 0.05.
Figure 3.1.3-2. Examples of EGF-regulated IE genes
Examples of UCSC genome browser views of average read density for protein-coding genes expressed from the top (Watson) strand of DNA – c-FOS (A) and bottom (Crick) strand of DNA – CTGF (B). Read densities on the Watson and Crick strands have positive and negative values respectively. For both genes, the scales on individual tracks were normalised to the maximum and minimum signal present in the view window. 1 – structure of genes detected by Cufflinks in MCF10A cell line. 2 – Gencode v.19 gene structure.
In order to further validate the experimental model, I investigated the relationship (based on the previous knowledge) between the differentially expressed genes. Pathway analysis using the IPA Ingenuity suite (Qiagen) showed that the top upstream regulators of the affected genes are growth factors (including EGF and platelet-derived growth factor (PDGF), see Table 3.1.3-1 top 20 regulators and Supplementary File 2 for all of the regulators). The analysis has also identified members of the MAPK pathways and a number of cytokines. Additionally, I analysed the functional link between genes affected by EGF stimulation. Figure 3.1.3–3A shows biological functions and diseases (Figure 3.1.3–3B) associated with the gene. The top biological functions were processes associated with cells growth, cell motility and regulation of gene expression. Unsurprisingly, the top associated diseases were cancer, neurological diseases and cell growth dependent conditions.
Table 3.1.3-1: Summary of the upstream regulator analysis of genes affected by stimulation of MCF10A cells with EGF.

<table>
<thead>
<tr>
<th>Upstream Regulator</th>
<th>Molecule Type</th>
<th>Predicted Activation State</th>
<th>Activation z-score</th>
<th>p-value of overlap</th>
<th>Number of Associated Genes</th>
<th>Proportion of EGF-Regulated Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDGF BB</td>
<td>complex</td>
<td>Activated</td>
<td>7.179</td>
<td>1.56E-77</td>
<td>62</td>
<td>40.26%</td>
</tr>
<tr>
<td>EGF</td>
<td>growth factor</td>
<td>Activated</td>
<td>6.01</td>
<td>9.37E-52</td>
<td>53</td>
<td>34.42%</td>
</tr>
<tr>
<td>TNF</td>
<td>cytokine</td>
<td>Activated</td>
<td>7.125</td>
<td>9.04E-48</td>
<td>74</td>
<td>48.05%</td>
</tr>
<tr>
<td>IL1B</td>
<td>cytokine</td>
<td>Activated</td>
<td>5.783</td>
<td>7.21E-43</td>
<td>56</td>
<td>36.36%</td>
</tr>
<tr>
<td>TGFBI</td>
<td>growth factor</td>
<td>Activated</td>
<td>6.251</td>
<td>7.12E-41</td>
<td>69</td>
<td>44.81%</td>
</tr>
<tr>
<td>NFkB (complex)</td>
<td>complex</td>
<td>Activated</td>
<td>5.401</td>
<td>7.07E-33</td>
<td>42</td>
<td>27.27%</td>
</tr>
<tr>
<td>P38 MAPK</td>
<td>group</td>
<td>Activated</td>
<td>5.161</td>
<td>1.45E-32</td>
<td>35</td>
<td>22.73%</td>
</tr>
<tr>
<td>TREM1</td>
<td>transmembrane receptor</td>
<td>Activated</td>
<td>4.592</td>
<td>1.36E-31</td>
<td>29</td>
<td>18.83%</td>
</tr>
<tr>
<td>IFNG</td>
<td>cytokine</td>
<td>Activated</td>
<td>5.783</td>
<td>1.64E-31</td>
<td>54</td>
<td>35.06%</td>
</tr>
<tr>
<td>F2</td>
<td>peptidase</td>
<td>Activated</td>
<td>5.403</td>
<td>1.66E-31</td>
<td>31</td>
<td>20.13%</td>
</tr>
<tr>
<td>TP53</td>
<td>transcription regulator</td>
<td>Activated</td>
<td>4.655</td>
<td>1.90E-31</td>
<td>56</td>
<td>36.36%</td>
</tr>
<tr>
<td>ERK</td>
<td>group</td>
<td>Activated</td>
<td>5.148</td>
<td>2.37E-31</td>
<td>30</td>
<td>19.48%</td>
</tr>
<tr>
<td>HGF</td>
<td>growth factor</td>
<td>Activated</td>
<td>4.756</td>
<td>6.38E-31</td>
<td>38</td>
<td>24.68%</td>
</tr>
<tr>
<td>CREB1</td>
<td>transcription regulator</td>
<td>Activated</td>
<td>5.531</td>
<td>7.04E-31</td>
<td>34</td>
<td>22.08%</td>
</tr>
<tr>
<td>RELA</td>
<td>transcription regulator</td>
<td>Activated</td>
<td>4.356</td>
<td>2.54E-30</td>
<td>34</td>
<td>22.08%</td>
</tr>
<tr>
<td>CSF2</td>
<td>cytokine</td>
<td>Activated</td>
<td>4.519</td>
<td>6.86E-30</td>
<td>36</td>
<td>23.38%</td>
</tr>
<tr>
<td>IRAK4</td>
<td>kinase</td>
<td>Activated</td>
<td>1.924</td>
<td>1.12E-29</td>
<td>21</td>
<td>13.64%</td>
</tr>
<tr>
<td>EGF</td>
<td>kinase</td>
<td>Activated</td>
<td>3.238</td>
<td>2.48E-29</td>
<td>31</td>
<td>20.13%</td>
</tr>
<tr>
<td>NR3C1</td>
<td>ligand-dependent nuclear receptor</td>
<td>Activated</td>
<td>-1.337</td>
<td>1.41E-28</td>
<td>40</td>
<td>25.97%</td>
</tr>
<tr>
<td>Pkc(s)</td>
<td>group</td>
<td>Activated</td>
<td>2.857</td>
<td>3.28E-28</td>
<td>27</td>
<td>17.53%</td>
</tr>
</tbody>
</table>
Figure 3.1.3–3. The analysis of functional correlation of EGF-regulated genes in MCF10A cell line

Ingenuity Pathways Analysis software was used to investigate the function correlation between 154 EGF-regulated genes in MCF10A cells. Data are presented as p-value of the correlation for A) molecular functions and B) disease categories. Only categories with p-value <0.01 are shown.
Finally, the defining feature of IE genes is their transient induction kinetics. As detailed in section 1.3.4, transcription of IE genes follows a wave-like pattern with rapid activation phase followed by rapid decrease in the level of mRNAs. In order to confirm whether the expression pattern of genes identified in the RNA-seq screen matches the expression pattern of the IE genes, I investigated the EGF-dependant expression of genes over a prolonged period of time using RT-qPCR. The expression of majority of the selected genes (FOS, CTGF, EGR2, PTGS2, Figure 3.1.3-4A-D) followed the expected transient expression pattern. The transcript levels reached maximum at 30-60 minutes after EGF stimulation. In the case of DUSP5 (dual specificity phosphatase 5), we can observe prolonged expression pattern with fairly high level of expression at 6 hours after stimulation with EGF (Figure 3.1.3-4E).

In summary, the initial investigation of the expression of protein-coding genes showed that the experimental model (MCF10A cells stimulated with EGF for 30 minutes) is an appropriate model to study and discover IE genes regulated by EGF. The protocol led to identification to 144 protein-coding genes that show characteristics of IE genes. Their expression increases significantly after EGF stimulation. The family-wide investigation showed that the function and disease association of the identified genes strongly correlate with information already present in the literature. Finally, detailed investigation of time-dependant expression of selected genes after EGF stimulation strongly resembles the expression of IE genes. Hence, the genes identified in the screen can confidently be classified as IE genes with transcription regulated by EGF. As a result, the experimental model can be confidently used for the purpose of identification and investigation of novel lncRNAs associated with EGF-dependant signalling pathways.
The expression of selected EGF-regulated genes in MCF10A cells was measured using RT-qPCR from total RNA. Cells were stimulated with EGF for 0, 15, 30, 45, 60, 75, 90, 120, 180, 240, 300 and 360 minutes. Data points for each replicated (n=3) with a line connecting average value (at each timepoint) are drawn. The expression was normalized to 0 timepoint using delta delta Ct method. Y-axis is log$_2$ transformed.

**Figure 3.1.3-4. Expression of EGF-regulated genes over time**

The expression of selected EGF-regulated genes in MCF10A cells was measured using RT-qPCR from total RNA. Cells were stimulated with EGF for 0, 15, 30, 45, 60, 75, 90, 120, 180, 240, 300 and 260 minutes. Data points for each replicated (n=3) with a line connecting average value (at each timepoint) are drawn. The expression was normalized to 0 timepoint using delta delta Ct method. Y-axis is log$_2$ transformed.
3.2. EGF-dependant expression of long noncoding RNAs

The initial analysis presented in section 3.1 provided evidence that the MCF10A cell line is a robust model to study the expression of IE genes. The combination of the high quality RNA-seq data and robust cellular model allows me to investigate whether lncRNAs play role in the EGF-dependant cellular response. The following sections focus on the identification of lncRNAs that are regulated by EGF-dependant signalling in the MCF10A cells.

3.2.1. Identification of long noncoding RNAs expressed in EGF-dependant manner

The key to quantitative studies of gene expression is the availability of well-annotated structure of reference genes. In the case on vast majority of human protein-coding genes, Refseq data base is the gold standard (Pruitt et al., 2014). However, since the field of lncRNAs is fairly young, the availability of large and well-curated databases is limited (e.g. Amaral et al., 2011; Volders et al., 2013). Currently, one of the largest data sets that contains annotated lncRNAs is GENCODE database (Harrow et al., 2012). The biggest strength and simultaneously the biggest limitation of GENCODE lies in its reliance on the transcripts that are expressed in multiple cancer cell lines sequenced by the ENCODE consortium (Djebali et al., 2012). For example, since the majority of ENCODE RNA-seq experiments are performed from unconditioned, stably grown cell lines, there is a possibility that EGF-induced lncRNAs have not been annotated in GENCODE. As a result, I decided to combine the GENCODE annotation with de novo assembly of the structure of genes present in our samples.

GENCODE version 19 contains 42014 unique genes with a length of at least 400 bp in at least one of gene’s isoforms. My analysis pathway (Figure 3.1.2-2) expanded that list to 51489 genes over 400 bp long. Differential expression analysis identified 482 differentially expressed genes. The expression of 317 of the genes increased (144 protein coding genes described in 3.1.3 and 173 non-coding genes) and 165 genes were downregulated (10 protein coding genes described in chapter 3.1.3 and 155 non-coding genes). Out of 173 upregulated non-coding genes, just over 50% (89
genes) were classified as long intergenic ncRNAs (lincRNAs) (Figure 3.2.1-1A) and 13 of them were previously described in GENCODE. Additionally, I observed 31 genes (9 intronic transcripts, 14 pseudogenes, and 8 antisense transcripts) that were previously annotated in GENCODE. Finally, a significant proportion of upregulated genes were classified as potential polymerase run-ons (transcripts that are present within 2 kb downstream of a known highly expressed transcript). In the case of downregulated transcripts, a majority (69%, 114 genes) was classified as lincRNAs (95 novel and 19 known, Figure 3.2.1-1B). Similarly to the upregulated genes, I also observed intronic transcripts (3), antisense transcripts (18), pseudogenes (5) and potential polymerase run-on (15). In order to maintain consistency across the thesis, in the following sections, when I refer to IncRNAs, I will include antisense transcripts, potential polymerase run-on transcripts, intronic transcripts and both classes of lincRNAs. When I refer to lincRNAs, I will include only known and novel lincRNAs.

Long non-coding RNAs are normally expressed at much lower level than protein-coding genes (Djebali et al., 2012). Similarly, the average expression value of EGF upregulated lincRNAs was equal to 1.61 FPKM at 0 minutes and 2.64 at 30 minutes, and it was much lower than the average expression of upregulated protein-coding genes (FPKM=29.92 at 0 minutes and 73.72 at 30 minutes, supplementary file 1). However, the EGF-induced increase in the expression was comparable between IncRNAs and protein-coding genes (on average 2.40-fold increase in expression of IncRNAs and 2.79-fold increase of protein-coding genes). In the case of downregulated IncRNAs, the average expression was equal to 0.55 FPKM at 0 minutes and 0.29 at 30 minutes, and the addition of EGF resulted in 0.48-fold reduction in expression (Supplementary file 1).

Manual investigation of upregulated IncRNA showed that the induction of both the intronic and antisense (located on the opposite strand of DNA and overlapping the reference protein-coding gene) lincRNAs tends to be coupled (follows the same level of induction) with the expression pattern of associated protein-coding genes. Furthermore, the investigation of the lincRNAs indicated that some of them might be located downstream of the TTS of known protein coding genes. In order to investigate the co-existence of protein-coding genes and lincRNAs, I investigated the global distribution of sequencing reads around the putative TSS of IncRNAs and
Figure 3.2.1-1. Distribution of differentially expressed genes

The distribution of EGF upregulated (A) and downregulated (B) genes in the MCF10A cells. A gene is classified as differentially expressed if p-value is smaller than 0.05, expression in at least one condition is above 0.1 FPKM and the absolute fold change is above 1.5 (i.e. 0.66 for downregulated genes). IncRNA – long non-coding RNA, lincRNA – long intergenic RNA. See section 3.2.1 or further details of the definitions of gene classes.
protein-coding genes (Supplementary file 3 contains profiles for all of the lncRNAs classes). Plots of the average density of the sequencing reads around the TSS of the upregulated lincRNAs indicated an increase in the density of sequencing reads upstream of the putative TSS (Figure 3.2.1-2A, right, sense strand). In comparison, the protein-coding genes did not have reads upstream of their TSS (Figure 3.2.1-2A, left). This indicates that at least some of the upregulated lincRNAs are expressed downstream of other genes.

In order to further investigate the pattern of lincRNAs expression, I plotted the expression pattern around the TSS on the level of individual genes (Figure 3.2.1-3). The profiles for the upregulated lincRNAs indicate presence of sequencing reads upstream of the TSS (on the same strand as the lincRNAs). However, the comparison of density across time points indicates that for the majority of lincRNAs, the increase in the read density takes place only downstream of the TSS (in the gene body, compare sense strand data for 30 minutes and 0 minutes on Figure 3.2.1-3). There was a small number of genes where the read density increased upstream of the TSS (in particular cluster 1, Figure 3.2.1-3); however, these gene had low density of reads downstream of the TSS indicating that they might be artefacts of automated gene structure annotation. The pattern for the protein-coding genes is much clearer (Figure 3.2.1-3). For majority of genes, there is a significant increase in the read density downstream of the TSS with a very low read density upstream of the TSS. However, similarly to the lincRNAs profiles, there is a small cluster of genes (cluster 2 for the upregulated protein-coding on Figure 3.2.1-3) where sequencing reads are present upstream of the TSS of the protein-coding genes. Finally, the expression pattern seems not to be continuous upstream and downstream of the putative TSS of lincRNAs (a decrease in the read density immediately upstream of the lncRNA TSS, Figure 3.2.1-2A, right).
Figure 3.2.1-2. Average profiles of RNA-seq read density around gene TSS
The average read density around the putative TSS of upregulated (A) and downregulated (B) protein-coding genes (left) and long intergenic non-coding RNAs (lincRNAs, right) identified in the RNA-seq experiment. X-axis indicates distance (nt) from the reference TSS (positive values are downstream of TSS). Grey and black lines represent reads on the sense strand (in direction) of the genes. Orange and red lines represent reads on the antisense strand (in opposite direction) of the genes. Arrows indicate the change in the density of reads on the sense strand after EGF stimulation. Read density was averaged across 50 bp bins in ±1000 bp window around TSS.
Figure 3.2.1-3. RNA-seq read density around TSS of individual genes

Read density (50 bp bins) in 2000 bp (-1000 to 1000) around the TSS of protein coding and long intergenic ncRNAs (lincRNAs, known and novel) identified in the RNA-seq experiment. Read densities were normalised using z-score per row across both time points and strand orientations. The read densities are split between sense and antisense DNA strands (in relation to the putative TSS of the gene. Gene expression profiles are clustered using hierarchical clustering. PCG – protein-coding genes.
The lack of inducible change in the density of the reads upstream of the upregulated lincRNAs combined with a decrease in read density near the TSS of lincRNAs supports the notion that, as a class, the identified lincRNAs are independent transcripts and not potential polymerase run-ons expressed downstream of other genes.

Similarly to the upregulated lincRNA, the downregulated lincRNA were located in close vicinity to protein-coding genes (manual inspection of the samples). However, in contrast to the upregulated lincRNAs, the downregulated lincRNAs were located on the opposite strand to the reads from the upstream genes. This indicates that they might originate from bi-directional promoters. The average expression profiles support this hypothesis (Figure 3.2.1-2B, right). On average, the read density decreases on the sense strand (in reference to the lincRNAs) and do not change on the anti-sense strand (despite much higher average density). On the individual gene level (Figure 3.2.1-3) the downregulated lincRNAs split into two broad categories: 1) lincRNAs expressed from bi-directional promoters (cluster 4, Figure 3.2.1-3) and 2) lincRNAs expressed from potentially, uni-directional promoters (cluster 5, Figure 3.2.1-3). In the case of downregulated lincRNAs expressed from bi-directional promoters, the read density on average decreased only on the sense strand and the density of the upstream antisense reads was not affected. This can be observed on the population level on Figure 3.2.1-2 (decrease from grey to black for lncRNA and no change on the other strand – orange and red) and on the individual gene profiles on Figure 3.2.1-3 (cluster 4). This indicates that the change in transcript level is selectively applied only to one of the transcripts. However, since the RNA-seq only measures the steady state level of RNA rather than the rate of transcription, reason for the decoupling of transcripts levels from the bi-directional promoters cannot be identified.

In order to provide further evidence that the changes in the expression of the transcripts were driven co-transcriptionally (rather than by changes in the stability of RNA), I investigated the RNApolII binding profiles around the putative TSS of MCF10A EGF-regulated genes using data from recently published work (Gardini et al., 2014). Using chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq), the authors investigated the RNApolII (antibody was specific to all forms of the polymerase) binding to chromatin at 0 and 20 minutes after EGF
stimulation of HeLa cells. It has previously been shown that the RNApolII tends to accumulate at the promoters of actively transcribed genes including IE genes (Core et al., 2008; O’Donnell et al., 2012). Additionally, the strength of the RNApolII binding signal positively correlates with the activity of the gene and increases both at the promoter and in the gene body (Guenther et al., 2007). Consistently with the previous observations, I also observed an increase in the density of RNApolII binding at the putative TSS of both protein-coding and lincRNA genes (Figure 3.2.1-4). In the case of upregulated genes, we can observe an increase in RNApolII binding (both in protein-coding and lncRNA genes) upon EGF stimulation (Figure 3.2.1-4A). In the case of protein-coding genes, the increase takes place both in the peak of paused polymerase at the promoter and in the body of the gene, but is much more modest in the upstream region (Figure 3.2.1-4A, left). In the case of lincRNAs, the change in binding takes place across the whole studied region including the TSS, the region downstream of TSS (within the gene body) and upstream of the TSS (Figure 3.2.1-4A, right). The RNA data suggested (Figure 3.2.1–3) that some of the lincRNAs upregulated by EGF stimulation are located immediately downstream of protein-coding genes. The binding of RNApolII upstream of the lincRNAs TSS is consistent with this possibility. However, the presence of a peak of RNApolII binding near the TSS of lincRNAs indicates paused RNApolII. The presence of paused RNApolII peak at the putative promoters and the decreases in the RNA-seq read density upstream of the putative promoters (Figure 3.2.1-2) indicate that the lincRNAs and upstream protein-coding genes are independent transcripts. Nevertheless, a strong conclusion cannot be drawn for all of the genes and further biochemical analysis on the level of individual genes is required to confirm the structure of individual lincRNAs and their relation to the upstream protein-coding genes.
Figure 3.2.1-4. Average profiles of RNApolIII ChIP read density around the TSS of EGF-regulated genes

Average read density (measured by ChIP-seq (Gardini et al., 2014)) of RNApolIII in HeLa cells around the putative transcription start sites (TSS) of gene whose expression is upregulated (A) or downregulated (B) after stimulation of MCF10A cells with EGF. Data are shown for protein-coding genes (left) and long intergenic non-coding RNAs (lincRNAs). The normalized read density was calculated in 50 bp bins ±1000 bp from the putative TSS. Arrows indicate the change in RNApolIII binding after EGF stimulation of HeLa cells.

Legend
- 0 minutes
- 20 minutes
In the case of the downregulated genes, the changes in binding of RNAPolII binding are not as strong. The binding near protein-coding downregulated genes decreases both at the TSS and in the body of the genes (Figure 3.2.1-4B), what is consistent with the decrease in the rate of transcription. In the case of downregulated lincRNAs, the change in binding near the TSS and in the gene body is small. Furthermore, the peak at the putative promoter near the TSS is wider (near 400 bp) than in the other classes of genes and extends into the upstream region. However, the change in binding does not take place upstream of the extended putative promoter. This can indicate that the transcription of the antisense transcripts from the bi-directional promoter is not affected.

In summary, the data presented in this section indicate that I have identified a panel of novel lincRNAs whose transcription is regulated by EGF-dependant signalling. In comparison to the protein-coding genes (where the transcription of vast majority increases), the transcription of lincRNAs can be up- or downregulated. In the case of the upregulated lincRNAs, the majority of lincRNAs tend to reside downstream of other genes; however only the transcription of lincRNAs changes after EGF stimulation indicating that they are indeed independent transcripts. On the other hand, the downregulated lincRNAs tend to split into two groups expressed from either bi-directional or uni-directional promoters. Further evidence supporting the discovered genes comes from the pattern of RNAPolII binding in the HeLa cells. However, it should be noted that HeLa cells is transcriptionally, highly permissive and has unstable genomic structure (Landry et al., 2013). As a result, caution should be taken during comparison of data obtained from MCF10A and HeLa cells. Taken together, these results indicate that the level of lincRNAs is regulated at the level of transcription in EGF-dependant manner.

3.2.2. Long intergenic non-coding RNAs show characteristics of IE genes

Having identified lincRNAs that are differentially expressed after EGF stimulation, I next wanted to determine the expression profile of selected lincRNAs over an extended time course after EGF stimulation. In the MCF10A cell line, the expression of protein-coding IE genes follows an expected, rapid and transient wave-like induction pattern (see section 1.3.4, Figure 3.1.3-4 and previously shown in (Amit et
The RNA-seq data only provided information for the 30 minutes time point; hence I selected 7 upregulated lincRNA and 17 downregulated lincRNA for validation by RT-qPCR. The upregulated transcripts showed various patterns of expression. Three followed transient induction with a peak in expression around 45 minutes after EGF stimulation, Figure 3.2.2-1, Cluster U2). Two had extended (up to 4 hrs) transient expression pattern (Figure 3.2.2-1, Cluster U1) and two had consistently high expression after stimulation (Figure 3.2.2-1, Cluster U3).

The downregulated genes showed two profiles of expression: one with minimum expression around the 45-60 minute time points (Figure 3.2.2-1, Cluster D2) and another with a minimum around 3-4 hrs after stimulation (Figure 3.2.2-1, Cluster D1).

One of the upregulated transcripts, XLOC_014669, showed the strongest induction (over 10-fold on linear scale, over 3.3-fold on log2 scale). In the RNA-seq data (Supplementary file 1) XLOC_014669 was the most induced transcripts not located near any other genes. Nevertheless, its expression did not follow classical IE gene expression profile. Its transcription increased immediately after EGF stimulation, but returned to basal level four hours post-EGF. When compared to the IE genes profiles (Figure 3.1.3-4), its expression profile was most similar to the expression of DUSP5.
Figure 3.2.2-1. Average profiles of lincRNA expression over an extended EGF time course
The expression of lincRNA expression was measured using the Fluidigm Biomark HD system (nanolitre volume RT-qPCR). Average of –ΔΔCt values are plotted (normalized to house-keeping genes and 0 minutes time point, n=3). Gene expression profiles were clustered using hierarchical clustering.
Previous ChIP-chip and ChIP-seq results from our laboratory indicate that ELK1 is an important transcription factor that directly regulates expression of (at least some) the IE genes in EGF-dependant manner (Boros et al., 2009; Odrowaz and Sharrocks, 2012). Hence, one might expect that ELK1 binding events will be enriched in the proximity of the newly identified genes regulated by EGF. Comparison of the ELK1 binding regions previously identified by Zaneta Odrowaz (Odrowaz and Sharrocks, 2012) with the EGF-regulated genes, showed that nearly 50% (63 out of 144) of EGF-induced protein coding genes are located within 100,000 bp from the ELK1 binding site (Figure 3.2.2-2A). Similarly, nearly 1/3 (28 out of 89) lincRNA are located with 100,000 bp from ELK1 binding sites. Both of those classes of gene are significantly enriched near the ELK1 sites when compared to the whole transcriptome (Figure 3.2.2-2A). Global distribution of the distance between ELK1 binding site and both classes of upregulated transcripts showed that protein-coding genes and lincRNAs are statistically located much closer to the ELK1 binding regions than the random genomic background (Figure 3.2.2-3A). The profile of protein-coding gene was shifted towards smaller distance for all of the genes. In the case of lincRNAs, the enrichment was less uniform. The genes in the bottom fifty percentiles were preferentially enriched near the ELK1 site and the top 50% of the most distant genes followed random genomic distribution.

In the case of the downregulated genes, the ELK1 binding sites are not enriched near the protein-coding nor lincRNAs (and Figure 3.2.2-3B). However, the distribution of the distance of the downregulated lincRNAs is not uniform across the entire range. 20% of downregulated lincRNAs, which are closest to the ELK1 sites, are located at distance smaller than random distribution would suggest (distance of $2^{10}$ (1024 bp) to $2^{16}$ (65536 bp), Figure 3.2.2-3B). Hence, the distribution of downregulated lncRNAs in 10,000 bp window is statistically significant (Figure 3.2.2-2B) but it is not in the 100,000 bp window.
Figure 3.2.2-2. Distribution of ELK1 binding sites in the proximity of EGF-regulated genes
Plots present total number (blue) of genes upregulated (A) and downregulated (B) by EGF in each indicated class of genes whose TSS is within 10 kbp (red) or 100 kbp (green) from the ELK1 binding regions. * - p-value <0.05 from χ² test performed against distribution of ELK1 binding regions in comparison to the whole transcriptome (on average 4.31% of all genes are within 10 kbp from ELK1 binding region and 22.36% are within 100 kbp).
Figure 3.2.2-3. Cumulative distribution of the distance between ELK1 binding sites and selected classes of genes.
Plots present the cumulative number of upregulated (A) and downregulated genes (B) in relation to the closest ELK1 binding regions. Distance was measured between the TSS of each of the genes in the class and the summit of closest ELK1 binding regions. Number of genes was normalized to the total number of genes in the class. A one-sided Kolmogorov-Smirnov test was used for statistical comparison.
Preferential distribution of ELK1 binding regions and EGF upregulated genes does not indicate causality, only a correlation. However, a direct link has been shown between the activity of ELK1 and some of the IE genes (see section 1.3.4). Hence, the preference of ELK1 binding regions to be located near the upregulated lincRNAs might indicate that at least some of them are regulated via ELK1-dependant mechanisms. XLOC_014669 – the most induced lincRNAs – has ELK1 binding site located within 3,200 bp upstream of its putative TSS. Investigation of XLOC_014669 genomic locus showed that its promoter region has characteristics of protein-coding gene promoters (H3K4me3 histone mark, RNApolII ChIP-seq signal from multiple ENCODE cell lines, discussed in further detail in section 4.1 (Djebali et al., 2012)). These results indicate that XLOC_014669 might be an interesting target for further characterisation.

3.2.3. Characterisation of the expression profile of EGF-regulated genes in the primary cancer samples

The results presented in section 3.1.3 indicate that the protein-coding gene upregulated by EGF stimulation are IE genes and the results from sections 3.2.2 indicate that lincRNAs are immediate-early lincRNAs (IE lincRNAs). The prevailing and expected view in the literature states that IE genes are likely to be overexpressed in primary cancer data (Healy et al., 2012; Krishna and Narang, 2008; Yang et al., 2013b). This expectation is derived from the fact that sustained MAPK signalling is observed in many cancers (Koboldt et al., 2012; Mitri et al., 2012; Slamon et al., 1987) (see section 1.3.5 for details) and logically it is assumed that the activation of MAPK pathways should lead to activation of transcription of IE genes. In order to further understand the biological significance of lincRNAs regulated by EGF, I investigated the expression pattern of both protein-coding genes and lincRNAs in primary cancer samples. Since IE lincRNAs have not been previously described and their expression pattern in the cancer cells is unknown, firstly, I will focus on the protein-coding IE genes identified in the MCF10A cells and I will use them as a benchmark to study the behaviour of the IE lincRNAs in cancer. Currently, the EBI Expression Atlas (http://www.ebi.ac.uk/gxa, (Petryszak et al., 2014)) contains easily accessible cancer expression data; hence I investigated the expression of MCF10A IE genes using data present therein.
Unexpectedly, the analysis of the expression of the whole panel of MCF10A IE genes showed that their expression changes only minimally in the majority of the protein-coding cancer samples (Figure 3.2.3-1A). The investigation of genes, that changed significantly (p<0.05) in the cancer studies, showed striking differences between different cancers (Figure 3.2.3-1B). For example, the majority of MCF10A IE genes were downregulated in the lung cancer samples, but upregulated in the primary pancreatic adenocarcinoma. A global view of the expression of MCF10A IE genes might mask individual changes in expression. Nonetheless, the expression of key IE genes associated with cancer development (members of EGR family of TF, JUN, FOS) did not show a clear directional pattern of expression (Figure 3.2.3-2).

The EBI Expression Atlas contains a limited number of cancer samples and it lacks data for some of the key cancers such as breast cancer. Nevertheless, the data showed an unexpected pattern of IE genes expression. The expression of IE genes does not follow a ‘one-fit-all’ expression pattern. Depending on the cancer, the expression of the IE genes induced by EGF in the MCF10A cell might increase or decrease and further investigation would be required to find underlying differences behind the expression patterns. However, these data suggest that if there is a significant change in the expression of the IE genes, most of them will follow the same pattern (Figure 3.2.3-1B). Hence, from the IE lincRNAs point of view, the IE protein-coding genes cannot be used a global benchmark; however, their expression can be used as a benchmark within a single cancer type. However, all of the comparison data available in the EBI Atlas originated from studies performed using microarrays. As a result, they do not contain information about the expression of EGF-regulated lincRNAs. Hence, in order to compare the expression profiles of EGF-regulated lincRNAs, I analysed the RNA-seq experiments from cancer patient data.
Figure 3.2.3-1. Expression of EGF-regulated genes in the primary tumour samples – EBI data
Boxplots represent fold change in the expression of MCF10A IE genes derived from the EBI expression atlas (Petryszak et al., 2014). A) Fold change in the expression of all of the MCF10A IE genes detected in cancers samples shown. B) Fold change in the expression of the MCF10A IE genes that changed significantly (p<0.05, from EBI statistics) in shown cancers. All of the fold change data are from the comparison of cancer samples with normal cells. * p-value <0.05 (Wilcoxon Signed Rank Test for differences, one sample, two tails).
Figure 3.2.3-2. Expression of selected IE genes in the primary tumour samples - EBI data
Data points represent fold change data (cancer to normal) of the expression of each indicated gene in the cancer data available from the EBI expression atlas (Petryszak et al., 2014). A) Fold change in the expression of the indicated MCF10A IE genes detected in the cancer samples. B) Fold change in the expression of the indicated MCF10A IE genes that changed significantly (p<0.05, from EBI statistics) in the cancer samples.
Currently, the biggest collection of cancer expression data is generated via collaborative efforts of The Cancer Genome Atlas (http://cancergenome.nih.gov/) and The International Cancer Genome Consortium (https://icgc.org). However, at the time of writing, we did not have access to their raw sequencing data required for meaningful analysis of lincRNA expression. As a result, I used two independent studies, one of breast cancer patients (Horvath et al., 2013) and another of lung cancer patients (Seo et al., 2012).

In both of the studies, the expression of the majority of the protein-coding IE genes was downregulated (134 out of 144 genes for the breast cancer and 109 out of 144 for the lung cancer, Figure 3.2.3-3, clusters 1 and 5). In the case of breast cancer data (Figure 3.2.3-3A, cluster 1), down-regulation was independent of cancer subtype. In the case of lung cancer data, paired samples (cancer and adjacent normal tissues) were available. This resulted in more controllable experimental conditions and a clear pattern of downregulation of many of the genes is observed (Figure 3.2.3-3B, cluster 5).

On the other hand, the expression of the lincRNAs upregulated by EGF in the MCF10A cells was much more diverse. In the breast cancer data, the expression of 44 lincRNAs was above 0.5 FPKM in at least one condition and, of them, 32 were downregulated (Figure 3.2.3-3A, cluster 3) in cancer and 12 were upregulated (cluster 4). Similarly, in the case of lung cancer, 89 EGF-induced lincRNAs were expressed with FPKM>0.1 in at least one condition and the majority of them (60 genes) were downregulated (Figure 3.2.3-3B, cluster 7). Out of the cancer upregulated lincRNAs, only five were common for both conditions (XLOC_006857, XLOC_012373, XLOC_014669, XLOC_018147 and XLOC_049015). Closer inspection of their expression showed that XLOC_014669 is the most highly induced in TNBC cancer (out of the five) and in the lung cancer samples (Supplementary files 4 and 5).
Figure 3.2.3-3. Expression of EGF-regulated genes in the primary tumour samples

Heatmaps of expression of protein-coding genes and lincRNAs upregulated after EGF stimulation of the MCF10A. A) Expression in primary breast cancer samples (Horvath et al., 2013). HER2–HER2-positive (Estrogen receptor- and progesterone receptor-negative) breast cancer, non-TNBC – triple positive (Estrogen receptor-, progesterone receptor- and HER2-positive) breast cancer, TNBC – triple negative (Estrogen receptor-, progesterone receptor- and HER2-negative) breast cancer. B) Expression in primary lung cancer samples and adjacent normal cells (Seo et al., 2012). Each row indicates a gene. The expression data are normalised per row A) to the average expression in the normal sample or B) as a z-score of the expression. On both panels, hierarchical clustering was used to cluster the gene expression profiles. For the lincRNA, only genes with FPKM (in at least one condition) above 0.5 (A) or above 0.1 (B) are shown. Arrow indicates XLOC_014669
The expression pattern of the EGF-upregulated IE genes (at least in case of selected studies) is strikingly different from the expected (increase in cancer) results. The difference might originate from the bias present in the experiments. In case of the breast cancer sample, normal sample are not fresh biopsies from patients but breast organoids (Horvath et al., 2013). In the case of lung adenocarcinomas, PCR was used to screen some of the samples for *EGFR* (point mutations in exons 18-21), *KRAS* (exon 2 point mutations) mutations and *EML4-ALK* fusions (known drivers of lung cancer) and the positive samples were not used in the study (Seo et al., 2012). However, the sequencing data showed that over half of the samples carried mutations in other regions of these genes. As a result, a slight bias exists in the samples and caution should be taken in the interpretation of the data.

Although the majority of EGF-regulated IE genes and lincRNAs were downregulated in the cancer samples, notable examples of lincRNAs that are strongly upregulated in both cancer studies were XLOC_006857, XLOC_012373, XLOC_014669, XLOC_018147 and XLOC_049015. In both cases this could be important during the tumorigenesis of these cancers.

### 3.3. Chapter summary

The results presented in this chapter showed that I have established an effective protocol for discovery of lincRNAs whose transcription is regulated by EGF. I identified 144 protein-coding genes which are induced by EGF in the MCF10A cell line. Within the list of identified genes, I found some of well-known IE genes (*FOS, JUN, EGR1-4*) (Amit et al., 2007; O’Donnell et al., 2012; Sharrocks, 2013). By comparing the expression of IE genes and ELK1 genomic occupancy, I showed that ELK1 binding is enriched in the proximity of EGF-regulated IE genes. Unexpectedly, the patterns of expression of EGF-regulated IE genes in the cancer samples were different to logical predictions (Healy et al., 2012; Krishna and Narang, 2008; Yang et al., 2013b). Most of the expression data showed that the expression of IE genes decreases in the cancer samples. Nevertheless, we only used small selection of samples and further global studies of the expression of EGF-regualted IE genes in cancer are needed. These studies however, are beyond the scope of this thesis.
The main aim of my work was identification of long non-coding RNAs regulated by EGF stimulation. During the initial analysis, I identified multiple classes of long non-coding RNAs; however, the expression of 89 EGF-upregulated long intergenic ncRNAs (lincRNAs) seemed to be the most similar to the expression profile of protein-coding IE genes. The expression patterns of the upregulated lincRNAs following established pattern (rapid, transient induction) of expression of protein-coding IE genes. Promoters of lincRNAs were enriched for EGF-inducible RNApolII binding events and ELK1 binding sites were also enriched near upregulated lincRNAs. These results suggest that the transcription of EGF-upregulated lincRNAs might be regulated by the EGF → MAPK (ERK2) → ELK1 signalling cascade.

In order to further understand the mechanism of the EGF-stimulation and role of lincRNAs in the regulation of cellular processes, I selected XLOC_014669 for further studies. The rationale behind XLOC_014669 selection is as follow:

- XLOC_014699 was one of the most induced genes after EGF stimulation,
- it distal (65 kbp) from any protein-coding genes and miRNAs (truly intergenic in nature)
- it has an ELK1 binding region in its proximity
- RNApolII binds its promoter and the binding increases after EGF-stimulation (in HeLa cells)
- its transcription is rapid and transient (decreases after 240 minutes post-EGF)
- it is strongly upregulated in breast and lung cancer samples

For the purpose of this thesis, I decided to name XLOC_014669 as an EGF-inducible ncRNA 1 (EIN1) and the following chapters will focus on its characterisation.
4. Characterisation of *EIN1* – EGF-inducible non-coding RNA

Long intergenic non-coding RNAs are a class of non-coding RNAs that, arguably; are the most studied. Section 1.4.3 includes description of notable, cancer-related lincRNAs. One of the key elements necessary for understanding of the function of lincRNAs is description of their cellular properties. Results presented in chapter 3 identified *EIN1* as the most highly induced lincRNA after EGF stimulation that is upregulated in breast and lung cancer samples. The following chapter will focus on characterisation of *EIN1*. I will describe its cellular localisation, the lack of protein-coding potential, the structure of its genomic locus and I will investigate how its expression is regulated by EGF-signalling.

4.1. Genomic locus of *EIN1*

4.1.1. Identification of genomic features associated with *EIN1*

*EIN1* is located on the Watson strand of human chromosome 14q24.3. Based on the RNA-seq data, the estimated coordinates can be defined as chr14:77,425,827-77,437,282 (hg19 coordinates). A Sashimi plot (a descriptive plot counting the number of reads that span the exon-exon junction) of the nuclear RNA-seq data indicates that there is very small amount of splicing present in the gene (Figure 4.1.1-1). Out of 1110 reads that mapped to *EIN1* locus, only 3 reads spanned the junctions (0.27%). In all of the sequencing libraries, 5.44% of the reads spanned the exon-exon junctions. As a result, I did not observe strong evidence supporting *EIN1* splicing. *EIN1* genomic locus overlaps two GENCODE genes (Harrow et al., 2012) – *RP11-7F17.7* – a spliced, uncharacterised long noncoding RNA and *RP11-7F17.1* – a non-coding pseudogene of *RPLP1* (60S acidic ribosomal protein P1). However, the spliced reads from the RNA-seq did not support the structure of *RP11-7F17.7* (reads did not connect first and the second, nor second and third exon, Figure 4.1.1-1). Based on these data, I concluded that in the nucleus, both in stimulated and non-stimulated cells, *EIN1* is mainly in the non-spliced form.
Figure 4.1.1-1. Sashimi plot of EIN1 structure and expression

Sashimi plot of the spliced junctions detected in the nuclear RNA-seq data. The coverage represents raw sequencing data for the EIN1 genomic locus. Lines indicate detected splice junction together with the number of reads supporting the junction. Only junctions with at least 2 reads are shown. The gene structure for gencode RP11-7F17.7 and RP11-7F17.1 is shown.
This view however, was challenged by subsequent results (see chapter 5). When I performed RNA-seq on longer time-courses after EGF stimulation (0-180 minutes), I observed additional splice forms of *EIN1* (Figure 4.1.1-2, the details of the experiment set up are presented in chapter 5). The new sequencing data supported at least five different isoforms of *EIN1* (bottom section of Figure 4.1.1-2). However, none of them was compatible with the structures presented by GENCODE data.

BLAST search of the genomic sequence of *EIN1* resulted in only a single complete sequence similarity with human genome (located at *EIN1* locus) and there were no similarities between the human genomic sequence and mouse genomic sequence. Investigation of the conservation of the region showed that most of the genomic sequence of *EIN1* gene is not conserved between human and mouse; however, there was good conservation of the promoter region (Figure 4.1.1-3). Within the core promoter two feature of TATA-containing promoter can be identified. 32 bp upstream of the putative TSS of *RP11-7F17-7* (the putative TSS of *EIN1* overlaps with the TSS of *RP11-7F17-7*) I observed a TATA-like sequence (TACAAAAG) and an Inr-like sequence (TCATACT) overlapped the TSS (Figure 4.1.1-3 and see section 1.2 for details of promoter structure).

Promoters of protein-coding genes as well as many lincRNAs are marked by specific well-established genomic features (Table 1.1.2-1). Trimethylation of lysine 4 on histone 3 (H3K4me3) is a well-known mark associated with promoters and acetylation of lysine 27 on histone 3 (K3K27ac) is associated with active promoters and enhancers (Shlyueva et al., 2014). In order to investigate the distribution of the histone marks near the *EIN1* gene body, I used the ENCODE data from the MCF7 cell line line (Figure 4.1.1-3). I observed a strong H3K4me3 peaks in the gene body of *EIN1* (around 1.5 kb downstream of TSS). An H3K4me3 binding peak was not present at the RNA-seq derived promoter of *EIN1*. Additionally, there was a strong mark for H3K27ac in both locations. On the other hand, I did not observe strong H3K36me3 (mark of actively transcribed genes) binding in the gene body of *EIN1*. Finally, the repressive marks (H3K9me3 and H3K27me3) were absent. The data from the MCF7 cell line should be interpreted with caution, as the cells were not stimulated with EGF; hence, the histone marks presented indicate steady state level. Nevertheless, the presence of positive marks of active promoters (H3K4me3 and H3K27ac) and lack of repressive marks (H3K9me3 and H3K27me3) is consistent
Figure 4.1.1-2 Sashimi plot of EIN1 structure and expression in total RNA-seq

Sashimi plot of the spliced junctions detected in the total RNA-seq data. The coverage represents raw sequencing data for the EIN1 genomic locus. Lines indicate detected splice junctions together with the number of reads supporting the junction. Only junctions with at least 2 reads are shown. The gene structure for gencode RP11-7F17.7 and RP11-7F17.1 is shown together with the gene structure of EIN1 deduced from the sequencing data.
Figure 4.1.1-3. Genomic features of EIN1 locus

A screenshot of the genomic features observed near the EIN1 genomic locus compiled from UCSC genome browser data. Data for the histone modifications (H3K3me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3) ChIP-seq are derived from ENCODE data from the MCF7 cell line (Dunham et al., 2012). RNApolII ChIP-seq data from HeLa cells (before and after EGF induction) are derived from (Gardini et al., 2014). The RNA-seq data are from this study. Human promoter sequence of RP11-7F17.7 (TSS±1000bp) was aligned with five primates, mouse and rat reference sequence using multiz aligner (Blanchette et al., 2004). Inr-like – initiator-like sequence.
with active promoter state and when combined with the presence of TATA-like and Inr-like sequences, supports the notion that *EIN1* is actively transcribed by RNApolII.

In order to further investigate the activity of *EIN1* promoter and its transcriptional potential, I investigated the RNApolII binding in HeLa cells before and after stimulation with EGF (Figure 4.1.1-3) (Gardini et al., 2014). Although the experiment was performed in HeLa cells, my previous analysis showed that the RNApolII binding in Hela cells increases in the vicinity of the EGF-regulated IE genes from MCF10A cells (Figure 3.2.1-4). In the case of *EIN1*, in the EGF-starved cells, I observed RNApolII pausing at the promoter of the gene (Figure 4.1.1-3). There were two peaks of binding. One was directly adjacent to the putative promoter of *EIN1* and the other was located around 1.5 kb downstream of RNA-seq derived TSS and it overlapped with H3K4me3 histone mark. After EGF stimulation, I observe an increase in RNApolII bind, both at the promoter and in the gene body of *EIN1*. This is consistent with RNApolII being released from a paused state and increased processivity along the gene body (Figure 4.1.1-3).

To summarise, *EIN1* is a poorly conserved gene, most likely with multi-exonic transcripts expressed from a conserved putative promoter. The histone mark data are consistent with marks observed for RNApolII genes. This observation, combined with the RNApolII binding in the promoter of unstimulated cells and induction of binding after EGF stimulation, strongly suggests RNApolII transcribes *EIN1*. There is a possibility that *EIN1* is expressed from another promoter located around 1.5 kb downstream of the core promoter; however, the depletion of *EIN1* presented in chapter 5 indicates that *EIN1* is expressed only from the promoter identified in this section.

4.1.2. Investigation of the structure of *EIN1* transcripts

The RNA-seq data indicated that *EIN1* might exist in multiple splice isoforms. RNA-seq data did not however provide compelling evidence about the structure of the genes. First of all, the identification of both 3’- and 5’-ends is impaired due to limited sequencing coverage of transcript ends. In order to identify the location of 5’-end of *EIN1* in MCF10A cells, I performed 5’ rapid amplification of cDNA ends (5’ RACE).
Firstly, in order to simplify the search for the 5’-end I perform PCR using cDNA as a template with three antisense primers (GSP1, GSP2 and GSP3) designed for a region 400-800 bp downstream of the 5’-end estimated from the RNA-seq data and from the ENCODE RP11-7F17.7 transcript (the RP11-7F17.7 5’ end is located on chromosome 14, position 77,425,981, Figure 4.1.2-1A). Additionally, I designed two sense primers located upstream (Anchor2) and downstream (Anchor1) of the putative TSS. PCR reactions performed on a cDNA library (produced with polyT primer from MCF10A stimulated with EGF for 60 minutes) should result in products only in the case of antisense primers paired with Anchor1 but not with Anchor2. Additionally, the efficiency of reaction would show that the primers and cDNA are of high enough quality for the 5’ RACE reaction. PCR with the Anchor1 and GSP1-3 against the cDNA template resulted in products of expected size (Figure 4.1.2-1B). PCR with Anchor2 and GSP1-3 against the cDNA template did not have detectible products (Figure 4.1.2-1B). In order to confirm that Anchor2 and GSP1-3 work in a PCR reaction, I performed PCR using genomic DNA (gDNA) as a template (Figure 4.1.2-1B). Only GSP2 and Anchor2 resulted in a product of expected size. From the above result it can be deduced that:

- Primers GSP1, GSP2 and GSP3 are specific against EINI transcripts (3 bands of expected size with cDNA template)
- Primers Anchor1 and Anchor2 are specific against the region of interest
- EINI 5’ end is located between primers Anchor1 and Anchor2 and near the 5’-end of RP11-7F17.7
Figure 4.1.2-1. Identification of EIN1 5’ ends

A) Schematic of the primers used for the identification of the EIN1 TSS. The location of potential TSS was based on the nuclear RNA-seq data. The distance between the putative TSS and 5’ end of the primer is shown. B) End-point PCR of selected primers using genomic DNA (gDNA) or polyA-tailed cDNA produced from MCF10A cells stimulated for 60 minutes as the templates. GSP1-3 – gene specific primer 1-3, Anchor1-2 – control primers.
Confirmation of the estimated location of *EIN1* 5’-end simplified the experimental design of the 5’ RACE protocol. In order to identify the 5’-end, I performed two rounds of PCR. The PCR with GSP2 and universal primer (Figure 4.1.2-2B) did not produce a single band. The second, nested, PCR using GSP3 and universal primer (Figure 4.1.2-2C) resulted in two bands. After sequencing only the longer (~500 bp long) band aligned to the region of interest and it identified the location of 5’-end of *EIN1* at position chr14: 77,425,979, just two nucleotides upstream of RP11-7F17.7 annotated 5’-end (Figure 4.1.2-2D). The shorted PCR fragment was an artefact that did not align to human genome. In order to further confirm the location of the 5’-end, I compared my data with data from FANTOM5 project (Forrest et al., 2014). FAMTOM5 project uses the Cap Analysis Gene Expression (CAGE) technique to identify the 5’-ends of all of the genes using next generation sequencing. Remarkably, the location of 5’-end from the FANTOM5 data was exactly the same as the results from the 5’ RACE (Figure 4.1.2-2D). Taken together, these results indicate that the 5’-end (and as a result – TSS) of *EIN1* is located at position chr14: 77,425,979.

The result of the identification of the 3’-end of *EIN1* transcripts was ambiguous. Most of the 3’ RACE techniques rely on cDNA priming using polyT primer with specific sequence at the 5’ end. However, the sequence of *EIN1* contains multiple polyA sequence stretches that results in mispriming of the cDNA-synthesis primer. As a result, at the time of writing, exact location of the 3’ end of *EIN1* transcripts is not known.
Figure 4.1.2-2. Identification of EIN1 5’ ends – 5’ RACE

A) Schematic of the 5’ RACE experimental workflow. B) After cDNA 3’ tailing with dCTP, samples were amplified with a Universal primer and gene specific primer 2 (GSP2) and analysed on an agarose gel. Subsequently, the PCR product was purified and amplified using a Universal primer and GSP3. C) Two bands were identified (arrowed) and sequenced. D) After alignment, only band 1 mapped to the region of interest. A nucleotide (bold) in the known genomic sequence indicates the TSS of EIN1. The location of the EIN1 TSS was overlaid with the RNA-seq and CAGE data from FANTOM5 (Forrest et al., 2014). CAGE all data – combined all CAGE reads from all studies cell lines. CAGE Top cell lines – data from the cell line with the highest read number in the window shown. ssDNA – single stranded DNA, dsDNA – double stranded DNA.
4.1.3. Expression profile of *EIN1* in human tissues

As the final element of *EIN1* characterisation, I investigated the expression of *EIN1* in normal tissues. I used data provided by the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013). GTEx provides RNA-seq for multiple human tissue samples collected from healthy tissues samples collected during autopsies. Currently, GTEx provides its data as FPKM values for GENCODE annotated gene (Harrow et al., 2012). In order to analyse expression data for *EIN1*, I used *RP11-7-F17.7* GENCODE gene, the closest match to *EIN1* transcripts, as a proxy (see section 4.1.1). *RP11-7-F17.7* was expressed in multiple tissues including multiple elements of the cardiovascular system (heart: left ventricle and atrial appendage, arteries: aorta, coronary artery and tibial artery and whole blood), in mammary glands, in multiple elements of gastrointestinal tract and in bladder (Figure 4.1.3-1). Furthermore, when we investigated the expression of *EIN1* locus in the developing human embryos, we observed that the only tissue expressing *EIN1* was left ventricle of the heart (Prof. Neil Hanley, University of Manchester, personal communication). Additionally, the GTEx project analysed lymphocytes transformed with Epstein-Barr virus (EBV). *RP11-7-F17.7* was also expressed in these EBV-transformed cell lines. In other tissues, *RP11-7-F17.7* was expressed at a limited level and it practically was virtually undetectable in the brain samples.
Figure 4.1.3-1. RP11-7F17.7 expression in normal tissues

Boxplot of the expression of RP11-7F17.7 in normal tissues. The midpoint of each boxplot is median expression. Data were downloaded from http://www.gtexportal.org/home/gene/RP11-7F17.7 (Lonsdale et al., 2013)
Another key feature of GTEx is the availability of data for *cis* expressed quantitative trait loci (*cis*-eQTLs). *Cis*-eQTLs are defined as genomic loci that are located within 100 kb from the TSS of the gene of interest and their presence correlates with changes in the expression of the gene (Lonsdale et al., 2013). Strikingly, in the case of *RP11-7F17*, there are 13 eQTLs that correlate with a decrease of its expression in several tissues (Table 4.1.3-1 and Figure 4.1.3-2). All of the *RP11-7F17.7* associated eQTLs are single nucleotide polymorphism (SNPs, Table 4.1.3-1) and the frequency of the minor allele is relatively high (around 40%, Table 4.1.3-1, MAF). Investigation of the ancestral alleles (with an assumption that dominant chimpanzee allele is ancestral) showed that all of the minor allele are derived alleles indicated novel evolutionary origin. Secondly, despite *RP11-7F17.7* expression in multiple tissue (Figure 4.1.3-1), the only tissues affected by eQTLs are left ventricle of the heart, tibial artery, aorta and esophagus muscularis. Of the SNPs downregulating *RP11-7F17.7*, the most significant regulators are rs12889775, rs10873298, rs10873299 and rs8181996. All of them are located in close proximity to the *RP11-7F17.7* TSS, within its gene body. Search of known diseases associated with *RP11-7F17.7* eQTLs yielded no results.

The expression data from normal tissues provides interesting observations, however unequivocal conclusions cannot be made. The expression of *RP11-7F17.7* is a proxy of *EIN1* expression and it is possible that in various tissues different isoforms of the gene are expressed. Secondly, the eQTL data are correlation based and no causality can be derived from it. However, using the expression data from adult and embryo tissues and the eQTL correlation data, it is tempting to speculate that *EIN1* is important for the development and maintenance of heart (and possible other elements of cardiovascular system). However, at the moment we are not able to test this hypothesis. The *EIN1* locus is poorly conserved in established model organism including mouse and rat. As a result, the remaining sections of the thesis will focus on characterisation of *EIN1* role in the cellular context.
Table 4.1.3-1: Summary of the SNPs significantly associated with EIN1 expression – cis eQTLs

<table>
<thead>
<tr>
<th>SNP Id</th>
<th>P-Value</th>
<th>Effect Size</th>
<th>Tissue</th>
<th>Genotype (major/minor allele)</th>
<th>Location (hg19)</th>
<th>Minor Allele Frequency (MAF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12889775</td>
<td>1.20E-08</td>
<td>-0.87</td>
<td>Artery Aorta</td>
<td>G/T</td>
<td>chr14:77426077</td>
<td>T=0.4657/2332</td>
</tr>
<tr>
<td>rs10873298</td>
<td>1.60E-07</td>
<td>-0.86</td>
<td>Artery Aorta</td>
<td>C/T</td>
<td>chr14:77426525</td>
<td>T=0.4383/2195</td>
</tr>
<tr>
<td>rs10873299</td>
<td>1.80E-07</td>
<td>-0.86</td>
<td>Artery Aorta</td>
<td>A/G</td>
<td>chr14:77426711</td>
<td>G=0.4377/2192</td>
</tr>
<tr>
<td>rs8181996</td>
<td>1.70E-07</td>
<td>-0.86</td>
<td>Artery Aorta</td>
<td>C/G</td>
<td>chr14:77427469</td>
<td>C=0.4355/2181</td>
</tr>
<tr>
<td>rs6574341</td>
<td>7.90E-07</td>
<td>-0.86</td>
<td>Artery Aorta</td>
<td>C/T</td>
<td>chr14:77422803</td>
<td>T=0.4643/2325</td>
</tr>
<tr>
<td>rs17810431</td>
<td>1.30E-06</td>
<td>-0.67</td>
<td>Artery Aorta</td>
<td>A/T</td>
<td>chr14:77430681</td>
<td>A=0.4758/2383</td>
</tr>
<tr>
<td>rs12889775</td>
<td>6.90E-30</td>
<td>-1.1</td>
<td>Artery Tibial</td>
<td>G/T</td>
<td>chr14:77426077</td>
<td>T=0.4657/2332</td>
</tr>
<tr>
<td>rs10873298</td>
<td>1.20E-29</td>
<td>-1.1</td>
<td>Artery Tibial</td>
<td>A/G</td>
<td>chr14:77426525</td>
<td>T=0.4383/2195</td>
</tr>
<tr>
<td>rs10873299</td>
<td>1.20E-29</td>
<td>-1.1</td>
<td>Artery Tibial</td>
<td>A/G</td>
<td>chr14:77426711</td>
<td>G=0.4377/2192</td>
</tr>
<tr>
<td>rs8181996</td>
<td>1.50E-29</td>
<td>-1.1</td>
<td>Artery Tibial</td>
<td>C/G</td>
<td>chr14:77427469</td>
<td>C=0.4355/2181</td>
</tr>
<tr>
<td>rs6574341</td>
<td>5.70E-21</td>
<td>-1.1</td>
<td>Artery Tibial</td>
<td>C/T</td>
<td>chr14:77422803</td>
<td>T=0.4643/2325</td>
</tr>
<tr>
<td>rs17810431</td>
<td>1.30E-18</td>
<td>-0.89</td>
<td>Artery Tibial</td>
<td>A/T</td>
<td>chr14:77430681</td>
<td>A=0.4758/2383</td>
</tr>
<tr>
<td>rs11622638</td>
<td>3.20E-18</td>
<td>-0.9</td>
<td>Artery Tibial</td>
<td>A/G</td>
<td>chr14:77429307</td>
<td>A=0.4575/2291</td>
</tr>
<tr>
<td>rs4399486</td>
<td>5.90E-11</td>
<td>0.83</td>
<td>Artery Tibial</td>
<td>C/T</td>
<td>chr14:77421660</td>
<td>C=0.3992/1999</td>
</tr>
<tr>
<td>rs4899629</td>
<td>7.20E-11</td>
<td>-0.69</td>
<td>Artery Tibial</td>
<td>C/T</td>
<td>chr14:77429105</td>
<td>T=0.3578/1792</td>
</tr>
<tr>
<td>rs76925617</td>
<td>7.60E-11</td>
<td>-0.69</td>
<td>Artery Tibial</td>
<td>G/T</td>
<td>chr14:77429104</td>
<td>T=0.3566/1787</td>
</tr>
<tr>
<td>rs17810764</td>
<td>7.00E-08</td>
<td>-0.63</td>
<td>Artery Tibial</td>
<td>A/G</td>
<td>chr14:77452666</td>
<td>A=0.3528/1767</td>
</tr>
<tr>
<td>rs11622645</td>
<td>2.90E-06</td>
<td>-0.55</td>
<td>Artery Tibial</td>
<td>G/T</td>
<td>chr14:77452204</td>
<td>T=0.3654/1830</td>
</tr>
<tr>
<td>rs10148692</td>
<td>3.50E-06</td>
<td>-0.54</td>
<td>Artery Tibial</td>
<td>A/G</td>
<td>chr14:77444317</td>
<td>A=0.4008/2007</td>
</tr>
<tr>
<td>rs10873298</td>
<td>1.30E-15</td>
<td>-0.99</td>
<td>Esophagus Muscularis</td>
<td>C/T</td>
<td>chr14:77426525</td>
<td>T=0.4383/2195</td>
</tr>
<tr>
<td>rs10873299</td>
<td>1.30E-15</td>
<td>-0.99</td>
<td>Esophagus Muscularis</td>
<td>A/G</td>
<td>chr14:77426711</td>
<td>G=0.4377/2192</td>
</tr>
<tr>
<td>rs8181996</td>
<td>1.30E-15</td>
<td>-0.99</td>
<td>Esophagus Muscularis</td>
<td>C/G</td>
<td>chr14:77427469</td>
<td>C=0.4355/2181</td>
</tr>
<tr>
<td>rs12889775</td>
<td>2.10E-15</td>
<td>-1</td>
<td>Esophagus Muscularis</td>
<td>G/T</td>
<td>chr14:77426077</td>
<td>T=0.4657/2332</td>
</tr>
<tr>
<td>rs11622638</td>
<td>1.10E-10</td>
<td>-0.82</td>
<td>Esophagus Muscularis</td>
<td>A/G</td>
<td>chr14:77429307</td>
<td>A=0.4575/2291</td>
</tr>
<tr>
<td>rs17810431</td>
<td>1.10E-10</td>
<td>-0.82</td>
<td>Esophagus Muscularis</td>
<td>A/T</td>
<td>chr14:77430681</td>
<td>A=0.4758/2383</td>
</tr>
<tr>
<td>rs6574341</td>
<td>5.80E-09</td>
<td>-0.83</td>
<td>Esophagus Muscularis</td>
<td>C/T</td>
<td>chr14:77422803</td>
<td>T=0.4643/2325</td>
</tr>
<tr>
<td>rs4399486</td>
<td>1.70E-06</td>
<td>0.78</td>
<td>Esophagus Muscularis</td>
<td>C/T</td>
<td>chr14:77421660</td>
<td>C=0.3992/1999</td>
</tr>
<tr>
<td>rs10873298</td>
<td>1.30E-14</td>
<td>-1.1</td>
<td>Heart Left Ventricle</td>
<td>C/T</td>
<td>chr14:77426525</td>
<td>T=0.4383/2195</td>
</tr>
<tr>
<td>rs10873299</td>
<td>1.30E-14</td>
<td>-1.1</td>
<td>Heart Left Ventricle</td>
<td>A/G</td>
<td>chr14:77426711</td>
<td>G=0.4377/2192</td>
</tr>
<tr>
<td>rs8181996</td>
<td>1.30E-14</td>
<td>-1.1</td>
<td>Heart Left Ventricle</td>
<td>C/G</td>
<td>chr14:77427469</td>
<td>C=0.4355/2181</td>
</tr>
<tr>
<td>rs12889775</td>
<td>2.30E-14</td>
<td>-1</td>
<td>Heart Left Ventricle</td>
<td>G/T</td>
<td>chr14:77426077</td>
<td>T=0.4657/2332</td>
</tr>
<tr>
<td>rs6574341</td>
<td>6.10E-12</td>
<td>-1</td>
<td>Heart Left Ventricle</td>
<td>C/T</td>
<td>chr14:77422803</td>
<td>T=0.4643/2325</td>
</tr>
<tr>
<td>rs17810431</td>
<td>8.70E-11</td>
<td>-0.89</td>
<td>Heart Left Ventricle</td>
<td>A/T</td>
<td>chr14:77430681</td>
<td>A=0.4758/2383</td>
</tr>
<tr>
<td>rs11622638</td>
<td>1.00E-10</td>
<td>-0.92</td>
<td>Heart Left Ventricle</td>
<td>A/G</td>
<td>chr14:77429307</td>
<td>A=0.4575/2291</td>
</tr>
<tr>
<td>rs4899629</td>
<td>7.40E-07</td>
<td>-0.74</td>
<td>Heart Left Ventricle</td>
<td>C/T</td>
<td>chr14:77429105</td>
<td>T=0.3578/1792</td>
</tr>
<tr>
<td>rs76925617</td>
<td>7.40E-07</td>
<td>-0.74</td>
<td>Heart Left Ventricle</td>
<td>G/T</td>
<td>chr14:77429104</td>
<td>T=0.3568/1787</td>
</tr>
</tbody>
</table>
Figure 4.1.3-2. SNPs associated with EIN1 expression – cis-eQTLs

Genome browser view of SNPs associated with RP11-7F17.7 expression. Only cis-eQTLs (SNPs located within 100 kb of TSS of EIN1 that affect its expression were considered). Data were downloaded from http://www.gtexportal.org/home/gene/RP11-7F17.7 (Lonsdale et al., 2013). Red dots – SNPs significantly associated with RP11-7F17.7, grey dots, SNPs associated with other genes. Only tissues with significant associations are shown. For details of SNPs see table 4.1.3-1.
4.2. Characterisation of the protein-coding potential and localisation of EIN1

An inherent feature of lincRNAs is a lack of protein coding potential. The ability to code for proteins can be investigated both computationally and biochemically. The following sections will summarise the investigation of EIN1 coding potential.

4.2.1. Computational investigation of coding potential of EIN1

In previous studies researchers assessed coding potential using phyloCSF (Lin et al., 2011), the presence of pfam domains (Guttman et al., 2010), ribosomal profiling (Guttman et al., 2013) or simple the length of the longest open reading frame (ORF) (Guttman et al., 2010). In order to comprehensively assess the coding potential of EIN1, I used all of the above approaches and additionally I investigated the recent data from global mass spectroscopy profiles of human tissues (Kim et al., 2014). The RNA-seq data indicated that EIN1 is at least partially spliced; however, there is a possibility that one of the functional isoforms of EIN1 is never spliced. As a result, in order to account for all of the possibilities, for the purpose of this analysis, I assumed that EIN1 did not have any splice forms and I also investigated all of the possible isoforms deduced from the RNA-seq data (see section 4.1.1).

The longest, potential ORF identified in the unspliced sequence of EIN1 was 158 amino acids (aa) long and it overlapped the sequence of the pseudogene gene found ~9 kb downstream of the 5’-end of EIN1. The longest, potential ORF for the spliced EIN1 forms was 79 aa long and was located ~4.8 kb downstream of the 5’-end of EIN1. ORFs did not span any of the potential exon-exon junctions. An arbitrary cut-off for the ORF length in the lincRNAs is 100 aa, however lincRNA with potential ORFs longer than 150 aa have been reported (Guttman et al., 2010; Maamar et al., 2013). In order to further validate the coding potential of EIN1, I calculated phyloCSF score for each ORF identified in the genomic sequence of EIN1 (and by proxy in all of the spliced forms) (Lin et al., 2011). PhyloCSF score measures the evolutionary rate of non-synonymous codon substitution over synonymous codon substitution in related species (in the case of homo sapiens – 29 mammals). A score below -10 (decibans scale) for a given ORF indicates lack of coding potential and a score above 10 indicates a coding transcripts. Figure 4.2.1-1A shows the cumulative
phyloCSF score for each ORF found in *EIN1*, *XIST* – a well-described long non-coding RNA and *GAPDH* – a highly expressed protein-coding gene. The highest phyloCSF score (10 decibans) was measured for a short -33 aa ORFs. The score for the longest ORF (located in the pseudogene sequence) was -190 decibans, indicating a strong non-coding potential. Similarly, the investigation of local scores calculated for each region of human genome shows strong non-coding potential of *EIN1* and *XIST* (Figure 4.2.1-1C and D) and strong coding potential only for exons of *GAPDH* (Figure 4.2.1-1B).

The final computation element of the *EIN1* non-coding potential investigation was a search for expressed peptides from global analysis of human tissues using mass spectroscopy (Kim et al., 2014). A search of all of the potential ORFs from the potential *EIN1* isoforms did not match any of the 293 700 peptides identified in the original study. Moreover, a search of all of the potential ORFs present in the unspliced *EIN1* yielded 2 six amino acid matches, neither of them was unique for *EIN1* and each of them also mapped to other highly expressed proteins of known function.

To summarise, the above results strongly suggests a non-coding status for *EIN1*, regardless of its splicing status. The presence of the pseudogene sequence resulted in the presence of a relatively long ORF in the sequence of *EIN1*. However, the RNA-seq data indicate that the pseudogene sequence is spliced out from the mature *EIN1* isoforms. Secondly, the sequence of all of the ORFs present in *EIN1* is poorly conserved with high level of non-synonymous substitutions (low phyloCSF scores). Finally, investigation of publicly available mass spectroscopy data did not yield in any significant hits. Nevertheless, it has been reported that lincRNA might harbour short polypeptides which are not conserved and hence are not detectable using established computational approaches (Ingolia et al., 2011; Ruiz-Orera et al., 2014). As a result, I also used biochemical methods to test the non-coding potential of *EIN1* by investigation of *EIN1* interaction with ribosomes.
Figure 4.2.1-1. Computational analysis of EIN1 coding potential

A) Cumulative phyloCSF score (Lin et al., 2011) of EIN1, GAPDH and XIST. B-D) Genome tracks (UCSC browser) of local phyloCSF score for B) GAPDH, C) XIST and D) EIN (gaps originate from poorly conserved regions) (Lin et al., 2011)
4.2.2. Biochemical investigation of the coding potential of \textit{EIN1}

Computational evidence indicates that \textit{EIN1} does not encode for a functional protein. However, in order to fully assess the translation potential of \textit{EIN1}, I investigated its interaction with ribosomes. This approach has been successfully used for investigation of the translation potential of lincRNAs (Guttman et al., 2013). Initially, I attempted to perform polysome extraction using sucrose density gradient (Halbeisen et al., 2009) using lysates extracted from MCF10A and MCF7 cell lines. However, I did not observe polysomal profiles from either of the cell lysates (HEK293T cells produced expected results). As a result, I performed fractionation of ribosomes with bound RNA from the remaining elements of cellular lysate (including chromatin and free mRNA) using sucrose cushion density centrifugation (Figure 4.2.2-1A) (Blobel and Sabatini, 1971). This technique relies on the fact that in the presence of cycloheximide (CHX, translation elongation inhibitor) the actively transcribed mRNA is ‘frozen’ with ribosomes and they can be separated from the soluble mRNA using 60% sucrose cushion. In the presence of high concentrations of magnesium ion chelating agent (EDTA) ribosomes dissociate and mRNA is released.

In the presence of CHX, around 20% of both 18S (main element of small eukaryotic ribosomal subunit) and 28S (main element of large eukaryotic ribosomal subunit) rRNA associated with the pellet fraction (Figure 4.2.2-1B). The level of ribosomal RNA in the pellet after EDTA treatment dropped to ~2%, indicating strong dissociation of ribosomes and release of actively translated mRNAs. Similarly, ~20% of \textit{GAPDH} mRNA was present in the pellet of CHX-treated samples and it decreased to ~2% in the EDTA-treated samples. This indicates that, \textit{GAPDH} mRNA actively interacts with ribosomes most likely in translation-dependant manner (Figure 4.2.2-1B). In the case of \textit{EIN1}, also around 20% of RNA was in the pellet of CHX-treated fraction (Figure 4.2.2-1B). However, after the EDTA-treatment the levels of \textit{EIN1} did not decrease in the pellet. This result is consistent with lack of interaction between ribosomes and \textit{EIN1}. The presence (and slight increases in the EDTA-treated samples) of \textit{EIN1} in the pellet of both EDTA- and CHX-treated samples might originate from its interaction with the chromatin or other high molecular weight cellular elements present in the samples (Méndez and Stillman, 2000).
Figure 4.2.2.1. Sucrose cushion density centrifugation analysis of EIN1 interaction with ribosomes

A) Schematics of sucrose differential centrifugation of MCF7 cells in the presence of the protein synthesis inhibitor (cycloheximide, 100 μg/ml, ‘freezes’ RNA on the ribosomes) or in the presence of 50 mM EDTA (dissociates ribosomes) B) Relative levels of RNA (RT-qPCR) in the ribosomal (pellet) and soluble (supernatant) fractions. Data represent average ±SEM of the proportion in fractions from two biological repeats.
4.2.3. The cellular localisation of EIN1

In order to understand a potential role of EIN1 in the cells, I investigated its cellular localisation. ENCODE RNA-seq data indicate that majority of lincRNAs are localised in the nucleus of cells (Djebali et al., 2012). Hence, I investigated the expression of EIN1 in MCF7 cells using RNA-seq performed on samples from nuclear and cytoplasmic fractions. Firstly, the profile of read distribution across the gene body of EIN1 is similar between the nuclear fraction of MCF10A and the ENCODE data from the nuclear fraction of MCF7 cells (Figure 4.2.3-1A). Secondly, the MCF7 data indicate that EIN1 transcripts tends to localise to the nucleus rather than cytoplasm (Figure 4.2.3-1A). The drawback of the ENCODE data is that the RNA samples originate from asynchronously grown cells; hence the expression level of EGF-inducible genes might not match the EGF-stimulated condition.

In order to further investigate the cellular localisation of EIN1, I performed biochemical fractionation of MCF10A cells. Cell lysates were collected 30 minutes after EGF stimulation. In order to confirm separation of the nuclear fraction from the cytoplasmic fraction, I investigated the levels of U4 – a small nuclear RNA required for splicing and 18S rRNA – a major component of the small ribosomal subunit normally found in the cytoplasm (Figure 4.2.3-1B). As expected, the majority of U4 was located in the nucleus (average cytoplasm to nucleus ratio – 0.06:1) and majority of 18S rRNA was in the cytoplasm (average cytoplasm to nucleus ratio – 8.24:1). These ratios confirm the efficiency of the fractions. In the case of the highly expressed protein-coding gene – GADPH – the majority of mRNA was localised in the cytoplasm (ratio of 12.38:1). In the case of EIN1, the majority of mRNA localised to the nucleus (ratio of 0.30:1, Figure 4.2.3-1B). This result is consistent with the ratio observed for the MCF7 cells observed for ENCODE RNA-seq data (Figure 4.2.3-1A) and confirms that the EIN1 transcript localises to the nucleus. Furthermore, the nuclear localisation can be used a proxy for EIN1 protein-coding potential. A strong localisation in the nucleus further supports the notion that EIN1 is not translated.
Figure 4.2.3.1. EIN1 preferentially localises to the nucleus in MCF10A and MCF7 cells lines

A) UCSC genome browser tracks of EIN1 expression in MCF10A (nuclear RNA-seq, this study) and MCF7 cells (cytoplasmic and nuclear fractions, ENCODE (Djebali et al., 2012)). B) RT-qPCR of nuclear and cytoplasmic fractions of MCF10A cells. MCF10A cells were starved of EGF for 48 hrs and subsequently stimulated with EGF for 30 minutes. Individual data points from 3 independent repeats were normalized to the nuclear RNA level (taken as 1) and are shown on both linear and log2 scale to show depletion of U4 and EIN1 from the nuclear fractions. Horizontal lines indicated mean value. * - p-value < 0.01, t-test with multiple testing correction.
Both, bioinformatics and biochemical studies of EIN1 coding-potential and localisation are consistent with a non-coding status of the gene. Bioinformatically, with low phyloCFS score, lack of long ORFs and lack of detectable peptides in mass spectroscopy data, EIN1 displays features of long non-coding RNA. Biochemically, EIN1 does not interact with ribosomes in a translation-dependant manner. The results of the fractionation indicated that EIN1 is preferentially associated with nuclear rather than cytoplasmic fraction. The nuclear localisation of EIN1 might also be an indicative of its function. Numerous nuclear lincRNAs have recently been identified and their role in regulation of gene expression has been suggested (Gutschner et al., 2013; Han et al., 2014; Vance and Ponting, 2014; Vance et al., 2014). Judging from the properties of IEN1, it can be suggested that it also plays a role in the regulation of gene expression. This hypothesis is explored in chapter 5.

4.3. Investigation EIN1 transcription regulation by ERK1/2 signalling pathway

4.3.1. Transcription of EIN1 in mammary gland cell lines and in cancer data

Over the course of this chapter, I compiled data from multiple cell lines (including MCF7 and MCF10A) and I assumed that EGF signal is transmitted to the nucleus via the ERK1/2 pathway. In order to validate whether the EGF-dependant inducible expression of EIN1 is independent of cell type, I stimulated a panel of mammary gland cell lines with EGF. MCF10A are normal, immortal mammary gland cells, MCF7 is non-metastatic breast cancer cell line and MDA-MB-231 is metastatic breast cancer cell line (see section 1.3.6 for further details). In both MCF10A and MCF7 cell lines the expression of EIN1 and FOS (a well described IE gene, see section 1.3.4) was dependant on EGF (Figure 4.3.1-1A and B). However, in MDA-MB-231 cells both of the genes did not respond to EGF treatment. Previous studies showed that in MDA-MB-231 cells the EGF receptor is highly overexpressed (Davidson et al., 1987; Fan et al., 1998) but the ERK1/2 pathway is not active (Biswas et al., 2000). As a result, the lack of EGF-inducible expression of FOS and EIN1 would be expected if their expression is dependent on ERK1/2 signalling.
Figure 4.3.1-1. EGF-dependant expression of *FOS* and *EIN1* in MCF10A, MCF7 and MDA-MB-231 cell lines and in primary cancer cells.

RT-qPCR of *FOS* (A) and *EIN1* (B) from MCF10A, MCF7 and MDA-MB-231 cell lines. MCF10A cells were starved of EGF for 48 hrs and subsequently stimulated with EGF for 30 minutes. Individual data points from 2 independent repeats were normalized to the 0 time point of MCF10A cells (taken as 1) and are shown on both linear and log2 scale. C) Boxplots of *EIN1* expression from breast cancer RNA-seq (Horvath et al., 2013) and D) lung cancer RNA-seq from patients biopsies (Seo et al., 2012).

* - p-value < 0.05, t-test with multiple testing correction.
On the other hand, the MCF10A (being normal cells) and MCF7 do not carry mutations or overexpression of any elements of the ERK1/2 pathway (Lacroix and Leclercq, 2004) and their responsiveness is as expected and of similar amplitude.

In order to further investigate the expression pattern of EIN1, I further analysed expression of EIN1 in breast and cancer samples (globally studied in section 3.2., Figure 4.3.1-1C and D). Among the breast cancer samples, EIN1 expression increased significantly in the triple negative breast cancer (progesterone receptor negative, estrogen receptor negative and with normal copy number of HER2, Figure 4.3.1-1C). EIN1 expression was also elevated in the panel of paired lung cancer samples (Figure 4.3.1-1D). The increase of EIN1 stands in contrast to vast majority of EGF-regulated protein-coding genes, as their expression decreased significantly in these breast and lung cancer samples (Figure 3.2.3-3).

This provides a confusing image of EIN1 expression. In the breast cancer cells models the expression is strongly dependant on EGF-stimulation, but only when ERK1/2 pathway is not changed from wild-type state. Similarly, in the cancer cells, the overexpression of EIN1 happens only in the triple negative breast cancer – cancer where most-likely normal ERK1/2 signalling is still active. However, in the case of FOS the expression decreases in cancer cells. The difference in expression pattern might originate from different response of FOS (and other IE genes) and EIN1 to the feedback loops activated by the ERK1/2 pathway.

4.3.2. EIN1 transcription is MEK1/2-dependant and independent of new protein synthesis

In the previous section I investigated the EGF-dependant expression of EIN1 in different cell lines and the data suggests a potential link between ERK1/2 pathway and EIN1 expression. In order to address this question, I investigated whether EIN1 transcription is directly regulated by ERK1/2 signalling pathway.

EGF, in addition to ERK1/2 pathway, might also activate other signalling pathways, including PI3/AKT pathway (see table 1.3.1-1) (Biswas et al., 2000). In order to test whether the EGF signal regulates the EIN1 transcription directly via the ERK1/2 signalling pathway, we investigated the effect of MEK1/2 inhibition on EIN1 transcription. As detailed in section 1.3.2, MEK1/2 is a selective activator of ERK1/2; hence, we used the U0126 inhibitor which selectively and potently blocks
activation of MEK1 and MEK2 (Figure 4.3.2-1A) (Davies et al., 2000; Favata et al., 1998). As expected, in the presence drug vesicle (DMSO) the expression was still inducible (Figure 4.3.2-1B). In the presence of U0126, EGF was not able to induce the expression of EIN1. This result is consisted with direct signal transduction via the ERK1/2 pathway.

In order to further investigate the signalling mechanism, I tested whether the EGF-dependant induction of EIN1 is independent of protein synthesis. MAPK signalling pathways transmit signal from the surface of the plasma membrane to the nucleus via phosphorylation of subsequent elements of the pathway. Hence, the signalling and IE gene transcription can be initiated without protein synthesis (see section 1.3.4 and Figure 4.3.2-2A). Indeed, in the presence of a protein-synthesis inhibiting concentration of CHX, I still observed EGF-dependant phosphorylation of ERK1/2 (Figure 4.3.2-2B). Secondly, since the stimulation of MCF10A cells with EGF triggers transcription of FOS, I also observed an increase of in the protein level (Figure 4.3.2-2B – FOS level in the CHX negative samples). After CHX treatment the levels of FOS protein decreased, which indicates that CHX was indeed blocking protein synthesis (Figure 4.3.2-2B – FOS level in the CHX positive samples). At the same time, the level of FOS mRNA increased (Figure 4.3.2-2C) further confirming the protein synthesis-independent ERK1/2 response. Comparison of FOS mRNA profiles (Figure 4.3.2-2C) showed that in normal physiological conditions the transcription was transient as previously observed in longer time course (Figure 3.1.3-4A). On the other hand, when protein synthesis was inhibited, the level of FOS mRNA increased for the whole duration of induction and I did not observe transcription inhibition (Figure 4.3.2-2C).
Figure 4.3.2-1. The expression of EIN1 in the presence of MEK inhibitor
A) Simplified ERK signalling pathway together with the point of inhibitory action of U0126. B) effect of MEK inhibition on the EGF-dependant expression of EIN1, n=2, * - p-value < 0.05. U0126 was used at a final concentration of 10 µM in 0.1% DMSO. RT-qPCR was performed on the nuclear fraction by Megan Moruzzi.
The expression of EIN1 is independent of protein synthesis

A) Schematic of the expression of immediate-early (IE) genes after stimulation with epidermal growth factor (EGF). The expression of IE genes occurs transiently after EGF stimulation and is independent of protein synthesis. B) Western blot analysis of the effect of cycloheximide (CHX, 25 ug/ml) on EGF-dependant protein synthesis. pERK - phosphoERK C) RT-qPCR analysis of the effect of CHX on RNA levels of FOS and EIN1 after EGF stimulation for indicated times. Data points represent mean ± SEM from two independent repeats.
This result is consistent with transcriptional inhibition triggered by transcription of IE genes – i.e. in the absence of protein synthesis; the inhibitory action of IE genes cannot be exercised (see section 1.3.4). On the other hand, previous analysis showed much longer induction of EINI expression (Figure 3.2.2-1 – XLOC_14669) which was sustained for around 180 minutes. Similarly, in the current experiment, in the absence of CHX, the transcription of EINI was still high 120 minutes after EGF-stimulation (Figure 4.3.2-2D). However, in the presence of the translation inhibitor, the transcription of EINI returned to near basal level within 120 minutes (Figure 4.3.2-2D). This result suggests, that in contrast to FOS transcription where signalling is terminated by translation if IE genes, EINI transcription requires a new regulatory stimulus to originate from the translation of IE genes.

The results presented in this section provide evidence that EINI transcription is directly regulated by the ERK1/2 MAPK signalling pathway. In the presence of a specific pathway inhibitor EGF-dependant transcription ceases. Secondly, data presented here showed that EINI transcription is independent of protein synthesis and it appears that some of the IE genes protein products are required for a sustained EINI expression.

4.4. Chapter summary

The aim of the work presented in this chapter was to characterise a long noncoding RNA whose transcription is dependent on EGF signalling. In the first section (4.1) I investigated the structure and genomic properties of EINI. EINI appears to be a poorly conserved lincRNA located in an intergenic region of chromosome 14. Its genomic locus contains features often associated with actively transcribed genes (histone H3K4 trimethylation and paused RNApolII). It has a well-defined 5’-end and its promoter contains key elements (TATA-like and Inr) associated with RNApolIII-dependant transcription. The RNA-seq data indicate that it can be spliced; however, the determination of the structure of different isoforms has been a significant challenge and, at the time of writing, it is an on-going project. Finally, using the human tissue expression data, I observed that EINI (or specifically its proxy gene – RP11-7F17.7) is expressed in a limited number of tissues and its transcription is associated with multiple eQLTs. However, as tempting as it might be,
the expression of *EIN1* cannot be tested in developmental context due lack of conservation across species.

Both bioinformatics and biochemical analysis strongly suggest that *EIN1* is a non-coding gene. Bioinformatically, it has a low phyloCSF score, and it has only short ORFs. Biochemically, it mainly localises in the nucleus and the data suggest that it does not bind ribosomes in a transcription-dependant manner. Finally, investigation of data from the global mass spectroscopy studies did not detect unique peptides sequences matching *EIN1* ORFs. It should however be noted that recent studies indicated that some of the lncRNAs might actually code for short, regulatory peptides (Anderson et al., 2015; Bazzini et al., 2014). An ultimate test of *EIN1* non-coding status would be an *in vitro* or *in vivo* transcription experiments, both successfully used in the past to show lack of coding-potential (Han et al., 2014; Maamar et al., 2013). However, these reactions require cloned versions of the gene. Hence the identification of the relevant splice variants is an important step in the characterisation of *EIN1*.

Furthermore, the data suggest that *EIN1* transcription (at least the initial induction) is regulated directly via the ERK1/2 pathway. The ERK dependant signalling is supported by the observation that inhibition of MEK1/2 kinase stops EGF-dependant *EIN1* transcription. Furthermore, the initiation of transcription of *EIN1* is independent of new protein synthesis – a characteristic feature of IE genes (O’Donnell et al., 2012). However, the profile of protein synthesis-independent transcription of *EIN1* is unusual. New protein synthesis is essential for the sustained induction of *EIN1* transcription (up to 240 minutes). The inhibition of protein synthesis stops the transcription within the first 120 minutes after EGF-stimulation. It is tempting to suggest that that although *EIN1* is an IE gene (transcribed immediate after EGF stimulation and in absence of protein synthesis), it is also a DE gene (sustained translation of IE genes is essential for *EIN1* sustained transcription). As a result, *EIN1* might be an interesting example of gene that is required at both stages of the transcriptional cascade downstream of EGF signalling.

Finally, *EIN1* has another unusual feature. Data presented in chapter 3 indicate that the majority of EGF-regulated IE genes follow the same pattern of expression – they are either up- or downregulated. However, *EIN1* was one of few genes whose levels were different to the consensus established by other IE genes. In the cancer data
tested th majority of EGF-regulated IE genes was downregulated, yet *EIN1* was upregulated.

The unusual behaviour of *EIN1*, both downstream of EGF stimulation in the cell lines and the patterns of its expression in the primary cancer samples suggests that *EIN1* can play an important role in the cellular processes normally disrupted in cancer, a function that might be different to the protein-coding IE genes. Hence, in the following chapter I will explore the function of *EIN1*. 
5. The effect of *EIN1* depletion on the EGF-inducible gene expression profile

5.1. Establishment of CRISPRi expression regulation system

Over the years a large number of techniques have been established in order to study gene function using reverse genetics approaches. In the case of protein coding genes, one of the most widely used techniques is RNAi. The question however is, whether RNAi approaches can be used to study lincRNAs. In case of *EIN1* the task has two obstacles, for knockdown to be effective, the technique has to inhibit the gene in the nucleus (*EIN1* is mainly nuclear) and the knockdown has to take place during or immediate after transcription (EGF-dependant transcription is rapid). In the past, an RNAi approach has been used to knockdown protein-coding immediate-early gene and the outcome downregulated the level of the target protein (Cheng et al., 2013). However, to the best of my knowledge, rapid degradation of mRNA has not been undertaken using RNAi, despite potential co-transcriptional activity of siRNA (Castel and Martienssen, 2013). In the case of lincRNAs, a number of studies have used RNAi in the past for knockdowns with varying success rates. For example, the knockdown of nuclear linc-HOXA1 was around 50% efficient using RNAi (Maamar et al., 2013) and the knockdown of cytoplasmic TINCR was 60% efficient (Kretz et al., 2013). However, even though inefficient, both knockdowns resulted in quantifiable phenotypes.

Recently, an RNA interference approaches based on the clustered, regularly interspaced, short palindromic repeat (CRISPR, (Jinek et al., 2012; Mali et al., 2013)) system has been developed. In contrast to the RNAi methods which degrade the mature form of the transcript, the CRISPR interference (CRISPRi) approaches developed by Weissman (Gilbert et al., 2013) and Zhang (Sanjana et al., 2014) laboratories control the expression of genes co-transcriptionally. If effective, the co-transcriptional regulation of expression gives an unparalleled opportunity to regulate the expression of genes whose transcription is tightly regulated. In the case of the lincRNAs, it also overcomes the limitation of RNAi methods associated with time
delay between transcription and knockdown of mature RNA. As a result of these advantages, I decided to test whether the CRISPRi approach can be used for the downregulation of *EIN1* expression.

5.1.1. The outline CRISPRi approach

In order to downregulate the transcription of *EIN1*, I combined the system presented by Gilbert et al. (Gilbert et al., 2013) with guide RNA delivery presented by Sanjana et al. (Sanjana et al., 2014). The CRISPRi system relies on two recent discoveries. The first one was the adaptation of the bacterial immune system (Cas9 protein from *Streptococcus pyogenes*) for the purpose of RNA-guided genome editing (Jinek et al., 2012; Mali et al., 2013) and the second is observation that two mutations in the catalytic domain (D10A and H840A) render Cas9 nuclease catalytically inactive (dead Cas9 - dCas9 (Qi et al., 2013)). Finally, by fusing dCas9 with Krüppel associated box (KRAB) from *ZNF10* (encoding zinc finger protein 10 (Vissing et al., 1995)), a strongly repressive, RNA-guided system was created (Figure 5.1.1-A) (Gilbert et al., 2013). In order to increase the efficiency of gene delivery I used second generation lentiviral delivery system (see materials and methods and (Tiscornia et al., 2006; Zufferey et al., 1998)). The knockdown protocol followed a two-step approach (Figure 5.1.1-1A). In the first transduction, I transduced normal MCF10A cells with dCas9-KRAB containing viruses and, using flow cytometry, I selected a population expressing dCas9-KRAB (blue fluorescent protein – BFP – was selection marker). The daughter cell line – termed MCF10A-KRAB – originated from a mix population of dCas9-KRAB positive cells. The subsequent round of transductions delivered guide RNAs into MCF10A-KRAB cells.
Figure 5.1.1-1. CRISPRi-driven regulation of gene expression

A) Schematics of the mode of action of CRISPRi. Cells express dCas9-KRAB protein together with the guide RNA designed against the promoter of the gene of interest. When expressed the dCas9-KRAB protein is targeted to the promoter and represses gene expression. B) Two-step lentiviral delivery of dCas9-KRAB and target specific guide RNA. LN₂ – liquid nitrogen.
5.1.2. Validation of the efficiency of EIN1 knockdown

The first step during the investigation of EIN1 function was validation of the knockdown performed using CRISPRi system. Initially, I designed six guide RNAs (gRNAs) targeting a 200 bp regions near the TSS of EIN1. Transduction of MCF10A-KRAB cells with each of the gRNAs resulted in over 90% decreases in the basal level of EIN1 transcript in all cases (Figure 5.1.2-1A).

The next key experiment was investigation whether the CRISPRi system is able to stop the inducible expression of EIN1. The transduction of MCF10A-KRAB cells with three different gRNAs against EIN1 followed by stimulation of cells with EGF stopped the EGF-inducible expression of EIN1 (Figure 5.1.2-1B). On average, I observed a 75% decrease in the level of EIN1 at 0 minutes time point and it did not change after EGF-stimulation. On the other hand, the expression of EIN1 in the control samples was similar to the profiles of expression I observed previously in the parental MCF10A cells (Figure 3.2.2-1). Finally, investigation of the mRNA of FOS showed that the knockdown of EIN1 had no effect on its expression (Figure 5.1.2-1B).

Taken together, these results show that CRISPRi-driven downregulation of gene expression is effective. The system is able to repress the EGF-inducible profile of EIN1 expression. In the absence of EGF signalling, the normally low expression of EIN1 decreased further and remained at a stable low level after EGF stimulation. The control condition where cells infected with lentiviral particle had no effect on the expression profile of protein-coding EGF-induced IE genes. This indicates that the system can be used for phenotypic studies.
Figure 5.1.2.1. Validation of CRISPRi-driven knockdown of EIN1

A) EIN1 knockdown. MCF10A cells stably transduced with dCas9-KRAB were transduced with guides targeting indicated regions near the promoter of EIN1 or control guide RNA. The EIN1 level was measured by RT-qPCR from continuously grown cells three days after transduction. B) EIN1 knockdown after EGF induction. MCF10A-KRAB cells were transduced with either of -137, -54 and -23 anti-EIN1 guide RNAs or with non-targeting control guide RNAs. Cells were grown in the presence of puromycin for ten days to select for MCF10A-KRAB stably expressing guide of interest. Subsequently, cells were starved of EGF for two days and finally, stimulated with EGF for indicated times. The RNA levels were normalised to the average expression at 0 minutes in the non-targeting sample using delta delta Ct method.
5.2. Investigation of EGF-regulated expression profile after *EIN1* downregulation

Results in the previous section showed that the expression of lincRNA, such as *EIN1* can be tightly downregulated using CRISPRi-driven system. The combination of *EIN1* nuclear localisation (section 4.2.3) and expression profile after EGF stimulation (section 4.3.2) indicates a possible role of *EIN1* in the regulation of gene expression downstream of EGF signalling pathway. In order to test this hypothesis, I performed an RNA-seq experiment where I investigated the effect of *EIN1* downregulation on the expression of protein-coding genes after EGF stimulation.

5.2.1. Validation of RNA-seq

Previous microarray-based investigations of gene expression after EGF stimulation showed that in the HeLa cells the expression of genes follows wave-like pattern of gene expression (Amit et al., 2007). Different classes of EGF-regulated genes have maximum expression at different times (see Figure 1.3.4-1 for an outline). The first wave of transcription (IE genes) takes place within the first 30 minutes after EGF stimulation and the subsequent wave follow with 30–60 minute long gaps. In order to observe the main classes of EGF-regulated genes, I decided to measure gene expression at four time points following EGF stimulation (0, 30, 90 and 180 minutes). Each time point should correspond to a different wave of expression. The CRISPRi system is a recent development and its accuracy has not been thoroughly tested; hence, in order to control for potential off-target effects, I used three gRNAs (-137, -54, -23, Figure 5.1.2-1) to deplete *EIN1* expression. Furthermore, as a control, I used two non-targeting gRNAs that do not target human genome. In order to have three control samples, I collected samples for one of the non-targeting gRNAs twice. Taken together, we sequence 24 RNA-seq libraries using polyA-tailed fraction of total RNA extracted from MCF10A cells.

On average each library was sequenced to over 23 million of paired-end reads (Table 5.2.1-1). Similarly to the nuclear RNA-seq analysis (Chapter 3), all of the reads with low quality and/or containing adapter sequence were removed and final, processed libraries had on average 22,240,039 paired-end reads (Table 5.2.1-1). On average, I observed over 97% mapping efficiency and over 75% of all of the paired-end reads
were non-redundant (Table 5.2.1-1). Investigation of the distribution of reads in the gene body showed that the read density is slightly skewed toward the 5’-end of the genes (Figure 5.2.1-1A). However, since all of the samples follow very similar profiles, the skewness should not have effect on the quantification of expression. Finally, I observed small value of coefficiency of variation for expression with \( \log_{10}(\text{FPKM}) \) above 0 (FPKM=1) (Figure 5.2.1-1B). However, the coefficiency of variation sharply increases below \( \log_{10}(\text{FPKM}) \) equal to 0 (FPKM=1) indicating that the differential analysis of genes with FPKM < 10\(^0\)=1 might not be possible (most-likely only large changes will be statistically significant). However, I expect that the expression of most of the protein-coding genes that are EGF-regulated will be above FPKM=1.
Table 5.2.1-1: Summary of the mapping statistics for RNA-seq after *EIN1* downregulation

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Condition</th>
<th>Time Point (minutes)</th>
<th>Pairs of reads before trimming</th>
<th>Pairs of reads after trimming</th>
<th>Mapping rate</th>
<th>Non-redundant mapped pairs</th>
<th>Redundancy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NT1</td>
<td>0</td>
<td>24,002,711</td>
<td>22,251,098</td>
<td>96.03%</td>
<td>15,427,445</td>
<td>30.67%</td>
</tr>
<tr>
<td>2</td>
<td>NT1</td>
<td>30</td>
<td>19,315,569</td>
<td>17,598,018</td>
<td>96.73%</td>
<td>13,043,828</td>
<td>25.88%</td>
</tr>
<tr>
<td>3</td>
<td>NT1</td>
<td>90</td>
<td>36,883,770</td>
<td>33,842,568</td>
<td>96.88%</td>
<td>22,726,814</td>
<td>32.85%</td>
</tr>
<tr>
<td>4</td>
<td>NT1</td>
<td>180</td>
<td>31,149,650</td>
<td>28,317,344</td>
<td>97.08%</td>
<td>19,216,808</td>
<td>32.14%</td>
</tr>
<tr>
<td>5</td>
<td>gRNA -054</td>
<td>0</td>
<td>10,671,199</td>
<td>9,921,256</td>
<td>97.14%</td>
<td>7,438,794</td>
<td>25.02%</td>
</tr>
<tr>
<td>6</td>
<td>gRNA -054</td>
<td>30</td>
<td>29,551,701</td>
<td>27,127,402</td>
<td>97.22%</td>
<td>18,544,032</td>
<td>31.64%</td>
</tr>
<tr>
<td>7</td>
<td>gRNA -054</td>
<td>90</td>
<td>28,129,460</td>
<td>26,046,292</td>
<td>97.30%</td>
<td>19,055,162</td>
<td>26.84%</td>
</tr>
<tr>
<td>8</td>
<td>gRNA -054</td>
<td>180</td>
<td>19,591,992</td>
<td>18,173,106</td>
<td>97.43%</td>
<td>13,732,833</td>
<td>24.43%</td>
</tr>
<tr>
<td>9</td>
<td>NT2</td>
<td>0</td>
<td>21,537,787</td>
<td>21,192,943</td>
<td>97.48%</td>
<td>15,444,658</td>
<td>27.12%</td>
</tr>
<tr>
<td>10</td>
<td>NT2</td>
<td>30</td>
<td>19,528,577</td>
<td>19,176,567</td>
<td>97.17%</td>
<td>14,844,711</td>
<td>22.59%</td>
</tr>
<tr>
<td>11</td>
<td>NT2</td>
<td>90</td>
<td>21,252,453</td>
<td>20,887,258</td>
<td>97.57%</td>
<td>15,327,551</td>
<td>26.62%</td>
</tr>
<tr>
<td>12</td>
<td>NT2</td>
<td>180</td>
<td>23,876,266</td>
<td>23,452,137</td>
<td>97.38%</td>
<td>17,610,269</td>
<td>24.91%</td>
</tr>
<tr>
<td>13</td>
<td>gRNA -137</td>
<td>0</td>
<td>19,975,327</td>
<td>19,633,725</td>
<td>96.71%</td>
<td>14,850,569</td>
<td>24.36%</td>
</tr>
<tr>
<td>14</td>
<td>gRNA -137</td>
<td>30</td>
<td>21,307,570</td>
<td>20,923,768</td>
<td>96.62%</td>
<td>16,251,519</td>
<td>22.33%</td>
</tr>
<tr>
<td>15</td>
<td>gRNA -137</td>
<td>90</td>
<td>26,773,670</td>
<td>26,286,830</td>
<td>96.89%</td>
<td>20,209,258</td>
<td>23.12%</td>
</tr>
<tr>
<td>16</td>
<td>gRNA -137</td>
<td>180</td>
<td>21,536,959</td>
<td>21,116,999</td>
<td>96.99%</td>
<td>16,915,752</td>
<td>19.90%</td>
</tr>
<tr>
<td>17</td>
<td>NT1_2</td>
<td>0</td>
<td>22,075,176</td>
<td>21,730,260</td>
<td>97.57%</td>
<td>17,080,911</td>
<td>21.40%</td>
</tr>
<tr>
<td>18</td>
<td>NT1_2</td>
<td>30</td>
<td>17,545,741</td>
<td>17,260,367</td>
<td>97.58%</td>
<td>13,784,882</td>
<td>20.14%</td>
</tr>
<tr>
<td>19</td>
<td>NT1_2</td>
<td>90</td>
<td>16,801,443</td>
<td>16,526,953</td>
<td>97.77%</td>
<td>13,723,738</td>
<td>16.96%</td>
</tr>
<tr>
<td>20</td>
<td>NT1_2</td>
<td>180</td>
<td>22,550,392</td>
<td>22,216,559</td>
<td>97.81%</td>
<td>17,609,608</td>
<td>20.74%</td>
</tr>
<tr>
<td>21</td>
<td>gRNA -023</td>
<td>0</td>
<td>20,006,368</td>
<td>19,687,132</td>
<td>96.99%</td>
<td>15,508,452</td>
<td>21.23%</td>
</tr>
<tr>
<td>22</td>
<td>gRNA -023</td>
<td>30</td>
<td>20,686,039</td>
<td>20,348,263</td>
<td>97.13%</td>
<td>16,381,934</td>
<td>19.49%</td>
</tr>
<tr>
<td>23</td>
<td>gRNA -023</td>
<td>90</td>
<td>21,684,180</td>
<td>21,309,709</td>
<td>97.30%</td>
<td>16,874,550</td>
<td>20.81%</td>
</tr>
<tr>
<td>24</td>
<td>gRNA -023</td>
<td>180</td>
<td>39,396,895</td>
<td>38,734,385</td>
<td>97.39%</td>
<td>28,499,040</td>
<td>26.42%</td>
</tr>
</tbody>
</table>

Total: 555,830,895 pairs mapped, 533,760,939 non-redundant pairs mapped, 97.17% of reads mapped, 24.48% redundancy rate.

Notes: Column names and properties are the same as Table 3.1.2-1 with the exception of condition column. NT1 – non-targeting gRNA repeat 1, NT1_2 – non-targeting gRNA repeat 2, NT2 – non-targeting gRNA, gRNA -023, gRNA -054 gRNA -137 – gRNAs against *EIN1* regions -23, -54 and -137 respectively (see Figure 5.1.1-2).
Figure 5.2.1-1. Quality control of RNA-seq

A) Read coverage in the gene body of 1000 randomly selected genes. Sample numbers correspond to samples from Table 5.2.1-1. B) Coefficiency of variation (CV\(^2\)) for each test condition based on the average of three gRNA per condition per time point.
The final element of the validation of the RNA-seq setup was confirmation that the RNA-seq data also detected a significant decrease in the expression of \textit{EIN1} after knockdown. Judging from the RNA-seq data, CRISPRi-driven downregulation of \textit{EIN1} resulted in almost complete removal of the transcript (Figure 5.2.1-2A). However, the RT-qPCR results showed a 75\% decrease in the level of \textit{EIN1} transcripts (Figure 5.1.2-1B). This indicates the level of \textit{EIN1} transcript might have been below the detection threshold for RNA-seq with the depth of sequencing I used. Furthermore, the quantification after knockdown supports notion that the transcription of \textit{EIN1} is driven from the putative promoter described in section 4.1.2. If an additional downstream promoter existed for \textit{EIN1}, I would expect partial transcription downstream of this additional promoter.

Finally, the level of \textit{EIN1} increased in the sample treated with non-targeting gRNAs in EGF-dependant manner (Figure 5.2.1-2B and C). The expression profile measured by RNA-seq differed from the RT-qPCR profile (compare Figure 5.2.1-2C and Figure 5.1.2-1B). The RT-qPCR showed high transcription which plateaus throughout the time course, whereas the level of \textit{EIN1} increased continuously in the RNA-seq data. Potentially, this can be explained by the fact that the RT-qPCR primers are located near the 5’-end of the transcripts, whereas RNA-seq data measures \textit{EIN1} level using its whole length. As a results, the RNA-seq based quantification can be affected by the delayed induction in the 3’ section of the transcript (Figure 5.2.1-2A).

Taken together, the above results show that the CRISPRi-driven knockdown of \textit{EIN1} was successful and the data can be used for differential analysis of gene expression.
Figure 5.2.1-2. Validation of EIN1 downregulation

A) Read coverage in the gene body of EIN1. The average coverage from three gRNA per condition, per time point was normalized to 10^10 sequenced nucleotides and displayed using UCSC genome browser. Only data for the Watson strand are presented. B) Quantification of EIN1 transcript level. Data are presented for each gRNA per condition. C) Quantification of EIN1 transcript level – averaged. Each point represents an average of FPKM value calculated by Cufflinks package with SEM from three gRNA per time point per condition.
5.2.2. Expression of protein-coding genes after EGF-stimulation

A time course RNA-seq experiment where one compares two conditions at multiple time-points is an example of multifactorial RNA-seq. Despite the fact that a number of standard RNA-seq tools can perform limited factorial studies (e.g. edgeR (Robinson et al., 2009)) or time series testing without conditions (Trapnell et al., 2012), to the best of my knowledge, a comprehensive approach to the time course analysis has not been presented (Spies and Ciaudo, 2015). Recently, an approach was proposed (DyNB, (Aijö et al., 2014)), however, it did not show significant improvement over differential analysis provided by edgeR or Cufflinks and has not been widely implemented. As a result, in order to maintain consistency between current time course analysis and the nuclear RNA-seq presented in chapter 3, I performed analysis using the Cufflinks package (Trapnell et al., 2012).

In order to maintain a logical flow to the analysis, firstly, I analysed the changes in the expression of protein-coding genes in the non-targeting control samples. This provided a reference, EGF-regulated expression profile (data are presented in this section). Subsequently, the established reference EGF-regulated profile of gene expression can be compared with the pattern of expression following depletion of EIN1 (presented in section 5.2.3). Due to the high number of pairwise comparisons, I only present a selection of figures relevant for the interpretation of data, however, more comprehensive data comparisons are provided as supplementary data.

Bioinformatically, the analysis of the RNA-seq data followed the workflow established for the analysis of the nuclear RNA-seq (Figure 3.1.2-2). Since multiple conditions were present in the experiment, the analysis included pairwise comparisons of every pair of experiments. For the EGF regulation, only comparisons between 0 minutes sample and other time points are presented. A gene was significantly affected if the p-value is above 0.05 and fold change was above 1.5 (0.58 on log2 scale) or below 0.66 (-0.58 on log2 scale).

Stimulation of cells with EGF triggers a wave-like pattern of gene expression (reviewed in section 1.3.4). The RNA-seq data from the samples treated with the non-targeting gRNAs showed a similar wave-like pattern of expression (Figure 5.2.2-1 and Supplementary File 6 – includes pairwise statistics for the differential expression and Supplementary Files 7 and 8– contain individual expression graphs.
for each gene). In the case of upregulated genes, majority of the genes showed maximum expression only at a single time point. There were 79 genes that were upregulated at 30 minutes, 328 at 90 minutes and 596 at 180 minutes for a total of 731 genes upregulated in at least one time point (Figure 5.2.2-1A and Supplementary File 6). Based on the time point of maximum induction genes can be classified into three waves of expression (wave 1 – maximum at 30 minutes, wave 2 – at 90 minutes and wave 3- at 180 minutes, Figure 5.2.2-1A) Similarly to the nuclear RNA-seq (chapter 3), list of the most enriched genes included well-known IE genes (EGR1, EGR2, EGR3, FOS JUN and CTGF). The maximum expression of EGR1, EGR2, JUN and FOS corresponded to wave 1 and CTGF and EGF3 were in wave 2. In the case of the downregulated, there were 8 downregulated genes at 30 minutes, 76 at 90 minutes and 351 at 180 minutes for a total of 372 genes (Figure 5.2.2-1B and Supplementary Files 6–8 ). Similarly to the upregulated genes, the downregulated genes also followed 3 waves of expression.

In the nuclear RNA-seq, I identified 144 upregulated protein-coding genes. The current analysis of total RNA yielded 79 genes upregulated at 30 minutes. 59 genes were detected in both experimental approaches. However, comparison between genes that were upregulated at either 30 or 90 minutes time-point (346 genes) showed much higher overlap (91 genes were detected in both experimental setups). Finally, inclusion of genes upregulated at 180 minutes increased the overlap to 98 genes.

The discrepancy between the experimental setup can originate from possible higher sensitivity of the nuclear RNA-seq. In the case of protein-coding genes the majority of mRNA resides in the cytoplasm, as a result the nuclear RNA-seq most likely measured the changes in the level of newly synthesised transcripts and the steady-state levels of mRNA present in the cells did not influence the quantification. This is supported by the average fold change in the nuclear RNA-seq – the expression of genes detected in both experiments on average increase 3.17-fold and the genes unique for the nuclear RNA-seq increased 2.13-fold. As a result, for total RNA-seq to detected significant change in the level of transcript, a higher level of newly synthesised mRNA must accumulate in the cell. The nuclear RNA-seq does not have that problem as it predominantly detects novel transcripts.
Figure 5.2.2-1. EGF-regulated expression of genes

Heatmaps of the expression of protein-coding genes which are upregulated (A) or downregulated (B) after EGF treatment in at least one time point in comparison to the 0 minutes time point. Data are z-score normalised across row and sorted by the timepoint of the maximum (A) or minimum (B) expression. Only changes with p-value < 0.05 and fold change above 1.5 or below 0.66 are shown. On the right, the expression profiles of average ±95% CI normalised expression are depicted. For each genes, the expression was normalised to 0 minutes time point.
Taken together, these results show that the MCF10A-KRAB cells maintained its EGF-regulated expression programme. Nevertheless, it should be noted that the MCF10A and MCF10-KRAB cell lines were not directly compared and furthermore, the nuclear RNA-seq was used to study MCF10A cells and total cell RNA-seq was performed for MCF10A-KRAB cells. As a result, the differences in the experimental procedures might have influenced the results. However, the changes are not biologically significant, as the MCF10A-KRAB cells treated with the non-targeting gRNAs produce an EGF-induced expression programme that is similar to previous studies (Amit et al., 2007; Tullai et al., 2007) and the key genes such as FOS, JUN, member of EGR and DUSP families are expressed at correct times in the programme. As a result, this programme can be used a benchmark to study the effect of EIN1 downregulation on EGF signalling.

5.2.3. Effect of EIN1 downregulation on the expression of EGF-regulated genes

The ultimate aim of the experiment was identification of deviations from normal EGF-induced expression pattern observed after EIN1 depletion. Firstly, the investigation of the differences in the level of protein-coding genes between samples treated with the non-targeting and anti-EIN1 gRNAs showed that the changes are not as substantial as the changes induced by EGF on the starved cells (Figure 5.2.3-1). The majority of changes were within 2-fold induction or reduction in the transcripts levels. However, despite the fact that I used 3 different guide RNAs to deplete EIN1, the changes were highly reproducible. This indicates that the CRISPRi system is specific against EIN1.

The analysis showed that there were 206 genes affected at 0 minutes time point (79 up- and 127 downregulated, with p-value < 0.05 and log₂(fold_change)>0.48 or log₂(fold_change)<0.48), 204 at 30 minutes (91 up- and 113 downregulated), 262 at 90 minutes (98 up- and 164 downregulated) and 341 at 180 minutes (146 up- and 195 downregulated) (Figure 5.2.3-1).
Figure 5.2.3-1. Scatter plots of the effect of EIN1 downregulation on the expression of EGF-regulated genes

Scatter plots represent the average expression of the protein-coding genes in the presence of non-targeting or anti-EIN1 gRNAs. The data for EIN1 expression were added to the list of protein-coding genes. Data are presented for each time point after EGF stimulation. Red dots – genes with $\log_2$(fold_change)>0.48 or $\log_2$(fold_change)<-0.48 and p-value below 0.05 are significant. On each panel a number of significant genes is shown.
After identification of the differentially expressed genes, I investigated whether they play role in the EGF-regulated signalling. The EIN1 depletion triggered both upregulation and downregulation in the profile of gene expression. Similarly, EGF-regulated expression programme leads to induction and reduction in transcripts levels (with the exception of 30 minutes time point where majority of genes are upregulated, section 5.2.2). As a result, I investigated the effect of EIN1 downregulation on the panel of all of the EGF-regulated genes at the specific time point.

Firstly, for each time point after EGF stimulation, I performed comparison with the effect of EIN1 depletion at each time point (e.g. the list of genes affected by EGF-stimulation at 30 minutes was compared with the lists of genes affected by EIN1 depletion at either 0, 30, 90 and 180 minutes). The comparisons showed that the biggest overlap was normally observed when the lists originated from the same time point (e.g. the overlap of EGF-regulated genes at 90 minutes with the effect of EIN1 knockdown at 90 minutes was the higher than either of the overlaps with the lists of genes from 0, 30 and 180 minutes after EIN1 depletion, see Supplementary File 9). At each timepoint, I observed that around 8–9% of the genes that were EGF-regulated also were EIN1-regulated at the same time point (Figure 5.2.3-2). Simultaneously, the proportion of EGF-regulated genes in the set of EIN1-regulated genes increased as the time course progressed (from 4% at 30 minutes to 24% for 180 minutes, Figure 5.2.3-2). Finally, the enrichment in the overlap between the lists of EGF and EIN1-regulated at the specified times was statistically significant when compared with random chance (see statistics in Figure 5.2.3-2). Taken together, these results indicated that there is a possibility that EIN1 plays a role in the regulation of the expression of EGF-regulated transcription programme. However, that role will be limited to a small number of EGF-regulated genes (under 10%).
Figure 5.2.3-2. Effect of EIN1 downregulation on the expression of EGF-regulated genes
Scatterplots of gene expression after stimulation of MCF10A cells with EGF. Comparisons between 0 minutes (non-stimulated cells) and 30 minutes (A), 90 minutes (B) or 180 minutes (C) are drawn. Genes affected by EIN1 downregulation (log₂(fold_change)>0.48 or log₂(fold_change)<-0.48 and p-value below 0.05) are marked with red dots. For each comparison data are summarised with Venn diagrams and hypergeometric p-value for the overlap being equal or larger than observed is calculated in R. An EGF-regulated gene is defined as having log₂(fold_change)>0.58 or log₂(fold_change)<-0.58 and p-value below 0.05, EIN1-regulated genes is defined as having log₂(fold_change)>0.48 or log₂(fold_change)<-0.48 and p-value below 0.05.
In order to further research that hypothesis, I investigated the expression profiles of individual genes. Due to the biggest overlap between EGF- and EIN1-regulated genes at the 180 minutes time point, I focused on these genes. Figure 5.2.3-3 shows the pattern of expression of genes that were affected by EIN1 depletion 180 minutes after EGF stimulation. Genes that were both EGF- and EIN1-regulated split into four clusters depending on the direction of EGF and EIN1 effects (Figure 5.2.3-3, clusters 1-4). Normalisation of the expression of each of the genes within the clusters showed that on average, the expression profiles of EGF-downregulated genes after EIN1 depletion are shifted and they consistently either increase or decrease throughout the EGF-controlled time course (Figure 5.2.3-4A and B). The EGF-upregulated genes follow, similar pattern, albeit opposite in direction (Figure 5.2.3-4C and D). The average expression profiles indicate that EIN1 depletion affected the magnitude of the response to EGF-stimulus (of some of the normally EGF-regulated genes) but it not change the kinetics of the response.

However, the pattern of expression of EGF-independent genes drew an opposite picture. As expected, there was little change in the expression profile of EGF-independent genes after EGF stimulation in the control samples (non-targeting genes in clusters 5 and 6 on figures 5.2.3-3 and 5.2.3-4). However, the depletion of EIN1 triggered changes in the time-dependant profiles of expression. Hierarchical clustering split the gene lists into downregulated (cluster 5 on Figures 5.2.3-3 and 5.2.3-4) and up-regulated (cluster 6 on Figures 5.2.3-3 and 5.2.3-4) genes after EIN1 depletion. At 180 minutes post-EGF stimulation, the majority (60%) of genes were downregulated. Similarly to the EGF- regulated genes, EIN1 depletion also triggered change in the magnitude of expression at each time point, however, as the as the time progresses, the magnitude of differences increase (Figure 5.2.3-4). The EGF-dependant changes in the magnitude of expression in absence of EIN1 expression might indicate that EIN1 works as an insulator of EGF-regulated expression programme. When EIN1 is present, EGF-stimulation triggers changes in the core expression programme and concordantly, EIN1 depletion triggers expansion in that programme.
Figure 5.2.3-3. EGF-dependant expression of genes that are affected by EIN1 downregulation at 180 minutes after EGF stimulation.

Heatmap of the RNA-seq expression data for genes which were affected by EIN1 knockdown at 180 minutes after EGF stimulation. Gene were manually sorted into 2 classes: EGF-regulated at 180 minutes (clusters 1-4) and EGF-independent at 180 minutes (clusters 5 and 6). Within the classes, genes were cluster using hierarchical clustering. Expression is presented as per-row z-score of log2(FPKM).
**Figure 5.2.3-4. Normalized expression of genes affected by EIN1 downregulation at 180 minutes after EGF stimulation.**

Normalised line plot of the average expression of downregulated genes (A, C, E) or up-regulated genes (B, D, F) after EIN1 downregulation. For each gene that was affected by EIN1 downregulation at 180 minutes after EGF stimulation, the log-transformed expression values were normalized to the 0 minutes time point of non-targeting (control) guide RNA experiment. Subsequently, the mean expression for each cluster of genes was calculated. Error ribbons indicate 95% confidence intervals. Gene clusters (1) to (6) correspond to clusters from figure 5.2.3-3.
In order to investigate the effect of \textit{EIN1} depletion on cellular processes I analysed data using Ingenuity Pathways suite. As a benchmark for comparisons, I used the effect of EGF-stimulation on the cells. Similarly (and expectedly) to the nuclear RNA-seq, the Pathways analysis showed that the key process associated with the EGF-regulated genes are gene expression, cell growth, differentiation and cellular movement (summarised in Supplementary files 10 to 12). The investigation of genes that were EGF-regulated and affected by \textit{EIN1} depletion did not show strong functional enrichment; however, the analysis of the whole set of \textit{EIN1}-regulated genes resulted in significant enrichment of the biological processes and disease normally associated with EGF signalling (summarised in Supplementary files 13 to 16 and detailed in Figures 5.2.3-5 and 5.2.3-6). The key cellular processes associated with genes affected by \textit{EIN1} depletion were movement, survival and proliferation (Figure 5.2.3-5). From the disease point of view, \textit{EIN1} depletion triggered change in the expression of genes associated with cancer, inflammation and cardiovascular diseases (Figure 5.2.3-6). Importantly, the statistical power of functional associated increases as the EGF stimulated time course progresses.

The most significantly enriched cellular process associated with \textit{EIN1} depletion was movement and the effect was strongest 180 minutes after EGF stimulation. Data from our laboratory indicate that EGF in an important regulator of cellular movement in the MCF10A cells (Odrowaz and Sharrocks, 2012). Further investigation showed that out of 341 genes affected by \textit{EIN1} depletion at 180 minutes after EGF stimulation, the function of 87 (25.5\%) is associated with cellular movement (listed in Supplementary files 6) and 24 of them were EGF-regulated. Strikingly, the directional analysis of the effect of changes in the expression of these genes showed that statistically they trigger downregulation of cellular movement.
Figure 5.2.3-5. Cellular processes associated with the genes regulated by EIN1.
Bar plot of the Fisher’s exact p-value for the enrichment of the genes whose expression changes after EIN1 downregulation at each time point. Data were compiled using the Ingenuity pathways suite.
Figure 5.2.3-6. Diseases associated with the genes regulated by EIN1.

Bar plot of the Fisher’s exact p-value for the enrichment of the genes whose expression changes after EIN1 downregulation at each time point. Data were compiled using the Ingenuity pathways suite.
In order to understand the function of *EIN1*, I investigated the changes in the level of genes associated with cellular movement. Out of 87 *EIN1*-regulated genes that play role in cellular movement, 47 were linked with a decrease in the process and 29 with an increase (for the remaining, a directional information was not available, Supplementary File 6). The expression of the majority (38) of the genes which are associated with the decrease in cellular movement, decreased after *EIN1* depletion. These genes can be split into two groups – EGF-regulated and EGF-independent in presence of *EIN1* (Figure 5.2.3-7 groups 1 and 2). An example of EGF-independent gene might be *DMTN* (dematin actin binding protein). The expression of *DMTN* is not strongly affected by EGF signalling in the presence of *EIN1* (Figure 5.2.3-7). However, after *EIN1* depletion, *DMTN* expression decreases after EGF stimulation. *DMTN* is an actin binding protein and its knockout reduces cellular motility (Mohseni and Chishti, 2008).

Additionally, it appears that *EIN1* can influence the expression of some of the EGF-regulated genes, as its depletion changed their EGF-dependant expression profile (Figure 5.2.3-7 groups 2). For example, the expression of *MMP3* (matrix metalloproteinase 3) is induced after EGF stimulation in presence of *EIN1*, however in the absence of *EIN1*, the EGF-induced stimulation is limited (Figure 5.2.3-7). *MMP3* is a protein involved in the breakdown of extracellular matrix required for cellular movement (Murphy and Gavrilovic, 1999) and its function has been associated with cellular migration during wound healing (Kametaka et al., 2007).

Judging from these data, it appears that *EIN1* might function as an insulator of the EGF-regulated expression programme in MCF10A cell line. In the presence of *EIN1*, the expression programme triggered by EGF stimulus is limited to the core response involving IE, DE and SR genes (Figure 5.2.3-7 groups 2 and 3). However, when *EIN1* is depleted the EGF-regulated expression profile expands (Figure 5.2.3-7 groups 1) and expression of new classes of genes is affected.
Figure 5.2.3-7. The effect of EIN1 on the EGF-regulated cellular movement programme.

The diagrams present a possible mode of action for EIN1 in relation to cellular movement in the presence (A) and in absence (B) of EIN1. Genes presented were identified as regulators of cellular movement by the Ingenuity Pathways suite. For simplicity, all of the genes shown are associated with stimulatory role on the cell movement and decrease in their expression leads to decrease in the cellular movement. On both panels: (1) EIN1-modulated genes, downregulated after EIN1 knockdown, independent of EGF-signalling in presence of EIN1, (2) EIN1-modulated genes, downregulated after EIN1 knockdown, induced by EGF-signalling and (3) EIN1-independent genes, not affected by EIN1 downregulation, induced by EGF-signalling. Green background indicates increase in the function/expression after EGF stimulation. Line plots represent average expression of selected genes over time.
5.3. Chapter summary

The aim of this chapter was to identify the role of \textit{EIN1} in the regulation of the EGF-regulated programme of gene expression. First of all, the results showed that CRISPRi approach to gene regulation is very potent. It was able to downregulate the transcription of \textit{EIN1} and importantly, the downregulation was sustained in the presence of EGF signalling. It should be noted that the CRISPR-based technique are very recent development and the knowledge about potential off-target effects is not fully understood (Kuscu et al., 2014; Polstein et al., 2015; Wu et al., 2014). However, data presented in this chapter indicate that the depletion was specific to \textit{EIN1}. The combination of its potency with the efficiency of construct delivery using the lentiviral system makes it a first choice technique for the purpose of gene regulation at the transcription level.

Secondly, using RNA-seq, I built an EGF-regulated signalling programme. Similarly to previous studies (e.g. Amit et al., 2007), I observed a wave-like pattern of gene expression. In the previous studies, most of the attention was paid to the EGF-dependant upregulation of gene expression; however, the results clearly showed that EGF-dependant transcriptional cascade also leads to downregulation of hundreds of genes. I observed that the biggest changes in the transcript levels take place 180 minutes after stimulation. It is not unexpected, since each preceding wave of transcription includes a number of transcription factors what causes expansion in the regulation.

Finally, I investigated the effect of \textit{EIN1} on the EGF-regulated expression programme and the result was perplexing. \textit{EIN1} is one of the most upregulated genes after EGF stimulation and a relatively high level of its transcript is maintained up to four hours after stimulation. However, its downregulation had only minor (yet statistically significant) effect on the expression of EGF-regulated genes. Around 10\% of all of the genes associated with the EGF-dependant transcriptional cascade were affected by \textit{EIN1} depletion. Closer examination of these genes showed that \textit{EIN1} depletion changed the magnitude of the expression of these genes but it did not affect their EGF-dependant expression profile.

However, the \textit{EIN1} depletion also affected the expression of seemingly EGF-independent genes. In the presence of \textit{EIN1}, the genes affected by EGF stimulation
were strongly associated with well-established functions of cellular growth, differentiation and movement (Amit et al., 2007; O’Donnell et al., 2012; Odrowaz and Sharrocks, 2012). However, the depletion of EIN1 seems to expand this profile of regulation. A large number of genes (over 200) were added to the EGF-regulated programme and it appears that the new set of genes decreases the influence of EGF on some of the downstream processes including cellular movement.

It can be suggested that EIN1 functions as an insulator of the EGF stimulated expression programme. In MCF10A cell, EIN1 modulates the expression of genes downstream of EGF by two mechanisms:

- it prevents changes in the level of genes that are not required for the core EGF function
- it maintains the level of some of the genes that are included into the core EGF stimulated transcriptional programme.

The bioinformatics analysis of the function of EIN1-regulated genes showed that, among others, they are regulators of cellular movement. Furthermore, it appears that the combined effects of changes might lead to decrease in the movement. MCF10A cells are an excellent model to study this process and in the future, I will test whether EIN1 depletion indeed affects the movement of cells.
6. Discussion

The aim of the work presented in this thesis was to identify and characterise long non-coding RNAs regulated by EGF signalling. RNA-seq data showed that EGF stimulation upregulates 173 IncRNAs and downregulates 155 IncRNAs (Chapter 3). Most of the subsequent studies focused on the upregulated IncRNAs. Using information about the induction profiles after EGF stimulation and expression data from the cancer samples, I selected EIN1 for further studies. In Chapter 4, I described multiple properties of EIN1, including its coding potential (it is non-coding), cellular localisation (predominantly nuclear), MAPK-dependant expression (via the ERK1/2 pathway) expression in normal (expressed in embryonic and adult cardiovascular system) and cancer (upregulated in breast and lung cancer) samples. Finally, in chapter 5, I investigated the cellular phenotype of MCF10A cells after EIN1 depletion. It appears that EIN1 knockdown triggers changes in RNA level of multiple genes associated with cellular movement and that EIN1 functions as an insulator of the EGF-controlled expression programme.

Measurement of IE gene expression using nuclear RNA-seq

In chapter 3 of the thesis, I investigated the expression programme downstream of EGF signalling. During the selection of RNA-seq protocol, I paid particular attention to its ability to detect novel transcripts. Eventually, I decided to use polyA enriched RNA-seq on the nuclear fraction. The selection of the nuclear fraction was based on previous studies which showed that IncRNAs tend to localise in the nucleus (Djebali et al., 2012) and since I was investigating newly transcribed genes, I expected the enrichment to be magnified. I chose polyA enrichment protocol for the preparation of the sequencing libraries. The choice of polyA enrichment over rRNA depletion methods (e.g. Ribozero) was a compromise between technique ability to detect full length transcripts and accurate quantification of genes. rRNA depletion methods tend to detect all of the transcripts including non-polyadenylated genes and pre-mRNAs (Zhao et al., 2014). However, at least in the case of actively transcribed genes, rRNA depletion can cause 5’ bias during quantification of the transcripts and due to inclusion of unspliced introns; it is difficult to assembly mature transcripts. On the other hand, polyA enrichment protocols allow accurate quantification of polyA-tailed transcripts and help assemble mature transcripts (Tariq et al., 2011), but they might...
cause 3’ bias (not observed, Figure 3.1.1-1A) and miss non-polyadenylated transcripts (Zhao et al., 2014). Interestingly, in the nuclear RNA-seq, I was able to detect some of the IncRNAs that have been previously described as non-polyadenylated. For example, MALAT1 expression was around 80 FPKM and NEAT1 — around 200 FPKM, making them some of the most expressed genes. Both of these IncRNAs are mainly non-polyadenylated (Yang et al., 2011); however, they contain short internal polyA sequences (Wilusz and Spector, 2010). This indicates that polyA-enriched RNA-seq can detect some classes of IncRNAs (containing internal polyA stretches); however, it is clear that I did not detect some classes of IncRNAs (such as eRNAs (Kim et al., 2010) and circRNAs (Memczak et al., 2013)) which are normally not polyadenylated.

Expression profile of IE genes

The nuclear RNA-seq experiment identified 144 protein-coding genes that were upregulated after EGF stimulation. Subsequent investigation of their expression showed that in many cancer biopsies, on average, their expression decreases or does not change. (Figures 3.2.1–1 to 3.2.3–3). These results contradict some of the expectations about the function of IE genes in cancer. It is often assumed that IE genes are upregulated in cancer (Healy et al., 2012); however, most of the knowledge about the IE genes is extrapolated from studies of FOS – a well-characterised proto-oncogene (O’Donnell et al., 2012). To the best of my knowledge, a systematic investigation of IE gene expression in cancer is not available. However, an investigation of the expression of DE genes in cancer showed that they tend to be downregulated in several epithelial tumours (Amit et al., 2007). Since many of the DE early genes are suppressors of MAPK pathways (Avraham and Yarden, 2011; Blüthgen et al., 2009), it is likely that they function as cancer suppressors. Since my data show that IE genes are downregulated in at least some of the cancers, it can also be suggested that many IE genes also function as cancer suppressors. Further studies involving much higher cohorts of patients and possible inclusion of data from patients with known MAPK-related mutations such as V600E BRAF mutation and HER2 overexpression might provide more information about the behaviour of IE genes in cancer.
The expression profile of long non-coding RNAs

The main aim of the project was the identification of IncRNAs that are regulated by EGF stimulation. Due to similarity between upregulated protein-coding genes and upregulated IncRNAs, the majority of this work focused on the EGF upregulated IncRNAs. However, in the original nuclear RNA-screen, I also identified 155 downregulated IncRNAs. RT-qPCR showed that the down-regulation of IncRNAs is sustained for around three to four hours (Figure 3.2.2-1). This indicates that they might play an important function in the regulation of EGF response, possible similar to the role of ID-miRs (see section 1.3.4) (Avraham and Yarden, 2012; Avraham et al., 2010). Moreover, I observed that the loci of many of the downregulated IncRNAs are upstream antisense RNAs (uaRNAs) in relation to the protein-coding genes (Figure 3.2.1–3). Over the recent years, uaRNAs have been identified as short and unstable transcripts expressed by RNApolII from bidirectional promoters (Preker et al., 2011; Scruggs et al., 2015); however, a function has not been associated with them. Normally, uaRNAs are rapidly turned-over by exosome complexes (RNA degradation system) and can be detected only after exosome knockdown (Preker et al., 2008). Our data suggests, that at least some of the uaRNAs are stable enough to be detected by polyA enriched RNA-seq in EGF-starved cells and that their level decreases after EGF stimulation. Furthermore, the observed decrease in the level of RNApolIII binding to these bi-direction promoters might indicate co-transcriptional regulation. However, since I did not observe EGF-mediated changes in the transcript level of the upstream antisense protein-coding genes, the changes in uaRNAs levels might be explained by increased exosome activity. To the best of my knowledge, a link between exosome-dependant RNA degradation and EGF signalling has not been shown and it would be interesting to investigate this process in more detail.

Investigation of the expression profile of EGF-upregulated intergenic IncRNAs (lincRNAs) shows that they exhibit properties remarkably similar to the protein-coding IE genes. Their expression profile is similar to protein-coding and follows wave-like patterns. Their genomic loci are enriched near ELK1 binding regions. Most of the EGF-upregulated lincRNAs seem to be downregulated in cancer; however, the pattern is not as clear as in the case of protein-coding genes (Figure 3.2.3-3). The remaining question is whether they play a significant role in cellular processes. Our study is the first that comprehensively surveys the transcription of
IncRNAs after EGF stimulation. Recently, the transcriptional dynamics of the EGF-induced genes have been studied using CAGE data in MCF7 cell line (Aitken et al., 2015). In this study, the authors identified lincRNAs that were precursors of ID-miRs and investigated their expression kinetics. The results showed that the expression of host genes for previously described ID-miRs (mir-155, mir-320a) precedes the expression of mature miRNAs. However, due to low sensitivity of CAGE data, the authors were not able to study other genes.

**EIN1 – EGF-induced long non-coding RNA**

The ultimate aim of the project was identification of lincRNAs whose expression is regulated by EGF and subsequent characterisation of the transcripts. I chose *EIN1* for further studies because:

- It is a long intergenic transcript with well-defined ends,
- Its putative promoter region contains elements (TATA-like box and Inr-like sequence) normally found in the promoter of genes normally expressed by RNApolII,
- An ELK1 binding region is located in the putative upstream regulatory region indicating direct regulation by EGF→ERK1/2→ELK1 pathway,
- It was the most induced lincRNA after EGF stimulation,
- In contrast to many EGF-regulated IE protein-coding genes and EGF-induced lincRNAs, its level was higher in cancer samples than in normal tissues.

The first step in the characterisation of *EIN1* was investigation of its coding potential. Both, bioinformatics and biochemical data indicated that *EIN1* is not actively translated. Nevertheless, since one cannot prove negative effects, further evidence can be provided to support the non-coding status of *EIN1*. The investigation of non-coding status of lncRNAs is important because recent reports suggested that some of the lncRNAs might harbour small, functional peptides (Anderson et al., 2015; Bazzini et al., 2014). In addition to the techniques employed in chapter 4, the non-coding status of lncRNAs can be investigated using *in vitro* and *in vivo* translation (Ji et al., 2003; Prensner et al., 2011). However, in order to perform these assays, cloned full length, mature transcript is required. Additionally, the non-coding status of lncRNAs can be inferred from their cellular localisation. In addition to
biochemical fractionations, one can perform RNA-FISH (Marín-Béjar et al., 2013; Sauvageau et al., 2013) which allows for identification of each copy of the transcripts. Furthermore, the RNA-FISH data can be used to infer the copy number and relative abundance of each of the isoforms of the gene.

Another important feature of EIN1 is the mode of its expression. To the best of my knowledge, EIN1 is the first lncRNAs whose transcription has been directly linked with ERK1/2 signalling. Data in section 4.3.2 show that the EGF induction signal is directly transmitted to nucleus via MEK1/2→ERK1/2 pathway and that EIN1 transcription is independent of protein synthesis. EGF, in addition to ERK1/2 pathway, can activate PI3/AKT signalling pathway (Amit et al., 2007; Turjanski et al., 2007; Yang et al., 2013b) and it was important to identify the signalling programme associated with the regulation of EIN1 transcription. The protein synthesis-independent transcription showed that EIN1 is indeed an IE gene (Cochran et al., 1984). Interestingly, the transcription of IE genes, such as FOS, tends to be hyper-activated and extended beyond normal time-scales in the absence of protein synthesis (Tullai et al., 2007); however, the transcription of EIN1 ceased around two hours post-EGF stimulation in the presence of cycloheximide (Figure 4.3.2-2). This unusual transcription pattern suggests that the transcription activation module present at the EIN1 promoter is not stable and it requires further input from new proteins that are synthesised as IE genes to promote prolonged EIN1 transcription.

**CRISPRi-driven gene depletion is effective against lincRNAs**

In the final results chapter, I investigated the effect of EIN1 knockdown on the EGF-regulated expression programme. Due to the reasons described in section 5.2.1, I decided to use CRISPRi-driven downregulation of EIN1 transcription (Gilbert et al., 2013). The system proved to be an effective suppressor of transcription. Using RNA-seq, the EIN1 transcript was not detected in any of the EGF-stimulated samples using RNA-seq (Figure 5.2.1-2). Over recent years, the number of CRISPR-derived techniques has grown dramatically. Currently, in addition to standard RNA-directed DNA cleavage and described here downregulation of transcription, CRISPR can be used for up-regulation of transcription (Konermann et al., 2014), for targeting lncRNAs to a specific locus – CRISPR-Display (Shechner et al., 2015) and it can be targeted to RNA instead of DNA (O’Connell et al., 2014). CRISPR-driven techniques are a very recent development and studies are ongoing investigating the
off-target effects of CRISRP (Kuscu et al., 2014; Polstein et al., 2015; Wu et al., 2014). The results show that RNA-guided CRISPR does have some off-targets. However, at least some of the off-targets can be predicted computationally, as they show homology with the designed target sequence (Kuscu et al., 2014). In order to overcome that limitation, in the experiments presented in chapter 5, I used 3 individual guide RNAs. It is unlikely that three independent guide RNAs would target exactly the same gene. Indeed, the RNA-seq data did not seem to contain genes, other than EIN1, that were highly affected in the experiments. Clear functional correlations between the genes that are modestly affected by EIN1 downregulation further supports that the majority of the changes come from the effect of EIN1 depletion on the cellular processes rather than off-target action of the CRISPRi system.

**EIN1 functions as an insulator of the EGF-induced transcription programme**

The depletion of EIN1 triggered changes in the level of multiple protein-coding genes. Data presented in chapter 5 seem to indicate that EIN1 works as an insulator of EGF signalling. In the presence of EIN1, the EGF-induced expression programme is limited to core genes that are required for cellular movement, growth and differentiation. The correct level of expression of some of the core EGF-regulated genes (e.g. HLX, MMP3) depends on the action of EIN1 what in turn results in an appropriate cellular response (Internal EIN1-dependant signalling, Figure 6.1.1.-1A). In the absence of EIN1, the activity of these genes changes in the manner that reduced the final cellular process such as cell movement (Internal EIN1-dependant signalling, Figure 6.1.1.-1B).

Simultaneously, EIN1 seems to influence the expression of genes that normally are outside of the core EGF-regulated expression programme. When the expression of EIN1 is triggered in the cells, it insulates the EIN1-modulated genes (EMGs, e.g. DMTN, ACKR3) from the EGF-dependant signalling (External EIN1-dependant signalling, Figure 6.1.1.-1A). However, the removal of EIN1 leads to EGF-dependant changes in the expression of EMGs (External EIN1-dependant signalling, Figure 6.1.1.-1B). Since it has been observed that the population of EGF-induced IE genes differs between cells lines (Amit et al., 2007), it can be expected that a cellular
element plays role of expression modulator and an lincRNAs, such as *EIN1*, might play a critical role in the selection of genes that are affected by EGF stimulation. The expression of the core IE genes, such as *FOS, EGR1, EGR3, JUN, FOSB* (detected in multiple EGF induction studies (Amit et al., 2007; Odrowaz and Sharrocks, 2012; Tullai et al., 2007)) is not modulated due to their central role in the propagation of the signal. However, the expression of the additional genes, which are responsible for fine-tuning of the cellular response downstream of EGF-signalling, is modulated and they can be isolated from the EGF-signalling by lincRNAs, such as *EIN1*.

Further work is required to dissect the cellular function of *EIN1*. RNA-seq only allows for measurement of the steady-state levels of transcripts. The changes in the level of over 300 genes can be explained by multiple mechanisms. Judging for nuclear localisation of *EIN1*, it can be speculated that, similarly to *lincRNA-p21* and *HOTAIR*, *EIN1* might function as a regulator of transcription (Huarte et al., 2010; Rinn et al., 2007). However, it cannot be ruled out that *EIN1* function through another mechanism, potentially involving target RNA degradation or change in their localisation. Some of the potential experiments to identify that role are described in the following section.

Finally, additional studies of the phenotype changes can be performed. MCF10A cells have a well-defined movement and 3D differentiation phenotypes and the effect of *EIN1* depletions and overexpression can be tested. Preliminary results from the downregulation experiments showed that *EIN1* knockdown reduces the mobility of MCF10A cells in the wound healing assay; however, the effect has to be studied in more details in order to draw final conclusions.
Figure 6.1.1-1. Proposed mechanism of *EIN1* action

A) In the presence of *EIN1*, the expression of EGF-regulated programme includes only core immediate-early (IEG) and delayed-early genes (DEG) and leads to changes in the secondary response genes (SRG). *EIN1* can influence the function of some of the SRG via unknown mechanism. Furthermore, the EGF-dependant regulation of *EIN1*-modulated genes (EMG) is limited in the presence of *EIN1* by an unknown mechanism. B) Depletion of *EIN1* removes its function and the EGF-dependant expression programme includes additional genes what ultimately leads to changes in the final cellular processes. EGF – epidermal growth factor, IEG – immediate-early gene, DEG – delayed-early genes, SRG – secondary response genes, EMG – *EIN1* modulated genes.
Future directions

In order to fully understand the molecular mechanism of \textit{EIN1} action, further biochemical and phenotypical studies are required. Key biochemical studies are:

- **Identification of \textit{EIN1} partners.** A key step to understand the mechanism of \textit{EIN1} action is identification of its partners. Since I lack prior knowledge about protein partners (although it can be suggested that some of the IE gene products might be \textit{EIN1} partners) of \textit{EIN1}, I cannot use RIP techniques; however, I can employ recently developed ChIRP-MS (Chu et al., 2015) technique. This technique takes advantage of improvements in mass spectroscopy field and allows for identification of protein partners of a given lncRNA.

- **Additionally,** the ChIRP technique can be used to study IncRNAs – chromatin interaction (Chu et al., 2012). One of the mechanisms for influencing RNA levels by IncRNAs is their interaction with chromatin (Huaré et al., 2010) and it is possible that \textit{EIN1} modulates the EGF-induced programme by its interaction with chromatin.

- **RNA-seq data suggests that \textit{EIN1} is expressed in multiple isoforms.** Hence, a key to understand its function is cloning of the full-length of mature transcript. This would allow for exogenous delivery of mature transcripts to cellular systems and use in systems such as CRISPR-Display (Shechner et al., 2015).

- **On the cellular level,** one can investigate the localisation and, potentially, differential expression of IncRNAs using RNA-FISH.

- **Finally,** the data suggests that \textit{EIN1} is directly regulated by the ERK1/2 pathway. However, it remains elusive what transcription factors play a role in \textit{EIN1} expression and identification of these TFs would provide further evidence for the regulatory loop associated with \textit{EIN1} function.

Secondly, the function of \textit{EIN1} can be studied from the phenotypical point of view:

- **The MCF10A cell line** is an excellent model of cellular movement and invasion (Odrowaz and Sharrocks, 2012; Wang et al., 2012). Previous studies from our laboratory showed that migration is dependent on the expression of IE genes via ELK1 (Odrowaz and Sharrocks, 2012).
• MCF10A cells can form 3D acini when grown in Matrigel (Debnath et al., 2002, 2003). Feature of the cells can be used for investigation of EIN1 role in cellular differentiation
• Finally, EIN1 role can be investigated in the organism. The MCF10A cell line do not form tumours when transplanted into nude mice (Soule et al., 1990).
• All of the above phenotypes can be investigated after downregulation of EIN1. Additionally, all of these processes can be studies using system where EIN1 transcription is upregulated. CRISPRi-driven system where dCas9 is fused with four viral protein 16 domains (4xVP16 or VP64) has been show to upregulated expression of multiple genes (Konermann et al., 2014) and it can be used to drive expression of endogenous EIN1.

Final Remarks

The field of lncRNAs is growing rapidly and novel, functional lincRNAs are discovered continuously. The results presented in this thesis show the transcription of EIN1 is tightly controlled by the EGF-induced transcription programme and potentially plays role in the key cellular processes such as movement and growth. EIN1 is an important example of an lncRNA with a well-described regulation of expression and it is the first lncRNA whose transcription is regulated by EGF-induced ERK1/2-pathway. Furthermore, our data indicate that EIN1 functions within the EGF programme. It works as an insulator of the programme. Taken together, EIN1 supports notion that lncRNAs are important regulators of cellular processes.
7. References


Bollag, G., Tsai, J., Zhang, J., Zhang, C., Ibrahim, P., Nolop, K., and Hirth, P.


International Human Genome Sequencing Consortium (2004). Finishing the


188 | P a g e


Sharrocks, A.D. (2013). Professor Andrew Sharrocks (Faculty of Life Sciences - The University of Manchester).


Simon, M.D., Wang, C.I., Kharchenko, P. V., West, J. a., Chapman, B. a.,


Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H.,


Wei, G., Wei, L., Zhu, J., Zang, C., Hu-Li, J., Yao, Z., Cui, K., Kanno, Y., Roh, T.-


